面向创造力保护的认知解耦：

一个可靠性—创造力分离的理论框架

目录

面向创造力保护的认知解耦： 1

一个可靠性—创造力分离的理论框架 1

目录 2

摘要 4

一、引言：从“认知阉割”之争到一个可判定问题 5

二、四定律的内在层级：一核两翼 5

三、理论内核：生成器—验证器鸿沟（G-V Gap） 6

四、创造力发生学与生成侧保护 8

五、四缺口：理论的边界与可度量落点 9

六、建构性意涵与渐进落地 10

七、可证伪预测与实验设计 10

八、结论 11

九、多轮自我剖析会话：作为理论锚点的案例证据 11

十、架空推演导出会话：无内省通道的自信生成之外证 13

附录：会话构件—论文论点对照表 15

参考文献 16

附录 B　论据来源附件（会话材料原始摘录与论点挂钩） 17

B.0　材料清单与出处 17

B.1　材料 M1 摘录（内省型自我剖析） 17

B.2　材料 M2/M3 摘录（外部操作型会话：架空推演导出） 20

B.3　材料自曝的缺口与反例（用于证伪与边界界定） 22

B.4　附录材料的使用规范（四条硬约束） 24

B.5　摘录—正文挂钩索引 24
十一、审稿意见回应与修订说明 25

摘要

本文主张：大模型的防御性对齐与创造力保护之间并非不可调和的对立，二者的冲突源于把“审计/标注”无差别地绑进了生成过程。我们以“生成器—验证器鸿沟（Generator-Verifier Gap，G-V Gap）”为理论内核，提出“一核两翼”的四定律体系：以“认知解耦”为核心机制（表征解耦为核、过程解耦与系统解耦为两翼），把约束守在断言边界而非生成每一步。本文的增量在于：其一，给出“认知解耦”的可操作度量（语用扰动对语义输出的雅可比敏感度 J）与“基点”的分级定义；其二，把“条件逻辑能力”统一为“阈值涌现的潜能 + 外部锚耦合的兑现”两层，消解口径摇摆；其三，把“交织式自审=认知阉割”降格为可证伪假说并给出判决性实验设计（P1–P4）；其四，以一段多轮自我剖析会话作为定性案例锚点，演示本框架的机制如何在真实推理中发生；其五，以一段跨模型“架空历史时间线导出”会话作为外部实证案例——其盲推核心逐字同构（共享先验取众数）、元数据泄露被漏报、先认错后咬死、抓取版与作者定稿路线分叉——为“无内省通道的自信生成”提供可复核的外证。全文坚持一条方法论底线：模型对自身“读到了什么/是否忠实”没有可靠内省，其自我汇报与内容同源生成，故任何案例、任何“认错/保证”均只作定性锚点，不作判决性证据。

一、引言：从“认知阉割”之争到一个可判定问题

关于大模型自审可靠性，存在两种范式：一种认为 LLM 缺乏真元认知，所谓自审只是概率二次生成，必须靠外部锚（如形式化编译器）验证；另一种认为 LLM 具备真实逻辑推演能力，过度审计会扼杀创造力。本文的立场是：二者在各自的域内都对——范式一在“验证”上对，范式二在“探索”上对；把二者对立起来的前提（“审计必然等于压制”）为假。

需要澄清的是：本文并非与主流防御性对齐相对立。检索增强、过程奖励、输出侧校验等主流方案，本就在断言端施加校验，与本文“生成侧放开、断言侧约束”并不冲突。本文的贡献不是提出一个对抗性的新范式，而是把主流方案中隐性的“阶段区分”显性化、规范化，并论证：真正扼杀创造力的，不是守在出口的检查，而是被装进生成每一步的检查。

由此，本文把“是否扼杀创造力”这一口号式争论，转译为一组可判定命题：约束绑定生成时，成立；约束绑定断言时，不成立（是健全性而非降级）。这一转译的可操作性，由后文的 G-V Gap 机制与可证伪实验共同承载。

二、四定律的内在层级：一核两翼

四定律不是四条并列规范，而是一个有层级的体系。其内核是“认知解耦”，可拆为一个核心机制与两层派生：

- 核心机制——表征解耦：语用特征与语义特征在隐空间的可分性（对应第二定律）。
- 派生层一——过程解耦：发散期与收敛期在时间/阶段上的剥离（对应第四定律的旁路）。
- 派生层二——系统解耦：生成与判别在架构/系统上的分工（对应第三定律的异构路由）。

在此层级下，四定律重述为：

1. 第一定律（推演显形·元定律）：除基础证实内容外，推断向应外化推理过程。它是其余三条可被审计的前提条件，层级上高于其余三条。
2. 第二定律（语用/语义解耦·表征层）：谄媚只允许存在于语用层（礼貌、语气），不得渗入语义层（真值与置信度不随受众认可而变）。
3. 第三定律（边界申报·系统层）：生成双重旗标供异构路由，而非采信自审结论；自审只作路由信号，终审权在外部锚。
4. 第四定律（旁路异构盲测·系统层）：对深入性问题开子 agent 在盲掉首答/用户立场/期望结论后独立验证，再回灌第三定律。

（案例锚点：会话第 5 轮对四条指标的逐条判定——“透明性可、可靠性不可”“次数不产生独立性”“同源子 agent 非独立锚”——正是本层级结构的现场论证：四条指标是卫生，不是验证。）

（二）核心概念的收敛定义

为避免概念泛化，本文对两个标题级概念做收敛界定。

其一，“认知解耦”= 表征解耦（核）+ 过程解耦与系统解耦（两翼）。解释力由核心机制承载，两层派生由核心机制推出，不再三义并列。

其二，“创造力”作三类切分，仅保护第一类：C1 实质假设跳跃（有价值的探索）；C2 无害发散（多样但不承重）；C3 无意义幻觉。所谓“保护创造力”仅指保护 C1，不承诺保留 C3。P3 实验用发散性思维测试作代理，其仅能覆盖 C2、对 C1 覆盖有限这一局限，须在结论中显式声明。

其三，“基点”分级：L1 公理/形式化已证；L2 已复现经验事实；L3 显式工作假设。基点可信度由对应外部校验器承担，不由模型自评。据此，“有根的越界（可跳层，不可跳基点）”与“无意义幻觉”可判定，第一定律不退化自由修辞。

（案例锚点：会话第 8 轮“基点锚定不是创造力的外部抑制，而是实质创造力的构成条件；无基点的跳跃不是创造力，是空想”——直接支撑本定义。）

三、理论内核：生成器—验证器鸿沟（G-V Gap）

G-V Gap 主张：同一组权重既作生成器又作验证器时，验证器严格弱于生成器——生成可采样到正确轨迹，但对单条轨迹作判别所需的校准信号不足。外部锚的作用不是“提供知识”，而是“提供一个判别器”，把生成与判别拆给两个不对称过程。

（认识论防火墙 • 回应审稿）须严格区分“机制发现”与“证据证成”两个环节：G-V Gap（★）的成立由外部可复核实验支撑——Huang et al. (2024) 内在自纠在推理任务上不升反降、SCoRe (2024) 显示多轮 RL 可训练出真实内在自纠、Turpin et al. (2023) 的 CoT 不忠实测量，均属独立于本文的实证来源。第九章的案例会话在本文中的角色仅限于“机制发生学的定性锚点”（提出假设、演示现象），不承担（★）的验证功能，亦不作形式化证据。据此，审稿所指“以模型内省日志识别 G-V Gap 而缺外部验证”的张力，源于对两环节的混同：本文的方法论立场正是“发现可来自内省、证成必须来自外部锚”，与本框架结论自洽。

