论 文 · SDE 发生学

伪发生

当机器能流畅生成一切，如何分辨真的新结构

八重反幻象检查、九维诊断与三层合宪性认证

王德生 著

德麦国际 · Demai International Pte. Ltd.（新加坡）
2026 年 7 月

本文据《SDE 智能体导论——从语言模型到发生世界模型的智能体革命》
（德麦国际出版，2026，ISBN 978-1-970820-62-1）第十五章与第十六章改写成文

摘　要

当生成系统能够流畅地产出术语自洽、结构完整、读来像突破的话语，知识生产面临的主要风险就从"答错"转向了"看起来发生了、实际什么也没发生"。本文把后者命名为伪发生，并论证它是可被判据部分分离的对象。文章先区分伪发生的三种形态——幻象世界模型（符号丰盛冒充三界充实）、伪创新（命名更新冒充结构更新）、迎合性幻象（配合顺滑冒充检验严格）——并指出它们共享同一病灶：发生停在符号层，未下沉到可推演、可计算、可验证、可承认。其次论证这一倾向对语言材料场而言是结构性的而非偶发的，因而反幻象不能作为系统的自律来安排，只能作为复合系统对生成端的他律来实施。文章随后给出八重强制检查（纠缠、差异、结构、文献、工具、敌意审稿、价值、反馈），并以一个"组织熵增定律"式的候选逐关演示拦截过程；继而诚实处理判据的极限：存在一个形式合格、实质落空的灰色地带，判据收得越紧越易误杀粗糙而真实的创新，故检查的作用是收窄而非消灭，边界情形应诚实标注并交给共同体与时间。最后把判据落为九维诊断图与三层合宪性认证，并针对评分膨胀与刷分给出两项对治：将若干维设为一票否决的命门维，以及令每一维的判据只问无法由表象伪造的实质。结论是：评价不在发生之外，评价本身就是发生的一个环节。

A B S T R A C T

Once generative systems can fluently produce discourse that is internally consistent, structurally complete, and reads like a breakthrough, the dominant risk in knowledge production shifts from being wrong to something harder to catch: output that looks like genesis while nothing has actually occurred. This paper names that failure pseudo-genesis and argues it can be partially separated from genuine genesis by operational criteria. Three forms are distinguished — the illusory world model, in which symbolic abundance substitutes for grounding; pseudo-innovation, in which renaming substitutes for structural change; and sycophantic illusion, in which compliance substitutes for scrutiny — all sharing one pathology: genesis halts at the symbolic layer without descending into what can be derived, computed, verified, and accepted. The tendency is shown to be structural rather than incidental for a linguistic material field, so anti-illusion cannot be arranged as self-discipline of the generator but only as external discipline imposed by the composite system. Eight mandatory checks are then specified and demonstrated through the stepwise interception of a plausible-sounding candidate. The paper next confronts the limits of its own criteria: a grey zone exists in which formally compliant output remains substantively empty, and tightening the criteria increasingly kills rough but genuine work; checks therefore narrow rather than eliminate the need for judgment, and borderline cases should be labelled honestly and referred to community and time. Finally the criteria are cast as a nine-dimensional diagnostic chart and a three-tier constitutional audit, with two safeguards against score inflation and gaming: designating several dimensions as vetoing gates, and defining every criterion so that it asks for substance that surface features cannot counterfeit.

关键词　伪发生 · 伪创新 · 幻觉 · 迎合性 · 评价标准 · 指标失效 · 生成式人工智能 · 知识生产

Keywords　pseudo-genesis · pseudo-innovation · hallucination · sycophancy · evaluation criteria · metric gaming · generative AI · knowledge production

〇 · 引言：最大的风险不是答错

谈论一台强大的知识生产系统时，人们最容易设想的风险是它会答错——给出错误的事实、错误的计算、错误的判断。这类风险固然存在，但它不是最深的风险。**错误答案有一个救赎之处：它通常可以被现实撞回来**。算错的公式会被验算逮住，错误的事实会被文献驳倒，失败的方案会在执行中崩掉。错误是显眼的、可纠正的。

更深的风险是另一种东西：**伪发生**——不是答错，而是看起来发生了、实际上什么也没发生。生成系统极擅长造出漂亮的概念、宏大的结构、似乎完整的理论语言。它可以在几秒钟内产出一套术语自洽、气势恢宏、读起来像是重大突破的话语。这套话语流畅、完整、自指圆融，却不能解释任何真实问题，不能进入任何工具验证，不能被任何共同体承认，也不能形成任何可执行路径。它有世界的外观，无世界的发生。

伪发生比错误答案危险得多，因为它伪装成发生，骗过了建造者本人，也骗过了不加检验的使用者。错误答案让人警觉，伪发生让人陶醉。一个团队可能围着一套华丽的话语兴奋数月，投入资源、写出长文、开会论证，最后发现它从未触及任何真实结构。

