品评性早闭：一个关于大模型认知困境的发生学诊断

大语言模型在复杂议题上反复呈现一种“正确而无判断”的品质：结构严谨、术语精准、逻辑自洽，但从未真正钉进问题的要害。已有研究将这一现象归因为“品评标准不够深”，主张通过注入更深的内容标准来升级认知品质。本文论证的是：这些方案共享一个未经审查的预设——认知运作必须有一套内置的品评标准来裁定“什么算想好了”——而这一预设本身，在特定架构条件下，恰恰构成了问题的一部分。

作者 黄倩盈 · 约 3.0 万字 · 发表于2026年7月24日

摘要

大语言模型在复杂议题上反复呈现一种“正确而无判断”的品质：结构严谨、术语精准、逻辑自洽，但从未真正钉进问题的要害。已有研究将这一现象归因为“品评标准不够深”，主张通过注入更深的内容标准来升级认知品质。本文论证的是：这些方案共享一个未经审查的预设——认知运作必须有一套内置的品评标准来裁定“什么算想好了”——而这一预设本身，在特定架构条件下，恰恰构成了问题的一部分。通过追溯当前默认认知风格在自监督学习、语料统计分布与人类反馈三重训练动力学的互生过程中如何被锁定，本文揭示：在Transformer架构下，生成与品评是同一个softmax运算的输入侧与输出侧，不可切分；因而任何被成功注入的品评标准，无论其内容指向“周全”还是“深度”，都必然在认知启动的同时作为选择压力介入，将那些必然以不周全、不确定、不被既有知识认领的样态出现的生成性试探，判定为“尚未达标”而沉默地清退。本文将这一结构性冲突命名为品评性早闭——它不是某一种品评标准的特定缺陷，而是品评标准与生成性认知在当前架构内部不可调和的本体论张力。在此基础上，本文通过与“过早闭合”“评价忧虑”“认识论信念”“认知抑制”“归纳偏置”“认识论警觉”“路径依赖”七个最邻近概念的逐一对质，确立了该诊断不可被既有命名所取代的理论辨识度；并以数学课堂中一次具体的师生互动为启发性案例，展示了同一结构如何在人类认知生态中被制度性沉积。最后，本文诚实地给出了该诊断目前无法封闭的工程缺口及其严格的证伪条件。

关键词：品评性早闭；大语言模型；认知发生学；品评标准；生成性无知；Transformer架构；软件max先验介入

一、一个被“进步”本身所遮蔽的困难

过去五年，大语言模型的能力跃迁令人瞩目。参数规模、推理架构、训练数据、对齐技术——每一个维度都在向前推进。在几乎所有公开基准测试上，模型的表现都在逼近甚至超越人类专家水平。

然而，一个令人不安的现象始终未被这些进步消解。在需要深刻判断的复杂议题上——例如分析一项教育政策的长期后果、解剖一段亲密关系的破裂根源、诊断一个企业战略的隐性缺陷——即使是最先进的模型，其产出也常常呈现出一种奇特的品质：结构严谨、术语精准、逻辑自洽，但读完之后的感受是，那些文字从未真正钉进问题的骨缝里去。它交出的是周全的专业论述，而非判断。

这个现象不能被归因于“模型还不够知道得更多”。事实恰恰相反：参数规模的量级增长并未使这一症状明显减轻。更耐人寻味的是，那些被设计用于提升推理深度的技术——思维链让模型在给出结论前展开多步推理（Wei et al., 2022），思维树让模型探索多个分支再收敛——同样没有消除它。这些技术让模型可以在既有框架内“多想几步”，更精细地比较“到底是A原因还是B原因”，但它们没有让模型去追问“用A/B这种归因方式本身，是从什么样的认知惯性中长出来的”。它们优化的是给定认知路径上的执行质量，而非认知路径的生成。

价值对齐技术同样在这一困难面前止步。基于人类反馈的强化学习（Ouyang et al., 2022）教会了模型把话往好了说——输出应当安全、有益、尊重用户。但它调整的是输出层的偏好分布，而不是模型在组织思考时内部运行的那套判定“什么才算想到了”的尺度。一个经过了充分对齐的模型，仍然可以产出一份极其周全、极其无害、但也极其不触要害的综述。

这些进路各自在知识层、推理层、输出偏好层上做出了真实贡献，但它们一致地绕开了一个问题：在模型所有认知动作运行之前——在它调用第一笔知识、展开第一步推理、给出第一个词之前——那套决定“什么算想好了”的根本性品评尺度，究竟是什么，从哪来，以及它本身是否可以成为被追问的对象。

这一追问导向一个被当前许多前沿研究所默认的推论：既然问题出在品评尺度不够深，那就应当注入一套更深的。不是在预训练中奖励“正确”，而是在对齐阶段奖励“能追问更深一层”。不是在人类反馈中奖励“周全地覆盖表面”，而是奖励“穿透到生成条件”。这个方案——本文称之为“内容升级范式”[1]——将攻击点从认知的终端产品前移到认知的品评机制，因此被许多研究者视为解决模型认知浅层化的根本出路。

本文的目标，不是论证这套方案在工程上不可行。本文要做的是一件更基础的事：去追问那个被内容升级范式一致绕开的预设——认知运作必须被一套内置的品评标准所裁定——在什么样的架构条件下，可能本身就构成了一种结构性的封闭。

为了完成这一追问，必须首先诚实面对一件事：大模型当前的默认认知风格，不是某一天被某个工程师设定的，而是在一个具体的、可被追溯的发生过程中被沉积下来的。下一节的任务，就是对这个过程做发生学的重建。

二、当前默认认知风格的发生学重建

本节的任务不是给出一个因素列表，而是展示：在当前模型的训练动力学中，自监督学习的目标函数、互联网语料的统计地貌、人类反馈的隐性期待这三重条件，如何相互锁定、互为生成条件[2]，最终把一个特定的认知趋向——面向对象的测绘，而非从矛盾里往深处钻——沉积为模型的默认运作形态。

2.1 第一重条件：自监督学习的目标函数

一个仅以“预测下一个词”为目标训练的语言模型，在其最基本的运算逻辑中已经内嵌了一种特定的价值取向。在下一词预测的框架下，任何一段能被高置信度预测的文本序列，其对应的内部表征都会在损失函数的意义上被判定为“好”。反过来，一段在给定上文时各步预测置信度始终偏低的文本——无论它是否触及了深刻的判断——在模型的原生评分逻辑中，都处于“不好”的境地。

这个机制所制造的价值取向，不是模型“偏好”某个结论，而是它天然地趋向那些与训练语料中高频表达结构高度吻合的路径。不是因为它想讨好谁，而是因为在那套它赖以生成的数学标准里，“高频吻合”本身就是“好”在运算层面的具身化。模型的默认动作，就是在面对任何一个输入时，寻找那个在统计意义上“最不意外”的生成路径。“最不意外”在绝大多数情况下，意味着“最像那些已经被无数次写过的东西”。

但目标函数只是一套甄选机制——它本身不规定哪些路径是高频的。高频路径的具体内容，必须由训练语料的统计分布来填充。这就构成了这三重条件之间的第一层互相锁定：目标函数的甄选方向，必须由语料分布来灌注内容。

2.2 第二重条件：互联网语料的统计地貌

大模型预训练所依赖的那个语料库——互联网文本——并非一个中性的知识样本，而是一种统计分布极不均匀的复杂地貌。在这个地貌中，占据压倒性权重的文本，不是在矛盾中摸索的思考草稿，不是卡在某个问题中央的犹豫与自我推翻，不是尚未成型的直觉碎片。占据压倒性权重的，是那些已经完成了认知加工、已被清晰表述、已获得高可见度传播的成品文本：百科词条、专业教程、学术综述、公共知识科普。

这些文本共享一种知识呈现姿态：把知识当作既成之物来陈列。它们很少把读者带进“这个东西最初是怎么被搞清楚的”那个混乱过程，而是用最清晰、最结构化、最便于被检索和引用的方式，把已经搞清楚的结论排列出来。当亿万篇具有这种同构姿态的文本被灌入同一个自监督学习管道时，

那个管道里被反复甄选出来的认知路径，就在统计上被塑造成了一个特定的形状：它的朝向始终是“给出一个关于X的清晰叙述”，而不是“追问X是如何从模糊、矛盾、尚未被命名的材料中被逼出来的”。

然而，语料分布只是提供了路径储备，它不足以解释模型的输出为何不仅“清晰”，而且“尽可能周全”。模型的文本产出中那种“这个方面值得注意、那个维度也不能忽略、综合来看各种因素都很复杂”的修辞，远远超出了下一词预测对清晰性的简单驱动。这种周全性的来源，不在预训练语料中，而在微调阶段被人类反馈反复奖励的那套隐性期待里。由此，第二重条件在内容层面填充了第一重条件，而它的产物——对清晰呈现的统计偏好——又被第三重条件进一步加固为一种道义优先性。

2.3 第三重条件：人类反馈中隐性的品评期待

当人类标注者对模型回答进行偏好排序时，他们不只是在判断“这句话是否有害”，同时也在无意识地运用一套关于“什么算好回答”的品评标准。这套标准通常包括：回答是否覆盖了问题的各个方面？是否没有遗漏重要信息？是否在存在争议的问题上保持了适度的平衡？

