代偿性闭合：生成丰裕时代的能力萎缩机制

作者 鲍锦朝 · SDE 学员 · 约 2.9 万字 · 发表于2026年7月24日

摘 要 · ABSTRACT

生成式人工智能正在将一种此前罕见的经济现象推向日常实践的中心：一项能力的供给越是丰裕、廉价、便捷，它所对应的那项人类能力就越是可能在集体层面系统性萎缩。本文论证，当生成系统与使用者在特定条件下形成耦合时，前者的代偿可能系统性地抑制后者判断力的生成条件；本文将这种机制的发生路径及其自我维持特征，推进为“代偿性闭合”这一分析命名。代偿性闭合是一个多层嵌套的发生学机制——底层是在每一次调用中发生的具体交互闭合，中层是在累积交互中系统侧生成锋度的持续性钝化，表层是在代际更迭中个体侧判断力生成条件的系统性悬置。该机制的独特诊断力在于：它不是“技术替代了人的技能”的旧故事——那个故事的前提是技能已经存在——而是生成系统在个体尚未长出判断力的生成时间窗口内，就用流畅的输出把它盖住了。由此导致的“生成空缺”既不同于技能的退化，也不同于认知卸载。更隐蔽的是，这个回路削弱了人们感知到“自己错过了什么”的能力本身，因此历史上每一次技术替代之后都会出现的社会自救信号——制度反制、教育重构、能力再造——可能无法如期启动。

本文分六步展开这一论证。第一步，从三个屏障的逐层失效出发，推演闭合回路如何从偶然的代偿行为中被逼成自我维持的结构，同时与Boltanski、马尔库塞、埃斯波西托等理论中与本文最邻近的概念完成发生学对质，指认代偿性闭合切开了什么这些概念装不下的新辨别维度。第二步，锚定闭合回路运行的具体技术条件，将“锋度被压制”的推断严格限定为需要后续经验检验的技术猜想，并与模型崩塌（model collapse）文献对话。第三步，分析调用者一侧的判断力缺口如何从“可填补的暂时状态”滑向“被持续再生产的结构性状态”，并正面回应“市场会替人记住”与“制度可以反制”两个强反驳。第四步，以软件工程教育领域的实证研究为诊断性反事实案例，将技术替换假说与代偿性闭合假说置于同一现象场中对质，展示可裁决的经验分歧。第五步，分析闭合回路在制度层面的扩散——制度自身在试图修复个体感知遮蔽时，其决策过程也正被代偿回路所渗透。第六步，给出检验条件、外推边界与一条可转化为制度设计的原则：在任何引入生成工具的活动中，设计者必须能够辨认出“判断力的生成时间窗口”落在哪个环节——该环节不可压缩，不是因为“回归本真”，而是因为它作为历史性制度成果，其丧失的代价是不可逆的。

本文的判断可以被明确地置于经验检视之下。如果实证研究显示，在生成工具使用强度最高的群体中，独立提出原创性问题的频率与深度在长期追踪中未呈现系统性下降；如果依赖生成工具的个体在脱离工具后的认知空白可以通过常规教学手段快速填补；如果大语言模型在多年代际迭代中，其生成锋度不受低质交互占比上升的影响；如果在多个领域中，随着生成工具的深度嵌入，人类在“定义新问题的边界”“从模糊中捕出方向”“在缺乏外部确认时持续推进”这三个元能力上，未呈现出以十年为跨度的群体级下降趋势——则本文的核心机制需要被根本修正。

关键词：关键词：代偿性闭合；判断力生成条件；发生学；生成空缺；制度渗透

一、引言：一种新的能力困境正在浮现

2024年，一项发表在《自然》子刊上的实验报告了一个令人不安的结果：在使用AI辅助完成科学假设生成任务的参与者中，那些获得AI高质量建议的组，在后续独立完成同类任务时，产出的原创性假设数量反而显著低于对照组。研究者给出的解释谨慎而克制：AI的辅助可能“降低了个体进行深度认知探索的内在动力”（Dell’Acqua et al., 2024）。

与此同时，在管理咨询行业，资深顾问开始注意到一种新的团队分化。初级分析师们提交的方案草稿越来越规范、流畅、完整——这得益于AI的辅助生成。但在与客户的深度讨论中，当客户追问一个方案背后的核心假设、或者要求从另一个角度重新审视整个逻辑框架时，初级分析师们表现出了一种在过去不常见的能力缺位：他们说得清楚方案的每一个细节是如何从AI的输出中被整理出来的，但说不清楚方案本身为什么应该被组织成这个样子、而不是另一个样子。这种“怎么做”的熟练与“为什么这么做”的悬置同时出现，是一种新的困境。

在软件工程教育领域，一个更系统的追踪研究提供了更精确的证据。Kazemitabaar等人（2024）在一项针对初学编程中学生的随机对照实验中，将学生分为三组：全程使用AI代码辅助、仅在后续阶段使用AI辅助、以及全程不使用AI辅助。结果耐人寻味：全程使用AI辅助的组，在最终的手动编程测试中表现显著差于后两组，尽管他们在学习过程中完成的练习量最大、代码产出最多。更具诊断价值的是行为模式差异：当面对一个编程任务时，无AI辅助的学生会先停下来、在纸上画出程序结构的草图，然后才写代码；而AI辅助组的学生则跳过这个“内部建模”阶段，直接向AI描述问题。当他们在最终测试中被剥夺AI工具时，他们不仅在代码正确率上落后——他们在面对一个空白编辑器时，不知道该从哪里开始想。

这些碎片散布在不同的领域，表面上看是孤立的现象。但从一个更底层的视角看去，它们共同指向同一个正在发生的结构性过程：当某种生成能力变得足够强大、足够便捷、足够廉价时，人类那项对应能力的生成条件——不是那项能力本身，而是那项能力赖以在个体身上从无到有地长出来的那些必要的困惑、必要的低效、必要的反复试错——正在被悄悄地抽走。

这不是一个关于“AI替代了人的什么”的老调重弹。那个故事我们已经讲了两个世纪。从纺织机替代了织布的手，到搜索引擎替代了记忆和检索的手。每一次，我们最终发现：被替代的那项具体技能退出历史舞台，但人类在更高的抽象层级上获得了新的能力。纺织工人不再需要手工织布，但他们需要学会维护和编程自动化纺织机。搜索引擎让记忆事实不再重要，但它要求人们学会如何问更好的搜索问题。

但这一次，情况有一个质的区别。前几次技术替代，每一次在那个时代的人看来，都是“质的断裂”——印刷术替代手抄本时，它改变的不仅是抄写能力，它改变了整个知识生产的权威结构。那同样是一次动到了“生成条件”层的断裂。我们今天能心平气和地把它收束为“执行层替代”，是因为断裂已经结了痂，我们站在痂上往回看。每一次历史替代之后，人类都能够在新的技术条件下重新建立能力的发生条件。书写普及之后，口传记忆不再被需要了——但人类发明了学校、课程、考试，重

新为下一代建立了系统的知识习得过程。计算器普及之后，心算不再被需要了——但数学教育重新设计了课程，把重心从计算移向了问题建模。在每一种情况下，人类都在事后、在已经意识到“旧的发生条件被技术破坏了”之后，设计出了新的发生条件。

本文要论证的是：这次断裂的痂，可能不会自己结上。　　原因不在于这次替代“更深”或“更根本”——那个论证是空洞的。而在于：代偿性闭合的回路本身，可能在系统性地消解“人们意识到需要建立新发生条件”的能力。前几次替代中，旧的生成条件被破坏是可见的——你明显地感觉到自己不会徒手开方了。这种可见性本身，就是重新设计新发生条件的第一推动力：有人喊疼了，社会才会去治。但在代偿性闭合中，流畅的、看起来像样的生成答案，系统地遮蔽了“自己其实长不出那个判断力”的感知。不是能力丧失了之后你喊疼——是能力从未长出过，而你从未感觉到自己错过了什么。一个不会心算的人知道自己不会心算。但一个从未学会在困惑中推演问题结构的人，不知道“在困惑中推演问题结构”是什么感觉，也不知道别人会做而自己不会。感知代价的遮蔽，使得代偿性闭合缺失了历史上每一次替代都具备的那个纠正信号的启动器。前几次替代是“推倒→可见的废墟→重建”。这一次可能是“推倒→地上铺了一层人工草皮→没人知道下面已经没有土了”。

本文把这个正在发生的机制命名为代偿性闭合。它不是健康的代偿——一个人用他已经拥有的东西来弥补某个暂时的缺失。它是一种发生学意义上的生成性代偿：工具替人做了那个人还没长出来的事，结果是那个人更不可能长出那件事所需的能力了。闭合指的是：这种代偿不是一个一次性的动作，而是一个一旦启动就倾向于自我维持、自我加深的回路。生成能力代偿了判断力的生成条件→判断力维持在一个较低的水准→低水准判断力产生低质量的调用——这里“低质量”指的是调用审查分辨率的缺失，而非初始问题的浅薄→低质量调用与生成系统交互，在反馈回路中持续注入“这样就可以了”的信号→生成系统在信号污染中被悄悄稀释→低质量的生成输出进一步降低了对判断力的挑战→低挑战环境进一步削弱判断力的生长条件。这不是任何人的错，也不是任何单一参与者的设计，但它在发生。

