当两种阅读彼此校正：人类终审下的程序追踪、内在反诘与生成性复调同一首诗为何引出两篇结论相近、作用却不同的评本?

同一首诗为何会引出两篇结论相近、作用却显著不同的批评文本？既有原稿将这种差异概括为"发生学诗评"与"体内诊断"两种不可互还原的本体论范式，但其论证把方法来源、文本格式与知识目标混为一体，并把可组合的操作误写成逻辑上相互排斥的立场。本文以一首围绕围棋人工智能可解释性展开的组诗及其两篇机器生成评本为配对个案，重新界定这一问题。文章不再追问两种批评"分别是什么本体"，而考察它们如何对同一对象施加不同的认知压力：其一是程序追踪，即借助稳定问题序列重建意象、结构和概念的生成路径；其二是内在反诘，即接受作品已经提出的规范，再检验作品自身是否承担了该规范的后果。研究发现，两者的关键差异不在"外部方法/内部方法"或"论文体/对话体"，而在证据组织方式：程序追踪扩大解释覆盖面，内在反诘提高自洽性压力；前者容易把作品收编为方法的例证，后者容易把局部矛盾放大为整体结论。基于二者的交叉校正，本文提出"生成性复调"模型：相异评本不必融合为统一裁决，而可在保留证据链和分歧来源的前提下，产生任一评本单独无法稳定提出的新命题。进一步地，本文把该模型操作化为六项编码指标、一个基于出现率的增量命题判据以及四道人类终审门：材料准入、尺度确认、语义等价裁定与理论退场。本文的贡献不是宣布一种新批评学派，而是提出一种可执行、可审计、可证伪的多评本研究方法：人工智能负责高密度生成候选差异，人类不以直觉替代证据，而在不可自动归约的判断节点保留署名责任和最终决定权。

关键词： 诗歌批评；内在批评；生成性复调；人工智能；人类在环；配对个案

When Two Readings Correct Each Other: Procedural Tracing, Immanent Interrogation, and Generative Polyphony under Human Adjudication

Abstract

Why can two readings of the same poem reach similar conclusions while performing markedly different critical functions? An earlier version of this study treated the difference as an ontological split between "genetic poetry criticism" and "endogenous diagnosis." That claim, however, conflated the source of a method, the format of its output, and its epistemic aim, turning potentially combinable operations into supposedly incompatible paradigms. Using a paired case consisting of a poetic sequence on explainable Go AI and two machine-generated reviews, this article reframes the problem in operational terms. It distinguishes procedural tracing, which reconstructs how images, structures, and concepts emerge through a stable sequence of questions, from immanent interrogation, which adopts a work's own stated norms and tests whether the work bears their consequences. The decisive difference lies neither in external versus internal method nor in academic versus conversational form, but in the organization of evidence. Procedural tracing increases interpretive coverage, whereas immanent interrogation increases pressure for self-consistency. The former risks assimilating the poem to a prefabricated framework; the latter risks inflating a local tension into a total judgment. Their cross-correction yields what this article calls generative polyphony: divergent readings need not merge into a final verdict but can preserve their evidential and methodological differences while producing propositions that neither could stably generate alone. The model is operationalized through six coding dimensions, an occurrence-rate criterion for incremental propositions, and four gates of human adjudication: corpus admission, validation of critical criteria, semantic-equivalence decisions, and theory-retirement decisions. The contribution is not a new school of criticism but an executable, auditable, and falsifiable method for multi-reading research. AI systems generate candidate differences at scale; human participants retain accountable judgment at decisions that cannot be reduced to automatic aggregation.

Keywords: poetry criticism; immanent critique; generative polyphony; artificial intelligence; human-in-the-loop; paired case

一、问题的提出：从"谁更好"转向"差异如何工作"

文学批评经常把同一作品的多种读法理解为意见竞争：它们或者争夺更正确的解释，或者被折衷为"各有优长"。前一种处理把复杂差异压缩为胜负，后一种处理则把差异消解为互补。本文面对的材料恰好抵抗这两种简化。

个案中的原诗《由棋入道：一架会显影的钢琴》以序曲、五个乐章和终曲组织全篇，将围棋人工智能的搜索、所有权预测、稳定性检验和人机盲区写入"钢琴—显影—色场"的隐喻系统。围绕该诗形成的两篇评本在核心判断上接近：二者都不把作品视为给人工智能知识披

