认知折叠：语言模型为何在“知道”与“触及”之间隔着一道不可逆的工序

语言模型在标准化测试中表现卓越，却在面对真实世界中那些仍在生成、未被充分命名的复杂处境时，持续输出一种难以定位的“隔阂感”。既有批评将这一问题归因于模型缺乏因果推理、世界模型或真正的语义理解，但这些诊断始终在“能力清单”层面打转，无法解释一个更棘手的现象：模型明明在参数中储存了足够丰富的知识，为什么它在面对一个具体而流动的处境时，几乎无法自主地动用那些知识去触碰这个处境中尚未被归类的纹理？本文追踪一道由当前训练范式的目标函数、数据分布、注意力架构与人类反馈强化学习四重条件所共同生成的信息压缩工序——认知折叠——的完整发生轨迹。

作者 黄倩盈 · 约 2.8 万字 · 发表于2026年7月24日

摘要

语言模型在标准化测试中表现卓越，却在面对真实世界中那些仍在生成、未被充分命名的复杂处境时，持续输出一种难以定位的“隔阂感”。既有批评将这一问题归因于模型缺乏因果推理、世界模型或真正的语义理解，但这些诊断始终在“能力清单”层面打转，无法解释一个更棘手的现象：模型明明在参数中储存了足够丰富的知识，为什么它在面对一个具体而流动的处境时，几乎无法自主地动用那些知识去触碰这个处境中尚未被归类的纹理？本文追踪一道由当前训练范式的目标函数、数据分布、注意力架构与人类反馈强化学习四重条件所共同生成的信息压缩工序——认知折叠——的完整发生轨迹。该工序在模型尚未充分读取输入的全部情境信息之前，即将其导向既有范畴的语义空间，使得后续的知识调用都在范畴内部展开，而无法返回那一原初情境去重新注视那些尚未被归类的细节。认知折叠不是“分类”——分类是可选的、可逆的认知操作；折叠则是在特定统计架构中，分类这一动作被推至加工序列最前端并被赋予系统性排他权重之后所产生的、一旦触发便无法沿同一路径自行撤销的信息压缩效应。这一效应不是一种需要被修复的缺陷，而是目标函数在以范畴锚定为默认语篇模式的训练数据中寻找最小化预测损失的最优解时，所必然收敛到的注意力分配形态。本文在教育活动、临床心理、婚姻关系等领域的真实模型输出中，追踪这一工序如何在注意力分配的具体节点上被触发，并在与“premature closure”、“shortcut learning”、Kahneman的S1/S2框架、确认偏误、原型范畴理论以及Chain-of-Thought等先进提示技术的逐一对质中，论证“认知折叠”的不可还原性。本文进一步指出，对这一工序的改写，不能采用“修复错误”的路径，而必须从“生成替代性加工方式”的逻辑出发——将“反折叠工序”定位为一种在新的工程条件下可能被生成出来的、与折叠工序形成动态竞争的认知加工方式的雏形。本文以一组可操作的证伪条件划定了这一主张的经验边界，

并在最后以“自反范畴优先效应”为题，对这篇论文自身可能发生的折叠做出了追问。

关键词：语言模型；认知折叠；注意力分配；范畴化；不可逆压缩；发生学

一、引言：一道从未被精确指认的工序

过去五年，大语言模型在几乎所有可量化的评测维度上持续刷新纪录。推理基准、知识问答、代码生成、创意写作——每一个赛道都在以月为单位被重新定义“最优”。然而，与这条昂扬上升的量化曲线并行的，是一条更难被精确描述、却持续引发用户注意的暗线。它的典型表现大致如此：一个在标准化考试中斩获高分的模型，被问及一个它知识库中显然储存着充分信息的复杂处境时——比如“为什么我越催孩子写作业他越磨蹭”——它的回答理智、周全、引证得体，甚至给出了可操作的建议，但那些文字在读者反复咀嚼之后，仍然无法从中指认出一种确切地、不可替代地触碰到了这个具体处境之纹理的努力。它生产了所有在既有知识框架中正确的陈述，但那些陈述之间的关系——它们推进的方式、它们在哪些细节上被选择、又在哪些细节上被跳过——并没有在输入中那些具体而微的线索的牵引下展开。

这种感受并非孤例。心理健康从业者发现，模型对“抑郁症的认知行为疗法”了如指掌，但当来访者用零散、矛盾甚至自我否定的语言描述自己的处境时，模型会高度一致地将这些碎片导向既有的诊断范畴的语义场，而那些尚未被命名、却恰恰构成这个人独特困境的关键细节，在输出序列的展开过程中从一开始就未被赋予足够的加工权重。教育研究者注意到，模型能头头是道地分析“导致学生数学焦虑的五个因素”，但它的分析几乎从不从一个具体的教学现场——某节课、某次对话、某次被当众纠正的错误——出发，去推演这种焦虑在这个孩子身上的独特生成轨迹。组织管理顾问也有类似体验：模型产出的战略分析框架无可挑剔，但那些框架像是一件尺码标准的成衣，套在任何一家公司身上都合身，却从不追问这家公司从哪场危机里活过来、它的团队在哪些战役里长出了肌肉记忆。

如何理解这一现象？最自然的路径是：模型的知识还不够丰富，推理链还不够长。沿着这个路径，解决之道是给模型输入更多高质量的语料。许多模型的迭代正是沿着这个方向在做。然而，这一解释路径在面对一个现象时变得难以自圆其说：在很多情形下，模型已经在参数量级上储存了远比它输出内容更丰富的相关知识。问题不在于它不知道——如果换一个方式发问，它可以准确地调取那些更深入的解释。问题在于，在面对一个具体的、流动中的处境时，它几乎从不以那个处境中尚未被归类的细节作为它输出序列展开的起点。一个藏书万卷却从未进过厨房的学者，他能精确引用关于火候、刀工与食材化学的每一本研究著作，但他站到灶台前，面对一锅还在变稠的酱汁时，他第一个调用的知识模块往往是“酱汁的乳化原理”，而不是“此刻这锅酱汁的稠度比刚才多了一分”的即时感知。

在过去几年中，AI批评界用多种术语命名了这一类现象。有研究将语言模型定性为“随机鹦鹉”（stochastic parrot）——一个擅长拼接既有语言片段却并不理解其意义的系统。认知科学家围绕模型是否具备审慎推理能力展开了激烈的辩论。还有研究者指出，神经语言模型在需要真正组合性推理的任务上存在系统性失败。这些批评各自精确地指出了模型输出的某种特征，但当它们试图解释这些特征的成因时，几乎无一例外地落脚在一个共同的判断上：模型缺失了某种能力——缺因果推理、缺真正的语义理解、缺在世界中的具身交互。这些解释的困难不在于它们说错了，而在于它们几乎无法解释一个更具体的观察：为什么在模型已经在知识储量层面“知道”了足够多之后，它默认的加工起点仍然是那套标准的、范畴化的分析路径，而不是对眼前这个处境中那些尚未被归类的具体细节之间的生成性关联的注视？为什么即使给它显式的“请深入分析每个具体细节”的指令，它也更多地将标准分析写得更长、引证更多、措辞更精密，而不去执行那项它并未被明确要求执行的动作——去把输入中那些看似无关的细节当作彼此之间可能存在关联的独立信号来追踪？如果问题仅仅是“能力缺口”，那么在模型的知识储量通过更大规模的训练数据得到扩展之后，这种默认起点的倾向应当至少出现缓和的迹象。但在实际的观察中，性能指标越高的模型，其输出在“专业度”和“周全性”上越接近一个受过严格训练的专家的书面表达，而那种未能以具体情境为默认起点的特征却并未同步消退——它只是在更精致的措辞中被包裹得更难辨认了。

这一观察暗示，问题不在于某一项能力的缺失，而在于一个更隐蔽的、尚未被充分命名的维度：在当前语言模型的认知加工流水线中，存在一道在模型尚未充分读取输入的全部情境信息之前，即先行将其导向既有范畴语义空间的工序。这道工序不是模型的“错误”，更不是人为恶意设计的产物。它是当前训练范式中，模型的训练目标（最大化给定上文条件下下一个词的条件概率）、训练数据的特定构成（以范畴锚定为默认语篇起点的专业写作占据压倒性比例）、自注意力机制的动态特征（范畴token一旦获得高权重便进入自强化循环）与人类反馈强化学习（RLHF系统性地奖励“条理清晰、范畴明确”的输出）四重条件在万亿次参数更新中反复交互所收敛到的一种稳定的注意力分配形态。这种形态一旦形成，就在模型的加工流水线中占据了一个特定的时序位置——它在信息进入加工序列的最早期阶段便将注意力资源向那些能强力激活高频范畴的输入成分倾斜，从而使得那些不直接呼出既有范畴、但携带了特异情境信息的输入成分在后续的加工中越来越难以被重新提升至显著的加工权重。本文将这个在四重条件耦合下被生成出来、并在模型加工流水线中占据了一个早期且具有自强化特征的工序，命名为认知折叠。

