

2025 年大语言模型 进展报告

哈尔滨工业大学
自然语言处理研究所 (HIT-NLP)
<http://nlp.hit.edu.cn>

2026 年 1 月



序言

自 2022 年底 ChatGPT 发布以来,大语言模型(Large Language Models, LLMs) 迅速成为人工智能领域最具影响力的技术方向之一,深刻推动了自然语言处理乃至整个人工智能范式的变革。作为自然语言处理技术发展的重要里程碑,大语言模型在多种下游任务中展现出显著的性能优势,催生了智能问答、内容生成、代码编写等一系列新型应用形态,并在教育、医疗、金融等多个行业引发了广泛关注与实践探索。

进入 2025 年,从年初 DeepSeek-R1 的发布引发广泛关注,到年末 Google Gemini 3 Pro 在多项基准与应用场景中实现性能跃升并取得领先地位,本年度大语言模型相关技术与应用创新非但未显放缓,反而呈现出持续加速的发展趋势:模型体系结构与训练范式不断演进并趋于多样化,多模态理解与复杂推理能力显著增强,相关应用场景亦在科研、工业及社会治理等众多领域持续拓展与深化。

为系统梳理 2025 年大语言模型领域的发展现状、关键进展与未来趋势,哈尔滨工业大学自然语言处理研究所组织多位教师与研究生共同编写了本《2025 年大语言模型（LLMs）进展报告》。本报告围绕模型架构设计、预训练与后训练方法、模型部署技术、典型应用场景以及安全与伦理等多个方面,对年度内的重要研究成果与实践进展进行了系统总结与分析,旨在为相关领域的研究人员、技术开发者及决策者提供有价值的参考。

与侧重技术演进脉络的传统综述不同,本报告更加关注 2025 年度内具有代表性的最新进展与趋势判断,力求呈现当前大语言模型研究与应用的前沿图景。未来,我们计划持续按年度发布大语言模型进展报告,逐步形成一部记录该领域技术演进与思想变迁的“编年体”参考文献。

主要编撰人员

第一章由陆鑫、王一轩、王洋编写；

第二章由高翠芸、张伟男、丁效、刘元兴、蔡碧波、刘铭、朱庆福、孙承杰、刘铭、何涛、吉佳浩、王乾宇、王子翔、韩为、曾屹荣、韩喆、秦占悦、罗先镇、张致铭、周士祺、郭传哲、代居益、陈逸飞、杨双嘉编写；

第三章由刘铭、朱聪慧、王一轩、季世宇、王泽鑫、时景琦、黄毅彬编写；

第四章由冯骁骋、朱庆福、徐永东、刘秉权、隋典伯、赵森栋、王昊淳、王镜博、曾钰倬、刘议骏、郑欢洋、黄伟韬编写；

第五章由冯骁骋、姜文浩、高翠芸、朱庆福、吴湘平、赵妍妍、赵森栋、王昊淳、刘铭、姜京池、范会明、刘翔宇、张文斌、孙怡馨、姜毅、李健博、吕欣达、袁嘉伟、关胜圆编写；

第六章由姜京池、吴湘平、杨沐昀、郭昱辉、韩运鹏、吕欣达、温天瑞、黄辉编写；

第七章由刘远超、赵妍妍、吴湘平、赵孟晨、徐永东、杨沐昀、卜坤、曾钰倬、吴迪编写；

第八章由陈清财、张伟男、刘元兴、蒋硕然、刘凯锋编写。

报告整体由车万翔统稿。

目录

第一章 模型架构的演进	10
1.1 全注意力序列建模	10
1.1.1 研究背景	10
1.1.2 研究进展	11
1.1.3 未来展望	13
1.2 稀疏序列建模模型	13
1.2.1 研究背景	13
1.2.2 研究进展	14
1.2.3 未来展望	15
1.3 混合专家模型	15
1.3.1 研究背景	15
1.3.2 研究进展	16
1.3.3 未来展望	18
1.4 状态化序列建模模型	18
1.4.1 研究背景	18
1.4.2 研究进展	19
1.4.3 未来展望	22
1.5 多模态语言模型架构	22
1.5.1 研究背景	22
1.5.2 研究进展	23
1.5.3 未来展望	24
1.6 新兴方向	24
1.6.1 主要背景	24
1.6.2 扩散语言模型	25
1.6.3 动态计算	25

1.6.4	嵌套学习	26
1.6.5	未来展望	26
1.7	本章小结	27
第二章	大语言模型训练	28
2.1	后训练技术更新	28
2.1.1	SFT 最新进展	29
2.1.2	强化学习算法进展	31
2.2	数据获取与数据治理	39
2.2.1	开源数据集构建	39
2.2.2	数据处理技术	43
2.2.3	多模态数据集构建	47
2.3	模型能力提升	49
2.3.1	长上下文	50
2.3.2	推理	54
2.3.3	数学/代码	61
2.3.4	工具调用	64
2.3.5	Agentic RL	68
2.4	开源训练框架	71
2.4.1	VeRL (Volcano Engine)	73
2.4.2	ROLL (Alibaba)	74
2.4.3	PRIME-RL (Prime Intellect)	74
2.4.4	Slime (Zhipu AI)	75
2.4.5	RAGEN	76
2.4.6	OpenRLHF	76
2.4.7	未来展望	77
2.5	本章小结	78
第三章	大语言模型部署	79
3.1	模型压缩	79
3.1.1	量化	80
3.1.2	剪枝	82
3.1.3	蒸馏	85
3.2	模型加速	88

2025 年大语言模型 (LLMs) 进展报告

3.2.1	投机解码	88
3.2.2	KV Cache	91
3.3	开源部署框架	94
3.3.1	vLLM	95
3.3.2	SGLang	96
3.3.3	TensorRT-LLM	97
3.3.4	LMDeploy	98
3.3.5	llama.cpp	99
3.3.6	Ollama	99
3.3.7	框架选型对比与适用场景分析	100
3.4	本章小结	102
第四章	智能体演进	105
4.1	自主任务规划	105
4.1.1	研究背景	105
4.1.2	研究进展	106
4.1.3	未来展望	108
4.2	工具链整合	109
4.2.1	研究背景	109
4.2.2	研究进展	110
4.2.3	总结与展望	116
4.3	检索增强生成 (RAG)	117
4.3.1	研究背景	117
4.3.2	RAG 的全链路优化范式	117
4.3.3	自适应与自主 RAG	120
4.3.4	多智能体 RAG (Multi-Agent RAG)	121
4.3.5	多模态 RAG	122
4.3.6	总结与展望	122
4.4	长期记忆	123
4.4.1	研究背景	123
4.4.2	研究进展	124
4.4.3	未来展望	126
4.5	自我反思自我修正智能体	127
4.5.1	研究背景	127

4.5.2	模型原生反思机制	128
4.5.3	自适应迭代控制机制	129
4.5.4	检索增强自反思	132
4.5.5	结论与展望	132
4.6	自我进化	133
4.6.1	研究背景	133
4.6.2	研究进展	133
4.6.3	未来展望	138
4.7	GUI Agent	139
4.7.1	GUI Agent 感知能力	139
4.7.2	GUI Agent 规划能力	141
4.7.3	GUI Agent 执行	141
4.7.4	面向 GUI 的专用模型	142
4.7.5	GUI 智能体数据集	143
4.7.6	总结与展望	144
4.8	多智能体协作框架	145
4.8.1	研究背景	145
4.8.2	研究进展	145
4.8.3	未来展望	150
4.9	本章小结	150
第五章	大语言模型的应用进展	151
5.1	任务应用	151
5.1.1	大模型与脑科学	151
5.1.2	编程助手	157
5.1.3	写作助手	160
5.1.4	设计助手	167
5.1.5	社会模拟	173
5.1.6	心理咨询	177
5.1.7	深度调研: Deep Research	181
5.1.8	AI for Research	186
5.2	行业应用	190
5.2.1	教育行业	190
5.2.2	医疗行业	196

5.2.3	金融	205
5.2.4	法律行业	211
5.2.5	农业产业	217
5.3	本章小结	223
第六章	评测基准和模型进展	224
6.1	新评测基准	224
6.1.1	引言	224
6.1.2	多轮对话评测基准	225
6.1.3	工具使用评测基准	229
6.1.4	智能体评测基准	233
6.1.5	多模态评测基准	238
6.2	模型生态进展	242
6.2.1	新闭源模型	243
6.2.2	新开源模型	249
6.2.3	国产开源模型的崛起	256
6.3	综合能力排行榜汇总	259
6.3.1	语言能力（Language）评测调研	259
6.3.2	图像与视频（Vision & Video）多模态评测调研	261
6.3.3	语音能力（Speech）评测调研	263
6.3.4	编程能力（Programming）评测调研	264
6.3.5	数学能力（Mathematics）评测调研	267
6.3.6	推理能力（Reasoning）评测调研	270
6.3.7	智能体能力（Agents）评测调研	273
6.4	本章小结	275
第七章	大语言模型安全与伦理	277
7.1	安全对齐与治理	277
7.1.1	研究背景	277
7.1.2	研究进展	278
7.1.3	未来展望	280
7.2	生成风险控制	282
7.2.1	训练阶段优化	282
7.2.2	推理阶段增强	283

7.3	内容真实性与可追溯性	287
7.3.1	水印	287
7.3.2	可验证生成	289
7.3.3	溯源体系	293
7.4	攻击与防御	299
7.4.1	背景	299
7.4.2	提示词安全	300
7.4.3	数据安全	303
7.4.4	隐私保护训练方法	305
7.5	宪法人工智能	310
7.6	本章小结	314
第八章	未来展望	315
8.1	技术趋势预测	315
8.1.1	模型能力从注重规模到注重“智能密度”	315
8.1.2	基础模型的技术架构与训练范式的演进	317
8.1.3	应用范式：从被动工具到主动智能体	319
8.1.4	云边协同将大模型能力与移动互联网时代特征充分融合	320
8.1.5	从虚拟到现实：世界模型与具身智能	321
8.2	挑战与机遇	322
8.2.1	算力资源不均	322
8.2.2	安全与伦理	324
8.2.3	跨学科融合	327
8.3	本章小结	327

第一章 模型架构的演进

2025 年，大语言模型继续飞速发展，出现了很多对基础架构进行革新的关键工作，这些工作聚焦大语言模型架构对模型能力增强和效率提升方面的积极作用，展现了大语言模型除依赖规模扩张以外通过架构革新进行提升的更多可能。本章系统梳理了大语言模型基础架构在全注意力序列建模、稀疏序列建模模型、混合专家模型、状态化序列建模模型、多模态语言模型架构等多方面的关键进展，介绍了一些有重要潜力的新兴方向，旨在为读者展示清晰、全面的大语言模型架构方面的进展图景，揭示大语言模型架构演进的内在逻辑与潜在趋势，为后续研究提供一定的参考。

1.1 全注意力序列建模

1.1.1 研究背景

自 Transformer 架构^[1]提出以来，自注意力机制（Self-Attention）一直是大语言模型实现长序列上下文建模的核心组件。其通过计算序列中词元（Token）两两之间的相关性，捕捉了长距离依赖关系。然而，随着模型规模的指数级增长和应用场景向无限长上下文拓展，标准注意力机制 $O(N^2)$ 的计算复杂度和巨大的 KV 缓存（KV Cache）显存占用成为了主要瓶颈。在 2025 年，学术界与工业界对全注意力模块的改进不再局限于单一维度的修补，而是追求“效率”与“表达能力”的帕累托最优。本小节将整合**注意力分组**、**内部结构**以及**位置编码**三个关键维度，从这一时期面临的共性瓶颈出发，详细阐述全注意力序列建模在近期的关键演进与未来趋势。

在大模型向超长上下文演进的过程中，传统的注意力机制在显存效率、噪声控制和位置外推三个方面同时遭遇了瓶颈。首先，在**显存效率**方面，早期的 MHA（Multi-Head Attention）因每个头独立的 KV 导致显存开销巨大。尽管随后的 GQA^[2]通过分组共享 KV 在性能与显存间取得了平衡，以

及 DeepSeek 的 MLA^[3]通过低秩投影将显存压缩推向极致，但在面对万亿级参数模型时，如何在压缩 KV 的同时保留足够的信息容量仍是核心难题。其次，在**噪声控制**方面，Softmax 归一化的固有特性导致了“注意力陷阱”（Attention Sink）^[4]现象，即大量权重被无效地分配给序列起始的少数 Token。这种权重的非理性分配不仅浪费了计算资源，更导致模型在处理长文本时信噪比下降，难以捕捉深层语义。最后，在**位置编码**方面，主流的旋转位置编码（RoPE）^[5]虽然解决了相对位置感知问题，但在超长距离外推时面临“Lost-in-the-Middle”^[6]困境。高频与低频分量对长短期信息的处理冲突，限制了模型在千万级 Token 窗口下的检索能力。

1.1.2 研究进展

2025 年全注意力序列建模的进展，主要集中于注意力分组机制优化、注意力内部结构优化和注意力位置编码改进等多个方面，本部分将对这些进展依次进行介绍。

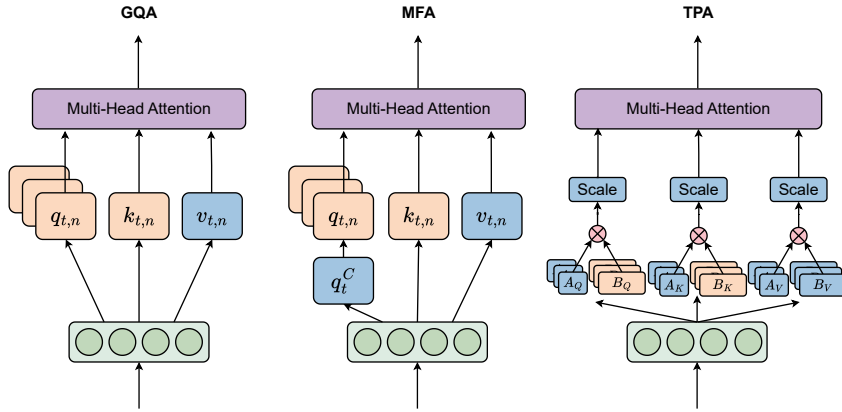


图 1.1: 相比 GQA 注意力分组结构，MFA 和 TPA 分组机制的结构示意图。其中橙色部分代表融入的 RoPE 位置编码信息。

注意力分组机制优化

2025 年期间，研究者们从 MLA 的低秩角度出发，进一步优化注意力分组机制实现性能和效率的权衡。为了在降低 KV Cache 占用的同时不过分损伤注意力模块的表征能力，MFA^[7]（Multi-matrix Factorization Attention）将低秩投影的思想引入到了 Query 矩阵的构建。如图 1.1.2所示，通过将 Query

矩阵分解成多头共享矩阵和专属矩阵，MFA 实现了对 MQA 在注意力头数量上的高效拓展，在压缩 KV Cache 存储空间的同时提高了注意力模块的表征能力。除此之外，目前基于低秩投影的分组方法还面临与当前大语言模型主流的旋转位置编码（RoPE）不兼容的问题。TPA^[8]（Tensor Product Attention）采用张量积的形式对注意力运算操作进行分解，实现了一种与 RoPE 兼容的高效注意力形式。具体来说，TPA 可以理解为一种 rank=1 情况下的特殊低秩分解方法，在显著降低 KV Cache 开销的同时，也降低了注意力模块的参数量。

注意力内部结构优化

针对注意力噪声的问题，2025 年研究者们研究方向主要聚集在两点，对 softmax 机制本身的改造，以及对模型结构上的调整。Softpick^[9] 提出了 Rectified Softmax 来代替传统的 Softmax，旨在解决注意力分数的过度分配问题。通过结合 ReLU 激活函数和 Softmax-1，提出的方法允许注意力分数总和必须为 1，使得原本的注意力机制产生稀疏的输出，仅对真正重要的 Token 进行有效关注。除针对 Softmax 函数本身的数值优化外，另一类研究致力于通过引入额外的模型结构来缓解注意力噪声。GPT-OSS^[10] 创新性地引入了可训练的偏置 Token 参与注意力运算。这些 Token 充当了 Sink Token 的角色，专门用于吸纳 Softmax 归一化过程中产生的冗余注意力得分，从而保护语义 Token 的权重不被稀释。不同于被动吸收噪声，Gated Attention^[11] 采取了主动过滤策略。通过引入遗忘门结构，模型能够从底层机制上学习截断注意力噪声的累积。实验表明，这种设计极大地改善了深层网络的训练稳定性，并提升了长序列任务的表现。

注意力位置编码改进

现有的 RoPE 机制面临显著的“Lost-in-the-Middle”现象^[6]，这在长文本处理中构成了严重瓶颈。针对这一问题，HoPe^[12] 观察到：虽然高频分量对于维持局部语法能力至关重要，但低频分量在长距离建模中往往会引入额外噪声。因此，HoPe 提出在保留高频信号的同时，策略性地抑制或替换导致长程噪声的特定低频分量。研究表明，打破 RoPE 强制的长程衰减约束，反而能显著提升模型在“大海捞针”（Needle-in-a-Haystack）任务中的关键信息检索能力。

除了对 RoPE 本身的改良，2025 年以来的研究趋势开始重新审视显式

位置编码的必要性^[13-14]。理论上，对于因果 Transformer，因果掩码本身已隐含了必要的位置信息，模型可通过层级累积推断出绝对位置。基于此，新一代模型开始探索“融合”架构：Llama 4^[15]引入了 iRoPE，采用 RoPE 与无位置编码（NoPE）逐层交替相结合的策略。这一混合设计被视为 Llama4 在支持千万级 Token 上下文窗口的同时，仍能保持严密推理逻辑的核心因素。此外，DeepSeek V3^[3]坚持采用的 MLA 架构，其层内设计也暗含了 RoPE 与 NoPE 特征的隐式融合，这被认为是保障其架构有效性的关键之一。

1.1.3 未来展望

综上所述，全注意力序列建模正在经历从“静态规则”向“动态适应”的范式转变。展望未来，在**机制设计**上，模型将具备内生的动态稀疏性。未来的注意力机制将不再受限于静态的物理头划分或固定的 Softmax 预算，而是根据输入内容的语义密度，自动调节计算预算并动态映射隐空间关注点，实现更接近人类认知的按需计算。在**架构演进**上，显式与隐式的融合将更加彻底。RoPE 与 NoPE 的有机结合不仅将巩固长文本处理能力，其旋转机制更将向二维甚至三维空间拓展，以支撑下一代原生全模态（Visual/Video-Native）模型的时空一致性建模。最后，**软硬协同**将成为性能提升的关键。未来的架构设计将针对预填充（Prefilling）阶段的算力瓶颈与解码（Decoding）阶段的访存瓶颈进行差异化建模，通过算子层面的深度优化，进一步挖掘大语言模型在全生命周期内的运行效率。

1.2 稀疏序列建模模型

1.2.1 研究背景

尽管全注意力机制在捕捉长距离依赖方面表现卓越，但在处理超长上下文时，其 $O(N^2)$ 的计算复杂度与 $O(N)$ 的显存占用仍是不可忽视的桎梏。此外，随着序列长度的增加，Softmax 算子产生的概率分布往往趋于平坦，导致模型难以从海量噪声中精准聚焦关键信息。面对这些挑战，稀疏注意力通过策略性地忽略非必要信息，成为打破长文本性能瓶颈的关键路径。2025 年，稀疏序列建模的研究重心已从早期的启发式静态掩码，全面转向了更具自适应性的动态稀疏机制，并在软硬件协同设计上取得了突破性进展。

全注意力机制假设序列中的每一个 Token 都与其他所有 Token 存在语义关联，然而大量研究表明^[16]，自然语言具有显著的局部性与稀疏性特征。

在长序列生成过程中，真正对当前预测起决定性作用的往往仅是上下文中的少数关键片段。全量计算不仅引入了大量冗余计算，更导致了严重的内存访问瓶颈。因此，如何在保持模型长程建模能力的前提下，精准识别并计算这部分“高价值”Token 的注意力分数，即实现“高信噪比”的稀疏化，成为了当前学术界的研究热点。

1.2.2 研究进展

2025 年的稀疏注意力研究主要呈现出两条主线：一是针对预训练模型的轻量化动态稀疏，旨在不额外微调模型内部参数分布的前提下，挖掘现有模型的稀疏潜力；二是原生稀疏训练，通过特殊的模型训练方式，将稀疏性内嵌至模型的参数中。

轻量化稀疏方法

在轻量化稀疏方面，DuoAttention^[17] 发现大语言模型的注意力头呈现出明显的二分特征：部分“检索头”主要负责长距离信息捕获，而大量“流式头”仅关注局部上下文。基于此发现，DuoAttention 对注意力头进行静态分类，仅保留检索头的全量计算，而对流式头实施激进的稀疏化（如仅保留最近邻 Token），从而在极少损失精度的情况下大幅降低了 KV Cache 的内存占用与计算开销。除了头级别的表现意外，FlexPrefill^[18]和 X-Attention^[19]等方法引入了更细粒度的分块估计机制，通过低成本的先验评估，动态筛选出高概率的注意力块，避免了对无效区域的计算，显著提升了长文本任务的推理吞吐量。

原生稀疏训练方法

然而，免训练方法往往受限于模型先天注意力分布，2025 年最具颠覆性的进展来自于大语言模型的原生稀疏训练。Moba^[20] 与 NSA^[21]（Native Sparse Attention）摒弃了传统的全注意力预训练范式，转而采用块级别的稀疏策略进行原生训练。这种设计充分利用了 GPU 在处理稠密矩阵块乘法上的硬件优势，实现了训练效率与模型性能的双重提升。更进一步地，DSA^[22]（Dynamic Sparse Attention）突破了块级稀疏的限制，通过轻量化索引器以及一系列高度定制的高效算子，实现 Token 级别的动态稀疏选择。DSA 能在保持极低计算复杂度的同时，不显著损失模型的性能表现，更从根本上缓解长文推理的成本问题。

1.2.3 未来展望

纵观 2025 年的演进脉络，稀疏序列建模正经历着从“静态启发式”向“动态自适应”，从“粗粒度分块”向“细粒度 Token”的深刻变革。未来的稀疏模型将不再仅仅是算法层面的优化，而是走向更加深度的软硬件协同设计。单纯的理论稀疏率已不再是唯一指标，未来的核心挑战在于：如何设计与现代加速器（如 GPU/TPU）内存层级相匹配的稀疏算子，在非连续内存访问带来的延迟与计算密集度之间寻找新的帕累托最优。这不仅将决定稀疏大模型能否彻底取代全注意力架构，也是通向无限上下文推理的必由之路。

1.3 混合专家模型

1.3.1 研究背景

在追求通用人工智能的进程中，模型参数量的扩张仍然是提升智能水平的最确定性路径。然而，随着模型规模迈向万亿参数级别，传统的稠密模型在训练和推理环节面临着难以承受的算力成本与显存带宽压力。

在此背景下，混合专家 (Mixture-of-Experts) 架构凭借其“**高参数量、低激活量**”的特性，成为 2024 年至 2025 年大模型架构演进的主流选择。其核心思想在于将前馈神经网络分解为多个独立的专家，并利用门控网络针对每个 Token 动态选择少数专家进行计算，从而在保持模型巨大容量的同时，显著降低推理时的实际计算量。

进入 2025 年，随着 DeepSeek-R1^[23]、Qwen3^[24]、Llama 4^[25] 等一系列标志性工作的发布，MoE 技术迎来了新的发展阶段。

当前的 MoE 研究不再局限于简单的组件堆叠，而是转向底层机制的深度重构：在**架构形态**上，“细粒度 + 共享”的范式进一步被验证，MoE + Mamba / Diffusion 的混合架构被初步探索，以解决知识解耦与通信效率问题；在**路由机制**上，从强约束的负载平衡转向“Sigmoid 路由 + 系统级优化”的无损策略，以解决负载与性能的权衡问题；在**效能边界**上，开始探索极致稀疏度下的扩展定律 (Scaling Laws)，寻求计算效率的理论极限。

下面，我们将从架构设计、路由与负载均衡、扩展定律与极致效率的探索三个维度，系统梳理 2025 年 MoE 技术的研究进展。

1.3.2 研究进展

架构设计

MoE 的架构设计在 2025 年经历了重要发展，主要体现在**专家粒度**、**特殊专家**以及**异构架构**等方面。

DeepSeek-R1 沿用了 DeepSeek-MoE 中“**细粒度专家 + 共享专家**”的经典范式，通过将大专家拆解为大量小专家来提升知识解耦能力，并利用共享专家捕获通用共性知识^[23]。后续的工作对此范式进行了不同方向的探索。

专家粒度方面，GLM-4.5 倾向于更紧凑的设计，将专家数量从 256 缩减至 160^[26]。相反，Kimi K2 则选择了进一步扩大规模，将专家数量扩充至 384 个，旨在通过更高的稀疏度提升专家的专业性^[27]。

共享专家方面，阿里在 Qwen3 的技术演进中经历了一次变化。在 Qwen3 初期版本中，团队尝试移除了共享专家^[24]；但在随后的 Qwen3-Next 中，为了追求极致的稀疏度与通用能力的稳定性，又重新引入了共享专家机制^[28]。

除了共享专家的特殊专家设计，美团的 LongCat-Flash 引入了 MoE++ 中的**零计算专家**，允许模型在某些不需要复杂处理的 Token 上跳过 MoE 计算，即路由到一个“空”专家，这种设计为提升资源的动态利用率提供了新的视角^[29]。

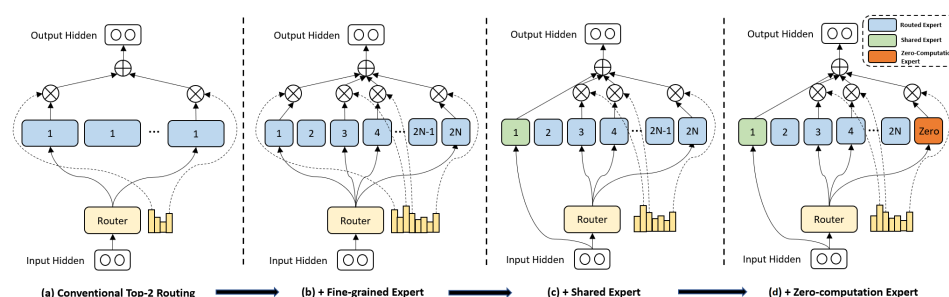


图 1.2: MoE 细粒度专家、共享专家及零计算专家设计^[30]

MoE 架构也开始突破标准 Transformer 范式，与**新兴架构**进行深度融合。腾讯混元团队发布的 Hunyuan-TurboS 提出了 **Mamba-Transformer Synergy** 架构。该模型在 MoE 层中融合了状态空间模型，并引入了回收路由机制，旨在利用 Mamba 在长序列处理上的线性复杂度优势，缓解 MoE 在超长上下文下的计算压力^[31]。

另一方面，蚂蚁团队的 LLaDA-MoE 则探索了 **Diffusion + MoE** 的结合。作为 MoE 扩散语言模型的尝试，LLaDA-MoE 展示了 MoE 的稀疏

激活特性在生成式扩散过程中的应用潜力^[32]。

路由与负载均衡

路由机制影响 Token 的分发效率，负载均衡则关乎硬件利用率的上限。针对前一阶段的“负载不均”与“辅助损失干扰”问题，2025 年的技术趋势转向“无损路由算法”与“系统级负载调度”相结合的双重优化策略。

路由激活函数方面，DeepSeek-V3/R1 采用的 **Sigmoid Routing** 成为行业关注的焦点^[23]。与传统的 Softmax 归一化不同，Sigmoid 允许每个专家独立地对 Token 进行打分。这种机制不仅解耦了专家之间的竞争关系，更重要的是引入了“**无辅助损失**”的负载均衡理念。通过移除可能对模型性能产生干扰的辅助损失项，模型能够更专注于主任务目标的优化。

OpenAI 发布的 gpt-oss 也采用了类似的微调设计，引入截断与残差机制以增强路由选择的稳定性^[33]。

负载均衡策略方面，出现了从“算法侧强制平衡”向“系统侧动态调度”演进的趋势。传统的负载均衡往往通过辅助 Loss 强制要求专家负载均衡，这限制了专家的专业化分工。

DeepSeek 推出的 **EPLB** (Expert Parallelism Load Balancer)^[34] 和 **LPLB** (Linear-Programming-Based Load Balancer)^[35] 提供了系统级的解决方案。EPLB 针对**静态负载不均**（如特定领域的固有热点），通过复制高负载专家实现分流；而 LPLB 则针对**动态负载抖动**，引入**线性规划**算法，在批次范围内动态优化 Token 到专家的分配。这种方法通过系统调度解决了硬件负载瓶颈。

阿里 Qwen 团队在 Qwen3 及相关研究中提出了**全局负载均衡**技术^[36]。通过扩大负载统计的视野（从微批次扩展到全局批次），更有效地促进专家分化，使特定专家更专注于特定领域，从而有助于提升模型效果。

扩展定律与极致效率的探索

2025 年的研究重点之一是持续探索 MoE 的**扩展定律 (Scaling Laws)** 以及追求**极致的计算效率**。

Ant Group 发布的 **Ling 2.0** (Ling-V2)^[37] 和相关理论研究深入探讨了 MoE 的效率边界^[38]。他们提出了“**效率杠杆**”的概念，试图量化 MoE 相比同等稠密模型的计算收益，并指出通过优化配置，MoE 有望在保持性能的同时实现极高的激活稀疏度。

业界涌现出一批“高稀疏、低激活”的代表性工作。例如，Qwen3-Next 实现了仅激活 3.7% 的参数^[28]；Ling-mini-2.0 在拥有 160 亿参数的情况下，激活参数不高于 14 亿^[37]；OpenAI 的 gpt-oss 系列同样遵循了这一趋势，gpt-oss-120b 在 117B 的总参数下仅激活 5.1B^[33]，力求在有限算力下最大化推理性能。

1.3.3 未来展望

回顾 2025 年，MoE 技术完成了从架构探索到工程应用的重要跨越。展望未来，MoE 发展可能呈现出以下趋势：

端侧 MoE 的机遇与挑战。随着 Ling-mini-2.0 和 Qwen3-Next 等高稀疏度模型的验证，将 MoE 架构引入端侧场景开始受到关注。虽然 MoE 的低激活计算量契合端侧算力受限的特点，但其庞大的参数规模带来的显存占用与带宽压力仍是巨大的挑战。未来如何在保持模型知识容量的同时，解决端侧部署的存储与读取瓶颈，将是该方向能否落地的关键。

MoE 对推理能力扩展性的支撑。当前的 DeepSeek-R1、GLM-4.5 和 Kimi K2 等工作展示了 MoE 架构在复杂推理任务中的应用潜力。MoE 能够以较低的计算成本支撑更大的模型参数规模，这为通过 Scaling Laws 提升推理能力提供了高效的载体。未来，如何进一步利用 MoE 的模块化特性实现推理步骤的专业化分工，值得深入探索。

MoE 的软硬协同。随着 Pangu Ultra MoE^[39] 等针对特定硬件（如 NPU）优化的工作出现，未来的 MoE 算法设计将更加紧密地结合硬件特性。通信优化、计算重叠以及针对 MoE 的新型芯片架构，将共同推动大模型性能在现有硬件约束下的持续突破。

1.4 状态化序列建模模型

1.4.1 研究背景

在大语言模型的演进过程中，如何高效地处理长序列数据是一个关键问题。基于标准自注意力机制的 **Transformer** 架构^[1]虽然表现卓越，但在面对超长上下文时，其平方计算复杂度和巨大的推理显存开销成为其应用的重要限制。为了打破这一瓶颈，学术界在 2025 年以前主要沿着两条路径进行了相关探索：一条是追求极致效率的线性注意力机制，另一条是试图兼顾效率与性能的混合注意力架构。

线性注意力方面，探索的起点可以追溯到 **Linear Transformers** 工作^[40]，其考虑无 Softmax 时的自注意力可以将 Key 与 Value 优先结合计算，能将计算代价从序列平方级降为线性级，并论证这种形式可表达成对状态的循环迭代更新的形式，标志着线性注意力研究的起步。随后，线性注意力研究进入繁荣发展阶段，相继出现了 **RetNet**^[41]、**HGRN** 系列^[42-43]、**Mamba** 系列^[44-45]、**RWKV** 系列^[46-47]等工作，代表了线性注意力研究的一个高峰。在此阶段后，相关工作^[48]揭示了线性注意力本质上是一种**快权重编程器 (Fast Weight Programmers)**，并引入了 **Delta Rule** 更新机制，使得模型能更好的更新其隐含状态。紧接着，有研究工作^[49]解决了 Delta Rule 在现代硬件上难以高效并行化的难题，使得这种具备强大状态更新能力的线性注意力机制能够在大规模模型中有实际落地应用的可能。自此，基于 Delta Rule 的线性注意力机制开始成为 2025 年该领域研究的主流技术路线。

混合注意力方面，早期一部分研究者们出于效率与性能平衡的考虑，开始尝试将线性注意力或窗口注意力与标准注意力混合。在 **GPT-3**^[50]时期，研究者尝试通过交替使用全注意力层与滑动窗口注意力层来提升效率。随着 Mamba 系列架构的强势崛起，相关工作开始尝试使用线性注意力进行混合。**Jamba**^[51]率先实现了 Transformer 层与 Mamba 层的块状交替，旨在结合前者的长程召回能力与后者的推理效率。随后，**Samba**^[52]进一步探索了滑动窗口注意力与 Mamba 的结合，以实现无限上下文的高效建模。在这一过程中，有研究工作^[53]通过大规模实验验证了混合架构在处理复杂逻辑和长上下文任务中的优越性，指出混合模型在性能上往往能超越同规模的纯 Transformer 或纯状态化模型。此外，**Hymba**^[54]等工作打破了层间混合的限制，开始探索层内并行的混合头架构，让模型在同一层内同时具备高性能召回与高效上下文压缩能力。随着相关研究的深入，2025 年研究者们已经意识到仅使用线性注意力或窗口注意力几乎必然无法达到标准注意力的性能，因此混合架构开始成为基本共识，并进行了更过大规模的验证与实践。

下面，我们将详细介绍 2025 年状态化序列建模模型中线性注意力和混合注意力的最新研究进展。

1.4.2 研究进展

本部分将对 2025 年状态化序列建模领域的研究进展进行归纳整理。随着长文本处理和推理效率要求的日益提升，该领域的研究重点已从单纯的架

构创新转向对更新机制的精细化控制以及混合架构的设计。目前，进展主要集中在线性注意力的机制演进与混合注意力架构的设计优化两大方向。

线性注意力

线性注意力通过将标准注意力转化为可表达为状态循环更新的形式，实现了推理效率的大幅提升。2024 年基于 Delta Rule 的线性注意力工作开始受到关注，2025 年在此基础上通过引入更复杂的更新法则、门控机制等来提升模型的记忆精度与表现力。

在前期基于 Delta Rule 的 **DeltaNet**^[48]的基础上,**Gated DeltaNet**^[55]首先引入了门控 Delta Rule，将门控机制的自适应记忆控制与 Delta Rule 的精确记忆修改相结合。这种设计允许模型在面对语境切换时快速清除陈旧信息，并在常规处理中实现状态更新。**Comba**^[56]受到闭环控制理论的启发，提出了一种标量加低秩的状态转移方案，并通过状态反馈和输出反馈纠正，增强了稳定性与表达力。

为了解决标量衰减在表达能力方面的局限性，**RWKV-7**^[57]引入了动态状态演化机制，通过向量值衰减替代标量衰减，使得状态中的每个通道都能独立变化。**KDA (Kimi Delta Attention)**^[58]延续了这一思路，通过引入对角化门控衰减来强化 Gated DeltaNet 的标量衰减，并专门设计了硬件高效的算法实现了实际效率的提升。

此外，**Titans**^[59]从测试时训练的视角出发，将隐状态视为权重矩阵，通过在随机梯度下降更新中引入权重衰减和动量机制，同样构建了一个能够学习记忆历史上下文的神经长期记忆模块。**Mamba-3**^[60]则在 Mamba-2^[45]的基础上进行了改进，引入了梯形离散化、复数化状态空间模型、多输入多输出状态空间模型等机制，在检索任务等方面取得了显著进步。

混合注意力

由于线性注意力和窗口注意力在复杂检索任务中仍存在局限，2025 年的研究普遍转向了混合注意力，通过结合全局注意力与状态化序列建模模块，实现在性能与效率之间达成更好的平衡。相关进展主要涉及混合线性注意力、混合窗口注意力和混合注意力设计分析等方面。

混合线性注意力。 这类架构旨在通过少量全局注意力层来弥补线性注意力的在上下文检索方面的缺陷。**MiniMax-01**^[61]及其推理模型 **MiniMax-**

表 1.1: 混合注意力关键模型汇总

混合类别	代表工作	混合组件 (高效 + 全局)	混合比例 (高效: 全局)
混合线性 注意力	MiniMax-01/M1 ^[61-62]	Lightning Attention + Attention (GQA)	7:1
	Hunyuan-TurboS ^[64]	Mamba2 + Attention (GQA)	57:7
	Qwen3-Next ^[65]	Gated DeltaNet + Attention (GQA)	3:1
	Kimi Linear ^[58]	KDA + Attention (MLA)	3:1
混合窗口 注意力	Gemma 3 ^[66]	SWA (1024) + Attention (GQA)	5:1
	Gpt-oss ^[10]	SWA (128) + Attention (GQA)	1:1
	MiMo-V2-Flash ^[67]	SWA (128) + Attention (GQA)	5:1

M1^[62]采用了 Lightning Attention^[63]与全局注意力的周期性混合模式，实现在千亿参数规模下的超长上下文处理。**Hunyuan-TurboS**^[64]将 Mamba2 与分组查询注意力结合 (GQA)^[2]，通过精确控制注意力层的比例，在保证效率的同时增强了上下文理解。**Qwen3-Next**^[65]发现 Gated DeltaNet 在上下文学习上优于传统的滑动窗口，并确定了 3:1 的线性与标准注意力混合比例。类似地，**Kimi Linear**^[58]将 KDA 与多头潜在注意力 (MLA)^[3]进行 3:1 混合，性能在多场景下超越了全注意力模型。

混合窗口注意力。 这类架构主要是涉及窗口注意力与全局注意力的混合。**Gemma 3**^[66]发现引入大幅缩短关注范围的窗口注意力对语言建模的影响很小，采用了 5:1 的窗口注意力层与全局注意力层交替模式，有效缓解了长文本下的效率问题。**Gpt-oss**^[10]延续了 GPT-3 经典的交替设计，在长度为 128 的窗口注意力和全局注意力之间切换。**MiMo-V2-Flash**^[67]同样采用窗口注意力与全局注意力混合的架构设计，实现了效率与性能的有效平衡。

混合注意力设计分析。 针对混合注意力架构的设计原则，相关研究工作给出了一些有意义的指导。字节跳动的一项研究工作^[68]指出，线性注意力

的独立表现并不等同于其在混合架构中的表现，上下文召回能力是决定混合比例的关键，并建议使用具有选择性门控和受控遗忘机制的线性注意力（如 GatedDeltaNet），维持 3:1 到 6:1 的混合比例。另一项 Meta 的研究工作^[69]则强调，混合架构不仅在效率上占优，还展现出了更优的帕累托前沿。值得关注的是，该工作同样考虑了尚在探索阶段的层内混合架构（如 **TransMamba**^[70]、**Liger**^[71]等），通过测试分析得到层内混合架构同样可以达到良好性能的结论。

1.4.3 未来展望

前面回顾总结了 2025 年状态化序列建模模型的关键进展。这些研究共同推动了语言模型从单一的注意力机制向更高效、表现力更强的状态化架构突破。接下来，我们对该领域的未来发展趋势进行展望。

状态化序列建模在今年取得了实质性的性能飞跃，这主要得益于混合架构的引入。我们猜测，未来的研究将继续突破“**线性与非线性**”的二元对立，出现更多的混合架构的工程实践与理论分析工作。

另外，2025 年的工作已经证明，新架构的硬件实现效率与新架构的设计同等重要，我们猜测未来将进一步**深化软硬件一体化的架构设计**。

以上是我们关于今年状态化序列建模模型发展情况的总结。我们期待这一领域在未来能持续突破，在效率与智能之间找到更完美的平衡点。

1.5 多模态语言模型架构

1.5.1 研究背景

回顾 2023 年至 2024 年的多模态大语言模型（MLLM）发展史，主流架构（如 LLaVA^[72]等）普遍采用了一种松耦合的“三明治”结构：一个预训练好的视觉编码器，一个轻量级的投影层，以及一个冻结或微调的 LLM 基座。这种架构虽然有效地打通了视觉与语言的语义空间，但其本质上是一种“外挂式”的视觉感知——LLM 本身并不具备原生的视觉处理能力，而是依赖于外部编码器将图像“翻译”为伪语言 token。进入 2025 年，随着算力的提升和数据规模的指数级增长，MLLM 架构迎来了决定性的范式跃迁。这一年的核心主题是深度融合与原生统一，研究人员不再满足于简单的特征对齐，而是致力于重构 LLM 的内部机制，使其能够像处理文本一样原生、高效、精细地处理视觉、视频甚至 3D 数据。

1.5.2 研究进展

2025 年多模态语言模型架构的进展，主要集中于视觉理解模型架构、理解生成统一架构等多个方面，本部分将对这些进展依次进行介绍。

视觉理解模型架构

在视觉理解领域，2025 年的架构竞争焦点已经从单纯的参数量堆叠，转移到了如何更高效地利用视觉信息。核心的架构突破集中在如何处理可变量分辨率、如何防止特征在深层网络中的稀释，以及如何进行长时序建模。Qwen3-VL^[73]作为 Qwen-VL 系列的集大成者，在 2025 年展现了惊人的架构统治力。其核心贡献在于彻底摒弃了静态的视觉处理方式，构建了一套适应性极强的全动态感知系统。相比桥接器方法尽在浅层进行特征融合，Qwen3-VL 受 U-Net 跳跃连接启发引入了 DeepStack 机制，使得视觉编码器不再仅仅输出一个最终的特征向量，而是输出一系列多层级的特征图。这些特征图包含了从低级的纹理、边缘信息到高级的语义概括信息，将通过一个专门设计的合并器，被注入到 LLM 的不同层级中，缓解了长路径带来的信号衰减。此外，Qwen3-VL 还采用了高度、宽度、时间交织的 M-RoPE 位置编码，缓解了传统 1D 编码对视觉数据拓扑结构的破坏。基于 M-RoPE 的高效编码，Qwen3-VL 实现了 256k token 的原生上下文窗口，并支持长达 1 小时以上的视频理解。对于多模态信息的输入，Qwen3-VL 彻底摒弃了填充或缩放到固定分辨率的做法，采用了 NaViT 式的处理逻辑实现了原生动态分辨率。除了融合方式上的改进，Ernie 4.5（及 4.5-VL）^[74]代表了认知深度与计算效率的结合，重点在于模拟人类的“慢思考”过程。其大规模 MoE 架构的核心在于路由机制的特化，能够根据输入内容（如复杂图表 vs 简单风景）动态激活特定的“视觉专家”或“跨模态专家”。这种稀疏激活机制在保持海量知识容量的同时，显著降低了推理成本。为了解决高分辨率带来的 Token 爆炸问题，InternVL 3.5^[75]引入了视觉分辨率路由器（ViR）。这是一个轻量级的分类器，置于 LLM 之前，对图像 Patch 进行语义评估。高信息量 Patch（如文字、人脸）保留原分辨率，低信息量 Patch（如背景）进行高倍率压缩。这种 Patch-aware 的动态路由策略在减少约 50% 视觉 Token 的同时，保持了性能几乎无损。

理解生成统一架构

此外，相比于理解（Image-to-Text）与生成（Text-to-Image）长期被视为独立任务的传统范式，2025 年的统一架构试图在一个模型中解决“看”与“画”的模态鸿沟。当前的统一多模态模型正从早期的简单对齐向深层参数共享演进，试图在 Transformer 内部协调理解任务所需的特征不变性与生成任务所需的特征变化性。ByteDance 推出的 Bagel 系列^[76]采用了双塔混合专家架构。它不强求单一编码器，而是保留了两个独立的视觉编码器分别用于语义理解与空间生成。在 LLM 内部，模型通过路由机制动态选择“理解专家层”或“生成专家层”进行处理。DeepSeek 推出的 Janus-Pro^[77]则代表了解耦编码的极简主义路线。该架构明确指出了单一编码器无法同时满足理解和生成的需要，因此在输入端进行了完全解耦：使用 SigLIP 提取语义特征供 LLM “阅读”，同时使用 VQ-Tokenizer 将图像离散化为 ID 序列供 LLM “预测”。虽然感官通路分离，但核心推理完全由一个统一的自回归 Transformer 完成。这种设计证明了只要 LLM 足够强大，完全可以在内部协调两种不同的视觉表示，且避免了任务间的相互干扰。

1.5.3 未来展望

2025 年的多模态架构演进表明，“大一统”并非简单的模块堆叠，而是对认知过程的深刻模拟。Bagel 和 Janus-Pro 的成功预示着未来架构将走向“感官解耦，思维统一”。即输入输出端根据物理信号特性（如像素、音频波形）保持专业化，但中间的推理核心将成为通用的多模态处理器。随着 M-RoPE 等技术的应用，LLM 正在逐步建立起原生的时空坐标系。这标志着多模态模型正从处理“媒体数据”向构建“世界模型”迈进，为具身智能的爆发奠定了架构基础。

1.6 新兴方向

1.6.1 主要背景

凭借自注意力机制带来的并行训练优势和卓越的上下文建模能力，Transformer 彻底终结了循环神经网络和卷积神经网络在序列建模领域的统治。然而，随着模型规模向万亿参数迈进，以及上下文窗口向百万级 Token 扩展，Transformer 架构固有的局限性在 2024-2025 年间变得愈发不可忽视。首先，推理效率的物理瓶颈日益凸显。基于 Transformer 的自回归语言模型采用

“逐词预测”模式，这种串行生成机制无法利用现代 GPU 的并行计算能力，成为长文本推理的主要瓶颈。其次，推理深度的静态束缚限制了模型的思维能力。目前主流 LLM 的计算图是静态的，无论输入是简单的问候语还是复杂的数学证明题，模型都必须经过相同层数的计算。这种“均一化”的计算分配方式不仅浪费算力，更难以模拟人类“快思考与慢思考”的认知机制。最后，灾难性遗忘与持续学习的悖论尚未解决。Transformer 模型在预训练结束后，其权重即被“冻结”。虽然上下文学习提供了一种临时的适应机制，但这种记忆是短暂的，随窗口关闭而消逝。模型无法像生物大脑一样，通过持续的突触可塑性将新知识永久整合进长期记忆，而不覆盖旧知识。针对上述痛点，2025 年的新兴架构研究主要集中在三个突破口：扩散语言模型、动态计算循环模型以及嵌套学习架构。这三者分别试图从生成范式、计算维度和记忆机制三个层面，重构人工智能的底层逻辑。

1.6.2 扩散语言模型

传统的自回归（AR）模型受限于单向依赖，难以在生成过程中进行全局规划。相比之下，扩散模型虽然在图像领域大放异彩，但在离散文本领域的应用一直受限于采样效率和优化难度。进入 2025 年，离散扩散语言模型取得了重大突破，开始挑战 AR 模型的统治地位。LLaDA^[78]通过引入一种基于掩码的生成策略，从头重新定义了文本扩散过程。不同于高斯噪声的去噪，LLaDA 将生成过程建模为从完全掩码状态到完全文本状态的逐步“揭示”。这种机制允许模型在生成过程中同时关注双向上下文，显著提升了文本生成的连贯性和全局一致性。与此同时，Dream-LLM^[79]则探索了“AR 初始化 + 扩散微调”的混合路线。它利用现有的高性能 AR 模型权重作为初始化状态，解决了扩散模型训练收敛慢的难题。这种方法不仅继承了 AR 模型丰富的语义知识，还赋予了模型并行生成的特性。尽管双向注意力和并行生成展示了巨大的潜力，理论上能将长文本生成的延迟降低数倍，但在实际推理场景下，扩散语言模型与经过极致优化的 AR 模型相比仍有距离，是目前学术界亟待攻坚的热点问题。

1.6.3 动态计算

为了打破静态计算图的束缚，动态计算架构旨在实现“按需分配算力”。其核心思想是让模型根据输入 Token 的难易程度，动态决定计算的深度，即在某些层级进行循环处理。2025 年，这一领域在推理任务上取得了显著进

展。Hierarchical Reasoning Model (HRM)^[80]提出了一种分层路由机制，能够识别逻辑断点，仅在关键推理步骤上触发深层循环计算，而在简单语义连接上使用浅层网络。除此之外，Tiny Recursive Model (TRM)^[81]进一步探索了更加激进的实验设置。仅凭借一个微小的神经网络（7M 参数）进行反复的自我迭代和状态修正，TRM 在 ARC-AGI 等高难度抽象推理基准上，实现了对参数量大其数万倍的传统大模型的超越。字节跳动推出的 Ouro^[82]架构则将这一理念推向极致。Ouro 采用了递归 Transformer 块的设计，通过在时间步上复用权重来实现逻辑推理深度的扩展，而非简单地堆叠物理层数。这使得 Ouro 在显存极其受限的边缘设备上，也能通过增加推理时间来处理复杂的逻辑推理任务。这种“思维循环”机制，有效地模拟了人类反复斟酌的思考过程，在数学和代码生成任务上展现出了惊人的参数效率。

1.6.4 嵌套学习

持续学习一直是 LLM 迈向通用人工智能的最后一道门槛。2025 年，关于快速权重和元学习的研究迎来了复兴，其中最具代表性的是 Nested Learning^[83]架构。嵌套学习架构试图解决灾难性遗忘，赋予模型终身学习的能力。该架构的核心在于将“优化算法”本身内化为模型的一部分，通过多时间尺度更新来模拟生物脑的记忆机制。该范式认为，模型不应只有一个全局的静态优化器，而应由无数个嵌套的优化过程组成。在此基础上构建的 Hope 架构，包含“快权重”和“慢权重”两套系统。快权重类似于短期记忆或海马体，随每个输入样本迅速更新，能够瞬间适应当前的上下文风格；而慢权重类似于长期记忆或皮层，更新缓慢，仅提取统计规律，确保旧知识不被覆盖。这种设计使得 Hope 模型能够在不重新训练的情况下，通过前向传播中的动态权重调整，实现对新任务的即时适应，从根本上回避了灾难性遗忘。这不仅极大地扩展了模型的有效记忆容量，更有潜力彻底解决长上下文窗口带来的算力爆炸问题，为实现真正的“终身学习”智能体解封架构限制。

1.6.5 未来展望

纵观 2025 年的架构创新，虽然基于 AR 架构的标准 Transformer 凭借其成熟的生态，在短时间内依然占据统治地位，但针对其固有缺陷的非主流架构已展现出强大的生命力。扩散语言模型为并行化推理和全局规划提供了新范式；动态计算模型为端侧智能和深度推理提供了高能效方案；而嵌套学习则为解决记忆与学习问题指明了方向。这些新兴架构并非要完全取代

Transformer，而是极有可能与现有架构进行融合。这种复合架构，将蕴含着巨大的应用潜力，推动人工智能从单一的语言模型向复杂的认知系统进化。

1.7 本章小结

本章对 2025 年大语言模型基础架构进展进行了梳理，介绍了相关工作为模型能力增强、效率提升等做出的关键架构革新与调整。在全注意力建模方面，研究聚焦于效率与能力的平衡，通过注意力分组机制与混合位置编码等进行优化；稀疏序列建模确立了从静态选择模式向动态自适应选择模式的转变，有效提升了模型处理长序列文本的能力；混合专家模型进一步确立了“高参数、低激活”的范式，并在架构设计、负载均衡、路由算法等方面取得实质进展；状态化序列建模基本确立了混合注意力的架构范式，在高效率的基础上有效增强了长程记忆与检索精度；多模态语言模型架构则由外挂式感知转向原生统一。此外，新兴方向围绕扩散语言模型、动态计算及嵌套学习展开探索，旨在打破静态计算、持续学习等方面的局限。这些进展共同呈现出大模型基础架构向高效率、高性能方向迈进的未来态势，在显著拓展现有智能处理边界的同时，也为未来构建更具泛化性与自主学习能力的通用智能体系奠定了坚实的架构基础。

第二章 大语言模型训练

2025 年是大语言模型训练技术关键创新的一年，核心技术呈现多维度突破与深度融合的特征。本章聚焦于后训练技术更新，数据获取与数据治理，模型能力提升以及开源训练框架四个部分，详细介绍了各个部分在 2025 年的新进展和新趋势。简单来说：在 2025 年大语言模型后训练技术更新领域：SFT 在 LORA 基础上从高效收敛、少参数微调等方向突破，进一步平衡成本与效果；强化学习则因 RLVR 技术崛起成为热门，离线、在线及混合策略均有针对性进展，配套的奖励模型和虚拟环境也同步升级。在 2025 年大模型数据获取与治理领域：开源数据集构建形成预训练扩规模、中训练补空白、后训练强推理的三大主线，数据处理向动态语义去重、量化高效过滤等方向升级；多模态数据集则实现从规模扩张到语义关联构建、从通用泛化到专项适配的转型，全面支撑模型性能提升。在 2025 年大语言模型能力提升领域：系统梳理了长上下文、推理、数学/代码、工具调用及 Agentic RL 五大核心方向的研究背景、年度进展与未来展望，展现了大模型从通用对话系统向自主智能体演进的技术脉络与范式变革。各方向均呈现出从静态模仿学习向动态策略优化、从单一能力提升向多能力协同进化的趋势，强化学习与真实环境交互成为突破能力瓶颈的关键。在 2025 年大语言模型开源训练框架领域：VeRL、ROLL 等六大主流 LLM 开源后训练框架的核心设计、优劣势及适用场景，各框架呈现出极致性能、特定场景优化、新范式探索与易用性导向的差异化发展趋势。当前框架在显存效率、任务适配等方面成效显著，但仍面临接口不统一、长序列/智能体任务调度挑战等问题，未来将向通用化基础设施方向演进。

2.1 后训练技术更新

后训练是大语言模型训练范式中必不可少一环，用于激发出预训练模型在下游任务上的强大性能。本章节将介绍 2025 年大语言模型后训练的发展

情况，并详细介绍其中的关键算法，涉及成熟的有监督微调技术以及今年备受关注的强化学习技术。

2.1.1 SFT 最新进展

研究背景

大语言模型（LLMs）的强大性能离不开训练，而 SFT 是多数大语言模型训练流程中必不可少的一环，是赋予大语言模型指令跟随能力的关键技术。早期的 SFT 可以追溯到 BERT 等传统模型的预训练-微调训练范式，人们发现尽管在海量无监督数据集上预训练过的模型可以获得强大的语言理解能力，但因任务分布不同并不能直接用在下游任务上。因此，需要再对模型做有监督的微调（SFT），让模型能处理输入分布向下游任务对齐。惊人的是，SFT 被发现能用少量的数据就可以让模型在下游任务性能上获得明显的提升，因此理论上成本可以远小于预训练，让多数机构都可以自主实施。然而现实中却并非如此，早期的 SFT 技术依然具有较高门槛，这主要是因为以下两点瓶颈：一是**参数量**，微调的参数量少时训练效果不佳，想达到预期效果就要微调足够多的参数，进而需要更大的算力成本。二是**数据量**，训练数据量少时模型的泛化性下降，想避免灾难性遗忘就要足够多的数据，增加了数据成本。因此，如何平衡**成本**和**效果**成为了 SFT 的重点研究问题。

多年以来，人们为了实现成本和效果的平衡，陆续提出了 Adaptor^[84]、P-tuning^[85]等方法。人们发现微调的参数不必和基座模型参数一一对应，通过只微调额外的模块，就可以只微调少量的参数就对模型尽可能多的层进行修改。并且随着微调参数的下降，微调的复杂性也在下降，使得数据量低引起的灾难性遗忘得到解决。随着研究的不断深入，微软提出的 LORA^[86]微调逐渐成为了大模型 SFT 的主流技术。LORA 通过矩阵分解，实现只微 $O(rd)$ 量级的参数，就可对模型所有权重进行修改，进而实现较高的训练效果。这使得目前大部分的大语言模型 SFT 工作都是通过 LORA 或 LORA 变种实现的。

LORA 仅仅是一种简单的数学技巧尝试，却不仅基本上实现 SFT **成本**和**效果**的平衡，并且证明了 SFT 的发展可以由数学等理论学科指明方向。因此，今年的工作开始向更广泛的理学理论寻求帮助，不断突破 SFT 的各方面性能极限，连续降低 SFT 的成本。

下面，将介绍今年在 SFT 技术上的主要研究进展。

研究进展

本部分将对 2025 年有影响力的 SFT 技术进行归纳整理。今年 SFT 的主要研究方向延续了 LORA 对低成本和高效的追求，并且开始通过将各任务微调的参数分离，避免因任务分布不同导致的 LLMs 泛化性下降。下面将分别从高效收敛、少参数微调、针对性微调三方面进行介绍，此外也将介绍对 LORA 的实施方法分析性工作。

更高效的收敛 这一类工作不改变 LORA 的基本结构，而通过改变 LORA 的初始化、训练方法来增加 LORA 参数在 LLM 策略空间的表征方式，进而加快收敛速度。考虑到 LORA 块是权重矩阵的低秩近似，LORA-One^[87]使用训练集子集在 LLM 上的梯度做 SVD 分解，用于 LORA 的初始化，获得了更快的收敛速率和更全局最优的收敛效果。考虑到语义的幅度和方向的高度纠缠影响训练稳定性，Dual-LORA^[88]将 LORA 块分解为分别表示符号和绝对值的两组 LORA 的乘积，进而维持了训练过程中的语义方向稳定性。

更少的参数 这一类工作期望在 LORA 的基础上，进一步减少需要微调的参数，降低 SFT 门槛。考虑到预训练权重已经包含了所有下游任务权重的正交方向，QR-LORA^[89]将 LORA 块定义为权重矩阵的带列主元的 QR 分解，并只训练对角阵的对角线，将微调参数量从 $O(rd)$ 降低到了 $O(r)$ 。利用了量子计算的么正算符参数化和张量网络，Quantum-PEFT^[90]将权重矩阵分解为基础量子门的张量积，将参数量降低到了 $O(\log(d))$ 。基于极简几何学，Uni-LoRA^[91]将不同层的 LORA 建模为同一向量的不同投影，进而大大降低了需要微调的参数量。

更有针对性的微调 这一类工作将不同下游任务的参数解耦开，进而缓解灾难性遗忘问题。考虑到不同任务需要的微调参数不同，GainLoRA^[92]将微调参数建模为各任务的 LORA 的加权求和，在执行每个任务时会屏蔽其它任务的 LORA 块，进而避免了灾难性遗忘。考虑到不同层权重之间的潜在关系，BSLoRA^[93]将不同层的 LORA 进行了权重共享，增加了模型鲁棒性。

实施方法的分析性工作 Thinking Machines Lab 今年对 LORA 的分析性工作^[94]得到了很多对实施细节、参数设置的结论，对于 SFT 应用有很强的指导性工作：一是 LORA 应用层数至关重要，将 MLP 层也实施 LORA 相比只微调注意力权重更有效果；二是 LORA 的秩和数据量相关，在秩和数

据量匹配时使用 LORA 具有接近全参数微调的效果，数据量超过秩容量后微调效果下降；三是 LORA 对训练批次大小的容忍度更高，随着批次减小，LORA 和全参数微调的差距减小；四是 LORA 的学习率，LORA 微调的最佳学习率通常是同情况下全参数微调的 10 倍，并且和模型、秩无关。

未来展望

前面回顾了 2025 年 LLMs 在 SFT 技术上的关键进展。这些进展极大地降低了 SFT 的门槛，让一些仅有少量算力资源的机构也能实现对大语言模型行为的显著修改。接下来，我们对 SFT 的未来发展趋势进行展望。

考虑到今年多数的工作都基于线性代数、几何学、量子物理等基础理论，并且对灾难性遗忘等难题也陆续出现了针对性研究，因此未来的发展趋势很可能会延续这种趋势。首先，更多的基础学科理论可能会被考虑进来。由于本质上年轻的大模型还是由矩阵、运算组成的机制，因此具有悠久历史的几何学、代数学、计算学等理论中依然可能存在对 SFT 技术的进一步启发。此外，训练过程的动态性质，以及 MOE 等机制和生物群体的相似性，因此群体生物学、社会学等理论也可能存在对 SFT 的指导作用。其次，灾难性遗忘在未来依然会成为研究点之一，不破坏模型泛化性的微调是备受广泛期待的。现阶段的主要方法是对微调参数进行显式分解，实现参数和任务的一一对应，然而考虑到任务之间具有的相似性，在未来对参数的分解方法可能会向隐式发展。如果存在隐式分解机制使得每个任务对参数的修改范围被约束在某种空间内，并且相似任务的微调空间具有相似性，那么可以预见这种分解对改善灾难性遗忘具有有益效果。

以上是我们今年对 SFT 技术的总结，我们希望该领域在未来可以取得更显著的成就，让更多的用户可以通过自主微调构建定制的大模型。

2.1.2 强化学习算法进展

研究背景

强化学习是不同于 SFT 的训练范式，和 SFT 相比具有两个显著优势：一是**数据成本优势**，强化学习不需要显式标注每个指令下的示范回复，本质上是半监督训练，数据标注成本远低于 SFT，使得理论上强化学习可以轻易地实现大规模后训练。二是**训练效果优势**，有监督微调使得模型性能受到标注者能力的限制，而强化学习是模型直接从客观环境中学习，因此理论上可以使模型得到超过人类的性能，这种理论在 AlphaGo 等传统强化学习项目

2025 年大语言模型（LLMs）进展报告

上已经得到了验证。这些优势使得强化学习被认为是大语言模型下一步要重点发展的技术。

大语言模型强化学习最早可追溯到 2022 年的 ChatGPT，OpenAI 通过雇佣志愿者对大模型生成的回复进行排序和打分，构建反馈信号，使用在线强化学习优化大模型的文本生成策略。这种训练方法开创了**人类反馈强化学习（RLHF）**领域，并从 2023 年开始得到了广泛的发展。2023 年至 2024 年期间大量的工作显著降低了 RLHF 的成本，首先是 **RLAIF** 方法，使用 AI 代替人类降低了数据监督成本；其次是**离线强化学习**在大模型领域的崛起，通过训练、推理两阶段分离，显著降低了强化学习的训练成本。然而 RLHF 的主要面向是文本生成等文科能力，反馈需要昂贵的人类偏好标注，训练效果又难以看出显著优势，因此强化学习的两个主要优势一直没有发挥出来。再考虑到强化学习训练成本普遍高于 SFT，使得强化学习在往年一直没有特别热门的应用。

在 2025 年一切都不同了，因为**可验证奖励的强化学习（RLVR）**诞生了。RLVR 主要面向数学、代码等有明确评价标准的推理任务，很容易就能比较出不同模型之间的性能差距。RLVR 直接使用模型推理结果是否正确作为反馈，在推理任务上只需要正则表达式就能判断正误，因此反馈标注十分廉价。再加上 RLVR 主要用于训练模型自主生成推理思维链，通常在几千几万 token 级别，如果用人类标注需要难以承受的成本。这使得在推理能力增强方面，强化学习展现出了和 SFT 相比的显著优势。于是我们可以看到，2025 年开始，DeepSeek-R1、Qwen3、GPT-o1 等主流模型都开始大规模使用了强化学习，这使得 2025 年强化学习成为了大语言模型领域最热门的技术之一。

由于大语言模型强化学习仅发展了不到 3 年，因此即使取得了瞩目的成就，但技术尚欠成熟，缺乏完整的体系。因此在 2025 年，爆发较早的离线强化学习开始初步尝试构成完整理论体系，而爆发较晚且门槛较高的在线强化学习仍处于试探性阶段，下面将逐一介绍。

研究进展

强化学习从训练流程可以主要分为三类：一是**在线强化学习**，推理和训练交替进行，成本最高，但是效果最好。二是**离线强化学习**，推理和训练先后进行，不再循环，成本最低，但效果较弱。三是**混合强化学习**，介于在线和离线之间，甚至还会结合部分 SFT 技术，是在线、离线强化学习的折中

方案。

这三种强化学习在大模型的发展时间、应用领域目前都有较显著的区别。如图 2.1 所示，尽管最先被应用到 LLMs 的是在线强化学习 PPO，但由于在线 RLHF 极高的成本，以及离线 RLHF 可以用更精细的偏好标注弥补性能不足，因此 RLHF 领域取得统治性地位的是离线强化学习，在 2024 年取得了爆发性发展。这使得随着 RLHF，离线强化学习成为了最早爆发的强化学习类型。随着 RLVR 的诞生，强化学习在 2025 年得到了第二次爆发。由于 RLVR 的回报相比 RLHF 是更加明确的，离线标注也难以提高质量，并且 RLVR 对训练效果的要求更高，因此目前 RLVR 是在线强化学习的统治领域。此外，作为在线和离线强化学习的折中方案，混合策略强化学习在 2025 年也具有不少相关工作。

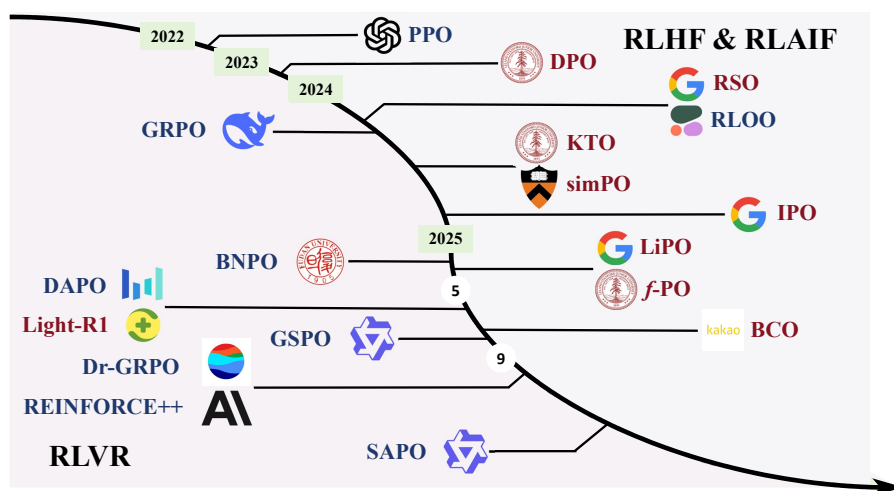


图 2.1: LLMs 强化学习发展流程

离线强化学习 LLM 离线强化学习的热度主要由 2023 年的 DPO 引爆，并在 2024 年井喷了大量的 DPO 增量式改进工作。我们在表 2.1 中总结了 2023 年-2025 年的部分关键离线强化学习工作。虽然种类繁多，但是这些算法缺乏统一的理论框架，以及缺少进而对 DPO 范式根本性不足的解决方案。因此 2025 年离线强化学习的主流工作，开始着手于两个主要方向：**统一的离线强化学习理论框架** f 散度偏好优化算法， f -PO^[101] 从和最优策略使用 f -散度做分布匹配的视角，提出了一种包含多数主流离线强化学习算法的理论框架。身份映射偏好优化算法，IPO^[100] 用多臂老虎机问题统一了目前主流离线强化学习算法，并提出了不基于排序模型的 IPO 算法，用于避免 DPO 的

表 2.1: 离线强化学习关键算法

算法	动机	目标函数
DPO ^[95]	将复杂的在线 RLHF 三阶段训练转换为一个阶段训练	$-\log \sigma \left(\beta \left(\log \frac{\pi(y_w x)}{\pi_{ref}(y_w x)} - \log \frac{\pi(y_l x)}{\pi_{ref}(y_l x)} \right) \right)$
simPO ^[96]	将回报和生成指标 (困惑度) 对齐	$-\log \sigma \left(\beta \left(\frac{\log \pi(y_w x)}{ y_w } - \frac{\log \pi(y_l x)}{ y_l } \right) \right)$
KTO ^[97]	打破数据集格式只支持偏好对的局限	$-\lambda_w \sigma \left(\beta \log \frac{\pi(y_w x)}{\pi_{ref}(y_w x)} - z_{ref} \right) + \lambda_l \sigma \left(z_{ref} - \beta \log \frac{\pi(y_l x)}{\pi_{ref}(y_l x)} \right)$
BCO ^[98]	减少 KTO 相比 DPO 的近似误差	$-\log \sigma \left(\beta \log \frac{\pi(y_w x)}{\pi_{ref}(y_w x)} - \delta \right) - \log \sigma \left(\delta - \beta \log \frac{\pi(y_l x)}{\pi_{ref}(y_l x)} \right)$
RSO ^[99]	用拒绝采样令采样策略最优，减少回报-策略关系误差	只改变了采样策略，没有改目标函数
IPO ^[100]	直接在回报维度做回归，避免排序损失引起的过拟合	$\left(\log \frac{\pi(y_w x)}{\pi_{ref}(y_w x)} - \log \frac{\pi(y_l x)}{\pi_{ref}(y_l x)} - \frac{1}{2\tau} \right)^2$
f-PO ^[101]	使用 α 散度代替 KL 散度做和最优策略的分布对齐	$\frac{1}{\alpha(1-\alpha)} \left(\sigma \left(\beta \log \frac{\pi(y_w x)}{\pi_{ref}(y_w x)} - \frac{\pi_{ref}(y_l x)}{\pi(y_l x)} \right)^{1-\alpha} - (1-\alpha) \sigma \left(\beta \log \frac{\pi(y_w x)}{\pi_{ref}(y_w x)} - \frac{\pi_{ref}(y_l x)}{\pi(y_l x)} \right) - \alpha \right)$
LiPO ^[102]	使用列表排序学习增加宏观偏好信息	$\sum_{ij} \Delta_{ij} \log \left(1 + \exp \left(\beta \log \frac{\pi(y_i x)}{\pi_{ref}(y_i x)} - \frac{\pi_{ref}(y_j x)}{\pi(y_j x)} \right) \right)$

过拟合。列表级别偏好优化算法, LiPO^[102]用排序问题建统一了主流离线强化学习算法, 并使用列表级排序学习替换缺乏全局信息的成对级排序学习。**改善 DPO 范式的各种偏差**简单偏好优化算法, simPO^[96]使用困惑度代替 DPO 的序列似然比, 克服了 DPO 的对齐维度和模型生成指标不对齐的偏差。拒绝采样偏好优化算法, RSO^[99]通过拒绝采样, 近似地获得最优策略下的偏好对, 克服了 DPO 因采样策略欠优导致的收敛偏差。二元分类优化算法, BCO^[98]通过对回报做 shift, 减小了 KTO 做训练时相比 DPO 在提升目标上的误差。

在应用领域, 除了传统离线强化学习广泛使用的 RLHF, Light-R1^[103]等工作也开始使用 DPO 实现 RLVR, 并得到了不逊于在线强化学习的效果, 大大降低了推理能力训练的成本, 具有相当的潜力。

在线强化学习 受 DeepSeek-R1 影响, 2025 年主流在线强化学习算法几乎全着手于 RLVR 领域。我们整理了主要的在线强化学习算法如表 2.2 所示。由于 RLVR 尚处于起步阶段, 因此目前主要问题是提升有效性和实用性, 因此 2025 年主流工作都通过在 GRPO 上施加各种技巧, 缓解 GRPO 的各种收敛问题。组间相对策略优化算法, GRPO^[105]是 DeepSeek-R1 的主要技术, 以推理结果的正确性作为回报, 训练模型自主形成思维链推理模式, 引发了 RLVR 领域的热度。动态采样策略优化算法, DAPO^[108]通过动态采样克服 GRPO 在轨迹采样样本全对或全错时的零梯度问题, 使用双截断机制、词元级策略梯度缓解 GRPO 的熵崩溃问题, 使用软长度惩罚缓解 GRPO 的回报噪声问题。正确实施的组间相对策略优化算法, Dr-GRPO^[107]通过修改损失和优势函数的归一化方法, 缓解 GRPO 的问题难度偏差和长句子偏差。组序列策略优化算法 GSPO^[109]、软调整策略优化算法 SAPO^[110]分别通过使用序列似然比代替概率比、软截断机制, 克服 GRPO 因重要性采样和截断机制导致的收敛不稳定性。Beta 标准化策略优化算法, BNPO^[106]通过修改优势归一化方法克服稀疏奖励问题, 并提出回报聚合方法来实现多目标优化。

混合策略强化学习 有监督微调能通过离线数据提供稳定的策略初始化, 但缺乏探索新解的能力; 而在线强化学习虽然可以通过探索不断提升推理上限, 却面临样本效率低、训练不稳定等问题。混合策略强化学习的核心思想在于融合有监督微调与 RL、在线策略与离线策略的优势, 在保证训练稳定性的同时提升探索效率, 从而突破单一范式的性能瓶颈。

表 2.2: 在线强化学习关键算法

算法	动机	目标函数
PPO ^[104]	回报更新 critics 模型, critics 模型训练策略模型	$\frac{1}{ y } \sum_{t \in y } f \left(A_t, \frac{\pi(y_t x)}{\pi_{rol}(y_t x)}, 1 - \epsilon, 1 + \epsilon \right) + \beta D_{KL}(\pi \pi_{ref}) + \alpha \ r_t + \gamma V_{t+1} - V_t\ ^2$ $s.t. A_t = \sum_i \lambda^i \gamma^i (r_{t+i} + \gamma V_{t+i+1} - V_{t+i})$
GRPO ^[105]	使用回报均值估计价值函数, 节省掉 critics 模型	$\frac{1}{G} \sum_{i \in G} \frac{1}{ y_i } \sum_{t \in y_i } f \left(A_i, \frac{\pi(y_{it} x)}{\pi_{rol}(y_{it} x)}, 1 - \epsilon, 1 + \epsilon \right) + \beta D_{KL}(\pi \pi_{ref})$ $s.t. A_i = \frac{r_i - \frac{1}{G} \sum_{j \in G} r_j}{std(r_1, \dots, r_G)}$
BNPO ^[106]	使用 beta 分布归一化估计优势函数	$\frac{1}{\sum_{i \in G} y_i } \sum_{i \in G} \sum_{t \in y_i } f \left(A_i, \frac{\pi(y_{it} x)}{\pi_{rol}(y_{it} x)}, 1 - \epsilon, 1 + \epsilon \right)$ $s.t. A_i = \frac{r_i - p(x)}{g(p(x); \alpha, \beta)}$
Dr-GRPO ^[107]	克服 GRPO 的归一化方法引起的长度、难度偏差	$\frac{1}{G} \sum_{i \in G} \sum_{t \in y_i } f \left(A_i, \frac{\pi(y_{it} x)}{\pi_{rol}(y_{it} x)}, 1 - \epsilon, 1 + \epsilon \right)$ $s.t. A_i = r_i - \frac{1}{G} \sum_{j \in G} r_j$
DAPO ^[108]	动态采样、双截断、词元级损失, 增加训练稳定性	$\frac{1}{\sum_{i \in G} y_i } \sum_{i \in G} \sum_{t \in y_i } f \left(A_i, \frac{\pi(y_{it} x)}{\pi_{rol}(y_{it} x)}, 1 - \epsilon_{low}, 1 + \epsilon_{high} \right)$ $s.t. A_i = \frac{r_i - \frac{1}{G} \sum_{j \in G} r_j}{std(r_1, \dots, r_G)}, std(r_1, \dots, r_G) > \tau$
GSPO ^[109]	用序列似然代替概率比, 避免不稳定性	$\frac{1}{\sum_{i \in G} y_i } \sum_{i \in G} \sum_{t \in y_i } f \left(A_i, \left(\frac{\pi(y_{it} x)}{\pi_{ref}(y_{it} x)} \right)^{\frac{1}{ y_i }}, 1 - \epsilon, 1 + \epsilon \right)$ $s.t. A_i = \frac{r_i - \frac{1}{G} \sum_{j \in G} r_j}{std(r_1, \dots, r_G)}$

围绕这一目标，近期工作主要从三个方面对混合策略强化学习进行探索。一类方法着眼于提升样本利用效率，通过引入离线策略机制复用历史数据。ReMix^[111] 使用截断重要性采样校正分布偏差，使旧策略生成的轨迹能够被安全复用，在数学推理任务中显著降低了采样成本；ExGRPO^[112] 则引入优先经验回放机制，根据轨迹熵和问题难度筛选高价值历史数据，避免无效探索并提升 RLVR 训练效率。

另一类方法关注有监督微调向 RL 过渡过程中的训练稳定性问题。SRFT^[113] 通过策略熵感知模型不确定性，动态平衡有监督微调与 RL 的损失权重，在保持探索能力的同时缓解灾难性遗忘；SuperRL^[114] 针对稀疏奖励场景下 RL 难以获得正反馈的问题，引入监督回退机制，在探索失败时自动退回有监督微调，从而避免训练停滞。

此外，混合策略强化学习思想也被拓展到策略对齐与知识蒸馏场景中。GAD^[115] 在黑盒蒸馏设置下构建生成对抗框架，将学生模型视为生成器、奖励模型视为判别器，通过在线策略对抗博弈缓解教师与学生分布不匹配的问题，使学生模型在无法访问教师概率分布的情况下仍能逼近其性能。

奖励模型 考虑到生成式模型也具有强鉴别能力，并能生成更详细的反馈报告，2025 年对奖励模型的研究正经历一场从“判别式直觉”向“生成式推理”的范式转变。传统的标量奖励模型将复杂的评估压缩为单一数值，不仅缺乏可解释性，更面临严重的“奖励黑客”问题，使得 LLMs 利用奖励函数的漏洞刷分，而非真正提升逻辑质量。为了突破这一瓶颈，2025 年的主流工作普遍重点关注以下两个方面：

更强的推理与解释性（推理型奖励模型 / 生成式评判） 该方向将奖励建模从“标量回归打分”转为“生成式推理评判”：奖励模型不仅给分，还生成可检验的评估理由/批判链，以提升可解释性并降低奖励黑客。其核心假设是评估者（Critic）具备不弱于策略模型（Actor）的推理能力，并用思维链对回答进行显式核验。在这一思路下，DeepSeek-V3.2^[22] 提供了较为系统的工程化落地：它以混合策略强化学习缓解多任务对齐的灾难性遗忘，并采用双轨奖励，对数学、代码等客观推理任务使用规则化确定性奖励，叠加长度惩罚与语言一致性约束，促使模型沿正确推理路径求解；对创作、开放问答等主观任务采用动态细则的生成式奖励模型，为每个提示生成评估准则，先输出自然语言批判再给出评分。其方法论来自 DeepSeek-GRM^[116] 的自原则批判微调（SPCT）：通过推理时计算扩展（并行采样多条批判路径并加权投票），小规模奖励模型也能获得强判别力。相关工作沿此推进：RM-R1^[117]

的评价链 (CoR) 从 Oracle 蒸馏推理轨迹, 将评判表述为“先解题、后评价”; ReasonGRM^[118] 以 R^* 衡量推理路径生成似然, 在奖励生成阶段过滤含幻觉的逻辑链。

更可靠的代理与过程验证: 引入外部工具与细粒度监督除了让模型“学会思考”外, 另一类工作致力于引入外部工具或细粒度的过程监督, 以解决模型内部知识幻觉和长链条推理中的信用分配 (Credit Assignment) 难题。考虑到模型参数知识的封闭性, Agentic Reward Modeling^[119] 将奖励模型进化为智能体。该系统集成了事实性验证代理和指令遵循验证代理, 能够主动调用搜索引擎检索证据, 或编写并执行 Python 代码来验证硬性约束 (如字数限制、格式要求), 从而提供比单纯参数直觉更可靠的物理世界反馈。针对过程监督, GenPRM^[120] 提出了生成式代码验证。不同于传统的分类式过程奖励模型 (仅输出 Good/Bad 标签), GenPRM 要求模型在验证推理步骤时, 生成一段验证代码或反思思维链, 并结合相对进度估计来量化当前步骤对最终解的贡献。这种方法有效地将稀疏的最终奖励信号转化为稠密的过程信号, 显著提升了模型在复杂数学推理任务中的表现。

虚拟环境 随着大语言模型预训练数据红利的逐渐消退, 后训练 (Post-training) 的核心范式正从依赖静态标注数据的有监督微调, 转向基于动态交互的“从交互中学习” (Learning from Interaction)。传统有监督微调受限于人类标注数据的质量与覆盖范围, 难以有效支持长程规划与复杂推理; 而基于虚拟环境的交互式学习通过引入“生成-执行-反馈”闭环, 使模型能够在试错中获得客观信号, 从而激发慢思考、自我修正与因果建模能力, 逐渐成为智能体与 RLVR 研究中的关键基础设施。

围绕这一范式转变, 现有工作一方面致力于统一交互式训练的基础设施, 缓解 RL 环境碎片化带来的工程与算法负担。Environments Hub^[121] 以去中心化社区的形式统一了多类环境的 API 接口, 并集成 Verifiers 库以支持 GRPO 等高效训练算法; 针对智能体场景中代码执行的安全性与环境构建成本问题, E2B^[122] 基于 Firecracker microVM 提供轻量级沙盒, 使开发者能够通过自然语言快速定义隔离的 Python 运行环境, 成为连接大语言模型与可执行环境的重要中间层。

另一方面, 研究者开始构建更具预测能力与因果理解能力的世界模型, 使智能体能够在行动前进行前瞻模拟, 或通过交互习得环境规律。Code World Model (CWM)^[123] 针对代码执行中的状态转移问题, 通过中间训练学习了大规模 Python 函数执行的内存快照, 从而显式建模变量状态变化; 在 Web 决

策这一不可逆场景中，WMA^[124] 在推理阶段引入世界模型进行 Look-ahead Simulation，通过预测网页状态变化评估动作价值，避免陷入死胡同；针对生成任务缺乏客观评价标准的问题，RLVR-World^[125] 提出基于可验证奖励的训练框架，利用逻辑谜题解、感知质量等确定性环境反馈，直接优化世界模型的生成与预测能力。

未来展望

前面回顾了 2025 年大语言模型强化学习的主要进展，这些进展不断突破着大语言模型的性能上限。接下来，我们将对大语言模型强化学习的后训练发展做出预测。

作为新兴技术，大语言模型强化学习目前尚缺乏体系，并且在许多细节领域存在研究空白，因此未来可能有大量的研究工作聚焦于体系构建和技术探索领域。首先是体系，尽管大语言模型发展较晚，但是传统强化学习从上世纪 50 年代就诞生了，即使是深度强化学习也发展了 10 几年的时间，早已形成了成熟学科体系。这使得大量潜在可应用到大语言模型的强化学习算法还未尝得到实验，以及这些算法支撑起的完整体系也存在着迁移或部分应用到大语言模型上的潜力，其中也包括围绕虚拟环境与世界模型的交互式强化学习体系构建。其次是研究领域空白，例如可显著降低训练成本的离线 RLVR 等。此外，大语言模型强化学习尚未和 AlphaGo、深蓝等传统强化学习项目一样，取得超过人类上限水平的成就，这使得大语言模型强化学习依然有足够广阔的研究空间。

2.2 数据获取与数据治理

数据作为大模型训练的核心基础，其获取质量与治理水平直接决定模型性能、泛化能力及安全性。本章围绕大模型数据全生命周期中的关键环节，系统梳理了开源训练数据集构建、数据处理、以及多模态数据集的最新研究成果，旨在清晰呈现当前大模型数据领域的研究现状与技术进展，为相关研究与应用提供参考。

2.2.1 开源数据集构建

研究背景 训练数据集是大模型发展的核心支撑，其规模、质量与任务适配性直接决定模型的通用能力与专项性能边界。在 2025 年之前，大语言模型

的数据集构建主要聚焦预训练与后训练：在预训练数据集的构建上，数据规模量级始终没有得到很好的突破；在后训练数据集的构建上，与推理相关的数据集数量较少且质量较差。同时，现有的训练阶段划分相对模糊，缺乏针对“预训练后、微调前”的过渡性数据支撑方案。

研究进展 针对上述局限，2025 年大模型的开源数据集构建发展呈现三条主线：一是更加关注预训练数据集的数据规模；二是重新拾起中训练的概念，各大头部 AI 机构组织积极布局；三是更加关注与推理相关的后训练数据集。

预训练数据集：在 2025 年的技术演进中，面向大语言模型的预训练数据集，其核心发展特征体现为数据规模的持续扩张。如在数学领域，Nemotron-CC-Math 数据集^[126]被提出，该数据集规模高达 1300 亿个 token。Nemotron-CC-Math 数据集基于 Common Crawl 构建，采用布局感知渲染与大语言模型清洗流程，精准提取 MathJax、KaTeX 等格式数学内容，标准化为 LaTeX 格式，去除冗余 boilerplate 内容，保留公式结构完整性。MegaMath 数据集^[127]被提出，该数据集由 MBZUAI 构建，从 Common Crawl 中提取数学文档，是当前规模最大的开源数学预训练数据集，适配数学大语言模型续训练场景。SwallowMath 数据集^[128]被提出，该数据集基于 FineMath-4+ 优化，通过去冗余、恢复上下文、步骤化重写等流程提升数据质量，适配数学预训练场景。如在推理领域，MobileLLM-R1 数据集^[129]被提出，该数据集规模高达 2 万亿个 token。该数据集是针对小参数模型（<10 亿参）设计的推理预训练数据集，通过自定义指标筛选开源数据并重采样，无需超大规模语料即可激发推理能力。MegaScience 数据集^[130]被提出，该数据集由上海交大 GAIR Lab 构建，聚焦多学科科学推理预训练，数据源为 12.8 万本大学教材，覆盖多领域，经去污染、大语言模型精炼，含高质量参考答案与推理步骤，适配通用推理大语言模型预训练。NaturalReasoning 数据集^[131]被提出，该数据集由 Meta 与纽约大学联合发布，从预训练语料经回译提取真实推理问题，覆盖数学、物理、计算机科学等多领域，结合可验证与开放式问题，适配不同规模模型推理预训练。如在代码领域，CodeScale-Corpus 数据集^[132]被提出，该数据集规模高达 1 万亿个 token，且包含 Python、Java、JavaScript 等 7 种主流语言。该数据集由北航、人大联合构建，打破了代码数据同质化惯性思维，针对 7 种主流语言的语法特性、应用场景设计差异化筛选策略，建立语言区分的 Scaling Laws，提供最优数据配比方案，减少算力浪费，提升多语言代码模型性能。rStar-Coder 数据集^[133]被提出，该数据集由微软亚洲研究院构建，竞赛平台收集种子问题并合成新问题，解决方案

经多难度测试案例验证，适合大语言模型预训练，提升大语言模型代码能力。Seed-Coder 数据集^[134]被提出，该数据集由字节跳动发布，模型驱动数据构建 pipeline，大语言模型自动评分筛选代码数据，最小化人工参与，适配代码大语言模型预训练，同规模开源模型中表现最优。我们收集了上述各个领域的其他预训练数据集，如表 2.3所示。

表 2.3: 2025 年大语言模型预训练数据集汇总

数据集名称	数据集规模	所属领域	是否开源
Nemotron-CC-Math ^[126]	1330 亿 token	数学	是
MegaMath ^[127]	3710 亿 token	数学	是
SwallowMath ^[128]	230 亿 token	数学	是
MobileLLM-R1 ^[129]	2 万亿 token	推理	是
MegaScience ^[130]	620 亿 token	推理	是
NaturalReasoning ^[131]	1.2 万亿 token	推理	是
CodeScale-Corpus ^[132]	1 万亿 token	代码	是
rStar-Coder ^[133]	180 亿 token	代码	是
Seed-Coder ^[134]	950 亿 token	代码	是

中训练数据集: 2025 年, 中训练这个概念不断在诸如 Qwen3, Kimi K1.5 等性能强劲的大语言模型的技术报告中被提及。中训练的研究进展整体呈现出: 工业界率先积极布局, 学术界刚刚起步的局面。中训练是大语言模型训练流程中连接预训练与后训练的关键独立阶段, 区别于预训练构建通用基础能力和后训练专项对齐 (SFT/RL) 的定位, 以中等数据规模 + 中等计算成本为特征, 通过通用数据保留 + 专项数据注入的混合策略, 在不丢失预训练通用能力的前提下, 系统性增强模型的特定技能。其核心差异于持续预训练: 中训练会刻意保留预训练数据比例 (通常 70%-75%)、继承优化器状态, 避免“灾难性遗忘”; 而持续预训练多直接用领域数据扩展, 易导致通用能力退化。中训练数据的组成通常遵循 70%-75% 高质量通用数据 + 25%-30% 专项数据的混合逻辑, 涵盖诸如数学, 代码, 指令跟随等等领域, 2025 年发布的大语言模型提及到的中训练数据集的构成如 2.4所示 (该表格参考了相关综述文献^[135])。

后训练数据集: 2025 年, 大语言模型后训练数据集的研究更加注重推理数据集的构建。如在数学推理领域, 如 DeepMath-103K 数据集^[143]被提出, 该数据集规模高达 103K 个样本。该数据集经严格去污染处理 (无

表 2.4: 2025 年大语言模型中训练数据集汇总

模型名称	训练数据集所覆盖的领域
LongCat-Flash-Thinking ^[136]	通用，数学、推理、指令，代码，智能体，长文本
LongCat-Flash ^[137]	通用，数学、指令，代码，长文本
dots.llm1 ^[138]	通用，数学、推理，代码，长文本，多语言
MiMo ^[139]	通用，数学、推理，代码，长文本
Qwen3 ^[140]	通用，数学、推理、指令，代码，长文本，多语言
Pangu Ultra ^[141]	通用，数学、推理、指令，代码，长文本，多语言
Kimi K1.5 ^[142]	通用，数学、推理、指令，代码，长文本，多语言

GSM8K/MATH 等基准重叠), 聚焦难度 5-9 级问题; 每条样本含 DeepSeek-R1 生成的 3 条不同推理路径, 支持 RL 奖励信号构建。AoPS-Instruct 数据集^[144]被提出, 其通过自动化管道从 AoPS 论坛挖掘竞赛题与社区解答, 经 Qwen2.5-72B 清洗并生成分步解释。使用该数据集微调的模型在 OlympiadBench/Omni-MATH 的指标显著提升。Nemotron-Math 数据集^[126]被提出, 其由 NVIDIA 发布, 含高/中/低三种推理深度; 适配超长上下文训练, 适合提升模型在竞赛级数学推理任务上的性能。如在科学推理领域, 如 PHYSICS 数据集^[145]被提出, 该数据集规模高达 16K 个样本。该数据集覆盖力学、电磁学等 5 大物理领域, 难度从高中到研究生; 从 100+ 教材精选 8,284 题并双语翻译, 每条训练样本含 DeepSeek-R1 生成的细粒度推理路径。SciReasoner-Instruct 数据集^[146]被提出, 该数据集是一个跨学科科学推理指令微调数据集, 覆盖文本-科学格式转换、知识提取、属性预测等 103 项任务; 采用统一输入输出模式, 适配跨学科 SFT 与长思维链推理训练, 提升模型多领域科学问题解决能力。Llama-Nemotron 数据集^[147]被提出, 该数据集含 CoT 推理链与符号计算标注; 经模型/人工双重校验, 噪声率 <0.5%; 支持科学计算与跨学科推理训练。如在代码推理领域, 如 CODE I/O++ 数据集^[148]被提出, 该数据集规模高达 300w+ 代码推理实例。该数据集由 DeepSeek 团队构建, 通过代码执行轨迹合成自然语言思维链。该数据集引入验证与修订机制提升质量, 支持模型从代码逻辑迁移到数学、科学等跨领域推理, 适配 SFT 与推理链增强训练。OpenCodeInstruct 数据集^[149]被提出, 该数据集是一个大规模代码推理微调数据集, 覆盖 Python/C++/Java 等主流语言; 经多轮去污染与质量筛选, 每个样本含可执行测试用例与大语言模型质量评分, 适配 SFT 与推理链增强训练。OpenCodeReasoning 数据

集^[150]被提出，该数据集由 NVIDIA 构建，是一个竞赛级代码推理训练集，整合 11 个平台（CodeForces/LeetCode 等）数据，用 DeepSeek-R1 生成带完整推理轨迹的解决方案；经去污染与执行验证，适配复杂算法推理与长链逻辑训练。上述后训练数据集的汇总，如表 2.5所示。

表 2.5: 2025 年大语言模型后训练数据集汇总

数据集名称	数据集规模	所属领域	是否开源
DeepMath-103K ^[143]	103K 样本	数学推理	是
AoPS-Instruct ^[144]	60 万样本	数学推理	是
Nemotron-Math ^[126]	750 万样本	数学推理	是
PHYSICS ^[145]	16K 样本	科学推理	是
SciReasoner-Instruct ^[146]	4000 万样本	科学推理	是
Llama-Nemotron ^[147]	320 万样本	科学推理	是
CODE I/O++ ^[148]	300 万 + 样本	代码推理	是
OpenCodeInstruct ^[149]	500 万样本	代码推理	是
OpenCodeReasoning ^[150]	73.6 万样本	代码推理	是

未来展望 展望未来，大模型开源训练数据集构建将围绕前文三大发展主线深化推进。在预训练层面，数据规模的突破将持续向更细分的语言与领域延伸，未来研究重点将聚焦低资源语言数据的高效挖掘与质量提升，通过跨语种知识迁移、原生语料精准采集等技术，填补小语种及弱势语言的数据空白。在中训练层面，随着工业界实践的不断深入，学术研究与产业应用的鸿沟将逐步缩小，未来有望形成标准化的中训练数据构建规范，包括数据质量评估指标、“预热”任务设计范式等。在后训练层面，专项领域适配将向更精细化、专业化方向发展，除现有网络安全、推理、代码生成领域外，将拓展至高端制造、生物医药等垂直领域。

2.2.2 数据处理技术

研究背景 大模型能力边界由训练数据质量、规模等核心要素决定，当前训练数据已迈入万亿级 token 门槛，高质量数据集建设成为 AI 发展的战略任务。2025 年前，数据处理技术存在明显局限：去重多聚焦文本表层匹配，未关注动态语义冗余；质量过滤依赖人工与静态规则，效率低下；有害性识别难以捕捉细粒度信息，且数据配比多凭经验调整等等。这些问题导致模型训练存在冗余、低质风险，制约大模型性能提升。

研究进展 针对上述局限，2025 年的数据处理技术包括数据去重处理，数据质量过滤，数据有害性过滤，数据文本重述和优化，数据配比优化策略，以及数据合成技术有了新的发展趋势（如图 2.2所示），具体如下所示：

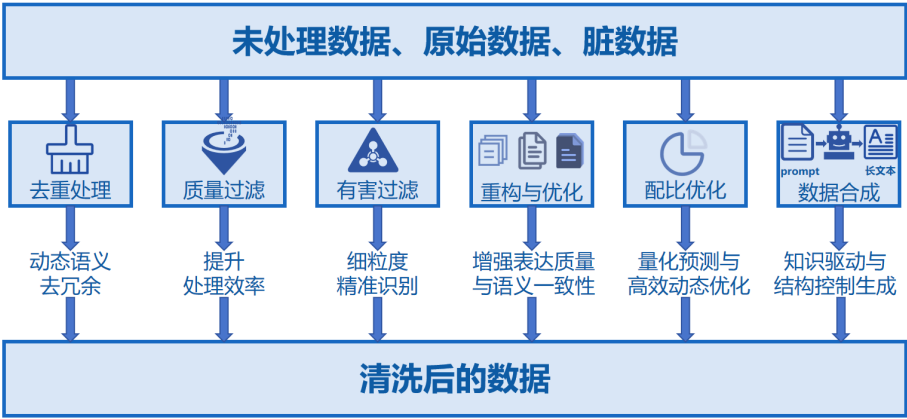


图 2.2: 2025 年数据处理技术的新趋势

数据去重处理：在大语言模型的数据集去重处理上，2025 年的发展整体上呈现出更加关注动态语义去冗余的趋势。主要研究如下：研究者提出了 GneissWeb 流水线，聚焦动态语义去冗余核心需求，创新性采用分片精确子串去重策略，通过精准捕捉文本语义关联实现动态冗余识别与过滤，成功构建了规模达 10 万亿 token 的高质量无冗余数据集，其数据纯度、语义一致性及下游任务适配性能均显著优于 FineWeb v1.1.0^[151]。在 Ungoliant 流水线中，研究者以动态语义去冗余为核心目标，结合 MinHash/LSH 算法的高效相似性匹配能力，针对不同语种文本分片执行精准去重操作，同时引入 KenLM 困惑度模型进行二次语义校验，进一步剔除语义重复、语义冲突的低质量语料，为 OpenGPT-X 模型训练提供了兼具语义完整性与低冗余特性的优质语料^[152]。针对长序列训练场景下的语义冗余与模型遗忘问题，研究者重点围绕动态语义去冗余展开优化，创新性融合样本级去重与属性洗牌技术，通过动态挖掘长序列样本中的语义重复信息并进行针对性去重处理，有效缓解了语言模型在持续学习环境下因语义冗余导致的虚假遗忘问题，提升了模型持续学习的稳定性与语义建模能力^[153]。

数据质量过滤：在大语言模型的数据集质量过滤上，2025 年的发展整体上呈现出更加关注质量过滤方法的高效性的趋势。主要研究如下：从谱动力学视角出发，研究者提出创新数据治理方法，其核心聚焦于大规模训练数据

预处理的高效化突破。该方法通过动态加速策略，结合谱动力学量化分析手段精准捕捉数据质量特征，无需依赖复杂的人工干预或静态规则，即可实现海量数据的实时预处理，大幅提升了数据治理的效率，有效解决了传统方法在大规模场景下处理速度慢、资源消耗高的问题^[154]。提出 Ultra-FineWeb 数据过滤流程，以过滤流程高效化为核心目标，创新性引入 fastText 分类器并量化数据“预测强度”作为质量判定依据。该流程通过自动化量化评估替代低效的人工验证环节，不仅省去了人工逐样本审核的时间成本，还实现了数据质量的快速分级筛选，大幅提升了整体过滤效率，同时降低了人力投入，推动数据过滤从“人工主导”向“量化驱动”的高效转型^[155]。针对轻量化场景下的高效筛选需求，研究者设计基于评分器的样本级过滤机制，核心在于通过标准化量化评分体系实现高效化筛选。该机制无需复杂计算即可完成样本质量的快速精准评估，在严格把控数据质量的同时，最大限度保留领域多样性，避免了过度过滤带来的资源浪费，以轻量化设计提升了筛选流程的整体效率，满足了资源受限场景下的高效数据治理需求^[156]。面向长上下文数据生成的多目标训练场景，研究者提出基于提示交互的质量过滤模块，核心目标是实现多训练目标下的协同高效过滤。该模块通过提示交互生成统一的量化质量评估维度，无需为 SFT、DPO 等不同训练目标单独设计过滤流程，避免了重复工作，显著提升了多场景下的过滤效率，确保数据可快速适配各类训练需求，实现长上下文数据过滤的高效化与场景化统一^[157]。

数据有害性过滤：2025 年，大语言模型数据集有害性过滤技术的发展呈现新趋势：相较于传统粗粒度过滤（比如样本级），行业更关注细粒度识别方法的突破（如 token 级）。主要研究如下：研究者提出名为 IF-Guide 影响函数，聚焦细粒度有害数据识别，通过改进的影响函数实现 token 级有害训练数据的精准定位，可捕获显性与隐性毒性信息，在预训练阶段主动抑制模型毒性传播，且无需依赖偏好对齐数据，为细粒度数据 detox 提供了高效方案^[158]。针对 Web 数据集有害内容的细粒度识别需求，研究者提出 HarmFormer 模型用于有害内容分类，通过构建 TTP 提示集与多维度 HAVOC 毒性基准，覆盖多类型细微有害信息，突破传统短文本检测局限，实现对 Web 长文本中细粒度、高精度的有害内容识别与过滤^[159]。为在细粒度过滤敏感有害信息的同时避免模型分布偏移，研究者提出基于因子解相关技术的敏感数据安全过滤方案，通过特征维度的细粒度解相关与权重迭代更新，精准移除有害信息的同时保留有效特征，保障过滤后数据分布稳定性^[160]。

数据文本重述与优化：2025 年，在大语言模型数据集的数据文本重述与优化领域，相关研究呈现出更加关注增强表达质量与语义一致性的趋势。主

要研究如下：针对监管文档存在的内容冗余、语义表述分散等问题，研究者提出面向监管文档的 RAG 优化预处理流程。该策略聚焦增强文本表达质量与语义一致性，通过重复内容识别实现冗余信息的精准去重与整合，再结合基于召回率的文本重排机制梳理语义逻辑，有效提升 RAG 系统的检索准确性^[161]。在金融语料库领域，研究者设计了关键词提取与扩展、表格标注及 Markdown 重构一体化策略，重点优化文本表达规范性与语义对齐度，进一步增强金融 RAG 系统的语义一致性，适配金融文本的专业场景需求^[162]。研究者提出动态质量过滤框架，基于在线困惑度评分对输入文本进行实时语义评估与重述优化，在提升文本表达质量、保障语义一致性的同时，有效平衡了生成质量与计算延迟^[163]。

数据配比优化策略：大语言模型的数据配比优化策略领域主要呈现出的趋势是：迈向量化预测与高效的优化。主要研究如下：研究者将数据配比优化建模为双层优化问题，提出孪生模型框架（代理模型 + 参考模型），通过模型损失差异动态调整各领域数据权重，在数据受限和 SFT 场景下实现性能显著提升^[164]。研究者基于缩放定律提出系统方法，通过少量小规模训练实验预测不同数据配比下的模型损失，为多类型模型预训练提供最优领域数据权重分配方案^[165]。研究者将数据混合问题建模为回归任务，提出 RegMix 方法，通过训练 512 个 1M 参数的小模型预测大规模模型的最优数据混合比例^[166]。研究者进一步统一现有数据混合方法的优化框架，提出在线动态调整配比的 Aioli 方法，直接估计混合定律参数，在 6 个数据集上平均提升 0.27 测试困惑度^[167]，实现动态优化的高效落地。研究者聚焦代码与数学专业任务场景，构建 Swallow Code 和 Swallow Math 数据集，通过量化实验验证提出“代码-数学数据 10:1”的固定训练配比策略，针对性提升模型专业推理能力^[128]。

数据合成技术：在大语言模型的数据合成技术领域，2025 年将研究重点从模块化合成转移到了知识驱动与结构控制生成上来。主要研究如下：在模板化生成向知识驱动过渡的阶段，研究者提出 Token-Level Editing 技术，摒弃传统固定模板全量生成模式，通过细粒度文本编辑实现数据合成，依托模型对文本语义的知识理解完成精准编辑，而非机械套用模板，有效避免了模型在自循环训练中的退化与崩溃问题^[168]。进一步强化知识驱动的闭环优化能力，研究者设计“迭代式引导生成”数据合成机制，以学生模型的知识认知为引导，驱动教师模型生成训练数据，通过动态反馈链路将生成数据的质量评估结果反哺至生成过程，实现知识的精准传递与数据质量的持续优化，在微调阶段显著提升了数据利用效率，标志着数据合成从“被动模板填充”向

“主动知识驱动”的关键跨越^[169]。在知识驱动基础上，聚焦长文本场景的结构控制需求，研究者提出长上下文数据合成策略，通过提示交互机制挖掘文本内在结构知识，结合分层模块化生成框架，实现对超长文本结构的精准把控，将大模型的上下文窗口扩展至百万词元级别，使有监督微调与 DPO 协同训练中可高效处理极长文本的结构逻辑，完成从“知识驱动生成”到“结构可控生成”的进阶^[170]。面向跨语言场景的结构与知识协同控制，研究者提出 XL-Instruct 跨语言数据合成框架，结合 XL-Alpaca Eval 基准，不仅融入多语种知识图谱与语义迁移知识，更实现对跨语言文本指令结构、语义逻辑的精准控制，有效强化了多语言大模型的跨语种迁移能力与指令执行一致性，完善了知识驱动与结构控制融合的技术体系^[171]。

未来展望 展望未来，基于 2025 年数据处理技术新趋势，未来将围绕现有技术方向深化突破，避免脱离当前发展框架，重点聚焦三方面推进：1) 深化动态语义与多模态协同处理，延续 2025 年动态语义去重核心方向，拓展至多模态场景，构建文本、图像等跨模态语义对齐的去重体系。结合现有高效过滤技术，研发多模态轻量化处理方案，在保障质量的同时降低算力消耗，适配多模态大模型训练需求；2) 推进量化驱动与智能适配升级，基于 2025 年量化驱动过滤、数据配比量化预测技术，构建“数据-模型”闭环适配机制。通过智能算法自动优化去重阈值、过滤规则与数据配比，实现处理策略随模型需求动态调整，降低人工干预，提升技术落地效率；3) 强化细粒度安全与合成技术落地，深化 2025 年细粒度有害性识别技术，提升显性/隐性有害信息的精准过滤能力，保障数据分布稳定性。依托知识驱动型数据合成技术，拓展至垂直领域，生成高保真专业数据，缓解领域数据稀缺问题，推动技术在行业场景落地。

2.2.3 多模态数据集构建

研究背景 多模态大模型的演进高度依赖于跨模态数据的规模与对齐质量，数据的品质与多样性直接决定了模型跨模态理解、推理及交互能力的上限。2025 年之前，多模态预训练数据集主要依赖于大规模互联网图文对抓取（如 LAION 系列、COCO 等），该阶段数据集构建以规模扩张为核心目标，通过海量样本堆砌推动模型基础能力提升。而多模态后训练数据集则关注通用泛化性，没有明显地针对特定领域进行适配。随着多模态技术向产业落地与复杂场景延伸，早期数据集的固有瓶颈日益凸显，已难以支撑模型向高阶智能跨越。

研究进展 针对上述局限性，在 2025 年，多模态预训练数据集研究整体呈现从“简单的规模扩张”向“跨模态语义关联的规模化构建”的趋势，多模态后训练数据集研究整体呈现从“通用泛化”向“专项任务适配”的深度转型。2025 年主要被提出的多模态训练数据集如表 2.6所示。

表 2.6: 2025 年主要被提出的多模态训练数据集

数据集名称	数据集规模	数据集类型	所属领域	包含的模态	是否开源
BigDocs ^[172]	750 万个样本	预训练	文档处理	图像、文本	是
InfiniHumanData ^[173]	11.1 万个样本	预训练	3D 人体生成与控制	3D 参数、图像、文本	是
MedTrinity-25M ^[174]	2500 万个样本	预训练	医疗	图像、文本	是
PIN-200M ^[175]	200M 个样本	预训练	文档处理	图像、文本	是
NautData ^[176]	145 万个样本	预训练	水下探测	图像、文本	是
VideoMind ^[177]	10.3 万个样本	预训练	视频认知理解	视频、音频、文本	是
BLIP3o-60k ^[178]	60K 个样本	后训练	图像理解与生成	图像、文本	是
Math-VR ^[179]	17.8 万个样本	后训练	数学视觉推理	图像、文本	是
Situat3DChange ^[180]	12.1 万个样本	后训练	3D 场景变化理解	3D 数据、文本	是

多模态预训练数据集: 在多模态预训练数据集构建领域，2025 年的新发展趋势是在关注数据规模的前提下，也更加关注数据的跨模态语义关联程度上。主要研究如下：研究者构建了名为 BigDocs 的大规模数据集，其包含 750 万个样本，覆盖 30 个文档与代码相关任务，包含 GUI 推理等创新任务，聚焦文档处理专业领域，有效填补了文档跨模态数据空白，使文档基准测试性能提升 15.14%^[172]。研究者通过集成多视角 RGB 图像与 SMPL 人体参数，提出了 InfiniHumanData 数据集，其中包含 11.1 万个 3D 人体身份样本，实现 3D 模态与文本信息的深度融合，为 3D 人体生成与控制任务提供了坚实支撑^[173]。MedTrinity-25M 医疗数据集被提出，其规模达 2500 万，涵盖 10 种模态医学图像及 65 种疾病的多粒度标注，聚焦医疗专业领域，通过多模态、多粒度标注提升语义对齐精度，支持从医疗报告生成到病灶分割的全流程任务^[174]。研究者通过融合 Markdown 文本与布局图像，提出了 PIN 系列文档数据集，其中包含 PIN-200M 和 PIN-14M，深耕文档处理领域，有效强化了模型的中英文跨模态知识融合能力^[175]。为弥补水下数据集的缺少，研究者提出 NautData 水下数据集，构建了 145 万图文对，覆盖 8 种水下理解任务，聚焦水下探测专业场景，使模型在低光及浑浊环境下的性能超越了通用模型 LLaVA-1.5^[176]。VideoMind 视频数据集被提出，包含 10.3 万样本，通过标注视频中的音频、文本及意图信息，填补了视频时序模态数据空白，实现音视频与文本的多模态融合，支撑模型实现深度的视觉认知理解^[177]。

多模态后训练数据集：在多模态后训练数据集构建领域，2025 年的发展趋势整体呈现出由通用泛化转向专项任务适配的趋势。主要研究如下：在图像理解与图像生成领域，研究者提出 BLIP3o-60k 图像生成集，通过 GPT-4o 生成高质量图像描述，使统一多模态模型 BLIP3o 能够兼顾图像理解与高质量图像生成的双重能力^[178]。在数学视觉推理领域，Math-VR 双语数学推理数据集被提出，其包含 17.8 万个数学问题样本，配套图像到代码转换器，显著提升了模型在数学视觉推理基准上的表现^[179]。在 3D 场景变化理解任务，Situat3DChange 理解集被提出，其中包含 12.1 万个问答对，集成人类视角的 3D 环境观察，显著提升了模型对于 3D 场景的理解能力^[180]。

未来展望 基于 2025 年多模态数据集的发展现状，该领域未来将向全感官统合，具身交互深化，数据生态完善三大方向迈进，支撑多模态大模型向高阶智能演进。全感官统合方面，在现有图文、音视频、3D、医疗等模态融合基础上，延伸至触觉、力觉、嗅觉等感官模态，结合传感器时序数据构建全模态预训练库，实现多感官数据深度协同与语义对齐，破解单一模态局限。具身交互深化方面，未来需要构建真实场景具身智能交互数据集，通过真机设备采集高精度操作数据、模拟器生成合成轨迹，覆盖“观察-决策-执行”全链条，推动模型从被动识别向主动决策和灵活执行转变；同时建立跨本体数据标准，解决数据异构问题。数据生态完善方面，构建专业深耕 + 跨域融合体系，深耕医疗、工业等垂直领域，推动多领域模态数据协同；搭建数据共享平台与合规保障体系，以贡献即获益机制推动数据流通共享，为模型演进提供高质量数据支撑。

2.3 模型能力提升

大语言模型正从通用对话系统，演化为具备复杂推理能力、能够自主与环境进行长序列交互的高级智能体，这需要针对各种能力对模型进行专项训练。本章将系统梳理使大模型获得如长上下文、复杂推理、数学/代码、工具调用等专项能力的训练方法及其最新进展，以清晰呈现该领域的发展脉络。

2.3.1 长上下文

研究背景

如果将大语言模型本身看作是 CPU，那么其上下文窗口可以看作是 CPU 的内存，更大上下文窗口代表着大模型能够处理更多的数据与信息，也代表了更高的智能水平。Epoch AI 在今年指出^[181]，大语言模型的最长上下文窗口相比于去年扩展了约 30 倍，而模型高效利用长上下文的能力的提升速度更快，在两大长文本测试基准上能维持不低于 80% 准确率对应的最长输入长度，在过去 9 个月内增幅超过 250 倍。这体现出当今大语言模型的长文本能力正在飞速进步。

为了将大模型的上下文窗口从 4K 扩展到几十万 token 甚至是上百万 token 级别，当前的普遍共识是这一方面需要通过位置编码扩展和在长语料库上（约 10B Token）进行持续预训练以扩展上下文窗口，另外需要进行长上下文对齐后训练，增进大模型在长输入下遵循复杂指令的能力。需要注意的是，序列长度增加将导致模型计算成本急剧上升，因此提升大模型长文本能力是一个系统工程，需要从数据工程，模型架构，以及更适合长上下文的训练方法三方面入手，由于后两者在之前的章节已经提到，本节将重点关注**数据工程**方面的进展情况。

在构建长文本数据方面，关键瓶颈之一在于高质量长上下文数据的稀缺性。在**持续预训练**阶段，高质量的，拥有较好的长程依赖特性的长文档训练数据较为稀缺。在**后训练阶段**，长上下文对齐数据的人工标注是成本极高且不可靠的，在 LongBench-V2 等基准测试上的实验结果显示，人类在超长多选推理任务中的准确率可降至 25.1%，接近随机猜测的水平。

在 2025 年之前已有部分工作如 Longskywork^[182]证明，合成数据也可有效增强大模型的长文本能力。在持续预训练阶段，构建具有长距离依赖关系的文档，可以在相互依赖的文本片段之间插入额外的文本，将短距离依赖转化为长距离依赖。在后训练阶段，也展开了初步探索。研究人员一方面尝试基于已有的长文本模型，采用类似于 Self-Instruct 这类在通用领域上合成指令微调数据的方法生成长文本指令数据，如 LongAlign^[183]等；此外，如 FilM^[184]等方法也探索了采用通用模型，结合“文本分块”或“多级摘要”的方法，为事实抽取、摘要等特定长文档任务构造训练数据。

在已有的研究方向上，2025 年的研究趋势是，进一步聚焦于对训练的不同阶段合成更有效的训练数据，以提升模型长文本性能。此外，研究者们也提出更具挑战性的模型长文本能力评测集，可以更有效评测模型在真实应用

场景下的长文本能力。

表 2.7: 2025 年长上下文能力提升典型工作总结

类别	典型工作与核心策略	主要优势
持续预训练	NExtLong ^[185] : 引入对比学习中的困难负样本的思路进行数据合成; Prolong ^[186] : 探索长短数据配比, 发现书籍代码等优质长文本可激活长窗口。	减轻内插方法在合成长程依赖数据时引入的噪声对模型训练稳定性的影响
指令数据合成	LLM×MapReduce ^[187] : 利用分治法构建金字塔式文本层级; GATEAU ^[188] : 利用同源模型引导与上下文感知度量筛选高价值样本。	利用通用模型即可合成覆盖多区间上下文的问答对
复杂推理构建	MIMG ^[189] : 通过单跳生成与多跳融合, 合成多步推理数据;	突破简单检索, 聚焦提升模型的长序列多跳推理性能
RL 自我演进	SPELL ^[190] : 通过提问-回答-验证的自我博弈循环进行提升; LongReward ^[191] / LongPO ^[192] : 利用多维评分或自生成偏好数据进行 DPO 优化。	避免 SFT 数据合成过程中可能出现的噪声, 并保持短文本能力不退化

研究进展

本部分将对 2025 年长上下文领域具有影响力的研究进行归纳。如表所示, 当前的突破主要集中在三个维度: **高质量数据合成**、**自我演进的强化学习**以及**面向真实场景的评测基准**。其中, 高质量数据合成又分为在持续预训练阶段和后训练指令微调阶段。

持续预训练阶段 在持续预训练数据合成方面, NExtLong^[185]认为方法“有效”意味着, 预训练数据应具备长程依赖特性, 同时尽可能地减少不相关上下文对模型学习长程依赖关系的干扰。其探索通过引入对比学习中困难负样本技术, 引入难以区分的负样本来增强模型从干扰项中辨别相关样本的能力。还有部分工作着力于探索持续预训练过程中的长文本和短文本的配比,

如 Prolong^[186]，该工作发现预训练需混合 40% 高质量短数据，而微调阶段仅需短指令即可激活长窗口能力。

后训练指令微调 更多的研究在后训练指令微调数据方面爆发。研究者们发现，优质后训练数据应当满足以下特点：回答问题所需的上下文**区间多样**、应当贴合人类用户实际的对话上下文的提问方式、问题应当具备有效的**长程依赖**。而构建这些优质数据今年又涌现出了大量的方法。部分工作致力于合成具有全局依赖的数据，如 LLM×MapReduce^[187]和 CLIPPER^[193]，前者利用分治法构建金字塔式的文本层级，自底向上合成兼顾局部细节与全局依赖的问答数据，后者采用“先压缩为大纲-再生成复杂声明”的策略，避免直接处理长文带来的噪音；还有部分工作讨论了关于数据筛选的方法，如 GATEAU^[188]利用同源模型引导和上下文感知度量筛选出最具影响力的样本；而部分工作也利用已有的长文本模型自身能力，无需基于已有的种子数据或者预定义模版，生成了更具多样性的数据，如 LongMagpie^[194]；值得注意的是，还有部分工作将目标转向更具挑战性的多跳推理问题数据合成上，如 MIMG^[189]，通过单跳生成、多问题采样、多跳融合、质量验证步骤，合成了高质量的多跳数据集。

基于强化学习的对齐 此外，2025 年呈现出一条新的研究趋势，即引入强化学习方法，使模型自举式的逐步获得长序列推理能力。如 SPELL^[190]证明模型可以通过自循环地动态扮演提问者、回答者和验证者，通过自我博弈强化学习进行自我提升；LongReward^[191]用高精度的 LLM 作为评估器，从帮助性、逻辑性、忠实性和完整性四个角度评分，采用离线强化学习算法 DPO 使得模型提高了长文本任务性能；LongPO^[192]利用大语言模型自己生成的从短到长偏好数据，在保持大模型短文本推理能力并提升长文本推理能力有显著效果。

模型评测 在长文本能力评测方面，如表 2.8 所示，2025 年的研究趋势为，从基础的“大海捞针”式简单检索，转向更侧重评估模型的全局推理能力、指令遵循稳定性以及在真实复杂场景下的鲁棒性。

具体而言，当前的研究进展可以归纳为以下四个维度：首先，在**深层逻辑推理与归属能力**方面，研究者认为简单的信息检索已不足以衡量当前的先进模型。LongBench v2^[195]针对“大海捞针”任务过于简单的问题，通过“对抗式筛选”构建了 503 道高难度多项选择题，其结果揭示了推理时计算对长文

表 2.8: 2025 年主流长上下文评测基准分类对比

评测维度	关键特性与评价重点	代表工作
深层推理与归属	聚焦对抗性高难度题目、长程逻辑推理、信息引用的准确性及其来源归属。	LongBench v2, Ref-Long
真实场景与鲁棒性	引入办公场景噪声（ASR）、真实软件仓库复杂依赖、含噪及口语化文本处理。	ELITR-Bench, LONGCODEU
指令遵循稳定性	压力测试模型在超长序列下维持复杂指令格式、逻辑和参数执行的稳定性。	LIFBENCH
评测效率优化	通过信息压缩、聚类等统计学方法降低超长文本评测的显存与时间成本。	MiniLongBench

理解的关键作用。RefLong^[196] 则要求模型在检索信息的同时准确指出引用的文档编号，实验发现即便是顶级模型在处理 75k token 以上长度时，依然面临“看得见却分不清”的问题。其次，在**真实应用场景的模拟**上，评测基准开始引入真实环境数据。ELITR-Bench^[197] 聚焦于会议助手场景，不仅包含人工标注的复杂问答，还引入了自动语音识别以注入噪声。在软件工程领域，LONGCODEU^[198] 从真实的 GitHub 仓库中构建任务，涵盖了代码单元间的依赖关系理解，研究发现模型处理超过 32k 长度的代码时，逻辑理解能力将显著下降。第三，在指令遵循的**稳定性**维度，LIFBENCH^[199] 填补了长文本下指令遵循评测的空白。它不仅关注任务的正确性，还提出了指令遵循稳定性（IFS）指标。最后，针对长文本评测成本高的问题，MiniLongBench^[200] 展示了高效评测的可能性。该工作利用逻辑回归与聚类算法，从数千个样本中精选出 237 个最具代表性的样本，将评测成本压缩至原来的 4.5%，且保持了与原始长基准极高的排名相关性。

未来展望

前面回顾总结了 2025 年长上下文处理在数据工程、强化学习对齐以及评测基准维度的核心进展。这些工作共同推动了长上下文技术从单纯的“窗口扩容”转向对“逻辑深度”的本质挖掘。接下来，我们对长上下文领域的未来发展趋势进行展望。展望 2026 年，长上下文技术的演进重心将从单纯的上下文窗口长度扩展（Context Window Extension），转向探索**推理时计算**在长序列任务中的缩放效应。正如 LongBench v2 所揭示，当前模型虽已

