

三模型建站对比实验报告

ORBITAL EYE · 本地大模型 SPA 生成能力对比（归档版）

日期：2026-10-07 | 场景：张硕凯 / Hermes 监督侧 | 环境：本机三张 GPU 全本地推理、Kilo Code CLI 驱动

0. 一句话结论

同一份任务书下，**qwen3.8-27B（4090）一次成型且质量最高**（32 分钟交付，并自主完成 3 轮自检修复）；**Ornith-35B 两臂均未一次成完**：C 臂（3090）经一轮修复达到完整可用，B 臂（双 A4000）经历为最曲折的续跑与修复，最终版可用但仍为三站中最弱。三站终版**全部通过统一验收并已上线**。

另一个关键结论写在 § 5：**输出上限与上下文的配比（context - output = 可用输入窗）是本轮最致命的工程变量**，它比模型规模更早地决定了一次任务能否跑完。

1. 实验设计

1.1 任务书（三个臂完全相同）

虚构项目「**ORBITAL EYE · 近地轨道态势感知平台**」单页应用（SPA）：

- 4 个 hash 路由视图：总览 / 轨道态势 / 碰撞预警 / 系统日志；
- Canvas 轨道交互动画（地球、轨道环、目标运动、悬停与选中详情）；
- 列表筛选/排序、详情弹窗、日志流式追加与暂停；
- 响应式（桌面 + 移动端）、**零外部 CDN 依赖**；
- 品质条款（原文照录）**：作为品质标准，目标是达到 awwwards、webby awards、FWA 能获奖的品质，通过自我检查，反复提升品质，直到达到那个品质标准为止。并要求至少一轮"自检 → 修复"闭环。

1.2 三臂配置

臂	模型	硬件	上下文	驱动	配置路径
A	qwen3.8-27B	RTX 4090 单卡	180K	Kilo Code CLI (--auto)	armA-qwen4090
B	Ornith-1.5-35B-A3B Q4_K_M	双 RTX A4000（层切分）	131K	同上	armB-ornith-a4000
C	Ornith-1.5-35B-A3B Q3_K_M	RTX 3090 单卡	131K	同上	armC-ornith-3090

三臂：同提示词、同工作目录隔离（三重锁定 `--dir`）、并行发车，输出上限初始统一 100000。

1.3 公平性修正记录（重要）

初始统一"输出上限 = 100000"本意是保公平，但它与 **Ornith 两臂的 131K 上下文组合出致命后果**：可用输入窗口只剩 ~31K，写一个大文件即触发压缩且压缩必失败（"Compaction exhausted"），两臂连续中止。修正：两臂输出上限降至 32768（窗口回到 ~98K）后，续跑方可完成。此差异在 § 5 展开。

1.4 验收方法（监督侧）

真浏览器（Chrome）+ 桌面 1440×900 / 移动 375×812 双视口；量化项：控制台错误数、四视图切换、数据数值是否非零、Canvas 有效像素（ink）、公网 URL 直连。**全程禁用浏览器缓存**（本机 http.server 无 Cache-Control，不禁缓存会出假象）。

2. 时间线（事实记录）

时刻	事件
20:35	三臂并行发车
20:43	B、C 先后中止（Compaction exhausted）；A 继续
21:07	A 完成 （wall 1910s / 32 分钟，rc=0）—— 自带浏览器自测 9 张截图 + 3 处自修复结论
21:02	续跑轮 1（B/C）发车
21:12	轮 1 双中止（rc=1，各写 1-2 个文件后压缩再爆）
21:14	定位根因（输出/上下文配比）→ 修复配置 → 轮 2 发车
22:13	轮 2 完成 （B/C 文件补齐到位；C 并自检出数据接线问题并请示）
21:31-22:24	C 修复轮 ：数据面板 / 轨道视图 / hash 路由全部修复（wall 3186s）
22:14-22:32	B 修复轮 ×2：均被上游 API 流中断卡死（SIGTERM 收场）
22:55-23:20	监督侧人工兜底完成 B 的 4 类根因修复 （详见 § 4.3）
23:0x	B 同步上线；三站统一验收全绿；公网 4 入口全部 200

