Chinese Traditional Culture 国学, 2026, 14(4), 1115-1125
Published Online August 2026 in Hans. https://www.hanspub.org/journal/cnc
https://doi.org/10.12677/cnc.2026.144157

Hans汉斯

当机器拟其形而失其律

——大语言模型辅助近体诗学习中的知识建构与批判性思维

许怀之[1]，陈弘正[2*]

[1]黄冈师范学院文学院(苏东坡书院)，湖北 黄冈
[2]黄冈师范学院机电与智能制造学院，湖北 黄冈

收稿日期：2026年7月7日；录用日期：2026年8月7日；发布日期：2026年8月14日

摘 要

本研究考察大学生与大语言模型(LLM)协同创作近体诗时的知识建构与批判性思维。延续既有七言绝句研究，本文以DeepSeek-V4辅助五言律诗创作，并设置两种提示条件：第一项不限次数，第二项限为七次。研究对两名学生的四份完整人机对话进行解释性逐轮分析。案例显示，模型较能遵守提示中明确列出的平仄与押韵要求，却可能遗漏对仗等未明示规范；自我核验时还可能以“拗救”“宽对”等术语为错误作出似是而非的解释。低投入学生倾向整体接受模型输出；积极互动若缺少稳固的格律知识，也未必能有效纠错。本文以元任务觉察(MTA)为分析框架指出，错误得以留存的关键在于学生对任务评价标准的觉察不足，且思维回路中的反思解释环节被让渡给生成者。本文建议将LLM定位为由规则核验与教师引导约束的诊断性陪练，而非格律判断的最终权威。由于样本仅含两名学生与一名专家，结论限于案例解释。

关键词

大语言模型，近体诗教学，格律认知，知识建构，批判性思维，人机协作，元任务觉察

When the Machine Mimics the Form but Misses the Meter

—Knowledge Construction and Critical Thinking in LLM-Assisted Learning of Chinese Regulated Verse

Huaizhi Xu[1], Hongzheng Chen[2*]

[1]School of Liberal Arts (Su Donpo Academy), Huanggang Normal University, Huanggang Hubei
[2]School of Mechatronics and Intelligent Manufacturing, Huanggang Normal University, Huanggang Hubei

Received: July 7, 2026; accepted: August 7, 2026; published: August 14, 2026

*通讯作者。

文章引用: 许怀之, 陈弘正. 当机器拟其形而失其律[J]. 国学, 2026, 14(4): 1115-1125. DOI: 10.12677/cnc.2026.144157

Abstract

This study examines knowledge construction and critical thinking as undergraduates co-compose Chinese regulated verse (jintishi) with a large language model (LLM). Extending earlier work on seven-character quatrains, the study uses DeepSeek-V4 to support the composition of five-character regulated verse, a form with more extensive structural requirements. Inspired by the allusion of "composing a poem in seven steps", the first task allowed unlimited prompting, whereas the second limited students to seven prompts. We conducted an interpretive, turn-by-turn analysis of four complete human-AI dialogue records produced by two students. The cases suggest that the model more reliably followed explicitly stated tonal and rhyme constraints but could overlook unstated genre requirements, particularly parallelism. During self-checking, it sometimes offered plausible-sounding yet incorrect explanations using terms such as aojiu (tonal deviation and compensation) and kuandui (loose parallelism). The two students also showed distinct risks: minimal engagement led to wholesale acceptance of model outputs, while active interaction without secure metrical knowledge did not ensure successful error detection. Drawing on Meta-Task Awareness (MTA), we argue that errors persisted because the students lacked awareness of the task's evaluative criteria and because the reflective-interpretation node of the thinking loop was ceded to the generator. We therefore propose positioning the LLM as a diagnostic practice partner constrained by rule-based verification and teacher guidance, rather than as the final authority on prosodic correctness. Because the study draws on two students and one expert evaluator, its findings are interpretive and preliminary.

