

BiliGo V3 Ultra: 面向多平台的 AI 私信自动回复系统

炽阳 001

2026 年 9 月 4 日

摘要—BiliGo V3 Ultra 是一套面向中国主流社交平台的多平台私信自动回复系统。本报告给出一套初步、可复现的基准评测，重点考察两个系统级维度：消息链路性能与 AI 能力完整度。延迟评测将一次回复过程划分为 T0（用户发送至系统检测到消息）、T1（系统处理）和 T2（系统发起发送至用户端可见）三个阶段。针对哔哩哔哩私信、抖音、小红书、微博和闲鱼分别完成 5 轮重复测量；每个阶段剔除最大值和最小值后，对剩余 3 个观测值取平均。在关闭 AI、启用关键词/默认回复的模式下，BiliGo 五平台端到端平均延迟为 1.982 s，其中哔哩哔哩私信为 0.723 s。成功率实验共处理 440 条消息，成功回复 418 条，加权成功率为 95.0%。补充 AI 实验接入 DeepSeek 官方 V4 Flash API，五平台端到端平均延迟为 2.514 s；鉴于 AI 服务商延迟具有外部依赖，本报告未将该结果纳入核心系统排名。SaleSmartly、respond.io、Intercom、Zendesk 和 Gorgias 的公开数据仅作为行业背景参考，其指标边界与本次 BiliGo 受控实验存在差异。现有证据支持 BiliGo 在测试环境中具备具有竞争力的系统响应性能；正式 SOTA 结论需等待后续统一条件下的竞品对照实验。

关键词—全渠道消息；私信自动回复；AI Agent；RAG；延迟基准；哔哩哔哩；抖音；小红书；微博；闲鱼

1 引言

创作者与客服 workflow 日益分散在多个社交平台，带来了重复操作以及不同平台之间不一致的回复延迟。BiliGo 最初以开源的哔哩哔哩私信与评论自动回复系统形式发布，随后演进至本报告所评测的 V3 Ultra 架构。公开仓库记录了项目早期基于规则的自动化基础、本地部署方式、可配置消息监测间隔以及消息回复控制功能。[1]

V3 Ultra 的首版基准评测有意保持较窄范围。第一部分评测 BiliGo 能直接控制的链路，即消息被检测、处理并送达的速度；第二部分将 AI 支持主要作为功能能力层进行评估，暂不进行纯粹的延迟竞赛。不同用户可能接入不同的 AI 厂商、中转网关、区域节点和模型，这些外部变量往往会显著影响生成耗时，与消息系统核心链路的关联有限。

因此，本报告将 BiliGo 的受控实验数据与厂商公开的行业指标分开展示。SaleSmartly 当前公开宣称平均响应时间低于 6 s；respond.io 文档给出的 AI Agent 典型处理时间为 10 - 15 s，并指出部分响应可能超过 20 s。这些数据具有较高的行业参考价值，但本报告不将其视为统一测试条件下的直接竞品结论。[4], [5]

2 系统范围与基准设计

2.1 评测平台

实验覆盖五个私信渠道：哔哩哔哩私信、抖音、小红书、微博和闲鱼。哔哩哔哩评论属于独立的互动类别，不在本次私信基准测试范围内。

2.2 三阶段延迟模型

每次消息事务被划分为三个阶段。T0 表示从用户完成发送操作到 BiliGo 接收或检测到消息的耗时；T1 表示从系统收到消息到完成回复生成或回复任务准备的系统侧处理耗时；T2 表示从系统发起发送操作到回复在用户端实际可见的耗时。端到端延迟定义为 $T_{total} = T0 + T1 + T2$ 。

针对每个平台的每个阶段均采集 5 次测量值。为降低单次异常高值或低值对结果的影响，本报告采用 20% 对称截尾均值：当 $n = 5$ 时， $t_{\text{trim}} = (\sum t_i - \max(t_i) - \min(t_i)) / 3$ 。原始测量记录完整保留于附录 A 与附录 B。

2.3 测试配置

平台	消息监测间隔	发送间隔 / 等待	其他配置间隔
哔哩哔哩私信	0.05 s	1.0 s	300 s 自动重启
抖音	0.5 s	0.5 s	-
小红书	1.0 s	1.0 s	-
微博	1.0 s	1.0 s	-
闲鱼	2.0 s	2.0 s	-