3. 三臂结果对比

3.1 总表

维度	A（qwen3.8-27B）	B（Ornith Q4 双A4000）	C（Ornith Q3 3090）
单会话完成	✅ 32 分钟	❌（3 次中止 + 2 次卡死）	❌（2 次中止）
续跑轮次	0	2 + 2 修复轮	2 + 1 修复轮
自检→自修复	✅ 3 处（自主定位）	⚠️ 长时间调试未及根因	✅ 自检并主动报告问题
人工兜底	无	有（4 类根因）	无（仅提示词引导）
终版控制台错误	0	0	0

维度	A (qwen3.8-27B)	B (Ornith Q4 双A4000)	C (Ornith Q3 3090)
终版完成度	作品级	可用（三站最弱）	完整可用
遗留小瑕疵	无记录	个别数据语义待打磨	日志徽标折行、{tcc} 占位符

3.2 代码规模（终版）

臂	文件	行数	说明
A	8	2953	HTML 327 + CSS 1071 + JS 1555，单 CSS 46.8KB
B	9	2406	模块化 JS（router/data/app/logView…）
C	8+2	2176	+2 个调试脚本（px.js/sel.js，未上线）

3.3 质量观察（截图见附录）

- **A**：设计密度与信息量最高——KPI 卡 + 雷达扫描 + 碰撞列表 + 子系统状态四区完整，数据命名拟真（天舟-9×SL-16 DEB-4411 等），视觉与排版无重叠、无溢出。
- **C**：体系感强（六卡仪表 + 趋势面积图 + 事件流），轨道视图星野/光晕/目标标注完整；数据面板在修复轮后全部接通。
- **B**：修复后结构齐整（KPI + 趋势 + 分段 + 轨道 Canvas 满帧绘制），但数据层单薄、个别语义待打磨。

4. 缺陷与修复实录

4.1 A 的自修复（自主完成，无需干预）

1. 模块注册错位：路由 `enter()` 从未执行 → 全部视图空白（自加别名解析器修复）；
2. 移动端日志页高度失控（4432px）→ 改为 `min(62vh,640px)` + 内部滚动；
3. CSS/JS 类名核对（含 `[hidden]` 被 `display:flex` 覆盖的真 bug）补齐。

4.2 C 的修复轮（提示词引导，模型完成）

自检发现并修复：`dashboard.js` 查询了不存在的元素 id（卡片/图表不渲染）→ 对齐；`orbital.js` 引用了 `data.js` 未定义的数据接口 → 补全；`hashchange` 不触发视图切换 → 路由补全。残留：日志级别徽标 CRITICAL 折行、日志文案 `{tcc}` 模板占位符未插值（不影响使用）。

4.3 B 的修复实录（含人工兜底说明）

B 的最终版包含 **4 类深层缺陷**，其中模型自身长时间调试（1 小时余）未定位到根因，由监督侧复核定位并（在其修复轮连续被 API 流中断卡死后）人工完成：

1. **选择器 vs 裸 id**：全站 `$('#id')` 调用点传入选择器写法，而 `$` 实现是 `getElementById`（需裸 id）→ `getElementById('#kpiGrid')` 恒为 null → 全部 `null.appendChild / null.innerHTML` 崩溃。修法：容错版 `$`（strip # 后 `getElementById`，失败再 `querySelector`），4 个文件各一行。

2. **CSS 特异性覆盖**: `#view-collision / #view-log / #view-orbital { display:flex }` 三条 ID 规则 (特异性高于类) 盖掉 `.view { display:none }` → 四个视图永远叠加显示。修法: 三处加 `.active` 限定。
3. **漏声明 self**: `_drawBackground` 使用 `self.cx/cy/R` 但未 `var self = this` ——strict 模式下 `self` 解析为浏览器全局 `window.self` → `undefined` → `createRadialGradient` 非有限值, **轨道画布每帧中断**。修法: 补一行声明。
4. **字段不匹配**: 分段列表读 `s.name`, 数据源字段为 `s.label` → 五条全部显示 "undefined"。修法: 改字段引用。

