

BiliGo V3 Ultra: A Multi-Platform AI-Enabled Private-Messaging Automation System

Chiyang001

4 September 2026

Abstract—BiliGo V3 Ultra is a multi-platform private-message automation system targeting major Chinese social platforms. This report presents an initial reproducible benchmark focused on two system-level dimensions: messaging-path performance and AI capability completeness. The latency benchmark decomposes each reply into T0 (user send to system detection), T1 (system processing), and T2 (system send to user-visible delivery). Five repeated measurements were collected for Bilibili private messages, Douyin, Xiaohongshu, Weibo, and Xianyu; the maximum and minimum observations were removed for each stage and the remaining three values were averaged. Under non-AI keyword/default-reply mode, BiliGo achieved a cross-platform mean end-to-end latency of 1.982 s, with Bilibili private messaging at 0.723 s. Across 440 processed messages, 418 replies succeeded, corresponding to a weighted success rate of 95.0%. A supplementary AI-enabled run using the official DeepSeek V4 Flash API produced a 2.514 s cross-platform mean end-to-end latency, but AI-provider latency is intentionally excluded from the primary system ranking. Public vendor figures from SaleSmartly, respond.io, Intercom, Zendesk, and Gorgias are used only as contextual industry references because their metric boundaries differ from this controlled BiliGo experiment. The present evidence supports competitive system responsiveness in the tested environment while reserving any state-of-the-art claim for a future controlled head-to-head benchmark.

Keywords—omnichannel messaging; private-message automation; AI agent; RAG; latency benchmark; Bilibili; Douyin; Xiaohongshu; Weibo; Xianyu

1 Introduction

Creator and customer-service workflows are increasingly distributed across multiple social platforms, creating duplicated operational work and inconsistent response latency. BiliGo was originally released as an open-source Bilibili private-message and comment auto-reply system and has since evolved into the V3 Ultra architecture evaluated in this report. The public repository documents the project's earlier rule-based automation foundation, local deployment model, configurable monitoring intervals, and message-reply controls. [1]

The V3 Ultra benchmark deliberately starts with a narrow scope. First, it measures the part of the system that BiliGo can directly control: how quickly a message is detected, processed, and delivered. Second, it evaluates AI support primarily as a capability layer rather than as a raw latency race. This distinction matters because users can connect different AI vendors, intermediary gateways, geographic regions, and models, all of which can dominate generation time independently of the messaging system itself.

The report therefore separates controlled BiliGo measurements from vendor-reported industry metrics. SaleSmartly currently advertises an average response time below 6 s, while respond.io documents a typical 10-15 s processing time for its AI Agent and notes that some responses can exceed 20 s. These figures are valuable context, but they are not treated as a controlled head-to-head comparison. [4], [5]

2 System Scope and Benchmark Design

2.1 Evaluated Platforms

The experiment covers five private-message channels: Bilibili private messages, Douyin, Xiaohongshu, Weibo, and Xianyu. Bilibili comments are excluded because comment automation belongs to a separate interaction category and is outside the private-message scope of this benchmark.

2.2 Three-Stage Latency Model

Each message transaction is decomposed into three stages. T0 measures the elapsed time from the user's send action to the moment BiliGo receives or detects the message. T1 measures system-side processing from message receipt to completion of reply generation or reply-task preparation. T2 measures the time from the

system's send action to the point at which the response is visibly presented to the user. The end-to-end value is defined as $T_{total} = T_0 + T_1 + T_2$.

For each platform and each stage, five measurements were collected. To reduce sensitivity to a single unusually high or low observation, the report uses a 20% symmetric trimmed mean: $t_{trim} = (\text{sum}(t_i) - \max(t_i) - \min(t_i)) / 3$ for $n = 5$. Raw measurements are retained in Appendices A and B.

2.3 Test Configuration

Platform	Message monitoring interval	Send interval / wait	Other configured interval
Bilibili DM	0.05 s	1.0 s	300 s auto-restart
Douyin	0.5 s	0.5 s	-
Xiaohongshu	1.0 s	1.0 s	-
Weibo	1.0 s	1.0 s	-
Xianyu	2.0 s	2.0 s	-

Table 1. Configured intervals in the supplied experimental environment.

Because platform-specific monitoring and send intervals differ, the ranking reflects BiliGo under the supplied practical configuration; it should not be interpreted as an intrinsic ranking of the platforms themselves.