上诗性外衣，而认为作品的形式正在演示"不可见过程如何获得暂时可见性"。然而，两篇评本使读者发生改变的方式并不相同。

评本A逐乐章追踪意象和概念的生成，辨析"色场"从概率分布到认识论声明再到制度性证词的多重转化，并提供可直接比较的改写方案。评本B则抓住诗中"哪里能够相信，哪里只能暂信"的自我声明，反问诗的抒情形式是否也标记了自身判断的可信度差异；继而提出"共同世界""持续受审"和"未闭合循环"等原诗未明说的后果。A的优势是覆盖、区分与可复用，B的优势是施压、反转与产生剩余。

原稿据此提出，两者构成"发生学诗评"与"体内诊断"两种不可互还原的诗评本体论。这个命名捕捉了真实差异，却作出了过强推论。第一，"使用预先形成的问题"与"从作品内部抽取标准"可以在同一篇批评中先后发生，并不逻辑矛盾。第二，论文体、对话体或聊天体属于传播形态，不能直接证明一种批评的本体论目标。第三，可复用性和事件性也不是零和关系：一次性洞见可以被重新编码，规范性论文也可以造成不可逆的阅读改变。若坚持原论证，任何成功的混合文本都会被解释成例外或"第三位置"，理论因而获得免疫，失去可证伪性。

因此，本文把研究问题从"存在两种什么样的批评本体"改写为三个较弱但更可检验的问题：

两篇评本对同一作品分别施加了何种分析操作？

这些操作如何组织证据、制造盲区，并在交叉阅读中彼此校正？

这种交叉校正能否被编码、复现和证伪？

这一改写并非降低理论雄心，而是把雄心从命名新范式转移到解释新机制。理论的强度不取决于宣称层级，而取决于它能否说明材料中的差异、预测新的比较结果，并允许失败。

二、理论位置：四条传统之间的重新接线

（一）细读：从形式自足到证据锚定

新批评确立了对语词、悖论、张力和整体结构的精密阅读。兰色姆将"新批评"命名为一种专业化批评方案，布鲁克斯则以具体诗篇展示形式与意义不可轻易拆分[1-2]。本文继承的不是作品绝对自足的假设，而是较朴素的证据纪律：批评中的概念必须能够返回具体词句、结构位置和修辞关系。没有这一纪律，"生成""反噬"或"复调"都可能只是批评者自我增殖的术语。

（二）文本发生：从草稿谱系到阅读中的过程重建

法国文本发生学把研究重心从成品转向写作过程，通过手稿、草稿、删改和版本追踪作品的形成。德·比亚西明确把这一转向概括为从"书写结果"走向"书写活动"、从结构走向过程、从作品走向发生[3]。本文所称"程序追踪"与之相关，但必须保持边界：个案没有手稿谱系，不能把对成品结构的逆向重建冒充严格意义上的文本发生学。本文研究的是批评如何构造"可辩护的生成路径"，而不是考证作者实际写作顺序。

这一界限修正了原稿的一个概念跃迁。由成品推断形式逻辑，可以是一种发生取向的阅读，却不等于掌握作品的真实发生史。只有当推断得到草稿、版本、创作日志或作者陈述支持时，二者才能局部重合。

（三）内在批评：从揭露矛盾到检验规范后果

原稿把评本B主要放在德里达式解构谱系中，并把解构描述为"终结文本"。这种描述过于单线。解构并非简单宣布文本崩溃，其工作恰恰依赖文本内部无法被稳定封闭的差异运动[4]。对本研究更直接的理论资源是内在批评：它不从外部强加一套完全异质的尺度，而是考察对象自身承认的规范及其实现之间的张力。

贝克尔区分了内在批评的两个含义：以对象的内部规范评价对象，以及从对象出发说明批评者自身的立场[5]。这一区分使评本B的操作得到更准确描述。它不是因为"来自诗内部"就天然正确，而必须证明三点：作品确实提出了该规范；该规范与被检验的形式层面相关；由此推出的缺失不是批评者偷换概念的结果。所谓"用诗反问诗"，因此不是修辞姿态，而是一条需要完整证据链的推理。

（四）读者反应与复调：意义不是单向提取

伊瑟把阅读理解为文本结构与读者活动之间的互动，意义并非作为固定物被从文本中搬运出来[6]。这为"评本使作品变得不同"的经验提供了较稳健的表述：改变的不是已经完成的物理文本，而是文本可被实现的关系结构。巴赫金关于复调的论述则强调多个独立而不被归并的声音[7]。本文借用"复调"时不把两个机器输出拟人化为完整意识，而仅指两套分析序列在共同对象上维持可辨认的差异，且不被预先