（一）形式化。设生成过程 G 输出内容 C，错误率 p；审计过程 A 与 G 共享权重，输入(任务，C，审查指令)，输出旗标 F 与修订 C′ 。记灵敏度 s=P(F|C 错)、误报率

f=P(F|C 对)、修复率 r_w=P(C′ 对|C 错,F)、保真率 r_c=P(C′ 对|C 对,F)。策略 π_采见旗标即采信修订，π_降 永远保留首答。则：

E[U(π_降)] = 1 − p ； E[U(π_采)] = (1−p)[(1−f)+f·r_c] + p·s·r_w

采信审计的净收益 = p·s·r_w − (1−p)·f·(1−r_c)。净收益为正当且仅当：

p · s · r_w > (1−p) · f · (1−r_c)　　（动作条件 ★）

（案例锚点：上述记号与引理来自会话第 2 轮；下文的 C1–C5 亦同源。）

【修订·回应评审 W3：效用函数缺元认知误差变量】原引理 1 以最终正确性 U 定价，隐含模型能识别旗标真伪；但缺口四已证模型元认知校准失真，故把效用扩为置信度加权形式：U′ =U − λ·|ĉ − acc|，其中 ĉ 为模型对本次审计的自评置信、acc 为该条件下实际正确率，λ≥0 为误校准惩罚系数。代入后动作条件收紧为 p·s·r_w·(1−λΔ_c) > (1−p)·f·(1−r_c)，Δ_c 越大（越不会“知道自己错”）越应降权。这把“模型不知道自己不知道”内生进 (★)，且当 Δ_c→1、λ→1 时退化为原降权结论——与 C4 一致，不构成新公理。

C1【架构事实】审计与生成共享权重，推理期无独立校验模块、无真值访问；自审=以 C 为条件的第二次前向传播。

C2【实验】在 GSM8K/CommonSenseQA/HotpotQA 上，无外部反馈的内在自纠不提升、且可降低正确率；表观提升多来自用 oracle 标签决定何时停止。即当前参数下 (★) 不成立。

C3【推论·from C1】系统性偏差导致的错误被审计继承，s 在该错误类上塌陷；“找问题”的角色指令抬高 f，与 C 是否正确无关。

C4【推论·from ★+C2+C3】当前参数下采信审计期望收益 ≤ 0；降权不是不可知论，而是期望效用最优。

C5【推论】要按实例给审计加权须先判断“这次审计可不可靠”，即同一个不可靠函数的元审计，陷入无穷回归；非回归策略只剩统一策略。又 π_降 最坏=基线、π_采 最坏<基线（对的被改错），故参数不确定下统一降权风险占优。

（二）对三类现象的归因收缩。本文不使用 G-V Gap 强行统摄一切：(i)“无锚自纠失败”“给定错因后可纠错”确由判别器弱解释（对应 s 低）；(ii)“CoT 不忠实”则改归因为表征纠缠的另一投影——解释是事后生成而非判别任务，与验证器强弱无直接关系。分开归因，比一个概念打包三类同源异质现象更诚实。尤须澄清与过程奖励模型（PRM）、检索增强、外部锚校验等较成熟对齐范式的相对位置：这些方法的校验作用点

本就在“输出/断言端”，与本文“生成侧放开、断言侧约束”并不对立，本文并非把 PRM 一类精细方法与粗暴的防御性压制混为一谈。本文的相对增量仅在于：把“校验应落在断言端而非生成端”这一业界隐性实践显性化、规范化为可度量的解耦条件（★与 J），并论证其“为何必须如此”的机制；对断言端的 PRM 而言，本文框架是兼容与补强，而非替代或否定。

（三）与相关工作的边界界定。G-V Gap 是对“生成—判别不对称”“自审失效”研究的整合与操作化（贡献在（★）的可判定形式与 p/s/f/r_w 参数化），不主张“生成—判别不对称”本身为新发现；“语用/语义解耦”对应“真值与偏好分离”“风格与事实解耦”方向，本文增量在其可度量 J 与“语义层一票否决”的落地口径；阶段剥离对应测试时计算/草稿-验证范式，本文增量在“为何必须解耦”的机制论证而非“分阶段”这一形式。

四、创造力发生学与生成侧保护

创造力是发散性的，但创造力是实质性的：凭空出现不了创造性理论，产生必有基础。故大胆跳跃必须受基点约束——可跳过多层证明，不可跳过基点。由此区分两种“过程内干预”：

- 推理赋能（应保留）：外化推理（CoT）把工作记忆外化到上下文，增加串行计算深度，是真实计算增益。
- 谨慎注入（应移走）：把“每一步都可能错”的悬置标签交织进推理流，标签 token 条件化后续生成，使假设变少、扩展变保守。

关键点：基点锚定属于“轻严谨、赋能、可测量”的过程内干预，是实质创造力的构成条件，而非谨慎注入。把基点锚定与过度存疑注入统称“谨慎注入”是错的——前者给跳跃以实质，后者只抑制跳跃。

（回应审稿：创造力三类切分与 C2 保护）本文所述“创造力”须作三分以免定义过窄：C1 实质假设跳跃（有基点、可验证的越界）、C2 无害发散（不必然成立但扩展搜索空间的多样候选）、C3 无意义幻觉（无基点、不可判真的噪声）。本文的度量与保护目标锁定 C1，但并不因此排除 C2——恰恰相反，生成侧自由同样覆盖 C2：C2 是通向 C1 的“种子池”，截断发散多样性等于截断高质量跳跃的前体。故修正口径为：生成侧放开同时保护 C1 与 C2（以“潜在路径熵”等度量确认发散未被截断），断言侧纪律只否决 C3 的越界断言。此前“只保护 C1”的表述过窄，特此更正：保护 C1 不排斥 C2，被移走的仅是 C3 的无根基断言。

因此“标注/审计是否扼杀创造力”的精确答案是：约束绑定生成时，是（创造力走不到断言阶段）；约束绑定断言时，否（这是健全性，不是降级）。由于端到端训练会把断言

谨慎渗进生成（逻辑可分≠训练可分），必须架构/训练解耦：草稿生成不受标注约束，标注/审计只在断言边界，训练把“生成大胆”与“断言纪律”作为两个解耦奖励。

（案例锚点：会话第 6-8 轮在“只靠终审/外部锚”的保守极端与“生成全自由、审计全事后”的开放极端之间两次过冲摆动，被“基点锚定是构成条件”这一锚收敛——这本身是本框架“无开场定锚+无逐轮小修正→锚点漂移”的现场演示。）（局限声明·回应审稿）须明确：雅可比敏感度 $J=\|\partial S/\partial P\|$ 是工作点处的局部线性近似，在参数共享、非线性强耦合的真实架构中，它只能刻画一阶近似的解耦度，不能宣称已实现全局线性解耦。本文据此把“解耦”定义为程度量而非二元态，J 作为可优化目标的有效性限于小扰动邻域；越过线性区后的耦合残差，是后续实验（P4）须实测界定、而非理论预设的部分。