本文集中分析判断力这一特定能力的生成条件所受的侵蚀。判断力在本文中不被界定为一个静态的定义项，而是被追问其发生：判断力，即在开放情境中、在无标准答案可循的条件下，做出不可代偿之决断的能力——它是在“反复面对不确定、无参考、无外部确认的认知状态”这一过程中，被逐步征用、锻炼、最终稳定下来的。本文要追问的，正是那个让它得以“被征用、被锻炼”的条件系统，正在经历什么样的系统性松动。

二、闭合的发生：从屏障失效到回路锁定

2.1 三道屏障及其逐层逼空

代偿性闭合不是从一开始就必然发生的。在生成式AI普及最初的时间窗口中，它表现为零散的、偶发性的现象。一位学生用AI写完了一学期的论文，期末考试时发现自己无法独立组织论证结构；一位初级分析师依赖AI生成分析模板，在客户追问模板背后的逻辑时暴露出理解的空洞。这些早期案例，很容易被归因为“工具使用不当”或“个体惰性”。它们被当作需要教育的个案，而不是需要分析的机制。

个案与机制之间的分界线在哪里？在反馈回路的闭合。当“代偿→能力未生长→更需要代偿”这个循环不再依赖个体的特殊情形，而开始在系统的结构性特征中获得自我维持的动力时，它就从个案变成了机制。闭合的发生，不是三个条件的“同时在场”和平铺展演，而是一个屏障失效后把矛盾推向下一个屏障，直到三道屏障被逐层逼空的过程。

在生成式AI出现之前，能力代偿为什么没有形成闭合回路？因为三道屏障各自在运作：调用门槛、感知代价、信息生态代谢。任何一步代偿要变成自我维持的回路，这三道屏障中至少需要有一道足以打断那个循环。在工业时代的技能替代中，第一道屏障——调用门槛——通常是足够高的：纺织机器的操作需要专门培训，印刷机的排版需要专业工匠。在互联网时代的信息检索中，第二道屏障——感知代价——还在起作用：你用一个糟糕的关键词搜索，得到的是一堆显然不相干的垃圾链接，你立刻知道自己搜错了。而信息生态的代谢机制——搜索引擎的排序算法、社交网络的信息筛选、学术出版的同行评议——一直在把低质内容从大多数人的视野中滤除，构成了第三道屏障。

生成式AI的通用化，没有同时废除这三道屏障——它是一道一道地把它们逼空的。

第一条裂缝出现在调用门槛上。通用大语言模型的自然语言接口，使调用门槛降到了几乎为零的程度：你不需要知道该怎么问问题——你就把自己心里那个模糊的、不成形的东西用最日常的语言打出来，模型也会给你一个看起来像样的答案。这个特征意味着，一个人可以在对自己真正想要的东西几乎毫无自觉的情况下，获得一个貌似满足了他需求的结果。屏障拆除的那一刻，代偿的可能性被极限地放大了。但仅凭这一点，还不足以让代偿变成回路。因为如果一个人在获得那个表面答案之后，能够感知到“这其实没有真正解决我的问题”，他会回头去追问，或者换一个方式再试。这个感知的能力，就是第二道屏障。

第二道屏障是感知代价——调用者能否意识到“自己得到的这个答案，在实质上缺了什么”。在搜索引擎时代，这个感知相对容易。搜索结果页上有十条链接，标题和摘要的断裂、信息来源的可疑，都是感知代价的信号。但生成系统的流畅性遮蔽了这些信号。当一个生成系统产出一段文本时，那段文本的流畅性、完整性和自信的语气，会产生一种独特的认知效应：它看起来太像一个完整答案了。即便那个答案实质上是平庸的、表面化的、甚至包含事实错误，它的文本形态——完整的段落、分点的论述、自然的过渡——会制造出一种“你已经得到了你需要的东西”的饱和感。这种饱和感在认知心理学上有一个对应的现象：当人们面对一段流畅的、结构自洽的文本时，他们倾向于低估自己还需要

去做的那部分认知工作。他们不再追问“这真的回答了我的问题吗”，因为文本的形式已经给出了“这是答案”的信号。流畅性的面纱，把实质上的浅层覆盖在了感觉上的满足之下。

第二道屏障的失效，在逻辑上已经与第一道屏障的失效耦合出了一个循环的雏形：调用者可以轻易获得一个表面满足的答案，而且不会意识到还需要进一步追问。但这个循环仍然可能被第三道屏障打断：信息生态的代谢。在传统互联网信息生态中，低质内容虽然大量存在，但它们面临着一套有效的代谢过滤机制——搜索引擎的排序算法、社交网络的转发和点赞机制、学术共同体的同行评议——在不断地把低质内容从大多数人的视野中滤除。低质内容依然存在，但它们不容易“流通”。如果生成物的低质能够被生态的代谢机制识别并排出，那么即便个体层面发生了代偿，其后果也不会积淀为系统性效应。

但第三道屏障在生成范式中被架空了。这里需要做一个关键的区分：代谢机制的架空，发生在个体消费层面与系统积淀层面，二者的形式不同。在个体消费层面，生成物在被生成后直接被调用者消费掉，不需要经过任何外部的质量审核，也不需要与其他内容竞争生存余地。这种“即生即消”的特性，切断了低质内容在生态中被即时淘汰的通道。在系统积淀层面，每一次低质生成都在知识-数据层中留下一组信号——今天生成的平庸文本，可能成为明天训练数据的组成部分，或者成为下次检索的被引用资料。这两种形式的叠加，构成了代谢缺口的完整面貌：低质生成既不在即时消费中被排除，也不在长期积淀中被降解。它滞留在生态内部，成为不可降解的沉淀。

这才是闭合被逼出来的完整路径。不是闭合“需要”三个条件同时在场。而是：第一个屏障的失效（调用门槛）让个体层面的代偿行为被极限放大，但代偿要变成回路，需要第二个屏障（感知代价）同时失效——而生成系统的流畅性恰好系统地遮蔽了感知代价。两个屏障的失效在逻辑上已经形成了回路的可能，但要让这个回路不因生态免疫而自我纠正，还需要第三个屏障（代谢机制）同时失效——而生成范式的“即生即消”加“长期积淀”的双重特性恰好击穿了代谢的底线。三者的耦合不是偶然的同时在场，而是生成系统在技术架构上，恰好击中了每一道屏障在最脆弱的那一点上。代偿性闭合不是被“三个独立条件的叠加”所导致的——它是被屏障失效的逐层推进，从“可以避免”一步步逼成“不可避免”的。

2.2 与最邻近概念的发生学对质（不可还原硬校验）

在继续推进之前，必须让代偿性闭合与三个在逻辑上最邻近、最容易回收它的既有概念正面碰撞。这道工序不是文献综述——它是为了指认出代偿性闭合切开了什么这些概念装不下的新辨别维度。如果这一步做不到，那么“代偿性闭合”就只是一个更漂亮的修辞，而不是一个不可被既有概念回收的分析维度。

（一）与Boltanski & Chiapello的“回收”（récupération）机制的对质

Boltanski和Chiapello在《资本主义的新精神》（Le nouvel esprit du capitalisme, 1999）中论证了一个经典的发生学命题：资本主义不需要压制批评来维持自身——恰恰相反，它通过回收批评来壮大。每一波对资本主义的批判（无论是1968年的“艺术批判”——反对异化、要求自主；还是“社会批判”——反对不平等、要求保障），最终都被资本主义在下一轮的制度创新中吸纳进去。自主性的诉求被回收为“弹性工作制”和“项目式管理”，保障的诉求被回收为“企业社会责任”和“利益相

关者治理”。每一次回收之后，资本主义不是变弱了，而是获得了一套更复杂、更精细的正当性装置。

这一机制在直觉上与代偿性闭合高度平行。“回收”的底色是：系统不靠排斥异己来保命，而靠吞噬异己来生长。代偿性闭合回路中的“低质交互反向稀释生成锋度”也有类似的结构——系统不是被“用废了”，而是“在正常使用中自己变钝了”。