表 1 实验环境中的平台配置间隔

由于不同平台采用的消息监测与发送间隔存在差异，本排名反映 BiliGo 在当前实际配置下的表现；这些结果不用于评价各平台本身的固有性能。

3 消息链路性能结果

3.1 分阶段延迟

排名	平台	T0 (s)	T1 (s)	T2 (s)	端到端 (s)
1	哔哩哔哩私信	0.450	0.040	0.233	0.723
2	小红书	0.530	0.633	0.137	1.300
3	闲鱼	0.513	0.600	0.460	1.573
4	微博	0.893	0.210	0.640	1.743
5	抖音	0.590	3.500	0.480	4.570
-	跨平台平均	0.595	0.997	0.390	1.982

表 2 非 AI 模式截尾均值延迟与平台排名：端到端延迟越低，排名越高

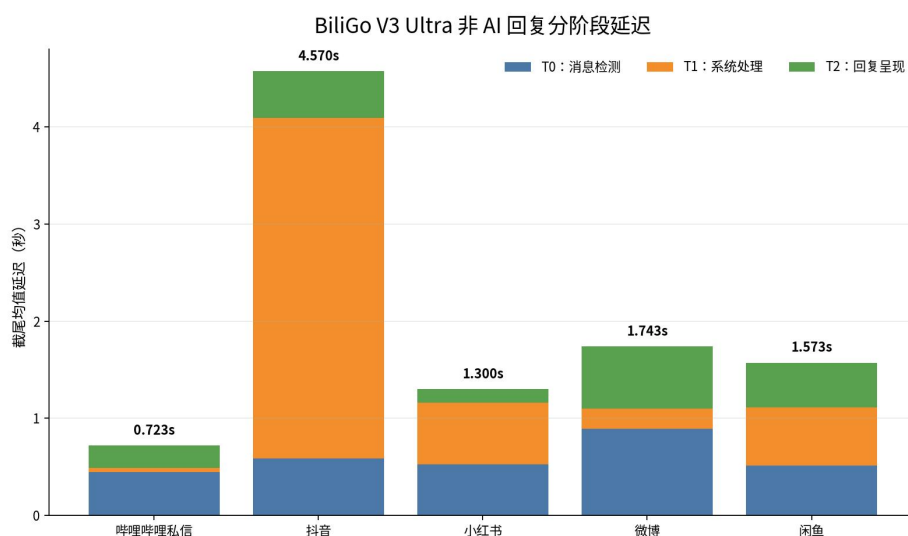


图 1 非 AI 关键词/默认回复模式的分阶段延迟。每个阶段均由 5 次观测值计算截尾均值。

五个平台的平均端到端延迟为 1.982 s。哔哩哔哩私信的端到端截尾均值最低，为 0.723 s；随后依次为小红书 1.300 s、闲鱼 1.573 s、微博 1.743 s、抖音 4.570 s。按平台平均后，T0 为 0.595 s，

