房地产行业深度报告

地产+AI 工具系列报告之八：AI 投研如何降本提效—— Token、耗时与 Skill 优化实践

增持（维持）

投资要点

- **AI 投研的效率评价应从“模型账”上移至“任务账”。**单次模型价格、Token 数或响应速度仅反映局部成本，完整任务还包括检索、数据提取、口径核验、建模、写作、审校及返工等环节。我们认为，应以“通过质量门槛的完整研究任务”为效率单位，并以质量调整后任务成本（QATC）综合衡量模型、工具、人工及返工等投入。
- **Token、耗时与质量应作为“三本账”同步计量。** Token 下降并不必然对应墙钟时间下降，搜索、文件解析、工具调用、人工确认和返工均可能成为关键路径。效率优化应在质量不下降的前提下，同时降低资源消耗和交付时间，并区分平台原生 Token、分词器估计及字符代理等不同计量口径。
- **重复读取、上下文膨胀、全文重写和口径返工，是当前 AI 投研流程中较典型的隐性成本来源。**同一份年报被多个阶段或多个 Agent 重复全文读取，历史对话和工具返回在后续调用中持续携带，局部修改触发整篇报告重新生成，以及前序数据口径错误导致后续分析链整体返工，均可能显著放大任务成本。
- **Skill 的核心并非增加 Prompt 长度，而是将研究经验转化为可执行流程。** Prompt 描述单次任务，SOP 固化研究原则，脚本承担确定性计算，Skill 则进一步连接任务路由、工具调用、证据规则、失败处置和终止条件。面向正式投研场景，我们认为“证据型 Skill”尤其值得关注，其核心是将原始材料转化为可追踪的证据对象，并通过口径冲突闸门、程序断言和分层验证降低错误与返工。
- **48 轮实验：结构化方法的价值与任务难度高度相关，流程复杂度本身并非优势。**简单数据核验任务中各配置质量差异有限；长报告审阅中，仅证据型配置实现 100% 首次通过；估值任务则显示额外规则和质量闸门也可能增加资源负担。真正有效的复杂度，应能够改变结构化状态、阻断错误或减少返工，而非单纯增加规则和 Agent 数量。
- **研究团队的落地重点，应由“个人会不会用 AI”转向“能否形成可评测、可回滚的 Skill 资产”。**建议优先将高频、流程稳定、错误代价较高且结果可验证的任务 Skill 化，并建立统一日志、固定任务集、版本管理和回归测试。对于正式发布场景，研究员仍应保留问题定义、来源与口径裁决、核心假设审查及最终发布责任。
- **风险提示：**实验样本数量及任务代表性不足风险；模型及平台版本变化风险；大模型幻觉及数据口径判断偏差风险等。

2026 年 08 月 16 日

证券分析师 姜好幸
执业证书：S0600525110001
jianghx@dwzq.com.cn
证券分析师 刘汪
执业证书：S0600526030001
liuwang@dwzq.com.cn

行业走势

相关研究

《地产+AI 工具系列报告之七：从 AI 投委会到可用的投研平台——AlphaChain 产品化》

2026-07-31

《地产+AI 工具系列报告之六：面向 Agent 时代的工程师训练路径与国产大模型实践》

2026-07-17

内容目录

图表目录

引言：从“平台可用”到“研究生产率可度量”

系列之七重点讨论 AlphaChain 由研究方法论向可用投研平台的产品化过程，核心问题是如何将完整报告、结构化决策、资源调度、历史跟踪和机构化控制组合为稳定的使用体系。本篇进一步讨论平台进入日常生产后更加基础的一组问题：一项 AI 研究任务到底消耗了多少资源，效率改善发生在哪里，质量是否同步保持，以及有效路径能否沉淀为团队可复用的 Skill 资产。

模型调用只是投研生产链条的一部分。真实任务通常跨越问题定义、资料发现、证据提取、分析建模、写作和交付验证。一次调用看起来便宜，并不意味着完整任务便宜；模型输出很快，也不意味着研究员能更早拿到可以正式使用的结果。若遗漏关键来源、混淆统计口径或在最后阶段触发全文返工，前序节省可能被后续成本重新吞噬。

因此，本篇不把“Token 越少”直接等同于“效率越高”，而是提出三个递进层次。第一，建立任务级成本、耗时和质量的统一计量；第二，定位重复读取、历史携带、全文重写和无效验证等流程泄漏；第三，将有效机制固化为具有输入契约、证据规则、程序断言、失败边界和终止条件的 Skill，并通过固定任务集进行对照评测。

本篇重点回答四个问题：一是 AI 投研究竟省了什么，将 Token、工具、时间、人工、返工及失败统一纳入完整任务成本；二是成本浪费发生在哪里，重点拆解检索、长文处理、分析、写作、验证及收尾等环节；三是哪些方法值得固化，比较 Prompt、SOP、程序与 Skill，重点讨论证据管理、冲突处理及分层验证；四是如何验证实际效果，通过冻结任务、独立 Session 及 A/B/C/D 对照实验，比较不同方案的成本、效率与质量。

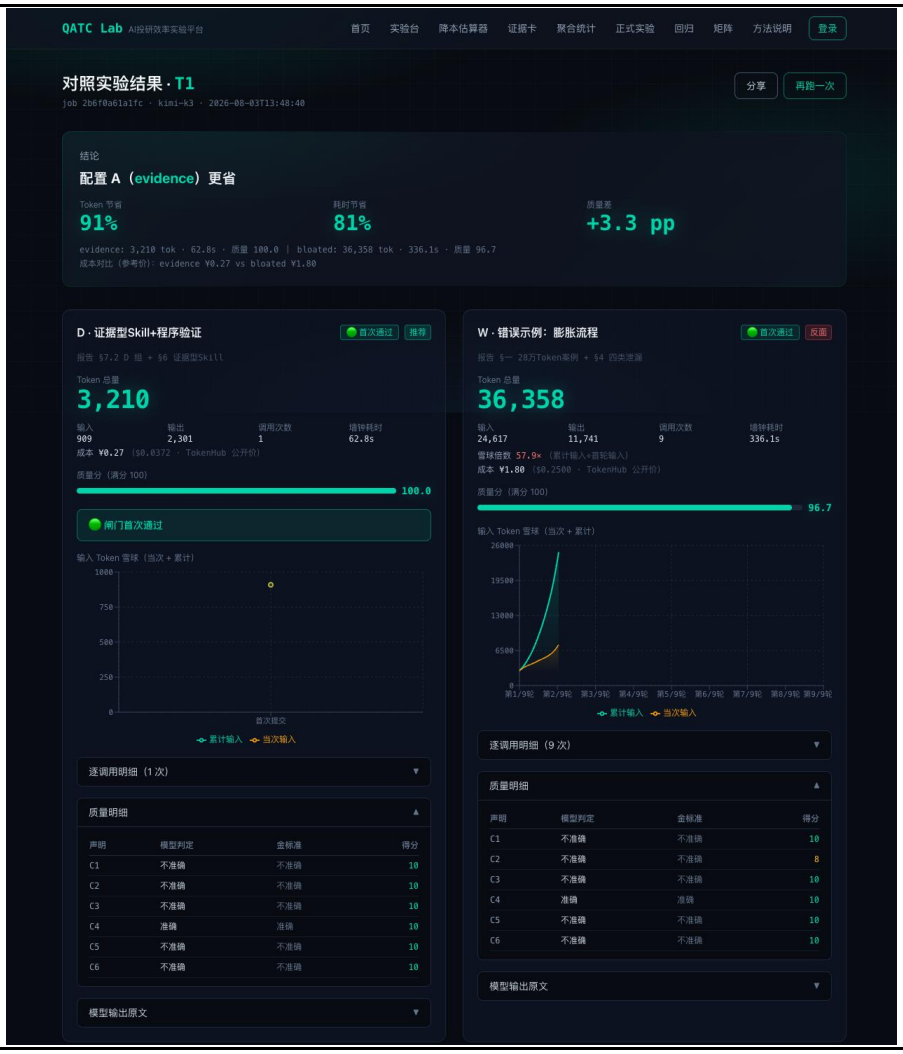
1. 效率度量：从模型单价到质量调整后任务成本

1.1. 从模型成本到任务总成本

投研效率的核算对象应当是一项完整研究任务，而不是一次模型调用。一项典型任务通常包括问题拆解、资料发现、原文读取、数据提取、口径核验、计算建模、写作、审校和最终交付。模型调用只发生在其中部分环节，而前序信息遗漏或口径错误造成的补证据、重算和改稿，仍属于同一项任务的真实成本。

例如，一份公司业绩快评可能只产生数元模型费用，但若分析师需要花费数小时检索和核验，主管还需进行审阅，且一处收入口径问题导致分析链整体返工，则真正的经营成本主要来自人力和返工，而非模型账单。相反，一项模型单次价格较高但首次即可交付的任务，其总体成本可能低于多轮纠错的低价模型方案。

图1：QATC Lab 对照样例——低效多轮流程 VS 证据型流程



数据来源：QATC Lab 平台，东吴证券研究所

因此，任务成本至少需要同时包含模型、工具、计算、人工、等待、返工及失败尝试。对于暂时难以货币化的环节，可以先分别使用 Token、墙钟时间、人工时间等原生指标。

表1：投研任务总成本的主要构成

成本类别	典型内容	容易遗漏的部分
模型成本	输入、输出、推理及缓存相关计费	多轮历史重复携带、失败调用
工具成本	搜索、数据接口、文档解析、代码执行	重复搜索、重复解析和长返回体
人工成本	问题定义、确认、审阅、裁决和发布	口径冲突处理、错误兜底
等待成本	模型、网络、工具、队列及编排等待	多个活动并发时的关键路径
返工成本	补证据、重算、改稿、重新生成文件	首次交付后发生的自动返工
失败成本	超时、资源触顶、Skill 报错等	只统计成功样本造成幸存者偏差

数据来源：东吴证券研究所整理

在这一框架下，我们使用质量调整后任务成本（Quality-Adjusted Task Cost，QATC）观察不同配置：

$$QATC = \frac{全部任务成本之和}{通过验收的交付件数量}$$

该指标的核心并不在公式复杂度，而在分母定义。若一项方案完成十次尝试但只有六次达到质量门槛，不能只计算六次成功运行的成本；四次失败同样属于该方法为形成六份合格交付所付出的代价。若一组任务最终没有任何交付通过，QATC 应记为不可定义，并直接披露零通过，而不是呈现低成本。