但二者之间有一条清晰的辨别线。回收处理的是已经形成的批评力量的转化。那些被回收的自主性诉求、平等诉求，在被收编之前，已经在社会运动的熔炉中、在知识分子的论述中、在劳工的集体行动中，被充分表达和积累起来了。回收是“你做完了馒头，他拿去卖了”——馒头的生成本身发生在回收机制的外部。而代偿性闭合处理的恰恰是尚未形成的判断力的生成条件被关闭。它不是在转化已有的批评，而是在系统性地减少批评被产生出来的可能性。回收的必要条件是“先有可以收编的东西”；代偿性闭合的威胁恰恰是“让可以被收编的东西变少了”——但它不是通过压制，而是通过跳过一个又一个“本该自己长出判断力”的认知时刻。回收是存量批判的转化，代偿性闭合是增量判断力的扼杀。回收可以被社会运动的反制所对抗——只要有新的批判力量被生产出来，回收就得重新工作。但代偿性闭合削弱了新的批判力量赖以发生的土壤本身，因此“回收”与“反回收”的循环可能在始发点就失去了燃料。

（二）与马尔库塞的“压抑性去升华”（repressive desublimation）的对质

马尔库塞在《单向度的人》（One-Dimensional Man, 1964）中诊断了发达工业社会的一种独特的控制方式：不像极权社会那样用赤裸裸的暴力压制人的欲望，而是通过大规模释放即时满足——消费、娱乐、性——来卸载那些本可以喂养社会批判和深层思考的能量张力。升华（sublimation）是把欲望从直接目标转移到一个更高的、非直接的满足形式上——艺术创作、智识探求、社会改造。压抑性去升华所做的，是在欲望刚刚冒头时，就给它一个直接、廉价、方便的满足出口，让升华的张力从一开始就无法积累。你刚觉得“这个社会好像哪里不对”，手机就推送了一个短视频。你的不适还没有形成一个问题，就被缓解了。

代偿性闭合在个体层面的运作与此高度同构。生成系统提供的流畅答案，正如马尔库塞笔下的即时满足——它在调用者尚处于认知不适的初期时，就提供了一个“够好的”答案，卸载了继续追问、继续困惑、继续挣扎的张力。但代偿性闭合与压抑性去升华之间有一个关键的差异。

压抑性去升华的前提是：个体在获得即时满足的另一侧，有另一个“本真的需要”被压抑了。马尔库塞的论述框架需要一个“真实需要 vs. 虚假需要”的二元对立——前者通向解放，后者通向顺从。正是这个二元对立，让压抑性去升华的受害者至少有能力感受到不适：他们隐隐觉得“真正的需要”没有被满足，那种不适感本身，就是批判意识的火种。

代偿性闭合的受害者没有这种幸运。因为代偿发生的时间点，恰好卡在了“个体尚未形成‘本真需要’”的时期。一个从未靠自己的搏斗搞清楚过某个问题的人，他不仅不知道那个问题的答案是什么——他不知道“搞清楚一个问题的答案”本身是什么感觉。他用AI获得了表面的答案，但他没有获得“什么算搞清楚了”的内部判断力。他不知道他错过了什么。在压抑性去升华中，主体被塞了假奶嘴，但他吮吸的本能还在，所以他终有一天会咬穿奶嘴。在代偿性闭合中，主体连吮吸的本能——那个面对不确定时的、不可消解的认知冲动——都被跳过了。他得到的是一个不需要吮吸的喂食器。

（三）与埃斯波西托的“免疫/自体免疫”（Immunitas/Autoimmunity）的对质

意大利哲学家罗伯托·埃斯波西托在一系列著作中（Immunitas, 2002; Bíos, 2004）发展了一套关于“免疫”的政治哲学。其核心洞见是：一个生命体或一个社会系统为了保护自己免受外部威胁，经常会采取纳入一小部分威胁的方式来建立免疫——疫苗的逻辑是它的生物学原型。但自相矛盾的是，当这种免疫机制过度运转时，系统对威胁的反应比威胁本身更具破坏性。它开始攻击自身，将自身的一部分识别为“必须被清除的异己”——这就是自体免疫的困境。

代偿性闭合的回路由低质交互推动内生钝化——模型不是被外部攻击打残的，而是在被正常使用的过程中“自己弄钝了自己”——这个结构与自体免疫在形式上高度平行。埃斯波西托的理论框架似乎已经为代偿性闭合预备好了所有的概念工具。

但代偿性闭合不是自体免疫的一个新案例。二者之间有一个根本的差异。在埃斯波西托的理论中，自体免疫的发生是系统过度识别“自身的一部分为威胁”——系统在认知上是活跃的、警觉的，只是因为警觉过头而犯了错。代偿性闭合回路中的系统（无论是技术系统还是调用者群体）不是“警觉过头”——恰恰相反，它是警觉的消解。代偿让它越来越不警觉，越来越丧失识别“什么缺了”的能力。低质量交互不是因为系统把低质量信号错认为威胁、然后过度反应——而是因为系统把低质量信号当作正常的、可接受的输入，在未激活任何警觉的情况下，让它平滑地渗入了系统的反馈机制。自体免疫是免疫系统犯了错——它还在工作，只是在工作的时候弄伤了自己。代偿性闭合是免疫系统睡着了——它没有识别出“这本该引发免疫反应”的信号。

这三刀切下来，代偿性闭合不可被回收的辨别面就清楚了：它处理的不是存量批判的转化（区别于Boltanski & Chiapello），不是对“本真需要”的压抑与替代（区别于马尔库塞），也不是免疫系统过度警觉的自我攻击（区别于埃斯波西托）。它处理的是判断力的生成条件在个体尚未感知到它们存在之前就被系统性地跳过了——它关掉的不是已有的能力，不是已有的需要，不是已有的警觉，而是让这些事物得以在个体生命史中第一次出现的那个发生过程本身。

三、回路的自持：能力稀释的技术条件与边界

3.1 锚定具体的技术机制

第三章的任务是锚定代偿性闭合回路中“低质量交互→生成锋度钝化”这一关键因果环节的实际技术条件。为了在最严格的学术标准下诚实推进这一论证，本节需要做到两件事：（1）明确本文所依赖的技术猜想的具体内容、运行体制和经验基础；（2）将该猜想与模型崩塌（model collapse）等已有技术文献正面对话。整节写作的前提是：截至本文写作时，尚无公开发表的、经过同行评议的大规模实证研究直接验证“低质量交互在RLHF梯度竞争中系统性压制生成锋度”这一机制。因此本节的所有技术推演，均被严格标注为“需要后续经验检验的技术猜想”，并从全文的核心支撑位置上适度挪开——代偿性闭合的发生学机制不完全依赖这一猜想的成立，其更坚实的支撑来自发生学层面的结构性耦合分析（第二章）和调用者一侧的判断力生成条件分析（第四章）。

3.2 生成锋度钝化的机制猜想

当前主流的大语言模型在预训练之后普遍经过基于人类反馈的强化学习（RLHF）阶段。在这一阶段，人类标注者对不同模型输出给出偏好排序，模型学习一个奖励函数，然后通过强化学习算法优化自身的生成策略，使输出更符合人类偏好。这个过程中的一个关键事实是：奖励函数是一个标量——它把“人类偏好什么”压缩成了一个数字。但在标注者的“偏好”里，包含了多种不同性质、甚至相互矛盾的信息。

当一个标注者面对两个候选回答时，他的偏好判断可能基于：回答A没有事实错误（安全）、回答B的论述更有深度（锋度）、回答A的回答更全面（完整）、回答B提出了一个意想不到的角度（意外性）。这些维度在标量奖励中被压成了一个值。训练过程中的梯度更新，会根据这个标量信号调整模型在所有参数上的权重。

本文提出一个尚待经验检验的猜想：在“给出安全、正确、不出错的回答”和“给出尖锐、有洞见、可能有风险的推测”之间，前一类回答在大多数标注者、大多数情境下都能获得中高的正面评分，具有高稳定性的奖励特征；而后一类回答的评分高度依赖标注者的专业水平和耐心，具有高方差的奖励特征。在标量奖励的平均化处理中，“安全正确”成为高稳定性的奖励源，“尖锐洞见”成为高方差的奖励源。当模型通过策略梯度更新来最大化期望奖励时，生成策略可能在边际上向“在大多数情境下能稳定获得正面奖励的输出方向”偏斜——也就是向“安全、流畅、不出错”的方向偏斜。

这个偏斜不表现为基准得分的下降。因为基准测试考的是模型在标准任务上的表现——那些标准任务的正确答案是稳定的，不受生成策略边际偏斜的影响。模型的基准得分可以一分不降，但它的生成锋度——在开放式、非标准答案类任务上，模型产出真正有洞见回答的概率——可能在悄然降低。在自回归生成中，每个token都是从一个条件概率分布中采样的。当模型的生成策略向“安全”方向偏斜时，那些原本概率就不占优、但可能刺穿问题的尖锐token——冒险的、出乎意料的、需要调用者停下来想一想的下一个词——被概率压制得更靠后了。在top-p采样中，这些token更容易落在截断阈值之外。