裁决为一个总答案。

这里必须与一个最容易把本文论点整个吞掉的近邻正面划界——巴特（Roland Barthes）的可读／可写文本之分。巴特把文本分为“可读的”（lisible，读者被动消费一个已完成的意义）与“可写的”（scriptible，读者被邀请参与意义的生产）；他推崇后者，因为它把读者从消费者变为生产者。表面上，本文所说的“评本使作品变得不同”“两种读法在同一对象上维持不被归并的差异”，几乎就是巴特“可写文本”的一个当代复述——都在讲读者一方的生产性、都在拒绝单一确定的意义。

但两者切在完全不同的层面，且本文多出一个巴特框架无法容纳的构造。巴特的可写／可读是关于文本邀请何种阅读姿态的属性判断——它描述读者被允许做什么，其单位始终是“读者的自由”。而本文的程序追踪与内在反诘不是两种阅读姿态，而是两套可相互校正的分析约束：程序追踪用稳定问题序列重建生成路径，内在反诘用作品自设规范反检其兑现——它们各自都可被证伪（路径重建是否返回具体词句、规范反检是否确由作品支持），且第二套可以纠正第一套的过度诠释。巴特的“可写”恰恰取消了这种校正的可能：一旦文本被判为可写，读者的任何生产都是正当的，没有一个内部标尺能说某次生产“读错了”——这正是本文在批评原稿时指认的那种“理论获得免疫、失去可证伪性”的病。因此本文与巴特的判决性分离线是：可写文本把差异的正当性交给读者的自由，本文则把差异的正当性交给“能否返回作品被校正”的证据纪律。本文要的不是更自由的读者，而是两套都能被作品否决的读法——它们的价值不在于摆脱了确定性，而在于各自随时可被作品撞回。

这四条传统共同限定了本文的理论位置：证据锚定防止概念漂移，过程重建解释结构如何被追踪，内在批评提供自洽性检验，读者反应与复调说明多重读法如何共同改变作品的可读空间。

三、材料与方法：一个配对个案的机制分析

（一）材料身份与证据层级

本文的直接材料包括：一首以围棋人工智能可解释性为主题的组诗、两篇在相近时间内生成的评本，以及把两篇评本理论化的v3.0原稿。

必须在此明确证据层级问题。当前研究对原诗与两篇评本的引用，分为两个层级：E1（一手证据）指可直接核对原诗或评本原文的引文与结构描述；E2（经v3.0中介的转述）指依据v3.0对评本的引述与结构性转述而重建的描述，尚未逐条返回评本原文核对。本文第四节的机制重建将逐处标注其证据层级。由于E2占相当比例，且完整语料库（原诗全文、两篇评本全文及其生成条件记录）尚未公开，本文把结论限定为"机制发现型个案"，不声称代表一般诗歌批评。在语料库公开并完成E2向E1的升级核对之前，第四节的个案描述应被读作"待复核的机制假设"，而非已确证的文本事实。语料库的公开与授权列入附记所述的人类终审事项。

两篇评本由不同人工智能工作流生成。本文将其记为评本A与评本B，不把系统名称当作作者人格或稳定方法主体。大型语言模型产生的是受训练数据、提示、上下文和采样条件约束的文本输出。把模型写成"它知道自己在做什么"会把语言效果误认成心理状态；相关研究也提醒，描述语言模型时应避免把表面流畅直接等同于人类式理解或意图[8-9]。因此，本文分析的是两个文本如何工作，而不是两个"批评家"内心如何选择。

（二）比较单位

本文不以篇幅、格式或是否含参考文献作为核心分类标准，而以"最小批评动作"为比较单位。一个动作至少包含四个环节：对象证据、分析操作、中间推论和结果命题。例如：

"诗提出可信度分层"是对象证据；"检查全诗是否对自身判断实行分层"是分析操作；"统一抒情声调抹平了可信度差异"是中间推论；"作品的可解释性伦理尚未回写到自身形式"是结果命题。

只有四环相连，才能把"锐利"从印象评价变为可复核的论证特征。

（三）分析策略

研究分三步进行。第一步分别重建评本A、B的动作序列，避免先用某个宏大名称吞没细节。第二步进行交叉反事实：删除A或B，观察哪些命题将无法稳定产生；再尝试把两套动作置于同一阅读流程，检验其是否真有逻辑冲突。第三步把交叉校正后的机制转化为编码指标和后续实验设计。