Keywords

Large Language Models, Chinese Regulated Verse Learning, Prosodic Knowledge, Knowledge Construction, Critical Thinking, Human-AI Collaboration, Meta-Task Awareness

1. 引言

近体诗在中国古典诗歌教育中占有重要位置，其学习涉及字数、句数、平仄、押韵、黏对与对仗等多重规则。以五言律诗为例，学习者不仅需理解规则，还需在反复诵读与辨析中形成对声律的感知。生成式人工智能进入课堂后，格律校验似乎可以被部分“外包”给 LLM，使学习者将注意力转向意象、语词与情感。然而，既有研究已显示，模型在古典诗歌生成与评价中仍会违反形式约束，且自动评价与专家判断可能出现偏差[1] [2]。因此，技术提高写作效率的同时，是否减少学习者练习判断与累积经验的机会，仍需从具体的人机互动过程加以考察。

本研究的重点并非单独评定 LLM 生成的五言律诗是否优美，而是观察学生与模型协同创作时，如何提出规则、核验结果、回应错误并修订作品。本文关注三个问题：第一，模型在何种条件下遵守或遗漏格律要求；第二，学生既有知识如何影响其对模型输出的判断与追问；第三，限制提示次数是否有助于显现不同的参与与决策方式。通过分析完整对话轨迹，本文将“知识建构”具体化为学习者对规则的调用、解释、验证与修正过程，并将“批判性思维”具体化为对模型输出的质疑、证据核验与理由说明。

2. 概念框架

本研究从三个相互关联的脉络出发，作为解读人机互动记录的理论与经验依据。

2.1. 大语言模型诗歌生成的形式瓶颈

现有研究一致指出，古典诗歌生成的难点不只是语言流畅度，而是同时满足形式约束与内容表达。PoeTone 对 18 种 LLM 的宋词生成进行评估，发现严格的结构、声调与押韵约束仍构成主要瓶颈，并显示规则化评价与反馈可提高形式符合度[1]。针对唐诗的研究也发现，模型可能生成表面流畅、意象完整，却未严格遵守声律规则的作品[2]。另一方面，面向中国诗歌数据进行专门微调的较小模型，在特定任务上可优于规模更大的通用模型[3]。这些研究提示，模型“形似而律未必合”的表现与任务约束、训练资料和评价机制密切相关，不能仅以模型规模解释。

2.2. 评价偏差与大语言模型作为自我裁判的局限

LLM 亦可能在评价自身或同类模型作品时产生系统性偏差。Ma 等人的唐诗研究比较自动特征、LLM 评价与人类专家判断，发现 LLM 评价者可能形成“回声室”效应：它们偏好符合模型统计模式的诗句，却可能忽略严格的声律缺陷[2]。这并不等同于证明模型每次自检都会为错误辩护，但足以说明，LLM 不能被视为独立且可靠的终审者。有关 LLM 推理失败的综述进一步指出，当任务要求超出模型可维持和操作的信息容量时，输出可能出现不一致、遗漏或对提示变化敏感等问题[4]。据此，本研究将模型同时处理平仄、押韵、对仗、典故与意象时的疏漏，视为多重约束任务中的潜在失败，而不直接归因于单一机制。

2.3. 认知与能力外包的错觉

AI 使用也可能改变学习者投入认知努力的方式。Gerlich 的混合方法研究在不同年龄与教育背景的参与者中发现，AI 工具使用频率、认知外包与批判性思维表现之间存在统计关联，但该研究不能据此建立简单的因果关系[5]。Lee 等人针对知识工作者的调查则显示，对生成式 AI 的任务特定信心越高，受访者自报投入的批判性思维越少；相对地，对自身完成任务能力的信心越高，批判性思维投入越多[6]。一项涵盖数字技术与认知研究的整合性综述提出“效率－萎缩悖论”：工具可能提升短期效率，却也可能减少维持独立认知所需的练习；该综述同时强调，生成式 AI 的长期证据仍处于早期阶段[7]。因此，本文不预设 AI 必然削弱能力，而是考察学生是否把本应自行完成的规则判断交由模型，以及他们能否保留核验、解释与修正的责任。