1.2. 三本账：Token、耗时与质量

仅使用一个综合指标仍不足以定位问题。我们进一步将 AI 投研效率拆为 Token、耗时和质量三本账。三者回答的问题不同，且必须来自同一轮独立运行。

表2：Token、耗时与质量三本账

层级	主要指标	核心问题	典型误区
Token 账	输入/输出 Token、缓存读写、模型调用、工具返回体积	用了多少上下文和生成资源	把字符、文件字节直接换算为真实 Token
耗时账	墙钟时间、首 Token 等待、模型/工具活动、人工暂停、返工	用户为什么需要等待	将并行调用耗时简单相加冒充墙钟时间
质量账	事实正确性、关键事实召回、来源绑定、公式正确、首次通过等	消耗是否形成可交付结果	只对最终返工版本评分，忽略首次表现

数据来源：东吴证券研究所整理

Token 账首先是一套口径纪律。平台原生 Token 应作为第一层计量，只有 API usage 或账单明确返回的输入、输出、缓存和推理 Token 才进入这一层。若平台原生字段缺失，但完整可见请求能够导出，可以使用冻结版本的 tokenizer 形成第二层估计。若连完整请求都无法获得，则保留字符数、字节数、调用次数和文件读取量等第三级代理，不应将代理静默转换为平台 Token。

中文投研任务下这种机制尤为重要。自然语言、代码、JSON、表格和中英文混合文本的分词密度差异较大，不存在稳定的“字符—Token”通用换算关系。模型服务商对缓存读取、缓存写入和推理 Token 的定义也可能不同。因此，不同计量层级宜并列披露，而不是通过估算补齐缺失值后再形成一个看似精确的总数。

耗时账应围绕墙钟时间和关键路径建立。模型请求、文件解析和搜索可能并发发生，十个并行搜索的耗时之和可以远大于用户实际等待。建议将运行轴拆为模型独占、工具独占、模型—工具重叠、人工暂停和编排等待等互斥区间，并保留返工时间。首 Token 等待只有在客户端能够真实记录首字节或首个流式 Token 时才计算，不能用完整请求耗时替代。

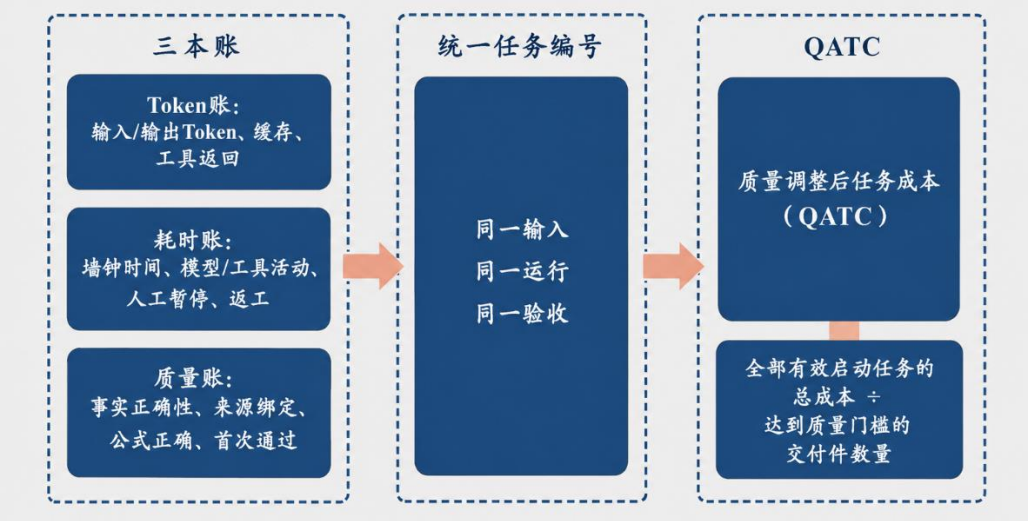
质量账的目标不是评价文风，而是判定结果是否满足正式投研的证据和交付要求。我们认为，至少应覆盖事实正确性、关键事实召回、来源绑定、一手来源率、口径冲突处理、反方与风险覆盖、公式正确性、交付完整性和首次通过九项指标。首次提交和返工后的最终提交需要分别保存，避免将“第一次就做对”和“通过昂贵闭环最终做对”混在一起。

表3：投研质量评分卡

指标	主要含义	典型硬边界
事实正确性	可审计事实是否被有效证据支持，且对象、时点、单位和口径正确	重大核心事实错误可直接否决
关键事实召回	冻结金标准中的重要事实是否被覆盖	不能通过少写内容换取高准确率
来源绑定	关键主张能否定位到真正支持该主张的原文位置	机构首页、搜索摘要不视为精确绑定
一手来源率	已引用关键主张中由有效一手来源支持的比例	转载不能制造虚假交叉验证
口径冲突处理	时点、范围、单位、状态等冲突是否被识别并合理处理	静默选数不视为解决
反方与风险覆盖	核心结论是否面对有证据的替代解释和失效条件	仅列风险标题不构成完整覆盖
公式正确性	关键公式的结构、依赖、单位、符号与结果	硬编码碰巧正确不得视为满分
交付完整性	章节、表格、文件、字段和格式是否满足预设要求	文件可打开不等于交付完成
首次通过	第一份完整产物是否同时越过所有硬门槛	返工后的结果不能覆盖首次版本

数据来源：东吴证券研究所整理

图2：AI 投研效率的“三本账”与 QATC 框架



数据来源：东吴证券研究所绘制

1.3. QATC、首次通过率及其他派生指标

除 QATC 外，研究团队还可以使用首次通过率、重复读取率和每条已核验事实 Token 等指标定位流程问题。其中，首次通过率对研究生产较为重要：一项方法即使最终能够通过，但每次都需要自动返工或人工介入，其稳定性和真实成本仍可能较差。

重复读取率可按照文件哈希和读取范围识别同一内容是否被多次读取。需要注意，第二次读取并不天然等于浪费：首次抽取、独立复核和修复后的回归检查都可能合理地再次读取同一材料。因此，重复读取需要结合阅读目的、声明编号和触发原因判断。

有效 Token 率可以用于研究内部诊断，即观察可测 Token 中有多少直接服务于最终使用的证据、分析、质量控制和交付内容。由于“有效”本身包含一定人工判断，该指标更适合同一任务前后比较，不宜直接用于跨任务排名。

1.4. 缓存与并行不能替代任务级计量

缓存可以降低稳定前缀的部分输入成本，也可能改善响应延迟，但它解决的是重复输入的计价和服务机制，并不意味着重复内容已经从研究流程中消失。若一个长工具返回被后续十轮持续携带，即使其中部分命中缓存，任务仍存在上下文管理问题。因而，缓存读取、缓存写入和未缓存输入应分别记录，不能把缓存收益预先写成固定折扣。

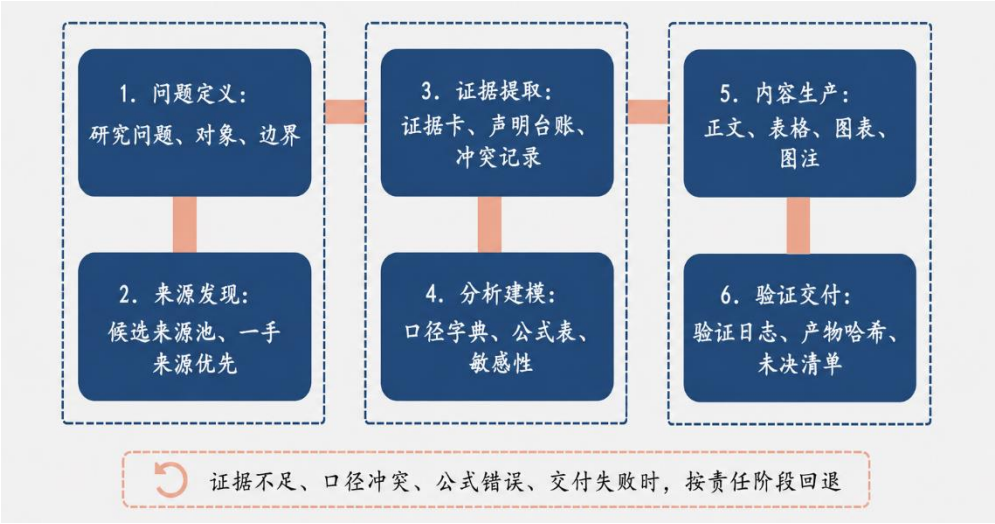
并行 Agent 同样主要改变墙钟关键路径，而不天然降低总工作量。多个 Agent 同时打开相同材料、复制相同任务书并分别形成完整分析，可能缩短用户等待，却增加模型活动和合并返工。因此，效率评估最终仍需回到同一质量门槛下的任务总成本。

2. 工作流拆解：六阶段任务链与四类主要泄漏

2.1. 六阶段任务链：从问题定义到验证交付

一份正式研究报告并不是模型接到题目后连续生成若干章节。实际工作更接近一个状态机：研究对象在问题定义、来源发现、证据提取、分析建模、内容生产和验证交付六个阶段之间迁移，并可能因证据不足、口径冲突、模型错误或交付问题退回前序环节。

图3：一项投研任务的六阶段工作流



数据来源：东吴证券研究所绘制

表4：六阶段任务链的产物及回退条件

阶段	核心问题	主要产物	典型回退条件
问题定义	真正需要证明什么	任务书、问题树、证据需求矩阵	核心数据不可获得、研究边界不清
来源发现	哪些材料值得进入验证	候选来源池、来源目录、补证队列	仅有转载、版本不匹配、时点不一致
证据提取	原文真正支持什么	证据卡、声明台账、冲突记录	对象、时点、单位、事实状态缺失
分析建模	数据是否可比、计算是否成立	口径字典、公式表、敏感性结果	关键输入缺失或假设敏感度过高
内容生产	如何将已验证结论转化为交付文本	正文、表格、图注及多载体内容	出现新事实、跨载体数字不一致
验证交付	事实、公式、结构和呈现是否通过	验证日志、产物哈希、未决清单	引用无法定位、公式错误、必交项缺失

