

多模型 Token 供给能力与性能验收测试报告

GLM-5.2 • MiniMax-M3 • Kimi K2.7 • Kimi K3(单机 8 卡 • 对外正式版 V2.3)

报告编号:TPM-2026-001-V2.3 | 测试日期:2026-08-18 ~ 08-19 | 密级:商密(仅限合作方) | 全部指标为交付态实机实测;标 * 项为按统一画像由实测锚换算 | V2.3 变更:GLM-5.2 主池升级为 128K+MTP 投机解码配置并全实测(总TPM 32.6 万, §4.1), 256K 长文以专用池承接(双池, §4.1 注)

一、报告说明

本报告面向流量合作方, 给出我方单台推理服务器(8×AMD Instinct MI308X, 单卡 192GB)上四个开源模型服务的 token 供给产能、并发承压、服务等级与规模承接能力, 用于流量外采/API 转售合作的验收评估。所有吞吐、延迟数据均为交付配置下的实机压测结果, 测试全程 0 错误;可安排合作方按第十一节方法现场复测。V2 版新增:各模型相对官方默认部署方案(或可研基准情景)的提升比例, 以及优化带来的经济性效益量化。

二、核心结论(供给能力总表)

模型	单机最高总TPM(输入+输出)	其中输出TPM	单机日总产能	较官方/基准方案	定位
MiniMax-M3	29.3 万(实测)	7.2 万	4.22 亿 token	方案级 13.7 倍(§六)	走量主力, 单流最快
GLM-5.2(128K+MTP 主池)	32.6 万(实测)	8.2 万	4.70 亿 token	+51%~+104%, 缓存+65%, MTP+7%	旗舰质量, 高价值流量; 256K 长文池另配(25.4 万口径)
Kimi K2.7	≈12.2 万*	3.0 万	≈1.76 亿*	+124%(vs 可研基准)	长上下文/代码
MiniMax-M2	≈9.9 万*	7.6 万	≈1.42 亿*	达盈亏线 1.17 倍	轻量高性价比
Kimi K3	专线口径	0.2 万	≈320 万(输出)	9.2 倍(4.06→37.25 tok/s)	超长上下文单流专线

总TPM 口径:输入+输出 token 全计(输入含缓存命中部分, 按缓存价计费);命中型画像 = 输入:输出≈3:1、高前缀命中(Agent/企业客服等典型流量)。M3 与 GLM-5.2 为该画像直接实测(N=256/N=128 档);标 * (K2.7/M2)为由实测吞吐锚换算。提升比例出处见 §六。

三、MiniMax-M3 分档承压实测(重点模型)

3.1 解码型画像(短输入, 输出 512 tok/请求)

并发N	聚合输出吞吐	RPM(请求/分)	p50 完成延迟	错误率
1	114 tok/s	16	3.9 s	0
8	425 tok/s	56	8.4 s	0
32	998 tok/s	122	14.0 s	0
128	1671 tok/s(≈10.0万输出TPM)	208	32.2 s	0
256(批量档)	1720 tok/s(≈10.3万输出TPM)	222	72.6 s	0

3.2 命中型画像(输入:输出≈3:1, 高前缀命中)

并发N	总TPM	其中输出	RPM	p50 完成延迟	错误率
64	16.2 万	4.0 万	78	44.4 s	0
128	24.5 万	6.0 万	119	58.6 s	0
256(批量档)	29.3 万	7.2 万	141	90.9 s	0

建议接入档位:交互型流量按 N≤128 承接(p50 ≤ 60s);批量/离线任务可用 N=256 档拉满产能。

3.3 方案演进对比(同机实测, 官方方案 → 我方交付方案)

方案	聚合输出吞吐峰值	单流(短输入/1k)	并发扩展性
官方参考部署(BF16·TP8·SGLang)	122.4 tok/s(并发32饱和)	69.4 tok/s	并发8~16即饱和
MXFP8 量化(SGLang)	149.1 tok/s	—	同上
MXFP8 + vLLM 引擎	165.8 tok/s	—	并行度提升 1.6 倍
我方交付方案(vLLM+调优+投机解码+前缀缓存)	1671 tok/s(N=128)	113.8 tok/s(+64%)	N=256 仍在爬升

口径注:前三行取自同机历史实测报告(8k 输入画像的最优操作点);末行为本轮交付方案在解码型画像的最优操作点。方案级 13.7 倍 (122.4→1671)反映"引擎+量化+并发扩展+投机解码+缓存"的综合效果;单流 +64% 为近似同口径可比项。KV 缓存净贡献另见 §六(同画像 A/B:总TPM +37%)。