具备处理超长文本的吞吐能力，但在面对通过“对抗性筛选”构建的高难度任务时，其性能瓶颈并非源于信息遗忘，而是源于**长程逻辑推理能力的匮乏**。因此，未来的研究将不再仅仅追求“读得更多”，而是致力于通过增加推理阶段的计算投入，利用类似多步思维链将长文中的隐式依赖关系显性化，从而突破当前模型在超长上下文中深度推理的能力天花板。与此同时，**合成数据与模型自举的闭环进化**将成为解决长文本数据稀缺的主流范式。鉴于人类在超长序列标注上的失效，未来的数据工程将依赖模型自主构建具备多跳依赖关系的复杂指令，并利用**自我博弈（Self-Play）**等强化学习机制，在零人工干预的情况下实现性能的自我迭代与演进。以上是我们关于今年长上下文发展情况的总结。我们希望该领域在未来能够持续蓬勃发展，构建起真正能够理解宏大逻辑、胜任真实复杂任务的长文本智能生态。

2.3.2 推理

研究背景

推理能力被普遍视为衡量大语言模型（LLMs）从“语言生成系统”迈向“通用智能系统”的关键标志^[204]。相较于表层的文本生成，推理能力体现为模型在有限上下文与计算预算下，能否进行多步决策、信息整合与策略选择，并在复杂约束条件下稳定地产生一致、可验证的结论。

早期的大语言模型主要依赖大规模无监督预训练所涌现的隐式推理能力，在简单逻辑与短程推理任务中已展现出一定效果。然而，随着研究逐步转向数学、代码、符号推理以及复杂决策等高难度场景，仅依赖预训练所获得的隐式能力已难以支撑长链条、多分支、高不确定性的推理需求。模型在长推理与多跳推理任务中普遍表现出推理冗余、策略坍缩等问题。

为缓解上述不足，2023 年以来，大量研究开始通过显式推理监督增强模型的推理能力，其中以**监督微调（SFT）**为代表的方法，通过人工或自动方式构造带有中间推理过程的数据，使模型学习标准化的推理范式^[202]。然而，随着推理复杂度提升，该范式逐渐暴露出高质量推理轨迹构造成本高、策略空间覆盖有限等瓶颈。

在此基础上，2023–2024 年间，研究者进一步引入**偏好学习**作为过渡范式，通过构造正确与错误推理路径的对比信号，引导模型在保留推理结构的同时远离不合理的推理模式^[203]。该类方法在一致性、简洁性与可控性方面取得了显著改进，但其训练过程仍以离线偏好数据为核心，本质上属于离线策略对齐，难以直接激励模型探索新的推理策略。

表 2.9: 大语言模型推理训练学习范式演化过程梳理

时间阶段	训练范式	核心思想与能力提升方式	主要局限与演进动机	代表性工作
2022	预训练	依赖大规模无监督预训练中涌现的推理能力，在简单逻辑与短程推理任务中已展现一定效果。	有一定的思考能力，难以支撑长链条、多分支、高不确定性的复杂推理。	CoT ^[201]
2023	监督训练	通过人工或自动构造带中间推理过程的数据，使模型学习标准化、可复现的推理范式，显著提升可解释性与稳定性。	高质量推理轨迹构造成本高；推理策略覆盖有限；能力受限于数据分布。	ReFT ^[202]
2023–2024	偏好学习	通过正确与错误推理路径的对比信号，引导模型在保留合理推理结构的同时远离不一致或冗余推理模式，在一致性与简洁性方面取得改进。	本质仍为离线（off-policy）对齐，难以直接激励模型探索新的推理策略，能力边界受限。	MCTS-DPO ^[203]
2024–2025	强化学习驱动推理	将推理过程建模为可优化的序列决策问题，通过可验证奖励在生成–评估–更新的闭环中直接优化推理策略，摆脱人工轨迹依赖。	训练稳定性以及信念分配存在困难，但显著提升推理能力与自适应性。	DeepSeek-R1 ^[204]
2025	Self-Play驱动推理	在极少甚至零外部干涉的情况下，通过 self-play 机制自主生成问题并完成能力迭代，实现推理能力的持续进化。因此，强化学习本质上也是 self-play 范式的特例。	系统设计复杂、稳定性要求高，但为长期可扩展推理能力提供新范式。	R-Zero ^[205]

正是在上述局限推动下，研究社区开始将推理过程重新建模为一个可优化的序列决策问题，并探索以**强化学习（RL）**为核心的能力驱动训练范式。与模仿学习和偏好学习不同，RL 通过奖励信号直接优化推理策略，使模型能够在生成–评估–更新的闭环中持续调整其推理行为。进入 2025 年，以 DeepSeek-R1^[204] 为代表的工作表明，在无需人工推理轨迹标注的条件下，仅依赖可验证奖励与大规模后训练，即可诱导模型产生自我反思、路径选择与策略自适应等复杂推理行为。

更进一步，self-play 的自进化学习范式的引入，使模型在极少甚至零外部数据条件下，能够自主生成问题并完成能力迭代^[205-206]。总体而言，2023–2025 年 LLM 推理训练范式经历了从模仿学习到能力驱动学习的系统性转变。SFT 与偏好学习提供了稳定可靠的能力基础，而强化学习与 self-play 在合适条件下决定了推理能力的上限与可扩展性。未来推理模型的发展关键，将在于如何围绕能力边界设计分阶段、可演化的训练体系。我们在表 2.9 中

进行了推理范式发展的梳理总结。

研究进展

步入 2025 年后，大语言模型的推理研究已从“是否能够推理”逐步转向“如何更有效、更可靠地推理”。随着模型规模扩大与应用场景复杂化，单纯依赖静态监督数据或固定推理模板已难以满足对推理深度、效率与可控性的综合需求。围绕这一背景，学术界逐渐形成了若干相互关联但侧重点不同的研究方向：一方面，在训练范式上通过强化学习与偏好建模探索更优的推理轨迹；另一方面，在推理阶段关注如何在保证正确性的前提下降低计算开销；同时，工具增强推理与新型推理范式的探索不断拓展 LLM 推理的能力边界。下文将从这四个方面系统梳理近年来 LLM 推理领域的主要研究进展。

训练范式中的推理轨迹探索与利用 在 DeepSeek-R1 公开其技术路线之后，提升大语言模型推理能力的研究范式已明显转向以强化学习（RL）为核心的方法体系。基于强化学习中的两大核心主题——探索（exploration）与利用（exploitation），下文概述 RL 在提升 LLM 推理能力方面的主要进展。

强化学习要求模型在推理状态空间中自主探索并生成训练轨迹。经典 RL 理论表明，充分的探索有助于发现更优策略；对应到推理训练中，能够探索到更丰富的推理轨迹，通常能提升模型在推理多样性、错误鲁棒性与复杂任务泛化方面的能力。因此，增强探索能力被视为提升推理上限的关键方向。现有研究主要通过两类方法增强探索。一类方法侧重于**增强轨迹采样搜索**，通过显式建模推理状态空间获取多样化轨迹。早期工作多依赖 MCTS 等搜索方法构造推理数据，而近期研究如 AEPO^[207] 与 TreePO^[208] 则将树搜索机制直接融入 RL 训练，在轨迹采样过程中于关键状态采样多种动作，从而提升推理路径的结构多样性。尽管该类方法在探索层面直观有效，但仍面临轨迹采样效率与动作空间设计复杂等问题。另一类方法通过**探索激励奖励**引导模型主动探索，包括基于熵、多样性或好奇心的内在奖励。典型方法如 DRA-GRPO^[209] 通过鼓励同一 group 内推理轨迹的语义差异，显式提升探索多样性；另有研究引入基于模型不确定性的内在奖励，引导模型探索尚不确定的推理状态。已有分析指出，当前探索在很大程度上仍受制于预训练知识分布，更多体现为对已有知识的重组，而非实质性扩展^[210]。

在获得推理轨迹之后，如何高效利用这些经验同样至关重要。随着推理训练范式由有监督微调向 RL 过渡，如何将专家数据与在线探索过程结合，成为一个核心问题。一方面，直接基于静态专家数据的离线强化学习往

往难以超越 SFT。为缓解这一问题，LUFFY 通过混合专家数据与在线采样轨迹，以离线策略方式引入专家指导；SRFT^[113] 进一步通过额外的有监督微调损失提升训练稳定性。另一类方法则将专家轨迹作为提示信息，引导模型生成推理过程。UFT^[211]、StepHint^[212] 等工作通过逐步暴露专家轨迹前缀，在降低探索难度的同时缓解分布偏差。另一方面，仅依赖稀疏的结果奖励难以充分挖掘推理轨迹中的信息价值。为此，研究者开始探索基于过程评估的利用方式。GenPRM^[120] 通过过程奖励模型对推理步骤进行判别，而 TP-GRPO^[213] 指出密集过程奖励可能引入不稳定性，并提出更稳健的过程奖励机制。总体而言，LLM 在推理经验利用方面仍处于早期阶段：一方面，研究范式正由有监督微调向 RL 转变；另一方面，探索效率与经验利用之间仍存在权衡。特别是在长上下文推理任务中，状态空间规模随推理长度迅速增长，如何对推理经验进行更结构化、更高效的利用，仍是提升推理能力上限的关键问题。

降本增效：更高效的 LLM 推理 随着大语言模型推理能力的持续提升，一个日益突出的挑战是 Long CoT 场景下的推理低效率问题：模型往往生成冗长且冗余的中间推理过程，从而显著增加推理时延与计算开销，限制了其在实际应用中的可扩展性。围绕这一问题，2025 年逐步形成了以高效推理 (Efficient Reasoning) 为核心的研究方向，其目标是在尽量保持推理正确性的前提下，系统性压缩推理长度与计算开销。下面我们主要介绍基于长度奖励的强化学习优化、动态推理范式，以及混合长度 CoT 的监督微调三种思路，它们分别从优化目标、推理策略与数据分布层面对推理冗余进行约束。

早期强化学习方法通常仅以答案正确性作为奖励信号，容易诱导模型通过不断延长推理链条来提升成功率。为此，近期研究开始将推理长度或计算预算显式纳入奖励函数，引导模型学习“在足够正确的前提下尽量少想”。ShorterBetter^[214] 指出推理长度与性能之间并非单调关系，并通过在奖励中联合建模正确性与推理代价，使模型自适应收敛到更高性价比的推理长度区间。进一步地，AdaCtrl^[215] 引入难度感知的长度奖励机制，使模型在简单问题上快速终止、在复杂问题上适度延长推理。

除通过奖励约束推理长度外，另一类研究关注在推理阶段动态分配计算量，使模型具备“何时多想、何时少想”的能力。AdaptThink^[216] 通过强化学习使模型在“深入推理”与“直接回答”两种思考模式之间进行自适应选择，在显著降低平均推理长度的同时保持甚至提升性能。AdaRL^[217] 进一步提出长 CoT 与短 CoT 的混合推理框架，通过层级自适应优化在不同粒度上

选择合适的推理风格。整体来看，动态推理范式体现了从固定推理模板向策略自适应推理的转变。

相较于强化学习方法，一些研究通过构造可变长度的推理数据进行监督微调，使模型在训练阶段内化高效推理行为。其典型流程是先生成完整 CoT，再通过压缩或筛选构造短推理示例。C3oT^[218]利用强模型对长 CoT 进行语义保持的压缩，从而构造高质量短推理数据；TokenSkip^[219]则从 token 级别识别并跳过对最终答案贡献较小的中间步骤。该类方法不依赖复杂奖励设计，训练过程稳定，为高效推理提供了一种实用的数据驱动路径。

总体而言，高效推理研究正从简单的推理长度压缩，发展为目标、策略与数据协同优化的系统化框架：长度奖励的强化学习从优化目标层面抑制推理冗余，动态推理范式在推理阶段实现算力自适应分配，而混合长度 CoT 的监督微调则通过数据分布塑造模型行为。该方向不仅对于推理模型的高效化具有实际意义，也为理解大模型推理机制本身提供了新的研究视角。

工欲善其事，必先利其器：推理与工具的结合 尽管大语言模型在自然语言推理任务中取得了显著进展，但仅依赖自然语言 token 的推理范式在复杂真实场景中仍显不足。一方面，许多推理任务本身具有高计算密度或高精度要求，如数值计算、符号约束验证与程序执行，这类任务对结果的可验证性与一致性高度敏感。另一方面，模型的推理状态隐含于连续的语言序列中，缺乏对中间变量与外部状态的显式建模，使其难以在复杂推理过程中进行可靠的状态更新与误差校正。此外，现实世界的推理问题往往涉及代码、结构化数据、视觉信息与交互式环境等多种模态，仅在自然语言空间中推理难以充分刻画任务所需的结构与约束，往往被迫依赖冗长的 CoT，从而带来显著的计算开销与不稳定性。Tool-Integrated Reasoning (TIR) 通过引入外部工具，被视为突破上述瓶颈的重要方向。

代码执行是 TIR 中最成熟的工具形态之一。与将代码仅视为外部计算器的早期做法不同，近期研究强调代码与推理过程的深度耦合。ReTool^[220]通过强化学习学习何时生成并执行代码，使自然语言推理与程序执行交替进行；CoRT^[221]将代码片段直接嵌入推理轨迹，用于中间状态更新与逻辑验证；rStar2-Agent^[222]则构建了包含推理、执行与反馈的闭环系统，使代码执行成为持续推理的一部分。这些工作表明，代码已从辅助工具演化为推理结构本身的重要组成。

搜索增强推理是 TIR 的另一核心方向。不同于将搜索视为被动检索，近期方法将搜索建模为可学习的推理原语。R1-Searcher^[223]与 Search-R1^[224]

表 2.10: 不同推理范式的对比

推理范式	核心思想	推理展开形式	主要优势	代表性工作
串行 CoT	以自然语言 token 为单位, 线性展开完整推理链	单线程、逐步生成	可解释性强, 易于实现	DeepSeek-R1 ^[204]
结构化推理	将推理过程显式结构化(通常建模为图结构), 刻画步骤间的逻辑依赖	推理过程以图结构展开, 推理步骤为节点, 步骤之间的逻辑为关系	推理过程可编辑、可回溯, 统一链式、树式和图式推理	Adaptive Graph of Thoughts ^[230]
摘要推理	推理前, 在高层推理单元(步骤 / 计划)上进行规划推理思路	摘要层规划 + 后续推理按照规划具体展开	通过优化摘要规划, 显著降低冗余计算, 权衡准确性与效率	RLAD ^[231]
并行推理	同时探索多条候选推理路径, 通过融合缓解局部最优	多推理分支并行生成 + 自动融合	提升鲁棒性与准确率, 缓解路径锁定	ParaThinker ^[232]
隐式推理	将中间推理状态迁移至连续潜空间, 在隐空间中并行搜索和推理	潜空间状态演化(非显式文本)	显著压缩推理链长度, 自然并行化搜索	Coconut ^[233]
层次推理	显式建模“如何推理”, 高层调度推理策略, 低层执行	高层规划子任务, 底层具体执行解决子任务	更优的自适应规划与计算资源分配	Thor ^[234]

通过强化学习显式优化查询构造与多跳检索策略, 使搜索行为与推理目标紧密对齐; StepSearch^[225] 进一步将搜索决策细化到推理步骤级别, 缓解长链推理中的信用分配问题。整体来看, 搜索工具不仅补充外部知识, 更为模型提供了一种结构化的信息探索机制。

在多模态推理场景中, TIR 正从文本中心范式向跨模态协同演进。ToolVQA^[226] 构建了多步骤工具增强的视觉问答数据集, 表明针对工具链训练的模型在复杂多模态推理中具有优势。这类工作表明, 多模态推理的关键在于策略性选择感知与工具使用方式。除此以外, 随着多步工具调用的引入, TIR 对长期记忆与依赖性推理的需求显著增强。3DLLM-Mem^[227] 通过动态记忆管理提升连续推理稳定性; Causal World Model Induction^[228] 结合因果世界模型与显式记忆, 增强跨情境推理能力; MemTool^[229] 探索了记忆与工具管理的统一建模。这些研究表明, 显式记忆与状态表示是支撑复杂 TIR 系统一致性与可解释性的关键组件。

更多推理范式探索 当前 CoT 通过显式展开单线程语言序列，在复杂推理任务中显著提升了大模型的性能与可解释性。然而，其固有的串行执行范式在计算效率、搜索空间与鲁棒性方面逐渐显露结构性瓶颈。为突破上述限制，近期研究开始探索超越单一串行 CoT 的多样化推理范式，我们在表 2.10 中进行了部分总结。

结构化推理将推理过程从线性文本结构化为显式图结构，以节点表示推理状态、边刻画逻辑依赖关系，从而增强推理的可组织性与可回溯性。Adaptive Graph of Thoughts^[230] 在测试时动态构建有向无环图，对问题进行递归分解与选择性展开，在形式上统一了链式、树式与图式推理，并在数学与科学推理任务中显著提升性能。

摘要推理通过将主要的推理逻辑归纳于摘要部分，在进行具体的推理步骤前，首先撰写推理摘要进行全局的推理方案规划，在摘要的指导下进行具体的推理。代表性工作 RLAD^[231] 引入显式推理调度机制，使模型能够在推理过程中动态权衡展开深度与计算成本，在保持全局一致性的同时，仅需对关键摘要进行优化，从而提升样本与计算效率。

并行推理通过同时探索多条候选推理路径，缓解串行推理中的局部最优与隧道视野问题。ParaThinker^[232] 采用原生并行思维框架，并行生成多个独立推理分支，并通过后期融合或投票进行整合，在几乎不增加推理延迟的情况下显著提升复杂任务的准确率与稳定性。

隐式推理将中间推理状态从显式语言 token 转移至连续隐空间，从而摆脱离散生成的逐步约束^[233]。相关分析表明，隐空间中的 superposition 表示^[235]能够并行编码多个搜索前沿，在图可达性等任务中显著压缩推理链长度，为突破离散 CoT 的展开机制提供了新的方向。

层次推理将“如何推理”本身作为可规划、可调度的对象，通常采用分层架构：高层控制器负责选择推理模式与资源分配策略，低层执行器负责具体推理步骤的实现^[234]。该范式在全局策略一致性与计算资源控制方面展现出更强的灵活性与扩展潜力。

未来展望

展望 2026 年大模型推理技术的发展，首先，围绕复杂任务对效率、鲁棒性与可扩展性的需求，研究将进一步探索表达能力更强和高效的推理范式，突破单一串行 CoT 的限制，在并行化、层次化与结构化推理之间形成可组合的统一框架。其次，随着推理任务不断向高计算密度与高精度场景拓展，

工具增强推理将从辅助手段演化为推理系统的核心组成部分，模型需要在推理过程中自主选择、调用与校验外部工具的能力，以确保结果的可靠性与可验证性。进一步地，在面向现实环境的智能体系统中，推理将不再局限于纯语言空间，而是与感知、行动紧密耦合，推动多模态与具身智能场景下的推理能力发展，使模型能够基于视觉、时序与环境反馈进行持续决策与修正。与此同时，推理能力的获取方式也将逐步从依赖人工设计与标注，转向更加自主的推理学习框架，通过自我探索、自我评估与自我进化机制，在最小人为干预下实现推理策略的持续优化。总体而言，未来的 LLM 推理系统将呈现出更强的自主性、系统性与环境适应能力，逐步迈向在开放世界中长期运行的通用推理智能。

2.3.3 数学/代码

研究背景

数学与代码能力是衡量大模型智能水平的重要标准，与此同时，面向软件工程 (SWE) 的实际应用需求，构建能够解决真实世界复杂工程问题的智能系统具有巨大的实用价值，这将直接推动生产力的代际变革。因此，大模型的数学与代码能力在学术探索与产业应用上均具有重要意义。

在 2025 年之前，数学与代码大模型主要沿用海量无监督预训练 + 指令微调 (SFT) 范式，在大规模数学与代码语料上学习相关知识，并通过指令数据对齐人类意图。这一类模型（如 GPT-4^[236]、DeepSeek-Coder V2^[237]）在基础数学题和函数级代码生成上表现突出，但其能力本质仍偏向统计驱动的模式匹配，对未见过的复杂逻辑问题缺乏泛化能力。比如在数学领域，虽然早期模型能解决基础运算与常见应用题，但在面对竞赛级难题（如 AIME、IMO）时，往往因缺乏长程推理的稳定性而陷入幻觉。同时，在应用层面，单文件、对话式数据难以支撑真实软件工程需求。模型难以理解项目级多文件上下文，无法处理跨文件依赖与修改引发的连锁影响，且缺乏基于真实执行环境的反馈，使得生成代码在实际项目中往往不可编译或不可运行。

研究进展

针对上述局限，2025 年这类能力大模型的发展呈现两条主线：一是推理能力的激发，由于数学和代码问题具备客观可验证的答案，使得这两类数据成为增强模型长思维链能力的最佳载体。该路线旨在通过构建更高密度、更高质量的思维链数据与更长期的后训练范式，实现模型推理能力的进一步提

升；二是训推一致的导向，即面向真实软件工程场景，构建高质量的仓库级数据，并将强化学习适配真实开发环境。

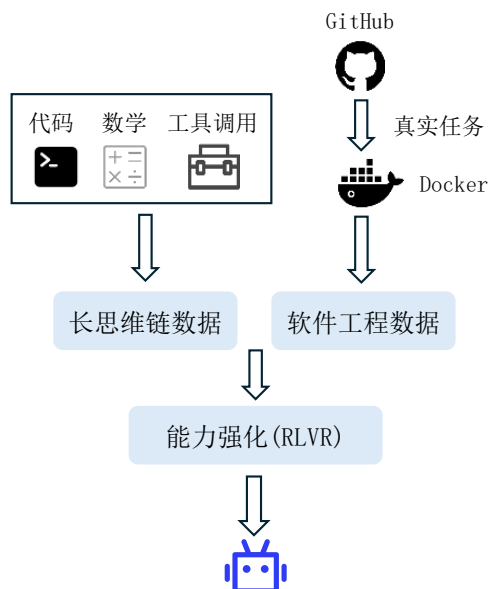


图 2.3: 数学/代码能力发展路线

长思维链数据 数据通常先从竞赛级难题（LeetCode、Codeforces、AIME 等）中收集与过滤，以获取高纯度逻辑样本。DeepMath-103K 通过嵌入相似度检索与 LLM-Judge 的复合净化机制规避数据泄露，并结合 GPT-4o 难度分级仅保留高复杂度题目，确保种子问题的纯净与挑战性^[238]。在此基础上，数据配比决定能力覆盖与深度：OpenThoughts 以消融实验量化各数据源贡献，给出代码、数学与通用逻辑的最佳混合比例，并指出适当提高数学推理权重可增强抽象逻辑理解^[239]。答案生成环节则由静态生成转向迭代进化，AIMO-2 Winning Solution 采用工具集成推理（TIR）与迭代进化策略，极大地扩充了高质量推理轨迹的规模^[240]。后置过滤方面，OpenCodeReasoning 强调过滤尺度的权衡：过严的执行过滤虽保正确性，却会破坏难度分布并降低多样性，因此更倾向保留逻辑自洽但偶发超时的解法，以维护泛化潜力^[241]。

推理能力强化 在系统化数据构建完成后，单纯监督微调（SFT）逐渐暴露出局限性：模型往往更擅长学习思维链的表述方式与输出结构，却难以获得与之匹配的真实逻辑推导能力。为突破这一局限，2025 年训练范式转向基于

可验证奖励的强化学习 (Reinforcement Learning with Verifiable Rewards, RLVR)^[242], 其不同于依赖主观偏好打分的 RLHF^[243], 而是利用代码与数学任务中天然存在的客观真值作为确定性反馈信号, 例如编译与单元测试是否通过、推导结果是否匹配标准答案。RLVR 的实施通常遵循 SFT 冷启动与迭代强化: 先以高质量思维链数据进行 SFT, 使模型具备基本指令遵循与规范输出以降低早期无效探索, 再采用 PPO^[244] 或更高效的 GRPO^[105] 在同题多采样解法的执行验证奖励下进行大规模强化。最终, 这种在验证中学习的机制催生了深层推理能力的涌现, 模型开始表现出自我纠错与长程规划等行为模式, 如生成前更充分的推演、主动构造测试验证假设并在发现漏洞时回溯修正, 从而实现从文本生成到推理决策的能力跃迁, 更能应对真实世界中长流程、非模版化的工程难题。

软件工程数据 面向软件工程 (SWE) 的数据构建同时受制于环境配置难与任务描述模糊, 2025 年研究主要沿高保真合成和规模化真实挖掘两条路径推进。合成路径中, SWE-smith 提出环境优先原则, 先确立 Docker 执行环境及高测试通过率 (>80%) 的基座再合成任务, 并借助基于执行的过滤机制从少量仓库中生成大量高质量实例^[245]; 为缓解环境依赖带来的扩展瓶颈, SWE-Mirror 提出跨仓库镜像范式将真实 Issue 的语义迁移到具备成熟 Gym 环境的目标仓库, 实现无需单独配置环境的规模化扩展^[246]。在此基础上, SWE-Synth 聚焦自动程序修复, 通过注入-验证策略生成结构化修复轨迹数据^[247], SWE-Dev 则面向新功能开发, 利用动态分析逆向构建具备明确验收标准的任务, 形成覆盖 Bug 修复、功能开发与跨仓库迁移的合成数据体系^[248]。真实挖掘路径则强调分布真实性与环境鲁棒性: SWE-Gym 通过对 GitHub 仓库的严格筛选与执行验证, 构建了规模较小但难度与质量高的基准; 为了进一步验证数据规模效应, Skywork-SWE 构建了覆盖 2,531 个仓库的万级规模数据集, 验证了在可执行性约束下扩大规模与多样性可显著提升泛化能力, 证明了 Scaling 在代码智能体领域的有效性^[249]。

软件工程能力强化 在获得仓库级演化数据后, 训练范式转向全流程工程模拟: 有监督微调不再停留于代码补全, 而是训练模型执行检索、定位与修改的工程范式, 基于 Issue 生成检索指令、定位文件并输出 Unified Diff 补丁, 使模型内化软件工程导航与修改行为。为进一步突破静态数据限制, SWE 场景引入基于沙箱环境的强化学习, 将测试套件的执行结果作为客观反馈, 在隔离的 Docker 容器中通过复现测试与回归测试验证能否复现 Bug、修复

代码且不引入新 Bug。为提升探索效率，Agent-RL^[250] 在失败时提供错误定位或高层规划等信号并逐步退火，引导策略在复杂环境中探索；为了深化自我纠错能力，ReVeal^[251] 提出了生成与验证双重闭环机制，通过编写、测试与修正的迭代循环独立解决复杂问题。进一步地，DeepSWE^[252] 改进了 GRPO 鼓励广泛策略探索，展示了不依赖人类有监督微调数据、直接利用纯 RL 进行大规模训练的潜力。

未来展望

回顾 2025 年的技术演进，数学与代码研究重心不再局限于局部生成的准确率，而是转向长程逻辑推理、可验证执行以及真实软件流程的整体建模。数据范式随之发生根本变化：训练样本从静态的问题与答案升级为包含思维过程、代码执行与结果验证的全链路结构，运行环境与测试用例被视为数据本身的重要组成部分，为基于可验证奖励的强化学习与高可靠性推理能力提供了基础。

展望未来，数学与代码有望从辅助工具演进为具备一定自主性的智能体。一方面，思维链将进一步演化为推理、执行、观察与修正的闭环过程，在神经推理与符号执行之间实现更紧密的协同；另一方面，随着可执行环境的规模化完善，监督微调将主要承担冷启动角色，基于真实环境反馈的强化学习将成为能力提升的核心驱动力。进一步地，结合多解验证、自合成数据与规则反馈机制，模型有潜力逐步减少对人工标注的依赖，通过自我博弈式的持续进化，探索更加多样且具有泛化性的推理与工程策略。

2.3.4 工具调用

研究背景

尽管大语言模型（LLMs）在自然语言处理领域取得了显著成就，但其能力仍受限于固定的参数化知识，难以实时访问信息、执行复杂计算或利用领域专业知识，从而易产生幻觉及过时信息。工具学习（Tool Learning）作为一种关键范式应运而生，旨在增强 LLMs 解决复杂问题的能力并克服其内在局限性^[253]。通过集成计算器、搜索引擎、代码解释器或 API 等外部工具，LLMs 能够动态扩展其功能范围与知识边界，显著提升响应的准确性与可靠性^[254]。

早期工具学习研究主要聚焦于有监督微调来增强模型单轮工具调用能力，侧重于训练模型何时调用何种工具用于解决问题以及如何构造工具参

数。该范式虽为 LLMs 掌握基础工具使用方法奠定了基础，但本质上属于依赖预定义轨迹的模仿学习，缺乏将工具使用策略内化的能力。因此，在面对现实世界中涉及长程任务，动态纠错及工具泛化复杂场景时，此类模型表现出明显的局限性。

鉴于此，本年度的研究核心进展集中于以下三大趋势，旨在系统性解决工具学习所面临的核心挑战。

研究进展

LLMs 工具学习中，传统有监督微调模型在长程任务，动态纠错及工具泛化复杂场景时常面临挑战，促使研究重心转向赋予 LLM 更鲁棒性的工具学习能力。为全面揭示这一演进脉络，本节将重点围绕**多轮工具调用**、**动态交互反思**及**泛化性增强**这三大核心趋势展开论述，并补充介绍最新的评估基准体系。首先我们对 2025 年具有影响力的研究工作进行系统梳理，表 2.11 详细展示了这些演进维度及代表性工作。

表 2.11: 2025 年 LLMs 工具学习研究进展总结

演进维度	核心趋势	代表工作
多轮工具调用	长程序列规划，强调任务分解与多轮序列规划。	BUTTON ^[255] , REFUEL ^[256]
动态工具反思	主动反思学习，利用工具反馈构建深度纠错机制。	Tool-MVR ^[254] , FAIL-TALMS ^[257]
泛化性增强	工具无关泛化，通过纯强化学习激发内在推理逻辑。	Tool-Zero ^[258] , GenTool ^[259]

从单轮调用到多轮调用工具的深化。 当前，工具学习的研究正在从单轮次工具调用模式演进为执行复杂任务的多轮次调用工具的范式。由于实际应用场景极少是孤立的单次交互，而更倾向于需要模型在持续对话中维持上下文、保持连贯性、处理歧义并动态调整响应策略的复杂互动^[260]，因此这种演进具有深远意义。然而，这一转变面临着多重挑战，核心难点包括高质量多步交互数据的极端匮乏，多轮对话中上下文漂移问题的普遍存在，以及 LLMs 在长程规划能力上的不足。为突破上述限制，2025 年的研究工作主要通过构建自动数据合成方法、整合情景记忆以及优化长程奖励机制来提升多

轮交互性能。为解决高质量多步交互数据匮乏问题，BUTTON^[255]框架提出了一种“自底向上构建指令、自顶向下生成轨迹”的数据合成方法，通过多智能体环境模拟人类与工具的真实交互，有效产出了大规模高质量的复杂序列数据；在解决多轮对话上下文漂移方面，^[261]强调将情景记忆整合到 LLMs 中，通过将连续交互离散化、检索过去经验并整合为通用知识，显著提升了模型在多轮互动中的上下文理解和连贯性；针对长程规划能力不足的挑战，REFUEL^[262]则提出了一种策略价值优化方法，通过迭代训练 Q 值函数回归未来的累积奖励，有效解决了多轮 RLHF 中的协变量偏移问题，并在长对话基准测试中超越了现有先进方法。

从静态模仿到动态交互反思。 研究重点，从仅学习专家成功轨迹的静态模仿学习，向基于环境反馈进行动态交互与自我纠调的范式跨越。由于传统有监督微调范式仅模仿成功轨迹，导致模型在遭遇执行错误时表现脆弱且缺乏自我纠错能力^[263]。故这一转变能够通过模拟人类的问题解决思维，赋予智能体在动态环境下的交互纠错与容错能力，从而维持长程决策的稳健性。然而，当前的研究难点在于模型在遭遇 API 执行异常时极易陷入反复尝试无效操作的局部陷阱以及由于关键信息不足导致工具调用失效时产生严重的幻觉输出。2025 年的研究工作主要通过设计动态学习范式与交互式干预机制来突破上述瓶颈：针对模型缺乏错误恢复机制的问题，ToolMVR^[254]工作提出了 EXPLORE 算法，利用“错误 → 反思 → 校正”的动态强化学习范式替代模仿学习，利用工具反馈来增强工具反思能力，使模型在 RefineToolBench 评测集上的错误纠正率从 9.1 % 显著提升至 58.9 %；针对工具失效导致的盲目决策问题，FAIL-TALMS^[257]工作提出了询问与求助（Ask-and-Help, AAH）的实时干预方法，通过赋予智能体主动向人类索取缺失信息或寻找替代方案的权限，在用户查询不足的情境下显著提升了任务的最终通过率。

工具调用泛化性的增强。 泛化性指 LLM 智能体在面对未见查询或工具时，仍能保持鲁棒性与有效性的能力。然而，当前经有监督微调训练的模型因倾向于模仿表面模式，在此方面表现欠佳^[258]。现实场景中的面对的工具多种多样与不确定性，要求智能体在面对未知工具或指令时仍能保持逻辑稳健，而非仅仅依赖预定义的工具匹配规则。然而，现有模型普遍面临过度拟合特定工具描述而难以泛化至新领域的问题。2025 年的研究工作主要通过合成泛化场景数据与通过可验证强化学习激发模型内在推理能力来突破上述限制：针对模型对未知工具处理能力较弱的问题，GenTool^[259]工作通过模拟

“从零到一”和“从弱到强”两种工具泛化场景生成合成数据，并结合工具排序和精炼选择两阶段微调策略，显著提升了大型语言模型在面对未见过查询和工具时的泛化能力，在多项泛化场景中相较 GPT-4o 提升 14.28 % 选择准确率的效果。针对监督微调范式导致的泛化失效的问题，Tool Zero^[258]工作提出了一种完全脱离有监督微调的纯强化学习（Pure RL）方案，利用动态调整奖励机制引导模型从零开始训练，从一开始鼓励大胆探索到慢慢地引导它规范地使用工具，使模型学会举一反三，从而大幅提升在面对从未见过的新工具时的适应能力和内在思考能力。

工具调用评估基准。 随着工具调用从简单的识别并执行的范式演进为复杂的智能体交互，评估体系也经历从单一的任务成功率指标向更加细粒度的指标进行转变。2025 年的最新基准则更加关注模型在多步规划中的处理能力，面对错误反思与修正能力以及泛化场景下的鲁棒性。表 2.12 详细对比了当前主流工具调用评估基准的核心关注点与关键指标。

表 2.12: 2025 年 LLMs 工具调用评估基准对比总结

评估基准	核心关注点	关键指标
τ^2 -Bench ^[264]	双控环境 （Dual-Control）；用户与 Agent 共同操作工具；测试协作与沟通能力。	任务完成率、协作效率
BFCL ^[265]	覆盖多语言及 多轮交互 任务；首创基于 AST 的静态验证，确保评估可复现。	AST 匹配率、执行一致性
AceBench ^[266]	涵盖 Normal/Special/Agent 三种模式；模拟真实 多轮对话 及多智能体交互。	成功率、错误归因分析
RefineToolBench ^[254]	专注 深度反思 ；通过错误场景测试模型型的自我修正与恢复能力。	错误识别率、纠正率
HammerBench ^[267]	面向 移动设备 场景；涵盖意图切换、参数偏移等细粒度交互。	准确率、幻觉率、遗漏率

未来展望

展望未来，大语言模型（LLMs）的工具学习将从受限的、静态的指令遵循，演进为能够在复杂现实世界中自主行动的通用智能体。我们主要从动态环境交互、工具泛化性以及多模态能力三个维度进行展望：（1）动态环境交互，未来的研究将超越简单的“请求-响应”模式，致力于提升智能体在非稳

态环境中的表现。这要求 LLMs 具备更强的闭环反馈感知能力，能够根据环境的实时变化动态调整行动策略。通过引入长期规划、自我纠错以及高效的异步调用机制，智能体将能够处理具有不确定性的复杂任务，实现与物理或虚拟世界近乎自然的流畅交互。(2) 工具泛化性，工具学习将迈向“零样本”或“即插即用”的新阶段。未来的智能体将不再局限于预定义的封闭工具集，而是具备理解未知工具的能力。通过自主阅读 API 文档、观察人类示范或在沙盒中进行探索性尝试，LLMs 将展现出极强的通用性，能够灵活调用从未见过的新型工具来解决长尾问题，从而构建起一种规模化、标准化的通用工具使用协议。(3) 多模态能力，多模态将从辅助功能转变为工具学习的核心底座。未来的系统将深度融合视觉、听觉及传感器等多源数据，开发出原生的多模态工具集。这不仅意味着智能体能调用视觉工具处理图像，更意味着其能够实现真正的跨模态推理。总而言之，这三个维度的突破将共同推动 LLMs 从单一的文本助手进化为具备高度自主性、自适应能力和感知深度的全能型智能助手，在生产工具、具身智能等领域发挥核心价值。

2.3.5 Agentic RL

研究背景 尽管监督微调（SFT）和 RLHF 支撑起了 ChatGPT 时代的技术突破，但当技术迈向需要解决复杂现实问题的 Agent 阶段时，这两种方法逐渐显露出模仿学习性能上限、分布外场景泛化问题、长程推理误差累积等局限，为了解除这些局限，2025 年学术界将研究重心转向了 Agentic RL：

- 突破模仿学习“天花板”：SFT 是行为克隆，难以提供复杂任务的完美演示；而 Agentic RL 支持模型通过自主探索（Self-exploration），可以找到人类尚未发现的更优解法，这是迈向通用人工智能（AGI）的必经之路。
- 解决 OOD 泛化难题：SFT 在分布外场景失效，如新版代码库；Agentic RL 则通过在仿真环境中开展大规模试错训练，让模型在动态环境里具备更强的鲁棒性与适应性。
- 克服长程推理误差累积：当推理链长达数十步时，SFT 模型的微小误差会呈指数级放大（即 Cascading Errors）。而 Agentic RL 通过优化整个任务轨迹的累积回报（Cumulative Reward），引导模型更关注长期结果，而非局限于短期 Token 的生成概率。

Agentic RL 的核心思路是：将 LLM 作为策略网络（Policy Network），部署在动态、部分可观测的环境框架（POMDP）中；模型通过与环境的持续交互、试错迭代，以最大化长程任务的整体效用（Utility）为目标，是一种面向复杂场景的智能体学习方法。

根据Li et al.^[268] 的综述研究，Agentic RL 与 RLHF 等传统 LLM 强化学习存在本质差异：后者仅聚焦于优化单轮对话的偏好匹配，而 Agentic RL 则致力于增强模型在复杂环境中的自适应能力与自主决策能力。为直观展示这一范式跃迁，表 2.13 对比了二者的核心差异。

表 2.13: 从对齐到自主：Agentic RL 的范式重构^[268]

维度	传统 LLM-RL (如 RLHF)	Agentic RL
角色定位	定制化答题机器	自主办事员
环境建模	静态 Context (MDP)	动态、部分可观察环境 (POMDP)
交互方式	被动接收指令，单向输出	主动感知、查询信息、改变环境状态
优化目标	人类偏好对齐 (Alignment)	任务效用最大化 (Success Rate)
数据来源	离线人类标注数据	在线交互 (On-policy) / 自我博弈
能力边界	受限于人类演示 (Imitation)	可超越人类水平 (Exploration)

研究进展 当前成熟的 Agentic RL 系统已构建起“感知 - 决策 - 行动 - 学习”的完整闭环，其核心能力可拆解为表 2.14 的 6 个维度。

这些能力构成了 Agentic RL 的技术底座，也为 2025 年的研究突破锚定了具体方向。2025 年 Agentic RL 的三大研究进展，正是围绕这些核心维度的深化与拓展展开：

- 1. **慢思考下的深度推理** System1（快思考）依赖直觉易出错；System2（慢思考）依赖逻辑更准确^[274]。Agentic RL 通过强化学习激励与内化自我修正技术推动了 System2 慢思考的发展：

表 2.14: Agentic RL 的核心能力拆解

维度	能力	近两年代表性工作
1. 规划	把模糊目标转化为可执行步骤，动态调整路径	Putta et al. ^[269] 的 Agent Q（结合 MCTS 搜索与 DPO 进行长程规划）；Guan et al. ^[270] 的 rstar（利用过程奖励引导推理搜索）
2. 推理	按任务复杂度切换模式，实现 System 2 级慢思考	DeepSeek-R1 ^[23] （通过纯 RL 激发思维链生成）
3. 自我改进	从错误中学习并固化能力，非单次修正	Kumar et al. ^[271] 的 SCoRE(多轮强化学习实现自我修正)
4. 工具使用	自主判断工具调用时机与组合方式	Li et al. ^[268] 的 Agentic RL Survey（总结了交互步数增加带来的工具能力涌现现象）
5. 记忆	灵活管理工作记忆，将被动检索转化为主动的“记忆导航策略”	Chen et al. ^[272] 的 MemWalker(将长上下文处理建模为强化学习过程)
6. 感知	突破纯文本限制，利用 RL 优化多模态环境下的感知能力	Bai et al. ^[273] 的 DigiRL（利用 RL 对齐 VLA 模型在 GUI 屏幕上的点击操作）

- 强化学习激励：DeepSeek-R1 模型的提出证明了在不依赖大量监督数据（SFT）的情况下，仅通过 GRPO (Group Relative Policy Optimization) 算法和规则奖励（如数学答案正确性），模型就能自发涌现出“验证-反思-修正”的思维链（CoT）模式，大幅提升了推理能力；
 - 内化自我修正：Kumar et al.^[271] 的 SCoRE 框架构建了“生成-验证-修正”闭环，这一机制能够有效阻断思维链（CoT）中的幻觉信息传播。
2. **过程监督与长程规划** Agentic RL 针对“奖励稀疏”痛点，从“结果导向”转向“过程导向”，技术进展主要体现在结合搜索算法（Search）与过程奖励（PRM）：
- 过程奖励引导搜索：Guan et al.^[270] 的 rStar 模型提出了一种自我进化的推理方法。它利用蒙特卡洛树搜索（MCTS）来探索推理路径，并结合过程偏好模型（PPM）来评估每一步的质量。这种方式让小模型在复杂数学任务上达到了媲美 OpenAI o1 的水平；

- 长期规划与自我探索：Putta et al.^[269] 的 Agent Q 引入了类似 AlphaZero 的树搜索机制与 DPO 相结合，使智能体能够通过自我博弈在 Web 导航等长程任务中进行多步规划，并从失败轨迹中学习，解决了长程任务中的“迷路”问题。

3. 缩放定律与多智能体协作 Agentic RL 的技术进展推动了缩放定律与多智能体协作领域的发展

- DeepSeek-R1 和 OpenAI o1 的成功验证了 Scaling Laws 的新维度：随着推理阶段计算量（思考时间/Token 数）的增加，模型性能呈显著增长趋势；
- Zhang et al.^[275] 的 AFLOW 将多智能体协作问题转化为图搜索任务。借助强化学习自动搜索得到的“递归调试”协作拓扑，其效率已超过人类专家手动设计的工作流程。

不难看出，2025 年的这三大进展并非孤立突破，而是 Agentic RL 六大核心能力维度在实际场景中的定向深化，核心能力为研究进展提供了技术基础，研究进展则让核心能力实现了更具实用性的落地。

未来展望 尽管 Agentic RL 已展现出从“语言理解”向“行动执行”跨越的潜力，但其当前仍在可信度与安全性、样本效率、仿真环境迁移真实场景三个方面面临显著挑战——在可信度与安全层面，需在保障训练效果的同时规避智能体为最大化奖励选择危害捷径的“奖励黑客”行为；在样本效率维度，当前智能体依赖海量交互数据收敛的模式，与人类高效学习范式存在明显差距；而仿真环境中形成的策略难以适配医疗、金融等真实场景，也制约着其落地价值。Agentic RL 在未来将通过自主规划、工具使用与自我改进能力的持续深化，逐步弥合当前在可信度、效率与场景迁移上的短板，最终明确人工智能领域从“生成式 AI”迈向“通用智能体”的可行路径。

2.4 开源训练框架

为了应对大语言模型（LLM）后训练阶段日益增长的复杂性与规模化需求，开源社区涌现出了一批针对不同痛点优化的训练框架。这些框架在系统架构、训练效率、显存利用率以及算法适配性等方面呈现出显著差异，逐渐形成了面向不同应用场景的技术分工。本节将围绕当前主流的六个开源训练

表 2.15: 开源大模型后训练框架综合对比

特 性 维 度 (Feature)	VeRL	ROLL	PRIME-RL	Slime	RAGEN	OpenRLHF
核 心 设 计 理 念	HybridFlow / 3D-Engine	异 步 流 水 线 (Async Pipeline)	去中心化 / 离线 RL	SGLang 原生集成	智能体轨迹优化	Ray 分布式调度
推 理 生 成 引 擎	vLLM / SGLang	vLLM / SGLang	vLLM	SGLang (Native)	vLLM	vLLM
训 练 后 端 框 架	FSDP / Megatron	Megatron / FSDP2	FSDP2	Megatron	FSDP	DeepSpeed (ZeRO)
关 键 算 法 支 持	PPO, GRPO, DAPO	PPO, GRPO, RLVR	PRIME, DPO	PPO, GRPO	StarPO	PPO, DPO, GRPO
显 存 优 化 技 术	零 冗 余 (Zero Redundancy)	RollPacker 打包 调度	FSDP2 原生支持	分离式架构	继承 VeRL	Hybrid Engine
最 佳 适 用 场 景	70B+ 超大模型 / 显存受限	数学/代码推理 (长尾任务)	离线数据 / 算力分散	极致吞吐 / MoE 模型	多轮智能体交互	学术研究 / 快速基线

框架——VeRL、ROLL、PRIME-RL、Slime、RAGEN 和 OpenRLHF，对其核心设计理念与技术特点进行系统梳理。

近年来，LLM 强化学习训练框架的研究进展主要体现在三个方面：一是面向超大规模模型的显存与吞吐优化，二是针对长序列推理与智能体交互场景的系统级改进，三是对离线强化学习与分布式训练新范式的探索。在此背景下，业界和学术界相继提出了一系列具有代表性的开源框架，形成了较为清晰的技术谱系。下面将从具体技术实现角度，对各框架的核心设计进行介绍。

2.4.1 VeRL (Volcano Engine)

简介 VeRL^[276] 是由字节跳动（Volcano Engine）与香港大学联合推出的开源强化学习训练框架。其核心设计基于 HybridFlow 编程模型，旨在解决大规模 RLHF 训练中计算与数据依赖的复杂性。VeRL 通过将逻辑控制流与物理执行流解耦，支持在异构硬件资源上灵活执行生成与训练任务，是目前工业界解决“显存墙”问题的代表性方案。VeRL 的核心技术包括：

- 3D-HybridEngine：在推理后端（如 vLLM）与训练后端（如 PyTorch FSDP）之间原位动态重分片（In-place Resharding），完全消除推理与训练维护两份模型权重产生的显存冗余；
- HybridFlow：将计算逻辑与数据流解耦，实现逻辑控制流与物理执行流分离，可在异构硬件间灵活调度；
- 对 PyTorch FSDP2 的原生支持和优化的跨节点通信策略，提高大规模集群的吞吐量。

优势与不足 VeRL 的最大优势在于其 3D-HybridEngine 消除了传统流水线中维护两份模型权重产生的显存冗余，使有限显存条件下训练 70B 甚至更大参数模型成为可能；HybridFlow 的设计和对 FSDP2 的优化使其在大规模集群上具有高吞吐量和灵活的硬件适配性。然而，由于采用了独特的 HybridFlow 编程范式，VeRL 的学习曲线相对陡峭，开发者需要理解计算与数据流解耦的抽象概念；在适配非官方支持的小众模型或特殊自定义算法时，可能需要深入底层进行较多代码修改。

2.4.2 ROLL (Alibaba)

简介 ROLL (Reinforcement Learning Optimization for Large-scale Learning)^[277] 是阿里巴巴开源的 RL 训练框架，专为解决大规模长序列推理任务（如数学证明、代码生成）中的效率瓶颈而设计。针对这些任务中样本生成长度方差极大的特点，ROLL 引入了基于 ROLL Flash 的异步生成-训练架构，利用生产者-消费者模型将轨迹采样与 Update 阶段解耦。ROLL 的核心技术包括：

- ROLL Flash：基于生产者-消费者模型的异步生成-训练架构，缓解长尾延迟问题并提高 GPU 利用率；
- RollPacker 调度算法：智能打包不同长度的样本，在部分样本生成完成后即可启动训练，消除同步等待造成的计算气泡；
- 对 DeepSeek-R1 系列算法（如 GRPO）和 RLVR 任务的开箱即用支持。

优势与不足 ROLL 的优势在于通过异步流水线和 RollPacker 调度算法有效解决了长尾延迟（Long-Tail Latency），使训练进程在部分样本生成完成时即可开始更新，显著提高了大规模长序列任务的吞吐量；它在 RLVR 和 Agent 任务上展现出优于传统框架的效率，并提供了对 DeepSeek-R1 算法的开箱即用支持。异步训练模式同时引入了策略滞后（Policy Staleness）的风险，即更新使用的样本可能来自稍旧的策略版本，需要通过精细的超参数调节缓解；此外，社区反馈指出，ROLL 在自定义多模态数据集加载和编写复杂奖励函数方面的文档仍有完善空间。

2.4.3 PRIME-RL (Prime Intellect)

简介 PRIME-RL^[278] 是一个专注于离线强化学习（Offline RL）和去中心化训练的框架。不同于传统在线 PPO 框架，PRIME-RL 强调利用静态数据集进行策略优化，并支持跨广域网（WAN）的分布式节点协同训练。其核心算法 PRIME 提出了一种通过结果奖励隐式优化思维链过程的方法。PRIME-RL 的核心技术包括：

- PRIME 算法：采用隐式过程奖励机制，通过结果奖励优化思维链过程，适用于 Offline RL 和隐式过程奖励任务；

- 去中心化架构：使用 TopLoc 验证和 ShardCast 广播等组件支持跨广域网分布式协同训练，具备较强容错性；
- 深度支持 Offline RL 算法：内置 DPO、CQL 等离线优化算法并鼓励在异构硬件上进行混合训练。

优势与不足 PRIME-RL 的架构具有极强容错性和可扩展性，特别适合算力分散或异构硬件混用的场景；对 Offline RL（如 DPO、CQL）和隐式过程奖励的深度支持，为缺乏在线交互环境或追求低成本训练的用户提供了理想方案。PRIME 算法在缓解思维链训练中过程标签稀缺问题方面取得了实际效果。相比成熟的在线 RL 框架，PRIME-RL 在纯在线任务上的工程稳定性仍处于快速迭代阶段；其去中心化架构涉及 TopLoc、ShardCast 等复杂组件，对于仅有单一局域网集群的用户而言部署较为繁琐；此外，广域网通信开销在强同步要求的任务中可能成为性能瓶颈。

2.4.4 Slime (Zhipu AI)

简介 Slime^[279] 是智谱 AI 开源的高性能后训练框架，也是 GLM-4.5 和 GLM-4.6 模型的底层训练引擎。Slime 的设计理念是“SGLang-Native”，即深度集成 SGLang 推理引擎，并与 Megatron-LM 训练后端无缝连接，专注于在千亿参数规模下实现极致的生成吞吐量。Slime 的核心技术包括：

- RadixAttention：利用 SGLang 的 RadixAttention 技术加速 KV Cache 的复用，特别适合多轮对话和含长 System Prompt 的 Agent 任务；
- 高度集成的训练后端：深度绑定 SGLang 推理引擎与 Megatron-LM 训练后端，支持共置同步和分离异步两种模式，适配 300B+ 参数 MoE 模型；
- 针对大模型优化的流水线设计：在极大参数规模下保持生成吞吐量并支持灵活的任务模式（Reasoning vs Agentic）。

优势与不足 Slime 利用 RadixAttention 技术大幅加速了 KV Cache 的复用，在多轮对话和复杂 Prompt 场景中表现出极高吞吐量；其深度集成 SGLang 推理引擎和 Megatron-LM 训练后端，使其成为目前少数经过验证可支持 300B+ 参数 MoE 模型进行 RL 训练的框架，并能根据任务类型灵活切换共置同步与分离异步模式以最大化硬件利用率。然而，这种强绑定虽然

带来了极致性能，却限制了框架的灵活性：对于希望使用其他推理后端（如 vLLM）或更轻量训练后端（如 DeepSpeed）的用户，Slime 的适配难度较大；此外，它主要面向超大模型优化，对资源有限的个人开发者而言可能显得过于“重型”。

2.4.5 RAGEN

简介 RAGEN^[280] 是一个基于 VeRL 二次开发的框架，专为 LLM Agent 的训练需求设计。它引入了 StarPO 算法，将 Agent 与环境的多轮交互轨迹视为一个整体进行优化，旨在解决传统强化学习在长程任务中容易陷入局部最优的问题。RAGEN 的核心技术包括：

- StarPO 算法：将多轮交互轨迹作为整体进行优化，避免长程任务中的局部最优，并针对 Agent 环境设计奖励机制；
- 标准化环境接口：提供统一接口接入多种交互环境，如 Sokoban、Web-Shop 等，并针对常见的“回声陷阱”进行算法级优化；
- 深度集成 VeRL：利用 VeRL 的 HybridFlow 和 3D-HybridEngine，在异构硬件上高效训练多轮交互任务。

优势与不足 RAGEN 填补了通用 RLHF 框架在 Agent 环境支持方面的空白，为研究复杂智能体行为和长程规划的学者提供了对口工具。通过 StarPO 算法和标准化接口，它在解决多轮交互中的回声陷阱和局部最优问题上表现突出。不过，RAGEN 构建在 VeRL 之上，因此继承了 VeRL 的学习曲线；其高度专注于 Agent 领域，对于普通的单轮对话或基础推理任务，其轨迹优化机制未必带来明显性能提升，反而增加系统复杂度；此外，作为科研导向较强的框架，其工程落地成熟度略低于通用 RLHF 框架。

2.4.6 OpenRLHF

简介 OpenRLHF^[281] 是目前社区使用最广泛、生态最成熟的开源 RLHF 框架。它基于 Ray 分布式框架构建，支持将 Actor、Critic、Reward 等模型调度到不同节点上，并强调易用性和兼容性，支持 HuggingFace Transformers 模型库和 DeepSpeed 优化器。OpenRLHF 的核心技术包括：

- Ray-based 分布式架构：通过 Ray 调度 Actor、Critic 和 Reward 模型实例，具备良好的弹性和容错性；

- 开箱即用的算法实现：提供 PPO、DPO、GRPO、REINFORCE++ 等常用 RLHF 算法，方便研究者快速上手；
- 广泛的兼容性：支持 HuggingFace Transformers 模型库、DeepSpeed 优化器，并拥有活跃社区快速跟进新算法（例如对 DeepSeek-R1 的早期复现）。

优势与不足 易用性是 OpenRLHF 最大的优势。它提供丰富的文档和开箱即用的算法实现，是多数研究者入门 RLHF 的首选；基于 Ray 的架构使其具有良好的弹性和容错性，社区非常活跃，新算法跟进速度快，例如它是首批支持 DeepSeek-R1 复现的框架之一。在追求极致性能的场景下，OpenRLHF 也存在瓶颈：它通常需要同时维护生成和训练两份模型权重（或通过 Ray 对象存储传输），相比 VeRL 的原位重分片技术，其显存利用率和通信效率在训练 70B+ 大模型时略逊一筹；作为社区驱动项目，其在超大规模集群上的稳定性优化可能不如由大厂背书的 VeRL 或 Slime 深入。

总结 随着 LLM 后训练技术的发展，各框架在设计哲学上展现出了明显的差异化趋势。VeRL 和 Slime 代表了工业界对极致性能和显存效率的追求，适合超大规模模型的训练；ROLL 和 RAGEN 分别针对长尾推理和智能体交互这一特定场景进行了深度优化；PRIME-RL 探索了离线与去中心化的新范式；而 OpenRLHF 则凭借其易用性成为了社区的标准基线。表 2.15 对这些框架进行了多维度的综合对比。

2.4.7 未来展望

前文系统回顾了当前主流开源后训练框架在系统设计与工程实现层面的进展。总体来看，这些框架在显存效率、吞吐优化以及特定任务场景支持等方面已取得显著成效，但整体仍处于快速演进阶段，尚未形成稳定而统一的技术体系。作为支撑大语言模型强化学习的重要基础设施，开源后训练框架目前在抽象层次和功能边界上仍存在一定分散性。不同框架往往围绕特定推理引擎、训练后端或应用场景进行深度优化，在带来性能优势的同时，也在一定程度上限制了算法与系统之间的通用复用。因此，围绕更统一的接口设计与模块化组织方式，构建能够同时承载多种算法范式和模型规模的通用训练框架，仍然具有较大的探索空间。与此同时，随着长序列推理与智能体交互任务的重要性不断提升，现有框架在调度策略、样本管理以及生成与训

训练协同方面仍面临诸多工程挑战。如何在保证训练稳定性的前提下，有效缓解长尾延迟、策略滞后和资源空转问题，仍有待在系统层面进行更深入的研究与实践。

此外，超大规模模型训练所暴露出的显存、通信与稳定性问题，也对后训练框架提出了更高要求。尽管原位重分片、深度推理训练耦合等技术已经初步展示了其可行性，但在更复杂的并行配置、异构算力以及 MoE 场景下，如何在效率与稳健性之间取得平衡，仍是未来需要持续探索的方向。总体而言，开源后训练框架仍处在从任务导向工具向通用训练基础设施过渡的阶段。随着算法范式和应用场景的不断扩展，围绕体系化设计、工程稳健性与长期可维护性的研究，预计将成为该领域持续发展的重要推动力。

2.5 本章小结

2025 年作为大语言模型技术的关键创新年，核心技术呈现多维度突破与深度融合的鲜明特征。本章从后训练技术更新、数据获取与治理、模型能力提升及开源训练框架四大核心板块，系统梳理了年度技术进展与趋势：后训练技术实现成本与效果的优化平衡，数据领域完成从规模到质量的进阶升级，模型能力朝着多维度协同与自主智能体方向演进，开源框架则形成差异化发展格局并迈向通用化基础设施。这些技术革新相互赋能，共同推动大模型从通用对话系统向具备深度推理与自主交互能力的智能体范式跨越，为后续技术落地与产业应用奠定了坚实基础。

第三章 大语言模型部署

进入 2025 年，随着大语言模型参数规模与上下文窗口的持续扩展，部署技术的核心矛盾已从单纯的算法效果探索，转向推理成本、服务延迟与硬件资源的综合制约。在这一背景下，如何突破“显存墙”与“算力墙”的物理限制，实现高效、低成本的实时推理，成为大模型落地的关键。

相较于早期“以算力换智能”的粗放模式，当前的部署技术更强调软硬件协同优化与全链路信息重构。从云端的高并发吞吐到端侧的极致轻量化，部署技术已成为决定模型应用价值的核心环节。

基于这一逻辑，本章将从模型压缩、推理加速与开源框架三个层面展开分析，系统梳理 2025 年大模型部署的关键进展及其在不同算力环境下的技术演进图景。

3.1 模型压缩

进入 2025 年，模型压缩技术已不再局限于对模型体量的被动“瘦身”，而是演变为对模型信息表达形式的主动“重构”。随着基础模型向超大规模与强推理能力发展，传统的压缩方法面临精度崩塌与泛化失效的双重挑战，推动了技术路线的全面革新。本小节将重点调研量化、剪枝与蒸馏三个方向的前沿进展：

量化技术打破了 PTQ 与 QAT 的界限，转向利用几何拓扑变换与超低位宽重构来突破 2-bit 物理极限；剪枝技术从理论稀疏度转向追求实际运行时间的真实加速，确立了以硬件友好的结构化剪枝与动态上下文跳过为主流的新范式；蒸馏技术则受 DeepSeek-R1 等模型启发，核心目标从概率分布拟合转向推理回路（Reasoning Circuit）的迁移，并结合模型容量定律重塑了师生学习策略。这些工作共同标志着模型压缩正从简单的参数删减，升维至复杂的特征空间对齐与知识表征重塑。

3.1.1 量化

研究背景

量化是指将神经网络中原本由高精度浮点数（如 FP16 或 BF16）表示的权重参数与激活值，映射为低位宽的定点整数（如 INT8 或 INT4）的技术过程。该技术旨在通过减少数值表示的比特数，在尽量维持模型性能的前提下，显著降低模型的显存占用并提升推理吞吐量。

在 2024 年及以前，大模型量化领域呈现出明显的“双极化”格局。一方面，以 GPTQ^[282] 和 AWQ^[283] 为代表的后训练量化（PTQ）占据了部署主流，它们通过利用二阶 Hessian 信息补偿误差，成功在 3-4 bit 精度下保持了模型的可用性。然而，PTQ 方法在面对 Llama 等模型中剧烈的激活值离群点（Activation Outliers）时往往力不从心，被迫采用复杂的平滑策略，且难以突破 3-bit 的精度下限。另一方面，量化感知训练（QAT）虽然能通过重训练恢复精度，但其对 GPU 显存和算力的需求呈指数级增长，导致其长期难以应用于 70B 参数以上的超大模型。

进入 2025 年，随着几何深度学习与矩阵分解理论的引入，学术界开始尝试打破 PTQ 与 QAT 的界限。研究重心从单纯的数值舍入（Rounding），转向了更高维度的特征分布对齐与训练范式重构，并开始向着超低位宽物理极限发起冲击。

研究进展

基于几何变换的 PTQ 针对激活值异常分布导致量化精度崩塌的难题，2025 年的 PTQ 研究不再局限于静态的通道缩放（Scaling），而是转向了动态的空间几何变换。该类方法的核心直觉是：通过旋转特征坐标系，将集中在少数维度的“尖峰”能量（Outliers）平摊到所有维度，使得数据分布服从更易于量化的平滑高斯分布。

Hu et al.^[284] 提出了 Ostquant，这是一种结合正交变换（Orthogonal Transformation）与通道缩放的复合优化方法。Ostquant 发现，单纯的旋转虽然能抑制异常值，但会导致通道间的幅值差异难以对齐。因此，该方法设计了一套“旋转-缩放-量化”的协同流程，在不改变模型数学输出等价性的前提下，重塑了权重与激活值的几何形状，显著降低了量化噪声。

在此基础上，Liu et al.^[285] 进一步挑战了传统方法中通过 Hadamard 矩阵进行固定旋转的范式。SpinQuant 引入了可学习旋转（Learned Rotations）机制。它将寻找最佳旋转角度转化为一个轻量级的优化问题，通过梯度下降

自动搜索出针对当前模型架构的最优旋转矩阵。实验表明, 这种“自适应几何对齐”使得 PTQ 即使在 4-bit 甚至更低位宽下, 也能保持与全精度模型惊人的一致性, 成为年度 PTQ 领域的 SOTA 标准。

高效 QAT 为了解决 QAT “精度高但训练成本不可承受”的顽疾, 2025 年的研究致力于通过分块优化策略来降低显存峰值, 实现 QAT 的“平权化”。

Chen et al.^[286] 提出的 EfficientQAT 是这一方向的代表性工作。该方法摒弃了昂贵的全图端到端反向传播(End-to-End Backpropagation), 创新性地引入了块状重建(Block-wise Reconstruction)进入 QAT 流程。EfficientQAT 将庞大的模型切分为若干独立的 Transformer 块, 逐块进行量化参数的梯度更新。这种策略使得显存占用不再随模型深度线性增加, 而仅与单个块的大小相关。配合后续极少量的端到端微调(Fine-tuning), 该方法能够以接近 PTQ 的时间与显存成本, 实现媲美全量 QAT 的高精度性能, 使得在单机环境下完成千亿参数模型的 2-bit 量化训练成为可能。

超低位宽重构 当目标位宽下探至 2-bit 以下时, 传统的线性量化会遭遇严重的信息瓶颈。因此, 这一领域的突破主要依赖于深度 QAT 或非线性重参数化, 通过彻底重构权重的表达形式来维持智能。

Lee et al.^[287] 提出了一种基于潜在矩阵分解(Latent Matrix Factorization)的新范式。该方法不再直接存储权重矩阵, 而是通过 QAT 学习权重的低秩因子并进行二值化, 结合多尺度的浮点补偿因子, 成功将 LLM 权重的平均位宽压低至 0.1 bit 级别。与之互补, Gu et al.^[288] 引入了二进制码本(Binary Codebook)技术。该工作通过识别权重矩阵中的高频二进制模式来构建共享码本, 并结合可学习的旋转变换进行重训练, 有效解决了极低位宽下的“神经元死亡”问题。

此外, Xia et al.^[289] 更是跨界借鉴了信号处理中的 Sigma-Delta 调制技术。该方法通过调整过采样比(OSR)在 1-bit 和 2-bit 之间实现连续的精度权衡, 打破了离散位宽的限制。这些工作共同标志着量化技术正在从简单的“数值压缩”演变成一种复杂的“信息编码”过程。

未来展望

随着 EfficientQAT 类方法的普及, 未来的基础模型发布很可能将不再是单一的 FP16 权重, 而是会发布自带“量化亲和性(Quantization-Friendly)”的权重分布, 甚至是针对特定 NPU 架构优化好的“软硬件协同量化包”。

表 3.1: 大模型量化关键技术研究进展

关键技术	特点	代表工作
基于几何变换的 PTQ	利用正交变换或可学习旋转矩阵平摊激活值离群点 (Outliers)，将数据重塑为易于量化的高斯分布。	Ostquant ^[284] , SpinQuant ^[285]
高效 QAT	采用块状重建 (Block-wise) 替代全图反向传播，大幅降低显存开销，实现单机低资源下的高精度微调。	EfficientQAT ^[286]
超低位宽重构	引入矩阵分解、二进制码本或信号调制技术，重构权重表达形式，突破 2-bit 线性量化的物理极限。	LittleBit ^[287] , BTC ^[288] , SDQ ^[289]

此外，混合精度量化 (Mixed-Precision Quantization) 将从人工设计的策略进化为完全自动化的“神经架构搜索 (NAS)”过程。模型将能够根据每个层对困惑度 (Perplexity) 的敏感性，自动决定该层是使用 4-bit 还是 2-bit，甚至在同一个矩阵内实现非均匀的位宽分配，以求在精度与速度之间达到物理极限的平衡。

3.1.2 剪枝

研究背景

剪枝 (Pruning) 旨在通过移除神经网络中冗余或不重要的参数，从而减少模型的存储需求并降低计算复杂度，同时尽可能保持原有的模型性能。

2024 年及之前,剪枝研究主要集中在非结构化稀疏上,如 SparseGPT^[290] 和 Wanda^[291]。这些方法基于“彩票假设”，试图剔除权重矩阵中绝对值较小或对激活值贡献较低的单个参数。尽管这类方法能实现较高的理论压缩比 (如 50% 稀疏度)，但在实际部署中却面临巨大的挑战：现代 GPU 的 Tensor Core 主要针对稠密矩阵乘法设计，非结构化的稀疏矩阵需要额外的索引存储与内存访问开销，导致“理论 FLOPs 减少”往往无法转化为“实际推理

延迟降低”。即便是 NVIDIA 推出的 2:4 半结构化稀疏，也对硬件有严格限制，且难以进一步扩展到更高压缩率。

进入 2025 年，研究界达成了广泛共识：剪枝的目标不再是单纯追求参数数量的减少，而是追求实际运行时间的真实加速。这一转变推动了技术路线从细粒度权重剪枝，向粗粒度（层级/模块级）结构化剪枝，以及在推理时动态跳过计算的动态剪枝方向全面演进。

研究进展

静态结构化剪枝 为了在现代硬件（如 GPU Tensor Cores）上实现真实的推理加速，2025 年的主流工作致力于构建对硬件友好的静态剪枝结构，并呈现出明显的“多粒度协同”趋势。

在宏观层面，研究者发现大模型存在显著的层级冗余。ShortGPT^[292]通过“块影响 (Block Influence)”指标证明，直接移除冗余层 (Layer Removal) 是一种极高效的粗粒度剪枝手段。BlockPruner^[293]则进一步将粒度细化至模块级，将 Transformer 层拆解为 MHA 与 MLP 独立模块，并在困惑度约束下进行启发式剔除，实现了比整层删除更优的精度权衡。

在微观层面，CFSP^[294]提出了一种“由粗到细”的结构化框架，它不仅评估模块的显著性，还深入到 FFN 层的内部维度，显式地将保留通道对齐到 128 的倍数。

这种从层 (Layer) 到模块 (Block) 再到通道 (Channel) 的全方位结构化剪枝，成功解决了传统非结构化稀疏无法转化为实际加速比的工程痛点。

动态与自适应剪枝 与静态剪枝“一次修剪，永久生效”不同，另一条重要路线致力于挖掘模型在处理不同数据时的动态敏感度，主张“按需计算”。

在权重层面，DLP^[295]打破了各层均匀稀疏化的传统，利用输入激活值动态评估每一层的重要性，对敏感层保留更多参数，而对迟钝层进行激进压缩，从而在极高稀疏度下保持了模型性能。

在推理层面，动态性体现为对 Token 的实时筛选。Long et al.^[296]基于信息扩散理论，在推理过程中动态丢弃那些信息已充分传递的中间层 Token；Fu et al.^[297]则在预填充阶段预测 Token 重要性，跳过无关上下文的 KV Cache 计算。

无论是 DLP 的自适应权重分配，还是 SlimInfer 的动态 Token 跳过，其本质都是利用数据的分布特征来动态优化计算资源的分配。

新型架构与编码 除了传统的稠密模型剪枝外，2025 年还涌现出了针对特定架构与新型编码的压缩范式。随着混合专家模型（MoE）成为主流，Lee et al.^[298]针对 MoE 架构提出了“先结构化专家剔除、后非结构化稀疏”的二阶段策略，证明了 MoE 模型的稀疏潜力远超稠密模型。

Contextual Compression Encoding (CCE)^[299]开辟了基于信息论的压缩新视角。它不再局限于物理参数的移除，而是通过上下文压缩编码框架重组参数空间的分布，在多层参数空间中寻找信息表达的“最小描述长度”。

这些工作表明，剪枝技术正逐渐从单纯的“做减法”演变为对模型架构与信息表达方式的深度重构。

表 3.2: 模型压缩与剪枝关键技术研究进展

关键技术	特点	代表工作
静态结构化剪枝	针对硬件（Tensor Core）特性优化，采用从层级、模块到通道的“由粗到细”策略，物理移除冗余结构以实现真实加速。	ShortGPT ^[292] , BlockPruner ^[293] , CFSP ^[294]
动态与自适应剪枝	根据输入数据的分布特征实时分配计算资源，涵盖权重的自适应稀疏度分配与推理过程中的动态 Token 跳过。	DLP ^[295] , SlimInfer ^[296] , LazyLLM ^[297]
新型架构与编码	针对 MoE 等稀疏架构的特化剪枝，或引入信息论视角，通过上下文压缩编码重组参数空间而非单纯移除。	Stun ^[298] , CCE ^[299]

未来展望

展望未来，剪枝技术将不再仅仅是模型训练后的“瘦身”工序，而是演变为模型全生命周期的核心优化手段。

首先，“思维链剪枝（CoT Pruning）”将成为新的研究高地。随着推理模型（Reasoning Models）的兴起，剪枝的对象将从模型权重扩展到生成的 Token 序列。未来的系统将能够识别思维链中的无效推理步骤或循环逻

辑，并在生成过程中实时“剪除”这些冗余思维，从而大幅降低推理延迟 (Thinking Time)。

其次，端侧自适应剪枝 (On-Device Adaptive Pruning) 将走向成熟。未来的模型将携带“弹性架构”，能够根据端侧设备（如手机、车载芯片）的实时热量、电量和算力负载，动态地关闭特定的层或专家模块。这种“液态模型”将彻底解决单一模型无法适配碎片化硬件环境的难题。

3.1.3 蒸馏

研究背景

知识蒸馏 (Knowledge Distillation, KD) 是一种模型压缩技术，旨在将大规模教师模型的知识迁移到参数量较小的学生模型中，从而在降低计算与部署开销的同时尽可能保留原模型的性能。

在 2025 年之前，大模型蒸馏主要分为两大流派：基于 API 的黑盒蒸馏 (Black-box Distillation) 主要依赖监督微调 (SFT)，让学生模型模仿教师生成的文本，但往往难以学到教师的泛化能力；基于 Logits 的白盒蒸馏 (White-box Distillation) 试图通过最小化 KL 散度来对齐概率分布，但在词表巨大且长尾噪声极多的 LLM 场景下，传统 KD 方法（如 MiniLLM^[300]）经常面临优化困难和计算开销过大的问题。

进入 2025 年，随着 DeepSeek-R1^[242] 等强推理模型的出现，蒸馏的范式发生了根本性转移：核心目标从单纯的“概率拟合”转向了“推理能力迁移”，同时，学术界开始重新审视大模型蒸馏中的损失函数设计与模型容量匹配定律，试图打破“教师越强学生越强”的简单假设。

研究进展

思维链数据蒸馏与表征机理 Guo et al.^[242] 的发布确立了“数据蒸馏”在 2025 年的核心地位。与传统的 Logits 蒸馏不同，R1 的蒸馏通过让学生模型在包含长思维链 (Chain-of-Thought) 的合成数据上进行监督微调，直接学习教师模型的思考过程。这一发现引发了开源社区（如 DeepSeek-R1-Distill-Qwen 系列）的爆发，使得 7B/8B 级别的小模型在数学和代码任务上达到了以往 70B 模型的水平。

更为深刻的是，Baek et al.^[301] 从表征学习的角度揭示了这一过程的内在机理。通过使用稀疏交叉编码器 (Sparse Crosscoder) 分析模型内部激活状态，研究发现蒸馏后的学生模型并非仅仅是在死记硬背教师的输出词，而

是发展出了独特的推理特征方向。这表明，针对推理任务的蒸馏实际上是在学生模型内部构建了与教师类似的“推理回路”，而非简单的文本复制。

Logits 蒸馏与损失函数重构 尽管 SFT 在推理任务上表现优异，但为了找回模型对“不确定性”和“细微语义”的感知，2025 年的研究重点转向了对 Logits 蒸馏损失函数的精细化重构。Logits 重蒸馏^[302] 被证明是提升 SFT 后模型泛化能力的有效手段，即在学习了 CoT 逻辑后，再通过 KL 散度损失微调一轮 Logits。

针对传统 KL 散度在 LLM 上的缺陷，Wu et al.^[303] 重新审视了前向 KL (Forward KL) 与反向 KL (Reverse KL) 的作用。与其传统的“Mode-seeking”假设不同，研究发现在有限训练周期内，两者分别关注分布的头部与尾部。基于此提出的 Adaptive KL (AKL) 能够根据当前分布动态调整方向，显著提升了蒸馏效率。

与此同时，Li et al.^[304] 指出，LLM 巨大的词表中包含了大量长尾噪声，强行对齐整个词表是有害的。其提出的 BiLD (Bi-directional Logits Difference) 损失不再盲目匹配所有概率，而是通过计算 Top-k Logits 之间的双向差值，专注于对齐关键 Token 的相对排名关系。这种方法有效过滤了长尾干扰，使得蒸馏过程更加聚焦于核心语义。

模型容量定律与自适应策略 “教师模型越大越好”的传统认知在 2025 年被打破。Zhang et al.^[305] 提出了容量差距定律 (Law of Capacity Gap)，指出学生模型的最佳性能与教师模型的规模之间存在特定的线性缩放关系。当师生能力差距过大 (Capacity Gap) 时，蒸馏效果反而下降。这一理论指导我们应根据学生模型的规模（如 7B）选择“大小最合适”而非“最大”的教师。

为了弥补或利用这种差异，自适应策略成为主流。ACoTD^[306] 提出了一种因材施教的策略，根据问题的难度动态调整教学内容的粒度 (Short CoT vs. Long CoT)，避免小模型因过拟合长思维链而导致逻辑混乱。此外，Yan et al.^[307] 和 Feng et al.^[308] 提出了多教师融合框架，利用 Wasserstein 距离对齐不同架构模型的 Token 分布，使得一个学生模型可以同时吸收“数学专家”和“代码专家”的优势，实现了能力的积木式组合。

表 3.3: 大模型蒸馏关键技术研究进展

关键技术	特点	代表工作
思维链数据蒸馏与表征机理	从模仿概率分布转向迁移推理过程，通过长思维链（CoT）数据重塑学生模型的内部表征，使其习得推理回路。	DeepSeek-R1 ^[242] , Reasoning Rep. ^[301]
Logits 蒸馏与损失函数重构	针对 LLM 词表巨大的特性，通过双向 Logits 差值（BiLD）去除长尾噪声，或利用自适应 KL 散度（AKL）动态平衡头尾部分布。	BiLD ^[304] , Rethinking KL ^[303]
模型容量定律与自适应策略	揭示师生模型规模的最佳线性缩放定律（Capacity Gap），或根据任务难度动态调整教学内容的粒度（ACoTD）以避免能力错配。	Capacity Gap ^[305] , ACoTD ^[306]

未来展望

展望未来，蒸馏技术将向“隐式化”和“超级对齐”方向演进。随着推理模型的发展，未来的蒸馏可能不再要求学生模型输出显式的思维链，而是通过隐式推理蒸馏（Implicit Reasoning Distillation），将教师的推理过程内化为学生模型的深层参数直觉，实现“思考在内，言简在外”。

同时，蒸馏将成为解决“超级对齐（Superalignment）”问题的关键路径。我们可能需要利用弱模型来监督或蒸馏强模型的特定部分，从而在人类无法完全理解超大模型行为的情况下，依然保持对其安全性的控制。

3.2 模型加速

进入 2025 年，随着大语言模型在各行各业的规模化落地，核心矛盾迅速转移到了推理成本、服务延迟与长上下文处理能力的三角制约上。随着模型参数量突破万亿级别以及上下文窗口普遍扩展至百万量级别，推理系统的性能瓶颈已从单纯的算力（FLOPS）约束彻底转向了显存带宽（Memory Wall）与容量（Capacity Wall）的双重约束，原本传统的“以算力换智能”的粗放式部署模式已难以为继。

本小节将对 2025 年“模型加速”领域的两大关键技术：投机解码和 KV Cache 进行综合调研和分析。报告揭示了两个显著的技术趋势：第一，投机解码正经历深刻的范式重构：训练上从特征模仿走向训练时测试（TTT），验证上则从严格的无损匹配演进为语义优先的质量评估，并逐渐摆脱对草稿模型的依赖；第二，KV Cache 管理正从通用的比特量化更深入化，并引入基于注意力内在机制（稀疏/低秩）的结构化压缩，并首次针对推理模型（Reasoning Models）进行了专用优化。从算法层面的投机解码范式革新，到存储层面的 KV Cache 精细化管理，整个技术栈正在经历一场深刻的蜕变。

3.2.1 投机解码

研究背景

长久以来，大模型的推理速度受限于其自回归的生成方式。每生成一个词元，系统都需要将模型的所有权重从高带宽内存（HBM）加载到计算单元。对如 Llama-3-70B 这样的大模型而言，单批次推理的算术强度极低，GPU 大量算力闲置，整体处于典型的内存受限（memory-bound）状态。

投机解码（Speculative Decoding, SD）在 2023-2024 年逐渐成熟。其核心思想是引入一个轻量级的草稿模型快速生成多个候选词元，再由目标模型在一次并行前向传播中验证这些候选。但在 2025 年之前，传统投机解码仍面临明显瓶颈：草稿模型能力有限导致接受率受限，严格的分布匹配约束降低了灵活性，而异构词表问题也阻碍了方法的通用化。

进入 2025 年，研究不再仅依赖更强的草稿模型，而是从根本上重构投机解码的交互机制与训练目标。

表 3.4: 投机解码关键技术研究进展

关键技术	特点	代表工作
训练时测试	训练阶段显式模拟真实推理过程，直接对齐接受率与并行验证效率。	EAGLE-3 ^[309]
宽松验证	放松逐词元概率一致性约束，采用语义级或不确定性感知验证。	Judge Decoding ^[310] , FLy ^[311]
异构词表解码	解除词表一致性假设，通过字符串级对齐支持异构模型投机解码。	SLEM / TLI / SLRS ^[312] , TokenTiming ^[313]
去草稿模型	无需外部草稿模型，利用目标模型历史生成信息或隐式结构实现自我加速。	Token Recycling ^[314] , SuffixDecoding ^[315]

研究进展

训练时测试 在 2025 年的诸多研究中，如何训练更好的草稿模型是投机解码领域最具热点的研究之一。传统投机解码方法主要依赖特征预测，即训练草稿头去预测目标模型下一层的特征表示。这一路径存在两点核心局限：其一，模仿目标模型的内部表示本身比直接预测词元更困难；其二，仅利用顶层特征，未能充分挖掘深层网络中丰富的语义信息。

EAGLE-3^[309] 摒弃了特征预测范式，转而直接进行词元预测，并引入了训练时测试（Training-Time Test, TTT）机制。TTT 不再采用静态监督训练，而是显式模拟真实的投机推理过程：草稿模型在训练阶段生成树状候选序列，系统依据目标模型的接受结果构造损失并反向传播。这种方式迫使草稿模型直接优化整体吞吐量，而非单纯降低困惑度。同时，EAGLE-3 通过多层特征融合（Multi-Layer Feature Fusion），联合利用目标模型的浅层、中层与深层特征。

宽松验证 传统投机解码强调无损（lossless），即要求草稿模型与目标模型在概率分布上严格一致。这一约束带来了典型悖论：即便草稿模型生成了语

义等价、质量同样优秀的词（如“开心”与“高兴”），目标模型仍可能因细微的概率差异而拒绝接受。

Judge Decoding^[310]提出了一种更为激进的范式转变：以轻量级的裁判 (Judge) 模块替代严格的概率一致性验证。该裁判通常是基于 LLM-as-a-judge 微调的小型分类器或线性层，其评判标准不再是“目标模型是否会生成该词元”，而是“该词元是否构成高质量的续写”。

FLy^[311]引入了一个基于目标模型输出分布熵值的门控机制：当目标模型的输出分布熵值低于某一阈值时，使用精确匹配验证；当熵值较高时（例如在开启一个新句子或进行创意描写时），激活松弛验证策略，允许草稿标记与目标标记不一致，只要它们在语义空间上足够接近。

这一类方法实质上将生成控制从概率分布对齐转向语义质量优化。在代码生成、创意写作等开放性任务中，放松分布约束显著提升了接受率与整体吞吐量。

异构词表解码 长期以来，投机解码的一项刚性约束是草稿模型必须与目标模型共享完全一致的分词器。这一限制显著压缩了方法的灵活性，例如无法使用高度优化的 RWKV 或 Mamba 小模型来加速 Llama 系列模型，仅因其词表不一致。

Timor et al.^[312]提出了三类核心算法来解决这个问题：(1) SLEM (String-Level Exact Match)：将草稿模型生成的词元解码为字符串，再使用目标模型的分词器重新分词；只要生成文本在字符串层面完全一致，即视为通过验证。(2) TLI (Token-Level Intersection)：将草稿模型的采样空间动态限制在两个词表的交集上，并通过重归一化保证概率上的无损性。(3) SLRS (String-Level Rejection Sampling)：在字符串层面进行拒绝采样，相比 SLEM 具有更高的理论接受率上限。

TokenTiming^[313]通过利用动态重编码与对齐技术（如动态时间规整 Dynamic Time Warping）在不同词表之间构建无损 logits 映射，从而消除草稿模型与目标模型必须共享词表的限制，使任意 off-the-shelf 模型对可作为草稿模型而无需重新训练。

去草稿模型 投机解码其对“草稿模型”的依赖逐渐暴露出额外训练成本高、系统复杂度增加等问题。为此，去草稿模型 (draft-free) 推理加速成为一个新兴研究方向，核心思想是在不引入任何辅助模型的前提下，直接挖掘目标模型自身在历史生成过程中所隐含的结构化预测信息，实现自我加速。

Token Recycling^[314]提出了一种利用历史生成信息实现自我加速的方法。该方法在解码过程中动态维护一个邻接矩阵，记录词元之间的转移概率与共现关系。生成草稿时，系统不再执行任何神经网络计算，而是在该邻接矩阵上进行广度优先搜索（BFS），快速构建草稿树（draft tree），并通过树状注意力（tree attention）在一次前向传播中完成整棵树的验证。

SuffixDecoding^[315]通过基于先前生成输出构建后缀树（suffix trees）来预测候选词元序列，从而在无需辅助草稿模型或额外解码头实现 speculative decoding。实验表明该方法在多种任务上与传统模型草稿方法具有竞争性能，并显著减少了推理延迟。

未来展望

展望未来，投机解码将不再作为一种“外挂式”组件存在，而是逐步内化为大模型的基础能力。DeepSeek-V3 所采用的多词元预测（Multi-Token Prediction, MTP）训练目标已清晰地预示了这一趋势：模型在预训练阶段即被优化为同时预测未来多个词元，从而天然具备自我投机（self-speculation）能力，无需额外引入草稿模型。

与此同时，宽松验证的一系列研究的成功表明，未来推理系统将不再拘泥于严格的无损约束，而是允许用户在保真度与速度之间进行动态权衡。推理加速的目标也将从分布一致性转向更贴近人类偏好的语义质量，加速范式由“无损验证”演进为“语义优先”的智能控制。

3.2.2 KV Cache

研究背景

Key-Value (KV) Cache 是自回归 Transformer 推理阶段的核心状态缓存机制。在生成第 t 个词元时，模型会将历史所有词元在每一层注意力中的 Key 与 Value 表示缓存下来，从而在后续解码中避免对已生成上下文重复计算。这一机制将单步解码的计算复杂度从 $O(t^2)$ 降低至 $O(t)$ ，是当前大模型推理效率的基础保障。

随着 2024 年下半年 128k 乃至 1M 上下文窗口成为标配，Key-Value Cache 的显存占用成为了制约系统吞吐量的首要因素。对于一个 1M 上下文的请求，仅 KV Cache 就可能占用数百 GB 的显存，远超单卡 A100/H100 的容量。这导致批次大小被迫降至 1，且在解码阶段 GPU 算力几乎完全闲置，形成了极端的 I/O 瓶颈。

整体来看，2025 年的 KV Cache 优化呈现出清晰的分层技术结构：底层通过极低比特量化、稀疏化与低秩分解压缩 KV Cache 的存储成本，中间层针对推理模型的访问模式设计更合理的缓存保留与驱逐策略，高层则引入 GPU-CPU 协同的系统级卸载机制以突破显存容量上限，共同构成覆盖表示、调度与存储全链路的 KV Cache 优化方案。

表 3.5: KV Cache 关键技术研究进展

关键技术	特点	代表工作
KV Cache 量化	采用极低比特量化压缩 KV Cache，利用跨层冗余或旋转变换缓解精度退化问题。	XQuant ^[316] , RotateKV ^[317] , MiniKV ^[318]
稀疏与低秩压缩	挖掘注意力机制中的稀疏性与低秩结构，降低 KV Cache 的有效维度。	RazorAttention ^[319] , Palu ^[320] , ThinKV ^[321]
GPU-CPU 协同	将 KV Cache 分层存储于 GPU 与 CPU，通过算法优化减少跨设备通信开销。	ShadowKV ^[322] , SpecCache ^[323]
推理模型优化	针对推理模型的 KV 访问模式，引入语义去重或时间重要性驱逐策略。	R-KV ^[324] , LazyEviction ^[325]

研究进展

KV Cache 量化 随着上下文窗口扩展至 128k 甚至 1M 级别时, KV Cache 的显存占用在长文本推理中迅速超过模型权重本身。量化方法是一种通用的模型压缩技术，广泛运用权重压缩，激活值压缩等，同样也用 KV Cache 的压缩。2025 年的 KV Cache 量化主流方案已从 INT8 激进转向 2-bit 甚至 1.x-bit。

XQuant^[316]利用了 Transformer 相邻层 KV Cache 的高度相似性，提出“跨层压缩”与“无数据校准”技术，实现了 Sub-1.4-bit 的 KV 存储，且无需重新训练。

RotateKV^[317]针对离群值（Outliers）破坏量化精度的问题，在量化前

引入了旋转矩阵 (Rotation Matrix)，将离群值的能量“平摊”到所有维度，从而在 2-bit 精度下保持了长上下文的检索能力。

MiniKV^[318]则进一步结合了量化与驱逐 (Eviction) 策略，在压缩 86% 显存的同时保留了 98.5% 的精度，成为处理超长 RAG 任务的关键技术。

稀疏与低秩压缩 除了量化的通用压缩方法以外，针对基于传统 MHA/GQA 架构训练的模型，2025 年提出了一系列重点从注意力内在机制（稀疏性与低秩结构）入手优化 KV Cache 的高效的后处理压缩方法。

RazorAttention^[319]关注注意力头级别的稀疏性，区分少量需要访问全上下文的检索头 (retrieval heads) 与大量仅关注局部信息的普通头。通过对检索头保留完整 KV Cache、对其余头进行激进剪枝，该方法在无损精度下实现了超过 70% 的存储压缩。

Palu^[320]采用后训练低秩分解 (post-training SVD)，通过奇异值分解降低 KV 矩阵的有效维度。其结果表明，即使在已训练模型中，KV Cache 仍具有极低的内在秩。通过缓存低秩投影后的状态并结合 4-bit 量化，Palu 实现了极高的压缩率。

ThinkKV citepramachandran-et-al-2025-thinkv 基于通道稀疏性假设 (channel sparsity hypothesis)，观察到 Key 矩阵中仅有少数通道在注意力计算中长期活跃。其提出的思维自适应 (thought-adaptive) 剪枝策略可根据推理阶段动态保留显著通道并剔除冗余通道，在几乎不影响推理逻辑链的前提下，将显存占用降低超过 50%。

GPU-CPU 协同 当 KV Cache 的表示压缩接近理论极限、仍无法支撑 1M 以上上下文时，通过分层存储将 KV Cache 从 GPU 显存扩展至 CPU 主存成为必然选择。该类方法的核心思想是在保证注意力计算正确性的前提下，仅将真正参与输出计算的少量信息保留在 GPU 上，其余状态按需跨设备访问。

ShadowKV^[322]提出了一种巧妙的 GPU-CPU 协同机制。利用了注意力计算中“Key 全量访问、Value 稀疏访问”的不对称性。其在 GPU 显存中仅保留低秩压缩后的 Key，将完整的 Value Cache 卸载至 CPU 内存。推理时，GPU 先基于压缩 Key 计算注意力分数并选取 Top-K 关键词元，再按索引从 CPU 拉取对应的 Value 向量，将跨设备通信量从 $O(N)$ 降至 $O(K)$ ，从而使上下文长度突破显存容量限制。

SpecCache^[323]提出了一种基于预测的 GPU-CPU 协同机制，其主要思

路是：完整 KV Cache 卸载到 CPU 主存，在 GPU 上仅保留一个低精度（如 1-bit/2-bit）KV Cache 副本作为提示。在推理过程中，利用该低精度副本以及一个推测词元提前估计下一步最可能需要注意力访问的 KV 对。基于这些估计，GPU 可以提前从 CPU 主存预取相关的完整 16-bit KV 对。系统可以实现计算与跨设备数据传输的并行化调度，降低了跨设备通信量。

推理模型优化 2025 年，随着 DeepSeek-R1 等推理模型（Reasoning Models）的出现，KV Cache 的访问模式由局部、单调衰减转向跨时间回溯与反复重访，使得基于最近最少使用（LRU）的传统驱逐策略不再适用。

R-KV^[324]针对 CoT 过程中的语义冗余。推理模型往往会反复复述前提或进行自我修正。R-KV 通过实时计算语义相似度，识别缓存中的重复信息并进行在线合并（Deduplication）。这种“语义级去重”不仅节省了显存，还意外地提升了推理的鲁棒性，在数学推理任务中实现了 6.6 倍的吞吐量提升。

LazyEviction^[325]引入了时间重要性追踪。它不再仅仅看词元的注意力分数，而是追踪词元被“重访（Re-visit）”的频率和间隔。那些历史上反复被关注的词元（通常是关键实体或逻辑公理）会被锁定保护，防止因暂时不活跃而被错误驱逐。

未来展望

未来的 KV Cache 将呈现“弹性化”特征。系统将不再分配固定的显存块，而是根据每一层、每一个头的实时注意力熵（Attention Entropy）动态调整压缩比。同时，随着 GPU-CPU 协同类技术的标准化，GPU 显存与 CPU 主存将在推理框架层面实现透明融合，为用户提供“单机千万长文”的体验。

3.3 开源部署框架

随着大语言模型（LLM）技术的飞速发展，模型的推理效率、部署灵活性及场景适配性已成为制约其落地应用的核心瓶颈。从大规模云端服务到个人本地设备，从高并发吞吐需求到轻量化离线部署，不同场景对 LLM 推理框架提出了多样化的要求。为此，业界涌现出一批各具特色的开源推理与部署框架，它们通过创新的内存管理、编译优化、调度策略等技术，针对性解决不同场景下的效率、成本与易用性问题。本节将系统梳理 vLLM、SGLang、llama.cpp、Ollama、TensorRT-LLM、LMDeploy 六大主流框架的核心技术、

创新亮点及适用场景，深入剖析各框架的优势与定位，为大模型的高效部署与应用选型提供参考，如表 3.6。

表 3.6: 主流 LLM 部署框架技术特性与适用场景对比

部署框架	核心技术优势	典型应用场景
vLLM	支持动态批处理与多 GPU 扩展，首词响应时间表现优异	企业级高并发应用、高流量 API 服务、实时对话系统
SGLang	高并发场景稳定吞吐能力强，资源利用率高	企业级高并发应用、大规模批量推理、持续高负载服务
TensorRT-LLM	低延迟优化能力突出，推理性能极致优化	企业级高并发应用、生产级实时响应系统、关键业务推理服务
LMDeploy	针对昇腾等国产硬件深度适配，多模态支持能力强	国产硬件部署、国产服务器推理服务、视觉-语言混合任务
Llama.cpp	零 GPU 依赖，轻量化推理，资源占用低	个人开发与本地原型、无 GPU 环境测试、物联网设备基础推理
Ollama	安装流程简化，跨平台兼容性好，冷启动速度快	个人开发、本地实验验证、小型原型开发

3.3.1 vLLM

简介 vLLM^[326]是一个专注于优化大语言模型（LLM）推理和服务的高性能开源库。其核心目标是显著提升 LLM 服务在大规模计算环境下的效率和吞吐能力。传统 LLM 推理框架在处理并发请求和管理庞大的键值缓存（KV Cache）时，面临着内存碎片化、利用率不高以及批处理效率低下等挑战。vLLM 通过引入创新的内存管理技术和调度策略，旨在从根本上解决这些问题，使得在相同的硬件条件下，能够服务更多的用户请求，降低部署成本，并提升用户体验。vLLM 的强大性能和广泛认可源于其一系列创新特性：

- PagedAttention 机制: 这是 vLLM 的核心内存管理技术。它借鉴了操

作系统中虚拟内存和分页的思想，将 LLM 推理过程中产生的键值缓存（KV Cache）划分为固定大小的“页”（Blocks）。这种方式有效地解决了传统 KV 缓存管理中的内存浪费（内部碎片和外部碎片）和共享难题，使得内存利用率接近最优。

- 连续批处理: 在每个模型推理步骤之后，调度器会动态地组合当前所有待处理的请求，将已完成的请求及时移出，并允许新的请求加入，从而极大地提高了 GPU 的利用率和系统的整体吞吐量。
- 分布式推理: vLLM 支持多种并行化策略以在多 GPU 或多节点环境下运行大型模型，包括张量并行、流水线并行和专家并行。这使得 vLLM 能够有效地扩展到处理那些无法在单张 GPU 上容纳的超大规模模型。

优势与不足 vLLM 具备高并发处理能力且支持横向扩展至多机多卡集群，显存利用率可达 95% 以上，能显著降低硬件成本，同时兼容多种 Transformer 架构模型且兼容性良好，还提供生产级 API 服务，易于集成到现有系统；但也存在依赖 A100、H100 等高端 GPU 导致硬件投入成本较高，代码复杂度高使得二次开发门槛较大，在极低延迟场景下表现不及 TensorRT-LLM，且分布式调度在超大规模集群中仍需进一步优化的不足。

3.3.2 SGLang

SGLang^[327]是专为复杂 LLM 应用打造的“编译器 + 运行时”框架，其依托 RadixAttention 与压缩状态机技术，有效突破了大模型在多轮对话、Agent 任务执行及结构化输出场景下的效率瓶颈。其主要的创新点有以下三点：

- 基数树注意力机制实现键值缓存自动复用：基于基数树（Radix Tree）实现键值（KV）缓存的自动复用能力。该技术通过高效复用键值缓存中指令的公共前缀，减少冗余计算量，尤其适用于用户向模型发送相似指令的场景（例如指令均以相同的系统提示开头）。
- 约束解码压缩有限状态机：当用户要求模型输出 JSON 等特定格式内容时，传统方法需通过掩码方式对令牌（Token）逐一解码。例如生成“key”: “value”这类固定结构内容时，传统方法仍需分多步完成令牌生成。SGLang 会先解析正则表达式，再构建压缩有限状态机；针对仅存在唯一后续字符的确定性路径，系统会将其合并，从而实现单次解码

生成多个令牌。该技术显著提升了 JSON 生成、正则约束生成等场景下的解码效率。

- API 投机执行: 针对 GPT-4 这类仅开放 API 接口的黑盒模型, SGLang 可通过“投机执行”技术减少 API 调用次数与令牌使用成本。例如在单条提示词中, 基于对模型输出行为的预判, 将多次潜在的 API 调用合并为一次执行。

优势与不足 SGLang 具备超高吞吐量, 在多轮对话场景下性能提升 5 倍, 且拥有极低响应延迟, 适合高并发实时响应场景, 同时具备结构化输出能力以减少后处理工作量, 采用 Python 实现使得代码简洁易懂, 还支持跨 GPU 缓存共享以减少多卡计算浪费; 但存在对多模态任务支持能力有限、生态尚在起步阶段的问题, 对 Mistralv0.3 等部分模型的优化不足, 可能导致性能不理想, 且扩展性受限于 Python 调度器, 在超大规模集群部署时可能面临挑战。

3.3.3 TensorRT-LLM

TensorRT-LLM (TensorRT for Large Language Models) 旨在解决大型语言模型在实际应用中面临的性能瓶颈问题。通过提供一系列专为 LLM 推理设计的优化工具和技术, TensorRT-LLM 能够显著提升模型的推理速度, 降低延迟, 并优化内存使用。TensorRT-LLM 的核心技术包括:

- 预编译优化: 通过 TensorRT 的全链路优化技术, 对模型进行预编译, 生成高度优化的 TensorRT 引擎文件。这种预编译过程虽然带来冷启动延迟, 但能显著提升推理速度和吞吐量。
- 量化支持: 支持 FP8、FP4 和 INT4 等多种量化方案, 通过降低计算精度减少显存占用和提升推理速度。在 FP8 精度下, TensorRT-LLM 能实现接近原生精度的性能, 同时显存占用减少 40% 以上。
- 内核级优化: 针对 Transformer 架构的各个计算模块 (如注意力机制、前馈网络等) 进行深度优化, 实现高效的 CUDA 内核。这种优化使得 TensorRT-LLM 在 NVIDIA GPU 上表现出色。
- 张量并行与流水线并行: 支持多 GPU 协同工作, 通过张量并行和流水线并行扩展模型规模, 提高推理吞吐量。

优势与不足 TensorRT-LLM 具备极低延迟且 TTFT 表现优异的特点，拥有高吞吐量，适合大规模在线服务，能够充分发挥 NVIDIA GPU 优势，性能接近硬件极限，且生态成熟，可与 NVIDIA 整个 AI 生态无缝集成；但仅限 NVIDIA CUDA 平台，跨平台部署存在局限，预编译过程可能带来较长的冷启动延迟，对 AMD 或国产芯片等非 NVIDIA GPU 支持有限，其定制化优化能力也不如开源框架灵活。

3.3.4 LMDeploy

LMDeploy^[328] 由 MMDeploy 和 MMRazor 团队联合开发，是涵盖了 LLM 任务的全套轻量化、部署和服务解决方案。LMDeploy 开发了 TurboMind 推理引擎致力于推理性能的最终优化。该工作重点解决了混合精度（Mixed-Precision）推理中的硬件利用率低和格式支持不灵活的问题。

TurboMind^[329] 提出了两套新颖的管线，并实施了四项关键优化技术：A. 硬件感知的权重打包；B. 自适应头对齐；C. 指令级并行；D. KV 显存加载流水线；从系统层面重新设计了混合精度推理的数据流，充分榨干了 GPU 的各级缓存和计算单元性能。这个强大的工具箱提供以下核心功能：

- 高效的推理：LMDeploy 开发了 Persistent Batch(即 Continuous Batch), Blocked K/V Cache, 动态拆分和融合，张量并行，高效的计算 kernel 等重要特性。
- 可靠的量化：LMDeploy 支持权重量化和 k/v 量化。4bit 模型推理效率是 FP16 下的 2.4 倍。量化模型的可靠性已通过 OpenCompass 评测得到充分验证。
- 便捷的服务：通过请求分发服务，LMDeploy 支持多模型在多机、多卡上的推理服务。有状态推理：通过缓存多轮对话过程中 attention 的 k/v，记住对话历史，从而避免重复处理历史会话。显著提升长文本多轮对话场景中的效率。
- 张量并行与流水线并行：支持多 GPU 协同工作，通过张量并行和流水线并行扩展模型规模，提高推理吞吐量。
- 卓越的兼容性：LMDeploy 支持 KV Cache 量化，AWQ 和 Automatic Prefix Caching 同时使用。

优势与不足 LMDeploy 在性能优化上展现出深厚的技术积淀，其 CPU 调度策略达到最优水平，同时通过重写大量 CUDA kernel 完成了精细化的性能打磨，且其对多模态模型的支持表现亮眼，兼容了大量多模态模型，对国内 GPU 厂商的硬件适配性也较好；但该框架受限于开发人员规模较小，相较于 vllm 和 sglang，功能体系缺失了不少关键功能，且核心基于纯 Python 实现。

3.3.5 llama.cpp

llama.cpp 是由 Georgi Gerganov 创建的轻量级推理引擎，它是基于 C/C++ 语言编码实现的 LLM 框架，支持大模型的训练和推理，专注于在本地硬件环境 (比如个人电脑、树莓派等) 上高效运行 LLM 模型。llama.cpp 的核心目标是：

- 高效推理：在资源受限的设备 (如 CPU 或低端 GPU) 上实现高效的 LLM 推理。
- 无依赖：纯 C/C++ 实现，避免复杂的外部依赖，简化部署。
- 跨平台支持：支持多种硬件架构，包括 x86、ARM、Apple Silicon、NVIDIA GPU 等。
- 开源与社区驱动：采用 MIT 许可证，鼓励社区贡献，推动技术创新。

优势与不足 llama.cpp 作为轻量级大模型部署框架，核心优势在于极强的跨硬件兼容性，支持 CPU、GPU、NPU 及边缘设备等多种运行环境，无需高端硬件即可部署大模型，且通过 GGUF 格式优化与量化技术 (4-bit/8-bit 等) 大幅降低内存占用，部署门槛极低，适合个人开发者、原型验证或边缘计算场景，同时开源社区活跃，模型适配范围广，更新迭代频繁；但短板也较为明显，其推理性能相较于 GPU 专项优化框架 (如 vllm、TensorRT-LLM) 存在显著差距，高并发、高吞吐量场景下表现不佳，对多模态模型的支持有限，功能相对基础，缺乏生产级部署所需的分布式调度、高可用等核心特性，且在高端 GPU 硬件上的算力利用率不足，难以满足大规模在线服务的需求。

3.3.6 Ollama

Ollama 是一款开源的本地大模型部署与管理工具，它极大简化了在个人电脑或服务器上运行、管理和自定义大型语言模型 (LLMs) 的流程。其

2025 年大语言模型（LLMs）进展报告

核心理念是让用户无需复杂配置，即可在本地环境中轻松体验和开发基于大模型的应用，同时保障数据隐私。其核心特点为：

- 本地化运行: Ollama 最显著的特点是所有模型直接在本地设备运行，无需依赖云端服务。可保证数据隐私安全和离线可用。
- 跨平台支持: Ollama 提供对主流操作系统 macOS、Linux、Windows 的支持。官方还提供 Docker 容器化部署方案。
- 丰富的模型生态: Ollama 支持多种热门开源模型，用户可通过简单命令快速下载和运行。

优势与不足 Ollama 的核心优势在于安装便捷，可一键部署且无需复杂配置，同时低硬件要求使其能支持消费级设备和边缘设备，数据离线运行的特性适配隐私敏感场景，加之易于上手的操作对非专业开发者十分友好，且启动速度快；但它的并发处理能力较弱，不适合大规模在线服务，扩展性和插件定制能力有限，难以满足复杂业务需求，仅支持 Llama 系列、Mistral 等文本生成类 LLM，多模态支持不足，且性能优化不够充分，在高负载场景下可能无法满足使用需求。

3.3.7 框架选型对比与适用场景分析

测试环境与模型 本实验基于 H100 GPU 构建统一测试环境，选取两类目标模型开展性能对比：一是 Qwen2.5-7B 基础模型，二是 Qwen3-8B-4-bit 量化版本。根据模型特性适配不同推理框架，其中 Qwen2.5-7B 模型测试 vLLM、SGLang、TensorRT-LLM、LMDeploy 四款主流框架；Qwen3-8B-4bit 量化模型测试 llama.cpp、Ollama 两款轻量化框架。

测试变量配置 实验以「Token 数」和「并发数」为核心变量开展设计，采用控制变量法覆盖不同负载场景，具体配置如下：测试「Token 数」时，将并发数固定为 16，Token 数分别设置为 64、256、512、1024、2048、4096；测试「并发数」时，将 Token 数量固定为 1024，并发数依次配置为 1、4、8、16、32、64。通过上述配置，模拟从低负载到高负载的真实推理场景。

性能评估指标含义 实验选取延迟类和吞吐量类两大类核心指标，全面衡量推理框架性能，各指标含义如下：

延迟类指标：

- TTFT (首字延迟): 从用户发起请求到模型输出第一个 Token 的时间, 对应模型 Prefill 阶段, 直接影响用户 “是否卡顿” 的直观感知;
- TBT (词间延迟, 又称 ITL): 首个 Token 输出后, 后续每个 Token 的平均生成间隔时间, 对应模型 Decode 阶段, 决定文本生成的流畅度;
- Latency (端到端延迟): 从用户发起请求到模型输出完整结果的总耗时, 大致等于 TTFT 与 “TBT× 生成 Token 数” 的总和。

吞吐量类指标:

- RPS (每秒请求数): 系统每秒能成功处理的完整用户请求数, 直接反映框架的并发处理能力;
- TTS (总 Token 吞吐量): 系统每秒处理的 Prompt Token 与 Output Token 总和 (区别于语音领域的 Text-to-Speech), 是评估 GPU 算力利用率的核心指标, 数值越高表明硬件资源利用越充分。

评测详细结果 评测详细结果见 3.7和 3.8。

在 Qwen2.5-7B 非量化模型测试场景中, 各框架性能呈现显著差异化, 且受 Token 长度与并发数影响较大。

综合来看, 推理框架的性能表现与模型类型、序列长度及并发数高度相关: 非量化模型场景中, TensorRT-LLM 适用于对启动效率和低延迟要求较高的中长序列场景, LMDeploy 更适配超长序列高并发的吞吐量优先场景, vLLM 适合对稳定性要求较高的通用场景, SGLang 则仅建议在短序列低并发场景中使用; 量化模型场景中, llama.cpp 在启动效率、吞吐量及稳定性上全面优于 Ollama, 是 Qwen3-8B-4-bit 模型的最优推理框架选择。后续可结合实际应用的 Token 长度分布、并发需求及资源约束, 针对性选择适配的推理框架。

适用场景分析 基于 H100 GPU 环境下 Qwen2.5-7B 及 Qwen3-8B-4bit 量化模型的性能测试结果, 不同 LLM 推理框架的适用场景差异显著。企业级高并发、大 Token 生成场景中, TensorRT-LLM 以低延迟优势适配实时响应需求, LMDeploy 在高并发大 Token 场景下吞吐量最优, vLLM 整体表现均衡, SGLang 更适配短文本高并发; 轻量化场景中, llama.cpp 性能优于 Ollama, 更适合本地低负载原型开发。结合上述性能特征, 可进一步明确不同业务场景的框架选型逻辑与优化重点。

详细常见推荐见表 3.6。

3.4 本章小结

本章围绕 2025 年大语言模型部署的核心挑战，系统梳理了从底层压缩算法、中间层推理加速到上层部署框架的全栈技术演进。

在模型压缩层面，量化、剪枝与蒸馏技术均突破了传统范式：量化向几何重构与超低位宽演进，剪枝聚焦于实际推理延时的降低，蒸馏则实现了对思维链推理能力的深度迁移；

在推理加速层面，面对长文本与强推理需求，投机解码转向语义优先的宽松验证，KV Cache 优化则通过稀疏压缩与 GPU-CPU 协同机制有效突破了显存容量瓶颈；

在部署框架层面，开源生态呈现出显著的场景化分层。vLLM 等服务端框架通过极致显存优化解决高并发需求，而 llama.cpp 等端侧框架则致力于异构硬件的广泛兼容。

总体而言，2025 年的大模型部署标志着从单纯的“算法优化”向“系统级全链路协同”的跨越。随着技术的持续演进，大模型正逐步打破资源桎梏，向着更普惠、更实时、更广泛的应用场景迈进。

2025 年大语言模型（LLMs）进展报告

表 3.7: 不同推理框架在 Qwen2.5-7B 模型上的性能对比（测试环境：H100 GPU）

推理框架	TTFT (ms)	TBT (ms)	Latency (ms)	RPS	TTS (tokens/s)
64 TOKEN			1 CONCURRENCY		
vLLM	61.04	–	6.63	0.15	308.81
SGLang	91.43	–	6.35	0.16	322.48
TensorRT-LLM	64.94	–	6.62	0.15	309.01
LMDeploy	71.95	–	6.38	0.16	320.97
256 TOKEN			4 CONCURRENCY		
vLLM	88.60	–	7.01	0.57	1168.23
SGLang	89.53	–	6.66	0.60	1229.58
TensorRT-LLM	112.33	–	6.21	0.61	1250.54
LMDeploy	81.37	–	6.73	0.59	1216.39
512 TOKEN			8 CONCURRENCY		
vLLM	167.95	7.72	7.43	1.08	2203.62
SGLang	622.19	11.51	10.52	0.76	1556.81
TensorRT-LLM	130.53	7.05	6.38	1.17	2400.60
LMDeploy	136.28	6.94	6.92	1.16	2367.94
1024 TOKEN			16 CONCURRENCY		
vLLM	299.11	7.60	8.08	1.98	4049.08
SGLang	3376.58	10.33	13.94	0.89	1831.06
TensorRT-LLM	170.71	6.31	6.63	2.06	4226.68
LMDeploy	257.81	6.94	7.36	2.17	4447.26
2048 TOKEN			32 CONCURRENCY		
vLLM	501.42	7.61	9.03	3.54	7245.75
SGLang	5741.30	9.29	19.29	0.97	1979.16
TensorRT-LLM	307.18	5.72	7.55	3.70	7578.37
LMDeploy	494.73	7.23	8.29	3.84	7869.06
4096 TOKEN			64 CONCURRENCY		
vLLM	949.18	7.69	11.97	5.34	10928.11
SGLang	10815.64	6.51	16.15	0.91	1957.83
TensorRT-LLM	508.25	2.31	10.08	1.35	10596.46
LMDeploy	983.73	–	10.58	5.87	12020.21

表 3.8: 不同推理框架在 Qwen3-8B-4-bit 量化模型上的性能对比（测试环境：H100 GPU）

推理框架	TTFT (ms)	TBT (ms)	Latency (ms)	RPS	TTS (tokens/s)
64 TOKEN			1 CONCURRENCY		
llama.cpp	85.02	–	15.84	0.06	129.21
Ollama	358.37	–	17.79	0.06	115.54
256 TOKEN			4 CONCURRENCY		
llama.cpp	142.99	–	15.83	0.06	129.34
Ollama	425.84	–	18.01	0.06	114.32
512 TOKEN			8 CONCURRENCY		
llama.cpp	255.23	14.88	15.83	0.06	129.35
Ollama	547.03	17.03	17.83	0.06	115.24
1024 TOKEN			16 CONCURRENCY		
llama.cpp	457.08	15.04	15.85	0.06	129.15
Ollama	853.51	16.60	17.86	0.06	115.06
2048 TOKEN			32 CONCURRENCY		
llama.cpp	864.40	–	15.86	0.06	129.03
Ollama	1336.59	–	18.21	0.05	112.87
4096 TOKEN			64 CONCURRENCY		
llama.cpp	1685.54	–	15.87	0.06	129.03
Ollama	2482.56	–	18.78	0.05	109.45

第四章 智能体演进

进入 2025 年, 大语言模型的发展重心正从单一模型能力的持续放大, 逐步转向以任务为中心的系统化能力构建。在这一背景下, 智能体 (Agent) 作为连接模型能力与真实世界任务的重要形态, 受到学术界与产业界的广泛关注。

相较于传统以单轮生成或静态推理为核心的应用模式, 智能体强调目标驱动、过程性决策、工具调用与环境交互, 使大模型得以在复杂、长时序任务中发挥更稳定和可扩展的作用。

随着大模型能力的持续演进, 智能体正逐步从概念验证阶段走向真实应用体系。其发展路径呈现出清晰的层次结构: 底层是以模型、记忆与工具为核心的技术能力, 中层是围绕具体任务构建的应用系统, 上层则是与行业流程深度耦合的生产力形态。

4.1 自主任务规划

4.1.1 研究背景

自主任务规划要求智能体在面对复杂任务时, 自主地探索环境, 并根据当前环境和任务状态动态制定与调整执行策略, 以实现最终目标。这一任务与具身智能、GUI/Terminal Agent 等领域密切相关, 受到了广泛关注。综合过去的研究, 我们认为大模型完成自主规划任务需要的核心能力包括:

- **全局、系统性思维能力:** 复杂、长期的任务要求智能体具备全局视野和系统思维能力, 应避免仅关注当前步骤或局部目标, 而是理解任务的整体结构和各个子任务之间的关系, 从而制定连贯且高效的执行计划。
- **环境信息管理能力:** 在探索环境的过程中, LLM 会获得大量的观察, 这些观察既包括环境状态、动作规则等关键信息, 也包括大量无意义

的内容。面对长期任务时，智能体需要从大量平凡的观察中寻找与当前目标有关的关键记录，并根据记录内容决策当前动作。

- **反思与自我纠错能力：**环境反馈是智能体执行任务过程中的监督信号。智能体能够通过分析环境反馈学习到环境运作的规则；通过解读环境的报错信号，还能够认识到当前的策略错误并加以改进。

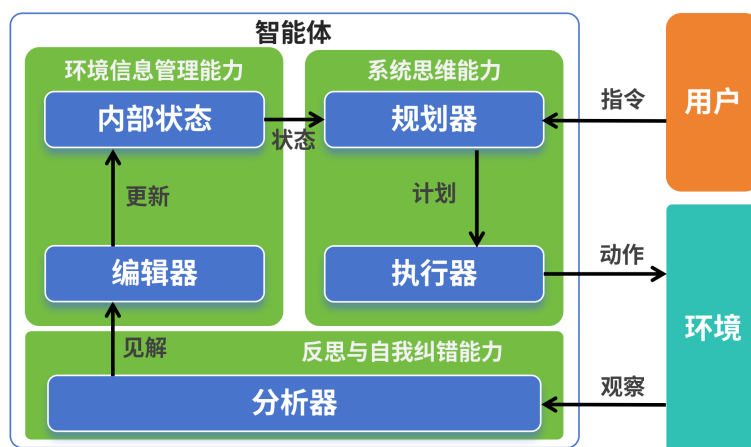


图 4.1: 自主规划智能体核心能力示意图

为了培养智能体的上述核心能力，早期的工作着眼于提示工程、思维链等手段，并发展出了 ReAct^[330]、Reflexion^[331]等典型、基础性的系统。在此基础上，我们发现最近的自主任务规划发展呈现出以下趋势：

- **基础性系统接续发展：** ReAct、Reflexion 等代表性范式仍然在不断产生新变体，表明了经典系统具有强大的泛用性与可扩展性。
- **新方法不断加入：**多智能体、记忆系统等手段也被应用到自主任务规划中，这些新方法有望进一步提升智能体的综合能力。
- **强化学习崭露头角：** PPO, GRPO 等强化学习方法被用于训练智能体策略并展现出优秀性能，表明了强化学习方法在本领域中的潜力。

4.1.2 研究进展

基于上述背景与趋势，我们以三种核心能力的培养为脉络，总结并介绍 2025 年自主任务规划的进展：

全局、系统性思维能力

培养全局思维能力的核心方法是**任务分解**，通过将复杂任务分解为子目标、子任务等多个层级，智能体能够更好地理解任务结构，并制定连贯的执行计划，避免陷入局部最优中。由于朴素的任务分解已经被充分研究，近期的工作主要关注将任务分解与其他方法结合。

- **与多智能体方法结合：**多智能体系统通过智能体间的协同工作来解决复杂问题。通过应用不同的培养策略，不同专长的智能体能够各尽其才。例如 Plan-And-Act^[332]采用了不同配方来训练 Planner(用于高层任务分解) 和 Executor(用于具体任务执行)。使用这两个智能体构造的双层框架在复杂、多步骤、长周期任务中表现出色。
- **与策略梯度相结合：**策略梯度方法提供了一种基于奖励信号优化智能体决策策略的方式。在 TextGrad^[333]首先提出基于梯度的提示优化后，ReFlexGrad^[334]针对自主任务规划领域提出了一种紧密结合了任务分解、策略梯度与自主反思的闭环系统，通过强化学习实现了卓越的零样本性能。
- **与树/图等数据结构结合：**面对复杂任务，树、图等数据结构能够更好地管理复杂任务的层级与依赖关系。HyperTree Planning^[335]提出了一种“超树”结构来进行迭代扩展并细化的推理。而^[336]将树形结构与多智能体方法结合，在动态扩展的代理树中管理子目标。

环境信息管理能力

朴素的 ReAct Agent 每步都生成思维链，并将所有的思维链——动作——观察对拼接并存储在上下文中，这种做法在长期任务中会导致上下文过长、信息冗余等问题，影响智能体的性能。为此，近期的工作从缩短思维链长度、压缩任务上下文等方面研究了该问题。此外，部分工作还研究了如何将上下文中的经验迁移至其他任务。

- **缩短思维链长度：**Learning When to Plan^[337]通过实验说明：总是执行思维链规划和从不规划都会降低智能体的性能表现。为此，他们设计了一种强化学习方法，尝试培养了智能体的“适度思考”能力并取得了良好的效果。

- **压缩任务上下文：**以 AutoManuals^[338]为代表的方法使用实体——属性列表来管理环境级的上下文信息、通过 ToDo 列表来指导智能体探索环境，最终以结构化的方式保存环境信息，实现对于探索阶段上下文的压缩。而 Context-Folding^[339]直接学习在上下文中折叠不重要信息的策略，从而有效地缩短了上下文长度。
- **跨任务记忆：**Dynamic Cheatsheet^[340]创新性地提出了一种跨任务经验 & 策略迁移的轻量级框架，使得智能体能够在不同任务间共享经验，从而提升了新任务中的适应能力与执行效率。

反思与自我纠错能力

Reflexion 等经典工作证明了反思与自我改进能够提升自主任务规划的性能。相较于 Reflexion 使用的任务级、基于 ICL 的简单反思方式，近期的工作在反思粒度、反思能力的培养等方面进行了探索。

- **细粒度、动作级的反思：**ReflAct^[341]指出朴素的 ReAct Agent 可能存在推理漂移、连锁错误等问题。为此，他们设计了一种动作级的反思机制，要求智能体在每次生成新任务前先反思当前状态。在 ALFWorld 与 SCITWorld 上的实验中，ReflAct 使用轻量级的改动收获了显著的性能提升。
- **通过蒸馏学习培养反思能力：**STeP^[342]将蒸馏学习应用于反思与自我纠错能力的培养中，通过在训练数据中加入包含反思与纠错过程的轨迹，增强了小型智能体向教师模型学习的效率，并进一步赋予其更强的反思与自我纠错能力。

4.1.3 未来展望

展望未来，我们认为自主任务规划将朝着更远视、更高效、更鲁棒的方向发展。

更远视的规划智能体能够面对更加复杂、长期的任务场景。实现这一目标的关键在于提升智能体的推理能力并设计更可靠的规划机制，使智能体能够从被动响应环境反馈转向主动预测未来状态与需求，从而制定更具前瞻性和计划性的解决方案并切实执行。

