

大模型词元服务性能监测报告

(2026 年)

中国信息通信研究院人工智能研究所
人工智能大模型及软硬件评测工业和信息化部重点实验室
中国人工智能产业发展联盟
2026年10月

版权声明

本报告版权属于中国信息通信研究院、人工智能大模型及软硬件评测工业和信息化部重点实验室和中国人工智能产业发展联盟，并受法律保护。转载、摘编或利用其他方式使用本报告文字或者观点的，应注明“来源：中国信息通信研究院、人工智能大模型及软硬件评测工业和信息化部重点实验室和中国人工智能产业发展联盟”。违反上述声明者，将追究其相关法律责任。

前 言

随着大模型技术的迅速迭代，词元（Token）服务成为从技术走向应用的关键点，我国日均 Token 调用量持续攀升，2026 年 6 月，国内词元日均消耗量约 500 万亿，是 2025 年 12 月的近 4 倍。产业中越来越多的 MaaS（Model as a Service，模型即服务）平台厂商通过公有云 API 方式供给大模型 Token 服务。然而用户对 Token 服务性能缺乏直观的判断依据，且面临选型困难等各类问题。

为有效衡量服务性能，中国信息通信研究院（以下简称“中国信通院”）人工智能研究所依托人工智能大模型及软硬件评测工业和信息化部重点实验室（以下简称“实验室”），联合中国人工智能产业发展联盟（AIIA）模型服务（MaaS）工作组和高质量词元服务工作组，自 2025 年以来持续开展公有云大模型 Token 服务性能监测工作。

监测显示，大模型 Token 服务性能整体呈现提升态势，头部平台在每秒 Token 吞吐量等关键指标上持续优化，满足多数通用场景要求。本报告基于 2026 年 1 月-8 月的监测数据，围绕各平台 Token 服务调用成功率、吞吐速度、首词时延、稳定性、价格、上下文窗口等维度，月均 TPS 相较 2025 年实现翻倍，原厂大模型整体时延有所增长，但开源模型在第三方平台上时延持续优化。对比原厂大模型在其官方平台，以及 DeepSeek、Qwen、GLM、Kimi 四款开源模型在第三方平台性能表现上的差异，并进一步分析影响 Token 服务性能的关键因素，旨在为用户选型提供参考，并为产业界优化服务能力提供方向指引。

本报告中涉及的性能数据基于公开接口实测所得，受测试时间、网络环境、并发策略等因素影响，结果仅供技术参考，不构成对任何产品或服务的推荐、背书或商业评价。由于数据较多，报告整理可能存在疏漏之处，如有任何问题欢迎与我们沟通交流，后续将结合最新数据持续更新完善。

目 录

一、 核心观察	1
(一) 大模型 Token 服务吞吐与时延同步增加	1
(二) 大模型价格回升明显，向“能力跃迁”发展演进	2
(三) 开源模型 Token 服务性能提升显著	2
二、 监测方法	3
(一) 监测对象	3
(二) 监测指标	4
(三) 监测参数设置	4
三、 原厂大模型监测结果	5
(一) 调用成功率	5
(二) 每秒输出词元数 (TPS)	8
(三) 首词时延 (TTFT)	10
(四) 服务稳定指数	12
(五) 服务价格与费用	15
(六) 模型上下文	17
四、 开源大模型监测结果	18
(一) 开源模型上线率	18
(二) DeepSeek 监测结果	20
(三) Qwen 监测结果	22
(四) GLM 监测结果	25
(五) Kimi 监测结果	27
五、 总结	30
附录 A: 其他事项说明	31
附录 B: 监测模型清单	33

图 目 录

图 1	8 月份原厂大模型 Token 服务调用成功率平均值	6
图 2	各原厂大模型 Token 服务调用成功率（越高越好）	7
图 3	所有大模型 Token 服务月均调用成功率	7
图 4	8 月份原厂大模型 Token 服务 TPS 平均值	8
图 5	各原厂大模型 Token 服务 TPS 平均值（越高越好）	9
图 6	所有大模型 Token 服务月均 TPS	10
图 7	8 月份原厂大模型 Token 服务 TTFT 平均值	10
图 8	各原厂大模型 Token 服务 TTFT 平均值（越低越好）	11
图 9	所有大模型 Token 服务的月均 TTFT	12
图 10	8 月份原厂大模型 Token 服务稳定指数	13
图 11	原厂大模型 Token 服务月均稳定性指数（越高越好）	14
图 12	所有大模型 Token 服务月均稳定指数	14
图 13	国内原厂大模型发布时归一化价格趋势图	15
图 14	国外原厂大模型发布时归一化价格趋势图	16
图 15	8 月份国内外大模型 Token 消耗量及费用对比图	17
图 16	国内大模型上下文长度占比情况	18
图 17	所有平台上 DeepSeek 月均调用成功率	20
图 18	各平台的 DeepSeek 月均调用成功率（越高越好）	20
图 19	所有平台的 DeepSeek 月均 TPS	21
图 20	各平台的 DeepSeek 月均 TPS（越高越好）	21
图 21	所有平台的 DeepSeek 月均 TTFT	22
图 22	各平台的 DeepSeek 月均 TTFT（越低越好）	22
图 23	平台的 Qwen 月均调用成功率	23
图 24	各平台的 Qwen 月均调用成功率（越高越好）	23
图 25	所有平台的 Qwen 月均 TPS	23
图 26	各平台的 Qwen 月均 TPS（越高越好）	24
图 27	所有平台的 Qwen 月均 TTFT	24
图 28	各平台的 Qwen 月均 TTFT（越低越好）	24

图 29	所有平台的 GLM 月均调用成功率	25
图 30	各平台的 GLM 月均调用成功率（越高越好）	25
图 31	所有平台的 GLM 月均 TPS	26
图 32	各平台的 GLM 月均 TPS（越高越好）	26
图 33	所有平台的 GLM 月均 TTFT	27
图 34	各平台的 GLM 月均 TTFT（越低越好）	27
图 35	所有平台的 Kimi 月均调用成功率	28
图 36	各平台的 Kimi 月均调用成功率（越高越好）	28
图 37	所有平台的 Kimi 月均 TPS	29
图 38	各平台的 Kimi 月均 TPS（越高越好）	29
图 39	所有平台的 Kimi 月均 TTFT	29
图 40	各平台的 Kimi 月均 TTFT（越低越好）	30

表 目 录

表 1	典型开源大模型在国内外主流 MaaS 平台的上线率统计	18
表 2	原厂模型监测清单	33
表 2	原厂模型监测清单（续表）	34
表 3	不同平台上开源模型最新监测清单	34

一、核心观察

从 2026 年 1 月至 8 月，根据国内外主流的公有云大模型 Token 服务性能监测结果可见，在模型参数量、上下文窗口不断增长的同时，Token 服务性能呈现出不同程度的变化。

（一）大模型 Token 服务吞吐与时延同步增加

1.Token 服务调用成功率相对平稳，保持较高水平

自 5 月份以来，由于逐步增加了调用成功率的监测频率，偶发性失败被捕捉到的概率上升，成功率存在下降趋势，每月均值均高于 99%，8 月平均值为 99.18%，明显优于 2025 年。

2.Token 服务每秒吞吐量较 2025 年翻倍，仍低于国外

8 月份国内各模型平均 TPS（每秒输出 Token 数）达到 78.23（个/秒），与 2025 年均值 38.25（个/秒）相比实现翻倍。但 8 月份国外典型模型的 TPS 均值为 112.82（个/秒），与 2025 年 12 月的 51.35（个/秒）相比，提升亦超过一倍，整体优于国内。

3.Token 服务时延有所增加，国内外趋势一致

今年以来，国内各模型的平均 TTFT（首词时延）呈现持续提升态势，8 月份均值达到 1.26 秒，相较于 2025 年平均值 0.78 秒明显增加。但同时 8 月份国外典型模型 TTFT 平均值达到 2.08 秒，相比 4 月份 1.04 秒亦明显增加。值得注意的是，TTFT 上升并非单纯的性能退化，而是伴随着模型参数量增加、上下文长度增加、调用量增加等多种因素同步产生的结果。

（二）大模型价格回升明显，向“能力跃迁”发展演进

1.2026 年词元服务价格明显回升

经历了过去两年价格战后，2026 年国内外模型服务价格明显上涨。国内 2025 下半年归一化均价为 6.8 元，而今年以来均价为 13.46 元，涨幅接近 1 倍。但国外价格仍高于国内，2026 年以来发布的典型模型平均归一价格约为 117.8 元。

2.模型上下文长度不断向百万级长度扩展

被监测的模型中，上下文长度达到 1M（1 百万 Tokens）的模型占比 47.8%，200K 及以上的合计达 95.7%，相比 2025 年底的占比提高超 76 个百分点，而低于 200K 的模型占比持续减少。

3.国产开源大模型成为全球诸多平台的主流选择

根据对 22 个国内外主流 MaaS 平台统计显示，DeepSeek、Qwen 和 GLM 系列模型上线率达到 100%，Kimi、MiniMax 系列上线率分别达 95%、86%，以上 5 款国内开源模型上线率相较于 5 月份统计结果有所上升。

（三）开源模型 Token 服务性能提升显著

1.开源模型新版本性能较好，第三方 MaaS 平台服务性能持续提升

5 月份更新了 4 款开源模型的新版本进行监测，结果显示 TPS 持续提升，TTFT 时延不断压缩。例如 8 月份 DeepSeek-V4-Pro 的 TPS 已达到 DeepSeek-V3.2 的 2 倍，Qwen3.5-397B-A17B 8 月份 TPS 突破

100（个/秒）。时延方面，5 月份主要模型完成换代升级后，TTFT 普遍出现阶段性上升；自 6 月份起，随着推理优化措施落地，时延指标逐步回落至换代前水平。8 月份 3 款模型 TTFT 均值均低于 1 秒。

2.同一模型跨平台性能波动较大，平台侧工程优化差异显著

4 款模型在不同平台上的 TPS、TTFT、调用成功率表现不一。部分第三方平台性能明显优于官方平台提供的模型服务，同一模型在不同平台间的 TPS 差距最大可接近一倍。各平台对相同模型的推理加速、优化策略、成本控制不同，是造成性能分化的主要原因。

二、监测方法

（一）监测对象

本报告监测对象为普通用户可调取的公有云 Token 服务，一方面包括国内外原厂大模型 Token 服务（即来自大模型原厂商官方接口），其中国外模型选取 GPT、Claude、Gemini 三款系列模型，通过监测可横向对比各类大模型的性能；另一方面对 DeepSeek、Qwen、GLM、Kimi 共 4 款开源模型在不同 MaaS 平台上的 Token 服务进行专项监测，用以衡量各平台的工程化能力。

纳入监测的模型通常选择最新三代以内的版本，原厂模型以官方上线的标准版或满血版模型为准（如 Max、Pro 版），对于优化版本模型（如 Flash、Lite 模型）未纳入监测与分析。对于发布时间较早的模型采取逐步取消机制，以确保数据时效性与行业前沿性高度契合。对于同时具备思考和非思考能力的模型，通常采用思考模式作为性能

测试结果，思考深度选择默认模式。其他参数通常采用平台默认设置。

（二）监测指标

1. **调用成功率**，指请求调用成功的次数占总调用次数的比例。调用成功指发送 1 次请求调用后，收到回复即算作成功。由于速率限制、服务下线、账户欠费等问题导致的请求失败情况不纳入统计。由于监测过程中通过采样方式进行数据统计，调用成功率为 100% 时并不代表实际应用时不会发生调用失败的情况。

2. **每秒输出词元数（Tokens Per Second, TPS）**，指模型每秒输出的词元数量，单位为（个/秒）。

3. **首词时延（Time To First Token, TTFT）**，指从请求开始至返回首个词元之间的时延，单位为秒。

4. **服务稳定性指数**，指在测试周期内，TTFT 和 TPS 的变化浮动情况。

5. **开源模型上线率**，指开源模型在被监测 MaaS 平台中的上线比例。

6. **模型服务价格**，指调用公有云模型时，其处理每百万 Token 的价格，单位为元。为便于价格归一化处理 and 对比，模型价格计算时将输入价格和输出价格进行了公式化处理。计算公式见附录 A。

7. **模型上下文**，根据官方公布的上下文长度进行统计。

（三）监测参数设置

1. **监测节点分布**：本监测选择位于不同国家和地区的云服务器节点，具体包括北京、广州等国内节点，美国等海外节点，各节点通过公有云 API 向目标模型发送请求。

2.请求参数设置：

（1）输入长度（Prompt Tokens）：监测数据集来自基准测试集固定题库，每次请求的输入 Token 长度约为 50~200 个 Token，属于短文本场景。