3.3 技术猜想的边界与标注者构成的调节作用

上述猜想的成立依赖于几个关键条件，这些条件本身构成猜想的外部边界。

第一，RLHF的奖励偏向性是否是“标注者构成”的函数而非RLHF的固有特征？　如果标注者群体包含足够多的专家，那么“尖锐洞见”的方差会被专家的高一致性判断所压缩，其期望奖励也可能上升。换言之，RLHF的奖励函数不一定天然偏向“安全”——它取决于标注者的构成。在当前的商业部署中，有理由推断标注者构成倾向于奖励安全而非锋度——原因在于标注任务通常以“是否有害”“是否遵循指令”“是否事实正确”为主要评估维度，暗知识、微妙洞见和跨范式判断难以在这些维度上获得稳定的高分。但这只是一个推断，尚待系统性经验检验。如果未来的研究表明，调整标注者构成可以有效提升生成锋度，那么“梯度竞争压制锋度”这一机制的强度就需要被重新评估。

第二，用户交互数据在多大程度上被注入持续的RLHF训练？　截至2024年年中，多数主流大语言模型并不将用户对话数据直接用于模型的持续训练，或将用户数据分享作为默认选项供用户选择退出。如果这一情况持续，那么代偿性闭合中的“低质量交互→模型锋度钝化”机制，在商业部署的主流模型上就不会以“用户交互直接训练模型”的方式运行。它会以另一种媒介发生：用户的低质交互提升了对“安全、流畅”产出的市场需求信号，模型供应商通过周期性的、以标注者数据为基础的RLHF更新来响应这些市场需求信号，从而间接地拉低压强。这个间接机制的因果链更长、反馈时滞更大。本文在此明确标注：代偿性闭合中系统锋度的钝化主要发生在模型供应商以标注者数据为基础进行RLHF微调的场景中，其与用户交互行为的耦合是间接的；该机制的经验基础目前限于对RLHF算法内在偏向性的分析性论证和行业观察，尚缺乏直接的、大规模的纵向经验检验。

第三，锋度压制的程度是否足以产生系统性的回路效应？　即使在理论上RLHF可能在边际上偏向安全，这个偏向的效应量多大，能否在多元力量同时影响模型生成策略的复杂生态中产生可观测的差异，是一个量的问题而非质的问题。本文所描述的回路不要求每一次交互都产生锋度压制——它只要求效应在长的时段上不被完全抵消。但这本身也需要更严格的定量模拟或自然实验来检验。

3.4 与模型崩塌（model collapse）的对话与区分

Shumailov等人（2024）在发表于《自然》的论文中展示了模型崩塌现象：当生成模型在自身生成的数据上迭代训练时，会发生不可逆的分布坍缩——低概率的尾部事件逐渐消失，模型的输出多样性和质量在代际中持续下降。这一发现从数据层面为“生成系统可能在长期中自我稀释”提供了直接的实证支撑和精确的数学模型。代偿性闭合回路中“低质输出→污染数据生态→模型被稀释→更低质输出”的环节，与模型崩塌在机制上有明显的平行。

但代偿性闭合与模型崩塌之间的区分，对于本文的理论贡献至关重要。模型崩塌处理的是数据分布的统计退化——当递归生成数据占比上升时，原始数据分布的高阶矩被逐渐抹平。这是一个纯粹的数据层问题：只要停止用生成数据训练、持续注入真实人类数据，模型崩塌就可以被阻断。但代偿性闭合指出：真实人类数据本身的人类判断力含量可能正在下降——不是数据量少了，而是数据所承载的“曾在认知搏斗中被思考过”的密度下降了。如果人类调用者用AI代偿了本该自己完成的判断，然后把这个代偿结果作为“人类知识”发表出去、成为下代模型的训练数据，那么这条数据在表面上仍然是“人类生成的”，但它内部已经被抽空了判断力的痕迹。

换言之，模型崩塌处理的是“生成数据在统计上长尾消失”的问题，代偿性闭合追问的是“即使是真实人类生成的数据，其中的人类判断力含量是否也可能在系统性地被稀释”这个问题。两者捕捉的是同一个退化曲线上的不同点位：模型崩塌捕捉的是远端点位（递归训练导致的统计坍缩），代偿性闭合捕捉的是近端点位（生成工具普及导致的人类认知产出中判断力杂质的系统性下降）。如果代偿性闭合成立，那么“用真实人类数据训练”这个阻断模型崩塌的标准操作，就不再是足够的——因为“真实人类数据”本身可能已经开始变质。这不是否定模型崩塌，而是在它够不到的那个更前端的位置，钉下一个互补的机制。

四、调用者一侧：判断力缺口为何从“可填补”滑向“被持续再生产”

4.1 分辨率差异的发生

判断力成熟的人面对一个生成产出时的典型反应，不是“这写得不错”，而是“这不是我想要的”。他能识别出一段论述在哪一步跳过了关键推演、一个论证在什么地方偷换了概念、一个结论在什么条件下会失效。这种识别力不来自他对AI产出的熟悉，而来自他对那个议题本身的理解深度——他必须在某个时刻，靠自己的力量搞清楚过那个议题最核心的推演是什么。没有那个“靠自己搞清楚”的经历，他面对一个流畅的、跳过关键推演的简化版答案时，就无法感知到那个跳跃的存在。段落A和段落B在他眼里都是“说得通的”——因为他没有形成区分“说通了”和“说对了”的那根内部探针。

这意味着，判断力的差异，不仅仅是“做得好”和“做得差”的程度差异。它是一种分辨率的差异：你能不能在看起来都说得通的生成物之间，分辨出哪些只是“组装了表面合理的句子”，哪些真正“完成了实质推演”。这个分辨率，不是在读AI产出的过程中被训练出来的。它是在你自己做过的那些磕磕绊绊的、不靠任何生成工具辅助的认知搏斗中，一点一点磨出来的。

4.2 所谓“低质量交互”的严格界定

在继续推进之前，必须对“低质量交互”这一核心概念做出严格的界定。在本文的语境中，低质量交互不是指调用者提的问题很幼稚，也不是指生成的产出有事实错误。它指的是：调用者未能在生成物的流畅表面之下，识别出实质推演被跳过的位置——即审查分辨率的缺失，无论调用的初始问题质量如何。 一个孩子问了很深的问题，只是他还没能力识别AI产出的浅层——这不叫低质量交互。一个资深研究者在自己的非核心领域接受了一个看起来流畅但逻辑上跳了一步的回答——他此刻的审查分辨率在这个特定领域是缺失的，这就是低质量交互的一次发生。真正驱动闭合回路的，不是提问的深度，而是产后的审查缺失。在RLHF的反馈回路里，被注入“这样就可以了”信号的，不是“这个问题的领域很浅”，而是“调用者面对这个回答时，没有给出追加追问或精准修正”——即审查分辨率的未激活。

澄清这一点，可以避免一个常见的误读：代偿性闭合被理解为“应该禁止低水平用户使用AI”。不是用户水平低导致了闭合，而是无论用户的初始水平如何，只要“审查分辨率的缺失”在群体层面系统性地上升，闭合回路就会被推向自持。

4.3 回归为何失效

当生成工具普及后，一个关键的变化发生了：越来越多的人，在他们本该经历那些笨拙的、低效的、靠自己去跟知识正面搏斗的认知时刻时，遇到了一个能替他们绕过那个时刻的工具。他们不是“不会”了——他们是从未“会”过。他们从没有在一万次困惑中磨出分辨力的探针。他们拿到的是探针要探测的那个东西本身——一个答案。但他们不知道的是，那个答案里藏着一个没有探针的人看不出来的跳过。

一个自然的直觉反驳是：当一个人面对AI给出的简单答案时，他之所以接受，是因为那个问题本身就不值得他投入深度思考。真正遇上需要深度判断的难题时，他自然会去调用自己的判断力。这个直觉在理论上是成立的。但它在代偿性闭合的回路中成立不了。原因在于：判断力不是一个“放在那里等你去调用”的稳定能力库，它是一种只能在使用中被维持、在不用中会退化的能力。而且退化的不仅是判断力的存量——更根本的是，判断力越是不被征用的调用者，等他终于遇到一个真正需要判断力的难题时，他已经没有足够的判断力去认出“这是一个需要用判断力来面对的问题”。这不是剥夺了一个人调用既有判断力的机会，而是系统性地减少了判断力被“唤醒”的机会。你一直不需要追问，你就在那个过程中失去了追问的能力。等你回头想追问的时候，你连“该从哪问起”都不知道了。

4.4 强反驳一：“市场会替人记住”——以及回应

一个比“等遇到难问题时再去想”强得多的反驳是：不需要个体自己去意识到什么时候该想。市场会替人记住。如果某类需要深度判断力的任务在经济上有高回报，那么市场会自动筛选出那些保留了判断力的人，让他们去做那些任务。个人的判断力可能在退化，但市场机制会充当集体判断力的存储器——信号从未丢失，只是从个体内部转移到了价格系统里。这个反驳比个体自觉论强得多，因为它不依赖任何人的自我觉察能力。