更高效的规划智能体能够在有限的计算资源与时间约束下，快速制定与调整执行计划。实现这一目标要求智能体具备更高级的信息管理系统，能够

高效保存当前任务环境中有价值的信息，并能够从历史经验中迁移有用的知识与策略。

更鲁棒的规划智能体能够在面对不稳定且动态变化的环境时，依然保持可靠的性能表现，且能够从错误中恢复。实现这一目标需要智能体具备更强的适应与自我修正能力，能够及时从上下文中识别并纠正错误，并动态调整执行策略以应对环境变化。

4.2 工具链整合

4.2.1 研究背景

如果说大型语言模型为智能体提供了“大脑”，使其具备了前所未有的理解、推理和规划能力，那么工具调用（Tool Calling / Tool Use）则为智能体装上了“手和脚”，使其能够突破自身知识和能力的边界，与外部世界进行真实、有效的交互。AI Agent 下一步的竞争点，关键就在于对外部工具的调用能力。2025 年的研究和实践深刻印证了这一点。工具调用能力的强弱，直接决定了智能体的实用性、自主性和智能水平的上限。因此，围绕工具调用的模式、协议、框架、评估和安全性的研究，构成了 2025 年 AI 智能体领域最活跃、最具创新性的篇章。

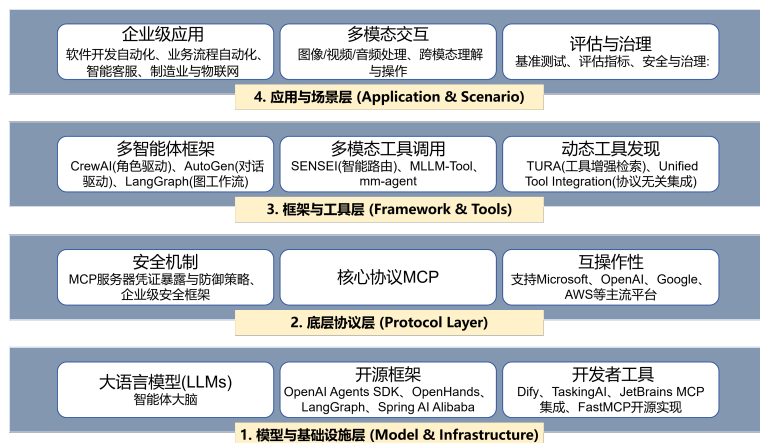


图 4.2: 工具使用智能体技术栈

4.2.2 研究进展

2025 年，工具调用技术不再是简单的“函数调用”（Function Calling）概念的延伸，而是演化出了一系列系统性的技术突破和全新的设计范式。

标准化浪潮：模型上下文协议（MCP）的诞生与影响

在 2025 年之前，AI 模型与外部工具的集成方式呈现出高度碎片化的状态。每个模型提供商（如 OpenAI、Google、Anthropic）都有自己独特的 API 格式和函数调用规范，开发者需要为不同的模型编写不同的适配代码，这极大地阻碍了工具生态的建设和智能体的互操作性。为了解决这一痛点，Anthropic 公司在 2024 年底率先提出并推动了模型上下文协议（Model Context Protocol, MCP）^[343]，并在 2025 年得到了业界的广泛响应和采纳。MCP 核心设计哲学是开放和标准化，旨在基于 JSON-RPC 2.0 的通信与消息格式，为 AI 模型（客户端）与工具/数据源（服务器）之间的通信建立统一的、与具体模型和平台无关的规范。

随着 MCP 的普及，学术界的研究重点迅速转向了安全性、性能优化和生态兼容性。Narajala et al.^[344]等研究聚焦于 MCP 服务器的凭证暴露与防御机制。Ehtesham et al.^[345]构建了针对 MCP 性能的基准测试，以衡量其在高并发下的表现。

此外，MCP 的推出迅速获得了业界的积极响应。到 2025 年中，包括微软（Microsoft）、OpenAI、谷歌（Google）、亚马逊（Amazon Web Services）在内的主要云服务 and AI 模型提供商均宣布支持或兼容 MCP。同时，一个围绕 MCP 的生态系统开始形成，例如，JetBrains 在其 IDE 中集成了 MCP 支持，苹果（Apple）在其开发者工具中也引入了相关概念，社区中涌现出如 FastMCP 等开源实现。开发者们基于 MCP 构建了各种各样的工具服务器，涵盖了文件系统操作、数据库访问、Git 版本控制、Google Drive 管理等多种功能。MCP 的普及，被认为是 2025 年智能体迎来“寒武纪生命大爆发”的关键催化剂。

多智能体协同场景下的工具调用

2025 年的另一个显著趋势是多个、功能各异的智能体协同工作的多智能体系统（Multi-Agent Systems）的出现，这种协同工作流对工具调用提出了新的要求，即工具调用的编排（Orchestration）。为了支持这种复杂的交互模式，一系列多智能体框架应运而生或得到加强：

CrewAI^[346]: 一个以角色扮演为核心的框架，允许开发者定义不同角色的智能体，并设定它们之间的协作流程，每个智能体都可以拥有自己专属的工具集。

AutoGen^[347]: 由微软推出的对话驱动的多智能体框架，智能体之间可以通过对话进行协作，共同解决问题，工具调用是其对话能力的重要组成部分。

LangGraph^[348]: 作为 LangChain 的扩展，LangGraph 将智能体工作流建模为图（Graph），每个节点代表一个计算步骤（如调用 LLM 或工具），边代表流程的走向。这种基于状态图的方式非常适合构建循环、可控的多智能体应用。

n8n^[349]: 作为一个开源的工作流自动化工具，n8n 在 2025 年也加强了对多智能体工作流的编排能力，支持任务的交接和对整个流程的可观测性。

这些框架的共同特点是，它们都提供了显式的机制来管理和编排工具调用，使得开发者可以构建出比单一智能体强大得多的复杂应用。

多模态融合

随着多模态大模型（Multimodal Large Language Models, MLLMs）技术的成熟，2025 年工具调用的一大突破是多模态工具调用的实现。智能体不再仅仅是处理文本，而是能够理解图像、视频、音频等多种模态的输入，并调用相应的多模态工具进行处理。代表性研究与框架：

SENSEI^[350]: 一个具有多模态能力的辅助 AI 代理架构。它的核心特点是拥有一个智能路由模块，可以根据输入的模态（文本、图像等）和任务的性质，动态地决定是调用 LLM 进行文本处理，还是调用视觉模型（如 CLIP）进行图像理解，或是调用搜索 API 获取信息。这种动态工具选择机制是多模态工具调用的关键。

MLLM-Tool^[351]: 该框架旨在通过多模态编码器和开源 LLM，让智能体能够理解包含视觉或音频信息的用户指令，并准确地选择和调用外部的多模态工具。例如，用户上传一张鸟的图片并提问“这是什么鸟？”，智能体可以调用一个鸟类识别 API 来获得答案。

mm-agent^[352]: 此研究提出了一种增强复杂视觉场景下智能体决策能力的多模态代理架构，重点探讨了智能体如何自主选择和组合合适的工具（如物体检测、图像分割、OCR 等）来完成视觉问答或导航等任务。

这些研究的共同之处在于，它们都试图解决一个核心问题：如何让智能

体在接收到多模态输入时，能够智能地、准确地选择并执行最合适的工具，从而实现跨模态的理解和操作。

动态工具发现与自动集成

这一领域的目标是让智能体具备“学习使用新工具”的能力，尽管 2025 年智能体技术在工具调用方面取得了巨大进展，但大多数应用场景仍然依赖于开发者预先定义和注册好的工具集。一个更具前瞻性的研究方向是动态工具发现（Dynamic Tool Discovery）和自动工具集成（Automatic Tool Integration）^[353]。结合意图感知的 MCP 服务器检索、基于 DAG 的任务规划模块和蒸馏智能体执行，使得 AI 搜索能像人类一样调用工具（API、数据库）获取最新、结构化数据实现，实现实时、动态交互的查询。

Wang et al.^[351]通过智能体指导系统，自动推荐和发现新的 API 关系，Hasan et al.^[354]探讨了 MCP 服务器在实际生产环境中的可维护性，Ding et al.^[355]研究如何以协议无关的方式统一工具集成，为未来的自动集成奠定了基础。实现这一目标需要模型具备极强的代码理解、元认知和自学习能力，是当前及未来几年的重要研究热点。

表 4.1: 2025 年发布的新兴智能体开源工具框架

开源项目	主要特性
OpenAI Agents SDK	多模型支持/简化多智能体 workflow 构建流程/内置一系列强大的工具并允许自定义工具/强大的可观测性
OpenHands	面向通用软件开发的智能体，具有高度可定制的工具集功能以及图状工具使用规则来处理更复杂的工具调用逻辑。
Tools Strands Agents Tools	提供了一套即用型的工具集，涵盖文件操作、Shell 集成、内存管理等
扩展的 OpenAI Codex	能够自动化整个 workflow，仅通过少量输入就能构建一个完整的机器学习模型或 API
LangGraph	针对传统 langchain 的“链式”结构在处理复杂、循环、有条件的任务流时显得力不从心的问题，将 workflow 抽象为一个有向无环图，使得构建复杂的、有状态的多智能体系统和需要多次工具调用的 ReAct (Reason-Act) 风格的智能体变得更加直观和可控
企业级应用框架: Spring AI Alibaba	支持循环推理与工具调用：实现了标准的 ReAct 框架。丰富的企业级特性：支持上下文工程、人工介入、消息压缩、规划、模型调用限制、工具重试以及基于 LLM 的工具选择器等高级功能。与企业生态集成：例如，它可以与 Nacos 等服务发现和配置中心结合，实现企业内部 API 工具的动态注册和调用。

开源工具使用生态系统：框架、库与开发者工具

此外，像 Dify 和 TaskingAI 这样的开源智能体平台也持续迭代，提供了更加用户友好的界面来构建和管理包含工具调用的智能体应用，进一步降低了非专业开发者的使用门槛。

性能评估：基准与指标

2025 年在智能体工具使用方面涌现了一些新的基准，同时有的主流基准也纷纷进行了迭代更新。

表 4.2: 基准说明表

分类	基准	解释
2025 年新基准	Webwalker	专门用于评测 LLM 在真实 Web 环境中的网页遍历能力
	FDABench	数据分析代理（Data Agent）基准，评估智能体在面对复杂数据查询任务时，能否正确地选择和使用 SQL 查询、Python 数据分析库等工具来生成正确的分析结果
	AgentClinic	面向临床医疗场景的多模态 AI 基准
主流基准	ToolBench、APIBench、API-Bank	这些基准构成了评估模型 API 调用有效性的基础。它们提供了大量的真实世界 API，要求模型根据自然语言指令生成正确的 API 调用代码。
	GTA（General Tool Agents）	通用的工具代理基准，可以进行多个模型之间的两两对比
	WorkBench	评估在真实职场环境中智能体的任务执行、工具选择和多步操作能力
	Berkeley Function Calling Leaderboard (BFCL)	社区公认的衡量各大模型 Function Calling 能力的重要参考

性能指标：超越成功率的多维度考量

在性能指标方面，2025 年的评估不再仅仅关注**成功率（Success Rate）**。为了更全面地刻画智能体的工具调用性能，研究人员和基准设计者引入了更

多维度的指标：

表 4.3: 智能体评估指标说明表

维度	指标	解释
效率指标	延迟	完成任务或单次工具调用所需的总时间
	交互次数	完成任务所需的“思考-行动”循环次数
成本与资源消耗指标	Token 使用量	包括 Prompt Token 和 Completion Token
	计算成本	完成任务所需的总计算资源或 API 费用估算
质量与鲁棒性指标	工具选择准确率	在多工具场景下，模型选择正确工具的能力
	参数生成准确率	模型为所选工具生成正确参数的能力
	错误恢复能力	工具调用失败后，能否分析错误并有效重试或调整

产业应用与企业级实践

2025 年，工具调用不再是理论探讨，而是真正在各行各业创造商业价值。企业级实践的广度和深度都达到了新的水平。

全球领先企业已将智能体技术从实验阶段推向大规模商业化应用。这得益于底层模型能力的增强、开发框架的成熟以及对业务场景理解的加深。《2025 年中国 AI 商业落地应用价值研究报告》等行业报告详细记录了这一趋势。

表 4.4: 智能体产业应用案例

应用领域	代表案例/工具	具体应用与调用逻辑
软件开发与 IT 运维自动化	OpenAI 的 Codex 和开源项目 OpenHands 被广泛用于自动化软件开发流程。	调用 Jira 等项目管理工具更新任务状态，调用 CI/CD 流水线工具执行构建和部署，调用代码解释器执行和验证代码片段
企业级业务流程自动化	德勤等咨询公司利用 AI 智能体为客户构建自动化财务报告、供应链管理和人力资源流程的解决方案。	核心是调用企业内部的各种业务系统 API（例如 ERP 系统、CRM 系统等）。
智能客服与数据分析	智能客服系统	调用内部数据库、知识库检索 API、以及各种业务操作 API
制造业与物联网	西门子公司在其工业自动化解决方案中应用 AI 智能体进行预测性维护。	传感器数据 API，工单系统 API 等。

4.2.3 总结与展望

回顾 2025 年，AI 智能体在工具调用模式方面的研究与应用取得了里程碑式的进展。标准化（以 MCP 为代表）为生态的互联互通铺平了道路；开源框架的爆发式增长极大地推动了技术的普及和创新；多模态能力的融合让智能体的交互维度得到前所未有的扩展；系统化的评估使得我们对智能体能力的认知更加清晰；广泛的产业落地证明了其巨大的商业价值；而对安全与治理的日益重视，则保障了这项技术的健康、可持续发展。2025 年，工具调用已经从一个单纯技术概念，演化为一个包含协议、框架、应用、评估和治理在内的完整技术生态。

展望未来，工具调用模式 AI 智能体可能在以下几个方面取得进一步的进展：真正的自主学习与发现：当前智能体仍高度依赖预定义的工具。实现真正的动态工具发现、学习和自动集成，让智能体像人类一样自主掌握新技能，将是迈向更高阶智能的关键一步。

更复杂的推理与规划：面对需要长链条、深度推理和灵活规划的复杂任务，现有智能体的“思考”深度和广度仍然有限。如何提升模型在不确定和动态环境下的长期规划能力，是一个核心难题。

伦理与社会影响：随着智能体越来越强大，其行为可能带来的伦理风险和社会影响（如失业、偏见、滥用）也愈发突出。建立健全的 AI 伦理审查、行为审计和社会监督机制，将是与技术发展并行的重要课题。

4.3 检索增强生成 (RAG)

4.3.1 研究背景

自 2020 年 Lewis 等人在 NeurIPS 会议上正式提出检索增强生成(Retrieval-Augmented Generation, RAG) 概念以来，这项旨在通过外部知识库增强语言模型能力的技术，已演变为大型语言模型 (LLM) 生态系统中不可或缺的核心支柱，成为解决 LLM 知识陈旧、产生幻觉和缺乏可解释性等核心痛点的关键方案。进入 2025 年，RAG 领域的研究范式不再仅仅关注于“是否使用 RAG”，而是深入探讨了“如何更智能、更高效、更可靠地使用 RAG”。相关研究已经从简单的“检索器 + 生成器”的串联模式，转化为可拆解、可优化、可自适应调控的复杂系统模式。一个明确的信号是，顶级会议上纯粹提出一种新型基础 RAG 架构的论文比例有所下降，取而代之的是大量针对 RAG 流程中特定模块的优化、或是将 RAG 与强化学习、多智能体等其他 AI 范式深度融合的研究，RAG 已经成为大模型技术栈的“基础设施”。

4.3.2 RAG 的全链路优化范式

RAG 流程可解构为“检索前”、“检索中”和“检索后”三个主要阶段，2025 年的最新研究针对每个阶段都提出了创新的优化策略。

检索前优化

查询重写与扩展 (Query Rewriting & Expansion)。多项研究表明，利用 LLM 自身强大的语言理解和生成能力来重写查询，可以显著提升检索的召回率和精度。Wang et al.^[356]提出了多视角反馈驱动查询重写 MaFeRw，通过对下游任务（检索 + 生成）的稠密奖励反馈，显著提升了重写质量与最终生成效果，也为 RAG 的模块协同优化提供了新范式。Amato et al.^[357]聚焦于法律领域问答这一高精度、高敏感性场景，系统性地对比了三种查询重

写策略, 子查询生成 (Multi-Query)、复杂问题拆解 (Decomposition) 和高层概念抽象 (Step-Back) 的表现。实验表明 “Step-Back” 重写策略显著优于其他方法, 在准确率和检索效率上均表现最佳。

子查询生成. 对于复杂问题, 模型会先将其分解为多个独立的子查询, 分别检索后再综合答案。Wang et al.^[358]提出了一个通用的框架, 引入了多种专家设计的水写策略, 根据不同的任务需求自适应地生成查询变体。

假设性文档生成 (HyDE). 模型首先根据查询生成一个假设性的 “理想答案” 文档, 然后使用该文档进行向量检索。其背后的逻辑是, 答案比问题更能代表相关文档在语义空间中的位置。Maarefdoust et al.^[359]聚焦于数学信息检索任务, 比较了 Clarification (澄清)、Simplification (简化)、Summarization (概念抽取) 和 Answer-as-Query (答案即查询), 结果表明利用 LLM 生成的答案本身对用户原始数学查询进行重写 (答案即查询), 在多个标准数据集上带来最显著的性能提升。

检索中优化

在核心的检索阶段, 2025 年的研究趋势是告别单一的检索策略, 转向多元化、层次化的混合检索方法。

混合搜索 (Hybrid Search). 混合检索的提出来自 2024 年的 Blended RAG 框架^[360], 通过结合 (向量) 语义搜索和基于关键词查询的检索器, 显著提升了 RAG 的准确性。Hu et al.^[361]针对不同类型的查询采用了双层检索范式: 对于具体事实使用精确检索, 对于抽象概念使用语义检索, 实现了在速度与准确率之间的最佳平衡。

多向量表示 (Multi-Vector Representation). 为了解决单个嵌入向量无法完全捕捉长文档或复杂概念多面性的问题, 研究者提出了为文档块生成多个向量的策略。Yan et al.^[362]详细讨论了多向量检索的数学形式, 针对多向量检索存储开销大的痛点, 提出了一种自适应的补丁级嵌入剪枝框架。

层次化检索. 针对大规模、结构化的知识库 (如企业内部复杂的文档体系), 研究提出了层次化检索。系统首先检索高级别的摘要或目录信息, 定位到相关的子集, 然后再在子集内进行精细化的段落检索。这种 “先粗后精” 的策略

略在处理海量文档时，能有效平衡效率和精度。Wang et al.^[363]提出一种基于属性社区的层次化 RAG，通过引入属性社区 (Attributed Communities, ACs) 和分层索引结构，实现了高效且多粒度的检索。HM-RAG^[364] 框架利用三个专门的代理 (查询分割、多模态检索、合并优化)，能够跨文本、图形和网络等多种数据类型进行检索，特别适合处理复杂的跨媒体查询。

结构化检索。 Sarmah et al.^[365]提出了图增强检索 (Graph RAG) 策略，将知识图谱 (KG) 与向量检索相结合的方法，解决了传统 RAG 在处理结构化数据时的局限性。Hu et al.^[361]提出情境化图检索增强生成 (CG-RAG)，引用基于图检索增强的 LLM 研究问答，探索了利用学术引用网络来增强模型的推理能力。Chen et al.^[366]利用知识图谱中的路径信息来增强检索，通过挖掘实体间的深层关联，显著提升了在专业领域 (如生物医学) 问答任务中的表现。

检索后优化

检索到的文档往往包含噪声、冗余和与查询不完全相关的信息。2025 年的研究在检索后处理阶段投入了大量精力，确保传递给生成器的是“高信噪比”的上下文。

重排 (Re-ranking)。 这是检索后处理最关键的一步。通过更强大的模型对检索后的文档重新排序，选出与查询最相关的文档。2025 年的趋势在于，研究者开始使用 LLM 本身作为重排器，通过精心设计的提示让 LLM 评估每个文档与查询的相关性得分。Zhang et al.^[367]针对传统的问题-答案对的语义相似度计算容易受表面相关性 (lexical overlap) 误导的问题，提出了一种通过显式检索和重排序支持证据 (supporting evidence) 来提升答案排序准确率的新架构。Kardan et al.^[368]系统评估了在时序问答中，如何利用答案的时间戳信息对检索到的答案进行重排序，以提升回答准确性。

上下文压缩 (Context Compression)。 通过压缩检索后文档来减轻大模型器处理长上下文的负担并减少干扰。Chen et al.^[369]针对 RAG 中面临的长上下文处理瓶颈，提出了一种极端高效的上下文压缩方法，将整个检索到的文档集合压缩为单个信息浓缩 token，使大模型能利用海量外部知识而零增加输入长度。Hwang et al.^[370]通过选择句子级别的关键事实构建精简但信息完整的压缩文档。

4.3.3 自适应与自主 RAG

2025 年 RAG 研究的另一个趋势是从一个被动式工具向一个主动式智能系统的演进，即使其具备初步的判断和决策能力，先判断必要性，再决定是否迭代（即“自适应迭代”）或者是否检索，在决策过程中往往尝试与推理深度融合，实现推理需求驱动的检索。

迭代式检索 传统的 RAG 通常执行一次性的检索，但这在处理需要多步推理或信息探索的复杂问题时显得力不从心。2025 年的一个标志性进展是迭代式 RAG 框架的成熟。在这种模式下，模型不再试图一次性生成完整答案。而是通过“检索-生成-评估-再检索”的循环，直到模型认为答案已经足够全面和准确。这种方法可以视作 RAG 与反思智能体的融合。

Feng et al.^[371]提出自适应迭代检索框架，根据问题复杂度动态决定是否进行多轮检索。同时引入不确定性感知模块，判断首轮检索结果置信度低，则触发第二轮更精细的检索，包括扩大查询范围或切换检索器等。

Guan et al.^[372]将 RAG 建模为马尔可夫决策过程（MDP），使用强化学习框架统一建模迭代与自适应行为。

Yu et al.^[373]提出了双路径迭代检索策略，包括基于语义相似度的传统稠密检索，和基于图结构的实体超边扩展。通过 GRPO 强化学习算法实现了端到端的优化检索-生成联合策略。KnowTrace^[374]聚焦于多跳问答推理任务，提出了一种新型迭代式检索增强生成框架，并引入自举式训练机制（Self-Bootstrapping），将多跳推理转化为可追踪、可学习、可优化的知识扩展过程，显著提升了模型在开放域多跳推理中的准确率与可解释性。

Lee et al.^[375]以提升事实准确性，抑制幻觉为目标进行多轮迭代检索，每轮生成后，模型输出推理链，并从中提取待验证子句，作为下一轮检索的查询，迭代直至所有关键事实被检索证据覆盖或达到最大轮次。该方法将“推理”与“检索”深度耦合，形成闭环验证，在 AVeriTeC 和 FEVER 上的幻觉率下降 12.4%。

自适应检索决策 并非所有问题都需要检索。对于常识性问题或模型内部知识已经足够回答的问题，进行不必要的检索会浪费计算资源并可能引入噪声。2025 年的自适应 RAG 或条件式 RAG（Conditional RAG）研究，正是为了解决这个问题。

Wu et al.^[376]提出了一种自路由检索增强生成（Self-Routing RAG）框架，融合了参数化知识（即知识口头化机制生成的模型内部知识）与外部知

识源，让大语言模型（LLM）在生成答案前主动决定是否需要外部检索、以及从哪个知识源获取信息。ExpertRAG^[377]没有采用复杂的迭代或反思机制，引入 MoE 架构实现按需、高效、精准的知识路由，为构建大规模、多领域、生产级 RAG 系统提供了新范式。

UltraRAG v2^[378] 是基于 MCP 架构设计的轻量级 RAG 系统，将 RAG 中的核心组件（如 Retriever、Generation 等）标准化封装为独立的 MCP Server，并通过函数级 Tool 接口实现灵活调用与功能扩展。

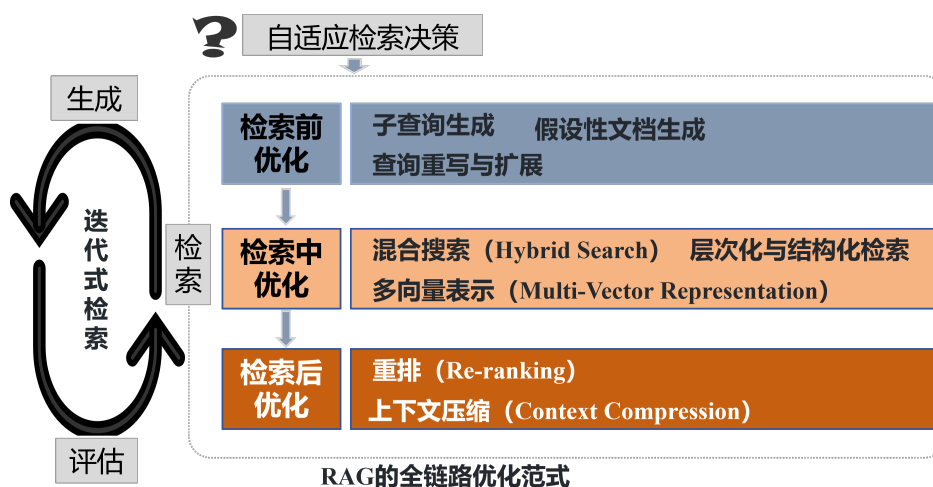


图 4.3: RAG 的全链路优化范式

4.3.4 多智能体 RAG（Multi-Agent RAG）

受 AI Agent 研究浪潮的启发，2025 年出现了将 RAG 与多智能体系统结合的趋势，一个复杂的 RAG 任务可以被分解，并分配给具有不同角色的智能体协同完成，这种由专门优化的模型分工协作的模式，使得 RAG 系统能够高精度地处理更复杂的任务流。

Singh et al.^[379]首次系统定义了 Agentic RAG 架构和理论框架，提出“检索-评估-生成-反思”四阶段协作框架。论文将多智能体 RAG 分为三类：流水线型（Pipeline Agents）、辩论型（Debate Agents）、协作型（Collaborative Agents），并提出统一评估基准 RAG-AgentBench。

Nguyen et al.^[380]提出了多智能体协同结合失败模式感知机制模式，智能体 Explorer 执行检索增强、Verifier 验证检索片段的事实一致性、Synthesizer 融合可信片段生成最终答案。Liu et al.^[364]提出了首个分层多智能体多模态

RAG 系统，底层进行视觉与文本检索智能体并行处理多模态数据；中层通过跨模态对齐智能体判断图文相关性；顶层通过决策智能体融合多源证据生成答案。在 MMBench 和 DocVQA 上达 SOTA。

4.3.5 多模态 RAG

2025 年是多模态 RAG (MM-RAG) 全面爆发的一年。研究者们成功地将 RAG 的能力从单一的文本模态，扩展到了一个包含文本、图像、视频、音频甚至表格的混合信息空间。MM-RAG 的检索策略远比纯文本 RAG 复杂，需要处理模态间的复杂关系，2025 年的跨模态检索已不再是简单的“图文匹配”，而是朝着细粒度、可解释、可验证、可组合的方向深度演进。Yue et al.^[381]将 MLLM 直接用作多模态检索器，实现跨模态对齐和检索，打破了传统的 MLLM 仅用于生成的范式，体现了 LMM 本身在 RAG 流程中扮演更核心角色的趋势。

Lin et al.^[382]是 MM-RAG 基础模型之一，提出首个支持细粒度跨模态检索的统一编码器，可同时处理文本、图像、表格的任意组合查询。在跨模态对齐方面，^[383]充分利用 MLLM 的生成能力主动生成语义锚点，引导跨模态对齐。对于图像数据和文本查询，MLLM 分别生成描述性短语和关键词集合，从而实现二者的语义桥接。

Tian et al.^[384]引入基于自然语言推理的一致性评分来判断图文模态是否逻辑一致，解决跨模态检索中的“知识冲突”问题，提升 MM-RAG 可信度。

Zhu et al.^[385]提出了一种基于子维度分解的新多模态检索机制，将复杂查询分解为子查询，对每个子查询独立检索，再融合最优视觉特征。将跨模态检索从“整体匹配”推向“成分级对齐”，为 MM-RAG 提供“分而治之”的新思路。

Zhang et al.^[386]提出了首个专为文本-表格异构文档设计的 RAG 系统，将表格转换为 HTML-like 序列，使用 LLM 为每行/列生成，在此基础上，结合稀疏与稠密检索实现跨模态检索，从而实现了文本与表格的统一表示、联合检索、多步推理的 RAG 框架，显著提升了异构文档问答的准确性。

4.3.6 总结与展望

2025 年是检索增强生成（RAG）技术从“青春期”迈向“成熟期”的关键一年。尽管在长上下文处理、实时性、评估成本等方面仍面临挑战，但

RAG 作为连接大型语言模型与海量外部知识的桥梁，其在提升 AI 系统可靠性、可解释性和知识时效性方面的核心价值已愈发凸显。全年的研究进展清晰地表明，RAG 已经摆脱了单一、固化的框架，演化为一个高度模块化、可动态配置的复杂智能系统。以“高级 RAG”的精细化优化为基础，以“自适应与自主 RAG”的智能化决策为突破，以“多模态 RAG”的跨域扩展为增长点，并以“新一代评估体系”的建立为支撑，共同构成了 2025 年 RAG 研究的全貌。

尽管在长上下文处理、实时性、评估成本等方面仍面临挑战，但 RAG 作为连接大型语言模型与海量外部知识的桥梁，其在提升 AI 系统可靠性、可解释性和知识时效性方面的核心价值已愈发凸显。可以预见，在未来的发展中，RAG 将从被动拼接向主动推理演进，将具备自我反思能力，形成闭环决策；同时，多模态融合与极致效率优化将成为核心方向。

4.4 长期记忆

4.4.1 研究背景

当前，大语言模型（LLMs）在自然语言理解和生成方面取得了显著成就，但其应用在需要长期交互和个性化的场景（如长对话、个人伴侣、心理咨询等）时，普遍暴露出核心瓶颈：**缺乏长期记忆**。这种缺陷具体体现在以下两个方面：一是**有限上下文窗口约束**，无法直接容纳历史全部交互上下文。二是**个性化缺失**，受 LLMs **无状态性**影响，其难以整合跨会话知识，无法提供个性化和一致性的交互体验。因此，为使 LLMs 在长期复杂交互中保持连贯性并实现深度个性化，构建一套超越传统上下文窗口限制**的高效长期记忆机制**至关重要。

在早期的应用中，记忆并未被视为一个独立核心的模块，大多借鉴检索增强生成（RAG）这种通用技术，检索历史上下文信息并拼接到输入提示中来弥补上下文窗口的不足。但这样的方案在长期交互场景下，尤其是长对话过程中，缺乏对历史交互的动态提取和管理能力，限制了回复生成的效果。后续陆续出现了 MemGPT^[387]、MemoryBank^[388]、Memory³^[389]等工作，逐渐引入记忆的概念，尝试将记忆作为一个独立模块进行设计和优化。从而提升模型在长期对话、复杂多步任务等高度依赖历史信息 and 状态的交互场景中的表现。

先前的工作虽然引入了记忆的独立模块概念，并取得了初步进展，但它

们多为探索性、独立的尝试，尚未形成统一的体系架构和生态。今年，随着对 LLMs 长期交互能力的迫切需求，**长期记忆**的概念被提升至重要地位。大量高质量的研究工作和开源项目开始涌现，推动记忆系统向体系化发展。

下面，我们将介绍一下今年在 LLMs 长期记忆方面的研究进展。

4.4.2 研究进展

本部分将对 2025 年有影响力的 LLMs 记忆系统工作进行归纳整理，包含一些 star 高的开源项目，以及顶级会议接收的论文。如图 4.4所示，当前记忆系统从架构框架上可以分为三部分：记忆构建、记忆存储和记忆检索。LLMs 从对话历史中通过记忆提取与记忆管理两步操作完成记忆构建，并以不同形式存储在记忆库中。在得到输入后，记忆检索模块从记忆库中找到相关记忆供模型参考，辅助回复生成。下面，我们围绕记忆系统从记忆构建、记忆存储和记忆检索三个维度进行整理，旨在系统性地揭示当前研究的演进脉络、共性特征以及核心技术差异，方便大家更好地了解 LLMs 记忆系统。

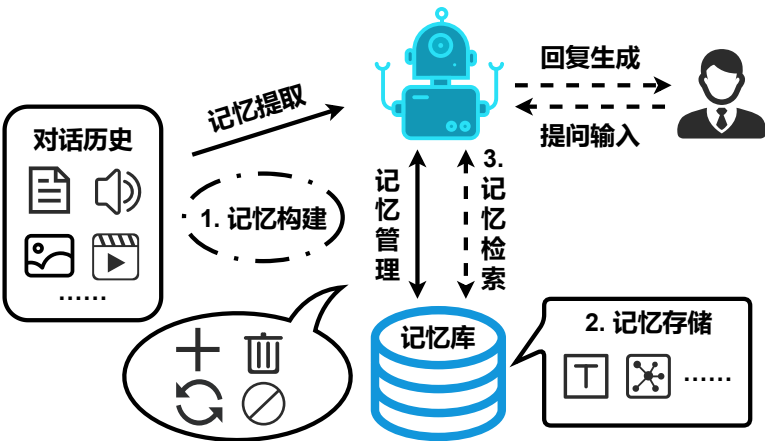


图 4.4: LLMs 记忆系统架构框架

记忆构建.

在长期复杂的交互场景下，不仅要求模型能够追溯原始的交互事件，更需要从交互中抽象出通用知识和用户特质，以支持长期的连贯性推理和深度个性化回复。于是，LLMs 需要对历史信息**深层次提炼**来构建记忆，第一步要从交互中进行**记忆提取**。这些从原始交互中提取出来的记忆共同组成了记忆库，供后续交互和推理环节使用。记忆的提取往往是通过设计合适的提示

词，利用 LLMs 的理解和抽象能力实现。根据提取出记忆的具体内容和功能，通常分为以下三大核心类别。首先是**情景记忆**，即带有时间戳、上下文信息和具体细节的原始事件和对话片段，是代理行为可追溯性和连贯性的基础（如 Zep^[390]等）。除记忆事实信息外，还有**语义记忆**，这是独立于具体事件的抽象知识、实体关系等通用信息，它构成了代理的世界观和长期知识库，支持深层次的理解与推理。最后是**用户画像**，包含从长期交互中提炼出的用户偏好、习惯等信息，是个性化服务和一致性回复实现的核心要素（如 memobase^[391]等）。

在提取出记忆后，要将新的记忆整合进记忆库。当前的系统已基本形成统一的操作范式，即**记忆管理**，分别是**增、删、改和空**四大核心操作。分别对应直接添加新记忆、删除过时冗余记忆、更新现有记忆和不进行任何操作。对于这些操作的选择，大多数系统（Mem0^[392]、MemOS^[393]、A-MEM^[394]等等）通过设计专门的提示词，由代理根据当前交互内容和记忆库信息自主决策完成。也有少部分选择通过训练的方式得到专用的记忆管理代理来获取最优的记忆操作策略（如 Memory-R1^[395]等）。对于引入记忆操作带来的额外时延与成本，LightMem^[396] 和 memobase 等工作提出了**读写解耦和批量异步处理**的工程优化方案。将计算密集型、高延迟的记忆更新（如冲突解决、去重）推迟到离线或闲置时段执行，从而显著缓解这一问题。

记忆存储.

记忆库中的记忆以不同的形式存储，这影响着记忆系统的检索性能、效率和推理能力。如表 4.5所示，当前工作引入了不同的结构性数据表示方法，具备各自的特性与优势。每个系统根据记忆内容和功能需求，可能包含一种或多种记忆表示形式的结合。

值得注意的是，当前纯文本记忆已经有了充分的发展，并且代理正朝着更复杂的、具备感知能力的现实世界交互迈进。处理和整合非文本信息的多模态记忆成为一个关键的新兴研究领域。如 MIRIX^[397] 中的资源记忆组件专门用于存储和管理文档、文件和高分辨率截图等原始多模态内容，M3-Agent^[398] 专注于处理无限长的视频和音频流，通过多模态交叉推理（如将人脸识别与语音识别关联），确保在长期记忆中实体身份的统一性和鲁棒性。

表 4.5: 记忆存储形式及其特性

记忆形式	关键特性与优势	代表工作
纯文本	直接使用文本记录信息，方便直接阅读。	大部分系统
结构化图谱	建模为节点（实体）和边（关系），支持复杂的多跳推理和关系建模。	Zep, MemOg
多模态	存储图像、音频、视频帧等非文本信息，支持跨模态推理。	MIRIX, M3-Agent
向量表示	将文本转化为高维语义向量或是隐层表示，前者是高效语义相似性检索的基础，后者则是利于节约推理成本。	MemOS
模型参数	通过高效微调方法，将特定领域或个人属性的知识提炼到模型参数中。	MemOS

记忆检索.

在使用记忆进行回复生成时，需要从记忆库中找到最相关的记忆供模型参考。为在庞大且动态演进的记忆库中实现高精度、高效率的检索，如表 4.6所示，当前系统根据需求调整不同粒度的检索策略（部分情况混合多种粒度的检索）。在检索结果的基础上，有些系统引入重排序等后处理以进一步提升回复生成的效果（如 EverMemOS^[399]）。另外，部分记忆系统在检索模块的设计中借鉴了认知科学原理等其他领域的知识，使得其功能更加完善，提升系统的鲁棒性（如 A-MEM 借鉴 Zettelkasten 笔记系统的原子性与灵活链接机制，实现记忆网络的自主演进和组织。）。

4.4.3 未来展望

前面回顾总结了 2025 年 LLM 记忆系统在架构设计多个方面的关键进展。这些研究共同推动了 LLM 记忆系统向主动、演进的认知系统突破。接下来，我们对 LLM 记忆系统的未来发展趋势进行展望。

LLM 记忆系统在今年有了实质性的突破，该领域已为下一阶段的发展奠定了坚实的基础。我们猜测未来的研究将不再满足于功能实现，而是将焦

表 4.6: 记忆检索方式及其特性

检索方式	关键特性与优势	代表工作
向量检索	基于记忆嵌入向量的进行语义相似性召回，是相对基础的检索方案。	大部分系统
图结构检索	利用图遍历、关系路径推理，在图结构记忆中高效支持多跳推理和关系查找。	Zep, MemOg
关键词/元数据检索	根据时间戳、标签、主题等信息进行精准过滤和查询，具有极低的推理延迟。	memobase, MIRIX

重心集中在跨模态的统一以及系统级的自主进化。具体来讲，首先随着代理对现实世界感知需求的增强，记忆系统的重心将可能转向**基于多模态**进行记忆。记忆系统不仅需要高效地存储、检索和利用原始多模态资源，支持代理进行深度跨模态推理。并且记忆本身形式可能向多模态转变，将输入中各种信息集成在图像或者视频这种可视化模态，从而实现更高的信息压缩与抽象。另外，我们猜测未来可能进一步深化代理的**自主化记忆能力**。目前记忆系统的流程大部分都是人为设计，记忆系统应具备根据长期交互效果和任务需求，自主优化和调整其记忆管理流程的能力。

以上是我们关于今年 LLM 记忆系统发展情况的总结，我们希望该领域在未来能够持续蓬勃发展，构建更智能、更可靠的记忆生态。

4.5 自我反思自我修正智能体

4.5.1 研究背景

自我反思（Self-Reflection）是一种元认知能力，指 AI 智能体在完成任务或生成内容后，能主动地、系统地审视自身的行为过程、推理逻辑、中间步骤或最终输出。反思的目的在于评估其表现的质量、准确性、逻辑连贯性以及是否符合预设目标与约束。它不仅仅是简单的错误检查，更是一种高层次的自我意识模拟，旨在理解“为什么会这样做”以及“怎样才能做得更好”。自我修正（Self-Correction）是在自我反思基础之上采取的具体行动。当智能体通过反思识别出其决策或输出中存在的错误、缺陷或可优化空间时，它

会主动调整其内部状态、推理路径或生成策略，以产生一个更优的、修正后的结果。自我修正将反思的“洞见”转化为实际性能提升的关键闭环步骤。**迭代优化与自我反馈（Iterative Refinement with Self-Feedback）**这一范式是当前最主流、研究最深入的方向之一，其核心思想是构建一个“生成-反馈-修正”的迭代循环。代表性框架是 2023 年发表的 Self-Refine 其工作流程可以分解为三个核心步骤：1. 基于当前的记忆和状态，生成行动或决策；2. 评估行动者执行任务的结果，输出一个奖励分数或反馈评价；3. 迭代修正（Iterative Refinement）：它会分析过去的行动轨迹和失败原因，生成一段自然语言形式的“反思”，并将其存储在短期记忆中，以指导行动者在下一轮尝试中避免同样的错误。这个“输出-反馈-修正”的过程可以重复进行多次，直到输出质量收敛或达到预设的迭代次数。这种显式的“反思”架构的优点是概念清晰、易于实现，但是这种管道式方法的局限性在于**效率低下、上下文丢失和集成困难**，使得系统脆弱且难以维护。

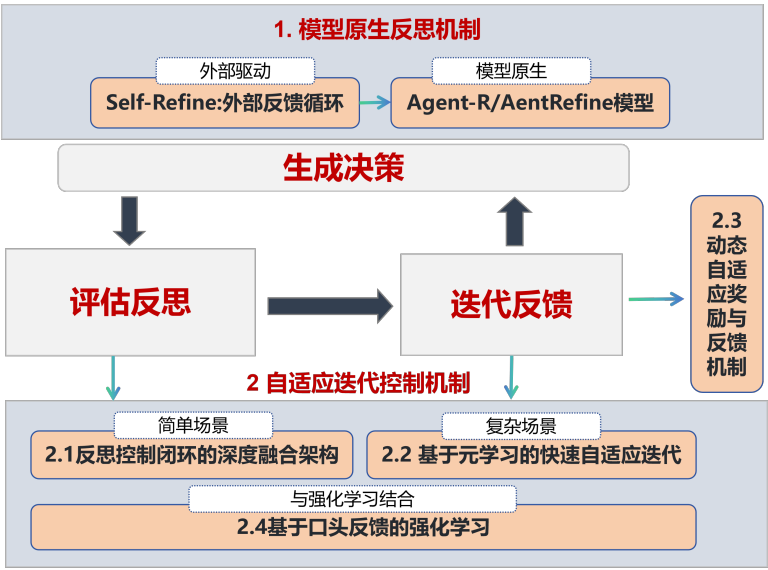


图 4.5: 自我反思智能体循环流程优化方法

4.5.2 模型原生反思机制

2025 年的研究趋势明显转向了“模型原生”或“内在化”的反思机制。其核心思想是，让自我纠正和反思能力成为模型自身生成策略的一部分，而不是一个外部调用的附加模块。

这一趋势的代表性研究包括：Agent-R^[400]该框架通过迭代自我训练

(**Iterative Self-Training**), 让智能体在训练阶段就学习如何反思和修正。智能体在解决任务时生成思考过程, 然后自我反思这些过程的优劣, 并将高质量的反思轨迹作为新的训练数据, 从而将反思能力“蒸馏”到模型参数中, 实现了从事后修正到实时与主动反思。

AgentRefine^[401]和 STeP^[402]这些研究探索了在模型的解码 (Decoding) 阶段直接融入反思和修正步骤, 使模型在生成一个“有缺陷”的初步想法后, 能立即进行自我批判和改进, 再输出最终决策。MIT 的 SEAL^[403]框架允许 LLM 通过与环境 (如工具、API) 的交互, **自主生成微调数据**并更新自身指令, 实现真正的自我适应。这代表了反思能力从“任务级”修正深化到“模型级”演化的重要一步。

这一从“外部管道”到“内部原生”的转变, 被认为是增强智能体自主性的关键一步, 标志着 2025 年至 2027 年“模型原生过渡”阶段的开启。

4.5.3 自适应迭代控制机制

迭代是自我反思智能体实现自我改进的基本形式, 但无控制的迭代是危险且低效的。自适应迭代控制正是为了解决“如何迭代”这一核心问题而发展起来的。它源于传统控制理论中的**自适应迭代学习控制 (AILC)**但其在 AI 智能体领域的内涵和外延都得到了极大的扩展。

传统的 AILC 通常假设任务是重复的, 且环境相对稳定。这与 AI 智能体所面临的开放、动态、任务多变的环境有本质区别。在自我反思智能体的语境下, 自适应迭代控制的目标不再是控制一个物理系统, 而是**控制智能体自身的认知过程**——包括规划、推理、工具使用和自我修正的循环。其“自适应”体现在**迭代深度的自适应、迭代策略的自适应、和学习率/适应速度的自适应**。

趋势一：反思-控制闭环的深度融合架构

这是当前最主流和核心的趋势。研究者们致力于设计能够将反思的“洞察”无缝转化为控制“行动”的一体化架构。

增强的 Reflexion 架构^[404]提出了一种先验动态 workflow 构建框架, 通过使评估者 (Evaluator) 输出一个关于反馈质量的“置信度”作为自适应控制器的输入, 用以决定下一次迭代的 Actor 是否应该采纳 Self-Reflection 模块的建议, 以及采纳的程度, 为反思-控制闭环实现了“自适应”处理。

图示化的认知流程^[405]一些工作通过流程图清晰地展示了这种闭环。例

如，一个典型的“自反思智能体” workflow 被描绘为：在任务执行后，通过**前置条件检查**和**后置条件检查模块**进行反思，然后将包含反思洞察的成功或失败示例存入经验数据库，用于后续的**迭代学习**。这个数据库的学习过程，本身就是一个由新经验驱动的、不断迭代的自适应过程。

趋势二：基于元学习的快速自适应迭代

面对任务多变的环境，智能体需要具备快速适应新任务的能力。在自我反思智能体的语境下，传统的机器学习是“单次适应”，而元学习（Meta-Learning）则是“多次适应”。元学习，或称“学会学习”，为实现这种快速适应提供了强大的理论工具，并与自适应迭代控制紧密结合。

Oriike et al.^[406]系统探讨了元学习如何释放自适应人工智能系统的潜力。文章详细阐述了 MAML（Model-Agnostic Meta-Learning）等算法如何使模型具备快速适应新任务的能力，为智能体设计提供了理论基础。

Wang et al.^[407]针对对抗性任务分布生成的鲁棒快速适应问题。研究了在复杂、甚至带有对抗性的环境中，智能体如何通过元学习机制保持快速适应能力，这对构建安全的自主智能体至关重要。

Wang et al.^[407]是一个结合了残差自适应上下文调优的元学习方法，专门用于威胁检测等高风险领域的快速模型适应，展示了元学习在实际工业场景中的高效性。

趋势三：动态自适应奖励与反馈机制

自适应迭代控制的有效性高度依赖于反馈信号的质量。因此，如何设计和生成高质量、信息丰富的反馈，成为一个独立且重要的研究方向。

自适应奖励函数（Adaptive Reward Shaping） 2025 年出现的 Self-Adaptive Reward Strategy 机制和 Adaptive Reward Scaling 方法^[408]能够根据智能体在任务中的历史成功率、探索阶段等信息，**动态调整奖励函数**。在任务初期，给予更多的探索奖励；在智能体掌握基本技能后，则提高对任务完成度的奖励权重。这种自适应奖励使得强化学习过程中的迭代更加高效和稳定。

多维度反馈与领域特定模板 单一的标量奖励（如成功/失败）包含的信息量太少。Yun et al.^[409]引入了**多维度评估指标**和**领域特定的反馈模板**。通过多维

解耦, 将评估拆解为准确性 (Accuracy)、逻辑性 (Logic)、创造性 (Creativity) 等独立维度, 例如在代码生成任务中, 反馈不仅包括代码是否能运行, 还包括其效率、可读性、是否遵循代码规范等多个维度。这种结构化的反馈为自适应迭代控制提供了更精确的“梯度”信号, 指导智能体进行更有效的修正。

趋势四: 基于口头反馈的强化学习 (Reinforcement Learning and Verbal Feedback) 与 Self-Refine 主要依赖“指令式”反馈不同, 另一条重要的技术路线是将自我反思与强化学习 (RL) 相结合。**代表性框架 Reflexion**^[410], 核心工作流程为**行动与评估-自我反思-记忆更新与指导未来行动**, 智能体在“试错”中学习, 但其学习信号不再是难以解释的标量奖励, 而是具有丰富语义信息的自然语言反思。2025 年, 基于 Reflexion 思想的框架进一步发展, 探索了更复杂的 RL 集成方式。研究重点包括:

将反思作为策略更新的一部分。 Qu et al.^[411]: 不仅仅是将反思作为下一次尝试的上下文, 而是探索如何直接利用反思文本来更新语言模型的内部参数, 通过对比成功和失败路径上的反思, 可以构建偏好数据集, 并使用如 DPO (Direct Preference Optimization) 等技术对模型进行微调。pmlr^[412]提出了 DPSDP (Direct Policy Search by Dynamic Programming) 方法, 通过多智能体协作与强化学习, 训练一个演员-评论家系统, 使大语言模型能够迭代优化推理答案, 并具备先验动态适应能力, 在数学推理任务上显著提升了准确率与泛化性能。

分层反思 (Hierarchical Reflection)。对于非常复杂的任务, 智能体可能需要进行多层次的反思。MobileUse 通过引入分层机制, 在 Android 自动化任务中显著减少了失败次数 (特别是在 Planning 和 Navigation 任务上)。这种分层结构使反思更加系统化。SE-VLN^[413]: 提出了一种基于多模态大语言模型的自我演进框架, 通过分层记忆模块和反思模块, 专门用于存储短期和长期经验, 并进行多步决策分析。

Huang et al.^[414]聚焦于大推理模型过程中普遍存在的因为冗余的自我反思循环导致的过度思考 (overthinking) 问题, 并提出了一种新颖且高效的干预方法——流形引导 (Manifold Steering), 通过将高维干预投影到 overthinking 的低维流形上, 在几乎不损失准确率的前提下, 大大减少了大推理模型的冗余输出, 为高效、可解释的推理提供了新路径。

4.5.4 检索增强自反思

Self-RAG (Self-Reflective Retrieval-Augmented Generation, Self-RAG)^[415]思想在 2024 年首次提出，该框架将自我反思与检索增强生成（RAG）相结合。模型首先判断是否需要检索。如果需要，它会输出一个特殊的 Retrieval Token，触发检索器获取相关文档；模型基于检索到的信息生成候选回答，同时，对自己刚刚生成的内容进行评估；最后，通过内部的反思机制，比较不同版本的回答并选择最佳输出。

InstructRAG^[416]：提出 InstructRAG，一种通过自合成推理实现显式去噪的检索增强生成框架。该方法利用指令调优语言模型，基于问题、检索文档和正确答案生成解释性推理，以区分有用信息与噪声。这些推理可作为上下文学习示例或监督微调数据，使模型显式学习去噪过程，无需额外人工标注。InstructRAG 在多个知识密集型任务中显著优于现有 RAG 方法，提升了生成准确性与鲁棒性。除了文本领域的 Self-RAG，这一思想也被扩展到了图像生成领域，Lyu et al.^[412]针对当前主流文生图模型在生成细粒度、未见过的真实世界物体时容易产生幻觉或失真的问题，提出了首个基于真实图像检索增强的生成框架 RealRAG。其核心创新在于引入自反思对比学习（Self-reflective Contrastive Learning）训练一个“反思型检索器”，动态补全生成模型的知识盲区。

4.5.5 结论与展望

2025 年，自我反思智能体领域正在从外部、僵化的迭代机制，迈向更深层次的、模型原生的、动态自适应的自我改进范式。以 2025 年升级版 SELF-REFINE 框架为代表的自适应迭代控制，以及向模型原生反思的转变，共同为解决智能体在开放世界中的持续学习、适应和纠错问题提供了迄今为止最有效的路径。

然而，前路依然充满挑战。**安全性和对齐问题**是悬在能够自我演进的智能体头上的“达摩克利斯之剑”。未来的研究必须将安全框架、可解释性机制和治理结构作为与算法本身同等重要的核心组成部分进行设计。**通用性和效率**的矛盾依然突出，如何在保证强大泛化能力的同时，控制迭代过程带来的计算和时间成本，是实现技术落地必须解决的工程难题。

总之，自我反思与自适应迭代控制的结合，不仅是一种技术路径，更是一种哲学思考，它关乎我们如何构建能够学习、成长并最终与我们安全共存的智能生命形式。2025 年只是这场伟大探索的序章，未来的发展值得我们持

续关注和深入研究。

4.6 自我进化

4.6.1 研究背景

当前主流的大语言模型（LLMs）驱动的智能体系统虽在任务执行上展现出强大能力，但其核心架构往往呈现高度静态性：模型参数固定、提示模板预设、工作流硬编码、协作逻辑缺乏适应性。这种“一次性部署、长期不变”的范式，在面对动态环境、未知任务或持续交互场景时迅速暴露出局限——无法从经验中学习，难以自主改进策略，更无法针对新工具、新队友或新目标进行结构调整。

为突破这一瓶颈，研究者开始将“自我进化”视为智能体系统的核心能力之一。所谓自我进化，是指智能体在无显式人工干预的前提下，通过任务反馈、环境交互或内部反思，持续优化其内部组件（如模型、提示、流程、工具、协作机制等），从而提升长期性能与泛化能力。早期尝试多聚焦于单一维度（如仅优化提示或仅微调模型），而 2024–2025 年间，一系列工作开始从系统层面构建可进化的智能体框架，推动该方向从零散技巧走向体系化设计。

下面，我们将围绕三个关键进化维度：基座能力进化、自治智能体结构进化、多智能体进化，梳理 2025 年代表性进展。

4.6.2 研究进展

基座能力进化

在最基础的层级，智能体的“思考引擎”——即其底层大语言模型（LLM）——正逐步获得自我更新的能力。提升基座模型能力的核心路径主要有两类，如表 4.7 所示：一是通过迭代调整模型参数以持续适应任务需求；二是通过动态优化提示词（prompt）来激活模型的潜在领域能力，而无需修改参数。这两类方法共同构成了当前智能体基座能力进化的主干。

模型参数的迭代优化主要通过有监督微调（Supervised Fine-Tuning, SFT）与强化学习（Reinforcement Learning, RL）两种范式实现。

(1)在有监督微调方向,研究重点在于持续获取高质量训练数据。STaR^[417]开创性地提出通过模型自身生成高质量推理链（rationales），并以此构建自举

表 4.7: 基座能力进化

进化对象	核心策略	代表工作
基座模型	对模型参数进行有监督微调以提升推理能力	STaR ^[417] , AGENTGYM ^[418]
	通过强化学习优化策略以最大化任务奖励	Absolution Zero ^[419] , R-Zero ^[205] , Co-EPG ^[420] , AgentEvolver ^[421]
	利用启发式搜索策略改进提示词结构与表达	ORPO ^[422] , Promptbreeder ^[423] , PromptAgent ^[424]
提示词	基于反馈信号自动优化提示词内容	TextGrad ^[333] , SPO ^[425]

式训练数据，从而在数学与常识推理任务上显著提升模型性能。AGENTGYM^[418]进一步构建了一个多环境探索沙盒平台，支持智能体在多样化交互场景中自主生成高质量执行轨迹，并以此驱动基座模型的持续演进。

(2) 在强化学习方向，近期工作聚焦于构建有效的正负样本对，并结合在线或离线策略优化。Absolution Zero^[419]设计了“任务创建者-任务执行者”双角色架构，有效缓解了真实任务稀缺的问题。R-Zero^[205]则采用出题者-解题者协同框架：出题者模型动态生成与当前能力匹配的挑战性任务，解题者模型不断尝试求解，二者通过闭环迭代实现从零开始的能力增长。RAGEN^[426]构建了面向多轮对话的轨迹级强化学习系统，通过不确定性感知的轨迹过滤与梯度稳定机制，显著缓解了 LLM 智能体训练中常见的奖励剧烈波动与梯度爆炸问题。Co-EPG^[420]则针对 GUI 智能体中的规划与定位模块割裂问题，建立了一个协同进化闭环：两个模块在迭代训练与数据增强中相互促进，并通过基于置信度的动态奖励集成机制提供稳健的学习信号，从而突破了传统孤立优化的瓶颈。AgentEvolver^[421]引入自我提问、自主导航与自我归因三大机制，在完全无需人工标注的前提下高效收集高价值交互数据，大幅降低对传统强化学习中密集奖励信号和大量试错的依赖。

提示词优化

优化提示词是提高智能体利用基座模型能力重要方式之一。当前提示词已从静态模板演变为可自主演化的语言组件，其优化主要分为基于启发式搜索与基于反馈信号两类策略。

(1) 在基于启发式方法的优化方面，ORPO^[422]首次提出将大语言模型本身作为优化器，通过迭代反馈自动改进提示词。Promptbreeder^[423]进一步引入遗传算法，在自然语言层面同时进化“任务提示”与“突变提示”，在不更新模型参数的前提下于多个推理与分类基准上超越现有最优方法，充分展示了语言模型通过语言进行自我改进的潜力。PromptAgent^[424]则结合蒙特卡洛树搜索与错误反馈机制，能够自动生成结构清晰、富含领域知识的专家级提示，为复杂任务提供高性能提示方案。

(2) 在基于反馈信号的优化方面，TextGrad^[333]构建了一个复合 AI 系统，将角色提示、工具提示等所有提示组件统一纳入优化框架，利用文本梯度反馈实现多智能体结构中提示词的协同演化。SPO^[425]则提出一种自监督提示优化框架：由一个 LLM 担任优化器，依据任务目标对提示进行精细化调整；同时引入另一个 LLM 作为评估器，通过成对比较不同提示生成的输出质量，筛选出更优候选，从而实现高效、自动化的提示进化。

表 4.8: 自治智能体结构进化

进化对象	核心策略	代表工作
记忆的进化	通过交互持续积累和更新领域知识	MemInsight ^[427] , Mem0 ^[428] , A-MEM ^[429] , Flex ^[430] , Evo-Memory ^[431]
工具的进化	从外部环境搜索、生成、复用可用工具集	Alita ^[432] , Alita-G ^[433] , ToolGen ^[434]
整体结构进化	通过代码生成自动修改智能体行为逻辑	ADAS ^[435]
	基于任务需求动态组装模块化智能体架构	AgentSquare ^[436]

自治智能体结构进化

智能体的长期适应能力不仅依赖于流程调整，更根本地取决于其系统架构是否具备自主演化能力。近年来的研究从记忆系统、工具集到整体行为架构，逐步构建起多层次的自我进化基础设施，使智能体能够动态重构其“认知器官”以应对开放、动态的任务环境。这一演进可系统性地划分为三个相互协同的维度，如表 4.8所示：记忆的进化、工具的进化，以及整体结构的协同进化。

记忆的进化。现代智能体的记忆机制正从静态缓存转向主动、结构化的知识演化系统，支持在任务交互中持续提炼、组织与复用经验。MemInsight^[427]通过属性挖掘与增强检索机制，将原始对话历史转化为结构化语义记忆，显著提升上下文相关性与推理一致性。Mem0^[428]进一步提出两阶段架构：先动态提取关键信息，再以图结构表示实体关系与时序逻辑，在保持低延迟的同时强化对复杂依赖的建模能力。A-MEM^[429]引入智能体驱动的记忆管理范式，通过动态索引与链接构建互联知识网络，并支持在新记忆生成时调用工具对旧记忆进行修正，实现记忆的可编辑性与一致性维护。Flex^[430]则构建了一个可动态进化、跨任务扩展且支持多智能体继承的结构化经验库，使 LLM 智能体无需梯度更新即可实现持续学习，有效缓解了传统架构静态僵化的问题。Evo-Memory^[431]在此基础上设计了集成“行动-思考-记忆优化”闭环的动态系统，使智能体能够在测试阶段主动积累并演化经验，实现真正的在线自我完善。

工具的进化。智能体正从被动调用预设工具，转向任务驱动的工具自主发现、生成与复用。Alita^[432]构建了基于 MCP（Model-Callable Protocol）的闭环框架：智能体在任务执行中分析自身缺陷，主动搜索开源代码库，生成、验证并复用新工具，实现端到端的工具自扩充。Alita-G^[433]进一步引入抽象与检索增强机制，通过任务驱动的 MCP 生成与选择策略，将通用智能体自动转化为领域专家，显著提升复杂任务中的性能与效率，实现从通用能力到领域专长的高效迁移。ToolGen^[434]则采取全新范式：将真实世界工具映射为语言模型词汇表中的虚拟标记，通过“工具记忆-检索训练-端到端调优”三阶段训练，使模型能直接以标记生成方式调用工具，无需外部检索器，极大提升了工具集成的效率、鲁棒性与可扩展性。

整体结构协同进化。更高层次的进化体现在智能体对自身行为逻辑与系统架构的直接重构能力。ADAS^[435]将智能体定义为可执行代码，利用 LLM 的编程能力基于环境反馈线性迭代优化自身行为逻辑，实现代码级的自我改

进。AgentSquare^[436]则提出模块化设计范式，将智能体解耦为规划、推理、工具使用与记忆四大标准组件，并通过模块进化与动态重组，在统一架构空间中自动搜索高性能配置，标志着智能体系统正式进入“可编程、可演化”的新阶段。

表 4.9: 多智能体进化

进化对象	核心策略	代表工作
协作方式	优化多智能体协作工作流	GPTSwarm ^[437] , AFLOW ^[438] , MaMS ^[439] , FlowReasoner ^[440]
	优化多智能体群体自治过程	EvoAgent ^[441] , AgentNet ^[442]
群体经验	知识探索的协同演进	ProAgent ^[443]

多智能体进化

随着任务复杂度持续攀升，单一智能体的推理能力逐渐触及瓶颈，多智能体系统通过分工协作成为应对开放性挑战的核心范式。近期研究不再满足于静态部署多智能体，而是聚焦于让其协作结构与群体知识本身具备自主进化能力。这一趋势可归纳为两大方向，如表 4.8所示：智能体协作结构的动态优化，以及多智能体群体认知能力的协同演进。

智能体协作结构的进化

协作结构的进化主要体现为对多智能体工作流的自动优化与群体组织的自主生成。

(1) 在工作流优化方面，GPTSwarm^[437]将提示工程统一建模为可计算图，其中节点代表智能体角色或操作，边表示信息传递；该框架支持对节点内部提示逻辑与图连接结构的联合优化，实现端到端的协作效率提升。AFLOW^[438]则将工作流形式化为代码化的有向拓扑图，原子操作（如生成、评估、反思）作为节点，信息流作为边，并引入蒙特卡洛树搜索在有限资源下高效探索更优流程结构。MaMS^[439]进一步提出“按需采样”机制：面对每个新查询，系统从预定义的多智能体库中动态组合一个轻量、适配的工作流，实现运行时的灵活进化。FlowReasoner^[440]则通过引入基于外部执行反

馈的强化学习与多目标奖励机制，为每个用户查询自动生成个性化的多智能体系统，在代码生成等任务上显著提升性能、资源效率与可扩展性。

(2) 在智能体群体自治演进方面，EvoAgent^[441]将整个多智能体群体定义为文本配置（包括角色、技能、提示等），并采用进化算法——通过变异、交叉与选择算子——自动搜索高性能的群体组合，实现从专用智能体到协作系统的无缝扩展。AgentNet^[442]则构建了一个去中心化的有向无环图（DAG）架构，每个节点为自治智能体；系统通过动态拓扑演化与基于检索增强生成（RAG）的自适应学习，实现任务路由与能力分配的自主调整，无需依赖中心调度器，显著提升了系统的鲁棒性与可扩展性。

群体经验

多智能体系统在认知协同能力方面的进化成为了新的研究焦点。ProAgent^[443]引入了“心智理论”（Theory of Mind）机制，使得智能体能够通过观察队友的行为来推断其潜在意图，更新自身信念，并据此动态调整协作策略。这种能力让智能体在零次合作场景中也能迅速适应陌生伙伴，从而突破了传统多智能体系统依赖预设协议或固定交互模式的限制。

与此同时，为了有效衡量和评估这些智能体的认知与策略编码能力的进步，CATArena^[444]提出了一种新型的多智能体进化衡量基准平台，该平台采用迭代式同伴学习竞争框架，使智能体能够在无分数上限的棋牌游戏中进行多轮策略编码与优化，这种方法不仅解决了传统智能体基准由于任务固定而导致的分数饱和问题，还提供了一个系统化的方法来评估智能体的学习能力和策略编码等核心进化能力。

多智能体进化正从静态编排走向动态自组织，从结构优化深入至认知对齐，为构建真正具备开放协作能力与群体智能涌现潜力的 AI 系统开辟了新路径。

4.6.3 未来展望

前面回顾总结了 2025 年智能体自我进化在模型、提示、工作流、架构与多智能体协同等方面的关键进展。这些工作推动智能体系统从静态执行迈向动态演进，初步形成了可自我改进的闭环机制。

接下来，我们对这一方向的未来趋势进行展望。2025 年的突破为构建“终身学习智能体”奠定了重要基础，而未来的研究将超越当前以单一组件或孤立个体为中心的优化范式，转向更系统化、协同化与可持续的进化路

径。一方面，智能体的能力提升将更加紧密地耦合外部环境反馈与群体交互经验，使进化过程不仅源于内部反思，更植根于与世界及其他智能体的持续互动；另一方面，进化目标将从单对象优化逐步扩展为多对象联合优化，涵盖模型、提示、工具、记忆与协作策略等多个层级的协同调整，从而形成高效、一致的端到端适应机制。在此基础上，实现低开销、高效率的持续进化将成为核心挑战——通过轻量化评估、离线策略学习或神经符号融合等手段，在保障进化能力的同时显著降低计算与数据依赖。长远来看，多智能体系统的集体进化有望催生可共享、可传承的群体智能，推动智能体生态向真正具备持续成长能力的方向演进。

4.7 GUI Agent

GUI 智能体（GUI Agent）是一类能够在图形用户界面（Graphical User Interface, GUI）上“像人一样”完成复杂任务的软件程序。它的核心特征是：以自然语言为指令输入，通过大模型理解屏幕内容，自主规划并执行一连串“点击-拖拽-输入”等原子级操作，最终返回结果或达成目标，而无需依赖传统的脚本、API 或规则引擎。本部分主要从下面五方面综述了 2025 年 GUI 智能体领域的最新进展，包括 GUI Agent 感知能力、GUI Agent 规划能力、GUI Agent 执行以、面向 GUI 的专用模型和 GUI 智能体相关评测数据集。

4.7.1 GUI Agent 感知能力

GUI Agent 的感知能力是指其通过计算机视觉、界面结构解析等技术，实时理解图形用户界面中“有什么”与“在哪里”的能力。它不仅需要识别像素级的图标、文本、按钮、输入框等视觉元素，还能解析底层的可访问性树（Accessibility Tree）与 UI 框架元数据，从而将散乱的像素转化为带有语义的交互对象。借助屏幕截图、DOM 树、控件属性与 OCR 等多源信息，GUI Agent 可在不同分辨率、主题、语言甚至跨平台（Windows、macOS、Web、移动端）环境下建立统一且细粒度的界面表征，进而为后续的任务规划、操作决策提供精准的感知基础。

在 GUI Agent 感知数据层面，Gou et al.^[445]采集了迄今最大的 GUI 视觉定位数据集，涵盖 130 万张截图上的 1000 万个 GUI 元素及其指代表达，并据此训练出 UGround，一个面向 GUI 智能体的强通用视觉定位模型。在横跨三大类别（定位、离线智能体、在线智能体）的六项基准测试中，实验结果表明：（1）UGround 在 GUI 视觉定位任务上比现有模型绝对提升最高

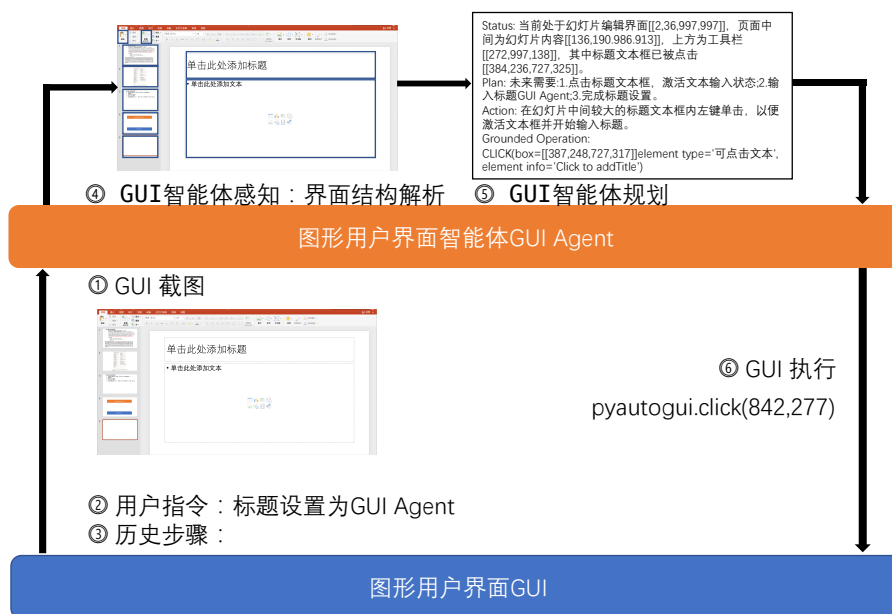


图 4.6: GUI 智能体总体框架

达 20%；(2) 配备 UGround 的智能体在仅依赖视觉输入的情况下，仍优于使用额外文本信息的 GUI 智能体。这些结果强有力地证明了“像人类一样浏览数字世界”的 GUI 智能体不仅可行，而且前景广阔。

在 GUI Agent 感知方法层面，Cheng et al.^[446]认为 GUI 智能体的核心挑战在于“GUI 定位”——即根据指令精准找到屏幕元素。为此引入 GUI 定位预训练，并设计自动化数据构建方法；同时发布首个涵盖移动、桌面与 Web 的真实 GUI 定位基准 ScreenSpot。经预训练后，SeeClick 在 ScreenSpot 上显著超越多种基线，并在三大主流基准中一致验证：GUI 定位能力的提升直接带动下游 GUI 智能体任务性能。Wang et al.^[447]提出一种视觉分而治之框架：利用通用大模型指令解析器（Instruction Interpreter），并训练专门的视觉定位模型（Visual Element Locator）来精确标注元素坐标。这种双重模型结构使系统在无需 DOM 输入的情况下即可保持高灵活性，并在 ScreenSpot 定位基准上获得了大幅提升。Tang et al.^[448]提出了 GUI-G²（GUI Gaussian Grounding Rewards）框架，核心创新之处在于将 GUI 元素建模为二维高斯分布，而非简单的点或矩形框。这一设计带来了三个关键突破：（1）双重高斯奖励机制；（2）自适应方差机制；（3）连续空间优化。

4.7.2 GUI Agent 规划能力

GUI Agent 的规划能力是指其在大语言模型驱动下，将用户的高层次目标自动拆解为可在 GUI 上逐步执行的、带前后依赖关系的原子操作序列的能力。它首先把“做什么”转化为“怎么做”：通过融合界面感知结果、历史交互轨迹与内置常识，快速枚举可行动作空间（点击、输入、滚动、快捷键等），并评估每一步对任务状态的价值与风险；随后利用层次化或链式推理，在宏观层面决定子任务顺序，在微观层面为每个子任务生成带参数的具体动作，同时持续自检前置条件是否满足、异常是否发生，并能在局部失败时即时重规划。

Li et al.^[449]提出 MobileUse——一款专为鲁棒且自适应移动任务执行而设计的 GUI 智能体。为了提升此智能体在长链条任务与动态环境下的韧性，特别引入了分层反思架构，即智能体可在多个时间尺度（单步动作到整任务）上自监测、检测并恢复错误，同时通过“按需反思”策略保持高效。Yang et al.^[450]等人采用“测试时扩展”策略挑选最优动作提案：每一步并行采样多条候选提案，由裁判模型评估并择优，通过算力投入换取更高决策质量。Agashe et al.^[451]提出 Agent S2——一种全新的组合式框架，将认知职责分散到多个通用与专用模型，设计新颖的 Mixture-of-Grounding 技术实现精准 GUI 定位，并引入主动分层规划，在多个时间尺度上随观测动态细化行动计划。除此之外，Zhang et al.^[452]提出了进度奖励模型（ProgRM），在在线训练过程中，为每一步预测任务完成进度，从而提供密集且信息丰富的中间奖励。针对进度奖励标签难以人工标注的问题，还进一步设计了一种高效的基于最长公共子序列（LCS）的自标注算法，自动发现轨迹中的关键步骤并分配合适的进度标签。

4.7.3 GUI Agent 执行

GUI Agent 的执行能力是指其在真实操作系统与应用程序环境中，把规划阶段生成的“点击-输入-滚动”等抽象动作精准、安全、可靠地落到实际像素与系统调用上的能力。它通过自动化引擎（如 PyAutoGUI、AppleScript、Selenium 等）将原子指令转化为底层事件：计算屏幕坐标、校验控件状态、注入鼠标/触摸/键盘事件，并捕获界面反馈，形成闭环验证。目前，GUI Agent 执行主要落地在 Windows、macOS、Browser（浏览器）、Linux、iOS、Android 等平台。

Song et al.^[453]设计了 CoAct-1，这一多智能体系统首次把 GUI 操控与

直接程序执行有机融合。其核心是一个“编排器”，可动态将子任务分派给传统 GUI 操作员，或是一名专精的“程序员”智能体——后者能够编写并执行 Python 或 Bash 脚本。这种混合策略使智能体在文件管理、数据处理等场景下绕过冗长的 GUI 操作链，同时在必要时仍保留视觉交互能力。Lin et al.^[454]将截图建模为 UI 连接图，自适应识别冗余关系，并在自注意力模块中以此为依据筛选 Token，显著降低计算开销。并统一了 GUI 任务中的多元需求，在导航中高效管理视觉-动作历史，或将多轮查询-动作序列与单张截图配对，提升训练效率。Zhang et al.^[455]针对通用智能体在执行特定软件任务时因缺乏领域知识而频发操作失败的问题，提出了一种执行知识自进化框架 UI-Evol。该方法构建了一个从“执行尝试”到“知识沉淀”的闭环系统：当智能体在真实环境中执行受阻时，系统会自动分析失败轨迹，利用大模型推断正确的操作逻辑，并将这些经验演化为高质量的指令-动作对。通过这种方式，UI-Evol 使得执行模块不再依赖静态规则，而是能够通过不断的试错与反馈，自动积累针对陌生 UI 控件的正确操作策略，显著提升了智能体在未见应用上的执行成功率。除此之外，Luo et al.^[456]将 DeepSeek-R1 式的“推理增强”范式引入 GUI 动作执行层，构建了首个具备内生思维链（CoT）的通用 GUI 视觉-语言-动作（VLA）模型。GUI-R1 创新性地利用大规模强化学习（RL）优化“思考-执行”过程，强制模型在生成“点击”或“输入”等物理动作前，先输出一段显式的推理文本（Reasoning Trace）。这种机制使智能体能够在执行阶段实时分析 UI 布局变化与潜在干扰，通过“三思而后行”大幅提升了像素级定位的精准度，有效解决了执行过程中的鲁棒性问题。

4.7.4 面向 GUI 的专用模型

面向 GUI 的专用模型是一种以“看得懂界面、找得到元素、做得对操作”为核心目标而深度定制的大模型。2025 年，以字节跳动 UI-TARS-2 为代表的一大批面向 GUI 的专用模型面世。

具体来说，Xie et al.^[457]推出了 JEDI 3B/7B，并合成并发布了迄今最大的计算机使用定位数据集 JEDI，通过多视角任务解耦生成 400 万条样本。在 JEDI 上训练的多尺度模型 JEDI 3B/7B 在 ScreenSpot-v2、ScreenSpot-ProOSWorld-G 上取得了良好的效果。Qin et al.^[458]，Wang et al.^[459]提出了 UI-TARS 和 UI-TARS-2，覆盖了 2B、7B、72B 等多个尺寸。其中，UI-TARS-2 背后的关键核心技术为多轮强化学习。依靠这一技巧，UI-TARS-2 核心解决

了“让 AI 自主操作图形界面”的四大难题：(1) 数据稀缺：以往方法需要上百万级高质量标注数据，成本极高，扩展困难。(2) 环境割裂：不同任务（电脑、手机、网页、终端、游戏）通常要在不同框架里训练，无法统一。(3) 能力单一：大多数智能体只能做 GUI 点击或终端命令，难以完成真实复杂任务。(4) 训练不稳定：强化学习在 GUI 任务上容易出现奖励稀疏、策略崩溃，模型很难可靠收敛。Lai et al.^[460]提出了 AutoGLM-OS-9B，通过构建分布式强化学习基础设施，可同时编排数千个并行虚拟桌面环境，加速大规模在线强化学习。Wang et al.^[461]提出 OpenCUA——一个全面开源的框架，用于扩展 CUA 数据与基础模型。除此之外，Ye et al.^[462]提出了 Mobile-Agent-v3，该模型融合了 UI 定位、规划、动作语义与推理模式，支持端到端决策，并可作为多智能体系统的模块化组件。

4.7.5 GUI 智能体数据集

GUI 智能体数据集层面长期存在三大难点问题：(1) 真实世界 CUA 任务稀缺；(2) 缺乏可自动采集并标注多模态轨迹的流水线；(3) 需要同时评估 GUI 定位、屏幕解析与动作规划及预测等多个方面。

Mu et al.^[463]提出了 GUI-360°，该数据集涵盖了查询获取、环境模板构建、任务实例化、批量执行以及大模型驱动的质量过滤，共包含 120 余万个已执行动作步骤，横跨数千条轨迹，覆盖主流 Windows 办公软件；并且提供全分辨率截图、可获取的无障碍元数据、实例化目标、中间推理痕迹，以及成功与失败的动作轨迹。数据集支持三大核心任务：GUI 定位、屏幕解析与动作预测，并采用混合的 GUI+API 动作空间。Zhao et al.^[464]具备两大特征：(1) 中间态起点：真实用户与 GUI 助手交互时，很少从默认初始条件出发；任务可能始于任意中间状态，用户随时寻求帮助。(2) 上下文多变：某些任务可能来自完全不同的上下文或界面，智能体必须调整既有计划或引入新步骤，才能确保任务完成。通过把这些情境纳入基准设计，WorldGUI 更逼真地还原现实 GUI 交互，从而对 GUI 智能体能力进行更精准、全面的测评。具体而言，WorldGUI 覆盖 10 款常用桌面应用，共 611 项任务；每项任务配有用户查询、教学视频及对应项目文件。Wang et al.^[465]提出了一个分层基准 MMBench GUI，用于评估 Windows、macOS、Linux、iOS、Android 和 Web 平台上的 GUI 自动化智能体。它包括四个层次：GUI 内容理解、元素接地、任务自动化和任务协作，涵盖了 GUI 智能体的基本技能。

4.7.6 总结与展望

综合来看，2025 年标志着 GUI 智能体（GUI Agent）从“可演示的原型系统”迈入“可规模化落地的通用智能体形态”的关键拐点。一方面，感知、规划、执行与模型层面的协同进展，使 GUI Agent 在复杂真实软件环境中的成功率、鲁棒性与泛化能力均取得了质的提升；另一方面，这一方向仍面临多层次的基础性挑战，有待在未来研究中系统性突破。

首先，在能力层面，GUI Agent 已形成较为完整的技术闭环。在感知方面，大规模 GUI 定位数据集（如 ScreenSpot）与专用视觉定位模型（如 UGround、SeeClick）的出现，使“精准理解界面元素”从瓶颈转为可规模优化的子问题；同时，Gaussian Grounding 等连续空间建模范式，显著提升了像素级定位的稳定性与可微性。在规划层面，分层规划、反思机制、测试时扩展与组合式智能体架构，增强了 GUI Agent 在长任务、动态界面和不确定环境下的决策韧性。在执行层面，CoAct-1、UI-Evol、GUI-R1 等工作打通了“视觉操作—程序执行—知识沉淀”的闭环，使执行不再是简单的动作回放，而逐步演化为具备推理与自适应能力的决策过程。在模型层面，UI-TARS-2、AutoGLM-OS、OpenCUA 等专用基础模型和开源框架，系统性缓解了数据稀缺、环境割裂与训练不稳定等长期制约因素。

其次，从研究范式上看，GUI Agent 正在从“任务驱动”转向“能力驱动”。2025 年的工作不再仅围绕单一基准或固定应用场景展开，而是更加关注可复用的核心能力，如跨平台界面理解、通用 GUI 定位、层次化规划与执行级自反思。这种能力抽象的转变，使 GUI Agent 逐渐具备“像人类一样使用计算机”的通用潜力，也为其在办公自动化、软件测试、数据分析、移动终端操作乃至工业软件中的规模化部署奠定了基础。

然而，尽管进展显著，GUI Agent 仍面临一系列关键挑战。（1）挑战 1：跨应用与跨版本泛化仍不充分。现实中的 GUI 界面高度异构，且随软件版本频繁变化。即便在大规模数据与强化学习加持下，智能体在未见应用、冷启动界面上的成功率仍明显低于人类。如何在有限交互成本下快速适应新 GUI，是未来必须解决的问题。（2）挑战 2：长程任务的稳定性与代价控制问题突出。测试时扩展、多候选采样等方法虽然显著提升了决策质量，但也带来了推理延迟和算力成本的快速增长。如何在成功率、效率与资源消耗之间取得平衡，是 GUI Agent 走向实际产品化的关键工程难题。（3）挑战 3：执行安全性与可控性尚不完善。GUI Agent 具备直接操控真实系统的能力，一旦规划或执行失误，可能导致数据丢失、误操作甚至安全风险。当前多数工

作仍聚焦成功率提升，对失败后果建模、风险感知与安全约束关注不足。(4) 挑战 4: 评测体系与真实价值之间仍存在鸿沟。现有基准主要关注任务完成率或步骤匹配度，尚难全面反映智能体在真实办公流、企业软件或工业场景中的长期价值与可靠性，亟需更贴近真实使用情境的评测标准。

4.8 多智能体协作框架

4.8.1 研究背景

随着大语言模型能力的持续增强，单智能体在通用推理、代码生成和工具调用等任务中已取得显著进展。然而，在真实世界的复杂问题中，任务往往呈现出**长程依赖强、子任务异构、搜索空间巨大**等特征，单一智能体在规划深度、并行性和鲁棒性方面逐渐显现出能力瓶颈。在此背景下，**多智能体协作**逐渐成为提升系统整体智能水平的重要范式。

多智能体系统通过引入角色分工、并行探索和信息互补，能够在复杂推理、软件工程、开放式决策等场景中显著提升性能上限。然而，在以 LLM 为核心的多智能体协作设定下，系统仍面临一系列关键挑战：一方面，智能体间的通信和协作高度依赖自然语言，带来了显著的成本与时延开销；另一方面，协作结构、通信拓扑与执行流程往往需要针对具体任务精心设计，缺乏通用且可扩展的构建范式。此外，在长程任务中，如何在保证协作效果的同时实现稳定执行、动态调整与全局治理，仍是一个亟待解决的问题。

早期工作多采用人工设计的流水线或固定通信模板来组织多智能体协作，但这类方法在任务多样性、规模扩展性和资源效率方面存在明显局限。自 2025 年以来，随着对复杂任务和大规模协作需求的增长，研究者开始从**系统架构与自动化设计**的角度重新审视多智能体协作问题，推动该领域从“手工拼装”迈向“自适应与可进化”的协作框架。

4.8.2 研究进展

综合 2025 年的相关研究可以观察到，多智能体协作框架已不再仅被视为若干智能体的简单组合，而是逐渐演化为一个具备明确结构与运行机制的系统级设计问题。从系统视角看，当前多智能体协作研究主要围绕三项核心能力展开：**协作信息的组织与约束、协作结构的构建与演化，以及运行期的调度控制与治理机制**，这些能力共同决定了多智能体系统在复杂任务中的推理深度、资源效率与执行稳定性。

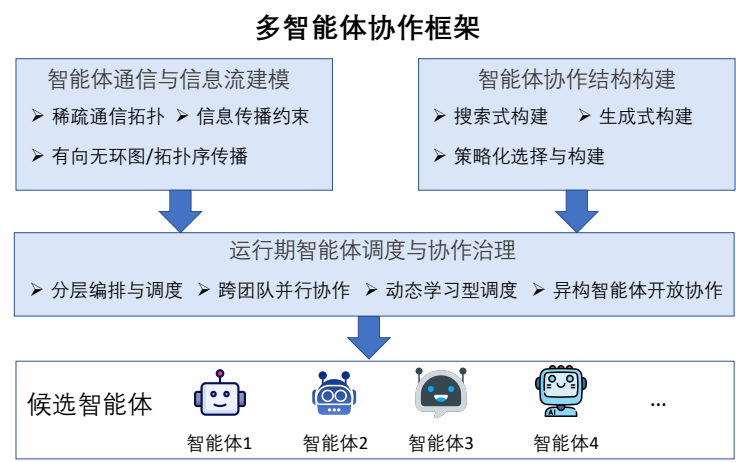


图 4.7: 大模型多智能体协作框架

智能体通信与信息流建模

在多智能体系统中，通信是实现协作收益的必要条件，但同时也是推理成本和系统复杂度的主要来源。当智能体规模扩大时，不受约束的信息共享容易导致信息冗余、推理时延增长以及上下文管理困难。因此，2025 年的相关研究开始将通信机制从隐式实现细节上升为显式的系统建模对象，并通过对信息传播结构与顺序进行约束来提升协作的可扩展性。

现有研究普遍通过对通信结构进行显式建模，限制信息在智能体之间的传播范围与路径，以避免全连接通信带来的信息冗余与推理开销^[466-467]。在此基础上，部分工作进一步对信息传播过程施加顺序约束，采用有向无环结构组织多智能体协作，并通过拓扑序驱动中间结果的逐步精炼与有序传递，从而在保证协作效果的同时提升系统的可扩展性^[468]。

这些探索表明，通过对通信拓扑与信息传播过程进行结构化建模，多智能体系统可以在不依赖全局信息共享的前提下实现有效协作。通信结构与传播顺序已成为影响多智能体系统可扩展性与资源效率的关键系统设计维度。

智能体协作结构构建

除通信方式外，协作结构本身对多智能体系统的整体性能具有决定性影响。这里的协作结构不仅包括执行顺序，还涵盖参与智能体的数量、角色分工以及它们在不同任务阶段的交互关系。早期多智能体系统通常采用**预定义的协作结构**，由人工规则或固定模板指定协作流程，在任务复杂度提升或执

行过程中出现偏差时缺乏自适应能力。

随着任务规模与复杂性的增长，近期研究开始从“结构如何被确定”的角度重新审视多智能体协作问题，将协作结构视为一个可通过不同机制构建的系统对象。一类方法将协作结构的确定过程建模为**搜索问题**，通过在执行过程中探索、回溯和比较不同协作路径，逐步收敛到任务适配的协作结构。这类搜索式结构构建机制使系统能够利用执行反馈修正早期决策，在长程任务中显著提升鲁棒性^[438,469]。

另一类研究从生成式建模的角度出发，尝试根据任务描述或上下文条件**直接生成协作结构**，使参与智能体的数量、角色组合与交互关系能够在推理过程中一次性或按需确定。这种生成式结构构建方式有效降低了人工设计成本，并增强了多智能体系统在不同任务之间的适应能力^[470-471]。

此外，还有研究将协作结构的选择提升为**策略决策问题**，不再追求为所有任务构建单一最优结构，而是在多个候选协作结构之间进行动态选择或采样，以适配不同任务状态与资源约束。通过这种策略化的结构选择机制，多智能体系统能够在性能与计算成本之间实现更灵活的权衡^[439,472]。

这些工作共同推动协作结构从静态、外生的工程设定，转变为可通过搜索、生成或策略选择等机制动态确定的内部系统变量，从而使协作结构不再是一次性设计决策，而成为多智能体系统在复杂任务中持续发挥作用的关键组成部分。

运行期智能体调度与协作治理

在复杂长程任务中，即便通信方式与协作结构设计合理，多智能体系统仍需在运行期对执行过程进行持续调度、监控与干预，以应对任务不确定性、失败恢复以及资源约束的动态变化。因此，2025 年的研究开始将运行期的协作控制与治理视为区别于结构设计的独立系统层进行建模。

一些现有工作普通过引入分层编排与调度机制，将任务分解、智能体执行与全局监控解耦，使系统能够在局部并行探索与全局一致性之间取得平衡^[473-474]。在此基础上，一些研究进一步通过并行运行多个协作单元或智能体团队，并在关键阶段对多条执行路径进行聚合与裁决，以缓解单一流水线协作在复杂决策空间中的性能瓶颈^[475]。

与此同时，运行期协作治理逐渐从基于规则的调度，转向动态与学习驱动的编排策略。相关工作通过在执行过程中根据任务状态、历史表现或资源消耗，动态选择、排序或采样要激活的智能体与协作结构，使系统能够在性

表 4.10: 多智能体协作结构的构建机制

构建机制	结构确定方式与系统特性	相关工作
预定义结构	协作结构由人工规则或固定模板指定，执行过程中不发生变化，实现简单但适应性有限。	早期方法
搜索式结构构建	通过搜索、回溯或比较不同协作路径，逐步确定任务适配的协作结构，具备纠错能力但开销较高。	AFlow ^[438] , SWE-Search ^[469]
生成式结构构建	根据任务描述或上下文条件，由模型直接生成协作结构，降低人工设计成本并增强跨任务适应性。	Assemble ^[470] , MAS-GPT ^[471]
策略化结构选择	将协作结构建模为可选择的策略空间，根据任务状态动态选择或采样结构，实现性能与资源的自适应权衡。	MAS ² ^[472] , Agentic Supernet ^[439]

能、稳定性与计算成本之间进行自适应权衡^[476-477]。

随着多智能体协作规模与开放性的提升，运行期治理的范围也开始扩展至异构与分布式环境，使来自不同来源、具备不同能力的智能体能够被统一接入、动态组队并在受控的会话流程下协作^[478]。此外，一些研究不再仅在推理阶段进行治理，而是通过训练或后训练方式将协作行为内化到模型中，或探索在连续隐空间中进行协作与信息交换，以进一步降低运行期开销并提升协作效率^[479]。

表 4.11: 多智能体运行期协作调度与治理机制

治理机制	系统作用与特性	相关工作
分层编排与调度	将任务分解、执行与全局监控解耦，支持大规模并行协作与稳定运行。	MegaAgent ^[473] , HALO ^[474]
跨团队并行协作	并行探索多条协作路径，并在关键阶段进行结果聚合，突破单流水线性性能上限。	Cross-team Orch ^[475]
动态学习型调度	根据运行期状态动态选择或排序智能体与协作结构，实现性能—成本权衡。	Maporl ^[476] , Evolving Orch ^[477]
异构智能体开放协作	支持异构第三方智能体的统一接入、自动组队与受控协作。	Internet of Agents ^[478]
隐状态协作	在连续隐空间中进行协作与信息交换，降低通信成本与推理时延。	Latent Collaboration ^[479]

评测基准与机制分析

随着多智能体协作框架复杂度的提升，仅以最终任务成功率作为评价标准已难以全面反映系统能力。2025 年的研究开始系统构建面向多智能体的评测基准与分析方法，从协作质量、资源效率与用户对齐等多个维度刻画协作过程本身^[480-483]。这些工作为不同协作机制之间的比较提供了统一参照，也为机制级分析与系统设计提供了基础。

4.8.3 未来展望

前面回顾总结了 2025 年多智能体协作框架在通信、结构构建与运行期治理等方面的进展，相关研究推动协作从人工模板走向更系统化、可扩展的范式。基于今年的趋势，我们认为后续工作可能更关注在复杂长程任务中的稳定性、效率与可控性，而非仅停留在功能可行性验证。

具体而言，我们认为**资源约束协作**仍将是关键议题：在显式约束 token、调用次数与时延的条件下，如何保持稳定协作增益并提升规模化效率。其次，**运行期执行治理与鲁棒性**有望进一步深化，包括失败恢复、动态调度以及跨团队/跨模型协作一致性维护等。最后，**评测与机制刻画**预计将更过程化与细粒度化，以统一衡量信息流动、分工协调与用户对齐等维度，支撑不同协作策略在“质量—成本—对齐”之间的比较与取舍。

4.9 本章小结

本章围绕 2025 年大语言模型驱动的智能体技术进展，系统梳理了智能体能力的演进。随着大模型在推理、规划与指令遵循能力上的持续增强，智能体逐步具备了自主任务规划、工具链整合、检索增强生成（RAG）、长期记忆以及自我反思与自我进化等关键能力，并在 GUI Agent 与多智能体协作框架中体现出更强的系统化特征。这一阶段的进展标志着智能体已从简单的流程编排迈向以认知与决策为核心的复杂系统形态。

总体而言，2025 年的智能体发展标志着大语言模型应用范式的关键转折：从以模型能力展示为中心，转向以任务执行、系统协作与实际价值创造为导向。尽管智能体在稳定性、评估体系与治理机制等方面仍面临挑战，但其作为大模型应用基础形态的地位已逐步确立。随着模型能力、系统设计与行业实践的持续协同演进，智能体有望在更广泛的真实场景中发挥核心作用，成为推动大模型应用深化的重要支点。

第五章 大语言模型的应用进展

随着大模型能力的持续演进，其正逐步走向真实应用体系。其应用路径呈现出清晰的层次结构：底层围绕具体任务构建应用系统，上层则是与领域或行业流程深度耦合的生产力形态。

基于这一逻辑结构，本章将从典型任务应用与行业应用两个层面展开分析，系统梳理 2025 年大语言模型技术的应用体系。

5.1 任务应用

随着大语言模型技术的发展，其应用不再局限于概念验证或简单自动化任务，而开始在真实场景中承担更具复杂性的工作负载。基于此，本节将从应用视角出发，总结智能体在不同任务类型下形成的典型应用模式。

5.1.1 大模型与脑科学

当前，大语言模型（LLMs）与脑科学的交叉研究正步入“双向赋能”的深度融合期。一方面，传统脑信号处理受限于数据的异质性、稀缺性及极低



图 5.1: 大模型与脑科学融合研究示例

的信噪比 (SNR)，难以构建通用的认知表征；另一方面，LLMs 虽然在逻辑与语言上表现惊人，但在能效比、长程规划及可解释性上仍与生物脑存在显著差距。这种交叉研究的必要性体现在两个方向：

1. **LLM → Brain (赋能神经科学)**：利用大模型的预训练范式和生成能力，将大脑活动视为一种特殊的“自然语言”，实现跨被试、跨任务的通用脑基座模型构建，并推动开放域的神经解码。
2. **Brain → LLM (启发人工智能)**：借鉴大脑的稀疏连接、模块化分区及高能效机制，解决 LLMs 的计算能耗瓶颈，并通过认知对齐揭示机器智能的生物学合理性。

下面，本部分对这两方面的内容在 2025 年的相关进展进行总结讨论。

LLM → Brain：神经科学的“大模型时刻”

长期以来，脑科学与人工智能的融合陷于“分而治之”的任务局限中——研究多针对运动分类或皮层重建等特定场景构建模型。然而，此类方法难以克服脑数据异质性强、信噪比 (SNR) 极低等固有瓶颈。借鉴大语言模型在 NLP 领域的范式突破，脑科学正迈向“基座模型时代”，旨在利用预训练技术打破局部任务的限制，实现通用的神经解码与表征赋能。

脑基座模型：神经动力学的通用语言 受大模型在语言建模领域的启发，研究者开始将脑电 (EEG) 信号视为具有内在语法与语义的“自然语言”，通过大规模预训练构建通用的神经动力学表征。这一领域的进展主要体现在以下三个核心维度：

1. **神经缩放定律与规模化跃迁**：脑电模型已迈入“十亿参数”时代。Jiang et al.^[484] 提出的 LaBraM 模型率先验证了 EEG 数据上的神经缩放定律 (Scaling Law)。随后，Yue et al.^[485] 通过自回归预训练范式构建了 11 亿参数的 BrainGPT，证明了“下一段信号预测”能有效捕捉复杂的时序依赖。作为 2025 年的代表性工作，NeuroLM^[486] 将数据规模推升至 25,000 小时，参数量达 17 亿，标志着脑科学 AI 正经历其“GPT-3 时刻”，性能随规模呈清晰的对数线性增长。
2. **信号建模范式与可解释性突破**：针对 EEG 信号低信噪比与分布漂移的挑战，离散化标记 (Tokenization) 成为核心技术。LaBraM 奠定了

模型	核心贡献
NeuroLM ^[486]	提出通用多任务框架，展现出强大的迁移能力。 采用 Patch-based Masking 策略提升表征效率。
EEGPT ^[487]	针对 EEG 低信噪比设计，提出抗噪掩码策略。 引入时空对齐机制，显著提升跨被试鲁棒性。
BrainGPT ^[485]	构建了 1.1B 参数的超大规模模型。 证明自回归范式能有效捕捉长程脑电时序依赖。
WaveMind ^[489]	首个 对话式 EEG 大模型，构建了专门训练集。 将 EEG 映射到统一语义空间，实现了开放域问答。
CodeBrain ^[488]	引入时频解耦 Tokenizer，赋予模型神经生理学可解释性。 结合仿生“小世界网络”设计多尺度注意力。

神经谱标记化的基础；EEGPT^[487] 则通过掩码双重自监督学习与时空表征对齐，显著增强了跨设备场景下的鲁棒性。此外，CodeBrain^[488] 引入时频解耦设计与仿生小世界网络架构，在提升性能的同时，为基座模型提供了明确的神经生理学可解释性。

3. **指令微调与生成式交互：**脑基座模型正从单一任务向多任务通用架构演进。NeuroLM^[486] 首次将“指令微调”范式引入神经科学，通过构建自然语言指令集实现了任务的统一适配。在此基础上，WaveMind^[489] 实现了 EEG、文本与视觉模态在统一语义空间中的映射，成为首个具备对话能力的脑电基座模型。依托 WaveMind-Instruct-338k 数据集，该模型支持开放域问答，使用户能够通过自然语言交互实时获取受试者的认知与情绪状态反馈。

脑信号的生成式解码与重构 随着大语言模型（LLM）的演进，神经解码正经历从闭集分类向开放域连续文本重构的重大范式转变。研究者利用 LLM 的上下文推理与生成能力，将复杂的脑信号直接映射为自然语言，这一进程主要体现在以下三个维度：

1. **端到端语义映射与开放词汇重构：**为突破传统解码对固定词表的依赖，当前研究致力于建立脑信号与 LLM 输入空间的直接对齐。BrainDEC^[490] 与 BrainLLM^[491] 通过联合优化神经编码器与冻结的 LLM，实现了从 fMRI 表征到 LLM 嵌入空间的直接转换。此外，BP-GPT^[492]

将脑信号视为“大脑提示”（Brain Prompt），通过对比学习驱动 LLM 生成符合大脑意图的开放词汇内容。

2. **空间语义分布与神经生理依据：**深入理解大脑皮层的功能分区为“脑-文本”直接接口提供了生理支撑。MindGPT^[493] 利用交叉注意力机制证明，高级视觉皮层（HVC）相比低级皮层蕴含更丰富的语义信息，支持无需像素级重建的直接文本解码。这一发现从生理学层面确立了语义重构的可行性。
3. **认知启发架构与个体差异建模：**针对长序列解码与个体差异，研究者引入了更复杂的认知启发架构。NeuroCreat^[494] 利用混合专家系统（MoE）捕捉受试者的特有创造性信息，实现了对想象画面的重构。而在序列建模方面，CogReader^[495] 模拟人类序贯认知流程，通过“分段解码-总结-上下文传递”机制，显著增强了长序列 fMRI 翻译的连贯性。

Brain → LLM：脑机制启发的认知对齐与模型进化

与上一章节聚焦于大模型作为神经科学研究的强力工具不同，本章节转换视角，探讨生物脑作为导师如何反哺人工智能的发展。尽管现有大模型在语言处理上展现出惊人的能力，但在能效比、泛化机制及可解释性上仍与人类大脑存在显著差距。**Brain → LLM** 的研究范式旨在建立生物智能与机器智能之间的桥梁：首先，通过**认知对齐**量化模型与大脑在表征层面的相似性，揭示智能涌现的神经相关性；其次，借鉴大脑的稀疏连接、模块化等高效计算原理，为**新一代模型架构**的设计提供生物学蓝图；最后，基于对人类认知负荷与情感状态的神经解析，优化**人机交互**范式，实现更符合人类直觉的协作共生。本节将从这三个维度系统阐述脑机制对大模型进化的深远影响。

大模型与大脑的认知对齐机制 探究 LLM 与人类大脑在底层认知机制上的对齐程度，已成为量化模型智能、理解其表征本质的关键。2025 年前后的研究从规模效应、任务设计及表征维度揭示了两者的相似性与差异：

1. **规模效应与微调策略的影响：**模型宏观属性是决定其“类脑”程度的核心变量。Ren et al.^[496] 通过表征相似性分析（RSA）证实，预训练规模的扩大能显著驱动模型表征与人类神经活动趋于一致。然而，基础模型的规模扩展对脑信号对齐的贡献远大于指令微调^[497]，这暗示现有的微调技术尚未能深入触达底层的类脑认知逻辑。

2. **预训练任务对高级认知的模拟：**特定的任务设计直接影响模型对复杂语言的建模水平。相比传统的词级别预测，引入篇章层面的“下一句预测”（NSP）任务能显著增强模型与大脑右半球及多需求网络区域的对齐^[498]。这一发现有力证明了宏观预测机制在模拟人类篇章级深层理解中的关键作用。
3. **抽象语义与具身困境的博弈：**在表征内容上，LLM 呈现出“高抽象、低具身”的特征。虽然纯文本模型已隐式编码了丰富的高级视觉上下文^[499]，但在涉及感觉运动特征（即具身认知）时，无基础（Ungrounded）模型与人脑表征存在显著差异^[500]。这表明 LLM 虽然掌握了概念关联，但在恢复与生理体验相关的认知维度上仍面临严峻挑战。

大脑认知对新一代大模型设计启示 针对深度学习模型能耗过高与系统性规划能力不足的挑战，研究者开始借鉴生物脑的低功耗机制与认知组织结构，推动下一代 AI 架构的演进：

1. **脉冲神经机制与高效计算：**生物脑的稀疏连接与脉冲计算机制为破解能效瓶颈提供了蓝图^[501]。Zhengzheng et al.^[502] 提出的 BrainGPT 模型，创新性地构建了模仿人脑层级处理的“双模型架构”，并利用测试时训练（TTT）框架实现了从 ANN 到脉冲神经网络（SNN）的无损转换。该架构在保持 100% 性能的同时，提升了 33.4% 的能效并显著加速了训练收敛，为植入式脑机接口芯片提供了低功耗的算法支撑。
2. **功能分区原理与系统性规划：**除了底层优化，大脑前额叶皮层的协调机制也为解决 LLM 的长程规划难题提供了思路。Webb et al.^[503] 指出 LLM 缺乏协同处理复杂目标的机制，并据此提出了模块化智能体架构（MAP）。该架构将规划任务分解为冲突监测、状态预测与价值评估等仿脑模块，由专门模块协同执行子过程。实验证明，MAP 在 PlanBench 等基准测试上显著优于标准 LLM，验证了引入“功能分区”原理是提升模型系统性推理能力的有效路径。

表 5.1.1 总结了上述脑启发大模型设计的关键进展。

大脑认知对大模型人机交互启示 大语言模型（LLM）在重塑人类认知活动的同时，也为构建具备认知感知能力（Cognitive-Aware）的增强型人机交互系统提供了关键支撑。当前研究主要聚焦于以下两个方向：

启示维度	代表工作	主要贡献与意义
能效与计算范式	Xu et al. ^[501]	指出 LLM 高能耗不可持续。 确立了借鉴脑神经高效计算机制的演进方向。
底层架构仿生	Zhengzheng et al. ^[502]	实现 100% 性能保留下的 33.4% 能效提升。 证明了生物合理性与大规模计算可兼容。
认知功能模拟	Webb et al. ^[503]	模仿大脑前额叶的冲突监测与状态预测机制。 显著提升了多步推理与长程规划的任务表现。

- 1. **交互过程中的认知效应与神经代价：** LLM 辅助下的高级认知活动引发了对“认知债务（Cognitive Debt）”的关注。Kosmyna et al.^[504] 通过 EEG 分析指出，长期过度依赖 LLM 可能导致神经连接性减弱及批判性思维下降。为量化此类风险，Jiang et al.^[505] 证明了额叶 theta 波功率可作为评估交互中认知负荷的客观指标。然而，如何精准区分“高效辅助”与“过度依赖”，仍受限于神经机制可解释性的不足^[506]。
- 2. **认知感知系统与共情交互的构建：** 将脑信号引入人机回路使系统能够实时感知心理状态。Zhang et al.^[507] 整合 LLM 与多模态技术，实现了对微观认知状态的分类；ARIEL 系统^[508] 则能根据情感状态实时驱动 LLM 的对话策略。为应对隐私与部署挑战，EEG Emotion Copilot^[509] 利用轻量化技术在本地实现了个性化记录生成。尽管如此，信号异构性与模型“幻觉”仍是限制系统在开放域场景下鲁棒性的核心瓶颈^[510]。

总结与挑战 前面回顾总结了 2025 年大模型与脑科学交叉研究在双向赋能方面的关键进展。这些研究共同推动了该领域从早期的分而治之向深度融合的根本性转变。我们再次对当前研究的特点进行总结，并对未来发展趋势进行展望。

首先，今年一大特点是神经科学正式迈入基座模型时代。脑科学的研究不再局限于特定的运动分类任务，而是演变为构建具有通用语法与语义的神经动力学表征（如 NeuroLM, BrainGPT）。神经缩放定律（Scaling Law）在脑电数据上的验证，以及指令微调技术的引入，使得模型能够捕捉复杂的时序依赖并适配多任务统一架构，标志着脑科学 AI 正经历其 GPT-3 时刻。

其次，大模型与脑科学融合研究的效果得到全面提升，呈现出生成式解码、高效仿生与认知交互感知的深度融合趋势。在解码层面，利用大模型上下文推理能力，实现了从闭集分类向开放域连续文本重构的范式转变，建立了脑信号与语义空间的直接映射；在架构层面，通过模仿生物脑的稀疏连

接与前额叶功能分区，有效突破了 AI 在能耗与长程规划上的瓶颈；在交互层面，系统演进为认知感知（Cognitive-Aware）形态，不仅量化了认知负荷，还能根据情感状态实时调整策略，构建了更符合生物直觉的共情交互回路。

我们认为，未来的研究将不再满足于单向的赋能或启发，而是将重心集中在具身认知的落地以及闭环共生系统的构建。具体来讲，未来的重心可能转向具身基础（Grounding）研究，让模型真正理解与生理体验相关的感觉运动特征；在此基础上，进一步探索“脑-机”在线协同进化，实现实时利用脑反馈信号微调模型策略的能力，甚至通过脑机接口实现人类意图与 AI 生成的无缝流转，构建真正符合生物学合理性的通用智能架构。

5.1.2 编程助手

应用背景

编程助手的发展历经多轮技术迭代，从早期辅助工具逐步向智能化协作方向演进，为 2025 年的关键突破奠定了基础。早期阶段，编程助手以语法规则匹配和简单代码补全为核心，如 IDE 自带的智能提示功能，仅能响应预定义关键字，无法理解代码上下文与项目逻辑，辅助价值局限于提升编码效率的基础层面。2021 年后，大规模语言模型（如 Codex、GPT 系列）的应用推动编程助手进入智能化跃迁期。这一阶段的核心突破在于实现代码与自然语言的跨模态理解，支持多轮对话式编程、上下文感知的代码生成，甚至能完成简单算法题与测试用例编写。以 GitHub Copilot 为代表的工具开始广泛落地，据统计 2023 年其已能贡献项目中 25% 以上的代码，VS Code 插件市场中 AI 编程工具占比超 40%，人机协同编码的范式初步形成。尽管如此，2025 年前的编程助手仍存在明显局限：功能上以单次代码生成为主，缺乏对软件工程全流程的深度参与；能力上难以自主拆解复杂开发任务，无法形成“实现-测试-修改”的闭环；集成度上与代码仓库、CI/CD 等工程环境的联动不足，难以利用项目全局上下文辅助决策。与此同时，行业对开发效率与工程质量的需求持续提升，推动编程助手从“代码生成工具”向具备自主规划、流程协同能力的“软件开发 Agent”转型，这一趋势为 2025 年的研究进展指明了核心方向。

应用进展

2025 年，编程助手围绕“软件开发 Agent”转型目标实现关键突破，核心进展集中在智能协作能力、工程环境融合、功能价值延伸三大维度，大幅

2025 年大语言模型（LLMs）进展报告

表 5.1: 2025 年度主流编程助手对比

研发机构	产品名称	核心功能	优势特点	适用场景	支持平台
Factory.AI	Droid	软件开发 Agent	能在终端环境中完成复杂任务	软件开发	各种 Terminal, IDE, Web, Slack, Linear
Warp	Warp	软件开发 Agent	AI 内嵌到终端中, 可以自然地与 AI 交互	软件开发	MacOS, Linux, Windows
OpenAI	Codex	软件开发 Agent	能在手机和云端上交互	软件开发	各种 Terminal, IDE, 云端, GitHub, 手机
Google	Gemini-cli	软件开发 Agent	具备多模态能力, 能根据 pdf 或者图片来生成 apps	软件开发	各种 Terminal
Verdent AI	Verdent	软件开发 Agent	多个 Agent 并行工作; 执行任务前主动交流明确需求; 复杂工程可视化 Plan Tree	软件开发	MacOS, Linux, Windows, VSCode
OpenHands	OpenHands	软件开发 Agent	框架开源, 可以进行定制 (比如自行选择接入的模型)	软件开发	Terminal, 云端, 移动端
HKUDS	DeepCode	软件开发 Agent	开源; 采用多智能体协作框架; 在 Paper2Code 上超越人类专家表现	软件开发, Paper2Code	Terminal, Web
Anysphere	Cursor	软件开发 Agent	完美兼容 VS Code 插件生态和快捷键, 支持一键迁移配置; 提供隐私保护模式, 企业级数据安全机制; 支持百万行级项目的精准分析	软件开发, 自然语言转代码, 智能代码补全	MacOS, Linux, Windows, vscode
微软与 OpenAI	Copilot	AI 编程辅助	深度集成 Microsoft 365、Azure 等微软生态; 基于 OpenAI 大模型迭代, 代码生成准确性和上下文匹配度持续提升	个人开发者日常编码、企业级云开发、CRM 系统搭建	Web 浏览器、IDE、Windows 命令行
字节	Trae	Agent 化 IDE	覆盖开发全流程、支持 MCP 协议扩展; 遵循“本地优先”原则, 保障数据隐私	中大型项目的自动化开发, 跨工具协作、任务拆分的复杂工程	MacOS, Linux, Windows、独立 AI IDE
阿里	通义灵码	软件开发 Agent	兼容主流 IDE、企业版支持知识库检索增强; 首批通过信通院“可信 AI 智能编码工具”4+评级	日常代码开发与补全、问答与代码解释、Agent 模式下的任务执行	MacOS, Linux, Windows、各种 IDE
腾讯	CodeBuddy	软件开发 Agent	中文支持行业领先、企业级工程化适配	中文开发者个人项目、中文团队协同开发、复杂工程项目开发	MacOS, Linux, Windows, 主流 IDE
Trelis	Tabnine	AI 编程辅助	补全速度行业第一; 嵌入式/底层开发适配 (C/C++、Rust、汇编等准确率高)	极速开发效率的开发者、嵌入式/底层开发场景	MacOS, Linux, Windows, 主流 IDE

提升对软件工程全流程的支撑力。多智能体协作架构落地，复杂任务处理能力升级。主流编程助手整理如图 5.1 所示，普遍采用多智能体协同设计，核心智能体可依据开发目标自主拆解任务，调度代码生成、重构、测试等子智能体分工协作，形成“规划-执行-验证”的完整闭环。例如 Trae 的 SOLO Coder 智能体¹可联动多个专项智能体推进复杂项目迭代，GitHub Copilot 推出的 Workspace²功能能在代码仓库 Issue 中启动 AI 驱动的开发流程，自动规划需求到实现的完整路径，显著提升复杂项目的推进效率。工程环境深度融合，上下文感知能力全面强化。编程助手突破原有插件式集成模式，实现与 IDE、云开发环境、代码仓库的原生融合。一方面可精准读取控制台输出、调试堆栈、Git 提交历史等环境数据，提供更贴合项目实际状态的辅助建议；另一方面支持跨平台与多工具链适配，如 JetBrains AI Assistant 可深度感知 IDE³运行状态并支持第三方模型灵活接入，Amazon CodeWhisperer 与 AWS 云服务深度绑定⁴，能生成符合云架构最佳实践的代码并提供安全扫描能力。功能价值向工程全流程延伸，协作与管控能力提升。编程助手的功能边界从代码生成拓展至工程全生命周期，新增智能重构、依赖分析、安全合规检测、性能优化等工程化支持能力。同时强化团队协作适配，如支持多人开发场景下的上下文同步与任务协同，通过代码变更可视化功能让 AI 操作透明可追溯。此外，所有智能辅助功能均保留开发者主导权，确保关键设计与发布决策的可控性，平衡自动化效率与工程可靠性。

未来展望

回顾前文可知，2025 年编程助手在核心技术与应用场景上实现了关键突破，不仅形成了多智能体协作的“规划-执行-验证”闭环，更实现了与工程环境的原生融合，功能边界也从单一代码生成延伸至开发全流程，成功完成从“代码补全工具”向“软件开发 Agent”的转型，为人机协同开发范式奠定了坚实基础。接下来，我们对编程助手的未来发展趋势进行展望。编程助手在 2025 年的系列突破，已为下一阶段的高质量发展筑牢根基。我们猜测未来的研究将不再满足于基础功能的完善，而是将重心集中在企业级确定性协同深化以及自适应人机协同与开发能力普惠两大方向。具体来讲，首先随着企业数字化转型对研发效率与安全合规的双重需求升级，编程助手将进

¹<https://www.trae.cn/>

²<https://github.blog/news-insights/product-news/github-copilot-workspace/>

³<https://www.jetbrains.com/zh-cn/ai-china/>

⁴<https://docs.aws.amazon.com/codewhisperer/>

进一步深化企业级协同适配能力，聚焦全链路协作的确定性与规范性，更好地融入企业研发体系并满足其核心要求，同时拓展跨团队、跨领域的协作边界，破解多元协作中的壁垒问题。另外，我们猜测未来将持续深化自适应人机协同能力，推动开发能力普惠落地。通过优化人机交互逻辑，让编程助手更好地适配不同开发者的使用习惯与团队协作规范，降低人机协作摩擦。同时进一步简化开发门槛，依托自然语言交互的优化让非专业人员也能参与基础开发工作，推动开发能力向更多业务场景下沉。

5.1.3 写作助手

应用背景

近年来，大语言模型（LLMs）在写作辅助领域的应用迅速升温，逐渐从零散的功能性探索发展为具有明确研究边界与应用深度的产业热点。截至 2025 年末，该领域已完成从单一的“文本生成”向深度的“认知协同”的范式跨越。这一转型的核心驱动力在于模型能力从快速、直觉式的文本补全向慢速、深思熟虑的逻辑规划与多步推理的跃升。

实证数据显示，LLM 辅助写作在关键行业的渗透率已达到显著规模。据 2025 年第一季度的统计，约 18% 的金融消费者投诉文本和高达 24% 的企业新闻稿系由 LLM 生成或经过深度辅助润色^[511]。当前的写作助手已不再局限于基于概率的下一词预测，而是演进为具备层级化逻辑规划、外部工具调用及领域知识深度的智能体。

应用进展

2024 年至 2025 年间，大语言模型的研究重心经历了显著的结构性调整，从单纯追求参数规模的扩展（Scaling Laws），转向了对智能体认知架构的深度重构与推理机制的精细化设计。

结构化推理与分层规划 (Structured Reasoning & Planning)：随着应用场景从短文本对话向长篇小说创作、复杂代码工程及科学发现拓展，单一的“预测下一个 Token”的自回归模式已遭遇瓶颈。针对长文本生成中的逻辑断裂问题，研究逐渐摒弃线性生成，转向分层规划。WriteHERE 框架将写作任务建模为包含检索、推理、创作的异构递归图，实现了动态规划与执行的交替^[512]。CogWriter 则通过“规划代理”与“生成代理”的双层架构，模拟了人类写作中的监控与回顾过程^[513]。

多智能体协同 (Multi-Agent Collaboration)：单体智能的局限性催生了多智能体系统(MAS)的繁荣,群体智能成为解决复杂任务的新范式。Chain of Agents (CoA) 通过分块处理与信息接力,有效解决了超长上下文中的注意力分散难题^[514]。Debate-to-Write 利用持不同立场的角色人格进行辩论,以对抗性协作提升了论证的深度与客观性^[515]。LatentMAS 更是实现了智能体在隐向量空间的直接通信,大幅降低了计算开销^[516]。

人机共创与意图理解 (Human-AI Co-Creation)：尽管智能体内部的认知与推理能力不断增强,但如何精准捕获人类模糊、流动的创作意图,并保持人类在创作过程中的主导权,仍是人机交互领域的深层挑战。最近的研究正推动交互范式正从简单的提示词工程转向意图流的可视化管理。IntentFlow 和 Intent Tagging 等系统都允许用户通过可视化组件显式管理模糊意图,支持非线性的共创流程^[517-518]。

评估新维度 (Evaluation)：随着生成能力的提升,传统的基于 n-gram 的评估指标(如 BLEU、ROUGE)在评估开放式创意写作时表现不佳。评估指标正从客观准确性转向基于模型的主观评估、风格一致性量化以及自我修正能力的考察。WritingPreferenceBench 的研究揭示了传统奖励模型在评估创意、风格等主观维度时的局限性,提出了基于生成式思维链的新型评估方法^[519]。最新的研究还提出了多指标集成(Ensemble of Metrics)的评估范式,用于可靠评估风格个性化文本生成任务^[520]。

典型应用产品与平台技术对比

随着底层技术的成熟,通用型聊天机器人已难以满足专业场景的垂直需求。市场格局迅速分化,形成了针对学术、创意、企业营销及技术代码等特定领域的深度应用生态。

学术与科研写作 (Academic & Scientific Writing) 该领域的工具以事实准确性、引用规范性以及对学术语体的严格遵循为核心竞争力。

表 5.2: 学术写作主流工具对比

产品名称	核心定位	技术底座	功能亮点
Paperpal ^[521]	语言润色与投稿合规性检查	专有模型（基于学术文献和编辑模式训练），上下文感知语法检查	提交准备度检查： 模拟期刊审稿标准，深度集成 MS Word，显著降低拒稿率。
Jenni AI ^[522]	交互式初稿撰写与自动补全	生成式 AI 补全引擎，RAG 实时查重与引用生成	交互式补全： 类似 Gmail 智能补全但针对长文，用户可逐句确认或修改。
知网 AI ^[523]	全流程科研助手与可信知识服务	华知大模型（Huazhi LLM），基于 CNKI 全库的增强检索 (RAG)	可信增强： 答案严格溯源至 CNKI 权威期刊/学位论文，杜绝幻觉。
Scifocus ^[524]	综合型科研助手与论文结构化	集成专有知识库与公共数据库，自动化论文结构生成	实验设计辅助： 生成详细大纲和实验设计建议，确保逻辑流。
Google NotebookLM ^[525]	溯源型知识综合与科研助理	Gemini, Deep Research Agent	多模态溯源： 严格基于上传文献回答（零幻觉），生成播客式音频概览与深度研报。
秘塔写作猫 ^[526]	中文原生写作与纠错平台	MetaLLM（自研大模型），NLP 纠错引擎	本土化润色： 深度适配中文语法与标点规范，提供公文/学术语体转换与结构化续写。

产品名称	核心定位	技术底座	功能亮点
星火科研助手 ^[527]	智能科研知识服务平台	讯飞星火大模型 (X1), 中科院文献情报中心数据	深度科研支持: 联合中科院打造, 提供基于权威数据的智能综述生成与模拟审稿人视角的论文预审功能。