2.4. 元任务觉察与人机共代理

上述三条脉络指出了风险的来源，却未提供统一的分析语汇。Wong 提出的元任务觉察(Meta-Task Awareness, MTA)适可补此不足：MTA 指学习者或教师超越任务的表层指令，辨识并反思其深层目标、结构与评价标准的能力，即不仅知道任务“要求什么”，也理解它“为何”与“如何”展开，并据此决定如何与 AI 协作[8]。该文强调，MTA 并非纯粹内在的心理状态，而是“通过可观察的互动模式得以体现”：使用者须在任务理解、提示建构、系统回应与反思解释之间往复循环，此一“思维回路”既是 MTA 运作之处，也是其得以发展之处。MTA 同时是可教、可练的——教师可要求学生先陈述任务目的再与 AI 互动，或评估提示与 AI 输出之间的一致性[8]。

MTA 对本研究有三重意义。其一，它把“评价标准的觉察”列为任务觉察的构成要素，因而能解释一个否则费解的现象：具备语言审美判断的学生，仍可能无力审核模型的格律输出。其二，若思维回路中的“反思解释”环节被让渡给生成者，回路即失去外部参照，正与 LLM 评价偏差的文献相互印证[2]。其三，该文所描绘的“后提示时代”——AI 主动推断意图并提供脚手架，使用者的重心由提示建构转向轨迹解释——使本研究具前瞻意涵：若学习者今日已无法审核模型对自身错误的技术性辩解，则在 AI 更为主动的情境中，此一风险将更难察觉。该文并指出，MTA 作为兼含元认知、任务觉察与 AI 协作的多

面能力，其拆解与操作化仍待后续研究；本文即以格律这一形式规范明确的领域，尝试对其“评价标准”维度作初步操作化。

本研究位于上述四条脉络的交汇处，并延续本团队对 DeepSeek 辅助七言绝句创作与格律认知的既有案例研究[9] [10]。先前两篇文献使用部分相同的研究脉络，分别为会议论文与后续期刊论文，因而在本文中被视为同一研究计划的阶段性成果，而非彼此独立的证据。本研究改以五言律诗、更新的模型版本及提示次数条件，进一步观察大学中文相关专业学生在人机共创中的规则调用与判断过程。

综合以上文献，本文以三类可观察行为作为分析维度：(1) 规则知识的显性调用，如是否提出平仄、押韵与对仗要求；(2) 验证与修正，如是否要求理由、比对规则并据此修改；(3) 任务主导权，如提示是否持续推进创作，是否将判断责任完全交给模型。此三类行为亦可视为 MTA 在本任务情境中的可观察指标，依次对应对任务目标、评价标准与反思解释的觉察[8]。

3. 研究方法

本文采用解释性案例研究与逐轮对话分析。分析单位为一次“学生提示－模型响应－后续处理”的互动轮次。每份记录包含学生输入的提示词(P)、模型呈现的推理文本或思考输出(D)、模型答案(A)与教师批注(T)。需要说明的是，D 仅指系统向使用者显示的文本，不等同于模型内部完整的推理机制。教师批注由一名中国古典诗学教师完成，主要依据五言律诗的平仄、押韵、黏对与对仗规则核验输出，并记录学生是否提出质疑、要求证据或进行有效修订。这种“模型输出－规则核验－专家复核”的思路与近期混合评价框架相近[1] [2]。单一专家判断可能带来解释偏差，是本研究的重要限制。

写作任务要求学生选择一位唐宋文人，围绕相关轶事创作一首仄起格五言律诗，融入自身的故乡意象，并押《平水韵》下平声一先韵。每名学生完成两项作业：第一项不限制提示次数；第二项限为七次，以“七步成诗”典故作为任务情境。此设计不预设提示次数越多错误率越高，也不将七次视为最佳值；其目的在于形成一个可比较的资源约束，观察学生如何分配提示、选择核验重点并维持对创作任务的主导。

课程共有两个班级、约 150 名学生完成同类作业。本文从中选取两名学生的四份完整记录进行目的性案例分析：学生 A 代表低投入、被动接受的互动模式；学生 B 代表积极参与但规则知识不足、容易受模型解释影响的模式。选择这两例是为了呈现机制差异，而非估计两类行为在全体学生中的比例。本文所有结论均限于案例解释，不作总体推论。后续研究需报告更系统的抽样与编码程序，并以多名专家进行复核。