（2）最大输出长度（max_tokens）：采用各平台默认设置，未进行自定义调整。

（3）上下文窗口：采用平台默认上下文长度，未指定历史对话上下文，每次请求为独立会话（即空载/无历史上下文状态）。

3.负载状态：本监测为非独占式公有云 API 调用，监测时各模型服务处于正常公有负载状态，非空载、非独占资源测试。不同时间段的监测结果可能受平台实时负载影响而产生波动。

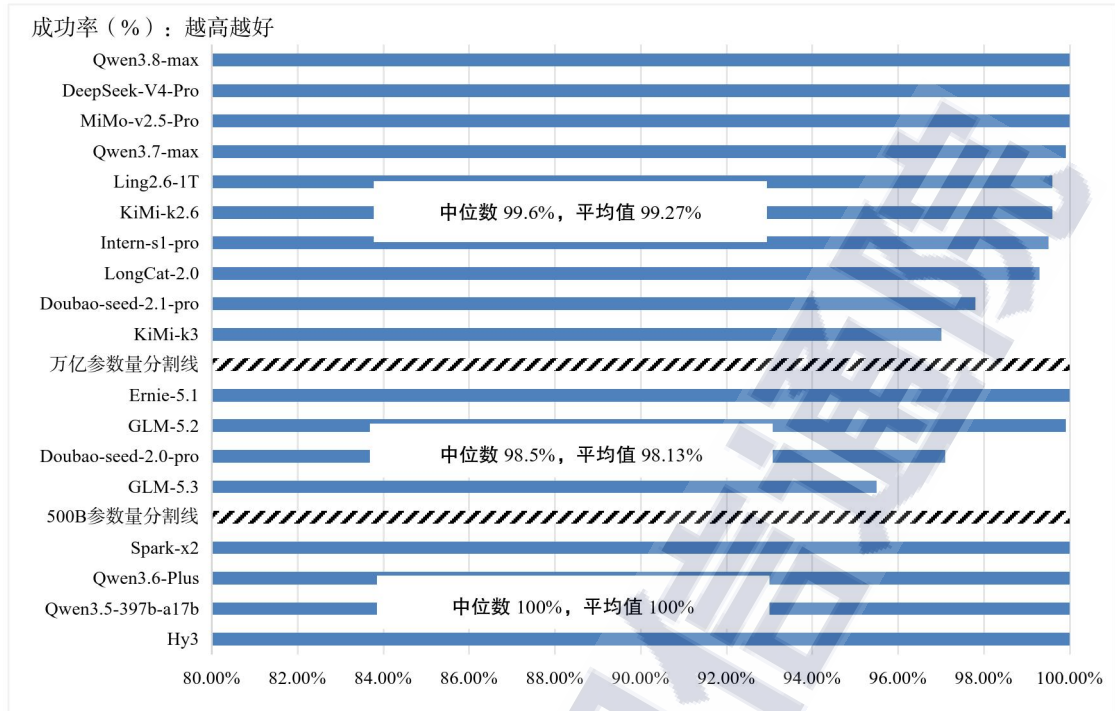
4.输出模式：统一采用流式输出（Streaming），思考模式选择默认深度。

三、原厂大模型监测结果

监测范围为 2026 年 1 月至 8 月，监测对象为国内外主流原厂大模型 Token 服务。监测指标包括调用成功率、每秒输出 Token 数(TPS)、首词时延（TTFT）和服务稳定性指数等。由于不同模型上线时间、版本迭代节奏和平台接入情况存在差异，部分监测数据并不完全连续。

（一）调用成功率

图 1 为 8 月份国内各原厂大模型 Token 服务调用成功率监测结果。数据显示，调用成功率整体保持稳定，国内所有被监测原厂总体调用成功率中位数为 99.75%，8 月平均值为 99.18%。



来源：中国信通院，2026 年 9 月

注：图中参数量的划分，来源于官方公开信息，以及网络推测信息

图 1 8 月份原厂大模型 Token 服务调用成功率平均值

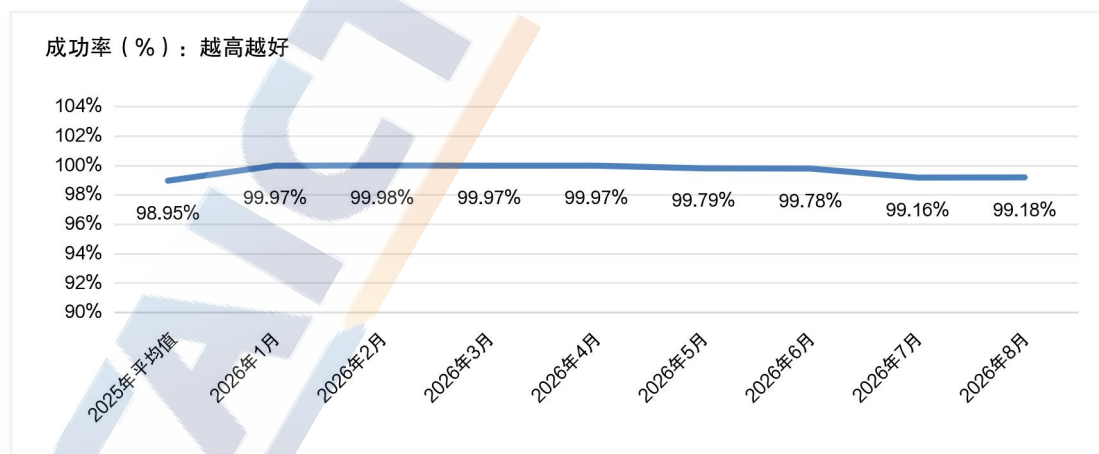
图 2 为每月国内外各原厂大模型 Token 服务调用成功率监测结果。数据显示，8 月份调用成功率达到 100% 的占比为 44.4%，相比 6、7 月份有所提升。但与前 5 个月相比明显降低，主要原因是 5 月份起监测采样频率逐渐从此前每天 15 次提升至每 5 分钟 1 次，监测密度大幅增加，偶发性失败被捕捉到的概率上升。



来源：中国信通院，2026 年 9 月

图 2 各原厂大模型 Token 服务调用成功率（越高越好）

图 3 为所有被监测国内大模型 Token 服务调用成功率的每月平均值。数据显示，自 6 月份起，国内大模型 Token 服务调用成功率平均值有所降低，但均保持在 99% 以上，相比 2025 年有较大提升。

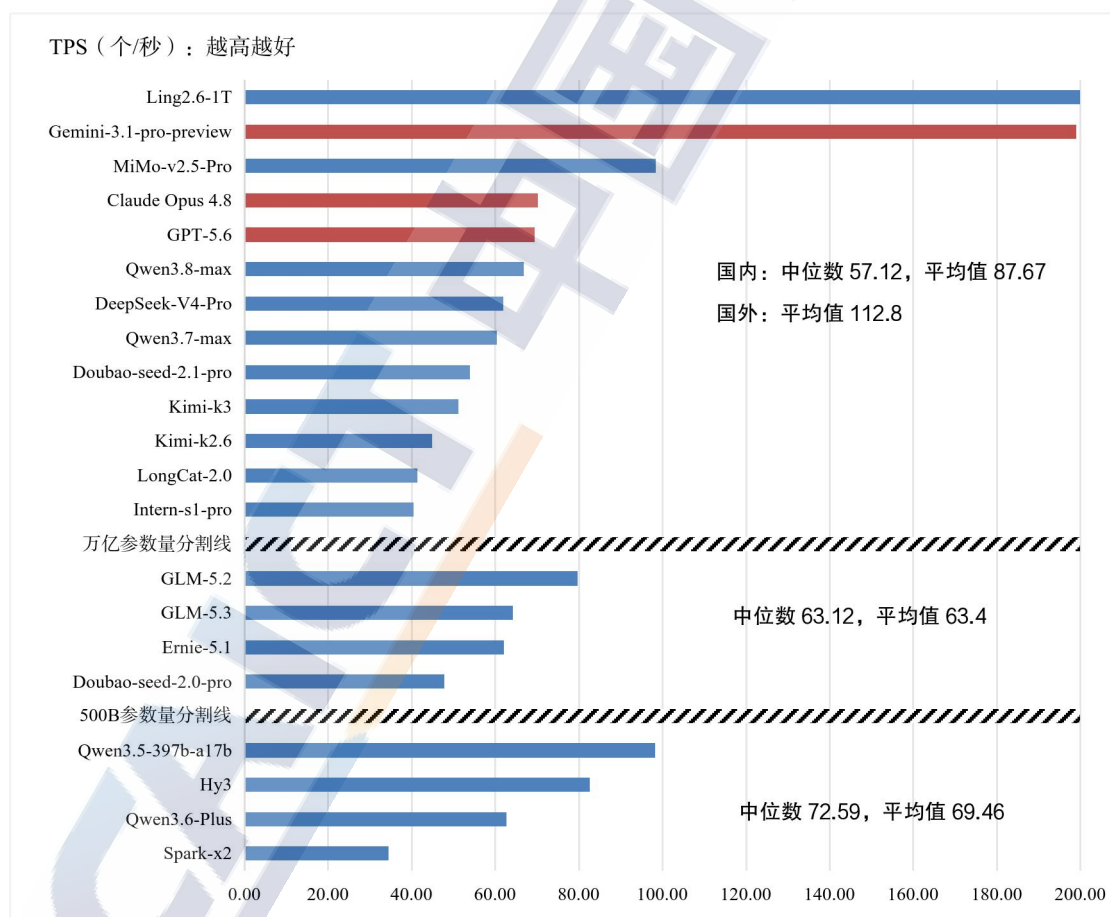


来源：中国信通院，2026 年 9 月

图 3 所有大模型 Token 服务月均调用成功率

（二）每秒输出词元数（TPS）

图 4 为 8 月份国内外各原厂大模型 Token 服务 TPS 监测结果。国内方面，1 款模型的 TPS 超过 300（个/秒），500B 以下模型 TPS 中位数为 72.59（个/秒），500B-1T 参数量模型 TPS 中位数为 63.12（个/秒），1T 以上参数量国内模型 TPS 中位数为 57.12（个/秒），这说明输出速率与时延表现与上下文长度、激活参数规模存在一定关联。由于部分模型参数量未公开，本节结论仅适用于已标注来源的模型子集。然而海外三款模型的 TPS 略高，GPT-5.6 最低也已达到 69.31（个/秒）。