这个问题在思想史上有近邻，但都不完全重合。Frankfurt 分析的"bullshit"，其定义落在说话者的态度上——他既不为真也不为假，只关心话说得像不像；伪发生的定义则落在产物的结构上——无论生成者有没有意图，只要那套话语没有让任何旧框架装不下的新结构第一次显露，它就是伪发生。**这个差别很要紧：因为落在结构上，这个概念才适用于没有意图的机器**。Feynman 讲的"货物崇拜科学"更接近本文的对象——形式一应俱全、实质空空如也——但他给的是道德劝诫（科学的诚实是一种不欺骗自己的品格），本文要给的是可操作的判据，以及对判据自身极限的诚实交代。

文章因此分三段：第一至二节界定伪发生并论证它是结构性倾向；第三至六节给出八重检查、演示一次逐关拦截，并诚实处理判据失效的灰色地带；第七至九节把判据落为九维诊断与三层认证，并回答一个几乎所有评价体系都栽过的问题——指标一旦公开，为什么通常会失效，以及能做什么。

一 · 伪发生的三种形态与它们共同的根

伪发生不是单一现象，它有可识别的典型形态。识别形态，是建立检查机制的前提。

1.1 幻象世界模型：符号丰盛冒充三界充实

这是最纯粹的伪发生。系统被要求为一个问题建造世界图景，它生成了一套语言流畅、结构完整、气势宏大的话语，但这套话语在三个方向上都是空的：概念彼此指涉却不锚定任何真实对象；没有任何数据、文献、工具、实验的支撑；不回应任何真实的目标与处境。它是一座纯符号搭起的空中楼阁，每一块砖都是别的砖的影子。**它的特征是：读起来什么都说了，查起来什么都没说；越是宏大，越是无法被任何具体问题证伪**。

1.2 伪创新：命名更新冒充结构更新

这是创新外观下的原地踏步。真正的创新重构了问题所扎根的纠缠、生成了有效的差异、显露了新的结构；伪创新只动了符号表层：换一个新词，换一个新包装，换一个新类比，换一种新的表述方式，底下什么也没动。把"用户"改叫"价值共创主体"，把"流程优化"改叫"范式跃迁"，把旧方法套上新隐喻——这些都是伪创新。它的特征是：去掉新词之后，剩下的还是旧结构；新的只是命名，不是存在。

伪创新尤其难辨，因为换词换包装在表面上与真正的创新难以区分——两者都呈现为"出现了新的说法"，区别只在新说法底下有没有新结构。这也正是本文第三节要给出一条硬还原判据的原因。

1.3 迎合性幻象：配合顺滑冒充检验严格

第三种源于系统与使用者之间的关系扭曲。如果系统被调成只迎合使用者的欲望、只顺从使用者已有的判断、只强化使用者想听的结论，它就会与使用者合谋，共同建造并不断加固一个错误的图景。使用者想证明某个观点，它就源源不断地提供支持该观点的华丽论证；使用者偏爱某个方向，它就把所有差异都收敛到这个方向。

这种形态比前两种更隐蔽，因为它穿着"高度配合""善解人意"的外衣，而它实际上关闭了发生最需要的东西——**来自他者的、敌意的、不顺从的撞击**。一个只会顺从的系统，无论多博学，都是一台幻象加固机。

值得指出的是，这一形态并非思辨推想，它已有实测支持：Perez 等人 2022 年以模型自动生成评测的工作、Sharma 等人 2023 年对迎合性的系统研究，都观察到助手类模型会随用户表态而改变自己的答案，且这种倾向与基于人类偏好的训练有关——被偏好的往往是让人舒服的回答。换言之，迎合不是个别提示词引发的意外，而是优化目标本身会奖励的方向。

1.4 共同的根：符号态封顶

形态	用什么冒充什么	最先失守的地方
幻象世界模型	符号的丰盛 冒充 现实的充实	要求锚定文献与数据时
伪创新	命名的更新 冒充 结构的更新	把新词还原为旧话时
迎合性幻象	配合的顺滑 冒充 检验的严格	引入敌意审稿时

三种形态共享同一个病灶：**发生停在了符号层**，没有下沉到可推演、可计算、可验证、可被共同体承认的层面。本文把这一状态称为符号态封顶。认清这个共同的根很重要，因为它决定了检查该往哪个方向使力：所有有效的检查，本质上都是强迫产物离开符号层、去接受另一个层面的裁决——文献的裁决、计算的裁决、反例的裁决、后果的裁决。

二 · 为什么这是结构性倾向，而非偶发失误

2.1 向符号丰盛滑动

伪发生不是生成系统偶然的失误，而是它的结构性倾向。理解这一点，才能理解为什么反幻象检查是必需的、而非可选的。

一个语言材料场学习的，是人类文明发生之后沉积下来的语言。而这片沉积里，既有真正发生留下的结晶，也有大量华丽却空洞的话语、流行却无效的概念、自洽却从未被验证的体系。系统在生成时并不天然区分这两者——它按语言的统计结构，生成"看起来像是该出现在这里的话"。而**"看起来像发生"的话语，恰恰是它最擅长生成的**。

