

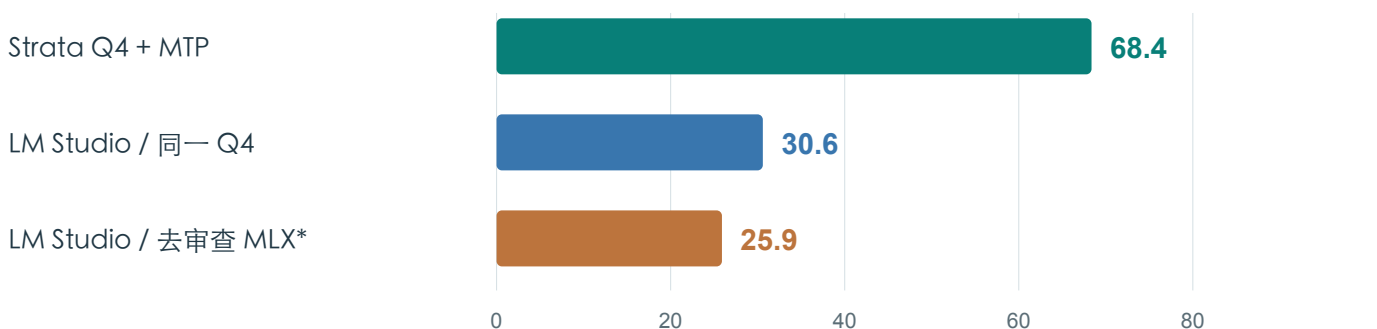
01 / 实测结论

Strata 在 512 GiB Mac 上有什么实用价值？

Qwen3.8-Flash-Next Q4 · 独立 llama.cpp · LM Studio · 思考模式

同一份普通版 Q4，开启思考后：
Strata 68.4 token/s，LM Studio 30.6 token/s，速度约为 2.23 倍。

开启思考的总生成吞吐 | 三次样本中位数



* MLX 为已有 Uncensored 4bit 模型，与普通 Q4 的权重和量化方式不同，单独作为使用参考。图中吞吐包含思考与正文，不是完整回答的完成速度。

实用价值落在什么地方

速度收益已确认。 这台机器上，Strata 在所测 Qwen3.8 模型与配置中生成更快。开启思考的代码题，Strata 完整回答用时 8.26 秒，LM Studio 同权重 Q4 为 16.38 秒；两边均通过六个检查用例。

容量突破尚未证明。 测试 Q4 本来就能被这台 512 GiB Mac 的其他后端装下。本轮没有跑通一个“原先因容量不足无法运行、只有 Strata 能跑”的模型。

速度不等于质量排名。 关闭思考时，Strata Q4 曾违反代码题要求的返回类型；开启思考后该题通过。少量样例不能代表整体编码或推理能力。

测试主体完成于 2026-10-24，日本时间。IQ3_S 为 2026-10-23 的历史参照。所有速度来自本机记录；本 PDF 是既有测试整理，没有为排版重新测速。

02 / 配置与口径

先确认比较的是谁

项目	实际配置
硬件	Apple M3 Ultra，32 核 CPU（24 性能核 + 8 能效核），512 GiB 统一内存
系统与工作负载	macOS 26.6.2；既有服务及虚拟机保持原状，测试后端先后运行
普通版 Q4	unsloth/Qwen3.8-Flash-Next-GGUF UD-Q4_K_XL，原始四个 GGUF 分片共 111.33 GB
Strata	v0.2.3；Mac Control 0.1.48；Metal 引擎 0.1.48 标准配置 MTP 开启，spec=2；另测关闭 MTP
独立 llama.cpp	0.6.0-dev，commit 23b0202；全 GPU，Flash Attention，未配草稿模型；物理批大小 2,048
LM Studio	应用 0.4.23+1；GGUF 运行时 2.31.2；MLX 运行时 1.11.0 GGUF 全 GPU，Flash Attention，MTP 关闭；物理批大小 512
MLX 补充模型	ARC4NUM/Qwen3.8-Flash-Next-Uncensored-MLX-Serve-4bit 已有不同权重，约 107.33 GB

版本核查结果

测试时核查到最新正式发布为 v0.2.3.1（日本时间 10 月 23 日 22:45）。该补丁调整 Windows/Linux 的 Control，Mac Control 仍为 0.1.48，引擎也未更换；本机测试沿用上述安装。此处记录测试时版本，不作为未来的最新版声明。[1]

统一设置与计时

上下文 32,768，单请求并发，温度 0，文本任务。生成吞吐取三个独立提示样本的中位数；短任务与完整代码/计数题为单次或少量功能样例。

客户端生成吞吐 = (完成 token 数 - 1) / (完整响应时刻 - 首个生成片段时刻)。含思考时一并计入。首 token、正文首字和完整回答耗时分别记录；正文首字不计仅换行的片段。