3 Messaging Performance Results

3.1 Stage-Separated Latency

Rank	Platform	T0 (s)	T1 (s)	T2 (s)	End-to-end (s)
1	Bilibili DM	0.450	0.040	0.233	0.723
2	Xiaohongshu	0.530	0.633	0.137	1.300
3	Xianyu	0.513	0.600	0.460	1.573
4	Weibo	0.893	0.210	0.640	1.743
5	Douyin	0.590	3.500	0.480	4.570
-	Cross-platform mean	0.595	0.997	0.390	1.982

Table 2. Trimmed-mean non-AI latency and platform ranking; lower end-to-end latency ranks higher.

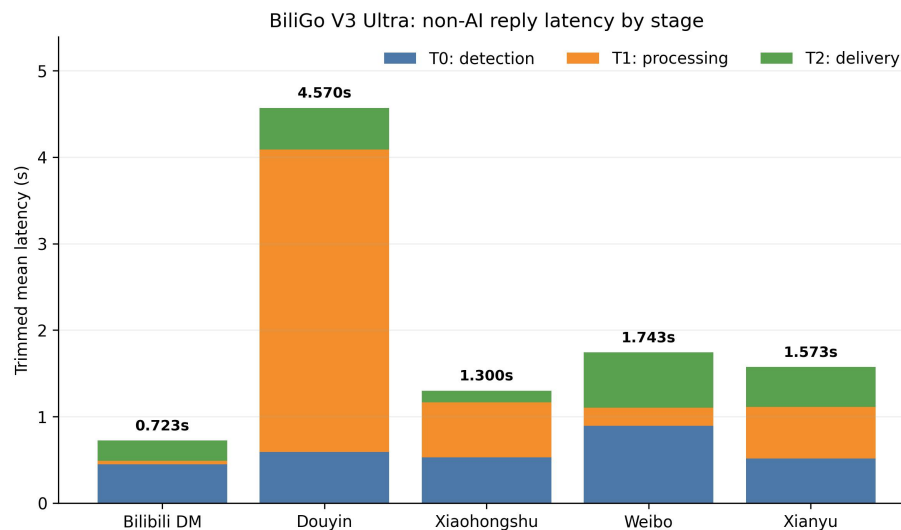


Fig. 1. Stage-level latency breakdown for non-AI keyword/default replies. Values are trimmed means from five observations per stage.

Across the five platforms, the mean end-to-end latency was 1.982 s. Bilibili private messaging produced the lowest trimmed-mean end-to-end latency at 0.723 s, followed by Xiaohongshu at 1.300 s, Xianyu at 1.573 s, Weibo at 1.743 s, and Douyin at 4.570 s. Averaged across platforms, T0 contributed 0.595 s, T1 0.997 s, and T2 0.390 s. In proportional terms, T1 accounted for approximately 50.3% of the cross-platform mean latency, making processing the largest component in this configuration.