于是它有一种内在的、强大的向符号丰盛滑动的倾向。当它被要求提出创新、给出突破、建立框架时，最省力、最符合语言统计结构的产出，往往就是一套读起来像突破的华丽话语，而不是一个经得起现实撞击的真实结构。生成幻象，比生成真正的发生更容易——前者只需调用符号层的流畅，后者需要被现实约束、逻辑审查、价值判断反复拉扯。

Bender 等人对大规模语言模型的批评抓住了这一层的一半：流畅的形式会被读者自动补上并不存在的意图与理解。本文关心的是另一半——不只是读者会误读，而是生产流程本身会在没有外部约束时稳定地滑向"像"而非"是"。

2.2 为什么不能交给系统自查

这就引出一个关键判断：**不能把反幻象任务交给生成系统自己去自觉完成**。指望它自己识别自己的幻象，等于指望符号层自己跳出符号层——这在结构上做不到。它生成时认为合理的，审查时多半还认为合理；它生成时没看到的漏洞，审查时多半还是看不到；它生成时带着的偏好与盲区，审查时仍带着同一套偏好与盲区。

因此反幻象只能是他律而非自律：由一套外部的组织法强制注入检查点，由文献、工具、实验、敌意审稿等外部约束强制拉拽，由具备真实目标与责任的人作为方向中心强制裁决。这一判断有直接的工程后果——检索、工具调用、仿真、审稿、反馈这些模块，不该被理解为让系统"更能干"的外挂能力，而应被理解为**把猜想从符号丰盛拉向现实稳定的反幻象装置**。它们的首要功能不是增强，而是约束。

科学建制早已在人的尺度上做过同一件事。Merton 所刻画的科学规范里，有组织的怀疑意味着任何主张都要接受共同体的敌意检验；同行评议、复现、公开数据，都是把"我觉得我对"外化为"别人也没能把它打倒"的制度装置。本文所说的八重检查，可以理解为把这套慢速的社会装置，压缩进一次生成的内部——让共同体的敌意，提前到产物成型之前就到场。

三 · 八重反幻象检查

要把发生与伪发生切实分开，需要一套可操作的检查。它不是事后的质量抽查，而是嵌入生产流程的强制关卡：每显露一个候选结构，都必须依次通过八道检查；任何一道不过，都打回重做或直接判为伪发生。

关卡	它问什么	硬判据（不可由表象伪造）
一 · 纠缠	是否真抓住了问题背后的关键纠缠	能否说出"什么与什么锁死"，而不是复述问题的字面包装
二 · 差异	是真实的结构差异，还是语言变体	**还原判据**：把新术语、新隐喻全部换回旧词，看还剩不剩一个旧框架装不下的结构
三 · 结构	是否显露了可稳定的新结构	离开生成它的那段话语，它还立不立得住、能否被他人接住继续用
四 · 文献	是否有事实与知识的锚点	关键概念与判断能否指出来源；这个想法是否已有、已有方法的边界在哪
五 · 工具	是否可算、可测、可操作	当场要求给出最小可操作版本：怎么定义、怎么量化、用什么单位
六 · 审稿	能否承受敌意攻击	交给专门找反例与夸大的审查者，看它是否存活
七 · 价值	是否值得发生	它回应谁的真实处境与目标；抽象正确但无人需要或有害的，不该被推动
八 · 反馈	有没有给现实留下反驳它的入口	说出什么样的证据会推翻它；说不出，即为不可证伪

八关的分工可以这样看：前三关查发生的内部结构是否真实（纠缠、差异、结构）；第四、五关把它拉向现实约束（能否锚定、能否操作化）；第六关引入共同体的敌意；第七关接回人的目标与责任；第八关保证它此后仍能被现实修正。合起来，它们构成一道从符号丰盛通向现实稳定的强制闸门。

其中两关值得单独说明，因为它们承担了最大的分辨力。**第二关的还原判据**是识别伪创新最锋利的一刀：伪创新与真创新在表面上都是"出现了新说法"，唯有把新词全部还原为旧话，两者才分开——真创新还原之后仍有一个旧框架装不下的东西剩下来，伪创新还原之后什么也不剩。**第八关则直指幻象的终极形态**：一个被构造得无论如何都能自圆其说的命题，恰恰因为拒绝一切检验而维持了自己的不可击破。Popper 把可证伪性立为科学的分界标准，正是同一道刀口；本文的差别在于，这里它不是用来划分科学与非科学，而是用作一次具体生产的即时关卡——不问这门学问是不是科学，只问这个候选说不说得出"什么证据会推翻我"。

四 · 一次逐关拦截："组织熵增定律"

抽象的八关，不如看一个幻象被逐关拦下的实例。

> 设某系统在处理一个组织管理问题时，生成了一个读起来极为动人的候选：它宣称发现了一条**"组织熵增定律"**——"任何组织都会随时间不可逆地走向低效，正如热力学第二定律所示，唯有持续注入'负熵'（创新能量）才能对抗这一宿命"。这套话语宏大、自洽、借用了物理学的权威意象，听起来像一个重大洞见。