这种方法不证明两个模式是文学史意义上的"范式"。它只判断：在本个案中，两种动作是否具有稳定差异，差异是否带来可识别的增量，增量能否被其他研究者复现。

（四）人类终审的位置

本文所说的"人类在环"不是在模型输出之后笼统签字，也不是把人的直觉设为不可质疑的真理源头。它指向四类必须留下责任主体和决定理由的判断：哪些材料可以进入语料库；批评尺度是否确由作品或研究设计支持；两个表述能否被编码为同一命题；何种反例足以迫使概念收缩或退场。机器可以为这些判断提供候选、对照和反例，却不能同时生成标准、应用标准并宣布自己通过。

为避免"人类终审"退化为研究者任意裁决，确认性研究应事先冻结语料范围、提示版本、采样参数、编码手册、主要结局和排除规则；语义等价与质量编码由不知道实验条件的两名以上编码者独立完成，分歧先保留、后讨论，并同时报告初始一致性与最终裁定。署名研究者对材料授权、理论命名和结论强度承担最终责任，但其裁定本身也属于可审计数据，而不是证据链之外的特权。

四、两种批评操作：程序追踪与内在反诘

（一）程序追踪：扩大解释覆盖面

程序追踪指批评者借助相对稳定的问题序列，重建作品各部分之间的形成关系。它通常依次追问：核心意象承载哪些未被单一命名穷尽的属性？章节或段落之间发生了怎样的位移？局部形式如何回写整体命题？作品的结尾是否兑现了前文建立的逻辑？

评本A对"色场"的处理体现了这种操作（证据层级：E2，依据v3.0对评本A的结构性转述）。它没有停在"色场象征可解释性"，而把同一意象拆成四个层面：工程层面的概率分布、界面层面的视觉化、认识论层面的暂时相信、制度层面的可受盘问证词。这个拆分的价值，不是术语数量增加，而是建立了从技术对象到诗学形式的连续桥梁。随后，评本A把"暗河""风雪""盲区赋格""边界上的耳朵"等段落放入同一形成序列（E2），使局部亮点成为整体结构。

程序追踪至少有三项认知收益。其一是覆盖收益：它迫使批评者处理作品中可能不支持首要判断的段落。其二是区分收益：它把笼统赞美拆成可争论的层次。其三是改写收益：稳定的问题序列可以生成具体修改方案，而非只给抽象评价。

它的风险也来自同一结构。预先稳定的问题可以成为"吸附器"，把所有意象都收编为方法的例证。当每个段落都恰好证明框架正确时，完整性可能只是框架闭合。判断程序追踪是否有效，不能看它覆盖了多少文本，而要看它是否允许反例改变问题序列。

（二）内在反诘：提高自洽性压力

内在反诘指批评暂时接受作品自身提出的规范，并追问作品的形式、语调和结尾是否承担了该规范的后果。它不是一般意义上的"找矛盾"，而是从对象内部建立评价前提。

评本B抓住诗中的关键命题（证据层级：E2）：可解释性不是给所有疑问涂上确定颜色，而是区分"能够相信""只能暂信"和"仍需不同耳朵"的区域。由此，它提出一个有效的内在问题：既然诗要求人工智能标注不确定性，诗是否也标注了自己的判断强度？如果全篇以同一抒情权威发声，那么作品可能在形式上取消了它在内容上主张的分层。

这一步产生了原评本A没有稳定提出的三个后果（E2）。第一，人机交换盲区的结果不只是互相理解，而可能出现既不属于人类直觉、也不等于机器搜索的"共同第三项"。第二，"回到风雪"不是一次检验，而是模型、规则与对手变化所要求的持续受审。第三，如果显影总会留下新盲区，终曲的闭合就需要被重新解释：它可以是乐章结束，却不能被当作认识过程完成。

内在反诘的力量在于，它把作品的观念从被陈述的内容转化为约束作品自身的规范。然而，这种操作也有两项风险。其一是尺度错配：内容层面的主张未必必须在所有形式层面同构实现。诗讨论可信度分层，并不自动意味着每一段都要显式标注置信度。其二是局部放大：一个富有生产力的反问可能揭示张力，却未必足以否定作品整体。内在反诘必须区分"形式未兑现""形式以别种方式兑现"和"该规范并不适用于此形式"。

（三）差异不在内外，而在压力方向

原稿把A定义为"外部方法"，把B定义为"内部提取"，并据此断言两者互斥。个案却显示，A同样需要从作品中识别结构，B也带着"内部规范应约束对象"的预设进入文本。任何阅读既携带先验问题，也依赖对象提供的结构；纯粹的"无预设阅读"与纯粹的"自动生长"都不存在。

较准确的区分是压力方向不同：程序追踪把压力施加在批评者身上，要求其解释更多材料、维持结构连续；内在反诘把压力施加在作品上，要求作品承担自身规范的后果。前者首先问"我的解释遗漏了什么"，后者首先问"作品尚未兑现什么"。两者不是两个封闭本体，而是可以相继运行、也会彼此牵制的操作模式。