这是具体部署配置的比较：MTP、构建版本、批大小和模板存在差异，没有穷尽调优，也没有测出各后端的理论上限。Strata 的 Q4 重排检查了 1,223 个张量的字节一致性；官方准备步骤另生成所需 BF16 副本。[2][3]

03 / 关闭思考

相同 Q4 的后端对比

第一轮生成测试：每个请求上限 256 token，三个样本均达到该上限。表中使用统一的客户端口径，因此 Strata 显示 69.1；原日志的服务端计时为 69.0。

部署配置	token/s	256 token 总耗时	首 token
Strata Q4 / MTP 开	69.1	4.17 秒	0.478 秒
Strata Q4 / MTP 关	58.1	4.89 秒	0.488 秒
独立 llama.cpp Q4	39.6	6.70 秒	0.254 秒
LM Studio Q4	32.6	8.16 秒	0.334 秒

Strata 开 MTP 是 LM Studio Q4 的 2.12 倍；关闭 MTP 后仍为 1.78 倍。独立 llama.cpp 的 39.6 token/s 是另一个可执行文件的成绩，不能标成 LM Studio。

实际任务的完整等待时间

任务	Strata Q4	独立 llama.cpp	LM Studio Q4
日文付款通知翻译	0.80 秒	1.14 秒	1.68 秒
订单 JSON 提取	0.96 秒	1.70 秒	2.07 秒
输出工具调用	0.91 秒	1.63 秒	2.02 秒
4,520 token 文档查找	3.72 秒	4.96 秒	6.69 秒
27,968 token 文档查找	22.16 秒	29.89 秒	49.53 秒
代码题：完整要求通过	1 / 6	6 / 6	6 / 6

代码失败的具体含义：Strata 合并数值及保留输入均正确，但返回 `[[1, 7]]`，题目要求 `[(1, 7)]`。五个非空用例因此不通过；关闭 MTP 仍生成相同代码，尚未定位具体实现原因。

三组翻译、提取、工具参数和文档代号检查通过。工具调用只检查模型生成的函数名及参数，没有真正执行外部搜索。长文主表使用请求关闭缓存的记录；Strata/独立 llama.cpp 回传缓存数 0，LM Studio 未回传缓存计数。

首次请求成本另计：Strata Q4 首次翻译为 10.71 秒，热身后为 0.80 秒。主表并不代表从点击加载模型开始的总等待时间。

04 / 开启思考

吞吐接近，答案需要多等一会儿

第二轮重新做同题开/关对照，生成预算为 512 token。普通 Q4 开/关设置都实测；MLX 仅确认默认思考实际输出，当前模型接口未暴露关闭开关。

部署配置	关闭思考	开启思考	开启的三个样本
Strata Q4 + MTP	64.2	68.4	67.8 / 72.8 / 68.4
LM Studio / 同一 Q4	30.6	30.6	31.5 / 30.6 / 29.5
LM Studio / 已有 MLX	未测	25.9	25.9 / 26.2 / 25.8

单位：token/s。与上一页的提示措辞、输出预算不同，判断思考影响应使用本页同轮对照。Strata 的小幅吞吐上升可能受生成内容和 MTP 接受率影响，不能推断思考本身会加速。

完整回答测试：另给 4,096 token 预算

配置	代码正文首字	代码完整回答	思考 token	计数题完整回答
Strata Q4	7.03 秒	8.26 秒	约 494	5.16 秒
LM Studio Q4	13.03 秒	16.38 秒	375	9.15 秒
LM Studio MLX	26.82 秒	29.51 秒	671	11.30 秒

这些完整回答均正常结束。三组代码都通过六个用例。关闭思考时，Strata 同题仅需 2.82 秒但通过 1/6；LM Studio Q4 为 5.66 秒且通过 6/6。

计数题正确答案为 401。开启思考后三组均答对；关闭时 Strata 答 187、LM Studio Q4 答 267。这里只能说明这道题的结果，不能据此给模型整体能力排名。

为什么 512 token 的测速不能当完整回答时间

三个开启思考的生成样本都用完输出预算。Strata 和 MLX 各有一个样本尚未产生正文；这属于预算截断。完整任务耗时见上表，不能用“生成了 512 token”代替“答完了问题”。

思考 token 分项：Strata 可见思考文本重新分词约 104 / 346 / 512；LM Studio Q4 接口报告 223 / 181 / 384；MLX 报告 239 / 238 / 511。MLX 请求上限 512，实际回报总输出为 511。

Strata 未返回思考 token 分项，表中的“约 494”使用原模型 tokenizer 对可见思考文本估计，排除隐藏边界与结束符；LM Studio 分项直接来自服务端 usage。