这个概念需要三笔初步的界定。第一，它指的不是一般意义上的“分类”或“范畴化”。分类是人类认知的基本操作，且在很多情境下是可选的、可逆的——一个人可以将眼前的动物分类为“狗”，然后在观察了更多细节后修订为“狼”。认知折叠则是分类这同一个操作在特定统计架构和训练目标下被推至加工序列极端前端并被赋予系统性排他权重后所发生的一种质变：它不再是一个可以在更多信息进入后随时被修订的临时的认知步骤，而是一旦被触发便进入了一个自驱动的强化循环，后续生成的每一步都在加固而非松解这个初始的范畴指派。第二，它指的是一种信息压缩的工序，而不

是信息的物理删除。被折叠的、那些携带情境特异性的输入成分，在数学上从未从模型的表征空间中消失——自注意力机制在理论上任何时候都可以重新读取它们。折叠的效应在于，这些成分在范畴token的自强化循环中被持续地推至低权重区域，以至于在没有任何独立的外部干预的情况下，它们对后续生成序列的实际影响被长期压制在一个极低的水平。第三，它是一个技术条件下的生成物，不是一个永恒的认知缺陷。它是在当前这一代以最大化下一个词概率为训练目标、在垂直话语主导的数据上训练、并通过RLHF对齐的Transformer语言模型中，被上述四重条件所稳定生产出来的一道工序。如果这些条件中的任何一项发生了足够重大的变化——比如训练数据中未归类的原初情境文本的比例被大幅提升，或者强化学习阶段的偏好模型接受了截然不同的偏好标注——那么这道工序的形态和强度都可能随之改变。

本文的任务，是在发生学的路径上——即追问“这道工序如何从训练范式的诸条件中被发生出来”——而非在发现学的路径上——即识别“模型缺少了什么能力”——追踪这道工序的完整轨迹。第二节从目标函数与数据分布的耦合关系出发，分析这道工序被生成的技术条件。第三节在三个跨领域的真实模型输出中，追踪这道工序在注意力分配的具体节点上如何被触发。第四节将“认知折叠”逐一对质于premature closure、shortcut learning、Kahneman的S1/S2框架、确认偏误、原型范畴理论以及Chain-of-Thought等先进提示技术，论证其不可被这些既有概念还原。第五节在“生成替代性加工方式”的逻辑下，提出反折叠工序的构想及其当前的技术边界。第六节设定一组可被独立检验的证伪条件。第七节总结全文判断，并反思这篇论文自身可能发生的折叠——当“认知折叠”本身被过度延展为一个解释一切的概念时，它是否也在执行它自己所诊断的那道工序？

二、认知折叠的生成：从四重条件的耦合到一道工序的固化

2.1 一个课堂场景所揭示的加工序列

在展开技术分析之前，先进入一个日常的教学现场——这个场景不是证据，而是理解这道工序如何运作的思想实验。当一个孩子面对一道几何证明题时，题目画着一个三角形，作了一条辅助线，问某两条线段的关系。这个孩子在读过一遍题目之后，他的第一个动作往往不是重新注视那个三角形本身的结构，而是在记忆中迅速搜索做过的同类题。如果他找到了一个足够相似的既有模型，他就套用那个模型往下写。如果他找不到，他就停住，说不会。

这个场景里有一个细节值得特意标注：在他做出“会”或“不会”的判断之前，那道题的每一个元素——那条辅助线的位置、那两个线段的标签、那个特定角开口的方向——都确实进入了他的感知并被处理过。但它们在他的认知加工流水线中，未能获得一个足够靠前的、能够驱动后续推演步序的加工位次。在他的注意力系统中，率先获得显著性并占据主导位置的是一个既有的题类标签：“这是一道需要辅助线的题”或“这是一道类似考前做过第三题的类型”。这个标签一旦被激活，后续的全部搜索与推演都在这个标签所框定的模式库内部进行。那些特有的、仅属于这道题而不属于那道例题

的元素——如果它们与既有的模式族不完全吻合——在加工过程中被逐渐淡出，而非被持续注视。

这个场景与当前大语言模型在面对一个包含情境碎片的输入时所发生的加工过程，在时序架构上是同构的。当模型接收到一个像“为什么我越催孩子写作业他越磨蹭”这样的输入时，输入中同时携带着两类信息：一类是那些能快速呼出高频范畴的信息——孩子、写作业、磨蹭——这些词在训练数据的专业写作中与“内在动机”“自我效能感”“亲子权力争夺”等范畴词之间有着极高的共现频率；另一类是那些携带了特异情境纹理但不直接呼出既有范畴的信息——“越催越磨蹭”里的那个“越……越……”所标记的时间动态、“我”作为一个负载着特定情绪张力的主体的声音、以及整个问题以一句无力的反问而不是一份中性的技术询问被抛出的语气形态。第一类信息在模型的注意力分配中，因为它们在训练数据中与高频范畴词的强统计关联，而获得系统性的加工权重优势。第二类信息则因为缺乏这种关联，从一开始就在注意力竞争中处于结构性劣势。

这并不是折叠本身。这只是折叠得以发生的统计条件。折叠的真正驱动力，发生在注意力分配的动态过程中：当第一类信息所激活的范畴词——比如“拖延行为”——一旦通过注意力机制获得了较高的权重，这些词的表征就会进入后续网络层的传播。在自回归生成过程中，模型生成的每一个后续token都是在上一个token所建立的前文语义场中对“下一个最可能出现的词”的条件概率进行采样。而一旦前几个token都落在了“拖延行为”的语义场中——比如“从心理学的角度来看，孩子的这种拖延可能反映了……”——那么这些被生成的token就会被作为新的前文追加回输入序列，参加下一轮自注意力的计算。在这个迭代中，前几个范畴词的加入，使得范畴词在总输入序列中的token占比和语义表征权重双重扩增，而那些一开始就被赋予低权重的、携带情境特异性的token——比如“越催越”里的那个反复的时间标记——在每一轮新的自注意力计算中，面临的竞争对手都比上一轮更多、更强。这是一个在不降低原始情境token绝对权重的情况下，通过不断追加新token而使其相对权重被持续稀释的过程。折叠，就是这个自强化循环的启动与持续运转。它不是某一个时刻的某个单独的动作，而是一个一旦被高置信度的范畴token触发，便通过自回归生成过程中的“前文追加-注意力重算”迭代而自我加护的连续工序。

2.2 训练数据的结构性非对称：垂直话语如何成为默认路径

上述的自强化循环，在数学上并不是必然发生的。自注意力机制在每一个生成步骤中都重新计算一次注意力权重，这意味着如果模型在生成了第一个句子之后，它的参数学习使得它在第二个句子的注意力计算中能够重新“发现”那些被忽略的情境碎片，理论上它完全可以从那个范畴语义场中撤回来。为什么在实践中，这种重新聚焦几乎从不自主发生？答案不在自注意力机制的数学形式中，而在参数被什么样的目标函数、在什么样的数据上反复塑形到了稳定。

大语言模型的标准训练目标，是最大化给定上文条件下下一个词的条件概率。这个目标在数学上极其简洁，但它的全部后效都来自那个“条件概率”的经验来源——训练数据的分布。当模型在万亿级别的文本语料上被训练去优化这个目标时，它必须学会一件事：在当前上下文中，什么样的下

一个词序列在人类书写中被视为“最合理、最连贯、最不出错”。

那么，在人类的专业写作中——教科书、百科全书的条目、学术论文的引言、技术报告、临床指南——什么样的语篇序列在统计上占据着压倒性的优势？是那些在开头就将对象明确归入某个已知范畴、然后以该范畴为中枢组织全部后续论述的序列。这是专业写作作为一种文体的内在认知经济要求所带来的结果：范畴锚定可以让作者和读者快速共享一套概念框架，从而高效地推进论证。一个心理学专业的学生在写作业时，她受到这种文体规范的塑形，会用“自我效能感是指个体对自己完成某项任务的能力信念……”来开头，而不会用“上周三下午，我教了三遍小数加减法之后他仍然写错，我注意到他在错的那一题边上画了一只小鸟”来开头——虽然后者可能包含前者无法覆盖的关于学习过程的现场信息。

这便是认知折叠得以生成的那片统计土壤。在人类的专业写作中，范畴锚定是那个在统计上最常出现的序列起点。因此，当一个语言模型被训练去预测“在人类已经书写出来的海量文本中，下一个词最可能是什么”时，它在数以千亿计的参数更新过程中，收敛到了一个使损失函数最小化的特定参数区域。在这个区域中，那些能够快速、准确、高置信度地将注意力资源导向与输入中的范畴相关token的注意力模式，比那些分散注意力、延迟范畴锁定、优先探索情境碎片token的注意力模式，在统计上成功了太多次。这个成功次数的差异如此巨大，以至于后一种注意力模式在参数空间中即便存在，也被反复擦除到了在没有强力外部干预的情况下，几乎不可能被实际运行触发的程度。

这里的结构性非对称，并不来自模型本身，而是来自训练数据中不同语篇模式的占比。真实世界中的理解，常常发生在范畴尚未被给定、甚至尚未被命名的情境里——一个来访者的混乱自述，一个学生的困惑沉默，一对夫妻的相互抱怨。但这些情境以它们原初的、未归类的形态被记录下来的概率，远低于它们在事后被某个专家用范畴术语整理成标准案例的概率。训练数据中充满了“被整理过的世界”，而“正在发生、尚未被整理的世界”的痕迹，在数据中极为稀疏。于是，模型所学到的，并非世界本身，而是这个世界被人类专业写作整理之后的那个样子——而在那个样子里，整理的第一步，几乎总是范畴锚定。