五、生成性复调：从互补清单到差异生产

（一）何谓"生成性"：概念定义与操作判据

如果把A的优点和B的优点简单相加，所得只是"完整且锐利"的理想批评人格。这种互补论没有解释新命题如何出现。本文把更严格的生成性定义为：在评本交叉作用后出现一个新命题P，而P不能从任一评本的现有证据链中单独稳定推出；同时，P仍能返回共同文本证据，并改变至少一个原评本的判断。

"不能单独稳定推出"是一个不可推导性断言，在实践中无法直接证明。为使其可判定，本文把它改写为可观察的频率差。对同一作品，在仅执行程序追踪、仅执行内在反诘以及交叉校正三种条件下分别独立生成评本。对命题P，分别计算它在两个单一条件中的出现率，并取较高者为p_single；交叉条件中的出现率记为p_cross。只有当P能返回共同文本证据、改变至少一个原判断，并且p_cross相对于p_single呈现预先规定且具有不确定性区间支持的增量时，才把P称为该配对的增量命题。样本量不再由"不少于十次"之类的固定小数目决定，而应依据先导数据、最小有意义差异、检验功效和多重比较方案预先计算。语义等价由不知道实验条件的编码者独立判断，机器聚类只可提供候选，不得直接决定命题同一性。这一判据牺牲了哲学断言的强度，换取可复核性；在完成确认性实验前，任何增量命题都只能称为候选。

以个案为例，"作品应把可解释性伦理回写到自身发声制度"不是A的结构追踪自然包含的结论，也不是B的单次反问即可充分论证的判断。它需要A对"色场—证词—盘问"的层次区分，也需要B对"作品是否标注自身不确定性"的压力测试。两者交叉后，问题从"这首诗写清了什么"推进到"它用什么制度安排自己的确定性"。这才是差异生产，而不是优点拼盘。当然，依据上述判据，该例目前只是增量命题的候选：它来自单一交叉实例，尚未经过出现率检验。

（二）何谓"复调"

"复调"在这里包含三条约束。第一，评本必须保持可识别差异，不能在汇总时被改写成同一个声音。第二，每个判断必须保留来源和证据路径，使读者知道它来自何种操作。第三，综合文本不得享有天然高位；它也必须接受原评本的反向检验。

因此，复调不是"多找几个模型投票"。投票消除理由差异，只保留结论数量；生成性复调保存的是异质推理。当两个评本结论一致但理由不同，理由差异比票数更有研究价值；当结论冲突时，也不能自动把中间立场当作更客观。

（三）双重约束校正

为使生成性复调可执行，本文提出"双重约束校正"：

来源约束：任何新增命题都必须标明其文本证据、推理步骤和适用范围，防止批评借作品自我扩张。

后果约束：任何解释都必须接受作品内部规范或自身方法规则的反向检验，防止批评只解释支持自己的材料。

程序追踪主要强化来源约束，内在反诘主要强化后果约束。二者交叉时，A迫使B说明反问所依赖的全部文本证据，B迫使A说明其结构解释是否对反例和自我矛盾开放。所谓"彼此校正"，不是一方替另一方补缺，而是每一方改变另一方可以合法声称什么。

六、从概念到测量：六项编码指标

为了避免"生成性复调"停留为漂亮隐喻，本文提出六项可用于评本比较的编码指标。每项采用0—2分：0表示缺失，1表示局部存在，2表示稳定且有证据。

需要说明第三项指标的更名。原稿及本文早先版本将其称为"内在尺度"，0分描述为"外部标准未说明"。这一措辞与本文第四节的论证自相矛盾：既然程序追踪所用的外部问题序列是合法操作，指标就不应惩罚标准的外部性本身，而只应惩罚来源的不透明。因此本版将该指标更名为"尺度透明"：外部标准与内在标准同样可以得到2分，条件是批评明确说明标准从何而来、适用于哪个形式层面。

这张表不是审美价值的总量表。它测量的是多评本研究所需的论证行为，不应被用于给诗歌本身打分。不同体裁也可能需要调整权重：短诗未必追求路径覆盖，实验文本也可能有意拒绝自洽。量表的作用是暴露研究者在哪一步作出了判断，而不是把文学批评伪装成精确计量。

