

Qwen3.8升级前16维核验与整理

结论与基线

升级前16维梳理完成，正式切换条件尚未满足。当前保持3.6；先完成接入适配与隔离验收。原dev2任务、ASR、向量服务保持。仅发出一次短健康推理，没有业务写库或模型切换。

本轮真实核验

17组主复核rg保存540行结果，另有19组当前调用盘点。33个当前app文件语法通过；11请求模板、7文件、20显式客户端调用点。7/9关键文件与旧测试一致；H5和导演打标不同。GPU与本地17份原件哈希一致。382条历史请求、48条扩展评分、52条消费者回放可复算。

16维范围

基础：版本基线、GPU依赖资源、端点身份、请求名与版本记录。协议：思考正文、HTTP/SSE错误、JSON及流程、语义评分。业务：视觉语音检索、业务规则、权限租户、预算耗时。验收：落库历史手机、测试隔离、回滚共存、证据与发布。已核基线、需整改和待真验分别标记；未宣称全通过。

优先整理项

1. 核对B3两条旧模型名、三处固定版本标记及各worker有效配置。
2. 分离思考和最终正文，处理流式4xx、坏帧、空final和length。
3. 按业务配置参数与预算，保留JSON/schema、权限及确认。
4. 冻结评分真值，增强眼别/否定/数值绑定。
5. 独立PG、Redis和测试机构，再做真HTTP/DB/GPU验收。
6. 补当前微信导航、手机展示、并发/长稳和回退恢复。

收益数字与失败保留

规则通过率：非思考29/40→35/40，思考37/40→39/40。不是总体业务准确率。56条AI语义审阅：非思考旧7通过/5失败/2不确定，新13通过/1失败；思考两代13/14。新版仍有错误眼别绑定。382条HTTP失败0，但空最终答案21、length6，重叠3。4L/导演并未稳定提升，真实全业务尚待验收。

证据补齐与剩余缺口

原证据清单为空，已补建499项私有哈希索引并保留原件。旧评分器与冻结hash不符，历史原因未证明。52条阶段副本已排除重复计数。原baseline、恢复备份、before与实验日志已取回。下一轮须预先保存请求UTC/ID、输入与评分器hash、原始流帧及有效参数。

资源、隔离与回滚

GPU—SSD可用183G；两环境依赖版本相同。原3.6配置/启动脚本hash一致，只读回滚check通过；ASR与向量进程启动ID一致。两代MTP不同，不宣称纯版本速度提升。测试夹具含建表、Redis清空和org_id=1清理，因此不能直接对共享库跑pytest/verify all。

交付与下一步

完整MD及私有证据索引位于dev2报告目录。先在最新test的独立worktree完成最小适配和隔离真验；保留现有导演任务。手机业务、生产所有消费者、并发长稳仍未验证。正式切换须依据这一轮完整业务结果。