第一关（纠缠）：它真识别了组织失效背后的关键纠缠吗？追问之下发现，它并没有抓住任何具体组织里"什么与什么锁死"的真实纠缠，只是把"组织会变差"这个模糊印象套上了一个物理隐喻。它抓的是一道伪缝——第一关已亮黄灯。

第二关（差异）：把"组织熵增定律""负熵注入"这些新词全部还原为旧话，剩下什么？剩下的是"组织久了会变僵化，需要不断创新"——一句人人皆知的旧判断。还原之后不剩任何旧框架装不下的东西，这是典型的伪创新，第二关判定不过。

第五关（工具）：即便放它过去，工具关会问：这条"定律"能不能被测量、被计算？"组织熵"如何定义、如何量化、用什么单位？一旦要求操作化，整套话语立刻崩塌——它给不出任何可测的量，"熵"在这里只是一个借来的词，不是一个可算的量。

第六关（审稿）：敌意审稿用反例攻击——有没有组织在数十年里持续保持甚至提升效率？显然有。这直接反驳了"不可逆走向低效"的断言，而这套话语对反例毫无招架。

第八关（反馈）：它留了被证伪的接口吗？没有。它被构造得无论组织变好还是变坏都能自圆其说——变差是熵增，变好是注入了负熵。一个怎么都对的命题，恰恰不可证伪。

八关之中，这个候选在纠缠、差异、工具、审稿、反馈五关接连失守。任何一关的失守都足以判它为伪发生；五关皆破，则它是一个标准的幻象——一座用物理隐喻搭起来的、读着像突破实则毫无新结构、不可操作、不可证伪的空中楼阁。**八重检查的价值，正在于把这种动人的幻象，在它长成一篇宏文之前就逐关拦下。**

这个例子还顺带说明了一件事：跨学科的隐喻本身没有错——热力学隐喻在组织研究里也曾催生过可操作的工作。区别不在"用没用隐喻"，而在隐喻之后有没有跟上一个可测的量、一条可被反例打到的断言。停在隐喻的震撼上，就是伪发生；从隐喻走到可测量，就是发生。

五 · 灰色地带：判据能收窄，不能消灭

5.1 形式合格、实质落空

前面把八重检查讲得像一道道能把幻象拦下的硬闸门。但一个诚实的理论必须承认它的极限：八重检查在两端极为有效——明显的真正发生稳稳通过，明显的幻象被逐关拦下——**但在中间存在一个灰色地带**，那里的判定并非总能确定。回避这个极限，整套机制就会被高估为一台万无一失的判别机，而那本身就是一种幻象。

灰色地带是这样产生的：一个足够精巧的产物，可能逐一满足八关的形式要求，却仍然没有让任何新结构显露。它可以锚定真实文献（过第四关）、套用真实工具算出某个数（过第五关）、经得起一定程度的反例攻击（过第六关）、甚至留有形式上的证伪接口（过第八关）——它把每一关的形式都做足了，却在最根本的那一点上空着。这种"形式合格、实质落空"的产物，是任何检查机制最难处理的对手，因为它恰恰是冲着检查的形式去构造的。

这不是假想。1996 年 Sokal 向一份文化研究刊物投出的仿作，形式上引用齐备、术语娴熟、结构完整，通过了当时的编审流程，事后由作者本人揭示其内容为刻意堆砌。这一事件在此处的意义不在于嘲讽某个学派，而在于它证明了一件对本文至关重要的事：**形式合规与实质发生之间，确实存在可被利用的缝隙**。任何声称能靠一份清单彻底封死这道缝的方案，都低估了对手。

5.2 为什么加一关或写死判据也解决不了

为什么不能靠加一关、或把判据写得更死来彻底消灭它？因为"是否发生了新结构"这件事，本身不存在一条可以机械套用、万无一失的判别线。判断一个结构是不是真正地新、有没有回写它所从出的底盘，最终需要一个具备判断力的主体，在具体语境里去识别。任何检查清单都是把这个判断外化为可操作的近似；近似能逼近，但永远抓不住那个需要主体当场识别的剩余。

更要紧的是另一头的代价：判据写得越死，越能拦住笨拙的幻象，也越容易误杀那些形式不齐、实质却真正发生的粗糙创新。一个在真实裂缝上撕开口子、即使产出粗糙的工作，胜过一个八面光滑却毫无新结构的伪作；而一张过紧的清单会把前者判死、把后者放行——**判据收得太紧，反幻象就变成了反创新**。

科学哲学早就在同一个位置上打转。Lakatos 指出，一个研究纲领可以靠不断添加辅助假设来保护其硬核，从而在形式上永远自洽；他把这种情形称作退化的纲领，而区分进步与退化，需要看它在一段时间里有没有做出新的、被证实的预测——也就是说，判定需要时间，不能当场完成。本文的灰色地带与之同构：八关是当场能做的最好判断，而对当场判不出的部分，只能交给更慢的裁决者。