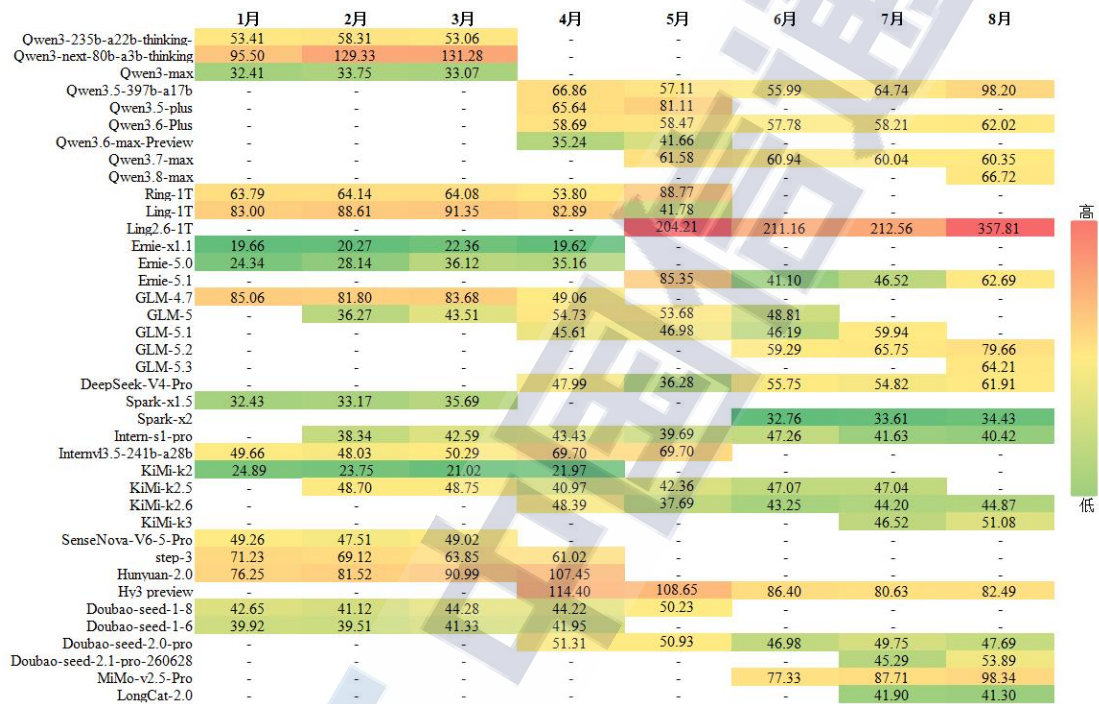


来源：中国信通院，2026 年 9 月

注：图中参数量的划分，来源于官方公开信息，以及网络推测信息

图 4 8 月份原厂大模型 Token 服务 TPS 平均值

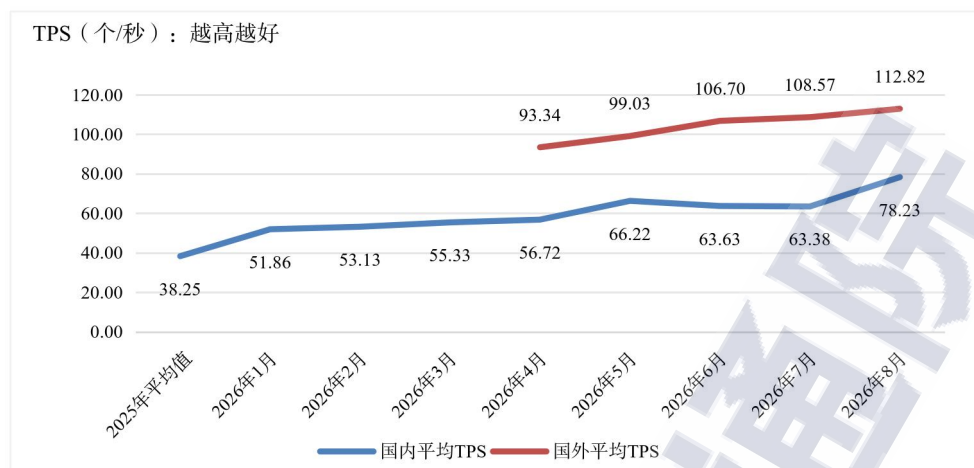
图 5 为国内各原厂大模型 Token 服务每月 TPS 监测结果。数据显示，多数新发布模型的 Token 服务 TPS 性能明显提升，而旧版本模型的 TPS 会逐渐下调并下线服务。同时，今年 4 月份开始已有模型 TPS 突破 100（个/秒），5 月份突破 200（个/秒），8 月份突破 300（个/秒），Flash 版本模型最高已突破 500（个/秒）。



来源：中国信通院，2026 年 9 月

图 5 各原厂大模型 Token 服务 TPS 平均值（越高越好）

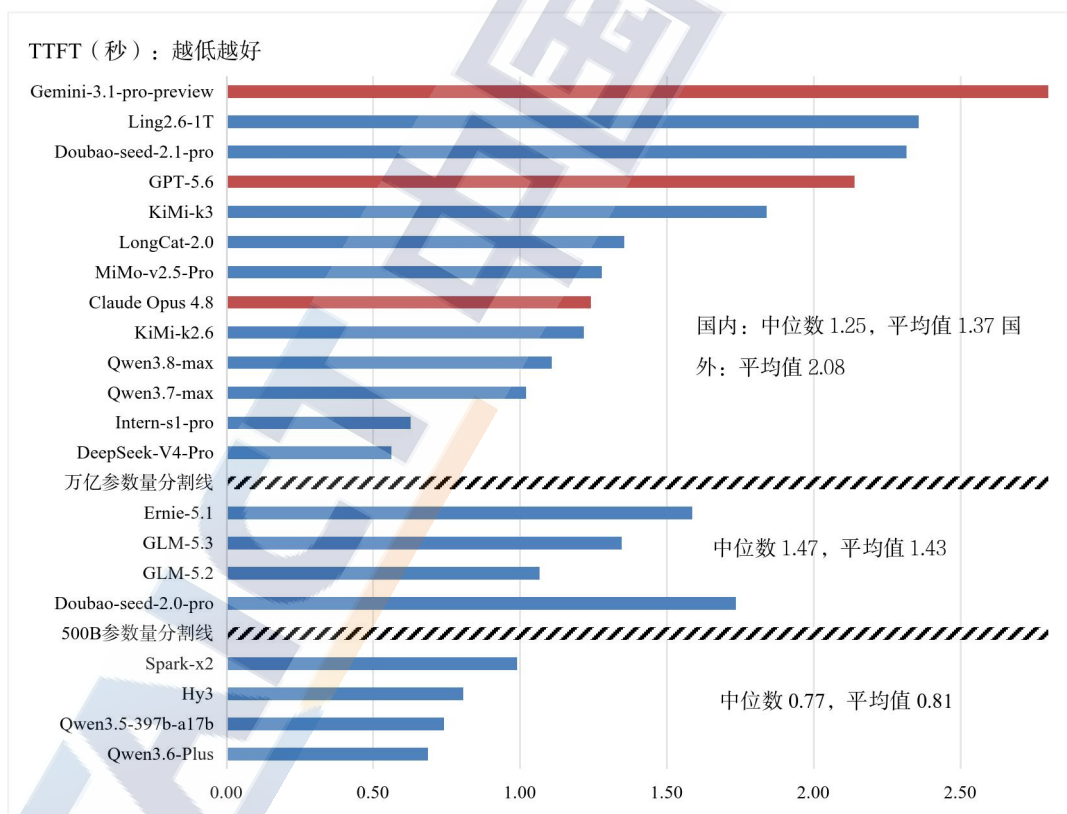
图 6 为每月所有被监测大模型 Token 服务的 TPS 平均值。数据显示，1 月至 8 月国内大模型 TPS 平均值稳步提升，8 月份 TPS 均值达到 78.23（个/秒），相比于 2025 年均值实现翻倍，但 TPS 中位数为 62.36（个/秒），由于 Ling2.6-1T 模型 TPS 达到 357.81（个/秒）远高于其他模型，因此平均值被整体拉高。而国外三款模型 8 月份 TPS 平均值提升至 112.82（个/秒），是国内均值的 1.4 倍。



来源：中国信通院，2026 年 9 月

图 6 所有大模型 Token 服务月均 TPS

（三）首词时延（TTFT）



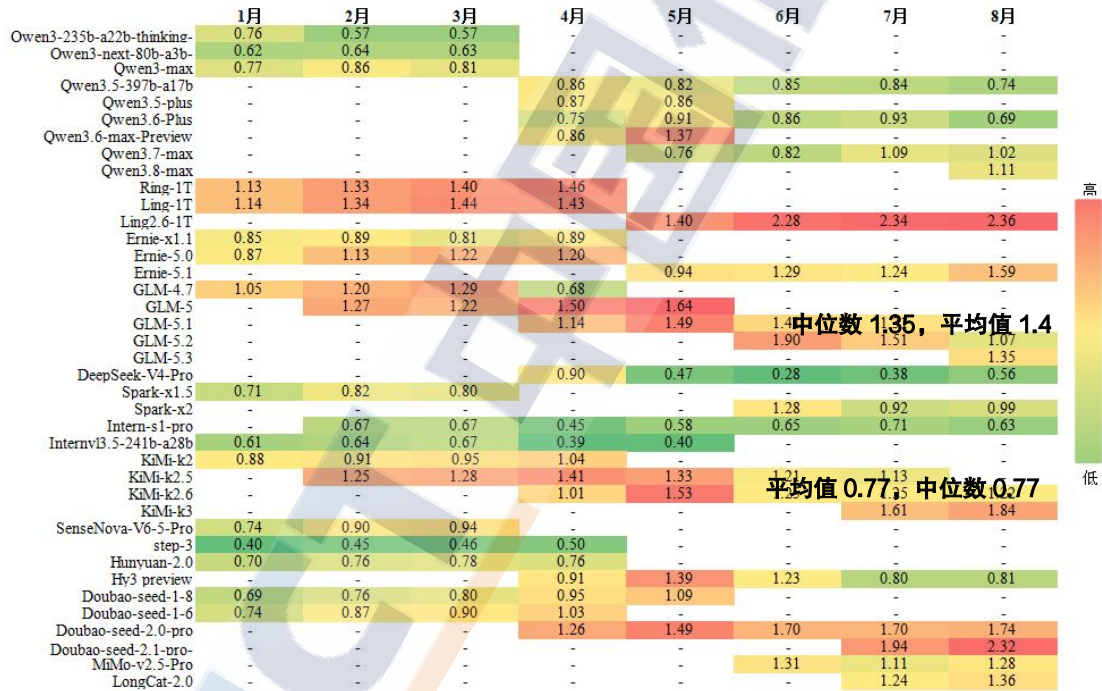
来源：中国信通院，2026 年 9 月

注：图中参数量的划分，来源于官方公开信息，以及网络推测信息

图 7 8 月份原厂大模型 Token 服务 TTFT 平均值

图 7 为 8 月份国内外各原厂大模型 Token 服务 TTFT 监测结果。数据显示，TTFT 与模型参数量有直接关系，国内方面，最低 TTFT 为 0.69 秒，最高为 2.36 秒，平均值为 1.26 秒，33.3% 的模型 TTFT 低于 1 秒。国外方面，三款模型 TTFT 均超过 1 秒，GPT-5.6 和 Gemini-3.1-pro-preview 均超过 2 秒。

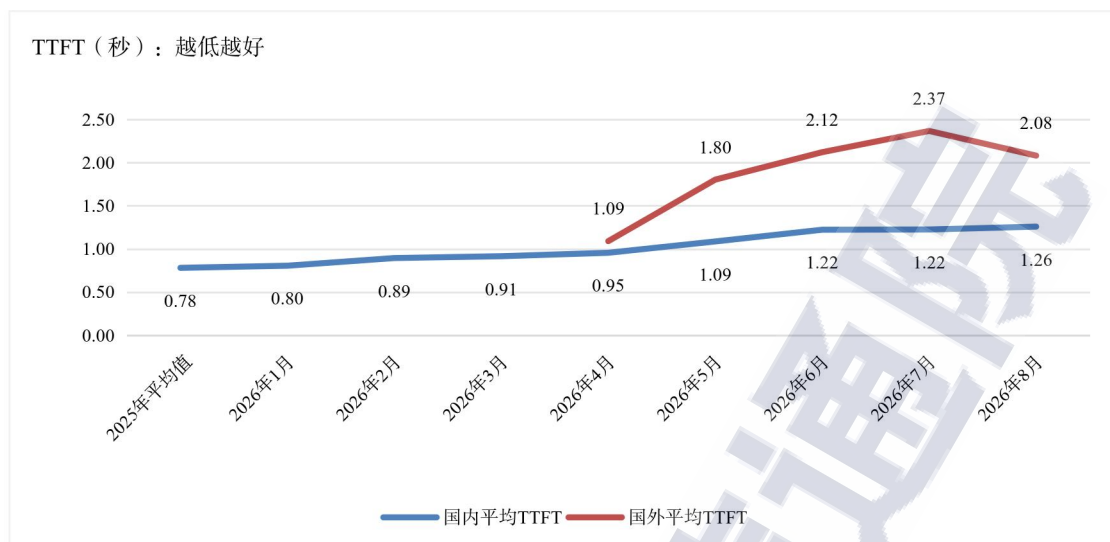
图 8 为国内各原厂大模型 Token 服务每月 TTFT 监测结果。数据显示，多数新发布的模型 Token 服务时延明显更高，大部分模型的 TTFT 大于 1 秒。



来源：中国信通院，2026 年 9 月

图 8 各原厂大模型 Token 服务 TTFT 平均值（越低越好）

图 9 为每月所有被监测大模型 Token 服务的 TTFT 平均值。数据显示，1 至 8 月 Token 服务的 TTFT 逐渐有所增加，8 月份 TTFT 均值已达到 1.26 秒，相较于 2025 年均值明显增加，这与模型参数量增大等因素有关系。6 月份以来，国外模型的 TTFT 平均值均高于 2 秒。