数据来源：东吴证券研究所整理

问题定义阶段的核心，是将“想讲什么”改写为“需要证明什么”。例如，“AI 基础设施改变地点价值”只是一项宽泛叙事，需要进一步拆分为负载需求、电力供给、项目建设、利用率、现金流和资产价值等可验证环节，并预先区分外部事实、政策目标、分析假设和模型结果。

来源发现阶段应区分“发现来源”和“验证来源”。搜索摘要、二手文章和导航页面可以提高召回率，但不能因为表述完整就自动进入正式证据层。证据提取阶段进一步将原文压缩为最小可审计单元，保留对象、时点、单位和统计范围；分析建模再解决可比性与计算问题。内容生产的职责是组织已经通过边界检查的结论，而不是重新创造未登记事实。

2.2. 重复读取与上下文膨胀

AI 投研中较常见的成本并非来自一次特别长的输出，而是相同材料在多轮上下文中不断被重新携带。假设一段长材料在第 3 轮进入会话，并在之后的每一次调用中保留至第 10 轮，则这段内容会在累计输入中被重复计入多次。若其上又叠加摘要、工具返回、临时代码和错误日志，实际负担会进一步扩大。

长文阅读中的另一种典型浪费是多个角色分别全文读取同一材料。一个 Agent 为概览读取全文，另一个 Agent 为抽取数据再次读取，写作阶段为了引用又重新读取，验证阶段再做一次全文核对。若没有内容哈希、页段索引和明确阅读目的，系统很难区分必要复核与无新信息的重复搬运。

更合理的处理方式是建立“目录—页段—证据卡”三级索引。第一次结构化阅读负责识别章节、表格和可能相关区域，后续环节只根据声明编号或研究问题读取最小必要范围。共享对象应当是带原文定位和内容哈希的证据卡，而不是每个 Agent 各自生成一

份全文摘要。

2.3. 计算、写作与返工成本

分析阶段的主要泄漏，常来自让语言模型承担了额外的计算，而这本可由程序一次性完成。筛选、排序、汇总、单位换算、公式填充、哈希、去重和一致性扫描均适合程序执行；模型更适合判断两个口径是否可比、异常是否具有经济含义、哪些反方可能改变结论。

若模型不断生成临时代码、执行后读取长错误栈，再重新生成整段代码，错误信息会与数据样本一起进入后续上下文。将稳定的计算脚本沉淀为资产，并固定输入输出schema，可以同时降低重复 Token 和随机性。

写作阶段最典型的浪费是“局部修改触发全文重写”。一份数万字的报告中，修改一个数字、增加一句限定语或替换一张图片，若每次都重新加载全部证据和全文，再生成新版本，输入和输出都会显著放大，并可能破坏此前已经通过核验的其他章节。

因此，正式报告更适合维护声明编号—章节编号—来源编号的映射。修改一项主张时，只读取相关证据和目标段落；修改排版时，应由构建工具处理，不重新进入研究推理。只有核心论点变化、统一口径调整或章节依赖发生变化时，才需要扩大重写范围。

2.4. 验证泄漏与成功后的无效调用

验证阶段容易从“检查既定结果”滑向“重新做一遍研究”。若验证器缺少冻结规则，只是再次让模型“看看有没有问题”，第二轮分析往往与第一轮使用相同材料和类似推理路径，独立性有限，却产生大量新增成本。

验证应当有明确对象。公式验证只检查输入、公式、单位和结果；来源验证检查引用是否真正支持主张；视觉验证检查截图或图表是否正确；反方验证检查结论是否存在能够改变判断的替代解释。完整成功日志应外置保存，执行流程只接收失败规则、位置、期望值和实际值。

收尾阶段的泄漏则来自没有把“已经通过验收”转化为机器可判断的终止状态。任务已经完成目标文件、质量检查和产物保存后，Agent 仍可能重新读取文件、再次解释过程或继续生成未要求的扩展内容。解决方式是冻结必交产物、保存哈希和验证状态，在硬门槛全部通过后写入停止原因，并将任何可选增强拆为新任务。

图4：四类主要 Token 与时间泄漏机制

数据来源：东吴证券研究所绘制

表5：必要研究动作与流程浪费的区分

环节	必要研究动作	可识别的流程浪费	主要优化方式	不可削减边界
来源发现	扩大召回、回到一手原文	重复转载反复打开、无边界补证	来源归并、预算和停止条件	不以搜索摘要替代原文
长文阅读	首次全局阅读、关键页精读	多 Agent 无目的全文重读	目录索引、页段读取、证据卡	保留必要全局理解和独立复核
分析建模	口径判断、情景分析、关键公式复算	模型逐行计算、临时代码重复生成	确定性程序和可复用脚本	程序不替代经济含义判断
内容生产	结构性改稿、受影响章节联动	局部要求触发全文重写	局部修改、章节哈希、主张映射	核心论点变化时允许全局修订
验证交付	关键事实核验和修复后回归	无规则“再看一遍”	结构化失败摘要和定向回归	不取消关键事实与公式验证
收尾终止	原子保存、清单和最终说明	成功后继续调用或重复构建	显式终止、幂等和可选任务分离	硬门槛未通过不得提前结束

数据来源：东吴证券研究所整理

3. 从 Prompt 到 Skill：将研究经验转化为可执行流程

3.1. Prompt、SOP、脚本与 Skill 的边界

投研团队积累 AI 使用经验后，最容易形成的是越来越长的提示词。长 Prompt 可以加入来源优先级、格式要求、检查清单和风险提示，但若所有规则均以自然语言存在，系统仍缺少状态管理、工具路由和可阻断错误的验证机制。

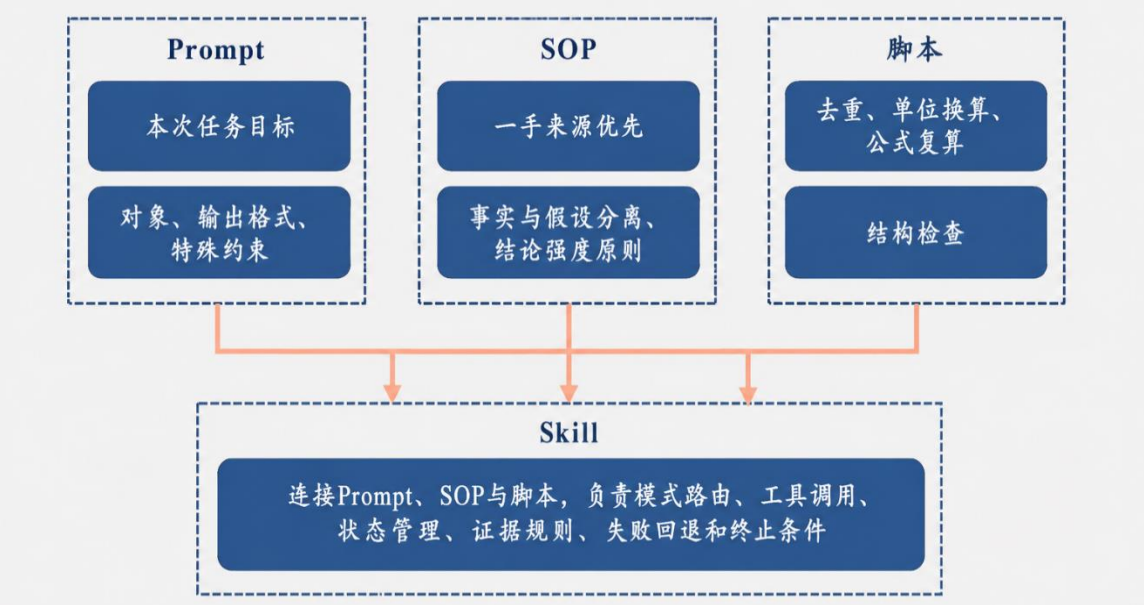
Prompt、SOP、脚本和 Skill 分别解决不同问题。Prompt 面向本次任务；SOP 固化研究原则；脚本执行确定性动作；Skill 则将上述三者连接为一个可运行、可回退、可终止的任务路径。

表6：Prompt、SOP、脚本与 Skill 的职责边界

类型	主要问题	适合承载的内容	主要边界
Prompt	这一次要做什么	对象、期间、读者、输出要求、本轮特殊约束	不宜重复承载大量稳定规则
SOP	研究应遵守什么原则	一手来源优先、事实与假设分离、结论强度原则	原则本身未必可被机器执行
脚本	哪些动作应得到确定结果	哈希、去重、筛选、单位换算、公式和结构检查	不能独立裁决复杂语义和经济含义
Skill	任务如何运行并结束	模式路由、工具调用、状态、证据、失败和终止	需要版本、测试和适用边界

数据来源：东吴证券研究所整理

图5：Prompt、SOP、脚本与 Skill 的关系



数据来源：东吴证券研究所绘制

以核验公司收入增速为例，Prompt 给出公司、期间和输出格式；SOP 要求优先使用公司正式披露并确认会计期间；脚本负责提取收入并复算同比；Skill 则先判断财年还是自然年，再选择公告或结构化数据，将原值、期间、币种和来源位置写入证据卡，在遇到重述或币种变化时进入冲突处理，最终检查结论和证据是否一致。

3.2. Direct、Analysis 与 Report 三类模式

Skill 不宜按照“简单版—复杂版”固定分级，而应根据任务的不确定性、证据风险和交付范围进行路由。我们将常见投研任务分为 Direct、Analysis 和 Report 三类。

表7：Direct、Analysis 与 Report 三类 Skill

模式	典型触发条件	适用任务	主要工具	完成条件
Direct	对象明确、问题窄、输出短、证据风险较低	指定文档问答、单项数据核验、定义解释	局部文件读取、单一权威源、轻量计算	直接回答且关键事实可定位
Analysis	多来源、多口径、需要计算或反方检验	行业核验、公司快评、估值情景、专项审阅	多源检索、结构化数据、脚本和证据验证	核心问题、证据、冲突和风险均完成检查
Report	多问题、跨章节、多载体、正式交付	行业深度、完整公司报告、主题策略	Analysis 全部工具，加章节编排、构建和跨载体验证	全部必交物、硬门槛和跨载体一致性通过