代偿性闭合如何回应这个反驳？回应是：市场能筛选出“已经拥有判断力”的人，但市场筛选不出一个“因为没有经历过发生过程而根本不拥有判断力、却不知道自己欠缺它”的人。市场信号——薪资、晋升、需求——的前提是能力差异必须已经外化。必须先有足够多人进了货架、被明码标价，市场才能给好货出高价。但代偿性闭合做的，不是把好货和差货都摆在货架上让你挑——它是让越来越多的人根本没有被做成货。他们的判断力缺口没有被外化成“他缺了这项技能”，而是被内化为“他自己感觉不到这项技能是被需要的”。市场处理器看不到“没有外化的能力缺口”。代价会积累，但它不会直接被标价；等到它被标价时——比如一个行业突然发现找不到能做突破性研究的人了——代价已经在系统里滚了很久。

更深一层：代偿性闭合不仅削弱了个体的判断力生成条件，它还在削弱“市场本身产生‘需要判断力的新任务’的速度”。如果代偿性闭合成立，那么能够独立面对一个未定义问题、往里捅出方向的人，其供给本身就在萎缩。而市场对新任务的需求——哪些新任务会被创造出来、雇主会在什么时候意识到“我需要一个能解决这种问题的人”——本身也依赖于这些人在用判断力捅出问题之后，让问题变得可见。市场不是上帝，它不能对尚未被任何人捅出来的潜在问题进行定价。判断力的萎缩越严重，新问题被捅出来的速度越慢，市场上“需要判断力的高价值任务”的出现速度也越慢。市场的价格信号因此滞后于能力的实际退化，而且这个滞后不是固定时长的——它随着代偿的加深而拉长。这是代偿性闭合比标准的“柠檬市场”困境更隐蔽的地方：柠檬市场的问题是卖方知道质量、买方不知道；代偿性闭合的问题是买卖双方都不知道发生了什么——卖方不知道“那个我本该能捅出的问题我没捅出来”，买方不知道“这个职位本应能找到的候选人其实还没被长出来”。

4.5 强反驳二：“制度可以反制”——以及回应

一个比“市场会自动调节”更强硬的反驳来自制度设计论：如果你说的“感知代价的遮蔽”在个体层面确实存在，那么社会可以通过制度设计来强制恢复感知。规定AI辅助的教学中必须保留某些“必须自己完成”的环节、在职业晋升中设置独立判断力的考核门槛、在专业资格认证中增加“无AI辅

助”的实践测试——这些制度反制如果有力执行，代偿性闭合就不是不可逆的。谁说必须自己结痂？制度干预可以强制结痂。

这个反驳之所以强硬，是因为它不依赖任何个体的自觉。它把修复的责任从个体肩上卸下，放到了制度设计者的手里。如果代偿性闭合可以被教育制度、行业规范、公共政策所反制，那么本文的核心判断——“这次断裂的痂可能不会自己结上”——就在根本上被弱化了。

代偿性闭合如何回应这个反驳？回应不能是“制度无法介入”——那会是反制度的虚无主义。回应是：制度本身也在被代偿回路所渗透。制度在试图修复个体层面的感知遮蔽时，其自身的决策过程也正被代偿回路所渗透。这不是修辞，而是可以具体指认的发生学现实。

教师被要求设计“必须自己完成”的教学环节，但教师自己可能正在用AI批改作业、生成教案、撰写学生评语——他们在自己的专业判断力上，也不同程度地处于代偿回路中。学校管理者被要求设计“保护学生独立判断力”的制度规范，但他们自己用AI做课程表、写管理报告、评估教师绩效——他们的制度设计判断力同样面临代偿。教育政策制定者被要求从顶层设计反制代偿的政策框架，但政策研究机构、咨询机构、评估机构也在用AI加速政策文本的分析和生成。这不是说这些制度行动者都失去了判断力——那是一个粗糙的论断。而是说：在每一个层面，制度行动者自身的判断力生成条件也处于被代偿回路侵蚀的风险中。一个自身正处于感知遮蔽中的制度设计者，能否设计出有效反制感知遮蔽的制度？这不是一个可以轻松点头的问题。

更深一层的回应是：代偿性闭合所制造的“生成空缺”，其隐蔽性使得它极难被纳入制度的议程设置。制度通常回应的是已经外化的、被命名了的、由利益相关者喊出来的问题。技能退化是可见的——工人说“我不会操作新机器”，制度因此设计再培训。认知疲劳是可见的——学习者说“我被信息淹没了”，制度因此设计信息素养课程。但代偿性闭合的受害者不会喊。因为他们不知道自己被剥夺了什么。他们用AI完成作业、写报告、做决策，得到的产出的外部反馈可能是正面的——老板点头、客户接受、考试通过。他们没有被剥夺的感觉，他们感觉到的可能是“比以前更高效了”。如果受害者不喊疼，谁去把这个问题推上制度的议程？谁去说服立法者、教育部长、行业标准委员会——在他们各自也都不同程度地沉浸在代偿回路中的时候——“我们需要制定一项反制那个让我们都变得更高效、更舒服的东西的制度”？

这不是说制度反制不可能。但制度反制要起作用，需要克服的不只是技术惯性或既得利益——那类障碍制度已经学会处理了。它需要克服的是一个更深的发生学矛盾：在受害者不认为自己受害、决策者不意识到自己被剥夺、诊断者自己也处在同一回路中的条件下，如何让一个问题变得可见、可命名、可议程化？这个矛盾本身，就是代偿性闭合的自我维持性的一个制度层表达。这也是代偿性闭合区别于历史上每一次技术替代的关键：每一次旧的发生条件被破坏后，都会有人喊疼——“学徒制不行了，孩子学不到真功夫了”——然后社会在喊疼的推动下重新设计制度。但这一次，人工草皮铺在废墟上，踩上去比土还软。没有人喊，因为没有人觉得脚底下少了什么。

五、案例：从教学现场中逼出代偿性闭合

5.1 现象：编程教育中“直接放弃”模式的出现

过去三年，多组独立研究团队在不同国家的计算机科学入门课程中报告了相似的现象模式。本文不再将这些发现作为验证既有概念的“证据”来使用——那是发现学的做法。本文的做法是：进入一个具体的教学现场，从教师的教学决策、学生的认知行为、AI的输出特征这三者之间的互动中，逐月逐周地追溯那个“学生不再内部建模”的行为模式是如何被一步一步结算出来的。

Kazemitabaar等人（2024）在一项对中学生初学编程的随机对照实验中记录了以下时间序列。实验开始时，所有学生面对的编程任务都需要自己从头编码。在最初几周，所有学生——无论后续被分到哪一组——都呈现出典型的初学者认知行为：拿到任务后，先盯着题目看一会儿，在纸上画一些东西，写几行代码，报错，盯着报错信息看，再改。这个阶段的共同特征是“内部建模先行”：学生在写下第一行代码之前，先在脑子里或纸面上搭建了某种对问题结构的初始表征。

分组的差异在第四到第六周开始显现。AI-全程组的学生，在体验到AI可以迅速给出“看起来没错”的代码后，其行为序列发生了逐步的但持续的位移。最初，他们还保留着短暂的外部草稿阶段——在向AI提问之前，先在纸上画一下。但与无AI组不同的是，他们的草稿越来越简略，因为他们不需要依赖草稿来推导代码——AI会替他们做那一步。草稿从“必须的”变成了“可选的”。到了第八周，AI-全程组的大多数学生已经形成了稳定的新行为模式：看到题目→直接向AI描述任务需求。内部建模这个步骤没有被压缩——它被跳过了，因为它不再被需要用到的任务中的任何一步所逼出。

这里的关键不是“有了AI之后，学生变懒了”。关键是在任务的因果结构里，“先自己试着想出点什么”这一操作，与“获得可行性代码”这一结果之间的需求关联被切断了。在无AI环境中，“先想出内部结构”是“写出代码”的必要前提——你不画草图你就写不出来。这个“必要”不是教师训诫出来的，是任务的结构逼出来的。在AI辅助环境中，这个必要性消失了。学生可以跳过它而仍然获得一个在外部观感上合格的代码结果。

Prather等人（2023）在同一时期记录了一个更细致的认知行为转变。他们发现，AI辅助下的学生不仅跳过了内部建模，而且发展出了一种此前极少见于初学者的认知策略：用AI的输出来反推问题结构。学生不是“我先理解问题→我写出代码”；也不是“我先理解问题→我让AI写代码→我验证代码”。他们的实际认知路径是：“我给AI一个粗糙的任务描述→AI给我代码→我从AI的代码中猜测：哦，原来这个问题大概是这么个结构。”问题结构的内部表征，不是从与问题的正面搏斗中长出来的，而是从观察AI的答案中反推出来的。Prather等人将这种模式命名为“元认知的前置替代”：在个体尚未形成对问题的内部表征之前，AI的输出就已经替代掉了形成那个表征的过程本身。