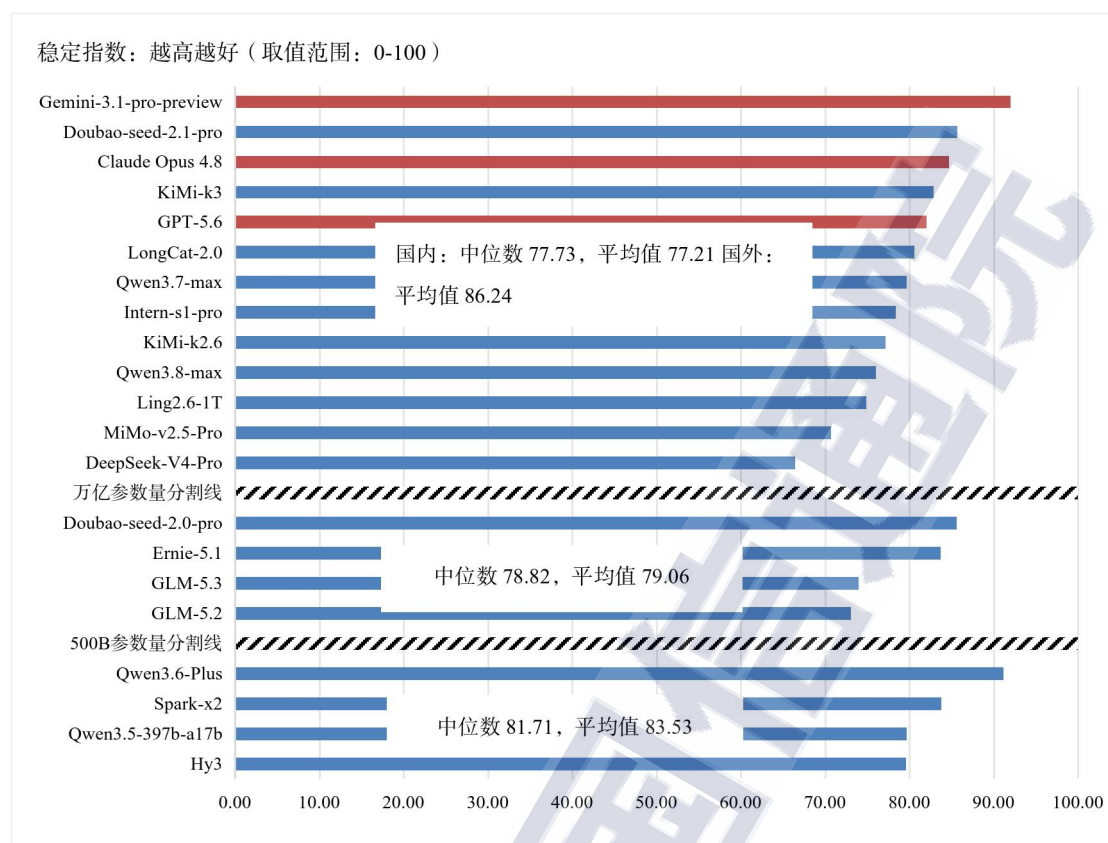


来源：中国信通院，2026 年 9 月

图 9 所有大模型 Token 服务的月均 TTFT

（四）服务稳定指数

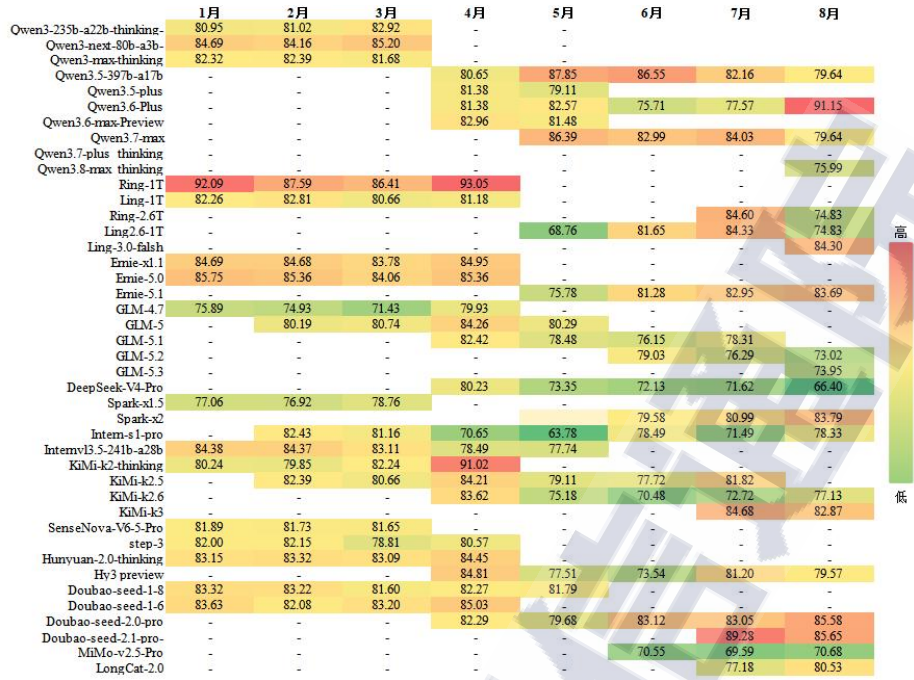
图 10 为 8 月份国内外各原厂大模型 Token 服务稳定指数监测结果。国内方面，最高稳定指数可达 91.15 分，约 38.9% 的模型达到 80 分以上优秀水平，Token 服务的输出和时延相对平稳。国外方面，Gemini-3.1-pro-preview 模型整体稳定性较好达到 92 分，而 GPT-5.6 稳定性指数也达到 82.3 分。



来源：中国信通院，2026 年 9 月

图 10 8 月份原厂大模型 Token 服务稳定指数

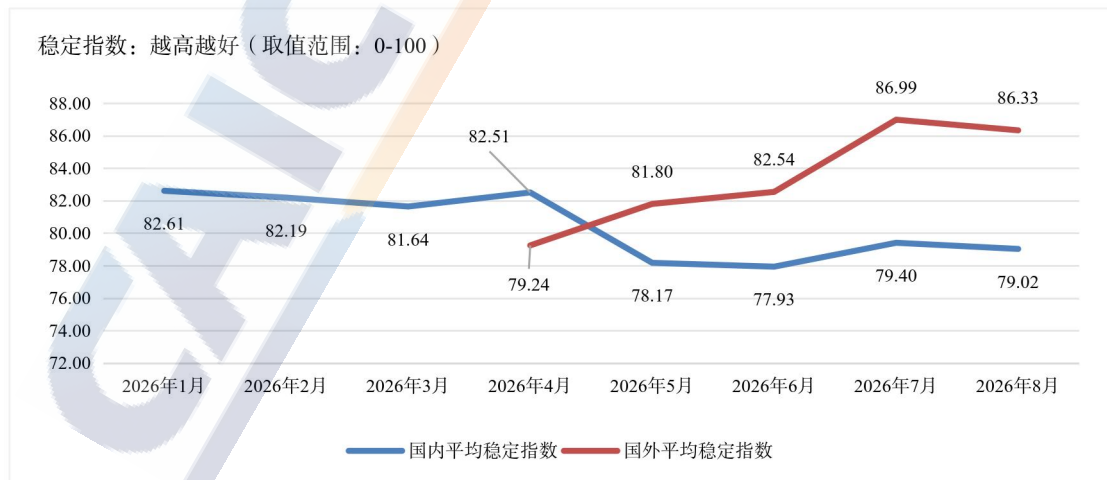
图 11 为每月各原厂大模型 Token 服务稳定性指数监测结果。数据显示，5 月份以来国内原厂大模型 Token 服务波动性明显，8 月份稳定性指数平均值为 79.02。而国外方面，8 月份稳定性指数均值达到 86.33 分，明显高于国内。



来源：中国信通院，2026 年 9 月

图 11 原厂大模型 Token 服务月均稳定性指数（越高越好）

图 12 为每月所有被监测大模型 Token 服务稳定性指数平均值。数据显示，1 至 4 月国内大模型 Token 服务稳定性保持平稳，均达到 80 分以上的较好水平，但 5 月份开始降低，主要由于部分热门模型在新版本发布后，每天不同时间段内 TPS 和 TTFT 变化较为剧烈。国外方面，稳定性持续上升，7 月份平均达到最高 86.99 分，高于国内。

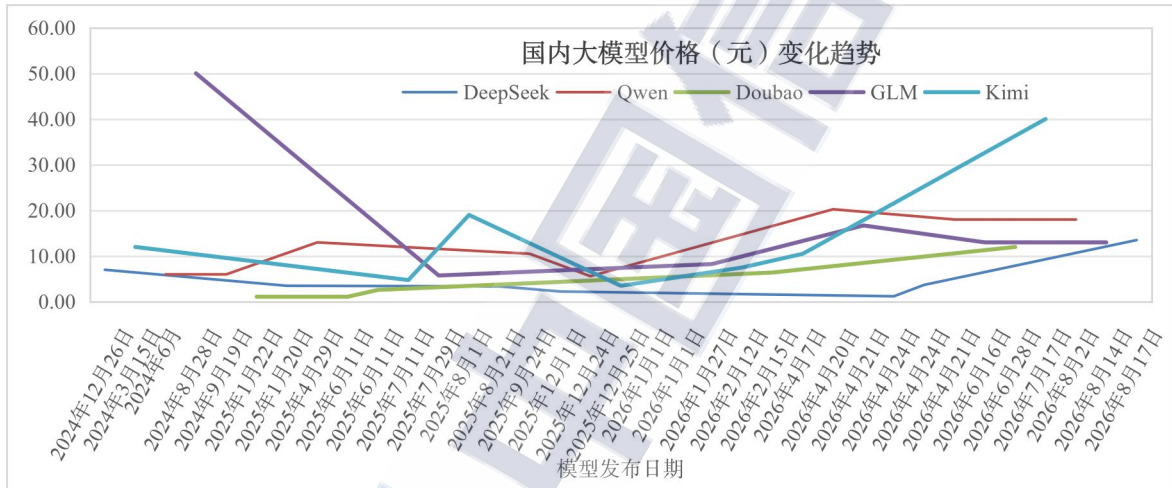


来源：中国信通院，2026 年 9 月

图 12 所有大模型 Token 服务月均稳定性指数

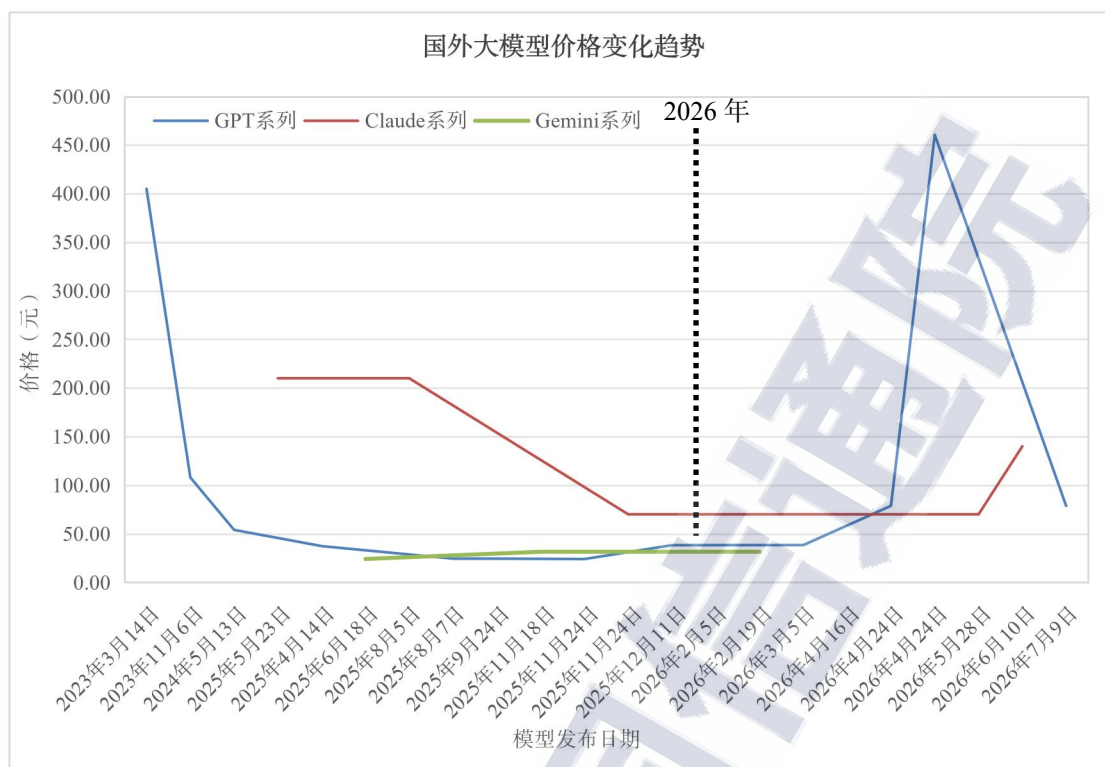
（五）服务价格与费用

图 13 和图 14 为国内外主流大模型归一化价格变化趋势。总体来看，2024 年以来每百万 Token 价格大幅度下降，国内上述 5 款原厂模型 2026 年以来新发布版本的 Token 价格普遍有所提升，归一化均价为 13.46 元，相比 2025 年下半年 6.8 元，涨幅达到 98%之多。同时，国内模型 Token 服务价格仍低于国外主流模型价格，2026 年以来上述 3 款国外模型新发布版本均价为 117.8 元。



