个人项目 / 人机协作开发 / 2026 年 10 月

让长任务可恢复、可核验

AI 写作工作流的设计与故障复盘

我用长篇中文修订检验多智能体工作流。任务需要保留原稿事实、对白和人物声音，同时改善空洞解释与重复描写。连续运行后，问题集中在两个地方：审核意见没有落实到正文；会话在上下文压缩阶段卡住，无法顺利续作。

在这个个人项目中，我同时承担需求方与验收方，负责提出要求、检查稿件，并据此调整分工和交付顺序。智能体辅助完成代码、日志分析、候选修订与测试，关键修改保留来源和对照。

从具体漏审处改流程

一份已交稿的开头有45段，逐段对照后发现全部沿用原文。旧报告一方面指出表演场景存在重复套路，另一方面又把这段标为“可原样继承”，没有列出必修项。气质突然变化、笼统称赞表演出色等文字因此留了下来。[1]

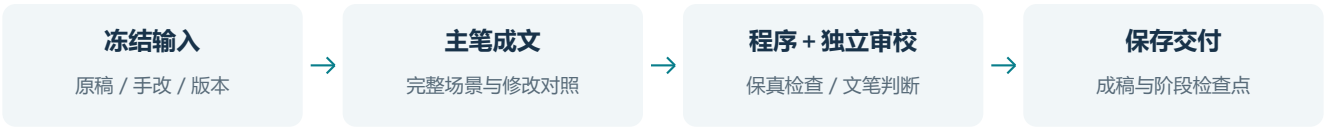
我把验收要求落到具体段落：每项问题记录原句、问题所在、可改范围和关闭条件。作者交稿时须回应必修项；程序检查回应是否齐全、指定文字是否发生变化，编辑再判断问题是否真正解决。只有修改声明、没有正文变化的交件会被拦下。

三个岗位围绕同一份候选工作

主笔负责成文，文笔编辑判断表达与节奏，设定审校核对事实、对白和前后依赖。总控合并一次返修清单，保留已经通过且逐字未变的范围；有问题才返修，达到约定上限后停止当前路径。

把确定性检查交给程序

输入包固定原稿、最新手改、段落编号和文件哈希。程序检查缺段、乱序、保护对白、引用位置及报告版本，独立审校负责语义和文笔。新岗位按职责读取当前场景和必要邻文，避免每个岗位都接收完整对话。



必修项回到主笔；复核绑定同一版本，正文与检查点分别保存。

实现环境：Python、本地文件与JSON记录、SHA-256版本校验；当前工具需要显式调用，尚未接入宿主形成自动拦截。

故障诊断与续作

压缩卡住后，先确认任务停在哪里

历史日志显示，任务达到自动压缩阈值后提交了约0.99 MB的请求。收到HTTP 200后，响应体在解码阶段断流，两次重试仍失败；用户继续操作，又进入同一压缩步骤。日志可以定位失败阶段，尚不能确定问题来自客户端、代理链路还是服务端。[2]

391次	31次	11次
工具回传	超过12000字符	输出被截断

以上是跨五个窗口的历史累计量，不是单个上下文窗口的容量。

减少进入会话的材料，保留完整原文

正文、来源表和长日志放在文件里，检索先定位，再按场景和行号读取。单次工具回传默认控制在8000字符内；长场景连续分次读取，出现截断就补读缺失范围。这个限制约束回传大小，不能用来缩短小说正文。

context_guard.py根据传入的统计检查回传规模、截断和压缩失败标记。它只处理显式传入的数据，宿主事件仍需另外采集；缺少统计字段时，不能把检查通过当作预算已经核清。

让进度可以从磁盘重新核对

每个任务只保留一份权威检查点，记录输入版本、完成范围、未决问题、运行中的操作和下一步。恢复规程要求更新记录时先写同目录临时文件，检查旧文件未被其他写入者修改后再替换，避免中断留下半份正式记录。

续接时核对真实文件、哈希、运行状态和已生成产物。没有收到回复的操作标为“结果待核实”，不能直接当失败重做。新会话只接收交接入口和当前依赖，不重新装入全部聊天历史。

给诊断设置退出条件

同一故障最多连续尝试三轮，每轮只改变一个关键变量；连续两轮没有新证据就提前退出，保存已排除方法和下一项验证。底层问题暂时无法定位时，继续不受影响的本地工作，避免在同一步骤无限重试。

当前验证范围止于本地预检与续作协议。远程断流尚未被独立复现，底层故障原因仍待排查。

验证结果与边界

用失败用例检查验收边界

测试同时包含应通过和应拦截的情况。退出码正常只能说明程序运行结束；交稿是否完整、文笔是否可用，仍需分别读取检查结果和正文。

检查对象	已留存的结果	能证明的范围
上下文预检	本次复跑7项测试通过；现行工具拒绝历史触发包、接受有界样例	本地已知故障的检查规则有效
低模型机械执行	GPT-5.6 Luna执行20例，12项机械失败、2项输入错误、6项机械通过，均符合预设	可以按冻结规则运行工具并报告结果
必修项回应覆盖	新增8项测试通过，包含漏回应、未改正文、只改空白等情况	防止漏交；不判断改后的文笔

机械通过仍会漏掉语义错误

20例测试中，故意改动未锁定金额、旁白说话主体和心理否定的三个样本仍能机械通过。它们划清了分工：低模型适合执行确定的核验，独立审校仍须逐段核对事实、人物动机和表达。工具始终保留“不可据此直接发布”的状态。[3]

场景化用有局部收益，整体提升仍待验证

在一段街头说书场景中，原稿只评价“贯口功底展现得淋漓尽致”。我要求按表演场景寻找参考，智能体从《老残游记》的听书描写中取用“说得极快却字字清楚”的关系，写出具体听感。完整27段样稿经过独立文笔与内容复核，确认了一处措辞化用的收益。[4]

早期18篇小样对照中，有无参考两组的文笔通过均为5/9，没有证明参考输入能稳定提升质量。整本去AI效果和成本下降也未获验证。当前交付保留完整稿、逐处对照、运行结果和失败样本，出现新的验收问题时，对应范围的旧放行结论需要重新核查。

下一轮只看实际交付

先完成一章修订及对照，按实际读感确认文笔标准，再扩展到全文。过程指标记录返修次数和读取范围，结果指标记录完成的正文与仍未关闭的问题。模型费用尚缺完整账单，不能用缓存比例或减少调用次数代替降本结果。

[1] 开头45段原文对照、旧文笔前检与root_opening_audit_gap.json。

[2] 运行诊断.md、trace_metrics.json；2026年10月5日7项本地测试复跑。

[3] luna_execution.json、root_luna_comparison.json；coverage_tests_run_final.json。

[4] 《老残游记》第二回；R042候选与同版两份独立审校记录。

写作实验记录截至2026年10月4日；上下文预检于10月5日复测。均为个人项目材料，不代表商业上线或最终验收。