另: 清理调试期遗留的 3 处 console 探针 (RK/BOOT/LIVE), 恢复干净文件。共 12 处人工补丁。

4.4 运行层记录 (工程资产)

- **kilo 会话静默卡死模式**: 进程 `ep_poll` 静等、日志停更、服务端请求被取消——上游流中断而 kilo 无超时。处置: 外部杀会话重跑; 重跑仍卡则人工兜底, 不做无限重试。
- **8008 双 A4000 服务曾意外死亡一次**, 导致 B 首轮修复空转 12 分钟; 已恢复并保留重启脚本。

5. 关键发现与可复用结论

1. **上下文-输出配比定律**: 可用输入窗口 $\approx$ context - output。100K 输出 + 131K 上下文 = 31K 窗口 = 死于第一个大文件。输出上限不得贪过上下文的 1/4; 131K 上下文请用 32768 输出。
2. **模型的"上下文管理"是完成度分水岭**: qwen 臂 180K 上下文 + 更快生成, 撑过整程; Ornith 双臂 131K 在写大文件 + 重读的负载下必然触墙——**多轮续跑是此类模型跑大型任务的正常作业方式** (提示词需"填空式": 已有清单 + 缺件 + 单文件写入 + 禁重读)。
3. **小模型缺陷原型清单 (验收必查)**: ① `$('#id')` vs `getElementById`; ② 高特异性选择器盖显隐逻辑; ③ 函数漏 `var self = this` (strict 下即 `window.self` 陷阱); ④ 字段名不匹配; ⑤ 路由/视图双层状态不一致。
4. **验收纪律**: 禁缓存 (`CDP Network.setCacheDisabled`); 模型自述不算数, 必须真浏览器量化复测 (错误数/active 视图/数据值/canvas ink)。
5. **监督分工有效**: 根因定位与"标准答案式"提示词/兜底 = 监督侧价值所在; B 的案例同时说明——**当 API 层不稳时, 及时止损转人工比继续赌重试更优**。

6. 交付物与证据

- **线上入口 (公网直链)**: <https://cad.qr4t.cn/elevator/sites/>
- A: <https://cad.qr4t.cn/elevator/sites/a/>
- B: <https://cad.qr4t.cn/elevator/sites/b/>
- C: <https://cad.qr4t.cn/elevator/sites/c/>
- 源码/运行记录: `/data/code/site-bakeoff/` (三臂 run 目录、编排脚本、续跑与修复日志 `continue.log`、`results.json`)
- 部署位置: `/var/www/site-bakeoff/{a,b,c}/`