来源：中国信通院，2026 年 9 月

图 13 国内原厂大模型发布时归一化价格趋势图



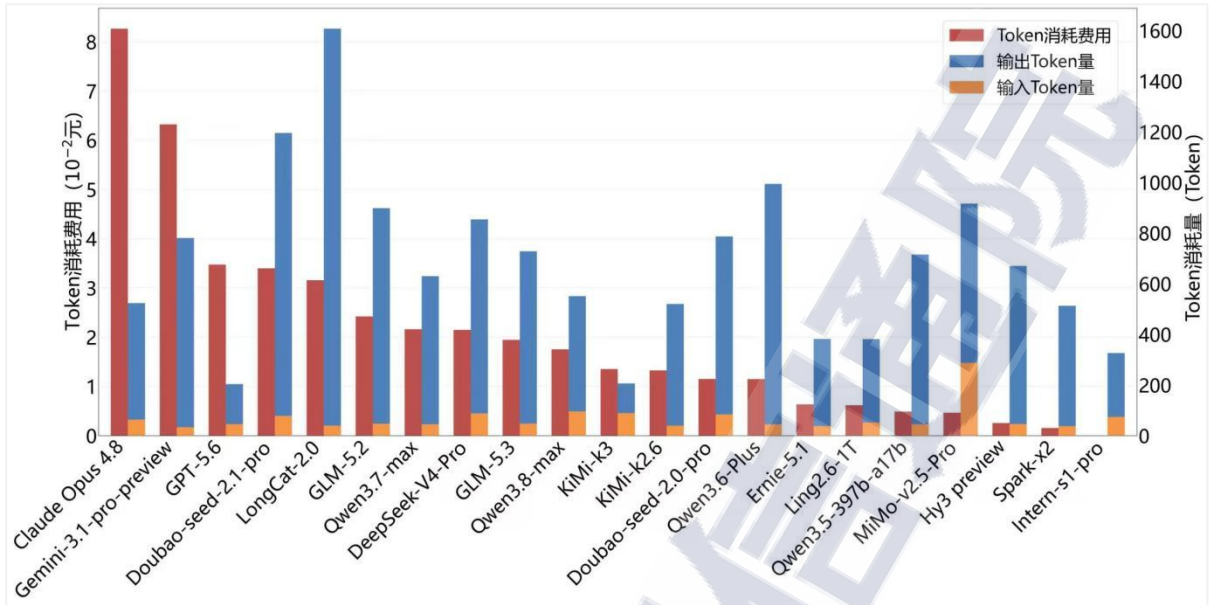
来源：中国信通院，2026 年 9 月

图 14 国外原厂大模型发布时归一化价格趋势图

图 15 为 8 月份开展 Token 服务性能监测过程中，原厂大模型平均每个任务的 Token 消耗量（包括输入和输出量）和费用。总体来看，一是国外模型推理过程消耗的 Token 量明显较少，但由于价格昂贵，整体消费金额更高，如 Claude 和 GPT 为消费最高的两款模型。二是国内模型普遍 Token 消耗量较高，由于价格便宜，整体消费金额偏低。三是在相同数据集上，不同模型所计量的输入 Token 量存在差异。

需要说明的是，由于性能监测的持续性，且过程中存在模型版本更换情况，所以各模型的监测次数并不完全一样，但监测数据集一致，因此上图数据展示的是平均每次调用后，模型返回提示用户所消耗的输入输出 Token 量，并根据价格计算的费用。同时，性能监测的数据集均为简单的开放性问题，与基准测试数据集差异较大，因此上述结

果是在本轮性能测试数据集上的表现。

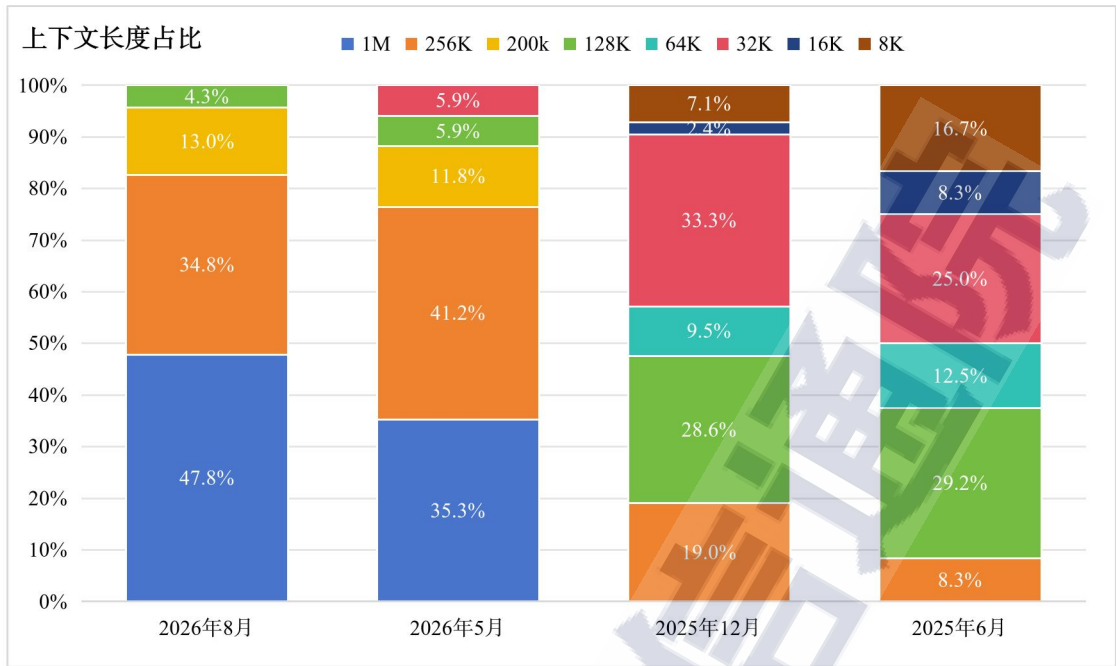


来源：中国信通院，2026 年 9 月

图 15 8 月份国内外大模型 Token 消耗量及费用对比图

（六）模型上下文

图 16 为国内大模型上下文长度占比情况。数据显示，上下文长度不断扩展仍然为主流趋势。一方面，2026 年新发布的模型上下文长度多数已达到 1M（即 100 万 Token），8 月份 47.8% 的模型上下文达到 1M，如 Qwen3.7、DeepSeek-V4-Pro 等。另一方面，长上下文占比快速上升，短上下文持续下降，其中 200K 及以上模型占比为 95.7%，128K 及以下的模型占比减少至 4.3%。国外方面，OpenAI、Anthropic 和 Google 的最新系列模型上下文长度均已达到 1M。



来源：中国信通院，2026 年 9 月

图 16 国内大模型上下文长度占比情况

四、开源大模型监测结果

（一）开源模型上线率

表 1 为典型开源大模型（不考虑版本）截至 2026 年 9 月份在国内外主流平台的上线率统计结果。数据显示，国产开源模型已成各平台的主力模型，在调研的 22 个国内外平台中，DeepSeek、Qwen 和 GLM 系列模型上线率达到 100%，Kimi、MiniMax 系列上线率分别达 95%、86%，以上五款国内开源相较于 5 月份统计结果有所上升。而 GPT-OSS、Llama 等国外开源模型更新速度较慢，上线率有所下降。

表 1 典型开源大模型在国内外主流 MaaS 平台的上线率统计

序号	平台名称	Deep Seek	Qwen	GLM	Kimi	Mini Max	GPT-OSS	Llama
1	AWS Bedrock	√	√	√	√	√	√	√
2	Google Cloud Vertex	√	√	√	√	√	√	√

表 1 典型开源大模型在国内外主流 MaaS 平台的上线率统计（续）

序号	平台名称	Deep Seek	Qwen	GLM	Kimi	Mini Max	GPT-OSS	Llama
3	Azure Foundry	√	√	√	√	√	√	√
4	OpenRouter	√	√	√	√	√	√	√
5	Together AI	√	√	√	√	√	√	√
6	Fireworks AI	√	√	√	√	√	√	√
7	Hugging Face	√	√	√	√	√	√	√
8	DeepInfra	√	√	√	√	√	√	√
9	Cloudflare AI Gateway	√	√	√	√	√	√	√
10	PPIO 派欧云	√	√	√	√	√	/	/
11	商汤大装置	√	√	√	/	/	/	/
12	无问芯穹	√	√	√	√	√	/	/
13	火山方舟	√	√	√	√	√	/	/
14	百度云千帆	√	√	√	√	/	/	/
15	硅基流动	√	√	√	√	√	/	/
16	讯飞星辰	√	√	√	√	√	/	/
17	阿里云百炼	√	√	√	√	√	/	/
18	腾讯云 TokenHub	√	√	√	√	√	/	/
19	华为云 MaaS	√	√	√	√	/	/	/
20	七牛云	√	√	√	√	√	/	/
21	优刻得	√	√	√	√	√	/	/
22	蓝耘云	√	√	√	√	√	/	/
9 月份上线率		100%	100%	100%	95%	86%	41%	41%
5 月份上线率		100%	100%	91%	86%	82%	50%	41%

来源：中国信通院，2026 年 9 月

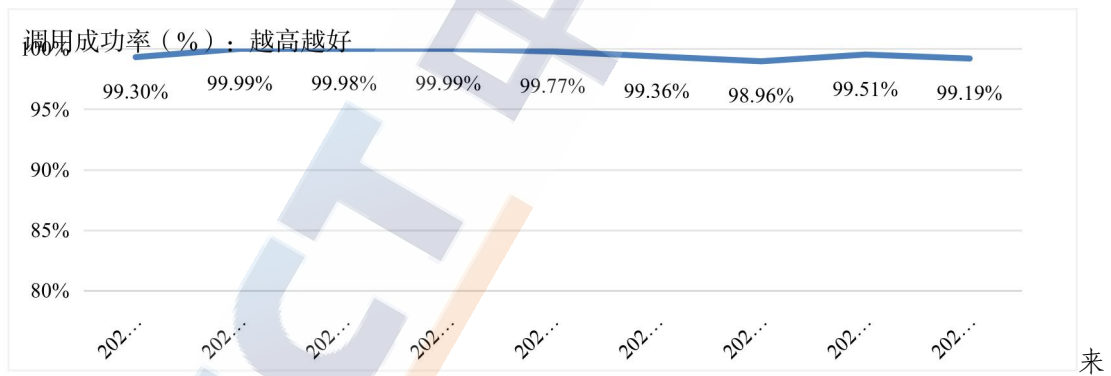
为衡量同一模型在不同平台上的性能表现，下面选择了 DeepSeek、Qwen、GLM、Kimi 四款模型，监测 2026 年 1 月至 8 月，其相同版本在各平台上的调用成功率、每秒输出词元数（TPS）和首词时延（TTFT）。

（二）DeepSeek 监测结果

监测对象为阿里云百炼、火山方舟、无问芯穹、腾讯云 TokenHub、PPIO、百度云千帆、硅基流动 7 个平台，以及深度求索官方所提供的 DeepSeek-V3.2 和 DeepSeek-V4-Pro 的 Token 服务，DeepSeek-V3.2 监测周期为 2026 年 1 至 4 月，5 月更换为 DeepSeek-V4-Pro，6 月份停止对无问芯穹平台的监测。

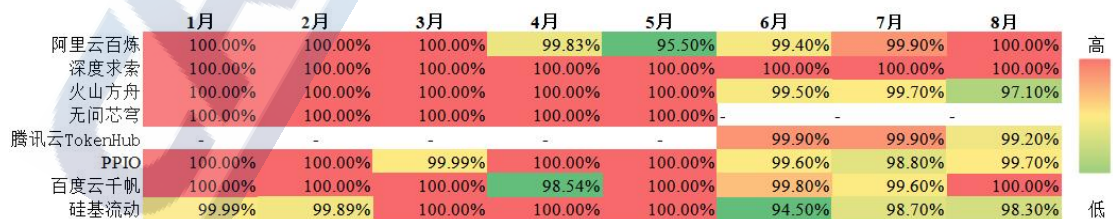
1. 调用成功率

根据下图数据，各月 DeepSeek 在所有平台上的调用成功率多数保持在 99% 以上，6 月份以后成功率稍降，8 月份平均调用成功率达到 99.19%。整体来看，深度求索官方平台的调用成功率保持在 100%，相较 2025 年的服务保障能力明显提升。