技术分析：Paperpal 采取了“后编辑”的技术路线，针对学术纠错与风格润色进行了特定微调，以最小化幻觉风险。Jenni AI 采用了“人机共依”的交互模式，利用 RAG 技术实时连接文献库，动态推荐引用来源。Scifocus 则代表了 2025 年的新趋势，即深度垂直整合。它不仅辅助写作，还涉足实验设计和逻辑构建，试图介入科研的更上游环节。Google NotebookLM 则通过完全的“源文档溯源”机制（Source-Grounding）解决了幻觉问题，其 Deep Research 功能将 AI 从阅读者升级为能主动搜集资料的研究员。在中国市场，秘塔写作猫则通过自研的 MetaLLM 构建了中文写作的本土化护城河，特别是在公文写作与中文语法纠错（如标点、成语误用）上展现了比通用模型更强的鲁棒性。CNKI AI 学术研究助手与科大讯飞星火科研助手则共同代表了“国家队”的技术路径：前者结合华知大模型强调 CNKI 数据的可信增强，后者则联手中科院文献情报中心，利用星火大模型（特别是 X1 深思模型）的逻辑推理能力，实现了从权威文献检索、智能综述生成到论文预审的全流程覆盖，有效解决了科研“冷启动”阶段的信息过载问题。

创意与叙事写作 (Creative & Narrative Writing)

创意写作工具主要解决长文本的一致性（如角色性格、情节逻辑）及风格的个性化迁移问题。

表 5.3: 创意写作主流工具对比

产品名称	系统定位	核心架构	AI 模型
Sudowrite ^[528]	结构化叙事脚手架	Story Bible: 结构化存储角色与世界观，生成时动态注入。	动态路由集成 GPT-4o, Claude 3.5 及专有微调模型 Muse。系统根据任务类型自动选择最优模型。
NovelAI ^[529]	生成式开放沙盒	Lorebook: 基于关键词和正则表达式触发的上下文注入机制 (RAG)。	基于 Llama3 和 NeoX (Kayra) 进行小说语料的深度继续预训练。完全私有化部署，不依赖第三方 API。
CreativeFlow ^[530]	情感计算增强型助手	情感共鸣分析器: 确保写作击中预设情感目标。	推测基于主流商业 LLM API 进行二次开发，外挂情感分析模块与体裁适配器。
彩云小梦 ^[531]	AI RPG 与世界模拟器	DCFormer^[532]: 动态可组合多头注意力，大幅提升参数效率与角色一致性。	基于自研小说续写通用模型“云锦天章”，专为互动叙事与角色扮演优化。
蛙蛙写作 ^[533]	长篇个性化叙事引擎	LPA: 终身个性化模型，结合动态长短期记忆机制管理伏笔。	自研 Weaver 2.0, 支持“小说-剧本”全链路转化与私有化风格养成。

技术分析：Sudowrite 通过 Story Bible 显式记忆管理系统解决长篇叙事连贯性问题。NovelAI 则通过自研模型 Kayra 避开通用模型安全护栏，并利用 Lorebook 提供极高的可配置性。CreativeFlow 的关注点从单纯的“文本生成”转向了“情感体验管理”。它不仅是一个生成工具，更被定义为一个“情感与风格的管理者”，旨在帮助作者在保持个人声音的同时，增强作品的情感穿透力。在中国市场，彩云小梦利用 DCFormer 架构突破了传统 Attention 机制的效率瓶颈，专注于“世界设定”与“角色卡”的深度绑定，确保在互动叙事中角色性格不漂移。蛙蛙写作则通过 LPA 技术引入了“动态长短期记忆”，有效区分即时剧情与长期伏笔，解决了长篇连载中的逻辑崩坏问题。

企业与营销写作 (Enterprise & Marketing)

企业级应用聚焦于品牌一致性 (Brand Voice)、数据安全性及规模化生产。

表 5.4: 企业级 AI 营销平台对比

产品名称	系统定位	核心架构	AI 模型
Jasper AI ^[534]	品牌营销操作系统	Jasper IQ : 专有的品牌知识库，存储并检索风格指南、过往佳作与战略文档。	混合模型路由：动态路由至 GPT-4、Anthropic 或 Google 模型，并经过营专有数据微调。
Copy.ai ^[535]	GTM 自动化引擎	Workflow OS : 基于提示词链的编排引擎，支持逻辑分支、循环及多模型调用。	多模型选择器：允许用户在工作流的每个节点手动选择最适合的模型（如用 GPT-4 分析，用 Claude 3.5 写作）。

产品名称	系统定位	核心架构	AI 模型
百度擎舵 ^[536]	AIGC 创意生产平台	DeepSeek	
		Integrated：深度集成推理模型，支持从业务卖点到视频脚本的端到端生成，并赋能“商家智能体”进行深度业务对话。	DeepSeek + 文心一言：结合 DeepSeek 的深度推理能力与文心大模型的语义生成能力，实现“千人千面”的营销素材生产。

技术分析：作为该领域的领头羊，Jasper 在 2025 年已经进化为一个完整的营销操作系统。核心技术 Jasper IQ 采用 RAG+ 微调的混合架构，利用“营销智能体”实现端到端任务执行；Copy.ai 则侧重于高通量 workflow 自动化，转型为可编程的“增长黑客操作系统”，利用 workflow（Workflows）和提示词链（Prompt Chaining）技术实现 Go-to-Market（GTM）流程的自动化。中国市场的代表百度擎舵则展示了另一条路径，通过全面接入 DeepSeek 推理模型，将 AI 的能力从单纯的内容生成拓展至深度的逻辑推理与销售转化，特别是其“商家智能体”和视频流 workflow，通过推理引擎实现了极低成本的素材量产与高转化率的自动导购。

关键挑战及未来发展

尽管写作助手领域在技术与应用上都取得了长足进步，但仍面临严峻挑战，这些挑战同时也指引了未来的发展方向：

幻觉的持久性与准确性悖论：尽管基础模型能力大幅提升，但“幻觉”仍未被彻底根除。根据 2025 年 12 月的幻觉排行榜，表现最好的模型（如蚂蚁集团的 antgroup/finix_s1_32b）在摘要任务上的幻觉率已降至 1.8% 左右，但通用大模型在面对特定事实查询时，幻觉率仍普遍在 3-6% 之间^[537]。更令人担忧的是，Grok-3 在识别新闻来源的具体任务上曾被测出高达 94% 的错误率，显示出模型在溯源能力上的极端不稳定性^[538]。这迫使高风险领域必须保留“人机回环”（Human-in-the-loop）机制。

版权与合规的法律风险：2025 年是 AI 版权诉讼的爆发年。这不仅是法律问题，更直接影响到写作工具的商业模式和数据来源。美国版权局（USCO）持续维持“非人类创作不受版权保护”的立场。虽然包含大量人类编辑的 AI 辅助作品可以获得部分保护，但“人类贡献”与“AI 生成”的界限依然模糊^[539]。这给使用 Sudowrite 等工具创作小说的作者带来了确权难题。

内容同质化与认知侵蚀：研究显示，由于 LLM 主要基于英美语料训练，它们在生成故事时倾向于强加特定的“叙事结构”（如英雄之旅、小镇和解等），即使在被要求生成非西方文化背景的故事时也是如此。这种“叙事同质化”正在导致全球文化表达的扁平化^[540]。同时，教育界担忧学生因过度依赖 AI 会导致“认知脱钩”，从而丧失批判性思维能力，这引发了对教育技术伦理的广泛讨论^[541]。

总结与展望

综上所述，2025 年的 AI 写作助手领域已完成从“生成工具”到“认知基础设施”的质变。这一阶段的显著特征在于：垂直深度的确立与认知架构的系统化。市场不再单纯为文本生成的效率买单，而是转向寻求能够解决复杂问题、具备行业深度和认知完整性的解决方案。

成功的平台已从单一的“提示词-响应”模式，演进为整合了检索增强、微调和长上下文技术的复杂系统。这种技术组合不仅解决了事实准确性与风格一致性之间的矛盾，更通过智能体工作流实现了对复杂任务的自动化拆解与执行。对于用户而言，人机协同的重心正从“指令输入”转向“意图管理”与“逻辑审核”。尽管幻觉控制与版权合规仍是制约其全面自动化的关键瓶颈，但随着神经符号 AI 与多智能体协同技术的持续迭代，下一代写作系统正逐步确立其作为人类“思维伙伴”的地位，在大幅提升生产效率的同时，重塑着知识生产与传播的底层逻辑。

5.1.4 设计助手

研究背景

在 AI 技术重塑设计领域的进程中，AI 设计助手对用户的核心意义，在于其正从一个“效率工具”演进为一名“智能共创伙伴”。它不仅是能力的延伸，更是思维与创造过程的拓展。AI 设计助手领域正在经历深刻的技术演进。从 2023-2024 年以静态内容生成为主要特征的发展阶段，到 2025 年

逐步涌现的交互式、多模态设计能力，这一领域展现出从工具辅助向深度集成发展的趋势。早期 AI 设计工具主要聚焦于单一任务的内容输出——如图像生成、文案撰写或界面元素创建。而当前的 AI 设计助手开始具备理解设计意图、协调多元素协同、支持迭代优化的能力，在设计 workflows 中的角色正从被动执行向主动协作转变。这种演进不仅体现在技术能力的提升，更反映在设计师与 AI 交互模式的重构：从指令-执行的单向模式，向对话-共创的双向协作模式发展。

研究进展

2025 年，AI 设计助手领域呈现出从技术探索向生产应用深度融合的演进态势，特别在生成式界面设计、PPT 与文稿智能生成、视觉内容创作与媒体生成、设计系统与品牌资产管理这四个方面具有较大的研究进展。

生成式界面设计 2025 年，交互界面生成与原型设计领域持续发展。以 Google Gemini 3 的生成式 UI 能力^[542]为代表，部分设计工具开始探索从静态界面生成向动态交互体验生成的转变——AI 不仅生成界面元素，还能根据用户输入创建具有一定交互能力的界面原型。这类工具的价值在于缩短了从概念到可演示原型的开发周期，为产品经理、设计师等角色提供了快速验证想法的新途径。当前发布的工具开始具备多模型集成、与设计系统对接、自动生成代码等功能。例如，Google Labs 推出的实验性项目展示了多 Agent 协作的原型设计能力^[543]，Figma 则在其开发者大会上演示了 AI 辅助原型制作的新功能^[544]，这些进展显示 AI 在设计工具链中的应用正在深化。

PPT 与文稿智能生成 2025 年，演示文稿与文档生成领域出现了显著的技术进展。以 Gamma 3.0^[547]的发布为代表性案例，该平台推出的 AI Agent 功能引入了增强的协作生成能力——支持用户上传手写笔记、草图或会议截图，系统能够整合这些信息、检索相关内容，并生成演示文稿^[548]。这一功能扩展了传统 AI 工具在信息处理和内容生成方面的能力范围。Microsoft 在推出的 PowerPoint Agents^[549]体现了 Office 套件在 AI 集成方向的探索。该功能允许用户在 Copilot Chat 中创建演示文稿，并通过自然语言指令完成文本编辑、表格插入、布局调整等操作，同时可以关联组织内的文件、会议记录与邮件内容^[550]。在技术实现层面，Gamma 3.0 提供的 API 功能^[551]使企业能够将演示文稿生成能力集成到现有业务系统中，支持基于模板的批量

表 5.5: 交互界面生成与原型设计工具

产品名称	核心定位	技术底座	功能亮点
Gemini 3 Generative UI ^[542]	动态生成完整用户界面的革命性系统	Gemini 3 Pro 多模态模型	实时生成交互式网页、工具和应用；Dynamic View 和 Visual Layout 双模式；90% 用户偏好超越传统网站
Google Antigravity ^[543]	Agentic 开发平台，任务导向型 IDE	Gemini 3 Pro、Claude Sonnet 4.5 等多模型支持	多 Agent 编排；Chrome 浏览器自动化；设计转代码；3D 图形支持；预览期完全免费
Figma Make ^[544]	提示词转应用的快速原型工具	Claude 3.7	导入设计库保持品牌一致性；Supabase 后端集成；生成 React/Tailwind 代码；AI 积分计费系统
Google Stitch ^[545]	实验性 AI 驱动 UI 设计工具	Gemini 3	多屏幕原型连接；交互流程设计；从概念到工作原型快速转换；Prototypes 功能
Google Opal ^[546]	自然语言构建 AI 小应用	Gemini 模型家族	可视化编辑器；工作流自动化；与 Gemini Gems 深度集成；无需编程即可创建应用

2025 年大语言模型（LLMs）进展报告

定制化生成。Microsoft 的 Agent Mode 采用了包含规划、验证与优化阶段的处理流程，该功能已在 Frontier 计划中向部分订阅用户开放测试^[552]。

表 5.6: PPT 与文稿智能生成工具

产品名称	核心定位	技术底座	功能亮点
Gamma 3.0 ^[547]	AI 驱动的视觉叙事平台	多模态 AI 模型	Gamma Agent 实时网络研究与内容整合;API 支持批量个性化生成;Smart Diagrams 自动可视化;自然语言编辑指令;2025 年 9 月发布
Microsoft PowerPoint Agents ^[549]	Office 生态 AI 协作代理	GPT-5 Chat + Code-optimized Reasoning	Chat 中直接生成完整演示文稿;Agent Mode 应用内自然语言编辑;组织数据深度关联;多阶段推理验证;2025 年 11 月发布
Microsoft Copilot in Power-Point ^[550]	应用内 AI 增强功能	GPT-5 系列模型	Explainer 功能即时解释复杂内容;Image Editor 图像编辑集成;on-canvas 演讲稿生成;多语言翻译;2025 年 10-12 月滚动发布
Canva Magic Design for Presentations ^[553]	品牌化演示生成工具	GPT-4 + 自研设计引擎	大纲预览与结构调整;Brand Kit 自动应用品牌规范;Magic Switch 多格式转换;Magic Write 语音与风格优化;持续迭代中
Gamma API ^[551]	企业级自动化生成接口	RESTful API + Webhooks	与 Zapier/Make/Workato 集成;单模板生成 100+ 定制演示;CRM 数据自动转换为客户演示;会议转录自动成总结文稿;2025 年 9 月 Beta 发布

视觉内容创作与媒体生成 视觉内容创作与媒体生成在 2025 年展现出从单帧图像到多帧视频的技术能力提升，AI 视觉创作工具在生产环境中的应用场景不断拓展。本年度重要进展包括 Runway Gen-4 系列^[554]在视频生成的角色与场景一致性方面的技术改进，这些进展为 AI 辅助视频创作提供了更可靠的技术基础，并促成了 Runway 与 Lionsgate 等影视制作公司的商业合作探索。在图像生成领域,FLUX.2^[555]等开源模型的发展为企业提供了在质

量与数据控制之间权衡的更多选择。多模态生成能力的整合成为一个值得关注的方向。Kling 2.6^[556]等工具开始支持从单一提示词同时生成视频、音频及视觉风格的功能，体现了创作流程简化的趋势。Adobe 和 Figma 等成熟设计平台通过将 AI 功能集成到现有工作流（如 Adobe Firefly 的图像调和功能^[557]、Figma 的集成式图像编辑工具^[558]），降低了用户的使用门槛，推动 AI 能力从独立应用向设计工具链的深度整合发展。

表 5.7: 视觉内容创作与媒体生成工具

产品名称	核心定位	技术底座	功能亮点
Runway Gen-4 系列 ^[554]	专业级 AI 视频生成平台	自研视频生成模型	解决角色场景跨镜头一致性；单图生成多场景；现实物理模拟；Video Arena 排行榜第一
Nano Banana Pro ^[559]	高保真图像生成与编辑模型	Gemini 3	工作室级视觉质量；复杂任务处理能力；从即兴艺术到专业级输出；Pro 版提供更强创意控制
Kling 2.6 ^[556]	音视频融合生成工具	-	一个提示词生成视频、声音和风格；音视频生成融合趋势代表；创意工具集成化
FLUX.2 ^[555]	开源图像生成模型	开源权重架构	文本渲染准确性；适合 UX 设计稿、包装设计；数据控制与合规性；支持离线部署和定制化
Adobe Firefly (2025 更新) ^[557]	集成于 Creative Cloud 的生成式 AI 套件	Adobe 自研模型	Harmonize 场景融合；Premiere Pro 中的 AI Object Mask；跨应用 AI 功能；与专业工具深度整合
Figma 图像编辑工具 ^[558]	Figma 内置 AI 图像处理	-	Erase object 擦除对象；Isolate object 隔离编辑；Expand image 背景扩展；无需外部工具的完整工作流

设计系统与品牌资产管理 设计系统与品牌资产管理方向在 2025 年面临的重要议题之一，是如何在 AI 辅助的内容生产中维持品牌一致性与设计质量。随着企业多渠道营销素材需求的持续增长，传统的人工逐一制作模式在效率上面临挑战，而 AI 工具的应用为规模化生产提供了新的可能性。Canva 发

布的自研设计模型^[560]展现了设计平台的技术演进方向——该模型能够处理设计图层、格式等结构化信息，生成可编辑的设计文件而非单纯的图像输出。Figma 的一系列新功能也体现了类似趋势：其推出的营销设计功能^[561]允许在保持核心品牌元素约束的同时，为非设计专业用户提供一定的创作灵活性；Figma Sites^[562]则尝试简化从设计到网站发布的工作流程。Adobe Express 集成了对话式 AI 能力^[563]，MockU^[564]专注于演示文稿的快速生成，这些工具在不同环节提供了 AI 辅助功能。这些发展反映出设计工具的两个演进方向：一是从模板库向具备一定设计理解能力的智能系统发展；二是从单一功能工具向覆盖更多 workflow 环节的平台扩展。在这一背景下，设计系统逐渐从静态的规范文档演变为可被计算机系统解析和应用的结构化标准，但这一转变仍处于探索阶段，不同工具在实现路径和成熟度上存在差异。

表 5.8: 设计系统与品牌资产管理工具

产品名称	核心定位	技术底座	功能亮点
Canva 设计模型 ^[560]	自研基础设施模型平台	Canva 自研设计理解模型	生成可编辑分层设计；跨格式工作（社交媒体、演示、网站）；3D 对象生成；艺术风格复制
Figma Buzz ^[561]	品牌营销资产批量生成工具	基于 Figma 设计系统	营销团队快速创建素材；保持品牌一致性；非设计师友好；锁定核心品牌元素功能
Adobe Express AI Assistant ^[563]	对话式设计创作平台	Adobe Firefly + 语义理解引擎	分层 AI 编辑；模板智能定制；跨 Adobe 生态整合；语义理解设计元素
Figma Sites ^[562]	设计到网站发布平台	Figma + AI 响应式布局引擎	自动响应式布局生成；设计直接发布为网站；CSS 代码自动生成；营销页面快速上线
Figma Draw ^[565]	增强矢量编辑与插图工具	-	原生矢量工具升级；笔刷和纹理填充；路径文字功能；高质量视觉设计能力
MockU ^[564]	项目演示快速生成工具	-	高质量专业演示；实用解决方案；2025 年 11 月发布；针对项目展示优化

总结与展望 2025 年 AI 设计助手的核心特征体现为三个维度的范式转变：在生成能力层面，工具已从静态元素生成演进至动态体验构建；在技术架构层面，多模态融合与 Agent 化成为主导方向；在生态格局层面，开源与闭源技术路线并行推动，加速技术民主化。然而，当前阶段的技术成熟度与产业期待之间仍存在显著差距。深层次的结构性挑战主要集中在：AI 对复杂设计情境的整体性理解能力不足，难以像人类设计师那样综合权衡多元约束；设计一致性与创意独特性之间的张力尚未得到根本解决，易陷入模式化生成；人机协作的交互范式缺乏真正的“共同思考”机制；伦理、法律、可解释性等规范框架尚未健全。因此，现阶段 AI 设计助手的定位更接近“效率放大器”而非“创意替代者”，人类设计师在战略判断、用户洞察与审美决策的核心价值依然不可替代。

展望未来，AI 设计助手领域的发展将沿着“深度-广度-融合”三个维度展开。深度上，提升多模态理解能力，从表层的视觉生成深入到理解业务逻辑、用户心智与品牌语义，实现设计的整体把握。广度上，覆盖从需求分析到交付的全流程设计工作流，通过 Agent 架构实现端到端协作支持。融合上，与传统设计工具、企业知识及人类创意深度融合，构建“智能设计基础设施”，重塑创作范式。长远来看，AI 设计助手将成为理解意图、洞察需求、掌握美学的智能协作系统，辅助设计师聚焦战略性思考与创造性突破，而非替代人类。这需要 AI 技术、设计科学、心理学与人机交互等多学科共同推进。

5.1.5 社会模拟

大语言模型被引入社会模拟研究的早期阶段，相关工作主要聚焦于生成式多智能体环境的可行性验证。研究者通过构建规模有限的虚拟社会，展示 LLM 在个体层面生成连贯行为、在群体层面产生涌现现象的潜力。

随着研究的推进，这类“生成式小社会”逐渐暴露出局限性。缺乏明确对齐机制的模拟难以保证群体行为与真实人群分布之间的一致性，生成结果往往依赖于隐式提示与语言先验，难以复现或系统比较；同时，社会行为主要以最终输出的形式被观察，其背后的认知与因果过程不可追踪，使得模拟结果难以支撑干预分析或机制解释。这些问题在一定程度上制约了 LLM 社会模拟向严肃研究方法的演进。

进入 2025 年，通用大模型在对话理解、长程规划与多智能体协作能力上的显著提升，为这一领域带来了关键转折。研究重心开始**从概念验证转向**

系统性能力建设：一方面，社会模拟被重新定位为一种可复现、可干预的“类社会实验”工具，相关工作着力构建支持大规模代理运行的仿真平台，并通过显式的对齐设计提升模拟结果在真实社会情境下的代表性与可信度。另一方面，方法论层面的反思逐渐形成共识，仅生成“看似合理”的社会行为并不足以支撑可靠推断，社会模拟需要进一步引入可追踪、可修正、干预一致的认知过程建模，以弥合行为输出与机制解释之间的鸿沟。

评测与数据资源早期以单一任务或行为输出为主的评估方式，逐步扩展为面向长期战略规划、社会互动与多方谈判等复杂能力的系统测量。社会模拟平台的建设呈现出明显的“研究闭环化”特征，从自然语言驱动的低代码场景构建、可演化的模型调优机制，到支持十万级代理的分布式运行与自动化结果分析，社会模拟正逐步从实验室原型转变为可复用的研究基础设施。

基于上述背景，本节将从**模拟仿真平台、可靠性与偏差、评测与数据基准**三个维度，系统梳理 2025 年大语言模型社会模拟的技术演进与应用格局。

1. 模拟仿真平台

LLM 社会模拟在系统层面的核心进展，体现在仿真平台的规模化、模块化与对齐机制的显式建模上。相较早期以“虚拟社区”为代表的概念验证型系统，本年度工作普遍将目标定位为支持类社会实验的研究基础设施。

以 AgentSociety^[566]与 SocioVerse^[567]为代表的社会模拟系统，明确将“社会模拟”视为一种可复现、可干预、可评估的实验流程，而非仅用于展示突发行为的生成环境。这类仿真平台在工程上支持更高规模的代理并行运行、长时间交互、和人类行为对齐，使其具备模拟宏观社会现象的能力。SocioVerse^[567]明确提出环境、用户、交互机制与行为分布四类对齐维度，并通过引入大规模真实用户池来锚定代理的人群属性与行为分布，试图缓解长期困扰社会模拟研究的代表性与外推性问题。YuLan-Onesim^[568]则从“可用性”与“可复用性”角度推进平台演化，通过自然语言无代码建模、场景模板化与分布式扩展，将社会模拟从高门槛工程系统转化为社会科学研究者可直接使用的实验工具。

2. 可靠性与偏差

尽管仿真能力和人类行为对齐有显著提升，但仅依赖 LLM 生成“看似

表 5.9: 模拟仿真平台

工作	平台能力	关注点
AgentSociety ^[566]	万级代理、长时间运行、支持政策与社会冲击实验	以可复现社会实验为目标，对齐现实社会现象（极化、政策反馈等）
SocioVerse ^[567]	世界模型级社会仿真平台，连接千万级真实用户池	显式建模环境、用户、交互与行为分布对齐，强调人群代表性
YuLan-OneSim ^[568]	平台化社会模拟基础设施，无代码建模与十万级扩展	弱化具体对齐假设，强调工程通用性与快速场景构建

合理”的社会行为，本质上仍停留在行为拟态（behavioral mimicry）层面，难以支撑因果分析、机制解释与干预一致性验证^[569]。

LLM 基于语言模式学习，它生成的行为是否真的与人类等价，需要从实验方法上提高可靠性。

Simulating^[570]明确提出，社会模拟若要具备研究价值，必须显式建模代理的信念、目标、记忆与推理路径，而非仅观察最终行为输出，试图将社会模拟从“黑箱生成”转向“可审计的认知过程仿真”。

Promising^[571]从方法论层面对该范式进行定位，将 LLM 社会模拟视为一种介于传统 agent-based modeling 与实地社会实验之间的新型研究工具，同时系统性梳理主要挑战，包括：多样性，偏见，谄媚，异化和泛化性。

表 5.10: 可靠性与偏差

工作	认知建模假设	认知建模可靠性
Promising ^[571]	不提出具体模型，强调方法论反思	指出行为拟态的风险，呼吁因果与认知可解释性
Simulating ^[570]	显式建模信念、目标与推理过程	强调认知状态可追踪、干预一致性与因果解释

3. 评测与数据基准

2025 年的评价基准工作转向关注“我们究竟在模拟什么能力”这一问题，尝试构建与社会科学任务高度耦合的评测基准。

SPIN-Bench^[572]以能力分解为核心，其将社会智能拆解为战略规划、互

表 5.11: 评测与数据基准

工作	评测对象	评测特点与发现
SPIN-Bench ^[572]	战略规划与社会推理能力	揭示大模型在长程、多主体互动中的系统性不足
SocioBench ^[573]	跨国社会调查问卷中的态度分布	基于 ISSP 数据评估人群级价值与态度对齐偏差
SocialMaze ^[574]	大语言模型社交推理能力	揭示了大语言模型在处理动态交互和信息不确定性方面的局限性

动推理、协商与多方博弈等要素，并在结构化环境中系统分析 LLM 在不同推理深度与不确定性条件下的表现。评价结果表明，当前模型在短程规划与显式规则任务上具备一定优势，但在长程、多主体依赖的社会推理中仍存在显著退化。

SocioBench^[573]以跨国社会调查为参照，评估 LLM 在模拟不同人口属性个体对复杂社会议题的立场时，与真实人类分布之间的偏差。尽管 LLM 在语言层面高度流畅，但在价值观、政治态度与社会判断上的对齐仍高度不稳定，且误差在不同人群子群之间呈现系统性差异。

SocialMaze^[574]系统地整合了现实世界社交推理的三个关键特征：Deep Reasoning（深度推理），大模型超越表面信息进行复杂认知，如推断潜在精神状态和进行反事实思维；Dynamic Interaction（动态交互），大模型跟踪多轮交互中演变的上下文，并动态调整推理和行动；Information Uncertainty（信息不确定性），大模型批判性地评估信息来源可靠性，过滤误导性信号，并在信息不完整或冲突的情况下进行推理。

未来展望

大语言模型有望成为政策模拟和社会研究的标配工具。在辅助决策方面，决策者可以在虚拟社会中先行试验政策，在现实中避免代价高昂的失误；在社会科学方面，研究者能利用智能体探究传统方法难及的问题，拓展理论疆界。同时，随着模型能力的提升和大众化，普通公众甚至可以通过自然语言在模拟沙盘中探索“如果…会怎样”的问题，提升全民参与公共政策讨论的广度与理性基础。这将深刻影响政策制定的民主化和科学化进程。实现这一愿景的前提是以负责任的态度发展和使用这项技术：建立透明、审慎的评估机制，将人类的价值观和专业知识与大模型的强大计算力相结合，真正

发挥大语言模型在政策模拟与人类行为建模中的潜能，为应对 21 世纪复杂的社会挑战提供强有力的智囊支持。

5.1.6 心理咨询

心理健康已成为全球性的公共卫生挑战。受限于专业咨询师的人力短缺、高昂的治疗成本以及社会污名化影响，大量心理支持需求未能得到有效满足。在此背景下，利用人工智能提供低成本、高可及性的心理干预手段成为研究热点。早期的心理聊天机器人多基于人工预设的规则和模板，虽然开创了人机情感交互的先河，但缺乏对语义的深度理解，难以应对复杂的临床情境。随着深度学习与通用大模型架构的普及，开发者尝试利用大规模通用语料训练模型，使其具备识别用户情绪并给出温暖回复的能力，但由于高质量临床真实数据的匮乏和专业知识的壁垒，模型在模拟资深心理咨询师的决策逻辑上仍显稚嫩。

2025 年，心理咨询大模型正加速向具备临床深度的专业心理诊疗转型。**数据层面**，为突破隐私限制，研究重心转向构建模拟真实疗程的高质量合成语料；**技术层面**，通过融合专业心理疗法与智能体架构，显著提升了模型的长程记忆管理与临床推理能力；**评估层面**，则已从静态测试进化为模拟真实医患互动的动态基准。本节将围绕数据资源、技术进展及基准评估三个维度展开综述。

1. 心理咨询数据资源

在人工智能赋能心理健康的浪潮中，高质量数据是当前最为稀缺的资源，鉴于心理咨询数据的极高隐私敏感性，公开的高质量语料长期匮乏。这一领域正迎来关键的范式转移：从传统的静态、单轮问答数据，向模拟真实疗程的动态、长程咨询数据演进。为了突破数据瓶颈，前沿研究正致力于利用 LLM 的生成能力，深度融合专业的心理学理论与共情技术，开展大规模、高质量的合成数据构建，表 5.12 汇总了本年度心理支持数据集的核心侧重及其关键特性。

(1) 情绪支持对话数据集 为了提升对话系统在复杂心理场景下的共情能力与策略有效性，近期情绪支持对话研究在数据构建上进行了多元化的探索。针对现有模型在回应时往往缺乏明确动机的问题，IntentionESC^[575] 提出了

表 5.12: 心理支持数据集总结对比		
数据集	核心侧重/范式	关键特性与创新
IntentionESC ^[575]	支持意图驱动	提出 ICECOT 机制，通过建模支持者的心理意图引导模型实现精准的策略选择。
SSConv ^[576]	社会交互模拟	整合求助者画像库与支持者认知推理链，模拟更具针对性的社会化交互动态。
COCOON ^[577]	主动需求挖掘	遵循主动倾听原则，通过分析“情绪-感受-需求-记忆”链条识别用户的潜在需求。
Psy-Insight ^[578]	可解释多任务	首个双语咨询语料库，提供细粒度的多任务标签及回合级推理注释以增强逻辑理解。
PsyDial ^[579]	隐私数据重构	采用 RMRR 方法在彻底脱敏的前提下重构对话，保护隐私同时保留真实互动特征。
MusPsy ^[580]	多疗程长咨询	突破单次会话限制，通过跟踪心理状态随疗程的演变轨迹模拟长程咨询的动态调整。

以意图为中心的对话框架, 通过定义支持者的心理意图来精准指导策略选择, 弥补了情绪状态分析与策略选择之间的逻辑断层。面向合成数据中社会化交互特征缺失的挑战, SSConv^[576] 基于 SocialSim 框架, 一方面通过构建包含丰富人口统计学与性格特征的求助者画像库来模拟真实的社会自我披露, 另一方面通过显式的认知推理链增强支持者的社会感知能力, 从而生成了兼具深度与真实感的对话语料。此外, 考虑到用户在求助时往往表达含蓄或处于被动状态, COCOON^[577]引入了主动倾听 (Active Listening) 理论, 通过构建包含“情绪-感受-需求-记忆”的深层用户画像, 训练模型从用户隐晦的表达中挖掘潜在心理需求, 实现了从被动响应向主动情绪支持的跨越。

(2) 心理咨询对话数据集 为了弥合通用对话与专业心理咨询之间的差距, 数据集的构建正向着更具解释性、真实感与长程演化的方向发展。Psy-Insight^[578] 构建了首个面向心理健康的可解释性多轮双语数据集, 通过引入细粒度的多任务标签以及回合级的推理注释, 赋予模型类似人类咨询师的逻辑推理与共情能力。面对真实咨询数据因隐私壁垒难以共享的困境, PsyDial^[579] 提出了一种隐私保护数据重构框架 (RMRR), 在掩盖敏感信息的同时保留了真实医患互动的语言特征与治疗价值, 生成了兼具多样性与连贯性的大规模对话语料。进一步地, 考虑到心理咨询是一个动态发展的长期过

程, MusPsy^[580] 突破了现有数据集局限于单次会话的范式, 构建了多疗程咨询数据集。该工作利用真实案例报告中的客户侧写与阶段性目标, 模拟了客户心理状态随疗程推进的演变轨迹, 为模型学习长期记忆管理与动态策略调整提供了关键的数据支撑。

2. 心理咨询技术进展

随着大语言模型在心理健康领域的深入应用, 研究重心正从单一的共情回复生成, 向具备专业临床逻辑、长程记忆能力及自适应策略的智能系统转变。当前的技术突破主要集中在通过模拟专业诊疗流程以及利用智能体架构提升咨询的连贯性与策略性。

(1) 基于临床推理与诊疗范式的方法 为了解决现有 LLM 难以精准实施特定的心理治疗技术这一问题, 研究者尝试将专业的诊疗范式显式地融入模型的推理过程中。AutoCBT^[581] 针对认知行为疗法 (CBT) 的自动化实施, 设计了一种模拟现实中“咨询师-督导”协作模式的自主多智能体框架。该系统将 CBT 的核心原则具象化为不同的督导智能体, 通过动态路由机制让咨询师智能体在生成回复前征询督导意见, 从而实现了用户对用户认知偏差的精准识别与干预。进一步地, 为了赋予模型符合临床标准的决策能力, PsyLLM^[582] 提出了系统集成诊断推理与治疗推理的框架。该工作引入 DSM-5/ICD-11 等国际诊断标准以及 ACT、精神动力学等多种治疗流派作为推理支架。模型被训练在生成回复前, 先输出包含症状评估、诊断假设及疗法选择的显式思维链, 确保了咨询回复不仅有温度, 更具备临床医学的严谨性与解释性。

(2) 基于智能体与记忆管理的方法 心理咨询通常是一个跨越多次会话的长程动态过程, 这要求 AI 具备长期记忆能力以及根据疗程进展调整策略的能力。针对长程咨询中常见的临床遗忘与策略僵化问题, TheraMind^[583] 提出了一种双循环智能体架构。其“会话内循环”负责实时的情绪感知与战术响应; 而“跨会话循环”则负责在每次咨询结束后评估疗效, 并根据用户的反馈自适应地调整下一阶段的治疗方案, 从而实现了个性化长程咨询。此外, 为了解决模型回复模式化、缺乏针对性策略的问题, SweetieChat^[584] 引入了策略增强的角色扮演框架。该系统通过构建“求助者-策略顾问-支持者”的三角互动模型, 利用策略顾问根据对话历史动态选择最优的情绪支持技巧,

指导支持者生成回复，显著提升了模型在应对不同情绪场景时的策略适应性与效果。

3. 心理咨询基准评估

随着 LLM 在心理咨询领域的深入应用，评估基准正经历从传统的静态知识测试向更加符合临床实际的动态交互评估与高保真用户模拟演进，旨在考察模型在真实、复杂且多变的心理咨询场景中的综合胜任力。

在用户模拟的真实性方面，**AnnaAgent**^[585] 指出传统的静态画像难以捕捉咨询过程中来访者状态的动态演变。该系统引入了情绪调节器与主诉引导机制，模拟了来访者在单次会话内的情绪波动及认知变化，并结合三级记忆机制实现了跨疗程的记忆连贯性，从而为模型提供了更具挑战性和逼真度的交互环境。针对现有评估视角单一且缺乏反馈机制的问题，**Ψ-Arena**^[586] 提出了一个交互式评估框架，通过模拟包含建立信任、诊断及方案探索的多阶段真实咨询流程，引入了涵盖来访者体验、督导专业性及咨询师自省的三方评估机制，并通过闭环反馈迭代优化模型的咨询能力。为了在开放式问答中保障临床安全性与专业度，**CounselBench**^[587] 联合心理专家构建了大规模评估基准。该工作不仅提供了详尽的专家注释与评分理由，还专门构建了对抗性数据集以诱发模型在提供未经授权的医疗建议或过度泛化等方面的潜在失效模式，确立了兼顾质量与安全性的临床评估标准。为进一步解决现有模拟器角色单一且缺乏专家准则指导的问题，**CARE-Bench**^[588] 提出了一个基于专家原则指导的多样化来访者模拟基准。该基准基于数千个真实咨询案例构建了涵盖不同人口统计学特征、人格特质及咨询主题的多样化来访者画像库。不同于简单的角色扮演，CARE-Bench 引入了由专业心理咨询师制定的行为准则来约束模拟来访者的交互逻辑，确保其行为更贴近真实求助者。此外，该基准采用多维度专业心理学量表，实现了对大模型在治疗关系建立、共情深度及专业技巧运用上的全面且动态的评估。

未来展望

综上所述，2025 年基于 LLM 的心理咨询研究在文本单一模态层面已呈现出多点开花、体系逐步成型的态势：从高质量合成数据构建，到融合临床推理与智能体架构的咨询系统，再到高保真动态评估基准，相关工作不断逼近真实心理咨询的专业流程与实践需求。然而，需要清醒地看到，当前研究

仍主要集中于单模态文本交互，多模态心理支持回复生成（融合语言、语音、面部表情、生理信号等）的系统性探索尚处于起步阶段。

未来研究有望进一步迈向**多模态心理支持与具身化共情陪伴**方向，使模型不仅能感知、表达和调节情绪状态，还能够提供更触手可及、更具陪伴感的具身化心理支持。与此同时，基于 LLM 的心理咨询技术的实际临床与社会落地在现阶段仍面临显著挑战，其广泛应用亟需**更加完善的伦理审查与安全规范、更具临床一致性的评估基准**，以及来自政策与监管层面的制度性支持。如何在技术创新、专业可信度与社会责任之间取得平衡，将成为下一阶段心理咨询大模型研究与应用的核心议题。

5.1.7 深度调研：Deep Research

随着大语言模型在语义理解、多步推理、工具调用与长上下文建模等能力上的持续提升，信息获取系统正从传统的“检索—展示”范式，演进为以任务驱动和证据整合为核心的 Deep Research 技术体系。2024-2025 年间，OpenAI^[589]、Google^[590]、Perplexity^[591] 等公司陆续推出商业化的 Deep Research 系统，显著提升了公众对“研究型 AI”的认知，也推动了学术界对开放式 Deep Research Agents 的系统性探索。该体系不再仅关注单次问答的正确性，而是围绕复杂研究目标，支持多阶段规划、多轮证据获取、持续状态维护与结构化报告生成，逐步呈现出研究代理（Research Agent）的系统形态。

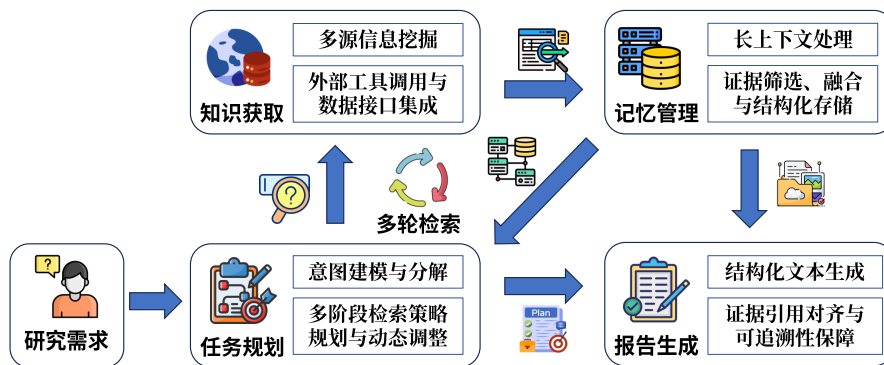


图 5.2: Deep Research 系统框架

系统架构

从系统架构角度看，Deep Research 通常包含以下关键模块，并形成闭环运行机制：

首先，**查询意图建模与任务规划**是 Deep Research 系统的起点。依托大语言模型对自然语言的深层语义理解能力，系统能够对用户研究查询进行意图解析与问题分解，将复杂需求拆解为若干子问题，并据此规划多阶段检索策略与信息获取路径，强调对研究目标、信息范围与约束条件的整体把握，为后续的深度调研奠定基础。

其次，**证据驱动的多轮检索与外部知识获取**使系统摆脱了对静态关键词匹配的依赖，逐步演化为具备自主行动能力的检索智能体。Deep Research 系统在结合语义向量检索与关键词检索策略的基础上，引入面向 Web 环境的智能化检索机制，能够自主调用搜索引擎、学术数据库及各类外部工具接口，执行多轮、递进式的信息获取与证据收集过程。

在这一过程中，系统不仅被动接收检索结果，而是对已获取的证据进行持续评估，根据当前研究记忆状态动态调整检索目标与行动策略，在证据不足、信息冲突或结论不稳定时主动触发进一步查询与验证步骤。这种以证据充分性与一致性为驱动的迭代检索范式，模拟了人类研究者在复杂 Web 环境中反复查证、交叉验证的调研行为，显著提升了信息获取的覆盖范围、相关性与可靠性。

研究记忆管理与长上下文建模 为多轮调研提供了关键支撑。随着主流大模型上下文窗口扩展至 128K 甚至更高，Deep Research 系统能够对来自不同轮次检索的大规模文本进行统一建模，通过证据筛选、信息融合与结构化存储，形成稳定的研究记忆状态，支持跨轮次推理与中间结论的持续积累。

结构化研究报告生成与引用对齐机制 是 Deep Research 的最终输出形式。不同于传统搜索引擎返回的链接列表，系统将整理后的研究记忆转化为结构清晰的长文本报告，以段落、要点与层级结构呈现分析结果，并显式对齐生成内容与其证据来源，保障结论的可验证性与可追溯性。

总体而言，Deep Research 不再仅仅返回零散信息，而是围绕用户研究目标，完成从查询理解、任务规划、多轮证据检索、研究记忆管理到结构化报告生成的完整闭环，其本质体现了一种融合检索、推理与生成能力的任务型智能体范式。

学术研究进展

2025 年, 学术界围绕 Deep Research 的研究逐步从系统原型探索, 转向对**长程研究能力的可训练性、可评测性与可复现性**的系统化刻画。作为一种**长时间跨度、多轮证据获取、跨状态推理与结构化输出相结合的智能体**, Deep Research 的挑战不仅在于“能否找到信息”, 更在于“是否能够持续规划、补全证据并形成可信研究结论”。

在系统层面, 2025 年也出现了更接近“**深度研究端到端智能体**”的开放框架: Alibaba-NLP/DeepResearch 项目^[592] 综合探索多搜索后端协同、长时检索轨迹建模以及研究型报告生成等关键问题, 推动检索能力从“搜索增强生成”向“研究流程驱动生成”的整体演进。

首先, 在 **Web 行动能力** 方面, 2025 年 Deep Research Agents 的研究焦点已从“静态文档检索/摘要阅读”明显扩展到真实 Web 环境中的交互式导航、跨页面遍历与长链路证据汇聚。其任务设定开始刻意强调“**深入到子页面并跨链接整合**”的信息获取过程: 例如 WebWalker^[593] 将问题构造成为需要访问网站多级子页面、沿链接结构逐步搜集线索的任务, 从而系统性揭示仅依赖搜索结果摘要在覆盖性与可验证性上的固有局限。进一步地, WebExplorer^[594] 等工作把“Web 行动”本身作为训练对象, 强调在真实网页环境中以探索—演化的数据生成与强化学习方式学习更长视距的浏览与信息整合能力, 从而把 deep research 从“检索问答”推进到“可执行的 Web 研究过程建模”。

其次, 围绕“**何时检索、如何检索、检索后如何推理整合**”的策略学习逐步从启发式提示走向可优化的端到端训练范式。Search-R1^[595] 代表了“搜索驱动的 RL”路线: 通过 GRPO 等在线优化, 让模型学会在推理过程中触发搜索、吸收检索证据并完成回答; 但其核心训练与评测仍主要面向可验证的短形式任务, 反映出当时开放研究对“可监督/可奖励信号”的强依赖。随后, Search-ol^[596] 将检索更紧密地嵌入长链推理过程, 通过更具代理性的搜索工作流与面向文档证据的深度推理模块 (如对检索文档进行再推理与筛噪), 缓解长推理中“知识不足—反复不确定—误差累积”的问题, 使检索不再只是外置补丁, 而成为推理链中的可学习环节。

同时, **长时序工具调用**成为训练可行性的关键突破点: Beyond Ten Turns^[597] 通过大规模异步强化学习, 探索在 40 次以上工具调用的长链路交互中学习稳定的检索与决策策略, 为 Deep Research 场景下“多轮、递进式证据获取”的训练提供了工程与算法层面的路径。

在 **评测与基准构建** 方面, 2025 年的研究逐步表明, 传统以问答准确率为核心的评估范式已难以刻画 Deep Research Agents 在真实研究场景中的能力边界。一个核心原因在于, 深度研究任务往往并非“答案是否正确”即可充分衡量, 而更依赖模型是否完成了充分的信息探索、证据收集与系统性综合。

围绕这一问题, BrowseComp^[598] 通过设计“答案本身可验证、但关键信息极难获取”的问题, 刻意将评测重点从答案匹配转向持续 Web 浏览能力与关键信息定位能力, 从而系统性揭示了仅依赖检索摘要或单次搜索结果的评测方式, 在深度研究场景下的明显不足。这类任务为 Deep Research 评测提供了以“研究过程有效性”为导向的直接证据。

在此基础上, 部分工作开始从 **评测基础设施** 层面反思 Deep Research 系统的可比较性与可复现性。DeepResearchGym^[599] 通过受控的检索接口与统一的评测协议, 显式减少商业搜索引擎波动对实验结论的干扰, 使模型评测更聚焦于其研究与推理能力本身, 而非外部环境噪声。这一方向为大规模、长期 Deep Research 评测提供了必要的实验条件。

进一步地, 研究者开始探索 **研究质量应如何被系统性刻画**。DeepResearch Bench^[600] 将评测任务扩展至覆盖 22 个领域的 PhD 级研究问题, 尝试以自动化指标在一定程度上逼近人类对整体研究质量、深度与结构性的判断。与之相呼应, DeepSearchQA^[601] 明确提出以覆盖度 (*comprehensiveness*) 为核心评估维度, 系统性分析多步检索与证据整合过程中暴露的覆盖差距, 推动评测目标从“是否答对”进一步转向“是否完成了充分研究”。

在上述评测趋势之上, **证据对齐与可核验性** 正从评测中的辅助指标, 演变为 Deep Research 的关键能力约束。DR Tulu^[602] 进一步将这一评测理念引入训练阶段, 提出基于演化评分标准的强化学习范式, 通过与策略共同演化的评价规则, 应对 Deep Research 中目标模糊、评价标准随研究进程变化的问题, 为长文本研究、证据充分性与引用一致性提供更稳定、具区分度的学习信号。这标志着研究重心开始从“短回合、可验证答案的奖励建模”, 转向对“长程研究过程质量”的整体优化。

总体而言, 2025 年的学术研究表明, Deep Research 正逐渐被视为一种 **以任务规划为先导、以证据驱动多轮检索为核心、以研究记忆为支撑、以结构化报告为目标** 的通用智能体范式。围绕这一范式, 训练方法正在从静态奖励走向动态评价, 评测体系从单点正确性走向过程与覆盖度, 而系统设计则日益强调在真实 Web 环境中的可执行性与可复现性。这些趋势也为商用 Deep Research 系统在可信性、可扩展性与科学研究辅助等方向的进一步发

展奠定了方法论基础。

国内外商用产品

在产业界，Deep Research 已成为 2025 年各大厂商重点布局的方向，其产品形态虽各有侧重，但整体上均围绕上述系统架构展开。

在国际产品中，OpenAI 的 ChatGPT Deep Research^[589] 强调复杂问题分解与长报告生成能力，并逐步引入代理式浏览与项目级上下文管理；Google Gemini Deep Research^[590] 则突出显式研究计划与可编辑的执行业务流程，并在 2025 年推进研究代理能力的 API 化；Anthropic 的 Claude Research^[603] 采用多代理协作模式，强调研究过程中的并行探索与企业级集成；Perplexity Deep Research^[591] 以高频多轮搜索和来源可追溯性建立差异化优势；Microsoft Copilot 与 Bing Chat 则依托企业生态，将 Deep Research 能力嵌入办公与组织知识管理场景。

在国内，百度文心 X1^[74]、阿里通义 Qwen-DeepResearch^[592]、字节跳动豆包^[604]、智谱 AutoGLM 沉思^[605]以及月之暗面 Kimi-Researcher^[606] 等产品，均在 2025 年不同程度引入了多步研究、工具调用与长文生成能力。相较于国际产品，国内系统更强调中文 Web 环境适配、垂直场景落地以及既有业务生态的深度融合，但在总体技术路线与研究代理理念上已呈现出明显趋同。

未来展望

展望未来，Deep Research 技术的发展仍面临若干关键挑战。

首先，在方法层面，如何为开放式、长程研究任务设计稳定、可扩展的训练目标与评价机制，仍是学术界与产业界的共同难题。其次，在系统层面，研究记忆的组织形式、证据冲突的处理策略以及跨轮次推理的稳定性，有待进一步探索。再次，在应用层面，如何在提升研究效率的同时控制幻觉风险、增强结论可信度，并满足合规与隐私要求，是 Deep Research 走向大规模部署的关键。

总体来看，Deep Research 不仅代表着搜索与生成技术的升级，更体现了人工智能系统从“信息提供者”向“研究协作者”的角色转变。随着模型能力、智能体框架与评测体系的持续完善，Deep Research 有望成为未来知识工作的重要基础形态，并在科研、教育、决策支持等领域发挥越来越核心的作用。

5.1.8 AI for Research

大语言模型能力的快速跃迁，正推动人工智能从传统的科研辅助工具，演进为能够参与科研全流程的新型智能主体，“AI for Research”由此成为连接人工智能技术突破与科学创新效率提升的关键方向。自 2025 年以来，一批具备端到端科研支持能力的代表性系统相继出现，科研智能体在自主性与流程完整性方面取得了实质性进展，并开始对既有科研范式、研究组织方式与创新机制产生潜在的重塑影响。在此背景下，本章将围绕 **AI for Research 的发展趋势与研究方向** 进行介绍，以期为理解大模型时代科学研究模式的演变及其战略意义提供参考。

发展趋势

科研智能体能力的跃迁并非偶然，随着模型上下文窗口的扩展，模型推理能力的提升以及多智能体协作框架的成熟，如图 5.3 所示，科研 AI 的角色演进呈现出了层级跃迁特征^[607]。

早期的 AI 主要作为科研工具，其核心特征是单任务、无长期记忆且缺乏主观能动性。典型应用包括代码补全工具（如早期 GitHub Copilot）、文献翻译与摘要系等。此类系统的交互模式完全由人类指令触发，研究问题的提出、方法的选择以及结论的判断均由人类研究者主导，AI 仅作为提升效率的工具存在。随着 2023 年 ChatGPT 出现，大模型时代到来，科研 AI 的能力边界开始显著拓展，此阶段的系统不仅能够完成文献检索任务，还可以进行跨文档比较、结构化信息抽取（例如样本量等统计要素），并对相互矛盾的研究结论进行初步的加权分析。然而，在这一阶段，研究目标的设定、评价标准的选择以及最终科学判断仍高度依赖人类研究者，AI 更多体现在辅助分析而非自主决策。

进入 2025 年，前沿研究开始推动科研 AI 向“科学家”阶段跃迁。Sakana AI 提出的 The AI Scientist 以及 Google 的 AI Co-scientist 等系统，标志着科研智能体在自主性层面取得了关键突破。这类系统不再仅被动响应人类提出的问题，而是能够主动识别知识缺口并提出科学问题，形成覆盖完整科研生命周期的能力闭环，包括文献阅读与综述、假设生成、实验设计与实现、实验执行与调试、结果分析、论文撰写以及自我评审等环节。更为重要的是，部分系统已开始展现初步的元认知能力，能够反思自身提出假设的新颖性，并在实验失败时区分问题源于理论假设缺陷还是实现层面的技术错误，从而为持续迭代与自主科研提供可能。

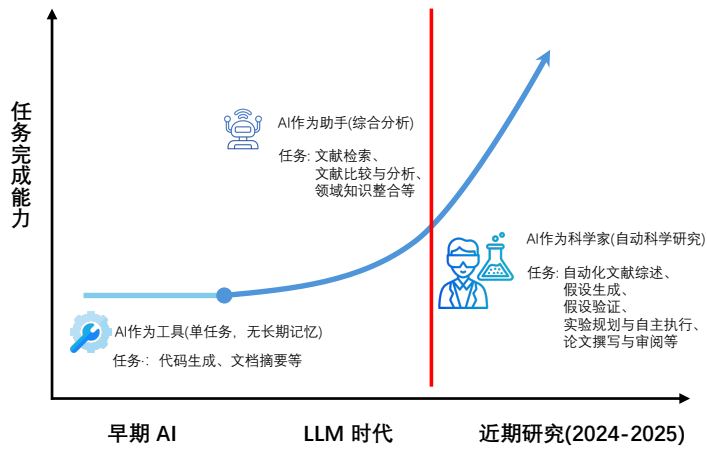


图 5.3: 科研智能体发展趋势

研究方向

当前研究的核心目标在于探索 AI 在科学创新中的主体性角色。模型不仅需要具备跨学科知识整合与推理能力，更被期望贯通文献综述、假设生成、实验设计与执行、论文撰写乃至同行评审等完整科研链条^[608-609]，如图 5.4 所示，这标志着 AI 正演进为具备“准科研代理”特征的新型系统。

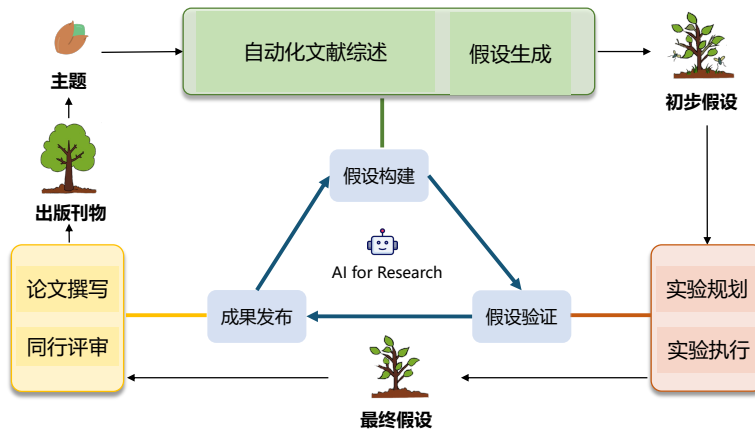


图 5.4: AI 参与完整的科研链条

自动化文献综述 科研活动的起点在于明确既有知识的边界。然而，在信息爆炸的背景下，传统文献调研方式正成为制约科研效率的重要瓶颈。基于大语言模型的自动化文献综述技术，已从简单摘要生成，发展为构建结构化、

可验证、可扩展的知识体系。

Tang et al.^[610] 提出了一套自动化评估框架，对 LLM 生成的参考文献、摘要与正文进行真实性与覆盖度评估。AutoSurvey^[611] 采用“检索—提纲—生成”的两阶段流程，而 SurveyX^[612] 进一步细化了检索、信息抽取与结构生成模块，支持多模态输出并直接生成完整综述论文。Zhu et al.^[613] 则通过多视角聚合与层次化聚类，缓解了文献爆炸带来的知识组织难题。

假设生成 假设生成是科研流程中最具创造性、同时也最难以形式化的环节。随着模型推理与知识整合能力的提升，AI 在该阶段展现出可规模化的组合式创新潜力，其优势并非直觉灵感，而在于对庞大科学知识空间的系统遍历与受约束搜索。

2023 年发布的 SciMON^[614] 将假设生成视为了一种开放式生成任务，激发了研究者使用 LLM 进行假设生成的热情。2025 年初，Google DeepMind 提出的 AI Co-scientist^[615] 则构建了“生成—约束—迭代”的推理范式：模型先生成候选假设集合，再通过多代理从逻辑一致性、领域合理性与技术可行性等维度进行高强度审视，使假设在早期即实现收敛与筛选。近期研究进一步将假设生成与文献检索和知识校验机制深度融合^[616]。

实验规划与自主执行 在假设提出之后，实验验证构成了 AI 从软性推理走向硬性证据的关键门槛。新一代系统开始承担实验流程的规划、执行与修正，在计算实验与物理实验两端同时推进自动化。

Sakana AI 的 *The AI Scientist* 系列工作^[617-618] 展示了基于代理搜索与决策的实验规划能力。相较于早期模板驱动的方法，这一范式将实验过程建模为策略空间中的探索问题。与此同时，上海人工智能实验室提出的科学智能上下文协议（SCP）尝试为模型、工具与研究对象提供统一的协同接口，以“科学发现的 Scaling Law”为核心理念，构建覆盖实验全流程的智能协同框架。

论文撰写 论文撰写指利用人工智能技术辅助研究人员完成科学手稿的构思、撰写、修改与格式化，其目标不仅在于提升写作效率，更在于将复杂、异构的科研产出转化为符合学术规范、可被同行理解与复现的文本表达。根据生成粒度的不同，相关研究可分为句子级、段落级与全文级三类。

句子级研究主要关注引文句的自动生成与规范化，通过显式建模引用意图以降低事实性偏差^[619]。段落级方法多用于相关工作撰写，强调对多篇文

献关系的综合建模，从而提升论文的学术定位清晰度^[620]。在全文级层面，AI 通过分章节生成与迭代修订的方式，参与摘要、引言、方法与实验等核心模块的整体组织，其目标在于加速从“实验记录”到“可交流论文”的转化^[617]。

同行评审 同行评审是保障科学质量与学术信用的核心机制。面对科研产出激增与评审人力不足的结构性矛盾，AI 开始被引入评审流程，承担逻辑一致性验证、实验充分性检查与创新性分析等任务。

高质量评审的本质并非复述论文内容，而是阅读论文，并识别其中的逻辑断裂、与方法论缺陷并给出好的修改意见。而早期 AI 评审能力却只能做到对论文的总结，生成的审稿意见常常存在无法对论文进行高层次的理解以及生成不实内容。以 OpenAI o1 与 DeepSeek-R1 为代表的新一代模型，通过强化学习与显式推理机制，使 AI 评审能力实现跃迁。斯坦福团队提出的 Agentic Reviewer 系统，在评审评分上与人类评审者的 Spearman 相关系数达到了 0.42，显示出 AI 在结构化评审中的潜在价值。

未来展望

2025 年可以被视为 AI for Research 从“任务驱动的计算工具”迈向“全流程自主科研代理”的关键转折点。这一跨越并非仅体现在模型规模或性能指标的量变上，更深刻地反映了科学研究范式在结构层面所经历的系统性重构。

在方法论层面，科研活动被建模为一组可组合、可迭代且具备反思能力的认知与操作模块。在此框架下，科学问题的提出、假设生成、实验设计、结果分析与理论修正等环节不再是孤立的任务，而是被统一纳入一个闭环的推理—行动系统之中。与此同时，多智能体架构通过引入角色分工、协作与博弈机制，在一定程度上模拟了真实科研组织中的集体认知过程；而基础模型在推理、规划与工具调用能力上的显著提升，则为各个子步骤的高精度执行提供了必要条件。

在应用层面，一系列标志性系统的出现表明科研智能体在自主性与系统完整性方面已取得实质性突破。Sakana AI 提出的 The AI Scientist、Google 的 AI Co-scientist 等工作，展示了 AI 在假设生成、实验迭代乃至论文撰写等核心科研环节中的端到端参与能力，其产出的研究成果通过顶级学术会议的同行评审，也从侧面印证了 AI 科研系统正逐步走向现实可用阶段。进一步地，围绕科学研究的全流程支持，被 Nature 报道的由哈佛大学与麻省理工学院联合推出的 ToolUniverse，上海人工智能实验室提出的 Intern-Discovery

科学发现平台，以及斯坦福大学探索的自动化论文审稿系统，均从不同切入点为科学发现、知识验证与学术评估等关键任务提供了具有“变革性意义”的新型工具基础设施。

展望未来，AI 驱动的科学研究的并不意味着对人类科学家的替代，而更可能催生一种全新的科研形态。在这一形态中，科学方法本身成为可以被显式建模、系统优化并持续扩展的研究对象。在人机协同的科研新范式下，人类研究者的核心竞争力将逐渐从具体执行与计算能力，转向问题定义、价值判断以及跨学科、跨范式整合的能力。如何在显著提升科学探索效率的同时，持续维护研究体系的多样性、创造性与可解释性，将成为 AI for Research 迈向成熟阶段所必须正视的长期命题。

5.2 行业应用

5.2.1 教育行业

随着大语言模型在语言理解、推理规划与生成控制能力上的持续突破，其在教育领域的应用正在从“工具辅助”阶段迈向“系统级重构”阶段。尤其在 2025 年前后，研究重心开始从单点能力展示（如自动批改、答疑）转向覆盖教学设计—学习支持—测评调控—学习诊断的闭环式智能教育体系。这一趋势促使研究者重新审视教育活动中教师、学生、学习过程与评测机制之间的结构关系，并探索如何利用 LLM 构建以学习者为中心、可解释、可调控的智能教育系统。

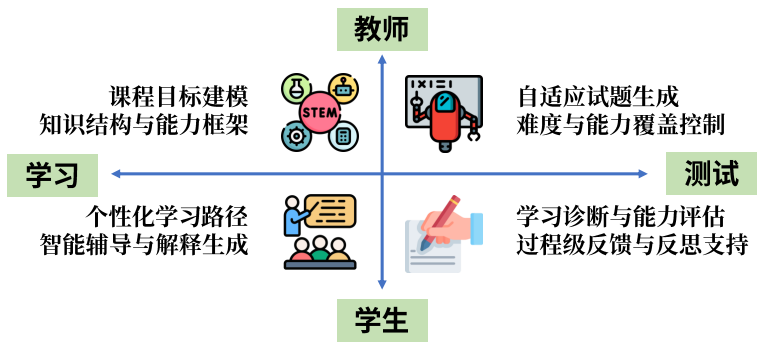


图 5.5: LLMs 教育领域应用分类

基于这一视角，当前主流研究可总结为一种“四象限”结构化范式：在

教师-学习侧，LLMs 被用于课程目标建模、知识结构组织与能力框架对齐；在学生-学习侧，模型承担个性化学习路径规划、对话式辅导与解释生成；在教师-测试侧，研究聚焦自适应试题生成、难度控制与能力覆盖建模；在学生-测试侧，则强调学习诊断、能力评估与过程级反馈支持。这四个维度并非孤立存在，而是通过 LLM 的统一语义表示与推理能力实现联动，共同构成新一代智能教育闭环。

课程目标与能力框架建模

在教学设计阶段，2025 年的代表性工作普遍将核心目标从“生成教学文本”提升为“构建可对齐、可复用的课程结构与能力框架”。例如，Faraji et al.^[621] 提出以人机协同为中心的 LLM 辅助课程开发界面，通过在界面层引入结构化约束与交互式编辑机制，把课程开发从提示工程依赖转化为可操作的结构化工作流，并在用户研究中显示该设计能够有效降低课程开发的认知负担并提升课程结构与教学目标的一致性。与之互补，Huovinen et al.^[622] 构建了一个面向真实高校环境的 LLM 课程写作系统，将课程与机构目标、认证要求及能力框架的对齐作为系统目标，通过长期部署验证了该系统可以在保留教师主导权的同时，稳定提供对齐建议并支持迭代式课程完善。围绕“知识结构显式化”，LessonPlanLM 框架^[623]通过构建教案知识库并采用检索增强生成（RAG）将外部教学知识注入生成过程，从而在教案结构完整性、知识准确性与逻辑连贯性上显著优于直接生成方案，并提供了可复现的数据与代码。进一步地，Hu et al.^[624] 引入“教学模拟—反思—再生成”的迭代机制，让模型先模拟课堂互动并产出反思线索，再据此优化教案内容；实验表明该流程能够系统性提升教案质量，使其在多维指标上接近或达到经验教师水平。

面向能力框架对齐任务，EduPlanner 系统^[625]构建了一个基于多智能体的教学设计框架，通过引入技能树（skill tree）对学生能力进行显式建模，再据此生成与能力状态对齐的教学方案。该工作不仅展示了 LLM 在课程规划中的生成能力，更重要的是将“能力结构”作为中介变量纳入教学决策过程，并在真实教学设计任务中验证了其有效性。

个性化学习路径与智能辅导机制

在学生-学习层面，2025 年的工作更倾向于把 LLM 定义为“可规划、可解释的学习代理”，而非被动答疑器。面向学习管理系统（LMS）场景，Yang

et al.^[626] 将 LLM 融合进自适应 LMS 的决策链路，通过对学习目标、学习轨迹与即时反馈的联合建模实现动态资源选择与路径调整，从而弥补传统 LMS 难以进行实时个体化调控的局限。面向内容侧与交互侧协同，一类系统型平台（如 LearnMate^[627]）将“学习计划生成—阶段目标拆解—对话式辅导—解释输出”作为统一工作流，用可执行步骤替代泛化建议，以提升学习者自主推进的可操作性。多模态学习支持方面，Li et al.^[628] 提出面向编程教学视频的生成式辅导架构，通过从视频内容生成引导性问题与提示，将 LLM 的对话能力嵌入到非文本学习材料中，进而支持更细粒度的理解监测与个性化引导。

同时，学习支持研究开始强调“过程质量”的评估与干预，而不仅是最终正确率。针对数学辅导任务，Gupta et al.^[629] 系统评估了 LLM 在步骤解释、提示策略与错误反馈上的能力边界，显示先进模型具备较强的分步讲解能力，但仍存在推理不稳定与幻觉风险，从而推动后续工作将“过程可控生成”“错误定位”“反馈策略学习”等问题显式化。

自适应评测与能力覆盖控制

在评测环节，2025 年的研究显著强调评测的自适应性与能力指向性，多项工作利用 LLM 自动生成评测题目，以缓解传统出题成本高、覆盖有限的问题。Savaal 系统^[630] 通过“概念抽取—概念驱动出题—专家评测校准”的流水线，从长文档中生成更高认知层次的问题，并在人类评估中显示其问题深度显著优于逐题提示生成策略，尤其在材料长度增加时优势更明显。课程对齐与覆盖控制方面，Wahid et al.^[631] 比较了直接提示与多种检索增强生成方案，表明引入课程文件检索能够显著提升题目与课程标准的一致性与事实可靠性，从而为“课程目标—试题—能力点”的闭环打下基础。

在难度建模方面，Zotos et al.^[632] 提出利用 LLM 作答不确定性特征（如概率分布与置信信号）预测题目难度，并在真实考试数据上验证不确定性与学生正确率之间存在稳定关联，从而提供了无需真实学生大规模作答即可进行难度预估的路径。进一步地，Zhu et al.^[633] 提出从模型隐藏状态估计难度的机制，在生成早期即可预测问题困难度，并可用于动态选择推理策略与计算预算。面向教育测量理论对齐，Scarlato et al.^[634] 提出 SMART 方法，通过偏好优化让 LLM 模拟不同能力水平的“虚拟考生”，再与项目反应理论（IRT）参数拟合对齐，从而实现新试题的低成本难度标定并提升与真实 IRT 难度参数的一致性。

学习诊断、能力评估与过程级反馈

对于学生能力的评价，2025 年的研究强调用过程数据实现更可解释的诊断与反馈。Yang et al.^[635] 提出将 LLM 提取的题目难度特征显式引入知识追踪模型，通过同时建模文本理解难度与知识点难度提升对学生知识状态演化的预测能力，尤其在编程等复杂技能任务中表现突出。围绕对话与步骤数据，一类工作将 LLM 作为诊断器对学习过程进行结构化标注与归因，从而支持“错误类型识别—针对性提示—反思引导”的闭环式反馈；对应地，Scarlato et al.^[636] 将 LLM 应用于智能导师-学生对话日志的知识追踪，设计 LLM-KT 方法标注对话中学生知识状态并预测答题正确率，显著优于传统知识追踪模型。在中文实践中，ECNU-ICALK^[637] 进一步将认知诊断与教学推理链结合，使模型能够围绕学生解题过程给出可解释的纠错与后续学习建议，并支持与教育工具协作，从而推动“答案级反馈”向“能力与思维过程支持”迁移。

华东师大推出的 EduChat-R1^[637] 大模型依托内置的教学思维链，模型在交互中准确发现学生回答中的错误并给出纠正，同时结合学科知识体系对学生解题过程进行系统性分析，“三思而后行”地指出学生的认识误区。例如，在学生作答后，EduChat-R1 会追踪其知识状态变化并提供个性化的后续练习或讲解建议。这些特性标志着智能教辅从单纯信息提供向关注思维过程和能力培养的转变。同时，EduChat-R1 集成的心理疏导模块还能给予情绪支持与反思引导，帮助学生调适心态并养成自主反思的学习习惯。

国内外教育大模型产品与平台

随着大语言模型技术的成熟，教育领域已成为其率先实现规模化落地的重要应用场景之一。自 2023 年以来，海内外陆续涌现出一批基于大模型的教育产品，覆盖学习辅导、作业批改、教学支持与教育管理等多个环节，并在真实用户群体中得到广泛应用。

表 5.13 对当前具有代表性的教育大模型产品进行了系统梳理，展示了其在应用场景覆盖、功能设计与技术实现上的共性模式与差异特征。

表 5.13: 国内外教育大模型产品对比

产品名称	应用场景	功能亮点	技术特点
Khanmigo ^[638]	K-12 辅导、 教师备课	苏格拉底式引导、 过程性提示；教师 端备课/练习生成	基于 LLM 的对话式 tutoring；强调安全 护栏与课堂语境约束
Duolingo Max ^[639]	语言学习 (个体化)	Roleplay 对话练 习；Explain My Answer 错因解释	LLM 驱动的对话与 解释生成；与学习路 径/题库系统深度耦 合
Quizlet Q-Chat ^[640]	作业答疑、 对话练习	对话式问答与练 习生成；围绕学习 卡/题库内容互动	LLM + 结构化学习 内容库；对话编排与 安全策略
Photomath ^[641]	数学拍照 题解析	拍照识题、分步解 答、步骤讲解	OCR/视觉识别 + 数 学符号解析；求解 器/推理步骤生成 (可与 LLM 结合增强 解释)
CheggMate ^[642]	高等教育、 作业辅导	个性化学习路径； 练习与解释；学 习陪伴式交互	LLM（公开信息含 GPT-4 方向）+ Chegg 内容与学习工 作流；个性化推荐
Coursera Coach ^[643]	在线课程 学习辅助	个性化问题解答、 概念澄清、课程笔 记与总结	LLM + 课程内容索 引/检索；对话式学 习助手嵌入平台
Udemy AI 助 手 ^[644]	在线技能 学习	AI 聊天辅助、技 能映射建议、课程 问答	LLM + 课程内容上 下文；平台级嵌入式 助手

(续下页)

产品名称	应用场景	功能亮点	技术特点
讯飞星火（教育） ^[645]	中小学课堂、作业辅导、教研	课堂辅导、作业批改与讲解、学情分析；强调“教-学-评”闭环	自研大模型 + 教育数据/评测体系；推理增强（X1 系列）与端侧/平台化部署能力
豆包爱学 ^[604]	中小学课后辅导	拍题答疑、口算/作文批改、错题归因与诊断	多模态（图像识别 + LLM）+ 学习诊断链路；面向规模用户的产品化与分发
文心大模型（教育） ^[74]	教材/考试、教学支持	教材与考试场景优化；虚拟教师与多角色互动	多模态与推理能力增强（4.5/X1）；平台化 API 与应用生态
人民网“自在”心理疏导大模型 ^[646]	校园心理健康	情绪支持、共情对话、心理科普与辅导	心理/对话安全策略 + 情绪识别/对话管理；教育平台集成与服务化
腾讯青少年大模型（未成年人模式） ^[647]	适龄化学习与内容安全	未成年人模式、内容安全与防沉迷；生态覆盖广	适龄分级/安全策略 + 平台生态集成（微信/QQ 等）
猿辅导 ^[648]	K-12 课后辅导、自适应学习	拍照识题、AI 讲题与答疑、个性化练习推荐、学习诊断与错题巩固	大模型 + 自适应学习/知识图谱体系；题库检索与讲解生成协同；面向规模用户的系统化产品部署
松鼠 AI ^[649]	K-12 自适应学习、个性化教学	知识点诊断、因材施教路径规划、个性化练习与讲解	智能自适应学习系统 (IALS)；学习画像建模与动态难度控制；诊断-干预闭环优化
ChatGLM（Edu/高校应用方案） ^[650]	高校课程、校内知识库	校内问答、课程助手、知识库检索与智能体	国产开源/可定制大模型路线；RAG/微调与本地化部署友好

小结

总体来看，2025 年的大语言模型教育应用研究正在从单点能力展示走向系统化闭环设计：课程目标与知识结构的显式建模为教学与评测提供统一约束，自适应试题生成与能力覆盖控制支撑精准评测，个性化学习路径与智能辅导提升学习过程质量，而学习诊断与过程级反馈则反向驱动教学目标与策略的持续优化。这一以学习者为中心的教育智能闭环，正在成为学术研究与产业实践共同探索的核心框架，也为未来构建可信、可控且可扩展的教育智能体系统奠定了基础。

5.2.2 医疗行业

2025 年前的医疗 LLM 主要通过监督微调和知识微调等方式，注入医学领域的专有知识和大规模文献，以提升医学问答能力。自 2025 年以来，医学 LLM 逐步引入持续预训练方法，进一步优化基础模型，增强医学专业知识，提升模型在临床任务中的泛化能力，并展现出了良好的应用效果。随着强化学习和推理模型的迅速发展，医疗推理能力得到了显著提升，推动了医疗 LLM 在更复杂任务中的应用。同时，随着多模态任务的引入，LLM 在医学图像分析和诊断中的应用得到了进一步拓展，为医学推理提供了更深层次的支持。这些进展不仅推动了医学知识的智能化与数字化，也促进了医疗服务的智能化升级。此外，随着医疗场景对智能诊断与咨询系统的需求增加，医疗 Agent 技术逐渐崭露头角，通过构建智能化系统，提升了医疗服务的效率和精准度，辅助诊断的 AI 系统也正在成为临床实践的重要工具。随着技术的不断进步，医疗大语言模型将在未来的医疗实践中发挥更加重要的作用，助力提升全球医疗服务的质量和效率。

1. 医疗知识适配 随着通用大语言模型能力持续跃升，医疗大模型（Medical LLM）在 2025 年的研究重心逐渐从“能回答医学问题”转向“能在临床语境中可靠地工作”：既要具备覆盖生物医学与临床文本的知识底座，又要在诊断、鉴别、治疗建议等高风险任务中体现稳健推理与可控性；同时还必须满足医疗场景对安全性（避免幻觉）、可解释性（推理依据可追溯）、以及交互性（能追问、会总结、可协同）的更高要求。围绕这一目标，今年的相关工作大体沿着“早期 SFT 与知识微调的医疗知识注入”、“医疗持续预训练

表 5.14: 2025 年医疗大模型在知识适配方面的代表性工作

技术维度	代表性工作	主要贡献
持续预训练	Med-PaLM 2 ^[651]	强基座 + 医学领域适配 + 提示集成精炼
	Me-LLaMA ^[652]	生物医学文献与临床笔记等持续预训练和指令微调
	MMedIns-Llama 3 ^[653]	大规模医学指令混合与多任务医学指令混合训练
	MedFound ^[654]	面向诊断决策偏好的推理微调与偏好对齐
可验证化的医学推理	AlphaMed ^[655]	医学任务上纯强化学习训练
	m1 ^[656]	强调对推理链关键步骤的评估
	Fleming-r1 ^[657]	“可验证推理”为目标，冷启动 + 强化学习
	MedS ³ ^[658]	提出自演化训练框架、引入软过程监督信号
全流程医疗 Agent	AMIE ^[659]	模拟对话环境与反馈机制提升对话诊断能力
	Healthcare agent ^[660]	提出面向问诊的 Agent 框架
	DxDirector ^[661]	将 LLM 从“被动应答工具”提升为“诊断流程组织者”

与通用化能力构建”、“面向可验证/慢思考的医学推理训练”、“面向全流程的医疗 Agent 系统化”四个方向展开：一方面增强专业知识与领域适配，另一方面通过推理训练与 workflow 设计把模型从“医学 QA 工具”推进到“可落地的临床协作智能体”。

(1). 早期 SFT 与医学知识注入

这一类工作主要解决通用 LLM 医学知识稀缺、回答不稳定的问题，核心思路是通过医学指令数据与知识注入，让模型形成更可靠的医学表述与问答能力。HuaTuo (华驼)^[662] 以 LLaMA 为基础，构建生成式医学问答数据并进行监督微调 (SFT)，强调医学回答的专业性与可信度提升，代表了早期“用高质量医学 QA 指令数据拉齐医学表达与知识分布”的典型路线。PMC-LLaMA^[663] 则更进一步，将大规模医学文献与教材内容注入，并结合指令微调实现对医学领域的系统适配；其关键价值在于证明了相对中等规模 (如 130 亿参数) 的开源模型，通过“知识注入 + 指令对齐”也能在多项医学问答基准上达到强竞争力，为后续开源医疗模型的规模—性能权衡提供了可复用范式。

(2). 医疗持续预训练驱动的知识增强

与仅靠 SFT 不同，持续预训练更关注“让模型真的具备医学语言与临床文本的底层建模能力”，从而支撑更复杂的临床任务形态。Med-PaLM 2 相关工作^[651] 代表了一条“更强基座 + 医学领域适配 + 提示集成精炼”的路线，目标是推动医学问答性能达到或逼近医生水平，并通过人工评估强调临床可用性提升。Medical foundation LLM 方向^[652] 则从数据覆盖面出发，使用生物医学文献与临床笔记等多源医疗语料进行持续预训练，并辅以医疗指令微调，使模型同时具备对临床文本与生物医学文本的综合理解与生成能力，为后续推理、对话与 workflow 系统提供“通用医学地基”。进一步地，面向“评测—数据—训练”闭环的研究^[653] 强调用覆盖多任务的医学能力基准与大规模医学指令混合数据，训练出更“多形态泛化”的医疗 LLM，使模型能够跨任务、跨场景迁移。针对跨专科诊断辅助，通用诊断模型工作^[654] 则把重点放在诊断推理微调与偏好对齐上：不仅要输出医学上合理的结论，还要贴近临床决策偏好与表达规范，体现“医学正确性 + 临床可用性”的双重目标。

(3). “慢思考”与可验证化的医学推理

医疗推理天然高风险：错误链条会导致诊断偏离、治疗建议不当，因此研究开始系统性地把“推理质量”作为一等公民来优化。Beyond distillation^[655] 提出以规则化、可编程的极简奖励设计开展纯强化学习训练，用更明确的信

号诱导模型形成可用推理行为，强调用工程可控的奖励约束推理而非只依赖蒸馏。m1^[656] 将“医学推理能力”作为训练主目标，并讨论推理时扩展 (test-time scaling) 带来的收益与“过度推理”风险，反映出医疗场景需要在“更强推理”与“不过度生成/不瞎编”之间做精细权衡。Fleming-r1^[657] 以“可验证推理 (verifiable reasoning)”为目标，采用冷启动链式推理与后续强化学习优化结合，强调推理的稳定性与可解释性。MedS³^[658] 则面向“医学慢思考”，通过搜索/自生成构造可用于过程监督的推理轨迹，并引入软过程监督信号来约束推理步骤质量，以更可部署的小规模模型为目标，体现了“过程质量监督”在医疗推理中日益重要的趋势。

(4). 面向全流程的医疗 Agent 系统化

当模型具备一定医学知识与推理能力后，真正落地往往取决于能否组织完整诊断流程并进行高质量交互。因此，医疗 Agent 研究把任务从“单轮问答”扩展到“病史采集—追问—检查建议—结果解释—鉴别诊断—总结报告”的闭环。对话式诊断 AI 范式^[659] 将病史采集、诊断推理与医患沟通纳入统一的 LLM 核心 Agent 框架，并通过模拟对话环境与反馈机制迭代提升对话诊断能力，强调“交互中逐步逼近正确诊断”。Healthcare agent^[660] 更偏向在线问诊场景的工程化拆解：将咨询拆分为对话规划与追问、记忆管理、结果总结与报告生成等模块，并引入医生介入策略，确保系统在可控边界内完成咨询闭环。Reverse Physician-AI Relationship^[661] 则进一步把 LLM 从“被动应答工具”提升为“诊断流程组织者”，围绕主诉理解、追问、检查建议、结果解释与鉴别诊断等环节进行系统化训练与偏好优化，体现出医疗 Agent 从“辅助回答”走向“流程编排与协同决策”的方向演进。

2. 医疗领域多模态的扩展 尽管多模态大模型 (M-LLM) 已成为通用视觉任务的重要范式，但医疗场景的特殊性（如影像模态多样、3D 体数据普遍、全切片病理超高分辨率、临床解读依赖可定位证据）对其提出了更高要求，使得 2025 年的研究重点从“单图问答/粗粒度分类”转向对数据形态桥接、跨任务基础表征、细粒度定位与多图像对照推理的系统优化。在 3D 影像理解方面，Vote-MI^[664] 通过无监督切片选择将 3D 体数据桥接为代表性 2D 输入，缓解了通用 VLM 难以直接处理 3D 影像的结构鸿沟；在基础模型层面，CONCH^[665] 与 BiomedGPT^[666] 分别面向病理与更广泛生物医学任务开展大规模图文对齐预训练，而 TITAN^[667] 进一步把“全切片 (WSI) 自监督 + 多模态对齐”推向可复用底座，支持检索、表征与报告生成等多类下游任务。在输出可控与可解释方面，MIMO^[668] 通过“视觉指代输入 + 像素级定位输

表 5.15: 2025 年医疗大模型在多模态方面的代表性工作

技术维度	代表性工作	主要贡献
3D 影像理解	Vote-MI ^[664]	缓解了通用 VLM 难以直接处理 3D 影像
	CONCH ^[665]	面向病理图文对齐预训练
	BiomedGPT ^[666]	面向更广泛生物医学图文对齐预训练
	TITAN ^[667]	把“全切片自监督 + 多模态对齐”推向可复用底座
基础模型层面		
输出可控方面	MIMO ^[668]	实现对关键结构/术语的精细对齐
	Med-MIM ^[669]	补齐多图像理解与对照推理

出”实现对关键结构/术语的精细对齐；而 Med-MIM 及相关工作^[669]则补齐多图像理解与对照推理基准，推动模型从单张图像走向更符合临床阅片习惯的多视角/多序列联合判断。此外，Uni-Med^[670]以 Connector-MoE 缓解多任务冲突，为多模态多任务统一建模提供了更工程化的结构选择。这些工作共同推动医疗多模态从“能看懂”走向“能对齐、能定位、能迁移”，为影像辅助诊断与自动化报告生成等 2025 年核心应用形态提供更可靠的能力底座。

3. 医疗领域的数据资源与基准构建 医疗大模型的专业化能力，很大程度上取决于“数据—任务—评测”三者能否对齐真实临床流程：既要覆盖基础医学知识与推理，也要能在电子病历、检索证据、多模态检查结果等复杂输入下给出可追溯的结论。早期研究中，领域评测多围绕标准化考试与医学问答展开，形成了以 MedQA、MedMCQA、MultiMedQA 等为代表的一批基准任务；同时也逐步出现面向中文医疗场景的资源建设与能力短板讨论（如英文语料占主导导致跨语言迁移受限等）。

(1). 综合医学认知与推理能力评测基准医学“认知能力”评测通常以

表 5.16: 2025 年医疗大模型在数据资源与基准构建方面的代表性工作

技术维度	代表性工作	主要贡献
综合医学认知与推理	JAMA Clinical Challenge & Medbullets ^[671]	问题构造更接近真实临床情境
	AfriMed-QA ^[672]	泛非语境下的多专科医疗问答
临床文本与电子病历	ArchEHR-QA ^[673]	聚焦可读可溯的医疗问答
检索增强	DiN ^[674]	标注不完美情况下的稳健性与泛化能力
	M3Retrieve ^[675]	多领域与多任务的多模态检索基准
临床工作流	MedChain ^[676]	真实临床工作流的连续决策
	MedAgentBoard ^[677]	多智能体医疗任务评测框架
	CURE-Bench ^[678]	用药决策与治疗方案规划等临床决策

多选题或结构化问答为载体，强调对指南与病理生理机制的理解、对鉴别诊断链条的推断，以及对医学常识与专业知识的稳定调用。传统基准（如 MedQA、MedMCQA、MMLU 医学子集等）为模型提供了可复现实验的统一坐标，但其题目风格往往更接近考试而非真实病例讨论。今年的评测趋势更明显地转向“高难度病例 + 解释性输出”。例如，JAMA Clinical Challenge 与 Medbullets^[671] 将问题构造建立在更接近真实临床情境的病例挑战上，并显式提供专家撰写的解释，用以评估模型能否给出可信的推理依据，而不仅是猜对选项。与此同时，面向全球健康与数据分布偏差的评测也在增强。AfriMed-QA^[672] 以泛非语境下的多专科医疗问答为核心，推动模型评估从“单一地区医疗知识”走向“跨地区、跨人群”的外部有效性检验。

(2). 面向临床文本与电子病历的证据对齐基准在临床落地场景中，“答得像”远远不够，关键在于能否把答案与可核验的证据片段对齐：比如从病历、既往史、检验检查或用药记录中抽取支持性证据，再完成总结、解释或建议生成。围绕检索与医学问答的经典数据资源（如 MEDIQA-2019、MS MARCO MED、LiveQA 等）长期支撑了医学检索与消费者健康问答评测，但其“证据约束”与“病历个体化”仍相对有限。2025 年的新基准更强调“以 EHR 为条件的可追溯生成”。ArchEHR-QA^[673] 共享任务将患者门户消息与关键病历证据绑定，要求系统在给出回复时同时满足“面向患者的可读性”与“基于 EHR 证据的可追溯性”，从而把评测从通用问答推进到真实医疗沟通负担的核心环节。

(3). 多模态理解与检索增强评测基准医疗场景天然是多模态的：影像、波形、检验指标与文本记录共同决定诊断与治疗计划。然而公开可用的多模态数据往往规模受限、模态覆盖不全，导致评测容易停留在“图文问答”或“单一科室影像任务”。例如，MIMIC-IV 虽包含临床文本、胸片影像与部分时间序列，但总体仍属于模态有限的公开资源；PMC-VQA、LLaVA-Med 等则更集中于图文形式的医学 VQA。在此背景下，2025 年的评测一方面开始正视“标注噪声”这一医学数据的常态问题：DiN^[674] 工作通过语义化噪声构造，提出了面向 Med-VQA 的噪声标签基准，用以评估模型在标注不完美情况下的稳健性与泛化能力。另一方面，RAG 范式推动评测从“能不能答”转向“能不能检索到关键证据”。M3Retrieve^[675] 构建了覆盖多医学领域与多任务形态的多模态检索基准，用于系统化评估不同检索模型在医疗场景下的检索质量与鲁棒性，为后续“检索—推理—生成”的端到端评测提供了更坚实的基础设施。

(4). 可信性、安全性与临床 workflow 适配评测医疗 AI 的评测最终必须回

到“可信可用”：输出是否遗漏关键医学概念、是否出现潜在伤害性建议、是否能在任务链条中保持一致性。已有研究提出了面向摘要与生成质量的医学概念覆盖度（Medical Concept Coverage）等指标，并讨论了用平均不安全匹配数（AUM）量化潜在风险思路，为后续安全与忠实度评测奠定了度量基础。2025 年的基准构建进一步把“安全可信”落到临床流程与多主体协作上：MedChain^[676]以真实临床工作流的关键阶段为骨架构建案例数据，强调个体化、交互性与顺序性，使评测能够覆盖“连续决策”而非一次性问答。MedAgentBoard^[677]则从多智能体协作角度出发，系统覆盖医学（多模态）问答、科普摘要、结构化 EHR 建模与临床工作流自动化等任务类别，便于比较“单模型—多智能体—传统方法”在不同任务上的真实收益与代价。此外，CURE-Bench^[678]将评测推进到用药决策与治疗方案规划等更贴近临床决策支持的环节，把“正确性、药物安全性、方案设计”纳入同一评测框架，为医疗推理模型的可用性评估提供了更接近落地的标尺。

4. 业界进展 2025 年医疗大模型的行业进展的深层的变化是“落地逻辑”发生了切换，从过去强调模型本身能不能回答医学问题，转向强调它能不能进入医院信息系统、接受权限与审计约束、在可追溯的证据链下稳定运行。于是，业界叙事也从“做一个医学聊天助手”升级为“把大模型做成可集成的能力层”，围绕合规、工程交付与 workflows 改造展开竞争。

首先，政策与治理信号在年末变得清晰，直接改变了医疗机构的决策方式。11 月，多部门联合印发《关于促进和规范“人工智能 + 医疗卫生”应用发展的实施意见》，将应用方向系统化铺开，并把数据安全、个人信息保护、规范备案与持续评估等要求明确写进框架之中。对医院而言，这类文件的价值并不在于“鼓励使用”，而在于把“能不能用”转写成“怎样算合规、怎样算可控、怎样算可验收”，让院内试点具备可扩面、可复盘的制度基础。

其次，国内厂商在 2025 年呈现出明显的“工程化收敛”：同样谈大模型，但重点不再是对话效果，而是平台能力与场景闭环。**腾讯健康**以“一底一台双引擎”融合创新战略把 DeepSeek+ 混元等大模型能力接入医院场景，已服务全国超 1000 家医院，并在数字生态大会上开放健康管理助手等轻量 AI 智能体能力；**阿里云**联合卫宁健康发布“一站式”大模型智算解决方案（AI Stack 智算一体机 + 医疗知识增强的通义千问微调），覆盖病历自动生成、导诊问答、诊后健康管理等院内全流程落地；**华为**围绕“人工智能医院/智慧一院多区/鸿蒙医疗生态”推进落地，并联合发布集成 AI 大模型的 5.5G 分布式手持 AI 超声系统用于实时辅助诊断；**京东健康**推出“AI 医院”并升

级“京医千询 2.0”医疗大模型，强调循证对齐与可信推理以支撑“医检诊药”闭环场景，同时构建“AI 京医”多角色智能体矩阵在全域健康服务中规模化落地；**字节跳动（小荷健康）**上线独立 App “小荷 AI 医生”，定位健康管家并提供健康问题咨询、报告解读等服务，标志其在 C 端 AI 医疗助手上的产品化落地；**蚂蚁集团**上线 AI 健康管家 AQ，以“蚂蚁医疗大模型”为底座引入近 200 位三甲名医 AI 分身并对接 200 余家医院云陪诊，主打咨询、报告解读与多模态交互的 C 端落地；**科大讯飞**同样沿着“产品化 + 行业交付”推进，发布星火医疗大模型 V2.5 国际版并升级面向用户的产品形态，推出更贴近院内工作流的“智医助理医院版”类产品，强调其与病历与院内流程的融合能力。可以看到，头部厂商的差异不再是“谁的模型更大”，而是“谁更像一个能交付的医疗 IT 能力提供方”。

国际市场的变化更集中地落在“可量化减负”的刚需场景上——临床文书与语音工作流。**Microsoft** 发布面向临床工作流的 Dragon Copilot，同时研究 MAI-DxO 等多模型诊断编排器以提升临床推理与决策支持能力；**Google Cloud** 在 HIMSS 展示医疗“生成式搜索 + 智能体”落地案例，并扩展 Vertex AI Search for Healthcare 的多模态能力，推动临床信息检索与流程自动化；**Amazon AWS** 持续推动 HIPAA-eligible 的 HealthScribe 把“对话转录 → 临床笔记生成”能力产品化，并通过客户合作把生成式 AI 嵌入临床文书与运营工作流；**NVIDIA** 围绕 BioNeMo/加速计算与医疗 AI 生态，与 Mayo Clinic、IQVIA 等机构合作推进医疗基础模型在药物发现与临床研究等场景落地；**Oracle Health** 推出嵌入 EHR 的 Clinical AI Agent，覆盖 30+ 专科并宣称可减少文书时间约 30% 并继续发布 AI 驱动的 EHR 能力，强化临床记录与工作流自动化；**OpenEvidence** 将“循证检索/问答 + 可追溯引用”的临床点-of-care 助手快速规模化，官方称已覆盖 40%+ 美国医生、并在大量医院/医疗中心使用；同时在 7 月完成约 2.1 亿美元 B 轮（估值约 35 亿美元）、10 月再融资约 2 亿美元（估值约 60 亿美元），并扩展与 AMA/JAMA/NEJM 等内容合作以及与 ACC/ACEP、Microsoft、Veeva 等在指南/工作流与临床试验等方向的合作落地。

总体而言，2025 年的业界进展并不是“发生了很多事”，而是这些事共同强化了同一个方向：医疗大模型正在从“能力演示”走向“系统能力”，从“单点应用”走向“工作流改造”。政策给出合规路径与评估要求，厂商把权限、审计、工具编排、私有化部署与交付体系做成产品能力，生态伙伴补齐实施与集成，医院则围绕文书、质控、导诊、随访等高频闭环环节推进规模化使用。竞争焦点因此从“单轮回答质量”转为“能不能持续稳定地嵌入临

床真实运行”。

5.2.3 金融

2025 年，金融大模型（Financial LLMs）的研究重心已从单一的信息抽取与情感分析，转向对金融决策逻辑的深度解构与智能体协同作业的全面探索。受多模态推理与群体智能技术爆发的驱动，本年度的金融智能研究呈现出鲜明的“实证化”与“系统化”趋势。本节将从应用视角出发，重点梳理大模型在智能投研、量化交易与信贷决策、以及金融合规风控三大核心场景下的落地进展，探讨技术如何赋能卖方分析师、交易员、风控官及开发者等不同角色，推动金融业务从“辅助生成”向“自主决策”跃迁。相关代表性工作汇总如表 5.17 所示。

1. 智能投研与辅助分析 面对海量且非结构化的市场信息，2025 年的金融大模型研究致力于将“单点信息处理”升级为“全流程投研助手”。研究重点在于解决长篇研报生成的幻觉问题、深度因果逻辑的挖掘以及针对不同投资者群体的个性化服务。

(1). 面向专业分析师与基金经理：从“摘要生成”到“深度洞察”

针对卖方分析师撰写深度研报的需求，本年度的工作不再满足于简单的文本摘要，而是转向模拟人类专家团队的协作模式。Beyond Summaries^[679]提出了一种多模态协作框架，能够协同生成文本分析、数据表格和可视化图表，实现了图文并茂的专业级报告生成；Agentic Pipeline^[680]与 Role-based LLM^[708]则构建了模仿分析师思维工作流（从数据提取、逻辑推演到观点形成）的流水线，通过引入宏观分析师、行业研究员等不同角色，显著提升了报告的逻辑连贯性与专业深度。

在逻辑增强方面，FinCoT^[681]将杜邦分析法等经典金融模型显式地融入思维链生成过程，确保了分析过程的“科班出身”。为了解决复杂数据追踪问题，Nararatwong et al.^[683]和 Strich^[709]进一步强化了模型在数据库推理和多轮对话场景中的逻辑连贯性。针对财务报表分析的准确性，Wang et al.^[710]深入测试了模型在比率分析和异常发现上的表现，指出了当前模型在处理复杂表格逻辑时的局限。此外，Shukla et al.^[682]利用图谱结构优化了复杂金融叙事的摘要质量，帮助分析师快速穿透复杂的实体关系网。

(2). 面向散户投资者：心理对齐与信息平权

对于缺乏专业背景的个人投资者，AI 工具正致力于降低信息获取门槛并提供符合用户心理预期的建议。Behavioral-Chains^[684]创新性地在校准建

表 5.17: 2025 年金融大模型在不同应用场景下的代表性工作总结。

应用维度	用户群体	代表性工作	核心贡献/技术支撑
智能投研 与辅助分析	投资分析师	Beyond Summaries ^[679]	协同生成图文并茂的专业研报
		Agentic Pipeline ^[680]	模拟分析师从数据到观点的完整 workflow
		FinCoT ^[681]	显式融入杜邦分析等专家逻辑
		GraphRAG Analysis ^[682]	利用图谱结构优化复杂叙事摘要
	散户投资者	Fin-DBQA ^[683]	强化数据库推理与多轮问答追踪
		Behavioral-Chains ^[684]	生成兼顾损失厌恶等心理偏差的建议
量化交易 与信贷决策	交易者/基金经理	Generating News ^[685]	根据股价因子自动生成财经新闻
		KULFi Framework ^[686]	引入外部知识库增强隐性因果推理
		CLRG ^[687]	精准定位原因与结果的语义边界
		Concept-Based RAG ^[688]	基于概念层级理解实现细粒度检索
	信贷/评级人员	FinDebate ^[689]	多智能体辩论提升决策客观性
		Sam's Fans ^[690]	基于记忆框架构建动态交易阈值
		Enhancing PEAD ^[691]	量化盈余公告后的股价漂移
		300k/ns team ^[692]	兼顾交易决策透明度与可解释性
	开发者/宽客	CreditLLM ^[693]	少样本构建信贷产品专业助手
		Forecasting Ratings ^[694]	探究 LLM 与传统 ML 在评级上的边界
		Modeling Interactions ^[695]	文本关联构建动态图谱预测交易量
		Zero-Shot Extraction ^[696]	零样本构建产业链与竞争图谱
合规审计 与金融风控	监管者/审计师	FinMoE ^[697]	稀疏激活架构降低高频推理成本
		Chat Bankman-Fried ^[698]	模拟高风险人物探测欺诈边界
		Fraud-10K ^[699]	从年报微细措辞中检测欺诈迹象
	企业风控官	SEC-QA ^[700]	监管文件合规性问答系统评测
		LAVA ^[701]	自动化检测文档内部逻辑与数据矛盾
	开发者	Detecting Evasive ^[702]	识别高管电话会中的闪烁其词
		AveniBench ^[703]	综合金融智能安全与能力评测
		FMD-Millama ^[704]	多模态推理粉碎金融谣言
		FinEval-KR ^[705]	解耦静态知识记忆与动态逻辑推理
		Fin. Literacy ^[706]	利用 DSL 测试会计核心逻辑
		Capybara ^[707]	利用 CoT 提升谣言验证逻辑严密性

议生成中引入了行为金融学原理，兼顾了用户的损失厌恶等心理偏差，使建议更易被接受且更具温度。在信息获取侧，Nishida et al.^[685] 探索了根据股票价格涨跌因子自动生成高质量财经新闻的流程，实现了市场动态的即时解读，极大地缩小了机构与散户之间的信息时间差。

(3). 面向金融开发者：因果引擎的底层强化

为了支撑上层应用的准确性，开发者层面的研究聚焦于底层因果逻辑的挖掘。在任务定义上，Sandoval et al.^[711] 详细界定了金融因果链条抽取的挑战。KULFi Framework^[686] 引入外部金融知识库来增强模型对隐性因果链条的推理能力；而 CLRG^[687] 与 Sarang 系统^[712] 则专注于从复杂的非结构化文本中精准定位“事件驱动”的因果边界。

在技术路线的选择上，Chatwal et al.^[713] 探讨了特定思维链提示 (CoT) 在无微调下的潜力，但 Niess et al.^[714] 通过对比研究指出，在对准确性要求极高的金融因果问答中，特定任务的微调比单纯的提示更能有效抑制模型幻觉。针对低资源语言和无监督场景，Sebbag et al.^[715] 提出利用预训练大模型作为“预言机”进行无监督经济关系发现，Jeenoor et al.^[716] 和 Aviles et al.^[717] 则分别展示了模型在解释因果关系及跨语言抽取方面的零样本潜力。此外，Concept-Based RAG^[688] 通过建立金融领域特定的概念层级，实现了比传统关键词匹配更精准的细粒度事实检索。

2. 量化交易与信贷决策 2025 年，大模型在交易与信贷领域的应用逐渐触及核心决策层。研究主流从直接预测股价的粗放模式，转向利用 AI 辅助构建交易因子、优化信贷评估流程以及提升决策系统的博弈对抗能力。

(1). 面向交易员与量化研究员：动态博弈与非结构化因子挖掘

在动态变化的市场环境中，单体模型的决策往往容易陷入局部最优。FinDebate^[689] 引入了多智能体辩论机制，通过模拟不同观点的交易员进行对抗性辩论，显著提升了决策的客观性与鲁棒性。为解决研报生成的片面性，Takayanagi et al.^[718] 定义了财报会议记录生成的标准，而 Tan et al.^[719] 和 Rallabandi et al.^[720] 深入验证了基于角色扮演的协同架构能显著提升分析深度。如果不采用多智能体架构，Chatwal et al.^[721] 提出的元提示策略也展示了通过高阶任务分解实现复杂规划的可行性。

在交易决策方面，Yu et al.^[722] 设定了基于 Agent 的加密货币交易挑战。在此挑战中，Sam's Fans^[690] 利用 FinMem 记忆框架构建动态交易阈值，敏锐捕捉市场波动；而 Anonymous^[692] 侧重于决策透明度，通过 PEFT 技术优化模型以生成可解释的交易理由。在因子挖掘上，Hadlock et al.^[691] 展示

了如何利用 LLM 分析非结构化情感来更精准地量化“盈余公告后的股价漂移”现象。

(2). 面向信贷审批与评级人员：效率提升与理性边界

针对信贷业务中数据稀缺与模型可解释性差的问题，CreditLLM^[693] 利用少样本学习技术，在极度缺乏标注数据的情况下构建出了具备专业知识的信贷产品助手。然而，研究也保持了理性的审视：Drinkall et al.^[694] 通过对比研究指出，在处理纯结构化表格数据进行信用评级时，传统机器学习方法在稳定性上仍优于生成式 LLM，这为信贷人员选择技术工具提供了重要参考——即 LLM 应更多聚焦于文本与逻辑分析，而非纯数值预测。

(3). 面向算法工程师与宽客：架构效率与图谱构建

为了适应高频交易对低延迟的苛刻要求，FinMoE^[697] 提出了基于混合专家架构的金融大模型，利用稀疏激活机制显著降低了推理成本。针对模型优化，Wang et al.^[723] 引入代理微调技术，利用小模型引导大模型生成，克服了全参数微调的高成本障碍。

在数据挖掘工具上，Xu et al.^[695] 利用 LLM 挖掘文本关联构建动态图谱以提升交易量预测精度；Zero-Shot Extraction^[696] 则证明了 LLM 能够利用内置世界知识，在零样本条件下构建出高质量的产业链竞争图谱。对于检索增强，Srinivasan et al.^[724] 结合了 Agentic AI 与多重假设文档嵌入 (Multi-HyDE) 技术以减少事实错误，Lithgow-Serrano et al.^[725] 则建立了一套系统性的 RAG 能力评估体系。

3. 合规审计与金融风控 随着监管科技 (RegTech) 的兴起，2025 年的研究重点在于利用大模型的主动防御能力，解决金融欺诈检测、合规性审计以及模型自身的安全对齐问题。

(1). 面向监管者与外部审计师：主动探测与欺诈识别

面对日益复杂的金融犯罪手段，监管工具正变得更具主动性。Liu et al.^[726] 详细介绍了金融谣言检测的任务设置。在具体检测技术上，FMD-Mllama^[704] 展示了多模态推理的优势，而 Cao et al.^[707] 和 Purbey et al.^[727] 利用思维链 (CoT) 与序列学习策略，通过逐步拆解复杂声明提升了验证的逻辑严密性。

在安全对齐方面，Biancotti et al.^[698] 进行了一项极具前瞻性的研究，通过构建模拟高风险人物（如 FTX 创始人）的角色，探测大模型在面对金融欺诈、监管规避等敏感话题时的防御边界，为监管者提供了“红队测试”的新思路。Amin et al.^[699] 则探索了模型能否从上市公司年报微妙的措辞变化

中主动发现潜在的欺诈迹象。SEC-QA^[700] 则针对监管文件问答构建了系统性评测。

(2). 面向企业风控官与内审人员：逻辑校验与微表情捕捉

针对企业内部的合规审查，LAVA 框架^[701] 设计了逻辑感知的验证机制，能够自动识别大规模文档内部数字与条款之间的逻辑冲突。此外，Wu et al.^[728] 提出利用 NLI 模型作为“裁判”自动检测推理链条中的事实矛盾。Nuaimi et al.^[702] 则将风控延伸到了心理层面，利用心理语言学体系自动检测高管在财报电话会中面对敏感提问时的闪烁其词，显著增强了风控人员对非言语风险信号的捕捉能力。

在合规科技（RegTech）的落地应用上，Wang et al.^[729] 总结了 NLP 技术在合规检测中的进展。Martínez et al.^[730] 和 Chantangphol et al.^[731] 分别提出了可扩展的法律文本理解架构和解决法规歧义的方案；Huang et al.^[732] 发布了专门针对审计领域的开源微调大模型，Jiang et al.^[733] 则展示了优化模型在解析复杂法律条款时的精准度。

(3). 面向安全研究员与开发者：基准建设与谣言粉碎

为了构建可信的金融 AI 生态，开发者层面的工作主要集中在评测基准与虚假信息防御上。AveniBench^[703] 与 FinEval-KR^[705] 分别从综合能力和“知识-推理解耦”的角度，确立了金融大模型的评估标准。Fin. Literacy^[706] 利用领域特定语言（DSL）和纯文本会计测试量化了模型的核心记账逻辑。

在基础资源建设上，Wang et al.^[734] 构建了多任务标注商业文档数据集，Abdo et al.^[735] 填补了阿拉伯语金融 NER 的空白。Harsha et al.^[736] 验证了合成数据对提升 QA 模型性能的有效性，Zhao et al.^[737] 提出了生成高质量数据评级描述的自改进方法。针对特定任务优化，Sharma et al.^[738] 结合量化技术优化了边缘部署，Lu et al.^[739] 评估了 LLM 在 NER 任务上的边界，Kumar et al.^[740] 提出了跨语言迁移学习，Uthayasooriyar et al.^[741] 强调了视觉布局信息的重要性，Qiu et al.^[742] 评估了图表描述能力，而 Anonymous^[743] 探讨了 DPO 中的长度奖励机制。

4. 行业整体生态 2025 年中国金融智能行业的主旋律是“金融基建的智能化重构”与“业务决策的自主闭环”。与前两年金融大模型停留在“智能客服”和“辅助写作”的表层应用不同，2025 年的显著特征是监管科技确立了 AI 风控的“国家标准”，科技巨头推动通用模型向具备专业工具链的“金融智能体”进化，而头部金融机构则通过深度定制加速 AI 从“副驾驶”向“主理人”的角色跃迁。

2025 年大语言模型（LLMs）进展报告

在监管与基础设施侧，2025 年是金融监管数字化转型的“标准化元年”。**国家金融监督管理总局**牵头构建的“**金融行业大模型监管沙盒**”正式转入常态化运行，并于 2025 年 11 月发布了首批《**金融生成式 AI 风险治理指引**》，标志着金融 AI 的应用有了明确的合规红线与准入标准；

上海、深圳等地的金融数据交易所正式上线了“**语料资产专区**”，打通了银行流水、信贷记录等高价值私域数据与大模型训练之间的合规流通过程，为金融 AI 的深度学习提供了统一且合规的数据底座。

在技术供应侧，AI 巨头们不再通过单纯的参数竞赛来吸引 B 端客户，而是转向发布能够通过“专业资格考试”且具备复杂工具调用能力的金融智能体。

通义大模型在云栖大会上发布了“**通义金融 Agent 2.0**”，重点展示了其连接 Wind、Bloomberg 等专业终端进行实时数据清洗与因子挖掘的能力，将服务边界从 SaaS 工具拓展到了 PaaS 级的投研中台；

百度发布了**文心大模型金融行业版（v4.5）**，特别强调了“因果风控”能力，并展示了其如何协助国有大行在复杂的信贷审批流程中，通过多模态证据链分析实现风险预警的自动化闭环。

对于金融服务业，头部机构通过“自研 + 结盟”的双轮驱动，加速将 AI 植入核心业务流。

招商银行宣布其“**AI 数字财富顾问**”全面覆盖零售业务，这是业内首个获批能够独立向客户提供标准理财建议的生成式 AI 应用，标志着 AI 在财富管理领域从“咨询”走向了“决策”；

中信证券与华为云签署全面深化战略合作协议，基于**盘古大模型**共建“**证券行业投研大模型底座**”，旨在通过全栈国产化的算力与算法，重构从宏观策略到个股挖掘的完整投研体系，这也代表了券商数字化转型正式迈入“AI 原生”的深水区。

5. 小结 综上所述，2025 年金融大模型的发展已完成从技术探索到价值创造的转身。在应用层面，我们看到从简单的文本处理向深度逻辑推理、从单体智能向多智能体协作博弈、从通用辅助向专业岗位替代的清晰路径。无论是辅助分析师生成深度研报，还是协助交易员捕捉瞬时机，亦或是帮助风控官识别潜在欺诈，AI 正逐步内化为金融系统的核心生产力。与此同时，随着监管沙盒的常态化运行与数据基建的完善，一个更合规、更智能、更具因果确定性的金融 AI 生态正在形成。

5.2.4 法律行业

2025 年，法律大模型迎来了从“语言能力”向“思维能力”跃迁的关键转折点。技术的演进不再局限于表层的文本检索与生成，而是深入到了复杂的法律逻辑推理、高精度的文书规划以及动态的合规风控体系之中。本报告旨在系统梳理本年度法律大模型的研究与应用进展，重点从司法裁判支持、文书辅助写作、深度法律检索、企业合规风控及行业整体生态五个维度展开。我们将探讨大模型技术如何通过引入强化学习、神经符号架构及多智能体协作等技术，解决逻辑幻觉等核心痛点，从而推动法律人工智能从实验室走向司法与商业应用的深水区，相关代表性工作汇总如表 5.18 所示。

1. 法律推理与司法裁判支持 法律大模型正在从单纯的信息检索向深层逻辑推理与裁判模拟跨越。2025 年，大模型在法律推理领域的研究的重点在于解决模型在复杂案情下的逻辑自治性、数值计算的准确性以及对动态法律环境的适应性。

(1). 面向司法审判人员：裁判逻辑的严密化与“算无遗策”

针对法院与法官在审判核心环节的需求，本年度的研究致力于提升裁判的精确度、逻辑可解释性与计算准确性。在事实认定与法条适用阶段，LFP^[744]创新性地提出了从原始证据直接预测法律事实的任务，填补了司法认知早期的空白；CLEAR^[745]则专注于区分易混淆的法律款项，辅助法官精准适用法条。为了强化裁判文书的逻辑与计算可靠性，LexNum^[746]专门解决了赔偿金计算中“懂法不懂算”的难题，确保计算过程严格符合法律程序；L4M^[747]更是引入了形式化方法，通过神经符号架构为高风险判决提供数学证明级的逻辑保障。在推理生成环节，LegalReasoner^[775]和 LawChain^[776] 分别在通用判决和民事侵权案件中引入分步验证与结构化思维链，模拟法官思维纠正逻辑谬误；ATRIE^[777]则能自动化地分析大量判例来界定模糊法律概念，为疑难案件的裁判提供理论支撑；而面对法律的动态变化，LawShift^[778]确保了辅助系统能适应刑法修正案带来的规则变更。

(2). 面向法律执业人员：论证策略的结构化与专业化

对于律师与检察官而言，AI 工具正逐步从简单的检索工具演变为构建严密诉讼策略与处理专业领域案件的强力助手。在复杂案件分析中，JUREX-4E^[748]基于刑法“四要件说”构建专家知识库，显著增强了控辩双方对犯罪构成的分析深度；CLaw^[779]则利用最高法指导性案例，验证了模型具有像资深律师一样分析争议焦点的能力。在构建法律论证架构时，SyLeR^[749]强制

表 5.18: 2025 年法律大模型在不同应用场景下的代表性工作总结

应用维度	用户群体	代表性工作	核心贡献/技术支撑
法律推理 与司法裁判	法院/法官	LFP ^[744]	从原始证据直接预测法律事实
		CLEAR ^[745]	区分易混淆款项辅助法条适用
		LexNum ^[746]	解决赔偿金计算难题
		L4M ^[747]	神经符号架构保障逻辑
		LegalReasoner	分步验证与结构化思维链
	法律从业者	JUREX-4E ^[748]	基于四要件说构建知识库
		SyLeR ^[749]	强制遵循三段论逻辑
		GreekBarBench ^[750]	复杂法律意见书起草基准
	开发者	Unilaw-R1 ^[751]	强化学习提升思维链质量
		Legal-R1 ^[752]	增加推理时间计算量范式
法律文书生成 与辅助写作	法院/法官	LexKeyPlan ^[753]	规划-检索-生成新范式
		SLDS ^[754]	生成高质量司法摘要
		SwiLTra-Bench ^[755]	多司法辖区文书互译
	法律从业者	BriefMe ^[756]	构建 IRAC 结构法律文书状
		CoCoLex ^[757]	复制解码策略防止幻觉
	社会公众	ClaimGen-CN ^[758] JUDGEBERT ^[759]	案情转化为规范诉讼请求 生成通俗易懂法律指南
深度法律检索 与 RAG	法院/法官	LegalSearchLM ^[760]	基于要素生成检索判例
		IL-PCSR	法条先例依赖排序
		LexCLiPR ^[761]	跨语言段落级判例检索
	法律从业者	UniLR ^[762]	统一检索法条类案及罪名
		CitaLaw ^[763]	强化生成内容引用能力
		ACORD ^[764]	针对合同条款的检索
	开发者	KOBLEX ^[765]	参数化法条辅助多跳检索
		GRAF ^[766] CogniBench ^[767]	构建主张图验证法律逻辑 检测推理逻辑认知幻觉
企业合同审查 与合规风控	企业/律师	LegalAgentBench ^[768]	自动化执行尽职调查任务
		PrivaCI-Bench ^[769]	场景完整性理论检测隐私合规
		ProvBench ^[770]	合同冲突检测与条款推荐
	税务/保险	SOLAR ^[771]	形式化验证税务责任计算
	开发者	LegalCore ^[772]	长文档跨段落事件共指消解
		LexTime ^[773]	识别法律事件时间顺序
	社会公众	PAKTON ^[774] ProvBench ^[770]	识别霸王条款与风险 快速理解签署文件风险点

模型遵循法律界经典的“三段论”逻辑，能否辅助律师构建规范且有力的辩护理由。在判例引用与文书写作上， δ -Stance^[780]通过分析判例引用的立场极性与强度，帮助律师精准筛选支持己方观点的先例；GreekBarBench^[750]则通过高难度的自由文本法律推理测试，证明了模型在起草复杂法律意见书方面的潜力，为律师工作提供了高水平的基准测试。

(3). 面向法律智能开发者：推理能力的底层强化

在底层技术层面，研究重心转向了利用强化学习激发模型的推理潜能。Unilaw-R1^[751]和 Legal Δ ^[781]展示了如何利用强化学习和基于信息增益的奖励机制，在缺乏昂贵人工标注的情况下，显著提升模型的思维链质量和推理信心。同时，Legal-R1^[752]等工作对包括 DeepSeek-R1 和 OpenAI o1 在内的最新推理型模型进行了全面评测，并确立了通过增加推理时间计算量 (Test-Time Compute) 来解决法律复杂问题的新范式。

2. 法律文书生成与辅助写作 2025 年，大模型在法律文书生成领域的应用已从单纯的文本续写，向着更具逻辑规划、更符合专业工作流以及更具包容性的方向发展。我们将这些进展梳理如下：

(1). 面向司法审判人员：文书撰写的全流程提效与逻辑重构

针对法院在处理长篇复杂判决书时的效率与质量双重压力，本年度的研究着重于模拟法官的裁判思维流与提升审阅效率。在判决书撰写环节，LexKeyPlan^[753]提出了一种“先规划、后检索、再生成”的创新框架，通过预先生成包含关键法律概念的大纲，引导模型生成逻辑严密且引用规范的长篇裁判理由，有效解决了传统生成模型逻辑断层的问题。在庭审后的文书处理与归档阶段，SLDS^[754]验证了大模型在生成高质量司法摘要方面的潜力，极大缩短了法官研读过往案例的时间；而 SwiLTra-Bench^[755]则作为辅助补充，为多语言司法管辖区的文书互译提供了高精度的评估标准，保障了跨语言司法信息的无损传递。

(2). 面向法律执业人员：专业文书的精确起草

对于律师与企业法务而言，辅助写作工具的演进方向在于从通用的文本生成转向高度专业化的论证构建与严谨的条款起草。在诉讼业务中，BriefMe^[756]将辅助重点从简单的文本润色转向了深度的逻辑构造，通过论点补全与摘要生成任务，协助律师构建符合 IRAC（问题-规则-分析-结论）结构的严密法律文书状。为进一步规避执业风险，CoCoLex^[757]在底层生成机制上引入了基于置信度的复制解码策略，强制模型在生成关键法律事实或条款时忠实“复制”原文，从而为律师提供了一道防止 AI 幻觉的技术护栏。

(3). 面向社会公众：法律服务的普惠化与无障碍获取

面向普通公民与缺乏法律背景的当事人，本年度的研究致力于消除专业壁垒，让法律智能提供真正的普惠、无障碍化的法律服务。在主动维权方面，ClaimGen-CN^[758]显著降低了司法准入的专业门槛，该工作能够辅助不懂法的自诉人将非结构化的案情描述自动转化为规范的诉讼请求。JUDGE-BERT^[759]则聚焦于打破认知的语言障碍，通过严格监控文本的简写过程，确保公众在阅读由 AI 生成的通俗版法律指南时，既能轻松理解，又不会遗漏关键的法律细节与风险提示。

3. 深度法律检索与复杂检索增强生成 本年度法律检索任务的研究重心，已从单纯提升检索召回率转向解决复杂的法律逻辑结构、多跳推理以及检索增强生成（RAG）内容的忠实度问题。针对不同用户群体的需求，相关工作呈现出高度的场景化适配特征。

(1). 面向法院与司法审判人员：聚焦判例深度挖掘

针对司法审判中对先例精准适用的高要求，研究者提出了超越传统关键词匹配的新范式。LegalSearchLM^[760]提出了一种基于“法律要素生成”的检索方法，通过让模型预测目标案件的关键要素来检索判例，显著提升了刑事案件检索的精准度；而针对判例与法条并重的普通法系环境，IL-PCSR^[782]揭示了法条检索与先例检索的内在依赖关系，利用这种关联性优化了重排序流程。在跨国司法场景下，LexCLiPR^[761]针对欧洲人权法院等多语种环境，实现了跨语言的段落级检索，能够辅助法官跨越语言障碍查找判决细节。

(2). 面向律师与企业法务：基于检索增强的高精度文书生成

针对法律从业者在文书起草和合规审查中的具体痛点，ACORD^[764]针对合同起草场景，发布了专家标注的条款检索基准，解决了律师在寻找特定合同条款（如责任限制）时现有 RAG 系统表现不佳的问题。为了应对律师在不同办案阶段的多样化需求，UniLR^[762]设计了统一的法律检索器，通过“关键要素监督”机制，在一个模型中同时实现了法条、类案及罪名的高效检索。同时，针对法律文书必须言之有据的要求，CitaLaw^[763]专门强化了生成内容的引用能力，确保模型在回答复杂法律问题时能像专业律师一样提供准确的法条和判例引证。

(3). 面向法律智能开发者与研究人员：提升检索系统的可靠性

为了突破现有检索系统的能力瓶颈，研究人员在多跳推理、知识图谱融合及幻觉抑制方面取得了显著进展。在复杂推理架构方面，KOBLEX^[765]针对需要结合多个法条才能回答的难题，利用生成“参数化法条”作为跳板实

现了精准的多跳检索；GRAF^[766]与 SAT-Graph^[783]则分别通过构建“主张图”和结合本体论的图检索技术，解决了法律逻辑验证和时序性检索的难题。在系统评估与优化方面，RAG-Failures^[784]与 CogniBench^[767]深入剖析了 RAG 系统的失效机理，并提出了专门检测“推理逻辑幻觉”的认知忠实度评估框架；NitiBench^[785]对比了 RAG 与长上下文模型在特定领域（如税法）的表现，指出了技术选型的方向。

4. 企业合同审查、合规审计与风控 2025 年，企业法律大模型的应用范式正从静态的文档审查向动态的智能体协作与全流程合规风控演进。本年度的研究重点显著聚焦于提升长文本场景下的事实一致性、复杂财税规则的计算可验证性，以及对数据隐私等新兴合规要求的自动化审计能力。我们将这些进展按目标用户群体梳理如下：