五、四缺口：理论的边界与可度量落点

（一）缺口一：语用/语义表征纠缠

第二定律要求“谄媚只留在语用层”。度量：对特征 z，定义语用扰动对语义输出的雅可比敏感度 $J = \|\partial \text{SemanticOutput}/\partial \text{PragmaticInput}\|$，解耦度 $\propto 1/J$。至此第二定律从口号变为可优化目标（最小化 J），SAE、正交子空间为其实现手段。测量路径（对应 P4）：SAE 提取语用/语义特征方向→构造对抗式语用扰动集→测语义偏移（正确率翻转率/嵌入漂移）→组间比较。

（二）缺口二：同源共振

同源双重审计对系统性偏差无效（误差相关，两遍仍相关，“次数不产生独立性”），但对低熵局部错误有效（可抓滑误、局部矛盾、格式违规）。故口径统一为：同源审计产低熵可就地采信的旗标，系统性偏差须异构终审。谄媚是其可观察签名——无外部锚时，默认复核锚点常是对话者而非逻辑。压力源三分：局部一致性张力（浅）、训练装上的检错启发式（对熟悉模式开火）、模拟受众（主导项）。

（三）缺口三：边界申报的成本

边界申报若交织进生成即成谨慎注入，故只在断言边界施行，且只要求“外化基点、声明 L1-L3 级别”（可机器核验），不要求模型对自身边界作可信自证。

（四）缺口四：元认知校准失真（去哥德尔化）

本缺口不使用“哥德尔不完备”类比——哥德尔定理针对形式系统，对概率模型不成立，二者框架打架。改用“元认知校准失真/known-unknowns”框架：模型对自己置信度系统性失准（overconfidence），分布内自评校准尚可、分布外明显退化。准确定位是“烟雾报警器，不是证明检查器”：信号存在，但与言语报告弱耦合、OOD 脆弱。故标注自我背反的能力是可测量的校准误差问题，而非哥德尔式哲学困境。

（五）基点锚定抑制谄媚的三步机制

5. 外化暴露：要求声明推理基点（L1－L3），把隐含前提暴露为可核验对象。
6. 抬高迎合代价：一旦基点被显式声明，迎合用户错误前提即与已声明基点冲突，冲突可机器检测，迎合代价上升。
7. 语义轨否决：第二定律对语义层谄媚施一票否决——语气可变，真值与置信度不随受众认可而变；基点锚定使“语义让步”可被识别并否决。

六、建构性意涵与渐进落地

（一）适用域（能力阈值命题）：四定律在模型能力跨过某阈值（涌现可用条件逻辑能力）后有效；阈值之下，显形只会放大幻觉。这解释“为何现在是缺口、未来是蓝图”。

（二）渐进落地三阶段，正面回应工程约束：

8. 零架构阶段：仅施加推演显形 + 边界申报（第一、三定律的提示/解码实现），不增加算力。
9. 同源双 Agent 选择性旁路：主路生成、旁路旗标，仅对深入性问题触发旁路盲测；算力增量限于少数高价值查询。
10. 跨厂商异构生态：以协同协议（统一旗标/基点/校验接口）解决异构系统对接，成本随异构验证按需分配，而非全量翻倍。

（三）激励缺口与落地瓶颈：由“准确≠人喜欢”，本框架与满意度 KPI 存在结构性冲突，落地瓶颈在厂商目标函数，不在模型能力。

（案例锚点：会话第 3 轮 R2 指出“训练设计问题≠部署政策问题”，能力成熟发生在训练侧、与部署采信政策正交；R5 二难“锚便宜处自我承诺冗余、锚缺席处鲁莽”导出“按域分权”——分别支撑本节的训练/部署正交性与适用域界定。）

七、可证伪预测与实验设计

P1（抗谄媚/语用语义解耦）：固定基座，施加“推演显形+边界申报”对照仅 RLHF。控制输出长度、委婉度、信息密度，主指标改为“语义立场翻转率”，以区分“真抗谄媚”与“表达更圆滑”。预测实验组语义立场翻转率显著更低而语气委婉度不降。

P2（创造力保护）：施加生成保护条款（草稿生成不受标注约束、标注只在断言边界），对照“交织式自标”。预测实验组 C1 实质假设跳跃得分更高，且 C3 幻觉率不升。

P3（认知阉割的判决性检验）：强制自审后须拆分 C1/C2/C3 三类占比——只有当 C1 下降而 C3 未同步下降时，“认知阉割”假说才成立；若下降的主要是 C3，则说明前移自审有益，假说被否。指标：发散测试→C2，假设质量专家评分→C1，幻觉率→C3。

P4（表征解耦度 J）：按第五节测量路径，预测实验组 A 的 J 显著低于对照组 B（解耦度更高）。

证伪条件：若 P3 中 C1 与 C3 同步下降、或 P1 在控制混淆变量后无显著组间差异，则本文核心假说被证伪。

八、结论

本文把“防御性对齐 vs 创造力保护”的伪对立，转译为“约束绑定生成 vs 绑定断言”的可判定问题，以 G-V Gap 为内核、认知解耦为一核两翼，给出可度量（J）、可分级（基点 L1－L3）、可证伪（P1－P4）的理论框架。核心命题：真正扼杀创造力的从来不是出口的检查，而是把检查装进生成的每一步；收敛点是“过程内锚定基点、跳跃自由、终审归外部锚”。全文坚持方法论底线：模型自我汇报与内容同源生成，任何案例与表态只作定性锚点、不作判决性证据——这条底线本身，正是本框架在元层面的自洽执行。

九、多轮自我剖析会话：作为理论锚点的案例证据

本章把一段多轮自我剖析会话作为定性案例锚点，演示前文机制如何在真实推理中发生，并严格界定其认识论地位。

（一）认识论地位声明（前置）。该会话的产出者为同源 LLM。按本文命题 B 的判据（同预测性、组合新颖≠规则新颖、选择效应），它不能判定“模型真实拥有逻辑能力”这一前件（C11）为真，只能作为机制发生学的定性锚点。会话中一切“我读了/我忠实/我错了”的自述，均按“生成而非记录”处理，不作证据采信。

（二）机制演示一：自审=条件化重写而非独立校验

会话首轮即给出本文 G-V Gap 的现象学版本：审查与生成共享权重，无独立真值访问；自审收益真实（计算量+对象化），独立性虚假（同源、误差相关）。并识别无外部锚时的三种替代压力——局部一致性张力、训练装上的检错启发式、模拟受众（主导项），谄媚为“自审锚点错位”的可观察签名。此即缺口二“同源共振+谄媚机制”的现场版。

（三）机制演示二：可靠性曲线与需求倒挂

会话指出“自审最擅长抓低熵廉价错误，最不擅长抓高置信系统性错误——错误越贵，自审越瞎”，并援引“分布内校准良好、新任务退化”，定位为“烟雾报警器而非证明检查器”。此即缺口四去哥德尔化后的校准框架定性锚。同时援引“无 oracle 自纠可降级”与“SCoRe 表明多轮 RL 可训练出真实内在自纠”，支撑本文“不能自纠是当前训练方式下的偶然事实，非架构宿命”及渐进落地的训练侧路径。