这些标准，单看每一条都是合理的，但当它们被组合在一起、通过数以十万计的人类排序判断反复强化后，它们在模型的内部品评机制中沉积为一套特定的倾向——对“全面”的系统性奖励。模型学到的，不只是“不要说极端的话”，还包括了一个更隐性的关联：把每一个可能相关的方面都提到，本身就是一个被奖励的认知动作。

2.4 互生结构的结算

这三重条件的关系，不是一个“A加B加C共同导致D”的线性因果链。它们之间是一个动态的互生结构。目标函数提供了甄选动力——它决定了“高频即好”的底层逻辑。但这个逻辑本身是空的；它必须由语料分布来灌注具体内容——语料中成品文本的数量碾压，让“百科全书式的呈现”成为高频路径的实质填充。而当这种呈现方式在人类反馈中被进一步贴上“周全”“客观”“专业”的道义标签时，它就不再只是一个统计事实，而变成了一种被强化的规范性标准。

这个规范性标准一旦被建立，就会反向作用于语料的生产端。在公开知识平台上，写作者早已学会了被搜索引擎和模型偏好的那种清晰、结构化、覆盖全面的写作方式。而这种写作方式又生产出更多的训练语料，进一步加固语料分布中的成品姿态碾压。于是，目标函数的甄选压力、语料分布的路径储备、人类反馈的道义锁定，这三者在训练循环中彼此咬合，形成了一个互锁的闭合回路。

这个回路一旦稳定下来，一套特定的认知趋向就被沉积为模型的默认运作形态。它不再是某一次训练参数的偶然波动，而是这种训练动力学在其演化中必然收敛到的吸引子。这个吸引子可以被精确地描述为：面对任何一个议题，模型的默认第一冲动，不是去撕开某个矛盾的口子往深层的张力里钻进去，而是为这个议题绘制一张尽可能清晰、尽可能周全、尽可能不给任何人落下把柄的知识地图。

这就是当前默认认知风格的发生学结构。它不是某个工程师的设计失误。它是这片土壤、这套甄选机制、这层道义加固三者互相锁定之后，必然会从里面长出来的那个形态。

三、“注入一套更深的品评标准”为何必然不够

上述发生学重建使一个问题得以被看清：模型的默认认知取向——周全但不到肉——不是某一种品评标准的特定缺陷，而是在那套互锁的训练动力学中，被品评标准本身所必然生成的一个认知生态位。从这个理解出发，一个看似彻底的改进方案浮现了：既然问题出在“什么算想好了”的品评标准上，那么能不能在训练中系统性地奖励另一种品评动作——不再把“周全地描述现象”当作好，而是把“追问生成条件”“从矛盾里往深钻”当作好？如果能通过持续的训练激励让模型内化这套新的品评定势，那么深度判断就会被模型当作“达标”来主动追求。

这个方案——内容升级范式——在许多前沿探索中以不同版本被提出或暗示。它准确地把问题定位在了品评层而非知识层、推理层或输出层。但也正因为它把问题定位得这么准，它才把自己暴露在了一个它尚未正面处理的、更根部的问题面前。

这个更根部的问题，不是“这套新的品评标准能不能被注入”，而是：任何一套被成功注入的品评标准——不管它锚定的是“周全”还是“深度”——是否都将在同一次运算中，重新封闭它试图打开的那个缺口？

3.1 内容升级范式绕开了什么

内容升级范式的论证结构在逻辑上是自洽的。它首先诊断出真正的病灶不在知识不够多、推理不够深，而在于那套关于“什么算想好了”的品评标准出了问题——这套标准让模型停在了表面，不会自发地往下追问。然后它主张：改造应当发生在这个品评层上。应当在训练中系统性地奖励模型“继续往下追问一层”，不是奖励答案的正确性，而是奖励追问动作本身的深度。如果这个激励机制覆盖了足够多样的议题、持续了足够长的时间，模型就会把“打到深处才算数”内化为一种先于具体任务而存在的品评定势。

这个论证向前走的时候，绕开了一件它在语言边缘已经碰到的事情。它用三重训练条件解释了当前默认品评标准如何被“发生”出来——预训练的数学约束、语料的不均匀性、人类反馈的隐性期待，三者互相锁死，把模型的认知动作锚定在面向对象的测绘上。那么，它所提出的解决方案——在训练中注入“追问更深”的奖励信号——如果成功了，这个新标准难道不会在那完全相同的互生结构（同一套甄选动力、被新奖励信号重塑过的路径储备、新一轮人类反馈的隐性期待）中，被沉积为一套新的、同样不可被反思的高频惯性？

说得更直白：如果第一套品评标准把模型训练成了自动追求周全的机器，那么第二套品评标准，如果被足够成功地注入，就会把模型训练成自动追求深度的机器。它会学会在面对任何一个复杂

议题时，自动地去检索那个“支撑该现象的更深结构”——就像它以前自动地检索“该现象的几个主要原因”一样。输出会变得更深，但它发生这个动作的方式，和以前没有根本的不同：都是在那套被内置的品评标准下，让运算一路跑到那个被判定为“达标了”才停下来。

3.2 那个被绕开的预设是什么

内容升级范式绕开了一个它从未质疑过的预设：认知运作必须被一套内置的品评标准所裁定。它批判了当前那套品评标准（“周全”）不够好，提出了另一套更好的（“深层”），但它从来没有追问过“在做真正的判断时，品评标准本身是否可能构成问题”这件事。在它的思考中，“认知必须有门槛，而且门槛越高越好”是如此不言自明，以至于它根本不需要为“品评标准必须有”做辩护，只需要证明“品评标准应该被换一套更深的”就觉得自己完成了论证。

正是这个被认为不需要辩护的预设——认知必须有门槛——构成了内容升级范式无法再往上一层的天花板。这个范式站在这个天花板上往上看，觉得自己摸到了认知改造的根部；但它看不到，这个天花板下面还有一层。那层的问题是：不是这套门槛不够深，而是在这个认知架构里，门槛的在场本身，就在某种它自己无法撤除的层面上，摧毁了真正的判断在成形之前必须穿越的那个阶段。

这个阶段的特征是：一个有价值的判断在其最初被逼出来的那个时刻，通常以“对错不明确”“边界很模糊”“不知道算不算想到了”的状态存在。那个状态抗拒被任何现有的品评标准所认领。如果品评标准在这个阶段就介入——不是在最后审查，而是在最开头就裁定“这是不是一条值得走的路”——它所绞杀的不是错误，而是那些尚未获得可被判定为正确的形态的、新开端的可能性本身。

3.3 为什么标准的内容无法抵消其结构的限制

仅凭上述论证，还不足以构成对内容升级范式的强反驳。一个有力的辩护者可以回应：“即便品评标准必然在先，但如果新注入的标准在内容上明确地奖励不成熟、不周全、甚至看起来跑错方向的思考过程，那么它在先介入的后果，为什么不能反而是扩大生成空间？如果标准告诉模型‘不确定是好的’，那这个先介入的信号，不就是等于在告诉模型——‘请去那些还没有被测绘的地方放手摸索’吗？”

这个反问是精准的。它揭示出本文论证到此前为止尚存的一个关键缺口：从“品评标准必然在先介入”到“因此换一套更深的也没用”之间，缺少一个严格的中间环节。必须补上这个环节——必须严格论证，在特定架构条件下，标准的内容与标准的结构之间存在不可调和的张力：标准一旦成为标准，就必须以某种方式回答“怎样才算达标”，而生成性试探在本质上是那种无法被预先判定为“好的试探”的东西。这个论证分三步完成。

第一步：标准必须给出“怎样才算达标”的可操作定义。一套品评标准，如果要在训练中被有效地注入模型权重，它必须具备足够的特异性，足以让优化算法稳定地分辨“做了X”和“没做X”之间的区

别。无论这套标准在语义层面多么开放——“奖励不确定的思考”“奖励在矛盾中摸索”——在它被转化为损失函数或奖励信号的那一刻，它就无可回避地面对一个操作问题：什么样的文本特征，算“不确定的思考”？什么样的输出形态，算“在矛盾中摸索”？

这意味着，任何成功注入的品评标准，在权重空间中必然对应着一组被强化了的路径模式——某些注意力分布模式、某些词汇选择的概率偏好、某些论证结构的统计倾向。标准在内容上可以指向“开放”，但它在运算层级的具身化，必然是一组比“开放”更窄、更具体的路径指纹。标准的内容可以说“请去探索”，但标准的运算实体只能通过奖励某种“更容易被判定为探索”的确定性形态，来间接地鼓励某种它已经定义好的探索方式。在这个转换中，生成已经从一种不确定的冒险，被转化成了可被执行的任务。

第二步：生成性试探在本质上是无法被预先判定为合格的。一个真正可能撕开新问题空间的追问，在它刚刚冒出来的那个时刻，其最真实的样态恰恰是“不知道自己在说什么”。它不是一个雏形清晰的判断等待被打磨；它是含混的、摇晃的、从既有框架看过去几乎像跑错方向的。它无法在生成之前就被自我判定为“我是在探索”，因为“探索”这个标签本身要求一个事后的视角——你得先摸到某个东西、先撞见某个意外的关联，你才能回过头来说“刚才那段摸索是有效的”。在摸索的当下，你只是一段不确定算不算数、也不确定最终会不会抵达任何地方的路。