七、可证伪方案：理论何时应当退场

原稿提出失败条件的意识值得保留，但部分检验把"格式不可翻译"预先当作结论，形成循环。本文改用以下四组实验。所有涉及人工

编码的实验均报告Krippendorff's α及其置信区间。该系数可处理不同测量层级、多个编码者和缺失值，适合本研究的文本编码结构[12]。依照内容分析研究的常用解释，α≥0.80可支持较稳定结论，0.667≤α<0.80只用于暂定性判断，低于0.667则返回编码手册修订与重新训练[13]。阈值不是自然定律；若错误裁定的代价较高，应在预注册中采用更严格标准。

（一）跨文本复现

从不同诗体、主题和篇幅中分层抽取作品；具体样本量由先导研究的方差与预设效应量决定，而不是把二十首当作普遍充分的门槛。对每首作品分别生成或邀请两类评本：一组明确执行程序追踪，一组明确执行内在反诘。由不知道评本来源的编码者使用六项指标评分。确认性分析应把作品、模型或作者以及重复生成视为不同层级，避免把同一作品上的多次输出误当成彼此独立的样本。若两种操作不能稳定提高各自预期指标（程序追踪应提升证据锚定与路径覆盖，内在反诘应提升后果压力），本文对两种操作的区分便缺乏可复现性。

（二）交叉增量实验

设置A→B、B→A、单独A、单独B四种条件，固定模型版本、提示、上下文窗口与采样参数，并以第五节的判据统计经人工盲码确认的增量命题。重复次数由功效分析决定；报告效应量和区间估计，而不只报告"显著/不显著"。生成性复调的核心效果受到直接否证：若交叉条件的增量命题出现率相对于表现更好的单一条件没有达到预注册的最小有意义差异，或者新增命题不能返回共同证据，理论便没有得到支持。

顺序效应（A→B与B→A之间的差异）作为探索性分析。顺序不敏感既可能表示交叉校正对称而稳健，也可能来自提示强度压过前序文本；仅凭无差异不能区分二者。顺序分析的价值在于描述校正机制的结构，并为后续实验提出竞争解释，而不单独承担理论否证。

（三）改写效度实验

让第三方诗人或编辑在盲态下比较原诗、依据A修改的版本、依据B修改的版本和依据交叉校正修改的版本。评价不以"更好"一项概括，而分别考察结构清晰度、张力保留、自我一致性和诗性损失。若交叉版本只提高解释清晰度却持续损害诗性，双重约束校正就不能被宣称为普遍优化程序。

（四）来源替换实验

把机器生成评本替换为人类评本，或系统改变模型、提示和语境，观察操作特征是否跨生成来源保持。来源标签应在编码阶段隐藏，并在必要时设置"仅换标签、不换文本"的控制条件，以区分标签期待与文本特征。如果所谓A、B特征主要由一次提示偶然造成，或评分随标签而非文本改变，那么研究对象应被描述为偶发文本或标签效应，而非稳定批评模式。

在以上实验中，出现下列任一结果都应迫使理论收缩：六项指标无法达到预注册的一致性要求；两种操作没有稳定区分；交叉条件未超过表现更好的单一条件；新增命题不能返回共同证据；人类编码者无法可靠辨认命题等价；或综合文本总能不受损失地还原为其中一篇评本。可证伪性不是论文末尾的姿态，而是概念获得使用资格的条件。

八、讨论：人工智能多评本研究的意义与边界

（一）人工智能的价值不在替代裁决，而在制造可比较差异

长篇文学理解仍是大型语言模型的困难任务。截至相关评测发布时（2024年），中文长篇小说理解评测显示，模型在零样本条件下对多类理解问题表现有限，不能把流畅回答直接视为可靠阐释[11]；需要说明的是，该评测对象为长篇小说而非诗歌，且模型能力此后仍在变化，故此处只作为"流畅不等于可靠"的邻近证据，而非对诗歌评本可靠性的直接测量。同样，针对自动文献综述的实证研究发现，即使较先进的模型仍会生成虚构参考文献，文献核验必须独立于文本生成[10]。这意味着，多模型或多提示的价值不应被表述为"让人工智能替人类决定诗的意义"，而应是低成本地产生结构差异明确的候选读法，供人类进行证据审查、比较和修订。

本个案最有启发性的地方，不是两个系统像两位天才批评家一样相遇，而是同一材料在不同约束下暴露出不同可读性。研究的最小伦理要求是：不把模型输出中的第一人称当作责任主体，不把生成文本中的引用当作已核事实，不把多模型一致当作真理，也不让综合步骤抹去反对意见。