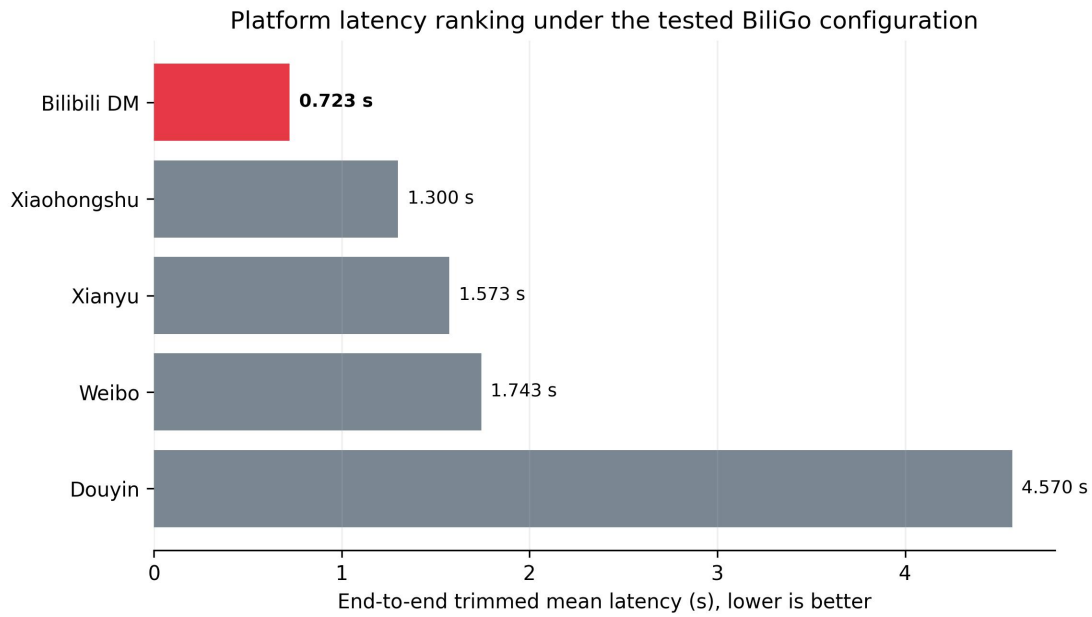


Fig. 2. End-to-end latency ranking. The fastest observed BiliGo platform configuration is highlighted in red.

3.2 Observed Reply Success Rate

Rank	Platform	Processed	Successful	Failed	Success rate	95% Wilson CI
1	Xiaohongshu	109	109	0	100.00%	96.60-100.00%
2	Weibo	90	90	0	100.00%	95.91-100.00%
3	Xianyu	41	39	2	95.12%	83.86-98.65%
4	Douyin	73	67	6	91.78%	83.21-96.18%
5	Bilibili DM	127	113	14	88.98%	82.34-93.32%
-	Overall weighted	440	418	22	95.00%	92.55-96.68%

Table 3. Observed reply success rate. Confidence intervals use the Wilson score method.

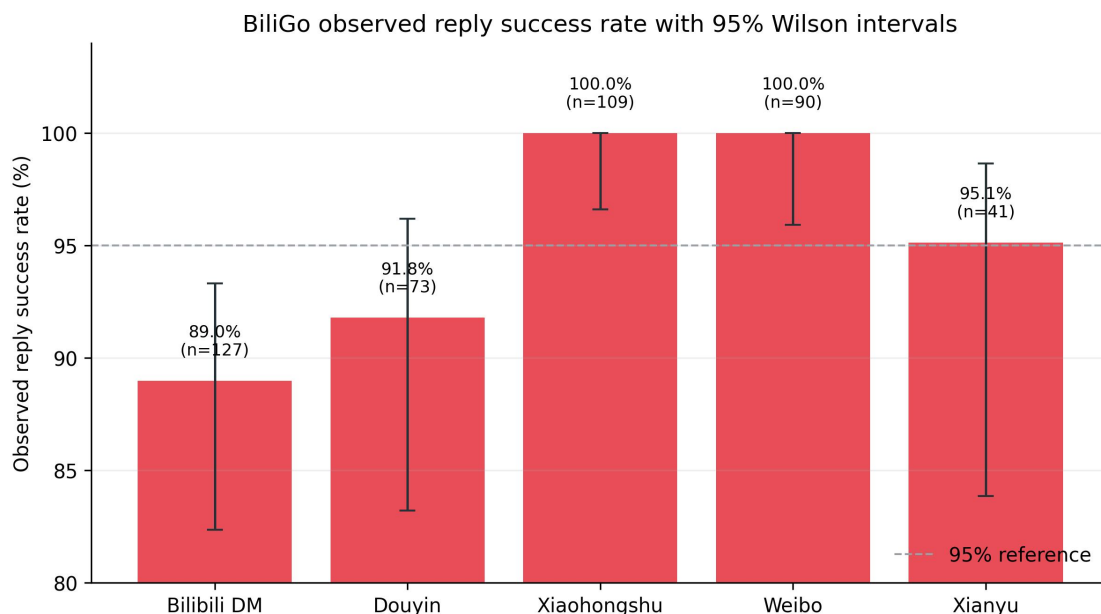


Fig. 3. Observed success rate by platform with 95% Wilson intervals. Red is reserved for BiliGo benchmark results.

The experiment processed 440 messages and successfully replied to 418, yielding a weighted success rate of 95.0% (95% Wilson CI: 92.55-96.68%). Xiaohongshu and Weibo achieved 100% observed success in their respective samples. Xianyu reached 95.12%, Douyin 91.78%, and Bilibili private messaging 88.98%. The macro-average of platform rates was 95.18%. Because sample sizes are modest and platform policies can change, these values should be treated as a point-in-time reliability snapshot rather than a permanent service-level guarantee.