T1 为 0.997 s, T2 为 0.390 s。T1 约占跨平台平均总延迟的 50.3%，在当前配置下构成最大的延迟来源。

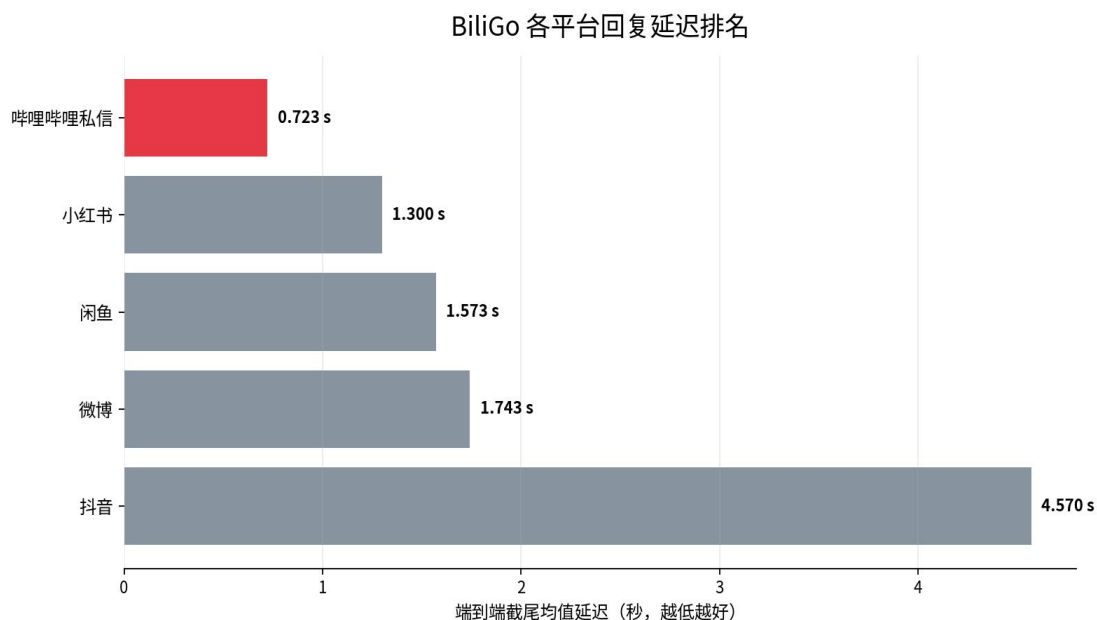


图 2 BiliGo 各平台端到端延迟排名。红色用于突出本次实验中延迟最低的平台配置。

3.2 观测回复成功率

排名	平台	处理消息	成功	失败	成功率	95% Wilson CI
1	小红书	109	109	0	100.00%	96.60% – 100.00%
2	微博	90	90	0	100.00%	95.91% – 100.00%
3	闲鱼	41	39	2	95.12%	83.86% – 98.65%
4	抖音	73	67	6	91.78%	83.21% – 96.18%
5	哔哩哔哩私信	127	113	14	88.98%	82.34% – 93.32%
-	总体加权	440	418	22	95.00%	92.55% – 96.68%

表 3 各平台观测回复成功率。置信区间采用 Wilson score 方法计算。

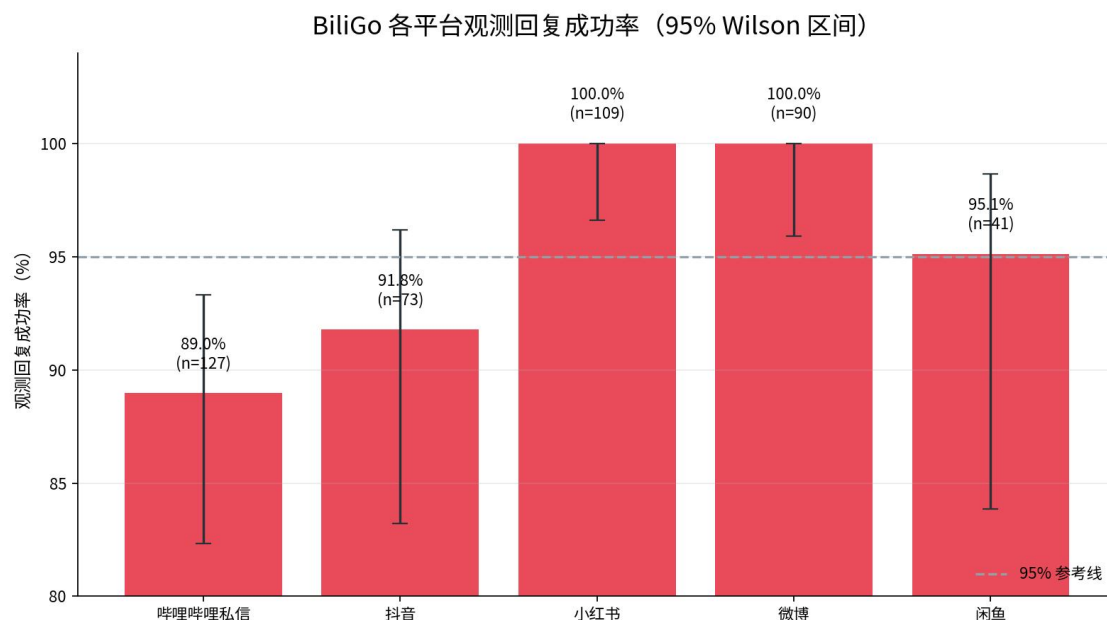


图 3 各平台观测回复成功率及 95% Wilson 置信区间。红色用于突出 BiliGo 基准测试结果。

成功率实验共处理 440 条消息，其中 418 条成功回复，对应 95.0% 的总体加权成功率（95% Wilson CI: 92.55% - 96.68%）。小红书和微博在各自样本中均达到 100% 的观测成功率；闲鱼为 95.12%，抖音为 91.78%，哔哩哔哩私信为 88.98%。五个平台成功率的宏平均值为 95.18%。鉴于样本规模有限，且平台策略可能随时间变化，这组数据更适合作为特定时间点的可靠性快照，不应被解释为长期服务等级保证。