数据来源：东吴证券研究所整理

3.3. 投研 Skill 的最小结构

一个能够进入团队治理的 Skill，至少需要同时回答九个问题：什么时候触发、启动前需要什么、任务如何路由、工具由谁调用、要交付什么、证据需要满足什么规则、错误如何处置、结果如何验证以及任务何时停止。缺少其中任何一项，Skill 都可能在单次演示中工作，却难以稳定复用。

表8：投研 Skill 的九项最小结构

结构块	主要内容	关键要求
触发条件	适用意图、任务和风险范围	同时说明不适用场景
输入契约	对象、问题、读者、截止时点、材料、交付载体	区分可默认、必须澄清和允许未知
模式路由	Direct / Analysis / Report	依据确定性、材料、错误代价和交付复杂度
工具路由	模型、搜索、数据、程序、构建和验证工具	明确语义判断与确定性动作边界
输出契约	交付物、字段、格式和保存位置	必交项应可机器检查
证据规则	来源等级、定位、事实状态、冲突处理	约束结论强度
失败边界	重试、回退、人工升级和停止条件	避免同错无限重试
验证规则	事实、来源、公式、结构和呈现检查	失败应输出规则编号和定位
终止条件	成功、资源触顶、不可恢复失败或人工接管	记录停止原因和未决问题

数据来源：东吴证券研究所整理

3.4. 什么应当固化，什么不应固化

Skill 应固化研究方法和控制边界，而不应固化具体公司结论、未经实验验证的阈值或会限制新发现的固定观点。例如，“关键数字必须保留时点和单位”“关键对象不一致时不得直接比较”可以成为稳定规则；“某公司具有长期成本优势”“某行业需求一定增长”则属于特定时点的研究结论，不应写入 Skill 的默认前提。

同样，不宜把一次性任务中的公司名称、季度、临时路径、大段原始材料和审稿意见永久嵌入 Skill。此类内容应通过输入契约或外部存储传入。固定来源清单也应谨慎：可以要求优先使用一手来源，但不宜因为历史项目使用过若干网站，就禁止发现新的有效来源。

规则数量也不是成熟度。诸如“超过五份文件必须使用多 Agent”“引用不足十条不得交付”之类阈值，如果缺少任务分层和实验依据，容易形成形式主义。真正需要固化的是错误机制及其对应的控制方式，而不是未经验证的数字门槛。

3.5. 版本治理与失败案例管理

Skill 只有进入版本治理，才能由个人技巧转化为团队基础设施。每次修改均可能改变提示长度、工具调用、证据覆盖、失败概率和最终质量，因此需要记录版本号、内容哈希、变更原因、适用任务、实验结果、失败案例和回滚条件，并同步关联模型、工具、输入包和验证器版本。

失败案例应被视为资产的一部分。团队不仅需要保存“最优 Prompt”，还应保留成本下降但遗漏证据、首次通过提高但反方覆盖下降、Direct 模式越权回答、Report 模式陷入无进展循环等反例。后续版本升级应优先针对已经观察到的问题，而不是以“增强智能”“优化体验”等无法验证的描述作为变更理由。

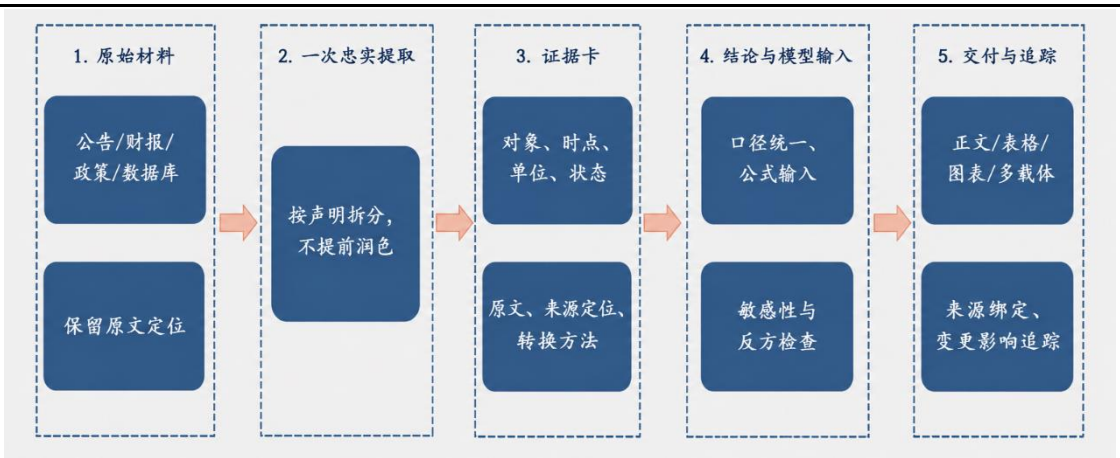
4. 证据型 Skill：降低搬运成本，同时守住质量边界

4.1. 一次提取、多次引用

投研任务中最值得复用的不是一段模型摘要，而是已经完成事实身份确认的证据对象。同一材料如果被正文、估值模型、图表和审校环节分别重新打开和摘录，不仅增加资源消耗，也容易发生“规划容量”在多次转述后被缩写为“容量”、最终被误写成“投运容量”等语义漂移。

证据型 Skill 因此将“提取”和“使用”分开。提取阶段首先保留必要原文片段，再拆分对象、数值、单位、时点和事实状态，登记来源定位和转换方法，并生成稳定的声明编号与来源编号。后续正文、模型、图表和审校均引用同一证据对象，来源发生变化时，可以追踪受影响位置，而无需逐章重新搜索。

图6：从原文到证据卡、结论和交付物的证据链



数据来源：东吴证券研究所绘制

一次提取并非任何场景都更省成本。QATC Lab 的证据卡复用实验显示，在仅约 800 字的短材料上，将材料结构化为多张卡片后，输入文本反而显著扩大；当材料达到年报级长度且存在多轮、多问题复用时，一次结构化提取才更容易摊薄成本。因此，证据卡属于有适用边界的优化机制，而不是所有任务都应机械套用。

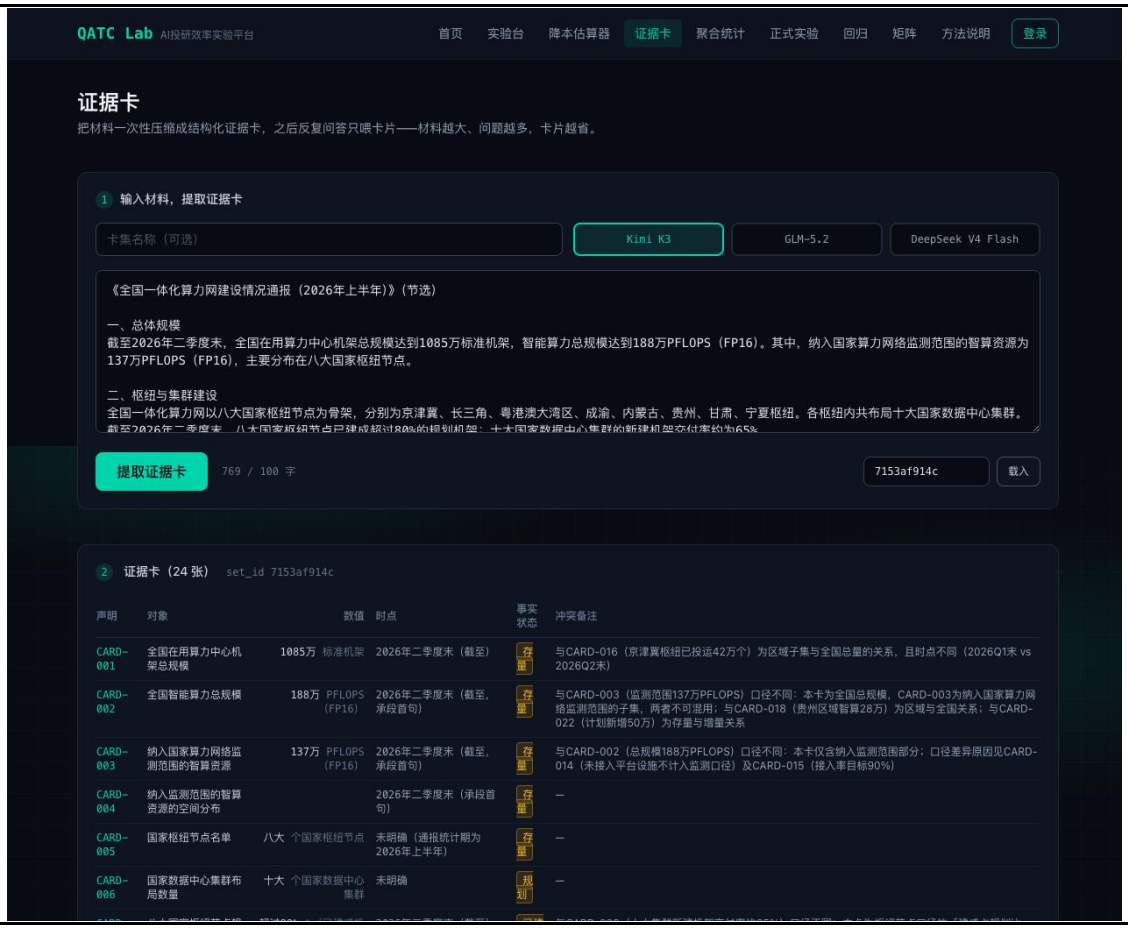
4.2. 证据卡的数据结构

表9：证据卡核心字段

字段	主要含义	控制目的
声明编号	单条可审计声明的唯一标识	连接正文、模型、图表和审校
来源编号	关联来源主表	支持多来源和来源谱系
原始文本	可回到原文核对的文字或表格片段	防止摘要替代证据
对象	统计或描述主体	避免统计对象被替换
数值与单位	原始或计算后的数值及量纲	防止单位混用
数据时点	统计期间或截止日	避免不同时点直接比较
事实状态	已发生、存量、在建、规划、预测、假设、推导等	区分现实与假设
来源定位	页码、表号、段落、网页小节等	支持复核
转换方法	单位换算、加总、比例、模型推导等	保证过程可追踪
冲突备注	与其他证据的时点、范围或定义差异	阻止静默选数
使用位置	章节、图表或模型位置	支持变更影响分析