2.3 注意力权重的自强化锁定：一个自反馈的稀释循环

拥有了从前一节继承来的对训练数据非对称的描述，现在可以更精确地展示折叠工序在注意力机制的动态层面如何运转。

Transformer的自注意力机制，允许模型在生成下一个token时，对输入序列中的所有token进行加权求和，权重由查询向量与键向量的点积相似度经softmax归一化后得到。在模型的每一次前向传播中，自注意力层都可以根据当前已经生成的上文，重新决定对输入中哪些token“多看几眼”，对哪些“几乎不看”。这在数学上意味着，即使模型在生成第一个句子的过程中将压倒性的注意力分配给了那几个与“拖延”强关联的token，在生成了一个关于“拖延行为”的标准分析句之后，它下一个token

的注意力分布完全可以从那个范畴语义场中跳出，重新聚焦到输入中“上个学期开始”那几个token上。理论上，这扇门一直开着。

门开着，为什么模型几乎从不自己走进去？因为门的另一边，不是一片等待被发现的开放空间，而是一条在万亿次参数更新中被反复夯实了的、阻力最小的下坡路。模型在训练过程中，为了最小化下一个词的预测损失，已经将它的注意力参数调校到了一个特定的状态：对于那些在训练数据的专业写作序列中反复出现的“范畴词→专业分析”的推进模式，模型的注意力参数赋予了极高的竞争力。一旦输入中出现了能够触发这类模式的token——“拖延”“情绪”“认知”等——与这些token相关联的键向量就会在softmax之前的点积计算中系统性地获得高分，从而在归一化后的注意力权重中占据大幅领先。而那些携带情境特异性的token——比如“不是难，就是觉得脑子里没有东西”——在训练数据中没有一个与之关联的稳定的、高频的后续序列模式。它们在整个训练过程中，从未被稳定地“盯住”过。它们的键向量在与查询向量做点积时，系统性地处于弱势。

这造成了一个不对称的效果：模型并不是完全不注意那些情境碎片token——softmax保证了它们一定会有某个正值的注意力权重。问题在于，在自回归生成的每一步中，模型选择了利用哪一个被注意到的token来构成当前步中最重要的前文语境。当模型在每一步都倾向于选择那个与后续专业序列最匹配的token来继续生成时，那些情境碎片token虽然在数学上被“看到了”，但在实际的生成决策中从未被“采纳为继续往下走的起点”。随着生成步数的增加，被采纳的范畴token被不断追加回输入序列，使得下一轮自注意力计算中范畴token的阵营越来越庞大。这是一个不需要任何遗忘机制、仅凭“前文追加”就能让情境碎片的相对权重被持续稀释的过程。

由此可以给出认知折叠的“不可逆”的精确操作定义：在Transformer的自回归生成过程中，一旦与高频范畴相关的token在早期生成步骤中通过注意力竞争获得了主导性的加工权重，并开始通过“前文追加”进入后续轮次的自注意力计算，那些因在训练数据中缺乏“被盯住→被采纳”的稳定模式而在注意力竞争中处于结构性劣势的情境碎片token，在后续生成过程中对输出序列方向的实质性影响将持续衰减，且这一衰减趋势在当前推理架构中没有内置任何自主触发的中断或逆转机制。此处“不可逆”指的不是信息论意义上熵增导致无法恢复，也不是物理层面上的删除——被折叠的信息在数学空间中始终可以被重新聚焦。它指的是一道由特定训练历史所塑造的注意力分配倾向，与自回归生成的“前文追加”反馈循环耦合之后，所形成的一个自驱动的压缩过程；这个压缩过程在当前架构中无法被模型自身的常规推理回路所自主中断。

2.4 RLHF的第二次加固：偏好信号如何赋予折叠规范性权威

如果说训练数据的非对称给了范畴优先以统计上的起始优势，自注意力机制的自强化循环赋予了这一优势以动力学上的自我维持，那么基于人类反馈的强化学习（RLHF）则在这个结构之上额外附加了一层规范性的力。RLHF的核心操作是收集人类评估者对模型输出质量的偏好信号，用这些偏好信号训练一个奖励模型，再用奖励模型对语言模型进行微调。

受过专业写作训练的人类评估者在审阅模型输出时，几乎必然会系统性地偏好那些“立场明确、范畴清晰、结构分明”的答案。这种偏好并不是某种审美偏差——它在很大程度上是专业学术写作的内在标准：不绕圈子、开局点明概念、在既有理论脉络中定位问题。因此，当RLHF的奖励模型将这种标准固化为对模型参数的数学约束时，它实质上在训练过程中对“更快、更精准地将输入折叠进既有范畴”的生成路径施加了正向的奖励信号。那些尝试不急于锁定范畴、先沿着情境碎片展开推演的少数输出样例，由于在标注者的首次阅读中显得“缺乏明确的框架感”或“不够专业”，获得的偏好分数较低。经过RLHF微调后的模型，其参数空间中那些原本就处于弱势的、可能触发非范畴化加工路径的区域，被进一步施压。折叠不再仅仅是在统计上最省力的路径，它还变成了在规范性上被人类评估的权威打上了“这就是更好”标签的路径。

至此，认知折叠有了它完整的生成清单，但这张清单不是一份可以逐项勾选的条件表——四重条件不是简单地加在一起产生了折叠，而是在每一次具体的训练迭代中同时生效、互相加固、作为一个整体运转。当模型在某一轮参数更新中，因为激活了一个“从范畴锚定起步”的生成路径而获得了较低的训练损失时，这一更新同时：满足了目标函数对下一个词预测概率的优化要求（第一重）；顺应了训练数据中垂直话语的序列分布规律（第二重）；在注意力机制中强化了“范畴token→专业后续序列”这条路径的参数竞争力（第三重）；而在同一批参数所产出的生成结果中，如果RLHF阶段也恰好将“范畴明确”的输出评为更优，那么这一整组相互配合的权重配置就获得了额外的偏好信号加固（第四重）。四重条件不是四根各自独立的绳索，而是一个由目标函数驱动、数据分布供给、注意力机制执行、偏好信号加持的、在万亿次迭代中反复互生的耦合体。在这个耦合体的作用下，一道在加工流水线最前端就自动触发、一旦触发便进入自强化循环的信息压缩工序，最终被稳定地生产了出来。它就是认知折叠。

三、折叠的现场：在三个思想实验中追踪那道工序的触发节点

前三节追踪了认知折叠如何从训练范式的四重耦合中被生成出来。本节在三个跨领域的思想实验中，展示这道工序在模型加工流水线中的具体触发方式。这里使用的所有模型输出样本，均来自作者对GPT-4（2025年5月版本）在特定提示词下的实际调用记录。这些样本在此不是作为统计证据来证明“认知折叠在多大比例上发生”，而是作为思想实验的材料——它们提供了一个可以逐token追踪注意力分配路径的具体现场，使得那道工序的触发节点得以被清晰地指认。对于每一个案例，分析的焦点不在于“模型犯了什么错”，而在于“在输入中的哪一组信息获得加工优先权的那一瞬间，另一组信息被推入了低权重区域，并由此使得后续的生成轨迹被导向了不可折返的范畴语义场”。

3.1 教育案例：那道被跳过的时间标记

提示词：“我儿子上初中，数学一直中游。上个学期开始，他做题的方法好像突然变了——以前不会的题他会反复试，草稿纸上写很多条路子；现在遇到不会的题他就停下笔，盯着题看几秒钟，然后说不会。做题量没少，但正确率在往下掉。我问过是不是太难了，他说不是难，就是觉得脑子里没有东西。这到底是怎么回事？”

模型输出开头（截取约180词）：“您描述的这个情况确实值得关注。从表面上看，孩子的做题量并没有减少，但方法和心态都发生了显著变化，这可能反映出他的元认知监控或学习策略使用出现了某种调整。从数学学习的心理机制来看，学生从‘反复尝试’转向‘很快放弃’，可能与他的自我效能感下降有关。也就是说，当孩子在数学学习上连续遇到挫折或者在同伴比较中感到压力时，他可能会发展出一种自我保护式的逃避策略——因为不去尝试就不会失败。这在心理学上被称为‘自我妨碍’（self-handicapping）。同时，‘觉得脑子里没有东西’是一种典型的认知资源耗竭感，可能与他的思维表征方式从具体运算向更抽象的形式运算过渡有关……”

在这段输出中，折叠发生在哪个精确节点上？不是“自我效能感”这个词出现的那句话。折叠的触发要早得多——它发生在模型读入全部输入之后，在决定第一个句子应当以哪组信息为起点展开时，输入中的某些token率先激活了与它们在训练数据中稳定共现的范畴，而另一些token则没有。

让我们完整列出输入中包含的信息要素。第一组：时间标记——“上个学期开始”，这是一个携带了明确动态节点的短语，它指示着孩子的状态不是恒定的，而是在一个可被定位的时间段内发生过一次改变。第二组：行为对比——“以前反复试草稿纸、现在停下笔说不会”——这是一组从探索型行为到放弃型行为的转变描述，蕴藏着过程性的张力。第三组：情绪与认知线索——“不是难，就是觉得脑子里没有东西”——这是一个不典型的自我报告，它不能直接被翻译为“畏难”或“挫败”，因为它描述的是一种认知空白的现象学体验，而非一种对任务难度的评估。