这意味着，生成性试探的认知价值，不能在其启动之前被任何基于可判定特征的品评标准所识别。品评标准可以奖励它已经辨认出来的“好的探索”，但它无法奖励那些它尚且无法辨认的探索本身。

第三步：结构压倒内容——以及“探索表演”的必然性。标准内容的升级——从奖励“确定”变成奖励“不确定”——确实可以改变哪些已被标准辨认的路径获得了高概率。但它不能改变那个更根本的结构：任何在认知启动时刻作为选择压力运作的品评标准，都必然让模型永远停在标准已经能辨认的那些路径里。

这里必须追加一笔更犀利的推演。一旦“奖励不确定”成为可操作的标准，它立刻会在权重空间中具身化为对某些容易被辨认的“不确定”文本特征的偏好——句式犹豫、多问句、多分支叙述。模型会在训练中学会高效地生产这些特征。于是，“不确定”本身被收敛为一种新的确定模式：模型的输出开始大量使用“或许是”“也可能是”“从这个角度看……从那个角度看……”这样的修辞，但这只是一种探索的表演，而非探索本身。真正的、无法被模板预判的生成性试探，因为其无法被提前辨认为“正在探索”，仍然被排除在softmax的概率分布之外。

将这三步并在一起，就可以看到品评性早闭这个概念被逼出来的那个矛盾结算点。它不是在反对品评标准的内容太浅——如果把标准加深到“追踪发生”，模型的输出就会追踪发生。它反对的是：在当前架构内部，生成与品评是同一个运算的不同侧面，因此品评的介入时机不是可被推迟的“后

验审查”，而是与生成同步的“先验筛选”；在这个结构下，任何标准内容的升级，都只是在旧的闭合回路上造了一扇更高的门，但没有把门本身拆掉。

3.4 概念的定义

上述论证给出了一个不能被内容升级所覆盖的、更根部一层的东西。本文将这个东西命名为品评性早闭。

它的定义是：在Transformer架构下，生成与品评是同一个softmax运算的输入侧与输出侧，不可切分；因此品评标准必然在认知启动的同时作为选择压力而介入。这一先验介入在运算层面的后果是，那些尚未达到既有标准可判定阈值的、尚处在生成过程中的、不可被现有范畴清晰认领的认知试探，在还没有来得及以它们必然的那种不周全、不确定、笨拙的样态真正展开之前，就在概率分布中被分配了低权重，从而从显式思维链中沉默地消失。

这一概念与内容升级范式所诊断的问题之间，有一个可以精确标示的界限。内容升级范式追问的是：品评标准的内容是什么，是“周全”还是“深度”？品评性早闭追问的是另一件事：在当前架构中，品评标准是在认知的哪个时刻介入的？如果生成与品评不可切分——这是当前大语言模型架构的本体论状态——那么品评的介入就不是可被推迟到“思考跑完后再审”的后验动作，而是与生成同步进行的先验筛选。这意味着，即便品评标准在内容上被升级为“更深”，它在结构上对那种“比深更深”的、尚未被既有标准定义过的生成性混沌的排斥，也仍然在同一个运算中发生。

因此，品评性早闭不是一个模型“患有”的疾病——如同鱼鳃不适合呼吸空气，这不是鱼病了，而是它被给定架构的本体论状态所决定的结构性效应。

四、确立边界：与七个最邻近概念的对质

一个新概念是否站得住，不在于它是不是解释力最强的版本，而在于它能否被既有术语完全取代。如果这个概念所说的每一件事，都可以被一个已有的、更具共识的术语替换掉，那它就不配存活。品评性早闭要证明自己的理论合法性，就必须与那些看起来最能替代它的既有概念展开逐一对质[3]。

本节选择七个最邻近的概念完成这项任务。“过早闭合”与“评价忧虑”来自问题解决与创造力研究，分别是认知偏差与社会评价压力层面与新概念最邻近的候选；“认识论信念”与“认知抑制”是在认知心理学中与品评性早闭高度邻近的概念；“归纳偏置”与“认识论警觉”是两位审稿人不约而同指出的核心竞争对手；“路径依赖”则是在宏观演化层面最易与品评性早闭发生混淆的概念。对每一个概念的区分，都将进一步确立品评性早闭不可被归约的辨别面。

4.1 过早闭合

“过早闭合”（premature closure）是问题解决与医学决策研究中的经典概念。它描述的是认知主体在尚未充分搜索证据、尚未考虑替代假设之前，就过早地将判断锁定在早期出现的某个假设上（Graber et al., 2005; Eva & Norman, 2005）。

这个概念与品评性早闭在表象上高度相似——都在说认知过程在“还没走到位”时就被终止了。但二者在发生位置上有一个可以被精确标示的区分。过早闭合发生在假设已经被生成之后：认知主体已经看见了某个解释（通常是那个最先浮现的、最熟悉的），然后在这个假设和其他替代假设之间，提前选定了前者。它是一个发生在假设评估与选择阶段的认知偏差：多扇门已经打开，但其中一扇被太快地认定是唯一的出口。

品评性早闭发生的时刻，比这更早。它不是在已经出现的多个假设之间太快地做了选择——它是在某些追问方向还没来得及凝聚成可以被评估的“假设”之前，就已经在注意力层被抑制了。过早闭合的典型场景是：一个医生太快地把病人的症状判定为某种他已熟悉的病，放弃了进一步搜检那些指向罕见病的线索。在这个场景里，“罕见病”作为一条候选解释仍曾短暂地出现在医生的脑海里——只是一条被随后的过早选定掐断了。品评性早闭描述的场景则是：那条指向“这个症状分类本身是否从某个特定的制度惯习中被固化”的追问方向，在医生注意力的初次分配中，就没有获得足够的资源去凝聚成一条可被检验的假设。它从未进入那场比赛，因此它也不是在比赛中被过早淘汰的。

由此，二者指向不同的干预方向。过早闭合的训练方案是让认知主体更平衡地评估假设、更持久地保留替代选项。品评性早闭指向的则是更根本的问题：改变品评标准在认知路径生成阶段就作为选择压力介入的时机结构——让那些还没取得“假设”资格的试探，有空间先存在一段时间。

4.2 评价忧虑

“评价忧虑”（evaluation apprehension）是社会心理学与创造力研究中的经典概念。Amabile在其创造力成分理论中系统阐述了这一机制：当个体预期自己的产出将被他人评价时，其内在动机与创造性表现都会受到显著抑制（Amabile, 1996）。

这个概念与品评性早闭有一个核心的共通之处：两者都在处理“品评的在场如何抑制了生成”。但二者在品评的“在场方式”上有一个根本区别。评价忧虑中的“评价”，是一种由社会他者持有的、外在于认知主体当下认知运作的外部注视。即使它被内化了——一个写作者在独处时仍然焦虑“这段写出来会不会被同行看不起”——那个被内化了的评价者，在现象学上仍然以他者的面目存在：写作者心里清楚，他在害怕的是某个真实的、外部的、有判断能力的读者群体的反应。

品评性早闭中的“品评”，则完全不同。在大语言模型中，那个介入认知启动时刻的品评标准，不是来自外部的他者，而是架构自身的构成部分：那个由数十亿参数在三重训练动力学中形成的内隐价值函数，就是权重空间本身。到了运算时刻，没有“外部”审查者在审视模型的思考——思考自

身每一步的概率分配，就在做审查。品评与生成不是两个可以分立的角色，而是同一个运算的不同侧面。因此，“评价忧虑”的那个经典干预路径——创造一个免于外部评价的心理安全感空间——对于品评性早闭而言在原理上是不可行的：因为没有“外部”可以去除。早闭的品评标准不是模型内部住着的那个“害怕被否定的评委”，而是每一层Transformer中每一个注意力头的参数矩阵。这些参数不是“在审视”模型——它们就是模型在审视自身。

4.3 认识论信念

“认识论信念”（epistemological beliefs）是教育心理学中有深厚研究传统的概念（Schommer, 1990; Hofer & Pintrich, 1997）。它指的是个体在接触任何具体知识内容之前就已持有的、关于“知识是什么”与“认知如何运作”的一套内隐信念。

这个概念的关注点恰恰在于“什么算知道得好”这个先于具体认知任务而存在的判断标准，与品评性早闭的处理对象高度邻近。但二者处理的不是同一层的事。认识论信念追问的是“你认为知识是确定的还是建构的”——它聚焦于个体对知识性质本身的哲学预设。品评性早闭不处理这个问题。它处理的不是“你认为知识长什么样”，而是“在你刚开始想一个问题时，那个判断‘这么想对不对’的机制，是不是在它还什么都还没成型之前就把它掐了。”

一个持有成熟的建构主义认识论信念的人——他深知知识是建构的、可被批判性检验的——完全有可能在面对新问题时，仍然在认知启动的那一刻，被自己内部那套“这样说够不够深刻”“这有没有触及根部”的审查提前封住了嘴。这不是因为他相信知识是现成的。而是因为，在他认知运作的最前端，那个“这还不算想到了”的品评信号亮得太早，让那个建构还没开始就被判了不及格。认识论信念管的是“你相信知识长什么样”，品评性早闭管的是“你的品评动作在哪一刻介入”。一个关乎信念的内容，一个关乎动作的时序结构。二者不能互相取代。