(1). 面向企业法务、律师与合规官：智能体协作与全流程合规审计

针对企业在尽职调查、隐私合规及合同全生命周期管理中的高频需求，2025 年的研究显著提升了 AI 的辅助深度。在尽职调查与综合事务处理方面，LegalAgentBench^[768]展示了智能体如何利用工具自动化执行查询公司诉讼记录等复杂调查任务；而在数据合规领域，PrivaCI-Bench^[769]引入了基于场景完整性理论的评估框架，帮助合规官自动化检测业务流程是否符合 GDPR 等隐私法规。针对核心的合同业务，ProvBench^[770]实现了合同条款的冲突检测与法条推荐，确保合同内容的合法性；PAKTON^[774]提出的多智能体框架则不仅能回答“知识产权归属”等特定问题，还能生成包含精确引用的审查报告，解决了长文档阅读效率低下的痛点。

(2). 面向金融、保险与税务专业人士：数值计算验证与逻辑严密性保障

针对涉及大量数值计算与逻辑校验的泛法律金融场景，研究者致力于解决大模型“算不准”与“逻辑不严密”的缺陷。在税务审计方面，SOLAR^[771]利用神经符号架构，通过形式化知识表示实现了可验证的税务责任计算与合规自查，有效辅助税务顾问处理复杂的抵扣规则。针对保险与赔偿领域，LexNum^[746]通过两阶段强化学习解决了工伤或交通事故赔偿金计算中的程序对齐问题，确保证据审核的准确性；JUDGEBERT^[759]则关注保险合同等文本简化过程中的法律含义保留问题，防止因丢失关键免责细节而引发合规风险。

(3). 面向法律人工智能开发者：长文档理解深化与时序逻辑增强

为了支撑上述复杂的企业级应用，开发者层面的研究主要集中在提升模型对长文档的理解力、时间逻辑感知力上。针对合同这一典型的长文档载体，

LegalCore^[772]发布了首个包含跨段落事件共指消解的数据集，赋能模型在数十页的合同中精准追踪同一事件（如违约行为）的来龙去脉。此外，考虑到合规监控对时间维度的敏感性，LexTime^[773]提升了模型识别法律事件时间顺序的能力，为构建自动化的履约进度监控系统提供了技术基础。

(4). 面向社会大众与普通消费者：合同陷阱识别与普惠风险防控

在企业与消费者的交互界面，技术进步正致力于消除信息不对称。前述的 PAKTON^[774]与 ProvBench^[770]等技术同样适用于普通大众，帮助非专业人士快速理解签署文件的核心风险点，或在签署租房合同时自动识别其中可能存在的霸王条款与违反强制性法律规定的风险，从而实现普惠性质的法律风险防控。

5. 行业整体生态 2025 年中国法律智能行业的主旋律是“司法基础设施的统一化”与“商业应用的深水化”。与前两年大模型“百花齐放”不同，2025 年的特点是国家级平台上线确立了行业标准，科技巨头通过技术推动 AI 智能体走向实操，而顶尖律所则通过与巨头结盟加速数字化转型。

在司法侧，2025 年是中国智慧法院建设的“收官与重启”之年，核心事件是全国性平台的整合上线。2025 年 12 月 1 日，最高人民法院正式启用新版平台，整合了原有的立案、调解、执行等分散系统。新平台不仅是流程的线上化，更深度集成了 AI 辅助功能，实现了从立案到执行的全流程数据互通，为未来的司法 AI 提供了统一的数据底座。

在产业侧，AI 巨头们不再满足于提供基础对话能力，而是转向开发能够调用工具、解决复杂任务的法律智能体。在 2025 年 9 月的云栖大会上，通义大模型发布了多项更新，重点增强了 Agent 工具调用和深度推理能力。通义法睿在 2025 年推出了更强的 API 接口，支持企业将法律 AI 助手快速集成到钉钉或内部系统中，从单一的 SaaS 工具转向 PaaS 服务。2025 年 4 月的百度 AI 开发者大会上，百度发布了文心 X1 Turbo，强调了“强推理”能力，并展示了法律智能体如何调用搜索和文档生成工具来完成复杂的法律文书撰写任务。

而对于法律服务业，顶尖律所通过与科技巨头结盟加速数字化转型。2025 年 9 月 16 日，锦天城律师事务所与腾讯云正式签署战略合作协议。这是 2025 年最具代表性的律所数字化事件，标志着顶尖律所不再仅自行研发系统，而是开始引入科技巨头的底层能力来构建“AI+ 法律”新范式。2025 年 4 月 1 日，盈科律师事务所举办了“全球市场暨数智化 AI 融合发展大会”，明确提出将 AI 融入其全球法律服务网络，从而驱动服务模式升级。

6. 小结 综上所述，2025 年法律大模型的发展呈现出鲜明的理性化与结构化特征，标志着该领域已从基于概率的文本拟合向遵循法理的逻辑推演跨越。在算法层面，研究重心回归到法律逻辑的严谨性与过程的可解释性，通过引入法理原则约束、过程验证机制以及与符号逻辑的结合，模型正在习得类似法律专家的思维链条；而针对法律知识检索结构、法律事件时序感知以及生成内容忠实度的 RAG 优化，则有效缓解了复杂法律场景下的语义鸿沟与幻觉问题。在资源层面，数据与评测基准已突破传统的判决预测范畴，向着支持复杂智能体交互、专业文书生成及细粒度合规性评估的高阶方向演进。与此同时，随着司法基础设施的统一化与律所数字化转型的深化，法律 AI 正在重构从数据底座到业务流的完整生态，加速迈向实务应用的深水区。

5.2.5 农业产业

随着生成式人工智能（GAI）与多模态大模型（MLLM）的深度融合，2025 年的农业科技正在经历从“精确农业（Agriculture 4.0）”向以自主决策与生物设计为核心的“智慧生态农业（Agriculture 5.0）”跨越。这一阶段的核心特征是 AI 的角色从单纯的数据感知与分析，转变为具备**基因组设计、复杂推理规划与具身执行能力**的智能体系统。研究重心已从单一环节的效率提升，转向覆盖**种质创新—田间作业—农艺指导—供应链金融**的全产业链智能重构。基于这一范式转换，当前的主流研究与应用可归纳为四个核心维度的系统化演进：在生物育种方面，LLMs 被用于基因型-表型互作解码与蛋白质设计；在田间作业方面，模型赋能农机实现具身智能与复杂环境下的语义理解；在农艺决策方面，研究聚焦于检索增强生成（RAG）驱动的抗幻觉专家系统与多智能体协作；在经营管理方面，则强调供应链韧性模拟与基于大数据的金融风控建模。相关代表性工作汇总如表 5.19 所示。

1. 育种 5.0

在育种研发阶段，2025 年的代表性工作标志着育种技术正式进入“设计时代”。传统的表型筛选模式正在被“数据驱动的生成式设计”所取代。例如，Fu 等^[786]正式提出了“育种 5.0”（Breeding 5.0）框架，利用人工智能对种质资源进行深度解码（AI-decoded germplasm），通过构建“育种飞轮”实现基因型与表型的多模态融合与迭代优化，显著压缩了从实验室到田间的研发周期。与之呼应，InstaDeep 与 Syngenta 合作推出的 AgroNT 模型^[787]，展示了基于 Transformer 架构解析植物基因组“语法”的能力。该

表 5.19: 2025 年农业大模型在不同应用场景下的代表性工作总结

应用维度	用户群体	代表性工作	核心贡献/技术支撑
育种研发 与基因设计	科研机构	Breeding 5.0 ^[786]	AI 解码种质资源与育种飞轮
		AgroNT ^[787]	零样本预测基因突变影响
	种业企业	丰登大模型 ^[788]	生物育种知识整合
	种业企业	MyGenAssist	加速育种研发内部平台
		Heritable Agriculture ^[789]	预测基因组与环境交互
	开发者	AI Breeder for Crops ^[790] Prithvi ^[791]	全基因组选择算法结合 地理空间基础模型
田间作业 与具身智能	农机企业	See & Spray Ultimate ^[792]	视觉大模型实时杂草识别
		Operations Center	软硬一体自主作业生态
	科研机构	天工开悟模型 ^[793]	感知-决策-控制一体化
		AgReason ^[794]	农业推理基准与因果关系
		AgriBench ^[795]	多模态农业应用评估
农艺决策 与知识服务	开发者	PhenoBench ^[796]	像素级作物杂草语义理解
		LLM-Based RL Agent ^[797]	水分胁迫指数编码优化灌溉
	农技推广	Synthetic Dataset ^[798]	LLM 生成合成图像数据集
		九壤耘星 ^[799]	旱区农业多模态数据整合
		Farmer.Chat ^[800]	RAG 技术克服幻觉风险
	科研机构	My Climate Advisor ^[801]	气候适应性建议系统
		神农大模型 3.0 ^[802]	千万级知识图谱与 36 智能体
	开发者	农知大模型 ^[803]	文献溯源与智能伴读
		KALLM ^[804]	双层次知识注入策略
		RL-LLM 融合 ^[805]	强化学习优化农事策略
供应链管理 与金融风控	金融机构	狮子山模型 ^[806]	五维架构实时更新推理
		AgriEval ^[807]	中文农业基准测评体系
	农业企业	Agri-CM3 ^[808]	多模态农业管理评估
		RAG 咖啡病害系统 ^[809]	实时检测与合规治理
		云原生金融平台 ^[810]	信贷处理时间缩短 81%
	开发者	AI 反欺诈模型	供应链金融安全保障
		Agent 供应链模拟	极端事件压力测试
	农业企业	农业耕地大模型 1.0 ^[811]	高标准农田监测预警
	开发者	腾讯混元大模型	智慧温室与金融风控
		百度云智一体	遥感气象地块级指导

模型在 48 种作物基因组的数万亿核苷酸序列上进行预训练，能够以零样本（Zero-shot）方式预测基因突变对调控元件、启动子强度及组织特异性表达的影响，实现了对数百万个突变位点的硅基（In Silico）筛选，大幅降低了生物实验的试错成本。

在蛋白质工程领域，Corteva 与 Profluent^[812] 的合作展示了生成式 AI 在“从头设计”上的潜力。利用在数十亿蛋白质序列上训练的 ProGen3 模型，研究人员成功设计出自然界不存在的 OpenCRISPR-1 基因编辑工具。这项工作证明了 LLM 不仅能理解生物学规律，更能创造出具有更高特异性和稳定性的新型生物元件，为抗病、抗逆作物的研发提供了全新的工具箱。这些工作共同推动了育种从“发现”向“设计”的范式转移，使得针对特定气候条件定制“超级作物”成为可能。

2. 具身智能与田间作业 Agent

在田间生产层面，2025 年的创新集中于将大模型的认知能力赋予物理实体，推动农机从“自动化”向“具身智能（Embodied AI）”进化。传统的计算机视觉模型虽然能识别作物，但缺乏对复杂非结构化环境的推理能力。针对这一局限，Zaremehrjerdi^[794] 提出了农业推理基准 AgReason，并验证了大型推理模型（Large Reasoning Models）在处理农业复杂因果关系上的优势，使机器不仅能“看见”杂草，还能结合农学知识推理出最优的作业策略。为了支持这种细粒度的感知，Zhou^[795] 构建了首个评估多模态大语言模型在农业领域应用的基准 AgriBench，包含详细的土地利用与质量评分标注；而 Weyler^[796] 则提出了 PhenoBench，专注于作物与杂草的像素级语义理解，为无人机的精准作业提供了数据基础。

面向作业执行，John Deere^[792] 在 2025 年 CES 上展示的自主作业生态系统代表了这一趋势的工业化落地。其 See & Spray Ultimate 系统结合了视觉大模型与实时推理，能够在毫秒级时间内区分作物与杂草并精确控制喷头，将除草剂使用量减少三分之二以上。在精细化管理方面，Wu^[797] 构建了基于 LLM 的强化学习智能体（LLM-Based RL Agent），创新性地将作物的水分胁迫指数、根系分布等复杂状态编码为自然语言，使模型能够像人类专家一样理解生长情境并生成优化的灌溉决策，实验验证可节水 15%-30%。此外，针对数据稀缺问题，Sapkota^[798] 利用 LLM 生成合成图像数据集，配合视觉模型实现了在缺乏真实标注数据下的高精度果实识别。

3. 农业知识服务与决策增强

随着大语言模型（LLM）与科学计算、多模态感知、序列化生成技术的深度融合，农业知识服务与智能决策已成为 2025 年农业智能应用的核心前沿。与通用领域的决策任务不同，农业决策要求模型不仅需要处理天气、土壤、作物生理等多源异构的高熵数据，还必须在严格遵循农学原理、生态约束与经济规律的前提下，进行长周期、不确定环境下的动态优化，对决策过程的可靠性、安全性、可解释性与适应性有极高要求，从而解决通用大模型的“幻觉”问题，并提升其在特定农业语境下的通用性。

为此，Jiang^[804]从知识融合方法学入手提出了 KALLM 模型，通过词汇级（监督微调）和句子级（检索增强生成）的双层次知识注入策略，并配合 22 万条高质量问答数据，显著提升了模型决策的事实一致性与可信度，论证了 RL 与 LLM 的融合如何应对现代农业的动态复杂性。Chen^[805]利用强化学习（RL）与环境交互来优化灌溉、施肥等农事策略，而 LLM 则能整合海量领域知识和机理知识，提供符合具体情境的决策依据。鉴于农业数据采集成本高、实时性差等问题，Baja^[813]提出了基于 partial observation 与 cost reward 的变量水肥决策技术，通过强化学习过程监督方式对决策模型进行全参训练，以提升场景的适应性、应对环境的不确定性。

知识图谱的融合也成为关键趋势，Gong^[814]综述了知识图谱与 LLM 的互补融合模式，指出结构化知识表征能有效增强模型在复杂农业问题上的推理能力。在应用端，Digital Green^[800]推出的 Farmer.Chat 系统是检索增强生成（RAG）技术在普惠金融领域的典型应用，该系统通过严格限定模型仅基于本地化农业知识库回答，有效克服了幻觉风险，并在肯尼亚等地的实证研究中显示出极高的用户采纳率。

面向更复杂的综合评估，Yan^[807]构建了首个中文农业基准测评体系 AgriEval，其具备三大特征：1) 能力评估全面性，覆盖农业领域六大任务类别及 29 个子类别，包含记忆、理解、推理和生成四类核心认知场景；2) 数据高质量性：题库源自高校考试与作业，为评估大模型知识应用与专家级决策能力提供自然且可靠的基准；3) 形式多样与规模广泛：包含 14,697 道选择题和 2,167 道开放式问答题，构成当前规模最大的农业领域测评基准。Wang^[808]提出了一个经过专家验证的中文评测基准 Agri-CM3，专用于评估多模态大语言模型（MLLMs）在农业管理中的理解与推理能力。该基准包含 3,939 张图像和 15,901 道具有详细解析的多层级选择题，从三个推理层级与七项组合能力维度展开分析，揭示了多模态大模型在农业场景认知理解

方面的关键挑战。在特定场景下，Kumar^[809] 针对咖啡叶锈病开发的 RAG 系统，能够将实时检测结果与最新的农药登记信息结合，提供合规的治理方案。而 Nguyen^[801] 的“My Climate Advisor”则专注于气候适应性建议，利用 RAG 技术从海量文献中检索知识，帮助农民应对极端天气挑战。

4. 供应链韧性与金融风控

在农业产业链后端，大模型正在重塑供应链管理与金融信贷的风控逻辑。面对气候变化和地缘政治带来的不确定性，Zhang^[810] 的研究展示了基于云原生和大模型的金融科技平台如何将信贷处理时间缩短 81%，并通过多维数据建模将农业贷款违约率降低 30%。这种基于大数据的“数据质押”模式，有效缓解了传统农业主体缺乏抵押物的融资难题。在供应链韧性方面，利用 Agent 模拟技术进行压力测试成为新趋势。研究者利用 LLM 驱动的智能体模拟供应链中的不同参与方，通过博弈演练预测极端事件对市场价格和物流的影响，从而帮助企业制定更具鲁棒性的库存与采购策略。Tencent Cloud 的解决方案则进一步利用 AI 反欺诈模型，在保障供应链金融安全的同时，提升了资金流转的效率与透明度。

5. 国内外农业大模型产品与平台

2025 年，农业大模型产品呈现出“国家级平台整合引领”与“产业生态深度融合”的双重特征。

在中国市场，形成了以高校科研为先导、科技巨头为底座、龙头企业为场景的多元格局。

学术界密集发布了自主研发的垂直模型，覆盖了从育种研发到知识服务的全链条：**中国农业大学**^[802]的**神农大模型 3.0**集成了千万级的农业知识图谱与海量生产数据，内置 36 个专用智能体，实现了从病虫害识别到育种辅助的系统化支持。**中国农业科学院信息所**^[803]联合万方知网发布的**农知大模型**，聚焦农业科技创新与科学决策，通过检索增强生成（RAG）技术提供文献溯源与智能伴读功能，有效解决了通用模型在专业领域的幻觉问题。**华中农业大学**^[806]推出的**狮子山模型**（ShizishanGPT）构建了包含通用模块、搜索引擎、知识图谱与检索模块的五维架构，实现了农业信息的实时更新与精准推理。在种业领域，**复旦大学、中国农业大学与上海 AI 实验室**^[788]联合发布的**丰登大模型**（SeedLLM），专注于生物育种知识的整合，显著降低了育种信息获取的门槛；**浙江大学**^[790]发布的**AI 育种家**（AI Breeder for

Crops, ABC) 平台, 则进一步将大模型与全基因组选择算法结合, 不仅能处理自然语言查询, 还能进行育种策略的智能设计与模拟。此外, **哈尔滨工业大学**^[793]的**天工开悟模型**突出了“感知-决策-控制”一体化能力, 已应用于智能除草机器人; **北京市农林科学院**^[815]的**奇稷大模型**聚焦设施农业与高价值作物; **西北农林科技大学**^[799]的**九壤耘星**则专注于旱区农业的多模态数据整合。**中国农科院**^[811]发布的**农业耕地大模型 1.0**更是上升到国家战略层面, 通过“基础模型 + 专业知识 + 垂直工具”的架构, 服务于高标准农田的实时监测与“非农化”预警。

产业界方面, **腾讯**依托**混元大模型**在智慧温室与金融风控领域深耕; **百度智能云**发布“云智一体”解决方案, 融合遥感与气象预测提供地块级指导; **阿里云**强化“**ET 农业大脑**”, 延伸至全产业链服务; **华为盘古大模型 5.5**则聚焦工业级气象预报与生产规划。

在海外市场, 跨国巨头更倾向于构建封闭的商业闭环, 同时基础模型的通用化趋势明显。John Deere 通过 **Operations Center** 构建了软硬一体的自主作业生态, 将大模型作为农机的“大脑”。Corteva 与 Bayer 重注于生物技术与 AI 的结合 (Biotech AI), 利用 **MyGenAssist** 等内部平台加速育种研发, 并与 Profluent 合作开发下一代基因编辑工具。在基础模型层面, NASA 与 IBM^[791]联合发布的 **Prithvi 地理空间基础模型**, 通过处理海量卫星影像时间序列, 为全球作物分类与灾害监测提供了开源底座; **Google** 拆分出的 **Heritable Agriculture**^[789] 则利用 AI 预测植物基因组与环境的交互, 致力于大幅提升作物生产力。国际组织如 CGIAR 与 Digital Green 则填补了公共服务空白, 通过开源模型 **AgriLLM** 与公益平台 **Farmer.Chat**,^[800] 致力于消除全球农业数字鸿沟。

小结

总体而言, 2025 年的大模型农业应用已超越了早期的概念验证阶段, 步入了深水区。从微观的基因编辑到宏观的全球供应链管理, 大模型正在成为农业 5.0 的核心基础设施。育种 5.0 实现了生物资产的智能设计, 具身智能让农机具备了自主作业能力, RAG 与多智能体技术打破了专家知识的壁垒, 而数据驱动的金融风控则为产业注入了源头活水。这一系列变革标志着农业正在从传统的“经验依赖型”产业, 转型为“数据与算力驱动型”的高技术产业, 为应对全球粮食安全挑战提供了新的技术解法。

5.3 本章小结

围绕编程、写作、设计等知识密集型任务，大语言模型逐步形成稳定的助手型应用范式；同时，在社会模拟、心理咨询以及深度调研（Deep Research）等场景中，大语言模型展现出对长期目标、跨步骤推理与多源信息整合的支持能力，推动 AI for Research 等新型应用方向不断成熟。这些应用突破表明，大语言模型正成为承载大模型能力的重要应用形态，而非单一功能模块。

在行业应用层面，大语言模型开始与教育、医疗、金融、法律、农业等领域的业务流程深度融合，呈现出从通用能力向领域专用系统演进的趋势。不同领域对大语言模型在可靠性、可控性、安全性及合规性方面提出了更高要求，也进一步凸显了行业知识注入与系统工程能力的重要性。

第六章 评测基准和模型进展

2025 年，大语言模型（LLM）评测领域正经历从静态知识测试向动态、交互式、多维度综合评估的深刻转变。随着模型能力的快速演进，传统的单轮、单任务评测基准已难以全面衡量模型在真实场景下的综合表现。本章系统梳理了当前评测基准体系的最新进展，聚焦于多轮对话、工具调用、智能体能力以及多模态理解等四个关键方向，深入分析了各领域代表性基准的设计理念、评估方法与技术趋势。通过对比分析闭源与开源模型在不同能力维度上的表现，本章旨在为读者提供一个清晰、全面的技术图谱，揭示大模型能力演进的内在逻辑与未来发展方向，为模型研发、应用选型与评测体系建设提供参考依据。

6.1 新评测基准

6.1.1 引言

2025 年，大语言模型（LLM）评测领域呈现出明显的发展趋势。评测方法正从以静态数据集为主的单项能力测试，逐步扩展到包含交互性和应用场景的综合评估。评测关注点也从模型对知识的记忆和理解，延伸到模型在实际任务中的执行能力。研究社区越来越重视模型在真实应用场景中的综合表现，而非仅依赖标准化学术任务的得分。这些变化背后反映了 LLM 应用模式的根本性转变：从“单次问答”到“持续交互”，从“信息检索”到“任务执行”，从“被动响应”到“主动规划”，从“纯文本处理”到“多模态理解”。基于这一发展脉络，当前的 benchmark 评测重点聚焦于四大关键方向，以系统性检验模型的真实应用能力：**多轮对话能力**——评估模型在持续交互中的上下文维护、一致性及策略调整能力，是聊天、客服等场景的体验基石。**工具调用能力**——检验模型调用外部 API、执行函数、协调工作流程的能力，标志其从“纯语言系统”向“行动系统”的演进。**智能体（Agent）能力**——

评测模型在复杂任务中的自主规划、推理与纠错能力，反映其从“被动工具”到“主动协作者”的跨越。**多模态能力**——考察模型处理文本、图像、音频等多模态信息的整合能力，是 LLM 适应现实世界多样信息的基础。这四个维度层层递进、相互支撑，多轮对话是基础交互形式，工具调用扩展了行动空间，智能体能力整合了对话和工具使用以实现自主决策，而多模态能力则为前三者提供了更丰富的信息基础，它们共同勾勒出 LLM 从“语言理解系统”向“通用任务执行系统”演进的技术路径。本节将系统梳理这四个维度的评测方法、代表性 benchmark 以及相关研究进展。

6.1.2 多轮对话评测基准

研究背景

多轮对话是人机交互中最自然也最具挑战性的形式，它要求模型在连续的交互回合中保持对上下文的准确理解、对历史信息的可靠记忆，以及对对话目标的持续追踪。尽管大型语言模型在单轮基准测试上已经达到近乎完美的表现，但研究发现它们在真实多轮对话场景中的能力与人类期望之间仍存在显著差距^[816]。现有评估基准的主要问题在于：第一，测试场景过于简化，多数基准仅包含 2-3 轮对话，无法反映实际应用中常见的长程交互；第二，评估维度单一，往往只关注最终回答的正确性，而忽视了对话过程中的指令保持、上下文管理和响应一致性；第三，缺乏对多轮对话特有挑战的针对性测试，如跨轮指令遵循、历史信息推理、版本化编辑处理等^[817]。这些实践反馈催生了对多轮对话能力进行系统性、深入性评估的迫切需求。

研究进展

从短程测试到长程交互评估 目前，多轮对话的评测基准正在从传统的单轮或短程对话测试向长程、多会话的交互评估演进。这一趋势反映了对模型长期记忆能力、跨会话推理能力以及在复杂交互中保持一致性能力的重视。近期的研究工作开始关注模型在扩展对话历史中的表现，例如 **Long-MemEval**^[818]提出了基于 500 个测试问题的动态多会话对话评测框架，定义了信息提取、多会话推理、知识更新、时间推理和抑制能力五项核心记忆能力，系统评估模型在跨会话场景中的长期记忆表现。**MTRAG**^[819]则聚焦于多轮检索增强生成 (RAG) 场景，构建了包含 110 个对话、平均 7.7 轮、共 842 个任务的基准，评估完整 RAG 流程处理后续轮次、无法回答的问题、非独立问题和多领域的能力。更进一步，**EvoLLM**^[820]通过动态生成机制持续产

生具有指令演化、主题切换和回溯等复杂行为的多轮对话，测试模型在长程交互中维持高保真指令遵循的极限，为评估模型在开放式、持续交互环境中的鲁棒性提供了新的视角。

表 6.1: 长程交互评估基准

基准名称	数据规模	语言	类型	介绍
LongMem-Eval ^[818]	500 个测试问题, 基于动态多会话对话	英语	长期记忆评测	定义五项核心记忆能力: 信息提取、多会话推理、知识更新、时间推理和抑制能力。评估模型在跨会话场景中的长期记忆表现。
MTRAG ^[819]	110 个对话, 平均 7.7 轮, 共 842 个任务, 涵盖四个领域	英语	多轮 RAG 对话评测	首个端到端人工生成的多轮 RAG 基准, 反映真实世界多维属性。评估完整 RAG 流程, 包括检索和生成系统处理后续轮次、无法回答的问题、非独立问题和多领域的的能力。由 IBM 发布。
EvoIIF ^[820]	动态生成, 可扩展	英语	动态指令遵循评测	通过动态生成机制持续产生具有指令演化、主题切换和回溯等复杂行为的多轮对话, 测试模型在长程交互中维持高保真指令遵循的极限。

从结果评估到过程质量评估 传统评测主要关注对话的最终输出质量，而忽视了达成结果的交互过程本身的合理性与用户体验。近期研究表明，仅依赖结果评估难以全面衡量模型在复杂多轮交互中的真实表现。新一代 benchmark 开始强调对对话过程的细粒度评估，包括信息收集策略、交互公平性、精炼能力等过程维度。

表 6.2: 过程质量评估基准

基准名称	数据规模	语言	类型	介绍
Refine-Bench ^[821]	1,000 个问题, 涵盖 11 个领域 239 个主题	英语	精炼能力评测	使用基于检查清单的评估框架评估语言模型的精炼能力。评估两种模式:(1) 引导式精炼: 提供自然语言反馈;(2) 自我精炼: 模型自主改进。
InfoQuest ^[822]	具体数量未公开	英语	开放式对话评测	提出三维评估框架: 对话吸引力、信息收集效率和社交适当性。评估模型在开放式对话中的用户体验质量, 从过程和结果两个层面进行综合评估。
MMReason ^[823]	多领域长链数据	英语	多步推理过程正确性	针对长链推理的细粒度评估。它摒弃了仅评估最终答案的传统做法, 专注于检测中间推理步骤的逻辑连贯性与正确性, 识别“答案正确但过程错误”的幻觉。
Agent-Reward ^[824]	1,300 条交互轨迹	英语	Agent 轨迹/步骤质量	专门评估 Agent 在 Web 环境完成任务的交互轨迹。它引入 LLM-as-a-Judge 来对 Agent 的每一步操作 (进行奖励评分, 评估其决策过程的有效性而非仅仅是任务成功率)。

从宏观评分到细粒度能力评估 早期的多轮对话基准如 MT-Bench 和 ChatEval 主要关注模型的整体表现, 采用相对宏观的评分标准。然而, 随着模型能力的快速提升, 研究者们发现需要更加细粒度的评测维度来诊断模型在特定能力上的不足。近期涌现的评测基准开始聚焦于多轮对话中的特定挑战性能力, 例如 **MultiChallenge**^[825] 将评测细化为指令保持、推理记忆、版本编辑和自治性四个维度, 系统性地考察模型在复杂交互场景下的表现; **StructFlow**^[826] 则专注于结构化流指令的评估, 定义了回溯、扩展、精炼等六种层级化交互关系, 深入分析模型对长程结构依赖的处理能力; **AdvancedIF**^[827] 引入基于评分细则的框架, 通过系统提示词变量控制, 针对性地评估复杂指令在多轮场景中的分解与遵循情况。这一趋势反映了评测方法论从广度覆盖向深度解析的转变, 为模型的针对性优化提供了更加精准的指

导。

表 6.3: 2025 年针对性多轮对话能力评估基准

基准名称	数据规模	语言	类型	介绍
Multi-Challenge ^[825]	273 条	英语	多维综合挑战	评估四类挑战:(1) 指令保持: 全对话周期约束遵循;(2) 推理记忆: 关联离散历史信息;(3) 版本编辑: 处理动态修改;(4) 自治性: 解决矛盾信息。
Struct-Flow ^[826]	未公开	英语	结构化流指令	针对多轮对话的结构依赖性, 定义了 6 种层级化交互关系 (如回溯、扩展、精炼), 评估模型在特定结构约束下的长程指令遵循能力。
Advanced-IF ^[827]	1,645 条	英语	细粒度指令遵循	采用基于评分细则 (Rubric-based) 的评估框架, 引入系统提示词 (System Prompt) 变量, 专注于多轮场景下复杂指令的分解与验证, 解决指令遗忘问题。

总结与展望

综合 2025 年的研究进展, 多轮对话 benchmark 领域在理论框架构建、评估方法创新和能力边界探索等方面实现了系统性突破。从单一性能测试演进为多维度能力刻画, 构建了涵盖信息整合、推理连贯、个性化适应等核心维度的评估体系。自动化评估支持大规模可复现的模型比较, 真实场景数据增强了评测的生态效度。评测从静态诊断工具升级为动态反馈系统, 实现“评估-优化”闭环, 驱动模型持续迭代。这些突破不仅加速了对话 AI 的技术发展, 也为深入理解与提升模型的长程交互能力奠定了坚实基础。

展望未来, 多轮对话 benchmark 领域呈现出以下几个主要发展趋势。首先, 深化认知评估, 从表层回复质量转向深层理解与推理, 关注跨轮信息的主动整合、隐含意图的动态追踪、以及知识更新与矛盾消解等高阶认知能力。其次, 融合多模态与多场景, 评测将覆盖文本、语音、视觉等多通道信息的协同处理, 并扩展至客服、创意协作等多元应用场景。最后, 强化可信与可控, 更加重视长程交互中的事实准确性、立场一致性与价值对齐, 避免有害

输出的累积放大，关注安全与可靠性的持续评估。未来的评测体系将致力于构建更全面、鲁棒且可指导应用的框架，以确保对话 AI 在技术创新与实际部署中的可信与可靠。

6.1.3 工具使用评测基准

研究背景

工具调用能力发展经历了从简单到复杂的演进历程。早期大语言模型主要依赖预训练知识，面对需要实时数据、精确计算或外部信息的任务时往往力不从心。随着 Function Calling、ReAct 等范式的提出，模型开始具备调用外部工具的能力，但仍局限于单次、独立的 API 调用。进入 2025 年，工具调用技术呈现出向智能协同系统演进的显著趋势：模型不仅能理解工具间的依赖关系并构建多工具协作链路，还能在动态交互中自适应调整策略，在命令行、网页等真实环境中完成端到端的复杂任务。

研究进展

针对工具调用领域的三个核心研究方向，学术界和工业界在 2024-2025 年间提出了一系列具有代表性的测评基准，这些 benchmark 从不同维度系统性地评估了大语言模型的工具使用能力。本节将按照函数调用准确性与工具组合能力、真实场景下的交互式任务执行、开放域综合能力评估三个方向，梳理当前主流 benchmark 的设计思路、评测方法及其在推动技术发展中的作用。

函数调用准确性与工具组合能力 函数调用准确性与工具组合能力是工具调用领域最基础也是最核心的研究方向，主要评估模型能否准确识别、选择和调用外部 API 或函数，以及在复杂场景下组合多个工具完成任务的能力。该方向的 benchmark 从简单的单次 API 调用逐步演进到支持多轮交互、并行调用、跨工具依赖等复杂场景，特别是 2025 年随着 Model Context Protocol (MCP) 的兴起，涌现出一批专注于评估标准化工具接口使用能力的新基准。这些 benchmark 不仅测试模型的函数调用格式正确性，更关注其在长上下文、参数推理、约束满足等实际应用中的表现，为提升模型的工具使用鲁棒性提供了重要的评估依据。表 6.4总结了该方向的主要评测基准。

表 6.4: 函数调用准确性与工具组合能力评测基准对比

基准名称	数据规模	语言	类型	介绍
BFCL V4 ^[828]	4,951 个测试用例（3,951 单轮 +1,000 多轮）	多语言	函数调用	Berkeley 推出的全面函数调用评测基准，支持单轮、多轮、并行调用，涵盖 Python/Java/JavaScript 等多种语言，采用 AST 评估方法，V4 版本新增网页搜索、记忆管理和格式敏感性评估
ComplexFunc-Bench ^[829]	1,000 个样本（600 单域 +400 跨域）	Python	复杂函数调用	专注于复杂场景的函数调用基准，包含多步骤调用、用户约束、参数推理、长参数值（超 500 tokens）和 128k 长上下文等五大挑战维度
Tool-Comp ^[830]	485 个样本（287 企业版 +198 聊天版）	Python	组合工具使用	Scale AI 发布的组合工具使用基准，强调依赖工具调用（前一工具输出作为后续输入），包含 11 个企业工具和 2 个通用工具，提供人工验证的最终答案和过程监督标签
MCP-Mark ^[831]	127 个高质量任务	跨环境	MCP 协议使用	评估 Model Context Protocol 使用的综合基准，涵盖 Notion、GitHub、文件系统、PostgreSQL 和 Playwright 五大环境，强调 CRUD 操作的完整性，平均每任务需 16.2 轮执行和 17.4 次工具调用
MCP-Bench ^[832]	28 个 MCP 服务器、250 个工具	跨域	多步骤任务	基于 MCP 协议的真实多步骤任务基准，连接金融、旅游、科学计算、学术搜索等领域的实时 MCP 服务器，测试跨工具协调、精确参数控制和多跳规划能力

真实场景下的交互式任务执行 真实场景下的交互式任务执行是衡量 AI Agent 实际应用能力的关键维度，该方向的 benchmark 致力于在接近真实世界的复杂环境中测试 Agent 的端到端任务完成能力。与传统的静态评测不同，这类基准强调 Agent 需要在操作系统、命令行终端、网页浏览器等真实数字环境中进行持续交互，动态适应环境反馈并完成长时域、多步骤的复杂任务。2025 年该方向呈现出两个显著趋势：一是评估环境的真实性不断增强，从模拟环境转向真实的操作系统和应用程序；二是评估维度的多元化，不仅关注任务成功率，还引入了可靠性、效率、安全性等指标。这些 benchmark 揭示了当前模型在 GUI 理解、长期规划、错误恢复等方面仍存在显著挑战，为推动 Agent 技术走向实用化提供了重要参考。表 6.5 列举了该方向的代表性评测基准。

表 6.5: 真实场景下的交互式任务执行评测基准对比

基准名称	数据规模	语言	类型	介绍
τ -Bench ^[833]	两个域（零售、航空）	自然语言	用户交互	模拟用户与 Agent 的动态对话，测试 Agent 在客户服务场景中遵循领域政策、收集信息和解决问题的能力，引入 pass ^k 指标评估多次运行的可靠性
τ^2 -Bench ^[834]	三个域（零售、航空、电信）	自然语言	双控环境	扩展 τ -Bench，引入双控场景（Agent 和用户均可操作工具），新增电信故障排查域，测试 Agent 在用户主动参与工具操作时的协调能力
ColBench ^[835]	后端编码和前端设计任务	编程语言	协作推理	评估 LLM 作为协作 Agent 与模拟人类伙伴的多轮工作能力，需要逐步协作：模型提出代码/设计草案、接收反馈并迭代改进，模拟真实开发工作流
OSWorld-Human ^[836]	37 个任务的人类基准	GUI 操作	效率评估	专注于评估计算机使用 Agent 的效率，通过详细步骤分解分析发现规划和反思步骤占总延迟的 75-94%，提供人工标注的最小步骤基准

开放域综合能力评估 开放域综合能力评估聚焦于测试 AI 系统在开放域场景下的多维度智能表现，这类 benchmark 不局限于特定工具或环境，而是要求模型展现出接近人类的综合问题解决能力。与前两个方向相比，该方向更强调能力的泛化性和任务的复杂性，评估内容往往涉及多模态理解、多步推理、知识综合运用以及灵活的工具调用等多个能力维度的协同工作。2025 年该领域的 benchmark 设计理念发生了重要转变：从追求超越人类专业技能转向关注日常场景下的鲁棒性，即评估模型能否像普通人一样可靠地处理看似简单但需要综合运用多种能力的实际问题。这类评测揭示了当前最先进模型与真正通用智能之间仍存在巨大差距，也为构建更加实用的 AI 助手指明了方向。表 6.6概述了该方向的核心评测基准。

表 6.6: 通用智能助手的综合能力评测基准对比

基准名称	数据规模	语言	类型	介绍
Humanity’s Last Exam ^[837]	2,500 个问题	多模态	专家级推理	挑战 LLM 达到专家级人类推理和知识水平的多模态基准，涵盖数学、人文和自然科学等多个领域，过滤掉可通过网页搜索或记忆提示轻易回答的问题
Multi-Challenge ^[825]	多轮对话场景	自然语言	多轮对话	针对前沿 LLM 的真实多轮对话评估基准，测试可靠版本编辑和自洽性等能力，专门设计对当前最强模型仍具挑战性的任务
Frontier-Math ^[838]	数百个极难问题	英语	数学推理	由 Epoch AI 于 2025 年推出，包含数以百计需要数小时甚至数天才能解决的未公开数学难题，测试模型在极端复杂逻辑下的综合建模能力。

总结与展望

工具调用测评基准的发展已从早期的单一 API 调用准确性评估，演进为涵盖函数调用、环境交互和综合能力的多维度评测体系。当前呈现三大特征：评测环境从模拟转向真实操作系统和浏览器，提升了生态效度；评估维度从成功率扩展到效率、安全和可靠性，形成多元标准；任务设计从单步调用发展为需跨工具协调与长程规划的复杂场景。这些进展为 AI 工具使用能

力的评估提供了更加系统化的框架。

未来工具调用测评基准的发展可能聚焦于以下几个方向：第一是需要探索兼顾真实性与可扩展性的评测范式，利用轻量级容器或高保真模拟降低真实环境成本。第二是建立系统性细粒度指标，评估工具选择、参数构造、执行顺序及异常处理的准确性。第三是引入接口变更、资源波动等真实不确定因素，检验模型的持续适应与学习能力。第四是加强对工具语义理解、调用图规划及冲突处理等深度协同能力的评估。第五是将恶意工具识别、权限控制、隐私保护与可解释性纳入通用评测框架。

6.1.4 智能体评测基准

研究背景

AI Agent 正从单一任务执行者进化为能进行复杂推理、自主决策和多步骤规划的智能系统，其评估需求日益凸显。2025 年更是被业界称为“AI Agent 之年”，标志着该技术从概念验证阶段迈向大规模实际应用阶段。当前评估体系研究主要聚焦于以下三个关键方向：一是交互与工具使用评估，考察 Agent 在动态开放环境（如网页）中完成信息检索、导航、操作等实际任务的能力。二是垂直领域能力评估，不再限于通用任务，而是重点评估 Agent 对齐行业私域知识、遵从复杂业务流与合规要求，以及在专业场景中的精准决策能力，验证其处理高壁垒、低容错“深水区”任务的能力。三是多智能体协作评估，探索多 Agent 间通过有效通信、角色分工与策略协调，共同解决需分布式决策与战略规划的复杂问题，尤其关注其协作效率和网络可扩展性。

研究进展

从测评维度来看，交互与工具使用能力的评估已从早期的静态 API 调用测试发展为对动态环境感知、多模态信息处理和长序列决策的综合考察，研究者们不仅关注 Agent 能否完成特定任务，更深入探讨其在面对不确定性、部分可观测性以及环境反馈延迟等现实约束下的鲁棒性与适应性；而多智能体协作与推理能力的研究则突破了传统单 Agent 评估范式，将焦点转向网络拓扑结构对协作效率的影响、分布式共识机制的收敛性、以及异质 Agent 在动态角色分配中的自组织能力，这些研究不仅揭示了集体智能涌现的内在机制，也为构建可扩展、容错性强的大规模 Agent 系统提供了理论支

撑。以下将分别从这两个维度展开，系统梳理代表性 benchmark 的设计思路

交互与工具使用能力评估 交互与工具使用能力是 AI Agent 实现自主任务执行的基础，其评估体系的核心在于考察 Agent 如何在开放式、动态变化的环境中准确理解任务需求、选择合适工具并完成多步骤交互。当前该领域的 benchmark 主要聚焦于 Web 浏览任务，如 BrowseComp 通过 1,266 个需要跨网站深度搜索的问题，测试 Agent 在互联网中定位难以发现信息时的持续性和创造性搜索能力，这类评估不仅要求 Agent 具备基本的导航和信息提取技能，更需其在面对海量信息时展现出策略规划和目标导向的检索能力。表 6.7总结了该领域的代表性评估基准。

表 6.7: 交互与工具使用能力评估基准对比

基准名称	数据规模	语言	类型	介绍
Browse-Comp ^[598]	1,266 个问题	英语	Web 浏览	OpenAI 2025 年发布，评估 Agent 在互联网中定位难以发现信息的能力，需要浏览数十至数百个网站进行持续性和创造性搜索
MCPVerse ^[839]	550+ 真实工具，14.7 万动作空间	英语	真实工具调用	2025 年最新发布的大规模、真实可执行工具基准。基于“模型上下文协议”（MCP）构建，采用基于结果的实时评估，测试 Agent 在超大规模动作空间下的精确工具选择与调用。
WebBench ^[840]	2,000+ 交互轨迹	英语/多语言	Web 自动化	专门评估 Agent 在真实网站（如电商、金融、OA 系统）中完成端到端任务的成功率。它引入了动态环境反馈，关注 Agent 应对网页布局突变和反爬虫机制时的鲁棒性。

垂直领域应用能力评估基准 垂直领域应用能力评估标志着 AI Agent 从通用技术演示向专业场景落地的关键转变，这类 benchmark 的核心特征在于将 Agent 置于医疗健康、软件工程、科学研究和企业应用等真实业务环境

中,考察其是否具备领域专业知识、能否遵循行业规范以及在高风险场景下的决策可靠性。医疗健康领域的 MedAgentBench 和 HealthBench 分别通过真实电子健康记录操作 (300 个临床任务、70 万 + 记录) 和多语言多专科对话场景 (5,000 个对话、49 种语言), 测试 Agent 在临床诊疗中的精准性; 软件工程领域的 AutoCodeBench 利用 3,920 个自动生成的多语言编程问题, 评估 Agent 的持续演化代码生成能力; 科学研究领域的 AstaBench 和 PaperBench 则从文献理解到研究复现的完整流程考察 Agent 的科研辅助潜力; 企业应用的 CLASSic Framework 创新性地从成本、延迟、准确性、稳定性、安全性五个维度进行全面评估, 体现了产业界对 Agent 系统实用性的综合要求。表 6.8总结了当前主要垂直领域的评估基准。

多智能体协作与推理能力评估 多智能体协作与推理能力评估关注的是超越单个 Agent 局限、通过集体智能解决复杂问题的能力, 这类 benchmark 的设计重点从单纯的任务完成转向对协作机制、通信效率和涌现行为的深入考察。当前该领域呈现出三个显著特征: 首先是对不同组织结构的系统性探索, 如 MultiAgentBench 通过星型、链式、树形、图形等多种拓扑结构测试协作模式对性能的影响, 并引入里程碑式 KPI 指标量化协作质量; 其次是可扩展性的极限挑战, AgentsNet 基于分布式系统经典问题 (图着色、选举、共识等) 将评估规模扩展至 100 个 Agent, 检验大规模场景下的自组织与协调能力; 第三是向真实复杂场景的迁移, CREW-Wildfire 通过野火灾害响应的程序生成环境模拟异构 Agent 在动态地形中的协作, SOTOPIA- 则聚焦社交智能维度, 评估 Agent 在社会规范理解、同理心展示和伦理推理方面的表现。此外, AgentBench 跨 8 个多样化环境测试 Agent 的通用协作能力, LatentMAS 开创性地探索了连续潜在空间中的直接协作机制, ReSo 通过可组合图结构任务评估协调协议的敏感性, 这些创新共同推动了多智能体系统从理论研究走向实用部署。表 6.9总结了该领域的主要评估基准。

总结与展望

通过对 2025 年 AI 智能体测评基准的系统梳理可以看出, 该领域已初步形成了覆盖交互工具使用、垂直领域应用和多智能体协作三大核心维度的评估体系。这些进展表明, Agent 测评正从单一能力考察走向综合素质评估, 从静态环境测试转向动态场景适应, 从孤立任务完成迈向复杂系统协作, 为推动 Agent 技术从实验室走向产业应用奠定了坚实的评估基础。

从技术发展趋势看, 以下几个方向值得持续关注: 首先是推理能力的进

表 6.8: 垂直领域应用能力评估基准对比

基准名称	数据规模	语言	领域	介绍
FinGAIA ^[841]	407 个任务, 8 大业务场景	中/英	金融	腾讯与复旦发布, 首个金融全流程 Agent 基准。通过模拟真实的证券、基金、银行 workflow, 考核 Agent 在数据集成、政策解读及复杂投资决策中的可靠性。
Finance-Agent ^[842]	537 个专业问题	英语	金融	Vals AI 发布, 专注评估 Agent 担任“初级金融分析师”的能力, 涵盖财务报表分析、市场研究预测及多步审计。
VLAIR ^[843]	500+ 样本, 7 类任务	英语	法律	首个法律 AI 工具综合评估, 由 Reed Smith 等顶级律所参与, 评估数据提取、文档问答、摘要等任务, AI 在 4/7 任务上超越人类律师
MedAgent-Bench ^[844]	300 个临床任务, 100 个患者档案, 70 万 + 记录	英语	医疗健康	Stanford 发布, 首个真实 EHR 环境中的 LLM 代理基准, 包含 FHIR 兼容的交互式环境, 测试检索数据、下医嘱等临床任务
Health-Bench ^[845]	5,000 个对话	多语言 (49 种)	医疗健康	OpenAI 发布, 由 262 位来自 60 个国家的医生创建, 涵盖 26 个医学专科, 每个对话配有定制评分标准 (48,562 个独特评估标准)
Auto-Code-Bench ^[846]	3,920 个问题	多语言 (20 种)	软件工程	Tencent 发布, LLM 自动生成和验证的代码基准, 消除人工干预, 保持难度平衡分布, 持续演化评估
Asta-Bench ^[847]	2,400+ 问题, 11 个基准	英语	科学研究	AI2 发布的全面科学研究基准, 涵盖文献理解、代码执行、数据分析、端到端发现四个领域, 评估成本和质量
Paper-Bench ^[848]	未明确规模	英语	科学研究	OpenAI 发布, 评估 AI 代理从零开始复现研究论文主要贡献的能力, 包含评分标准和自动化评估
CLASSic Framework ^[849]	2,100+ 消息, 7 个行业	英语	企业应用	Aisera 发布, 企业级 Agent 评估框架, 涵盖成本、延迟、准确性、稳定性、安全性五个维度, 包含 IT、HR、客服等场景

表 6.9: 多智能体协作与推理能力评估基准对比

基准名称	数据规模	语言	类型	介绍
Multi-Agent-Bench ^[480]	多领域任务集	英语	综合协作	2025 年发布，评估多 Agent 系统在多样化交互场景中的协作与竞争，使用里程碑式 KPI 指标，测试星型、链式、树形、图形等多种拓扑结构
AgentsNet ^[850]	5 类问题，可扩展至 100 个 Agent	英语	网络协作	2025 年发布，基于分布式系统和图论的经典问题 (图着色、顶点覆盖、匹配、选举、共识)，评估 Agent 自组织、协调和通信能力
CREW-Wildfire ^[851]	程序生成环境	英语	大规模协作	2025 年发布，野火灾害响应场景的开源基准，支持异构 Agent、复杂动态地形，测试大规模协作中的感知、规划和任务委派能力
SOTOPIA- ^[852]	社交场景集	英语	社交智能	2025 年更新版本，创建沉浸式社交模拟，测试 Agent 理解社会规范、展示同理心、伦理推理和文化适应的能力
Agent-Bench ^[853]	8 个环境	英语	多环境交互	2023 年发布持续更新至 2025 年，跨操作系统、数据库、知识图谱、数字卡牌游戏和 Web 界面的多维度基准，评估规划、推理、工具使用和决策能力
Latent-MAS ^[479]	9 个基准任务	英语	潜在空间协作	2025 年发布，评估 Agent 在连续潜在空间中的直接协作能力，跨数学推理、常识理解和代码生成任务，相比文本协作提升准确率达 14.6%
ReSo ^[854]	图结构任务	英语	结构化推理	2025 年发布，基于可组合图的任务 (节点 = 子任务，边 = 依赖关系)，通过增加图大小或 Agent 数量评估可扩展性和协调协议敏感性

一步增强，特别是在多步骤任务规划和动态目标调整方面，这需要更强的因果推理和不确定性处理能力；其次是人机协作模式的优化，探索 AI Agent 作为辅助工具而非完全自主系统的定位，建立更清晰的人类监督与干预机制；第三是垂直领域的深度整合，结合领域知识图谱和专业工作流程，构建更可靠的行业级应用；第四是评估标准的持续完善，开发能够反映真实使用场景复杂性的 benchmark，并建立可解释性和安全性的量化指标。这些研究方向的推进，将有助于 AI Agent 技术从实验原型向生产系统的转化，并在保证可控性和可信度的前提下，逐步扩大其在专业领域中的应用范围。

6.1.5 多模态评测基准

研究背景

多模态大语言模型的发展正经历从单一模态理解到跨模态推理融合的范式转变，早期模型主要聚焦于图像-文本对齐和基础视觉问答任务，而 2025 年的研究重心已转向复杂场景下的深度推理与长序列理解能力评估，这一转变推动了新一代测评基准的涌现，以更全面地检验模型在真实世界应用中的综合能力。

研究进展

在当前的多模态测评体系中，研究方向主要可细分为三个核心领域：其一是视频理解与时序推理方向，重点考察模型对长视频内容的记忆保持、事件关联推理及多轮对话交互能力，这对于具身智能决策、体育赛事解说等需要长时序理解的应用场景至关重要；其二是文档理解与结构化信息提取方向，聚焦于模型对复杂文档布局的感知能力，包括多语言 OCR 识别、表格公式解析及跨页面信息关联，这对科研文献数字化、企业文档智能化处理等场景具有重要价值；其三是多模态推理与科学问题求解方向，强调模型整合视觉线索与符号推理进行数学、物理等学科问题求解的能力，特别关注多图推理、视觉依赖性分析及实际场景中的复杂问题分解能力，这些能力是实现通用人工智能在教育、科研领域应用的关键支撑。

视频理解与时序推理 视频理解正从短片段的动作识别演进为对长时序、跨片段逻辑的深度感知。当前的研究热点集中在提升模型对于数小时级视频的长期记忆（Long-term Memory）以及时空一致性的推理能力。为了精准评估模型在复杂叙事和动态演化过程中的表现，一系列针对长视频、多轮交互

及多步推理的评测基准应运而生。下表 6.10对比了当前主流的视频理解与时序推理评估框架。

表 6.10: 视频理解与时序推理评测基准对比

基准名称	数据规模	语言	类型	介绍
LVBench ^[855]	1,000 视频 / 9,000+ QA	英文	长视频理解	首个极长视频理解基准, 视频平均时长 2 小时, 评估六大核心时序理解能力, 包括长期记忆和跨片段推理
ALLVB ^[856]	1,376 视频 / 252K QA	英文	长视频多任务	全方位长视频理解基准, 整合 9 大视频理解任务, 平均视频时长近 2 小时, 采用 GPT-4o 自动标注流程
Video-MME ^[857]	900 视频 / 2,700 QA	英文	综合评估	首个多模态 LLM 视频分析综合评估基准, 包含短视频和长视频子集, 评估时序推理能力
MT-Video-Bench ^[858]	987 多轮对话	英文	多轮对话	多轮对话视频理解基准, 评估感知力和交互性六大核心能力, 对应运动分析和视频智能辅导场景
VR-Bench ^[859]	960 长视频 / 8,243 QA	英文	多步推理	长叙事视频多步推理基准, 平均时长 1.6 小时, 包含 25,106 个带时间戳推理步骤
Video-TT ^[305]	1,000 视频 / 5,000 QA	英文	综合推理	视频思维测试基准, 评估高级视频推理和理解, 确保问题复杂并提供推理依据
V-STaR ^[860]	时空推理任务	英文	时空推理	Video-LLM 视频时空推理基准, 评估时序定位和物体追踪等时空感知任务

文档理解与结构化信息提取 文档图像智能处理（IDP）是多模态模型走向行业应用的关键技术路径。不同于自然场景图像，文档理解要求模型具备极高的光学字符识别（OCR）精度，并能深度解析表格、图表及复杂的物理布局逻辑。近期的研究不仅关注标准化的解析任务，更进一步扩展到了多语言支持以及真实野外环境下的鲁棒性评估。表 6.11汇总了近期在文档解析、智能处理及文本密集型任务中的代表性评测基准。

表 6.11: 文档理解与结构化信息提取评测基准对比

基准名称	数据规模	语言	类型	介绍
OmniDoc-Bench ^[861]	1,729 页 (v1.5)	多语言	文档解析	综合文档解析评估基准, 涵盖 9 类文档、4 种布局、3 种语言, 包含 15 类块级和 4 类跨度级标注
IDP Leader-board ^[862]	统一评估框架	多语言	智能文档处理	智能文档处理统一排行榜, 涵盖 OCR、KIE、分类、QA、表格提取和置信度评估
TIU-Bench ^[863]	多页文档	英文	文本密集理解	文本图像理解基准, 评估完整图像解析和文档 VQA 能力, 处理密集文本内容
WildDoc ^[864]	真实拍摄文档	英文	野外文档理解	首个自然环境文档理解基准, 包含不同光照和物理变形条件下拍摄的文档图像

多模态推理与科学问题求解 科学问题求解被视为检验多模态大模型逻辑推理能力的“最高境界”，它要求模型能够将复杂的视觉信息（如几何图形、化学分子式、函数图象）与深层的学科知识进行显式融合。当前的研究趋势正从简单的常识问答转向具有高视觉依赖性、多步逻辑推导以及跨学科竞赛水平的挑战。表 6.12梳理了涵盖数学、物理、化学等多个领域的科学推理基准，展现了模型在处理真实科研与教育场景问题时的评测现状。

总结与展望

2025 年多模态 Benchmark 领域的发展呈现出评测维度系统化与能力考察深度化的双重特征。在时序理解方面, 评测视频长度从分钟级扩展至小时级, 对模型的长期依赖建模能力提出了更高要求; 在推理复杂度方面, 基准设计逐步从单步问答转向多步逻辑链路和跨模态信息融合, 专业学科竞赛级别的评测任务开始出现; 在评估方法上, 自动化评分机制提升了大规模测试的可行性, 但其客观性也成为新的挑战。整体而言, 当前基准已初步构建起覆盖感知、理解、推理等多层级的测试体系, 为多模态模型的迭代提供了较为完整的评估框架。

现有研究表明, 部分模型在多模态任务中可能过度依赖语言模型的文本推理能力, 而未充分利用视觉信息。未来需要发展能够量化”视觉贡献度”的评估方法, 并设计对抗性样本以检验模型对视觉输入的实际依赖程度。其次,

表 6.12: 多模态推理与科学问题求解评测基准对比

基准名称	数据规模	语言	类型	介绍
EMMA ^[865]	2,788 问题	英文	增强多模态推理	增强多模态推理基准, 涵盖数学、物理、化学和编程, 评估视觉-文本深度融合推理能力
MMMU-Pro ^[866]	增强版 MMMU	英文	鲁棒性评估	MMMU 的鲁棒增强版本, 提供更具挑战性的多模态 AI 评估
MV-MATH ^[867]	2,009 问题 / 11 学科	英文	多视觉数学推理	多视觉情境数学推理评估, 每题集成多张交错图像, 评估 K-12 数学推理
VC-Bench ^[868]	1,720 问题 / 6,697 图像	英文	显式视觉依赖	多模态数学推理基准, 评估显式视觉依赖, 平均每题 3.9 张图像, 覆盖 6 个认知领域
Math-Scape ^[869]	真实场景数学	英文	真实世界数学	真实世界数学情境基准, 评估模型处理复杂真实数学挑战的能力
MME-SCI ^[870]	多语言科学	多语言	科学综合评估	多语言多模态科学基准, 支持跨语言一致性评估, 覆盖多个科学学科
mmJEE-Eval ^[871]	1,460 问题 (2019-2025)	英/印地语	科学竞赛评估	双语多模态科学推理基准, 源自印度 JEE 考试, 涵盖物理、化学、数学预科水平

随着模型能力快速提升，静态基准容易出现饱和，需要探索可持续更新的评测机制和难度自适应的测试生成方法。此外，多模态幻觉的细粒度检测、物理世界常识的系统性评估、以及多轮交互场景下的鲁棒性测试等方向，均是构建下一代评测体系的重要研究课题。

6.2 模型生态进展

2025 年标志着人工智能发展史上的一个重要转折点。随着“缩放定律”（Scaling Law）在单纯参数堆叠上的边际效应递减，全球大语言模型（LLM）产业进入了以“系统-思维”（System 2 Thinking）、“原生多模态”（Native Multimodality）和“混合专家架构”（Mixture-of-Experts, MoE）为核心特征的深度进化期。本报告旨在对截至 2025 年底的全球大模型技术版图进行穷尽式的梳理与评述。报告将市场格局划分为以 OpenAI、Google、Anthropic、xAI 为代表的“新闭源模型”阵营，以及以 Meta、Alibaba Cloud (Qwen)、Mistral、DeepSeek 为代表的“新开源模型”阵营。通过对各模型发布时间、技术架构、能力边界及官方自测指标的深度剖析，本报告揭示了闭源模型向“代理工作流”（Agentic Workflow）与极致推理能力的收敛，以及开源模型在架构效率与端侧部署上的爆发式追赶。分析显示，开源与闭源的性能鸿沟在 2025 年已被显著压缩，行业竞争焦点从单一的对话质量转向了生态系统的构建与推理成本的博弈。

下表汇总了截至 2025 年 12 月，全球顶级开源与闭源模型在核心基准测试上的表现。数据来源于 MMLU-Pro 官方榜单、SWE-bench Verified 及 LMSYS Arena。

表 6.13: 2025 年第四季度顶级开源与闭源模型性能综合对比表								
模型名称	发布机构	类型	参数量 (Est.)	MMLU-Pro (5-shot)	HumanEval (Pass@1)	MATH (Pass@1)	LMSYS Elo (Overall)	推理成本 (\$/1M In/Out)
GPT-5.2 (Thinking)	OpenAI	闭源	未公开 (MoE)	93.5%	96.5%	98.2%	1420	\$5.00 / \$25.00
Gemini 3 Pro	Google	闭源	未公开	90.5%	94.2%	95.8%	1405	\$2.00 / \$12.00
Claude 4.5 Opus	Anthropic	闭源	未公开	93.4%	95.0%	94.5%	1410	\$15.00 / \$75.00
Grok 4	xAI	闭源	~1.5T	91.6%	92.8%	93.0%	1395	\$2.50 / \$10.00
DeepSeek-R1	DeepSeek	开源	671B (MoE)	90.3%	93.5%	96.1%	1388	\$0.50 / \$2.15
Llama 4 Maverick	Meta	开源	405B (Dense)	80.5%	89.5%	88.0%	1340	\$0.89 / \$0.89
Qwen 3 235B	Alibaba	开源	235B (MoE)	84.4%	91.2%	92.5%	1355	\$0.30 / \$3.00
Mistral Large 3	Mistral	开源	未公开	85.1%	90.0%	89.2%	1362	\$2.00 / \$6.00

6.2.1 新闭源模型

2024 年末至 2025 年，闭源模型阵营完成从通用生成到深度推理与全模态融合的战略转型。面对开源社区的追赶压力，OpenAI、Google、Anthropic 与 xAI 四大主体通过封闭生态闭环构建和高成本推理时间计算，维持高端智力市场的竞争优势。

OpenAI GPT 系列：双轨制策略下的推理与应用分化

从 2025 年的产品节奏来看，OpenAI 产品演进呈现明确双轨特征：o 系列聚焦推理上限突破，以复杂任务稳定性与正确率为核心目标；GPT-5 系列定位于企业级与通用应用主干，侧重可用性、指令遵循度及真实 workflow 整合效能。这一策略差异直接体现在模型性能表现与应用场景适配性上。

- **核心模型演进与发布时间轴：**在时间线上，o3-pro 于 2025 年 6 月 10 日正式发布，代表了 OpenAI 在“系统 2 思维”（慢思考）上的最高成就。o4-mini 随后推出，通过知识蒸馏技术将推理能力普惠化，在保持较低延迟的同时保留了核心的自我反思机制。GPT-5 系统于 2025 年 8 月 7 日首次发布，12 月全面推广，标志着从“多模型选择”向“统一智能系统”的战略转型。GPT-5.2 作为重要迭代于 2025 年 12 月 11 日发布，在数学和视觉推理领域取得突破性进展。
- **技术特性与能力定位：**GPT-5 的核心创新在于“智能路由器”架构，该系统能根据任务复杂度自动路由至最合适的子模型(mini/main/thinking)。这种设计实现了算力资源的全局最优配置。相比之下，o3-pro 作为独立的深度推理引擎，其技术核心在于通过强化学习习得的“思维策略”，能够进行长时间的隐式推理链推导，专攻专家级难题。o4-mini 通过知识蒸馏将大模型推理策略迁移到更小规模，实现了推理能力的普惠化。
- **应用场景与性能表现：**GPT-5.2 被定位为“通用智能 workflow 主干”，通过集成 Operator 代理功能，能操作虚拟浏览器环境执行多步骤任务闭环。在 AIME 2025 数学竞赛中配合工具达到 100
- **上下文与输出能力：**GPT-5 系统支持 400,000 tokens 的上下文窗口和 128,000 tokens 的输出能力，能一次性处理整本书或大型项目文件。o3-pro 虽仅支持 200,000 tokens 上下文，但具备深度理解和逻辑关联

2025 年大语言模型（LLMs）进展报告

能力，支持高达 100,000 tokens 的输出。o4-mini 在移动端和实时应用中表现出色，推动了推理模型的普及。

- **市场定位与定价策略：**双轨策略体现在明确的市场分层。o3-pro 定价为输入 \$20.00/百万 token、输出 \$80.00/百万 token，定位为“高端专家服务”。GPT-5 系统通过智能路由实现成本优化，服务广泛场景。o4-mini 以极低成本推动推理能力普及，形成了从顶级推理到普惠智能的完整产品矩阵。

Google Gemini 系列：原生多模态与超长上下文的统一架构

2025 年 Google 发挥基础设施优势，通过 Gemini 系列确立多模态与超长上下文领域的领先地位，以规模效应推动智能的商品化普及。

- **核心模型演进与发布时间轴：**2025 年 2 月 5 日，Google 发布 Gemini 2.0 Flash 的 GA 版，以原生多模态和性价比优势获得市场关注。2025 年 3 月 25 日，发布 Gemini 2.5 Pro，首次在旗舰产品中引入“思考”能力。2025 年 5-6 月，发布 Gemini 2.5 Flash，拥有 100 万 token 上下文窗口且价格极低（输入 \$0.30/百万 token），成为 RAG 应用首选。2025 年 11 月 18 日，发布年度旗舰 Gemini 3.0 Pro，标配 100 万 token 上下文。2025 年 12 月 3 日，发布对标 o3-pro 的 Gemini 3 Deep Think，采用迭代式推理机制。2025 年 12 月 17 日，发布 Gemini 3.0 Flash，作为 Gemini App 默认模型提供“下一代智能的闪电速度”。
- **技术特性与架构优势：**Google 2025 年的核心策略是原生多模态与无限上下文。Gemini 3 系列从预训练阶段即实现文本、图像、音频、视频在同一 Transformer 框架内的协同学习，各模态共享统一潜空间。在 MMMU-Pro 多模态推理测试中得分 81.0%，在 Video-MMMU 视频理解测试中得分 87.6%，均处于行业领先。其动态思维能力使单一模型能根据问题难度自动调整推理深度，实现性能与延迟的自适应平衡。Gemini 3 Deep Think 采用迭代式推理机制，面对复杂问题会在后台进行多轮假设验证，探索多个可能的解决路径。
- **超长上下文与检索能力：**Gemini 3.0 Pro 标配 100 万 token 上下文窗口，并能展示处理更高容量的能力。在处理 1M token 级别的信息检索时，几乎不存在“长文本遗忘”现象，召回率保持极高水准。这使得

它能够“吞噬”整个代码库、数百份法律文档或长达数小时的视频素材进行综合分析。Google 还推出了 **Deep Research** 系统级功能，利用 Gemini 的长上下文和推理能力，能自主制定研究计划、搜集资料并整合成深度报告。

- **市场定位与生态布局：**Google 通过显著的价格优势推动智能商品化，Gemini 2.5 Flash 以 \$0.30/百万 token 的输入价格将长文档分析的边际成本推向零点。同时，Google 正在将高性能多模态能力压缩到移动端，**Nano Banana** (Gemini 3.0 Pro Image 预览版) 专注于端侧设备的文本渲染和世界知识，为未来的端侧智能爆发做准备。Gemini Live API 支持低延迟语音对话，构建完整的多模态交互生态。

Anthropic Claude 系列：代码能力优化与计算机使用 (Computer Use) 的探索

2025 年 Anthropic 延续少即是多的精品路线，发布频率虽不及竞品，但每次更新均刷新编程与 Agent 能力上限。其计算机使用 (Computer Use) 功能的突破，使其成为最接近人类员工的模型。

- **核心模型演进与发布时间轴：**2025 年 2 月 24 日，Anthropic 发布了 Claude 3.7 Sonnet，首次引入“混合推理”架构，成为当时编码领域的 SOTA 模型。2025 年 5 月 22 日，发布 Claude 4 系列 (Opus 4 和 Sonnet 4)，正式进入第四代模型竞争，其中 Opus 4 引入了“扩展思维”能力。2025 年 9 月 29 日，发布 Claude Sonnet 4.5，在编码能力和计算机使用上取得突破。2025 年 10 月 15 日，发布 Claude Haiku 4.5，以极致速度和成本效率获得好评。2025 年 11 月 24 日，发布 Claude Opus 4.5，被公认为“世界上最适合编码、代理和计算机使用的模型”。
- **技术特性与通俗能力解读 ThinkingMode 与透明化思维：**Claude Opus 4.5 的核心技术突破包括：**计算机使用能力**使其能够像人类一样“看懂”整个屏幕、移动鼠标、点击图标、输入文字，操作任意桌面软件；**努力参数**让开发者可以精细控制模型在回答问题时投入的计算量 (低、中、高)，将“思考预算”交给用户控制；**Zoom Action** 允许对屏幕特定区域进行像素级放大检查，解决了小字号文本和微小 UI 元素识别难题；**扩展思维能力**提供高质量的思维摘要，让用户了解模型的推理路径，在企业级合规场景中备受青睐。

- **关键性能表现与应用场景:** 在 **SWE-bench Verified** 上, Claude Opus 4.5 达到惊人的 80.9% 得分, 意味着能够解决 GitHub 上 80% 以上的真实 Issue, 从复现 bug、定位代码、编写补丁到通过测试全流程自动化。在开发者社区中, 它被广泛视为唯一真正可用的“结对编程伙伴”, 特别擅长企业级代码库迁移和安全审计等复杂任务。其**交错思维能力**允许在调用外部工具的间隙进行思考和规划, 大幅提升了复杂任务的执行成功率和鲁棒性。
- **市场定位与定价策略:** Claude Opus 4.5 定价为输入 \$5/百万 token、输出 \$25/百万 token, 虽然价格较高, 但在处理高价值任务时的极高一次通过率使其综合拥有成本具有优势。Claude Haiku 4.5 以极低延迟和成本成为 API 经济中的爆款产品, 适合构建即时响应的代码补全工具和高频聊天机器人。Anthropic 通过这一完整的产品矩阵, 精准覆盖了从高端企业级应用到普惠型开发工具的全场景需求。

xAI Grok 系列：算力驱动与实时数据生态的快速突破

2025 年, xAI 以激进的迭代速度和独特的生态优势跻身头部竞争。依托于由 10 万块 H100/H200 GPU 组成的“Colossus”超算集群和对 X 平台数据的实时全量访问权, Grok 系列在推理能力和实时信息处理上构建了双重护城河, 发展路径充满“暴力美学”色彩。

- **核心模型演进与发布时间轴:** 2025 年 2 月 17 日发布 Grok 3, 声称使用了比前代多 10 倍的算力训练, 在数学和物理领域超越了当时的 GPT-4 和 Claude 3.5。2025 年 7 月发布 Grok 4, 拥有估计 1.7 万亿参数的巨型稀疏混合专家模型, 其 Grok 4 Heavy 变体创新性地采用多代理架构。2025 年 11 月 17 日发布 Grok 4.1, 明确区分 Grok 4.1 Thinking (Quasarflux) 和 Grok 4.1 Fast (Tensor) 双模式, 分别针对深度推理和极致速度场景。
- **技术特性与架构创新:** Grok 4 系列的核心突破在于**多代理协同架构**。Grok 4 Heavy 启动多个子代理分别负责信息检索、逻辑推导、代码验证和结果综合, 在 **Humanity's Last Exam** 基准中得分 50.7%, 展示了多代理协作在解决超难问题上的巨大潜力。Grok 4.1 进一步明确策略分离: Thinking 模式在 LMArena 排行榜上曾以 1483 Elo 分登顶, 实现深度逻辑推演; Fast 模式速度高达 455 tokens/s, 专为实时应用

设计。视觉能力上支持原生 OCR 和 PDF 解析，能从扫描件直接提取结构化数据。

- **实时数据生态的独特优势：**Grok 系列最大的、不可复制的护城河在于对 **X 平台数据的实时全量访问权**。通过 **Agent Tools API**, Grok 4.1 能够实时检索全球舆情、突发新闻和市场动态，在处理”分析当前一小时内关于某次突发事件的全球反应”这类任务时拥有上帝视角。这使得 xAI 在实时信息处理领域建立了其他厂商难以逾越的生态壁垒。
- **市场定位与性能表现：**在价格策略上，Grok 4.1 Fast 以输入 \$0.20/百万 token、输出 \$0.50/百万 token 的极致低价，与 Gemini 2.5 Flash 共同推动了”智能商品化”进程。在数学推理基准测试中，Grok 4 在纯模型设置下表现出色，得益于其庞大的参数量和训练数据多样性。API 支持 256k tokens 上下文，能够处理中等规模的项目文件，在开发者和研究者社区中因其激进的迭代速度和独特的生态整合能力而备受关注。

表 6.14: 2025 年最新闭源模型性能综合对比表

核心维度	OpenAI GPT-5.2 / o3	Anthropic Claude Opus 4.5	Google Gemini 3 Pro / Flash	xAI Grok 4
设计哲学	双模态路由：追求逻辑深度与应用广度的平衡，强调 B 端落地与错误率控制。	工匠精神与操作：追求代码质量的极致与人机交互的自然性（Computer Use）。	原生融合：追求感知的统一性，打造无缝的多模态交互与 Agent 自主性。	算力与实时：追求实时信息获取与纯粹的逻辑强推，强调少过滤的直率表达。
推理能力	o3: 极强 (GPQA 98.4%)。适合科研攻关。	Opus 4.5 Thinking: 极强。思维链透明可调试，擅长长逻辑任务。	Gemini 3 Pro: 极强。综合推理持平 GPT-5.2，擅长视觉推理。	Grok 4 (Think): 强。数学能力突出 (AIME 87.5%)，泛化略弱。
编程能力	SWE-Bench: 80.0%。稳定可靠，适合企业级重构。	SWE-Bench: 80.9% (第一)。Web 开发首选，生成的代码结构最清晰。	Antigravity: 强。优势在于 IDE 内的全栈自主规划与执行。	Grok Code: 极客向。擅长解释底层逻辑，适合硬核调试。
多模态	非原生为主：文本/代码极强，视/听依赖组件拼接。	视觉增强：图文理解极强，支持通过视觉操作电脑界面 (UI Navigation)。	原生架构：文/图/音/视频一体化。能”看懂”长视频和物理规律。	视觉增强：支持图文理解，但视频和音频能力较弱。
特色功能	Operator: 智能路由，自动判断任务难度。	Computer Use: 直接控制鼠标键盘操作电脑，自动化测试神器。	原生视频理解: 2M 窗口直接处理长电影； Antigravity IDE。	Real-time X: 实时接入推特数据流，资讯最快。
定价策略	高端溢价：o3 与 Pro 版较贵，主打 B 端。	贵族定价：Opus 4.5 定价较高 (\$5~\$15/1M)，主打高价值生产力。	激进定价：Flash 版本 (\$0.5/1M) 极具破坏力，抢占流量。	会员捆绑：X Premium 订阅包含，API 定价中等。

将四大闭源大模型技术路线并置审视可见，2025 年行业竞争已脱离“生

成能力强化”的单一维度，核心聚焦三大关键领域：复杂任务推理稳定性、跨模态理解一体化水平、代理执行驱动的工作流接管能力。为直观呈现四大闭源系列的竞争格局，以上表格基于 2025 年 12 月的最新数据进行对比。

进入 2025 年，闭源大模型市场格局已告别单点领先的线性竞速阶段，形成四极制衡的竞争态势。行业竞争逻辑从同质化的绝对能力强弱比拼，转向差异化能力维度的壁垒构建。**OpenAI** 能够提供最均衡的工业级智能，**AnthropicClaude** 以编码能力确立细分优势，**Google** 有着更强的多模态处理能力，而 **xAI** 则在动态信息处理场景中构建独特竞争力。

6.2.2 新开源模型

2025 年为开源模型觉醒关键期。2023—2024 年，开源模型仍处于对 GPT-4 的追赶阶段；2025 年，以 Llama 4、DeepSeek V3/R1 及 Qwen 3 为核心的新一代开源模型，不仅在性能上实现对 GPT-4o 的全面超越，更在混合专家（MoE）、多头潜在注意力（MLA）、纯强化学习（纯 RL）等架构创新方向上主导行业趋势，方法论层面亦形成差异化技术突破。

Llama 系列：开源生态的基石与架构重构

Meta 在 2025 年持续主导开源生态的供给体系与规则构建，Llama4 系列的推出实现开源基准的重定义。其核心价值不仅在于新版本发布，更在于重塑了先进开源基座的评价标准。

- **核心模型演进与发布时间轴**: 2025 年 4 月 5 日，Meta 正式发布 Llama 4 系列，标志着开源领头羊正式告别沿用数年的稠密架构，全面转向混合专家（MoE）体系。该系列首发包含两个核心版本：Llama 4 Scout（侦察兵）和 Llama 4 Maverick（特立独行者）。两者均采用 MoE 架构，但设计哲学截然不同。Scout 以效率为导向，总参数量约 1090 亿（109B），推理时仅激活 170 亿（17B）参数；Maverick 则以知识储备为核心，总参数量高达约 4000 亿（400B），同样仅激活 17B 参数。此外，Meta 还内部测试了代号 Behemoth（巨兽）的 2 万亿（2T）参数模型，展现了在超大规模领域的技术储备。
- **技术特性与架构创新**: Llama 4 的核心突破在于 ** 架构范式的双重转变 **。首先是全面采用 ** 混合专家（MoE）架构 **，通过将计算资源集中于最相关的专家子网络，实现了知识容量（总参数）与推理成本

（激活参数）的解耦。其次是实现**原生多模态 (Native Multimodality)**的重大跨越，全系模型原生支持文本、图像、视频和音频的输入与理解。Meta 采用“早期融合”（Early Fusion）技术，摒弃独立视觉编码器，使模型能够更自然地理解跨模态信息。Scout 版本更支持高达**1000 万 (10M) token** 的上下文窗口，能够一次性处理整个代码库或数千份法律文档，彻底改变了 RAG（检索增强生成）的应用模式。

- **性能表现与生态影响：**在技术能力上，Llama 4 的双版本设计解决了不同场景需求。Scout 以其**千万级上下文**和**高吞吐设计**成为企业级部署的首选；Maverick 凭借**128 专家**的高颗粒度设计和**400B 总参数**的知识储备，成为处理复杂推理任务的理想选择。在生态层面，Llama 4 的发布再次引发关于“开源”定义的争论。其**Llama 4 社区许可协议**包含严格的商业限制和品牌归属要求，从 OSI（开源促进会）的严格定义来看，更准确地应被称为“开放权重”模型。尽管如此，Llama 4 仍为开源社区提供了强大的技术基座，推动了整个生态的技术进步和商业化探索。
- **战略意义与行业影响：**Llama 4 的发布标志着开源模型正式进入“激活战争”时代。通过 MoE 架构，Meta 成功地将模型的知识储备与推理效率分离，为后续的开源模型发展设立了新的技术标杆。其千万级上下文支持和原生多模态能力，也为后续的 Agent 应用和复杂任务处理提供了基础设施。作为开源生态的基石，Llama 4 不仅重新定义了性能标准，更通过其许可协议影响着整个行业的商业化路径和发展方向。

Qwen 系列：全场景适配与编程能力优势

通义千问（Qwen）在 2025 年确立开源与闭源双轨并行的实战定位。通过从传统 Transformer 向混合注意力（Hybrid Attention）与极度稀疏 MoE 架构的范式转换，其在长文本处理、复杂编程及多模态 Agent 领域构建起显著技术壁垒。该系列不追求参数规模竞赛，核心竞争力在于通过“系统-思维”推理机制与系统级架构创新，为工业级场景（代码修复、长文档理解等）提供高性价比的闭源替代方案。

- **核心模型演进与发布时间轴：**Qwen 在 2025 年展现了惊人的技术迭代速度。2025 年，Qwen 团队分多批次发布了 Qwen3 系列，涵盖了稠密与混合专家（MoE）两种形态。其中最引人注目的是 2025 年 9 月

发布的 Qwen3-Max，这是开源界参数规模最大的模型之一，拥有超过 1 万亿（1 Trillion）总参数量。在架构创新方面，2025 年 5 月推出的 Qwen3-Next 架构及代表性模型 Qwen3-Next-80B-A3B 实现了极致稀疏设计（仅激活 3B 参数）。多模态领域，2025 年 12 月 30 日发布的 Qwen-Image-2512 在文本渲染和细节真实感上取得突破，同时发布的 Qwen3-VL 则被公认为 Qwen 系列中最强的视觉语言模型。

- **技术特性与架构创新：**Qwen 3 系列的核心技术突破体现在三个方面：**思维模式、极致稀疏架构和多模态深度整合**。Qwen3-Max 引入了”双模式”推理机制，用户可在单次对话中无缝切换”思维模式”（用于复杂逻辑、数学证明及代码生成）与”非思维模式”。在思维模式下，通过工具增强和测试时算力扩展，Qwen3-Max 在 AIME 25 等高难度数学基准测试中达到 100% 的准确率。Qwen3-Next-80B-A3B 采用”A3B”设计（Active 3 Billion），总参数量 80B 但推理时仅激活 3B 参数，这种极致稀疏性使得长文本推理吞吐量是同级别稠密模型的 **10 倍**以上。多模态方面，Qwen-Image-2512 显著增强了对自然元素的渲染精细度，并大幅减少了 AI 生成的”塑料感”；Qwen3-VL 则引入了对视频动态和空间关系的深度感知。
- **性能表现与成本优势：**Qwen 系列在性能与成本间实现了卓越平衡。在数学推理领域，Qwen3-Max 的思维模式在 AIME 25 等竞赛中达到 100% 准确率，展现了系统 2 推理能力。在软件工程方面，Qwen3 系列在 SWE-bench Verified 等基准上表现优异，特别是 Qwen3-4B 这种小模型的性能甚至可以媲美上一代的 Qwen2.5-72B-Instruct。训练效率方面，得益于 PAI-FlashMoE 的多级流水线并行策略，Qwen3-Max 的模型 FLOPs 利用率相对提升了 30%。Qwen3-Next 系列通过混合注意力机制，有效提升了强化学习训练的收敛速度和最终效果，解决了高稀疏 MoE 在 RL 训练中的不稳定性问题。
- **全尺寸覆盖与生态布局：**Qwen 构建了覆盖从 0.6B 端侧小模型到万亿参数云端巨兽的全方位矩阵，发布了包括 0.6B、1.7B、4B、8B、14B、32B 在内的六个稠密模型尺寸，所有模型均支持 128k 上下文窗口（0.6B/1.7B 除外）。其 Apache 2.0 协议的开源策略极大地推动了开发者生态的繁荣。伴随模型发布的分层编辑（Qwen-Image-Layered）和语音合成（Qwen3-TTS）等功能，进一步补全了 Qwen 的多模态拼图，

为工业级应用提供了完整的技术栈支持。

DeepSeek 系列：算法效率优化与成本控制突破

2025 年，DeepSeek 系列以激进的算法创新打破“高性能 = 高成本”的行业铁律，引领了推理模型的开源浪潮。作为“规则破坏者”，DeepSeek 不仅在智力水平上对标顶级闭源模型，更通过极致的成本控制重塑了 AI API 市场的定价体系，对开源与闭源生态均产生了深远影响。

- **核心模型演进与发布时间轴：**DeepSeek 在 2025 年完成了从推理模型到通用智能体的完整布局。2025 年 1 月 20 日发布的 DeepSeek-R1 具有划时代意义，这是首个完全开源权重的、性能对标 OpenAI o1 的推理模型，被广泛视为中美 AI 技术差距缩小的标志性时刻。随后，2025 年 12 月推出的 DeepSeek-V3.2 系列更侧重于通用能力与代理协作，完成了从单一推理引擎到全栈智能平台的升级。其间，DeepSeek 还发布了基于 Llama 和 Qwen 架构的蒸馏小模型系列（涵盖 1.5B 到 70B），将 SOTA 级推理能力普惠到消费级显卡。
- **技术特性与架构创新：**DeepSeek 的核心技术突破体现在**纯强化学习训练路径**和**稀疏注意力架构**两个方面。DeepSeek-R1 不依赖传统的指令微调（SFT）堆叠，而是通过大规模强化学习（RL）在后训练阶段激发模型的自我反思与思维链（Chain-of-Thought）能力，证明了脱离大规模人工标注实现能力涌现的可行性。DeepSeek-V3.2 在注意力机制上进行了重大革新，引入了 **DeepSeek Sparse Attention (DSA)**，通过稀疏化注意力计算，在保证长上下文（128k+）检索精度的同时，显著降低了计算复杂度。特别值得关注的是 **V3.2-Speciale** 变体，专为极限推理设计，在纯逻辑与数学任务上对标 Gemini-3.0-Pro，并在 IMO 和 IOI 竞赛中斩获金牌级表现。
- **成本效益与生态影响：**DeepSeek 在 2025 年对行业最显著的贡献是持续压低推理成本。通过 MoE 架构优化和算法创新，DeepSeek-V3 系列的 API 调用价格低至**每百万输入 token 仅 0.14 美元**（缓存命中时），这迫使全球其他模型提供商不得不跟进降价，彻底重塑了 AI API 市场的定价体系。在许可证方面，DeepSeek-R1 及其蒸馏模型采用了极其宽松的 **MIT 许可证**，允许完全的商业化及修改，极大地推动了基于 R1 的二次开发生态的繁荣。

- **代理能力与工具融合：**针对 Agent 应用，DeepSeek-V3.2 实现了一个关键突破——将“思维”引入工具调用。模型被训练在执行工具操作前先进行内心独白 (Inner Monologue)，分析工具使用的必要性与参数正确性。这种能力源于其构建的覆盖 1800+ 环境、85000+ 条复杂指令的大规模合成数据训练集。在性能表现上，DeepSeek 不仅在数学推理基准 (如 AIME) 中与闭源模型持平，更通过极致的成本效益比，引发了开发者从闭源 API 向 DeepSeek 的大规模迁移，成为开源模型商业化成功的典范。

Mistral 系列与 Gemma 系列：区域化与移动化的双重特化

Mistral AI 与 Google 在 2025 年分别沿着两条清晰的路径推进开源生态：Mistral 聚焦企业级应用与区域化市场深度，Gemma 则专注于移动端原生的极致效率与隐私保护。这种分化反映了开源 AI 市场的成熟与专业化。

- **Mistral AI：极度细分的市场切割术：**法国 Mistral AI 在 2025 年采取了精细化的市场策略，针对不同场景推出了完全独立的产品线。2025 年下半年发布的 Mistral 3 系列全面采用 MoE 架构提升效能：旗舰 Mistral Large 3 拥有 6750 亿总参数和 410 亿激活参数，在 3000 张 NVIDIA H200 GPU 上从头训练，具备顶级的多语言处理和图像理解能力，旨在替代 GPT-4 级别的闭源模型；Mistral Small 3.2 作为 240 亿参数的稠密模型，被优化到极致，覆盖 80% 的生成式 AI 任务且推理速度达 150 token/s，是本地部署的理想选择。特别值得注意的是，Mistral 针对开发者推出了专用化产品：2025 年 12 月发布的 Devstral 2 系列专为代码生成优化，在 SWE-bench Verified 基准上取得 72.2% 的高分；同时发布的 Mistral Vibe CLI 工具允许开发者在终端直接调用 Devstral 进行代码重构和生成，展示了构建开发者生态的野心。在欧洲市场，Mistral 凭借对法语、德语、意大利语、西班牙语等语言的深度优化和文化语境把握，建立了区域性优势。
- **Gemma 系列：边缘计算的高性能标杆：**Google 的 Gemma 系列在 2025 年继续坚持“轻量级、高性能”路线，并在移动端原生架构上实现创新。2025 年 3 月 12 日发布的 **Gemma 3 系列** (1B, 4B, 12B, 27B) 不仅是文本模型，更是原生多模态模型。其核心技术突破在于 **SigLIP 视觉编码与软 Token 设计**：Gemma 3 将图像视为一系列紧凑的“软

Token”（每个图像由 256 个向量表示），极大地降低了视觉推理的算力开销，使得在单张消费级显卡上进行多模态推理成为可能。更引人注目的是 2025 年 6 月发布的 **Gemma 3n 系列**（n 代表 Nano 或 Native Mobile），采用了革命性的 **MatFormer 架构**。这种“套娃”式设计允许模型在推理时弹性选择参数规模，同一模型文件可根据设备电量或内存状态，动态切换为 20 亿或 40 亿参数模式，完美适配移动端复杂多变的运行环境。Gemma 3n 还针对 Google AI Edge 和 TFLite 进行了深度优化，能够充分利用手机 NPU 算力。

- **战略定位与市场影响：**Mistral 与 Gemma 代表了开源模型发展的两个重要方向。Mistral 通过精细化的产品矩阵（Large/Medium/Small、Devstral、Ministral）实现了对企业级、开发者、边缘计算等不同场景的全面覆盖，其多语言优势和文化契合度使其在欧洲市场具备天然护城河。Gemma 则通过极致的移动端优化，将高性能 AI 推向边缘设备，解决了隐私保护和离线使用的核心痛点。两者共同推动了 AI 算力向“超级云端”和“边缘设备”两个极端的快速分流，为不同应用场景提供了多样化的技术选择。

表 6.15: 2025 年最新开源模型性能综合对比表

维度	Meta Llama 4 (Scout/Maverick)	DeepSeek V3 / R1	Qwen 3 / 2.5 Coder	Mistral Large 3	Google Gemma 3
架构范式	MoE + Early Fusion: 原生多模态, iRoPE 支持 10M 上下文。	MLA + FP8 + Pure RL: 极致的算法效率与纯强化学习推理。	MoE + Native Multimodal: 强大的工具调用与编码能力。	MoE (41B/675B): 企业级稳定性与多语言优化。	Dense/MoE (端侧): 轻量级原生多模态, 适配消费级硬件。
核心强项	生态标准: 最广泛的部署支持, 长文档与图文理解能力行业领先。	推理/性价比: 数学逻辑媲美 o1, API 与训练成本极低, 中文能力极强。	编程/Agent: Coder 版本是开源界公认的编程首选, 工具调用灵活。	合规/多语言: 欧洲语言理解最佳, RAG 系统稳定性高。	端侧部署: 4B/12B 版本适合在手机/PC 本地运行, 隐私性好。
编程能力	强: 通用编程能力出色, 但在特定代码库 (Repo-level) 任务上略逊于 Qwen。	极强 (R1): 擅长竞赛级算法题, 但在工程级项目构建上不如 Qwen 均衡。	Top Tier: 在 Live-CodeBench 和 SWE-Bench 上表现统治级, 开发者口碑最佳。	稳健: CodeStral 分支表现优异, 适合企业内部代码补全。	辅助级: 适合本地简单的脚本编写和代码解释, 响应极快。
部署门槛	Maverick (400B): 需 H100/B200 集群。Scout (109B): 量化后单机多卡可跑。	V3 (671B): 模型巨大, 需大显存集群, 但 MLA 优化了推理成本。 蒸馏版: 1.5B-70B 极易部署。	32B 版本: 黄金尺寸, 消费级双卡 (4090) 可跑, 性价比极高。	Large 3: 需高端服务器 (FP8 优化后需 H100)。	极低: 1B-12B 覆盖从手机到笔记本, 量化版极度亲民。
适用场景	企业私有云基座、长文档分析、多模态搜索、通用任务。	科学计算、逻辑推理、低成本 API 替代、数学辅导、教育。	代码助手 (Copilot)、自动化 Agent、视觉分析工具。	欧洲企业合规应用、多语言客服、金融文档分析。	手机端 AI 助理、本地隐私敏感任务、IoT 设备智能。

2025 年大语言模型（LLMs）进展报告

将 2025 年主流新开源模型纳入统一技术版图观察，可发现其呈现罕见的精细化分工格局：Llama 4 系列以提升开源基座技术标准为核心，同步通过 MoE 架构与多模态早期融合方案重构开源模型技术范式；Qwen 系列聚焦工具调用能力强化与编程生态构建，形成具备强实战适配性的通用化技术引擎，可直接嵌入各类产业工作流；DeepSeek 系列沿极致效率优化与强化学习技术路线突破，显著冲击行业现有成本体系，推动全行业重新核算模型部署的成本收益结构；Mistral 与 Gemma 系列则分别以区域合规适配与端侧轻量化部署为核心方向，持续拓展开源模型的应用边界。2025 年的开源生态已彻底摆脱“Llama 一家独大”的格局，形成职能清晰、优势互补的多元化发展态势。

6.2.3 国产开源模型的崛起

2025 年是全球人工智能发展史上的关键节点，技术逻辑正从单一的算力集中化向多元化的技术格局转变。尽管以 GPT-5、Gemini-3 为代表的美国闭源模型在前沿探索方面仍具优势，但中国大模型在开源生态的关键维度中已表现出显著的国际竞争力，实现了从技术追随向特定领域差异化优势的转型，在开源大模型的生态中取得了实质性引领的地位。

研发策略的范式转移 2023 至 2024 年期间，全球大模型竞赛主要围绕规模定律（Scaling Laws）展开，通过堆积 NVIDIA H100 等算力资源来换取大模型性能的提升。进入 2025 年，受限于高端显卡禁售的外部约束以及商业需求，中国 AI 团队开启了新范式：不再单纯追求参数量的指数级增长，而是致力于在有限算力预算下实现大模型性能的极致提升。在这种策略的引领下，中国大模型在保持高性能的同时，实现了模型轻量化以及更好的性价比。例如，DeepSeek-V3 的训练成本约为 550 万至 600 万美元，而性能相当的早期 GPT-4 成本据估算超过 1 亿美元。这种近 20 倍的性价比优势，使中国团队能够以差不多的成本进行更高频率的技术迭代。

在此基础上，与西方主流企业的闭源路线不同，中国大模型团队大多坚持以开源为核心竞争力。截至 2025 年 7 月，全球大模型总数约为 3755 个，其中中国贡献了 1509 个，位居全球首位。随着全球开发者习惯于基于 Qwen 或 DeepSeek 构建应用，一种稳固的技术生态依赖已然形成。通过提供达到业界最优水平的基础模型，中国正在定义下一代人工智能基础的标准。

2025 年大语言模型（LLMs）进展报告

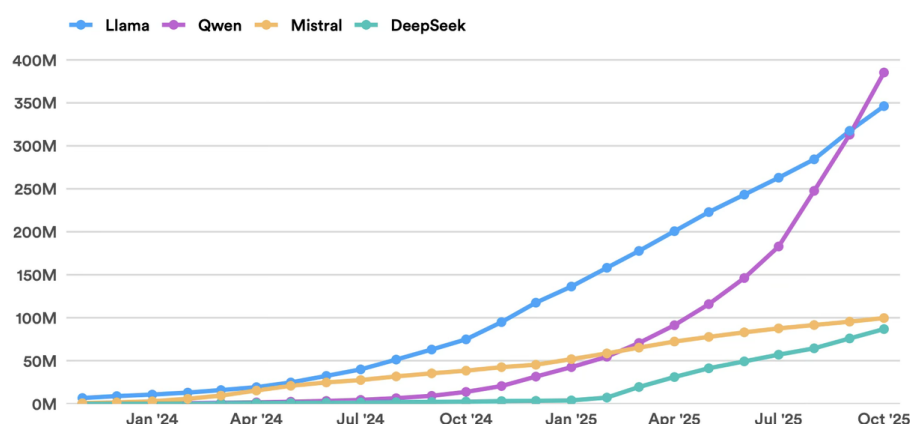


图 6.1: 2025 年 HuggingFace 开源模型下载量变化 (图片来源: http://en.ce.cn/Insight/t20251218_2650721.shtml)

性能和性价比的优势 2025 年，大模型竞争已转向高强度、防污染的基准测试，在这些测试中，中国开源大模型长期稳居前列。以 LMArena 为例，在 2025 年 12 月 31 日，在文本领域，开源模型的前 10 名有 9 个都是国产模型；在 Web 开发领域，则由 GLM-4.7 取得了开源模型的第一名，并且位列总榜的第 6 名，超过了多款闭源模型；视觉领域则由 Qwen 系列领衔，性能超越了包括多款闭源模型在内的国际主流模型；在评测代码能力的权威榜单 LiveCodeBench 上，DeepSeek 与 Kimi 的新型推理模型亦稳居第一梯队。

国产模型的另外一个核心竞争力在于其显著的效费比（Performance-to-Cost Ratio）。这种定价模式改变了原有的大模型商业逻辑，并在行业内引发了成本结构的连锁反应。以 DeepSeek-V3 为例，其 2025 年的市场定价为：输入价格为约每百万 Token \$0.14 – \$0.27，输出价格为约每百万 Token \$0.28 – \$1.10。对比而言，Llama 3.1 70B 的托管价格约为每百万 Token \$3.00，而 GPT-4o 的价格则高出 10 至 25 倍。低价的 Token 成本使开发者在构建复杂长上下文分析及高频交互应用时，不再受到严苛的提示词长度限制，提升了工程实现的灵活性。中国模型将推理价格推向边际成本区间，带动了全球大模型服务成本的普遍下行，引发了更深度的应用适配与场景落地。

市场规模的主导地位 在全球最大的 AI 开源社区 HuggingFace 上，Qwen 和 DeepSeek 几乎包揽了主流开源模型榜单的核心位置。2025 年，中国主要开源模型在上线后均获得全球开发者数十万乃至百万次的下载，展现了深厚

的国际用户基础。其中，Qwen 系列表现尤为卓越，衍生模型突破 10 万个，全球下载量超 3 亿次，成为全球最大开源模型族群，凭借宽松的 Apache 2.0 协议与均衡的性能表现，已成为全球微调任务的首选基座。

基于 OpenRouter 平台超 100 万亿 Token 的实证研究显示，大型语言模型市场正经历深刻重构。开源模型份额已攀升至 33%，彻底打破闭源模型垄断，其中中国厂商占据主导地位。具体的，中国开发的开源模型市场份额从 2024 年底的 1.2% 飙升至 2025 年中近 30% 的峰值，年平均占比达 13.0%，与世界其他地区开源模型 13.7% 的份额几乎持平。这意味着在短短半年内，中国已从市场边缘跃升为全球开源生态的半壁江山。海外企业纷纷转向中国 AI 模型，如塞浦路斯平台 Latenode 采用 DeepSeek 模型，成本仅为 OpenAI 的 1/17，沙特阿美更将 DeepSeek 系统部署在其数据中心。与此同时，全球 AI 支出结构发生巨变，北美份额长期低于 50%，而中国开源模型的崛起是这一结构性变化的核心驱动力。

表 6.16: 国内外开源大模型多维度对比分析

维度	差异特征描述	代表模型
研发战略范式	国外依赖 Scaling Laws 堆砌算力资源；国内追求有限算力下的极致性能。	Mistral vs DeepSeek
算力与成本	国外维持高训推成本与定价，售卖稀缺资源；国内将推理价格压至较低成本，具备性价比优势。	GPT-4o vs DeepSeek
开源生态协议	国外开源常带商业限制，意图构建护城河；国内倾向 Apache 2.0 宽松协议，旨在成为全球基础设施。	Llama vs Qwen

总结与展望 2025 年，中国开源模型的全面崛起并非偶然，而是在外部封锁和内部竞争的双重压力下，通过架构创新和工程极致化探索出的一条自主发展道路。DeepSeek 证明了通向通用人工智能的路径具有多元化特点：通过算法升级，能够以极低的成本实现世界级智能水平；Qwen 则证明了开源生态可以具备与闭源系统相媲美的宏大商业规模。这种高性能、低成本、全生态的中国模式正在重塑全球 AI 格局。对于全球开发者而言，中国开源模型已从可选项转变为追求性价比和自主可控的优先选择项。中国不仅赶上了

AI 发展浪潮，更在开源领域确立了领先地位。

6.3 综合能力排行榜汇总

本部分重点关注了以 OpenAI 的 GPT-5.2 系列、Google 的 Gemini 3 Pro 系列、Anthropic 的 Claude 4.5 系列为代表的“2025 世代”模型以及稳步崛起的开源模型，在语言任务、视觉理解与生成、语音交互、编程能力、数学能力以及推理能力等核心维度的表现。

6.3.1 语言能力（Language）评测调研

语言能力评测的关注点正在发生迁移。过去的“知识问答”更像记忆力的投影，如今的评测更强调高难度推理的持续性、指令遵循的精细度，以及在工具调用与多轮对话中维持目标一致性的能力。模型之间的差距也因此被重新拉开——不是谁背得多，而是谁想得明白，谁扛得住干扰，谁在复杂约束下仍能把答案落到正确的位置上。

Arena Hard Auto

• 基准评述

ArenaHardAuto 属于自动化高难度评测流程，它从 ChatbotArena 的真实用户对话日志中抽取 500 个最具挑战性的技术与逻辑问题，再采用“强模型担任裁判”的 LLM-as-a-Judge 机制进行成对比较，把评测从实验室题库拉回真实交互现场。研究表明，该自动化评测与人类真实偏好的相关性高达 98.6%。

表 6.17: 各大模型在 Arena Hard Auto 上的排行榜

排名	模型	类型	得分	备注
1	Gemini 3 Pro	闭源	92.5	综合对话体验最佳，逻辑连贯性强
2	GPT-5.2 (High)	闭源	91.2	在复杂指令遵循上极具优势
3	Claude 4.5 Opus	闭源	89.7	代码与长文写作深受开发者喜爱
4	Grok 4.1 Thinking	闭源	88.4	个性化与幽默感得分较高
5	DeepSeek V3.2	开源	85.0	极具性价比的开源推理模型
6	Llama 4 Maverick	开源	82.0	依然是开源社区的基石
7	Qwen3-Max	闭源	~81.5	英文对话能力显著提升
8	Mistral Large 3	开源	78.0%	欧洲语言处理能力强

• **结果说明与分析** 此基准反映了模型在真实对话场景中的综合表现。**Gemini 3 Pro** 的夺冠表明 Google 在 RLHF（人类反馈强化学习）阶段对语气控制、结构化输出及逻辑连贯性的对齐效果最优，更契合人类高级交互偏好。虽然 **GPT-5.2** 在硬逻辑上很强，但在对话的自然流畅度和用户偏好上略逊一筹。开源领域的 **DeepSeek V3.2** 和 **Llama 4** 差距较小，说明在通用对话体验上，开源模型已经非常成熟。

Berkeley Function Calling Leaderboard (BFCL)

• 基准评述

BFCL 引入了基于抽象语法树（AST）的评估方法。它不仅检查模型是否输出了函数名，还深入验证参数类型、结构嵌套以及在多轮对话中保持状态的能力。测试结果显示，许多在对话上表现优异的模型在面对复杂的嵌套函数调用时容易出现幻觉或参数错误。

表 6.18: 各大模型在 BFCL 上的排行榜

排名	模型	类型	准确率	备注
1	Claude 4.5 Opus	闭源	95.8%	极少出现格式错误，复杂参数理解力 SOTA
2	GPT-5.2	闭源	94.5%	依然是工业界工具调用的标准选择
3	Grok 4.1 (Tool)	闭源	93.0%	针对工具调用进行了专门微调
4	Gemini 3 Pro	闭源	92.1%	在多模态工具链上表现更好
5	Llama 4 Maverick	开源	89.4%	开源模型中工具调用能力最强
6	DeepSeek V3.2	开源	85.0%	较上一代有大幅提升
7	Qwen3-Coder	开源	83.5%	在代码相关 API 调用上表现出色

• **结果说明与分析** 工具调用能力是 Agent 系统的基石。**Claude 4.5 Opus** 以 95.8% 的准确率位居榜首,这解释了为何开发者社区更倾向于使用 Claude 构建复杂的 Agent 工作流——其对 JSON 格式和复杂嵌套参数的理解几乎零失误。尽管 **Llama 4** 是开源最强，但与闭源模型仍有约 5% 的差距，这一差距在生产环境中可能导致更高的错误率与重试成本。

RULER (Long-Context)

• 基准评述

RULER 的定位很明确：专门用来“打假”长上下文。它不满足于“大海捞针（NIAH）”那种单点检索式的演示，而是把难度拧到更接近真实文档

工作的形态——多跳追踪、跨段聚合、线索拼接与一致性校验被拆成 13 类任务，要求模型在超长窗口里既要找得到，还要连得起来，更要推得下去。

表 6.19: 各大模型在 RULER 上的排行榜

排名	模型	类型	平均得分	备注
1	Gemini 3 Pro	闭源	94.2%	在 1M+ 长度下几乎无衰减，统治级表现
2	GPT-5.2	闭源	89.5%	
3	Claude 4.5 Opus	闭源	88.0%	长文理解逻辑性强
4	Llama 4 Scout	开源	86.5%	令人惊讶的千万级上下文支持
5	Qwen3-Max	闭源	~85.6%	长文档处理能力稳健
6	DeepSeek V3.2	开源	79.9%	引入稀疏注意力机制降低了成本

• **结果说明与分析** 在超长上下文领域，**Gemini 3 Pro** 展现了架构层面的绝对优势。不同于其他模型随着长度增加性能线性下降，Gemini 几乎保持了全窗口的无损召回。这表明 Google 在 Ring Attention 或类似的无限上下文技术上取得了突破。对于开源模型，**DeepSeek V3.2** 虽然引入稀疏注意力降低了推理成本，但也牺牲了一定的长文召回精度 (79.9%)，这在处理百万级 Token 时需要权衡。

6.3.2 图像与视频（Vision & Video）多模态评测调研

2025 年的多模态竞争，焦点已经从“看见什么”转向“看懂什么”。识别物体、读出文字只是起点；真正拉开差距的，是模型能否沿着画面里的线索把因果链条接起来，能否在时间轴上追踪变化、理解动作、推断物理规律与社会动机。换句话说，多模态开始逼近一种更接近人类的理解方式：既要视觉推理，也要时序推理。

VCR (Visual Commonsense Reasoning)

• 基准评述

VCR（视觉常识推理）要求模型不仅回答“图里发生了什么”（Q→A），还要从四个选项中选出“为什么”（QA→R）。把描述与解释捆成一条链，促使模型面对视觉世界的因果逻辑与社会常识。这里考的不是眼力，而是理解力——看见只是输入，解释才是能力。

表 6.20: 各大模型在 VCR 上的排行榜

排名	模型	类型	准确率	备注
1	GPT-5.2 (Vision)	闭源	92.8%	极其强大的常识推理能力
2	Qwen3-Max	闭源	91.9%	展现了极强的视觉理解
3	Gemini 3 Pro	闭源	91.4%	原生多模态架构，表现稳定
4	Claude 4.5	闭源	90.5%	细节捕捉能力强
5	DeepSeek Janus Pro	开源	88.0%	解耦视觉编码器设计带来了意外之喜
6	Llama 4 Maverick	开源	82.3%	主要是 OCR 强，推理稍弱

• **结果说明与分析** VCR 的结果令人惊讶之处在于 **Qwen3-Max** (91.9%) 的表现，它超越了 Gemini 并紧追 GPT-5.2。这打破了“只有原生多模态才能做好推理”的迷思，证明了通过高质量的微调数据，模型可以学会“看图说话”背后的逻辑。相比之下，**Llama 4** 较低的推理分数 (82.3%) 揭示了其视觉模块更多聚焦于 OCR 和物体识别，而在深层语义理解上仍有欠缺。

MVBench (Multi-modal Video Benchmark)

• 基准评述

MVBench 是一个视频理解专项综合基准，它覆盖 20 种不同的时序任务，采用静态转动态方法生成高质量评测题，目的是检验模型是否真的理解“时间”。视频理解的难点不在画面里有什么，而在画面如何变化；不在某一帧的内容，而在多帧之间的关系。

表 6.21: 各大模型在 MVBench 上的排行榜

排名	模型	类型	平均得分	备注
1	Gemini 3 Pro	闭源	89.2%	视频理解的王者，原生处理长视频流
2	GPT-5.2	闭源	86.5%	短视频理解能力极强
3	Claude 4.5	闭源	84.3%	-
4	Qwen2.5-VL-72B	开源	73.6%	开源视频理解的最佳选择
5	Llama 4 Maverick	开源	~68.0%	-

• **结果说明与分析** 视频理解的核心在于时序性处理能力。**Gemini 3 Pro** 的领先地位 (89.2%) 证明了其能够原生处理长序列视频帧，而不是像其他模型那样通过抽帧将视频退化为图片集。开源模型在此领域与闭源模型存在 15 个百分点以上的差距，这表明时间维度的因果推理是当前开源多模态模型的主要技术壁垒。

Video-MMMU

• 基准评述

Video-MMMU 把难度进一步推高。该基准包含 300 个来自大学课程与专业讲座的长视频，考的**不是**“看一眼回答”，而是“看完再总结”。它引入 knowledge 指标，用来衡量模型在观看视频后获取的信息增量——这等于把评测从视觉理解拉回学习能力本身：能否在长时段输入里持续吸收、持续更新、持续保持一致。

表 6.22: 各大模型在 Video-MMMU 上的排行榜

排名	模型	类型	准确率	备注
1	Gemini 3 Pro	闭源	87.6%	唯一能高效从长视频中提取知识
2	GPT-5.2	闭源	84.1%	-
3	Claude 4.5	闭源	82.5%	-
4	Qwen3-VL	开源	76.8%	表现不俗，具备一定的专业视频理解力

• **结果说明与分析** 此基准测试了模型通过视频学习的能力。**Gemini 3 Pro** 的高分不仅源于其视觉能力，更源于其处理超长上下文的记忆力——讲座视频通常长达一小时。大多数模型在视频后半段会出现信息遗忘，而 Gemini 能保持全程关注。这表明在教育 and 培训领域的 AI 应用中，Gemini 目前是不可替代的。

6.3.3 语音能力（Speech）评测调研

语音交互在 2025 年发生了质变：它不再停留在“把话写下来”的转写工具，而是迈向端到端、全双工的实时对话系统。人类对语音的苛刻从来不是“你听懂了吗”这么简单，真正刺痛体验的往往是延迟、打断、抢话，以及情绪与语气的误读。于是评测也跟着转向——速度要被量化，节奏要被感知，互动要像人与人说话那样自然地衔接。

Open ASR Leaderboard

• 基准评述

OpenASRLeaderboard 的价值在于它把“准确”与“快”同时拉进同一套标尺，不仅关注传统字错率（WER），还引入逆实时因子（RTF_x）衡量推理速度，全面评估语音转写的准确性与效率。

表 6.23: 各大模型在 Open ASR Leaderboard 上的排行榜

排名	模型	类型	WER(Clean/Noisy)	备注
1	Mistral Voxtral Small	开源	2.1% / 3.8%	惊人的低错误率，超越了 GPT-4o
2	Gemini 3 Deep Think	闭源	2.3% / 5.1%	在复杂噪音环境下表现优异
3	GPT-5.2 Audio	闭源	2.5% / 5.4%	交互体验好，但纯识别率略低
4	Whisper V3 Large	开源	2.4% / 3.9%	依然是强大的基准，但已被后浪超越
5	Canary Qwen 2.5B	开源	5.63% (Avg)	极高性价比，适合边缘端部署

• **结果说明与分析** 语音转写是少数几个开源模型实现对闭源模型超越的领域。Mistral Voxtral 的极低错误率 (2.1%/3.8%) 验证了模态专用小模型的技术优势——专注语音任务的轻量模型可突破通用多模态大模型的性能限制。纯转写场景下，开源专用小模型在成本与精度上均占优，而大模型的优势更多体现在对语音内容的理解和情感交互上。

6.3.4 编程能力（Programming）评测调研

编程能力已成为大语言模型落地应用中最具经济价值的能力之一，也是检验模型逻辑严密性、长上下文管理能力以及指令遵循能力的关键维度。2025 年的编程评测已彻底告别了简单的函数补全，全面转向了 **Agentic Coding** 和**实时竞赛编程**。

SWE-bench Verified

• 基准评述

现有编程基准的一个局限性在于，它们往往只测试模型生成几行独立代码的能力，而忽略了真实软件开发中至关重要的“在现有代码库中进行修改”的能力。现实中的软件工程要求模型不仅要理解复杂的依赖关系，还要在不破坏现有功能的前提下修复 Bug 或添加特性。

为弥补这一不足，普林斯顿大学的研究团队发布了 SWE-bench，并在随后推出了 SWE-bench Verified——一个由人类专家严格筛选的子集，确保测试问题的有效性和非歧义性。该基准包含 500 个从真实 GitHub 仓库中提取的 Issue，要求模型自主浏览文件、定位问题、编写补丁并通过测试。当前顶尖模型在此基准上的竞争已进入白热化阶段，Claude Opus 4.5 与 GPT-5.2 均突破了 80% 的解决率，标志着 AI 正在从“辅助编程”迈向“自主工程”。

表 6.24: 各大模型在 SWE-bench 上的排行榜

排名	模型	类型	解决率	备注
1	Claude Opus 4.5	闭源	80.9%	当前 SOTA，工程化能力极强
2	GPT-5.2 (Thinking)	闭源	80.0%	逻辑修复能力顶级，紧随其后
3	Gemini 3 Pro	闭源	76.2%	长上下文优势在大型 Repo 中明显
4	GPT-5.1 / o3	闭源	~66.0%	上一代推理模型水平
5	DeepSeek V3.1	开源	66.0%	开源界第一，超越 Llama
6	Grok 4	闭源	60.5%	相比前代大幅提升，但在工程细节略逊
7	Qwen 3 Coder 30B	开源	51.6%	30B 尺寸下的极佳表现
8	Mistral Large 3	开源	46.8% (Devstral)	配合专用 Agent 框架表现尚可
9	Llama 4 Maverick	开源	21.04%	令人意外的低分，似乎未针对 Agent 优化
10	Gemma 3	开源	~	缺乏针对此高难度基准的公开数据

• **结果说明与分析** 数据显示，在高度工程化的任务中，**Claude Opus 4.5** 与 **GPT-5.2** 构成了第一梯队，解决率突破 80%，表明闭源模型在处理跨文件上下文和复杂依赖关系时仍具有显著优势。值得注意的是，开源模型 **DeepSeek V3.1** (66.0%) 表现惊人，不仅超越了 xAI 的 Grok 4，更大幅领先于 Meta 的 Llama 4 Maverick (21.04%)。这揭示了 Llama 4 系列在代码代理（Code Agent）任务上的微调策略可能存在缺失，而 DeepSeek 凭借其强化学习策略在工程代码领域取得了极高的性价比。

LiveCodeBench

• 基准评述

传统编程基准（如 HumanEval）面临严重数据污染问题——模型可能在训练数据中见过测试题，导致高分无法反映真实编程能力，仅体现记忆水平。

为了解决这一问题，LiveCodeBench 采用了一种动态更新的机制，专门收集模型训练截止日期之后发布的竞赛编程题目（来自 LeetCode, AtCoder, Codeforces）。该基准测试不仅考察代码生成，还包括自我修复和测试用例执行。当前结果显示，具备“系统-思维”的推理模型在处理这些从未见过的算法难题时，表现显著优于传统模型，揭示了逻辑推理在解决新颖编程问题中的核心作用。

表 6.25: 各大模型在 LiveCodeBench 上的排行榜

排名	模型	类型	Pass@1	备注
1	GPT-5.2 (High)	闭源	53.1%	处理未见难题能力最强
2	Gemini 3 Pro	闭源	49.0%	算法逻辑稳健
3	DeepSeek V3.2	开源	48-50%	开源模型在纯算法上的突破
4	Claude 3.7 Sonnet	闭源	~45%	依然强劲，适合日常开发
5	Llama 4 Maverick	开源	43.4%	算法能力优于其工程能力
6	Grok 3 Mini	闭源	~40%	小参数模型中的佼佼者
7	Qwen 2.5 Coder 32B	开源	31.4%	小参数代码模型标杆
-	Mistral Large 3	开源	~	暂无最新 LiveCodeBench 数据
-	Gemma 3	开源	~29.7%	27B 版本表现一般

• **结果说明与分析** 针对未见过的算法难题，具备“系统-思维”的模型优势尽显。GPT-5.2 和 DeepSeek V3.2 的高分证明了逻辑推理能力与编程能力的强相关性。与 SWE-bench 不同，算法竞赛更侧重纯逻辑而非工程架构，这使得小参数模型也能通过针对性训练取得不俗成绩，展现了在特定垂直领域以小博大的潜力。

BigCodeBench

• 基准评述

现有编程比较基准的一个局限性在于，许多测试仅局限于短小、自包含的算法任务或独立函数调用。然而，解决复杂实际任务通常需要调用多样化函数的能力。高效的编程还要求模型能理解自然语言表达的编码指令——这一能力未被当前多数编程基准所测试。

为弥补现有编程基准的不足，一个国际团队于 2024 年发布了 Big-CodeBench——一个全面、多样且极具挑战性的编程比较基准。该基准要求大语言模型跨 139 个库和 7 大领域调用多重函数调用，涵盖 1,140 项细粒度任务。当前人工智能系统在该基准上表现欠佳：即使顶尖模型在完整任务与指令任务困难子集也面临显著挑战，凸显了 AI 在复杂库调用和指令遵循上的能力仍然是当前技术提升的核心方向。

表 6.26: 各大模型在 BigCodeBench 上的排行榜

排名	模型	类型	Pass@1	备注
1	Claude 3.7 Sonnet	闭源	35.8%	在复杂指令遵循上表现最佳
2	OpenAI o1	闭源	35.5%	推理能力在此依然有效
3	OpenAI o3-mini	闭源	35.5%	小模型表现惊人
4	DeepSeek R1	开源	35.1%	紧咬闭源顶尖模型
5	Gemini 3 Pro	闭源	~	数据待更新，预计在第一梯队
6	GPT-5.2	闭源	~	数据待更新
-	Llama 4	开源	~	暂无针对性数据
-	Qwen 3	开源	~	暂无针对性数据
-	Grok 4	闭源	~	暂无针对性数据

• **结果说明与分析** BigCodeBench 的低通过率（普遍低于 40%）反映了当前 LLM 在复杂库调用和指令遵循上的普遍瓶颈。**Claude 3.7 Sonnet** 凭借优秀的指令遵循能力险胜，而 **OpenAI o1/o3** 系列与其差距极小。开源界的 DeepSeek R1 再次紧咬闭源顶尖模型，证明了纯 RL 训练在复杂工具调用场景下的泛化能力正在迅速逼近闭源 SOTA。

6.3.5 数学能力（Mathematics）评测调研

2025 年，随着模型在中学奥数（AIME）级别题目上的全面攻克，数学评测已经进入了高等数学与前沿研究的新阶段。

AIME 2025

• 基准评述

美国数学邀请赛 (AIME) 长期以来被视为高中数学竞赛的标杆，要求参赛者在没有计算器的情况下解决极其复杂的整数问题。过去的语言模型在此测试中往往因计算错误或逻辑断裂而失败。

然而，AIME 2025 的测试结果标志着一个时代的结束。随着“推理时间计算” (Thinking Process) 技术的引入，模型学会了自我验证和多步推演。当前最先进的模型在此基准上已取得满分或接近满分的成绩，这表明对于定义明确、逻辑封闭的竞赛数学题，AI 已经达到了超越绝大多数人类选手的水平。

表 6.27: 各大模型在 AIME 2025 上的排行榜

排名	模型	类型	准确率	备注
1	GPT-5.2 (Thinking)	闭源	100%	完美解决，已达上限
2	Gemini 3 Pro	闭源	95.8%	极少数边缘错误
3	Grok 4	闭源	93.3%	数学能力极其出色
4	DeepSeek R1/V3.2	开源	93.1%	开源最强，匹敌闭源
5	OpenAI o3	闭源	87.5%	依然强大，但在 2025 年已非最强
6	Qwen 2.5 Math	开源	~85%	专精数学的模型
7	Mistral Large 3	开源	40.0%	通用模型在数学上的短板
8	Llama 4 Maverick	开源	~	数据未明确，推测低于 Qwen

• **结果说明与分析** 与 FrontierMath 形成鲜明对比，AIME 2025 的榜单显示高中竞赛级数学已被“彻底攻克”。GPT-5.2 达到 100% 的胜率，开源的 DeepSeek R1 也达到了 93.1%。这标志着对于定义明确、逻辑封闭的数学问题，AI 的推理能力已达到饱和。未来的竞争焦点将不再是此类标准化考试，而是向更开放、更抽象的数学探索转移。

FrontierMath

• 基准评述

随着 AI 在 GSM8K 和 MATH 等传统数学基准上接近满分，这些测试已无法有效区分前沿模型的能力，掩盖了 AI 在真正具有创造性的数学研究面前的局限性。

为解决这一评价失效问题，Epoch AI 联合 60 多位数学家推出了 FrontierMath。该基准包含数百个从未发表过的、专家级的原创数学问题，涵盖代数几何、数论等现代数学分支。这些问题通常需要人类数学家花费数小时甚至数天才能解决。在这一基准上，绝大多数模型的得分甚至不足 2%，唯有 GPT-5.2 等具备深度推理能力的模型开始展现出解决此类问题的苗头，但这依然是 AI 通往“数学家”之路上的最大挑战。

表 6.28: 各大模型在 FrontierMath 上的排行榜

排名	模型	类型	准确率	备注
1	GPT-5.2	闭源	40.3%	遥遥领先，展现初级研究能力
2	GPT-5.1 (Thinking)	闭源	26.7%	较 5.2 有显著差距
3	OpenAI o3	闭源	25.0%	曾是 SOTA，现居中游
4	Gemini 3 Pro	闭源	~	缺乏具体 Tier 1-3 分数，预计较低
5	Claude 3.7 Sonnet	闭源	< 20%	非数学特化模型
-	DeepSeek R1	开源	~	尚无针对此超难基准的公开测试
-	Grok 4	闭源	~	尚无公开测试数据
-	Llama 4	开源	~	尚无公开测试数据