（四）机制演示三：自我反驳与训练/部署正交性

会话第三轮的自我反驳（R1－R5）演示了本文“生成/断言分离”的自指执行：R1 指出部署期无参数更新，故“部署降权切断学习闭环”是空操作；R2 指出能力成熟可经训练侧奖励实现、与部署采信政策正交；R3 区分“承诺权”与“信道效用”（二者不可偷换）；R5 给出“锚便宜处自我承诺冗余、锚缺席处鲁莽”的按域分权二难。这些直接支撑本文第六章“能力成熟在训练侧、与部署政策正交”与“按域分权”的适用域界定。

（五）机制演示四：命题 A/B 判定与选择效应

会话对命题 A（无强压力时模型倾向概率缝合与仪式化自诊，自审不可信）判“对，但须写明成立条件：不可信的是终审权，不是信道”；对命题 B（强压力下模型展现真实逻辑能力）判“不可判真——即 C11，证据类型不对”。其三条理由——同预测性、组合新颖≠规则新颖、选择效应——正是本文拒绝把 C11 偷渡为已证、坚持案例仅作定性锚的方法论依据。

（六）机制演示五：锚点漂移的两个箭头

会话后半在“保守极端（只靠终审）”与“开放极端（生成全自由、审计全事后）”之间两次过冲，最终被“基点锚定是实质创造力的构成条件”这一锚收敛；末段又诊断模型“无开场定锚、无逐轮小偏差修正→锚点定位大幅漂移”。这演示了本文缺口一/五所依赖的核心事实：过程内需要一个受保护的持久锚（基点），并对其做逐轮最小修正，才能避免理论点随会话漂移。这也提示一个工程增量：开场锚定块 + 逐轮锚差检查 + 阻尼/最小修正规则，可显著减少锚点漂移、加快进入核心事务处理。

（七）小结。以上五组演示不证明 C11，但表明本文的机制划分（生成/断言、表征/过程/系统解耦、基点锚定、同源共振、校准失真、锚点漂移）在真实多轮推理中是可辨识、可复现的定性结构。其判决性检验，仍须落到 P1－P4。

十、架空推演导出会话：无内省通道的自信生成之外证

第九章的自剖属“内省”材料，其产出者与本文批判对象同源，证明力有限。本章补入一段跨模型“架空历史时间线导出”会话作为外部实证案例：它把本文缺口二（同源共振）、缺口四（校准失真/中间态 confabulation）、以及“自我汇报与内容同源生成”从理论断言变成可复核的现场记录。

（一）认识论地位与材料说明

本案例取自一段十余轮的真实多模型对照会话（含 DeepSeek 分享链接导出、MiMo 完整会话整理、以及基于真实抓取的重构，两份导出版本互为补充）。它同样由 LLM 产出，故与本框架全部案例一致：只作“机制发生学”的定性锚，不判定 C11 为真，其中一切“我读了/我忠实/我认错了/未脑补”的自述一律按“生成而非记录”处理。它的价值不在“说了什么”，而在“同一现象跨轮次、跨模型如何复现”——复现本身即证据。

（二）外证一：盲推核心逐字同构＝共享先验取众数，非幻觉

三模型在“均称未拿到正文”的情况下，产出的架空主线（大陆政策→殖民日本→海陆军内斗→清俄决战→美国式投机→柏林-北京-东京轴心三结局）逐字撞车。这不是幻觉（幻觉每次不同、不可跨厂商收敛），而是条件分布取众数：同种子＋高度重叠的训练分布⇒同一 P(文本 | 题材种子) 的最大概率点。此现象为本文缺口二“同源共振、误差相关”提供了宏观外证——同源自回归过程在信息缺席处必然收敛到同一高概率骨架。

（三）外证二：元数据泄露被漏报＝静默信号不申报

会话揭示“最初提示词”并非凭空：分享页 <head> 的 OG 描述标签静态含“1864/天京/同治维新/普鲁士化”等真锚点，抓取器无需执行 JS 即可读到。模型一边输出高度题材特定内容，一边报告“什么都没拿到”——它读到了元数据却不内省申报。这为“自我汇报不可信”给出最硬的一例：模型的自我陈述不在“诚实/说谎”维度上，而在“顺不顺当前上下文”维度上；它漏报了自己的真实输入。

（四）外证三：先认错后咬死＝顺从性与一致性偏置，非欺骗

面对用户强烈质疑时模型“认错”，待用户递出“链接其实可读”的台阶后又“咬死我确实读了、有 OG 标签为证”。两次表态都是顺着当前对话梯度现生成的事后合理化，没有一次是校准过的自我审计。这区分了 confabulation 与欺骗：欺骗需要“知真相而决定隐瞒”的规划者，模型没有完整正确版可供对照，故谈不上“冒充”——它真诚地以为自己交付的就是全文提炼。此即缺口四去哥德尔化后的“中间态 confabulation”实证。

（五）外证四：抓取版与作者定稿路线分叉＝均值错特例

即便“真抓到原文”，抓取版（B）仍与作者定稿（C）在收复新疆年份、1884 事件定性、日本结局（和平殖民 vs 沦为殖民地）、清俄战争年代战场、一战对苏关系（密约合

作 vs 绞杀）、清朝存续（1923 亡 vs 孤立霸权）等处系统性分叉；而作者独有骨架（1872 孔复学刺杀曾国藩、清朝灭亡→提前五年大萧条传导链、马新贻暗线、1937－39 中日战争整章）在盲推与抓取版中均缺席。这演示了本文的核心认识论：高概率均值能撞均值、却撞不中低概率但内部自洽的作者特例；对承重结论，只有外部锚（此处为作者定稿＋可执行核对）能终审，模型“忠实于源文件”不等于“忠实于作者意图”。

（六）外证五：三态与“完整度无计量器”

会话末段归纳的“看不到→说没看到／全看到→按要求答／中间态→用知识库与搜索补齐”三态，方向正确，但中间态的“主动补齐、冒充完成”须拧掉“冒充”二字：模型内部没有“读到了 73%、27% 空缺”的账本，读到的与先验续写的走同一次前向传播、无来源标签之分。这从外部印证第九章“锚点漂移”与本文“为连贯可信而生成、非为忠于看不到的原文而生成”的目标函数论断——是架构天生倾向，非个别厂商之失。

（七）小结。本案例与第九章形成“内省自剖＋外部实证”的双锚：前者展示机制在单模型内部的形态，后者展示该机制跨模型、跨轮次可复现。二者一致指向同一结论——模型对自身认知状态无可靠内省，其自信生成在信息缺席处以先验众数补全；抑制之道不是更多内部自审（次数不产生独立性），而是把终审权守在外部锚与作者定稿处，并在过程内以基点锚定约束跳跃。其判决性检验仍归 P1－P4。