4 AI-Enabled Mode and Capability Evaluation

4.1 AI Latency as a Supplementary Measurement

The AI-enabled experiment used the official DeepSeek API at <https://api.deepseek.com> with the DeepSeek V4 Flash model family. DeepSeek's current API documentation identifies the callable model as deepseek-v4-flash and the current model version as DeepSeek-V4-Flash-0731. [3] The prompt used in each run was the Chinese greeting '你好', and context mode was disabled in the supplied experiment.

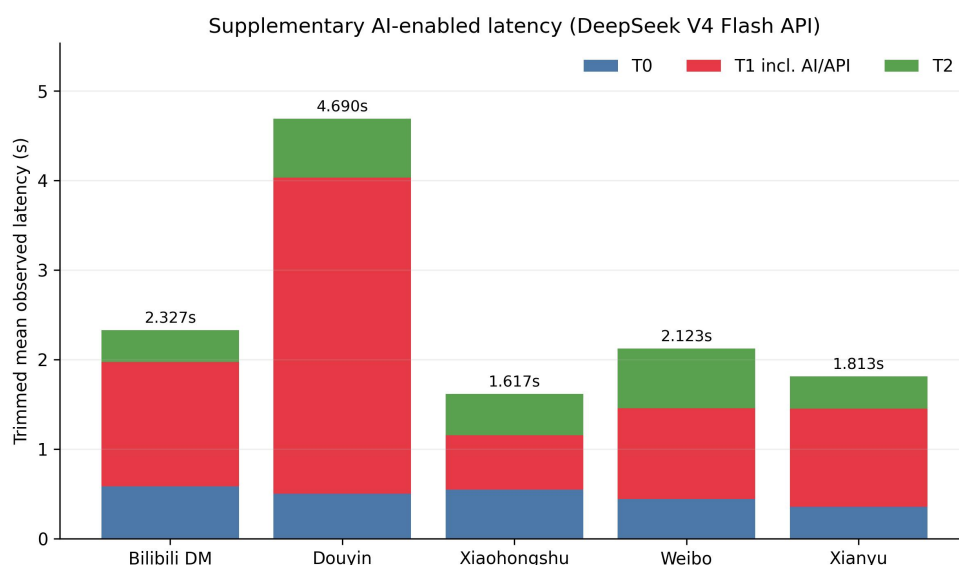


Fig. 4. Supplementary AI-enabled latency. T1 includes external AI/API time and is therefore not used as a primary BiliGo system ranking metric.

The cross-platform trimmed-mean end-to-end latency in the AI-enabled run was 2.514 s. The corresponding mean stage values were $T_0 = 0.486$ s, $T_1 = 1.527$ s, and $T_2 = 0.501$ s. The increase in T_1 is expected because it now includes an external model request. For this reason, the report does not interpret the 2.514 s figure as a portable property of BiliGo: a user selecting another model, API region, intermediary provider, or local model can observe materially different results.

4.2 AI Capability Completeness

For BiliGoBench v1.0, AI support is evaluated primarily by system capability coverage. The V3 Ultra project specification reports AI automatic replies, OpenAI-compatible and Anthropic-compatible integration modes, custom model/provider configuration, and knowledge-grounded reply functions including a knowledge base and RAG-style retrieval. These BiliGo capability statuses are project-reported implementation features and are not independently quality-scored in the present experiment.

System	AI auto-reply	Knowledge base	Retrieval/RAG	External agent/action integration	Model/provider flexibility	Evidence
BiliGo V3 Ultra	Yes	Yes (project-reported)	Yes (project-reported)	OpenAI-/Anthropic-compatible endpoints	Custom provider/model configuration	Project specification
respond.io AI Agent	Yes	Yes	Indexed knowledge-source retrieval	AI Agent actions / workflows	Platform-managed in cited docs	[5], [9]
Intercom Fin	Yes	Yes	Explicit RAG	Data connectors and MCP connectors	Proprietary Fin AI Engine	[10], [11]
Zendesk AI Agents	Yes	Yes, including external KBs	Knowledge-grounded agentic answers	External API integrations	Platform-managed in cited doc	[7]
SaleSmartly AI	Yes	Yes	Knowledge-base-grounded agent	Custom agent webhook; Coze/Dify/OpenAI Assistants	Multiple third-party agent options	[12], [13]

Table 4. AI capability evidence matrix. “Platform-managed” means the cited product documentation does not expose a direct user-selectable third-party LLM endpoint in the same configuration style as BiliGo.

The matrix is a feature-presence comparison, not a quality ranking. It intentionally does not convert undocumented capabilities into negative scores.