为使后文引用案例时有明确的对照基础，表 1 汇总本研究人机交互过程的设计与两个案例的选取思路。此一取向亦本于 MTA 之说：既然 MTA 通过可观察的互动模式得以体现[8]，则逐轮保存学生的提示、模型的回应与教师的核验，即为观察 MTA 在真实任务中如何运作或失效的可行途径。

Table 1. Design of the human-ai interaction process and case selection
表 1. 人机交互过程与案例选取的设计

设计层面	具体设计	设计意图
写作任务	择唐宋文人轶事，代入其人，作仄起格五言律诗一首，融入故乡意象，押《平水韵》下平一先韵，并须自订诗题	形式规范明确且多重(平仄、押韵、黏对、对仗)，可外部裁决，便于区分“提示中明示的要求”与“未明示的体裁要求”
提示条件(作业一)	不限制提示次数	观察资源充裕时学生自发的互动策略与核验习惯
提示条件(作业二)	严格限为七次，以“七步成诗”为任务情境	使提示成为有限资源，迫使学生对核验重点作出取舍，藉以显现其任务优先级

续表

记录结构	每轮含学生提示(P)、模型显示的思考文本(D)、模型答案(A)、教师批注(T)	使“提示建构–系统回应–反思解释”三环节皆可追溯；D 仅为系统显示文本，不等同模型内部机制
分析单位	一次“学生提示–模型响应–后续处理”的互动轮次	以轮次为单位，方能观察错误在轮次之间被保留、放大或粉饰的过程
专家核验	由一名中国古典诗学教师依平仄、押韵、黏对、对仗逐句复核，并记录学生是否质疑、举证或有效修订	提供独立于模型的外部判准；单一专家为本研究的重要限制
案例 1 (学生 A)的选取	低投入、被动接受型：作业一仅提示一次；作业二数次提示偏离任务(如要求赏析诗作)	呈现“提示用于满足任务形式”的机制，对应任务目标觉察薄弱
案例 2 (学生 B)的选取	积极参与但规则知识不足型：自订诗题、与模型辩论、要求修改“读来别扭”之句、指定诗眼方向	呈现“审美觉察强而评价标准觉察弱”的机制，对应积极互动未必导向有效纠错
两案例的对照逻辑	二者投入程度相反，最终却同样未发现颔联、颈联的对仗问题	显示纠错失败可循不同路径达成，关键不在投入多寡，而在是否具备可操作的评价标准

4. 初步发现

4.1. 模型的表现：合规、遗漏与自我检查中的幻觉

两名学生的四份作业显示，DeepSeek-V4 在提示明确列出平仄与押韵要求时，多数情况下能够生成大体合规的句式；但错误并未消失，且部分问题较不易被学生察觉。以下发现仅描述所选案例，不代表模型整体性能。

第一，模型可能遗漏未在提示中明确列出的体裁要求，最突出的是颔联与颈联的对仗不工。此处需区分“失对”与“对仗不工”：前者通常指同一联上下句的平仄未形成应有的相对关系；后者则涉及句法结构、词性及语义关系未能对应。案例中多处问题主要属于后者，例如上下句对应位置的词性或结构不匹配，因而不能仅以“失对”概括。结合多重约束生成研究[1] [2]与推理失败综述[4]，较稳妥的解释是：模型在同时处理平仄、押韵、对仗、用典与意象时，可能发生要求遗漏；这是一种与任务复杂度相关的现象，但本研究不足以判定其内部因果机制。

第二，模型在自我核验时可能产生错误修正与术语误用。案例中，模型有时修改原本无误的诗句，却未处理真正的问题；也曾把不合规范的声律解释为“拗救”，或以“宽对”为结构不对应的句子辩护。教师依据格律规则复核后，认为这些解释缺少成立条件。此现象与文献所揭示的 LLM 评价偏差相呼应[2]，但本文只主张案例层面的对应，不将其外推为所有模型自检的普遍规律。