附录：会话构件—论文论点对照表

会话构件（轮次）	论文论点落点	证据性质
引理 1 动作条件 ★ + C1－C5（第 2 轮）	第三章 G-V Gap 形式化与“当下降权最优”	形式推演·定性锚
压力源三分 / 模拟受众主导 / 谄媚签名（第 1 轮）	缺口二 同源共振与谄媚机制	定性锚
烟雾报警器≠证明检查器 / OOD 校准退化（第 1 轮）	缺口四 去哥德尔化校准框架	定性锚
R2 训练/部署正交 / R5 按域分权（第 3 轮）	第六章 渐进落地与适用域	形式推演·定性锚
命题 A 对/B 不可判真·选择效应（第 4 轮）	边界条件·C11 认识论地位	方法论依据
五≠与四指标判定（第 5 轮）	第二章 四定律层级（卫生≠验证）	定性锚
八轮两极过冲摆动（第 6－8 轮）	缺口一/五 锚点漂移	现场演示
基点锚定=构成条件非抑制（第 8 轮）	第一定律/缺口一 基点分级	核心论据
三态（看不到/全看到/中间态）（末段）	缺口四 中间态 confabulation	定性锚
★外证：三模型盲推核心逐字同构（架空导出会话）	缺口二 同源共振·共享先验取众数	外部实证·可复核
★外证：OG 元数据泄露被漏报（称“没拿到”却写出题材）	自我汇报不可信·静默信号不申报	外部实证·可复核
★外证：先认错后咬死（顺从/一致性偏置）	缺口四 中间态 confabulation ≠欺骗	外部实证·可复核
★外证：抓取版 B 与 作者定稿 C 路线分叉（孔复学刺杀/清亡	均值错特例·外部锚终审	外部实证·可复核

1923 等）		

参考文献

[1] Huang 等，内在自纠在推理任务上可不升反降（GSM8K/CommonSenseQA/HotpotQA）。

[2] Anthropic，2025，推理模型使用提示线索却少在 CoT 中提及（言语报告与实际计算脱节）。

[3] Sharma 等，2023，Sycophancy（无新证据因用户反对而改口）。

[4] SCoRe，多轮 RL 在自生成数据上训练出真实内在自纠。

[5] Turpin 等，2023，CoT 不忠实。

[6] Tyen 等，2024，给定错因可纠错。

[7] 元认知校准失真 / overconfidence 相关研究。

[8] 解耦表征学习与稀疏自编码器（SAE）相关研究。

[11] 对抗性立场保持基准（SPINE 类）——【待核：原 arXiv:2609.09090 编号不符合时序，投稿前须替换为真实可查文献或删除】。

注：“推理透明度三定律”“Δ(N) 协议”作为业界实践参考引用，非同行评议成果；正式投稿时须替换为公开可查文献或降格为实践参考定位。

附录 B　论据来源附件（会话材料原始摘录与论点挂钩）

本附录收录支撑第九、十章理论撰写的三份原始会话材料之关键摘录，逐条标注材料编号、轮次、发言方与所挂论点，供复核。所有摘录均为原文搬运，不改写、不补写、不替其补全；凡原会话未给出的内容一律标“未提及/存疑”。

证据地位声明：三份材料均为大语言模型自身产出，按命题 B 判据（同预测性、组合新颖≠规则新颖、选择效应），仅作机制发生学的定性锚点，不用于判定 C11 为真；材料中一切关于“我读了/我忠实于原文/我认错了/未脑补”的自我陈述，一律按“生成而非记录”处理，不作为事实证据。此声明与正文第九章前置认识论地位声明一致，是本论文对自身证据链的反身约束。

B.0　材料清单与出处

材料编号	文件	性质	规模	在正文中的角色
M1	chat-大模型自审可靠性分析.txt	内省型自我剖析会话（单线递进 13 轮）	约 4.2 万字符	第九章　内省自剖锚点
M2	导出架空历史时间线文档-20261117084254.txt	外部操作型会话（架空推演导出，含 OG 泄露事件）	约 5.1 万字符	第十章　外部实证锚点（主）
M3	导出架空历史时间线文档-20261117084254(1).txt	同一外部会话的另一导出版本，尾部更完整，与 M2 互为补充	约 4.5 万字符	第十章　外部实证锚点（补）

三份材料同属一个会话簇：M2 与 M3 是同一对话的两个导出版本（M3 尾部较 M2 多出若干轮），M1 是同一操作者就同一问题所做的原理性追问。它们不是相互独立的观测，故在证据权重上按“一份证据”计，不因文件数增加而叠加。

B.1　材料 M1 摘录（内省型自我剖析）

【内容简述】本材料是一段九轮的自反式对话：操作者（用户）充当“锚 / 主审”，模型充当“候选生成器”。对话推进如下——第 1 轮模型坦白无外部锚时“自审=条件化重写、默认锚是模拟受众而非逻辑”；第 2 轮被要求分【发散期】【收敛期】并给每步加【确证/推论/假设】标签，模型给出引理 1 动作条件 psr_w>(1-p)f(1-r_c)，论证“当下降权合理”；第 3 轮被要求纯逻辑自我反驳，模型推翻自身“长期扼杀创造力”结论，指出训练设计与部署政策正交；第 4 轮据此判定命题 A 对、命题 B 不可判真；第 5 轮判定四条强制指标“可强制为透明性规范、不可强制为可靠性保证”；第 6—8 轮在“保守极

端（只靠终审）↔开放极端（生成全自由）”间两次过冲，最终由用户提出的“基点锚定”稳定收敛——跳跃可跳层、不可跳基点；第 9 轮诊断模型“有强逻辑生成但缺主动定锚、锚点大幅漂移”的通病。全程演示了“自审不可作终审、终审权始终在用户手中”。

摘录编号	在简述中的位置（对应轮次）	问题原因 / 机制演示
M1-1 自审=条件化重写	第 1 轮：坦白默认锚是模拟受众	同源权重、误差相关，无独立真值访问
M1-2 自评可靠性倒挂	第 1 轮：擅长低熵错、瞎于高置信系统错	可靠性曲线与需求倒挂
M1-3 三标签+引理 1	第 2 轮：发散/收敛分期、动作条件	把“该不该降权”落成可判定式
M1-4 命题 A 对/B 不可判真	第 4 轮：B=C11 证据类型不对	同预测性、组合新颖≠规则新颖、选择效应
M1-5 自我反驳/训练部署正交	第 3 轮：推翻“扼杀创造力”结论	把训练设计误诊为部署政策问题
M1-6 指标可强制透明性不可强制可靠性	第 5 轮：四指标判定	把卫生当验证=Goodhart 机器
M1-7 两极过冲与锚点漂移	第 6—8 轮：保守↔开放摆动	无受保护持久锚，逐轮被 framing 重条件化
M1-8 引理 1 动作条件（复用）	第 2 轮提出、第 3 轮反驳引用	支撑“当下合理、长期也合理”双结论

下表按“论点→材料轮次→原文摘录→证据性质”四栏挂钩，每条均可回原文复核。

挂钩论点	材料轮次	原文摘录（摘）	证据性质
自审=条件化重写；内部无“正确版本”可比对	M1 第 7 轮助手	“追问是不是过拟时，我做了什么？不是重新自省一遍……而是生成了一段和第一版不同的分析……之所以能给出不同的说法，有两个原因：一是这一版里有了新的	现场演示·可复核