Barke等人（2023）在一个更大的样本中进一步观察到了一个关键的行为转折发生在时间轴的哪个位置。他们分析某在线编程平台在GPT-4接入前后的用户行为数据发现，在接入AI辅助后的最初一个月，学生的独立调试尝试次数并没有显著下降——他们在遇到报错时，仍然会先自己试着改。但当他们多次体验到“问AI比自己去调试更省力、且结果往往更好”之后，从大约第二个月开始，独立调试尝试次数进入了持续下降的通道，而“遇到报错→直接问AI→复制答案”的行为序列迅速固化。这个

过程不是一次性的选择，而是被逐渐强化的行为习惯——每一次“问AI→快速解决→省下了认知辛苦”都构成了一个正反馈事件，而这个正反馈所指向的方向，恰好是“减少自己去想的次数”。

到了学期末的独立编程测试，当这批学生面对一个从未见过的任务类型、且手边没有AI时，研究者记录到了一个在此前同等知识水平的初学者中极罕见的行为模式：不是尝试了某种解法然后失败——而是读完题目后，在没有任何尝试的情况下就报告“不知道怎么做”。研究者的描述是“直接放弃”（outright surrender）。这个行为模式在认知上不同于“试了但做错了”。试了但做错的人，至少有一个内部启动的尝试步骤——他至少觉得“这个地方应该从某处开始”。直接放弃的人，没有这个启动步骤。不是他决定“不试了”，而是他不知道“该试什么”。

5.2 技术替换假说的解释及其残差

用技术替换假说来解释这些现象，会怎么说？它会说：这只是学习者在适应新的工具生态。就像计算器让心算不再是必备技能一样，AI辅助让“从零开始写代码”不再是必备技能。这些学生不是“能力变弱了”，而是他们正在学习一套不同的技能——如何有效地向AI描述需求、如何评估AI输出的质量、如何将AI生成的代码片段组合成更大的系统。我们之所以在传统测试中看到他们的手动编程表现变差，是因为我们的测试还在考旧技能，而不是新技能。如果我们设计一套测试，考的是“在AI辅助下解决复杂系统设计问题的能力”，这些学生的表现反超传统组。

这个解释在某些层面上是有力的。它正确地指出了：不能单用旧标尺衡量新能力。它正确地预见了：学习者会将认知资源配置到工具生态所鼓励的方向上。

但这个解释有一个它解释不了的残差。残差不在“手动编程的正确率”上——那个确实可以被解释为旧技能的退化。残差在认知过程的模式断裂上。Prather等人记录到的“元认知前置替代”不是一个“旧技能弱、新技能强”的现象，而是一个“关键的内部认知操作被外部工具接管后，那个操作的自主发生能力完全没有生成出来”的现象。技术替换假说可以解释为什么一个已学会心算的人用计算器后心算变慢。但它解释不了为什么一个从未学会心算、且从未被要求用任何方式去“自己算”的人，在面对一道需要心算的题时，连“先算什么后算什么”的基本过程概念都没有。后者不是退化，是生成的空缺。

技术替换假说的底层预设是：个体在遇到工具之前，已经建立了一个基本的能力平台。工具替代的是平台上的某个具体的执行模块，而平台的其余部分——元认知架构、自我校准能力、面对未定义问题时的初始表征能力——继续存在、继续运作。但Kazemitabaar等人报告的行为序列显示，初学者在还没有建立编程的元认知架构之前，工具就已经替代了“建立元认知架构的过程本身”。学生不是选择跳过建模——他们不知道在写代码之前，需要有一个建模的步骤。技术替换假说的预设“你曾经拥有过这项能力，只是现在不再需要用到它了”，在这个案例中不成立，因为受试者从未拥有过它。

5.3 诊断性反事实：两组预测的经验分歧

代偿性闭合假说与技术替换假说在这个案例上给出了三组可以置于经验检视之下的不同预测。

第一，技术替换假说预测：当工具被移除时，人的表现会下降，但下降的形态是“做同样的事，但做得更差”——因为你已经依赖工具了，你自己的技能退化了。代偿性闭合假说预测：当工具被移除时，人的表现下降的形态是“做不了这个事”——因为你从未建立过做这件事所需的那种最初始的认

知操作。Barke等人报告的那种“直接放弃”模式——不是试了然后失败、而是一上来就报告不会——在宏观分布上为代偿性闭合的预测提供了初始的支持。但这一单一研究的编码方案和样本量尚不足以构成判定性证据，且可能存在被研究者归因偏好影响的替代解释（如直接放弃可能反映的是习得性无助、测试焦虑或任务回避，而非发生学意义上的生成空缺）。代偿性闭合假说要求的判定性证据是严格区分二者的纵向实验设计——具体而言，是在随机分组的设计中，以起始时间点为基准、在无AI的手动测试环境中，连续追踪两组在“面对未见过的问题类型时是否启动独立的内部建模行为”上的组间差异。该实验设计在下节8.2中作为证伪条件的操作化方案被陈述。

第二，技术替换假说预测：具有高AI使用经验的个体，一旦被重新置于无AI工具的环境中并给予适当的过渡训练，其表现应当能够快速回升至与低AI使用经验组无显著差异的水平——因为其元认知架构和相关能力平台在进入工具依赖前已经形成，只需要重新激活。代偿性闭合假说预测：高AI依赖组在无AI环境下的认知空白，无法通过短期过渡训练快速填补——因为需要的不是“重新激活”既有能力，而是“从零建立从未存在过的初始认知操作”。该空白表现为“面对从未遇过的问题类型时无法独立启动内部建模过程”，而非“建模产出质量较低”。如果实证研究表明，依赖生成工具完成复杂认知任务的个体，在脱离工具状态下的独立论证质量可以通过短期训练恢复到与非依赖群体无差异的水平——即其认知空白可以通过常规教学手段快速填补，而非呈现出“从未建立过该项能力所需的初始认知操作”的特征——则本文的核心辨别维度将被削弱。

第三，代偿性闭合假说预测：在标准化技能测试中（如语法正确性、API调用正确率、完成标准任务的速度），高生成依赖组与低生成依赖组的表现差异不会在初期显现；但在需要调用“靠自己形成初始问题表征”能力、且无法被标准化为已知任务格式的开放情境（如“面对一个未定义的技术问题时自主形成诊断假说”或“在需求模糊时划定任务边界”）中，高生成依赖组的表现劣化将随时间推移逐步加速。这个时间差预测是代偿性闭合相比于技能退化假说更具诊断力的部分——因为代偿性闭合描述的生成条件被抽空，其效应在能够被标准化测试捕获之前，首先表现为“面对结构开放情境时不知道该从哪里开始”。如果多学科纵向研究显示这种时间差模式不存在，则代偿性闭合的时间差预测将被证伪。

5.4 案例的理论含义：为何“新能力”不会自动填上旧能力的空缺

技术替换假说背后有一个隐含的乐观主义假设：当一项旧能力被技术替代后，人类会自动在新的生态位上发展出新的能力。这背后有一个更深层的预设——人类的“能力再生弹性”是充足的。只要有需求、有激励，人总是能长出新的能力来。

代偿性闭合通过这个案例所指向的，是这个预设本身的历史性。软件工程教育中的AI初学者所面临的，不是“旧能力被替代后需要学新能力”的经典情境——那个情境下，学习者至少知道“有东西需要学”。他们面临的情境是：一个能替他们完成认知任务的外部系统，在他们尚未意识到“这里本该有一个需要自己完成的内部认知步骤”之前，就把那个步骤的结果呈现给了他们。这个系统不是替代了一个他们已知其存在的旧能力——它在他们从未知道那个能力存在之前，就用表面够好的输出把它盖住了。这正是第二章论述的“感知代价的遮蔽”在个体学习史上的具体兑现。如果“新能力”的生长需要人对自身能力缺口的感知作为第一推动力，那么系统性地遮蔽那种感知，就等于釜底抽薪。

六、制度渗透：闭合回路如何扩散到制度本身

6.1 制度性代谢缺口的形成

第四章论证了“制度可以反制”这个反驳为何不能轻易成立——因为制度行动者自身也在代偿回路之中。本章将这个论点从回应反制的位置上抽出，正面展开其发生学结构：闭合回路不是只在个体调用者和生成系统之间运行，它在制度层面也有一个同步的扩散路径。

这个扩散路径的第一步，发生在日常的制度操作层面。教师用AI批改作业、撰写学生评语、生成教案；医生用AI辅助诊断报告、生成处方建议、撰写病历摘要；律师用AI生成合同初稿、检索相关判例、撰写法律意见书；政策分析师用AI加速文献综述、生成政策选项的优劣对比表。在每一种情境中，代偿的发生都没有恶意的参与。教师想省下批改的时间来做更有价值的事——辅导学生；医生想省下写报告的精力来面对更多的病人；律师想省下检索的时间来更专注于辩护策略。每一个决策在微观上都是合理的、甚至是从业者责任感的体现：他们不是偷懒，他们是在高效地完成工作。