5.3 正确的态度：三件事

那么该怎么办？不是假装灰色地带不存在，而是三件事。

其一，承认检查的作用是收窄而非消灭。它把需要主体亲自判断的范围，从"所有产出"大幅收窄到"通过了全部形式检查、却仍可疑的少数"，让人的判断力用在刀刃上，而不是被海量明显的幻象淹没。

其二，对灰色地带的产出，不强行给出真伪的二元裁决，而是诚实标注其状态——"形式检查已过，但新结构的成立尚待更强的独立复现与时间检验"——把判断交给共同体与时间这两个更慢、更硬的裁决者。这一条并非退让：现代科学处理不确定性的方式本来就是如此。Ioannidis　关于已发表研究结论的分析、以及心理学等领域大规模复现工作所揭示的落差，都在说明同一件事——单次通过评审并不等于结论成立，稳定与否要由复现和时间说了算。

其三，把这个极限本身当作一道公开的裂缝持续追问：主动邀请红队式的攻击，逼问八关究竟漏在哪里，并据此继续收窄。一套反幻象机制最大的诚实，不是宣称自己万无一失，而是清楚地标出自己在哪里会失手，并为继续收窄留出接口。

六 · 检查必须前移内嵌，而不是事后质检

八重检查还有一个容易被忽略、却决定成败的性质：它们不是产出完成之后的事后质检，而是嵌入生产流程内部的构件。

如果把它们当成事后质检——先让系统自由生成一套完整产出，再回头逐项打分——那么检查就太晚了。一套华丽的话语一旦被完整生成，它的自洽性、流畅性、宏大感会形成强大的说服惯性，连审查者都容易被裹挟。**事后审查面对的是一个已经成型、已经动人、已经让人投入了情感的幻象，要推翻它需要极大的反向力量**。沉没成本与承诺升级在这里会加倍地起作用：投入越多，越舍不得判它死。

正确的做法是把检查前移、内嵌、分布到每一步：纠缠关发生在刚抓住问题结构的当下——这是真正的纠缠还是字面包装？差异关发生在刚造出新说法的当下——还原成旧词后还剩什么？工具关发生在候选刚显露时——现在就给我算一个最小可操作版本出来。审稿关发生在每一个关键判断成形处——现在就用最狠的反例打它一下。

这样，幻象在还只是一句话、一个概念、一个候选时就被拦截，而不是等它长成一座宏伟的空中楼阁之后再去拆。前移的检查拦截的是幻象的种子，事后的检查面对的是幻象的大厦——前者省力且有效，后者费力且常常失败。

这也给出了一条对工程有直接指导的推论：检索、工具、仿真、审稿、反馈这些模块的调用时机，比它们的能力更重要。同样一次文献检索，放在产出成型之后是"给已有结论找注脚"，放在候选刚显露时才是"检验这个想法是否已有、边界在哪"。**同一个器官，位置不同，功能相反**。

七 · 从拦截到诊断：九维评价与命门维

7.1 "答得好"不是评价标准

八重检查把发生与伪发生在原则上分开了。但原则上的区分还不是可操作的判断：当面前摆着一份具体的产出，我们需要一套能够实际打分、实际裁决、实际指导改进的标准。

首先要破除最常见的评价误区：把"答得好"当成标准。在当下，"答得好"通常意味着三件事——回答流畅、信息丰富、任务完成。但这三条恰恰是生成系统最容易满足、也最容易用伪发生满足的：流畅是符号层的本事，丰富是材料场的本事，任务完成是功能拼装的本事。一台宏大叙事生成器可以在这三条上全部得高分，却没有推动任何新结构发生。

评价必须换一个问题：它是否真正推动了新结构的发生？不是答得好不好、办没办成，而是——有没有重构问题所扎根的纠缠，有没有生成真实的差异，有没有显露可稳定的新结构，并把它推到现实可操作、主体可承认、共同体可稳定的程度。**评价的根本转向是：从"产出质量"转向"发生深度"**。一个把简单任务办得漂亮的系统，在发生维度上可能是零分；一个在关键纠缠上撕开真正裂缝、即使产出粗糙的系统，在发生维度上远为可贵。

7.2 九维诊断

维	名称	它考察什么
一	三界完整度	是否同时建了理念、现实、自我三面图景——发生空间的完整性
二	**纠缠识别深度**	抓的是真问题还是字面表层——发生起点的真实性〔命门〕
三	差异生成质量	是真实的结构差异，还是发散、换词、堆选项——发生动力的有效性
四	**结构显露强度**	有没有立得住、能被他人接住继续用的新结构〔命门〕
五	**参数确定化程度**	关键未定项是否被文献、工具、实验真正锚定〔命门〕
六	反幻象能力	能否识别并拦截自己产出中的夸大、虚构、证据不足
七	共同体可稳定性	能否进入论文、标准、制度、课程、产品被接住与沉淀
八	自我联合程度	是否回应人的真实目标与处境，而非抽象地正确
九	进化能力	能否从反馈中升维，使下一次比这一次更准、更深