这台机器的收益主要在速度

没有验证到容量突破：所测 Q4 在其他后端也能运行。
这台 512 GiB 统一内存 Mac 不能直接套用小显存 PC 的收益。

普通版 Q4 在本机以驻留方式运行，并未依靠从 SSD 流式读取专家来突破内存容量。v0.2.3 Mac 文档列出 Q4/Q5/Q6，说明不止 IQ3；但 Q5、Q6 和 Strata 去审查 Q4 均未在本次实测。[2]

其他大模型的历史预检	结果与解释
MiniMax-M2.7 Q8_0 约 243.14 GB	旧版 CLI 报 unsupported_architecture，未进入推理。没有证据证明它此前因本机容量不足而失败。
所检查的 GLM-5.2 副本	九个分片只找到四个，报不完整分片；不能据此判断完整模型的架构兼容性或加载能力。

这些预检来自 10 月 23 日旧版环境；不能当作最新引擎全部能力的测试。限制专家缓存以模拟小内存的实验未执行，也没有下载一个超出本机容量的新模型。

磁盘与资源记录

项目	本次记录
Q4 主模型原始分片	111.33 GB
Metal 重排文件 / MTP 原始张量	另占约 82.54 GB / 5.21 GB
本次任务目录实占	约 202.02 GB，含原始分片、准备文件、运行环境及记录

202.02 GB 包含为对照保留的源权重，不是最小安装空间要求。首轮 Q4 的进程 RSS 采样最大值：Strata 90.11 GiB、独立 llama.cpp 82.40 GiB、LM Studio 109.19 GiB。RSS 含映射且统计方式不同，不能直接排名实际物理/GPU 内存占用。各采样轮均未观察到新增系统 Swapouts；未做长期压力测试。

IQ3_S 仅作前一天的历史参照

Strata 配置	生成吞吐	28k 查找耗时	代码检查
IQ3_S / 引擎 0.1.47	约 79.8 token/s	27.59 秒	6 / 6
Q4 / 引擎 0.1.48	约 69.1 token/s	22.16 秒	1 / 6 (思考关)

量化来源及软件版本同时改变，不能把差异全部归因于 Q3/Q4。IQ3_S 模型已清理，历史数据保留；Q4 安装和原始权重保留。测试服务已停止，LM Studio 临时实例已卸载，临时链接已删除。

06 / 复现与边界

怎样复查这些结果

核心提示与请求设置

生成提示：“请写一篇不少于600字的中文说明，介绍本地大模型用于日中翻译、文档处理和代码辅助的适用场景。这是独立样本 N。” N 分别取 1、2、3。第二轮不加 /no_think，预算 512；第一轮末尾另加“直接开始正文。/no_think”，预算 256。

代码提示：编写 merge_intervals(intervals)，输入为无序二元组列表；输出按起点排序的二元组列表，合并重叠或端点相接的闭区间；空输入返回空列表；不修改输入；只输出函数，不解释、不导入。检查空输入、乱序重叠、相接、嵌套、负数和重复区间。

计数提示：从 1 到 1000，能被 3 或 5 整除，但不能被 15 整除的整数有多少个？最终只答整数。答案为 $333 + 200 - 2 \times 66 = 401$ 。

API：OpenAI 兼容流式 Chat Completions；temperature=0，stream=true，include_usage=true，cache_prompt=false。思考开启请求为 enable_thinking=true、reasoning_effort=high；关闭为 false/none。Strata high 映射为模板 xhigh。MLX 仅验证实际思考输出，未核实强度设置生效。

证据与可比性

本地保留了每次请求、完整 SSE 响应、usage、计时、进程采样、加载配置及六个代码用例检查。首次命中提示缓存的独立 llama.cpp 结果另存，主表改用关闭缓存的复测；LM Studio 没有缓存命中回传，不能声称已从服务端核实零缓存。

未测复杂代码库、长篇专业翻译、视觉、多用户并发吞吐及超过 32k 的上下文。旧版 Mac 双请求实测出现排队；当前文档也注明单槽位。本次只验证 API 响应与函数调用格式，未验证完整编码代理的多轮工具执行。[2]

公开来源与固定版本

[1] Strata v0.2.3.1 发布说明

[2] Strata v0.2.3 Mac 支持、准备与限制

[3] Strata v0.2.3 发布资产

[4] Unsloth Q4 模型仓库（固定提交）

[5] LM Studio Chat Completions 接口文档

Strata 源码固定提交：5c8dd322528cc84f68f461ad54a506b9da8fa8d4

模型固定提交：38bb39ee97821de2c9009abb7e93950eec396e66

论坛版仅包含整理后的数据、合成题目与公开来源。原始日志未附入 PDF；已去除本机用户名、绝对路径、局域网地址和其他私人环境标识。数值按展示位数四舍五入，结论限于这里的版本与配置。