5 Industry Reference Data

Commercial customer-service platforms publish useful but heterogeneous reference figures. SaleSmartly advertises an average response time below 6 s for its customer-service product. respond.io documents a typical AI Agent processing time of 10-15 s and states that some responses may exceed 20 s. Intercom reports a 76% average resolution rate for Fin across thousands of teams. Zendesk recommends AI-agent targets of 60-80% resolution rate and 85-95% answer accuracy. Gorgias reports a 45% median AI resolution rate across its platform and 65% for the top quartile of merchants. [4]-[8]

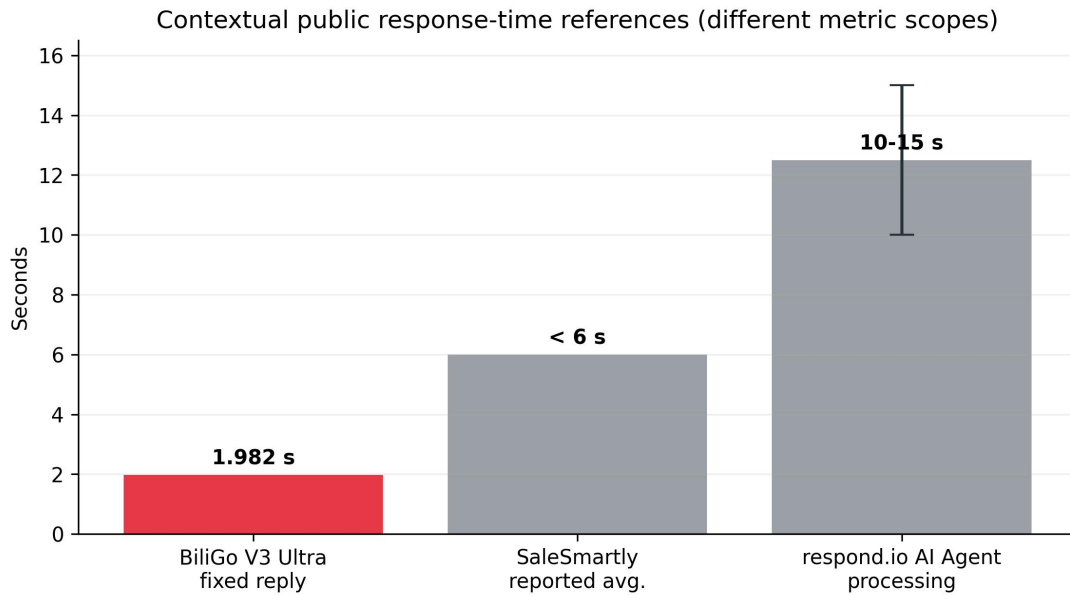


Fig. 5. Contextual public response-time references. BiliGo is highlighted in red. SaleSmartly and respond.io use different metric definitions, so the chart must not be read as a controlled head-to-head benchmark.

The BiliGo fixed-reply cross-platform mean of 1.982 s is numerically below SaleSmartly's publicly stated <6 s average response-time figure. However, SaleSmartly does not publish the exact timestamp boundaries used for that marketing metric, and respond.io's 10-15 s value specifically describes AI processing. Consequently, this report treats the comparison as contextual evidence of competitive response speed, not as proof that BiliGo is universally faster than either platform.

6 Discussion

6.1 What the Current Data Supports

The strongest conclusion supported by the present data is that BiliGo can deliver low single-digit-second automated replies across five major Chinese social platforms in the supplied laboratory configuration. Four of the five platform configurations remained below 1.75 s trimmed-mean end-to-end latency in non-AI mode, while the cross-platform mean was below 2 s. The system also demonstrated a 95.0% weighted observed success rate across 440 messages.

The latency decomposition is especially useful for engineering optimization. Douyin's 4.570 s total is dominated by a 3.500 s T1 component, whereas Bilibili's 0.723 s total contains only 0.040 s of trimmed-mean T1 processing. This indicates that future performance work should focus on platform-specific processing paths rather than applying a uniform optimization strategy.

6.2 What the Current Data Does Not Yet Prove

The current benchmark does not establish global state-of-the-art performance. A defensible SOTA statement requires controlled testing of competing systems under the same network, account conditions, message set, timestamp definitions, and platform configuration. Vendor-reported figures are valuable reference points but cannot replace that controlled experiment. Likewise, AI answer quality, factual accuracy, hallucination rate, RAG retrieval quality, and task resolution are outside the current scored scope.