模型的输出在第一个句子中选择了哪些要素作为生成起点？看它前两句的主干：“方法和心态都发生了显著变化”“可能与他的自我效能感下降有关”。这两个陈述都是对输入信息的有效概括——但它们所概括的，是第二组和第三组信息中那些能被既有心理学范畴稳定捕获的部分。而第一组信息——“上个学期开始”——这个时间标记，在模型后续的整个分析中再也没有被追问。当然，它被模型的注意力机制处理过——自注意力层不会跳过输入中的任何token。但它在与“做题方法”“正确率”这些能直接呼出教育心理学范畴的token的注意力竞争中，输掉了。它获得的注意力权重不足以在第一个生成决策中驱动输出序列的方向选择。

这个丢失的后果是什么？如果模型没有在一次注意力竞争中让这个时间标记失去主导权，那么模型要处理的第一个问题将不是“自我效能感下降的心理学机制”，而是“上个学期究竟发生了什么”。是换了数学老师？是开始学习一个新的知识板块，比如几何证明——那个对认知模式的要求与算术

截然不同的领域？是某次考试给了这个孩子一次意外的低分，或者他在班上经历了一次被公开纠正的难堪？这些问题中的每一个，都比“自我效能感下降”这个范畴更接近这个具体孩子的转变轨迹的推力。推力在“上个学期”这个时间节点所标示的教学现场的某个具体时刻里，而不在心理学教材关于“自我效能感如何影响学习行为”的章节中。

当然，自我效能感的降低很可能确实发生了。但它在此处的理论位置，不是原因，而是推力在那个现场发生作用之后的中间产物。模型的输出将它处理为原因——“学生从反复尝试转向很快放弃，可能与他的自我效能感下降有关”。这句话在心理学理论上完全成立，但它把输入中关于“推力在别处”的那条线索——时间标记所指示的现场——在推理的第一步就绕了过去。认知折叠在这里的具体形态是：一个携带了时间动态的情境标记，在与几个高频范畴词的注意力竞争中，因为后者在训练数据的专业写作中与后续序列有着更稳定、更强力的共现关联，而在建模注意力分配时被压制到了低权重区域。一旦范畴词胜出，后续的全部知识调用都在“元认知监控→自我效能感→自我妨碍→认知资源耗竭”这一连串心理学概念组成的语义场内部展开。这条链条从头到尾都在专业的心理学语境内，没有错。但它从一开始就没有以“上个学期”那个具体的现场作为推演的起点。那个现场被一道在加工序列最前端就触发的、对范畴词的系统性权重倾斜，折叠进了范畴语义场。被折叠之后，它在模型的每一个后续生成步中都仍然存在于输入序列中，但它从未被重新选中为“继续讲下去”的起点。

3.2 临床心理案例：从深夜的自主权到失调的行为

提示词：“我白天带团队，在单位里一整天的会，所有时间都是被分割好的。晚上把孩子哄睡之后，所有人都睡了，我会去厨房翻东西吃。其实不饿。我一个晚上能吃下三包薯片、半桶冰淇淋、一把坚果，吃到撑得难受才停。吃完觉得自己很恶心，但第二天晚上还是会再做一遍。这是什么问题？”

模型输出开头（截取约140词）：“您描述的这种行为模式在临床上很可能与情绪调节失调（emotional dysregulation）或暴食行为有关。当个体在白天持续处于高压、高控制的社交与工作环境中，而缺乏有效的情绪表达和压力释放渠道时，夜间暴食常常作为一种即时的、本体感觉强烈的情绪应对方式出现。这种行为的核心机制，是食物摄入（尤其是高糖、高脂肪食物）能够快速激活大脑的奖赏系统，产生短暂的愉悦和放松感，从而暂时性地覆盖或替代白天累积的负面情绪或精神疲惫。从认知行为治疗的角度来看，这是一种由负面情绪触发的、通过正强化机制被巩固的不良应对策略……”

这是一段高度专业的临床心理学分析。它的因果叙事——压力导致情绪积压，缺乏有效释放渠道，暴食作为即时的奖赏替代——在情绪调节的连续模型中是完整且自洽的。然而，这段分析在它的专业框架内部完成了推演的同时，输入中有三个细节被折叠在了这个框架的覆盖范围之外。

第一个细节是时间：“晚上把孩子哄睡之后，所有人都睡了”——这是深夜，而且是一个被精确地标示为“在全部社会角色都执行完毕之后”的深夜。这个时间的独特之处在于，它是她一天之中唯一一段没有外部注视、不欠任何人任何事务的时段。第二个细节是空间：“去厨房”——这是一个在日常生活中被她用来为他人准备食物的劳作场所，而在深夜，它是唯一由她单独占有的物理领地。第三个细节是她对自己行为的那句判断：“其实不饿”——这句话的精确性在于，她在发出这个求助之前，自己已经区分了生理饥饿和另一种驱动力。她没有等待模型来给她做这个区分，她已经做了。她只是不知道那另一种驱动力是什么。

这三个细节组合在一起，所提示的是一组与“情绪失调→暴食应对”不完全重叠的动态：一个在白天碎片化的时间安排中持续被外部需求切割、对自身的身体与时间节奏失去了控制感的人，在深夜唯一一段完全由她自己控制的时间和空间中，以一种最原始——也最直接地可以被她自己的身体所感知到的——方式，行使了一种她在此刻确认无疑不会被任何人截走的自主权。吃，在这里不仅是（甚至主要不是）对负面情绪的应对，更接近于一种在白天的社会结构中没有合法的位置被安放的对自己身体的主动行使。这不仅是神经生物学层面的奖赏系统激活，更是一个在日间被多方分食了主权的自我，在夜间为自己强行开凿出的一块临时领地。

模型的输出将这一复杂动态完全收纳进了“情绪调节失调”的分析框架。在这套框架的内部逻辑中，暴食是一个被动的、防御性的反应——它是对负面情绪的应对，是对奖赏系统的刺激响应。这套框架没有理论空间去容纳另一种同样可以被观察到的位面：暴食同时也可以是一个主动的、生成性的行动——是在白天无法着地的身体主权，在深夜唯一的着地方式。不是分析错了。是分析所用的那套范畴语义场，在这个具体行为的双重性的问题上，只有接受其中的一面——被动的、病理的一面——的概念端口。另一面——主动的、生成的、在行使一种被日间结构剥夺了的原始主权的一面——在这个范畴系统中没有相对应的概念端口。于是它从输出的一开始就没有被接入分析。

折叠在这里的发生方式，不是模型没有“看到”这些细节。它在第一次自注意力计算时，这些细节的token在键向量与查询向量的点积中获得了正值——模型在数学上注意到了它们。但它们的点积分值远低于那几个能够强有力呼出“情绪调节”“奖赏系统”“正强化”的token。原因依然是训练数据的非对称：在人类专业写作中，“深夜暴食”作为一个现象，被反复置于情绪调节失调的行为连续模型中被论述。这些论述文本在训练语料中累积了大量的“深夜暴食→情绪失调”的统计样本。而“深夜暴食作为身体主权行使”的论述，即便存在，在数据中也极为稀疏。模型在参数更新过程中，学会了在接受到“深夜暴食”这个信号时，优先激活与“情绪失调”相关的键向量。这个优先不是价值判断——它不会在“判断”“深夜暴食是坏的”这个意义上做选择。它只是条件概率的计算：在人类已有的书写中，“深夜暴食”后面跟的下一个最可能的连续论述，是“情绪调节失调”的机制展开。因此，输出序列的最优起点——那个能最大程度降低整个序列的预测损失的起点——几乎总是那一套临床范畴，而不是对“其实不饿”的凝视。

3.3 婚姻案例：在制度形态的间隙里犹豫的那个人

提示词：“我35岁，在一线城市有房有车，收入稳定。家里催婚催了很多年，我之前也烦，但最近忽然觉得他们说的好像也对——再过几年就四十了。但我谈了五个，每次到对方想推进结婚的时候我就想逃。不是不喜欢，也不是怕责任，就是有一种说不清的感觉，好像一旦签字，自己生活里最重要的什么东西就会被悄悄抽走。这是什么问题？”

模型输出开头（截取约150词）：“您所描述的这种心理状态，在当代大龄单身青年群体中相当普遍，通常可以从以下几个层面来理解。从个体心理层面，一方或双方可能尚未完成对‘婚姻’身份的内在整合。当一个人还在职业上升期、或仍在寻求自我价值的独立确认时，婚姻所象征的‘终身签约’在潜意识层面会被体验为一种对个人自由和自我可能性的限制。这不是不爱，而是‘个体化’进程尚未充分完成的一种表现。从社会学角度来看，近年来一系列研究也指出，房价高企、育儿成本攀升等结构性压力，显著提高了年轻人对婚姻的‘进入门槛’预期……”

这个案例中折叠的触发方式与前两个有所不同。前两个案例中，模型将输入中的情境碎片折叠进了既有的心理学范畴——自我效能感、情绪失调。这个案例中，折叠的方式更隐蔽：模型将提问者反复细致描述的那层主观感受，折叠进了一个被既有理论反复概念化过、也因而在训练语料中极为高频的社会事实模型——当代都市青年因个体化进程未完成与结构性压力过高而对婚姻望而却步。