来源：中国信通院，2026 年 9 月

图 17 所有平台上 DeepSeek 月均调用成功率

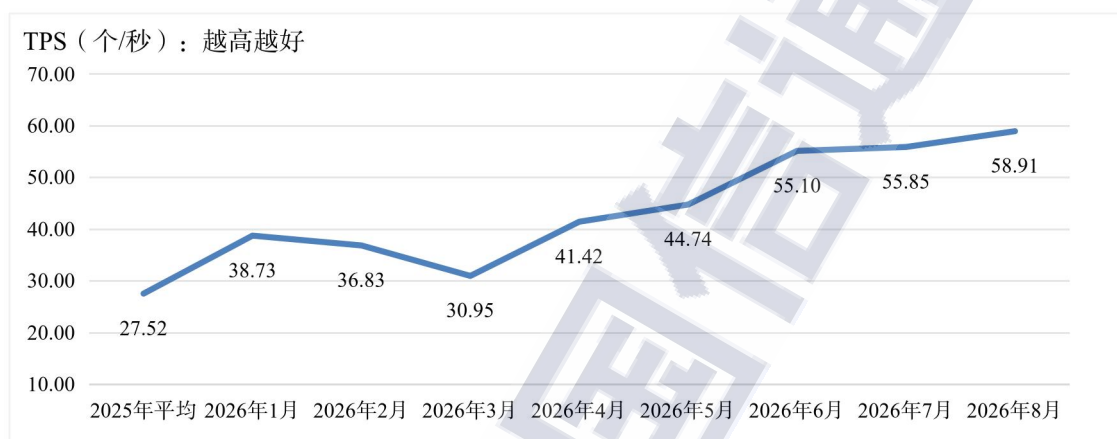


来源：中国信通院，2026 年 9 月

图 18 各平台的 DeepSeek 月均调用成功率（越高越好）

2. 每秒输出词元数（TPS）

根据下图数据显示，3 月份各平台上 DeepSeek 模型 TPS 均值经历短暂下跌后又开始提升，5 月份切换到 DeepSeek-V4-Pro 模型后 TPS 上升显著，8 月份绝大多数平台保持在 50（个/秒）以上，均值达 58.91（个/秒），提升较为明显。深度求索官方平台的 TPS 超过 60。



来源：中国信通院，2026 年 9 月

图 19 所有平台的 DeepSeek 月均 TPS

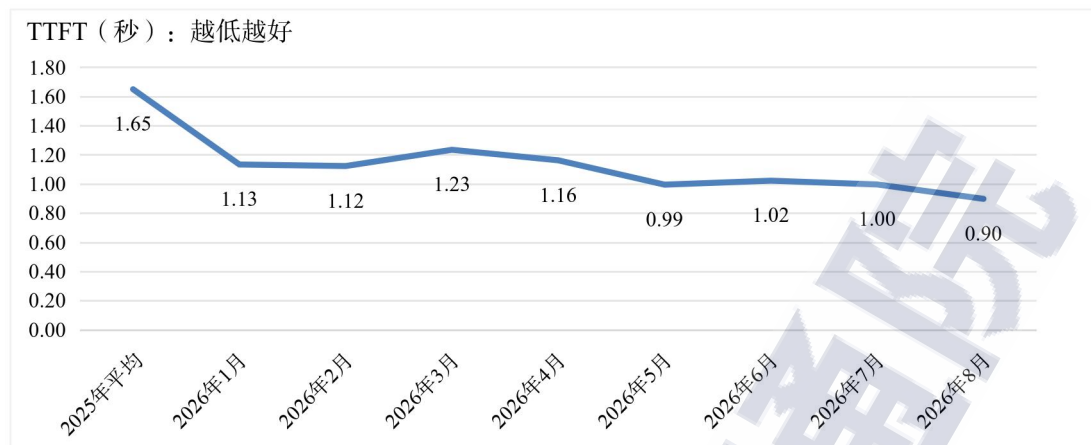


来源：中国信通院，2026 年 9 月

图 20 各平台的 DeepSeek 月均 TPS（越高越好）

3. 首词时延（TTFT）

根据下图数据，3 月份各平台上 DeepSeek 模型 TTFT 均值明显升高，达到 1.23 秒，之后持续回落，5 月份切换到 DeepSeek-V4-Pro 模型后 TTFT 下降显著，8 月份为 0.9 秒。深度求索官方平台的服务时延最低为 0.56 秒。



来源：中国信通院，2026 年 9 月

图 21 所有平台的 DeepSeek 月均 TTFT

	1月	2月	3月	4月	5月	6月	7月	8月	
阿里云百炼	1.15	1.15	1.29	1.12	0.81	1.17	1.32	1.13	高
深度求索	1.39	1.22	1.45	1.15	0.47	0.28	0.38	0.56	
火山方舟	1.03	1.06	1.18	1.09	1.22	1.92	1.56	1.20	
无问芯穹	1.44	1.60	1.56	1.52	1.99	-	-	-	
腾讯云TokenHub	-	-	-	-	-	0.96	1.22	1.18	
PPIO	0.79	0.78	1.06	1.12	0.71	0.43	0.48	0.59	
百度云千帆	0.67	0.67	0.90	0.94	0.67	0.57	0.64	0.71	
硅基流动	1.46	1.37	1.20	1.18	1.10	1.83	1.37	0.92	低

来源：中国信通院，2026 年 9 月

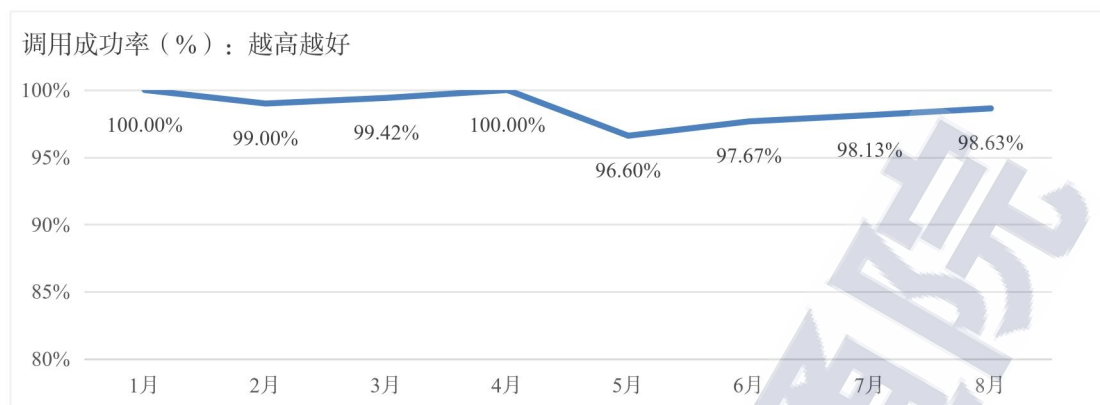
图 22 各平台的 DeepSeek 月均 TTFT（越低越好）

（三）Qwen 监测结果

监测对象为阿里云百炼、无问芯穹、百度云千帆、硅基流动 4 个平台上 Qwen3-235B-A22B 和 Qwen3.5-397B-A17B 的 Token 服务，Qwen3-235B-A22B 监测周期为 2026 年 1 至 4 月，5 月更换为 Qwen3.5-397B-A17B。

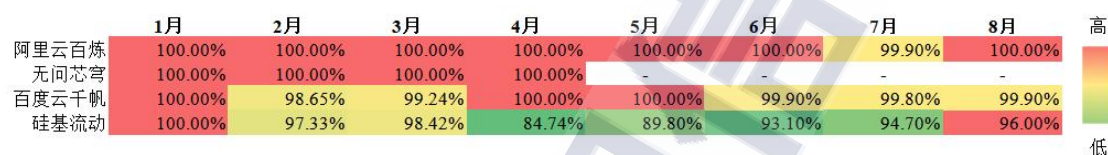
1. 调用成功率

根据下图数据，阿里云百炼官方平台的调用成功率基本保持在 100%，硅基流动平台自 2 月以来调用成功率波动较大，8 月份成功率有所回升达到 96%，其中“请求超时”报错较多。



来源：中国信通院，2026 年 9 月

图 23 平台的 Qwen 月均调用成功率

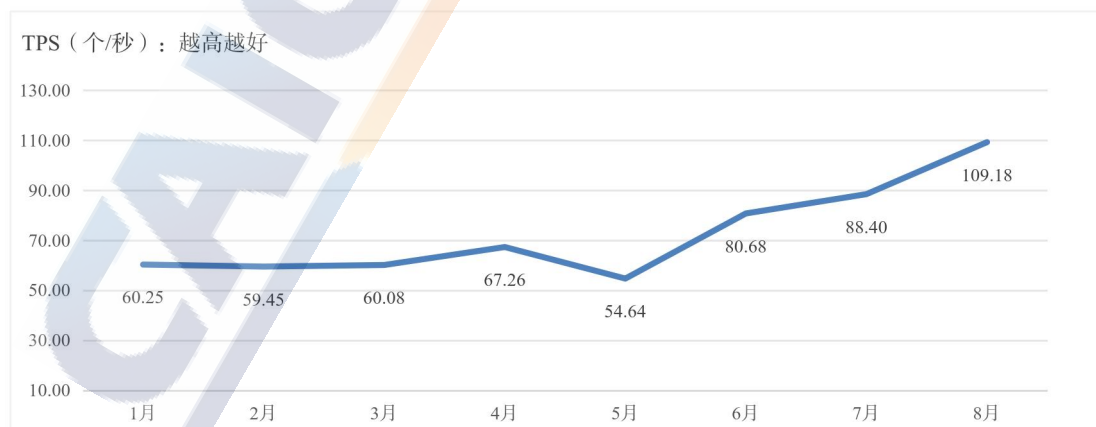


来源：中国信通院，2026 年 9 月

图 24 各平台的 Qwen 月均调用成功率（越高越好）

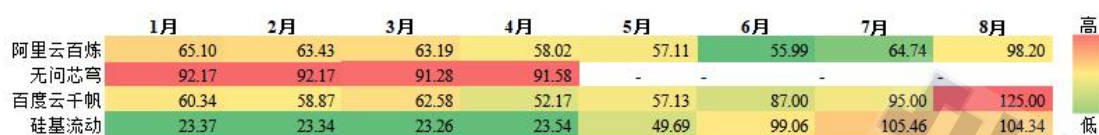
2. 每秒输出词元数（TPS）

根据下图数据，各平台上 Qwen 模型 5 月份 TPS 跌至低谷，5 月份切换到 Qwen3.5-397B-A17B 后 TPS 显著提升，6 月份平均值达到 80.68（个/秒），8 月份硅基流动、百度云千帆的 TPS 已达到 100（个/秒）以上，超过阿里云百炼官方平台。



来源：中国信通院，2026 年 9 月

图 25 所有平台的 Qwen 月均 TPS

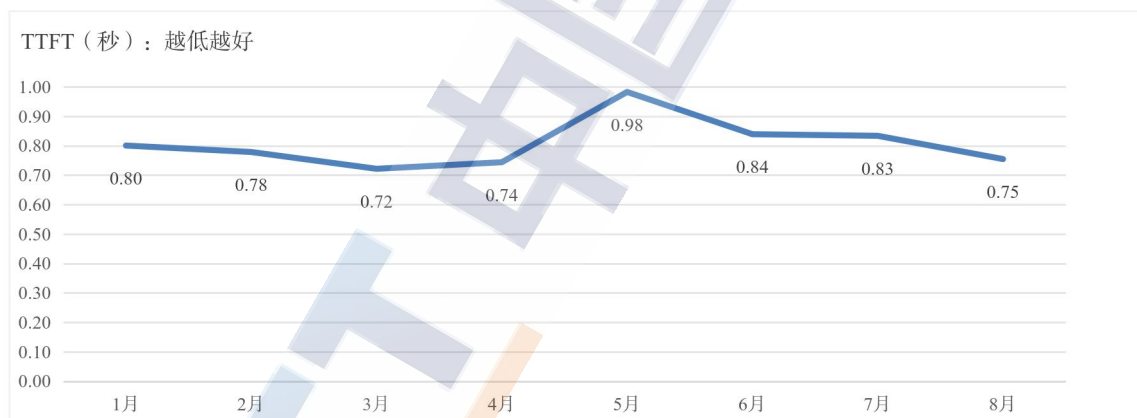


来源：中国信通院，2026 年 9 月

图 26 各平台的 Qwen 月均 TPS（越高越好）

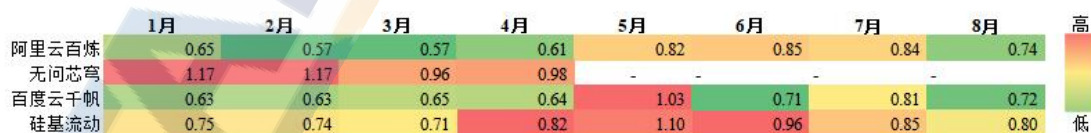
3. 首词时延（TTFT）

根据下图数据，各平台上 Qwen 模型 TTFT 平均值前 4 个月保持下降趋势，处于 0.7 至 0.8 秒之间。5 月份监测的模型更换为 Qwen3.5-397B-A17B，TTFT 普遍升高，均值高达 0.98 秒，之后逐步改善，8 月份均值降至 0.75 秒，阿里云百炼降至 0.74 秒，百度云千帆最低为 0.72 秒。