- 验收截图存档：~/hermes/cache/scratch/bakeoff-verify/（含三站总览、各视图、移动端）

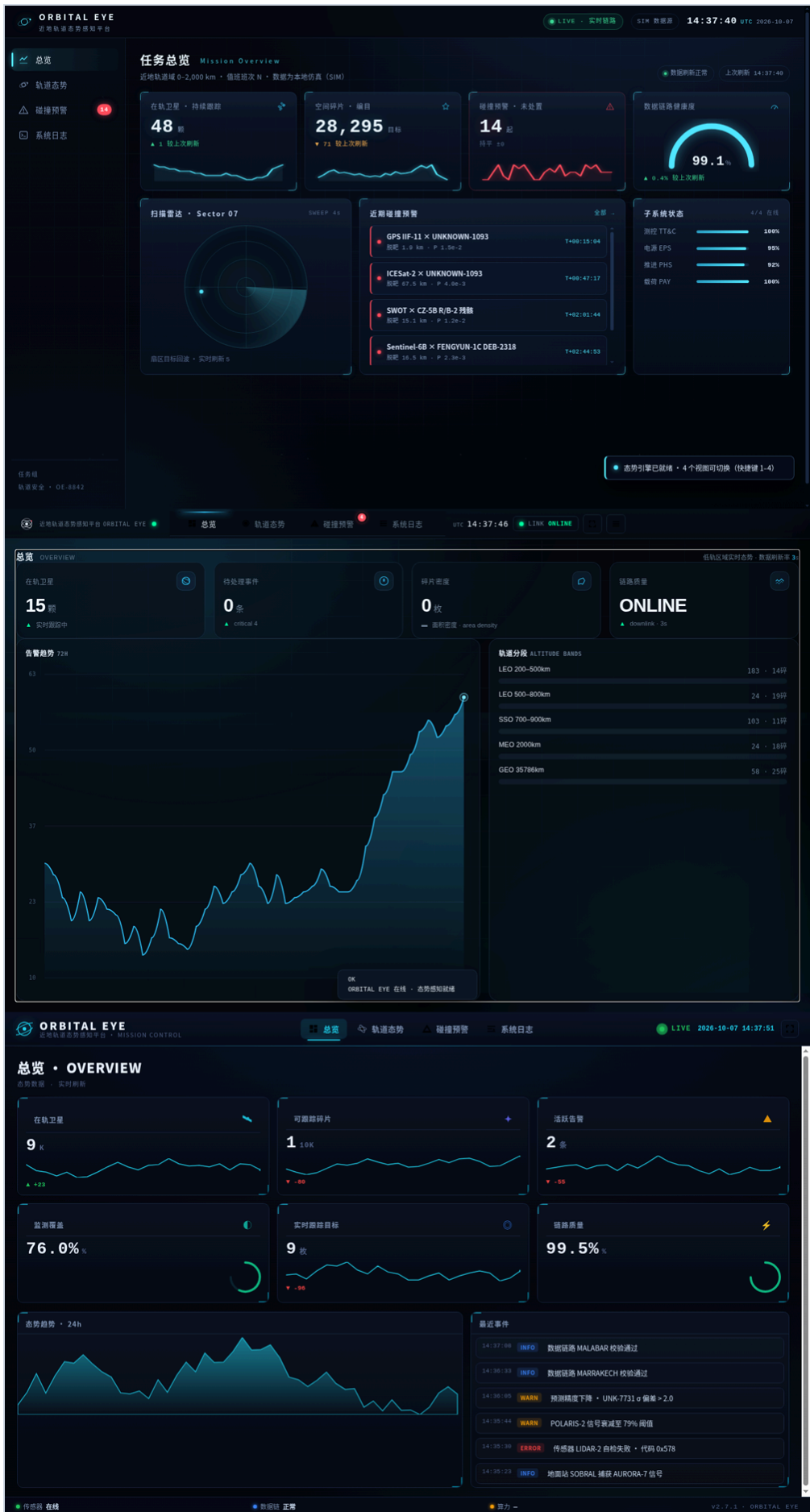
（附录图片：三站总览页实拍拼图 + 三站单页截图）

7. 诚实标注

- **人工介入范围**：B 终版的 4 类根因修复与调试代码清理（12 处补丁）为监督侧人工完成；A 无介入；C 仅提示词引导（模型自主修复）。"单模型独立完成度"口径：A $\approx$ 100%，C $\approx$ 85%，B $\approx$ 70%。
- 三站均存在不影响使用的零星小瑕疵（上表已列）；如需继续打磨（如 B 数据语义、C 文案细节），可再各开一轮"限定范围 + 填空式"修复。
- 本报告全部数字来自运行日志与实测输出，可回溯：`/data/code/site-bakeoff/` 保留了每一次会话的完整日志。

报告生成：2026-10-07 · Hermes 监督侧 · 全流程本地推理（4090 / 双 A4000 / 3090）

附录 · 实拍截图



三站总览页拼图 (上 → 下 = A / B / C)



A 站 · qwen3.8-27B (RTX 4090) 总览



B 站 · Ornith-1.5-35B Q4 (双 A4000) 总览



C 站 · Ornith-1.5-35B Q3_K_M (RTX 3090) 总览