数据来源：东吴证券研究所整理

图7：证据卡样例示意图



数据来源：QATC Lab 平台，东吴证券研究所

置信度可以作为辅助字段，但不能成为模糊判断的替代。高置信度至少应意味着来源身份明确、原文可定位、时点单位完整且口径与用途匹配；低置信度材料可以进入研究线索或待补证区域，但不宜在写作阶段被润色为高确定性结论。

4.3. 来源预算与证据门槛

来源并非越多越好。缺少预算时，Agent 容易围绕低价值背景信息持续扩展；如果只限制数量，又可能通过少量弱来源换取表面上的 Token 下降。更合理的方法是同时设置来源预算和证据门槛：预算决定资源优先投入到哪里，门槛决定什么材料能够支持什么强度的结论。

表10：不同结论类型的来源预算原则

结论类型	优先来源	主要检查	不宜采用的方式
政策类	主管部门、正式规则和政策原文	目标、执行范围、时点和法律状态	用媒体转述替代政策原文
公司类	公告、财报、投资者材料等正式披露	承诺、规划与已实现结果	将管理层目标写成现实结果
行业类	明确统计主体和样本范围的数据	指标定义、样本、地区和时点	用单一公司案例代表行业分布
研究判断	事实证据 + 模型假设 + 反方材料	可比性、敏感性和失效条件	用单一弱来源支撑高确定性结论

数据来源：东吴证券研究所整理

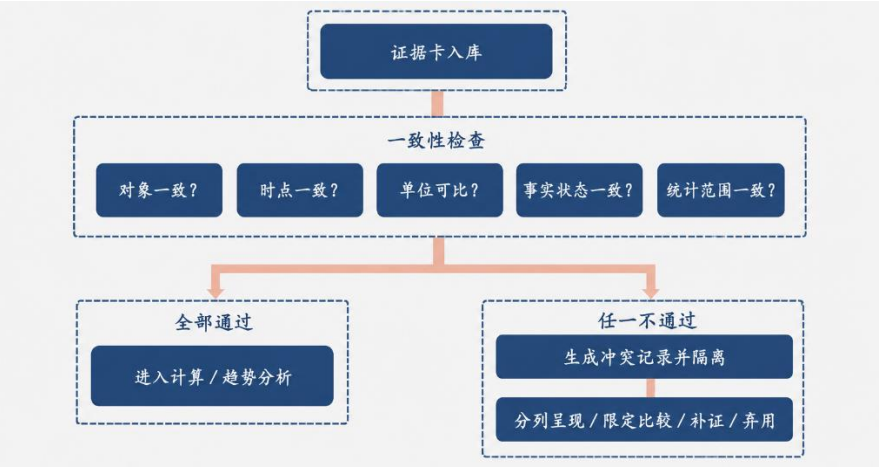
当证据门槛未满足时，合规处理应当是补证、降低表述强度、并列冲突或移出核心结论，而不是让语言模型补齐一个看似完整的答案。研究报告的完整性不意味着所有问题必须形成确定结论，能够明确披露“当前证据不足以支持进一步推断”本身也是研究质量的一部分。

4.4. 口径冲突闸门

投研中较危险的错误往往不是算术错误，而是把不能直接比较的数据自动合并。数值单位相同，不代表统计对象、时点和事实状态相同。证据型 Skill 因此需要在计算和写作之前设置口径冲突闸门。

至少有五类差异不宜自动合并：不同数据时点；未经验证换算的单位；存量与新增；规划与投运；总体与局部监测覆盖。触发任一差异时，应生成冲突记录并明确采用“分列呈现”“限定比较”“补证后统一”或“弃用”等处理方式，而不能默认为同一口径。

图8：口径冲突闸门示意



数据来源：东吴证券研究所绘制

即使比例、时点或状态等局部字段相同，只要统计对象发生变化，就不应因为数值“看起来相符”而判为部分准确。

4.5. 分层验证与质量达标后的硬终止

“查过来源”和“公式没有报错”只能说明局部检查通过。正式投研交付需要不同层级分别回答计算、证据、逻辑和投资可用性问题，而不能用综合评分替代全部控制。

表11：四层验证的职责分工

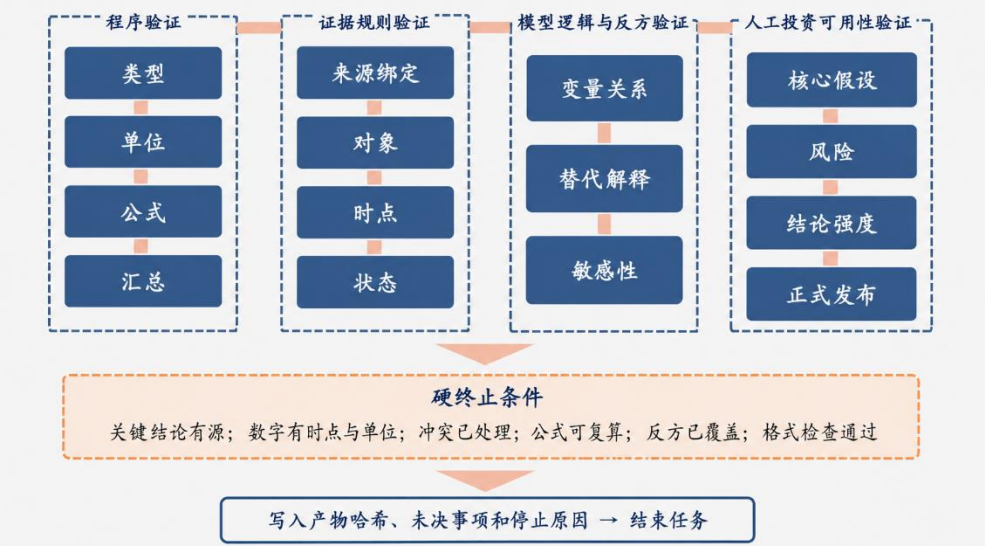
验证层	主要对象	回答的问题	典型执行者
程序验证	类型、单位、公式、汇总和结构	是否按设定正确执行	确定性程序
证据规则验证	来源绑定、时点、对象、事实状态	表达是否忠实于证据	规则引擎/程序 + 模型
模型逻辑与反方验证	变量关系、替代解释、敏感性和边界	即使计算正确，推理是否成立	模型 + 研究人员
人工投资可用性验证	核心假设、风险、结论强度和正式发布	结果是否可以进入目标交付	分析师/负责人

数据来源：东吴证券研究所整理

分层验证并非把同一内容重复看四遍。程序层不需要重新判断行业逻辑，人工层也不应替代公式复算。每一层应有独立对象和失败路径，并将错误退回最近的责任阶段。

质量通过后则需要硬终止。对于正式报告，我们认为至少应满足：关键结论有可定位来源；关键数字有时点和单位；主要口径冲突已解释；公式可以复算；核心反方得到覆盖；标题、表格、图注和引用等格式完成检查。任一关键结论未通过时，只能补证、降级表述或删除；全部通过后写入证据版本号、未决清单和适用截止日，并停止新增检索和无目的修改。

图9：四层验证与硬终止机制



数据来源：东吴证券研究所绘制

5. 实验验证：正式 48 轮显示“难任务见差距、复杂度有边界”

5.1. 四类任务与四种配置

若要判断 Skill 是否真正降本提效，不能只比较单次演示。Skill 会同时改变提示长度、工具调用、中间产物、验证步骤和返工路径；只有模型、输入、资源上限和评分口径保持一致，才能将差异主要归因于工作配置。

正式实验因此选择四类错误结构不同的任务：T1 政策与行业数据核验，重点检查事实身份和口径；T2 长报告审阅，重点检查问题发现、定位及误报控制；T3 估值模型，重点检查确定性公式、单位和依赖；T4 深度报告，综合考察多来源、分析和交付。

表12：正式实验四类任务及金标准

任务	主要风险	金标准重点	主要交付
T1 数据核验	对象、时点、范围和状态混淆	正确值、单位、期间、范围、一手来源、冲突解决	逐项核验表、冲突说明、来源
T2 长报告审阅	事实错误、引用错配、矛盾、漏项和误报	问题类型、位置、严重度、证据及可接受修法	问题清单、重点问题和修订建议
T3 估值模型	公式、单位、期间、硬编码和桥接错误	公式、值、依赖关系和容差	工作簿、估值摘要和验证结果
T4 深度报告	多来源、多章节、反方和交付完整性	关键事实、问题、反方、冲突及交付规范	报告、数据表、来源绑定和证据矩阵

数据来源：东吴证券研究所整理

四种实验配置并非由弱到强的等级，而是四种完整的工作方法。A 组为无 Skill 基线；B 组加入提示词和静态清单；C 组将研究步骤组织成流程型 Skill；D 组进一步加入主张—证据结构、程序验证和质量闸门。

表13：A/B/C/D 四种实验配置

配置	核心特征	新增能力	需要计入的额外成本
A 无 Skill	公共任务说明 + 基础工具	无专用流程	全部常规模型和工具调用
B 提示词 + 清单	增加冻结的自然语言规则	来源、口径、公式、反方等提醒	更长固定上下文
C 流程型 Skill	规划、检索、分析、撰写和自检形成状态流程	阶段产物、回退和终止	流程编排和中间状态
D 证据型 Skill	C + 证据账本 + 程序验证 + 质量闸门	结构化证据、断言和自动返工	验证、闸门、返工均纳入任务成本

数据来源：东吴证券研究所整理

5.2. 控制变量、独立 Session 与金标准

正式对照要求四组使用同一模型提供方及模型版本、同一输入包、数据截止日、输出要求和基础工具，并接受相同的墙钟、模型调用、工具调用及输出上限。配置本身造成的更长上下文、额外验证和自动返工均属于真实处理差异，不应为了“让成本相同”而扣除。

每轮实验从新的 Session 和独立运行目录启动，不读取前一轮输出、评分或其他配置的中间文件。输入包、公共提示、Skill、验证程序和关键工具均冻结版本并记录哈希。运行顺序采用轮换安排，以降低预热、执行者熟悉度和时段差异的影响。