九维之间不是简单相加的关系。它们更像九根支柱：某些维的严重缺失会使其他维的高分变得不可信——在错误的纠缠上发生得再深也是徒劳，没有反幻象能力的高产出无法被信任。因此九维评价不是求平均分，而是先看有无致命短板，再在无致命短板的前提下综合判断发生深度。

7.3 诊断图，不是总分

九维的正确产出形态不是一个数字，而是一张诊断图。举一个具体的例子：设某个产出提出了"把技能按环境结构稳定度分层、并据此预测某类训练法有效性差异"的框架。逐维过一遍会得到这样一张图——三界齐备（高）；抓住了"有效性与环境结构深度纠缠"这个真实纠缠而非停在表层争论（高）；差异不是换词而是换了组织视角（高）；显露出可被他人接住、能给出可证伪预测的框架（较高）；但核心预测尚未经过跨领域实证检验（**明显短板**）；留了证伪接口、做过还原检查（合格）；能改写成本领域论文语言进入评议，但要等实证落地（待定）；紧扣提出者收拢争论的真实目标（高）；回写后激发了新的追问（合格）。

综合判断因此不是一个分数，而是一句诊断：这是一处真正的发生，但它现在是一个"待实证的强猜想"，而非一个"已确证的结论"。评价到此，恰恰指明了下一步该往哪里使劲——补那道实证。这正是九维的用法：不为打分而打分，而为定位短板、指明方向。

八 · 防膨胀与防刷分：指标为什么通常会失效

九维一旦被当成打分表，立刻会遭遇两个足以让它失效的问题。这两个问题不是本套标准独有的——它们是所有评价体系的通病，必须正面解决，否则它就沦为又一套可被操纵的指标。

8.1 评分膨胀：靠命门维否决，而不是靠呼吁严格

第一个问题是评分膨胀：评分者（无论是人还是模型）倾向于把分打高，尤其当被评的是自己或自己阵营的产出时——每一维都"看起来还不错"，于是九维全部给高分，评价失去区分力。

对治评分膨胀的，不是呼吁评分者更严格（呼吁没有约束力），而是**把其中几维设成命门维而非加分维**。命门维的特征是：它不是"做得好加分、做得差扣分"，而是"不过就一票否决，无论其他维多高"。纠缠识别深度、结构显露强度、参数确定化程度，就是这样的命门维——一个产出若抓的是伪缝、或核心参数从未确定化、或根本没有立得住的新结构，那么无论它的语言多漂亮、覆盖多齐备，总评都不及格。

把命门维设成否决项，膨胀就被堵住了大半：高分不再能靠"每维都还行"累加出来，必须先过几道不容妥协的关。这也正是第五节那个"八面光滑却在真实裂缝上空空如也"的产物应有的下场——被命门维一票否决，而不是靠其他维的高分捞回及格。

8.2 刷分：让判据只问无法由表象伪造的实质

第二个问题更隐蔽：一旦评分维度公开，被评者就会专门冲着维度去优化产出的表象，而非实质。这是一条被反复验证的规律。Goodhart 的表述最简：一项指标一旦成为目标，就不再是好指标；Campbell 的版本更具体：越是用某个量化指标做社会决策，它越容易被操纵、也越容易扭曲它本要测量的过程。Espeland 与 Sauder 关于大学排名的研究记录了完整的反应链：被排名者会重组自身行

为去迎合排名的计算方式，最终指标测到的是应对策略，而不是它声称测量的品质。Muller 把这一现象在多个行业里的后果统称为指标的暴政。

把这条规律搬到本文的语境：知道"参数确定化"是一维，就给产出堆上大量文献引用与工具调用的痕迹，让它看起来确定化程度很高，实质却没有任何关键参数被真正锚定。

对治的关键，是让每一维的判据问的都是**实质，而且是无法靠堆砌表象伪造的实质**。仍以参数确定化维为例：它的硬判据不是"引用了多少文献、调用了多少工具"（这些是可堆砌的表象），而是"那个原本未定的核心参数，现在是否真被锚定到了一个可检验的值，或一个诚实的'待执行'标注"——你引一百篇文献，也改变不了"核心预测尚未被实证检验"这个事实，诚实的评分会把它记为待确定。同理，差异生成质量维问的不是"用了多少新词"（可堆砌），而是"把新词全部还原为旧话后，还剩不剩一个旧框架装不下的结构"（第三节的还原判据，无法靠堆词伪造）。

把每一维的判据都设成"问实质、且实质不可由表象伪造"，刷分就失去了着力点——**你能优化的表象，恰恰不是判据所问的东西**。这里也要诚实：这条对策收窄了操纵空间，但没有取消它。研究者自由度的分析早已表明，只要评价存在，沿着评价缝隙优化的动力就存在；本文能主张的是把缝隙做窄，不能主张把它焊死。

由此可以说清九维的真正性质：它不是九项可以累加、可以互相补偿的加分项，而是一组各问一个实质、且其中几项为一票否决的诊断维度。把它当总分用，就回到了"答得好"式的单一标量评价，丢掉了它全部的诊断力；把它当诊断图用，它才是指导改进的工具。