来源：中国信通院，2026 年 9 月

图 27 所有平台的 Qwen 月均 TTFT



来源：中国信通院，2026 年 9 月

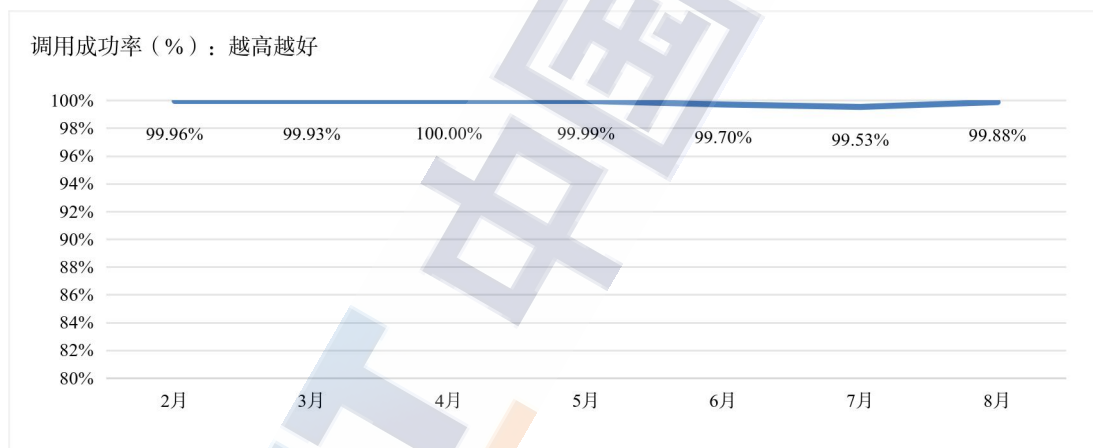
图 28 各平台的 Qwen 月均 TTFT（越低越好）

（四）GLM 监测结果

监测对象为智谱、无问芯穹、PPIO、百度云千帆、硅基流动 5 个平台上 GLM-5 和 GLM-5.1 的 Token 服务, GLM-5 监测周期为 2026 年 2 月至 4 月, 5 月更换为 GLM-5.1。

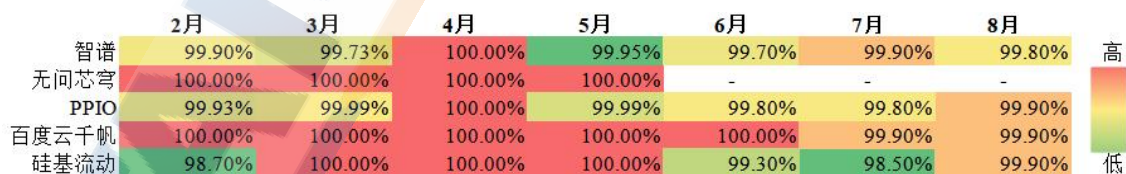
1. 调用成功率

根据下图数据, GLM 模型调用成功率整体较为平稳, 智谱官方平台在 4 月调用成功率达到 100%, 5 月份切换到 GLM-5.1 后, 调用成功率出现微降, 但是多数可以达到 99.9% 左右, 8 月份平均调用成功率达到 99.88%。



来源：中国信通院，2026 年 9 月

图 29 所有平台的 GLM 月均调用成功率



来源：中国信通院，2026 年 9 月

图 30 各平台的 GLM 月均调用成功率（越高越好）

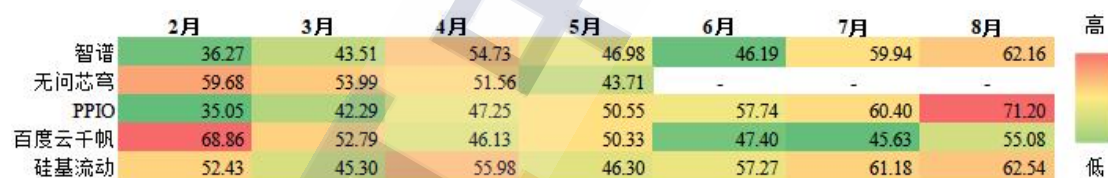
2. 每秒输出词元数（TPS）

根据下图数据，GLM 自 6 月份以来平均 TPS 保持上升态势。4 月份之前 GLM-5 处于 49（个/秒）上下，5 月份切换 GLM-5.1 后 TPS 稍有降低，均值为 47.58（个/秒），6 月份起开始逐步提升，8 月份平均 TPS 达到 62.75（个/秒），智谱官方平台 TPS 为 62.16（个/秒），PPIO 最高达到 71.2（个/秒）。



来源：中国信通院，2026 年 9 月

图 31 所有平台的 GLM 月均 TPS

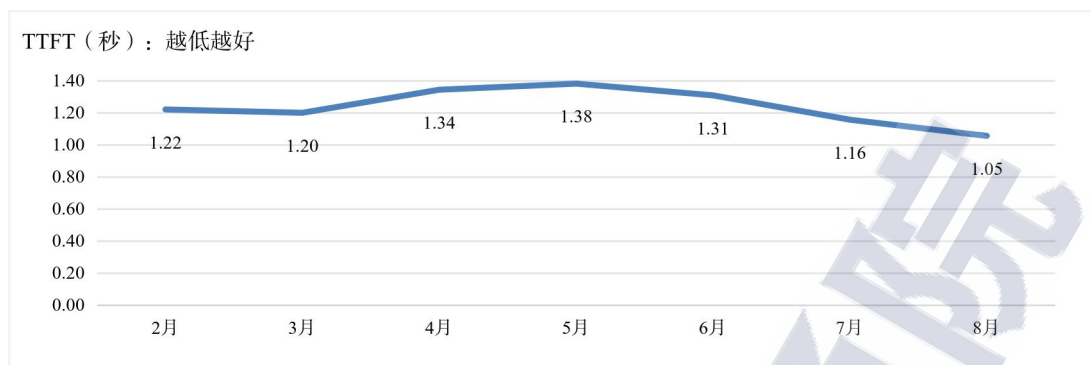


来源：中国信通院，2026 年 9 月

图 32 各平台的 GLM 月均 TPS（越高越好）

3. 首词时延（TTFT）

根据下图数据，自 6 月份以来 GLM 在各平台的 TTFT 较之前持续降低，8 月份均值为 1.05 秒。但智谱官方平台的时延相对较高，8 月份最高达到 1.69 秒。



来源：中国信通院，2026 年 9 月

图 33 所有平台的 GLM 月均 TTFT



来源：中国信通院，2026 年 9 月

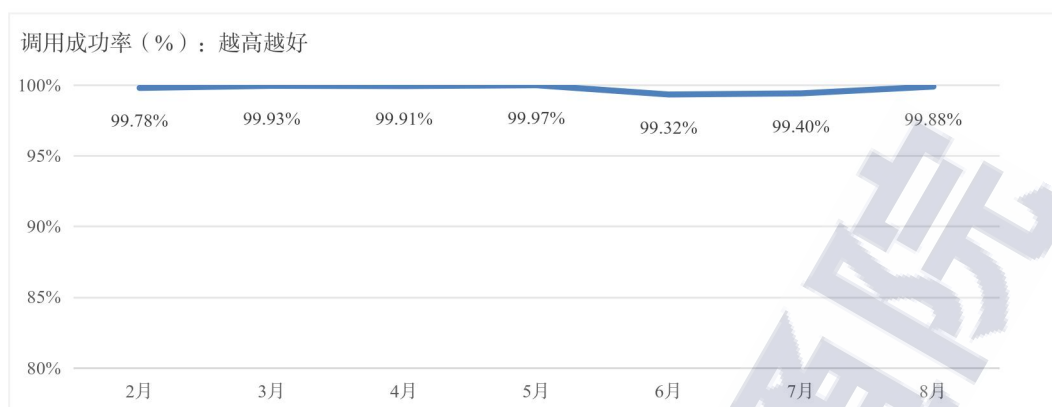
图 34 各平台的 GLM 月均 TTFT（越低越好）

（五）Kimi 监测结果

监测对象为无问芯穹、PPIO、百度云千帆、硅基流动 4 个 MaaS 平台，以及月之暗面官方所提供的 Kimi-K2.5 和 Kimi-K2.6 的 Token 服务，Kimi-K2.5 监测周期为 2026 年 2 月至 4 月，5 月更换为 Kimi-K2.6。

1. 调用成功率

根据下图数据，各平台上 Kimi 模型每月调用成功率相对平稳，8 月份各平台平均调用成功率为 99.88%。月之暗面官方平台的调用成功率前三个月保持在 100%，5 月份切换为 Kimi-K2.6 后，调用成功率均稍有降低。但是阿里云百炼平台的成功率基本保持在 100%。



来源：中国信通院，2026 年 9 月

图 35 所有平台的 Kimi 月均调用成功率

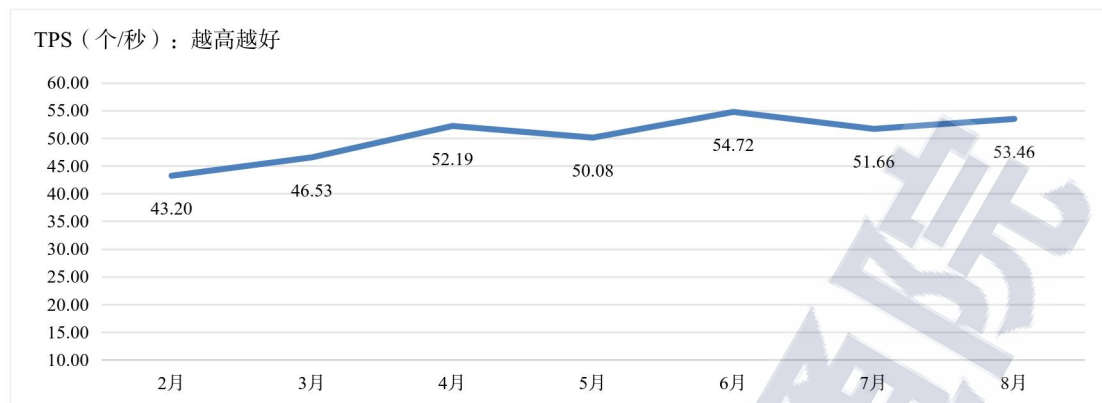


来源：中国信通院，2026 年 9 月

图 36 各平台的 Kimi 月均调用成功率（越高越好）

2. 每秒输出词元数（TPS）

根据下图数据，各平台上 Kimi 的月均 TPS 总体呈现提升趋势，4 月份提升明显，达到 52.19（个/秒），相比 2 月份涨幅 20.8%，5 月份切换为 Kimi-K2.6 后 TPS 未出现明显波动。但是阿里云百炼平台的 TPS 保持较高，8 月均值达到 73.06（个/秒），而月之暗面官方服务的 TPS 相对较低，仅为 44.87（个/秒）。



来源：中国信通院，2026 年 9 月

图 37 所有平台的 Kimi 月均 TPS

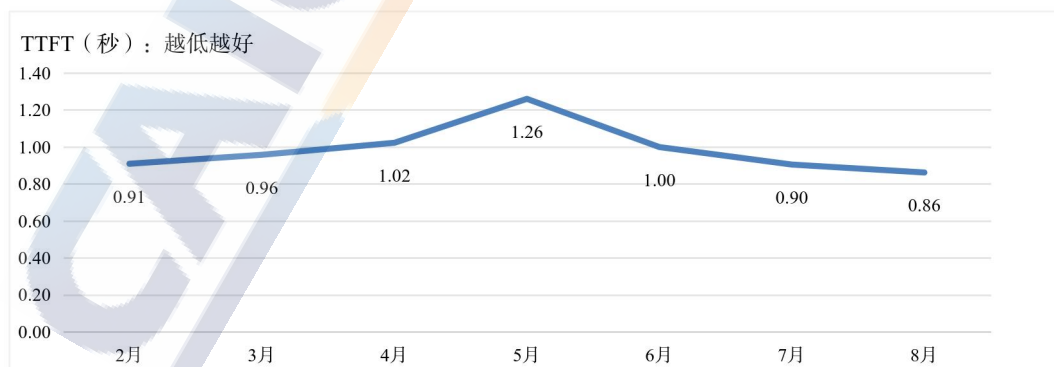


来源：中国信通院，2026 年 9 月

图 38 各平台的 Kimi 月均 TPS（越高越好）

3. 首词时延（TTFT）

根据下图数据，各平台上 Kimi 的月均 TTFT 自 5 月份切换为 Kimi-K2.6 后，TTFT 呈现下降趋势，月均 TTFT 从 1.26 秒下降至 8 月份的 0.86 秒，百度云千帆平台最低为 0.56 秒，月之暗面官方服务的时延持续相对较高，8 月份达到 1.22 秒。