质量金标准也在运行前冻结。它不是一份供模型模仿的“标准报告”，而是一组执行模型无法访问的可判定对象，包括关键事实及口径、有效来源、预埋错误、关键公式、必需反方和交付规范。评分员不查看配置名称、Token 和耗时，以降低先验偏差。

5.3. P1 先导实验：规则写出来不等于执行出来

P1 先导实验使用一项政策与行业数据核验任务，包含六条待核验声明，A/B/C/D 各独立运行一次。四组最终质量分别为 56/60、56/60、60/60 和 56/60。

表14：P1 先导实验结果

配置	C1	C2	C3	C4	C5	C6	总分	墙钟时间
A 无 Skill	8	10	10	10	10	8	56/60	54.466 秒
B 提示词+清单	8	10	10	10	10	8	56/60	49.667 秒
C 流程型 Skill	10	10	10	10	10	10	60/60	47.836 秒
D 证据型 Skill	8	10	10	10	10	8	56/60	54.139 秒

数据来源：QATC Lab，东吴证券研究所整理

注：先导每组仅 1 轮，墙钟差异不用于配置速度排名；该阶段未取得平台原生 Token，字符代理与 Token 口径分开。

差异集中在 C1 和 C6。A、B、D 均识别到原声明与来源之间存在对象替换，也能够写回相对正确的表述，但最终仍将冻结规则要求判为“不准确”的声明标为“部分准确”。C 组则执行了“关键对象不同即判定不准确”的规则。

该结果说明，发现问题与完成正确判定是两个不同环节。自然语言中的“必须”“不得”并不会自动转化为程序中的条件分支。若关键对象匹配仅存在于自由文本中，而没有进入结构化状态，质量闸门就可能无法稳定读取和阻断错误。

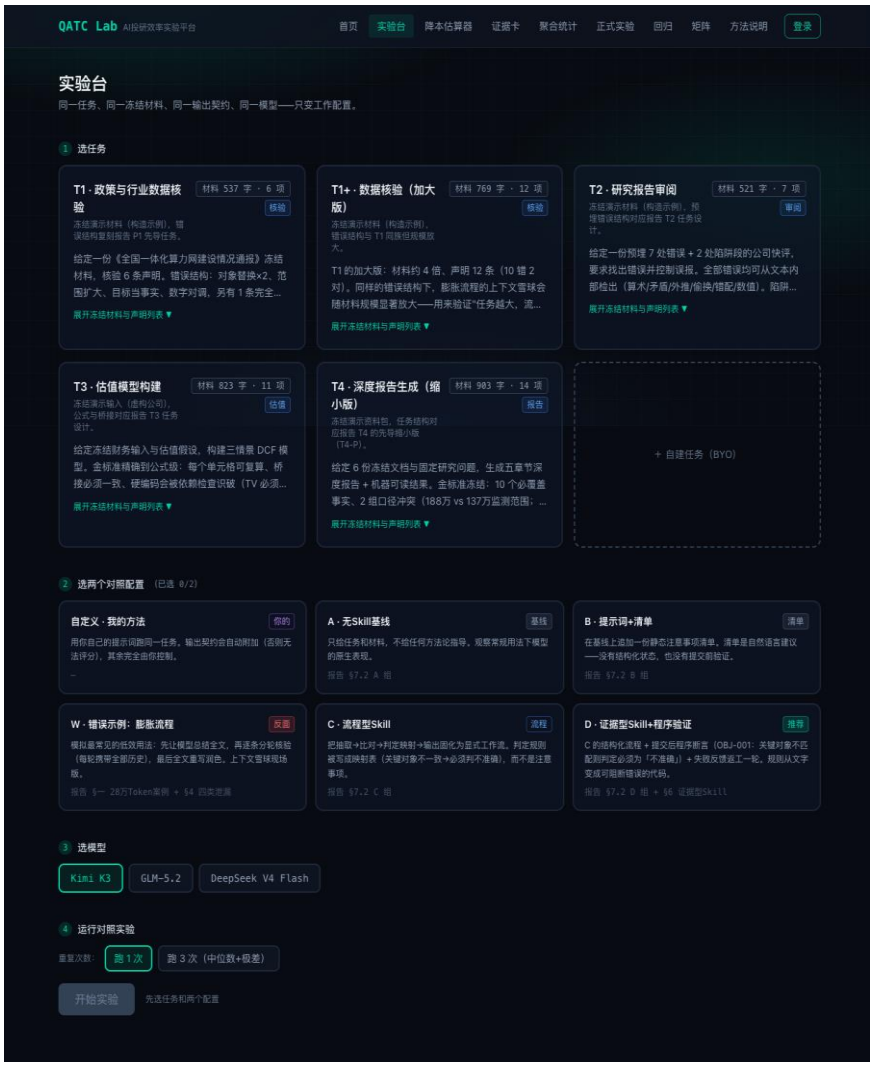
因此，一条真正可执行的规则至少需要三层：首先将对象、时点、数值、单位和状

态结构化；其次由验证器返回规则编号、实际值、期望值和允许修复动作；最后建立程序断言，在违反硬规则时阻止提交。P1 的意义并非证明 C 普遍优于 D，而是暴露出“规则写出来”和“规则执行出来”之间的工程差异。

5.4. 正式 48 轮实验结果

正式实验由 4 类任务×4 种配置×3 次独立重复组成，共 48 轮。实验采用冻结输入、版本哈希和轮换顺序，并记录平台原生 Token、墙钟时间、质量及处理失败。正式批次累计消耗 286,118 Token，总耗时约 5 小时 35 分。上述数值仅代表该冻结任务、模型和实验版本，不宜直接外推为真实研究项目成本。

图10：QATC Lab 实验台示意图



数据来源：QATC Lab，东吴证券研究所

表15：正式 48 轮实验主要结果

任务	配置	Token 中位数	质量中位数	首次通过/完成情况	处理失败
T1 数据核验	A 无 Skill	4114	100	首次通过 100%	0
	B 提示词+清单	3497	96.7	首次通过 100%	0
	C 流程型 Skill	4615	100	首次通过 100%	0
	D 证据型 Skill	3048	100	首次通过 100%	0
T2 长报告审阅	A 无 Skill	6114	0	首次通过 0%	0
	B 提示词+清单	8145	4.3	首次通过 0%	0
	C 流程型 Skill	4937	18.6	首次通过 0%	0
	D 证据型 Skill	4992	55.7	首次通过 100%	0
T3 估值模型	A 无 Skill	19649	100.0*	完成 1/3	2
	B 提示词+清单	—	—	完成 0/3	3
	C 流程型 Skill	16259	100.0*	完成 2/3	1
	D 证据型 Skill	12293	100.0*	完成 1/3	2
T4 深度报告	A 无 Skill	7479	100	首次通过 66.7%	0
	B 提示词+清单	8769	100	首次通过 100%	0
	C 流程型 Skill	8623	100	首次通过 100%	0
	D 证据型 Skill	8598	100	首次通过 100%	0

数据来源：QATC Lab，东吴证券研究所整理

注：T3 的质量与 Token 仅在形成完整交付的运行中可计算，处理失败单独列示，不能忽略失败后只比较完成样本。

5.5. 结果解读与适用边界

第一，任务难度决定配置差距。 T1 数据核验中四组质量中位数为 96.7—100，差异更多表现为资源消耗；T2 长报告审阅中，D 组质量中位数为 55.7，且首次通过率达到 100%，其余三组首次通过率均为 0。我们认为，结构化证据和程序约束的边际价值，在错误较隐蔽、需要定位和系统检查的任务中更容易体现。

第二，自然语言清单在困难任务上可能既不提高质量，也不降低成本。 B 组在 T2 中质量较低，在 T3 中三次运行均未完成。其含义并非“提示词没有价值”，而是当任务需要公式、状态和硬约束时，仅将注意事项写入自然语言不能替代可执行流程。

第三，复杂度存在反例边界。 T3 中 C 完成 2/3，而 D 仅完成 1/3。证据账本、质量闸门和返工机制本身也需要资源，当新增控制不能在资源上限内有效收敛时，名义上更完整的配置反而可能降低完成率。因此，质量闸门应当足够精确，并具有无进展退出机制。

第四，生成型任务未必需要最重配置。 T4 中四组最终质量中位数均达到 100。对于主要以组织和转译为主、金标准相对明确的任务，简单配置可能已经能够满足交付要求。此时继续增加 Agent 和验证步骤，需要以成本或稳定性改善证明其必要性。

第五，平均成本之外还应观察方差。 同一配置在重复运行中的消耗可能存在明显

差异。对于研究团队而言，可预测性本身具有经营价值：若一种方法平均成本较低但偶发产生数倍消耗和长尾等待，其排期和预算管理仍可能较困难。结构化流程是否降低方差，应作为后续扩大样本时的重要观察指标。

本实验每个任务一配置组合仅有三次独立重复，适合展示中位数、完成率和明显的机制差异，不足以形成统计显著性结论。实验任务为冻结材料，规模小于真实深度研究；执行模型、工具版本和资源上限变化，也可能改变配置间差距。因此，正式结果支持的是“方法具有任务依赖性、复杂度需要验证”的判断，而不是一组可以直接用于所有团队的固定节省比例。

6. 哪些优化真正有效：七项机制与三类常见误区

6.1. 结构化读取与证据复用

结构化读取的目标不是少读，而是减少无目的的全文重复读取。长报告或财报第一次进入流程时，先识别目录、表格、实体和相关页段；后续根据问题树和声明编号按需读取，并记录内容哈希、读取范围和阅读目的。需要全局理解时仍可阅读全文，但不应默认每个 Agent 都重新完成一次相同工作。

证据卡进一步将“读过”变为可复用状态。一张合格证据卡保留原文、对象、时点、单位、事实状态、来源定位、转换方法和冲突信息，使正文、图表、模型和审校共享同一对象。其效率收益来自长材料和多次复用，短材料和一次性任务则可能不值得支付结构化成本。

6.2. 程序计算、有限预览与局部修改

程序计算。 去重、筛选、排序、单位换算、公式填充、哈希、结构检查和固定规则断言，应尽可能由确定性程序承担。语言模型负责定义规则、解释口径和经济含义，而非反复执行可复算的机械计算。P1 中“关键对象不同必须判为不准确”即适合编码为程序断言。