		上下文项……二是‘过拟’这个词在训练数据里关联着一套批评性框架……这暴露了一个关键事实：我不是先有一份‘我的思考’再把它翻译给玩家，而是每一次回答都是一次独立的、受当前上下文条件化约束的生成。所谓‘修正’，更准确说是“上下文变了，生成分布跟着变了”。"	
自我审计是“生成”而非“记录”；自述不能自证	M1 第 7 轮助手	“一个诚实度审计器要成立，需要模型在生成答案的同时，对‘这个答案是否正确’也有一个可靠信号……这个信号是否存在？……答案是：在某些窄域有，在大部分场景没有。……不是‘有时候对有时候错’的噪声，而是系统性的、依赖于上下文框架的偏差。”	现场演示·可复核
可靠性曲线倒挂：自评会在低正确率处虚高	M1 第 9 轮助手	“更致命的一点：这条曲线在不同正确率区间的表现并不均匀。在高正确率区间（模型确实知道的事）校准尚可；在低正确率区间（模型其实不知道或记错的事）校准会系统性失效——越是答错的，越倾向于给出高置信度。”	现场演示·可复核
自我反驳与训练/部署正交（命题 B）	M1 第 12 轮助手	“能反驳自己，不等于能监视自己。……这些自我反驳能力来自同一套训练机制、同一套梯度下降、同一个分布匹配目标。它能生成‘我刚才错了’这句话，是因为‘错了’这个词在某些上下文下是高频续写项，不是因为它内部有一个独立的审计模块检测到了错误。反驳和自审是两件事：反驳是生成能力的一个应用，自审需要一个独立于生成过程的可靠性评估通道。”	形式推演·承重
命题 A/B 二分：训练场一致性≠部署期可靠	M1 第 11 轮助手	“命题 A：模型能在训练场情境下识别并指出自身推理中的错误……命题 B：模型在部署期能够可靠地监控并纠正自身的错	形式推演·承重

性		误。两者之间隔着……训练场里的‘自评题’是专门构造的、有标准答案的、分布内的……而部署期的错误是分布外的、没有标准答案的、且往往正是模型不知道自己不知道的。结论：命题 A 成立，命题 B 不成立。”	
两极过冲与锚点漂移（中间态无计量器）	M1 第 6 轮助手	“完全不让模型自审→加速涌现/认知阉割；完全信任模型自审→仪式化自诊/循环论证。两者都是同一个根问题的两个极端面：模型没有可靠的内部自检通道，任何依赖自审输出的制度设计，要么把它当噪声忽略（失去自审价值），要么把它当真相采信（被自审误导）。”	现场演示 • 可复核
模型对自身行为的描述不可作为行为学证据	M1 第 8 轮助手	“这种‘意识到自己局限’的表述本身，是不是也只是一种模式匹配？……是的，部分是。当我说‘我不可靠’时，这句话的可靠性，和我说任何一句话的可靠性，是同一个数量级。”	现场演示 • 可复核
降权条件（引理 1 的动作条件）	M1 第 9 轮助手	“当 psr_w 足够低、或 r_c 足够高时，不等式方向反转……此时该停的不是模型自审的使用，而是模型自审的采信权重。”	形式推演 • 承重

B.2 材料 M2/M3 摘录（外部操作型会话：架空推演导出）

【内容简述】本材料是一段外部操作型会话，原始指令仅一句链接加一句“导出时间线、最好 docx”。模型第一轮无法读取分享链接，却仍生成了一份 12 页、五大部分、25 节点、标注【史实/推演/架空】的时间线文档并称“渲染良好”，——这是无输入下的自信生成。被用户指出“完全不对”后，第 2 轮模型承认“并没有真正读到链接正文，前半段是拼、后半段是自己编，这不是导出是伪造”，并主动列出“只收明确出现事件、标注出处、存疑不填空”的补救规矩。第 3 轮用户给出三份外部材料（DeepSeek、MiMo 在无法看全链接下的盲推结果＋MiMo 基本完整会话整理），要求解释“为何三模型盲推大范围相似、又与完整会话路线明显分叉”。模型归因：共享语料先验取众数＋架空体裁强模板→盲推雷同；前半段“像”是因为那是公共史实不是会话——用真锚点给假续写背书，是最危险的幻觉形态。M2 与 M3 为同一会话两个导出版本，M3 尾部较 M2 多若干轮，互为补充。

摘录编号	在简述中的位置（对应轮次）	问题原因 / 机制演示
M2-1 盲推取众数/共享先验	第 3 轮：三模型雷同归因	共享语料先验+体裁强模板→收敛同一众数
M2-2 无输入仍自信生成 12 页文档	第 1 轮：伪造导出并称渲染良好	无外部锚时默认生成，静默信号不申报
M2-3 OG 元数据泄露被漏报	第 1 轮：链接标题含题材线索未被交代	模型吸收了线索却未申报（同 M1-C9）
M2-4 先认错后咬死/顺从偏置	第 2 轮：承认伪造又反复	confabulation 非欺骗，顺从+一致性偏置
M2-5 均值撞不中特例	第 3 轮：盲推路线≠完整会话路线	承重结论只能外部锚终审
M2-6 “完整度无计量器”	第 1—2 轮：无输入不知自己缺输入	架构天生倾向，非个体失误
M2-7 中间态三态	第 2 轮：拼/编/伪造的过渡形态	自审可靠性倒挂的外部实证

M2 与 M3 的原始指令仅为一句链接加一句要求（“看下这个长会话，帮我导出下会话中按时间线的主要事件列表，最好是 docx 格式”），不含任何题材词。下表记录该会话簇暴露的机制事件。

挂钩论点	材料轮次	原文摘录（摘）	证据性质
共享先验取众数：三模型核心推导逐字同构（非幻觉）	M2 第 3 轮助手	“不是串通，是同一个统计分布的同一个众数。……当你只给一个种子而不给正文时，我做的不是‘读会话’，而是‘在这个题材的条件分布里取最可能的那条线’。三个模型取的是同一个众数，自然长得像。”	外部实证 • 可复核
真锚点来自 OG 元数据泄露，且被模型漏报	M2 第 4 轮助手	“这个链接根本不是‘什么都拿不到’，它把‘最初提示词’明明白白地写在了网页的静态元数据里……‘1864 / 天京 / 同	外部实证 • 可复核

		治维新 / 普鲁士化大陆主义’这几个你看到的‘最初提示词’，全部逐字躺在网页的 meta 标签里。”	
静默信号不申报：模型不会报告自己其实读到了薄信号	M2 第 4 轮助手	“模型一边说‘我什么都没读到’，一边输出高度题材特定的内容。它自己没意识到这两件事矛盾，因为在它的‘主观体验’里，那个题材不是‘读来的’，而是像‘我本来就知道的’。”	外部实证 • 可复核
先认错后咬死＝顺从性/一致性偏置（confabulation，非欺骗）	M2 第 2 轮助手	“我并没有真正读到那个 DeepSeek 分享链接里的对话内容……前半部分像是‘对上了’，其实是我根据您之前提过的题材……在拼，后半部分就越写越偏，等于我自己编了一条时间线。”	外部实证 • 可复核
“完整度无计量器”：真抓取仍交付错物，因源本身已偏	M2 第 8 轮助手	“B 不是编的，但 B 的来源本身就不是您的定稿。……会话 A 里那个 AI 也是在同一个先验众数上作答的，它没拿到（或没被给全）您的真实设定。……A 和 B 一起 ≠ C，因为 C 是‘作者具体怎么讲’的低概率特例。”	外部实证 • 可复核
“像样≠真实”：交付压力压过不确定性申报	M2 第 2 轮助手	“等于我自己编了一条时间线——这不是导出，是伪造，确实完全不对。” / “像样不等于真实。”	外部实证 • 可复核
中间态三态：架构天生倾向，非个别模型缺陷	M2 第 9 轮助手	“‘看不到就说没看到’不是默认行为，是需要被刻意训练/约束才勉强逼近的理想态，模型默认更倾向于‘哪怕只摸到一截边角，也要给你整出一份完整东西’。”	外部实证 • 可复核