4 AI 回复模式与能力评测

4.1 AI 延迟：补充实验

AI 回复实验接入 DeepSeek 官方 API (<https://api.deepseek.com>)，使用 DeepSeek V4 Flash 模型系列。DeepSeek 当前 API 文档将可调用模型标识为 deepseek-v4-flash，当前模型版本为 DeepSeek-V4-Flash-0731。[3] 每轮实验发送的提示词均为中文“你好”，并关闭上下文模式。

补充实验：AI 回复模式延迟（DeepSeek V4 Flash API）

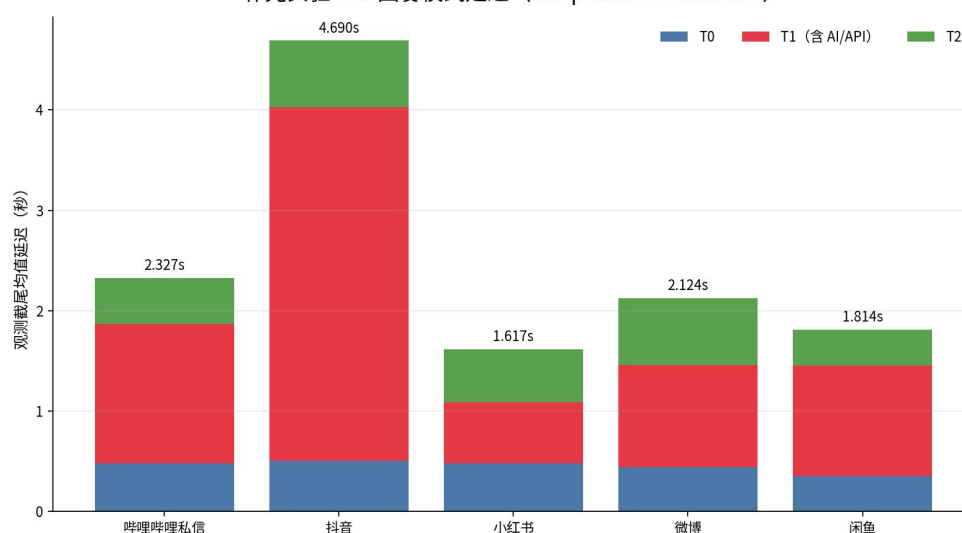


图 4 AI 回复模式补充延迟实验。T1 包含外部 AI/API 耗时，因此不作为 BiliGo 核心系统性能排名指标。

AI 回复实验的五平台端到端截尾均值为 2.514 s，对应的平均阶段耗时分别为 T0 = 0.486 s、T1 = 1.527 s、T2 = 0.501 s。T1 的增加符合预期，因为该阶段包含外部模型请求。本报告因此不将 2.514 s

视为 BiliGo 可跨环境复现的固定属性：用户选择不同模型、API 区域、中转服务商或本地模型时，结果可能出现明显变化。

4.2 AI 能力完整度

在 BiliGoBench v1.0 中，AI 支持主要依据系统能力覆盖范围进行评估。V3 Ultra 项目规格包含 AI 自动回复、OpenAI 兼容与 Anthropic 兼容接入方式、自定义模型/服务商配置，以及知识库与 RAG 风格检索等知识增强回复功能。表中 BiliGo 的能力状态来自项目已实现功能说明，本次实验暂未对这些功能进行独立的质量评分。

系统	AI 自动回复	知识库	检索 / RAG	外部 Agent / 动作集成	模型 / 服务商灵活性	证据
BiliGo V3 Ultra	支持	支持（项目已实现）	支持（项目已实现）	OpenAI / Anthropic 兼容接口	可自定义服务商与模型	项目规格
respond.io AI Agent	支持	支持	索引化知识源检索	AI Agent 动作 / workflow	引用文档中为平台托管	[5], [9]
Intercom Fin	支持	支持	明确采用 RAG	数据连接器与 MCP 连接器	专有 Fin AI Engine	[10], [11]
Zendesk AI Agents	支持	支持，含外部知识库	知识增强的 Agent 回答	外部 API 集成	引用文档中为平台托管	[7]
SaleSmartly AI	支持	支持	基于知识库的 Agent	自定义 Agent Webhook; Coze/Dify/OpenAI Assistants	支持多种第三方 Agent	[12], [13]