四、GLM-5.2 / Kimi K2.7 验收锚(已通过 48k 真实负载验收)

模型	验收口径	聚合输出吞吐	较优化前/基准	TTFT(首token)	上下文
GLM-5.2	N=128 混合命中(交付锚)	870 tok/s	+51%(577→870)	冷 1.56s / 命中 0.13s	256K
GLM-5.2	纯共享命中(上限)	1178 tok/s	+104%	命中 0.13s	256K
Kimi K2.7	48k 真实负载验收	497 tok/s	+124%(基准222→497)	48k冷 29s / 命中 0.37s	256K
Kimi K3	超长上下文单流专线	37 tok/s(单流)	9.2 倍(4.06→37.25)	冷 3.95s / 命中 0.76s	详见K3专项报告

4.1 GLM-5.2 命中型画像分档实测(主池 128K+MTP 投机解码配置, 与 M3 同口径)

并发N	总TPM	其中输出	RPM	p50 完成延迟	错误率
32	14.7 万	3.7 万	72	25.1 s	0
64	22.5 万	5.7 万	111	32.4 s	0
128(交付档)	32.6 万	8.2 万	160	47.2 s	0
128(256K 长文池口径)	25.4 万	6.4 万	125	61.6 s	0
128(无命中对照, 256K)	15.4 万	3.7 万	73	107.3 s	0

主池构成(均为同画像严格 A/B 实测):较 256K 单池口径 +28%, 其中上下文档位 128K 释放显存 +20%、MTP 投机解码再 +7%(N=32/64 档 MTP 增益 +25%/+19%)。缓存净贡献(256K 口径 A/B):输出 +72%、总TPM +65%、延迟 -43%;无命中对照 619 tok/s 与历史无缓存锚 614 吻合, 交叉验证通过。双池说明:主池承接 ≤128K 上下文流量(绝大多数场景);>128K 长文由 256K 专用池承接(扩容节点部署, 单机口径 25.4 万)。

五、服务可用性与恢复能力

模型	冷启动/故障恢复(至可服务)	较优化前提升	说明
MiniMax-M3	≤ 4 分钟	2.3 倍(8:27→3:38)	分钟级恢复, 支持滚动升级
GLM-5.2(128K+MTP 主池)	≤ 7 分钟(实测 6:21)	1.9 倍(12:10→6:21)	同上
Kimi K2.7	≤ 16 分钟	1.45 倍(22:34→15:33)	同上
Kimi K3	≤ 29 分钟	—(瓶颈在计算侧)	专线模型, 建议常驻

以上为实测端到端就绪时间(含全部初始化)。生产建议 N+1 冗余部署, 故障切换期间由冗余节点承接, 对外无感。7×24 监控与告警已具备。

六、与官方默认方案的性能对比(提升比例汇总, 均同机实测)

模型	对比基线(出处)	基线值	交付值	提升
MiniMax-M3	官方参考部署 BF16-SGLang(同机实测报告)	聚合峰值 122.4 tok/s	1671 tok/s	13.7 倍(方案级)
MiniMax-M3	同画像无缓存 A/B 对照(本轮实测)	总TPM 17.8 万	24.5 万	+37%(缓存净贡献)
MiniMax-M3	无投机解码 A/B 对照(本轮实测)	单流 65.4 tok/s	113.8 tok/s	+74%
GLM-5.2	标准部署口径初测(可研报告)	577 tok/s	870~1178 tok/s	+51%~+104%
GLM-5.2	同画像无缓存 A/B 对照(本轮实测)	总TPM 15.4 万	25.4 万	+65%(缓存净贡献)
GLM-5.2	256K 单池口径 A/B(本轮实测)	总TPM 25.4 万	32.6 万	+28%(128K档位 +MTP投机)
Kimi K2.7	可研基准情景(参考报告 V4)	222 tok/s	497 tok/s	+124%
Kimi K3	适配起点(K3 验收报告 §3.4)	单流 4.06 tok/s	37.25 tok/s	9.2 倍
全模型	故障恢复时间(优化前实测)	8:27~22:34	3:38~15:33	1.45~2.3 倍

方案级倍数含画像差异(见 §3.3 口径注);A/B 对照行为严格同画像同并发对比。所有基线均有对应实测报告可溯源, 支持现场复测。

七、经济性效益对比与提升比例

7.1 单 token 供给成本(按单机 5 年 TCO ¥235.9 万、日成本 ¥1,311 折算)