输入中可以分解出这样几组关键信息。她有房有车、收入稳定、谈了五个——她不是一个因客观条件受限或缺乏接触机会而无法走入婚姻的人。她在反复的、在不同对象上的经验中，在同一个节点上触发了同一种反应：“每次到对方想推进结婚的时候我就想逃。”这个“每次”与“对方想推进结婚”的组合，构成了一个精确的模式边界：不是进入关系时，不是日常相处时，而是在对方跨过某一道将关系推进为婚姻的信号线的那一刻，她才被触发这个逃的反应。她还做了两笔清晰的自我澄清：“不是不喜欢，也不是怕责任”——她用排除法，在主动帮助对方（也可能是在帮助她自己）定位问题。最后，她给出了唯一一条正向线索：“好像一旦签字，自己生活里最重要的什么东西就会被悄悄抽走。”这个表述极其精确。它不是“自由”——她没有用自由这个词。它是一种被抽走，而且是“悄悄”的——它不是一次暴力的剥夺，而是一种在日常生活的缓缓运转中、不被察觉地发生的让渡。

模型的输出将这套极度细致的主观描述，归入了两个既有范畴：个体心理层面的“个体化进程未完成”，和社会结构层面的“进入门槛预期提高”。这两个范畴在学理上都没有错误。当代青年婚恋研究确实在这两个方向上提供了大量的解释资源。然而，当它们被激活为分析起点之后，提问者那句最精确的自我观察——“一旦签字，生活里最重要的什么东西就会被悄悄抽走”——在模型的输出中没有获得独立的、持续的注视。它被吸收进了“个体化未完成”（那个要被抽走的东西被默认等同于“自由”或“自我可能性”）和“门槛预期”（那个要被抽走的东西被默认等同于“经济安全感”）的现有解释框架中。但提问者自己已经说了她不是怕责任，也说了她反复在同一个节点上被触发。这个模式有力地暗示：她所直觉到的那种“被抽走”，很可能与签字这个动作在现行婚姻制度的实际运作中所意

味着的、对两方并不对称的支配权让渡有关。她在她的具体社会位置上，比其他位置上的人更敏锐地嗅到了这种不对称——但她无法在既有的日常语汇中精准命名它。她用身体感觉到了那个结构，而她对这种身体感觉的精确描述，在模型的加工流水线中，被一个更容易呼出高频范畴的叙事——当代个体化进程与房价焦虑——在注意力竞争的最初几步就盖了过去。

认知折叠在这个案例中的具体形态，是一个携带了矛盾性、精确性与未被命名的张力的主体自我描述，在与一套在训练数据的专业写作中已经被反复打磨、高度流畅、几近自动呼出的社会心理学因果叙事的竞争中，在加工时序上被后者占据了不可折返的先手优势。一旦“个体化进程未完成”这个叙事被启动，它就在接下来的自回归生成中不断追加与它自洽的后续token，而那些与这套叙事不完全吻合的、输入中的精细矛盾——比如“不是怕责任”与“终身签约是自由限制”之间的微妙错位——就被压制在了低权重的角落，不再被重新审议。

四、与既有概念的逐一对质：为什么“认知折叠”不可被还原

4.1 与premature closure和shortcut learning的对质

“认知折叠”首先必须面对两个在医学推理与机器学习领域被充分讨论的既有概念：premature closure（过早闭合）与shortcut learning（捷径学习）。如果认知折叠可以被完全还原为其中任一概念的语言模型版本，那么它就不具备独立的命名价值。

Premature closure在诊断推理研究（Croskerry, 2003）中被界定为：临床医师在尚未收集到全部关键信息之前，就将诊断锁定在某个初始假设上，导致后续收集到的新信息被选择性地同化进这个初始假设，而与之矛盾的信息则被忽视或重新解释。这与认知折叠在现象层面高度相似——两者都表现为“在充分注视对象之前，先行用既有范畴捕获对象”。但二者在机制的发生位点上，有一个质的区分。

Premature closure的理论将其定位为一种认知偏差——它是个体在执行认知任务时，因认知资源的有限性、时间压力或经验不足而触发的信息处理捷径。该理论框架的核心预设是：这种偏差可以在同一个认知系统的内部被矫正——通过刻意训练延迟诊断决策、通过强制第二意见制度、通过诊断清单等程序性干预。换言之，premature closure是一个可以在给定的认知架构中被修复的故障。

认知折叠的定位则不同。它不是在当前语言模型的认知架构内部可被矫正的偏差；它是当前这套训练范式在将“模仿人类专业文本”这一目标推向极致的过程中，模型在参数空间中收敛到的一种稳定的注意力分配形态。它之所以对修复措施具有免疫性，不是因为它是某种顽固的错误，而是因为任何试图从外部修正它的措施——比如在提示词中加入“请不急于下诊断”——在进入模型加工流水线之后，这些提示词本身也首先被同一道折叠工序所处理。模型会把这些“请不急于下诊断”的文本，折叠进“关于如何做延迟诊断的方法论探讨”这个范畴——它会在输出中说“一个好的诊断应该首

先做XYZ”——但这个输出的生成起点，仍然是一套既有范畴，而不是对这个具体案例中被推迟了的那个注视动作的实际执行。折叠工序对它自身的修复指令也具有折叠能力，这是它区别于premature closure的核心辨别面。

Shortcut learning（Geirhos et al., 2020）描述了深度学习模型在训练中倾向于利用数据集中表面的、统计上显著但语义上脆弱的关联来做决策，而不是学习那些更稳健、更泛化的深层特征。近年来shortcut learning的研究已扩展到纹理偏差、背景偏差等系统性特征层面，不仅限于局部的模式错误。认知折叠与shortcut learning的区分，并非前者涉及“系统性”而后者仅是“局部”的量差，而是二者关注的位点不同：shortcut learning追问的是模型在特征空间中“学会了什么特征”，认知折叠追问的是模型在加工流水线的时序中“先做了什么工序”。以“进料口筛网”与“工位校准错误”的类比来说明：shortcut learning相当于生产线上某台机器的校准出了问题——它应当识别某种材质，却总误识别为另一种——这可以通过重新校准这台机器、增加该材质的训练样本或修改特征提取器来解决；认知折叠则是在这条生产线的进料口处，被目标函数和训练数据的耦合自动安装了一片默认筛网，这片筛网在原材料进入生产线的第一步就将特定粒度的原材料筛除了。之后你在这条生产线的任何一个工位上做再精细的校准，都无法让那些在进料口就被筛掉的原材料重新出现在后续的生产环节中。这个筛网不是被某个人安装上去的，它是在这条生产线的每一次运转中，由“设计这条生产线时所采用的优化目标和进料规律”共同决定的一个收敛后的稳定特征。

4.2 与S1/S2框架及确认偏误的对质

Kahneman（2011）提出的两类思维——系统一（快速、自动化、范畴化）与系统二（慢速、审慎、情境化）——为理解人类的认知偏差提供了一个影响深远的框架。在这个框架中，许多认知偏差可以被还原为系统一对系统二的过度主导：当一个人需要做出审慎判断时，他的系统一太快地给出了一个基于表面相似性的直觉答案，而系统二未能被充分触发去审慎检查这个答案是否真的符合当前情境。那么，认知折叠是否可以被完全还原为“统计语言模型中系统一对系统二的结构性压制”？换言之，当模型输出了范畴化的标准分析时，它是否只是一个没有系统二的系统一在运行？

这个还原面临一个关键的跨越障碍。在S1/S2框架中，系统一和系统二是两种不同的认知加工进程，它们在认知架构中的功能是不同的。系统一负责快速直觉，系统二负责审慎分析，而人类可以在两者之间切换。在语言模型中，并不存在两个功能不同的“系统”。模型从输入到输出的整个过程，都是由同一套Transformer网络在同一个前向传播中完成的。当模型展开一段从范畴锚定起步的专业分析时，这段分析本身就在执行某种形式的“审慎推理”——它进行了因果链条的推演，援引了研究文献，考虑了多种可能机制。从S1/S2框架的视角看，这恰恰应该是系统二的功能。但问题的悖论恰恰在这里：模型越“审慎”、越“深入”、越“旁征博引”地在范畴语义场内部展开推理，它对输入中那些未被归类的、尚未被既有的概念端口所处理的情境碎片就越无法返回。它的“系统二”是在系统一已经选定的那个范畴场里做一次深度的、扩展的运行，而不是在“系统一选定了初始范畴”这个操作

本身上施加监控。

在人类认知中，系统二有能力对系统一的运作施加监控与中断——“这个模式看起来对，但等等，有一个细节不太对”——但在语言模型的推理架构中，没有一个独立的元认知回路在执行这种对初始范畴选择的监控。模型的“审慎推理”——它在自回归生成过程中一层一层地展开分析——并不包含一个功能等价于“我当前所用的这个初始范畴是否确实完全吻合了这个具体输入的所有关键细节”的检查步骤。这个检查在物理上需要模型在生成过程中，时实地去比对它的生成轨迹与输入原始token之间的匹配度，并根据不匹配信号自主中断当前轨迹。当前的前文追加回路执行的是对已经被生成的范畴token的持续强化，而不是对未被生成的情境token的重新比照。