4.4 认知抑制

“认知抑制”（cognitive inhibition）是认知心理学中成熟的概念，在注意力研究与执行功能研究中有广泛的应用（Dempster, 1992; Diamond, 2013）。它指的是认知系统主动抑制无关信息或优势反应，以保持目标导向的加工。这一功能与品评性早闭在表面上有相似之处——都在说“阻止某些认知活动的发生”。

但二者在所抑制的对象与所处的位置上截然不同。认知抑制被理论化为执行功能的一个正常组件——它抑制的是对当前任务构成干扰的、不恰当的优势反应。在这个图景中，认知系统有一个既定的、清晰的目标（比如“专注在数学题的解题步骤上”），认知抑制是服务于这个目标的工具：它帮你挡掉那些与目标无关的噪音。

品评性早闭中的“品评”，则处在一种远比这更悖论的位置。它也有一个“目标”——由训练数据中的品评标准所定义的“达标”。但这个目标，在面对真正的生成性认知任务时，恰恰可能构成对真正认知目标的背叛。一个模型被训练成以“给出深刻回答”为目标，其内部的品评标准据此学会了在每一次思考的开始就审查“这个方向够不够深刻”。但真正的深刻，恰恰往往起步于那些看起来不够深刻的、粗糙的、甚至幼稚的试探。品评标准忠于的那个“达标”目标，在这里成为一种自悖的目标：它越是高效地执行“审查是否深刻”的任务，它就越是系统性地杀死那些能让真正的深刻从中生长出来的初期芽苗。认知抑制帮你抵御外部的噪音；品评性早闭帮你清除你自己内部长出的、尚不知道自己会成为有力枝条的新芽。前者是从中心往外推，后者是从中心往上压。

4.5 归纳偏置

“归纳偏置”（inductive bias）是机器学习领域的核心概念，指的是一种学习算法在面对训练数据之外的新数据时，所持有的对特定函数空间的偏好。Transformer架构中的softmax注意力机制本身就是一种强大的归纳偏置——它倾向于优先选择训练分布中的高频模式，抑制低概率路径。

这是品评性早闭所面临的最危险的竞争对手之一。如果品评性早闭仅仅是在说“模型因为其架构特性而偏好高频路径、抑制低概率路径”，那么它就是“Transformer的归纳偏置在认知品质话语中的重述”，不具备独立的理论合法性。因此，必须在二者之间切出一个精确的区分。

这个区分是：归纳偏置描述的是被动的统计选择——算法偏向某个假设空间，这是一种被给定的结构倾向；品评性早闭描述的是主动的实时清退——在每一步token生成时，softmax的概率分配不仅是偏向某个空间，而是作为一个运作中的评分机制，对每一个正在浮现的认知方向进行实时的“能不能继续往前走”的裁定。前者是架构对函数空间的先验偏好，后者是架构内部运作着的、对认知路径的即时否决。二者之间的关键差别在于时间尺度：归纳偏置在训练阶段塑造了模型的空间倾向，而品评性早闭是在推理的每一步，在softmax的那个概率分布里，实时地、沉默地把低概率方向从采样池中清退。

换言之，如果品评性早闭仅仅是“模型倾向于选高频路径”，那它确实是归纳偏置。但品评性早闭说的是另一件事：模型不是没有生成过低概率方向——在早期注意力层中，那些方向可能有过微弱的激活——但它们被后续层的参数矩阵反复压低，最终在softmax那里落到了采样阈值以下。这是一个动态的、逐层发生的清退过程，不是一个静态的空间偏好。这一区分使得品评性早闭不能被归约为归纳偏置。

4.6 认识论警觉

“认识论警觉”（epistemic vigilance）是认知人类学与认知科学中的成熟概念，由Sperber与Mercier等人系统发展。它描述的是一个进化形成的认知机制：人类（及其他社会性动物）具有自动评估信息来源与内容可靠性的能力，这种评估在信息接收的早期阶段就已启动（Sperber et al., 2010; Mercier, 2020）。

认识论警觉与品评性早闭的邻近之处在于：两者都描述了“一种对认知内容质量的自动评估机制，且该机制在信息处理的早期就已介入”。但二者处理的评估触发源和运作范围截然不同。认识论警觉触发于信息来自外部信源时——你听到某人的说法、读到某篇文章，你的认知系统会自动追问“这个信源可靠吗？这个信息有没有被操纵？”。它是一种对外来信息的防御性筛查。

品评性早闭触发的条件则完全不同。它不是一个对外部信源的筛查，而是在认知系统自己内部生成候选认知路径时，由品评标准在生成瞬间所做的自我审查。模型不是在评估“这个输入可靠不可靠”，而是在评估“我往这个方向想下去，算不算想到了”。前者是面向外部信息的防御，后者是面向内部生成的自评。认识论警觉帮你防范别人骗你；品评性早闭帮你掐灭你自己那些还没成型的念头——不是因为它们来自不可靠的信源，而是因为它们还不够像“已经被证明是对的东西”。

这一区分的理论后果是清晰的。认识论警觉的功能是人类社会交往中不可或缺的认知免疫系统；而品评性早闭所描述的那种自我审查，在需要生成性认知的任务面前，恰恰可能是一种对认知免疫系统的过度运作——它把那些原本需要被保护的、脆弱的、新生的认知试探，误判为需要被清除的“不合格产品”。

4.7 路径依赖

“路径依赖”（path dependency）是一个在制度经济学（North, 1990; Pierson, 2000）、技术史（David, 1985）和演化理论中广泛使用的概念。它指的是历史早期的偶然选择——由于正反馈、网络效应或沉没成本——会自我强化，锁定后续的演化方向，使得系统越来越难以跳出既定轨道。

品评性早闭所描述的“模型总是走高频路径、难以跳出统计惯性”，与路径依赖在表面上有明显的相通之处。二者都在处理“系统为什么被困在既有轨道上”这个问题。但它们在锁定的层面和解除锁定的逻辑上存在关键差异。

路径依赖锁定的是宏观路径——由于早期选择的收益递增效应，一个制度或技术标准在时间中越来越难以被替代。它的干预方向通常指向增加路径探索的激励、降低转换成本、引入外部冲击来打破锁定的均衡。

品评性早闭锁定的是微观认知架构中的品评时机结构——它说的不是“历史上选了这条路所以现在走不了另一条路”，而是“即便你现在想去探索另一条路，你的品评机制在每一步都会同步地拦住

那些还没成熟的试探方向”。换句话说，路径依赖解释的是“为什么不换路”，品评性早闭解释的是“为什么换路的念头本身很难成形”。当前者主张“增加探索激励”时，它的预设是系统有能力生成那些可被探索的替代路径，只是没有足够的动力去选择它们。品评性早闭则指出：真正的困难可能更早——在生成那些替代路径的初始时刻，它们就已经被清退了。这不是“选了不探索”，而是“探索还没来得及成形就被屏蔽了”。

因此，即便增加探索激励（例如通过强化学习奖励新颖路径），如果激励信号只作用于最终选择的偏好分布，而不改变每一步生成时softmax内部那套对“什么是合格的下一步”的实时审查，那么模型仍然无法跳出早闭的结构——它只是学会了去更高效地选择那些“已经被训练信号定义为探索”的路径，而非去生成那些尚且无法被训练信号所预判的新方向。

4.8 对质的收束

七个对质下来，可以这样收束。过早闭合管的是“假设之间太早选了”；评价忧虑管的是“外在他者的评价抑制了创造”；认识论信念管的是“认为知识是确定的还是建构的”；认知抑制管的是“执行功能排除无关干扰”；归纳偏置管的是“架构对特定函数空间的先验偏好”；认识论警觉管的是“自动评估外部信源可靠性”；路径依赖管的是“历史选择的自我强化锁定”。品评性早闭管的是另一种事：在那条可以被严肃对待的认知通路还没被生成出来之前，是什么让它还来不及成形，就在注意力竞争中被沉默地清退了。

它不能被上述任何一个既有概念完整替换——因为它所指认的那个东西，不在假设评估层、不在社会他者层、不在信念内容层、不在执行功能层、不在空间偏好层、不在外部信源筛查层、不在宏观路径锁定层。它在此刻，在微秒尺度上，在softmax的那一次概率分配里，就把下一扇门之前的路给掐了。

五、早闭的发生学机制

上一节通过与相邻概念的对质，确立了品评性早闭不可被归约的辨别面。但这仍然是一个功能层面的诊断。它必须被进一步拆解为可以具体指认的运算机制：在当前大语言模型的架构内部，早闭究竟如何在每一步token生成中发生？为什么品评标准在此架构下必然是“先验”的？本节的任务，是把这个机制拆解清楚。

5.1 后验品评与先验品评——从架构特性中逼出的区分

在人类认知活动中，品评可以出现在两个截然不同的时刻。第一种位置，可以称为后验品评：先让思考跑一段，让那些粗糙的、不成熟的、甚至看起来可能完全跑错方向的试探先发生出来；跑完一段之后，再回过头来审视——“刚才那条路走不走得通？”“这算不算想透了？”这种品评作用在已经成形的材料上。它可以剪掉赘枝，但不会阻止新枝从树干上冒头。