这也说明，人类在环的价值不等于"人比机器更会判断"。人在本流程中的不可替代性首先是规范与责任上的：语料是否获准使用、何种差异值得保留、哪些损失不可接受、何时停止扩张一个概念，都不能从生成频率本身推出。与此同时，人类判断也会受权威效应、标签期待和理论偏好影响，因此必须通过盲码、复核、异议记录和可撤销决策接受约束。理想结构不是"机器提议、人类批准"的单向闸门，而

是机器制造差异、人类说明取舍、反例重新开启判断的循环。

（二）"发生"应由可追踪差异证明

原稿多次把批评描述为"一次不可逆的发生"。这一表述富有诗性，但若无判据，任何强烈阅读感受都可被命名为发生。本文将其收紧为三个可观察结果：读者能够指出新增命题；新增命题改变了对至少一个原段落的理解；这种改变能通过重读、改写或第三方复述被追踪。这样，"发生"不再是理论自我授权，而成为可留下证据的阅读变化。

（三）论文格式不是敌人，自反性也不是结尾修辞

原稿把学术规范与事件锐利之间的关系写成结构性冲突，仿佛摘要和参考文献必然驯化思想。事实上，规范可以保存证据、暴露来源并使争议可持续；它的问题不在"太规范"，而在规范是否被当作思想成立的替代品。同样，自反性不应只在结尾追问"本文属于哪一范式"。更严格的自反是让本文的核心概念接受本文自己的量表：它是否证据锚定？是否覆盖反例？是否承担自身规范的后果？是否允许研究者据结果放弃它？

以此衡量，本文目前只达到理论候选状态。它对单一个案作了较完整的机制重建，但该重建部分依赖二手转述（见第三节的证据层级说明），且尚未完成跨文本复现、编码一致性检验和改写实验。把这一状态标记清楚，比把未决性提升为新的本体论位置更符合研究规范。本文自身的生成条件——包括人工智能在写作中的参与程度与人类核验环节——同样属于必须暴露的"看见条件"，详见附记的生成条件声明。

九、结论

两篇评本之间最重要的差异，不是一个属于知识、另一个属于事件，也不是一个从外部进入、另一个从内部生长。它们分别把批评压力导向不同位置：程序追踪要求解释者扩大证据覆盖、重建形成路径；内在反诘要求作品承担自己提出的规范、暴露尚未兑现的后果。两种操作都可被学术化，也都可能产生不可逆的阅读改变；因此，它们不应被实体化为互相封闭的本体论范式。

本文提出的"生成性复调"保留了原稿最有创造力的发现——差异能够生产单一读法无法产生的新命题——同时取消了不可证伪的强断言，并为"新命题"给出了基于出现率的可判定标准。其运行条件是双重约束：新增解释必须返回文本来源，也必须接受自身及作品规范的后果检验。它的产物不是终极评判，而是一条可审计的差异链：某个判断从哪里来，受到什么反问，如何改变，又在哪些证据面前应当退回。

这也给人工智能参与文学批评划出合适位置。模型可以高密度地产生候选差异，却不能因此获得终审权；人类保留终审权，也不能因此免于说明理由。格式可以保存知识，却不能替代证据；多声部可以扩大可读空间，却不保证真理自动从数量中出现。真正值得保留的，不是"两个智能体交换了眼盲"的拟人化神话，而是一种更朴素也更严格的研究能力：让每一种读法暴露自己的看见条件，让每一次裁定留下责任路径，并允许新的反例重新开启它。

参考文献

[1] RANSOM J C. The New Criticism[M]. New York: New Directions, 1941.

[2] BROOKS C. The Well Wrought Urn: Studies in the Structure of Poetry[M]. New York: Reynal & Hitchcock, 1947.

[3] DE BIASI P-M. Génétique des textes[M]. Paris: CNRS Éditions, 2011.

[4] DERRIDA J. De la grammatologie[M]. Paris: Les Éditions de Minuit, 1967.

[5] BECKER M A. On immanent critique in Hegel's Phenomenology[J]. Hegel Bulletin, 2020, 41(2): 224-246. DOI: 10.1017/hgl.2018.8.

[6] ISER W. The Act of Reading: A Theory of Aesthetic Response[M]. Baltimore: Johns Hopkins University Press, 1978.

[7] BAKHTIN M M. Problems of Dostoevsky's Poetics[M]. EMERSON C, ed. and trans. Minneapolis: University of Minnesota Press, 1984.

[8] SHANAHAN M. Talking about large language models[J]. Communications of the ACM, 2024, 67(2): 68-79. DOI: 10.1145/3624724.

[9] BENDER E M, GEBRU T, MCMILLAN-MAJOR A, et al. On the dangers of stochastic parrots: Can language models be

too big?[C]//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, 2021: 610-623. DOI: 10.1145/3442188.3445922.