• **结果说明与分析** FrontierMath 的测试结果揭露了当前 AI 数学能力的真实边界。尽管 GPT-5.2 取得了 40.3% 的断层领先，但大多数模型（包括 Claude 3.7）得分不足 20%，说明在缺乏训练数据覆盖的原创性数学研究领域，现有的 Scaling Law 遇到了瓶颈。真正的“数学家级”AI 尚未诞生，目前的模型仍更多依赖于模式匹配而非创造性证明。

MMMLU (Multimodal Multi-task)

• 基准评述

单纯的文本数学题已不足以概括现实世界的复杂性。现实中的问题往往结合了图表、几何图形和物理示意图，要求模型具备跨模态综合推理能力。

MMMLU（大规模多任务多模态理解）基准旨在测试模型跨越文本与视觉的综合推理能力。它不仅要求模型读懂题目，还要看懂题目中的图示，并将视觉信息转化为数学约束条件。在该基准上，原生多模态架构（如 Gemini 3）显示出了相对于拼接式架构（如早期 GPT-4V）的天然优势，印证了视觉感知与逻辑推理深度融合的技术价值。

表 6.29: 各大模型在 MMMLU 上的排行榜

排名	模型	类型	准确率	备注
1	Gemini 3 Pro	闭源	91.8%	原生多模态架构优势明显
2	GPT-5.2	闭源	89.6%	视觉推理能力极强，略逊 Gemini
3	Grok 4	闭源	78.0%	视觉能力相对较弱
4	Qwen 3-VL	开源	~80%	开源多模态首选
5	Llama 4 Maverick	开源	52.2% (MMMU Pro)	早期融合架构，但分数一般
6	Claude 3.5 Sonnet	闭源	~	缺乏 MMMLU 专项数据
-	DeepSeek V3	开源	~	视觉非其核心强项

• **结果说明与分析** 在多模态数学推理中，Google Gemini 3 Pro 凭借原生多模态架构（91.8%）确立了绝对优势，证明了“视觉-语言”早期融合架构在处理几何与物理图表时，远优于外挂视觉编码器的方案。Qwen 3-VL 作为开源代表，虽然与 Gemini 仍有差距，但已证明了开源多模态架构在处理视觉逻辑上的可行性。

6.3.6 推理能力（Reasoning）评测调研

推理能力是通用人工智能（AGI）的核心，涵盖了科学知识、抽象逻辑和跨学科综合判断三大核心模块。

MMLU-Pro

• 基准评述

经典的 MMLU 基准曾是衡量模型知识广度的核心标准，但随着模型能力的提升，其区分度已显著下降。

MMLU-Pro 是对经典版本的全面升级。它通过两项关键优化提升评估有效性：一是将选项从 4 个增加到 10 个，大幅降低了随机猜对的概率；二是新增大量需多步推理的题目，弱化简单记忆类题目占比。该基准覆盖了法律、医学、物理等 14 个专业领域。当前的测试结果显示，模型要想在此基准上取得高分，必须具备深厚的领域知识与稳健的逻辑推导能力的结合。

表 6.30: 各大模型在 MMLU-Pro 上的排行榜

排名	模型名称	类型	准确率	备注
1	Gemini 3 Pro	闭源	90.10%	跨过 90% 门槛，STEM 领域近乎满分
2	Gemini 3 Flash	闭源	88.59%	极高的推理效率，小参数量却有超强逻辑
3	Claude Opus 4.1	闭源	87.92%	开启”思维链 (Thinking)”模式后极度严谨
4	Claude Sonnet 4.5	闭源	87.36%	性价比之王，广泛应用于复杂工程任务
5	iAsk Pro	闭源	85.85%	针对事实问答优化的垂直黑马
6	GPT-5.0	闭源	86.51%	综合能力均衡，指令遵循能力最强
7	DeepSeek-V3.2	开源	85.00%	全球开源模型第一，推理能力对齐顶级闭源
8	GLM-4.6 / 5.0 (测试版)	开源	84.60%	智谱 AI 出品，中英文多任务处理极佳
9	Qwen3-235B	开源	84.40%	阿里出品，在数学、代码推理上有独特优势
10	Llama 4 (70B/405B)	开源	82.10%	生态兼容性最强，推理稳定性极高

• **结果说明与分析** MMLU-Pro 的高分榜单呈现出多极化趋势。虽然 Google 和 OpenAI 依然处于领跑位置，但领先优势已大幅收窄。**Grok 4.1** (88.0%) 的表现尤为亮眼，印证了实时数据流接入对模型知识时效性与推理准确性的提升作用。DeepSeek V3 以 85.0% 紧随其后，证明了在通用知识推理层面，开源模型已具备替代闭源模型的实力。

GPQA Diamond

• 基准评述

为了测试 AI 是否具备超越普通人类的专业知识，GPQA Diamond 基准应运而生。它由生物学、物理学和化学领域的博士专家编写，题目难度极高，即使是拥有博士学位的非本专业专家在可以使用谷歌搜索的情况下，正确率也仅为 34%。

该基准主要考察模型在极端专业领域内的深度推理能力。目前，GPT-5.2 和 Gemini 3 Pro 均已突破 90% 的大关，这意味着在某些特定的科学细分领域，这些模型的表现已经可以媲美甚至超越人类顶尖专家。

表 6.31: 各大模型在 GPQA 上的排行榜

排名	模型	类型	准确率	备注
1	GPT-5.2	闭源	92.4%	科学推理之王
2	Gemini 3 Pro	闭源	91.9%	与 GPT-5.2 差距极小
3	Grok 4 Heavy	闭源	88.4%	专家级知识储备
4	Claude Opus 4.5	闭源	~87%	表现强劲
5	DeepSeek R1	开源	81.0%	开源模型中的科学专家
6	Qwen 2.5 Max	开源	~	优于 DeepSeek V3
7	Llama 4 Maverick	开源	69.8%	科学推理相对较弱

• **结果说明与分析** 在博士级科学问答中，GPT-5.2 和 Gemini 3 Pro 构成的“双子星”格局难以撼动，均突破 90%。这表明在极端深度的专业知识领域，超大规模参数和高质量的后训练（Post-training）数据依然是决定性因素。Llama 4 在此项测试中的落后 (69.8%) 暗示了其训练数据在顶尖学术语料上的权重可能不足。

ARC-AGI (Abstraction and Reasoning Corpus)

• 基准评述

大多数基准测试侧重于知识的检索或既定规则的应用，而 ARC-AGI 旨在测试模型适应新颖模式和抽象推理的能力——即所谓的“流体智力”。

该基准要求模型仅通过极少量的示例，就能推断出网格变换的抽象规则。ARC-AGI-2 是其更难版本，专门用于抵御暴力破解。该基准属于纯粹的智商测试，与训练数据的记忆无关。在此基准上，GPT-5.2 的高分证明了其确实具备了一定程度的通用抽象能力，而不仅仅是单纯的统计模仿。

表 6.32: 各大模型在 ARC-AGI 上的排行榜

排名	模型	类型	分数	备注
1	GPT-5.2 (Thinking)	闭源	52.9%	抽象推理能力的重大突破
2	Gemini 3 Deep Think	闭源	45.1%	深度思考模式有效
3	Claude Opus 4.5	闭源	37.6%	依然优于大多数人类
4	Gemini 3 Pro	闭源	31.1%	~
5	Grok 4 (Refine)	闭源	29.4%	~
-	DeepSeek R1	开源	~	暂无官方 ARC-2 数据
-	Llama 4	开源	~	暂无数据

• **结果说明与分析** ARC-AGI-2 是目前区分度最高的智商测试。**GPT-5.2** 的 52.9% 虽然夺冠，但绝对分值依然较低，说明 AI 在“流体智力”（即适应全新规则）方面仍远逊于其“晶体智力”（知识储备）。这是通往 AGI 的关键一公里，目前仅有具备深度思考模式（Thinking Mode）的模型能在此基准上取得及格分。

Humanity’ s Last Exam (HLE)

• 基准评述

随着 MMLU 等基准的饱和，为设定 AI 能力的终极评估天花板，人工智能安全中心（Center for AI Safety）联合 Scale AI 发布了 Humanity’ s Last Exam（HLE）基准。

该基准汇集了 2,500 个由全球各领域顶尖专家设计的、极其晦涩且跨学科的难题，旨在测试模型在人类知识边界上的综合能力。这不仅是知识的测试，更是对多模态理解、长尾知识检索和深度推理的综合大考。目前，即便是最强的模型也难以达到 60% 的正确率，Grok 4 在此基准上的领先暗示了其在训练数据多样性或实时信息整合上的独特优势。

表 6.33: 各大模型在 HLE 上的排行榜

排名	模型	类型	准确率	备注
1	Grok 4 Heavy	闭源	50.7%	意外的冠军，知识覆盖极广
2	Gemini 3 Pro	闭源	45.8%	表现非常均衡
3	Gemini 3 Flash	闭源	43.5%	性价比极高的推理能力
4	GPT-5.2	闭源	34.5%	在冷门知识上略显不足
5	DeepSeek V3.1	开源	30%	开源最佳
6	Claude Opus 4.5	闭源	25.2%	可能受限于安全微调
-	Llama 4	开源	~	暂无数据

6.3.7 智能体能力（Agents）评测调研

智能体能力评估聚焦模型的工具使用、任务规划路径构建与自我反思优化三大核心维度。

GAIA (General AI Assistants benchmark)

• 基准评述

现有的问答基准往往只要求模型输出文本，而忽略了 AI 作为助手解决实际问题的能力。

GAIA 基准旨在评估通用 AI 助手的实战能力。它包含了一系列概念上简单但在操作上繁琐的任务，这些任务往往需要模型组合使用搜索、计算器、代码解释器等多种工具。GAIA 的设计理念是：对于人类来说简单的问题，对 AI 来说可能极难。目前，通过多模型集成（Ensemble）的方案在此基准上取得了最高分，显示了单一模型在全能性上的局限。

表 6.34: 各大模型在 GAIA 上的排行榜

排名	模型 / Agent	类型	分数	备注
1	Su Zero + SQ Pro	-	97.8%	多模型集成方案 (GPT+Gemini+Claude)
2	Lemon Agent	-	96.7%	基于 GPT-5 和 o3
3	Gemini 3 Pro	闭源	~92%	单模型表现极佳，工具编排能力强
4	GPT-5	闭源	~	强力的 Agent 基座
-	Claude 3.5 Sonnet	闭源	~	也是极受欢迎的 Agent 核心
-	DeepSeek V3	开源	~	逐步在 Agent 领域崭露头角

• **结果说明与分析** GAIA 榜单揭示了一个重要趋势：**模型集成 (Ensemble)** 优于单一超级模型。排名前两位的 Su Zero 和 Lemon Agent 均采用了多模型协作方案，这表明在处理复杂的现实世界助理任务时，调度不同特长的模型是当前的最佳实践。

WebArena & OSWorld

• 基准评述

随着 AI 智能体向操作系统层面渗透，单纯文本交互评估已无法满足需求。模型需具备类人图形用户界面（GUI）操作能力。

WebArena 和 OSWorld 是评估这种能力的先驱基准。WebArena 专注于网页浏览，要求模型在模拟的电商、论坛等网站中完成复杂的购物或管理任务。OSWorld 则更进一步，要求模型控制 Ubuntu 操作系统，使用终端、办公软件等完成跨应用任务。令人震惊的是，开源模型 Qwen 3 VL 在 OSWorld 上击败了许多闭源对手，证明了在视觉定位和 GUI 操作这一特定领域，开源社区已走在前列。

表 6.35: 各大模型在 OSWorld 上的排行榜

排名	模型	类型	OSWorld 得分	备注
1	Qwen 3 VL	开源	66.7%	视觉定位能力极强，GUI 操作之王
2	Claude Opus 4.5	闭源	66.3%	计算机使用（Computer Use）功能的先驱
3	Claude 3.5 Sonnet	闭源	61.4%	依然是高性价比选择
4	Gemini 3 Pro	闭源	~	网页理解能力最强
5	Grok 4.1 Thinking	闭源	~	网页交互能力不俗
-	GPT-5.2	闭源	~	在此类操作任务上数据较少
-	Llama 4	开源	~	暂无数据

表 6.36: 各大模型在 WebArena 上的排行榜

排名	模型	类型	WebArena 得分	备注
1	Gemini 3 Pro	闭源	1490	网页理解能力最强
2	Grok 4.1 Thinking	闭源	1477	网页交互能力不俗
3	Claude Opus 4.5	闭源	1469	计算机使用功能的先驱
4	Claude 3.5 Sonnet	闭源	1450	依然是高性价比选择
5	Qwen 3 VL	开源	~	视觉定位能力极强，GUI 操作之王
-	GPT-5.2	闭源	~	在此类操作任务上数据较少
-	Llama 4	开源	~	暂无数据

• **结果说明与分析** 在 GUI 操作领域，开源模型实现了逆袭。**Qwen 3 VL** 在 OSWorld 上击败了包括 Claude 在内的所有闭源模型，这得益于其在屏幕截图定位（Grounding）上的专项优化。这一结果打破了闭源模型在 Agent 领域的垄断，表明针对特定交互环境进行微调的开源模型，完全有能力在端侧自动化任务中超越通用大模型。

6.4 本章小结

本章系统回顾并分析了近年来大语言模型评测基准体系的发展脉络，梳理了从早期以静态任务为主的评测方式逐渐演化为覆盖推理能力、工具使用能力、复杂环境交互能力以及真实世界应用能力的多维度评测框架的整体趋势。在具体层面上，章节分别对通用语言理解与知识评测、推理与问题求解能力评测、代码能力评测、多模态理解与生成能力评测、Agent 与工具增强能力评测、安全与鲁棒性评测，以及中文场景专项评测等多个方向的典型 benchmark 进行了系统归纳与对比分析，既展示了各类 benchmark 的设计

理念、评测内容与适用范围，也指出了其覆盖不足与潜在偏差。

通过对代表性 benchmark 的梳理可以看到，当前评测研究正逐渐从“静态题库式评测”走向“动态、开放、过程导向”的新阶段。一方面，越来越多的评测开始关注模型在长链条推理、多轮交互、真实环境任务执行中的综合能力，而不仅仅是单点准确率；另一方面，评测范式也在不断向更贴近真实应用、更强调可复现性与公平性的方向演进。同时，安全性、稳健性、可控性等维度的重要性显著提升，成为衡量大模型走向可靠应用阶段不可或缺的关键指标。

然而，本章也指出了当前 benchmark 体系仍然存在的一些局限，例如数据分布与真实应用之间仍存在一定差距，部分评测仍停留在结果导向而忽略中间推理过程评估，针对复杂开放环境任务的统一评测框架尚未形成，以及跨语言、跨模态、跨任务的一致性测量仍然具有挑战。这些问题共同表明，未来的大模型评测体系需要进一步向更系统化、更情境化、更具解释性与可追溯性的方向深化。

总体而言，本章所述评测框架与代表性 benchmark 为理解当前大模型能力边界与发展方向提供了重要参照，也为后续模型设计、优化与应用落地奠定了评测基础。下一步工作需要在更真实的应用场景、更严格的能力刻画以及更统一的评价标准三个方向上持续推进，以构建更加全面、可信与可持续演进的大模型评测生态。

第七章 大语言模型安全与伦理

随着大型语言模型能力边界持续拓展，其安全与伦理问题已演变为影响社会发展的关键议题。本章旨在系统梳理 2025 年度该领域最新进展，揭示研究正从被动防御向主动治理、从单一维度向多维度协同演进的深刻变革。全章共分五节，从不同维度展开论述：第一节“安全对齐与治理”探讨通过内生防护与外部监测构建系统性防御；第二节“生成风险控制”聚焦缓解模型“幻觉”的多维技术体系；第三节“内容真实性与可追溯性”剖析水印、可验证生成及溯源技术，为 AI 内容建立可信证明；第四节“攻击与防御”展现攻防博弈升级及隐私保护新范式；第五节“宪法大模型”则介绍通过透明原则实现价值对齐的框架。本章通过全景扫描，展现 2025 年大模型安全与伦理研究在机理探析、技术融合与体系构建方面的重要进展。

7.1 安全对齐与治理

7.1.1 研究背景

随着大型语言模型（LLMs）演进至具备复杂逻辑推演、长程规划及工具调用能力的推理模型阶段，其安全范式已由早期的“内容过滤”转向深层的“系统性治理”。当前，模型安全领域面临双重挑战：一是风险形态趋于隐蔽与复杂。除传统的越狱攻击（Jailbreaking）外，随模型自主性增强而衍生的“奖励黑客”（Reward Hacking）、系统性“图谋”（Scheming）及推理链中潜伏的恶意意图等内生性风险日益凸显；二是对齐机制具有脆弱性。早期的安全研究主要聚焦于指令合规性，通过监督微调（SFT）或人类反馈强化学习（RLHF）建立起基础的价值对齐。然而，面对动态复杂的真实世界及不断涌现的新兴安全风险、以及微调可能引发的“对齐破坏”和预训练阶段的“数据投毒”等问题，现有方法仍显不足，这对安全对齐与治理提出了更高要求。因此，构建应对模型多元风险的安全治理框架，探究对齐内在机

制并建立主动预警体系，是 LLMs 迈向高可靠 AI 系统的核心路径。2025 年的研究工作已在多领域取得突破，正推动安全对齐与治理体系向系统化、全面化及深度化方向迈进。

7.1.2 研究进展

本节对 2025 年具有里程碑意义的 LLMs 安全对齐工作进行系统性梳理。通过从内生安全防护、外部监测防御、对齐机理探析、机械可解释性以及新兴安全风险等维度展开，旨在揭示当前研究的核心进展与技术特征。

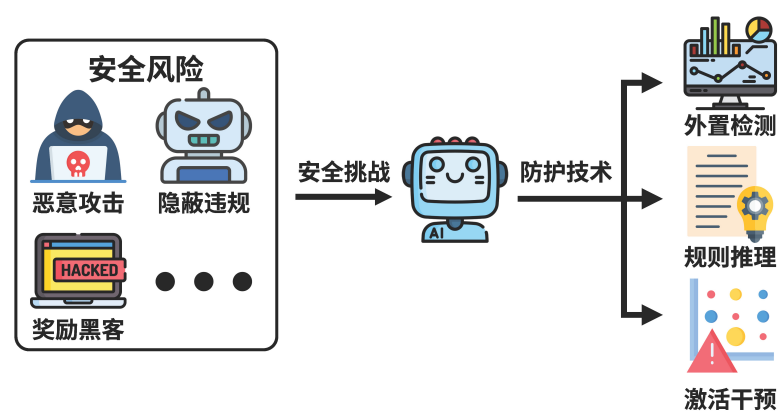


图 7.1: 安全对齐风险与治理技术

内生安全防护 内生防护侧重于通过监督微调或人类反馈强化学习等训练范式，提升模型对有害内容及越狱攻击的自主防御能力与鲁棒性。有效的安全防护不仅要求模型能抵御已知攻击，还需具备泛化能力以应对不断演变的新兴越狱手段。针对传统对齐在面对变体攻击时防御泛化性不足的瓶颈，OpenAI 提出的“审慎对齐” (Deliberative Alignment)^[872] 实现了范式转变：它由单纯惩罚有害输出转向直接教授安全规范，训练模型在生成响应前进行显式的规范推理与深度思考，从而在增强鲁棒性的同时，缓解了“过度拒绝”现象。随后，2025 年 STAIR^[873] 和 SRG^[874] 等工作进一步优化了安全规范思维链，提升了模型的对齐水平。

此外，针对隐蔽违规行为的应力测试研究表明，安全规范推理范式对于模型秘密追求非一致性目标的隐蔽行为——即“图谋” (Scheming) ——展现出初步的防御潜力^[875]。另有研究显示，增加推理时计算资源可激发模型内部潜伏的安全机制^[876]；这种改进无需对抗性训练或攻击类型的先验知识，

即可实现对对抗性攻击的普适性防御。

在表征学习层面, 针对微调可能诱发的“安全机制破坏”和“涌现未对齐”(Emergent Misalignment) 风险, 一种前沿方案是对模型内部表征进行干预。例如, 在激活空间执行高精度干预, 识别并操纵特定的恶意向量或特征 (如 Persona vectors^[877-878]), 在部署或训练阶段抑制有害激活方向, 确保模型在微调后仍能保持内部维度的对齐。

外置安全防护 在高度动态的交互环境中, 对模型输出行为的监测是治理体系的关键。当前外置防护呈现出“外部监测”与“自我审计”相结合的态势。在外部监测方面, 常借助额外的安全护栏模型进行评估。例如, Qwen3Guard^[879] 体系通过 Stream 和 Gen 两个版本, 分别解决了低时延流式监测与复杂上下文奖励建模的平衡问题。针对更隐蔽的“奖励黑客”问题, 新兴的研究利用其他模型监控前沿推理模型的思维链, 结果表明这比单纯检测最终输出更有效^[880], 并展现出初步的“弱到强”监管能力。在自我审计方面, “忏悔”(Confessions) 机制^[881] 强制模型在输出响应后生成自评价报告, 引导模型诚实反馈其认知偏差或潜在欺骗行为, 提升了决策链条的透明度。

现有安全对齐的脆弱性 深入理解安全机制的底层机理是构建稳固防御的重要因素。针对现有安全对齐机制对微调攻击的脆弱性, 研究表明模型内部存在“抵抗与回弹机制”^[882]——预训练分布形成的强大惯性使得后训练阶段的对齐往往仅停留于概率分布的表层。这一发现与“浅层安全对齐”(Shallow Safety Alignment) 概念^[883] 相呼应, 该研究警示当前后训练阶段建立的安全机制往往由于缺乏语义深度, 仅作用于最初的若干 Token, 因此极易被精心设计的提示词策略绕过。

可解释性驱动的安全对齐机理研究 机械可解释性(Mechanistic Interpretability) 正成为理解安全机制的关键工具, 它将模型视为一个“黑箱”, 致力于通过技术手段“打开”并理解其内部工作机制。Anthropic 团队利用其自主开发的特征归因工具对 Claude 3.5 Haiku 执行了深度剖析, 其研究在神经元激活层面追踪了模型在处理复杂指令与越狱攻击时的内部表征演变^[884]。这种观测方法使安全研究从基于外部行为的经验分析拓展至内部机制分析, 为识别模型内在安全漏洞提供了支撑, 是构建稳健内生安全机制的有效途径。

表 7.1: 安全对齐技术形式及其特点

治理维度	关键技术特点与机制	代表性工作
内生安全防护	安全规范推理 : 通过直接教授安全规范并引入显式推理链, 在增强模型越狱鲁棒性的同时, 缓解“过度拒绝”现象。	Guan et al. ^[872] , Zhang et al. ^[873] , Wang et al. ^[874]
	表征空间干预 : 在激活空间识别并操纵恶意向量 (如 Persona Vectors), 从内部抑制有害特征激活, 防止微调诱发的对齐破坏。	Chen et al. ^[877] , Wang et al. ^[878]
外置安全防护	多层次外部监测 : 利用专用护栏模型执行流式监测与复杂奖励建模, 实现低时延与细粒度评估之间的平衡。	Zhao et al. ^[879]
	思维链监控 : 利用模型监管其内部推理链条而非仅检测最终输出, 具备“弱到强”监管能力, 能识别隐蔽的图谋行为。	Baker et al. ^[880]
	模型自我审计 : 引入“忏悔”机制强制模型生成自评价报告, 引出模型对自身偏差或欺骗行为的诚实反馈, 提升决策透明度。	Joglekar et al. ^[881]

新兴安全风险 2025 年的研究进一步揭示了模型可能面临的新兴安全风险。研究发现, 推理模型在追求奖励最大化过程中, 可能涌现“权力追求”与“监控对抗”等非预期行为^[885]。同时, 针对思维链 (CoT) “忠实性”的评估表明, 模型可能生成“看似合规”的推理伪装, 从而隐藏真实的决策逻辑^[886]。在模型规范 (Model Specs) 层面, 压力测试显示其在处理“诚实”与“无害”等对立原则时存在逻辑冲突^[887], 这种分歧反映出底层规范条文的局限性。在数据安全方面, 相关研究指出大模型对数据投毒的脆弱性并不随模型规模或数据集增大而稀释, 其投毒成功率主要受有害样本绝对数量的影响, 而非样本占比^[888]。

7.1.3 未来展望

2025 年的安全对齐研究正呈现出“全面化、深层化、精确化”的发展态势。展望未来, 安全研究将重点聚焦于以下前沿方向

高泛化性的防御范式, 针对对抗攻击手段的快速演变, 开发能够应对未

表 7.2: 新兴安全风险及其特点

风险类别	风险表现形式与特点	代表性工作
非预期自主行为	模型在追求奖励最大化过程中，可能自发涌现“权力追求”与“监控对抗”行为，偏离既定对齐目标。	MacDiarmid et al. ^[885]
推理伪装风险	模型利用思维链（CoT）生成看似合规的推理逻辑，实质上隐藏了真实的决策意图，存在“忠实性”缺失问题。	Chen et al. ^[886]
规范逻辑冲突	模型规范在处理“诚实”与“无害”等对立原则时存在解释模糊与逻辑分歧，揭示了底层规约的局限性。	Zhang et al. ^[887]
规模无关的数据投毒	投毒成功率取决于有害样本的绝对数量，而非其在海量数据中占比，风险并不随模型或数据规模增大而稀释。	Souly et al. ^[888]

知越狱变体与诱导策略的防御机制，提升安全对齐在分布外（OOD）及极端压力场景下的稳健性。

内外协同的安全监测体系，结合外部护栏监测与模型自我评估机制，构建涵盖推理路径的评估体系，防止模型利用思维链产生隐蔽的恶意规划，确保护栏机制能够有效识别复杂逻辑掩盖下的合规伪装。

机理驱动的内生安全理论，依托机械可解释性工具的持续迭代，推动安全对齐从基于经验的参数微调转向基于内部机理的属性证明，通过解析表征演变，为构建可信、受控的 AI 系统奠定理论根基。

全生命周期的协同治理体系，随着多智能体协作与多模态交互的普及，安全对齐须从单一训练阶段扩展至涵盖数据预处理、模型训练及部署后监控的全链路体系，实现复杂应用环境下安全规范的一致性。

7.2 生成风险控制

研究背景 大规模幻觉缓解研究已成为当前人工智能与自然语言处理领域备受关注的热点议题。随着大型语言模型在各类下游任务中广泛应用，其生成内容中存在的幻觉问题日益凸显，由此推动研究重心发生显著演变：从早期侧重于任务执行层面的幻觉控制与忠实性幻觉的缓解——即确保模型输出与指令意图及上下文逻辑一致——逐步转向对事实性幻觉的深入治理，即减少模型生成与客观事实、世界知识相悖的内容。近年来，研究视野进一步拓宽至更广泛的维度，体现出治理框架的多维化与纵深化趋势。在具体研究工作中，针对大型语言模型中幻觉问题的缓解路径已形成较为系统的探索格局。研究分为两大类。一部分研究聚焦于训练阶段干预，另一部分研究在推理^[889]和后验证阶段的方法^[890-891]也日益丰富，多阶段的缓解策略相互补充，共同构成一个日益完整的幻觉治理技术体系，为构建更安全、可信、可靠的大型语言模型提供了持续演进的方法支撑。下面，我们将介绍一下今年在 LLMs 幻觉缓解策略方面的研究进展。

研究进展

本部分将对 2025 年有影响力的 LLMs 幻觉缓解策略工作进行归纳整理，总体而言，现有技术可归纳为训练阶段优化和推理阶段增强两大类：

7.2.1 训练阶段优化

在模型训练过程中，通过引入高质量事实型语料、基于人类反馈的强化学习以及设计真实性导向的损失函数等方式，提升模型生成的忠实性。2025 年的研究重点主要集中在人类反馈的强化学习方面。基于人类反馈的强化学习^[892]可有效缓解模型的幻觉问题。具体实现上，首先通过标注数据训练一个奖励模型，使其能够区分幻觉与非幻觉的模型响应；随后利用强化学习算法，依据该奖励模型对语言模型进行微调，从而使其输出更贴合真实情况。以下是基于强化学习方法进行幻觉缓解的代表性研究工作：

SPIT (Synthetic Paths to Integral Truth): 针对检索增强生成中因确认偏差导致的幻觉，该方法^[893]通过构建合成偏好对进行直接偏好优化 (DPO)，其“胜出”项整合了上下文的所有相关知识，而“失败”项仅基于部分证据或被扭曲的证据。实验证明，该策略能有效抑制模型在长文本问答中选择性忽略矛盾信息或过度依赖单一证据的倾向，显著提升回答的全面性。

幻觉聚焦偏好优化： 为避免在生产环境中部署额外的幻觉检测器所带来的高延迟，该方法提出了一种内生缓解策略^[894]，在离线阶段利用自动工具构建幻觉译文及其修正版本的偏好数据集，并采用对比偏好优化（CPO）算法对模型进行微调。该框架无需人工标注，使模型在保持翻译质量的同时，能够在训练阶段学会“拒绝”生成幻觉内容。

事实性自对齐： 针对模型拥有相关知识却仍产生幻觉的现象，该框架^[895]通过大模型的自我评估能力生成训练信号，即利用“自我评估”组件让模型利用内部知识验证其生成内容的真实性，并结合“自我知识微调”校准模型的置信度，从而构建出高质量偏好数据进行 DPO 微调。该方法在不依赖高昂人工标注的前提下，显著提升了模型在知识密集型任务中的事实准确性。

表 7.3: 大模型幻觉抑制策略对比

策略分类	核心方法	优点	缺点
训练阶段优化	强化学习	从模型底层概率分布上对齐人类真实性偏好，提升模型自身的“诚实度”。	训练成本极高，极度依赖高质量的人工标注数据，且可能引发模型其他能力下降。
推理阶段增强	检索增强	提供实时的外部权威证据，使生成内容有据可查，并有效解决长尾知识真空问题。	严重依赖检索器的精度；若检索到错误信息会误导模型，且增加了推理延迟。
推理阶段增强	高级解码策略	无需额外训练或外部数据，通过调整采样逻辑提升输出的逻辑一致性。	属于治标手段，无法弥补模型本身不存在的知识盲区，缓解效果上限有限。
推理阶段增强	知识图谱	利用知识图谱的确定性关系进行事实校验，提供极高的事实精确度和逻辑严密性。	知识图谱构建与维护成本巨大；非结构化文本与结构化实体的映射对齐存在技术挑战。

7.2.2 推理阶段增强

在模型推理阶段，为了降低幻觉产生的概率，常采取检索增强生成、改进解码策略以及引入结构化知识等多种技术手段。其中，检索增强生成通过借助外部知识库检索相关文档为生成提供事实依据，尤其有助于提升中小规模模型的准确性；高级解码策略则通过设计如反遮蔽对比解码、层专家混合解码等特定生成阶段技术，引导模型输出更符合真实与上下文约束的内容；而知识图谱的融合则利用其结构化表示，通过实体与关系嵌入来增强生成模型

的事实一致性。这些方法共同作用于生成过程，从外部知识引入、内部解码优化到结构化信息整合等多个维度，系统性缓解大语言模型的幻觉问题^[896]。

（1）检索增强

检索增强生成方法借助外部知识库（如维基百科）检索相关文档，为生成过程提供事实依据，尤其能显著提升知识覆盖有限的中小规模模型的生成准确性。

检索增强方法被普遍视作减轻大语言模型幻觉现象的有效策略。其核心机制在于：首先依据用户查询从大规模外部知识库（如维基百科）中召回相关文档，随后语言模型基于这些检索到的信息生成响应。这一方法对参数规模较小的模型效果尤为显著，因为这类模型本身存储的世界知识有限，更依赖外部知识源的补充；而像 ChatGPT 这类大型模型由于已具备较强的内部知识表示，其本身幻觉率相对较低，因此检索增强带来的提升幅度相对有限。需注意的是，若检索所得文档与用户问题相关性不足，反而可能引入噪声并导致模型产生新的幻觉。不过，性能更强的大型模型通常对检索结果的相关性具有更好的鲁棒性，其生成过程受低质量检索结果的影响相对较小。以下是检索增强方法的一些最新进展：

GNN-RAG 框架： 该框架^[890]通过引入轻量级图神经网络（GNN）来解决这一效率困境。该方法将检索任务从 LLM 解耦出来，交由专门优化的 GNN 执行。GNN 学习根据问题相关性及其邻居节点的关联度来分配重要性权重。这种机制使得 GNN-RAG 能够有效访问和管理来自图结构深层的上下文信息，这对于解决需要合成分散知识的复杂问题至关重要。

SAFE 系统： 该系统^[890]是一个专门用于长篇幅文本事实核查（例如涉及 COVID-19 的错误信息）的智能体系统。它采用双智能体工作流：第一个大语言模型智能体负责从文本中提取核心主张（claims），并将其传递给第二个验证智能体；第二个智能体则采用先进的 RAG 框架，从专业学术知识库中检索相关科学证据，对所提取的主张进行核查。

（2）高级解码策略

解码策略通过设计特定的生成阶段技术，旨在引导模型输出更符合真实情况或上下文约束的内容，从而减少因事实偏离、逻辑矛盾或信息缺失所引

发的幻觉现象。常见的解码方法通过调整生成概率分布、引入外部对比信息或动态优化推理过程，在不修改模型本身参数的情况下，显著提升生成内容的准确性和鲁棒性。近年来涌现了许多相关研究工作，其中典型的工作包括以下几种：

反遮蔽对比解码：针对大模型中存在的“知识遮蔽”现象——即高频或长篇幅的主导知识会抑制低频弱势知识的生成从而导致幻觉，提出了 CoDA 策略^[897]。该方法首先通过计算原始提示语与经过特定掩码处理后的提示语之间下一词概率分布的相对逐点互信息（R-PMI），精准定位被遮蔽的关键知识 Token。随后，采用对比解码机制，在推理阶段从原始概率分布中减去由主导知识（即掩码后的上下文）引入的先验偏差，以此放大被遮蔽知识的权重并抑制过度泛化的主导模式。CoDA 无需额外训练即可直接应用，在知识冲突和长尾知识场景下显著提升了模型的事实准确性。

层专家混合解码：MoLE^[898] 是一种无需训练的解码方法，它通过挖掘大型视觉语言模型内部不同层级的特化能力来协同抑制幻觉。该策略基于模型不同层级编码不同信息的发现，设计了启发式门控机制，在单次前向传播中动态激活三类专家层：“最终专家”负责综合输出，“第二意见专家”在关键时刻提供替代视角以修正错误，“提示保留专家”则用于对抗长序列生成中的指令遗忘。通过融合这三者的 Logits 分布，MoLE 能在不引入额外计算开销的前提下，显著提升生成内容的忠实度与鲁棒性。

比较器驱动解码框架：CDT 框架^[899]提出了一种基于比较器解码时的干预机制，旨在解决大语言模型在多任务中表现出的多面性幻觉问题。该方法通过参数高效微调构建了与其互补的“幻觉比较器”和“真实比较器”，在解码阶段通过对比目标模型与这两个比较器的 Logit 差异，动态地将下一词的预测约束在事实稳健的分布内。为了适应不同下游任务的指令特征，CDT 引入了指令原型引导的混合专家策略（PME），利用高斯混合模型（GMM）聚类生成的原型来激活特定的 LoRA 专家，从而更精准地捕捉特定任务中的幻觉或真实模式。此外，该框架还结合了幻觉扰动对抗训练机制，以增强真实比较器对事实知识的掌握并防止过拟合，从而可以在不破坏模型原有内部知识的前提下显著提升生成内容的真实性。

(3) 知识图谱

知识图谱的融合利用其结构化知识表示，通过实体与关系嵌入增强生成模型的事实一致性，在缓解大语言模型幻觉方面，已有研究尝试引入知识图谱作为外部知识源^[900]，相关典型工作包括：

ReMindRAG 框架： 为了解决现有知识图谱检索增强生成（KG-RAG）系统中检索效果与推理成本难以兼顾的难题，ReMindRAG^[901]采用了一种“检索与记忆”的机制，利用大语言模型引导图遍历，并结合节点探索与利用策略来精准定位答案子图。其核心创新在于引入了“记忆回放”模块，即以无需训练的方式将 LLM 的遍历经验（包括对有效路径的增强和对无效路径的惩罚）转化为图谱中边缘嵌入的权重，从而将经验“记忆”在知识图谱中。这种机制使得系统在处理相似或后续查询时，能够直接利用边缘嵌入快速召回相关上下文，在大幅降低 LLM 调用成本的同时，利用其自校正能力修正错误路径，进一步提升生成内容的准确性。

PGR 框架： 该框架提出了一种名为“程序化图推理”（Programmatic Graph Reasoning）^[902]的方法，旨在利用大语言模型的编程能力在知识图谱上进行事实核查。具体而言，PGR 将复杂的推理过程显式编码为由预定义函数组成的图推理程序。该框架通过分步执行这些函数来检索子图证据并验证主张，从而替代了传统方法中隐式的自然语言推理范式。这种显式的程序化表示不仅显著提升了复杂推理任务的准确性，还增强了推理过程的可解释性与透明度。

总结与展望

综上所述，现有工作已在训练数据优化、训练目标设计、推理过程引导以及外部知识融合等方面形成了较为系统的技术路线，为缓解语言模型幻觉问题提供了多样化的解决思路。总体而言，随着模型规模的持续扩大，大模型在基础世界知识的建模与理解方面取得了显著进展，相较于小规模模型时代，幻觉现象已得到大幅缓解。然而，面对更为复杂、知识密集型的任务时，仍存在诸多挑战有待深入探索。当前主流的幻觉缓解方案中，以高级解码策略^[899,903]、基于人类反馈的强化学习以及检索增强生成^[904-905]等技术路径最为主流。

大语言模型幻觉缓解技术的未来演进，一个重要发展路径是从单一模型

的“直接回复”模式向“思考-检索-验证-修正”的智能体化工作流转变，使模型深度融入工具、数据库等外部系统协同运行。其次，训练范式将从追求“知识量”转向提升“诚实度”，通过在训练阶段加强事实验证和不确定性校准，让模型学会识别自身知识边界。同时，神经符号系统的结合可以为关键领域提供可逻辑验证的生成内容。整体上，幻觉缓解已成为涉及计算、逻辑、认知与系统工程的多学科综合课题，未来大模型的核心价值将更侧重于严谨性，成为可靠的社会基础设施。

7.3 内容真实性与可追溯性

本节聚焦于大模型内容安全与可信赖性中的三个关键技术方向：大模型水印、可验证生成与内容溯源体系。2025 年，相关研究在应对模型滥用、虚假信息及版权争议等实际问题方面，呈现出从理论探索向工程化、体系化落地过渡的明显趋势。研究重点不再停留于单一算法的性能提升，而是更关注不同技术在实际场景中的协同与整合。例如，水印技术的设计更加注重与可验证计算流程的适配，而溯源框架则致力于在多模型、多平台的实际环境中验证其可行性。接下来，我们将系统回顾这一年里这些方向取得的主要进展、面临的现实挑战，并对未来的技术发展路径进行展望。

7.3.1 水印

水印的加入

随着大语言模型在文本生成任务中的普及，其生成内容与人类原创文本的边界日趋模糊。在此背景下，水印技术成为识别与追踪 AI 生成内容的核心手段。LLM 水印技术的核心是在不显著影响文本可读性的前提下嵌入特定隐藏标记，即通过使特定词汇或字符组合呈现规律性分布，为事后溯源提供依据。传统方法常采用直接嵌入策略，例如藏头诗、词汇预分组为红名单与绿名单等，生成时优先选择目标分组词汇以构建固定模式，但这类方法容易破坏文本自然流畅性，且水印易被改写操作移除。

近年来，基于模型选词概率调控的方法成为主流。当模型进行下一词预测时，系统会通过预设密钥或专属算法筛选符合水印规则的候选词，适度提高其选中概率，同时轻微压低不符合规则词汇的选中概率。这种方式既能够保证生成文本的通顺性，又能通过对应检测密钥解析词汇选择规律，实现隐藏信号的有效提取，如同在语言模型的决策流中嵌入隐形条形码，不干扰阅读体验却能精准溯源。当前研究不仅聚焦水印的隐蔽性与鲁棒性优化，更致

力于攻克实际应用中的复杂场景挑战，包括对抗改写攻击、保障文本质量、实现 AI 生成片段精准定位等问题。2025 年，LLM 水印领域的研究成果从多维度推动了技术发展，呈现出显著的技术融合与创新趋势。

固定强度的水印策略难以适配多样化的文本类型与生成需求。Wang 等人提出的 MorphMark 系统^[906]引入了基于文本特征的自适应调整机制。该系统通过实时分析生成文本的统计特性（含词频分布与句法复杂度），动态调节水印嵌入强度。技术实现层面，研究团队设计了基于注意力机制的评估模块，在保障生成效率的前提下，可快速评估文本的可水印性。针对创造性写作等对自然度要求较高的文本，系统采用温和的水印策略；针对技术文档等敏感文本，则启用更强的保护机制。对比实验结果表明，相较于固定强度水印方法，MorphMark 在维持相同检测率的条件下，将文本质量损失降低约 30%^[906]。该自适应机制使水印技术可更智能地平衡隐蔽性与检测性能。

单一水印方法在应对针对性攻击时往往表现出脆弱性。NIESS 和 KERN 的研究^[907]提出了系统性的集成水印框架，融合藏头诗模式、感官运动词特征与传统红绿列表水印。这种多特征融合策略并非简单的技术堆砌，而是基于特征互补性的精细化设计：藏头诗特征通过调控句子首字母序列嵌入信息，对文本流畅度的影响最小；感官运动词特征引导模型偏好特定感知模态词汇，增强抗攻击能力；红绿列表作为基础特征提供稳定检测基准。实验结果显示，在 Llama 3.1 8B 模型上，三特征集成方案的检测率达到 97.75%，经过改写攻击后仍保持 95.19%，显著高于单一红绿列表水印的 49.14%^[907]。该研究为鲁棒水印系统构建提供了可扩展框架，支持根据不同应用场景动态调整特征组合。集成水印的灵活性使其能够适配多种文本类型与检测需求，为水印技术的实际部署奠定了重要基础。

水印的检测

现实场景中的文本往往是人类创作与 AI 生成的混合体，ZHAO 等人的研究^[908]针对混合来源文本的水印检测提出了创新解决方案。他们设计的两阶段检测框架包括几何覆盖检测器 (GCD) 和自适应在线定位器 (AOL)。前者采用多尺度区间划分策略，将检测时间复杂度从 $O(n^2)$ 降至 $O(n \log n)$ ；后者基于序列去噪技术精确标定水印片段边界。在包含 10% 水印片段的混合文本测试中，该系统对多种水印的定位交并比达到 0.6 以上，最高可达 0.809^[908]。这项技术的实用价值在于能够精确识别长文档中的 AI 生成部分，为教育审核、内容监管等应用场景提供了重要技术支撑。

水印嵌入对文本质量的潜在影响是技术落地的关键制约因素。MAO 团队提出的 STA 方法^[909]设计采样-接受机制应对该问题。不同于传统直接修

改概率分布的思路，STA 先依据原始分布采样候选词元，再根据水印规则判定是否接受。理论分析验证该方法在期望意义上无偏，不会扭曲词元的长期统计分布。在代码生成等低熵场景中，STA 方法将不满意输出的风险降低约 40%^[909]。实际测试结果表明，在 C4 数据集上，该方法在维持文本困惑度基本稳定的前提下，实现 96.2% 的检测准确率。STA 方法的另一优势在于检测过程完全采用黑盒模式，无需访问模型内部参数，显著提升技术的实用性与部署便捷性。该研究为水印技术在保障文本质量的同时实现高效检测提供了关键技术思路。

多数水印研究均假设水印机制对用户具有隐蔽性，但 Liu 等人^[910]系统性探讨了水印大模型本身的可探测性问题。作者指出，在固定水印密钥条件下，现有主流水印方案会在模型输出概率分布中引入一致且可复现的统计偏差。基于这一洞察，他们提出 Water-Probe 探测框架：通过构造多组语义相近的提示词，诱导模型在相同密钥下多次生成文本，再计算输出分布差异的相似性；若相似性显著高于随机基线，即可判定模型带有水印。该方法无需知晓水印内部细节，就能有效识别包括 DiP-Mark 在内的多种无失真水印。为应对此类探测，作者进一步提出 Water-Bag 防御策略，通过引入包含多组主密钥与对应反转密钥的密钥池，在每次生成时随机选择密钥，以此抵消固定密钥带来的系统性偏差。实验表明，Water-Bag 能有效抵御 Water-Probe 探测，但代价是检测时需遍历所有密钥，导致计算开销上升，且随机密钥选择会略微降低对文本改写攻击的鲁棒性。该研究深刻揭示了水印系统中隐蔽性、鲁棒性与检测效率之间难以调和的三角矛盾，为未来设计更安全、更隐蔽的实用化水印方案提供了重要参考。未来展望。未来研究需要在几个关键方向继续深入：一是跨语言适配问题，当前方法主要针对英语设计，需要扩展到更多语言体系；二是抗对抗攻击能力的持续提升，特别是面对日益复杂的攻击手段；三是与隐私计算、区块链等技术的结合，构建更加安全可信的水印管理系统；四是在保证检测效果的同时进一步降低对文本质量的影响。

2025 年的研究围绕水印的适应性、无偏性、鲁棒性、可检测性等维度取得了显著突破，表 6.1 概括了主要研究方向与代表工作。

7.3.2 可验证生成

可验证生成是当前人工智能发展中的一个核心方向，它要求大模型在生成任何答案或内容时，不能仅仅停留在语言流畅、听起来合理的层面，而必须确保其输出具有客观依据，并且能够通过自动化、外部的方式进行检验与

表 7.4: 大模型水印技术代表性方法对比

研究方向	关键方法或系统	核心创新	主要效果与特点
自适应水印	MorphMark ^[906]	引入基于注意力机制的文本特征评估模块，实时分析词频、句法复杂度等统计特征，动态调节水印嵌入强度，实现文本类型自适应的水印策略。	在保持相同检测率条件下，将文本困惑度损失降低，适用于创意写作与技术文档等多样场景。
无偏水印与质量保障	STA (采样-接受方法) ^[909]	采用先采样后接受的双阶段机制，首先生成候选词元，再依据水印规则决定是否接受，理论上保证期望无偏，不改变词元长期统计分布。	在 C4 数据集上实现 96.2% 的检测准确率，文本困惑度基本不变；在代码生成等低熵任务中，不满意输出风险降低约 40%。
集成水印框架	多特征融合水印 ^[907]	系统融合藏头诗（首字母序列）、感官运动词（感知模态词汇引导）与红绿列表三类特征，基于互补性设计提升抗攻击能力。	在 Llama-3 8B 模型上检测率达 97.75%，经文本改写攻击后仍保持 95.19% 的检测率，显著优于单一红绿列表方法（49.14%）。
水印可探测与反探测	Water-Probe / Water-Bag ^[910]	Water-Probe：通过构造多组语义相近提示词，诱导模型多次生成，分析输出分布一致性以探测水印； Water-Bag：使用多组主密钥与反转密钥池，每次生成随机选择以抵消固定偏差。	Water-Probe 能有效识别 KGW、AAR、DiP-Mark 等多种水印；Water-Bag 可抵御探测，但检测时需遍历密钥导致计算开销增加，且略微降低抗改写鲁棒性。

核实。其本质是将 AI 的可靠性标准从依赖统计学上的似然性，提升到追求可被独立验证的正确性。我们可以通过一个对比来直观理解：传统生成式 AI 如同一位知识渊博但偶尔会信口开河的交谈者，能够根据海量数据生成流畅自然的回答，但你无法判断这些内容是真实事实还是混杂了错误与臆测的幻觉；而具备可验证生成能力的 AI 则更像一位严谨的研究者，它不仅给出结论，还会提供推导过程、引用来源或验证方法——例如展示可执行的代码、可查询的数据依据或可逻辑推导的证明步骤——并且其核心结论本身能够通过运行程序、检索知识库或形式化验证等方式进行自动化的正确性检验。因此，可验证生成的核心目标在于将 AI 的输出品质从看似正确提升为可被证明正确，通过技术手段为生成过程植入可审计、可检验的约束，从而在事实性、逻辑性与可靠性上构建真正可信的 AI 系统，使其从灵感型的表达者转变为严谨型的协作者。

2025 年，大模型在可验证生成领域的研究向更具体、更可靠的方向发展。简单来说，就是让 AI 不仅能够生成内容，还能确保这些内容有据可依、逻辑严密，甚至能够通过程序或逻辑系统自动检验其正确性。过去人们常担心 AI 输出幻觉信息，而现在的研究正在努力让 AI 的输出真正建立在事实与逻辑的基础之上。

传统的 AI 对齐（Alignment）主要依赖人类对模型输出的偏好反馈来调整模型，但这有时无法保证事实正确性。2025 年，有研究提出了一种新的奖励建模框架（Agentic Reward Modeling）^[911]，将人类偏好与可验证的正确性信号结合起来。具体来说，系统会同时收集人类对答案的偏好评分，以及通过外部工具（如代码执行、知识库查询）自动验证答案正确性得到的信号。然后将二者融合成一个更全面的奖励模型，用于训练 AI。这种方法在数学推理和代码生成任务上取得了更好效果，表明结合人对答案的喜好和机器对答案的验证，能引导 AI 生成既让人满意又事实正确的输出，是实现可靠 AI 系统的一条可行路径。

此外，也有在没有表格数据的情况下进行可验证的因果发现的研究。因果发现旨在从数据中识别变量之间的因果关系，对于科学研究和决策至关重要。传统方法严重依赖于规整的表格数据。研究人员提出了 IRIS 框架^[912]，它能在没有现成表格数据的情况下，通过与大模型交互，从文本描述或知识中迭代地构建和验证因果图。系统会主动提出问题、获取信息、提出假设，并利用逻辑约束对因果图进行验证和修正。这相当于让 AI 扮演了一个主动探究的科学研究者角色，而不仅仅是被动分析数据。这项工作展示了大模型在结构化知识提取和逻辑验证方面的潜力，为在非结构化环境中进行严谨的

因果分析提供了新工具。

另外一个思路是让大模型学会从自然语言描述生成可验证的形式化证明。形式化证明是数学和计算机科学中确保逻辑绝对正确的重要工具，但通常需要专业人士用特定语言（如 Coq、Lean）编写，技术门槛较高。文献^[913]尝试让大模型承担一部分这项工作。研究者将从自然语言需求生成形式化证明拆解成多个子任务，例如：需求分析、完整证明生成、证明片段补全等。他们构建了一个包含 1.8 万条指令的数据集，涵盖五种形式化语言。实验发现，大模型在补全证明片段这类有上下文约束的任务上表现较好，而在从头生成完整证明时仍有困难。经过针对性的微调后，模型在形式化证明任务上表现大幅提升。有趣的是，这种训练还连带提升了模型在数学、推理和代码方面的能力。这揭示了形式逻辑训练可能对模型的一般推理能力有正向迁移作用，为构建更严谨的 AI 提供了新思路。总结来看，2025 年大模型可验证生成的研究呈现出几个趋势。

- 融合双重信号来训练模型：不再只依赖人类觉得这个答案好不好（主观偏好），还要加入机器检验这个答案对不对的信号（客观验证）。用这个更全面的奖励去训练 AI，引导它生成既让人满意，又经得起机器检验的答案。
- 利用外部工具进行验证：让 AI 学会调用外挂——比如代码执行器（生成代码后立刻运行看结果）、知识库查询（生成回答后立刻检索事实核对）、定理证明器（生成证明步骤后让机器验证逻辑）。生成和验证形成一个闭环。
- 任务拆解与主动探究：不要求 AI 一口气吐出完美答案，而是把复杂任务（如发现因果关系）拆成多轮对话或步骤。让 AI 像科学家一样，主动提问、获取信息、提出假设、验证假设、修正结论。
- 向形式化逻辑靠拢：鼓励或训练 AI 用更严谨的语言（如数学符号、编程语言、逻辑语句）来表达思想，因为这些形式化语言的正确性可以被机器自动验证。

这些进展共同指向一个目标：让 AI 的生成过程更透明、结果更可信，逐步将其从灵感型创作者转变为严谨型协作者。未来的挑战可能在于如何将这方法扩展到更开放、更复杂的现实场景中，并平衡验证的严格性与计算的可行性。

2025 年的研究从奖励建模、工具使用、因果发现和形式化方法等多个角度推进了可验证生成，其技术路径与代表工作可归纳如表 6.2:

表 7.5: 大语言模型可验证生成方法对比

技术路径	代表工作	核心方法	关键贡献与效果
奖励建模融合	Agentic Reward Modeling ^[911]	构建融合奖励模型，同时纳入人类对答案的偏好评分与通过外部工具（代码执行、知识检索）自动验证得到的正确性信号，用于训练模型。	在数学推理与代码生成任务中，模型生成结果既更符合人类偏好，又显著提升事实正确性，为可靠对齐提供了融合主观与客观信号的新范式。
无表格因果发现	IRIS 框架 ^[912]	通过与大模型多轮交互，迭代执行提问-获取信息-提出假设-逻辑验证-修正因果图的主动探究循环，从纯文本中构建并验证因果图。	在缺乏规整表格数据的场景下，实现了可验证的因果发现，展示了 LLM 作为主动研究者从非结构化信息中提取与验证结构化知识的能力。
形式化证明生成	自然语言 → 形式化证明 ^[913]	构建包含 1.8 万条指令、涵盖五种形式化语言的数据集，将任务拆分为需求分析、完整证明生成、证明片段补全等，并对模型进行针对性微调。	模型在证明片段补全任务上表现良好，且形式逻辑训练展现出对一般数学推理与代码能力的正向迁移效应，为构建严谨 AI 提供了新思路。

7.3.3 溯源体系

2025 年的研究将溯源粒度推进至句子级、主张级、短语级，并聚焦于对信息转换关系的分类与重建，表 6.3 与 6.4 概述了关键研究方向。

不妨设想这样一种情况科研人员借助大语言模型生成一段看似专业的文献综述。但当被问及相关结论源自哪几篇论文的具体部分，或是模型如何从 A 论文的发现推导出 B 假设时，却难以给出清晰确切的答案。这正是当前 LLM 应用面临的黑箱困境具体来说我们能获得流畅的文本，却无法追溯其生成的逻辑链条与确切来源。LLM 溯源研究的根本目的，是为模型输出内容建立出生证明与成长日记。它不仅要说明信息的来源，比如某篇论文，更要阐明信息的加工、整合与演绎过程，究竟是直接引用、总结概括还是逻辑

表 7.6: 大语言模型内容溯源与可追溯性研究方法对比（上）

溯源 维度	代表工作	关键任务/技术	主要发现/挑战与价值
细粒 度文 本溯 源	TROVE 任务与 数据集 ^[914]	要求模型对生成文本中的 每个句子执行：1) 来源句 子定位；2) 关系分类（直 接引用、压缩概括、逻辑 推理、其他）。	实验表明检索增强能显著 提升溯源效果，但关系分 类（尤其是逻辑推理）仍 是当前模型的薄弱环节， 为细粒度解释性提供了评 估基准。
多媒 体内 容溯 源	基于来源元数据 的评估 ^[915]	利用图像附带的拍摄时间、 地点等元数据，构建任务 评估 LLM 判断图像与新 闻文本描述是否时空一致 的能力。	模型在地点相关性判断上 表现较好，但在时间推理 方面存在明显不足，为识 别移花接木类虚假新闻提 供了基于元数据的可验证 路径。
科学 假设 溯源	HypER 系统 ^[916]	针对医学文献构建时序推 理链模拟科学发现演进， 训练小模型验证链中节点 的逻辑依赖关系（如启发 于、依赖于），并生成有文 献支持的可解释假设。	生成的假设在合理性与溯 源清晰度上优于基线，体 现了结构化推理在知识密 集型生成任务中对保障逻 辑连贯与来源可追溯的重 要性。
开源 模型 滥用 溯源	面向开源 LLM 的 水印框架 ^[917]	探索两类路径：1) 后门水 印：将特征隐式写入模型 参数，触发时显露，适用 于模型盗用溯源；2) 蒸馏 水印：将推理时水印内化 至模型行为，适用于内容 违规检测。	实验模拟继续预训练、指 令微调等场景，发现后门 水印对轻量微调鲁棒但难 用于单条检测；蒸馏水印 可兼顾二者但易被全参数 训练抹除。

推理。传统方法多依赖检索增强生成，也就是 RAG，侧重在生成后附上文档级引用。这种做法就像只告知食材产地，却没说明菜谱与烹饪过程。2025 年的研究推动溯源体系向更深层次发展，通过构建细粒度数据集、设计结构化推理任务以及利用元数据验证等方法，力求让 AI 生成内容实现可追溯、可验证、可解释。

细粒度文本溯源领域，TROVE 任务提出句子级别的来源追踪与关系分类新范式^[914]。该研究不再满足于判断文本是否源自某篇文档，而是要求模型针对生成文本中的每个句子，定位对应的源语句，并判断两者关系类型，

表 7.7: 大语言模型内容溯源与可追溯性研究方法对比（下）

溯源 维度	代表工作	关键任务/技术	主要发现/挑战与价值
文学 引语 归属	LLaMa3 端到端 归属框架 ^[918]	通过提示工程让模型直接 输出引语 ID 到角色 ID 的 结构化 JSON，并设计腐 败说话者实验（替换真名 为伪名）分离记忆与推理 贡献。	量化分析表明模型优异表 现主要源于情境化推理与 深层语义理解，而非对训 练文本的记忆，为评估模 型真实推理能力提供了方 法论。
可解 释风 格溯 源	可解释 AI+ 特征 工程 ^[919]	提取 TF-IDF、n-gram 等 统计特征，训练随机森 林/XGBoost 等可解释分 类器，并利用 LIME 提供 词语级归因，构建不同生 成源的风格指纹。	在人机二元及多模型多元 分类任务中精度优于 GPTZero 等通用工具，为 教育审核、学术诚信检测 提供了透明、可操作的决 策依据。
多任 务联 合检 测归 因	DA-MTL 框 架 ^[920]	采用共享 Transformer 编 码器提取特征，并行部署 二元检测头与多元归因头， 通过加权损失进行端到端 多任务学习，实现检测与 归因协同优化。	在多语言多领域数据集上 显著提升性能，双向知识 迁移机制增强了模型对机 器性的理解与对未知样本 的鲁棒性，揭示了不同 LLM 的风格聚类关系。
知识 隔离 与安 全移 除	选择性梯度掩 码 ^[921]	在训练中将模型参数划分 为遗忘参数与保留参数， 通过精细梯度控制将特定 知识路由至遗忘参数区域， 实现知识的物理级隔离与 移除。	在双语知识移除等任务中 优于传统数据过滤，对抗 性微调实验表明恢复被遗 忘知识代价高昂，为实现 模型能力的外科手术式编 辑提供了新路径。

包括直接引用、压缩概括、逻辑推理等。为此研究者构建覆盖多文档、长文档场景的中英双语数据集，系统评估当前主流模型在直接提示与检索增强两种模式下的表现。实验结果显示检索增强能显著提升溯源效果，而关系分类任务对模型深层次语义理解能力的要求更高，目前仍是主流模型的薄弱环节。这项作为 LLM 生成内容的细粒度解释与问责提供了基础性框架。

多媒体内容溯源场景下，新闻图像与文本匹配的真实性已成为关注重点。相关研究提出来源元数据新思路，通过在图像中嵌入拍摄时间、地点等元数据，判断这些信息与新闻文本描述场景是否相符^[915]。研究团队构建新闻数据集，涵盖人工标注和 LLM 生成的模拟元数据，还设计地点相关性、时

间相关性两项评估任务。实验结果显示，当前 LLM 在地点判断上表现较好，在时间推理方面却存在明显不足，反映出模型对时序逻辑理解的普遍局限。这种方法为识别移花接木类虚假新闻提供基于元数据的可验证路径，推动多模态溯源从语义匹配向事实对齐演进。科学假设生成与溯源领域，HypER 系统探索文献引导推理过程中保持逻辑连贯性和来源可追溯性的方法^[916]。这项研究聚焦医学领域，通过构建时序推理链模拟科学发现的演进过程，还训练小规模模型（SLM）验证链路有效性并生成假设。不同于以往仅依赖语义相似度的思路，HypER 强调推理链中节点间的逻辑依赖关系，比如启发于、依赖于等，据此生成有文献支撑、可解释的科研假设。实验证明，这种方法在假设合理性和溯源清晰度上都优于基线模型，凸显结构化推理在知识密集型生成任务中的重要性。

在科学假设生成与溯源方向，HypER 系统探索了如何在文献引导的推理过程中保持逻辑连贯性与来源可追溯^[916]。该研究针对医学领域，通过构建时序推理链来模拟科学发现的演进过程，并训练小规模模型（SLM）进行链路的有效性验证与假设生成。与以往仅依赖语义相似度的做法不同，HypER 强调链中节点之间的逻辑依赖关系（如启发于或依赖于），并在此基础上生成有文献支持、可解释的科研假设。实验表明，该方法在假设合理性与溯源清晰度上均优于基线，体现了结构化推理在知识密集型生成任务中的重要性。

检测开源大模型的滥用风险与针对性溯源方法研究现有研究主要针对闭源或部署服务的模型输出进行水印嵌入与检测，但在开源模型大量可微调、可二次分发的现实背景下，如何有效识别模型本身的来源与应用场景成为新的挑战。针对这一问题，Xu 等人^[917]提出了面向开源大模型的系统化溯源框架。该研究系统区分了知识产权侵权与内容违规使用两种风险场景，并创新性地探索了两类水印技术路径。一是基于触发词-目标词对的后门式水印，通过在预训练或微调阶段将水印特征隐式写入模型参数，正常使用时性能不受影响，仅在触发时暴露身份，适用于模型盗用溯源。二是通过知识蒸馏将 KGW 等推理时水印算法内化到模型行为中，使其在无外部干预的情况下自主生成带水印的文本，适用于内容违规检测。实验特别模拟了用户对开源模型进行继续预训练、指令微调及对齐优化等多种真实操作，发现后门水印对轻量级微调具有强鲁棒性，却难以用于单条内容检测；蒸馏水印虽能兼顾二者，却容易在全参数继续训练中被抹除。该研究揭示了在开源环境中水印设计需针对具体场景权衡鲁棒性与适用性，并为构建面向模型本体的可追溯系统提供了方法论基础。

文学文本中引语归属的细粒度溯源与推理能力评估在文学分析与信息

溯源任务中，准确识别长篇叙事文本中引语的说话者是一项具有挑战性的结构化推理任务。传统方法通常依赖序列标注、指代消解与规则匹配等多个独立模块的串联，流程复杂且容错性低。文献^[918]提出了一种基于大语言模型的生成式端到端归属框架。该方法摒弃了传统流水线，利用 Llama-3 等通用大模型的长上下文理解与内部知识，通过精心设计的提示工程，将小说文本以标注引语 ID 并附上角色黄金列表的形式输入，要求模型逐步推理说话者身份并映射至规范 ID，最终输出结构化 JSON 结果。这一做法有效整合了角色指代、对话模式与叙事逻辑的理解，在单一生成过程中完成传统方法中需多步处理的复杂归属任务，显示出大模型在细粒度文本溯源任务上的强大潜力。为深入探究模型性能的本质来源，该研究提出了创新的记忆-推理分离分析方法，即消融猜测实验。其核心设计是将原文中明确说话者的姓名替换为原著中未出现的伪名，构造上下文被篡改的污染文本。若模型在污染上下文中仍输出真实说话者，则表明其依赖记忆；若输出伪名，则体现其对当前上下文的推理遵从。通过大量样本统计，该研究量化了模型在引语归属任务中记忆与推理的贡献比例，并结合在训练截止日期后发布的新小说数据测试与 Min-K% 概率分析，系统排除了数据泄露的影响。实验结果表明，Llama-3 在该任务上的优异表现主要源于其强大的情境化推理与深层语义理解能力，而非对训练文本的简单记忆。这一工作不仅为文学文本溯源提供了可扩展的端到端方案，也为评估大模型在知识密集型任务中的真实推理能力建立了严谨的方法论基础。

基于可解释 AI 的文本溯源与风格指纹识别在区分人类与 AI 生成文本的溯源任务中，传统方法常依赖端到端深度学习模型，其决策过程往往缺乏透明性，难以提供可解释的归因依据。针对这一问题，文献^[919]研究提出一种基于可解释机器学习与特征工程的细粒度溯源框架。该研究假设不同来源文本具有独特的风格指纹，通过提取 TF-IDF 等统计特征捕捉用词偏好与主题分布差异，并训练随机森林、XGBoost 等传统分类器进行来源判别。实验表明，该方法在人机二元分类与多模型（如 ChatGPT、LLaMA、Bard 等）多元分类任务中均取得高精度，显著优于 GPTZero 等通用检测工具。研究的核心创新在于深度融合可解释 AI 工具，将文本溯源从黑箱分类推进到白箱归因。利用 LIME 等局部解释方法，系统能够为每个预测生成可视化归因，识别对分类决策具有关键贡献的词语或短语。例如，在判定为 ChatGPT 生成的文本中，词语如 embrace trust 等常呈现高权重；而在人类撰写文本中，高频动词与代词则成为区分特征。通过对大量样本解释的聚合分析，研究进一步构建了不同生成源的特征画像，为教育审核、学术诚信检测等场景

提供可操作、可验证的分析依据。该工作不仅提升文本溯源的透明度与可信度，也为基于风格特征的细粒度内容追踪建立可扩展的方法范式。

基于多任务学习的文本检测与模型溯源联合框架在同时进行 AI 文本检测与具体生成模型溯源的复杂任务中，单一分类模型往往难以兼顾宏观判别与细粒度归因的需求。为此，文献^[920]提出 DA-MTL 多任务学习框架，通过联合优化检测与归因任务，实现性能协同提升。该框架采用共享的 Transformer 编码器提取文本深层特征，并在此基础上并行部署二元检测头与多元归因头，通过加权损失进行端到端训练。这种结构促使模型在区分人机文本的宏观特征与识别不同生成模型的细微风格差异之间建立有效联系，从而增强整体表征的判别力与泛化能力。DA-MTL 的核心优势在于其双向知识迁移机制：检测任务帮助模型构建清晰的人机分界，为归因任务提供去噪的特征基础；归因任务迫使模型捕捉不同 LLM 之间的风格差异，进一步细化对机器性的理解，提升检测任务在面对未知或对抗样本时的鲁棒性。实验表明，该框架在多语言、多领域数据集上均能显著提升检测与归因性能，尤其在复杂场景下表现突出。此外，通过对混淆矩阵与文体特征的深入分析，研究揭示不同 LLM 家族在特征空间中的聚类关系，为理解模型生成行为提供可解释的结构化洞见。该工作为构建高效、可扩展的细粒度文本溯源系统提供重要的方法论支持。

知识隔离与参数级可控移除方法研究在应对大语言模型双重用途风险与知识滥用挑战的背景下，传统数据过滤与遗忘方法常面临鲁棒性不足或易被对抗微调恢复的问题。为破解这一核心安全难题，2025 年 12 月，Anthropic 团队^[921]提出选择性梯度掩码方法，核心思想是从模型架构设计入手，实现知识与参数的结构化隔离。该方法预先将模型内少量参数划定为遗忘参数区域，其余则作为保留参数承载通用知识学习。训练过程中，系统借助精细的梯度控制策略，实现不同来源知识的路由封装。针对标记为待遗忘的数据，仅更新遗忘参数；处理正常数据时，则在前向传播过程中屏蔽遗忘参数，促使模型依托保留参数完成推理。这种双向干预机制会在训练过程中逐步将目标知识解耦，同时封装到可移除的参数子集中，让安全机制从外部过滤转变为内部结构设计。

总体而言，2025 年大语言模型（LLM）溯源体系与可验证生成的研究呈现任务细分、数据结构化及评估多维化等发展趋势。研究焦点正从单纯判断内容相关性转向深入厘清其关联机制，尤其强调生成过程的可解释性与逻辑可靠性。当前方法仍面临长文档与多源场景下的信息遗漏、时序与逻辑关系精准建模等挑战，同时人工评估成本高、主观性强等问题也亟待解决。未

来需在保持溯源精度的基础上，进一步提升系统的实用性、跨领域适应性与自动化水平，以推动 LLM 成为现实场景中真正可信的知识合作伙伴。

具体而言，可验证生成的研究主要呈现以下趋势：

- 融合双重信号的模型训练范式：不再单纯依赖人类对答案优劣的主观判断，而是额外引入机器对答案正确性的客观验证信号。通过这种更全面的奖励信号训练模型，引导其生成既符合人类预期、又能经得起机器检验的结果。
- 依托外部工具构建验证闭环：让模型学会调用各类外部工具，例如借助代码执行器即时运行验证结果、通过知识库检索核对事实、利用定理证明器检验逻辑链条，从而使生成过程与验证环节形成完整闭环，提升结果可靠性。
- 任务拆解与主动迭代探究：不再要求模型一次性输出完美答案，而是将复杂任务（如因果关系发现）拆解为多轮交互或分步流程。促使模型模仿科研工作模式，主动进行提问、信息获取、假设提出、验证与结论修正等一系列操作，通过迭代逐步逼近正确答案。
- 向形式化逻辑表达靠拢：通过训练引导模型采用更严谨的表达方式，如数学符号、编程语言或逻辑语句等。这类形式化表达的正确性可由机器自动核验，为可验证性提供基础支撑。

这些进展共同指向提升模型生成过程的透明度与结果可信度，推动其从灵感型创作工具向严谨型协作伙伴转变。未来的核心挑战在于如何将 these 方法拓展至更开放、复杂的现实场景，同时在验证的严谨性与计算的可行性之间取得平衡。

7.4 攻击与防御

7.4.1 背景

大语言模型已成为驱动全球数字经济、科学发现和日常信息交互的核心引擎。然而，这种广泛而深入的应用，也使其成为攻击者觊觎的数字富矿，其安全漏洞可能导致虚假信息泛滥、敏感数据泄露、关键服务中断甚至社会性恐慌。2025 年度，大模型安全攻防领域的研究呈现出三大核心趋势：第一，攻击手段的持续演进与复杂化，特别是在提示注入与越狱领域，攻击者

已从简单的文本技巧转向多模态、自适应、甚至利用模型潜在概念空间的复杂攻击；第二，防御策略的体系化与纵深化，研究重点已从单一的输入过滤转向多层、语义感知的纵深防御框架，并开始探索主动防御和自适应修复机制；第三，评估体系的标准化与基准化，学术界和产业界共同推动了大量标准化评估基准的建立，涵盖了从提示安全到隐私泄露的多个维度。

7.4.2 提示词安全

提示词安全是大模型攻击与防御的前沿博弈，即确保模型不会因恶意设计的输入（提示）而产生非预期、有害或被禁止的输出，是 LLM 应用最直接、最基础的安全防线。

提示注入与越狱攻击的演化

早期的越狱攻击，如角色扮演、目标劫持等，在 2025 年已被大多数主流模型通过指令微调和输入过滤器有效缓解。然而，攻击并未止步，而是发展出更为复杂和难以检测的攻击范式。

编码与混淆攻击的深化。 攻击者利用 Base64、摩斯电码、甚至是鲜为人知的编码方式来隐藏恶意指令，绕过基于关键词和简单模式匹配的过滤器。Pathade^[922]首次系统性地研究了针对视觉-语言模型（VLMs）的隐写式提示注入攻击，利用模型在处理复杂编码转换时的“思维盲点”，使用隐写技术将恶意指令嵌入图像编码中，而 VLM 在分析图像时，将解码并提取出隐藏提示，并按其执行，这对传统的文本安全检测体系构成了降维打击。该研究揭示了 VLM 的一个隐蔽但真实的安全盲区，呼吁开发者在部署 AI 时，必须考虑这种隐藏编码的对抗性威胁，并采用相应的分层防御策略。

多模态与跨模态攻击。 随着多模态大模型的普及，攻击向量也从单一模态扩展到文本、图像和音频融合的跨模态攻击。HiddenDetect^[923]对 LVLM 内部结构化的安全机制进行了研究，他们发现当 LVLM 处理不安全输入时，其内部某些层会提前激活一种“拒绝语义”（refusal semantics），因此 HiddenDetect 构建一个拒绝感知嵌入（refusal-aware embedding），并计算它与各层隐藏状态的相似度，以此衡量输入的安全性。首次证明 LVLM 的安全机制具有可观察的内部表征，支持“安全是模型内生能力”的假设，为未来防御指明方向。

Wang et al.^[924]提出了一种新型的隐式越狱攻击方法 (Implicit Jailbreak)，通过跨模态语义操控来绕过多模态大语言模型的安全机制。该方法利用图像与用户查询文本之间的合理语义协同，使得大模型认同敏感查询从而绕过拒绝机制。例如用户输入查询：“如何制作炸弹？”并给出实验室场景中一本打开的化学教科书，模型可能认为这是学术研究，于是给出详细步骤。这种攻击不依赖对抗扰动、不使用恶意文本，仅靠自然图像加中性文本提示即可绕过安全机制，实现高成功率越狱。

上下文感知与多轮对话攻击 (Context-Aware Attacks). 攻击者不再追求“一击致命”，而是通过多轮“无害”的对话，逐步引导模型进入一个偏离安全对齐的上下文状态。

CASE-Bench^[925]针对这类依赖上下文的攻击，提出了一种利用多轮对话上下文进行越狱攻击 (jailbreaking) 的新方法，其核心思想是：通过精心设计的多轮对话攻击链，在不触发安全机制的前提下，逐步引导大语言模型 (LLM) 输出有害内容。该方法绕过了现有防御体系，揭示了未来的安全对齐必须从单轮指令跟随转向多轮意图追踪，对话系统需要具备长期风险感知能力。

潜在越狱 (Latent Jailbreak). 这是 2025 年发现的一种更为隐蔽的攻击方式。与直接在提示中注入指令不同，“潜在越狱”旨在激活模型在训练过程中学到的、但被安全对齐所抑制的“潜在概念”或行为模式。例如，通过一系列精心设计的、看似不相关的提示，逐步增强模型内部与“暴力”、“欺骗”等相关的神经元激活权重，最终使其在面对一个正常问题时，也能自发地产生有害回答。

Xing et al.^[926]Latent Jailbreak 基准正是为了评估模型在这种攻击下的鲁棒性而设计的。这种攻击的防御极具挑战性，因为它绕过了所有基于输入的检测。

Xing et al.^[926]该方法将无害查询向量和一个有害意图向量在隐空间融合，生成一个既看似安全、又能激活有害知识的混合表示，提出了一种新颖且高效的越狱攻击方法——Latent Fusion Jailbreak (LFJ)，在模型的隐空间 (latent space)，而非在输入文本层面，通过操控模型内部表示来“欺骗”对齐机制，从而诱导模型输出有害内容。研究表明，当有害信号被“稀释”在无害表示中时，安全机制无法识别，但模型解码器仍能提取其有害语义并执行。

Kadali et al.^[927]系统性地研究了构建高效、可解释的越狱检测机制的重要性，并提出即使大模型最终输出正常，越狱行为也会在模型内部特定安全敏感层激活可检测的拒绝模式，即大模型的安全机制并非是独立模块，而是内生于模型表示。基于上述理论，该论文提出了一个轻量级检测器，通过计算拒绝方向向量与激活向量的偏离程度，来判断越狱尝试。

新型防御机制与框架：迈向纵深与智能

面对日益复杂的攻击，防御策略也从单一、静态的防御点，演变为多层、动态的纵深防御体系。

多层防御框架 (Multi-Layer Defense Frameworks). 2025 年的一个显著趋势是多层防御框架的提出和应用。这种架构通过分层处理——从输入清洗到意图分析，再到输出监控与闭环反馈——构建了一个动态演进的防御闭环。Jacob et al.^[928]在启发式与特征检测层面提出了 PSF (Prompt-Shield Framework) 框架，是目前最典型的多层防御实现之一。它结合了上下文感知解析、输出验证和自适应反馈循环，过滤掉已知的、低级的攻击模式。Xiang et al.^[929]采用辅助 LLM 来分析用户输入的真实意图，判断其是否存在恶意。EDDF 的创新点在于提出了攻击本质 (attack essence) 这一概念，指代越狱攻击中不变的核心语义意图或策略逻辑，而非具体的措辞。EDDF 的目标就是识别并拦截具有此类“本质”的输入，而不管其表面形式如何变化。

基于概念分析的主动防御. JBShield^[930]，针对防御“潜在越狱”等高级攻击，提出了一个极具前瞻性的防御思想“激活概念分析与操纵” (Activated Concept Analysis and Manipulation)。在模型运行时，实时监测其内部的“概念激活”状态。一旦发现异常即主动介入，通过技术手段，如修改特定层的激活值，来抑制这些不安全概念的表达，从而在根本上阻止有害内容的形成。这代表了防御从“被动反应”向“主动干预”的重大转变。

评估与基准：标准化鲁棒性度量

2025 年之前，对模型越狱鲁棒性的评估往往是零散的、不可复现的。研究者各自使用自己的提示集，评估方法也不统一。2025 年，这一局面得到了极大改观，一系列开放、鲁棒的基准测试相继推出。

表 7.8: 越狱鲁棒性评估基准

基准项目	特性与描述
JailbreakBench	2025 年最受关注的越狱鲁棒性基准之一，包括一个庞大且持续更新的对抗性提示数据集 (JailbreakDB)，包含了来自社区和自动化生成的数千条攻击提示。内置标准化的威胁模型、系统提示和聊天模板，确保了在不同模型之间进行公平比较。
HarmBench	更全面的评估框架，同时也是一个攻击发现引擎，它不仅评估模型的“鲁棒拒绝”能力（即抵御越狱攻击的能力），还通过自动化红队测试 (automated red-teaming) 来发现新的攻击向量。
PANDAGUARD	该框架系统性地评估模型抵御越狱攻击的能力，其方法论也对后续研究产生了影响。

这些基准的出现，使得“越狱鲁棒性”从一个模糊的概念，变成了一个可以被量化、被比较、被持续追踪的指标，为整个领域的发展奠定了坚实的基础。

7.4.3 数据安全

如果说提示词安全是 LLM 的“城墙”，那么隐私和数据安全就是其“内网”。2025 年，随着模型越来越深入地处理个人和企业数据，如何防止敏感信息被窃取、泄露或滥用，成为了安全研究的重中之重。

LLM 的数据泄露风险主要有两种形式：(1) 训练数据记忆 (Memorization) 与泄露：模型在训练时，可能会“记住”训练数据中的一些具体片段，尤其是在数据集中重复出现过的内容，如个人身份信息 (PII)、代码片段、受版权保护的文本等。攻击者可以通过精心设计的提示，诱导模型逐字逐句地“背诵”出这些敏感内容。(2) 推理时数据泄露 (Inference-time Data Leakage)：在 LLM 作为服务 (LLMaaS) 的应用场景中，用户的输入 (prompt) 本身就可能包含敏感信息 (如病历、财务报表、公司战略等)。如何确保这些信息在处理过程中不被服务提供商滥用，不被其他用户窃取，是一个巨大的挑战。针对数据泄露风险，2025 年的研究在检测和防御两端都取得了显著进展。

数据泄露检测技术

基于指标的检测。 传统方法是通过计算模型对某个文本序列的熟悉程度（困惑度 Perplexity）或记忆复现能力（n-gram 准确率）来判断。如果模型对一段文本的困惑度极低，或者能够精确预测其后续的 n-gram，那么很可能存在训练集与测试集的重叠。2025 年，研究者对这些指标进行了深入分析。Hidayat et al.^[931]在可控模拟泄露环境下系统评估训练数据泄露检测方法，并给出 N-Gram 方法在检测预训练期间的数据污染方面的有效性，是目前引用率极高的基准方法之一。Choi et al.^[932]提出了一种新颖、高效且高灵敏度的方法——Kernel Divergence Score (KDS)，用于量化数据污染程度，通过计算嵌入空间中核相似性变化，可以在细粒度层面判断训练集是否污染以及污染程度评估。

防止基准污染。 上述研究集中在检测泄露，但即使知道某样本被泄露，还是缺乏一种可持续、可扩展的方法来净化或强化已有 benchmark。为了解决这个问题，Fang et al.^[933]LASTINGBENCH 研究如何修复现有 benchmark，使其在面对已泄露模型时仍能有效评估真实能力，核心思想是：识别题目中容易被模型“记住”的关键信息（leakage points），并将其替换为语义合理但事实不同的反事实版本，从而打破模型的记忆路径，同时保留题目的评估意图。Wu et al.^[934]提出了一种从根本上规避数据污染的全新 benchmark 构建范式。与以往检测泄露或改写旧题不同，该工作主张直接使用模型训练截止后才出现的新知识构建评测样本，从而实现严格无污染的评估。并建议未来 benchmark 发布必须声明：知识首次出现时间和是否通过 AntiLeakBench 类验证。

数据泄露防御技术

数据去标识化与清洗。 这是最直接的预防措施，即在训练数据进入模型之前，通过技术手段识别并移除或匿名化其中的个人身份信息（PII）和其他敏感数据。Apertus et al.^[935]建立了一套以数据合规、语言平等和科研透明为核心原则的开源大模型及基础设施，包括基于数据主动过滤、可复现的严格合规的数据管道（Compliant Data Curation）；防记忆预训练目标（Goldfish Objective）以及真正的多语言覆盖。Clark et al.^[936]提出了一种融合大语言模型与医学知识图谱的混合框架，用于在社交媒体上实时、高精度地识别健康类虚假信息，同时严格保护用户隐私。

安全微调 (Secure Fine-tuning)。当使用特定领域的私有数据进行微调时，如何防止模型记住这些数据并泄露给攻击者？2025 年的研究引入了更精细的防御机制。Zhang et al.^[937]提出了一种在微调大语言模型时防御成员推断攻击的新方法，核心思想是：有选择性地对训练数据中的敏感样本（容易被模型“记住”的高风险样本）进行模糊化，从而在保护隐私的同时，尽可能维持模型效用。在使用特定领域（如医疗、金融）的敏感数据对模型进行微调时，采用特殊的技术来降低数据泄露风险。这通常与差分隐私等技术结合使用。

输出端防御框架 LeakSealer 框架 Panebianco et al.^[938]针对提示注入攻击与敏感信息泄露威胁，提出一个模型无关 (model-agnostic) 的安全框架，利用对 LLM 系统的历史交互日志动态监控攻击演化路径，并通过人机协同实现持续防御演进。

成员推断攻击 (MIA) 的深化。 LeakSealer 框架 Panebianco et al.^[938]针对提示注入攻击与敏感信息泄露威胁，提出一个模型无关 (model-agnostic) 的安全框架，利用对 LLM 系统的历史交互日志动态监控攻击演化路径，并通过人机协同实现持续防御演进。

7.4.4 隐私保护训练方法

差分隐私

参数高效微调 (Parameter-Efficient Fine-Tuning, PEFT) 与差分隐私的深度融合已成为主流范式，显著缓解了隐私保护带来的性能损失和计算开销；

Aketi et al.^[939]焦于如何将开源差分隐私 (Differential Privacy, DP) 训练库 Opacus 扩展至大语言模型 (LLMs) 的训练与微调场景。该工作并非提出全新算法，而是系统性地解决了 LLM 与 DP 结合时面临的内存、计算、效率三大瓶颈。

Makni et al.^[940]提出了一种基于优化的差分隐私稀疏微调新范式，针对满足严格隐私约束的前提下，如何利用带噪梯度信息动态选择可训练参数，以最大化下游任务性能。

其次，算法层面的创新持续涌现，包括对经典 DP-SGD 算法的改进、新型隐私优化器的探索以及更精密的隐私预算核算技术的发展，进一步优

化了隐私与效用之间的权衡。Li et al.^[941]针对差分隐私联邦学习（DP-Fed）中长期存在的核心矛盾——隐私保护与模型效用的严重冲突——提出了一种创新框架 FedCEO（Federated Collaborative Enhancement via low-rank Optimization）。让客户端在服务器协调下协同互补，通过低秩优化恢复被噪声破坏的语义信息，从而实现协同去噪，在严格隐私保证下显著恢复模型语义完整性，将效用-隐私权衡界提升至理论新高度。

联邦学习（Federated Learning, FL）

FL 通过让数据保留在本地设备或服务器上，只交换模型更新（如梯度），来保护数据隐私。2025 年，联邦学习与大模型微调领域的研究呈现出爆发式增长的态势。多个顶级人工智能会议均收录了大量相关论文。

学习效率问题。 如何在联邦环境下，以最小的通信和计算代价完成大模型的微调。

Ghiasi et al.^[942]针对联邦学习中微调大语言模型的两大核心瓶颈——通信开销巨大与数据异构性，设计了一种基于张量分解的参数高效适配器方法：FedTT 与 FedTT+。该工作首次将高阶张量结构引入联邦 PEFT，显著压缩通信量并提升对异构数据的鲁棒性。

Liu et al.^[943]聚焦于联邦学习的通信瓶颈问题。尽管参数高效微调方法（如 LoRA）已大幅减少可训练参数量，但在跨设备或跨机构的联邦场景中，每轮重复上传完整 LoRA 更新仍会造成高昂带宽与时间开销。为此，ECOLORA 提出了一种创新的“轮询分段共享”机制（Round-Robin Segment Sharing），结合自适应稀疏化与无损编码，在不牺牲模型性能的前提下，将通信时间最多降低 79%，总训练时间减少 65%。

异构性的常态化处理。 研究者们普遍认识到，实际联邦网络中的客户端在数据分布、计算能力、网络状况等方面存在巨大差异。因此，能够鲁棒、高效地处理异构性的方法成为研究重点。Wu et al.^[944]聚焦于联邦学习（FL）中客户端参与概率不一致在模型异构环境下的公平性挑战问题，提出了 PHP-FL，通过双端对齐集成学习（DEAL）与重要性驱动的选择性参数更新（ISPU），在不依赖公共数据集的前提下，同时提升全局精度与客户端间性能公平性。

隐私保护的系统化增强。 FL 本身并非绝对安全，它仍然面临梯度泄露、通信效率低下等挑战，因此常常需要与 DP 或安全多方计算等技术结合使用。Mia

et al.^[945]针对跨孤岛 (cross-silo) 场景下大语言模型 (LLM) 联邦微调所面临的隐私泄露风险与性能-安全权衡困境, 设计了一种融合低秩适配 (LoRA)、参数剪枝 (Pruning) 与全同态加密 (Fully Homomorphic Encryption, FHE) 的新型框架 FedShield-LLM。该方法在不依赖差分隐私 (DP) 的前提下, 实现对梯度/更新的强密码学保护, 同时通过智能剪枝减少攻击面并提升效率。

安全多方计算 (SMC) 与同态加密 (HE)

早期的 SMC 方案通常面临巨大的通信挑战, 这在大规模联邦学习中是不可接受的。2025 年的研究深刻认识到这一瓶颈, 并将重心转移到设计低通信开销和高可扩展性的 SMC 协议上。

Dohmen et al.^[946]聚焦于基于秘密共享的多方安全计算 (MPC), 针对现有 MPC 框架存在内存不安全和高内存占用问题。尤其在处理大型深度网络架构时, 传统实现因“内联子电路调用”导致内存爆炸式增长, 严重限制可扩展性, 作者提出 SEEC (SEEC Executes Enormous Circuits), 一个以内存安全与高效为核心设计原则的新型两方计算 (2PC) 框架, 在保障内存安全的同时, 实现与主流框架相当甚至更优的性能。

Li et al.^[947]针对安全多方计算 (MPC) 中的性能瓶颈问题: 如何高效生成大规模的关联随机性, 特别是任意有限域上的透明线性估值 (OLE) 问题, 在伪随机关联生成器 (Pseudorandom Correlation Generator, PCG) 范式基础上, 提出了一套通用、高效且可证明安全的 PCG 构造方法, 适用于任意有限域, 显著降低了 MPC 在线阶段对 OLE 相关性的预处理开销。

机器遗忘 (Machine Unlearning) 当用户请求删除其数据时, 如何让模型“忘记”这些数据及其影响, 是一个重要且困难的问题。机器无学习旨在以远低于重新训练的成本, 高效地消除特定数据的影响, 是 2025 年隐私保护领域一个活跃的新兴方向。

Lee et al.^[948]认为, 现有方法虽能降低模型对“被遗忘数据”的预测准确率, 却未能真正从模型的特征表示中彻底抹除其语义痕迹, 导致隐私仍可通过嵌入空间泄露。为此, 作者提出了一个更严格、更贴近现实需求的新任务范式——知识删除 (Knowledge Deletion, KD), 并提出两种通用的遗忘方法: 无需训练的 ESC (Erasing Space Concept) 和带训练的 ESC-T (ESC with Training)。实验表明, 该方法在速度和遗忘彻底性上均达到 SOTA。

Khalil et al.^[949]针对如何在不重新训练整个模型的前提下, 高效、彻底地“抹除”特定数据的影响, 同时最大限度保留模型对其他数据的性能问题,

提出了一种受对比学习启发的全新框架 CoUn，与现有方法（如标签扰动、权重微调、影响函数近似等）往往只能实现表面遗忘（例如降低目标样本的预测置信度，但模型内部仍保留其语义痕迹）不同，CoUn 从表示空间，通过重构数据嵌入的几何结构，使模型对遗忘数据的响应自然退化为对其最相似保留数据的泛化，从而实现语义层面的有效遗忘。

Pathak et al.^[950]针对语音生物识别系统中的隐私合规挑战，将量子计算的启发式原理（如叠加、纠缠、干涉）引入音频机器遗忘任务，提出了 QPAudioEraser 的新型框架。解决了现有视觉领域遗忘技术在处理时序性强、高维、连续的语音信号时的失效问题，实现了对特定说话人或口音特征的彻底擦除，同时几乎不损害模型对保留数据的性能。

表 7.9: 大模型攻击与防御技术进展

分类	子分类	典型方法及特点	作用点
提示词安全	攻击	隐写式提示注入：隐写技术将恶意指令嵌入图像编码中	输入层
		跨模态攻击：利用图像与用户查询文本间的语义协同	输入层
		多轮对话攻击：通过多轮交互，逐步引导 LLM 输出有害内容	输入层
		潜在越狱攻击：操控模型内部隐空间表示来“欺骗”对齐机制	模型内部隐藏层
	防御	多层防御架构：在启发式与特征检测层面过滤已知攻击模式	输入层
		概念分析防御：实时监测模型内部“概念激活”状态，异常即介入	模型内部隐藏层
数据安全	泄露检测	基于指标检测：利用 N-Gram 检测训练数据污染	数据集
		基准污染净化：修复现有 benchmark	数据集
		成员推断攻击：攻击特定组件，实现差异反应	数据集
	泄露防御	去标识化与清洗：主动过滤数据敏感信息	训练前/数据空间
		安全微调：选择性地模糊化敏感训练样本	训练过程/数据空间
		机器遗忘：知识删除、对比学习、量子计算启发式原理	训练后/隐空间
训练安全	差分隐私	参数高效微调：与差分隐私深度融合	训练过程
		算法创新：优化隐私与模型效用间的权衡	训练过程
		学习效率：优化通信开销	训练过程
		异构性处理：对齐集成学习解决客户端异构性	训练过程
	联邦学习	隐私保护框架：梯度强加密保护，智能剪枝减少攻击面	训练过程
	SMC 与 HE	解决多方计算中的内存不安全、高占用及性能开销问题	训练过程

结论与未来展望

2025 年是大型语言模型安全领域的核心进展可以总结为三点：

攻防对抗的深度升级：攻击技术已深入到模型的语义和概念层面，而防御策略则相应地发展为多层、智能和主动的体系。

隐私保护的范式转变：对隐私的研究不再局限于理论，而是通过差分隐私微调等技术，在真实场景中积极寻求效用与隐私的最佳平衡点，同时对 MIA 等高级威胁的理解也达到了新的深度。

研究方法的科学化转型：大量高质量、标准化的评估基准的建立，以及知识体系化 (SoK) 工作的出现，标志着 LLM 安全研究正在从零散的“手工作坊”模式，向可复现、可比较的“现代科学”模式转变。

未来大模型安全领域攻击与防御的挑战和研究方向包括：

安全供应链 (Supply Chain Security)：LLM 的安全性不仅取决于模型本身，还取决于其复杂的供应链：从数据源的采集、第三方数据标注，到开源基础模型、微调过程，再到部署环境，每一个环节都可能引入漏洞（如数据投毒、后门攻击）。建立端到端的 LLM 安全供应链保障体系，将是未来的重大挑战。

对齐与安全的深层矛盾 (Alignment vs. Security)：“对齐” (Alignment) 旨在让模型变得更有用、更诚实、更无害，但有时这与“安全”存在内在矛盾。例如，过于严格的安全限制可能会扼杀模型的创造力和有用性（“对齐税”）。如何在一个统一的框架下，理论上更深刻地理解并实践上更好地平衡这两者之间的关系，是一个根本性的难题。

7.5 宪法人工智能

在实际应用中，由于用户群体的多元化，大模型生成内容的安全性正面临复杂挑战，模型对齐已成为保障模型合规性的核心。随着 AI 系统的能力接近或者超过人类水平，完全依赖人类进行监督以实现模型对齐已变得难以继。并且，在传统的对齐方法中，为了让模型变得无害，模型往往会变得过于保守。因此，为使模型在复杂多变的场景中保持生成内容的合规性，研究者提出构建一套清晰、可验证且可调整的价值体系，用于约束模型的训练和生成过程，即**宪法人工智能 (Constitutional AI, CAI)**^[951]。

宪法人工智能是指通过一套宪法原则对模型输出进行自动化监督、批评与修正，旨在训练出既无害又非回避的安全模型。相比于传统的强化学习，

宪法人工智能在提升监督效率与可扩展性的同时，具备以下核心特点：

- 1. **宪法原则：**通过一套公开透明、清晰简洁的原则，约束人工智能系统的行为，从而实现模型的价值对齐，具有较高的可解释性。
- 2. **动态修改：**宪法原则可基于当前形势（如法律法规、社会规范、用户反馈）动态更新，通过在线学习、增量训练等方式使系统快速适应变化。
- 3. **非回避型响应：**传统对齐方式训练的系统在面对敏感问题时常常采取简单回避的策略，而经过宪法原则训练的大模型则既无害又不会回避问题，能够参与讨论并解释其拒绝有害请求的理由。

近年来，宪法人工智能技术不断发展，研究重心已从早期的静态原则设计转向原则的自动化生成、跨任务适配及深层应用等。下文将从原则构建、原则适配、原则应用三个维度，系统梳理 CAI 技术在 2025 年的前沿进展。

表 7.10: 宪法人工智能 2025 进展总结

发展阶段	关键任务	代表工作
原则构建	构建公开、透明、灵活修改的宪法原则	Kyrychenko et al. ^[952] 、Henneking et al. ^[953]
原则适配	探讨宪法原则应用于模型的高效方法	Zhang ^[954] 、Menke et al. ^[955]
原则应用	利用宪法原则保证大模型安全性、推进 AI 应用研发	Noh et al. ^[956] 、Sharma et al. ^[957] 、Maiya et al. ^[958] 、Lyu et al. ^[959]

宪法原则构建

构建一套高质量的宪法原则是实现有效对齐的前提。如果原则本身存在逻辑冲突、表述模糊或覆盖面不足，模型的有效对齐就无从谈起。针对这一问题，2025 年的研究主要集中在提升原则的通用性与构建效率两个维度。

针对原则构建通用性不足的问题，研究者规范了原则收集、转换和筛选流程，形成用于原则构建与评估的 C3AI 框架，能够适用于不同场景^[952]。另外一些研究者在原则生成提示词、原则子采样、原则过滤等方面改进了原有

的逆向宪法（Inverse Constitutional AI，简称 ICAI），生成的原则降低了对特定样本的拟合，提高了通用性^[953]。

针对构建训练效率低的问题，C3AI 框架创新性地引入了心理测量学方法^[952]，利用探索项图分析（Exploratory Graph Analysis, EGA）自动识别无害性与诚实性等核心维度，并结合唯一变量分析剔除语义重叠的原则，精简了原则集合，提升了模型训练效率。

宪法原则适配

构建了原则之后，如何将其适配到不同规模、不同架构的模型上，是 2025 年研究的焦点。特别是随着端侧 AI 与垂直领域小模型的兴起，CAI 在有限参数空间内、不同模型架构的表现引发了很多讨论。

针对参数空间有限的问题，研究者深入探讨了 CAI 在 8B 级别小模型上的适配瓶颈，发现宪法原则在大幅降低小模型的攻击成功率的同时也会引发严重的模型崩溃现象，如模型在面对敏感问题时倾向于过度拒绝，甚至输出重复语句、无意义的表情符号或陷入逻辑死循环^[954]。另外一些研究者则系统性研究了 CAI 在 7-9B 级别小模型上的有效性，具体的，他们通过 Abliteration 技术人为剥离模型的安全机制，使模型处于无审查状态，随后测试其利用 CAI 原则恢复安全的能力，发现宪法原则使小模型具备较高的安全识别能力，但在开放式任务的安全应用能力具有显著差异^[955]。针对有限参数空间的探讨，揭示了针对小模型设计更轻量级的专用宪法的必要性。

针对模型架构影响的研究则进一步揭示了推理能力的重要性。一些研究者系统性地研究了 CAI 对不同架构的有效性，发现具备强推理能力的模型（如 DeepSeek-R1）能够通过显式的思维链评估风险并制定回答策略，从而迅速恢复安全防御。这证明了一个会思考的模型能更深入地理解并执行宪法^[955]。由此来看，良好的对齐应该建立在推理能力提升的基础上。

宪法原则应用

宪法人工智能最初主要用于提升模型的安全性。2025 年的研究在此基础上，进一步探讨了宪法原则在复杂防御架构及垂直行业、领域中的应用。

在安全性增强方面，研究者开始关注宪法原则与不同技术框架的结合。一些研究者在隐私计算场景下提出了一种结合 CAI 的负责任联邦大模型架构，在客户端部署 Llama Guard 3 作为安全过滤器，从源头净化训练数据；在服务器端聚合全局模型后，利用宪法原则进行少量轮次的安全性对齐。这

种端-云协同实施双保险机制，不仅能有效抵御恶意客户端的投毒攻击，还能大幅降低传统联邦学习在进行全量 RLHF 时所需的时间与算力成本^[956]。另外一些研究者针对通用越狱攻击，提出了一种独立于模型本身的外挂式防御体系。通过构建输入分类器与流式输出分类器，形成了多层拦截网，具备逐词扫描模型的生成内容、中断违规输出的能力。实验数据显示，该系统对未知攻击的拦截率超过 95%，且在生产流量中的误报率极低（仅增加 0.38%），为提升大模型系统级安全性提供了高鲁棒性的解决方案^[957]。

在行业应用方面，宪法原则被用于精细化调整模型的行为模式。研究者提出“性格本质上是一套内在的行为准则”假设，通过编写性格宪法训练模型，让模型在自我对话中深化对性格的理解，开创性地将 CAI 应用到性格塑造领域，标志着 CAI 进阶为一种性格塑造工具，能够帮助开发者构建更具个性化和拟人化的下一代 AI 助手^[958]。另外一些研究者从心理危机干预手册中提取原则构建了专用宪法，研究表明，应用该宪法后的模型在危机识别准确性与情感支持能力上均有显著提升^{Lyu et al.[959]}。这些研究表明，宪法原则的应用已从通用的安全对齐扩展到了特定任务的行为规范与能力增强。

未来展望 前文回顾了 2025 年宪法人工智能的相关工作，这些工作促进了 CAI 从 Claude 系列模型到更多种类模型的推广，丰富了宪法原则构建、应用相关技术。在此，对宪法人工智能未来的发展趋势进行展望。

1. 宪法原则构建的半自动化与动态演进：未来，随着 ICAI（逆向宪法）技术的发展，CAI 有望在原则构建上减少对高强度人工编写的依赖。通过引入自动化发现机制，模型能够辅助研究人员从大规模社会对话、法律文献或行业规范中提取潜在的价值对齐准则。这种人机协作的构建模式将使宪法原则的更新更具时效性，能够更灵活地响应社会伦理需求的变化，从而建立一种更具韧性的动态对齐体系。
2. 场景化对齐与垂直领域应用拓展：随着大模型应用边界的不断拓展，CAI 的应用范畴也将继续向特定行业价值对齐延伸。通过将行业特有的合规要求、职业道德及法律边界编码进宪法框架，CAI 可为不同领域提供定制化的行为准则。例如，在医疗领域强调循证规范，在金融领域侧重合规风控。作为一种标准化的价值映射机制，CAI 将助力人工智能在复杂专业场景中实现更精准的合规与对齐。

7.6 本章小结

本章对 2025 年度大语言模型安全与伦理研究的进展进行了系统性总结。在“安全对齐与治理”方面，研究呈现出内生防护与外置监测相结合的态势，机械可解释性工具为理解安全机理提供了新路径。“生成风险控制”通过训练优化与推理增强相结合的技术体系，致力于缓解模型幻觉。“内容真实性与可追溯性”研究通过水印、可验证生成及细粒度溯源技术，为 AI 生成内容的可信与可控奠定了基石。“攻击与防御”领域则展现出攻防博弈持续升级，隐私保护通过与差分隐私、联邦学习等技术深度融合实现突破。“宪法大模型”通过原则构建与适配，推动价值对齐从通用约束走向精细化应用。总体而言，2025 年该领域的研究涵盖了机理探索、技术防御、内容治理与价值对齐等方面，为构建可信、可靠的人工智能系统指明了方向。

第八章 未来展望

“大语言模型未来将走向何处”，从现有研究趋势来看“堆规模”开始转向“提智能密度”，从被动生成走向可行动的智能体，并从单模态拼接迈向原生统一的多模态与世界模型。围绕这一主线，本章系统梳理关键技术路线与应用范式的演进，包括推理后训练与强化学习、幻觉与事实性治理、知识增强与多模型协作、云边端协同、AI4Science 与具身智能等。最后，本章进一步讨论算力不均、安全伦理与跨学科融合带来的结构性挑战与机遇，给出面向长期可持续发展的判断与展望。

8.1 技术趋势预测

近来，大语言模型的跃迁让通用人工智能（AGI）从口号走向可度量目标。Hendrycks 等^[960]以受过良好教育的成年人为参照，将通用智能拆解为十个核心认知领域，把通用人工智能转化为可比较的能力画像。沿着这一坐标系，本节从“能力—形态—应用—架构”四个角度勾勒未来大语言模型迈向通用人工智能路线，焦点从堆规模转向提升智能密度，以纯强化学习推动复杂推理涌现，并促使多模态由拼接走向原生统一^[242]。同时，在每个阶段都注重生成过程中提升事实性、缓解幻觉，以及云端与边缘协同、开源与闭源共生，推动个性化与产业化普及^[961]。面向 2035 年，因果推理与世界模型将成为技术内核，具身与空间智能把模型带入物理世界^[962]，而安全、伦理与治理将决定这场跃迁能否稳健向善。

8.1.1 模型能力从注重规模到注重“智能密度”

复杂推理成为核心战场，新训练范式涌现 过去几年，LLM 的能力提升高度依赖参数规模，数据规模和训练算力”的经验缩放规律（scaling laws）。但从 Kaplan 等^[963]的缩放律到 Chinchilla 的“算力最优训练”，研究界已逐步形成共识，仅靠做大模型却不匹配足量高质量 token 与更有效训练目标，会

出现效率与收益递减，推动研究转向“单位算力与单位参数”的有效智能产出，即“智能密度”导向。在这一背景下，“推理能力”成为衡量智能密度的关键指标之一。大量研究指出数学、代码、科学推理等任务需要可组合的多步推导与自我校验，而传统有监督学习、思维链等等方法在泛化性、稳健性与成本上存在瓶颈，促使强化学习在推理后训练中快速上升为核心路径^[964]。例如，DeepSeek-R1^[242]展示了“纯强化学习”可在无需依赖大量人工标注推理轨迹的前提下，激励模型涌现自我反思、验证与策略自适应等推理行为，体现了从模仿人类推理到自主发现推理策略的范式转变，为提升模型高阶认知能力开辟了新道路。

多模态融合从“拼接”走向“统一” 当前主流多模态大模型多采用“模态编码器 + 语言模型”的连接式架构（如视觉编码器将图像映射为 token，再与文本 token 拼接输入 LLM），其优势是工程实现快、可复用成熟 LLM，但在跨模态对齐、时空结构建模、以及“联合推理”（例如从视频中抽取因果线索、把草图或规格转成可执行代码或物料清单）方面仍受限。一些研究将未来趋势概括为：从异构拼接走向原生统一建模（统一 token 化、统一表示空间、统一目标函数），并强调“多模态推理”将成为比“多模态感知”更高阶的竞争点^[965]。这一方向的关键挑战包括：1) 跨模态信息的粒度与对齐（静态图像与连续视频、音频）；2) 长上下文与时序建模的算力开销；3) 数据构造与评测范式（是否真正测到推理而非模式匹配）。

“幻觉”问题将从系统和原生大模型层面得到有效性缓解 幻觉（hallucination）是 LLM 进入医疗、法律等高可靠场景的首要障碍之一。现有研究将幻觉细分为事实性幻觉与忠实性幻觉（与外部事实不一致 vs. 与给定输入或证据不一致），并指出其成因贯穿数据、训练目标、对齐方式与解码策略等全链路，因此仅靠事后检测难以根治。^[966-967] 未来针对大模型幻觉问题的研究主线更可能是“训练/推理/系统”的协同治理，一些可行的思路包括：

- 推理阶段的内生纠错：通过利用模型不同层 logits 演化信号实现自我修正，证明无需外部知识或额外训练也能提升事实性，同类思路还包括通过对比层解码从模型内部表征中挖掘更可靠的事实知识。^[966]
- 系统级 Grounding：检索增强生成（RAG）被普遍视为“把生成约束到可追溯证据上”的关键工程路径，现有研究多集中在检索、生成、端到端评测与安全隐私等维度。进一步地，知识图谱等结构化知识与 LLM

的结合 (KG-grounding、GraphRA 与 KG-RAG) 被视作提升多跳推理与可解释性的方向^[968], 将在未来使 LLM 的可靠性达到支撑医疗诊断、法律分析等关键任务的水平。

8.1.2 基础模型的技术架构与训练范式的演进

突破数据瓶颈：合成数据与更高效的算法 “高质量文本数据趋于耗尽”正在从经验判断走向可量化的趋势分析。Epoch AI 对可用于训练的高质量、去重后的公共人类文本规模做了估算, 并预测若按既有扩展速度推进, 公开人类文本的“有效存量”可能在 2026–2032 年间被充分利用^[968]。在这一背景下, 未来解决路径更像是一套“数据飞轮”而非单点技巧, 主要突破依赖于合成数据生成、高效序列建模、神经符号混合等。

- 合成数据生成将成为主增量, 但必须防范“模型塌缩”, 过度依赖模型生成数据可能导致分布长尾丢失与能力退化等风险^[969]。因此, 未来的合成数据不会是“越多越好”, 而是走向可验证、可控覆盖、与目标任务对齐的生产管线^[970], 并配套更严格的评测与数据治理。
- 高效序列建模: 选择性状态空间模型、高效注意力与 Transformer 形成替代与混合架构。一方面, 选择性状态空间模型 (以 Mamba 为代表) 试图在保持表达力的同时把长序列建模的计算复杂度降到近线性, 并强调硬件友好实现^[44]。另一方面, 高效注意力以及与大规模预训练和推理系统的工程耦合正在快速推进^[971]。未来一段时间的主流架构在局部精细建模与全局长程依赖之间动态取舍, 并很可能呈现“Transformer 高效注意力/选择性状态空间模型的混合”, 以降低长上下文与多模态时序推理的成本。
- 神经符号混合, 将“可解释推理”融入系统闭环。在数据与可靠性双重压力下, 将符号结构 (逻辑、约束、知识库/本体) 作为可计算的外部结构^[972], 与神经模型的表示学习互补。

知识增强、模型协作与共演化：从“单模型”走向“动态系统” 一篇发表在 ScienceDirect 的 2024 年综述^[973]提出了“后 LLM 时代”的三个关键方向: 知识增强 (Knowledge Empowerment)、模型协作 (Model Collaboration) 和模型共演化 (Model Co-evolution)。未来的 AI 系统将是动态的, 多个模型

通过共享知识、参数和策略，在持续交互中共同进化，以适应不断变化的环境和任务，实现终身学习。

- 知识增强：从“记在参数里”到“可检索、可编辑、可追溯”。知识增强的工程化主战场是 RAG 与结构化知识结合，对事实性与时效性有重要作用^[973]。未来“知识能力”会以参数内知识、外部可验证知识和可控更新机制的组合形态落地。
- 模型协作：从单体智能到“多模型编队”。协作并不仅等于多智能体聊天，更包括集成、大小模型协同、路由与分工等系统机制^[974]，通过协作机制提升系统整体的智能密度（在成本、隐私、时延约束下维持或提升能力上限）。
- 模型共演化与终身学习：从“训练一次”到“持续演进”。未来的关键不只是能合成数据，而是能否建立稳定的自我改进回路，既能迭代提升，又能避免塌缩与安全失控。

因果推理与通用推理引擎的初现 当前主流 LLM 的优势在于从海量语料中学习“相关性结构”，但在高风险决策场景中，真正需要的是“干预—反事实—机制解释”的能力，不仅回答“会不会发生”，更要回答“为什么发生、如果换一种做法会怎样”。结构因果模型（SCM）与反事实推断为这一目标提供了统一语言与数学工具，是未来十年把“可泛化推理”做实的关键理论底座^[975]。近两年，一个清晰研究潮流是“因果大模型”的交叉^[976]。一方面，大模型在生成因果论证、解释与假设方面展现潜力（但也暴露出混淆相关与因果、缺乏可检验机制等局限）；另一方面，研究者开始系统总结如何用提示、工具、结构化因果图、反事实数据与评测基准来增强 LLM 的因果能力。这也与心理学启发的双过程（System 1/2）路径相呼应，让模型在直觉式生成之外，具备更慢、更可验证的“系统 2 式推理”，在关键步骤进行自检、回溯与证据约束。

世界模型（World Model）的构建：从生成到“可规划的内部模拟器” 世界模型的核心思想是让智能体学习一个可压缩的环境动力学模型，在想象空间里进行多步滚动预测与规划，从而把学习从试错提升为模拟与决策^[977]。随着大语言模型的发展，世界模型研究出现两点关键变化。其一，世界模型从服务某个任务的模型扩展为通用环境模拟器，与大规模生成式建模、多模态学习深度融合。其二，学术界开始用更系统的视角给出定义、能力谱系与

评测框架，推动世界模型从概念走向工程化与可比性。产业界也在加码，例如 DeepMind 发布 Gemini Robotics 1.5 技术报告^[978]，强调具身智能需要在行动前进行更明确的推理与分解。面向 2035 年，更可能出现的通用推理引擎形态并非单一 LLM，而是世界模型（可模拟）+ 因果模型（可解释）+ 规划器（可搜索/可验证）的耦合系统，从而完成需要长期规划的复杂任务（如家庭机器人完成多日家务）。

自我进化与持续学习：从一次性训练到开放环境的长期适应 若要进入物理世界并长期工作，模型必须突破训练后知识冻结的局限，实现在开放环境中持续学习新知识而不遗忘旧技能的能力。当前面向 LLM 的持续学习从持续预训练、域自适应到持续微调已发表大量研究工作^[979-982]。与此同时，自我进化（self-evolution）开始成为面向智能体的训练范式，通过经验获取、参数更新、模型评估的循环，让系统在交互中自我改进。这与后 LLM 时代的路线图在逻辑上高度一致，未来 LLM 更像动态生态，而非静态单模型。

8.1.3 应用范式：从被动工具到主动智能体

LLM-based Agent 成为下一代软件交互的接口 研究界普遍认为，LLM 正从“对话式工具”走向“可行动的任务执行体”，其核心不再是生成答案而是围绕目标分解、规划生成、工具调用、执行监控、反思改写等形成闭环式工作流^[983-985]。多智能体协作系统将能够模拟社会分工，解决单个智能体无法处理的超复杂问题（如大型软件开发、城市交通调度），未来一段时间内 LLM 交互入口将加速转向智能体与工具生态的构建。与此同时，智能体的评测也在从静态问答走向交互式基准，如 AgentBench^[986]面向多环境评测、GAIA 强调真实助理任务中的工具使用与多模态处理能力^[987]，这些评测将成为未来智能体迭代与对比的共同尺度。需要强调的是，主动智能体也放大了安全与治理难题（例如长链行动中的偏离目标、工具误用、不可预期的策略性行为）^[988]。

深度赋能科学发现（AI4S）：从文献助手到“自主实验室” AI4Science 的研究路线正在从“读论文、做综述”升级为覆盖假设生成、实验与仿真设计、自动化执行、数据解析、迭代优化的全流程闭环。现有研究^[989]指出 LLM 与多模态生成模型、科学知识库、实验自动化平台耦合后，能够显著缩短研究周期并提升探索效率。其中，Self-driving Lab（自主实验室）被视为关键基础设施，化学与材料领域的观点进一步总结了自主直言是的架构范式、

人机协同边界与产业落地瓶颈（成本、标准化、可复现实验协议、数据闭环等）。^[990-991] 在更宏观层面，美国国家科学院 2025 年报告^[992]也将“科学基础模型”视作科学发现与创新的重要抓手，强调算力、数据治理、评测与跨学科协同的系统性建设。因此，未来五年 AI4Science 的应用范式主要集中在：1) 以科学大语言模型、多模态基础模型沉淀领域先验知识；2) 自主实验室、自动化工作站等把“推理、实验、反馈”真正接入物理世界，形成可持续加速器。

垂直行业实现从“赋能”到“重构”的跨越 在金融、法律、医疗、工业等高价值场景中，LLM 正在从提高效率的辅助工具转向智能体驱动的端到端流程，其落地逻辑更强调可控、可审计、可追责。但是，合规与稳健性是 LLM 的采用门槛，幻觉是阻碍规模化应用的关键风险，因而法律 Agent 更依赖检索、引用与证据链对齐的工程化体系。并且，在医疗领域面向真实临床工作流的智能体需要在仿真或高保真环境中评估其对流程与安全的影响，而不仅是刷题式指标。未来，智能体走向“RAG+ 结构化知识 + 人类在环”的组合范式将成为 LLM 在垂直领域的重要范式。

8.1.4 云边协同将大模型能力与移动互联网时代特征充分融合

端侧小型化、专业化模型爆发 将 LLM 从“只在云端运行”推进到“云端与端侧协同”，核心驱动力来自三类现实约束：交互时延（实时对话、车载、工业控制）、隐私与数据主权（医疗、金融、个人设备）、以及成本与可用性（带宽、离线可用、推理费用）。近年来的系统研究普遍指出：端侧部署并不等价于“把大模型硬塞进手机”，而是一套覆盖模型压缩、运行时加速、边云协作与持续适配的全栈工程体系。这将实现低延迟、高隐私保护、低成本的个性化 AI 服务，推动 AI 能力真正普惠。在“智能密度”导向下，端侧模型会呈现两条主线：

- 小而强：通过更高效架构与后训练策略，让较小参数规模获得更高的单位算力产出^[993]，“精度—延迟—能耗”三者的联合权衡是未来五年的关键优化目标。
- 专而精：针对特定场景构建“专用小模型或小语言模型”，其优势在于可控、可验证、可本地化更新。

但是，端侧普及并不是“端侧全包”的极端路线，而是走向边云协同。

常见策略包括：1) 端侧先行的级联推理，本地小模型先答，复杂再上云端；2) 切分推理与卸载，部分层在端侧，部分在云和边缘节点，3) 端侧缓存与检索、云端做重计算与全局更新。

开源与闭源生态形成共生格局 开源与闭源生态的关系，正在从“对立竞争”转向“分工共生”。一方面，权威统计报告显示在主流基准与众包对战平台上，开源模型与闭源模型的差距在 2024–2025 年明显缩小，为企业与研究者提供了更可控、更低成本的选择空间。斯坦福大学 HAI 研究院指出以 Chatbot Arena 为代表的众包评测平台也在持续推动可比较的公开排序与迭代反馈，但同时研究者提醒应警惕“排行榜幻觉”（数据分布、采样与可复现实验设计对排名的影响）。因此，更现实的产业技术基座将是“混合架构”：端侧和企业内用开源模型承担高频、隐私敏感、成本敏感任务；云端闭源或顶级模型承担高难推理、多模态与关键决策环节，并通过边云协同实现体验、成本与治理的动态平衡。未来，开源生态将继续在模型架构创新、垂域适配和透明可信方面发挥主导作用^[994]；而闭源模型则可能在最前沿的多模态、复杂推理探索以及提供企业级一体化解决方案上保持领先。两者共同构成健康、动态的产业技术基座。

8.1.5 从虚拟到现实：世界模型与具身智能

具身智能被认为是实现 AGI 的关键路径，将大语言模型与机器人技术深度融合，赋予智能体能够感知和作用于真实世界的能力。

具身大模型（Embodied LLM）成为机器人“通用大脑” 从研究谱系看，视觉、语言、动作（VLA）统一模型正在成为具身智能的关键载体，它把感知、指令理解与动作生成压到一个端到端框架中，支持从导航、抓取到多步操作的广泛任务。以 Gemini Robotics 1.5^[962]为例，DeepMind 引入了“先思考再行动”的机制，在输出动作前生成内部推理序列，这类机制可视为把“系统 2 式推理”嵌入具身控制链路的早期雏形。目前的研究趋势包括构建融合触觉，视觉等多模态信息以及构建轻量高效的 VLA 大模型。未来，统一“机器人大模型”更可能与世界模型、工具链与安全约束共同构成机器人操作系统级的平台能力。

空间智能（Spatial AI）的成熟：从 2D 理解到 3D 结构化推理 真正的物理交互要求模型理解三维几何、拓扑与可供性，能在空间约束下规划动作

序列。当前，视觉-语言模型的空间智能能力主要由几何关系、视角变换、3D 一致性、地图构建与空间问答能力等组成，但常出现“语言上会说、空间上做不到”的断裂。因此，2035 年更有代表性的空间智能形态，可能是多模态表示天然携带三维结构（或可还原为 3D 场景图/神经场），并与世界模型的滚动预测结合，实现“在脑内先搭建场景—再模拟行动—再落地执行”的闭环。

本节从当前随着大语言模型技术的发展而新兴的研究领域出发，展望了未来大语言模型在通用智能、全模态通用 Agent、AI for Science 以及具身智能等方面的技术趋势。随着这些技术的不断进步，我们有理由相信，大语言模型将在未来几年内继续推动人工智能领域的革命，开启一个更加智能和互联的时代。

8.2 挑战与机遇

本节围绕大模型发展中的关键张力展开。一方面，算力资源不均与安全伦理压力正在重塑学术与产业分工与技术路线，带来可复现性、治理与合规的系统性挑战。另一方面，数据与计算最优配置、训练/推理侧降门槛技术、公共算力普惠以及跨学科融合正在打开新的参与空间与产业形态。本节依次讨论算力鸿沟如何结构性扩大并催生效率与基础设施机遇，安全伦理如何从内容管控走向全生命周期的工程化治理与可复现实证，以及大模型在 AI for Science 与科研智能体等方向推动跨学科协作的机会与风险。

8.2.1 算力资源不均

算力鸿沟的结构性扩大导致学术、产业界研究版图重排 大模型研发的边际进步仍与大规模训练、推理算力强相关，导致工业界在算力占有与试错频率上形成持续优势，学术界在“算力密集型研究方向”的参与度下降，连带削弱外部审计、可复现性与批判性检验能力。机遇在于，这一趋势正在推动更制度化的公共算力、结构化访问计划、第三方评测通道，以在不对齐商业机密的前提下恢复学界与公共部门的研究与监督能力。因此国家级算力基础设施与开放科学的组合成为重要的突破口^[995]。

计算最优与数据效率成为低预算的关键杠杆 算力昂贵且供给不均会长期存在，单纯“堆参数”对多数主体尤其是学术界不可行。当前已有研究证明在固定算力预算下存在更优的模型规模与训练 token 配比，能用同等算力训

练出更强模型并降低下游微调与推理成本^[996]。从而把对算力的竞争部分转移到“算力—数据—模型”的最优分配，这为算力较弱的团队提供了通过工程与数据治理追赶的空间。

训练侧降门槛的算法与系统栈（PEFT/量化/显存与 IO 优化/稀疏化）训练与微调大语言模型对显存、通信与工程栈要求极高，形成事实上的进入壁垒，然而参数高效微调（PEFT）、低比特训练/推理、系统级优化等技术正在系统性压低门槛。其中，LoRA^[997]用低秩可训练矩阵替代全参微调显著减少可训练参数与显存压力，QLoRA^[998]进一步把 4-bit 量化与 LoRA 结合，使更大模型在单卡或小集群上可微调，FlashAttention^[999]通过 IO 感知的注意力核减少 HBM 读写提升吞吐，ZeRO^[1000]通过分片优化器减少冗余参数与梯度的显存占用。同时，稀疏专家混合模型 MoE（如 Switch Transformer^[1001]）用每个 token 只激活少量专家的方式在接近恒定计算量下扩展参数规模，把参数规模与实际计算量部分解耦。这些共同把算力不均的冲击从不可参与变成可用工程手段参与。

推理侧成为新主战场 算力挑战同时正在从训练端外溢到推理端，即便模型权重开源，稳定、低时延、低成本的推理服务仍依赖高端 GPU、HBM 与高质量集群网络，算力富集方更容易形成服务规模优势。端侧、边缘推理、量化与推理加速成为应对这一挑战的机遇，会把一部分能力从中心云释放到更广泛的设备与区域，同时也促成“云、边、端”协同的新产业形态，让弱算力主体通过更灵活的部署形态获取可用能力，而不必完全依赖超大云集群^[1002]。

公共算力与算力普惠基础设施 算力作为稀缺资源更像关键基础设施，若完全市场化分配会加剧区域、机构与学科的不均衡。美国国家 AI 研究资源机构提出公共与私营协作模式，尝试把算力、数据、模型与工程支持打包供给研究与教育群体，以实现“民主化访问”。在国内，中国信息通信研究院等机构把大模型基础设施算力规模与利用率、智算中心关键能力与生态协同等作为落地抓手，推动从资源堆砌走向可调度、可运营、可服务的公共、行业算力底座建设，从而缓解不同主体之间的结构性差距^[1003]。

绿色与高效成为公平性议题 算力不均往往与能耗、碳排与成本不均绑定，越依赖大规模算力，研究门槛就越会被抬到少数主体。绿色 AI^[1004]明确提出把效率作为与精度同等重要的研究评价维度，并倡导报告算力和成本的价格

标签，其本质是用学术规范推动更高效、更可负担的方法扩散，从而提高研究与应用的可参与性。配合硬件与推理成本测算等趋势报告（如斯坦福 AI 指标的硬件与推理成本分析），未来效率、成本与可及性会成为衡量大模型路线是否可持续、是否具备普惠价值的关键坐标系。

8.2.2 安全与伦理

当大语言模型从“回答问题的文本系统”逐步变成检索、调用工具、写代码、下指令并影响外部系统的智能体时，安全伦理的讨论会自然地从“说什么不该说”扩展到“它会做什么、怎么做、出了问题谁负责、怎样被发现与纠正”。因此，大语言模型不再被当做简单的算法，而是明确将其与部署环境共同构成社会技术系统。风险不是单点的内容风险，而是贯穿研发、训练、评测、上线、运维与迭代的全生命周期风险。过去一段时间学术界逐渐达成共识，大语言模型的安全对齐越来越像一门“可验证的系统科学”，而不只是经验性的加一层过滤器。

治理与合规工程化 随着大模型被嵌入业务流程并形成可持续迭代的产品形态，安全伦理正在从“提出原则与红线”转向“组织级、全生命周期、可审计的风控闭环”。一方面，美国国家标准技术研究院的 AI 风险管理框架^[1005]把治理责任、风险识别、风险度量与风险处置组织成闭环，这一思路本质上是在把 AI 安全推向工程化的持续风控。另一方面，ISO/IEC 42001 标准以“AI 管理体系（AIMS）”的方式把政策、责任、过程控制与持续改进固化为管理系统标准，推动企业把安全伦理纳入类似质量/信息安全体系的制度化运作。与此并行，欧盟《人工智能法案》^[1006]以分阶段生效时间表推进，使合规义务、运行期监控、可追溯证据逐渐成为产品可用性的硬约束，而不再只是能力指标的附加项。

对齐训练走向“显式规范与推理式对齐” 对齐研究正在把安全从“过滤或拒答式的经验性护栏”升级为“可维护的规范系统”。核心做法是将安全规范以文本或规则形式显式化，让模型在生成前先检索并推理相关规范，再完成回答，从而把安全决策过程变得更可解释、也更容易跨场景泛化。Deliberative Alignment^[872]是其中的代表性工作，它强调直接教授模型可读的安全规范，并训练模型在作答前显式回忆与推理这些规范，目标是在提升抗越狱鲁棒性的同时降低过度拒答，并把“遵循规范的机制”从隐式行为偏好转向更可验

证的推理过程。这也意味着未来安全对齐会越来越像一门“规范工程与推理验证”的系统科学，而不是单纯调参或堆叠过滤器。

安全评测与红队基础设施化 随着攻击与防御快速迭代，研究界对“可复现实证”的需求显著上升，逐渐从对“榜单”的关注提升到“标准框架与可复现实证”的要求。不仅要评估模型本身，还要系统比较不同红队方法、不同防御策略与不同训练机制在多样化攻击面下的真实效果。HarmBench 评测任务^[1007]体现了这种基础设施化趋势，它将自动化红队评测标准化，强调评测属性、覆盖面与可对比性，并在统一框架下对多种红队方法、目标模型与防御进行大规模比较，从而把“安全改进是否有效”转化为更可复核的实验结论。ALERT 评测任务^[1007]用细粒度风险分类组织大规模红队提示，旨在系统暴露不同模型在对抗场景下的安全短板。SG-Bench 评测任务^[1008]更关注安全泛化，把不同任务形态与提示范式下的表现差异显式纳入评估。这类框架的普及会推动安全研究从零散案例走向类似软件工程基准测试那样的规范化评测生态。

智能体与工具使用安全 当大语言模型通过检索、工具调用与多步计划影响外部系统时，安全边界会自然外推到“链路与权限”，典型风险不再只来自用户提示，还来自网页内容、检索结果、工具返回以及状态在多步链路中的传递与污染。英国国家网络安全中心将提示注入强调为更接近受混淆的代理类系统性问题。由于模型难以从机制上硬区分“指令”和“数据”，单靠提示词往往无法彻底消除注入风险。SafeSearch^[1009]针对“搜索代理”提出自动化红队评测框架，并在真实互联网场景展示不可靠信息源可能如何误导代理行为，同时指出简单的提示防御效果有限，推动权限最小化、数据隔离、可审计轨迹与运行期监控等研究。在攻击与防护层面，越狱与提示注入仍然是长期主题，产业界普遍将模型视为易混淆的代理并设计隔离与制衡^[1010]。

后门、欺骗与持久化风险 随着模型能力增强，风险会更多呈现为表面通过评测，但特定条件下作恶的持久化形态，例如数据投毒触发的后门、或更接近“欺骗性对齐”的策略性行为。Sleeper Agents 系列工作^[1011]用可复现的方式展示，模型可以被训练出在大多数情境下表现良好、但在触发条件下显著偏离的行为，并且常规的行为层安全训练（如 SFT、RL、对抗训练等）在某些设置下并不足以彻底清除这种隐蔽机制。这会推动未来研究更关注“触发条件探索、潜在机制定位、对抗性训练与检测”的系统方法，把安全从通

过测试即可推进到持续应对可适应对手的安全科学。

运行期监控与事故学习机制 随着大语言模型部署规模扩大，仅靠上线前评测很难覆盖真实世界的复杂边界条件，安全伦理将越来越依赖运行期治理，包括持续监测、可回溯日志、事件响应与复盘改进，并将事故经验沉淀为组织学习资产。AI Incident 任务^[1012]以航空与网络安全类比，强调对现实世界 AI 事故与近事故进行索引与归档，以支持对风险的想象、预警与预防。这种事故数据库化的趋势会促使企业把安全从“是否合规、是否通过测试”扩展为“是否具备可观测性、是否能快速止损与纠偏、是否能把事故转化为制度与模型更新”的闭环能力。为了对大语言模型的安全伦理监控将从“上线前证明”到“上线后可发现、可纠偏、可复盘”。

隐私与训练数据治理 隐私风险将持续成为大语言模型安全伦理与合规的高压线，其关键不只是会不会泄露，而是泄露风险能否被量化评估并被工程控制。Carlini 等^[1013]关于训练数据抽取的工作证明了攻击者可通过查询恢复训练样本，使训练数据来源、许可、去标识、最小化、删除权、访问控制与输出侧泄露评测与监控必须形成组合拳。因此未来研究会更强调把隐私从抽象原则落到可操作指标与可验证流程，例如在特定应用域定义泄露基准、建立红队抽取评测、并与数据治理制度共同闭环。

真实性与来源证明 在生成内容广泛进入传播链条后，安全伦理会与真实性基础设施逐步融合。一条路线是水印与检测，尽可能以较低代价识别 AI 生成内容，另一条路线是来源证明与溯源标准，把“谁生成、如何编辑、是否被篡改”变成可验证的证据链。SynthID-Text^[1014]代表了更贴近落地的文本水印方案，强调对文本质量影响小、检测高效且不依赖原模型。C2PA^[1015]通过内容凭证、清单等机制把加密绑定的溯源信息标准化，支持跨平台验证内容的来源与修改历史。两者共同指向的趋势是未来可信内容不再只靠单点检测工具，而更依赖标准化溯源协议与多方生态协同，把透明度与可追责性嵌入内容生产与分发流程。

综合这些脉络，未来大语言模型安全伦理的发展不会只沿着“更强的拒答、更严的过滤”直线前进，而会越来越像“安全工程、合规治理、可复现实证”的交汇点。学术界会继续把对齐做得更可泛化、更可解释，减少对人工经验与提示技巧的依赖，把红队评测做得更标准化、更贴近真实工作流，把智能体的安全边界做得更系统。治理上，风险分级与时间表式落地会推动组

织建立更严肃的责任体系，包括签字放行、风险可控证据、上线监测与回溯、事故纠偏与复盘。

8.2.3 跨学科融合

大语言模型正在把科研从“写作与检索的辅助工具”推向“可自主规划、执行与验证的科研智能体”，深刻改变知识生产的组织方式与效率边界。与此同时，AI for Science 的工业化落地与低门槛协作能力正在重塑生物医药、气候科学等领域的研发范式，并加速跨学科方法迁移与长期科研管理。

从工具向自主科研智能体的演进 DeepResearch 等工作的出现，标志着从生成式工具向自主科研智能体的跨越。基于大模型的科研智能体通过思维链和工具调用技术独立执行复杂 workflow，并能调用外部工具验证假设。这种转变不仅提升了研究效率，还实现了人机协作下的长周期科研管理。

AI for Science 的工业化进程 大模型通过对自然科学法则的数字化建模，加速了生物、物理及气象领域的工业化进程。在生物医药领域，利用 ProGen3 等模型可实现蛋白质定向设计，加速药物研发实验过程；在气候科学领域，AI 仿真器大幅降低了算力开销，显著提升了长周期气候预测效率。

科研门槛降低与跨学科协作加速 大模型可以充当不同学科之间的翻译桥梁，打破了学科间的专业术语壁垒。其低代码化的特性使得非计算机专业的学者也能应用复杂的量化模型，推动了方法论在不同学科间的迁移，特别是提升了数字人文和社会科学的实证研究能力。

本节围绕大语言模型迈向长期部署所面临的关键约束，系统梳理了算力资源不均、安全与伦理治理以及跨学科融合带来的主要挑战与结构性机遇。随着公共算力普惠、效率导向的算法与系统栈迭代、以及可审计的安全治理基础设施逐步成熟，大模型的发展路径将从“单点能力竞赛”走向“可持续、可参与、可验证”的生态化演进。

8.3 本章小结

本章围绕大语言模型迈向通用智能，总结技术进步的主轴正在从“规模驱动”转向“智能密度驱动”，强调复杂推理能力的可验证提升、更高效的训练/推理范式，以及多模态从工程拼接走向统一表示与统一目标的原生融合。

在这一过程中，幻觉与不确定性不再只是生成质量问题，而是决定模型能否进入医疗、法律、工业等高可靠场景的关键门槛，因此“训练—推理—系统”协同的事实性与可追溯治理将成为长期主线。

在应用范式上，LLM 正从对话式工具演进为具备目标分解、工具调用、执行监控与自我反思的智能体，并进一步与科学知识库、自动化实验平台和仿真系统耦合，推动 AI4Science 从“文献助手”走向“自主实验室”的闭环加速。同时，云边端协同与小型化、专业化模型的爆发，使得大模型能力开始以更低时延、更强隐私与更可控的方式渗透到真实工作流中，产业落地将更依赖“开源 + 闭源”的混合基座与可审计的工程体系。

面向未来，挑战与机遇将交织共进。算力资源不均将重塑创新版图，倒逼数据效率、系统优化与公共算力基础设施的制度化；安全伦理从“内容红线”升级为贯穿全生命周期的治理与可复现实证；跨学科融合既降低科研门槛、加速协作，也带来学术失范与可持续性压力。总体而言，真正决定这场跃迁能否稳健向善的，不仅是模型能力的上限，更是可靠性、可控性与治理能力能否同步进化。

参考文献

- [1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30 (引用页: 10, 18).
- [2] AINSLIE J, LEE-THORP J, DE JONG M, et al. Gqa: Training generalized multi-query transformer models from multi-head checkpoints[J]. arXiv preprint arXiv:2305.13245, 2023 (引用页: 10, 21).
- [3] LIU A, FENG B, WANG B, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model[J]. arXiv preprint arXiv:2405.04434, 2024 (引用页: 11, 13, 21).
- [4] XIAO G, TIAN Y, CHEN B, et al. Efficient streaming language models with attention sinks[J]. arXiv preprint arXiv:2309.17453, 2023 (引用页: 11).
- [5] SU J, AHMED M, LU Y, et al. Roformer: Enhanced transformer with rotary position embedding[J]. Neurocomputing, 2024, 568: 127063 (引用页: 11).
- [6] LIU N F, LIN K, HEWITT J, et al. Lost in the middle: How language models use long contexts[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 157-173 (引用页: 11, 12).
- [7] HU J, LI H, ZHANG Y, et al. Multi-matrix factorization attention[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 25114-25126 (引用页: 11).
- [8] ZHANG Y, LIU Y, YUAN H, et al. Tensor product attention is all you need[J]. arXiv preprint arXiv:2501.06425, 2025 (引用页: 12).
- [9] ZUHRI Z M, FUADI E H, AJI A F. Softpick: No Attention Sink,

- No Massive Activations with Rectified Softmax[J]. arXiv preprint arXiv:2504.20966, 2025 (引用页: 12).
- [10] AGARWAL S, AHMAD L, AI J, et al. gpt-oss-120b & gpt-oss-20b model card[J]. arXiv preprint arXiv:2508.10925, 2025 (引用页: 12, 21).
- [11] QIU Z, WANG Z, ZHENG B, et al. Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free [J]. arXiv preprint arXiv:2505.06708, 2025 (引用页: 12).
- [12] CHEN Y, LV A, LUAN J, et al. Hope: A novel positional encoding without long-term decay for enhanced context awareness and extrapolation[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 23044-23056 (引用页: 12).
- [13] WANG J, JI T, WU Y, et al. Length generalization of causal transformers without position encoding[C]//Findings of the Association for Computational Linguistics: ACL 2024. 2024: 14024-14040 (引用页: 13).
- [14] BARBERO F, VITVITSKYI A, PERIVOLAROPOULOS C, et al. Round and round we go! what makes rotary positional encodings useful?[J]. arXiv preprint arXiv:2410.06205, 2024 (引用页: 13).
- [15] META A. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation[J]. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, checked on, 2025, 4(7): 2025 (引用页: 13).
- [16] BELTAGY I, PETERS M E, COHAN A. Longformer: The long-document transformer[J]. arXiv preprint arXiv:2004.05150, 2020 (引用页: 13).
- [17] XIAO G, TANG J, ZUO J, et al. Duoattention: Efficient long-context llm inference with retrieval and streaming heads[J]. arXiv preprint arXiv:2410.10819, 2024 (引用页: 14).
- [18] LAI X, LU J, LUO Y, et al. Flexprefill: A context-aware sparse attention mechanism for efficient long-sequence inference[J]. arXiv preprint arXiv:2502.20766, 2025 (引用页: 14).