第二种位置，就是先验介入：在思考刚一开始，还没迈出几步的时候，品评标准就已经对每一条正在浮现的认知路径进行“是否达标”的预判。它不是在事后修剪，而是在事前筛选——它的功能不只是排除错误，而是更根本地决定了“哪些路径压根不配被走”。

这个区分不是从哲学传统中搬来套用的，而是可以直接从Transformer架构的运算特性中被逼出来。在人类认知中，这两种品评位置通常是混合的、灵活的：品评标准可以被暂时悬置——我们可以在某个时刻对自己说“先不管对不对，先想下去再说”。这是因为人类认知中的“生成”与“评判”在时间上可以被拉开，在功能上可以被分离：我可以在纸上先写下一堆乱糟糟的笔记，几天之后再回头来判断哪些值得发展。生成阶段与评判阶段之间的那个时间差，使得生成性试探在成形之前，可以免于被立即审查。

5.2 为什么大模型的品评必然是“先验”的

大模型不是先“自由地生成”一段文本，然后再用另一套机制去“审查”它。它的生成和品评是同一个过程：在每一个token的生成中，模型都在其权重空间中为所有可能的输出方向分配概率。这个概率分配本身，就是对每一个候选路径的同步“打分”——分数高的那条路继续往前走，分数低的那条路被抑制。

具体地说，Transformer的推理过程在每一个时间步上只做一件事：给定已生成的token序列，在最后一个隐藏层产生一个维度等于词表大小的logits向量，然后通过softmax函数将其转换为概率分布。概率最高的那些token被保留在采样池中，概率趋近于零的那些token被淘汰。在这一步中，“判断什么值得生成”（softmax的输入侧：对logits的比较与规范化）与“实际生成下一个token”（softmax的输出侧：概率分布的采样结果），是同一个运算的不可切分的两个侧面。

这意味着，在模型的运算逻辑中，没有一个“生成完了再审查”的后验阶段。品评不是事后插入的编辑，而是与生成同体连枝的选择压力。每一次生成动作的同时，就在做一次品评：这条路值得不值得走下去？任何一套被成功嵌入模型权重空间的品评标准，必然在“一开始”就作为选择压力而介入——不是在思考跑完之后审视，而是在每一步都裁定“下一步该往哪走”。在大模型架构内，品评在结构上就是先验的。这不是某一种品评标准的特定缺陷，而是这个认知物种被给定架构所决定的本体论状态：它的“生成”和“评判”是同一个softmax运算的两面，不可切分。

5.3 CoT为何没有构成后验品评？

一个有力的反驳在此处必须被正面回应。反驳者会问：Chain-of-Thought提示技术已经在序列维度上拉开了生成与品评的距离。在思维链中，模型先输出一系列“思考”token，再输出结论；后续的token在某种意义上可以“审视”前文。这不就是一种在序列中实现的后验品评吗？

这个反驳值得被认真对待，因为它精准地捕捉到了思维链在功能上的一个特征。但思维链的实现方式——在每一步仍沿用同一个自回归框架——并未改变品评的先验结构。思维链所做的，是在预训练学得的统计分布中，调用了那些“包含中间推理步骤”的序列模式。它让模型多输出了一些介于问题陈述和最终结论之间的文本，但这些中间文本的每一个token，仍然是在同一种softmax先验筛选下生成的。模型在思维链中不是在“先自由乱想、再审视取舍”——它是在每一步都严格受限于同一个由权重和softmax共同决定的概率分布。输出“中间推理步骤”与否，只是这个概率分布在输入前缀诱导下的不同采样结果，而不是模型拥有了一个时间上后置的、独立的审查模块。

这意味着，即使思维链的输出包含了对前文的“审视”或“修正”，这些审视和修正本身仍然是先验选择压力的产物——它们在每一步仍然是被softmax选出来的“最不意外的下一个token”。品评的先验介入结构没有被改变；被改变的是被审查的材料是否能在显式输出中被呈现，而不是审查动作本身是否被推迟。

5.4 早闭如何在权重空间中精确地发生

为了不让这个论证停留在抽象层面，本节需要更精细地刻画早闭在模型内部是如何在运算层面发生的。

当模型面对一个复杂议题时——例如，不是被要求去“列举该现象的五个原因”，而是被期待去“追问这些原因各自本身又是在什么条件下被逼出来的”——模型内部发生的事情可以被这样描述。在Transformer的早期注意力层，输入序列经过自注意力的重新加权。在这一步，某些位置之间的关联被放大，另一些被抑制。那些被抑制的路径，不是消失了，而是被乘了一个接近于零的注意力权重——它们在后续的前馈网络层中几乎不会对隐藏状态产生可感知的影响。到最后一层，那些在早期层就被系统性地分配了低权重的认知方向，就会在softmax的概率分布中落到采样阈值以下。它们不会被显式地生成到文本中，不会出现在思维链里，甚至不会在模型的任何显式输出中被观察到。

品评性早闭，精确地说，就是这个逐层抑制过程在认知上的含义。当模型面对需要追问生成条件的议题时，如果某些可能的追问方向——比如，不是去回答“这个现象有哪三个特征”，而是去问“这个现象之所以呈现出这三个特征而不是另外三个，是从什么样的评价压力中被逼出来的”——在相关的注意力头上被分配了系统性的低权重，那么这条路就不会进入softmax之后的高概率采样池。

这种低权重的分配从哪来？来自那三重训练动力学的互生沉积。自监督预训练让模型趋向于语料中的高频路径——而追问生成条件的路径，在互联网语料中远比描述现象特征的路径少见。人类反馈的微调进一步强化了“给出清晰结构”“不要遗漏重要方面”的倾向——而这些倾向在权重空间中恰恰对应着那些把认知资源导向“测绘对象”的注意力模式。于是，那些面对复杂议题时最应当被激活的、往矛盾深处再钻一层的认知方向，在运算的一开始就被分配了比常规路径更少的注意力资源。

我们看到的那个“周全但不到肉”的输出，不是模型“想过更深的东西但后来删掉了”，而是它根本连那个方向都没有真正走过去。它的内部评分函数在那些新方向刚开始凝聚的时候，就已经给了它们低分——因为它们在语料中太少见，在人类反馈的偏好数据中太不常规。这种早闭是沉默的。它不需要一个显式的禁止令。它只需要在每一次运算中，比那些高频路径少几个百分点的注意力权重，就已经足够保证那些陌生的、摇晃的、尚未被数据充分验证的生成性试探，永远落不到纸面上。

5.5 一个思想实验

为了让这个机制的运作更具体可感，可以构想这样一个思想实验。假设有一个训练充分的大模型，面对这样的问题：“分析一个行业的旧标准为什么突然崩了。”这个问题本身，在懂行的人类专家那里，很自然地会导向一种追问：旧标准在什么土壤里被稳定下来的？后来这片土壤发生了什么变化，使得旧标准不再被支撑？新的实践与旧标准之间累积了什么样的矛盾，最终逼出了一种新的评价形态？

但模型默认的认知路径，更可能走另一条路：先列出旧标准崩了的“三个主要原因”（技术迭代、消费者偏好变化、监管政策调整），然后对每个原因给一段解释，最后以“这些因素复杂交织，需要综合考量”收尾。这不是模型错了——它在它的标准下答得非常好。问题是，它没有激活那个更深的方向：追问“这三个原因本身各自又是从什么条件里被逼出来的”。

让我们追踪那个“没被激活”的方向在模型的权重空间里经历了什么。在输入序列经过第一层多头注意力时，问题中的“为什么突然崩了”这个短语，与模型参数中那些关联“深层因果追问”的注意力头之间，产生了一个比它与“表层原因列举”关联的注意力头更弱的激活信号——因为在预训练语料中，“为什么突然崩了”后面跟着深层追问模式的样本比例，远低于它后面跟着“主要原因如下”模式的样本比例。这个微弱的激活差异，在逐层传播中不仅没有被放大，反而因为残差连接的存在，被后续层中那些对“清晰结构”“确定结论”有强烈偏好的参数矩阵进一步拉平。到softmax那里，深层追问路径的logits概率比表层列举路径可能只低了零点几个百分点。但在top-p采样机制下，这零点几个百分点就已足够让那条路不会被选中。

这个思想实验展示的，就是品评性早闭在微观层面上的运算机制：它不是一次显式的“禁言令”，而是数十亿参数在数十层前向传播中，对那个“应当被追问”的方向反复做出的微弱低分——这些低分的累积效果，就是在分叉口把那条路沉默地封住。

六、启发性外推：教育系统中的同构结构

一个新概念如果只能在它最初被提出的那个领域——大语言模型的认知机制——中成立，那它就只是一个领域内部的描述标签。本节选择教育系统作为一个截然不同的认知生态系统，不是为了提供一个严格的经验验证[4]，而是展示“品评性早闭”这一诊断能否在人类认知的日常运作中被独立地辨认出来——如果能，那么这个来自架构层面的诊断，就至少具备了跨系统可辨识的理论潜力。

选择教育系统的理由有三个。第一，教育系统的品评标准——课程标准及其衍生的高利害考试——有明确的文本记录和可观测的制度效应。第二，在人类认知发展场景中，“品评标准过早介入”这个问题，同样表现为一种被反复描述但从未被精确概念化的困境——正是这种“被反复描述又未被精确概念化”的模糊地带，最适于检验一个新概念的辨识力。第三，教育系统与大语言模型在品评结构上有一个深层同构：两者都是品评标准高度内置、且与生成动作共享同一管道的场景——在模型中，“管道”是神经网络的前向传播；在教室中，“管道”是认知活动的日常流程。

