

Qwen3.8能力与历史限制评估

结论与当前状态

新版有可用优势，部分历史接入设置需要调整。

本轮不支持全局放开所有限制。

3.6 正常环境已恢复，3.8 独立环境保留。

原 dev2 后端健康，ASR 与向量进程保持。

线上产品规则没有修改。

核心能力对照

20 类用例，各重复两次，使用相同输入与评分。

澄清一条年龄/日期歧义后重新比较。

推荐非思考：3.6 29/40，3.8 35/40。

开启思考：3.6 37/40，3.8 39/40。

原始未澄清成绩仍完整保留。

前轮能说明旧参数下的兼容性，不能代表模型上限。

空输出改写的重新判断

同一条角膜塑形镜改写，用20个新种子成对复验。

3.6 原参数 20/20，仅改历史思考开关 20/20。

3.8 原参数 8/20，仅改历史思考开关 20/20。

请求只改变 `preserve_thinking`，其他设置一致。

模板渲染也确认了历史标记差别。

该坏例不能直接归因于新版不懂业务。

结果只覆盖这条坏例，不表示所有空输出都已解决。

哪些优势确实测到

80 token 的短改写，关闭历史思考保留：

眼别更正：3.6 0/3，3.8 3/3。

同为4096 token思考预算的140条记录筛选：

3.6 0/2，3.8 2/2。

旧版此题的失败属于预算耗尽。

旧版开启思考也能解决部分眼别题。

因此不把这些结果说成旧版永远做不到。

真实业务提示词与接收边界

真实4L提示词，推荐非思考：3.6 6/8，3.8 6/8。

真实4L提示词，开启思考：3.6 6/8，3.8 5/8。

导演组使用现有提示词，并测试取消JSON强制模式。

这不是原完整接口的通过率。

导演字段及流程，推荐非思考：3.6 3/4，3.8 3/4。

导演字段及流程，开启思考：3.6 1/4，3.8 0/4。

4L成绩经过既有后处理，只表示项目规则一致。

它不是临床正确率。

固定句逐字差别、服务补年龄前缀已单列。

放宽格式后仍有坏JSON，不能据版本升级取消保护。

现有服务没有分离思考文本；直接打开会影响长度门及JSON解析。

模型最终答案正确，不等于应用能直接接收。

现有默认超时60秒、4L输出上限2000。

本轮长推理放宽预算，仅用于能力诊断。

历史限制的处理建议

全局关闭思考：改为按业务选择，先补正确的思考/最终答案分离。

短改写80 token：本轮不是空输出根因，不建议一概放开。

历史思考保留：普通查询改写可显式关闭；多步智能体另测。

生成预算：复杂任务与短分类分开设置，并保留超时与回退。

固定句：拆开统一产品口径与模型兜底，不能整批删除。

JSON字段校验、缺失值保护、确认流程、权限与幂等：保留。

B1忽略与防重复1.2来自5月12日LoRA改动。

5月15日切到3.6后仍保留，需单独链路复测。

旧图像输入限制也已定位，不能直接解除。

本轮未证明扩大整个服务窗口具有业务收益。

新版思考三档与参数取舍

相同三条问题，每档两次，相同8192 token预算。

low：6/6，空输出 0，预算耗尽 0，中位耗时 5.517 秒。

medium：6/6，空输出 0，预算耗尽 0，中位耗时 4.648 秒。

xhigh：6/6，空输出 0，预算耗尽 0，中位耗时 10.335 秒。

档位选择应看完整任务结果，不能只按名称判断。

另测了单位识别及时间线的惩罚参数。

推荐参数是起点，仍需按业务验证。

范围、社区与证据

本轮新增382条完整记录的真实推理请求。

覆盖30类核心或真实提示词场景。

另含修正与单因素复验。

这是定向业务样本，不是通用榜单。

还包括20个新种子的模板成对复验。

使用GPU—实际AWQ部署，vLLM 0.20.2。

两代MTP设置不同，不据耗时宣称纯版本速度提升。

一次实验调度超时后续跑，未计为模型失败。

没有生产患者写入、真实微信发送或手机业务验收。

原始请求、响应、评分及恢复证据完整保留。

上一轮社区生成500等既有链路问题仍未修复。

本轮不宣告全业务替换验收通过。

社区线索来自官方模型卡、Kodesage和上游问题216/236。

社区不同精度、不同引擎的分数没有直接套用。