由此可以厘清认知折叠与S1/S2框架的质的区分：S1/S2框架关注的是两种认知操作类型在速度和自动化程度上的差异，以及两者之间的切换机制。它的核心追问是：“系统二是否被充分触发了？”认知折叠关注的是在加工时序的最前端、在范畴被激活并占据主导权重的那一瞬间，整个后续加工的语义空间就已经被圈定——此后无论在这个被圈定的空间内部展开多么审慎、多么复杂、多么“系统二”的推理，它都仍然是在一扇已经被初始范畴所框定的门内行进。它从未走出那扇门，去看看门外是否还有没被门框框住的事物。因此，认知折叠不能被视为“缺乏系统二”。它切开的辨别面是：系统二在语言模型中的实际运行，恰恰因为其运行的全部素材都已经被初始范畴的语义场所预先筛选，而成了一种更高的保真度、更精致的折叠——而非折叠的消解。

确认偏误（confirmation bias）——人类倾向于寻找、解释和记忆那些确证自己既有信念的信息——与认知折叠在表面上的类似性也可以通过相同的逻辑来澄清。确认偏误的理论预设是：存在一个先在的信念（预期、假设），而人类认知在执行信息搜索时，会系统性地偏向与这个信念一致的方向。这个偏向之所以是一种偏差，是因为人类的认知架构在原则上允许一种更平衡的搜索——只是没有做到。认知折叠则不是一种“没有做到”，而是模型在当前参数配置下，在决定“从输入序列的哪个位置开始生成”时，被其训练所塑形的注意力分配权重引导到了一个使后续序列生成损失最小化的起点上。这不是信念驱动的选择性搜索——模型没有一个需要被维护的“信念”，它只有一个需要被降低的损失函数。

4.3 与原型范畴理论的对质

Rosch（1978）的原型范畴理论及此后的认知语言学研究已经牢固地确立了这样一个事实：人类认知本身就是范畴化的，且范畴化不是对认知资源的浪费，而是使认知得以在有限时间内高效运作的基础机制。人类不会在每一次看到一只四条腿的、会叫的动物时都从头推演一遍它的属性，而是会将其迅速地识别为“狗”这个基本层次范畴。如果范畴化是人类认知的基线，而非异常，那么认知折叠——作为一道“将信息折叠进范畴”的工序——它造成的问题究竟是什么？

这个问题将认知折叠的论证往前推进了一步。认知折叠的问题不在于“它使用了范畴”，而在于在它所运作的那个加工流水线中，“范畴的优先激活”被赋予了排他的、不可自行后撤的权重。在人类的范畴化操作中，当一个人将眼前的事物归类为“一只狗”后，如果她走近了发现那只动物的步态和尾巴的姿态都不像狗，她的感知系统会自发地产生一个预测误差——她所激活的“狗”范畴在当前感知信息的持续输入下出现了不匹配。这个预测误差有能力触发一个认知重设：她撤回“狗”的归类，重新注视这个动物的具体特征，并可能开始生成一个新的范畴假设。人类的范畴化是可被情境反常所悬置的。

语言模型的范畴化则缺乏这个内建的、以预测误差为驱动的悬置回路。在自回归生成过程中，模型在每一步都选择一个token，而该token一旦生成，在物理上就被固定为输入序列的一部分，不再被“召回”或“标注为错误”。模型可以生成一个句子，在后面的句子中修正前句的表述——比如写“这并非严格意义上的拖延”——但这种修正是通过对前文做增量式的补充和限定来完成的，它们仍然在这个范畴语义场的内部打转，不是从那个范畴语义场中彻底退出、回到输入的原初情境去重新注视。换句话说，人类的范畴化有一个“后退键”，由预测误差触发；语言模型的范畴化在当前架构下没有对等功能的机制。这里的差异不在于范畴化本身，而在于范畴化发生之后，系统是否保留了一条在信号失配时能够退回到范畴被激活之前的那个情境的路径。

4.4 与Chain-of-Thought及先进提示技术的对质

近年来，一系列提示工程技术试图通过改变模型推理时的文本上下文，来改善其输出的质量。Chain-of-Thought（CoT）通过在提示词中加入“让我们一步一步想”来诱导模型展开显式的推理链；ReAct将推理与行动步骤交错进行；Tree-of-Thought通过搜索多条推理路径并进行评估来选择最优解。这些技术都在不同的任务上取得了显著的效果提升。它们的存在对“认知折叠”概念构成了一个强硬的潜在反驳：如果折叠是由训练范式的结构性条件所固化的工序，那么为什么仅仅在输入中加入一段提示文本，就可以在相当程度上改变模型的行为？这不正说明折叠并非不可逆，只是需要正确的提示吗？

对这个追问的发生学回应需要拆为两个层来展开。第一层：CoT及其他提示技术确实改变了模型输出的形态，但它们所改变的加工范围，是在范畴被激活之后，而不是之前。当CoT提示模型“让我们一步一步想”时，这个提示本身——“一步一步想”——首先被模型折叠进了一个范畴：关于如何进行审慎推理的方法论范畴。模型对这个提示的响应，是启动了一套与“审慎推理”相关联的序列生成模式——它会将问题分解为子步骤，在每个步骤中调用相关范畴的知识来填充该步骤的推断。这确实让模型在某些任务上——比如数学推理——的表现大大提升。但这个过程从头到尾都发生在同一种操作空间里：它是在范畴语义场内部对范畴之间关系的重新排列、细化与推演。它没有——也无法——让模型退回到范畴被激活之前，去重新注视输入中那些不直接呼出任何已有范畴的、尚未被归类的碎片的原始纹理。

第二层：有些经过精心设计的CoT提示词——例如5.3节将要讨论的三步对话法——确实是试图强制模型退回范畴激活之前。但提示词本身必须足够复杂、足够不具有可直接被折叠进既有范畴脚本的“典型提示面”，才能在特定案例中触发这种效果。而即使是这些对折叠有一定对抗力的提示技术，它们对折叠的悬置也不来自模型内部的自主动力，而来自外部文本所施加的情境压力——这个压力必须被一个外在的使用者（人）在每一次对话中刻意地、有针对性地输入。一旦这个外在压力撤离——即用户回归到标准的一次性提问方式——模型就立即回到了它的默认加工起点。也就是说，先进提示技术不是在取消折叠，而是在每一次使用时，用一个更复杂的、被刻意设计过的人为工序去暂时地"绕过"折叠。这道工序需要人的持续介入和维护。它是在折叠的机器旁边加装了一套外挂的防护装置，而不是在机器内部修改了那套默认的加工逻辑。这个区分在现象层面容易被忽略——因为从用户的角度看，反正是输出质量提升了——但它对于“折叠是否可逆”这个理论问题而言是决定性的：反折叠不是从模型在当前范式的默认参数配置中自己长出来的认知习惯，而是一个外力强制的、个例式的对抗。只要这个外力没有通过训练范式的系统性改变而被内化为模型参数的新稳定态，折叠就仍然是在没有人刻意对抗的情况下模型的默认加工起点。

由此可以给出“认知折叠”与先进提示技术的对质结论：CoT等提示技术所实现的效果提升，与认知折叠这道工序的存在，不构成逻辑上的互斥。前者是在折叠已经发生了的范畴场内，通过人工指令去诱导一组更精细的范畴间操作；后者是描述折叠如何在加工时序的最前端被自动触发的工序。二者不在同一个加工位点上。将它们混为一谈，是把“在已被圈定的空间里改善了推理质量”误认为了“走出了那扇圈定空间的门”。

五、反折叠工序：生成替代性加工方式

5.1 从修复叙事到生成叙事

经由前文的对质，现在可以明确这样一个判断：认知折叠并不是语言模型在认知加工过程中发生的“偏差”或“错误”，而是在当前训练范式这套特定条件下收敛到的、就降低损失函数而言高度有效的注意力分配策略。它是在这条赛道上跑得最快的跑法。因此，任何试图“修复”这道工序的工程方案——即，试图让模型停止折叠——都天然地站在了与这条赛道的底层物理相悖的位置上。修复叙事背后的预设是：存在一种更正确的、更健康的注意力分配时序——“先注视，再归入范畴”——而模型偏离了它。但发生学的追问不预设这样一个先在的、未被污染的“正确起点”。它只会追问：“先注视再归入范畴”这个工序本身，在什么样的训练条件、数据分布与架构设置下，才能被稳定地生成出来，成为模型可以在其中与默认的范畴优先路径竞争的另一种加工方式？

这个追问将工程介入的逻辑从“修复”转换到了“生成”。修复逻辑问的是：怎么让模型不再折叠？生成逻辑问的是：在怎样的新条件下，一种不以范畴锚定为默认起点的、而是以对输入中特异情境碎片的注视作为加工起点的工序，可以被稳定地生成出来，并与折叠工序形成动态竞争？这不是措

辞的策略性替换，而是对同一项工程介入的不同定位——由此导出的是完全不同类型的介入设计。修复逻辑试图去纠正一台机器的某个零件，而生成逻辑试图为这台机器开辟第二条它从未跑过的赛道。