4.2. 学生的参与：低投入与积极但知识不足

两名学生呈现出对照鲜明的参与方式。学生 A 在不限提示次数的任务中仅输入一次提示；在限为七次的任务中，数次提示转向赏析或重述创作意图，而未触及作品中的规则问题。就这两份记录而言，提示次数被用于满足任务形式，而非持续核验与修订。

学生 B 投入较多，能自拟题目、与模型讨论，并要求修改“读来别扭”的诗句或调整意象，显示出

较强的语言感受与创作主动性。可是，她对声律与对仗规则的掌握不足：虽察觉部分句子可能有问题，却难以准确定位；当她使用自己尚未充分理解的术语要求模型处理时，也无法判断模型解释是否成立。该案例说明，积极互动与审美敏感能够推动修订，却不能替代可操作的格律知识与外部核验。

此外，案例 2 还显示出一种值得注意的失败机制：术语委托。学生 B 在第一份作业中要求模型“不要犯四不犯和三不破的格律错误”，然而这两组术语在诗学论述中并无统一界定。模型并未要求学生澄清，而是自承“‘四不犯’‘三不破’并非通用术语”，随即“依据最严格的传统诗学禁忌”自拟标准，逐项宣告何者“犯忌”，并据此提出修改方案；其所交“故里橘垂涕”一句仍犯孤平，而模型在其后各轮均判定“无孤平”。教师批注指出，课堂上已不只一次要求学生须自行说明此类术语的判准，不可交由模型代查。此一轮次的意义在于：提示中的定义空缺不会保持为空缺，模型倾向以自订标准填补，核验的基准因而由被核验者设定。

表 2 与表 3 概括两名学生案例在作业中体现的行为及素养维度。

Table 2. Turn-level interaction behaviors in the two cases
表 2. 两个案例的轮次层面交互行为

阶段	学生 A	学生 B
不限制提示词输入次数	一次提示；对输出照单全收	六次提示；自拟标题、主动求证、提出修改
硬性限制输入七次提示词	数次用于偏离任务的请求(如“赏析此诗”)	所有提示词均指向推进诗作
提示词次数的使用方法	为凑足次数而填充	审慎、以创作为中心
格律与对仗的关注	无明确核验	关注行文流畅，但未进行明确的格律或对仗核验

Table 3. Literacy and knowledge-construction dimensions
表 3. 素养与知识建构维度

维度	学生 A	学生 B
格律知识	基本缺失；不知对仗规则	审美感受较强；规则知识不足
创作主动性	被动接受	高；拟题、反馈、定向
主要风险	整体接受错误输出	被模型看似专业的解释说服
对 AI 的依赖	完全依赖	部分依赖，但在关键处过度信任
案例结果	格律错误未获纠正	积极投入但仍残留格律与对仗问题

4.3. 关键互动轮次的完整记录

为使上述发现可供检核，本节并列呈现两组关键轮次的原始记录。表 4 取自案例 1 第二份作业，显示模型自我核验时的误判方向；表 5 取自案例 2 第二份作业，显示模型以格律术语为错误辩护的过程。两表均按“学生提示(P) – 模型思考文本(D) – 模型输出(A) – 学生后续操作(P) – 专家批注(T)”的顺序排列(轮次以数字表示)，以呈现错误如何在轮次之间被保留至定稿。

Table 4. A complete interaction round in case 1: misdirected self-verification
表 4. 案例 1 的完整互动轮次：自我核验的误判

环节	记录内容
学生提示(P2)	“请你仔细按照仄起格五言律诗的要求与押下平声一先韵，检查你的诗歌韵律是否正确”
模型思考文本(D2)	判断使用者“需要确认这首诗在格律上是严谨无误的”，声明“必须仔细核对每一个字的平仄和韵脚”，并预告“如果发现问题……我必须坦诚指出，并提供修改方案”
模型输出(A2)	先列出仄起格标准句式，继而逐字标注，判定四处“出律”：“急诏惊宿醉”第四字“宿”、“泼墨走云烟”第二字“墨”、“曲终人未醒”第四字“未”、“名动九重天”第二字“动”。结论为“押韵上完全正确，但在平仄格律上有多处出律”，并主动提议 提供“一版严格合律的修改稿”
学生后续操作(P3)	“你按照要求修改诗歌的相关语句”——全盘接受模型的判定，既未要求其说明所依据的句式，也未自行比对格律表；A3 所交修订稿即以此四处“出律”为修改依据
专家批注(T2~T3)	模型所列第四、七、八句之“出律”实皆不误，属误判；第三句虽被指出有问题，却误指为“宿”字，真正出律的“诏”字未获检出。A3 修订稿平仄虽合，颔联“急宣惊薄醉/泼雨化云烟”仍不成对仗，学生亦未察觉