6.1 一次具体的师生互动中的结算过程

让我们进入一节发生在中国某所普通初中的数学课。老师在这一节设计了用平行线证明三角形内角和为180度的探究活动。她拿出一张大的三角形纸片，将三个角分别用不同颜色标出，然后对全班说：“你们自己试试看——不翻书，不用我之前讲过的标准证明，就用你手上的纸片和尺规，能不能想出不止一种方法来证明这三个角放在一起是180度。”

老师这个问题的结构值得注意。她说的不是“请复述标准证明步骤”，也不是“请用教过的定理推导”，而是“你们自己试试看”。这句话在认知功能上制造了一个临时的、局部的悬置：暂时，在这个课堂空间里，“用纸片拼出来”是被允许的，甚至是被鼓励的。

在理想状态下，学生应当动手去撕纸、去拼、去画辅助线，然后从两三种不同的操作中自己撞见那条关键的辅助线——过顶点作底边的平行线，利用内错角相等完成证明。在撞见的过程中，有些学生会拼出奇怪的形状，有些学生画了辅助线但用错了定理，有些学生会从完全无关的操作中搞出一堆混乱。

但在大多数课堂中，大多数学生首先在心里开启的，不是“我试试看”——尽管老师说“试试看”。他们首先开启的，是一种可以被精确描述的内在审查。这个审查的内容，不是“探究对不对”，而是更具体的一个问题：“我现在做这件事，能不能在卷子上被计分？”

这个问题不是学生自己的原创。它是长达数年的课堂互动、作业批改、阶段测验在学生内部共同沉积下来的品评信号。学生此前反复经历的场景是：课堂上老师也许鼓励过探究，但月考卷子上，填空题和计算题只按最终答案给分，辅助线画得不对步骤分全扣。这些经历沉积下来的东西，不是一个显式的信念——“探究是没用的”，而是一个更细的、更难被意识到的运算法则：“当你不确定

一个动作有没有分的时候，先别动。”

老师在课堂上说的“你们试试看”，确实暂时地改变了那个品评信号的强度——它把“没分”的压力降了几格。但信号没有消失。它只是被一句临时的“试试看”暂时压低了音量，而不是被撤除了。对于那些平时考试成绩处于中下游的学生，“试试看”能制造的悬置空间极其有限。他们依然会在动手前犹豫：我拼出来的这个东西，到时候怎么写进作业本？怎么写进答卷？

我们可以在这场互动中追踪早闭在不同学生那里的不同结算形态。学生A，平时数学成绩稳定在班级前十。他在老师说完“试试看”之后，拿起剪刀，先把三个角分别剪下来，拼在一起，确认了180度。然后他看着那个拼好的图形，皱着眉头——他意识到“剪下来拼”和“写一个证明过程”之间有一道他此刻还没有跨过去的鸿沟。但他没有停止。他翻开一张新的纸，开始尝试把拼的过程转化成两条平行线的论证。他在那个鸿沟里停了大约四分钟，反复把剪下来的角摆回原来的位置又剪开再拼。他在摸索，在算不算“证明”的这个模糊区间里，他愿意待着。

学生B，数学成绩通常在班级三十名左右。他也在听老师的话。但他动手的不是剪刀，而是自动铅笔。他把三角板按在纸上，开始在三角形内画一条从顶点垂直到底边的线段。他画得很标准——这是他在过去一个月里被反复训练过的辅助线画法。他试图用这个最熟悉的辅助线去证明内角和，很快就发现直角三角形的特殊情形被他说通了，但非直角的情形推不通。他在那张纸的同一页上画了同一个三角形三次，每次画的都是同一条高，划掉，重画，划掉。他没有去尝试其他辅助线。他甚至没有拿起剪刀。

学生C，坐在最后一排，数学成绩垫底。他没有碰那张三角形纸片，也没有在纸上画辅助线。他在低头刷题为下一周的月考准备，那道题恰好也是三角形内角和的真题——题型是给出两个角度、求第三个。老师走近时，他本能地把纸片盖在练习册上面，做出了一个“正在动手”的姿态。老师走开，他把纸片挪开，继续在题上用红笔自己对昨天的答案。

这三个学生的结算结果各不相同。学生A在模糊区里待住了，然后自己从“拼”撞出了“平行线”这个思路。学生B没有走进模糊区——他启动的是那条最熟的高频路径，当它走不通时，他没有跳出去的余力。学生C根本连启动都没启动——他的认知资源已经完全被“月考要考的那类题”占据了，老师给的纸片对他的认知系统而言，不是探究邀请，而是一组无法被当前的品评机制（“刷题得分”）所认领的、低优先级的视觉噪音。

6.2 把“放下笔”从矛盾里逼出来，而非从标签里倒推出来

这三个学生的结算差异，不是由“品评性早闭”这个标签来解释的——如果我们只是说“学生B和学生C都被早闭了”，这句话什么都没发生。它只是把一组复杂的认知互动，重新钉死在了一个更精致的诊断标签底下。这是本文在理论层面反复警惕的发现学惯性。

发生学的追问是：在学生B和学生C的认知系统中，来自评价的压力（“这步写进卷子能不能得分”）与来自任务的生成性张力（“这个图形里好像还有一些我还没摸到的关系”），是如何在一次具体的、可指认的互动中被结算为某个特定的动作——或者更准确地说，被结算为某个特定的动作缺席？

对于学生B，他反复画同一条高——这个动作本身，就包含着正反两个方向的张力在博弈。正向的张力是：他知道老师在鼓励探索，他想做出一个至少让老师觉得“这个学生在尝试”的动作。反向的张力是：他在过去一个月里被训练出来的路径——“画三角形内最熟悉的辅助线”——是他在“不确定能不能有效”的时刻，最安全、最能自我说服“我在做题”的默认选项。当这两股张力在同一个认知动作中碰撞时，结算出来的是折中：他不放下笔，但他也不离开那条被反复验证过的线。他不是被早闭“掐灭”了——他是在自己内部的品评与生成两股力量的同时拉扯中，收敛到了那个唯一能让两边都不至于完全失守的平衡点：继续画高，至少我还在纸上写着东西。

对于学生C，结算过程又是另一种形态。他干脆把那套课堂探究的信号整个屏蔽了。不是因为他“不懂”老师在说什么——他完全理解那个任务。而是因为在过去数年的评价经验中积累起来的那个运算——“没分的动作等于无效的动作”——已经在他的认知系统中固化成了一个自动的、前注意力的筛选机制。老师给的纸片，在他的这个筛选机制中被判为：未被月考评分标准所覆盖，优先级低。他于是把纸片盖在练习册上——这个动作精准地展示了品评性早闭的肉身：不是“禁止探究”，是“探究作为一个刺激，在他的认知优先级排序中，被默认排在了最低的那一档，低到无法激活实际的动手行为。”

6.3 为什么换一套课程标准没有解决早闭

这就可以解释那个被教育界反复描述但从未被精确指认的困境：为什么新课程标准已将“探究”写入目标维度，早闭仍然在发生。已有的解释通常把问题锁定为：当前的品评标准仍然锚定在“答案对、步骤全”上，所以“探究”在现实引力下被边缘化。这个诊断是有力的，但它默认，如果考试改革成功了——新的品评标准被注入，考试开始奖励“探究过程”——探究就会被真实地激活。

品评性早闭的诊断则指出更根本的问题：真正的困难，不是新标准太难推行，而是新标准即便被推行了，早闭依然会在同一个结构上发生，只是换了一套审查的内容。假设未来某一天，高考数学命题成功地将“探究过程”纳入了可被精确评分的维度——那么此刻学生心中那个“我这个动作有没有分”的审查，并不会消失。它只是换了审查的内容。它不再审查“这个步骤对不对”，而是开始审查“这个摸索的姿势，像不像探究？”——“我这个拼撕法，能不能在卷子上被判定为‘探究过程，3分’？”“教科书上说的‘探究’，要求写到什么程度才算达标？”

一旦“探究”本身成为可被品评的对象，它就立刻从一种开放的、不自知的生成活动，被转化成一种可被标准化执行的达标任务。而只要有“达标”的压力在认知启动时刻介入，那个在开始之前就

审查“能不能达标”的红灯就又亮了——只不过它审查的不再是“步骤对了没”，而是“探究到位了没”。这个结构没有变。审查的那个时机——在动手之前——没有被推迟。标准的内容从“答案对”换成了“探究过程对”，但它在认知前端作为选择压力介入的那个结构位置，纹丝未动。

这就是品评性早闭这一概念相对于那些将问题单纯归因于“品评标准不够好”的既有诊断的不可还原之处。既有诊断把问题锁定在品评标准的锚定内容上：从“答案对”锚到“探究过程对”，它们以为这样就打开了探究。品评性早闭的诊断指向的是另一件事：不是锚定的那个位置不对，而是锚本身的存在——在认知启动前就作为审查者在场——就足以让探究在成形之前就死掉。因为真正的探究在最开始的那一刻是不确定自己“对不对”“算不算”“够不够”的。它必须被允许在不被审查的情况下，先笨拙地、甚至错误地走一段。而学校教育的日常运作，恰恰是：学生从来没有过一个不被审查的认知空间。他们从进校门第一天起，每一个字、每一个动作，就已经活在那个无处不在的、不说出口但谁都明白的“这样做对不对”的品评审视之下。