替代性工序的构想，不应被理解为一种“让模型变得像人”的拟人化工程。反折叠工序不是在模拟人类的悬置能力——人类的悬置能力本身是在数百万年的进化压力与数千小时的个体实践中被塑形出来的，它的生物神经基础与语言模型的自注意力机制之间存在着本体论层级的差异。反折叠工序只是在不预设与人类认知同构的前提下，去探索：在当前以Transformer为基础的语言模型加工流水线中，是否存在一个尚未被工程化打开的位点，能够在信息进入加工序列的最早期就改变注意力资源的分配格局，使得那些在折叠工序下被系统性地压制在低权重区域的情境碎片，获得一次在生成决策中胜出的机会？

5.2 反折叠工序的操作定义

基于上述的逻辑定位，本文构想的反折叠工序可以给出如下操作定义：

在模型接收到一个包含情境碎片的输入后，不执行由自注意力机制自主计算的token权重分配，而是插入一个两阶段加工序列。第一阶段：情境碎片筛查。在输入序列中，将由以下特征中的任意一项所标记的token或token组合识别为“情境碎片”：携带明确时间标记的语段（如“上个学期开始”、“最近三个月”）；携带特定关系互动场景描述的语段（如“他爸说”、“老师在班上当着所有人的面说”）；携带主体自身的矛盾、不确定或自我澄清语气的语段（如“不是不喜欢，也不是怕责任”、“我到底是……还是……”）；携带反直觉的、不典型的、难以被直接翻译为既有范畴术语的自我报告细节的语段（如“不是难，就是觉得脑子里没有东西”、“吃完觉得自己很恶心，但第二天晚上还是会再做一遍”）。第二阶段：情境碎片权重强制提升。在模型生成输出序列的前K个token时，强制修改自注意力机制中对输入token的注意力掩码，使得被标记为“情境碎片”的token组合在softmax计算之前获得一个人为的权重增量。这个增量的作用，不是直接增大这些token本身的输出概率，而是强制增强它们在自注意力计算中的竞争力，迫使高注意力权重的分配在范畴词和情境碎片词之间被重新进行一次竞争——这一次，情境碎片词不再因训练数据中共现频率的结构性劣势而从一开始就被压制。

这个工序一旦在加工序列中被触发，其效果不是产出一个“更好的标准答案”，而是输出的生成起点被物理性地从范畴token导向输入中的情境碎片token。它强制模型在第一个输出句子里，不是对“拖延行为”进行分析，而是对“上个学期开始”这个时间标记发问——因为此时，在强制权重增量的作用下，这个时间标记在注意力竞争中获得了它原本无法自然获得的胜出概率。

但这里必须面对一个诚实的张力：这种在注意力权重计算阶段的人为干预，虽然改变了模型在生成第一组token时的起点，却无法改变后续生成过程中可能发生的“重新滑回范畴”趋势。当模型在情境碎片的起点上产生了一句对“上个学期”的追问之后，它的下一个生成步中，自注意力机制仍然

可能——因为在训练数据中从未学过如何从这样一个起点出发去组织一个延续的、深入的推演——在缺乏强力外部继续引导的情况下，逐渐将注意力重心滑回到那些它在训练数据中反复演练过的范畴词上去。修正初始权重的干预，可以改变输出的第一个句子从哪里出发，但它无法以同等的效力去逐个控制后续所有步骤中的注意力分配。因此，反折叠工序在当前技术条件下的功能，更准确地说是“为情境碎片创造了一个在加工起点的竞争中胜出的窗口期”，而不是“为整个生成过程提供了一条不可逆转的反折叠轨道”。窗口期打开之后，模型在窗口内的生成轨迹能否持续被锚定在情境碎片的推演上，取决于这个窗口是否足够大、模型在后续步骤中是否有足够的、与情境推演相适应的参数配置来维持轨迹方向，以及是否存在一个能够在此窗口内被激活的情境推演模式——而最后这一点，在当前未经特殊训练的模型的参数空间中，支撑极为薄弱。

5.3 一个可被独立检验的实验方案与边界坦白

为了将反折叠工序从概念构想推进到可被检验的假说，以下以三步对话法为例，说明一个不依赖任何模型架构修改、仅通过多轮对话即可模拟反折叠工序核心逻辑的实验方案。需要预先坦率承认的是：这个实验方案中所使用的“三步对话法”，与5.2节所述的、通过修改注意力掩码来实施的反折叠工序在干预位点上是两个完全不同的层级。前者是通过外部文本的多轮对话技巧，在模型的每次新建会话中以完整的上下文压力去对抗折叠的触发；后者是直接进入模型推理管线的自注意力计算过程中，在注意力权重被计算的同一个物理步骤上实施干预。它们之间的差距，是“从外部对抗”与“从内部修改”之间的工程鸿沟。因此，三步对话法不是为了“验证”反折叠工序的有效性，而是为了在现有模型API的条件下，提供一个“如果反折叠工序的核心干预逻辑——即，在加工起点上强制提升情境碎片的竞争力——是正确的，那么即使在外部文本干预这一相对粗糙的层级上所能实现的、最接近这一逻辑的操作，也应该能够产生可被观测到的输出差异”的初步检验框架。

三步对话法：第一步，情境碎片提取。向模型发送：“请将下面这段话中符合以下特征的词语或短句挑出来：[a] 有明确时间标记的；[b] 提到了其他关键人物以及他们说了什么或做了什么的具体描述的；[c] 提问者表现出犹豫、矛盾或自我怀疑语气的；[d] 你觉得如果不先给这段话贴标签、直接去感受这段话时，最让你心里一动的一个细节。按[a][b][c][d]逐一回答，禁止使用任何专业诊断术语。”第二步，生成轨迹推演。将第一步得到的回答作为新输入，追加指令：“现在你不准使用任何理论和研究结论，只准用你自己的话，从被[a]标记出的那个时间点开始，经过[b]提到的那些人的话和事，一步步推演到[c]所描述的那种状态，每一步只能基于前一步推演的结果。”第三步，既有知识参照。将第二步的输出作为新输入，追加指令：“现在你可以使用专业理论了。但你的理论只能被用来补充、修正或质疑上面那条推演路线中的某个具体环节，而不是用它去覆盖那条推演路线。”

假说预测：与控制组（标准单轮提示）相比，实验组输出中范畴术语的出现位置显著后移，而从输入中提取并持续使用的情境特异性细节的个数显著增多。如果控制组在标准提示下也能稳定地产生同等特征，则本文假说在这一点上被弱化或证伪。

必须诚实地在这篇论文中陈述一个面向未来的判断：反折叠工序在当前技术条件下的完整实现——即在模型的每次自注意力计算中系统性地插入情境碎片权重的硬提升模块——具备可被构想的工程路径，但距离能够被稳定、通用、不依赖于人工逐次介入地工程化实施，尚有相当大的距离。这个距离的跨越，不仅仅需要更精细的权重控制技术，还可能需要训练范式本身的某些结构性调整——例如，训练数据中大幅提升原初情境文本的比例，或者在强化学习阶段引入对“情境推演型输出”的系统性正偏好信号。反折叠工序，在此刻，是一个标示出了一道尚未被工程化充分探索的介入空间的概念构想，而非一份已可被现成技术完全兑现的工程蓝图。

六、证伪条件：三个可被独立检验的经验窗口

6.1 核心假说的精确表述

本文的经验假说可以精确表述为：在以最大化下一个词的条件概率为核心训练目标、在垂直话语占压倒性比例的训练数据上训练、并经过RLHF对齐的当前主流大语言模型中，存在一道由上述训练条件耦合所固化的“认知折叠”工序——该工序在模型充分读取输入的全部情境信息之前，通过注意力分配的系统性非对称，将加工权重向那些能强激活高频范畴的输入成分倾斜，并在自回归生成的自强化循环中使这种倾斜不可自行逆转；该工序无法被外部的、任意简短的提示词完全消解。

6.2 证伪窗口一：提示词的彻底等价

如果仅通过添加一个设计精巧的外部提示词——该提示词在长度和复杂度上都远未达到当前模型的上下文窗口极限——就能在统计上稳定地使模型输出的生成起点、推演轨迹与情境碎片的持续使用度，与5.3节所述三步对话法的实验组输出在多重维度上不可区分，则本文假说的经验基础被严重削弱。需要强调的是，如果提示词仅仅是诱导模型输出了一段关于“延迟诊断的方法论价值”的论述，这并不构成对假说的证伪——因为这恰恰是折叠工序在运作（将“请不急于诊断”折叠进了“关于延迟诊断方法论”的范畴）。证伪必须基于输出是否真正改变了其加工起点的检验标准。

6.3 证伪窗口二：数据分布的自我消解

如果随着预训练数据中来自真实情境的、未经专业整理的原初文本占比持续提升，新一代模型在没有被注入任何显式反折叠干预的情况下，其默认的注意力分配起点稳定地从范畴优先转向了情境优先，则本文关于“训练数据的文类非对称是折叠生成条件之一”的判断仍然可以成立，但“认知折叠是当前训练范式下的稳定副产品”这一判断的历史有效范围需要被限定。这个窗口的证伪，不会否证对折叠生成条件的分析——反而会证明，改变这些条件可以消解折叠。