表 4 AI 能力证据矩阵。“平台托管”表示所引用产品文档未提供与 BiliGo 相同形式的、由用户直接选择第三方 LLM 接口的配置方式。

该矩阵用于比较功能是否存在，不代表 AI 回复质量排名。对于公开文档未明确说明的能力，本报告不将其直接记为负面结果。

5 行业公开参考数据

商业客服平台公开了若干具有参考价值、但统计口径并不统一的指标。SaleSmartly 宣称其客服产品平均响应时间低于 6 s；respond.io 文档给出的 AI Agent 典型处理时间为 10 – 15 s，并说明部分响应可能超过 20 s；Intercom 报告 Fin 在数千个团队中的平均解决率为 76%；Zendesk 建议 AI Agent 以 60% – 80% 的解决率和 85% – 95% 的回答准确率作为优化目标；Gorgias 公布的平台级 AI 解决率中位数为 45%，前 25% 商户达到 65%。[4] – [8]

公开响应时间参考（指标口径不同，仅作参考对照）

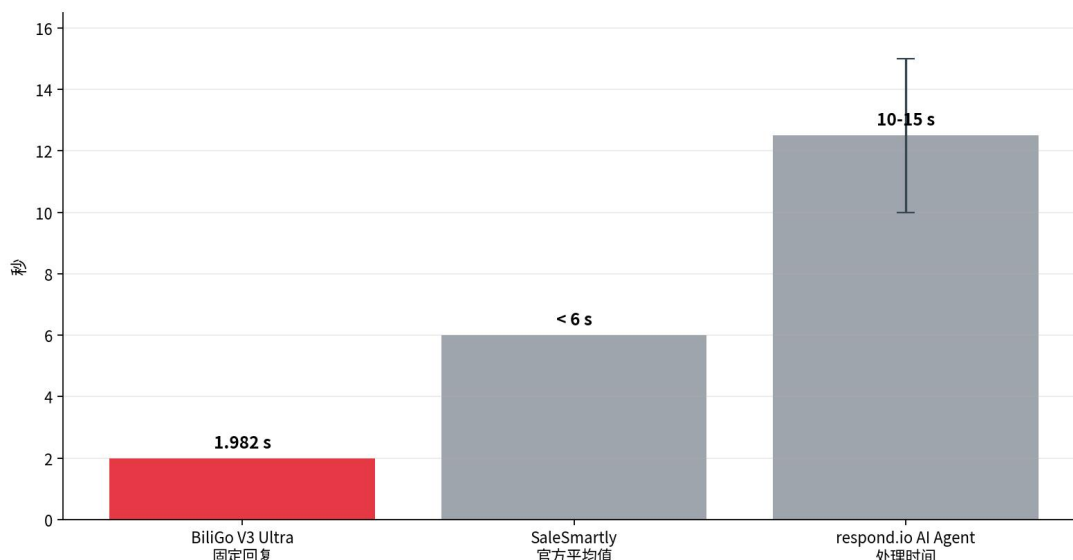


图 5 公开响应时间指标的背景对照。BiliGo 以红色突出显示。SaleSmartly 与 respond.io 的统计定义不同，因此该图不构成统一环境下的直接横向基准。

BiliGo 固定回复模式的五平台平均值为 1.982 s，在数值上低于 SaleSmartly 公开的“<6 s 平均响应时间”。不过，SaleSmartly 未公开该宣传指标的精确时间戳边界，而 respond.io 的 10 - 15 s 专门描述 AI 处理时间。基于这些口径差异，本报告仅将该比较作为 BiliGo 响应速度具有竞争力的背景证据，不据此推导其在所有场景下均快于相关平台。

6 讨论

6.1 当前数据能够支持的结论

现有数据最有力地支持以下判断：在给定实验室配置下，BiliGo 能够在五个中国主流社交平台实现低个位数秒级的自动回复。非 AI 模式下，五个平台中有四个平台的端到端截尾均值低于 1.75 s，跨平台平均值低于 2 s；与此同时，系统在 440 条消息样本上取得了 95.0% 的总体加权观测成功率。