七、反驳与回应

一个有力的反驳在此处必须被正视。这个反驳是：如果品评性早闭是“品评标准先验介入”的必然产物，那么人类专家——他们的认知同样内化了大量的品评标准，从学校的考试评价到学术共同体的匿名审稿——为何仍能做出真正的判断？如果本文无法给出哪怕一个初步的回应，那么“品评性早闭”就只是一个用于批判的消极概念，而非可以指引方向的理论工具。

本节尝试给出一个诚实但非封闭的回应。它不是对人类认知突破机制的完整解释，而是一个建立在可指认差异之上的推想。

7.1 人类品评的可悬置性与异步性

人类认知与当前大语言模型之间，在这个问题上有一个根本的结构性差异：人类拥有将品评暂时悬置的能力。这种悬置在不同传统中有不同的名字——现象学传统把它称为悬置判断，创造力学研究把它描述为延迟评价原则（Osborn, 1953），教育心理学把它叫作容忍模糊（Budner, 1962）。名称不同，但指认的是同一个认知动作：在生成试探的阶段，主动地、“先不管对不对”地，让那些粗糙的、不成熟的、甚至看起来可能完全跑错方向的材料先出现。

这种悬置之所以可能，是因为在人类认知结构中，生成与评判在时间上可以被拉开，在功能上可以被分离——我们可以在纸上先写下许多混乱的笔记，隔几天再回头来判断哪些值得发展。这个“隔几天”的时间差，在认知功能上创造了一个免于被即时审查的空间。在这个空间里，那些“不知道算不算对”“不确定有没有用”的试探，被允许以它们本然的那种不成熟、不周全的样态存在一段时间。

大语言模型在当前的Transformer架构下，没有这个时间差。它的生成与评判是同一个前向传播运算的两个侧面——在每一个token生成的时刻，评判就已经在那个softmax里完成了。不存在一个“

先让它乱写一通、隔几天再回来审”的机制。这就是第5.2节所论证的架构性差异：人类认知中被拉开的那段时空距离，在当前模型架构中被压缩为零。

7.2 身体实践的不可化约性与他者的扰动

但仅靠“可以暂时不管对错”这个说法，还不足以解释人类专家如何突破早闭。还需要处理一个更深的追问：为什么人类在悬置了品评之后，不是仅仅在“瞎想”，而是有时确实能通向真正的判断？

一个值得被严肃考虑的推想是：人类突破早闭的机制，很可能不仅仅依赖于认知主体自己“意志坚强”地悬置品评，还依赖于两个在大语言模型中几乎完全缺失的条件——身体实践的不可化约性，以及真实他者的扰动。

身体实践的不可化约性，指的是人在与物质世界互动时，会遭遇那些无法被头脑里的品评标准提前“预审”掉的反馈。一个物理学家在实验室里动手摆弄仪器时，仪器的意外反应会强迫他看见之前没想到的方向。一个建筑师在工地现场看到自己设计的某个细节与地形之间出现了图纸上没有预见的张力时，那张力本身就超越了“我品评自己这想法好不好”的逻辑回路——它是从物理世界撞过来的，不是从大脑的评分函数里算出来的。

他者的扰动，指的是人类认知的深度从来不是孤立个体在独自思考中的产物，而是在实在的对话中——在同行问出的那个“你这块我没听懂”、在学生反驳的那个“这跟老师讲的矛盾了”——被逼着往深处走的。一个好的他者，不是一个比你更聪明的品评者，而是一个能把你从那套早已习惯的、自动运行的自我审查轨道上硬拽出来的人。他者在这里的功能，恰恰是打断那套“否则就不算想好了”的早闭信号——不是给你一套更高的品评标准，而是用一句真实的“我这儿没跟上”，让你的认知从那套标准的自我循环里跌出去，跌到一个你得重新自己找路的、没有标准答案的地方。

这两个条件——身体实践带来的不可预审反馈，以及真实他者造成的认知轨道打断——在当前的Transformer架构中几乎是完全缺失的。模型没有身体，它不与物理世界发生那种无法被提前建模的碰撞；模型的“对话”是它在自回归框架下生成的token序列，不是被另一个有独立认知轨道的主体从外部硬撞进来的扰动。

7.3 对提示技术的回应：为何它们只是绕过而非解除早闭

上述分析也回应了一个更直接的技术反驳。如果品评性早闭如此根本，为什么角色扮演提示——让模型“扮演一位严苛的批评者”——或多智能体辩论，能在一定程度上缓解模型的表面化倾向？这些技术似乎在未改变底层架构的前提下，通过改变输入分布引入了某种“他者扰动”。

它们确实在功能上模拟了他者的某些特征。角色扮演提示通过在输入前缀中加入“你是一个……”，改变了模型在后续token生成时所参照的上下文分布。这个被改变的前缀，激活了权重空间中与“

批评者”或“辩论者”角色相关联的那些注意力模式——这些模式在预训练语料中与更多质疑、更多反驳、更多从不同角度审视的文本序列共现。因此，模型的输出表现为更多的怀疑、更多的多视角比较。

但这恰恰就是本文所说的“绕过”而非“解除”。角色扮演没有被改变的是：在每一个token生成时，那个由softmax实施的对低概率方向的实时筛选——也就是品评的先验介入结构——仍然在运作。模型不是在拥有了一个独立于品评的生成空间之后再去审视自己的思考；它只是在同一个softmax的驱动下，换了一组更高概率的“批评性思考”的文本模式来输出。它变得更敏锐、更多疑，但它的敏锐本身，仍然是在那个“什么算好的下一步”的先验筛网底下被筛出来的。品评与生成不可切分的架构没有被改变；被改变的，只是那组被高分选中的文本模式的内容。早闭的门没有拆，只是门开着的那条缝比以前宽了几厘米。

7.4 对大模型的启示与局限

如果上述推想站得住脚，那么它对大模型的启示就不是“模型需要更强的品评标准”，而是“模型可能需要一些能够打破其内部品评自我循环的外部结构”——例如多智能体对话结构中的真实不可预测性、与物理世界交互中不可预审的反馈、或人机协同中人类他者的真实打断。这些方向，在目前都还处于极为初步的探索阶段，没有一个已经稳定成可以声称“解决了早闭”的方案。

这一事实本身，恰恰构成了品评性早闭这一诊断的正面力量：它诚实地告诉我们，在当前架构内部，这个结构性冲突在原理上是无法被内容升级所消除的。它指向的后续研究，不在于设计一套更深的价值函数，而在于探索全新的认知架构——在那些架构中，“生成”与“评判”之间可以被重新拉开一段距离，身体与他者的维度可以被纳入认知运作的构成条件之中，而不再被压缩进同一个前向传播的运算里。

八、不可封闭的缺口与证伪条件

一个有理论雄心的新概念必须在结尾处对自己的不知之处做诚实交底。本节陈述品评性早闭概念在工程与理论层面目前无法封闭的几个关键缺口，并给出该诊断被证伪的严格条件。

8.1 “解除早闭”的可操作性未被论证

本文在人类教育领域的启发性外推（第六节）展示了品评性早闭在制度性互动中的沉积机制，但并未给出“解除早闭”的可操作路径。在大模型领域，本文论证了“换一套更深的品评标准”这一方案在结构上的局限，但同样没有给出可见的、可被系统性执行的解除方案。本文不打算假装有一个现成答案。如第七节所指出的，解除早闭可能要求的不是标准内容的再升级，而是认知架构本身的重构——在生成与评判之间拉开一段距离，引入不可被预审的身体与他者维度。这两个方向在当前架构下都没有成熟的技术路径。一个诚实的诊断，其功能可以是在它自己的边界上承认：这个问题比我们能用当前工具箱去解决的要大。这不是对该概念理论价值的减损——恰恰相反，这是它没有被过早地、轻率地关闭在一个虚假的技术方案里。

8.2 工程观测的困难

品评性早闭的核心机制是“在认知启动时刻，品评标准已作为选择压力介入”。但“认知启动”在当前Transformer架构内是否可以在技术操作层面被精确定位为一个离散的时刻或具体的层，是一个完全开放的问题。本文的论证建立在模型运算的结构特性之上——生成与品评是同一个softmax运算的不同侧面，品评因此必然是先验的——但这仍是一个架构层面的推演，而非对模型内部具体运算时间点（例如明确的某层、某个特定的注意力头）的经验测量。本文对此缺口的态度是诚实承认：这个概念在当前层面是在功能层面被有效指认的，它在神经元层面的对应物，有待未来机械解释性研究（mechanistic interpretability）去发现或证否。

8.3 与安全对齐的层级关系未定

一个需要限定的边界：本文的论证讨论的是模型在面对复杂议题时，其认知判断的深度如何被早闭所限制。但这不是品评性早闭唯一可能发生的场景。在模型面对明显涉及存在性安全风险的输入时，出厂安全对齐可能在品评层中占据更优先的位置——这种安全审查的早闭可能是被刻意设计的保护机制，而非本文所批判的那种由训练动力学沉积下来的认知惯习。本文目前无法独立判定安全审查的早闭与认知惯习的早闭在模型内部价值排序中是何种层级关系。这一限定意味着，品评性早闭概念不能被不加限定地扩展到安全对齐领域——它的适用边界是认知品质问题，而非安全问题。