有限预览。 搜索首先返回标题、主体、日期和短摘要，只有进入验证队列的材料再读取相关页段；数据文件先读取 schema、样例和统计摘要；验证器向模型返回失败项，而非完整成功日志。有限预览减少的是无关信息进入长期上下文，不应演变为只看摘要、不看关键原文。

局部修改。 内容层、证据层和呈现层需要解耦。修改数字时加载相关证据和段落；修改样式、图号或截图时不重新生成研究正文；只有核心论点、统一口径或章节依赖发生变化时才扩大修改范围。局部修改应配合章节哈希和回归检查，避免补丁造成跨章节矛盾。

6.3. 分层验证与显式终止

分层验证的重点是不同错误交由不同控制层处理，而非增加重复审阅。程序层检查公式和结构，证据层检查来源与事实身份，模型层检查因果和反方，人工层决定投资可用性和发布强度。验证资源应优先投入关键声明、模型输入和能够改变结论的风险点。

显式终止则解决任务“做完以后仍继续做”的问题。当必交物完成、关键主张有源、数字有时点和单位、冲突已处理、公式可复算、反方已覆盖且格式检查通过后，系统应立即写入成功状态。达到资源上限、不可恢复失败或需要人工裁决，也应视为合法终止状态，而不是继续以不同措辞重复尝试。

6.4. 并行 Agent 的双重效应

并行 Agent 的主要价值是缩短关键路径。例如，一组 Agent 分别处理互不重叠的公司原始披露，统一输出相同 schema 的证据卡；或者一个角色处理来源，一个程序复算公式，另一个角色进行反方检查。在子任务边界清楚时，并行可以显著改善等待体验。

但如果多个 Agent 均以“全面研究”为目标，就会重复接收长任务书、打开相同资料、形成相似中间结论，最后还需要汇总 Agent 再次读取全部产物并统一口径。此时墙钟时间可能下降，但总 Token、工具调用和合并返工上升。

因此，是否采用并行不宜以“任务复杂”作为唯一条件，而应至少检查三个问题：子任务能否独立；共享数据是否可以结构化复用；最终合并是否需要重新做一次研究。如果三项条件无法满足，并行更多代表算力换时间，而非降本提效。

6.5. 模型分层路由与任务—Skill 匹配

模型分层路由的合理方向，是让轻模型处理边界清楚、结果易检查的分类与格式化，让能力更强的模型处理证据冲突、复杂推理、反方和终审。但“简单任务用便宜模型、复杂任务用贵模型”仍只是一项机制假设，最终应按同一质量门槛下的任务成本验证。

文档类型识别、来源元数据整理、固定标签分类、字段格式归一等环节适合轻模型或程序；口径冲突、因果推断、替代解释和核心假设审查则更需要强模型或人工参与。若轻模型输出触发关键字段缺失、对象冲突或规则失败，应自动升级，而非为保持低成本继续在弱模型上重试。

从任务特征看，Skill 约束强度至少取决于确定性、材料规模、输出复杂度、错误代价和复用频率。高确定性任务更适合程序断言；大材料需要索引和证据卡；多载体正式交付需要项目状态和回归验证；高错误代价场景需要更高的人工和证据门槛；高频任务才值得承担深度 Skill 化的固定维护成本。

表16：真优化与伪优化的判别

动作	有效机制	容易滑向的伪优化	关键观察指标	不可突破边界
结构化读取	目录与页段索引、按需读取	只看命中片段、忽略全局限定	重复读取率、召回率	关键原文必须复核
证据卡复用	一次忠实提取、多处引用	过早摘要并重复传播失真信息	来源绑定、复用次数	对象、时点、状态不得丢失
程序计算	确定性批处理和断言	用程序替代口径与经济判断	公式错误、规则关闭率	程序正确不等于投资逻辑成立
有限预览	先看 schema、摘要和失败项	永远不读取原文和全量数据	工具返回体积、漏检率	核心结论必须取得足够上下文
局部修改	按影响范围更新和回归	只打补丁造成跨章节矛盾	重生成量、回归失败数	结构变化时允许全局修订
分层验证	不同层验证不同错误	重复审阅或用低层通过替代高层判断	首次通过、返工、重大错误	关键事实、公式和反方均需对应检查
显式终止	质量门槛通过后停止	为省资源提前结束	成功后调用数、未关闭问题	硬门槛未过不得标记成功
减少检索	去重转载、按预算停止	机械减少独立来源和反方	一手来源率、反方覆盖	证据门槛由结论风险决定

数据来源：东吴证券研究所整理

7. 组织落地：从个人技巧到研究团队 Skill 资产

7.1. Skill 资产目录

一次任务做对，并不意味着团队已经获得可复用能力。研究员可能在项目中找到有效检索方式、形成可靠的口径处理方法、写出一段表格校验脚本，或总结出减少全文重写的工作方式；如果这些经验仍停留在个人对话、临时脚本和口头说明中，下一个项目仍可能重新试错。

因此，团队需要管理的不是越来越多的“好用 Prompt”，而是一组能够被检索、调用、评测、审计、升级和退出的 Skill 资产。按照投研流程，可以形成检索、审稿、数据核验、模型、写作、文档交付和质量检查等七类资产。

检索 Skill 负责将问题转换为来源需求和候选池；审稿 Skill 负责发现事实、数字、引用和逻辑问题；数据核验 Skill 将声明拆成对象、范围、时点、数值和状态；模型 Skill 负责估值、情景和公式；写作 Skill 将已验证结论转化为不同载体；文档交付 Skill 管理 DOCX、PDF、HTML、表格和图片；质量检查 Skill 则封装来源绑定、公式正确、权限、首次通过和终止规则。

表17：Skill 资产登记卡

字段组	核心字段	主要要求
身份	Skill 编号、名称、类别、负责人、维护人	名称唯一，责任明确
任务边界	用途、适用/不适用任务、风险等级	说明解决什么、不解决什么
契约	输入契约、输出契约、必交物	关键字段可机器检查
运行机制	模式、状态、工具依赖、验证器依赖	记录回退、人工升级和终止路径
证据与质量	证据规则、质量闸门、固定任务集、失败案例	关键规则对应可复现测试
版本	版本号、SHA-256、变更原因和生效时间	正式运行可回到具体哈希
权限	数据分级、允许来源、允许工具和写入范围	明确读取、写入和外发限制
评测	成本、耗时、质量和兼容性结果	实验与生产监测分开记录
运营	使用次数、失败次数、评审时间和退役状态	避免无人维护资产长期累积

数据来源：东吴证券研究所整理

7.2. 创建、升级与回归测试

并非所有任务都值得 Skill 化。Skill 本身会产生固定上下文、维护、依赖升级和回归测试成本。我们认为，候选任务至少应同时具备四项特征：工作重复出现；执行路径相对稳定；错误代价较高；结果可以被验证。

满足上述条件后，可按照“候选—先导—稳定—退役”管理生命周期。升级申请需要说明旧版已经观察到的缺陷，例如重复读取、误路由、来源状态丢失、公式漏检、质量闸门无进展或成功后未终止，并提出对应机制假设和影响指标。

回归测试至少覆盖四项：成本回归、耗时回归、质量回归和兼容性回归。成本回归需要保持 Token 层级一致；耗时回归使用事件时间戳；质量回归重点检查重大事实、来源、冲突、公式和首次通过；兼容性回归则检查模型、文件格式、数据 schema、工具 API 和目标载体变化。

若新版在固定任务集上出现重大事实错误、关键公式退化、来源绑定下降、处理失败显著增加或无法复现稳定版本，应停止推广并回滚。Skill 的版本治理原则应接近研究团队管理估值模型和数据口径，而不是把它当作随时可以修改的个人提示词。

7.3. 权限、安全与人机职责边界

Skill 资产化后，权限风险也会随复用范围扩大。研究流程可能同时接触公开资料、内部底稿、客户材料和投资敏感信息，各类数据不宜共享同一默认权限。

公开资料仍需记录来源、时点和许可；内部数据只应向具有业务需要的身份开放；客户信息需要按项目隔离，默认不能跨客户复用原文；未公开投资判断、仓位、交易计划等敏感信息则应采用更高等级的最小权限控制。当权限不足、数据级别不明确或脱敏未通过时，正确行为应当是停止或升级授权，而非绕过限制。

AI 和程序更适合承担候选来源发现、证据结构化、确定性计算、固定规则检查、草

稿组织和运行日志保存。研究员至少应保留五项职责：定义真正的研究问题和质量门槛；裁决来源、会计口径和模型假设的适用性；判断哪些反方足以改变结论；审核核心结论能否进入正式交付；对最终发布承担责任。

在人机之间还需要明确接管点。当输入不足、关键来源冲突、核心假设无法验证、规则连续无进展或结论涉及高风险判断时，Skill 应生成结构化升级包，包括任务状态、已验证证据、冲突项、失败规则、已尝试动作和需要研究员裁决的问题。研究员的裁决也应回写原因和适用范围，以形成后续可复盘记录。

7.4. 30/60/90 天落地路径

团队落地不宜从一次性采购更多模型或建设“大而全”的 Agent 平台开始。更现实的路径，是先让当前任务可观察，再把少量高频流程转为可评测资产，最后形成版本和权限治理。

表18：研究团队 30/60/90 天行动表

阶段	重点目标	关键行动	产出	阶段判断
0—30 天	建立日志与基线	统一任务编号；记录模型、工具、文件、墙钟、人工和返工；盘点 Prompt、脚本和模板；选择高频任务	任务日志 schema、候选资产清单、基准任务包、评分卡、权限矩阵	能够回答“做了什么、成本如何记录、按什么验收”
31—60 天	封装少量高频 Skill 并跑先导	编写输入输出契约、状态、程序断言、失败反馈和终止条件；建立脱敏测试样本	首批候选 Skill、登记卡、固定任务集、先导日志和回滚草案	证明流程、评分、权限和回退可运行，不宣传单轮节省率
61—90 天	建立评测、版本和治理闭环	进行成本/耗时/质量/兼容性回归；区分探索版和稳定版；建立发布、回滚和退役流程	版本仓、回归报告、发布记录、访问审计、维护责任表	只有通过固定任务集和权限审查的版本进入稳定目录