模型(画像)	优化前成本	优化后成本	降幅
MiniMax-M3(命中型, 同并发 N=128)	¥5.12 /M token	¥3.72 /M token	-27%(批量档 ¥3.11/M, 最高 -39%)
GLM-5.2(命中型, 同并发 N=128, 实测)	¥5.90 /M token	¥2.79 /M token(主池)	-53%
GLM-5.2(48k 验收画像)	¥0.49 /M token	¥0.33 /M token	-34%

注:前两行为同一命中型画像(3:1), 可横比——GLM 与 M3 优化后供给成本相当;第三行为 48k 高命中计费口径(计费token基数大故单价低), 不可与前两行横比;降幅均为各自同画像前后对比。

7.2 单节点毛利(可研测算引擎口径:48k 画像、97.7% 命中、8 折外采)

模型	优化前毛利	优化后毛利	提升	备注
GLM-5.2	+72.6%(锚577)	+81.8%(锚870)	+9.2 pp	盈亏平衡 158 tok/s, 实测达 5.5 倍
Kimi K2.7	+5%(基准222)	+58%(实测497)	+53 pp	从踩线到厚利
MiniMax-M2	—	+14.7%(实测1264)	达盈亏线(1078)1.17 倍	薄利, 主因牌价低
MiniMax-M3	—	牌价待核后另出	—	产能口径见 §三

7.3 规模承接的硬件节省(以 300 万总TPM 为例, 5 年 TCO 口径)

承接方案	优化前需要	优化后需要	5 年 TCO 节省
MiniMax-M3(命中型)	17 台(¥4,010 万)	11 台(¥2,595 万)	¥1,415 万(-35%)
GLM-5.2(128K+MTP 主池实测)	22 台(¥5,190 万)	10 台(¥2,359 万)	¥2,831 万(-55%)

7.4 合作方采购视角:相对直接采购官方 API, 我方牌价折扣直接节省 5%~15%(具体见商务报价函);相对自建同等产能, 免去每台 ¥182 万 Capex 与部署运维, 且自建按官方默认/未优化方案的产能仅为我方交付态的 1/13.7(M3 方案级口径)~1/1.6(GLM 同画像实测口径)。

八、规模承接换算(需要多少台)

目标规模(总TPM)	MiniMax-M3(命中型)	GLM-5.2(128K+MTP 主池)	备注
50 万 TPM	2 台	2 台	含向上取整
100 万 TPM	4 台	4 台	
300 万 TPM	11 台	10 台	建议 +1 台冗余;GLM >128K 长文 另配 256K 池

如流量为纯输出型(不计输入), 按输出TPM 换算:M3 每台 10.0~10.3 万输出TPM, 300 万输出TPM 约需 30 台(+冗余)。具体以双方确认的流量画像为准, 我方可按贵方画像先行免费联测标定。

九、计费与结算口径

- 1. 计量:按 OpenAI 兼容接口返回的 usage 字段逐请求计量(输入 token / 缓存命中输入 token / 输出 token 三段分计), 双方可各自留存日志月度对账;
- 2. 计价:输入、命中输入、输出分别计价, 命中输入享缓存折扣价;具体单价与折扣以商务报价函为准(另附);
- 3. 结算:人民币结算, 可开增值税专用发票;支持月结;
- 4. 联测:正式接入前提供免费联测额度, 用于贵方验证吞吐、延迟与计量一致性。

十、合规与服务承诺

- 1. 模型合规:四个模型均为官方公开发布权重, 遵循其开源许可部署于我方自有算力;
- 2. 数据安全:请求与返回内容不用于任何训练, 不向第三方披露;可按合同约定日志留存期限, 支持签署保密协议;
- 3. 服务等级:承诺月度可用性目标与故障响应时效, 具体 SLA 条款随商务合同约定;
- 4. 可复测:本报告全部指标(含 §六 对比基线)可在合作方见证下现场复测。

十一、验收方法(可复现口径)

压测经 OpenAI 兼容 /v1/chat/completions 接口进行:每并发档持续 90~120 秒滚动发压, 以 usage 字段统计窗口内全部输入/输出 token, 吞吐=Σtoken/墙钟, TPM=吞吐×60, RPM=完成请求数/分钟;延迟为请求完成墙钟的分位数;错误率=非 200 响应占比。解码型画像:短输入+512 输出;命中型画像:输入:输出≈3:1、高前缀命中。启动/恢复时间为服务进程拉起至接口可用的端到端墙钟。

§六 官方基线出处:MiniMax-M3 BF16-SGLang 同机实测报告、GLM-5.2/K2.7/K3 可研与验收报告(编号另附), 均可提供原始日志。经济测算:单机 5 年 TCO \$327,651(¥235.9 万, 汇率 7.2), 测算引擎经 21 项中间数反算校验。