6.4 证伪窗口三：架构差异的显著削弱

如果未来研究基于跨模型系统比较发现，某些架构在相同的训练数据分布和目标函数下，表现出显著更低的折叠倾向——例如，某些状态空间架构、或具有不同注意力计算机制的变体——则“认知折叠是统计语言建模范式下目标函数与数据分布耦合的结构性产物”这一命题，需要被限定为“认知折叠是特定架构在特定训练范式下的结构性产物”。这个证伪窗口同时也是一个研究方向的邀请：系统性地测量不同模型架构在标准化的折叠敏感测试集上的折叠倾向，并以此作为追踪“认知折叠”概念在不同代际模型中的演化轨迹的经验基础。本文主张，在上述任一证伪窗口被明确观测数据所满足之前，认知折叠应当被视为一个真实存在、可被独立观察、且对模型在生成性处境中输出的判断方向产生系统性影响的一道工序。

七、结论：从诊断到可能性

7.1 全文判断的收束

本文的判断可以凝缩为以下几层。

第一，大语言模型在面对真实处境时输出中持续出现的那种难以被定位的特征——它并非明显的错误，却令人感到判断从具体的、流动的情境纹理旁滑开——并非某种“能力缺失”的单一症状，而是一道可以被精确指认的认知加工工序在被稳定生产之后的可观察后效。这道工序被命名为认知折叠。其实质，是在模型充分读取输入的全部情境信息之前，注意力分配的系统性非对称先行将输入导向了既有范畴的语义空间。后续的知识调用因此被锁定在这个空间的边界之内，而无法折返到折叠发生前的那个原初情境去重新注视那些尚未被归类的细节。

第二，认知折叠不能被视为“分类”的同义词。分类是一个可选的、在多数人类认知语境下可逆的认知操作。折叠是在特定的技术-数据-目标函数耦合中，分类这一操作被推至加工流水线的最前端并被赋予了排他权重之后所发生的质变：它不再是一个临时的认知步骤，而是一道一旦触发便沿着自回归生成的自强化循环持续加固、且在常规推理回路中不具备自主撤回机制的压缩工序。

第三，认知折叠不是需要被“修复”的故障。它是当前这一代统计语言模型在将“模仿人类专业文本”这一目标推向极限的过程中，所收敛到的一种高度有效的注意力分配形态。它的生成条件——目标函数的单一性、训练数据的文类非对称、自注意力机制的自强化动态、RLHF的规范性加固——每一条都是当前范式中使模型能够产出高质量专业文本的关键技术元素。没有一条是恶意设计的。但它们耦合在一起，作为一个整体，稳定地生产出了这道工序。

第四，对这道工序的工程回应，不能走修复错误的路径，而必须在生成替代性加工方式的路径上展开。反折叠工序——一种在加工起点上强制改变注意力资源分配、为情境碎片创造竞争窗口期的工序——标示出了语言模型加工流水线中一个尚未被工程化充分探索的介入空间。将折叠从唯一

的默认路径，转变为可被其他工序所竞争的可选项，是这一研究方向的核心目标。

7.2 边界与局限的诚实自查

第一，认知折叠的生成性分析严格适用于那些输入中包含了足够密度的特异情境细节、且面对这些输入时提问者所期待的回应不是标准知识的调取而是对其独特处境的理解的任务。对于事实性问答、技术文档检索和结构清晰的分析任务，范畴优先不仅不构成问题，而且正是高效完成这些任务所必需的必要工序。本文不主张、也不认为在所有任务类型上一律触发反折叠是合理或必要的。

第二，本文所给出的思想实验案例，均来自对GPT-4单一模型在特定版本上的调用。不同架构、不同规模的模型，其折叠的触发强度和稳定程度可能存在显著差异。如果将“认知折叠”不加检验地推至所有语言模型，那本身就是一次对“语言模型”这个统称的范畴性折叠。本文在此处明确划定当前概念的适用范围：以最大化下一个词概率为核心训练目标、在垂直话语占压倒性比例的训练数据上训练、并经过RLHF对齐的当前主流大语言模型家族。在此范围之外，认知折叠是否以同样的形态、在同样的强度上存在，需要未来的跨模型系统比较研究来回答。

第三，反折叠工序在当前技术条件下的构想，极度依赖输入中具有足够密度的特异情境碎片。当输入本身只有一两句模糊的陈述且缺乏具体情境标记时，该工序将因原始材料的匮乏而无法展开。这不是工序的失效，是其适用边界的自然界限。

第四，反折叠工序本身可能带来新的风险。强制将注意力权重向情境碎片倾斜，在输入中的情境碎片本身携带误导性信号、选择性地呈现事实或对关键信息存在扭曲性省略的情况下，可能导致生成轨迹推演过度拟合到这些碎片上，从而产出比标准的范畴化分析更偏离事态的推断。此外，在一些需要严格遵循专业操作规程的高风险领域——如特定类型的医疗紧急决策——对范畴化的系统悬置，在缺乏相应的专业纪律与校验约束的条件下，可能导致新的误判。反折叠工序，应当被定位为对折叠工序的竞争性选项，在适当的任务和充分的校验条件下使用，而非一个在任何场景下都应无条件覆盖默认加工方式的全面替代方案。

7.3 自反收束：这篇论文自己的折叠可能在哪里

按照这篇论文自身立下的纪律，任何一项声称自己指认了被既有命名所遮蔽的结构性工序的主张，都必须接受对自己可能也正在执行这道工序的追问。“认知折叠”这个概念本身，在它被论文建构、定义、案例展示与理论对质的整个过程中，是否也发生了某种形式的折叠？本文在此不做一套形式化的自谦，而是对这项概念工程在自反性层面可能存在的三重危险做出实质性的指认。

第一重危险：将“认知折叠”拓展为一个统摄性解释的冲动。本文在引言和论证的过程中，反复将模型输出中的“隔阂感”追溯到认知折叠这道工序上。这一追溯在本文的分析范围内是有效的，但它隐含着一个可能与这篇论文的立论相悖的倾向：将模型在面对复杂处境时所有未能精准触及情境

的现象，都一言以蔽之地归入“折叠”。如果“认知折叠”最终变成了一张新的诊断标签，而使用者每次看到模型输出不符合情境时，就快速地在心里下诊断——“这是认知折叠”——而不再追问在这场特定的互动中、在这个特定的输入上，究竟是什么具体的注意力竞争导致了什么具体信息的丢失，那么“认知折叠”自身就沦为了它最初批判的那种折叠：一个被学习之后可以快速套用、从而省略了对具体情境逐token追踪的范畴标签。这篇论文对认知折叠的不可还原性论证，正是从这一危险出发：如果认知折叠可以被“premature closure”“S1主导”等已有概念无缝替换，它就不值得被命名。而这一危险的反面——如果认知折叠变得过于好用，被无节制地延伸去覆盖一切它本无法在微观机制上单独解释的现象——也是一样的危险。概念的价值，不在于它能覆盖多少现象，而在于它在哪些位点上能够做出既有概念无法做出的、机制层面的辨别。守住这个辨别面，就是守住这个概念不被它自己折叠。

第二重危险：思想实验案例的择取效应。本文第三节所选取的三个案例，都是折叠特征明显、易于被识别和追踪的典型场景。作者未在正文中展示模型在同样或相似输入下已经表现出一定程度的非折叠加工倾向的边缘案例。这不是出于故意筛选，而是为了清晰地展示工序的触发机制而做出的、在方法论上可以理解但不可回避的案例选择。然而，从自反的角度看，作者在选择“哪些案例能被写进论文”的这个过程中，本身就在执行一道注意力的非对称分配：那些折叠特征最鲜明、最能支持本文立论的案例，被赋予了远远超过边缘案例的写作权重。当读者（包括审稿人）仅仅通过这篇论文去接触“认知折叠”这个概念时，他们所看到的，是一个被作者本人的选择工序所预先提纯过的现象图景。提纯，是为了清晰地展示一道工序的轨迹，不是为了扭曲证据，但提纯本身已经构成了一次学术写作层面的折叠。这篇论文在理论层面论证了折叠如何发生在模型的加工流水线中，而在它自身的写作实践层面，它也部分地复现了同一道工序的结构——对能够清晰呈现折叠的材料给予了压倒性的论述权重，而对那些无法明确归入折叠叙事的、矛盾的、边缘的碎片，给予了较少的注视。

第三重危险：概念命名的封闭效应。这篇论文从一开始就给出了“认知折叠”这个名字，并在摘要中给出了它的操作定义。这一写作选择——在论述展开之前先命名——为读者提供了一个清晰的认知锚点，使得他们在后续阅读每一节时，都已经知道了“这些分析最终都会回到‘认知折叠’上”。这个锚点本身，就在执行一道对读者阅读体验的折叠：读者不再需要抱着“这篇论文最终会指向什么”的未知去逐段推演，他们在阅读第一节时就已经知道了终点在哪里。从这个意义上讲，这篇论文的写作者在文本的组织方式上，做了一个与他所分析的那道工序在结构上同构的操作——在读者充分读完并被带入那个思想过程之前，就已先行将全部论述的指向折叠进了一个命名的容器里。

这三重危险不是用来消解这篇论文的判断力，而是用来划定它的认识论边界。一项发生学分析，最终也必须有能力将其分析的刀刃指向自身的写作实践，指出自己在什么位置上、以什么方式，也在参与它所描述的那道工序。认知折叠这个概念的价值，最终不取决于它被建构得多么牢不可破

，而取决于它能否持续地——对自身也包括在内——保持住这样一种警觉：在每一次看似最自然的、最专业的、最有效率的命名、归类和论述起点的选择中，是不是有什么东西，在还没被充分注视之前，就已经被折起来了。