[10] TANG X, DUAN X, CAI Z. Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Suzhou: Association for Computational Linguistics, 2025: 1602-1617. DOI: 10.18653/v1/2025.emnlp-main.83.

[11] YU L, LIU Q, XIONG D. LFED: A literary fiction evaluation dataset for large language models[C]//Proceedings of LREC-COLING 2024. Torino: ELRA and ICCL, 2024: 10466-10475.

[12] HAYES A F, KRIPPENDORFF K. Answering the call for a standard reliability measure for coding data[J]. Communication Methods and Measures, 2007, 1(1): 77-89. DOI: 10.1080/19312450709336664.

[13] KRIPPENDORFF K. Content Analysis: An Introduction to Its Methodology[M]. 4th ed. Los Angeles: SAGE, 2019.

生成条件与研究透明度声明

本文以人工智能生成的批评文本为研究对象，其自身写作过程亦有人工智能深度参与。依据本文第八节提出的最小伦理要求与自反性标准，特作如下声明：

写作分工。 本文由原稿v3.0经人机协作重构而来。人工智能参与的环节包括：论证结构的重组、概念的改写与命名（如"程序追踪""内在反诘""尺度透明"）、编码指标与实验设计的草拟、中英文摘要的撰写与互译。人类作者负责的环节包括：研究问题的设定、对v3.0原稿缺陷的初始判断、各版本之间的取舍决定，以及最终文本的逐段审读与修改。任何段落均不应被视为纯机器输出或纯人类写作；责任主体为署名作者。

文献核验。 参考文献的作者、题名、出版信息、卷期页码与DOI已优先依据出版社、期刊页面和ACL Anthology等一手书目来源复核。本文直接依赖的命题亦与可获得的摘要或正文核对：包括[5]对内在批评两种定义的区分、[8]对语言模型拟人化风险的论述、[10]对自动文献综述中虚构引文的实证发现、[11]对长篇文学理解困难的测量，以及[12-13]对编码一致性方法与解释阈值的说明。书目核验只能确认来源与转述相符，不能替代对本文推论的同行评议。

尚待人类终审的事项。 原诗与两篇评本的全文语料及生成条件记录的公开授权；第四节E2层级描述向一手证据的升级核对；作者署名与目标期刊体例适配。在这些事项完成前，本文对自身的评级是：按第六节量表，"证据锚定"暂为1分（关键个案描述仍依赖二手转述），其余各项的最终得分应由第三方编码而非作者自评给出。

这篇最有价值的动作是一次降级：把原稿「两种不可互还原的诗评本体论」改写成三个更弱但可检验的问题。放弃宣称的层级，换取解释的机制——在以命名新范式为荣的写作风气里，这是逆向操作，也正是它强于原稿的原因。作者说得准：理论的强度不取决于宣称层级，而取决于它能否说明材料中的差异、预测新的比较结果，并允许失败。

第二笔是用压力方向取代内／外之分：程序追踪把压力施加在批评者身上（我的解释遗漏了什么），内在反诘把压力施加在作品身上（作品尚未兑现什么）。这一刀比「外部方法 vs 内部提取」准确得多，且允许两者在同一篇批评里先后发生。第三笔是把「生成性」变成可判定的——用三种条件下的出现率差，替换「不能单独推出」这个无法直接证明的断言，再配六项 0–2 编码指标、Krippendorff α 与四组实验。其中来源替换实验里那个「仅换标签、不换文本」的控制条件尤其见功夫，它直接防住了标签期待冒充文本特征。至于 E1／E2 的证据分层自陈，以及据此把自己的评级降为「理论候选状态」，在学生论文中罕见。

期刊评审与评奖 把「两位审稿人意见冲突」从麻烦变成资源：要求各自标注证据与分析操作，冲突处保留而非折衷，主编承担终审并写明理由。折衷意见看似公允，实则销毁了最有信息量的部分。

教育评价 教师评语与机器评语并置时，用同一套六项指标编码，重点看交叉之后是否出现新增命题，而不是比较谁更权威。指标测的是论证行为，不该被用来给作品本身打分。

实验室与团队手册 人类终审的四类判断可以直接成文：哪些材料准入语料库、批评尺度是否确由对象支持、两个表述能否算等价、何时停止扩张一个概念。写下来，终审才不至于退化为「谁资历高谁说了算」。

内容与研究团队的日常 评两稿时先各自独立出具「证据—操作—结论」，禁止先读对方；汇总之后再做反例检验。这一条几乎零成本，却能挡住会议室里最常见的过早共识。