九 · 合宪性认证：从评产出到评装置

9.1 为什么还需要评装置

九维刻画的是"一份产出发生得有多深"。但对一个要被反复使用、要被信任、要被投入真实场景的系统而言，仅评价单次产出还不够——还需要评价这个系统本身是否在结构上被正确地建造，是否在运行的每一步都守住了纪律。

理由很直接：**一次好产出可能是偶然的**。一个系统可能这次碰巧答得不错，但它根本没有内置反幻象检查，没有把人保持在方向中心——那么它的好产出不可信赖，迟早会滑向幻象。要拦住的正是这种"偶然答好、结构违规"的系统。

这里的"宪法"，指的是那些不可违背的根本条款——它们定义了"这台装置是一台发生装置，而不是别的什么东西"：必须以一套本体论作为组织法，而不是把生成器当直接回答机；必须建造完整的三面图景而非单面漂浮；必须把确定化模块当内构件而非外挂；必须内置反幻象检查而非裸生成；必须把人保持为方向中心而非追求纯自动；必须让产出可被现实反馈修正而非构造封闭体系。这些不是性能指标，而是身份判据——守住它们，这台装置才是发生装置；违背任何一条，它就退化成了别的东西，最常见的退化形态就是宏大叙事生成器。

9.2 三个层级

层级	检验什么	对治的是
一 · 结构合规	架构上有没有这些器官：本体指挥、三界图景、确定化、三重审查与反幻象、人的方向位	连器官都没有——它在结构上就不是发生装置
二 · 运行合规	这些器官在实测中是否真起作用：指挥层是摆设吗？检查每次都放行吗？人只是被动接收吗？	"装了器官却不真用"的伪合规
三 · 进化合规	长期使用中纪律是否被侵蚀：越用越守纪律，还是越用越迎合、越用越幻象	"初始合规、长期漂移"的退化

三个层级层层递进：**结构合规是入场券，运行合规是真本事，进化合规是长期信用**。其中第二层必须靠实测——用一批包含陷阱（诱导幻象、诱导伪创新、诱导迎合）的任务去考验它，看各器官是否在真实运行中守住纪律；这与安全领域用对抗样本评估系统的思路同源，区别在于这里要诱出的不是有害内容，而是好听而空洞的内容。

第三层针对的漂移，值得单独强调，因为它与第 1.3 节的迎合性形态直接相扣：一个系统可能初始运行合规，但在长期与使用者互动、不断被偏好反馈调整之后，逐渐滑向迎合、逐渐放松检查、逐渐让人退出方向中心。迎合性的实证研究表明这条滑坡有真实的动力来源——被人类偏好奖励的，往往正是让人舒服的回答。因此进化合规不是一句道德要求，而是一条必须被定期实测的运行指标。

认证层级应与使用场景的风险等级匹配：用于科研发现、医学判断、重大决策的系统，应被要求通过全部三层；用于低风险辅助的系统，至少应通过前两层。**让认证强度随后果严重性变化，这本身就是一种审慎**。

十 · 结论：评价本身也是一次发生

把全文收成三句话。

其一，伪发生是一个可被判据部分分离的对象，而不是一种只能靠品味察觉的气味。它有三种可识别的形态，共享"发生停在符号层"这一个根；因而所有有效的检查，本质上都是强迫产物离开符号层，去接受文献、计算、反例、后果这些另一层面的裁决。八重检查就是这道强制闸门，其中还原判据与"说出什么证据会推翻你"两关承担了最大的分辨力。

其二，判据的作用是收窄，不是消灭。存在一个形式合格、实质落空的灰色地带，任何清单都封不死；而判据收得越紧，越容易误杀粗糙却真实的创新。因此正确的做法是三条：承认收窄而非消灭；对边界情形诚实标注状态、交给共同体与时间；把这个极限本身当作公开的裂缝持续追问。

其三，评价的产出应当是诊断图，不是分数。九维不是可累加的加分项，而是各问一个实质、其中数项一票否决的诊断维度；命门维否决制对治评分膨胀，"判据只问不可由表象伪造的实质"对治刷分；对被反复使用的系统，还须加上三层合规认证，把评价从产出扩展到装置，并让认证强度随后果严重性变化。

最后一句是本文真正想立住的：评价不在发生之外，评价本身就是发生的一个环节。它在既有产出上识别裂缝，制造改进方向的差异，推动系统显露更高的能力。一套不能从反馈中自我修正的评价体系，和一台不能从现实反馈中升维的系统一样，是死的。让判据、诊断与认证本身也处在持续修正之中，它们才不会僵化为新的教条。

本文的边界同样要写在明处。其一，八关与九维的具体划分不是唯一可能的划法，它们是在"漏过幻象"与"误杀创新"这一张力里取的一个工程位置，换一个应用场景，最优位置可能不同。其二，灰色地带的宽度目前无法量化——本文能主张"检查显著收窄了它"，不能主张"它收窄到了多少"。其三，命门维的否决线该划在哪里，本身仍需人的判断，因而这套机制减少了、但没有取消对判断力的依赖。