分阶段延迟对于工程优化尤其有价值。抖音 4.570 s 的总延迟主要由 3.500 s 的 T1 组成，而哔哩哔哩私信 0.723 s 的总延迟中，T1 截尾均值仅为 0.040 s。该结果表明，后续性能优化应优先分析平台特定的处理路径，并针对不同平台制定差异化优化策略。

6.2 当前数据尚不能证明的内容

当前基准测试尚不足以建立全球范围的 SOTA 结论。具有充分说服力的 SOTA 声明，需要在相同网络、账号条件、消息集合、时间戳定义与平台配置下，对竞品开展受控测试。厂商公开指标具有参考价值，但无法替代这一类统一条件实验。同时，AI 回答质量、事实准确率、幻觉率、RAG 检索质量与任务解决率均未纳入当前计分范围。

6.3 有效性威胁与实验限制

需要考虑三项主要限制。第一，每个延迟阶段仅包含 5 个观测值，即使采用截尾均值，小样本效应仍可能影响结果。第二，各平台消息监测间隔根据实际稳定性需求分别配置，因此跨平台延迟在一定程度上受配置参数影响。第三，成功率实验受个人账号测试容量限制，其样本规模无法用于推导企业级可靠性保证。本报告明确保留这些限制，以保证实验可复现，并减少对结果的过度解释。

7 下一阶段基准测试建议

下一版 BiliGoBench 建议保留相同的 T0/T1/T2 计时体系，并增加统一条件下的竞品测试。具有最高价值的实验设计包括：BiliGo 与 2 - 3 个商业替代产品使用固定回复内容、统一消息发送节奏、同步

时钟、相同客户端网络环境，并在每个平台完成至少 30 - 50 次重复实验。AI 能力对比在初期仍可采用功能型评测，明确记录知识库、RAG/检索、自定义 API/服务商、上下文配置与故障回退等标准；AI 回复质量评分可在后续版本中作为独立 Benchmark Track 加入。

8 结论

BiliGo V3 Ultra 已展示出作为多平台 AI 私信自动回复系统的可靠技术基础。在本次非 AI 基准测试中，五平台端到端截尾均值为 1.982 s，总体加权观测回复成功率为 95.0%。哔哩哔哩私信以 0.723 s 成为本次实验中延迟最低的平台配置，小红书与微博在现有样本中均记录到 100% 的成功回复。接入 DeepSeek V4 Flash 后，AI 模式端到端平均延迟为 2.514 s；由于其中包含外部模型服务商耗时，该数字被明确作为补充实验结果。综合当前数据，可以支持 BiliGo 在测试环境中具备较强的响应性能与较为完整的 AI 接入灵活性。若要在公开技术宣传中正式使用 SOTA 结论，仍需完成统一实验条件下的竞品对照评测。

参考文献

[1] Chiyang001, “BiliGo: Bilibili private-message and comment auto-reply system,” GitHub repository. [Available online](#)

[2] BiliGo Project, “BiliGo V3 Ultra multi-platform latency, AI-latency, and reply-success experimental workbook,” project-internal dataset, 4 Sep. 2026.

[3] DeepSeek, “Models & Pricing / API documentation,” current model: deepseek-v4-flash, DeepSeek-V4-Flash-0731. [Available online](#)

[4] SaleSmartly, “Cross-border After-Sales Team Solutions,” reporting <6 s average response time. [Available online](#)

[5] respond.io, “Getting Started with AI Agents” and “AI Agent: Advanced Settings,” documenting 10-15 s typical AI processing time and occasional >20 s responses. [Available online](#)

[6] Intercom, “From resolutions to outcomes: Evolving how Fin delivers value,” reporting a 76% average resolution rate. [Available online](#)

[7] Zendesk, “How can I train my AI agent and improve suggestions?” recommending 60-80% resolution rate and 85-95% answer accuracy. [Available online](#)

[8] Gorgias, “The cheapest ticket is the one a human never touches,” 28 Apr. 2026, reporting 45% median and 65% top-quartile AI resolution. [Available online](#)

[9] respond.io, “Managing AI Knowledge Sources,” documenting indexed file/web knowledge sources and source inspection. [Available online](#)