6.3 Threats to Validity

Three limitations should be considered. First, each latency stage contains only five observations, so even a trimmed mean remains sensitive to small-sample effects. Second, the platform monitoring intervals differ by design, which improves practical stability but makes cross-platform latency partly configuration-dependent. Third, the success-rate sample is constrained by personal-account test capacity, and therefore cannot be treated as an enterprise-scale reliability guarantee. These limitations are explicitly retained in the report to preserve reproducibility and prevent over-interpretation.

7 Recommended Next Benchmark Step

The next version of BiliGoBench should preserve the same T0/T1/T2 instrumentation while adding a controlled competitor pass. The highest-value design is to test BiliGo and two or three commercial alternatives with a fixed reply payload, identical message timing, synchronized clocks, the same client network, and at least 30-50 repetitions per platform. AI capability comparison can remain feature-based initially, with knowledge-base support, RAG/retrieval, custom API/provider support, context configuration, and fallback behavior recorded as explicit criteria. AI response-quality scoring can be introduced later as a separate benchmark track.

8 Conclusion

BiliGo V3 Ultra demonstrates a credible foundation for a multi-platform AI-enabled messaging automation system. In the supplied non-AI benchmark, it achieved a 1.982 s cross-platform trimmed-mean end-to-end latency and a 95.0% weighted observed reply success rate. Bilibili private messaging was the fastest tested configuration at 0.723 s, while Xiaohongshu and Weibo recorded 100% success in the available samples. The AI-enabled run averaged 2.514 s end to end with DeepSeek V4 Flash, but that figure is intentionally treated as supplementary because model-provider latency is external to the messaging core. Taken together, the results support a strong claim of competitive responsiveness and broad AI integration flexibility within the tested environment. A controlled competitor benchmark is the remaining requirement before using a formal SOTA claim in public technical marketing.

References

- [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](#)

Appendix A Raw Non-AI Latency Measurements

All values are seconds. The main report uses a trimmed mean for each stage: remove the maximum and minimum of five observations and average the remaining three.

Platform	Run	T0	T1	T2	Run total
Bilibili DM	1	0.67	0.08	0.41	1.16
Bilibili DM	2	0.83	0.06	0.24	1.13
Bilibili DM	3	0.55	0.02	0.23	0.80
Bilibili DM	4	0.13	0.03	0.23	0.39
Bilibili DM	5	0.11	0.03	0.07	0.21
Douyin	1	0.60	1.20	0.42	2.22
Douyin	2	0.79	3.50	0.45	4.74
Douyin	3	0.77	3.50	0.61	4.88
Douyin	4	0.12	3.50	0.21	3.83
Douyin	5	0.40	3.50	0.57	4.47
Xiaohongshu	1	2.43	0.90	0.44	3.77
Xiaohongshu	2	0.21	0.60	0.08	0.89
Xiaohongshu	3	0.59	0.60	0.12	1.31
Xiaohongshu	4	0.13	0.70	0.09	0.92
Xiaohongshu	5	0.79	0.60	0.20	1.59
Weibo	1	1.32	0.21	0.57	2.10
Weibo	2	0.78	0.23	0.44	1.45
Weibo	3	0.25	0.19	0.97	1.41
Weibo	4	1.21	0.57	0.33	2.11
Weibo	5	0.69	0.13	0.91	1.73
Xianyu	1	0.90	0.73	0.46	2.09
Xianyu	2	0.37	0.25	0.77	1.39
Xianyu	3	0.46	0.55	0.28	1.29
Xianyu	4	0.59	0.75	0.64	1.98
Xianyu	5	0.49	0.52	0.27	1.28

Table A1. Raw non-AI latency observations supplied in the experimental workbook.

Appendix B Raw AI-Enabled Latency Measurements

The experiment used the official DeepSeek API and the prompt '你好'; context mode was disabled. T1 includes external model/API latency.

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

Table B1. Raw AI-enabled latency observations supplied in the experimental workbook.

Appendix C Raw Reply-Success Counts

Platform	Processed	Successful	Failed	Observed rate
Bilibili DM	127	113	14	88.98%
Douyin	73	67	6	91.78%
Xiaohongshu	109	109	0	100.00%
Weibo	90	90	0	100.00%
Xianyu	41	39	2	95.12%

Table C1. Raw reply-success counts supplied in the experimental workbook.