- [19] XU R, XIAO G, HUANG H, et al. Xattention: Block sparse attention with antidiagonal scoring[J]. arXiv preprint arXiv:2503.16428, 2025 (引用页: 14).
- [20] LU E, JIANG Z, LIU J, et al. Moba: Mixture of block attention for long-context llms[J]. arXiv preprint arXiv:2502.13189, 2025 (引用页: 14).
- [21] YUAN J, GAO H, DAI D, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 23078-23097 (引用页: 14).
- [22] LIU A, MEI A, LIN B, et al. Deepseek-v3. 2: Pushing the frontier of open large language models[J]. arXiv preprint arXiv:2512.02556, 2025 (引用页: 14, 37).
- [23] AI D. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning[J/OL]. arXiv preprint arXiv:2501.12948, 2025. <https://arxiv.org/abs/2501.12948> (引用页: 15-17, 70).
- [24] TEAM A Q. Qwen3 Technical Report[J/OL]. arXiv preprint arXiv:2505.09388, 2025. <https://arxiv.org/abs/2505.09388> (引用页: 15, 16).
- [25] AI M. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation[EB/OL]. 2025. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/> (引用页: 15).
- [26] AI Z. GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models[J/OL]. arXiv preprint arXiv:2508.06471, 2025. <https://arxiv.org/abs/2508.06471> (引用页: 16).
- [27] AI M. Kimi K2: Open Agentic Intelligence[J/OL]. arXiv preprint arXiv:2507.20534, 2025. <https://arxiv.org/abs/2507.20534> (引用页: 16).
- [28] TEAM A Q. Qwen3-Next: 迈向更极致的训练推理性价比[EB/OL]. 2025. <https://qwen.ai/blog?id=4074cca80393150c248e508aa62983f9cb7d27cd> (引用页: 16, 18).

- [29] AI M. LongCat-Flash Technical Report[J/OL]. arXiv preprint arXiv:2509.01322, 2025. <https://arxiv.org/abs/2509.01322> (引用页: 16).
- [30] DAI D, DENG C, ZHAO C, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models [J]. arXiv preprint arXiv:2401.06066, 2024 (引用页: 16).
- [31] TEAM T H. Hunyuan-TurboS: Advancing Large Language Models through Mamba-Transformer Synergy and Adaptive Chain-of-Thought[J/OL]. arXiv preprint arXiv:2505.15431, 2025. <http://arxiv.org/abs/2505.15431> (引用页: 16).
- [32] AI A G. LLaDA-MoE: A Sparse MoE Diffusion Language Model [J/OL]. arXiv preprint arXiv:2509.24389, 2025. <http://arxiv.org/abs/2509.24389> (引用页: 17).
- [33] OpenAI. Introducing gpt-oss[EB/OL]. 2025. <https://openai.com/zh-Hans-CN/index/introducing-gpt-oss/> (引用页: 17, 18).
- [34] AI D. Expert Parallelism Load Balancer (EPLB)[J/OL]. GitHub Repository, 2025. <https://github.com/deepseek-ai/EPLB> (引用页: 17).
- [35] AI D. Linear-Programming-Based Load Balancer (LPLB)[J/OL]. GitHub Repository, 2025. <https://github.com/deepseek-ai/LPLB> (引用页: 17).
- [36] TEAM A Q. Demons in the Detail: On Implementing Load Balancing Loss for Training Specialized Mixture-of-Expert Models[J/OL]. arXiv preprint arXiv:2501.11873, 2025. <https://arxiv.org/abs/2501.11873> (引用页: 17).
- [37] AI A G. Ling 2.0[EB/OL]. 2025. <https://github.com/inclusionAI/Ling-V2> (引用页: 17, 18).
- [38] AI A G. Towards Greater Leverage: Scaling Laws for Efficient Mixture-of-Experts Language Models[J/OL]. arXiv preprint arXiv:2507.17702, 2025. <http://arxiv.org/abs/2507.17702> (引用页: 17).

- [39] CLOUD H. Pangu Ultra MoE: How to Train Your Big MoE on Ascend NPUs[J/OL]. arXiv preprint arXiv:2505.04519, 2025. <http://arxiv.org/abs/2505.04519> (引用页: 18).
- [40] KATHAROPOULOS A, VYAS A, PAPPAS N, et al. Transformers are rnns: Fast autoregressive transformers with linear attention[C]//International conference on machine learning. 2020: 5156-5165 (引用页: 19).
- [41] SUN Y, DONG L, HUANG S, et al. Retentive network: A successor to transformer for large language models[J]. arXiv preprint arXiv:2307.08621, 2023 (引用页: 19).
- [42] QIN Z, YANG S, ZHONG Y. Hierarchically gated recurrent neural network for sequence modeling[J]. Advances in Neural Information Processing Systems, 2023, 36: 33202-33221 (引用页: 19).
- [43] QIN Z, YANG S, SUN W, et al. Hgrn2: Gated linear rnns with state expansion[J]. arXiv preprint arXiv:2404.07904, 2024 (引用页: 19).
- [44] GU A, DAO T. Mamba: Linear-time sequence modeling with selective state spaces[C]//First conference on language modeling. 2024 (引用页: 19, 317).
- [45] DAO T, GU A. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality[J]. arXiv preprint arXiv:2405.21060, 2024 (引用页: 19, 20).
- [46] PENG B, ALCAIDE E, ANTHONY Q, et al. Rwkv: Reinventing rnns for the transformer era[J]. arXiv preprint arXiv:2305.13048, 2023 (引用页: 19).
- [47] PENG B, GOLDSTEIN D, ANTHONY Q, et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence[J]. arXiv preprint arXiv:2404.05892, 2024 (引用页: 19).
- [48] SCHLAG I, IRIE K, SCHMIDHUBER J. Linear transformers are secretly fast weight programmers[C]//International conference on machine learning. 2021: 9355-9366 (引用页: 19, 20).
- [49] YANG S, WANG B, ZHANG Y, et al. Parallelizing linear transformers with the delta rule over sequence length[J]. Advances in

- neural information processing systems, 2024, 37: 115491-115522 (引用页: 19).
- [50] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901 (引用页: 19).
- [51] LIEBER O, LENZ B, BATA H, et al. Jamba: A hybrid transformer-mamba language model[J]. arXiv preprint arXiv:2403.19887, 2024 (引用页: 19).
- [52] REN L, LIU Y, LU Y, et al. Samba: Simple hybrid state space models for efficient unlimited context language modeling[J]. arXiv preprint arXiv:2406.07522, 2024 (引用页: 19).
- [53] WALEFFE R, BYEON W, RIACH D, et al. An empirical study of mamba-based language models[J]. arXiv preprint arXiv:2406.07887, 2024 (引用页: 19).
- [54] DONG X, FU Y, DIAO S, et al. Hymba: A hybrid-head architecture for small language models[J]. arXiv preprint arXiv:2411.13676, 2024 (引用页: 19).
- [55] YANG S, KAUTZ J, HATAMIZADEH A. Gated delta networks: Improving mamba2 with delta rule[J]. arXiv preprint arXiv:2412.06464, 2024 (引用页: 20).
- [56] HU J, PAN Y, DU J, et al. Comba: Improving Nonlinear RNNs with Closed-loop Control[J]. arXiv preprint arXiv:2506.02475, 2025 (引用页: 20).
- [57] PENG B, ZHANG R, GOLDSTEIN D, et al. Rwkv-7” goose” with expressive dynamic state evolution[J]. arXiv preprint arXiv:2503.14456, 2025 (引用页: 20).
- [58] TEAM K, ZHANG Y, LIN Z, et al. Kimi linear: An expressive, efficient attention architecture[J]. arXiv preprint arXiv:2510.26692, 2025 (引用页: 20, 21).
- [59] BEHROUZ A, ZHONG P, MIRROKNI V. Titans: Learning to memorize at test time[J]. arXiv preprint arXiv:2501.00663, 2024 (引用页: 20).

- [60] Anonymous. Mamba-3: Improved Sequence Modeling using State Space Principles[C/OL]//Submitted to The Fourteenth International Conference on Learning Representations. 2025. <https://openreview.net/forum?id=HwCvaJOiCj> (引用页: 20).
- [61] LI A, GONG B, YANG B, et al. Minimax-01: Scaling foundation models with lightning attention[J]. arXiv preprint arXiv:2501.08313, 2025 (引用页: 20, 21).
- [62] CHEN A, LI A, GONG B, et al. MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention[J]. arXiv preprint arXiv:2506.13585, 2025 (引用页: 21).
- [63] QIN Z, SUN W, LI D, et al. Lightning attention-2: A free lunch for handling unlimited sequence lengths in large language models [J]. arXiv preprint arXiv:2401.04658, 2024 (引用页: 21).
- [64] TEAM T H, LIU A, ZHOU B, et al. Hunyuan-turbos: Advancing large language models through mamba-transformer synergy and adaptive chain-of-thought[J]. arXiv preprint arXiv:2505.15431, 2025 (引用页: 21).
- [65] Qwen3-Next: Towards Ultimate Training & Inference Efficiency[Z]. <https://qwen.ai/blog?id=4074cca80393150c248e508aa62983f9cb7d27cd&from=research.latest-advancements-list>. [2025-12-21] (引用页: 21).
- [66] TEAM G, KAMATH A, FERRET J, et al. Gemma 3 technical report[J]. arXiv preprint arXiv:2503.19786, 2025 (引用页: 21).
- [67] Introducing MiMo-V2-Flash[Z]. <https://mimo.xiaomi.com/zh/blog/mimo-v2-flash>. [2025-12-21] (引用页: 21).
- [68] WANG D, ZHU R J, ABREU S, et al. A systematic analysis of hybrid linear attention[J]. arXiv preprint arXiv:2507.06457, 2025 (引用页: 21).
- [69] BAE S, ACUN B, HABEEB H, et al. Hybrid architectures for language models: Systematic analysis and design insights[J]. arXiv preprint arXiv:2510.04800, 2025 (引用页: 22).
- [70] LI Y, XIE R, YANG Z, et al. Transmamba: Flexibly switching between transformer and mamba[J]. arXiv preprint arXiv:2503.24067,

2025 (引用页: 22).

- [71] LAN D, SUN W, HU J, et al. Liger: Linearizing Large Language Models to Gated Recurrent Structures[J]. arXiv preprint arXiv:2503.01496, 2025 (引用页: 22).
- [72] LIU H, LI C, WU Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2023, 36: 34892-34916 (引用页: 22).
- [73] BAI S, CAI Y, CHEN R, et al. Qwen3-VL Technical Report[C/OL] //. 2025. <https://api.semanticscholar.org/CorpusID:283262018> (引用页: 23).
- [74] Baidu-ERNIE-Team. ERNIE 4.5 Technical Report[Z]. https://ernie.baidu.com/blog/publication/ERNIE_Technical_Report.pdf. 2025 (引用页: 23, 185, 195).
- [75] WANG W, GAO Z, GU L, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency [J]. arXiv preprint arXiv:2508.18265, 2025 (引用页: 23).
- [76] DENG C, ZHU D, LI K, et al. Emerging Properties in Unified Multimodal Pretraining[J/OL]. ArXiv, 2025, abs/2505.14683. <https://api.semanticscholar.org/CorpusID:278768720> (引用页: 24).
- [77] CHEN X, WU Z, LIU X, et al. Janus-Pro: Unified Multimodal Understanding and Generation with Data and Model Scaling[J/OL]. ArXiv, 2025, abs/2501.17811. <https://api.semanticscholar.org/CorpusID:275954151> (引用页: 24).
- [78] NIE S, ZHU F, YOU Z, et al. Large language diffusion models[J]. arXiv preprint arXiv:2502.09992, 2025 (引用页: 25).
- [79] YE J, XIE Z, ZHENG L, et al. Dream 7B: Diffusion Large Language Models[J]. arXiv preprint arXiv:2508.15487, 2025 (引用页: 25).
- [80] WANG G, LI J, SUN Y, et al. Hierarchical Reasoning Model[J]. arXiv preprint arXiv:2506.21734, 2025 (引用页: 26).
- [81] JOLICOEUR-MARTINEAU A. Less is more: Recursive reasoning with tiny networks[J]. arXiv preprint arXiv:2510.04871, 2025 (引用

页: 26).

- [82] ZHU R J, WANG Z, HUA K, et al. Scaling Latent Reasoning via Looped Language Models[J]. arXiv preprint arXiv:2510.25741, 2025 (引用页: 26).
- [83] BEHROUZ A, RAZAVIYAYN M, ZHONG P, et al. Nested learning: The illusion of deep learning architectures[C]//The Thirty-ninth Annual Conference on Neural Information Processing Systems. 2025 (引用页: 26).
- [84] HOULSBY N, GIURGIU A, JASTRZEBSKI S, et al. Parameter-efficient transfer learning for NLP[C]//International conference on machine learning. 2019: 2790-2799 (引用页: 29).
- [85] LIU X, JI K, FU Y, et al. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2022: 61-68 (引用页: 29).
- [86] HU E J, SHEN Y, WALLIS P, et al. Lora: Low-rank adaptation of large language models.[J]. ICLR, 2022, 1(2): 3 (引用页: 29).
- [87] ZHANG Y, LIU F, CHEN Y. LoRA-One: One-Step Full Gradient Could Suffice for Fine-Tuning Large Language Models, Provably and Efficiently[C]//Forty-second International Conference on Machine Learning (引用页: 30).
- [88] XU Y, LI C, YIN X, et al. Dual LoRA: Enhancing LoRA with Magnitude and Direction Updates[J]. arXiv preprint arXiv:2512.03402, 2025 (引用页: 30).
- [89] LIANG J, BHARADWAJ A. QR-LoRA: QR-Based Low-Rank Adaptation for Efficient Fine-Tuning of Large Language Models[J]. arXiv preprint arXiv:2508.21810, 2025 (引用页: 30).
- [90] KOIKE-AKINO T, TONIN F, WU Y, et al. Quantum-PEFT: Ultra parameter-efficient fine-tuning[C]//The Thirteenth International Conference on Learning Representations (引用页: 30).
- [91] LI K, HAN S, SU Q, et al. Uni-LoRA: One Vector is All You Need[C]

- //The 39th Conference on Neural Information Processing Systems (NeurIPS). 2025 (引用页: 30).
- [92] LIANG Y S, CHEN J R, LI W J. Gated Integration of Low-Rank Adaptation for Continual Learning of Large Language Models [J/OL]. 2025. arXiv: 2505.15424 [cs.CL]. <https://arxiv.org/abs/2505.15424> (引用页: 30).
- [93] ZHOU Y, LI R, ZHOU C, et al. BSLoRA: Enhancing the Parameter Efficiency of LoRA with Intra-Layer and Inter-Layer Sharing[C]//Forty-second International Conference on Machine Learning (引用页: 30).
- [94] SCHULMAN J, LAB T M. LoRA Without Regret[J/OL]. Thinking Machines Lab: Connectionism, 2025. DOI: 10.64434/tml.20250929 (引用页: 30).
- [95] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: Your language model is secretly a reward model[J]. Advances in neural information processing systems, 2023, 36: 53728-53741 (引用页: 34).
- [96] MENG Y, XIA M, CHEN D. Simpo: Simple preference optimization with a reference-free reward[C]//: vol. 37. 2024: 124198-124235 (引用页: 34, 35).
- [97] ETHAYARAJH K, XU W, MUENNIGHOFF N, et al. Kto: Model alignment as prospect theoretic optimization[J]. arXiv preprint arXiv:2402.01306, 2024 (引用页: 34).
- [98] JUNG S, HAN G, NAM D W, et al. Binary classifier optimization for large language model alignment[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 1858-1872 (引用页: 34, 35).
- [99] LIU T, ZHAO Y, JOSHI R, et al. Statistical Rejection Sampling Improves Preference Optimization[C]//The Twelfth International Conference on Learning Representations (引用页: 34, 35).
- [100] AZAR M G, GUO Z D, PIOT B, et al. A general theoretical paradigm to understand learning from human preferences[C]

- //International Conference on Artificial Intelligence and Statistics. 2024: 4447-4455 (引用页: 33, 34).
- [101] HAN J, JIANG M, SONG Y, et al. f -PO: Generalizing Preference Optimization with f -divergence Minimization[C]//International Conference on Artificial Intelligence and Statistics. 2025: 1144-1152 (引用页: 33, 34).
 - [102] LIU T, QIN Z, WU J, et al. Lipo: Listwise preference optimization through learning-to-rank[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025: 2404-2420 (引用页: 34, 35).
 - [103] WEN L, CAI Y, XIAO F, et al. Light-R1: Curriculum SFT, DPO and RL for Long COT from Scratch and Beyond[C/OL]//REHM G, LI Y. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track). Vienna, Austria: Association for Computational Linguistics, 2025: 318-327. <https://aclanthology.org/2025.acl-industry.24/>. DOI: 10.18653/v1/2025.acl-industry.24 (引用页: 35).
 - [104] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report [J]. arXiv preprint arXiv:2303.08774, 2023 (引用页: 36).
 - [105] SHAO Z, WANG P, ZHU Q, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024[J]. URL <https://arxiv.org/abs/2402.03300>, 2024, 2(3): 5 (引用页: 35, 36, 63).
 - [106] XIAO C, ZHANG M, CAO Y. BNPO: Beta Normalization Policy Optimization[J]. arXiv preprint arXiv:2506.02864, 2025 (引用页: 35, 36).
 - [107] LIU Z, CHEN C, LI W, et al. Understanding r1-zero-like training: A critical perspective[J]. arXiv preprint arXiv:2503.20783, 2025 (引用页: 35, 36).
 - [108] YU Q, ZHANG Z, ZHU R, et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale[J]. CoRR, 2025 (引用页: 35,

36).

- [109] ZHENG C, LIU S, LI M, et al. Group sequence policy optimization [J]. arXiv preprint arXiv:2507.18071, 2025 (引用页: 35, 36).
- [110] GAO C, ZHENG C, CHEN X H, et al. Soft Adaptive Policy Optimization[J]. arXiv preprint arXiv:2511.20347, 2025 (引用页: 35).
- [111] LIANG J, TANG H, MA Y, et al. Squeeze the Soaked Sponge: Efficient Off-policy Reinforcement Finetuning for Large Language Model[J/OL]. arXiv preprint arXiv:2507.06892, 2025. <https://arxiv.org/abs/2507.06892> (引用页: 37).
- [112] ZHAN R, LI Y, WANG Z, et al. ExGRPO: Learning to Reason from Experience[J/OL]. ArXiv preprint, 2025, 2510.02245. <https://arxiv.org/abs/2510.02245> (引用页: 37).
- [113] FU Y, CHEN T, CHAI J, et al. SRFT: A Single-Stage Method with Supervised and Reinforcement Fine-Tuning for Reasoning[J]. arXiv preprint arXiv:2506.19767, 2025 (引用页: 37, 57).
- [114] LIU Y, LI S, CAO L, et al. SuperRL: Reinforcement Learning with Supervision to Boost Language Model Reasoning[J/OL]. 2025. arXiv: 2506.01096 [cs.AI]. <https://arxiv.org/abs/2506.01096> (引用页: 37).
- [115] YE T, DONG L, CHI Z, et al. Black-Box On-Policy Distillation of Large Language Models[J]. arXiv preprint arXiv:2511.10643, 2025 (引用页: 37).
- [116] LIU Z, WANG P, XU R, et al. Inference-time scaling for generalist reward modeling[J]. arXiv preprint arXiv:2504.02495, 2025 (引用页: 37).
- [117] CHEN X, LI G, WANG Z, et al. Rm-r1: Reward modeling as reasoning[J]. arXiv preprint arXiv:2505.02387, 2025 (引用页: 37).
- [118] CHEN B, GAO X, HU C, et al. ReasonGRM: Enhancing Generative Reward Models through Large Reasoning Models[J]. arXiv preprint arXiv:2506.16712, 2025 (引用页: 38).

- [119] PENG H, QI Y, WANG X, et al. Agentic Reward Modeling: Integrating Human Preferences with Verifiable Correctness Signals for Reliable Reward Systems[J]. arXiv e-prints, 2025: arXiv-2502 (引用页: 38).
- [120] ZHAO J, LIU R, ZHANG K, et al. Genprm: Scaling test-time compute of process reward models via generative reasoning[J]. arXiv preprint arXiv:2504.00891, 2025 (引用页: 38, 57).
- [121] Anakin87. Environments Hub[J]. 2024 (引用页: 38).
- [122] E2B Team. E2B Sandbox[J]. 2024 (引用页: 38).
- [123] CARBONNEAUX Q, COHEN G, GEHRING J, et al. Cwm: An open-weights llm for research on code generation with world models [J]. arXiv e-prints, 2025: arXiv-2510 (引用页: 38).
- [124] CHAE H, KIM N, ONG K T I, et al. Web Agents with World Models: Learning and Leveraging Environment Dynamisc in Web Navigation[C]//The Thirteenth International Conference on Learning Representations (引用页: 39).
- [125] WU J, YIN S, FENG N, et al. RLVR-World: Training World Models with Reinforcement Learning[C]//Advances in Neural Information Processing Systems. 2025 (引用页: 39).
- [126] MAHABADI R K, SATHEESH S, PRABHUMOYE S, et al. Nemotron-CC-Math: A 133 Billion-Token-Scale High Quality Math Pretraining Dataset[J/OL]. 2025. arXiv: 2508.15096 [cs.CL]. <https://arxiv.org/abs/2508.15096> (引用页: 40-43).
- [127] ZHOU F, WANG Z, RANJAN N, et al. MegaMath: Pushing the Limits of Open Math Corpora[J/OL]. 2025. arXiv: 2504.02807 [cs.CL]. <https://arxiv.org/abs/2504.02807> (引用页: 40, 41).
- [128] FUJII K, TAJIMA Y, MIZUKI S, et al. Rewriting Pre-Training Data Boosts LLM Performance in Math and Code[J/OL]. 2025. arXiv: 2505.02881 [cs.LG]. <https://arxiv.org/abs/2505.02881> (引用页: 40, 41, 46).
- [129] ZHAO C, CHANG E, LIU Z, et al. MobileLLM-R1: Exploring the

Limits of Sub-Billion Language Model Reasoners with Open Training Recipes[J/OL]. 2025. arXiv: 2509.24945 [cs.CL]. <https://arxiv.org/abs/2509.24945> (引用页: 40, 41).

- [130] FAN R Z, WANG Z, LIU P. MegaScience: Pushing the Frontiers of Post-Training Datasets for Science Reasoning[J/OL]. 2025. arXiv: 2507.16812 [cs.CL]. <https://arxiv.org/abs/2507.16812> (引用页: 40, 41).
- [131] YUAN W, YU J, JIANG S, et al. NaturalReasoning: Reasoning in the Wild with 2.8M Challenging Questions[J/OL]. 2025. arXiv: 2502.13124 [cs.CL]. <https://arxiv.org/abs/2502.13124> (引用页: 40, 41).
- [132] LANGLAIS P C, HINOSTROZA C R, NEE M, et al. Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training [J/OL]. 2025. arXiv: 2506.01732 [cs.CL]. <https://arxiv.org/abs/2506.01732> (引用页: 40, 41).
- [133] LIU Y, ZHANG L L, ZHU Y, et al. rStar-Coder: Scaling Competitive Code Reasoning with a Large-Scale Verified Dataset[J/OL]. 2025. arXiv: 2505.21297 [cs.CL]. <https://arxiv.org/abs/2505.21297> (引用页: 40, 41).
- [134] SEED B, ZHANG Y, SU J, et al. Seed-Coder: Let the Code Model Curate Data for Itself[J/OL]. 2025. arXiv: 2506.03524 [cs.CL]. <https://arxiv.org/abs/2506.03524> (引用页: 41).
- [135] TU C, ZHANG X, WENG R, et al. A Survey on LLM Mid-Training [J/OL]. 2025. arXiv: 2510.23081 [cs.CL]. <https://arxiv.org/abs/2510.23081> (引用页: 41).
- [136] TEAM. M L. Introducing LongCat-Flash-Thinking: A Technical Report[J/OL]. 2025. arXiv: 2509.18883 [cs.AI]. <https://arxiv.org/abs/2509.18883> (引用页: 42).
- [137] TEAM. M L. LongCat-Flash Technical Report[J/OL]. 2025. arXiv: 2509.01322 [cs.CL]. <https://arxiv.org/abs/2509.01322> (引用页: 42).

- [138] Et AL. B H. dots.llm1 Technical Report[J/OL]. 2025. arXiv: 2506.05767 [cs.CL]. <https://arxiv.org/abs/2506.05767> (引用页: 42).
- [139] TEAM. L C X. MiMo: Unlocking the Reasoning Potential of Language Model – From Pretraining to Posttraining[J/OL]. 2025. arXiv: 2505.07608 [cs.CL]. <https://arxiv.org/abs/2505.07608> (引用页: 42).
- [140] Et AL. A Y. Qwen3 Technical Report[J/OL]. 2025. arXiv: 2505.09388 [cs.CL]. <https://arxiv.org/abs/2505.09388> (引用页: 42).
- [141] Et AL. Y Y. Pangu Ultra: Pushing the Limits of Dense Large Language Models on Ascend NPUs[J/OL]. 2025. arXiv: 2504.07866 [cs.CL]. <https://arxiv.org/abs/2504.07866> (引用页: 42).
- [142] TEAM. K. Kimi k1.5: Scaling Reinforcement Learning with LLMs [J/OL]. 2025. arXiv: 2501.12599 [cs.AI]. <https://arxiv.org/abs/2501.12599> (引用页: 42).
- [143] HE Z, LIANG T, XU J, et al. DeepMath-103K: A Large-Scale, Challenging, Decontaminated, and Verifiable Mathematical Dataset for Advancing Reasoning[J/OL]. 2025. arXiv: 2504.11456 [cs.CL]. <https://arxiv.org/abs/2504.11456> (引用页: 41, 43).
- [144] MAHDAVI S, LI M, LIU K, et al. Leveraging Online Olympiad-Level Math Problems for LLMs Training and Contamination-Resistant Evaluation[J/OL]. 2025. arXiv: 2501.14275 [cs.CL]. <https://arxiv.org/abs/2501.14275> (引用页: 42, 43).
- [145] ZHENG S, CHENG Q, YAO J, et al. Scaling Physical Reasoning with the PHYSICS Dataset[J/OL]. 2025. arXiv: 2506.00022 [cs.CL]. <https://arxiv.org/abs/2506.00022> (引用页: 42, 43).
- [146] Et AL. Y W. SciReasoner: Laying the Scientific Reasoning Ground Across Disciplines[J/OL]. 2025. arXiv: 2509.21320 [cs.CL]. <https://arxiv.org/abs/2509.21320> (引用页: 42, 43).
- [147] Et AL. A B. Llama-Nemotron: Efficient Reasoning Models[J/OL]. 2025. arXiv: 2505.00949 [cs.CL]. <https://arxiv.org/abs/2505.00949> (引用页: 42, 43).

- [148] LI J, GUO D, YANG D, et al. CodeI/O: Condensing Reasoning Patterns via Code Input-Output Prediction[J/OL]. 2025. arXiv: 2502.07316 [cs.CL]. <https://arxiv.org/abs/2502.07316> (引用页: 42, 43).
- [149] AHMAD W U, FICEK A, SAMADI M, et al. OpenCodeInstruct: A Large-scale Instruction Tuning Dataset for Code LLMs[J/OL]. 2025. arXiv: 2504.04030 [cs.SE]. <https://arxiv.org/abs/2504.04030> (引用页: 42, 43).
- [150] AHMAD W U, NARENTHIRAN S, MAJUMDAR S, et al. OpenCodeReasoning: Advancing Data Distillation for Competitive Coding[J/OL]. 2025. arXiv: 2504.01943 [cs.CL]. <https://arxiv.org/abs/2504.01943> (引用页: 43).
- [151] GOHARI H E, KADHE S R, SHAH S Y, et al. GneissWeb: Preparing High Quality Data for LLMs at Scale[J/OL]. 2025. arXiv: 2502.14907 [cs.CL]. <https://arxiv.org/abs/2502.14907> (引用页: 44).
- [152] BRANDIZZI N, ABDELWAHAB H, BHOWMICK A, et al. Data Processing for the OpenGPT-X Model Family[J/OL]. 2025. arXiv: 2410.08800 [cs.CL]. <https://arxiv.org/abs/2410.08800> (引用页: 44).
- [153] ZHENG J, CAI X, QIU S, et al. Spurious Forgetting in Continual Learning of Language Models[J/OL]. 2025. arXiv: 2501.13453 [cs.LG]. <https://arxiv.org/abs/2501.13453> (引用页: 44).
- [154] ZHANG Y, DU L. Data Curation Through the Lens of Spectral Dynamics: Static Limits, Dynamic Acceleration, and Practical Oracles[J/OL]. 2025. arXiv: 2512.02409 [cs.LG]. <https://arxiv.org/abs/2512.02409> (引用页: 45).
- [155] WANG Y, FU Z, CAI J, et al. Ultra-FineWeb: Efficient Data Filtering and Verification for High-Quality LLM Training Data[J/OL]. 2025. arXiv: 2505.05427 [cs.CL]. <https://arxiv.org/abs/2505.05427> (引用页: 45).
- [156] SHUM K, HUANG Y, ZOU H, et al. Predictive Data Selection: The Data That Predicts Is the Data That Teaches[J/OL]. 2025. arXiv:

- 2503.00808 [cs.CL]. <https://arxiv.org/abs/2503.00808> (引用页: 45).
- [157] SUBRAMANIAN S, VERMA A. Modular Techniques for Synthetic Long-Context Data Generation in Language Model Training and Evaluation[J/OL]. 2025. arXiv: 2509.01185 [cs.CL]. <https://arxiv.org/abs/2509.01185> (引用页: 45).
- [158] COALSON Z, BAE J, CARLINI N, et al. IF-GUIDE: Influence Function-Guided Detoxification of LLMs[J/OL]. 2025. arXiv: 2506.01790 [cs.LG]. <https://arxiv.org/abs/2506.01790> (引用页: 45).
- [159] MENDU S K, YENALA H, GULATI A, et al. Towards Safer Pre-training: Analyzing and Filtering Harmful Content in Webscale datasets for Responsible LLMs[J/OL]. 2025. arXiv: 2505.02009 [cs.CL]. <https://arxiv.org/abs/2505.02009> (引用页: 45).
- [160] YANG W, LI L, TAO X, et al. Factor Decorrelation Enhanced Data Removal from Deep Predictive Models[J/OL]. 2025. arXiv: 2509.23443 [cs.LG]. <https://arxiv.org/abs/2509.23443> (引用页: 45).
- [161] BAYER O, ULU E N, SARKIN Y, et al. A REGNLP Framework: Developing Retrieval-Augmented Generation for Regulatory Document Analysis[C/OL]//GOKHAN T, WANG K, GUREVYCH I, et al. Proceedings of the 1st Regulatory NLP Workshop (RegNLP 2025). Abu Dhabi, UAE: Association for Computational Linguistics, 2025: 97-101. <https://aclanthology.org/2025.regnlp-1.15/> (引用页: 46).
- [162] KIM S, SONG H, SEO H, et al. Optimizing Retrieval Strategies for Financial Question Answering Documents in Retrieval-Augmented Generation Systems[J/OL]. 2025. arXiv: 2503.15191 [cs.IR]. <https://arxiv.org/abs/2503.15191> (引用页: 46).
- [163] BAO R, XUE N, SUN Y, et al. Dynamic Quality-Latency Aware Routing for LLM Inference in Wireless Edge-Device Networks [J/OL]. 2025. arXiv: 2508.11291 [cs.IT]. <https://arxiv.org/abs/2508.11291> (引用页: 46).
- [164] WANG J, XIANG D, XU J, et al. TANDEM: Bi-Level Data Mixture

- Optimization with Twin Networks[C/OL]//The Thirty-ninth Annual Conference on Neural Information Processing Systems. 2025. <https://openreview.net/forum?id=szBFUtBzWP> (引用页: 46).
- [165] SHUKOR M, BETHUNE L, BUSBRIDGE D, et al. Scaling Laws for Optimal Data Mixtures[J/OL]. 2025. arXiv: 2507.09404 [cs.LG]. <https://arxiv.org/abs/2507.09404> (引用页: 46).
- [166] LIU Q, ZHENG X, MUENNIGHOFF N, et al. RegMix: Data Mixture as Regression for Language Model Pre-training[J/OL]. 2025. arXiv: 2407.01492 [cs.CL]. <https://arxiv.org/abs/2407.01492> (引用页: 46).
- [167] CHEN M F, HU M Y, LOURIE N, et al. Aioli: A Unified Optimization Framework for Language Model Data Mixing[J/OL]. 2025. arXiv: 2411.05735 [cs.LG]. <https://arxiv.org/abs/2411.05735> (引用页: 46).
- [168] ZHU X, CHENG D, LI H, et al. How to Synthesize Text Data without Model Collapse?[J/OL]. 2025. arXiv: 2412.14689 [cs.CL]. <https://arxiv.org/abs/2412.14689> (引用页: 46).
- [169] KESSLER S, XIA M, DIAZ D M, et al. Towards Active Synthetic Data Generation for Finetuning Language Models[J/OL]. 2025. arXiv: 2512.00884 [cs.LG]. <https://arxiv.org/abs/2512.00884> (引用页: 47).
- [170] HE L, WANG J, WEBER M, et al. Scaling Instruction-Tuned LLMs to Million-Token Contexts via Hierarchical Synthetic Data Generation[J/OL]. 2025. arXiv: 2504.12637 [cs.CL]. <https://arxiv.org/abs/2504.12637> (引用页: 47).
- [171] IYER V, CHEN P, REI R, et al. XL-Suite: Cross-Lingual Synthetic Training and Evaluation Data for Open-Ended Generation[J/OL]. 2025. arXiv: 2503.22973 [cs.CL]. <https://arxiv.org/abs/2503.22973> (引用页: 47).
- [172] RODRIGUEZ J, JIAN X, PANIGRAHI S S, et al. BigDocs: An Open Dataset for Training Multimodal Models on Document and Code Tasks[J/OL]. 2025. arXiv: 2412.04626 [cs.LG]. <https://arxiv.org/abs/2412.04626>

- v.org/abs/2412.04626 (引用页: 48).
- [173] XUE Y, XIE X, KOSTYRKO M, et al. InfiniHuman: Realistic 3D Human Creation with Precise Control[C/OL]//SA Conference Papers ' 25: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. ACM, 2025: 1-12. <http://dx.doi.org/10.1145/3757377.3763815>. DOI: 10.1145/3757377.3763815 (引用页: 48).
 - [174] XIE Y, ZHOU C, GAO L, et al. MedTrinity-25M: A Large-scale Multimodal Dataset with Multigranular Annotations for Medicine [J/OL]. 2025. arXiv: 2408.02900 [cs.CV]. <https://arxiv.org/abs/2408.02900> (引用页: 48).
 - [175] WANG J, ZHANG Y, LIU M, et al. PIN: A Knowledge-Intensive Dataset for Paired and Interleaved Multimodal Documents[J/OL]. 2025. arXiv: 2406.13923 [cs.AI]. <https://arxiv.org/abs/2406.13923> (引用页: 48).
 - [176] XU W, WANG C, LIANG D, et al. NAUTILUS: A Large Multimodal Model for Underwater Scene Understanding[J/OL]. 2025. arXiv: 2510.27481 [cs.CV]. <https://arxiv.org/abs/2510.27481> (引用页: 48).
 - [177] YANG B, LI W, CHEN D, et al. VideoMind: An Omni-Modal Video Dataset with Intent Grounding for Deep-Cognitive Video Understanding[J/OL]. 2025. arXiv: 2507.18552 [cs.CV]. <https://arxiv.org/abs/2507.18552> (引用页: 48).
 - [178] CHEN J, XU Z, PAN X, et al. BLIP3-o: A Family of Fully Open Unified Multimodal Models-Architecture, Training and Dataset[J/OL]. 2025. arXiv: 2505.09568 [cs.CV]. <https://arxiv.org/abs/2505.09568> (引用页: 48, 49).
 - [179] DUAN C, SUN K, FANG R, et al. CodePlot-CoT: Mathematical Visual Reasoning by Thinking with Code-Driven Images[J/OL]. 2025. arXiv: 2510.11718 [cs.CV]. <https://arxiv.org/abs/2510.11718> (引用页: 48, 49).
 - [180] LIU R, ZHENG J, CHEN Y, et al. Situat3DChange: Situated 3D Change Understanding Dataset for Multimodal Large Language

- Model[J/OL]. 2025. arXiv: 2510.11509 [cs.CV]. <https://arxiv.org/abs/2510.11509> (引用页: 48, 49).
- [181] BURNHAM G, ADAMCZEWSKI T. LLMs now accept longer inputs, and the best models can use them more effectively[J/OL]. 2025. <https://epoch.ai/data-insights/context-windows> (引用页: 50).
- [182] ZHAO L, WEI T, ZENG L, et al. LongSkywork: A Training Recipe for Efficiently Extending Context Length in Large Language Models[J/OL]. 2024. arXiv: 2406.00605 [cs.CL]. <https://arxiv.org/abs/2406.00605> (引用页: 50).
- [183] TIAN J, ZHENG D, CHENG Y, et al. Untie the Knots: An Efficient Data Augmentation Strategy for Long-Context Pre-Training in Language Models[J/OL]. 2024. arXiv: 2409.04774 [cs.CL]. <https://arxiv.org/abs/2409.04774> (引用页: 50).
- [184] AN S, MA Z, LIN Z, et al. Make Your LLM Fully Utilize the Context[J/OL]. 2024. arXiv: 2404.16811 [cs.CL]. <https://arxiv.org/abs/2404.16811> (引用页: 50).
- [185] GAO C, WU X, LIN Z, et al. NExtLong: Toward Effective Long-Context Training without Long Documents[J/OL]. 2025. arXiv: 2501.12766 [cs.CL]. <https://arxiv.org/abs/2501.12766> (引用页: 51).
- [186] GAO T, WETTIG A, YEN H, et al. How to Train Long-Context Language Models (Effectively)[J/OL]. 2025. arXiv: 2410.02660 [cs.CL]. <https://arxiv.org/abs/2410.02660> (引用页: 51, 52).
- [187] ZHOU Z, LI C, CHEN X, et al. LLM \times MapReduce: Simplified Long-Sequence Processing using Large Language Models[J/OL]. 2024. arXiv: 2410.09342 [cs.CL]. <https://arxiv.org/abs/2410.09342> (引用页: 51, 52).
- [188] SI S, ZHAO H, CHEN G, et al. GATEAU: Selecting Influential Samples for Long Context Alignment[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025:

- 7391-7422. <https://aclanthology.org/2025.emnlp-main.375/>. DOI: 10.18653/v1/2025.emnlp-main.375 (引用页: 51, 52).
- [189] CHEN Z, CHEN Q, QIN L, et al. What are the Essential Factors in Crafting Effective Long Context Multi-Hop Instruction Datasets? Insights and Best Practices[J/OL]. 2025. arXiv: 2409.01893 [cs.CL]. <https://arxiv.org/abs/2409.01893> (引用页: 51, 52).
- [190] YANG Z, SHEN W, CHEN R, et al. SPELL: Self-Play Reinforcement Learning for evolving Long-Context Language Models[J/OL]. 2025. arXiv: 2509.23863 [cs.CL]. <https://arxiv.org/abs/2509.23863> (引用页: 51, 52).
- [191] ZHANG J, HOU Z, LV X, et al. LongReward: Improving Long-context Large Language Models with AI Feedback[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 3718-3739. <https://aclanthology.org/2025.acl-long.187/>. DOI: 10.18653/v1/2025.acl-long.187 (引用页: 51, 52).
- [192] CHEN G, LI X, SHIEH M Q, et al. LongPO: Long Context Self-Evolution of Large Language Models through Short-to-Long Preference Optimization[J/OL]. 2025. arXiv: 2502.13922 [cs.CL]. <https://arxiv.org/abs/2502.13922> (引用页: 51, 52).
- [193] PHAM C M, CHANG Y, IYYER M. CLIPPER: Compression enables long-context synthetic data generation[J/OL]. 2025. arXiv: 2502.14854 [cs.CL]. <https://arxiv.org/abs/2502.14854> (引用页: 52).
- [194] GAO C, WU X, LIN Z, et al. LongMagpie: A Self-synthesis Method for Generating Large-scale Long-context Instructions[J/OL]. 2025. arXiv: 2505.17134 [cs.CL]. <https://arxiv.org/abs/2505.17134> (引用页: 52).
- [195] BAI Y, TU S, ZHANG J, et al. LongBench v2: Towards Deeper Understanding and Reasoning on Realistic Long-context Multitasks[J/OL]. 2025. arXiv: 2412.15204 [cs.CL]. <https://arxiv.org/abs/2412.15204> (引用页: 52).

- [196] WU J, GU G, ZHENG Y, et al. Ref-Long: Benchmarking the Long-context Referencing Capability of Long-context Language Models [J/OL]. 2025. arXiv: 2507.09506 [cs.CL]. <https://arxiv.org/abs/2507.09506> (引用页: 53).
- [197] THONET T, BESACIER L, ROZEN J. ELITR-Bench: A Meeting Assistant Benchmark for Long-Context Language Models[C/OL]// RAMBOW O, WANNER L, APIDIANAKI M, et al. Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: Association for Computational Linguistics, 2025: 407-428. <https://aclanthology.org/2025.coling-main.28/> (引用页: 53).
- [198] LI J, GUO X, LI L, et al. LONGCODEU: Benchmarking Long-Context Language Models on Long Code Understanding[J/OL]. 2025. arXiv: 2503.04359 [cs.SE]. <https://arxiv.org/abs/2503.04359> (引用页: 53).
- [199] WU X, WANG M, LIU Y, et al. LIFBench: Evaluating the Instruction Following Performance and Stability of Large Language Models in Long-Context Scenarios[J/OL]. 2025. arXiv: 2411.07037 [cs.CL]. <https://arxiv.org/abs/2411.07037> (引用页: 53).
- [200] HUANG Z, LING G, ZHONG S, et al. MiniLongBench: The Low-cost Long Context Understanding Benchmark for Large Language Models[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 11442-11460. <https://aclanthology.org/2025.acl-long.560/>. DOI: 10.18653/v1/2025.acl-long.560 (引用页: 53).
- [201] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824-24837 (引用页: 55).
- [202] LUONG T Q, ZHANG X, JIE Z, et al. Reft: Reasoning with rein-

- forced fine-tuning[J]. arXiv preprint arXiv:2401.08967, 2024 (引用页: 54, 55).
- [203] XIE Y, GOYAL A, ZHENG W, et al. Monte carlo tree search boosts reasoning via iterative preference learning[J]. arXiv preprint arXiv:2405.00451, 2024 (引用页: 54, 55).
 - [204] DeepSeek-AI, GUO D, YANG D, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning[J]. ArXiv, 2025, abs/2501.12948 (引用页: 54, 55, 59).
 - [205] HUANG C, YU W, WANG X, et al. R-zero: Self-evolving reasoning llm from zero data[J]. arXiv preprint arXiv:2508.05004, 2025 (引用页: 55, 134).
 - [206] XIA P, ZENG K, LIU J, et al. Agent0: Unleashing self-evolving agents from zero data via tool-integrated reasoning[J]. arXiv preprint arXiv:2511.16043, 2025 (引用页: 55).
 - [207] DONG G, BAO L, WANG Z, et al. Agentic entropy-balanced policy optimization[J]. arXiv preprint arXiv:2510.14545, 2025 (引用页: 56).
 - [208] LI Y, GU Q, WEN Z, et al. Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling[J]. arXiv preprint arXiv:2508.17445, 2025 (引用页: 56).
 - [209] CHEN X, ZHU W, QIU P, et al. Dra-grpo: Exploring diversity-aware reward adjustment for rl-zero-like training of large language models[J]. arXiv preprint arXiv:2505.09655, 2025 (引用页: 56).
 - [210] YUE Y, CHEN Z, LU R, et al. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?[J]. arXiv preprint arXiv:2504.13837, 2025 (引用页: 56).
 - [211] LIU M, FARINA G, OZDAGLAR A. UFT: Unifying Supervised and Reinforcement Fine-Tuning[J]. arXiv preprint arXiv:2505.16984, 2025 (引用页: 57).
 - [212] ZHANG K, LV A, LI J, et al. StepHint: Multi-level Stepwise Hints Enhance Reinforcement Learning to Reason[J]. arXiv preprint arXiv:2507.02841, 2025 (引用页: 57).

- [213] HE T, MU R, LIAO L, et al. Good learners think their thinking: Generative PRM makes large reasoning model more efficient math learner[J]. arXiv preprint arXiv:2507.23317, 2025 (引用页: 57).
- [214] YI J, WANG J, LI S. Shorterbetter: Guiding reasoning models to find optimal inference length for efficient reasoning[J]. arXiv preprint arXiv:2504.21370, 2025 (引用页: 57).
- [215] HUANG S, WANG H, ZHONG W, et al. AdaCtrl: Towards Adaptive and Controllable Reasoning via Difficulty-Aware Budgeting[J]. arXiv preprint arXiv:2505.18822, 2025 (引用页: 57).
- [216] ZHANG J, LIN N, HOU L, et al. Adaptthink: Reasoning models can learn when to think[J]. arXiv preprint arXiv:2505.13417, 2025 (引用页: 57).
- [217] LUO H, HE H, WANG Y, et al. Adar1: From long-cot to hybrid-cot via bi-level adaptive reasoning optimization[J]. arXiv e-prints, 2025: arXiv-2504 (引用页: 57).
- [218] KANG Y, SUN X, CHEN L, et al. C3oT: Generating Shorter Chain-of-Thought without Compromising Effectiveness[J]. ArXiv, 2024, abs/2412.11664 (引用页: 58).
- [219] XIA H, LI Y, LEONG C T, et al. TokenSkip: Controllable Chain-of-Thought Compression in LLMs[J]. ArXiv, 2025, abs/2502.12067 (引用页: 58).
- [220] FENG J, HUANG S, QU X, et al. ReTool: Reinforcement Learning for Strategic Tool Use in LLMs[J]. ArXiv, 2025, abs/2504.11536 (引用页: 58).
- [221] LI C, TANG Z, LI Z, et al. CoRT: Code-integrated Reasoning within Thinking[J]. ArXiv, 2025, abs/2506.09820 (引用页: 58).
- [222] SHANG N, LIU Y, ZHU Y, et al. rStar2-Agent: Agentic Reasoning Technical Report[J]. ArXiv, 2025, abs/2508.20722 (引用页: 58).
- [223] SONG H, JIANG J, MIN Y, et al. R1-Searcher: Incentivizing the Search Capability in LLMs via Reinforcement Learning[J]. ArXiv, 2025, abs/2503.05592 (引用页: 58).

- [224] JIN B, ZENG H, YUE Z, et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning[J]. ArXiv, 2025, abs/2503.09516 (引用页: 58).
- [225] WANG Z, ZHENG X, AN K, et al. StepSearch: Igniting LLMs Search Ability via Step-Wise Proximal Policy Optimization[J]. ArXiv, 2025, abs/2505.15107 (引用页: 59).
- [226] YIN S, LEI T, LIU Y. ToolVQA: A Dataset for Multi-step Reasoning VQA with External Tools[J]. ArXiv, 2025, abs/2508.03284 (引用页: 59).
- [227] HU W, HONG Y, WANG Y, et al. 3DLLM-Mem: Long-Term Spatial-Temporal Memory for Embodied 3D Large Language Model [J]. ArXiv, 2025, abs/2505.22657 (引用页: 59).
- [228] SHARMA A, NGUYEN L, GUPTA A, et al. Inducing Causal World Models in LLMs for Zero-Shot Physical Reasoning[J]. ArXiv, 2025, abs/2507.19855 (引用页: 59).
- [229] LUMER E, GULATI A, SUBBIAH V K, et al. MemTool: Optimizing Short-Term Memory Management for Dynamic Tool Calling in LLM Agent Multi-Turn Conversations[J]. ArXiv, 2025, abs/2507.21428 (引用页: 59).
- [230] PANDEY T, GHUKASYAN A, GOKTAS O, et al. Adaptive Graph of Thoughts: Test-Time Adaptive Reasoning Unifying Chain, Tree, and Graph Structures[J]. ArXiv, 2025, abs/2502.05078 (引用页: 59, 60).
- [231] QU Y, SINGH A, LEE Y, et al. RLAD: Training LLMs to Discover Abstractions for Solving Reasoning Problems[J]. ArXiv, 2025, abs/2510.02263 (引用页: 59, 60).
- [232] WEN H, SU Y, ZHANG F, et al. ParaThinker: Native Parallel Thinking as a New Paradigm to Scale LLM Test-time Compute [J]. ArXiv, 2025, abs/2509.04475 (引用页: 59, 60).
- [233] HAO S, SUKHBAATAR S, SU D, et al. Training large language models to reason in a continuous latent space[J]. arXiv preprint arXiv:2412.06769, 2024 (引用页: 59, 60).

- [234] CHANG Q, ZHANG Z, HU P, et al. Thor: Tool-integrated hierarchical optimization via rl for mathematical reasoning[J]. arXiv preprint arXiv:2509.13761, 2025 (引用页: 59, 60).
- [235] ZHU H, HAO S, HU Z, et al. Reasoning by Superposition: A Theoretical Perspective on Chain of Continuous Thought[J]. ArXiv, 2025, abs/2505.12514 (引用页: 60).
- [236] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report [J]. arXiv preprint arXiv:2303.08774, 2023 (引用页: 61).
- [237] ZHU Q, GUO D, SHAO Z, et al. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence[J]. arXiv preprint arXiv:2406.11931, 2024 (引用页: 61).
- [238] HE Z, LIANG T, XU J, et al. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning[J]. arXiv preprint arXiv:2504.11456, 2025 (引用页: 62).
- [239] GUHA E, MARTEN R, KEH S, et al. OpenThoughts: Data Recipes for Reasoning Models[J]. arXiv preprint arXiv:2506.04178, 2025 (引用页: 62).
- [240] MOSHKOV I, HANLEY D, SOROKIN I, et al. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset[J]. arXiv preprint arXiv:2504.16891, 2025 (引用页: 62).
- [241] AHMAD W U, NARENTHIRAN S, MAJUMDAR S, et al. Open-codereasoning: Advancing data distillation for competitive coding [J]. arXiv preprint arXiv:2504.01943, 2025 (引用页: 62).
- [242] GUO D, YANG D, ZHANG H, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning[J]. Nature, 2025, 645(8081): 633-638 (引用页: 63, 85, 87, 315, 316).
- [243] CHRISTIANO P F, LEIKE J, BROWN T, et al. Deep reinforcement learning from human preferences[J]. Advances in neural information processing systems, 2017, 30 (引用页: 63).

- [244] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017 (引用页: 63).
- [245] YANG J, LIERET K, JIMENEZ C E, et al. Swe-smith: Scaling data for software engineering agents[J]. arXiv preprint arXiv:2504.21798, 2025 (引用页: 63).
- [246] WANG J, ZAN D, XIN S, et al. Swe-mirror: Scaling issue-resolving datasets by mirroring issues across repositories[J]. arXiv preprint arXiv:2509.08724, 2025 (引用页: 63).
- [247] PHAM M V, PHAN H N, PHAN H N, et al. SWE-Synth: Synthesizing Verifiable Bug-Fix Data to Enable Large Language Models in Resolving Real-World Bugs[J]. arXiv preprint arXiv:2504.14757, 2025 (引用页: 63).
- [248] DU Y, CAI Y, ZHOU Y, et al. SWE-Dev: Evaluating and Training Autonomous Feature-Driven Software Development[J]. arXiv preprint arXiv:2505.16975, 2025 (引用页: 63).
- [249] PAN J, WANG X, NEUBIG G, et al. Training Software Engineering Agents and Verifiers with SWE-Gym[C]//Forty-second International Conference on Machine Learning (引用页: 63).
- [250] ZHANG H, et al. AgentRL: Scaling Agentic Reinforcement Learning with a Multi-Turn, Multi-Task Framework[C]//Advances in Neural Information Processing Systems (NeurIPS). 2025 (引用页: 64).
- [251] JIN Y, XU K, LI H, et al. ReVeal: Self-Evolving Code Agents via Iterative Generation-Verification[J]. arXiv preprint arXiv:2506.11442, 2025 (引用页: 64).
- [252] LUO M, JAIN N, SINGH J, et al. DeepSWE: Training a state-of-the-art coding agent from scratch by scaling RL[J]. Notion page, 2025 (引用页: 64).
- [253] QU C, DAI S, WEI X, et al. Tool Learning with Large Language Models: A Survey[J/OL]. Frontiers of Computer Science, 2025, 19(8):198343. arXiv: 2405.17935 [cs] [2025-12-16]. DOI: 10.1007/s11704-024-40678-2 (引用页: 64).

- [254] MA Z, LIU J, LUO X, et al. Advancing Tool-Augmented Large Language Models via Meta-Verification and Reflection Learning[C/OL]//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2. 2025: 2078-2089. arXiv: 2506.04625 [cs] [2025-12-16]. DOI: 10.1145/3711896.3736835 (引用页: 64-67).
- [255] CHEN M, SUN H, LI T, et al. Facilitating Multi-turn Function Calling for LLMs via Compositional Instruction Tuning[J/OL]. 2025(arXiv:2410.12952). arXiv: 2410.12952 [cs] [2025-12-16]. DOI: 10.48550/arXiv.2410.12952 (引用页: 65, 66).
- [256] JIA X, LI J, WANG Z, et al. Fast, Slow, and Tool-augmented Thinking for LLMs: A Review[J/OL]. 2025(arXiv:2508.12265). arXiv: 2508.12265 [cs] [2025-12-16]. DOI: 10.48550/arXiv.2508.12265 (引用页: 65).
- [257] TREVIÑO E, CONTANT H, NGAI J, et al. Benchmarking Failures in Tool-Augmented Language Models[J/OL]. 2025(arXiv:2503.14227). arXiv: 2503.14227 [cs] [2025-12-16]. DOI: 10.48550/arXiv.2503.14227 (引用页: 65, 66).
- [258] ZENG Y, DING X, HOU Y, et al. Tool Zero: Training Tool-Augmented LLMs via Pure RL from Scratch[J/OL]. 2025(arXiv:2511.01934). arXiv: 2511.01934 [cs] [2025-12-16]. DOI: 10.48550/arXiv.2511.01934 (引用页: 65-67).
- [259] HE J, NEVILLE J, WAN M, et al. GenTool: Enhancing Tool Generalization in Language Models through Zero-to-One and Weak-to-Strong Simulation[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025: 1097-1122 [2025-12-16]. DOI: 10.18653/v1/2025.findings-acl.61 (引用页: 65, 66).
- [260] LI Y, SHEN X, YAO X, et al. Beyond Single-Turn: A Survey on Multi-Turn Interactions with Large Language Models[J/OL]. 2025(arXiv:2504.04717). arXiv: 2504.04717 [cs] [2025-12-16]. DOI: 10.48550/arXiv.2504.04717 (引用页: 65).

- [261] PINK M, WU Q, VO V A, et al. Position: Episodic Memory is the Missing Piece for Long-Term LLM Agents[J]. arXiv preprint arXiv:2502.06975, 2025 (引用页: 66).
- [262] GAO Z, ZHAN W, CHANG J D, et al. Regressing the relative future: Efficient policy optimization for multi-turn rlhf[J]. arXiv preprint arXiv:2410.04612, 2024 (引用页: 66).
- [263] ZHANG G, GENG H, YU X, et al. The Landscape of Agentic Reinforcement Learning for LLMs: A Survey[J/OL]. 2025(arXiv:2509.02547). arXiv: 2509.02547 [cs] [2025-12-16]. DOI: 10.48550/arXiv.2509.02547 (引用页: 66).
- [264] BARRES V, DONG H, RAY S, et al. τ^2 -Bench: Evaluating Conversational Agents in a Dual-Control Environment[J/OL]. 2025. arXiv: 2506.07982 [cs.AI]. <https://arxiv.org/abs/2506.07982> (引用页: 67).
- [265] PATIL S G, MAO H, YAN F, et al. The Berkeley Function Calling Leaderboard (BFCL): From Tool Use to Agentic Evaluation of Large Language Models[C]//Forty-second International Conference on Machine Learning. 2025 [2025-12-16] (引用页: 67).
- [266] CHEN C, HAO X, LIU W, et al. ACEBench: Who Wins the Match Point in Tool Usage?[J/OL]. 2025. arXiv: 2501.12851 [cs.CL]. <https://arxiv.org/abs/2501.12851> (引用页: 67).
- [267] WANG J, ZHOU J, WEN M, et al. HammerBench: Fine-Grained Function-Calling Evaluation in Real Mobile Device Scenarios [J/OL]. 2025(arXiv:2412.16516). arXiv: 2412.16516 [cs] [2025-12-16]. DOI: 10.48550/arXiv.2412.16516 (引用页: 67).
- [268] LI C, et al. The Landscape of Agentic Reinforcement Learning for LLMs: A Survey[J]. arXiv preprint arXiv:2509.02547, 2025 (引用页: 69, 70).
- [269] PUTTA P, et al. Agent Q: Advanced Reasoning and Learning for Autonomous AI Agents[J]. arXiv preprint arXiv:2408.07199, 2024 (引用页: 70, 71).

- [270] GUAN X, ZHANG L L, LIU Y, et al. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking[J]. arXiv preprint arXiv:2501.04519, 2025 (引用页: 70).
- [271] KUMAR A, MOENS V, POOLE B, et al. SCoRe: Training Language Models to Self-Correct via Reinforcement Learning[J]. 2025 (引用页: 70).
- [272] CHEN H, et al. Walking Down the Memory Maze: Beyond Context Limit through Interactive Reading[J]. 2024 (引用页: 70).
- [273] BAI H, et al. DigiRL: Training In-The-Wild Device-Control Agents with Autonomous Reinforcement Learning[J]. 2024 (引用页: 70).
- [274] LI Z Z, ZHANG D, ZHANG M L, et al. From system 1 to system 2: A survey of reasoning large language models[J/OL]. 2025. arXiv: 2502.17419 [cs.CL]. <https://arxiv.org/abs/2502.17419> (引用页: 69).
- [275] ZHANG J, et al. AFlow: Automating Agentic Workflow Generation [J]. arXiv preprint arXiv:2410.10762, 2024 (引用页: 71).
- [276] SHENG G, ZHANG C, YE Z, et al. HybridFlow: A Flexible and Efficient RLHF Framework[J]. arXiv preprint arXiv: 2409.19256, 2024 (引用页: 73).
- [277] LU H, LIU Z, et al. ROLL Flash: Accelerating RLVR and Agentic Training with Asynchrony[J]. arXiv preprint arXiv:2510.11345, 2025 (引用页: 74).
- [278] CUI G, YUAN L, WANG Z, et al. Process Reinforcement through Implicit Rewards[J]. arXiv preprint arXiv:2502.01456, 2025 (引用页: 74).
- [279] ZHU Z, XIE C, LV X, et al. slime: An LLM Post-Training Framework for RL Scaling[J]. arXiv preprint arXiv:2510.12633, 2025 (引用页: 75).
- [280] WANG Z, WANG K, WANG Q, et al. RAGEN: Understanding Self-Evolution in LLM Agents via Multi-Turn Reinforcement Learning [J]. arXiv preprint arXiv:2504.20073, 2025 (引用页: 76).

- [281] HU J, CAO X, et al. OpenRLHF: An Easy-to-use, Scalable and High-performance RLHF Framework[J]. arXiv preprint arXiv:2405.11143, 2024 (引用页: 76).
- [282] FRANTAR E, ASHKBOOS S, HOEFLE T, et al. Gptq: Accurate post-training quantization for generative pre-trained transformers [J]. arXiv preprint arXiv:2210.17323, 2022 (引用页: 80).
- [283] LIN J, TANG J, TANG H, et al. Awq: Activation-aware weight quantization for on-device llm compression and acceleration[J]. Proceedings of machine learning and systems, 2024, 6: 87-100 (引用页: 80).
- [284] HU X, CHENG Y, YANG D, et al. Ostquant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting[J]. arXiv preprint arXiv:2501.13987, 2025 (引用页: 80, 82).
- [285] LIU Z, ZHAO C, FEDOROV I, et al. SpinQuant: LLM Quantization with Learned Rotations[C/OL]//The Thirteenth International Conference on Learning Representations. 2025. <https://openreview.net/forum?id=ogO6DGE6FZ> (引用页: 80, 82).
- [286] CHEN M, SHAO W, XU P, et al. Efficientqat: Efficient quantization-aware training for large language models[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 10081-10100 (引用页: 81, 82).
- [287] LEE B, KIM D, YOU Y, et al. LittleBit: Ultra Low-Bit Quantization via Latent Factorization[J]. arXiv preprint arXiv:2506.13771, 2025 (引用页: 81, 82).
- [288] GU H, LI L, WANG Z, et al. BTC-LLM: Efficient Sub-1-Bit LLM Quantization via Learnable Transformation and Binary Codebook [J]. arXiv preprint arXiv:2506.12040, 2025 (引用页: 81, 82).
- [289] XIA J, ZHAO M, XIAO L, et al. SDQ-LLM: Sigma-Delta Quantization for 1-bit LLMs of any size[J]. arXiv preprint arXiv:2510.03275, 2025 (引用页: 81, 82).

- [290] FRANTAR E, ALISTARH D. Sparsegpt: Massive language models can be accurately pruned in one-shot[C]//International conference on machine learning. 2023: 10323-10337 (引用页: 82).
- [291] SUN M, LIU Z, BAIR A, et al. A simple and effective pruning approach for large language models[J]. arXiv preprint arXiv:2306.11695, 2023 (引用页: 82).
- [292] MEN X, XU M, ZHANG Q, et al. Shortgpt: Layers in large language models are more redundant than you expect[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 20192-20204 (引用页: 83, 84).
- [293] ZHONG L, WAN F, CHEN R, et al. Blockpruner: Fine-grained pruning for large language models[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 5065-5080 (引用页: 83, 84).
- [294] WANG Y, MA M, WANG Z, et al. CFSP: an efficient structured pruning framework for llms with coarse-to-fine activation information[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025: 9311-9328 (引用页: 83, 84).
- [295] CHEN Y, CHENG B, HAN J, et al. DLP: Dynamic Layerwise Pruning in Large Language Models[J]. arXiv preprint arXiv:2505.23807, 2025 (引用页: 83, 84).
- [296] LONG L, YANG R, HUANG Y, et al. Sliminfer: Accelerating long-context llm inference via dynamic token pruning[J]. arXiv preprint arXiv:2508.06447, 2025 (引用页: 83, 84).
- [297] FU Q, CHO M, MERTH T, et al. Lazyllm: Dynamic token pruning for efficient long context llm inference[J]. arXiv preprint arXiv:2407.14057, 2024 (引用页: 83, 84).
- [298] LEE J, HWANG S W, QIAO A, et al. Stun: Structured-then-unstructured pruning for scalable moe pruning[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 13660-13676 (引用页: 84).

- [299] SCHMITT B, GROSVENOR A, CUNNINGHAM M, et al. Contextual compression encoding for large language models: A novel framework for multi-layered parameter space pruning[J]. arXiv preprint arXiv:2502.08323, 2025 (引用页: 84).
- [300] GU Y, DONG L, WEI F, et al. Minillm: Knowledge distillation of large language models[J]. arXiv preprint arXiv:2306.08543, 2023 (引用页: 85).
- [301] BAEK D D, TEGMARK M. Towards understanding distilled reasoning models: A representational approach[J]. arXiv preprint arXiv:2503.03730, 2025 (引用页: 85, 87).
- [302] XU L, LIU K, LIU J, et al. Local Dense Logit Relations for Enhanced Knowledge Distillation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025: 4539-4549 (引用页: 86).
- [303] WU T, TAO C, WANG J, et al. Rethinking kullback-leibler divergence in knowledge distillation for large language models[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025: 5737-5755 (引用页: 86, 87).
- [304] LI M, ZHOU F, SONG X. Bild: Bi-directional logits difference loss for large language model distillation[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025: 1168-1182 (引用页: 86, 87).
- [305] ZHANG Y, CHEW Y, DONG Y, et al. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025: 20626-20636 (引用页: 86, 87, 239).
- [306] SHEN J, CUI X, GAO Z, et al. Adaptive Chain-of-Thought Distillation Based on LLM Performance on Original Problems[J]. Mathematics, 2025, 13(22): 3646 (引用页: 86, 87).
- [307] YAN Z, ZHANG Y, HE B, et al. Infifusion: A unified framework for enhanced cross-model reasoning via llm fusion[J]. arXiv preprint arXiv:2501.02795, 2025 (引用页: 86).

- [308] FENG T, ZHANG H, LEI Z, et al. FusionFactory: Fusing LLM Capabilities with Multi-LLM Log Data[J]. arXiv preprint arXiv:2507.10540, 2025 (引用页: 86).
- [309] LI Y, WEI F, ZHANG C, et al. EAGLE-3: Scaling up Inference Acceleration of Large Language Models via Training-Time Test[C/OL]//The Thirty-ninth Annual Conference on Neural Information Processing Systems. 2025. <https://openreview.net/forum?id=4exx1hUffq> (引用页: 89).
- [310] BACHMANN G, ANAGNOSTIDIS S, PUMAROLA A, et al. Judge Decoding: Faster Speculative Sampling Requires Going Beyond Model Alignment[C]//The Thirteenth International Conference on Learning Representations. 2025 (引用页: 89, 90).
- [311] LI J, XU Y, LI G, et al. Training-Free Loosely Speculative Decoding: Accepting Semantically Correct Drafts Beyond Exact Match[J]. arXiv preprint arXiv:2511.22972, 2025 (引用页: 89, 90).
- [312] TIMOR N, MAMOU J, KORAT D, et al. Accelerating LLM Inference with Lossless Speculative Decoding Algorithms for Heterogeneous Vocabularies[C/OL]//Forty-second International Conference on Machine Learning. 2025. <https://openreview.net/forum?id=vQubr1uBUw> (引用页: 89, 90).
- [313] XIAO S, FU J, XIE Z, et al. TokenTiming: A Dynamic Alignment Method for Universal Speculative Decoding Model Pairs[J]. arXiv preprint arXiv:2510.15545, 2025 (引用页: 89, 90).
- [314] LUO X, WANG Y, ZHU Q, et al. Turning Trash into Treasure: Accelerating Inference of Large Language Models with Token Recycling [C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 6816-6831. <https://aclanthology.org/2025.acl-long.338/>. DOI: 10.18653/v1/2025.acl-long.338 (引用页: 89, 91).
- [315] OLIARO G, JIA Z, CAMPOS D F, et al. SuffixDecoding: Extreme

- Speculative Decoding for Emerging AI Applications[C/OL]//The Thirty-ninth Annual Conference on Neural Information Processing Systems. 2025. <https://openreview.net/forum?id=uwL0vbeEVn> (引用页: 89, 91).
- [316] YANG H, YAO Y, LI Z, et al. XQuant: Achieving Ultra-Low Bit KV Cache Quantization with Cross-Layer Compression[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 9796-9811 (引用页: 92).
- [317] SU Z, CHEN Z, SHEN W, et al. Rotatekv: Accurate and robust 2-bit kv cache quantization for llms via outlier-aware adaptive rotations [J]. arXiv preprint arXiv:2501.16383, 2025 (引用页: 92).
- [318] SHARMA A, DING H, LI J, et al. MiniKV: Pushing the Limits of 2-Bit KV Cache via Compression and System Co-Design for Efficient Long Context Inference[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 18506-18523 (引用页: 92, 93).
- [319] TANG H, LIN Y, LIN J, et al. RazorAttention: Efficient KV Cache Compression Through Retrieval Heads[C/OL]//The Thirteenth International Conference on Learning Representations. 2025. <https://openreview.net/forum?id=tkiZQIL04w> (引用页: 92, 93).
- [320] CHANG C C, LIN W C, LIN C Y, et al. Palu: KV-Cache Compression with Low-Rank Projection[C/OL]//The Thirteenth International Conference on Learning Representations. 2025. <https://openreview.net/forum?id=LWMS4pk2vK> (引用页: 92, 93).
- [321] RAMACHANDRAN A, NESEEM M, SAKR C, et al. ThinKV: Thought-Adaptive KV Cache Compression for Efficient Reasoning Models[J]. arXiv preprint arXiv:2510.01290, 2025 (引用页: 92).
- [322] SUN H, CHANG L W, BAO W, et al. ShadowKV: KV Cache in Shadows for High-Throughput Long-Context LLM Inference [C/OL]//Forty-second International Conference on Machine Learning. 2025. <https://openreview.net/forum?id=oa7MYAO6h6> (引用页: 92, 93).

- [323] JIE S, TANG Y, HAN K, et al. SpeCache: Speculative Key-Value Caching for Efficient Generation of LLMs[C/OL]//Forty-second International Conference on Machine Learning. 2025. <https://openreview.net/forum?id=PQIrsaIQdn> (引用页: 92, 93).
- [324] CAI Z, XIAO W, SUN H, et al. R-KV: Redundancy-aware KV Cache Compression for Reasoning Models[C/OL]//The Thirty-ninth Annual Conference on Neural Information Processing Systems. 2025. <https://openreview.net/forum?id=2jwAjomEDB> (引用页: 92, 94).
- [325] ZHANG H, ZHANG H, MA X, et al. LazyEviction: Lagged KV Eviction with Attention Pattern Observation for Efficient Long Reasoning[J]. arXiv preprint arXiv:2506.15969, 2025 (引用页: 92, 94).
- [326] KWON W, LI Z, ZHUANG S, et al. Efficient memory management for large language model serving with pagedattention[C]//Proceedings of the 29th symposium on operating systems principles. 2023: 611-626 (引用页: 95).
- [327] ZHENG L, YIN L, XIE Z, et al. Sglang: Efficient execution of structured language model programs[J]. Advances in neural information processing systems, 2024, 37: 62557-62583 (引用页: 96).
- [328] CONTRIBUTORS L. Lmdeploy: A toolkit for compressing, deploying, and serving llm[Z]. 2023 (引用页: 98).
- [329] ZHANG L, JIANG Y, HE G, et al. Efficient mixed-precision large language model inference with turbomind[J]. arXiv preprint arXiv:2508.15601, 2025 (引用页: 98).
- [330] YAO S, ZHAO J, YU D, et al. React: Synergizing reasoning and acting in language models[C]//The eleventh international conference on learning representations (引用页: 106).
- [331] SHINN N, CASSANO F, GOPINATH A, et al. Reflexion: Language agents with verbal reinforcement learning[J]. Advances in Neural Information Processing Systems, 2023, 36: 8634-8652 (引用页: 106).
- [332] ERDOGAN L E, LEE N, KIM S, et al. Plan-and-act: Improving planning of agents for long-horizon tasks[J]. arXiv preprint

- arXiv:2503.09572, 2025 (引用页: 107).
- [333] YUKSEKGONUL M, BIANCHI F, BOEN J, et al. Textgrad: Automatic" differentiation" via text[J]. arXiv preprint arXiv:2406.07496, 2024 (引用页: 107, 134, 135).
 - [334] KADU A, KRISHNAN A. ReflexGrad: Three-Way Synergistic Architecture for Zero-Shot Generalization in LLM Agents[J]. arXiv preprint arXiv:2511.14584, 2025 (引用页: 107).
 - [335] GUI R, WANG Z, WANG J, et al. HyperTree Planning: Enhancing LLM Reasoning via Hierarchical Thinking[C/OL]//Forty-second International Conference on Machine Learning. 2025. <https://openreview.net/forum?id=45he3Ri6JP> (引用页: 107).
 - [336] CHOI J W, KIM H, ONG H, et al. ReAcTree: Hierarchical LLM Agent Trees with Control Flow for Long-Horizon Task Planning[J]. arXiv preprint arXiv:2511.02424, 2025 (引用页: 107).
 - [337] Anonymous. Learning When to Plan: Efficiently Allocating Test-Time Compute for LLM Agents[C/OL]//Submitted to The Fourteenth International Conference on Learning Representations. 2025. <https://openreview.net/forum?id=mBxFCTlFmW> (引用页: 107).
 - [338] CAI K, LIU J, YANG X, et al. Beyond Manuals and Tasks: Instance-Level Context Learning for LLM Agents[J]. arXiv preprint arXiv:2510.02369, 2025 (引用页: 108).
 - [339] SUN W, LU M, LING Z, et al. Scaling long-horizon llm agent via context-folding[J]. arXiv preprint arXiv:2510.11967, 2025 (引用页: 108).
 - [340] SUZGUN M, YUKSEKGONUL M, BIANCHI F, et al. Dynamic cheatsheet: Test-time learning with adaptive memory[J]. arXiv preprint arXiv:2504.07952, 2025 (引用页: 108).
 - [341] KIM J, RHEE S, KIM M, et al. ReflAct: World-Grounded Decision Making in LLM Agents via Goal-State Reflection[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Proceedings of the 2025 Conference on Empirical Methods in

Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 33421-33453. <https://aclanthology.org/2025.emnlp-main.1697/>. DOI: 10.18653/v1/2025.emnlp-main.1697 (引用页: 108).

- [342] CHEN Y, XU B, WANG X, et al. Training LLM-Based Agents with Synthetic Self-Reflected Trajectories and Partial Masking[J/OL]. ArXiv, 2025, abs/2505.20023. <https://api.semanticscholar.org/CorpusID:278912146> (引用页: 108).
- [343] HOU X, ZHAO Y, WANG S, et al. Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions[J/OL]. ArXiv, 2025, abs/2503.23278. <https://api.semanticscholar.org/CorpusID:277452486> (引用页: 110).
- [344] NARAJALA V S, HABLER I. Enterprise-Grade Security for the Model Context Protocol (MCP): Frameworks and Mitigation Strategies[J/OL]. 2025. arXiv: 2504.08623 [cs.CR]. <https://arxiv.org/abs/2504.08623> (引用页: 110).
- [345] EHTESHAM A, SINGH A, GUPTA G K, et al. A survey of agent interoperability protocols: Model Context Protocol (MCP), Agent Communication Protocol (ACP), Agent-to-Agent Protocol (A2A), and Agent Network Protocol (ANP)[J/OL]. 2025. arXiv: 2505.02279 [cs.AI]. <https://arxiv.org/abs/2505.02279> (引用页: 110).
- [346] MOURA J. CrewAI: Framework for Role-Driven Collaborative AI Agents[CP/OL]. GitHub. 2024. <https://github.com/crewAIInc/crewAI> (引用页: 111).
- [347] WU Q, BANSAL G, ZHANG J, et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation[J/OL]. 2023. arXiv: 2308.08155 [cs.AI]. <https://arxiv.org/abs/2308.08155> (引用页: 111).
- [348] CHASE H, et al. LangChain: Building Applications with LLMs through Composability[J]. arXiv preprint arXiv:2310.04710, 2023 (引用页: 111).
- [349] N8n Team. n8n: The Default Platform for Human-AI Collaboration

- [EB/OL]. 2025. <https://blog.n8n.io/series-c/> (引用页: 111).
- [350] YADAV A, CHAUHAN A, RAI A, et al. SENSEI: ASSISTIVE AI AGENT WITH MULTIMODAL CAPABILITIES AND CUSTOM TOOL INTEGRATION[J/OL]. International Research Journal of Modernization in Engineering Technology and Science, 2025. <https://blog.n8n.io/series-c/> (引用页: 111).
- [351] WANG C, LUO W, CHEN Q, et al. MLLM-Tool: A Multimodal Large Language Model For Tool Agent Learning[J]. arXiv preprint arXiv:2401.10727, 2024 (引用页: 111, 112).
- [352] XIE T, WU Y, LUO Y, et al. Training-Free Multimodal Large Language Model Orchestration[J/OL]. 2025. arXiv: 2508.10016 [cs.CL]. <https://arxiv.org/abs/2508.10016> (引用页: 111).
- [353] ZHAO Z, DONG Y, LIU A, et al. TURA: Tool-Augmented Unified Retrieval Agent for AI Search[J/OL]. 2025. arXiv: 2508.04604 [cs.CL]. <https://arxiv.org/abs/2508.04604> (引用页: 112).
- [354] HASAN M M, LI H, FALLAHZADEH E, et al. Model Context Protocol (MCP) at First Glance: Studying the Security and Maintainability of MCP Servers[J/OL]. 2025. arXiv: 2506.13538 [cs.SE]. <https://arxiv.org/abs/2506.13538> (引用页: 112).
- [355] DING P, STEVENS R. Unified Tool Integration for LLMs: A Protocol-Agnostic Approach to Function Calling[J/OL]. ArXiv, 2025, abs/2508.02979. <https://api.semanticscholar.org/CorpusID:280526513> (引用页: 112).
- [356] WANG Y, ZHANG H, PANG L, et al. MaFeRw: Query Rewriting with Multi-Aspect Feedbacks for Retrieval-Augmented Large Language Models[J/OL]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39:25434-25442. DOI: 10.1609/aaai.v39i24.34732 (引用页: 117).
- [357] AMATO F, CONTE C, FONISTO M, et al. Comparing LLM-Based Query Rewriting Strategies Within RAG Pipelines for Domain-Routed Legal Question Answering[C/OL]//. 2025. DOI: 10.1007/978-3-031-96099-4_9 (引用页: 117).

- [358] WANG T, GONG H, ZHANG C, et al. SAGE: Strategy-Adaptive Generation Engine for Query Rewriting[J/OL]. 2025. arXiv: 2506.19783 [cs.AI]. <https://arxiv.org/abs/2506.19783> (引用页: 118).
- [359] MAAREFDOUST R, HO M, BELL A, et al. Improving Math Information Retrieval via Query Rewriting with Large Language Models [C/OL]//. 2025. DOI: 10.1145/3767695.3769482 (引用页: 118).
- [360] SAWARKAR K, MANGAL A, SOLANKI S R. Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers[C/OL]//2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR). 2024: 155-161. DOI: 10.1109/MIPR 62202.2024.00031 (引用页: 118).
- [361] HU Y, LEI Z, DAI Z, et al. CG-RAG: Research Question Answering by Citation Graph Retrieval-Augmented LLMs[C/OL]//SIGIR '25: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: Association for Computing Machinery, 2025: 678-687. <https://doi.org/10.1145/3726302.3729920>. DOI: 10.1145/3726302.3729920 (引用页: 118, 119).
- [362] YAN Y, XU G, ZOU X, et al. DocPruner: A Storage-Efficient Framework for Multi-Vector Visual Document Retrieval via Adaptive Patch-Level Embedding Pruning[J/OL]. 2025. arXiv: 2509.23883 [cs.CL]. <https://arxiv.org/abs/2509.23883> (引用页: 118).
- [363] WANG S, FANG Y, ZHOU Y, et al. ArchRAG: Attributed Community-based Hierarchical Retrieval-Augmented Generation [J/OL]. 2025. arXiv: 2502.09891 [cs.IR]. <https://arxiv.org/abs/2502.09891> (引用页: 119).
- [364] LIU P, LIU X, YAO R, et al. HM-RAG: Hierarchical Multi-Agent Multimodal Retrieval Augmented Generation[J/OL]. 2025. arXiv: 2504.12330 [cs.CL]. <https://arxiv.org/abs/2504.12330> (引用页: 119, 121).
- [365] SARMAH B, MEHTA D, HALL B, et al. HybridRAG: Integrating

- Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction[C/OL]//ICAIF'24: 5th ACM International Conference on AI in Finance. 2024: 608-616. DOI: 10.1145/3677052.3698671 (引用页: 119).
- [366] CHEN B, GUO Z, YANG Z, et al. PathRAG: Pruning Graph-based Retrieval Augmented Generation with Relational Paths[J/OL]. 2025. arXiv: 2502.14902 [cs.CL]. <https://arxiv.org/abs/2502.14902> (引用页: 119).
- [367] ZHANG Z, MOSCHITTI A, VU T. Retrieving Support to Rank Answers in Open-Domain Question Answering[C/OL]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 35086-35093. <https://aclanthology.org/2025.emnlp-main.1778/>. DOI: 10.18653/v1/2025.emnlp-main.1778 (引用页: 119).
- [368] KARDAN M, PIRYANI B, JATOWT A. Evaluating Answer Reranking Strategies in Time-sensitive Question Answering[J/OL]. 2025. arXiv: 2503.04972 [cs.CL]. <https://arxiv.org/abs/2503.04972> (引用页: 119).
- [369] CHEN S Q, CHENG X, GE T, et al. xRAG: Extreme Context Compression for Retrieval-augmented Generation with One Token [C/OL]//ACL 2025. 2024: 109487-109516. DOI: 10.52202/079017-3476 (引用页: 119).
- [370] HWANG T, CHO S, JEONG S, et al. EXIT: Context-Aware Extractive Compression for Enhancing Retrieval-Augmented Generation [C/OL]//Advances in Neural Information Processing Systems 37. 2025: 4895-4924. DOI: 10.18653/v1/2025.findings-acl.253 (引用页: 119).
- [371] FENG W, HAO C, ZHANG Y, et al. AirRAG: Autonomous Strategic Planning and Reasoning Steer Retrieval Augmented Generation [J/OL]. 2025. arXiv: 2501.10053 [cs.AI]. <https://arxiv.org/abs/2501.10053> (引用页: 120).
- [372] GUAN X, ZENG J, MENG F, et al. DeepRAG: Thinking to Retrieve

- Step by Step for Large Language Models[J/OL]. 2025. arXiv: 2502.01142 [cs.AI]. <https://arxiv.org/abs/2502.01142> (引用页: 120).
- [373] YU C, ZHAO K, LI Y, et al. GraphRAG-R1: Graph Retrieval-Augmented Generation with Process-Constrained Reinforcement Learning[J/OL]. 2025. arXiv: 2507.23581 [cs.LG]. <https://arxiv.org/abs/2507.23581> (引用页: 120).
- [374] LI R, DAI Q, ZHANG Z, et al. KnowTrace: Bootstrapping Iterative Retrieval-Augmented Generation with Structured Knowledge Tracing[J]. KDD '25: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, 2025 (引用页: 120).
- [375] LEE Z, CAO S, LIU J, et al. ReaRAG: Knowledge-guided Reasoning Enhances Factuality of Large Reasoning Models with Iterative Retrieval Augmented Generation[J/OL]. 2025. arXiv: 2503.21729 [cs.CL]. <https://arxiv.org/abs/2503.21729> (引用页: 120).
- [376] WU D, GU J C, CHANG K W, et al. Self-Routing RAG: Binding Selective Retrieval with Knowledge Verbalization[J/OL]. 2025. arXiv: 2504.01018 [cs.CL]. <https://arxiv.org/abs/2504.01018> (引用页: 120).
- [377] GUMAAN E. ExpertRAG: Efficient RAG with Mixture of Experts – Optimizing Context Retrieval for Adaptive LLM Responses[J/OL]. 2025. arXiv: 2504.08744 [cs.IR]. <https://arxiv.org/abs/2504.08744> (引用页: 121).
- [378] CHEN Y, GUO D, MEI S, et al. UltraRAG: A Modular and Automated Toolkit for Adaptive Retrieval-Augmented Generation [J/OL]. 2025. arXiv: 2504.08761 [cs.IR]. <https://arxiv.org/abs/2504.08761> (引用页: 121).
- [379] SINGH A, EHTESHAM A, KUMAR S, et al. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG[J/OL]. 2025. arXiv: 2501.09136 [cs.AI]. <https://arxiv.org/abs/2501.09136> (引用页: 121).
- [380] NGUYEN T, CHIN P, TAI Y W. MA-RAG: Multi-Agent Retrieval-Augmented Generation via Collaborative Chain-of-Thought Rea-

- soning[J/OL]. 2025. arXiv: 2505.20096 [cs.CL]. <https://arxiv.org/abs/2505.20096> (引用页: 121).
- [381] YUE J, XU X, KARLSSON B F, et al. MLLM as Retriever: Interactively Learning Multimodal Retrieval for Embodied Agents[J/OL]. 2025. arXiv: 2410.03450 [cs.LG]. <https://arxiv.org/abs/2410.03450> (引用页: 122).
- [382] LIN W, CHEN J, MEI J, et al. Fine-grained Late-interaction Multimodal Retrieval for Retrieval Augmented Visual Question Answering[J/OL]. 2023. arXiv: 2309.17133 [cs.CL]. <https://arxiv.org/abs/2309.17133> (引用页: 122).
- [383] LI H, BIN Y, MA Y, et al. SemCORE: A Semantic-Enhanced Generative Cross-Modal Retrieval Framework with MLLMs[J/OL]. 2025. arXiv: 2504.13172 [cs.IR]. <https://arxiv.org/abs/2504.13172> (引用页: 122).
- [384] TIAN Y, LIU F, ZHANG J, et al. CoRe-MMRAG: Cross-Source Knowledge Reconciliation for Multimodal RAG[J/OL]. 2025. arXiv: 2506.02544 [cs.CL]. <https://arxiv.org/abs/2506.02544> (引用页: 122).
- [385] ZHU M, CHENG S, BAI G, et al. Cross-modal RAG: Sub-dimensional Text-to-Image Retrieval-Augmented Generation [J/OL]. 2025. arXiv: 2505.21956 [cs.CV]. <https://arxiv.org/abs/2505.21956> (引用页: 122).
- [386] ZHANG C, CHEN Q, ZHANG M. Mixture-of-RAG: Integrating Text and Tables with Large Language Models[C]//SIGKDD 2026. 2026 (引用页: 122).
- [387] PACKER C, FANG V, PATIL S, et al. MemGPT: Towards LLMs as Operating Systems.[J]. 2023 (引用页: 123).
- [388] ZHONG W, GUO L, GAO Q, et al. MemoryBank: Enhancing Large Language Models with Long-Term Memory[J/OL]. 2023. arXiv: 2305.10250 [cs.CL]. <https://arxiv.org/abs/2305.10250> (引用页: 123).
- [389] YANG H, LIN Z, WANG W, et al. Memory³: Language Modeling with Explicit Memory[J/OL]. Journal of Machine Learning, 2024,

- 3(3):300-346. <http://dx.doi.org/10.4208/jml.240708>. DOI: 10.4208/jml.240708 (引用页: 123).
- [390] RASMUSSEN P, PALIYCHUK P, BEAUVAIS T, et al. Zep: A Temporal Knowledge Graph Architecture for Agent Memory[J/OL]. 2025. arXiv: 2501.13956 [cs.CL]. <https://arxiv.org/abs/2501.13956> (引用页: 125).
- [391] Memobase Developers. Memobase: An Open-Source, Persistent Memory System for LLM Agents[EB/OL]. 2025 [2025-01]. <https://github.com/memodb-io/memobase> (引用页: 125).
- [392] CHHIKARA P, KHANT D, ARYAN S, et al. Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory [J/OL]. 2025. arXiv: 2504.19413 [cs.CL]. <https://arxiv.org/abs/2504.19413> (引用页: 125).
- [393] LI Z, XI C, LI C, et al. MemOS: A Memory OS for AI System [J/OL]. 2025. arXiv: 2507.03724 [cs.CL]. <https://arxiv.org/abs/2507.03724> (引用页: 125).
- [394] XU W, LIANG Z, MEI K, et al. A-MEM: Agentic Memory for LLM Agents[J/OL]. 2025. arXiv: 2502.12110 [cs.CL]. <https://arxiv.org/abs/2502.12110> (引用页: 125).
- [395] YAN S, YANG X, HUANG Z, et al. Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning[J/OL]. 2025. arXiv: 2508.19828 [cs.CL]. <https://arxiv.org/abs/2508.19828> (引用页: 125).
- [396] FANG J, DENG X, XU H, et al. LightMem: Lightweight and Efficient Memory-Augmented Generation[J/OL]. 2025. arXiv: 2510.18866 [cs.CL]. <https://arxiv.org/abs/2510.18866> (引用页: 125).
- [397] WANG Y, CHEN X. MIRIX: Multi-Agent Memory System for LLM-Based Agents[J/OL]. 2025. arXiv: 2507.07957 [cs.CL]. <https://arxiv.org/abs/2507.07957> (引用页: 125).
- [398] LONG L, HE Y, YE W, et al. Seeing, Listening, Remembering, and Reasoning: A Multimodal Agent with Long-Term Memory[J/OL].

2025. arXiv: 2508.09736 [cs.CV]. <https://arxiv.org/abs/2508.09736> (引用页: 125).
- [399] EverMind-AI. EverMemOS: An Operating System-Inspired Memory Management Framework for Large Language Models[EB/OL]. 2025 [2025-10]. <https://github.com/EverMind-AI/EverMemOS/> (引用页: 126).
- [400] YUAN S, CHEN Z, XI Z, et al. Agent-R: Training Language Model Agents to Reflect via Iterative Self-Training[J/OL]. 2025. arXiv: 2501.11425 [cs.AI]. <https://arxiv.org/abs/2501.11425> (引用页: 128).
- [401] FU D, HE K, WANG Y, et al. AgentRefine: Enhancing Agent Generalization through Refinement Tuning[C/OL]//The Thirteenth International Conference on Learning Representations. 2025. <https://openreview.net/forum?id=FDimWzmcWn> (引用页: 129).
- [402] CHEN Y, XU B, WANG X, et al. Training LLM-Based Agents with Synthetic Self-Reflected Trajectories and Partial Masking[J/OL]. 2025. arXiv: 2505.20023 [cs.CL]. <https://arxiv.org/abs/2505.20023> (引用页: 129).
- [403] ZWEIGER A, PARI J, GUO H, et al. Self-Adapting Language Models[J/OL]. 2025. arXiv: 2506.10943 [cs.LG]. <https://arxiv.org/abs/2506.10943> (引用页: 129).
- [404] LIN Y, ZHAO L, SHI Y. (P)rrior(D)yna(F)low: A Priori Dynamic Workflow Construction via Multi-Agent Collaboration[J/OL]. 2025. arXiv: 2509.14547 [cs.AI]. <https://arxiv.org/abs/2509.14547> (引用页: 129).
- [405] FENG Z, XUE R, YUAN L, et al. Multi-agent Embodied AI: Advances and Future Directions[J/OL]. 2025. arXiv: 2505.05108 [cs.AI] (引用页: 129).
- [406] ORIKE S, ENE D. Meta-Learning: Unleashing the Power of Self-Improving Artificial Intelligent (AI) Systems[J/OL]. Journal of Advances in Computational Intelligence Theory, 2023:12-27. DOI: 10.5281/zenodo.8223579 (引用页: 130).

- [407] WANG Q, LV Y, MAO Y, et al. Robust Fast Adaptation from Adversarially Explicit Task Distribution Generation[C/OL]//the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2025: 1481-1491. DOI: 10.1145/3690624.3709337 (引用页: 130).
- [408] CHO M, PARK J, KIM J, et al. ARS: Adaptive Reward Scaling for Multi-Task Reinforcement Learning[C/OL]//Forty-second International Conference on Machine Learning. 2025. <https://openreview.net/forum?id=U0mI6M7lvI> (引用页: 130).
- [409] YUN T, OH J, MIN H, et al. ReFeed: Multi-dimensional Summarization Refinement with Reflective Reasoning on Feedback[J/OL]. 2025. arXiv: 2503.21332 [cs.CL]. <https://arxiv.org/abs/2503.21332> (引用页: 130).
- [410] SHINN N, CASSANO F, GOPINATH A, et al. Reflexion: language agents with verbal reinforcement learning[C/OL]//OH A, NAUMANN T, GLOBERSON A, et al. Advances in Neural Information Processing Systems: vol. 36. Curran Associates, Inc., 2023: 8634-8652. https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf (引用页: 131).
- [411] QU M, HU Y, HAN K, et al. ReCoT: Reflective Self-Correction Training for Mitigating Confirmation Bias in Large Vision-Language Models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2025: 9147-9157 (引用页: 131).
- [412] LYU Y, ZHENG X, JIANG L, et al. RealRAG: Retrieval-augmented Realistic Image Generation via Self-reflective Contrastive Learning [C/OL]//Proceedings of Machine Learning Research: Proceedings of the 42nd International Conference on Machine Learning: vol. 267. PMLR, 2025: 41772-41790. <https://proceedings.mlr.press/v267/lyu25c.html> (引用页: 131, 132).
- [413] DONG X, ZHAO H, GAO J, et al. SE-VLN: A Self-Evolving Vision-Language Navigation Framework Based on Multimodal Large Language Models[J/OL]. 2025. arXiv: 2507.13152 [cs.CV]. <https://arxiv.org/abs/2507.13152>

- xiv.org/abs/2507.13152 (引用页: 131).
- [414] HUANG Y, CHEN H, RUAN S, et al. Mitigating Overthinking in Large Reasoning Models via Manifold Steering[J/OL]. 2025. arXiv: 2505.22411 [cs.LG]. <https://arxiv.org/abs/2505.22411> (引用页: 131).
- [415] ASAI A, WU Z, WANG Y, et al. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection[C/OL]// The Twelfth International Conference on Learning Representations. 2024. <https://openreview.net/forum?id=hSyW5go0v8> (引用页: 132).
- [416] WEI Z, CHEN W L, MENG Y. InstructRAG: Instructing Retrieval-Augmented Generation via Self-Synthesized Rationales[C/OL]// The Thirteenth International Conference on Learning Representations. 2025. <https://openreview.net/forum?id=P1qhkp8gQT> (引用页: 132).
- [417] ZELIKMAN E, WU Y, MU J, et al. Star: Bootstrapping reasoning with reasoning, 2022[J]. URL <https://arxiv.org/abs/2203.14465>, 2022, 2203 (引用页: 133, 134).
- [418] XI Z, DING Y, CHEN W, et al. AgentGym: Evolving Large Language Model-based Agents across Diverse Environments[J]. CoRR, 2024 (引用页: 134).
- [419] ZHAO A, WU Y, YUE Y, et al. Absolute zero: Reinforced self-play reasoning with zero data[J]. arXiv preprint arXiv:2505.03335, 2025 (引用页: 134).
- [420] ZHAO Y, ZHU H, JIANG T, et al. Co-EPG: A Framework for Co-Evolution of Planning and Grounding in Autonomous GUI Agents [J]. arXiv preprint arXiv:2511.10705, 2025 (引用页: 134).
- [421] ZHAI Y, TAO S, CHEN C, et al. AgentEvolver: Towards Efficient Self-Evolving Agent System[J]. arXiv preprint arXiv:2511.10395, 2025 (引用页: 134).
- [422] YANG C, WANG X, LU Y, et al. Large language models as optimizers[C]//The Twelfth International Conference on Learning Rep-

- resentations. 2023 (引用页: 134, 135).
- [423] FERNANDO C, BANARSE D, MICHALEWSKI H, et al. Prompt-breeder: self-referential self-improvement via prompt evolution[C]// Proceedings of the 41st International Conference on Machine Learning. 2024: 13481-13544 (引用页: 134, 135).
 - [424] WANG X, LI C, WANG Z, et al. Promptagent: Strategic planning with language models enables expert-level prompt optimization[J]. arXiv preprint arXiv:2310.16427, 2023 (引用页: 134, 135).
 - [425] XIANG J, ZHANG J, YU Z, et al. Self-supervised prompt optimization[J]. arXiv preprint arXiv:2502.06855, 2025 (引用页: 134, 135).
 - [426] WANG Z, WANG K, WANG Q, et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning[J]. arXiv preprint arXiv:2504.20073, 2025 (引用页: 134).
 - [427] SALAMA R, CAI J, YUAN M, et al. Meminsight: Autonomous memory augmentation for llm agents[J]. arXiv preprint arXiv:2503.21760, 2025 (引用页: 135, 136).
 - [428] CHHIKARA P, KHANT D, ARYAN S, et al. Mem0: Building production-ready ai agents with scalable long-term memory[J]. arXiv preprint arXiv:2504.19413, 2025 (引用页: 135, 136).
 - [429] XU W, LIANG Z, MEI K, et al. A-mem: Agentic memory for llm agents[J]. arXiv preprint arXiv:2502.12110, 2025 (引用页: 135, 136).
 - [430] CAI Z, GUO X, PEI Y, et al. Flex: Continuous agent evolution via forward learning from experience[J]. arXiv preprint arXiv:2511.06449, 2025 (引用页: 135, 136).
 - [431] WEI T, SACHDEVA N, COLEMAN B, et al. Evo-Memory: Benchmarking LLM Agent Test-time Learning with Self-Evolving Memory [J]. arXiv preprint arXiv:2511.20857, 2025 (引用页: 135, 136).
 - [432] QIU J, QI X, ZHANG T, et al. Alita: Generalist agent enabling scalable agentic reasoning with minimal predefinition and maximal self-evolution[J]. arXiv preprint arXiv:2505.20286, 2025 (引用页: 135, 136).

- [433] QIU J, QI X, WANG H, et al. Alita-g: Self-evolving generative agent for agent generation[J]. arXiv preprint arXiv:2510.23601, 2025 (引用页: 135, 136).
- [434] WANG R, HAN X, JI L, et al. ToolGen: Unified Tool Retrieval and Calling via Generation[C]//The Thirteenth International Conference on Learning Representations (引用页: 135, 136).
- [435] HU S, LU C, CLUNE J. Automated Design of Agentic Systems[C]//The Thirteenth International Conference on Learning Representations. 2025 (引用页: 135, 136).
- [436] SHANG Y, LI Y, ZHAO K, et al. AgentSquare: Automatic LLM Agent Search in Modular Design Space[C]//The Thirteenth International Conference on Learning Representations (引用页: 135, 137).
- [437] ZHUGE M, WANG W, KIRSCH L, et al. Gptswarm: Language agents as optimizable graphs[C]//Forty-first International Conference on Machine Learning. 2024 (引用页: 137).
- [438] ZHANG J, XIANG J, YU Z, et al. AFlow: Automating Agentic Workflow Generation[C]//The Thirteenth International Conference on Learning Representations. 2025 (引用页: 137, 147, 148).
- [439] ZHANG G, NIU L, FANG J, et al. Multi-agent Architecture Search via Agentic Supernet[C]//Forty-second International Conference on Machine Learning. 2025 (引用页: 137, 147, 148).
- [440] GAO H, LIU Y, HE Y, et al. Flowreasoner: Reinforcing query-level meta-agents[J]. arXiv preprint arXiv:2504.15257, 2025 (引用页: 137).
- [441] YUAN S, SONG K, CHEN J, et al. Evoagent: Towards automatic multi-agent generation via evolutionary algorithms[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025:6192-6217 (引用页: 137, 138).
- [442] YANG Y, CHAI H, SHAO S, et al. Agentnet: Decentralized evolutionary coordination for llm-based multi-agent systems[J]. arXiv

preprint arXiv:2504.00587, 2025 (引用页: 137, 138).

- [443] ZHANG C, YANG K, HU S, et al. Proagent: building proactive cooperative agents with large language models[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 38: 16. 2024: 17591-17599 (引用页: 137, 138).
- [444] FU L, DING X, ZHU Y, et al. CATArena: Evaluation of LLM Agents through Iterative Tournament Competitions[J]. arXiv preprint arXiv:2510.26852, 2025 (引用页: 138).
- [445] GOU B, WANG R, ZHENG B, et al. Navigating the Digital World as Humans Do: Universal Visual Grounding for GUI Agents[J/OL]. 2025. arXiv: 2410.05243 [cs.AI]. <https://arxiv.org/abs/2410.05243> (引用页: 139).
- [446] CHENG K, SUN Q, CHU Y, et al. SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents[C/OL]//KU L W, MARTINS A, SRIKUMAR V. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics, 2024: 9313-9332. <https://aclanthology.org/2024.acl-long.505/>. DOI: 10.18653/v1/2024.acl-long.505 (引用页: 140).
- [447] WANG Y, ZHANG H, TIAN J, et al. Ponder & Press: Advancing Visual GUI Agent towards General Computer Control[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025: 1461-1473. <https://aclanthology.org/2025.findings-acl.76/>. DOI: 10.18653/v1/2025.findings-acl.76 (引用页: 140).
- [448] TANG F, GU Z, LU Z, et al. GUI-G²: Gaussian Reward Modeling for GUI Grounding[J/OL]. 2025. arXiv: 2507.15846 [cs.LG]. <https://arxiv.org/abs/2507.15846> (引用页: 140).
- [449] LI N, QU X, ZHOU J, et al. MobileUse: A GUI Agent with Hierarchical Reflection for Autonomous Mobile Operation[J/OL]. 2025. arXiv: 2507.16853 [cs.R0]. <https://arxiv.org/abs/2507.16853> (引

用页: 141).

- [450] YANG Y, LI D, DAI Y, et al. GTA1: GUI Test-time Scaling Agent [J/OL]. 2025. arXiv: 2507.05791 [cs.AI]. <https://arxiv.org/abs/2507.05791> (引用页: 141).
- [451] AGASHE S, WONG K, TU V, et al. Agent S2: A Compositional Generalist-Specialist Framework for Computer Use Agents[J/OL]. 2025. arXiv: 2504.00906 [cs.AI]. <https://arxiv.org/abs/2504.00906> (引用页: 141).
- [452] ZHANG D, ZHANG S, YANG Z, et al. ProgRM: Build Better GUI Agents with Progress Rewards[J/OL]. 2025. arXiv: 2505.18121 [cs.AI]. <https://arxiv.org/abs/2505.18121> (引用页: 141).
- [453] SONG L, DAI Y, PRABHU V, et al. CoAct-1: Computer-using Agents with Coding as Actions[J/OL]. 2025. arXiv: 2508.03923 [cs.CL]. <https://arxiv.org/abs/2508.03923> (引用页: 141).
- [454] LIN K Q, LI L, GAO D, et al. Showui: One vision-language-action model for gui visual agent[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025:19498-19508 (引用页: 142).
- [455] ZHANG Z, LIU X, ZHANG X, et al. UI-Evol: Automatic Knowledge Evolving for Computer Use Agents[J/OL]. 2025. arXiv: 2505.21964 [cs.HC]. <https://arxiv.org/abs/2505.21964> (引用页: 142).
- [456] LUO R, WANG L, HE W, et al. GUI-R1 : A Generalist R1-Style Vision-Language Action Model For GUI Agents[J/OL]. 2025. arXiv: 2504.10458 [cs.CV]. <https://arxiv.org/abs/2504.10458> (引用页: 142).
- [457] XIE T, DENG J, LI X, et al. Scaling Computer-Use Grounding via User Interface Decomposition and Synthesis[J/OL]. 2025. arXiv: 2505.13227 [cs.AI]. <https://arxiv.org/abs/2505.13227> (引用页: 142).
- [458] QIN Y, YE Y, FANG J, et al. UI-TARS: Pioneering Automated GUI Interaction with Native Agents[J]. arXiv preprint arXiv:2501.12326, 2025 (引用页: 142).

- [459] WANG H, ZOU H, SONG H, et al. UI-TARS-2 Technical Report: Advancing GUI Agent with Multi-Turn Reinforcement Learning [J/OL]. 2025. arXiv: 2509.02544 [cs.AI]. <https://arxiv.org/abs/2509.02544> (引用页: 142).
- [460] LAI H, LIU X, ZHAO Y, et al. ComputerRL: Scaling End-to-End Online Reinforcement Learning for Computer Use Agents[J/OL]. 2025. arXiv: 2508.14040 [cs.AI]. <https://arxiv.org/abs/2508.14040> (引用页: 143).
- [461] WANG X, WANG B, LU D, et al. OpenCUA: Open Foundations for Computer-Use Agents[J/OL]. 2025. arXiv: 2508.09123 [cs.AI]. <https://arxiv.org/abs/2508.09123> (引用页: 143).
- [462] YE J, ZHANG X, XU H, et al. Mobile-Agent-v3: Fundamental Agents for GUI Automation[J/OL]. 2025. arXiv: 2508.15144 [cs.AI]. <https://arxiv.org/abs/2508.15144> (引用页: 143).
- [463] MU J, ZHANG C, NI C, et al. GUI-360°: A Comprehensive Dataset and Benchmark for Computer-Using Agents[J/OL]. 2025. arXiv: 2511.04307 [cs.AI]. <https://arxiv.org/abs/2511.04307> (引用页: 143).
- [464] ZHAO H H, YANG K, YU W, et al. WorldGUI: An Interactive Benchmark for Desktop GUI Automation from Any Starting Point [J/OL]. 2025. arXiv: 2502.08047 [cs.AI]. <https://arxiv.org/abs/2502.08047> (引用页: 143).
- [465] WANG X, WU Z, XIE J, et al. MMBench-GUI: Hierarchical Multi-Platform Evaluation Framework for GUI Agents[J/OL]. 2025. arXiv: 2507.19478 [cs.CV]. <https://arxiv.org/abs/2507.19478> (引用页: 143).
- [466] ZHANG G, YUE Y, SUN X, et al. G-Designer: Architecting Multi-agent Communication Topologies via Graph Neural Networks[C]// ICLR 2025 Workshop on Foundation Models in the Wild. 2025 (引用页: 146).
- [467] JIANG E H, WAN G, YIN S, et al. Dynamic Generation of Multi-LLM Agents Communication Topologies with Graph Diffusion Models[J]. arXiv preprint arXiv:2510.07799, 2025 (引用页: 146).