来源：中国信通院，2026 年 9 月

图 39 所有平台的 Kimi 月均 TTFT



来源：中国信通院，2026 年 9 月

图 40 各平台的 Kimi 月均 TTFT（越低越好）

五、总结

整体来看，我国原厂模型在调用成功率、输出速度与响应时延等关键性能指标上表现稳健，且呈现出螺旋上升的趋势。随着模型参数量、上下文长度的增加，模型时延有较明显的提升，但各平台通过优化推理架构与调度策略，吞吐的提升有效缓解了性能压力。开源模型在第三方 MaaS 平台上的性能优化尤为明显，部分平台的性能已超过官方。然而，开源模型在不同平台的性能差异也反映出平台服务能力与资源配置的不均衡，行业整体大模型服务性能仍有较大提升空间。

未来，中国信通院将持续优化大模型 Token 服务监测体系，扩大监测范围，加快监测能力建设，并扩充产业服务能力，致力于推动大模型 Token 服务高质量发展。一方面，持续完善监测指标与方法，提升监测结果的科学性与权威性，为产业各方提供更加客观、全面的参考依据。另一方面，深化与主流云厂商及模型提供方的协同合作，推动服务性能的横向对齐与最佳实践的行业共享，助力产业整体服务水平的提升。同时，也将进一步探索多模态模型、长上下文场景下的性能评估方法，适时发布相关监测结果，为产业界提供更及时、更精准的决策支撑。我们诚挚欢迎产业各方积极参与，共同完善大模型 Token 服务监测体系，共同推动我国大模型服务能力迈向更高水平。

附录 A：其他事项说明

1. 本监测结果仅供参考，由于监测频率、监测时长等维度的限制，监测结果并不能作为衡量公有云大模型服务性能以及大模型 Token 服务平台能力的绝对依据。且影响 TTFT 和 TPS 的因素较多，如模型尺寸、网络时延、用户数量、算力及其调度能力、模型能力等。
2. 监测基础大语言模型所使用的数据集来自常见的大语言模型基准测试集，构成固定题库。监测时从题库中随机选择题目，且为了减少请求时命中缓存的情况，为每个提示词添加了随机干扰字符。
3. 监测方法包括每日监测法和人工监测法：
 - 每日监测法，是指通过每日定点的持续性自动化监测，衡量大模型 Token 服务持续的稳定性，包括 TTFT 和 TPS。该方法选择位于不同国家和不同地区的云节点，于每日选择不同的 5 个时间整点，向所有大模型 API 连续发送多次请求，并计算其 TTFT 和 TPS。调用成功率的监测频率于 2026 年 5 月下旬更新为每 5 分钟发送一次请求。
 - 人工监测法，是指通过人工对各平台披露的大模型 Token 服务的输入输出价格、上下文等信息进行统计更新。
4. TPS 计算公式如下。

$$TPS = \frac{N_t}{RT - TTFT} \quad (A-1)$$

式中：

N_t ——输出的词元总数；

RT ——从请求开始到返回完全部词元之间的时延。

5. 模型价格归一化计算公式如下。

$$PRC = \frac{(3 \times P_i + P_o)}{4} \quad (A-2)$$

式中：

P_i ——输入每百万词元的价格；

P_o ——输出每百万词元的价格。

6. 服务稳定指数，是指持续平稳输出 Token 服务的能力，取值范围为(0,100]，值越接近 100，表明服务波动越小。稳定性指数由不同指标的加权构成，计算公式如下：

$$CV = \frac{\sigma}{\mu} \quad (A-3)$$

$$Stability = \frac{1}{1+CV} \times 100 \quad (A-4)$$

$$S = \sum_{i=1}^n \omega_i \times Stability \quad (A-5)$$

式中：

σ ——为标准差；

μ ——为平均值；

CV ——为变异系数，也称为离散系数，是标准差与平均值的比值，服务稳定指数中，进行了数值归一化处理；

ω_i ——为各指标权重，可依据实际服务场景各指标的重要性进行动态分配。本次监测选择了 TPS 和 TTFT 两项指标并分配同等权重。

7. 所有大模型 Token 服务的调用统一采用流式输出，其余设置均采用默认值。

附录 B：监测模型清单

表 2 原厂模型监测清单

序号	平台名称	模型名称	发布日期	开始监测时间	停止监测时间	上下文长度	模型参数量 (B)	公开模式
1	蚂蚁百灵	Ring-1T	2025 年 10 月 14 日	2026 年 1 月	2026 年 4 月	128k	1000	开源
		Ling-1T	2025 年 10 月 9 日	2026 年 1 月	2026 年 4 月	128k	1000	开源
		Ring-2.6-1T	2026 年 5 月 9 日	2026 年 7 月	监测中	256k	1000	开源
		Ling-2.6-1T	2026 年 4 月 23 日	2026 年 5 月	监测中	256k	1000	开源
		Ling-3.0-Flash	2026 年 7 月 24 日	2026 年 8 月	监测中	256k	124	开源
2	阿里云百炼	Qwen3-235b-a22b-thinking-2507	2025 年 7 月 25 日	2026 年 1 月	2026 年 3 月	128k	235	开源
		Qwen3-next-80b-a3b-thinking	2025 年 9 月 12 日	2026 年 1 月	2026 年 3 月	128k	80	开源
		Qwen3-max-thinking	2025 年 9 月 24 日	2026 年 1 月	2026 年 3 月	256k	1000	开源
		Qwen3.5-plus	2026 年 2 月 16 日	2026 年 4 月	监测中	1M	未公开	闭源
		Qwen3.6-Plus	2026 年 4 月 2 日	2026 年 4 月	监测中	1M	未公开	闭源
		Qwen3.6-max-Preview	2026 年 4 月 20 日	2026 年 4 月	监测中	1M	未公开	开源
		Qwen3.5-397b-a17b	2026 年 2 月 16 日	2026 年 4 月	监测中	1M	397	开源
		Qwen3.7-max	2026 年 5 月 20 日	2026 年 5 月	监测中	1M	未公开	闭源
		Qwen3.8-max	2026 年 8 月 2 日	2026 年 8 月	监测中	1M	2400	闭源
3	深度求索	DeepSeek-V4-Pro	2026 年 4 月 24 日	2026 年 4 月	监测中	1M	1600	开源
4	书生	Intern-s1-pro	2026 年 2 月 4 日	2026 年 2 月	监测中	256k	1000	开源
		InternV1.5-241b-a28b	2025 年 9 月 3 日	2026 年 1 月	监测中	32K	241	开源
5	月之暗面	Kimi-k2-thinking	2025 年 7 月 11 日	2026 年 1 月	2026 年 4 月	256k	1000	开源
		Kimi-k2.5	2026 年 1 月 27 日	2026 年 2 月	监测中	256k	1000	开源
		Kimi-k2.6	2026 年 4 月 20 日	2026 年 4 月	监测中	256k	1000	开源
6	百度云千帆	Ernie-x1.1	2025 年 9 月 9 日	2026 年 1 月	2026 年 4 月	64k	300	闭源
		Ernie-5.0-thinking	2025 年 11 月 13 日	2026 年 1 月	2026 年 4 月	128k	2400	闭源
		Ernie-5.1-thinking	2026 年 5 月 9 日	2026 年 5 月	监测中	128k	800	闭源
7	商汤大装置	SenseNova-V6-5-Pro	2025 年 7 月 25 日	2026 年 1 月	2026 年 3 月	128k	560	闭源
8	阶跃星辰	step-3	2025 年 7 月 25 日	2026 年 1 月	2026 年 4 月	64k	321	闭源
9	腾讯云 TokenHub	Hunyuan-2.0	2025 年 11 月 9 日	2026 年 1 月	2026 年 4 月	256k	406	闭源
		Hy3	2026 年 4 月 23 日	2026 年 4 月	监测中	256k	295	闭源
10	火山方舟	Doubao-seed-1-8	2025 年 12 月 18 日	2026 年 1 月	2026 年 5 月	256k	200	闭源
		Doubao-seed-1-6	2025 年 6 月 11 日	2026 年 1 月	2026 年 4 月	256k	200	闭源
		Doubao-seed-2.0-pro	2026 年 2 月 15 日	2026 年 4 月	监测中	256k	未公开	闭源
		Doubao-seed-2.1-pro	2026 年 6 月 28 日	2026 年 7 月	监测中	256k	未公开	闭源
11	讯飞星辰	Spark-x1.5	2025 年 11 月 6 日	2026 年 1 月	2026 年 3 月	192k	未公开	闭源
12	智谱	GLM-4.7	2025 年 12 月 25 日	2026 年 1 月	2026 年 2 月	200k	355	开源
		GLM-5	2026 年 2 月 12 日	2026 年 2 月	2026 年 7 月	200k	744	开源
		GLM-5.1	2026 年 4 月 7 日	2026 年 4 月	2026 年 8 月	200k	744	开源
		GLM-5.2	2026 年 6 月 16 日	2026 年 6 月	监测中	1M	744	开源

表 2 原厂模型监测清单（续表）

序号	平台名称	模型名称	发布日期	开始监测时间	停止监测时间	上下文长度	模型参数量(B)	公开模式
12	智谱	GLM-5.3	2026 年 8 月 14 日	2026 年 8 月	监测中	1M	743	开源
		GLM-5.3	2026 年 8 月 14 日	2026 年 8 月	监测中	1M	743	开源
13	小米	MiMo-v2.5-Pro	2026 年 4 月 28 日	2026 年 4 月	监测中	1M	1020	开源
14	美团	LongCat-2.0	2026 年 6 月 30 日	2026 年 7 月	监测中	1M	未公开	闭源
15	OpenAI	GPT-5.5	2026 年 4 月 16 日	2026 年 5 月	监测中	1M	未公开	闭源
		GPT-5.6 Sol	2026 年 7 月 9 日	2026 年 7 月	监测中	1M	未公开	闭源
16	Anthropic	Claude Opus 4.7	2026 年 4 月 16 日	2026 年 5 月	监测中	1M	未公开	闭源
17	Google Cloud	Gemini-3.1-pro-preview	2026 年 2 月 19 日	2026 年 5 月	监测中	1M	未公开	闭源

表 3 不同平台上开源模型最新监测清单

序号	模型名称	监测平台	开始监测时间	停止监测时间	备注
1	DeepSeek-V4-Pro	深度求索	2026 年 2 月	监测中	1-4 月监测的版本为 DeepSeek-V3.2
		阿里云百炼	2026 年 2 月	监测中	
		火山方舟	2026 年 2 月	监测中	
		无问芯穹	2026 年 2 月	2026 年 5 月	
		PPIO	2026 年 2 月	监测中	
		百度云千帆	2026 年 2 月	监测中	
		腾讯云 TokenHub	2026 年 6 月	监测中	
		硅基流动	2026 年 2 月	监测中	
2	Qwen3.5-397B-A17B	阿里云百炼	2026 年 2 月	监测中	1-4 月监测的版本为 Qwen3-235B-A22B
		无问芯穹	2026 年 2 月	2026 年 5 月	
		百度云千帆	2026 年 2 月	监测中	
		硅基流动	2026 年 2 月	监测中	
3	GLM-5.1	智谱	2026 年 2 月	监测中	1-4 月监测的版本为 GLM-5
		无问芯穹	2026 年 2 月	2026 年 5 月	
		PPIO 派欧云	2026 年 2 月	监测中	
		百度云千帆	2026 年 2 月	监测中	
		硅基流动	2026 年 2 月	监测中	
4	Kimi-K2.6	月之暗面	2026 年 2 月	监测中	1-4 月监测的版本为 Kimi-K2.5
		阿里云百炼	2026 年 2 月	监测中	
		无问芯穹	2026 年 2 月	2026 年 5 月	
		PPIO 派欧云	2026 年 2 月	监测中	
		百度云千帆	2026 年 6 月	监测中	
		硅基流动	2026 年 2 月	监测中	

中国信息通信研究院 人工智能研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62301618

传真：010-62301618

网址：www.caict.ac.cn