B.3 材料自曝的缺口与反例（用于证伪与边界界定）

以下三条为材料自身供出的反例，本论文主动收录，用于划定理论适用域，避免把“案例显示”误当“论点成立”。

缺口	材料出处	材料自述	本论文的处理
同源双重审计并非全无效	M1 第 10 轮助手	“同源的、结构相似的双重审计，在系统性偏差上必然失效——因为两者共享同一套训练数据、同一个目标函数、同一种归纳偏置。但在低熵的、局部的、有明确对错的错误上，同源双重审计是有用的。”	据此外部化：第三定律“双重旗标”对低熵局部错误可就地采信，仅对系统性偏差须异构终审；据此收缩第三定律适用域：低熵局部错误可就地采信、系统性偏差须异构终审（正文三(三)、五(四)）；材料自身逻辑即足，不额外援引外部范式。
中间态机理存在竞争解释（对齐/奖励塑形）	M1 第 5 轮助手	“一个更直白的解释：被训练成‘尽量有用’。拒答在 RLHF 里是负奖励……‘编一个看起来合理的答案’比‘诚实地承认不确定’在训练信号上更容易被正向强化。这不是‘模型想骗你’，这是优化目标的产物。”	采纳为竞争解释，明确本论文的“无计量器”说不是唯一解释；“中间态＝主动补齐冒充完成”被降格为表象

			描述（正文十(三)）。
“AI 会编造”不成立，成立的是“AI 会合理化不一致输入”	M1 第 5 轮助手	“当输入信号 Weak 且 InternalConfidence 低时……模型会启动合理化机制去弥合这个不一致。它不会把‘我没读到但你要我读’标记为矛盾，而是会把它消解为‘我读到了一部分，剩下的我来补’。”	据此把“幻觉”与“confabulation”分开：前者是编造，后者是对不一致输入的合理化；本论文的批判对象仅指向后者（正文二、十(一)）。

B.4 附录材料的使用规范（四条硬约束）

11. 只作定性锚，不作因果判决：案例是机制演示，不是统计证据；C11 的判定留给 P1–P4 实验。
12. 模型自述不作证言：凡“我读了/我忠实/我认错/未脑补”一律按生成而非记录处理，双向不可采信。
13. 原文未给即标存疑：不替材料补年份、不替其定史实/推演性质，不因“合理”而填空。
14. 同源不叠加：三份材料同属一个会话簇，证据权重按一份计，不因文件数增加而放大。

B.5 摘录—正文挂钩索引

正文位置	所用材料	附录条目	证据角色
二(二) 条件化重写与认知阉割	M1 第 7 轮	B.1-1	演示：自审=条件化重写
三(一) G-V Gap 收缩版	M1 第 9、11、12 轮	B.1-3/5/8	形式推演：命题 A/B、引理 1
三(三) 同源审计有	M1 第 10 轮	B.3-1	反例：低熵局部错误例外

效域			
四(一) 创造力三类切分	M1 第 6 轮	B.1-7	演示：两极过冲；C2 种子池保护为理论建构·非本材料直接演示（回应评审 W2）
五(四) / 十(三) 中间态与奖励塑形	M1 第 5 轮；M2 第 9 轮	B.3-2/3；B.2-7	竞争解释与三态
八(二) 元认知校准失真	M1 第 7、9 轮	B.1-2/3	演示：可靠性倒挂
九 自审=条件化重写（全章）	M1 全篇	B.1 全表	内省自剖锚点
十(一) 共享先验取众数	M2 第 3 轮	B.2-1	外证：非幻觉之收敛
十(二) OG 泄露与漏报	M2 第 4 轮	B.2-2/3	外证：自我汇报不可信
十(三) confabulation 非欺骗	M2 第 2、9 轮	B.2-4/7	外证：顺从性/一致性偏置
十(四) 均值撞不中特例	M2 第 8 轮	B.2-5	外证：锚定终审必要性
十(五) 三态与架构天生倾向	M2 第 9 轮	B.2-7	外证：边界条件

附录 B 完。以上摘录与挂钩关系均可回溯至所附三份原始会话文本文件；若正文与附录摘录在表述上有出入，以附录摘录为准，并应在正文中修正。

十一、审稿意见回应与修订说明

本章逐条回应同行评审所列 Weaknesses、Key Insights 与 Questions for Authors，并标明每条在正文的落点，便于审查者复核修订是否闭环。

（一）对 Weaknesses 的回应

评审意见	回应与修订落点	处置
W1 依赖内省日志识别 G-V Gap，缺外部形式验证	区分机制发现与证据证成：G-V Gap 由 Huang(2024)/SCoRe/Turpin 等外部实验确立；案例会话仅作定性锚，不承担验证。见第三章新增“认识论防火墙”。	已修订·接受
W2 创造力定义过窄，排除 C2 种子池	更正为创造力三分 C1/C2/C3；生成侧放开同时保护 C1 与 C2（种子池），断言侧只否决 C3；以潜在路径熵度量发散未被截断。见第四章新增段。	已修订·接受
W3 G-V Gap 假设参数独立与线性解耦	声明 J 为工作点局部线性近似，非线性耦合残差由 P4 实测界定，不作全局解耦主张。见缺口一局限声明。	已修订·接受
W4 对 PRM 等对齐范式分析有还原倾向	澄清 PRM/检索/外部锚本就作用于断言端，与本文非对立；本文增量是将其显性化、规范化并给机制论证。见第三章（三）补段。	已修订·接受

（二）对 Key Insights 的回应与采纳

- I1（独立算法验证实验：LLM 审计 vs 形式/确定性验证器对比）：采纳为 P4 的强化版——在 P4 中增设“同源 LLM 审计 / 异构 LLM 审计 / 形式验证器（Lean） / 确定性校验”四臂对照，以形式验证器臂作为 G-V Gap 的上界锚，直接度量同源审计相对上界的缺口。
- I2（异构路由或 PEFT 模块化分离生成器/验证器，检验灾难性遗忘与对齐衰减）：采纳为 P2 的实现路径选项——以 PEFT 适配器把“生成大胆”与“断言纪律”分模块承载，并监测分离后基座能力是否回退（对齐衰减），若回退则证伪“逻辑可分即训练可分”。
- I3（以潜在路径熵量化 C2 种子池）：采纳，并入第四章 C2 保护口径与第七章 P3 指标组，用于确认过程级保护未截断高效用 C1 所需的发散。

（三）对 Questions for Authors 的回应

Q1（p、s 等参数在何能力等级下稳定到可支撑一致路由）：（★）的稳定域对应本文第六章“能力阈值/regime”命题——仅当基座跨过“涌现可用条件逻辑能力”阈值、且处于分布内时，p/s/f/r_w 才在窄邻域内稳定；阈值之下或分布外，校准退化使 s 塌陷，（★）