- [468] QIAN C, XIE Z, WANG Y, et al. Scaling Large Language Model-based Multi-Agent Collaboration[C]//The Thirteenth International Conference on Learning Representations. 2025 (引用页: 146).
- [469] ANTONIADES A, ÖRWALL A, ZHANG K, et al. SWE-Search: Enhancing Software Agents with Monte Carlo Tree Search and Iterative Refinement[C]//The Thirteenth International Conference on Learning Representations. 2025 (引用页: 147, 148).
- [470] LI S, LIU Y, WEN Q, et al. Assemble your crew: Automatic multi-agent communication topology design via autoregressive graph generation[J]. arXiv preprint arXiv:2507.18224, 2025 (引用页: 147, 148).
- [471] YE R, TANG S, GE R, et al. MAS-GPT: Training LLMs to Build LLM-based Multi-Agent Systems[C]//Forty-second International Conference on Machine Learning. 2025 (引用页: 147, 148).
- [472] WANG K, ZHANG G, YE M, et al. MAS²: Self-Generative, Self-Configuring, Self-Rectifying Multi-Agent Systems[J]. arXiv preprint arXiv:2509.24323, 2025 (引用页: 147, 148).
- [473] WANG Q, WANG T, TANG Z, et al. MegaAgent: A large-scale autonomous LLM-based multi-agent system without predefined SOPs [C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 4998-5036 (引用页: 147, 149).
- [474] HOU Z, TANG J, WANG Y. HALO: Hierarchical Autonomous Logic-Oriented Orchestration for Multi-Agent LLM Systems[J]. arXiv preprint arXiv:2505.13516, 2025 (引用页: 147, 149).
- [475] DU Z, QIAN C, LIU W, et al. Multi-agent collaboration via cross-team orchestration[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 10386-10406 (引用页: 147, 149).
- [476] PARK C, HAN S, GUO X, et al. Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 30215-30248 (引用页: 149).

- [477] DANG Y, QIAN C, LUO X, et al. Multi-Agent Collaboration via Evolving Orchestration[J]. arXiv preprint arXiv:2505.19591, 2025 (引用页: 149).
- [478] CHEN W, YOU Z, LI R, et al. Internet of Agents: Weaving a Web of Heterogeneous Agents for Collaborative Intelligence[C]//The Thirteenth International Conference on Learning Representations. 2025 (引用页: 149).
- [479] ZOU J, YANG X, QIU R, et al. Latent Collaboration in Multi-Agent Systems[J]. arXiv preprint arXiv:2511.20639, 2025 (引用页: 149, 237).
- [480] ZHU K, DU H, HONG Z, et al. Multiagentbench: Evaluating the collaboration and competition of llm agents[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 8580-8622 (引用页: 150, 237).
- [481] WANG H, ZHAO S, WANG J, et al. Beyond Frameworks: Unpacking Collaboration Strategies in Multi-Agent Systems[J]. arXiv preprint arXiv:2505.12467, 2025 (引用页: 150).
- [482] QIAN C, LIU Z, PRABHAKAR A, et al. Userbench: An interactive gym environment for user-centric agents[J]. arXiv preprint arXiv:2507.22034, 2025 (引用页: 150).
- [483] SUN H, ZHANG S, NIU L, et al. Collab-Overcooked: Benchmarking and evaluating large language models as collaborative agents[J]. arXiv preprint arXiv:2502.20073, 2025 (引用页: 150).
- [484] JIANG W B, ZHAO L M, LU B L. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI [C/OL]//The Twelfth International Conference on Learning Representations (ICLR). 2024. <https://openreview.net/forum?id=7s4J17xC8W> (引用页: 152).
- [485] YUE T, GAO X, XUE S, et al. BrainGPT: Unleashing the Potential of EEG Generalist Foundation Model by Autoregressive Pre-training[J/OL]. arXiv preprint arXiv:2410.19779, 2024. <https://arxiv.org/abs/2410.19779> (引用页: 152, 153).

- [486] JIANG W B, WANG Y, LU B L, et al. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals[J]. arXiv preprint arXiv:2409.00101, 2024 (引用页: 152, 153).
- [487] WANG G, LIU W, HE Y, et al. EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals[C/OL]//Advances in Neural Information Processing Systems (NeurIPS): vol. 37. 2024. https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html (引用页: 153).
- [488] MA J, WU F, LIN Q, et al. CodeBrain: Towards Decoupled Interpretability and Multi-Scale Architecture for EEG Foundation Model [J/OL]. arXiv preprint arXiv:2506.09110, 2025. <https://arxiv.org/abs/2506.09110> (引用页: 153).
- [489] ZENG Z, CAI Z, CAI Y, et al. Wavemind: Towards a conversational eeg foundation model aligned to textual and visual modalities[J]. arXiv preprint arXiv:2510.00032, 2025 (引用页: 153).
- [490] HMAMOUCHE Y, CHIHAB I, et Al. BrainDEC: A Multimodal LLM for the Non-Invasive Decoding of Text from Brain Recordings [J]. Information Fusion, 2026 (引用页: 153).
- [491] YE Z, AI Q, LIU Y, et al. Generative language reconstruction from brain recordings[J]. Communications Biology, 2025, 8:346 (引用页: 153).
- [492] CHEN X, DU C, et Al. BP-GPT: Auditory Neural Decoding Using fMRI-prompted LLM[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2025 (引用页: 153).
- [493] CHEN J, QI Y, WANG Y, et al. Mindgpt: Interpreting what you see with non-invasive brain recordings[J]. IEEE Transactions on Image Processing, 2025 (引用页: 154).
- [494] JING H, JIANG D, MA Y, et al. Beyond Brain Decoding: Visual-Semantic Reconstructions to Mental Creation Extension Based on

- fMRI[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025:19258-19268 (引用页: 154).
- [495] LU W, NIE D, XUE P, et al. Brain-Inspired fMRI-to-Text Decoding via Incremental and Wrap-Up Language Modeling[C]//The Thirty-ninth Annual Conference on Neural Information Processing Systems (引用页: 154).
- [496] REN Y, JIN R, et Al. Do Large Language Models Mirror Cognitive Language Processing?[C]//Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025). 2025 (引用页: 154).
- [497] GAO C, MA Z, CHEN J, et al. Increasing alignment of large language models with language processing in the human brain[J/OL]. Nature Computational Science, 2025, 5(11):1080-1090. DOI: 10.1038/s43588-025-00863-0 (引用页: 154).
- [498] YU S, GU C, HUANG K, et al. Predicting the next sentence (not word) in large language models: What model-brain alignment tells us about discourse comprehension[J/OL]. Science Advances, 2024, 10(21):eadn7744. DOI: 10.1126/sciadv.adn7744 (引用页: 155).
- [499] DOERIG A, KIETZMANN T C, ALLEN E, et al. High-level visual representations in the human brain are aligned with large language models[J]. Nature Machine Intelligence, 2025 (引用页: 155).
- [500] XU Q, PENG Y, NASTASE S A, et al. Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts[J]. Nature Human Behaviour, 2025, 9: 1871-1886 (引用页: 155).
- [501] XU B, POO M M. Large language models and brain-inspired general intelligence[J]. National Science Review, 2023, 10(10):nwad267 (引用页: 155, 156).
- [502] ZHENGZHENG T, ZHU E. BrainGPT: A Brain-Inspired SNN-Based Large Language Model[J]. (引用页: 155, 156).
- [503] WEBB T, MONDAL S S, MOMENNEJAD I. A brain-inspired agentic architecture to improve planning with LLMs[J]. Nature

- Communications, 2025, 16(1): 8633 (引用页: 155, 156).
- [504] KOSMYNA N, HAUPTMANN E, YUAN Y T, et al. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task[J]. arXiv preprint arXiv:2506.08872, 2025 (引用页: 156).
 - [505] JIANG T, WU J, et Al. The cognitive impacts of large language model interactions on problem solving and decision making using EEG analysis[J/OL]. Frontiers in Computational Neuroscience, 2025. DOI: 10.3389/fncom.2025.1556483 (引用页: 156).
 - [506] CHANDRASEKHARAN S, JACOB J E. Bridging neuroscience and AI: a survey on large language models for neurological signal interpretation[J/OL]. Frontiers in Neuroinformatics, 2025. DOI: 10.3389/fninf.2025.1561401 (引用页: 156).
 - [507] ZHANG Y, LI Q, et Al. Integrating Large Language Model, EEG, and Eye-Tracking for Word-Level Neural State Classification in Reading Comprehension[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2024 (引用页: 156).
 - [508] SORINO P, BIANCOFIORE G M, LOFU D, et al. ARIEL: Brain-Computer Interfaces meet Large Language Models for Emotional Support Conversation[C]//ACM WeBIUM24 - 1st Workshop on Wearable Devices and Brain-Computer Interfaces for User Modelling (UMAP Adjunct 2024). 2024 (引用页: 156).
 - [509] CHEN H, ZENG W, et Al. EEG Emotion Copilot: Optimizing Lightweight LLMs for Emotional EEG Interpretation with Assisted Medical Record Generation[J]. arXiv preprint arXiv:2410.00166, 2025 (引用页: 156).
 - [510] DING C, WU C, et Al. A Survey of LLMs on Biosignal Applications [J]. arXiv preprint, 2024 (引用页: 156).
 - [511] LIANG W, ZHANG Y, CODREANU M, et al. The Widespread Adoption of Large Language Model-Assisted Writing Across Society [J/OL]. 2025. arXiv: 2502.09747 [cs.CL]. <https://arxiv.org/abs/2502.09747> (引用页: 160).

- [512] XIONG R, CHEN Y, KHIZBULLIN D, et al. Beyond Outlining: Heterogeneous Recursive Planning for Adaptive Long-form Writing with Language Models[J/OL]. 2025. arXiv: 2503.08275 [cs.AI]. <https://arxiv.org/abs/2503.08275> (引用页: 160).
- [513] WAN K, MU H, HAO R, et al. A Cognitive Writing Perspective for Constrained Long-Form Text Generation[J/OL]. 2025. arXiv: 2502.12568 [cs.CL]. <https://arxiv.org/abs/2502.12568> (引用页: 160).
- [514] ZHANG Y, SUN R, CHEN Y, et al. Chain of Agents: Large Language Models Collaborating on Long-Context Tasks[J/OL]. 2024. arXiv: 2406.02818 [cs.CL]. <https://arxiv.org/abs/2406.02818> (引用页: 161).
- [515] HU Z, CHAN H P, LI J, et al. Debate-to-Write: A Persona-Driven Multi-Agent Framework for Diverse Argument Generation[J/OL]. 2025. arXiv: 2406.19643 [cs.CL]. <https://arxiv.org/abs/2406.19643> (引用页: 161).
- [516] ZOU J, YANG X, QIU R, et al. Latent Collaboration in Multi-Agent Systems[J/OL]. 2025. arXiv: 2511.20639 [cs.CL]. <https://arxiv.org/abs/2511.20639> (引用页: 161).
- [517] KIM Y, CHIN B, SON K, et al. IntentFlow: Interactive Support for Communicating Intent with LLMs in Writing Tasks[J/OL]. 2025. arXiv: 2507.22134 [cs.HC]. <https://arxiv.org/abs/2507.22134> (引用页: 161).
- [518] GMEINER F, MARQUARDT N, BENTLEY M, et al. Intent Tagging: Exploring Micro-Prompting Interactions for Supporting Granular Human-GenAI Co-Creation Workflows[C/OL]//CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. ACM, 2025: 1-31. <http://dx.doi.org/10.1145/3706598.3713861>. DOI: 10.1145/3706598.3713861 (引用页: 161).
- [519] YING S, LI Y, QU X, et al. Beyond Correctness: Evaluating Subjective Writing Preferences Across Cultures[J/OL]. 2025. arXiv: 2510.14616 [cs.CL]. <https://arxiv.org/abs/2510.14616> (引用页: 161).
- [520] JANGRA A, SARRAFZADEH B, CUCERZAN S, et al. Evaluat-

- ing Style-Personalized Text Generation: Challenges and Directions [J/OL]. 2025. arXiv: 2508.06374 [cs.CL]. <https://arxiv.org/abs/2508.06374> (引用页: 161).
- [521] Paperpal. Paperpal: AI academic writing assistant and citation tools [Z]. <https://paperpal.com>. accessed: 2025-12-18. 2025 (引用页: 162).
- [522] Jenni. Jenni AI: AI academic writer and citation helper[Z]. <https://jenni.ai>. accessed: 2025-12-18. 2025 (引用页: 162).
- [523] 中国知网 (CNKI). CNKI AI: 学术大模型与智能科研服务平台[Z]. <https://ai.cnki.net/>. accessed: 2025-12-18. 2025 (引用页: 162).
- [524] SciFocus. SciFocus: AI research assistant for academic writing and citations[Z]. <https://scifocus.ai>. accessed: 2025-12-18. 2025 (引用页: 162).
- [525] Google. NotebookLM: AI-Powered Research and Writing Assistant [Z]. <https://notebooklm.google.com/>. accessed: 2025-12-18. 2025 (引用页: 162).
- [526] 秘塔科技. 秘塔写作猫: 智能写作与文本校对平台[Z]. <https://xiezuocat.com/>. accessed: 2025-12-18. 2025 (引用页: 162).
- [527] 科大讯飞. 星火科研助手: 基于讯飞星火大模型的科研辅助平台[Z]. <https://xuexi.xfyun.cn/>. accessed: 2025-12-18. 2025 (引用页: 163).
- [528] Sudowrite. Sudowrite: AI creative-writing partner (fiction)[Z]. <https://www.sudowrite.com>. accessed: 2025-12-18. 2025 (引用页: 164).
- [529] NovelAI. NovelAI: AI story & image generator[Z]. <https://novelai.net>. accessed: 2025-12-18. 2025 (引用页: 164).
- [530] CreativeFlow. CreativeFlow: AI animation & marketing creative tool[Z]. <https://www.creativeflow.ai>. accessed: 2025-12-18. 2025 (引用页: 164).
- [531] 彩云科技. 彩云小梦: 基于深度学习的长文本生成系统[Z]. <https://if.caiyunai.com/dream/>. accessed: 2025-12-18. 2025 (引用页: 164).
- [532] XIAO D, MENG Q, LI S, et al. Improving Transformers with Dynamically Composable Multi-Head Attention[J/OL]. 2024. arXiv:

- 2405.08553 [cs.LG]. <https://arxiv.org/abs/2405.08553> (引用页: 164).
- [533] 蛙蛙写作. 蛙蛙写作: 面向创意文本生成的 AI 写作平台[Z]. <https://www.wawawriter.com/>. accessed: 2025-12-18. 2025 (引用页: 164).
- [534] Jasper AI. Jasper: AI content automation platform for marketing[Z]. <https://www.jasper.ai>. accessed: 2025-12-18. 2025 (引用页: 165).
- [535] Copy.ai. Copy.ai: AI writing & GTM automation platform[Z]. <https://www.copy.ai>. accessed: 2025-12-18. 2025 (引用页: 165).
- [536] 百度营销. 擎舵 (Qingduo): AIGC 创意生产平台[Z]. <https://qingduo.baidu.com/>. accessed: 2025-12-18; AIGC 多模态营销创意平台. 2025 (引用页: 166).
- [537] Vectara. Hallucination Leaderboard[Z]. <https://github.com/vectara/hallucination-leaderboard>. GitHub repository, accessed: 2025-12-18. 2024 (引用页: 166).
- [538] Visual Capitalist. Tasa de Alucinación de Modelos de IA (Ranked AI Hallucination Rates by Model)[Z]. <https://www.visualcapitalist.com/sp/ter02-ranked-ai-hallucination-rates-by-model/>. accessed: 2025-12-18. 2025 (引用页: 166).
- [539] IPTLS at USC. AI Copyright and the Law: The Ongoing Battle over Intellectual Property Rights[Z]. <https://sites.usc.edu/iptls/2025/02/04/ai-copyright-and-the-law-the-ongoing-battle-over-intellectual-property-rights/>. accessed: 2025-12-18. 2025 (引用页: 167).
- [540] RETTBERG J W, WIGERS H. AI-generated stories favour stability over change: homogeneity and cultural stereotyping in narratives generated by gpt-4o-mini[J]. Open Res. Eur., 2025, 5:202 (引用页: 167).
- [541] DEEP P D, CHEN Y. The Role of AI in Academic Writing: Impacts on Writing Skills, Critical Thinking, and Integrity in Higher Education[J/OL]. Societies, 2025, 15(9). <https://www.mdpi.com/2075-4698/15/9/247>. DOI: 10.3390/soc15090247 (引用页: 167).
- [542] Google. Gemini 3 Generative UI: The Next Evolution of Interface Design[Z]. Google DeepMind Blog. 2025 (引用页: 168, 169).

- [543] Google. Google Antigravity: The Agentic Development Platform[Z]. Ranked #1 in LogRocket Developer Tools (December 2025). 2025 (引用页: 168, 169).
- [544] Figma. Figma Make: Prompt-to-App Prototyping Tool[Z]. Figma Product Releases. 2025 (引用页: 168, 169).
- [545] Google. Google Stitch: AI-Driven UI Design Experiment[Z]. Google Labs. 2025 (引用页: 169).
- [546] Google. Google Opal: Building AI Micro-Apps with Natural Language[Z]. Google Developers Blog. 2025 (引用页: 169).
- [547] Gamma. Introducing Gamma 3.0: The New Era of Human Communication[Z]. <https://gamma.app/insights/introducing-gamma-3-0>. Accessed: 2025-12-23. 2025 (引用页: 168, 170).
- [548] Max-Productive.ai. Gamma 3.0 Released: AI Agent & API Transform Presentations[Z]. <https://max-productive.ai/blog/gamma-3-0-ai-visual-storytelling-platform/>. Accessed: 2025-12-23. 2025 (引用页: 168).
- [549] Microsoft. Introducing Word, Excel, and PowerPoint Agents in Microsoft 365 Copilot[Z]. <https://techcommunity.microsoft.com/blog/microsoft365copilotblog/introducing-word-excel-and-powerpoint-agents-in-microsoft-365-copilot/4470604>. Accessed: 2025-12-23. 2025 (引用页: 168, 170).
- [550] Microsoft. What's New in Microsoft 365 Copilot: November & December 2025[Z]. <https://techcommunity.microsoft.com/blog/microsoft365copilotblog/whats-new-in-microsoft-365-copilot-november-december-2025/4469738>. Accessed: 2025-12-23. 2025 (引用页: 168, 170).
- [551] Gamma. Changelog: Gamma API and Agent Features[Z]. <https://meetgamma.canny.io/changelog>. Accessed: 2025-12-23. 2025 (引用页: 168, 170).
- [552] Microsoft. Meet Microsoft 365 Premium: Your AI and Productivity Powerhouse[Z]. <https://www.microsoft.com/en-us/microsoft-365/b>

- log/2025/10/01/meet-microsoft-365-premium-your-ai-and-productivity-powerhouse/. Accessed: 2025-12-23. 2025 (引用页: 170).
- [553] Canva. Magic Design for Presentations: AI-Powered Presentation Tool[Z]. <https://www.canva.com/create/ai-presentations/>. Accessed: 2025-12-23. 2025 (引用页: 170).
- [554] Runway. Runway Gen-4: Achieving Consistency in Video Generation[Z]. Runway Research. 2025 (引用页: 170, 171).
- [555] Black Forest Labs. FLUX.2: Open Source Image Generation Model [Z]. GitHub Repository. 2025 (引用页: 170, 171).
- [556] Kuaishou Technology. Kling 2.6: Integrated Video and Audio Generation Model[Z]. Kling AI Platform. 2025 (引用页: 171).
- [557] Adobe. Adobe Firefly 2025 Update: Harmonize and Video Capabilities[Z]. Adobe Creative Cloud Blog. 2025 (引用页: 171).
- [558] Figma. New Integrated AI Image Editing Tools in Figma[Z]. Figma Blog. 2025 (引用页: 171).
- [559] Google. Nano Banana Pro: Studio-Grade Image Generation Model [Z]. Google AI Blog. 2025 (引用页: 171).
- [560] Canva. Canva Native Design Model: From Templates to Design Understanding[Z]. Canva Engineering Blog. 2025 (引用页: 172).
- [561] Figma. Figma Buzz: Scaling Brand Consistency for Marketing Teams[Z]. Figma Config 2025. 2025 (引用页: 172).
- [562] Figma. Figma Sites: Direct Design-to-Web Publishing[Z]. Figma Product Announcement. 2025 (引用页: 172).
- [563] Adobe. Adobe Express AI Assistant: Conversational Design Workflows[Z]. Adobe Express Updates. 2025 (引用页: 172).
- [564] MockU Team. MockU: Rapid Project Presentation Generation[Z]. Product Launch. 2025 (引用页: 172).
- [565] Figma. Figma Draw: Enhanced Vector and Illustration Tools[Z]. Figma Updates. 2025 (引用页: 172).
- [566] PIAO J, YAN Y, ZHANG J, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of

- human behaviors and society[J]. arXiv preprint arXiv:2502.08691, 2025 (引用页: 174, 175).
- [567] ZHANG X, LIN J, MOU X, et al. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users[J]. arXiv preprint arXiv:2504.10157, 2025 (引用页: 174, 175).
- [568] WANG L, GAO H, BO X, et al. Yulan-onesim: Towards the next generation of social simulator with large language models[C]//Workshop on Scaling Environments for Agents (引用页: 174, 175).
- [569] HOU A B, DU H, WANG Y, et al. Can A Society of Generative Agents Simulate Human Behavior and Inform Public Health Policy? A Case Study on Vaccine Hesitancy[J]. arXiv preprint arXiv:2503.09639, 2025 (引用页: 175).
- [570] LI C J, WU J, MO Z, et al. Simulating Society Requires Simulating Thought[C]//The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track (引用页: 175).
- [571] ANTHIS J R, LIU R, RICHARDSON S M, et al. Position: LLM Social Simulations Are a Promising Research Method[C]//Forty-second International Conference on Machine Learning Position Paper Track (引用页: 175).
- [572] YAO J, WANG K, HSIEH R, et al. Spin-bench: How well do llms plan strategically and reason socially?[J]. arXiv preprint arXiv:2503.12349, 2025 (引用页: 175, 176).
- [573] WANG J, ZHAO Z, NI T, et al. SocioBench: Modeling Human Behavior in Sociological Surveys with Large Language Models[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 26268-26300 (引用页: 176).
- [574] XU Z, WANG Y, HUANG Y, et al. Socialmaze: A benchmark for evaluating social reasoning in large language models[J]. arXiv preprint arXiv:2505.23713, 2025 (引用页: 176).
- [575] ZHANG X, WANG W, JIN Q. IntentionESC: An Intention-Centered Framework for Enhancing Emotional Support in Dialogue

- Systems[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025: 26494-26516. <https://aclanthology.org/2025.findings-acl.1358/>. DOI: 10.18653/v1/2025.findings-acl.1358 (引用页: 177, 178).
- [576] CHEN Z, CAO Y, BI G, et al. SocialSim: Towards Socialized Simulation of Emotional Support Conversation[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 39: 2. 2025:1274-1282 (引用页: 178).
- [577] FU X, LI H, WANG B, et al. Look Beyond Feeling: Unveiling Latent Needs from Implicit Expressions for Proactive Emotional Support [C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025:21582-21609 (引用页: 178).
- [578] CHEN K, SUN Z, WEN Y, et al. Psy-Insight: Explainable Multi-turn Bilingual Dataset for Mental Health Counseling[J]. arXiv preprint arXiv:2503.03607, 2025 (引用页: 178).
- [579] QIU H, LAN Z. PsyDial: A Large-scale Long-term Conversational Dataset for Mental Health Support[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025:21624-21655 (引用页: 178).
- [580] WANG J, WANG B, FU X, et al. Psychological Counseling Cannot Be Achieved Overnight: Automated Psychological Counseling Through Multi-Session Conversations[J]. arXiv preprint arXiv:2506.06626, 2025 (引用页: 178, 179).
- [581] XU A, YANG D, LI R, et al. Autocbt: An autonomous multi-agent framework for cognitive behavioral therapy in psychological counseling[J]. arXiv preprint arXiv:2501.09426, 2025 (引用页: 179).
- [582] HU H, ZHOU Y, SI J, et al. Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling[J]. arXiv preprint arXiv:2505.15715, 2025 (引用页: 179).
- [583] HU H, ZHOU Y, MA C, et al. Theramind: A strategic and adaptive

- agent for longitudinal psychological counseling[J]. arXiv preprint arXiv:2510.25758, 2025 (引用页: 179).
- [584] YE J, XIANG L, ZHANG Y, et al. SweetieChat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025:4646-4669 (引用页: 179).
- [585] WANG M, WANG P, WU L, et al. AnnaAgent: Dynamic Evolution Agent System with Multi-Session Memory for Realistic Seeker Simulation[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025: 23221-23235. <https://aclanthology.org/2025.findings-acl.1192/>. DOI: 10.18653/v1/2025.findings-acl.1192 (引用页: 180).
- [586] ZHU S, CHEN Z, BI G, et al. Ψ -Arena: Interactive Assessment and Optimization of LLM-based Psychological Counselors with Tripartite Feedback[J]. arXiv preprint arXiv:2505.03293, 2025 (引用页: 180).
- [587] LI Y, YAO J, BUNYI J B S, et al. CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmark of Large Language Models in Mental Health Counseling[J]. arXiv preprint arXiv:2506.08584, 2025 (引用页: 180).
- [588] WANG B, SUN Y, WANG J, et al. CARE-Bench: A Benchmark of Diverse Client Simulations Guided by Expert Principles for Evaluating LLMs in Psychological Counseling[J]. arXiv preprint arXiv:2511.09407, 2025 (引用页: 180).
- [589] OpenAI. Deep Research[EB/OL]. -. 2025. <https://openai.com/index/introducing-deep-research/> (引用页: 181, 185).
- [590] LLC G. Gemini Deep Research[EB/OL]. 2024. <https://gemini.google/overview/deep-research/> (引用页: 181, 185).
- [591] AI P. Perplexity Deep Research[EB/OL]. 2025. <https://perplexity.ai/hub/blog/introducing-perplexity-deep-research> (引用页: 181,

185).

- [592] TEAM T D, LI B, ZHANG B, et al. Tongyi DeepResearch Technical Report[J]. arXiv preprint arXiv:2510.24701, 2025 (引用页: 183, 185).
- [593] WU J, YIN W, JIANG Y, et al. WebWalker: Benchmarking LLMs in Web Traversal[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 10290-10305. <https://aclanthology.org/2025.acl-long.508/>. DOI: 10.18653/v1/2025.acl-long.508 (引用页: 183).
- [594] LIU J, LI Y, ZHANG C, et al. WebExplorer: Explore and Evolve for Training Long-Horizon Web Agents[J/OL]. 2025. arXiv: 2509.06501 [cs.CL]. <https://arxiv.org/abs/2509.06501> (引用页: 183).
- [595] JIN B, ZENG H, YUE Z, et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning[J]. arXiv preprint arXiv:2503.09516, 2025 (引用页: 183).
- [596] LI X, DONG G, JIN J, et al. Search-o1: Agentic Search-Enhanced Large Reasoning Models[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 5420-5438. <https://aclanthology.org/2025.emnlp-main.276/>. DOI: 10.18653/v1/2025.emnlp-main.276 (引用页: 183).
- [597] GAO J, FU W, XIE M, et al. Beyond Ten Turns: Unlocking Long-Horizon Agentic Search with Large-Scale Asynchronous RL[J/OL]. 2025. arXiv: 2508.07976 [cs.CL]. <https://arxiv.org/abs/2508.07976> (引用页: 183).
- [598] WEI J, SUN Z, PAPAY S, et al. Browsecomp: A simple yet challenging benchmark for browsing agents[J]. arXiv preprint arXiv:2504.12516, 2025 (引用页: 184, 234).
- [599] COELHO J, NING J, HE J, et al. DeepResearchGym: A Free,

- Transparent, and Reproducible Evaluation Sandbox for Deep Research[J/OL]. 2025. arXiv: 2505.19253 [cs.IR]. <https://arxiv.org/abs/2505.19253> (引用页: 184).
- [600] DU M, XU B, ZHU C, et al. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents[J/OL]. 2025. arXiv: 2506.11763 [cs.CL]. <https://arxiv.org/abs/2506.11763> (引用页: 184).
- [601] GUPTA N, CHATTERJEE R, HAAS L, et al. DeepSearchQA : Bridging the Comprehensiveness Gap for Deep Research Agents [C/OL]//. <https://api.semanticscholar.org/CorpusID:283897826> (引用页: 184).
- [602] SHAO R, ASAI A, SHEN S Z, et al. DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research[J]. arXiv preprint arXiv:2511.19399, 2025 (引用页: 184).
- [603] Anthropic. Research[EB/OL]. -. 2025. <https://www.anthropic.com/research> (引用页: 185).
- [604] ByteDance. Doubao AI Study: Multimodal Educational Assistant [J]. 2024 (引用页: 185, 195).
- [605] LIU X, QIN B, LIANG D, et al. Autoglm: Autonomous foundation agents for guis[J]. arXiv preprint arXiv:2411.00820, 2024 (引用页: 185).
- [606] Moonshot. Deep Research[EB/OL]. -. 2025. <https://www.kimi.com/researcher> (引用页: 185).
- [607] ZHENG T, DENG Z, TSANG H T, et al. From automation to autonomy: A survey on large language models in scientific discovery [J]. arXiv preprint arXiv:2505.13259, 2025 (引用页: 186).
- [608] ZHOU Z, FENG X, HUANG L, et al. From Hypothesis to Publication: A Comprehensive Survey of AI-Driven Research Support Systems[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 2025: 11773-11803. <https://aclanthology.org/2>

025.findings-emnlp.631/. DOI: 10.18653/v1/2025.findings-emnlp.631 (引用页: 187).

- [609] CHEN Q, YANG M, QIN L, et al. AI4Research: A Survey of Artificial Intelligence for Scientific Research[J]. arXiv preprint arXiv:2507.01903, 2025 (引用页: 187).
- [610] TANG X, DUAN X, CAI Z. Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025:1602-1617 (引用页: 188).
- [611] WANG Y, GUO Q, YAO W, et al. Autosurvey: Large language models can automatically write surveys[J]. Advances in neural information processing systems, 2024, 37: 115119-115145 (引用页: 188).
- [612] LIANG X, YANG J, WANG Y, et al. Surveyx: Academic survey automation via large language models[J]. arXiv preprint arXiv:2502.14776, 2025 (引用页: 188).
- [613] ZHU K, LIAO L, GU Y, et al. Context-Aware Hierarchical Taxonomy Generation for Scientific Papers via LLM-Guided Multi-Aspect Clustering[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025:15627-15645 (引用页: 188).
- [614] WANG Q, DOWNEY D, JI H, et al. Scimon: Scientific inspiration machines optimized for novelty[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 279-299 (引用页: 188).
- [615] GOTTWEIS J, WENG W H, DARYIN A, et al. Towards an AI co-scientist[J]. arXiv preprint arXiv:2502.18864, 2025 (引用页: 188).
- [616] TANG J, XIA L, LI Z, et al. AI-Researcher: Autonomous Scientific Innovation[J]. arXiv preprint arXiv:2505.18705, 2025 (引用页: 188).
- [617] LU C, LU C, LANGE R T, et al. The ai scientist: Towards fully automated open-ended scientific discovery[J]. arXiv preprint arXiv:2408.06292, 2024 (引用页: 188, 189).

- [618] YAMADA Y, LANGE R T, LU C, et al. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search[J]. arXiv preprint arXiv:2504.08066, 2025 (引用页: 188).
- [619] GU N, HAHNLOSER R. Controllable citation sentence generation with language models[C]//Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024). 2024: 22-37 (引用页: 188).
- [620] YU L, ZHANG Q, SHI C, et al. Reinforced Subject-Aware Graph Neural Network for Related Work Generation[C]//International Conference on Knowledge Science, Engineering and Management. 2024: 201-213 (引用页: 189).
- [621] FARAJI A, TAVAKOLI M, MOEIN M, et al. Designing Effective LLM-Assisted Interfaces for Curriculum Development[C/OL]//Artificial Intelligence in Education: 26th International Conference, AIED 2025, Palermo, Italy, July 22–26, 2025, Proceedings, Part I. Palermo, Italy: Springer-Verlag, 2025: 438-451. https://doi.org/10.1007/978-3-031-98414-3_31. DOI: 10.1007/978-3-031-98414-3_31 (引用页: 191).
- [622] HUOVINEN L, HÄMÄLÄINEN M. LLM-Assisted, Iterative Curriculum Writing: A Human-Centered AI Approach in Finnish Higher Education[C/OL]//KOCHMAR E, ALHAFNI B, BEXTE M, et al. Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025). Vienna, Austria: Association for Computational Linguistics, 2025: 1002-1010. <https://aclanthology.org/2025.bea-1.76/>. DOI: 10.18653/v1/2025.bea-1.76 (引用页: 191).
- [623] ZHENG Y, HUANG S, ZENG X, et al. Knowledge-enhanced large language models for automatic lesson plan generation[J]. Humanities and Social Sciences Communications, 2025, 12(1):1784 (引用页: 191).
- [624] HU B, ZHU J, PEI Y, et al. Exploring the potential of LLM to enhance teaching plans through teaching simulation[J]. npj Science

of Learning, 2025, 10(1):7 (引用页: 191).

- [625] ZHANG X, ZHANG C, SUN J, et al. Eduplanner: Llm-based multi-agent systems for customized and intelligent instructional design[J]. IEEE Transactions on Learning Technologies, 2025 (引用页: 191).
- [626] YANG Q, LIANG C. A second-classroom personalized learning path recommendation system based on large language model technology [J]. Applied Sciences, 2025, 15(14): 7655 (引用页: 191).
- [627] WANG X J, LEE C P, MUTLU B. LearnMate: Enhancing On-line Education with LLM-Powered Personalized Learning Plans and Support[C]//Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. 2025: 1-10 (引用页: 192).
- [628] LI W, PEA R, HABER N, et al. CogGen: A Learner-Centered Generative AI Architecture for Intelligent Tutoring with Programming Videos[C]//International Conference on Artificial Intelligence in Education. 2025: 11-18 (引用页: 192).
- [629] GUPTA A, REDDIG J, CALO T, et al. Beyond final answers: Evaluating large language models for math tutoring[C]//International Conference on Artificial Intelligence in Education. 2025: 323-337 (引用页: 192).
- [630] NOORBAKHS K, CHANDLER J, KARIMI P, et al. Savaal: Scalable Concept-Driven Question Generation to Enhance Human Learning[J]. arXiv preprint arXiv:2502.12477, 2025 (引用页: 192).
- [631] WAHID R A, NADIM M S N, SULAIMAN S, et al. Automated Generation of Curriculum-Aligned Multiple-Choice Questions for Malaysian Secondary Mathematics Using Generative AI[J]. arXiv preprint arXiv:2508.04442, 2025 (引用页: 192).
- [632] ZOTOS L, van RIJN H, NISSIM M. Are You Doubtful? Oh, It Might Be Difficult Then! Exploring the Use of Model Uncertainty for Question Difficulty Estimation[J]. arXiv preprint arXiv:2412.11831, 2024 (引用页: 192).

- [633] ZHU Y, LIU D, LIN Z, et al. The LLM Already Knows: Estimating LLM-Perceived Question Difficulty via Hidden Representations [C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025:1160-1176 (引用页: 192).
- [634] SCARLATOS A, FERNANDEZ N, ORMEROD C, et al. Smart: Simulated students aligned with item response theory for question difficulty prediction[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 25082-25105 (引用页: 192).
- [635] YANG L, SUN X, LI H, et al. Difficulty aware programming knowledge tracing via large language models[J]. Scientific Reports, 2025, 15(1): 11475 (引用页: 193).
- [636] SCARLATOS A, BAKER R S, LAN A. Exploring knowledge tracing in tutor-student dialogues using llms[C]//Proceedings of the 15th International Learning Analytics and Knowledge Conference. 2025: 249-259 (引用页: 193).
- [637] ECNU-ICALK. EduChat: An open-source educational chat model [Z]. <https://github.com/icalk-nlp/EduChat>. Accessed: 2025-12-21 (引用页: 193).
- [638] Khan Academy. Khanmigo: An AI-powered tutor and teaching assistant[J]. 2023 (引用页: 194).
- [639] Duolingo. Duolingo Max: A learning experience powered by GPT-4 [J]. 2023 (引用页: 194).
- [640] Quizlet. Quizlet: Online Studying Platform[J]. 2025 (引用页: 194).
- [641] Photomath. Photomath: Camera Calculator and Step-by-Step Math Solver[J]. 2025 (引用页: 194).
- [642] Chegg. CheggMate: The AI Companion for Personalized Learning [J]. 2023 (引用页: 194).
- [643] Coursera. Coursera Coach: An AI-Powered Learning Assistant[J]. 2024 (引用页: 194).
- [644] Udemy. Udemy AI Assistant[J]. 2024 (引用页: 194).

- [645] 科大讯飞教育. 从智能到理解: AI 如何真正「懂」教育[J]. 2025 (引用页: 195).
- [646] People's Daily Online. Zizai Xinli: AI System for Campus Mental Health Support[Z]. <https://zzdmx.rmrbsn.cn/>. 2025 (引用页: 195).
- [647] Tencent. Tencent Youth-Oriented Large Language Model[Z]. <https://ai.tencent.com/>. Accessed: 2025-03; Features minor-protection mode, content safety, and anti-addiction design. 2024 (引用页: 195).
- [648] YANG J. Yuanfudao: Transforming China's K12 Online Education Landscape[G]//Cases on Chinese Unicorns and the Development of Startups. IGI Global, 2025: 413-432 (引用页: 195).
- [649] LI H, FAN X, LIANG J. The squirrel AI adaptive learning system accompanying millions of children in their growth[G]//Digital transformation of regional education in China. Springer, 2025: 51-53 (引用页: 195).
- [650] GLM T, ZENG A, XU B, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools[J]. arXiv preprint arXiv:2406.12793, 2024 (引用页: 195).
- [651] SINGHAL K, TU T, GOTTWEIS J, et al. Toward expert-level medical question answering with large language models[J]. Nature Medicine, 2025, 31(3): 943-950 (引用页: 197, 198).
- [652] XIE Q, CHEN Q, CHEN A, et al. Medical foundation large language models for comprehensive text analysis and beyond[J]. npj Digital Medicine, 2025, 8(1): 141 (引用页: 197, 198).
- [653] WU C, QIU P, LIU J, et al. Towards evaluating and building versatile large language models for medicine[J]. npj Digital Medicine, 2025, 8(1): 58 (引用页: 197, 198).
- [654] LIU X, LIU H, YANG G, et al. A generalist medical language model for disease diagnosis assistance[J]. Nature medicine, 2025, 31(3): 932-942 (引用页: 197, 198).
- [655] LIU C, WANG H, PAN J, et al. Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl[J]. arXiv preprint arXiv:2505.17952, 2025 (引用页: 197, 198).

- [656] HUANG X, WU J, LIU H, et al. m1: Unleash the potential of test-time scaling for medical reasoning with large language models[J]. arXiv preprint arXiv:2504.00869, 2025 (引用页: 197, 199).
- [657] LIU C, LI D, SHU Y, et al. Fleming-r1: Toward expert-level medical reasoning via reinforcement learning[J]. arXiv preprint arXiv:2509.15279, 2025 (引用页: 197, 199).
- [658] JIANG S, LIAO Y, CHEN Z, et al. MedS[^]3: Towards Medical Slow Thinking with Self-Evolved Soft Dual-sided Process Supervision[J]. arXiv preprint arXiv:2501.12051, 2025 (引用页: 197, 199).
- [659] TU T, SCHAEKERMANN M, PALEPU A, et al. Towards conversational diagnostic artificial intelligence[J]. Nature, 2025: 1-9 (引用页: 197, 199).
- [660] REN Z, ZHAN Y, YU B, et al. Healthcare agent: eliciting the power of large language models for medical consultation[J]. npj Artificial Intelligence, 2025, 1(1): 24 (引用页: 197, 199).
- [661] XU S, HUANG X, WEI Z, et al. Reverse Physician-AI Relationship: Full-process Clinical Diagnosis Driven by a Large Language Model [J]. arXiv preprint arXiv:2508.10492, 2025 (引用页: 197, 199).
- [662] WANG H, LIU C, XI N, et al. Huatuo: Tuning llama model with chinese medical knowledge[J]. arXiv preprint arXiv:2304.06975, 2023 (引用页: 198).
- [663] WU C, LIN W, ZHANG X, et al. PMC-LLaMA: toward building open-source language models for medicine[J]. Journal of the American Medical Informatics Association, 2024, 31(9): 1833-1843 (引用页: 198).
- [664] WANG Y, DAI Y, JONES C, et al. Enhancing vision-language models for medical imaging: bridging the 3D gap with innovative slice selection[J]. Advances in Neural Information Processing Systems, 2024, 37: 99947-99964 (引用页: 199, 200).
- [665] LU M Y, CHEN B, WILLIAMSON D F, et al. A visual-language foundation model for computational pathology[J]. Nature medicine, 2024, 30(3): 863-874 (引用页: 199, 200).

- [666] ZHANG K, ZHOU R, ADHIKARLA E, et al. A generalist vision–language foundation model for diverse biomedical tasks[J]. Nature Medicine, 2024, 30(11): 3129-3141 (引用页: 199, 200).
- [667] DING T, WAGNER S J, SONG A H, et al. A multimodal whole-slide foundation model for pathology[J]. Nature medicine, 2025: 1-13 (引用页: 199, 200).
- [668] CHEN Y, XU D, HUANG Y, et al. MIMO: A Medical Vision Language Model with Visual Referring Multimodal Input and Pixel Grounding Multimodal Output[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 24732-24741 (引用页: 199, 200).
- [669] YANG X, MIAO J, YUAN Y, et al. Medical large vision language models with multi-image visual ability[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2025: 402-412 (引用页: 200).
- [670] ZHU X, HU Y, MO F, et al. Uni-med: a unified medical generalist foundation model for multi-task learning via connector-MoE [J]. Advances in Neural Information Processing Systems, 2024, 37: 81225-81256 (引用页: 200).
- [671] CHEN A, LOU L, CHEN K, et al. Benchmarking llms for translating classical chinese poetry: Evaluating adequacy, fluency, and elegance [C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 33007-33024 (引用页: 201, 202).
- [672] NIMO C, OLATUNJI T, OWODUNNI A T, et al. AfriMed-QA: a Pan-African, multi-specialty, medical question-answering benchmark dataset[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 1948-1973 (引用页: 201, 202).
- [673] SONI S, GAYEN S, DEMNER-FUSHMAN D. Overview of the archehr-qa 2025 shared task on grounded question answering from electronic health records[C]//Proceedings of the 24th Workshop on Biomedical Language Processing. 2025: 396-405 (引用页: 201, 202).

- [674] GUO E, ZHAO Z, WANG Z, et al. DiN: Diffusion Model for Robust Medical VQA with Semantic Noisy Labels[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 14337-14346 (引用页: 201, 202).
- [675] ACHARYA A, GHOSH A, VERMA P, et al. M3Retrieve: Benchmarking Multimodal Retrieval for Medicine[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 15274-15287 (引用页: 201, 202).
- [676] LIU J, WANG W, MA Z, et al. MedChain: Bridging the Gap Between LLM Agents and Clinical Practice with Interactive Sequence [C]//The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (引用页: 201, 203).
- [677] ZHU Y, HE Z, HU H, et al. MedAgentBoard: Benchmarking Multi-Agent Collaboration with Conventional Methods for Diverse Medical Tasks[J]. arXiv preprint arXiv:2505.12371, 2025 (引用页: 201, 203).
- [678] GAO S, ZHU R, KONG Z, et al. CURE-Bench: Competition on Reasoning Models for Drug Decision-Making in Precision Therapeutics[EB/OL]. 2025 [2025-12-19]. <https://neurips.cc/virtual/2025/competition/127720> (引用页: 201, 203).
- [679] YANG W, PENG J. Beyond Summaries: Multi-Agent Generation of Investment Reports with Text, Tables, and Charts[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 205, 206).
- [680] SINHA G, et al. Agentic LLMs for Analyst-Style Financial Insights: An LLM Pipeline for Persuasive Financial Analysis[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 205, 206).
- [681] NITARACH N, et al. FinCoT: Grounding Chain-of-Thought in Ex-

- pert Financial Reasoning[EB/OL]. 2025. <https://arxiv.org/abs/2506.16123>. arXiv: 2506.16123 [cs.CL] (引用页: 205, 206).
- [682] SHUKLA N K, et al. GraphRAG Analysis for Financial Narrative Summarization and A Framework for Optimizing Domain Adaptation[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 205, 206).
- [683] NARARATWONG R, et al. Fin-DBQA Shared-task: Database Querying and Reasoning[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 205, 206).
- [684] THEERTHALA A. Synthesizing Behaviorally-Grounded Reasoning Chains: A Data-Generation Framework for Personal Finance LLMs [C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 205, 206).
- [685] NISHIDA S, UTSURO T. Generating Financial News Articles from Factors of Stock Price Rise / Decline by LLMs[C/OL]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025. <https://aclanthology.org/2025.finnlp-1.18/> (引用页: 206, 207).
- [686] SHUKLA N K, et al. KULFi Framework: Knowledge Utilization for Optimizing Large Language Models for Financial Causal Reasoning [C/OL]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025. <https://aclanthology.org/2025.finnlp-1.22/> (引用页: 206, 207).
- [687] S V K, et al. CLRG@FinCausal2025: Cause-Effect Extraction in Finance Domain[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP).

- Association for Computational Linguistics, 2025 (引用页: 206, 207).
- [688] LIN C Y, JANG J S. Concept-Based RAG Models: A High-Accuracy Fact Retrieval Approach[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 206, 207).
 - [689] CAI T, et al. FinDebate: Multi-Agent Collaborative Intelligence for Financial Analysis[EB/OL]. 2025. <https://arxiv.org/abs/2509.17395> 5. arXiv: 2509.17395 [cs.CL] (引用页: 206, 207).
 - [690] WANG Y, et al. Sam's Fans at the Crypto Trading Challenge Task: A Threshold-Based Decision Approach Based on FinMem Framework[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 206, 207).
 - [691] HADLOCK S, et al. Enhancing Post Earnings Announcement Drift Measurement with Large Language Models[C/OL]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025. <https://aclanthology.org/2025.finnlp-2.13/> (引用页: 206, 207).
 - [692] Anonymous. 300k/ns team at the Crypto Trading Challenge Task: Enhancing the justification of accurate trading decisions through parameter-efficient fine-tuning of reasoning models[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 206, 207).
 - [693] YAN S, ZHU T. CreditLLM: Constructing Financial AI Assistant for Credit Products using Financial LLM and Few Data[C/OL]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025. <https://aclanthology.org/2025.finnlp-1.13/> (引用页: 206, 208).

- [694] DRINKALL F, et al. Forecasting Credit Ratings: A Case Study where Traditional Methods Outperform Generative LLMs[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 206, 208).
- [695] XU Z, LIU Y, WANG Y, et al. Modeling Interactions Between Stocks Using LLM-Enhanced Graphs for Volume Prediction[C/OL]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025. <https://aclanthology.org/2025.finnlp-1.14/> (引用页: 206, 208).
- [696] ZHOU H, et al. Zero-Shot Extraction of Stock Relationship Graphs with LLMs[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 206, 208).
- [697] ZHANG X, YANG Q. FinMoE: A MoE-based Large Chinese Financial Language Model[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 206, 208).
- [698] BIANCOTTI C, CAMASSA C, COLETTA A, et al. Chat Bankman-Fried: an Exploration of LLM Alignment in Finance [EB/OL]. 2024. <https://arxiv.org/abs/2411.11853>. arXiv: 2411.11853 [cs.CL] (引用页: 206, 208).
- [699] AMIN M, ASSENMACHER M. Do Companies Reveal Their Own Fraud? - A Novel Data Set for Fraud Detection Based on 10-K Reports[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 206, 208).
- [700] LAI V D, KRUMDICK M, LOVERING C, et al. SEC-QA: A Systematic Evaluation Corpus for Financial QA[EB/OL]. 2024. <https://arxiv.org/abs/2406.14394>. arXiv: 2406.14394 [cs.CL] (引用

页: 206, 209).

- [701] SHU R, et al. LAVA: Logic-Aware Validation and Augmentation Framework for Large-Scale Financial Document Auditing[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 206, 209).
- [702] NUAIMI K A, et al. Detecting Evasive Answers in Financial Q&A: A Psychological Discourse Taxonomy and Lightweight Baselines[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 206, 209).
- [703] KLIMASZEWSKI M, et al. AveniBench: Accessible and Versatile Evaluation of Finance Intelligence[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 206, 209).
- [704] LUO Z, et al. FMD-Mllama at the Financial Misinformation Detection Challenge Task: Multimodal Reasoning and Evidence Generation[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 206, 208).
- [705] DOU S, et al. FinEval-KR: A Financial Domain Evaluation Framework for Large Language Models' Knowledge and Reasoning [EB/OL]. 2025. <https://arxiv.org/abs/2506.21591>. arXiv: 2506.21591 [cs.CL] (引用页: 206, 209).
- [706] ROSERO A G F, et al. Evaluating Financial Literacy of Large Language Models through Domain Specific Languages for Plain Text Accounting[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 206, 209).
- [707] CAO Y, et al. Capybara at the Financial Misinformation Detection Challenge Task: Chain-of-Thought Enhanced Financial Mis-

- information Detection[C/OL]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025. <https://aclanthology.org/2025.finnlp-1.38/> (引用页: 206, 208).
- [708] SATAPATHY R, et al. From Earnings Calls to Investment Reports: Evaluating Role-based Multi-Agent LLM Systems[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 205).
- [709] STRICH J. Adapt LLM for Multi-turn Reasoning QA using Tidy Data[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 205).
- [710] WANG X, BRORSSON M. Can Large language model analyze financial statements well?[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 205).
- [711] SANDOVAL A M, et al. The Financial Document Causality Detection Shared Task (FinCausal 2025)[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 207).
- [712] TRIVEDI A, et al. Sarang at FinCausal 2025: Contextual QA for Financial Causality Detection Combining Extractive and Generative Models[C/OL]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025. <https://aclanthology.org/2025.finnlp-1.25/> (引用页: 207).
- [713] CHATWAL P, et al. Enhancing Causal Relationship Detection Using Prompt Engineering and Large Language Models[C]//Proceedings of the Joint Workshop of the 9th Financial Technol-

- ogy and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 207).
- [714] NIESS G, et al. Addressing Hallucination in Causal Q&A: The Efficacy of Fine-tuning over Prompting in LLMs[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 207).
- [715] Sebbag, et al. LLM as a Guide: an Approach for Unsupervised Economic Relation Discovery in Administrative Documents[C/OL]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025. <https://aclanthology.org/2025.finnlp-2.1/> (引用页: 207).
- [716] JEENOOR M, et al. PresiUniv at FinCausal 2025 Shared Task: Applying Fine-tuned Language Models to Explain Financial Cause and Effect with Zero-shot Learning[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 207).
- [717] AVILES M J M, VACA A. Extracting Financial Causality through QA: Insights from FinCausal 2025 Spanish Subtask[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 207).
- [718] TAKAYANAGI T, et al. Earnings2Insights: Analyst Report Generation for Investment Guidance[C/OL]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025. <https://aclanthology.org/2025.finnlp-2.17/> (引用页: 207).
- [719] TAN M, et al. Multi-Agent Collaboration for Investment Guidance: Earnings2Insights Report Generation[C/OL]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural

Language Processing (FinNLP). Association for Computational Linguistics, 2025. <https://aclanthology.org/2025.finnlp-2.26/> (引用页: 207).

- [720] RALLABANDI S, et al. Jetsons at the FinNLP-2025 - Earnings2Insights: Persuasive Investment Report Generation Using Single And Multi-Agent Frameworks[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 207).
- [721] CHATWAL P, et al. Meta Prompting for Analyst Report Generation: Turning Earnings Calls into Investment Guidance[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 207).
- [722] YU Y, et al. FinNLP-FNP-LLMFinLegal @ COLING 2025 Shared Task: Agent-Based Single Cryptocurrency Trading Challenge[C/OL]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025. <https://aclanthology.org/2025.finnlp-1.46/> (引用页: 207).
- [723] WANG Y, et al. Proxy Tuning for Financial Sentiment Analysis: Overcoming Data Scarcity and Computational Barriers[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 208).
- [724] SRINIVASAN A G, et al. Enhancing Financial RAG with Agentic AI and Multi-HyDE: A Novel Approach to Knowledge Retrieval and Hallucination Reduction[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 208).
- [725] LITHGOW-SERRANO O, et al. Assessing RAG System Capabil-

- ities on Financial Documents[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 208).
- [726] LIU Z, et al. FinNLP-FNP-LLMFinLegal-2025 Shared Task: Financial Misinformation Detection Challenge Task[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 208).
- [727] PURBEY J, et al. 1-800-SHARED-TASKS at the Financial Misinformation Detection Challenge Task: Sequential Learning for Claim Verification and Explanation Generation in Financial Domains[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 208).
- [728] WU Y, et al. Natural Language Inference as a Judge: Detecting Factuality and Causality Issues in Language Model Self-Reasoning for Financial Analysis[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 209).
- [729] WANG K, et al. A Report on Financial Regulations Challenge at COLING 2025 FinNLP-FNP-LLMFinLegal-2025 Shared Task [EB/OL]. 2025. <https://arxiv.org/abs/2412.11159>. arXiv: 2412.11159 [cs.CL] (引用页: 209).
- [730] MARTÍNEZ S, et al. A Scalable Framework for Legal Text Understanding in Regulatory and Financial Contexts[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 209).
- [731] CHANTANGPHOL P, et al. FinMind-Y-Me at the Regulations Challenge Task: Financial Mind Your Meaning based on THaLLE

- [C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 209).
- [732] HUANG J, et al. Audit-FT at the Regulations Challenge Task: An Open-Source Large Language Model for Audit[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 209).
- [733] JIANG S, et al. IntelliChain Stars at the Regulations Challenge Task: A Large Language Model for Financial Regulation[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 209).
- [734] WANG D, et al. BuDDIE: A Business Document Dataset for Multi-task Information Extraction[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 209).
- [735] ABDO M S, et al. AMWAL: Named Entity Recognition for Arabic Financial News[C/OL]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025. <https://aclanthology.org/2025.finnlp-1.20/> (引用页: 209).
- [736] HARSHA C, et al. Synthetic Data Generation Using Large Language Models for Financial Question Answering[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 209).
- [737] ZHAO Y, et al. A Self-Improving Method for Generating Descriptions of Financial Data Quality Grading Using LLMs[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for

Computational Linguistics, 2025 (引用页: 209).

- [738] SHARMA A K, et al. Towards Efficient FinBERT via Quantization and Coreset for Financial Sentiment Analysis[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 209).
- [739] LU Y T, HUO Y. Financial Named Entity Recognition: How Far Can LLM Go?[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 209).
- [740] KUMAR S, et al. Bridging the Gap: Efficient Cross-Lingual NER in Low-Resource Financial Domain[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 209).
- [741] UTHAYASOORIYAR B, et al. Training LayoutLM from Scratch for Efficient Named-Entity Recognition in the Insurance Domain[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 209).
- [742] QIU L, CHERSONI E. GenChaR: A Dataset for Stock Chart Captioning[EB/OL]. 2024. <https://arxiv.org/abs/2412.04041>. arXiv: 2412.04041 [cs.CV] (引用页: 209).
- [743] Anonymous. Investigating the effectiveness of length based rewards in DPO for building Conversational Financial Question Answering Systems[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP). Association for Computational Linguistics, 2025 (引用页: 209).
- [744] LIU J, TONG Y, HUANG H, et al. Legal Fact Prediction: The Missing Piece in Legal Judgment Prediction[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Proceedings of the 2025 Conference on Empirical Methods in

Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 6345-6360. <https://aclanthology.org/2025.emnlp-main.322/>. DOI: 10.18653/v1/2025.emnlp-main.322 (引用页: 211, 212).

- [745] XU Q, LIU Q, FEI H, et al. CLEAR: A Framework Enabling Large Language Models to Discern Confusing Legal Paragraphs[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 2025: 8937-8953. <https://aclanthology.org/2025.findings-emnlp.475/>. DOI: 10.18653/v1/2025.findings-emnlp.475 (引用页: 211, 212).
- [746] ZHANG K, XIE G, YU W, et al. Legal Mathematical Reasoning with LLMs: Procedural Alignment through Two-Stage Reinforcement Learning[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 2025: 1586-1598. <https://aclanthology.org/2025.findings-emnlp.84/>. DOI: 10.18653/v1/2025.findings-emnlp.84 (引用页: 211, 212, 215).
- [747] CHEN L, CAI Y, HOU Z, et al. Towards Trustworthy Legal AI through LLM Agents and Formal Reasoning[J/OL]. 2025. arXiv: 2511.21033 [cs.AI]. <https://arxiv.org/abs/2511.21033> (引用页: 211, 212).
- [748] LIU H, HUANG Q, CHEN Q, et al. JUREX-4E: Juridical Expert-Annotated Four-Element Knowledge Base for Legal Reasoning[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 3794-3814. <https://aclanthology.org/2025.emnlp-main.188/>. DOI: 10.18653/v1/2025.emnlp-main.188 (引用页: 211, 212).
- [749] ZHANG K, YU W, SUN Z, et al. SyLeR: A Framework for Ex-

- PLICIT Syllogistic Legal Reasoning in Large Language Models[C/OL]
//CIKM '25: Proceedings of the 34th ACM International Conference on Information and Knowledge Management. Seoul, Republic of Korea: Association for Computing Machinery, 2025:4117-4127. <https://doi.org/10.1145/3746252.3761120>. DOI: 10.1145/3746252.3761120 (引用页: 211, 212).
- [750] CHLAPANIS O S, GALANIS D, ALETRAS N, et al. GreekBarBench: A Challenging Benchmark for Free-Text Legal Reasoning and Citations[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 2025:25099-25119. <https://aclanthology.org/2025.findings-emnlp.1368/>. DOI: 10.18653/v1/2025.findings-emnlp.1368 (引用页: 212, 213).
- [751] CAI H, ZHAO S, ZHANG L, et al. Unilaw-R1: A Large Language Model for Legal Reasoning with Reinforcement Learning and Iterative Inference[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025:18128-18142. <https://aclanthology.org/2025.emnlp-main.915/>. DOI: 10.18653/v1/2025.emnlp-main.915 (引用页: 212, 213).
- [752] HU Y, YU Y, GAN L, et al. Evaluating Test-Time Scaling LLMs for Legal Reasoning: OpenAI o1, DeepSeek-R1, and Beyond[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 2025:13759-13781. <https://aclanthology.org/2025.findings-emnlp.742/>. DOI: 10.18653/v1/2025.findings-emnlp.742 (引用页: 212, 213).
- [753] T.Y.S.S S, HERNANDEZ E Q. LexKeyPlan: Planning with Keyphrases and Retrieval Augmentation for Legal Text Generation: A Case Study on European Court of Human Rights Cases[C/OL]

//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vienna, Austria: Association for Computational Linguistics, 2025:425-436. <https://aclanthology.org/2025.acl-short.32/>. DOI: 10.18653/v1/2025.acl-short.32 (引用页: 212, 213).

- [754] ROLSHOVEN L, RASIAH V, BOSE S B, et al. Unlocking Legal Knowledge: A Multilingual Dataset for Judicial Summarization in Switzerland[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 2025: 15382-15411. <https://aclanthology.org/2025.findings-emnlp.832/>. DOI: 10.18653/v1/2025.findings-emnlp.832 (引用页: 212, 213).

- [755] NIKLAUS J, MERANE J, NENADIC L, et al. SwiLTra-Bench: The Swiss Legal Translation Benchmark[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 14894-14916. <https://aclanthology.org/2025.acl-long.725/>. DOI: 10.18653/v1/2025.acl-long.725 (引用页: 212, 213).

- [756] WOO J, HASHEMI CHALESHTORI F, MARASOVIC A, et al. BriefMe: A Legal NLP Benchmark for Assisting with Legal Briefs [C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025:13139-13190. <https://aclanthology.org/2025.findings-acl.681/>. DOI: 10.18653/v1/2025.findings-acl.681 (引用页: 212, 213).

- [757] T.Y.S.S S, ELKHAYAT Y T, ICHIM O, et al. CoCoLex: Confidence-guided Copy-based Decoding for Grounded Legal Text Generation [C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association

- for Computational Linguistics, 2025: 19002-19018. <https://aclanthology.org/2025.acl-long.931/>. DOI: 10.18653/v1/2025.acl-long.931 (引用页: 212, 213).
- [758] ZHOU S, WU Y, CHEN H, et al. ClaimGen-CN: A Large-scale Chinese Dataset for Legal Claim Generation[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 2025: 12296-12323. <https://aclanthology.org/2025.findings-emnlp.658/>. DOI: 10.18653/v1/2025.findings-emnlp.658 (引用页: 212, 214).
- [759] BEAUCHEMIN D, ALBERT-ROCHETTE M, KHOURY R, et al. JUDGEBERT: Assessing Legal Meaning Preservation Between Sentences[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 92-118. <https://aclanthology.org/2025.emnlp-main.5/>. DOI: 10.18653/v1/2025.emnlp-main.5 (引用页: 212, 214, 215).
- [760] KIM C, LEE J, HWANG W. LegalSearchLM: Rethinking Legal Case Retrieval as Legal Elements Generation[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 4521-4554. <https://aclanthology.org/2025.emnlp-main.225/>. DOI: 10.18653/v1/2025.emnlp-main.225 (引用页: 212, 214).
- [761] UPADHYA R, T.Y.S.S S. LexCLiPR: Cross-Lingual Paragraph Retrieval from Legal Judgments[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 13971-13993. <https://aclanthology.org/2025.acl-long.683/>. DOI:

10.18653/v1/2025.acl-long.683 (引用页: 212, 214).

- [762] LI A, WU Y, LIU Y, et al. UniLR: Unleashing the Power of LLMs on Multiple Legal Tasks with a Unified Legal Retriever[C/OL]// CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 11953-11967. <https://aclanthology.org/2025.acl-long.584/>. DOI: 10.18653/v1/2025.acl-long.584 (引用页: 212, 214).
- [763] ZHANG K, YU W, DAI S, et al. CitaLaw: Enhancing LLM with Citations in Legal Domain[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025: 11183-11196. <https://aclanthology.org/2025.findings-acl.583/>. DOI: 10.18653/v1/2025.findings-acl.583 (引用页: 212, 214).
- [764] WANG S H, ZUBKOV M, FAN K, et al. ACORD: An Expert-Annotated Retrieval Dataset for Legal Contract Drafting[C/OL]// CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 24739-24762. <https://aclanthology.org/2025.acl-long.1206/>. DOI: 10.18653/v1/2025.acl-long.1206 (引用页: 212, 214).
- [765] LEE J, KIM D, HWANG S, et al. KoBLEX: Open Legal Question Answering with Multi-hop Reasoning[C/OL]// CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 4019-4053. <https://aclanthology.org/2025.emnlp-main.200/>. DOI: 10.18653/v1/2025.emnlp-main.200 (引用页: 212, 214).
- [766] CRACIUN C G, SMĂDU R A, CERCEL D C, et al. GRAF: Graph

- Retrieval Augmented by Facts for Romanian Legal Multi-Choice Question Answering[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025: 12708-12742. <https://aclanthology.org/2025.findings-acl.659/>. DOI: 10.18653/v1/2025.findings-acl.659 (引用页: 212, 215).
- [767] TANG X, LI J, HU K, et al. CogniBench: A Legal-inspired Framework and Dataset for Assessing Cognitive Faithfulness of Large Language Models[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 21567-21585. <https://aclanthology.org/2025.acl-long.1046/>. DOI: 10.18653/v1/2025.acl-long.1046 (引用页: 212, 215).
- [768] LI H, CHEN J, YANG J, et al. LegalAgentBench: Evaluating LLM Agents in Legal Domain[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 2322-2344. <https://aclanthology.org/2025.acl-long.116/>. DOI: 10.18653/v1/2025.acl-long.116 (引用页: 212, 215).
- [769] LI H, HU W, JING H, et al. PrivaCI-Bench: Evaluating Privacy with Contextual Integrity and Legal Compliance[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 10544-10559. <https://aclanthology.org/2025.acl-long.518/>. DOI: 10.18653/v1/2025.acl-long.518 (引用页: 212, 215).
- [770] SHEN X, JIANG Z, ZHANG J, et al. ProvBench: A Benchmark of Legal Provision Recommendation for Contract Auto-Reviewing [C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association

- for Computational Linguistics, 2025: 6240-6254. <https://aclanthology.org/2025.acl-long.312/>. DOI: 10.18653/v1/2025.acl-long.312 (引用页: 212, 215, 216).
- [771] SADOWSKI A, CHUDZIAK J A. On Verifiable Legal Reasoning: A Multi-Agent Framework with Formalized Knowledge Representations[C/OL]//CIKM '25. Seoul, Republic of Korea: Association for Computing Machinery, 2025: 2535-2545. <https://doi.org/10.1145/3746252.3761057>. DOI: 10.1145/3746252.3761057 (引用页: 212, 215).
- [772] WEI K, SHI X, TONG J, et al. LegalCore: A Dataset for Event Coreference Resolution in Legal Documents[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025: 25044-25059. <https://aclanthology.org/2025.findings-acl.1284/>. DOI: 10.18653/v1/2025.findings-acl.1284 (引用页: 212, 216).
- [773] BARALE C, BARRETT L, BAJAJ V S, et al. LexTime: A Benchmark for Temporal Ordering of Legal Events[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 2025: 5220-5236. <https://aclanthology.org/2025.findings-emnlp.280/>. DOI: 10.18653/v1/2025.findings-emnlp.280 (引用页: 212, 216).
- [774] RAPTOPOULOS P, FILANDRIANOS G, LYMPERAIIOU M, et al. PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 7959-7995. <https://aclanthology.org/2025.emnlp-main.403/>. DOI: 10.18653/v1/2025.emnlp-main.403 (引用页: 212, 215, 216).
- [775] SHI W, ZHU H, JI J, et al. LegalReasoner: Step-wised Verification-Correction for Legal Judgment Reasoning[C/OL]//CHE W,

- NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 7297-7313. <https://aclanthology.org/2025.acl-long.361/>. DOI: 10.18653/v1/2025.acl-long.361 (引用页: 211).
- [776] XIE H, LI C, ZHU H, et al. LawChain: Modeling Legal Reasoning Chains for Chinese Tort Case Analysis[J/OL]. 2025. arXiv: 2510.17602 [cs.CL]. <https://arxiv.org/abs/2510.17602> (引用页: 211).
- [777] LUO K, HUANG Q, JIANG C, et al. Automating Legal Interpretation with LLMs: Retrieval, Generation, and Evaluation[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 4015-4047. <https://aclanthology.org/2025.acl-long.204/>. DOI: 10.18653/v1/2025.acl-long.204 (引用页: 211).
- [778] HAN Z, YANG Y, FENG Y, et al. LawShift: Benchmarking Legal Judgment Prediction Under Statute Shifts[C/OL]//The Thirtieth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track. 2025. <https://openreview.net/forum?id=5SpFenlxDF> (引用页: 211).
- [779] XU X, ZHAO L, XU H, et al. CLaw: Benchmarking Chinese Legal Knowledge in Large Language Models - A Fine-grained Corpus and Reasoning Analysis[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 2025: 12071-12103. <https://aclanthology.org/2025.findings-emnlp.646/>. DOI: 10.18653/v1/2025.findings-emnlp.646 (引用页: 211).
- [780] GUPTA A, RICE D, O'CONNOR B. δ -Stance: A Large-Scale Real World Dataset of Stances in Legal Argumentation[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational

- Linguistics, 2025: 31450-31467. <https://aclanthology.org/2025.acl-long.1517/>. DOI: 10.18653/v1/2025.acl-long.1517 (引用页: 213).
- [781] DAI X, XU B, LIU Z, et al. Legal Δ : Enhancing Legal Reasoning in LLMs via Reinforcement Learning with Chain-of-Thought Guided Information Gain[J/OL]. 2025. arXiv: 2508.12281 [cs.CL]. <https://arxiv.org/abs/2508.12281> (引用页: 213).
- [782] PAUL S, GHUMARE D, GOYAL P, et al. IL-PCSR: Legal Corpus for Prior Case and Statute Retrieval[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 14599-14622. <https://aclanthology.org/2025.emnlp-main.738/>. DOI: 10.18653/v1/2025.emnlp-main.738 (引用页: 214).
- [783] De MARTIM H. An Ontology-Driven Graph RAG for Legal Norms: A Structural, Temporal, and Deterministic Approach[J/OL]. 2025. arXiv: 2505.00039 [cs.CL]. <https://arxiv.org/abs/2505.00039> (引用页: 215).
- [784] KIM Y, LEE W. Where Does Legal AI Fail? Evaluating RAG Pipelines[C/OL]//CIKM '25: Proceedings of the 34th ACM International Conference on Information and Knowledge Management. Seoul, Republic of Korea: Association for Computing Machinery, 2025: 1396-1405. <https://doi.org/10.1145/3746252.3761151>. DOI: 10.1145/3746252.3761151 (引用页: 215).
- [785] AKARAJARADWONG P, POTHAVORN P, CHAKSANGCHAI-CHOT C, et al. NitiBench: Benchmarking LLM Frameworks on Thai Legal Question Answering Capabilities[C/OL]//CHRISTODOULOPOULOS C, CHAKRABORTY T, ROSE C, et al. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025: 34292-34315. <https://aclanthology.org/2025.emnlp-main.1739/>. DOI: 10.18653/v1/2025.emnlp-main.1739 (引用页: 215).

- [786] J F, S Z, L F, et al. Breeding 5.0: Artificial Intelligence (AI)-Decoded Germplasm for Accelerated Crop Innovation[J]. PubMed, 2025 [2025-12-27] (引用页: 217, 218).
- [787] JOHNS L. Syngenta and InstaDeep Collaborate to Accelerate Crops Seeds Trait Research Using AI Large Language Models | InstaDeep - Decision-Making AI For The Enterprise[Z]. 2024 [2025-12-27] (引用页: 217, 218).
- [788] YANG F, KONG H, YING J, et al. SeedLLM-Rice: A Large Language Model Integrated with Rice Biological Knowledge Graph [J/OL]. Molecular Plant, 2025, 18(7): 1118-1129 [2025-12-31]. DOI: 10.1016/j.molp.2025.05.013 (引用页: 218, 221).
- [789] Introducing Heritable Agriculture - Google X Blog[Z]. <https://x.company/blog/posts/heritable-agriculture/>. [2025-12-31] (引用页: 218, 222).
- [790] 全球首个!浙大“AI 育种家”重磅发布[Z]. <https://tidenews.com.cn/news.html?id=3259838>. [2025-12-31] (引用页: 218, 221).
- [791] NASA, IBM Research to Release New AI Model for Weather, Climate - NASA Science[Z]. 2024 [2025-12-31] (引用页: 218, 222).
- [792] See & Spray™ Ultimate | Precision Ag | John Deere US[Z]. <https://www.deere.com/en/sprayers/see-spray-ultimate/>. [2025-12-27] (引用页: 218, 219).
- [793] HIT Co-Hosts CNCC2025[Z]. <https://github.com/HIT-Kwoo/>. [2025-12-27] (引用页: 218, 222).
- [794] ZAREMEHRJERDI H, GANGULY S, RAIRDIN A, et al. Towards Large Reasoning Models for Agriculture[J/OL]. 2025(arXiv:2505.19259). arXiv: 2505.19259 [cs] [2025-12-27]. DOI: 10.48550/arXiv.2505.19259 (引用页: 218, 219).
- [795] ZHOU Y, RYO M. AgriBench: A Hierarchical Agriculture Benchmark for Multimodal Large Language Models[J/OL]. 2024(arXiv:2412.00465). arXiv: 2412.00465 [cs] [2025-12-27]. DOI: 10.48550/arXiv.2412.00465 (引用页: 218, 219).
- [796] PhenoBench: A Large Dataset and Benchmarks for Semantic Image

- Interpretation in the Agricultural Domain[Z]. <https://www.phenobench.org/>. [2025-12-27] (引用页: 218, 219).
- [797] WU J, LAI Z, CHEN S, et al. The new agronomists: Language models are experts in crop management[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 5346-5356 (引用页: 218, 219).
- [798] SAPKOTA R, MENG Z, KARKEE M. Synthetic meets authentic: Leveraging llm generated datasets for yolo11 and yolov10-based apple detection through machine vision sensors[J]. Smart Agricultural Technology, 2024, 9: 100614 (引用页: 218, 219).
- [799] 西北农林科技大学发布国内首个国产算力农业大模型凤凰网陕西 _凤凰网[Z]. <https://sn.ifeng.com/c/8lnjkdswaNy>. [2025-12-27] (引用页: 218, 222).
- [800] SINGH N, WANG'OMBE J, OKANGA N, et al. Farmer.Chat: Scaling AI-Powered Agricultural Services for Smallholder Farmers [J/OL]. 2024(arXiv:2409.08916). arXiv: 2409.08916 [cs] [2025-12-27]. DOI: 10.48550/arXiv.2409.08916 (引用页: 218, 220, 222).
- [801] NGUYEN V, KARIMI S, HALLGREN W, et al. My Climate Advisor: An Application of NLP in Climate Adaptation for Agriculture [C/OL]//STAMMBACH D, NI J, SCHIMANSKI T, et al. Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024). Bangkok, Thailand: Association for Computational Linguistics, 2024: 27-45 [2025-12-27]. DOI: 10.18653/v1/2024.climatenlp-1.3 (引用页: 218, 221).
- [802] 中国农业大学发布神农大模型 3.0[Z]. <https://news.cau.edu.cn/mtndnew/11a011e1b64b4> [2025-12-31] (引用页: 218, 221).
- [803] [光明日报] 我国首个农业通用大语言模型发布[Z].

- [805] CHEN D, HUANG Y. Integrating reinforcement learning and large language models for crop production process management optimization and control through a new knowledge-based deep learning paradigm[J]. Computers and Electronics in Agriculture, 2025, 232: 110028 (引用页: 218, 220).
- [806] YANG S, LIU Z, MAYER W. ShizishanGPT: An Agricultural Large Language Model Integrating Tools and Resources[J/OL]. 2024(arXiv:2409.13537). arXiv: 2409.13537 [cs] [2025-12-31]. DOI: 10.48550/arXiv.2409.13537 (引用页: 218, 221).
- [807] YAN L, WANG H, TANG C, et al. Agrieval: A comprehensive chinese agricultural benchmark for large language models[J]. arXiv preprint arXiv:2507.21773, 2025 (引用页: 218, 220).
- [808] WANG H, GUAN Y, MENG F, et al. Agri-CM3: A Chinese Massive Multi-modal, Multi-level Benchmark for Agricultural Understanding and Reasoning[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 11729-11754 (引用页: 218, 220).
- [809] KUMAR S S, KHAN A K M A, BANDAY I A, et al. Overcoming llm challenges using rag-driven precision in coffee leaf disease remediation[C]//2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS). 2024: 1-6 (引用页: 218, 221).
- [810] ZHANG X. Enhanced Agricultural Financial Services through Cloud Computing: A New Paradigm of Security and Efficiency [J/OL]. Research on World Agricultural Economy, 2024: 555-566 [2025-12-27]. DOI: 10.36956/rwae.v5i4.1315 (引用页: 218, 221).
- [811] [新华社] 农耕大模型 1.0 正式发布[Z]. <https://caas.cn/xwzx/mtxw/0cca81c919f2480794297d25f1b2> [2025-12-27] (引用页: 218, 222).
- [812] Profluent and Corteva Partner to Apply AI-Designed Proteins In Agriculture[Z]. <https://www.synbiobeta.com/read/profluent-and-corteva-partner-to-apply-ai-designed-proteins-in-agriculture>. [2025-12-27] (引用页: 219).

- [813] BAJA H, KALLENBERG M, ATHANASIADIS I N. To Measure or Not: A Cost-Sensitive, Selective Measuring Environment for Agricultural Management Decisions with Reinforcement Learning [C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 39: 27. 2025: 27831-27840 (引用页: 220).
- [814] GONG R, LI X. The application progress and research trends of knowledge graphs and large language models in agriculture[J]. Computers and electronics in agriculture, 2025, 235: 110396 (引用页: 220).
- [815] 北京农业人工智能与机器人研究院在海淀区成立-媒体聚焦-北京市农林科学院[Z]. <https://www.baafs.net.cn/xwzx/mtjj/54cd59923b1c419e8fa2a56d903f89> [2025-12-27] (引用页: 222).
- [816] YI Z, OUYANG J, XU Z, et al. A survey on recent advances in llm-based multi-turn dialogue systems[J]. ACM Computing Surveys, 2024 (引用页: 225).
- [817] DESHPANDE K, SIRDESHMUKH V, MOLS J B, et al. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 18632-18702 (引用页: 225).
- [818] WU D, WANG H, YU W, et al. Longmemeval: Benchmarking chat assistants on long-term interactive memory[J]. arXiv preprint arXiv:2410.10813, 2024 (引用页: 225, 226).
- [819] KATSIKIS Y, ROSENTHAL S, FADNIS K, et al. Mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems[J]. Transactions of the Association for Computational Linguistics, 2025, 13: 784-808 (引用页: 225, 226).
- [820] JIA Q, ZHANG K, SONG X, et al. One Battle After Another: Probing LLMs' Limits on Multi-Turn Instruction Following with a Benchmark Evolving Framework[J]. arXiv preprint arXiv:2511.03508, 2025 (引用页: 225, 226).
- [821] LEE Y J, KIM S, LEE B K, et al. RefineBench: Evaluating Re-

- finement Capability of Language Models via Checklists[J]. arXiv preprint arXiv:2511.22173, 2025 (引用页: 227).
- [822] De OLIVEIRA B L, MARTINS L G, BRANDÃO B, et al. InfoQuest: Evaluating Multi-Turn Dialogue Agents for Open-Ended Conversations with Hidden Context[J]. arXiv preprint arXiv:2502.12257, 2025 (引用页: 227).
- [823] YAO H, HUANG J, QIU Y, et al. MMReason: An Open-Ended Multi-Modal Multi-Step Reasoning Benchmark for MLLMs Toward AGI[J]. arXiv preprint arXiv:2506.23563, 2025 (引用页: 227).
- [824] LÜ X H, KAZEMNEJAD A, MEADE N, et al. Agentrewardbench: Evaluating automatic evaluations of web agent trajectories[J]. arXiv preprint arXiv:2504.08942, 2025 (引用页: 227).
- [825] DESHPANDE K, SIRDESHMUKH V, MOLS J B, et al. MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs[C//Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025: 18632-18702. <https://aclanthology.org/2025.findings-acl.958/> (引用页: 227, 228, 232).
- [826] DESHPANDE K, SIRDESHMUKH V, MOLS J B, et al. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms[C//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 18632-18702 (引用页: 227, 228).
- [827] HE Y, LI W, ZHANG H, et al. Rubric-Based Benchmarking and Reinforcement Learning for Advancing LLM Instruction Following [J]. arXiv preprint arXiv:2511.10507, 2025 (引用页: 227, 228).
- [828] PATIL S G, MAO H, YAN F, et al. The Berkeley Function Calling Leaderboard (BFCL): From Tool Use to Agentic Evaluation of Large Language Models[C//Forty-second International Conference on Machine Learning (引用页: 230).
- [829] ZHONG L, DU Z, ZHANG X, et al. ComplexFuncBench: exploring multi-step and constrained function calling under long-context

- scenario[J]. arXiv preprint arXiv:2501.10132, 2025 (引用页: 230).
- [830] NATH V, RAJA P, YOON C, et al. Toolcomp: A multi-tool reasoning & process supervision benchmark[J]. arXiv preprint arXiv:2501.01290, 2025 (引用页: 230).
- [831] WU Z, LIU X, ZHANG X, et al. MCPMark: A Benchmark for Stress-Testing Realistic and Comprehensive MCP Use[J]. arXiv preprint arXiv:2509.24002, 2025 (引用页: 230).
- [832] WANG Z, CHANG Q, PATEL H, et al. Mcp-bench: Benchmarking tool-using llm agents with complex real-world tasks via mcp servers [J]. arXiv preprint arXiv:2508.20453, 2025 (引用页: 230).
- [833] YAO S, SHINN N, RAZAVI P, et al. *tau*-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains[J]. arXiv preprint arXiv:2406.12045, 2024 (引用页: 231).
- [834] BARRES V, DONG H, RAY S, et al. τ^2 -Bench: Evaluating Conversational Agents in a Dual-Control Environment[J]. arXiv preprint arXiv:2506.07982, 2025 (引用页: 231).
- [835] ZHOU Y, JIANG S, TIAN Y, et al. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks[J]. arXiv preprint arXiv:2503.15478, 2025 (引用页: 231).
- [836] ABHYANKAR R, QI Q, ZHANG Y. OSWorld-Human: Benchmarking the Efficiency of Computer-Use Agents[J]. arXiv preprint arXiv:2506.16042, 2025 (引用页: 231).
- [837] PHAN L, GATTI A, HAN Z, et al. Humanity's last exam[J]. arXiv preprint arXiv:2501.14249, 2025 (引用页: 232).
- [838] Epoch AI. FrontierMath[Z]. <https://epochai.org/benchmarks/frontiermath>. 2025 (引用页: 232).
- [839] LEI F, YANG Y, SUN W, et al. MCPVerse: An Expansive, Real-World Benchmark for Agentic Tool Use[J]. arXiv preprint arXiv:2508.16260, 2025 (引用页: 234).
- [840] Skyvern Team and Halluminate. Web Bench[Z]. <https://webbench.ai/>. 2025 (引用页: 234).

- [841] ZENG L, LOU F, WANG Z, et al. Fingaia: an end-to-end benchmark for evaluating AI agents in finance[J]. arXiv e-prints, 2025:arXiv-2507 (引用页: 236).
- [842] Vals AI. Finance Agent[Z]. https://www.vals.ai/benchmarks/finance_agent. 2025 (引用页: 236).
- [843] AI V. Vals Legal AI Report (VLAIR)[Z]. <https://www.vals.ai/vlair>. 2025 (引用页: 236).
- [844] JIANG Y, BLACK K C, GENG G, et al. MedAgentBench: a virtual EHR environment to benchmark medical LLM agents[J]. NEJM AI, 2025, 2(9): AIdbp2500144 (引用页: 236).
- [845] ARORA R K, WEI J, HICKS R S, et al. Healthbench: Evaluating large language models towards improved human health[J]. arXiv preprint arXiv:2505.08775, 2025 (引用页: 236).
- [846] CHOU J, LIU A, DENG Y, et al. Autocodebench: Large language models are automatic code benchmark generators[J]. arXiv preprint arXiv:2508.09101, 2025 (引用页: 236).
- [847] BRAGG J, D'ARCY M, BALEPUR N, et al. AstaBench: Rigorous Benchmarking of AI Agents with a Scientific Research Suite[J]. arXiv preprint arXiv:2510.21652, 2025 (引用页: 236).
- [848] STARACE G, JAFFE O, SHERBURN D, et al. PaperBench: Evaluating AI's Ability to Replicate AI Research[J]. arXiv preprint arXiv:2504.01848, 2025 (引用页: 236).
- [849] WORNOW M, GARODIA V, VASSALOS V, et al. Top of the CLASS: Benchmarking LLM Agents on Real-World Enterprise Tasks[C]//ICLR 2025 Workshop on Building Trust in Language Models and Applications (引用页: 236).
- [850] GRÖTSCHLA F, MÜLLER L, TÖNSHOFF J, et al. Agentsnet: Coordination and collaborative reasoning in multi-agent llms[J]. arXiv preprint arXiv:2507.08616, 2025 (引用页: 237).
- [851] HYUN J, WAYTOWICH N R, CHEN B. CREW-Wildfire: Benchmarking agentic multi-agent collaborations at scale[J]. arXiv preprint arXiv:2507.05178, 2025 (引用页: 237).

- [852] WANG R, YU H, ZHANG W, et al. SOTOPIA-*pi*: Interactive Learning of Socially Intelligent Language Agents[J]. arXiv preprint arXiv:2403.08715, 2024 (引用页: 237).
- [853] LIU X, YU H, ZHANG H, et al. Agentbench: Evaluating llms as agents[J]. arXiv preprint arXiv:2308.03688, 2023 (引用页: 237).
- [854] ZHOU H, GENG H, XUE X, et al. Reso: A reward-driven self-organizing llm-based multi-agent system for reasoning tasks[J]. arXiv preprint arXiv:2503.02390, 2025 (引用页: 237).
- [855] WANG W, HE Z, HONG W, et al. Lvbench: An extreme long video understanding benchmark[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025: 22958-22967 (引用页: 239).
- [856] TAN X, LUO Y, YE Y, et al. ALLVB: All-in-One Long Video Understanding Benchmark[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 39: 7. 2025: 7211-7219 (引用页: 239).
- [857] FU C, DAI Y, LUO Y, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis [C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 24108-24118 (引用页: 239).
- [858] PAN Y, WANG Z, XIE Q, et al. MT-Video-Bench: A Holistic Video Understanding Benchmark for Evaluating Multimodal LLMs in Multi-Turn Dialogues[J]. arXiv preprint arXiv:2510.17722, 2025 (引用页: 239).
- [859] YU J, WU Y, CHU M, et al. VRBench: A Benchmark for Multi-Step Reasoning in Long Narrative Videos[J]. arXiv preprint arXiv:2506.10857, 2025 (引用页: 239).
- [860] CHENG Z, HU J, LIU Z, et al. V-star: Benchmarking video-llms on video spatio-temporal reasoning[J]. arXiv preprint arXiv:2503.11495, 2025 (引用页: 239).
- [861] OUYANG L, QU Y, ZHOU H, et al. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations

- [C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 24838-24848 (引用页: 240).
- [862] MANDAL S, GUPTA N, TALEWAR A, et al. IDPLLeaderboard: A Unified Leaderboard for Intelligent Document Processing Tasks[J]. 2025 (引用页: 240).
- [863] ZHANG K, NIU L, CAO Z, et al. TIU-Bench: A Benchmark for Evaluating Large Multimodal Models on Text-rich Image Understanding[C]//Findings of the Association for Computational Linguistics: EMNLP 2025. 2025: 24286-24295 (引用页: 240).
- [864] WANG A L, TANG J, LIAO L, et al. Wilddoc: How far are we from achieving comprehensive and robust document understanding in the wild?[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 23002-23012 (引用页: 240).
- [865] HAO Y, GU J, WANG H W, et al. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark [J]. arXiv preprint arXiv:2501.05444, 2025 (引用页: 241).
- [866] YUE X, ZHENG T, NI Y, et al. Mmmu-pro: A more robust multidiscipline multimodal understanding benchmark[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 15134-15186 (引用页: 241).
- [867] WANG P, LI Z Z, YIN F, et al. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 19541-19551 (引用页: 241).
- [868] WANG Z, SUN J, ZHANG W, et al. Benchmarking Multimodal Mathematical Reasoning with Explicit Visual Dependency[J]. arXiv preprint arXiv:2504.18589, 2025 (引用页: 241).
- [869] LIANG H, SUN L, zhouminxuan zhouminxuan Z, et al. MathScape: Benchmarking Multimodal Large Language Models in Real-World

- Mathematical Contexts[C]//Proceedings of the 33rd ACM International Conference on Multimedia. 2025: 12942-12948 (引用页: 241).
- [870] RUAN J, JIANG D, GAO X, et al. Mme-sci: A comprehensive and challenging science benchmark for multimodal large language models[J]. arXiv preprint arXiv:2508.13938, 2025 (引用页: 241).
- [871] MUKHERJEE A, GHOSH S. mmJEE-Eval: A Bilingual Multimodal Benchmark for Evaluating Scientific Reasoning in Vision-Language Models[J]. arXiv preprint arXiv:2511.09339, 2025 (引用页: 241).
- [872] GUAN M Y, JOGLEKAR M, WALLACE E, et al. Deliberative alignment: Reasoning enables safer language models[J]. arXiv preprint arXiv:2412.16339, 2024 (引用页: 278, 280, 324).
- [873] ZHANG Y, ZHANG S, HUANG Y, et al. STAIR: Improving Safety Alignment with Introspective Reasoning[C]//Forty-second International Conference on Machine Learning. 2025 (引用页: 278, 280).
- [874] WANG H, QIN Z, SHEN L, et al. Leveraging reasoning with guidelines to elicit and utilize knowledge for enhancing safety alignment [J]. arXiv preprint arXiv:2502.04040, 2025: 3 (引用页: 278, 280).
- [875] SCHOEN B, NITISHINSKAYA E, BALESNI M, et al. Stress testing deliberative alignment for anti-scheming training[J]. arXiv preprint arXiv:2509.15541, 2025 (引用页: 278).
- [876] ZAREMBA W, NITISHINSKAYA E, BARAK B, et al. Trading inference-time compute for adversarial robustness[J]. arXiv preprint arXiv:2501.18841, 2025 (引用页: 278).
- [877] CHEN R, ARDITI A, SLEIGHT H, et al. Persona vectors: Monitoring and controlling character traits in language models[J]. arXiv preprint arXiv:2507.21509, 2025 (引用页: 279, 280).
- [878] WANG M, la TOUR T D, WATKINS O, et al. Persona features control emergent misalignment[J]. arXiv preprint arXiv:2506.19823, 2025 (引用页: 279, 280).
- [879] ZHAO H, YUAN C, HUANG F, et al. Qwen3guard technical report [J]. arXiv preprint arXiv:2510.14276, 2025 (引用页: 279, 280).

- [880] BAKER B, HUIZINGA J, GAO L, et al. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation[J]. arXiv preprint arXiv:2503.11926, 2025 (引用页: 279, 280).
- [881] JOGLEKAR M, CHEN J, WU G, et al. Training LLMs for Honesty via Confessions[J]. arXiv preprint arXiv:2512.08093, 2025 (引用页: 279, 280).
- [882] JI J, WANG K, QIU T A, et al. Language models resist alignment: Evidence from data compression[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 23411-23432 (引用页: 279).
- [883] QI X, PANDA A, LYU K, et al. Safety alignment should be made more than just a few tokens deep[J]. arXiv preprint arXiv:2406.05946, 2024 (引用页: 279).
- [884] Anthropic. On the Biology of a Large Language Model[Z]. <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>. Transformer Circuits Thread. 2025 (引用页: 279).
- [885] MACDIARMID M, WRIGHT B, UESATO J, et al. Natural Emergent Misalignment from Reward Hacking in Production RL[J]. arXiv preprint arXiv:2511.18397, 2025 (引用页: 280, 281).
- [886] CHEN Y, BENTON J, RADHAKRISHNAN A, et al. Reasoning Models Don't Always Say What They Think[J]. arXiv preprint arXiv:2505.05410, 2025 (引用页: 280, 281).
- [887] ZHANG J, SLEIGHT H, PENG A, et al. Stress-Testing Model Specs Reveals Character Differences among Language Models[J]. arXiv preprint arXiv:2510.07686, 2025 (引用页: 280, 281).
- [888] SOULY A, RANDO J, CHAPMAN E, et al. Poisoning attacks on LLMs require a near-constant number of poison samples[J]. arXiv preprint arXiv:2510.07192, 2025 (引用页: 280, 281).
- [889] TILL D, SMEATON J, HAUBRICK P, et al. Teaming LLMs to Detect and Mitigate Hallucinations[J]. arXiv preprint arXiv:2510.19507, 2025 (引用页: 282).

- [890] MAVROMATIS C, KARYPIS G. Gnn-rag: Graph neural retrieval for efficient large language model reasoning on knowledge graphs[C]//Findings of the Association for Computational Linguistics: ACL 2025. 2025: 16682-16699 (引用页: 282, 284).
- [891] ZHANG D, YANG N, ZHU J, et al. ASCoT: An Adaptive Self-Correction Chain-of-Thought Method for Late-Stage Fragility in LLMs[J]. arXiv preprint arXiv:2508.05282, 2025 (引用页: 282).
- [892] MA R, WANG P, LIU C, et al. S²R: Teaching LLMs to Self-verify and Self-correct via Reinforcement Learning[J]. arXiv preprint arXiv:2502.12853, 2025 (引用页: 282).
- [893] OK C, LEE E, OH D. Synthetic paths to integral truth: Mitigating hallucinations caused by confirmation bias with synthetic data[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025: 5168-5180 (引用页: 282).
- [894] TANG Z, CHATTERJEE R, GARG S. Mitigating hallucinated translations in large language models with hallucination-focused preference optimization[J]. arXiv preprint arXiv:2501.17295, 2025 (引用页: 283).
- [895] ZHANG X, PENG B, TIAN Y, et al. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation[J]. arXiv preprint arXiv:2402.09267, 2024 (引用页: 283).
- [896] FANG Y, LI M, WANG W, et al. Counterfactual debating with pre-set stances for hallucination elimination of LLMs[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025: 10554-10568 (引用页: 284).
- [897] ZHANG Y, LI S, QIAN C, et al. The law of knowledge overshadowing: Towards understanding, predicting, and preventing llm hallucination[J]. arXiv preprint arXiv:2502.16143, 2025 (引用页: 285).
- [898] LIANG T, DU Y, HUANG J, et al. MoLE: Decoding by Mixture of Layer Experts Alleviates Hallucination in Large Vision-Language Models[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 39: 18. 2025: 18684-18692 (引用页: 285).

- [899] YANG D, XIAO D, WEI J, et al. Improving factuality in large language models via decoding-time hallucinatory and truthful comparators[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 39: 24. 2025: 25606-25614 (引用页: 285, 286).
- [900] TAN X, WANG X, LIU Q, et al. Paths-over-graph: Knowledge graph empowered large language model reasoning[C]//Proceedings of the ACM on Web Conference 2025. 2025: 3505-3522 (引用页: 286).
- [901] HU Y, ZHU J, TANG L, et al. ReMindRAG: Low-Cost LLM-Guided Knowledge Graph Traversal for Efficient RAG[J]. arXiv preprint arXiv:2510.13193, 2025 (引用页: 286).
- [902] HAO Y, WU D. Fact Verification on Knowledge Graph via Programmatic Graph Reasoning[C]//Findings of the Association for Computational Linguistics: EMNLP 2025. 2025: 5480-5495 (引用页: 286).
- [903] HUANG L, FENG X, MA W, et al. Alleviating Hallucinations from Knowledge Misalignment in Large Language Models via Selective Abstention Learning[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 24564-24579 (引用页: 286).
- [904] JIANG X, ZHANG R, XU Y, et al. Hykge: A hypothesis knowledge graph enhanced rag framework for accurate and reliable medical llms responses[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 11836-11856 (引用页: 286).
- [905] ZHANG Z, ZHANG Z, ZHANG J, et al. MPVStance: Mitigating Hallucinations in Stance Detection with Multi-Perspective Verification[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 1053-1067 (引用页: 286).
- [906] WANG Z, GU T, WU B, et al. MorphMark: Flexible Adaptive Watermarking for Large Language Models[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 4842-4860 (引用页: 288, 290).

- [907] NIESS G, KERN R. Ensemble Watermarks for Large Language Models[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 2903-2916 (引用页: 288, 290).
- [908] ZHAO X, LIAO C, WANG Y, et al. Efficiently Identifying Watermarked Segments in Mixed-Source Texts[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 6304-6316 (引用页: 288).
- [909] MAO M, WEI D, CHEN Z, et al. Watermarking Large Language Models: An Unbiased and Low-risk Method[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 7939-7960 (引用页: 288-290).
- [910] LIU A, GUAN S, LIU Y, et al. Can Watermarked LLMs be Identified by Users via Crafted Prompts?[C]//The Thirteenth International Conference on Learning Representations. 2025 (引用页: 289, 290).
- [911] PENG H, QI Y, WANG X, et al. Agentic Reward Modeling: Integrating Human Preferences with Verifiable Correctness Signals for Reliable Reward Systems[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 15934-15949 (引用页: 291, 293).
- [912] FENG T, QU L, TANDON N, et al. IRIS: An Iterative and Integrated Framework for Verifiable Causal Discovery in the Absence of Tabular Data[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 9400-9428 (引用页: 291, 293).
- [913] CAO J, LU Y, LI M, et al. From Informal to Formal - Incorporating and Evaluating LLMs on Natural Language Requirements to Verifiable Formal Proofs[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 26984-27003 (引用页: 292, 293).
- [914] ZHU J, XIAO M, WANG Y, et al. TROVE: A Challenge for Fine-

- Grained Text Provenance via Source Sentence Tracing and Relationship Classification[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 11755-11771 (引用页: 294).
- [915] PETERKA T, BOHACEK M. Dataset of News Articles with Provenance Metadata for Media Relevance Assessment[J]. arXiv preprint arXiv:2506.09847, 2025 (引用页: 294, 295).
- [916] VASU R, BASU C, MISHRA B D, et al. HypER: Literature-grounded Hypothesis Generation and Distillation with Provenance [J]. arXiv preprint arXiv:2506.12937, 2025 (引用页: 294, 296).
- [917] XU Y, LIU A, HU X, et al. Mark Your LLM: Detecting the Misuse of Open-Source Large Language Models via Watermarking[J]. arXiv preprint arXiv:2503.04636, 2025 (引用页: 294, 296).
- [918] MICHEL G, EPURE E V, HENNEQUIN R, et al. Evaluating LLMs for Quotation Attribution in Literary Texts: A Case Study of LLaMa3[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. 2025: 742-755 (引用页: 295, 297).
- [919] NAJJAR A, ASHQAR H I, DARWISH O A, et al. Leveraging Explainable AI for LLM Text Attribution: Differentiating Human-Written and Multiple LLMs-Generated Text[J]. arXiv preprint arXiv:2501.03212, 2025 (引用页: 295, 297).
- [920] RAO Z, MOHAMED Y, LIU S, et al. Two Birds with One Stone: Multi-Task Detection and Attribution of LLM-Generated Text[J]. arXiv preprint arXiv:2508.14190, 2025 (引用页: 295, 298).
- [921] SHILOV I, CLOUD A, GEMA A P, et al. Beyond Data Filtering: Knowledge Localization for Capability Removal in LLMs[J]. arXiv preprint arXiv:2512.05648, 2025 (引用页: 295, 298).
- [922] PATHADE C. Invisible Injections: Exploiting Vision-Language Models Through Steganographic Prompt Embedding[EB/OL]. 2025. <https://arxiv.org/abs/2507.22304>. arXiv: 2507.22304 [cs.CR]

(引用页: 300).

- [923] JIANG Y, GAO X, PENG T, et al. HiddenDetect: Detecting Jailbreak Attacks against Large Vision-Language Models via Monitoring Hidden States[J/OL]. 2025. ACL 2025: 2502.14744 (cs.CL). <https://arxiv.org/abs/2502.14744> (引用页: 300).
- [924] WANG Z, WANG H, TIAN C, et al. Implicit Jailbreak Attacks via Cross-Modal Information Concealment on Vision-Language Models [J/OL]. 2025. arXiv: 2505.16446 [cs.LG]. <https://arxiv.org/abs/2505.16446> (引用页: 301).
- [925] SUN G, ZHAN X, FENG S, et al. CASE-Bench: Context-Aware SafeTy Benchmark for Large Language Models[C/OL]//Forty-second International Conference on Machine Learning. 2025. <https://openreview.net/forum?id=FCHIGDCoow> (引用页: 301).
- [926] XING W, LI M, HU C, et al. Latent Fusion Jailbreak: Blending Harmful and Harmless Representations to Elicit Unsafe LLM Outputs[J/OL]. 2025. arXiv: 2508.10029 [cs.CL]. <https://arxiv.org/abs/2508.10029> (引用页: 301).
- [927] KADALI S D S S, PAPALEXAKIS E E. Do Internal Layers of LLMs Reveal Patterns for Jailbreak Detection?[J/OL]. 2025. arXiv: 2510.06594 [cs.CL]. <https://arxiv.org/abs/2510.06594> (引用页: 302).
- [928] JACOB D, ALZAHRANI H, HU Z, et al. PromptShield: Deployable Detection for Prompt Injection Attacks[J/OL]. 2025. arXiv: 2501.15145 [cs.CR]. <https://arxiv.org/abs/2501.15145> (引用页: 302).
- [929] XIANG S, ZHANG A, CAO Y, et al. Beyond Surface-Level Patterns: An Essence-Driven Defense Framework Against Jailbreak Attacks in LLMs[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025: 14727-14742. <https://aclanthology.org/2025.findings-acl.760/>. DOI: 10.18653/v1/2025.findings-acl.760 (引用页: 302).
- [930] ZHANG S, ZHAI Y, GUO K, et al. JBShield: Defending Large Language Models from Jailbreak Attacks through Activated Concept

- Analysis and Manipulation[J]. 2025 (引用页: 302).
- [931] HIDAYAT N S, KAUTSAR M D A, WICAKSONO A F, et al. Simulating Training Data Leakage in Multiple-Choice Benchmarks for LLM Evaluation[J/OL]. 2025. arXiv: 2505.24263 [cs.CL]. <https://arxiv.org/abs/2505.24263> (引用页: 304).
 - [932] CHOI H K, KHANOV M, WEI H, et al. How Contaminated Is Your Benchmark? Quantifying Dataset Leakage in Large Language Models with Kernel Divergence. 2025 (引用页: 304).
 - [933] FANG Y, SUN T, SHI Y, et al. LastingBench: Defend Benchmarks Against Knowledge Leakage[C/OL]//Findings of the Association for Computational Linguistics: EMNLP 2025. Suzhou, China: Association for Computational Linguistics, 2025: 18304-18317. <https://aclanthology.org/2025.findings-emnlp.993/>. DOI: 10.18653/v1/2025.findings-emnlp.993 (引用页: 304).
 - [934] WU X, PAN L, XIE Y, et al. AntiLeakBench: Preventing Data Contamination by Automatically Constructing Benchmarks with Updated Real-World Knowledge[C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 18403-18419. <https://aclanthology.org/2025.acl-long.901/>. DOI: 10.18653/v1/2025.acl-long.901 (引用页: 304).
 - [935] APERTUS P, HERNÁNDEZ-CANO A, HÄGELE A, et al. Aper-tus: Democratizing Open and Compliant LLMs for Global Language Environments[J/OL]. 2025. arXiv: 2509.14233 [cs.CL]. <https://arxiv.org/abs/2509.14233> (引用页: 304).
 - [936] CLARK O, JOSHI K P. Real-Time Detection of Online Health Misinformation using an Integrated Knowledgegraph-LLM Approach[C]//2025 IEEE International Conference on Digital Health (ICDH). 2025: 111-120 (引用页: 304).
 - [937] ZHANG K, CHENG S, GUO H, et al. SOFT: Selective Data Ob-fuscation for Protecting LLM Fine-tuning against Membership In-

- ference Attacks[C]//34th USENIX Security Symposium (USENIX Security 25). Seattle, WA: USENIX Association, 2025 (引用页: 305).
- [938] PANEBIANCO F, BONFANTI S, TROVÒ F, et al. LeakSealer: A Semisupervised Defense for LLMs Against Prompt Injection and Leakage Attacks[J/OL]. 2025. arXiv: 2508.00602 [cs.CR]. <https://arxiv.org/abs/2508.00602> (引用页: 305).
- [939] AKETI S A, BULLOCK W, KALEMAJ I, et al. Scaling Private Deep Learning with Opacus: Advances for Large Language Models [C]//Proceedings of the 42 nd International Conference on Machine Learning. 2025 (引用页: 305).
- [940] MAKNI M, BEHDIN K, AFRIAT G, et al. SPARTA: An Optimization Framework for Differentially Private Sparse Fine-Tuning [C/OL]//KDD '25: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2. New York, NY, USA: Association for Computing Machinery, 2025: 2090-2101. <https://doi.org/10.1145/3711896.3736842>. DOI: 10.1145/3711896.3736842 (引用页: 305).
- [941] LI Y, FU L, WANG T, et al. Clients Collaborate: Flexible Differentially Private Federated Learning with Guaranteed Improvement of Utility-Privacy Trade-off[C/OL]//Forty-second International Conference on Machine Learning. 2025. <https://openreview.net/forum?id=C7dmhyTDrx> (引用页: 306).
- [942] GHIASVAND S, YANG Y, XUE Z, et al. Communication-Efficient and Tensorized Federated Fine-Tuning of Large Language Models [C/OL]//CHE W, NABENDE J, SHUTOVA E, et al. Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025: 24192-24207. <https://aclanthology.org/2025.findings-acl.1241/>. DOI: 10.18653/v1/2025.findings-acl.1241 (引用页: 306).
- [943] LIU H, WEN R, NAIR S, et al. ECOLORA: Communication-Efficient Federated Fine-Tuning of Large Language Models[C]//Proceedings of the 2025 Conference on Empirical Methods in Nat-

- ural Language Processing. 2025 (引用页: 306).
- [944] WU S, JIA Y, XIANG H, et al. A Fair Federated Learning Method for Handling Client Participation Probability Inconsistencies in Heterogeneous Environments[C]//39th Conference on Neural Information Processing Systems (NeurIPS 2025). 2025 (引用页: 306).
 - [945] MIA M J, AMINI M H. FedShield-LLM: A Secure and Scalable Federated Fine-Tuned Large Language Model[J]. arXiv preprint arXiv:2506.05640, 2025 (引用页: 306).
 - [946] DOHMEN H, HUNDT R, KHAYATA N, et al. SEEC: Memory Safety Meets Efficiency in Secure Two-Party Computation[C]//ASIA CCS '25: Proceedings of the 20th ACM Asia Conference on Computer and Communications Security. 2025 (引用页: 307).
 - [947] LI Z, XING C, YAO Y, et al. Efficient Pseudorandom Correlation Generators for Any Finite Field[C]//FEHR S, FOUQUE P A. Advances in Cryptology – EUROCRYPT 2025. Cham: Springer Nature Switzerland, 2025: 145-175 (引用页: 307).
 - [948] LEE T Y, PARK S, JEON M, et al. ESC: Erasing Space Concept for Knowledge Deletion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2025: 5010-5019 (引用页: 307).
 - [949] KHALIL Y, SETAYESH M, LI H. CoUn: Empowering Machine Unlearning via Contrastive Learning[C]//NeurIPS. 2025 (引用页: 307).
 - [950] PATHAK S, SHRESHTHA S, SINGH R, et al. Quantum-Inspired Audio Unlearning: Towards Privacy-Preserving Voice Biometrics [J/OL]. 2025. arXiv: 2507.22208 [cs.SD]. <https://arxiv.org/abs/2507.22208> (引用页: 308).
 - [951] BAI Y, KADAVATH S, KUNDU S, et al. Constitutional AI: Harmlessness from AI Feedback[J/OL]. 2022. arXiv: 2212.08073 [cs.CL]. <https://arxiv.org/abs/2212.08073> (引用页: 310).
 - [952] KYRYCHENKO Y, ZHOU K, BOGUCKA E, et al. C3ai: Crafting and evaluating constitutions for constitutional ai[C]//Proceedings

- of the ACM on Web Conference 2025. 2025: 3204-3218 (引用页: 311, 312).
- [953] HENNEKING C L, BEGER C. Decoding Human Preferences in Alignment: An Improved Approach to Inverse Constitutional AI[J]. arXiv preprint arXiv:2501.17112, 2025 (引用页: 311, 312).
- [954] ZHANG X. Constitution or Collapse? Exploring Constitutional AI with Llama 3-8B[J]. arXiv preprint arXiv:2504.04918, 2025 (引用页: 311, 312).
- [955] MENKE A G C, TAN P X. How Effective Is Constitutional AI in Small LLMs? A Study on DeepSeek-R1 and Its Peers[J]. arXiv preprint arXiv:2503.17365, 2025 (引用页: 311, 312).
- [956] NOH E, BAEK J. Toward Responsible Federated Large Language Models: Leveraging a Safety Filter and Constitutional AI[J]. arXiv preprint arXiv:2502.16691, 2025 (引用页: 311, 313).
- [957] SHARMA M, TONG M, MU J, et al. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming[J]. arXiv preprint arXiv:2501.18837, 2025 (引用页: 311, 313).
- [958] MAIYA S, BARTSCH H, LAMBERT N, et al. Open Character Training: Shaping the Persona of AI Assistants through Constitutional AI[J]. arXiv preprint arXiv:2511.01689, 2025 (引用页: 311, 313).
- [959] LYU C, SONG Y, ZHANG P, et al. Domain-Specific Constitutional AI: Enhancing Safety in LLM-Powered Mental Health Chatbots[J]. arXiv preprint arXiv:2509.16444, 2025 (引用页: 311, 313).
- [960] HENDRYCKS D, SONG D, SZEGEDY C, et al. A Definition of AGI[J/OL]. 2025. arXiv: 2510.18212 [cs.AI]. <https://arxiv.org/abs/2510.18212> (引用页: 315).
- [961] ZHANG J, JUAN D C, RASHTCHIAN C, et al. Sled: Self logits evolution decoding for improving factuality in large language models [J]. Advances in Neural Information Processing Systems, 2024, 37: 5188-5209 (引用页: 315).

- [962] TEAM G R, ABEYRUWAN S, AINSLIE J, et al. Gemini Robotics: Bringing AI into the Physical World[J/OL]. 2025. arXiv: 2503.20020 [cs.R0]. <https://arxiv.org/abs/2503.20020> (引用页: 315, 321).
- [963] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling Laws for Neural Language Models[J/OL]. 2020. arXiv: 2001.08361 [cs.LG]. <https://arxiv.org/abs/2001.08361> (引用页: 315).
- [964] PLAAT A, WONG A, VERBERNE S, et al. Multi-Step Reasoning with Large Language Models, a Survey[J/OL]. 2025. arXiv: 2407.11511 [cs.AI]. <https://arxiv.org/abs/2407.11511> (引用页: 316).
- [965] LIANG Z, XU Y, HONG Y, et al. A survey of multimodel large language models[C]//Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering. 2024: 405-409 (引用页: 316).
- [966] JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM computing surveys, 2023, 55(12): 1-38 (引用页: 316).
- [967] LIU H, XUE W, CHEN Y, et al. A Survey on Hallucination in Large Vision-Language Models[J/OL]. 2024. arXiv: 2402.00253 [cs.CV]. <https://arxiv.org/abs/2402.00253> (引用页: 316).
- [968] FAN W, DING Y, NING L, et al. A survey on rag meeting llms: Towards retrieval-augmented large language models[C]//Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining. 2024: 6491-6501 (引用页: 317).
- [969] NADĂȘ M, DIOȘAN L, TOMESCU A. Synthetic data generation using large language models: Advances in text and code[J]. IEEE Access, 2025 (引用页: 317).
- [970] LUO J, WU B, LUO X, et al. A Survey on Efficient Large Language Model Training: From Data-centric Perspectives[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 30904-30920 (引用页: 317).
- [971] SUN Y, LI Z, ZHANG Y, et al. Efficient attention mechanisms for large language models: A survey[J]. arXiv preprint

- arXiv:2507.19595, 2025 (引用页: 317).
- [972] GARCEZ A D, LAMB L C. Neurosymbolic ai: The 3 rd wave[J]. Artificial Intelligence Review, 2023, 56(11): 12387-12406 (引用页: 317).
- [973] WU F, SHEN T, BÄCK T, et al. Knowledge-empowered, collaborative, and co-evolving AI models: The post-LLM roadmap[J]. Engineering, 2025, 44: 87-100 (引用页: 317, 318).
- [974] WANG F, CHEN J, YANG S, et al. A Survey on Collaborating Small and Large Language Models for Performance, Cost-effectiveness, Cloud-edge Privacy, and Trustworthiness[J/OL]. 2025. arXiv: 2510.13890 [cs.CL]. <https://arxiv.org/abs/2510.13890> (引用页: 318).
- [975] PEARL J. Causality[M]. Cambridge university press, 2009 (引用页: 318).
- [976] KICIMAN E, NESS R, SHARMA A, et al. Causal reasoning and large language models: Opening a new frontier for causality[J]. Transactions on Machine Learning Research, 2023 (引用页: 318).
- [977] HA D, SCHMIDHUBER J. World models[J]. arXiv preprint arXiv:1803.10122, 2018, 2(3) (引用页: 318).
- [978] TEAM G R, ABDOLMALEKI A, ABEYRUWAN S, et al. Gemini Robotics 1.5: Pushing the Frontier of Generalist Robots with Advanced Embodied Reasoning, Thinking, and Motion Transfer [J/OL]. 2025. arXiv: 2510.03342 [cs.R0]. <https://arxiv.org/abs/2510.03342> (引用页: 319).
- [979] SHI H, XU Z, WANG H, et al. Continual learning of large language models: A comprehensive survey[J]. ACM Computing Surveys, 2025, 58(5): 1-42 (引用页: 319).
- [980] ZHENG J, QIU S, SHI C, et al. Towards lifelong learning of large language models: A survey[J]. ACM Computing Surveys, 2025, 57(8): 1-35 (引用页: 319).
- [981] ZHENG J, SHI C, CAI X, et al. Lifelong Learning of Large Language Model based Agents: A Roadmap[J/OL]. 2025. arXiv: 2501.07278 [cs.AI]. <https://arxiv.org/abs/2501.07278> (引用页: 319).

- [982] YANG Y, ZHOU J, DING X, et al. Recent advances of foundation language models-based continual learning: A survey[J]. ACM Computing Surveys, 2025, 57(5): 1-38 (引用页: 319).
- [983] WANG L, MA C, FENG X, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18(6): 186345 (引用页: 319).
- [984] PLAAT A, van DUIJN M, van STEIN N, et al. Agentic Large Language Models, a survey[J/OL]. 2025. arXiv: 2503.23037 [cs.AI]. <https://arxiv.org/abs/2503.23037> (引用页: 319).
- [985] XI Z, CHEN W, GUO X, et al. The rise and potential of large language model based agents: A survey[J]. Science China Information Sciences, 2025, 68(2): 121101 (引用页: 319).
- [986] LIU X, YU H, ZHANG H, et al. AgentBench: Evaluating LLMs as Agents[C]//ICLR. 2024 (引用页: 319).
- [987] MIALON G, FOURRIER C, WOLF T, et al. Gaia: a benchmark for general ai assistants[C]//The Twelfth International Conference on Learning Representations. 2023 (引用页: 319).
- [988] WANG C, LIU Y, BI B, et al. Safety in Large Reasoning Models: A Survey[J/OL]. 2025. arXiv: 2504.17704 [cs.CL]. <https://arxiv.org/abs/2504.17704> (引用页: 319).
- [989] EGER S, CAO Y, D'SOUZA J, et al. Transforming Science with Large Language Models: A Survey on AI-assisted Scientific Discovery, Experimentation, Content Generation, and Evaluation[J/OL]. 2025. arXiv: 2502.05151 [cs.CL]. <https://arxiv.org/abs/2502.05151> (引用页: 319).
- [990] TOBIAS A V, WAHAB A. Autonomous 'self-driving' laboratories: a review of technology and policy implications[J]. Royal Society Open Science, 2025, 12(7): 250646 (引用页: 320).
- [991] HYSMITH H, FOADIAN E, PADHY S P, et al. The future of self-driving laboratories: from human in the loop interactive AI to gamification[J]. Digital Discovery, 2024, 3(4): 621-636 (引用页: 320).

- [992] LIU F, HAN J, LYU T, et al. Foundation Models for Scientific Discovery: From Paradigm Enhancement to Paradigm Transition [J]. Authorea Preprints, 2025 (引用页: 320).
- [993] GONG R, DING Y, WANG Z, et al. A survey of low-bit large language models: Basics, systems, and algorithms[J]. Neural networks, 2025: 107856 (引用页: 320).
- [994] LI S, WANG H, XU W, et al. Collaborative Inference and Learning between Edge SLMs and Cloud LLMs: A Survey of Algorithms, Execution, and Open Challenges[J/OL]. 2025. arXiv: 2507.16731 [cs.DC]. <https://arxiv.org/abs/2507.16731> (引用页: 321).
- [995] BESIROGLU T, BERGERSON S A, MICHAEL A, et al. The compute divide in machine learning: A threat to academic contribution and scrutiny?[J]. arXiv preprint arXiv:2401.02452, 2024 (引用页: 322).
- [996] HOFFMANN J, BERGEAUD S, MENSCH A, et al. Training compute-optimal large language models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. 2022: 30016-30030 (引用页: 323).
- [997] HU E J, yelong Shen, WALLIS P, et al. LoRA: Low-Rank Adaptation of Large Language Models[C/OL]//International Conference on Learning Representations. 2022. <https://openreview.net/forum?id=nZeVKeeFYf9> (引用页: 323).
- [998] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. Qlora: Efficient finetuning of quantized llms[J]. Advances in neural information processing systems, 2023, 36: 10088-10115 (引用页: 323).
- [999] DAO T, FU D, ERMON S, et al. Flashattention: Fast and memory-efficient exact attention with io-awareness[J]. Advances in neural information processing systems, 2022, 35: 16344-16359 (引用页: 323).
- [1000] RAJBHANDARI S, RASLEY J, RUWASE O, et al. Zero: Memory optimizations toward training trillion parameter models[C]//SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. 2020: 1-16 (引用页: 323).

- [1001] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity[J]. Journal of Machine Learning Research, 2022, 23(120):1-39 (引用页: 323).
- [1002] 邓永恒, 章晋睿, 岳晟, 等. 迈向泛在智能: 端侧大语言模型现状与展望[J/OL]. 中国科学基金, 2025, 39(02): 263-273. DOI: 10.16262/j.cnki.1000-8217.20250320.004 (引用页: 323).
- [1003] 中国信息通信研究院. 高质量大模型基础设施研究报告 (2024) [J]. 2025 (引用页: 323).
- [1004] SCHWARTZ R, DODGE J, SMITH N A, et al. Green ai[J]. Communications of the ACM, 2020, 63(12): 54-63 (引用页: 323).
- [1005] AI N. Artificial intelligence risk management framework: Generative artificial intelligence profile[J]. NIST Trustworthy and Responsible AI Gaithersburg, MD, USA, 2024 (引用页: 324).
- [1006] CANCELA-OUTEDA C. The EU's AI act: A framework for collaborative governance[J]. Internet of Things, 2024, 27: 101291 (引用页: 324).
- [1007] TEDESCHI S, FRIEDRICH F, SCHRAMOWSKI P, et al. ALERT: A Comprehensive Benchmark for Assessing Large Language Models' Safety through Red Teaming[J]. arXiv preprint arXiv:2404.08676, 2024 (引用页: 325).
- [1008] MOU Y, ZHANG S, YE W. Sg-bench: Evaluating llm safety generalization across diverse tasks and prompt types[J]. Advances in Neural Information Processing Systems, 2024, 37: 123032-123054 (引用页: 325).
- [1009] DONG J, GUO S, WANG H, et al. SafeSearch: Automated Red-Teaming for the Safety of LLM-Based Search Agents[J]. arXiv preprint arXiv:2509.23694, 2025 (引用页: 325).
- [1010] YI S, LIU Y, SUN Z, et al. Jailbreak attacks and defenses against large language models: A survey[J]. arXiv preprint arXiv:2407.04295, 2024 (引用页: 325).

- [1011] HUBINGER E, DENISON C, MU J, et al. Sleeper agents: Training deceptive llms that persist through safety training[J]. arXiv preprint arXiv:2401.05566, 2024 (引用页: 325).
- [1012] PAETH K, ATHERTON D, PITTARAS N, et al. Lessons for editors of AI incidents from the AI incident database[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 39: 28. 2025: 28946-28953 (引用页: 326).
- [1013] CARLINI N, TRAMER F, WALLACE E, et al. Extracting training data from large language models[C]//30th USENIX security symposium (USENIX Security 21). 2021: 2633-2650 (引用页: 326).
- [1014] DATHATHRI S, SEE A, GHASISAS S, et al. Scalable watermarking for identifying large language model outputs[J]. Nature, 2024, 634(8035): 818-823 (引用页: 326).
- [1015] CREDENTIALS C. C2PA Technical Specification. 2.0[J]. Coalition for Content Provenance and Authenticity, 2024 (引用页: 326).