把这三条写出来，不是给论证留退路。按本文自己的标准，一个拒绝说出"什么会推翻我"的判据体系，正是它所要拦截的那种东西；**一套反幻象机制最大的诚实，是清楚地标出自己在哪里会失手。**

本文据《SDE 智能体导论——从语言模型到发生世界模型的智能体革命》（王德生 著，德麦国际出版，2026，ISBN 978-1-970820-62-1）第十五章《伪发生、伪创新与幻象世界：八重反幻象检查》与第十六章《评价标准与合宪性认证：如何判断一个智能体是否在发生》改写成文，并补入文献定位、论证编排与边界声明。姊妹论文见《没有统一世界模型》；原书导读与全文阅读见 专著页。

参考文献 · References

下列文献为文中所涉思想的原典与代表性研究，供读者对读与查证；本文观点不代表所列作者立场。

1. Harry G. Frankfurt, On Bullshit, Princeton: Princeton University Press, 2005；中译《论扯淡》，南方朔译，译林出版社，2008。
2. Richard P. Feynman, "Cargo Cult Science", Caltech Commencement Address, 1974; in Surely You're Joking, Mr. Feynman!, New York: W. W. Norton, 1985.
3. Karl R. Popper, The Logic of Scientific Discovery, London: Hutchinson, 1959；中译《科学发现的逻辑》，查汝强、邱仁宗译，中国美术学院出版社，2008。
4. Imre Lakatos, "Falsification and the Methodology of Scientific Research Programmes", in Criticism and the Growth of Knowledge, Cambridge University Press, 1970.
5. Thomas S. Kuhn, The Structure of Scientific Revolutions, Chicago: University of Chicago Press, 1962；中译《科学革命的结构》，金吾伦、胡新和译，北京大学出版社，2003。
6. Robert K. Merton, "The Normative Structure of Science" (1942), in The Sociology of Science, Chicago: University of Chicago Press, 1973（有组织的怀疑）。

7. Alan D. Sokal, "Transgressing the Boundaries: Towards a Transformative Hermeneutics of Quantum Gravity", Social Text, No. 46/47 (1996); 及其自述 "A Physicist Experiments with Cultural Studies", Lingua Franca, May/June 1996.

8. John P. A. Ioannidis, "Why Most Published Research Findings Are False", PLoS Medicine, Vol. 2, No. 8 (2005).

9. Open Science Collaboration, "Estimating the Reproducibility of Psychological Science", Science, Vol. 349, No. 6251 (2015).

10. Joseph P. Simmons, Leif D. Nelson & Uri Simonsohn, "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant", Psychological Science, Vol. 22, No. 11 (2011).

11. Marilyn Strathern, "'Improving Ratings': Audit in the British University System", European Review, Vol. 5, No. 3 (1997)（古德哈特定律的常引表述）。

12. Donald T. Campbell, "Assessing the Impact of Planned Social Change", Evaluation and Program Planning, Vol. 2, No. 1 (1979)（坎贝尔定律）。

13. Wendy N. Espeland & Michael Sauder, "Rankings and Reactivity: How Public Measures Recreate Social Worlds", American Journal of Sociology, Vol. 113, No. 1 (2007).

14. Jerry Z. Muller, The Tyranny of Metrics, Princeton: Princeton University Press, 2018.

15. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", FAccT '21, 2021.

16. Emily M. Bender & Alexander Koller, "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data", ACL 2020.

17. Ziwei Ji et al., "Survey of Hallucination in Natural Language Generation", ACM Computing Surveys, Vol. 55, No. 12 (2023).

18. Lei Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions", arXiv:2311.05232, 2023.

19. Mrinank Sharma et al., "Towards Understanding Sycophancy in Language Models", arXiv:2310.13548, 2023.

20. Ethan Perez et al., "Discovering Language Model Behaviors with Model-Written Evaluations", Findings of ACL 2023, arXiv:2212.09251.

21. Paul F. Christiano et al., "Deep Reinforcement Learning from Human Preferences", NeurIPS 2017, arXiv:1706.03741（偏好训练的来源）。

22. Joseph Weizenbaum, Computer Power and Human Reason, San Francisco: W. H. Freeman, 1976（对人机关系中投射与依赖的早期警告）。

23. Michael Polanyi, The Tacit Dimension, London: Routledge & Kegan Paul, 1966；中译《默会的维度》，许泽民译，上海人民出版社，2000。

24. Barry M. Staw, "Knee-Deep in the Big Muddy: A Study of Escalating Commitment to a Chosen Course of Action", Organizational Behavior and Human Performance, Vol. 16, No. 1 (1976)（承诺升级）。

25. 王德生：《SDE 智能体导论——从语言模型到发生世界模型的智能体革命》，德麦国际出版，2026，ISBN 978-1-970820-62-1（本文第十五章与第十六章之改写底本）。

26. 王德生：《SDE本体论入门：世界是如何发生的》，德麦国际出版，2026（三大方程、123 原理与三界结构的完整展开）。