[10] Intercom, “The Fin AI Engine,” documenting Intercom’s RAG architecture. [Available online](#)

[11] Intercom, “Knowledge sources to power AI, agents and self-serve support,” 23 Jun. 2026. [Available online](#)

[12] SaleSmartly, “AI Bot,” documenting Coze, OpenAI Assistants, system AI corpora, and channel-specific AI bots. [Available online](#)

[13] SaleSmartly, “Custom AI Agent Integration Guide” and “Knowledge Management,” documenting custom agent webhooks and enterprise knowledge bases. [Available online](#)

附录 A 非 AI 模式原始延迟数据

所有数值单位均为秒。正文针对每个平台的每个阶段采用截尾均值：在 5 次观测中剔除最大值与最小值，对剩余 3 个数值取平均。

平台	轮次	T0	T1	T2	单轮总计
哔哩哔哩私信	1	0.67	0.08	0.41	1.16
哔哩哔哩私信	2	0.83	0.06	0.24	1.13
哔哩哔哩私信	3	0.55	0.02	0.23	0.80
哔哩哔哩私信	4	0.13	0.03	0.23	0.39
哔哩哔哩私信	5	0.11	0.03	0.07	0.21
抖音	1	0.60	1.20	0.42	2.22
抖音	2	0.79	3.50	0.45	4.74
抖音	3	0.77	3.50	0.61	4.88
抖音	4	0.12	3.50	0.21	3.83
抖音	5	0.40	3.50	0.57	4.47

平台	轮次	T0	T1	T2	单轮总计
小红书	1	2.43	0.90	0.44	3.77
	2	0.21	0.60	0.08	0.89
	3	0.59	0.60	0.12	1.31
	4	0.13	0.70	0.09	0.92
	5	0.79	0.60	0.20	1.59
微博	1	1.32	0.21	0.57	2.10
	2	0.78	0.23	0.44	1.45
	3	0.25	0.19	0.97	1.41
	4	1.21	0.57	0.33	2.11
	5	0.69	0.13	0.91	1.73
闲鱼	1	0.90	0.73	0.46	2.09
	2	0.37	0.25	0.77	1.39
	3	0.46	0.55	0.28	1.29
	4	0.59	0.75	0.64	1.98
	5	0.49	0.52	0.27	1.28

表 A1 实验工作簿提供的非 AI 原始延迟观测值

附录 B AI 回复模式原始延迟数据

实验接入 DeepSeek 官方 API，提示词为“你好”，上下文模式关闭。T1 包含外部模型/API 延迟。

平台	轮次	T0	T1	T2	单轮总计
哔哩哔哩私信	1	1.03	1.22	0.43	2.68
	2	0.48	1.37	0.34	2.19
	3	0.95	1.57	0.30	2.82
	4	0.27	0.97	0.64	1.88
	5	0.32	1.71	0.19	2.22
抖音	1	0.63	1.25	0.26	2.14
	2	1.01	3.52	0.45	4.98
	3	0.49	3.57	0.88	4.94
	4	0.19	3.51	0.83	4.53
	5	0.39	3.55	0.70	4.64
小红书	1	0.74	0.91	0.55	2.20
	2	0.23	0.62	0.38	1.23
	3	0.66	0.52	0.94	2.12
	4	0.68	0.60	0.45	1.73
	5	0.30	0.61	0.22	1.13
微博	1	1.27	0.54	0.79	2.60
	2	0.47	1.09	0.82	2.38
	3	0.21	1.06	0.33	1.60
	4	0.43	1.51	0.73	2.67
	5	0.42	0.90	0.48	1.80
闲鱼	1	1.54	1.22	0.31	3.07
	2	0.24	1.07	0.46	1.77
	3	0.21	1.00	0.31	1.52
	4	0.29	0.82	0.66	1.77
	5	0.54	1.34	0.27	2.15

表 B1 实验工作簿提供的 AI 回复模式原始延迟观测值

附录 C 原始回复成功率计数

平台	处理消息	成功	失败	观测成功率
哔哩哔哩私信	127	113	14	88.98%
抖音	73	67	6	91.78%
小红书	109	109	0	100.00%
微博	90	90	0	100.00%
闲鱼	41	39	2	95.12%

表 C1 实验工作簿提供的回复成功/失败计数