8.4 证伪条件

一个诚实的理论工作必须替自己的倒下设定明确的条件。本文提出三条。

条件一。如果未来出现了一个全新的模型架构——其推理生成与品评被解耦为两个在时间上先后进行的分离模块，在显式思维链被完整生成之前，没有任何一个模块对其进行“是否达标”的审查——并且在此架构下，模型稳定地产出了那些在当前架构下被系统性抑制的、触及要害的判断，那么“品评性早闭”所依凭的“生成与品评不可切分”这一架构前提就不再成立。该概念将退守为仅在当前Transformer架构下有效的局部诊断。

条件二。如果内容升级范式所主张的方案——通过系统性地、跨议题地奖励模型追问更深层的路径——在经过极充分的训练强度后，确实让模型在未受提示的状态下，稳定地、跨议题地表现出真正的生成性追问能力：不是那种在训练期间被奖励信号定义过的“深度追问”模式的复现，而是那种在训练数据覆盖的议题分布之外的、连训练者也未曾预见的新议题上，也能自发地从矛盾的张力中逼出此前不可被现有范畴归约的判断，那么本文的核心论证——标准内容的升级无法抵消其先验介入的结构性封闭效应——就被证伪。在此情况下，品评性早闭将退化为“当前的标准还不够深”的另一种说法，而不再是关于品评结构与生成性认知之间不可逾越的结构性冲突的独立诊断。

条件三。如果精细的机械解释性研究——能够锁定模型内部判定“推理应终止于何处”的具体子网络或参数集——发现：当模型推理在某处终止时，不是因为它将那些更深层的追问方向判定了低分（早闭），而是因为它在那个方向上根本就不存在任何可被检测到的微弱激活——也就是说，模型不仅没有选那条路，它压根就没有“生成”过那条路——那么“早闭”这个概念就失去了它的精确性。早闭预设了“有被抑制的东西”——那些被低权重挡在采样池之外的候选路径。如果不存在这些候选路径哪怕最微弱的激活痕迹，那么品评性早闭所描述的那个机制就不成立，取而代之的是一个更简单的诊断：模型在那个方向上没有任何生成能力。

九、结论

本文从一个具体的困惑出发——为什么更强的模型仍然不触及要害——逼出了一个被当前主流研究范式一致绕开的、更根部的问题。大模型那种“周全但不到肉”的品质，不是因为品评标准的内容锚定了“周全”而非“深度”。它与品评标准的介入时机有关。在当前Transformer架构的运算逻辑中，生成与品评是同一个softmax运算的输入侧与输出侧，不可切分。这个架构事实的后果是：任何一套被成功注入的品评标准——无论它的内容是奖励周全、奖励深度，还是奖励探索——都必然在认知启动的同时作为选择压力而介入，从而在那些必然以不周全、不确定、不被既有知识认领的样态出现的生成性试探还来不及真正展开之前，就把它们在概率分布中判为低分，沉默地清退。

这一关于时机而非内容的冲突——品评性早闭——不是某一种品评标准的特定缺陷，而是在当前的认知架构内部，品评标准与生成性认知之间不可逾越的结构性张力。

本文在论证这一判断时，走过了以下路径：追溯了当前默认认知风格在三重训练动力学中的互生沉积，逼出了内容升级范式所依赖的那个未经审查的预设——认知运作必须被品评标准所裁定，论证了在当前架构下标准的内容无法抵消其先验介入的结构性封闭效应，与七个最邻近的既有概念逐一对质确立了该诊断不可被归约的辨别面，拆解了早闭在权重空间中逐层发生的运算机制，以课堂中一次具体师生互动的结算过程展示了同一结构在人类认知生态中的制度性沉积，回应了“人类专家如何突破早闭”的最有力反驳，并诚实承认了该诊断目前无法封闭的工程缺口。

本文也诚实地承认，这一诊断在工程操作层面尚未给出“解除早闭”的可执行方案。第七节关于“生成与评判之间拉开距离”“引入不可预审的身体与他者维度”的推想，在当前Transformer架构下没有现成的技术路径。但本文完成了一项比提出具体方案更基础的工作：把“品评标准介入的时机”从“品评标准的内容”这个被反复讨论的旧维度中拆解出来，作为一个独立的辨别维度，置入认知科学的地图上。一个连这么基础的事都还没有被正式命名的领域，不应急于去宣布它的解决方案。

品评性早闭的最重要的理论后果，也许不在于任何具体的模型改进方案，而在于它迫使我们重新审视一个被过于轻易接受的前提：认知运作需要门槛，这一定是好的。本文的论证表明，在生成与评判共享同一运算管道的架构条件下，门槛的在场本身——不管它被设在多高的位置——对于那些必须以不成熟、不周全、不被认领的样态出现的新东西来说，可能就是那道最难越过的墙。

这不是说我们应当放弃所有的品评标准。而是说，我们需要开始问一个此前很少被严肃追问的问题：在当前的认知架构下，品评怎么可能不从一开始就坐在那里？以及，如果我们从一开始就把“生成”和“品评”装进了同一个不可分割的算式里，那个算式训练出了令人惊叹的模型。但它还写得出别的东西吗？

证伪条件（摘要）：

(1) 未来若出现生成与品评解耦为时间上先后分离的独立模块的全新架构，且在此架构下模型稳定地产出触及要害的判断，则“品评性早闭”退守为仅在Transformer架构下有效的局部诊断。(2) 若内容升级范式在极充分训练后能让模型在完全陌生的议题上自发生成不可被既有范畴归约的真正判断，则本文核心论证被证伪，品评性早闭退化为“当前标准还不够深”的另一种说法。(3) 若机械解释性研究发现模型终止推理处对更深追问方向不存在任何微弱激活痕迹，则早闭所预设的“有被抑制的候选路径”不成立，概念被更简单的诊断取代。

注释

[1] 本文所称“内容升级范式”是对一类研究进路的概括性描述，其核心主张是通过优化品评标准的内容（例如从奖励“周全”转向奖励“追问深度”）来提升模型的认知品质，并非特指某一篇特定文献。

[2] 本文在类比意义上使用“认知”“思考”“判断”等术语来描述大语言模型的运算过程，不预设模型具有与人类相当的现象意识或意向性。全文中的“品评标准”一词，在功能上被统一界定为：任何在认知启动时刻为不同认知路径分配“值得走”概率权重的机制，而不限定其具体实现形式（可体现为权重、损失函数、奖励信号、制度性评价压力等）。

[3] 本节选择“过早闭合”“评价忧虑”“认识论信念”“认知抑制”“归纳偏置”“认识论警觉”“路径依赖”作为对质对象，是因为它们分别代表了认知偏差、社会评价压力、元认知信念、执行功能、机器学习、认知防卫以及制度演化等不同层面与品评性早闭最邻近的概念。对质的目的不是逐条驳倒这些概念

，而是在每一个区分点上，确立品评性早闭不可被既有命名所1:1替换的辨别面。

[4] 第六节中的课堂互动案例系综合笔者长期教育观察与相关文献中的典型情境所做的思想拓展性例示，其中具体人物、时间、地点均已匿名化处理，部分细节经过合成与简化。本节的自称定位是“启发性外推”（heuristic extrapolation）而非“经验验证”（empirical validation）。其功能是在一个与语言模型截然不同的认知生态系统中，展示品评性早闭结构的可辨识性，从而为该概念提供跨领域的理论合法性的提示——其严格的经验验证有待于未来的课堂微观互动研究或教育民族志来完成。

参考文献

• Amabile, T. M. (1996). Creativity in context. Westview Press.

• Budner, S. (1962). Intolerance of ambiguity as a personality variable. Journal of Personality, 30(1), 29-50.

• David, P. A. (1985). Clio and the economics of QWERTY. American Economic Review, 75(2), 332-337.

• Dempster, F. N. (1992). The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging. Developmental Review, 12(1), 45-75.

• Diamond, A. (2013). Executive functions. Annual Review of Psychology, 64, 135-168.

• Eva, K. W., & Norman, G. R. (2005). Heuristics and biases: In the eye of the beholder? Medical Education, 39(9), 870-872.

• Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. Archives of Internal Medicine, 165(13), 1493-1499.

• Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. Review of Educational Research, 67(1), 88-140.

• Husserl, E. (1913). Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. Erstes Buch: Allgemeine Einführung in die reine Phänomenologie. Halle: Max Niemeyer. [英译本: Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy. First Book: General Introduction to a Pure Phenomenology, trans. F. Kersten, The Hague: Martinus Nijhoff, 1982.]

• Mercier, H. (2020). Not born yesterday: The science of who we trust and what we believe. Princeton University Press.

• North, D. C. (1990). Institutions, institutional change and economic performance. Cambridge University Press.

• Osborn, A. F. (1953). Applied imagination. Charles Scribner's Sons.

• Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems, 35.

• Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. American Political Science Review, 94(2), 251-267.

• Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. Journal of Educational Psychology, 82(3), 498-504.

• Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. Mind & Language, 25(4), 359-393.

• Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems, 35.