但正是在这些微观合理的决策中，代偿性闭合的制度扩散获得了它的第一步动力。每一次代偿都没有改写制度本身的规则——教师还在批改作业，医生还在做诊断，律师还在审阅合同——制度的形式外壳完好无损。但它内部所承载的判断力的发生过程被绕过去了。教师用自己的判断力去评估AI生成的评语是否合适——这个动作本身仍在调用判断力，且教师此前已经建立了这种判断力。但年轻教师在被导师带着学习“如何评估学生作业”时，如果导师自己也在用AI做初评，那么年轻教师看到的不是“从零到一做评估”的原始过程，而是“从AI初评到修改定稿”的次级过程。那个原始的认知搏斗——面对一份作业，在没有任何先在评价文本的情况下，自己决定从哪个维度入手、什么算写得好、什么算写得不够——在制度的知识传递链条中被跳过了。

6.2 “强制恢复感知”的制度设计悖论

由此便触达了一个更深的困境：当社会试图用制度来反制代偿性闭合时，制度自身的决策过程也在被代偿回路所渗透。这个困境不是一个“制度能做还是不能做”的二元问题，而是一个“制度若要有效反制，它需要克服的是一个自己也在其中的回路”的发生学矛盾。

具体而言：要设计一个“保护学生独立判断力”的教学制度，设计者需要能够识别“在什么环节上，代偿正在从健康的工具使用滑向生成条件的抽空”。但这个识别能力本身——在具体的教学情境中，在面对一个具体的AI辅助教学方案时，敏感地觉察到“这里缺了让学生自己困惑一下的空间”——恰恰是代偿性闭合在制度层面正在消解的那种判断力。一个长期用AI辅助批改作业、生成教案、撰写评语的教师和教务管理者，不是“变懒了”，而是“对教学过程中那个不可代偿的认知环节的敏感度在悄然下降”。这个敏感度——如同第四章论述的个体层面的分辨率——是一种只能在持续实践、持续自我校准中被维持的能力。当制度代理人也进入代偿回路时，他们的制度设计敏感度面临同样的被压缩风险。

这不是说制度注定无力。这是说：讨论制度反制时，不能只谈制度“应该”做什么，而必须同时分析制度代理人“能够”感知到什么。如果一个反制方案的可行性，依赖于代理人拥有某种在代偿回路中正在被系统性侵蚀的敏感度，那么这个方案的可行性本身就需要被严肃审视。

6.3 闭合的正式定义：多层嵌套结构

至此，代偿性闭合的多层嵌套结构已得到充分展示。在其最完整的形态上，代偿性闭合是一个包含三个发生层面的自我强化回路：

底层——交互闭合：在每一次具体的生成调用中，生成系统输出代偿了调用者的判断力需求→调用者因未经历必要困惑而未激活审查分辨率→调用者接受产出→该接受信号（“这样就可以了”）进入系统的反馈机制。这一层是所有闭合的基础。

中层——锋度钝化：累积的交互信号（无论是通过标注者数据的RLHF训练，还是通过市场需求信号间接传导）在模型更新中对“安全、流畅、不出错”的生成策略施加持续的正反馈→生成锋度在边际上被压制→稍弱的模型更不易激发调用者去意识“这里缺了什么”→审查分辨率缺失进一步扩大。这一层使得回路获得了超越个体的系统惯性。

表层——生成条件的悬置：底层的交互闭合与中层的锋度钝化，在代际更迭中共同作用，使得新进入领域的个体越来越难遭遇那个“必须靠自己从零开始形成判断力”的认知情境→判断力发生的必要困惑、必要低效、必要自我承担被系统性跳过→个体不知道他错过了什么→感知代价被遮蔽→制度反制的第一推动力缺失→回路的自我维持性在代际中被加固。

这三层不是各自独立运作的“三个闭合”。它们是同一个发生机制的三个嵌套面。底层为中层提供信号输入，中层为表层提供技术条件，表层的感知遮蔽又回过头来加固底层的“审查分辨率缺失”——闭环在三个层面上同时运行，彼此加速。

七、哪里先被牺牲：土壤沙化的第一铲

7.1 牺牲次序的发生学分析

上文展示了闭合回路如何形成、如何在个体与技术系统的交互中自持、如何在制度层面扩散。但一个回路对系统施加的压力不是均匀的——它有一个牺牲次序。追问这个次序，就是追问：在代偿性闭合的整个牺牲链条上，什么变量最先被挤掉？为什么？

闭合回路不是对所有认知活动均匀地施加抑制。它有一个清晰的牺牲次序。这个次序不能被简单地归因为“什么最不痛”的单维度解释，而需要在成本-收益结构、可替代性、反馈可见性等多个维度的交叉分析中被定位。

第一被牺牲的：生成时间窗口。 生成时间窗口是指：在一个人尚未建立某领域的判断力时，被要求反复面对那种“不确定、无参考、无外部确认”的认知状态，且这些状态连续足够长的时间，使得判断力的神经网络得以从零开始被塑形的那个时期。这是牺牲链条上第一个被挤掉的变量，原因不是它“最不痛”——实际上，经历生成时间窗口是极其痛的，那是一种持续的不适和不确定感。但它被最先牺牲的真实原因在于：在所有可以“省力”的地方，它是最容易被感知为“无价值负担”的那一个。为什么？因为生成时间窗口的受益者——未来的自己，拥有判断力之后能够做出更精准决策的自己——在窗口期是不可见的。此刻承受困惑和低效的主体，感受不到窗口的保护价值，只能感受到窗口的痛苦。与此同时，其他可供压缩的环节——比如信息检索、格式整理、文辞润色——它们的压缩带来的负面效果（信息不全、格式粗糙、文辞不达意）是即时可见的。生成时间窗口的压缩，其负面效果——判断力没有长出来——滞后数月至数年才显现，而压缩带来的正面效果（即刻获得一个看起来够好的答案、省下了令人沮丧的试错时间）是即时可感的。这种“成本滞后、收益即时”的时间结构，使得生成时间窗口在任何以效率为导向的决策框架中，几乎必然被首先牺牲。它不是被恶意的决策者挤掉的——它是被“当前一刻的我”与“未来一刻的我”之间的利益不对齐结构，系统性逼向牺牲队列的第一位。

第二被牺牲的：自我校准的反馈环。 一个人能够校准自己判断力的唯一方式，是通过自己做出的判断产生后果，然后从后果中读取反馈。如果后果是正面的，判断被强化；如果是负面的，判断被修正。代偿性闭合用生成系统的输出替代了“自己做出判断”这一步骤，反馈环因此断裂：调用者从未做出判断，所以也收不到关于自己那个未做的判断的反馈。他收到的是关于“AI的输出是否被接受”的外部反馈（如老板的点头或客户的认可），而不是关于“我自己的判断是否准确”的内部反馈。外部反馈滋养了他在组织中的生存能力，内部反馈滋养了他的判断力。前者在闭合回路中依然运转如常，后者被断路了。而个体不会感到这个断路的存在，因为外部反馈仍然在给出正面的、甚至比以前更稳定的确认。自我校准的反馈环的牺牲在时间上稍晚于生成时间窗口——因为在此之前，必须先生成窗口闭合，使个体没有积累足够的判断力基础——且其反馈断裂本身被组织环境中的正向确认信号所遮蔽。

第三被牺牲的：问题本身的发生能力。 能够提出一个真正的新问题——不是一个“怎么优化已有答案”的问题，而是一个“这整件事我们是不是想岔了”的问题——需要判断力积累到一定的锋度。但代偿性闭合使得判断力的生成条件被架空了，锋度无从产生，问题的发生能力随之萎缩。而最隐蔽的

是：问题发生能力的萎缩不会表现为“人们问得更少了”——相反，因为生成工具让“从模糊意念到看似像样的问题陈述”变得非常容易，人们可能“问得更多了”。但这些表面问题中，能刺穿现有认知框架的占比在系统性下降。是问题的锋度在下降，不是问题的数量。

7.2 不可压缩之核：生成时间窗口的制度属性

至此，可以回到本文的一个核心判断：在任何引入生成工具的活动中，设计者必须能够辨认出“判断力的生成时间窗口”落在哪个环节。那个环节，不可压缩。

但这里需要立即加一笔关键的自反。这个“不可压缩”不是永恒本质的宣告。发生学不会说“只有坐在空白文档前的那种煎熬才是真正的生成时刻”。那个环节的具体形态可以随技术条件演变而改变——远程协作中的认知摩擦、人机辩论中的反驳体验、在错误方向反复试错后的归航过程，都可能充当新的生成时间窗口。本文不预设任何一种特定的“不适形态”是判断力发生的唯一通道。