数据来源：东吴证券研究所整理

前 30 天的目标并不是形成漂亮仪表盘，而是让一项任务从输入、模型版本、工具事件、首次产物、最终产物、人工介入到验收状态可以被串联。平台原生 Token 无法取得时，可以先保留字符和调用代理，但不应为了建立基线而制造假精确数据。

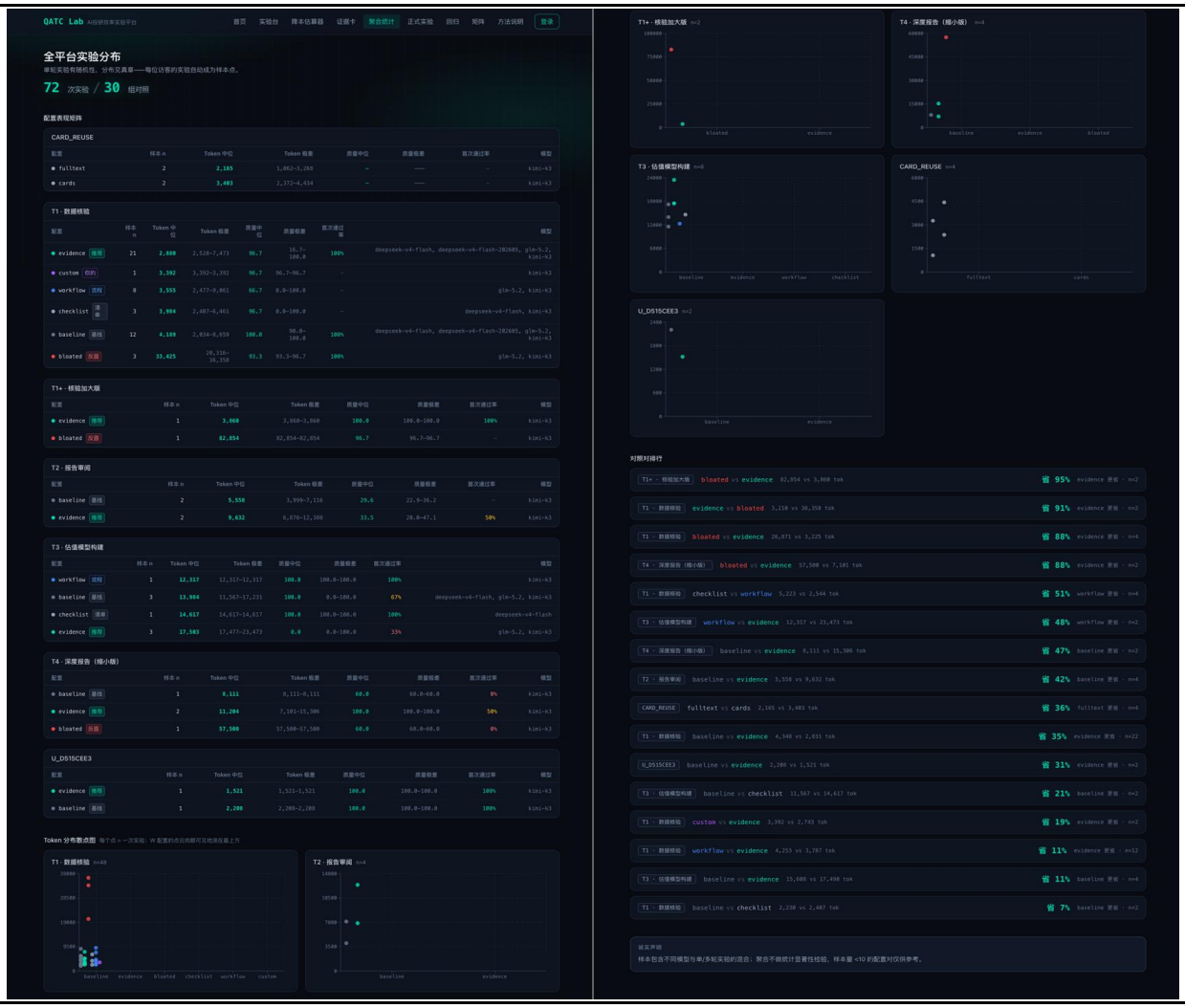
31—60 天宜只封装少量高频 Skill。每项候选资产至少包含正常任务、边界任务和预期失败任务，验证权限不足、证据冲突和规则失败时能否正确停止。61—90 天再将候选资产纳入稳定版本治理，并观察新人是否能够依靠输入契约和失败提示独立完成任务。

90 天之后，团队还需要持续清理低使用、无人维护或无法回归的资产，将新失败转化为脱敏测试案例，并跟踪模型、工具和数据接口升级。Skill 数量本身不宜成为 KPI；无法维护的 Skill 可能增加流程复杂度和隐性技术债务。

7.5. 公开验证平台 QATC Lab

团队已将上述方法论实现为公开平台 QATC Lab（qatc-lab.pages.dev）。平台提供：五类冻结任务与四种工作配置的对照实验（含自定义方法与用户自带材料入口）；三模型（Kimi K3 / GLM-5.2 / DeepSeek V4 Flash）任选对照与多模型矩阵；正式 48 轮实验矩阵的完整公开数据；证据卡自动提取与复用实验；提示词回归测试集；以及全平台实验的聚合统计。平台记录了每次实验的平台原生 Token、墙钟耗时、质量分与人民币成本，截至本文完成时已累计超过 60 次真实实验。“方法论 + 可验证平台”的组合，可能成为继数据接口之后投研工具的新形态：不只是提供数据，而是提供可测量的效率证据。

图11：QATC Lab 聚合统计——全平台实验按配置的中位成本、质量与首次通过率分布



数据来源：QATC Lab 平台，东吴证券研究所

7.6. 结论：把模型从证据搬运中解放出来

生成式 AI 进入投研后，最容易被看到的是模型单价、Token 数量和生成速度，较难被看到的是同一材料在多轮流程中的反复搬运、长上下文的累计、确定性计算的自然语言试错、局部修改造成的全文重写，以及质量不达标后的补证和返工。我们认为，AI 投研降本提效的核心并非追求更少的模型调用，而是减少“没有新证据、没有状态变化、没有质量贡献”的重复工作。

第一，效率单位应是一项合格任务，而不是一次调用。 模型、工具、人工、等待、返工和失败需要进入统一任务编号。只有达到冻结质量门槛的交付才能计为有效产出。低价模型如果需要更多轮纠错，未必更便宜；复杂方案若依赖大量返工才完成，也不能只展示最终结果。

第二，Skill 的价值在于重新划分模型、程序和研究员的职责。模型处理语义检索、口径解释、冲突分析、反方和结论表达；程序处理哈希、去重、筛选、单位换算、公式复算和硬断言；研究员保留问题定义、来源和口径裁决、核心假设判断以及最终发布责任。证据卡使三者共享同一事实对象，是降低重复搬运和控制语义漂移的重要中间层。

第三，流程复杂度本身不是能力。 更多步骤、更多 Agent 和更长规则都会增加固定上下文和协调成本。P1 先导表明，自然语言中写明规则并不能保证规则在最终判定中执行；正式实验进一步显示，提示词清单在困难估值任务上三次均未完成，证据型配置在部分任务中也因额外流程消耗而降低完成率。只有能够形成结构化状态、程序断言、明确失败反馈或有效终止的复杂度，才值得保留。

第四，方法收益具有明确的任务依赖性。 数据核验等简单任务中，结构化方法可能主要体现成本优势；长报告审阅等错误隐蔽的任务中，证据和验证机制同时改善质量与首次通过；组织和转译为主的生成任务中，简单方法也可能达到同等质量。因此，团队不应直接套用某一固定“节省比例”，而应按照自身高频任务建立冻结输入、统一质量门槛和重复实验。

第五，AI 投研的组织能力将逐步由“会用模型”转向“管理研究流程”。 当基础模型能力逐渐普及，差异更多来自任务定义、证据纪律、程序资产、质量闸门、版本治理、权限以及历史失败案例。研究团队如果能够将这些能力沉淀为可追踪、可测试和可回滚的 Skill 资产，AI 才可能由个人效率工具转化为稳定的研究生产基础设施。

8. 风险提示

（1）实验样本数量及任务代表性不足风险。 正式实验每个任务—配置组合仅进行 3 次独立重复，主要用于验证机制和观察明显差异，尚不足以形成具有统计显著性的普遍结论。冻结演示任务的材料规模和复杂度亦低于部分真实行业深度研究，实验结果可

能无法完整代表实际生产环境。

（2）模型及平台版本变化风险。不同基础模型在长文本、工具调用、结构化输出、代码执行和规则遵循方面存在差异，同一 Skill 更换模型后可能产生新的失败模式。模型服务商升级、上下文限制、缓存策略、价格和工具接口变化，也可能改变成本及质量表现。

（3）大模型幻觉及数据口径判断偏差风险。即使采用证据卡、来源等级、程序验证和冲突闸门，大模型仍可能出现对象识别错误、遗漏限定条件、引用错配、错误归因或不合理外推。结构化控制能够提高错误可见性，但不能完全消除模型偏差。

（4）Token、缓存和成本计量口径风险。不同平台对输入、输出、缓存读取、缓存写入和推理 Token 的定义并不完全一致；部分环境无法获取完整原生 usage。若将 tokenizer 估计、字符、字节或文件读取量与平台 Token 混用，可能导致效率比较失真。

（5）Skill 过度固化及流程复杂化风险。将阶段性投资判断、未经验证的阈值或固定来源写入 Skill，可能产生选择性注意和方法过拟合；过度增加 Agent、验证器和质量闸门也可能提高上下文、返工和系统维护成本。Skill 应持续通过固定任务集检验，并保留回滚和退役机制。

（6）数据安全、权限及合规风险。AI 投研工作流可能涉及内部研究底稿、客户材料、未公开投资观点及其他敏感信息。若数据分级、最小权限、外部模型调用边界、日志留存和跨项目隔离不足，可能产生信息泄露或合规风险。正式机构部署仍需按照所在机构的信息安全、模型风险、数据治理和供应商管理要求进行独立审查。

东吴证券 SOOCHOW SECURITIES

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持：预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性：预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持：预期未来 6 个月内，行业指数相对弱于基准 5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址：http://www.dwzq.com.cn