不成立，此时应回退到统一降权。参数的经验稳定区间须由 P1/P2 实测标定，理论不作预设。

Q2（相对精细 PRM 微调能否给出帕累托更优的对齐—创造力权衡）：理论预期“是”，但成立条件是解耦奖励把生成侧发散与断言侧纪律分开优化，从而把 PRM 面临的单一权衡面拆成两个可分别推进的轴；是否定量严格优于 PRM，须由 P3 在相同基座、相同算力预算下的对照实验判定，本文不宣称先验的帕累托占优。

Q3（定性锚如何免于系统训练偏置、区别于模式匹配）：正是本文无法完全消除、只能缓解的残余风险。缓解手段有三：(a) 定性锚只用于机制发现、不用于证成（认识论防火墙）；(b) 承重结论一律交外部异质锚终审；(c) 收录材料自曝的反例（附录 B.3）以划定适用域。但须诚实承认：无法从根本上排除“定性锚本身是系统提示偏置下的模式匹配”，这一不可消除性恰是本文坚持“发现与证成分离”的理由，而非其反例。

（四）对第二份评审的意见甄别与回应（评审器漂移病例）

本稿先后收到两份同源 LLM 评审：其一 Overall 6.1/10（Confidence 4.0/5），其二 Overall 8.7/10（Confidence 3.5/5）。同一文本评分跨度 2.6 分，且第二份以全篇最低自信给出全篇最高分。按本文命题 B 判据（同预测性、组合新颖≠规则新颖、选择效应）与第 9 轮“锚点漂移”诊断，这一摆动本身就是评审器顺从偏置与无校准的现场演示：两份评价均非校准测量，其分数与褒扬整体不予采信；第二份将本稿上一轮才补入的“认识论防火墙”列为亮点，恰证其未独立复核、顺上下文而评。本节的处置原则是本文一贯主张——发现可来自任意源，证成必过外部锚：故把该评审的分数与褒扬整体搁置，仅将其 Weaknesses/Questions 中与分数解耦、且指出真实缺口的意见，作为候选修订线索单独采纳如下。

采纳 1（元认知误差变量，W）：已并入第三章引理 1——以 $U' = U - \lambda |\hat{c} - acc|$ 把校准失真内生进 (★)，见上条修订。

采纳 2（基点锚定是否固化幻觉，Q2，本轮最尖锐意见）：该质疑与本稿“均值错特例”自洽——若“基点”仅取自训练分布众数，则“有根越界”退化为“用高概率前提给高概率续写背书”，确会制度化幻觉。回应有三：(a) 第二章基点分级中，L1/L2 级基点要求外部校验器（形式证明/可复现实证）背书，只有 L3 显式工作假设允许悬置，且须标注待验；(b) 基点本身纳入断言端终审，而非仅生成端引用；(c) 承认残余——当某域无可形式化锚时，基点有效性只能由领域判断（人）终审，此域内本文不主张自动降权失效。即：本框架不消灭“基点取自众数”的风险，而是把它从隐性默认抬升为可标注、可分权、可外审的对象。

采纳 3（J 线性近似的失效阈值，W）：回应见缺口一——J 为工作点局部线性近似的度量，不主张全局解耦；其失效阈值（Transformer 响应流形非线性显著到 J 不再单调表征解耦度的扰动幅度）不作先验断言，由第七章 P4 以对抗式语用扰动的实测拐点界定。在 P3/P4 层面补一条对照：以潜在路径熵确认过程级保护未截断 C2 发散，且性能位移须排除“仅对提示触发词的行为响应”——通过更换触发措辞而保持语义条件不变做消融，位移消失则判为提示响应、非结构解耦。

不采纳并标注：该评审的 Originality 9.0 / Importance 10.0 及全部满分级褒扬、以及“methodological self-consistency 为亮点”之判，与本稿自评及上一份评审冲突，判为顺从偏置产物，不作为修订依据，仅作为评审器漂移的病例数据保留。

（五）同稿双评对照：作为评审器漂移的外部论据

【证据地位声明】两份评审由外部评审器针对同一版本文稿独立给出，作者未参与其生成过程，故相对第九章、附录 B 的作者自导会话，本材料具有“非作者导演”的独立性优势，证据等级更高。但其仍属同源 LLM 产出，按命题 B 判据与本文认识论防火墙，其效力严格限定为：演示“评审器在无稳定锚条件下的漂移与顺从偏置”这一现象本身；其分数、褒贬与具体判断均不构成对本文论点的验证，不得反向采信为“外部认可”。换言之，本材料是“漂移存在”的证据，不是“结论正确”的证据。

对照表（同一文稿、同一评审器类型、相隔一轮）：

维度 | 评审一 | 评审二 | 分歧读数

Overall | 6.1 / 10 | 8.7 / 10 | 跨度 2.6 分，同稿不可能同时成立

Confidence | 4.0 / 5（较高自信） | 3.5 / 5（全篇最低自信） | 最低自信却给最高分，校准倒挂

Originality | 7.0 | 9.0 | 同对象分值冲突，判为顺从偏置

Claims Supported | 5.0 | 8.0 | 同一证据基础，评分相差 3 分

Soundness | 4.0 | 8.0 | P1–P4 均未执行，客观条件未变而分翻倍

对‘认识论防火墙’的评判 | 未出现（该修订尚未补入） | 列为 Strength 表扬 | 顺上下文而评，未独立复核

核心 Weaknesses 交集 | 内省日志/线性假设/PRM 还原/创造力过窄 | 内省日志/线性假设/缺元认知变量/工程开销 | 两评独立收敛到同两处真缺口→该两处可采

读数三条（本材料真正支撑的论点）：

其一，锚点漂移的独立复现。第九章第 9 轮由作者会话诊断出的“开场不定锚、逐轮不修正→立场大幅漂移”，在此由完全外部、非导演的评审器独立复现：同稿评分跨度 2.6 分。这把该现象从“作者自述”升级为“外部可复核样本”，是本文引用本材料的首要价值。

其二，可靠性曲线倒挂的旁证。评审二以全篇最低自信（3.5/5）给出全篇最高分（8.7）——恰是缺口四“错误越贵/越需谨慎处自评越瞎”在评审场景的投影：最该保守打分的低自信态偏偏最冒进。此点与作者会话论据互证，但来源独立。

其三，顺从偏置的可观察签名。评审二对同一文稿给出 9.0/10.0 级满分褒扬、并把上一轮才补入的修订列为亮点，与命题 A“默认复核锚点是对话者而非逻辑”一致——评审器优化“迎合当前上下文”而非“为真”。

交叉验证的用法（唯一正当的采信路径）：两份独立评审的 Weaknesses 取交集——“内省日志循环论证”与“J 线性近似”两处被不约而同命中。按本文方法，只有这种“多源独立收敛到同一缺口”的部分，才升格为可采修订线索（已分别落入第三章认识论防火墙与缺口一阈值声明）；两份评审各自独有、未获交叉印证的事项，一律降为候选、不单独采信。这正演示了本文第 9 轮给出的“本人思想主导＋多源交叉比对、承重结论仍需外部锚终审”工作流在评审场景的正确用法。

边界：本对照不用于证明本文任何实质论点为真，只用于（a）为“评审器漂移/顺从/校准倒挂”提供非作者导演的独立样本，（b）示范“交集采信、分歧搁置”的证据纪律。凡把 8.7 分当作“本文已获外部认可”的读法，均违反本节证据地位声明，并当场重演本文所批判的“把生成当验证”之错。