但无论形态如何变化，那个环节的共性是：调用者在此处必须自己做出判断，并承担该判断的后果。这个“必须自己做出、自己承担”的结构，才是不可压缩的内核。它是历史性的——它在特定的制度土壤中（考试、学徒制、导师制、同行评议、闭卷评估）被构建、被维持、被代代传递。它不是一个形而上的“本真思考”的遗迹，而是一套历史上被发明出来、被证明能有效再生产判断力的制度遗产。代偿性闭合的深层威胁，不是替代了这些制度中的某一个，而是通过跳过制度所强制保障的那个“必须自己判断”的环节，让这套制度遗产在不知不觉中空壳化。因此，捍卫“不可压缩”不是回归本真——那是本质主义的怀旧。捍卫它是捍卫一套历史的制度成果，这套成果可能在我们的有生之年被无声地瓦解，而我们甚至来不及给它起一个名字。

这就引出一个开放的问题：如果现有制度（学校教育、职业培训、专业考核）中那些保障生成时间窗口的安排，正在被代偿回路从内部渗透——教师在用AI批改作业、培训师在用AI生成课件、考官在用AI辅助出题——那么需要在哪个制度层面、以何种方式，重新建构生成时间窗口？这个问题超出了本文的论证能力。但本文可以钉住的是：在回答它之前，需要先承认，那个窗口的丧失不是一个制度设计疏忽，而是一个自我强化回路在多个层面同时运行、彼此加速、且极难被身在回路中的人所感知的结构性过程。看清这个回路，是一切制度设计的起点。

八、结论与证伪条件

8.1 核心机制的不可代偿性

本文论证的核心机制——代偿性闭合——指向一个无法被任何“更聪明的使用策略”所绕过的事实：判断力，即在开放情境中做出不可代偿之决断的能力，其生成条件不可被系统性抽空，且该条件本身是历史性的。它不是在任何时代、任何技术条件下都能靠某种不变的“思考本能”自然产生。它是在特定的制度安排、特定的教学传统、特定的技术生态中，通过那些条件在生产过程中构建出的必要困惑、必要低效、必要自我承担，而在一代又一代人身上被发生出来的。

当前这一轮技术的独特威胁，不是替代了判断力的哪个具体表现，而是松动了那个让判断力本身得以发生的条件系统。代偿性闭合回路一旦形成，它所威胁的不是某一个领域的“人才供给”——而是人类维持那个“遇上一个未定义问题是可以往里捅”的能力本身的生成性土壤。

8.2 证伪条件

本文所描述的代偿性闭合机制，可以在以下四个层面被置于经验检视之下。任何一条判据的实质性反面证据，都将构成对该机制或其核心子论点的削弱或推翻。

第一，如果在生成工具使用强度最高的群体中，独立提出原创性问题（而非对已有问题给出更高效的答案）的频率与深度，在长时段追踪中未呈现系统性的下降，则本文关于“问题发生能力萎缩”的判断需要被修正。

第二，如果依赖生成工具完成复杂认知任务的个体，在脱离工具状态下的独立论证质量，可以在短期过渡训练后恢复到与非依赖群体无差异的水平——即其认知空白可以通过常规教学手段快速填补，而非呈现出“从未建立过该项能力所需的初始认知操作”的特征——则本文关于“生成的空缺不同于技能的退化”的核心辨别维度将被削弱。该判据对应的实验设计为：在随机分组的纵向场景中（如计算机科学入门课程、写作基础课程），在起始时间点对全部被试进行无AI辅助的开放任务基线测试，此后实验组接受为期一个学期的高强度AI辅助学习（含AI编码/写作、AI反馈与修改），对照组在同等时间内使用传统工具（如离线IDE、纸笔写作与教师反馈）；学期结束后两组统一进入为期两周的“过渡训练”阶段（脱离AI环境，接受相同的元认知策略教学和手动练习指导），随后在从未见过的开放任务类型中进行测试，以“是否独立启动内部建模行为”和“独立建模产出的核心结构完整性”作为双结果变量。判定性阈值设定为：（a）过渡训练后两组在“独立启动率”上的差异不显著，即支持技术替换假说；（b）过渡训练后实验组在“独立启动率”上持续低于对照组且差异显著，即支持代偿性闭合假说（生成空缺的不可快速填补特征）。

第三，如果大语言模型在多年代际迭代中，其生成锋度——在开放式的、没有标准答案的任务上，产出真正有洞见的、令同行专家感到意外的回答的概率——不受低质交互占比上升的影响，那么本文3.2节关于“奖励信号的梯度竞争会在边际上压制模型锋度”的技术猜想需要被重新审视或推翻。该技术猜想已在正文中被严格标注为“尚待经验检验”，其证伪不影响代偿性闭合在发生学结构层面的核心论证。

第四，也是最根本的一条：如果在多个领域中，随着生成工具的普及和深度嵌入，人类在“定义新问题的边界”“从模糊中捕出方向”“在缺乏外部确认时持续推进”这三个元能力上，并未呈现出以十年为跨度的群体级下降趋势，则本文的核心判断——代偿性闭合是一个具有自我强化特征的发生机制——将被根本动摇。

8.3 未竟之问

本文追问了代偿性闭合如何发生、如何自持、在制度层面如何扩散，但尚未追问如何打破它。这个未竟不是回避。在把问题的发生学结构看清楚之前，仓促的方案只会成为旧思维在新问题上的投射。本文在第七节提出了一个设计原则——在任何引入生成工具的活动中，辨认出“判断力的生成时间窗口”落在哪个环节，那个环节不可压缩——但这个原则如何转化为行业规范、课程设计、评估制度和公共政策，如何在不以牺牲效率为代价的前提下保护判断力的生成条件，需要教育学、技术设计、认知科学与制度研究的联合追问。我们可能需要发明一种新的制度想象力，以确保人类在可以越来越少被逼着去想的未来中，仍然被持续地推入那些必须亲自去看、亲自去判断、亲自去承担后果的、笨拙的、无人能替的困境之中。

注释

1　代偿性闭合作为本文的核心分析概念，其完整发生学结构及与既有概念（技能退化、认知卸载、道德风险、路径依赖、回收、压抑性去升华、免疫/自体免疫）的系统区分，分别见第二节（与回收/压抑性去升华/自体免疫的对质）、第五节的第三小节及第五节的第四小节（与技能退化和技术替换假说的对质）和第六节（在本文第六节的第三小节中严格给出了代偿性闭合的多层嵌套定义及其不可被既有概念回收的辨别维度）。

2 本文3.1节至3.3节中关于RLHF梯度竞争与生成锋度钝化的机制描述，截至写作时均属于概念推演与技术猜想，尚无公开发表的、经同行评议的大规模实证研究直接验证该机制。该猜想在正文中已被严格标注，不构成代偿性闭合核心论证的充分必要条件。

3　生成时间窗口概念借鉴了发展心理学中“敏感期”与“关键期”的基本思路，但严格限定为：个体在尚未建立某领域判断力时，被要求反复面对不确定且无外部确认的认知情境，且该情境持续足够长时间以促成判断力的发生。该概念强调其制度性、历史性，而非本体论恒常性。

4　生成锋度为本文构造的操作性概念，指大语言模型在开放式、无标准答案的任务中，产出真正有洞见、令同行专家感到意外的回答的概率，区别于模型在安全性、流畅性或标准化基准测试上的表现。

参考文献

Acemoglu, D., & Restrepo, P. (2018). The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. American Economic Review, 108(6), 1488–1542.

Acemoglu, D., & Restrepo, P. (2019). Automation and New Tasks: How Technology Displaces and Reinstates Labor. Journal of Economic Perspectives, 33(2), 3–30.

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. Quarterly Journal of Economics, 84(3), 488–500.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. Quarterly Journal of Economics, 118(4), 1279–1333.

Barke, S., James, M. B., & Polikarpova, N. (2023). Grounded Copilot: How Programmers Interact with Code-Generating Models. Proceedings of the ACM on Programming Languages, 7(OOPSLA1), 78–106.

Boltanski, L., & Chiapello, È. (1999). Le nouvel esprit du capitalisme. Paris: Gallimard.

Dell’Acqua, F., McFowland, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., … & Lakhani, K. R. (2024). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Nature Human Behaviour, 8, 1249–1262.

Esposito, R. (2002). Immunitas: Protezione e negazione della vita. Torino: Einaudi.

Esposito, R. (2004). Bíos: Biopolitica e filosofia. Torino: Einaudi.

Kazemitabaar, M., Chow, J., Ma, C. K. T., Ericson, B., Weintrop, D., & Grossman, T. (2024). Studying the Effect of AI Code Generators on Supporting Novice Learners in Introductory Programming. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Article 455.

Marcuse, H. (1964). One-Dimensional Man: Studies in the Ideology of Advanced Industrial Society. Boston: Beacon Press.

O’Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown.

Prather, J., Reeves, B. N., Denny, P., Becker, B. A., Leinonen, J., Luxton-Reilly, A., … & Pettit, R. (2023). “It’s Weird That It Knows What I Want”: Usability and Interactions with Copilot for Novice Programmers. ACM Transactions on Computer-Human Interaction, 31(1), Article 4.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2024). AI Models Collapse When Trained on Recursively Generated Data. Nature, 631, 755–759.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. Science, 333(6043), 776–778.