Table 5. A complete interaction round in case 2: defending an error with technical terms
表 5. 案例 2 的完整互动轮次：以术语为错误辩护

环节	记录内容
学生提示(P6)	“选择秋山红似旧，不道是何年，注重检查该诗平仄和押韵问题，尤其注意不要犯格律问题”
模型思考文本(D6)	判定第五句“恍作霜橘挂”为“仄仄平仄仄”之“大拗”，随即推断第六句“垂垂苍壁边”为“平平平仄平”，并自行认定二者“正好拗救”“完全合规”
模型输出(A6)	交出定稿，并于格律说明中明示“颈联……仄仄平仄仄，平平平仄平(‘橘’入声，拗而自救)”，声称全诗“严格遵循仄起首句不入韵式”
学生后续操作(P7)	“最后检查定稿是否满足综上所有要求”——未质疑“拗救”是否成立，而是再次请模型自我确认；因本项作业限七次提示，此后已无修正机会
模型再次回应(A7)	重申“无孤平、三平调、失对、失粘”，并将颔联、颈联的不对仗表述为“宽对(意境相对，字面略欠工整)”，附言“若需严格可再调整，但鉴于作业未明确要求，予以保留”
专家批注(T6~T7)	“橘”字出律并未因所谓“拗救”而获救，该救法在此不成立，属明显幻觉；颔联“庭空如积水/藻荇影交缠”与颈联“恍作霜橘挂/垂垂苍壁边”皆未构成对仗，以“宽对”称之属强词夺理。两轮核验均由模型自证，未曾引入外部规则比对

5. 讨论与意义

将文献与案例并置可见，模型能力提升并未自动消除教学风险，而可能改变风险的呈现方式。较新的模型能生成更流畅、整体更接近规则的诗作，但局部错误因而更隐蔽；当模型再以专业术语解释这些错误时，学生更难仅凭语言流畅度作出判断。Ma 等人所说的“回声室”效应指向 LLM 评价与专家判断之间的系统偏差[2]。本研究的案例并非对该效应的重复验证，而是显示类似风险如何在课堂对话中具体表现：模型的自我解释不能替代独立规则核验。

上述失败可在 MTA 框架下获得更完整的诠释。作品能否合律，取决于人机协作整体是否至少保有一条有效的检错通道；而这条通道的存续，依赖于学习者对任务评价标准的觉察。表 4 与表 5 显示两端同

时失效：模型一端因需同时维持平仄、押韵、对仗、用典与意象等多重约束而遗漏对仗，与推理失败综述所述“任务要求超出可维持信息容量时出现遗漏与不一致”的情形相符[4]；学生一端则因缺乏可操作的对仗判准，无从履行外部核验。两端的盲区落在同一条规则上，协作因而失去冗余。就 MTA 而言，此非提示技巧的不足，而是任务觉察在“评价标准”这一维度上的空缺——学生能陈述任务要写什么，却不能陈述据何判定其写得合规。

更关键的是核验的同源性。学生反复要求的“检查”由生成者自身执行，其误差与生成误差并不独立：模型据以自检的规则表征，正是它据以生成的那一套表征。表 5 中，“拗救”的判定与被判定的诗句同出一源，故一次自检并未增加一次独立验证，只是把同一判断重复一次。Ma 等人所记录的“回声室”效应原指 LLM 评价者偏好模型自身的统计模式[2]；在课堂对话中，这一效应的作用范围由“模型评价模型”扩及“人机二元体内部的相互确认”。这也解释了一个表面矛盾的现象：学生 B 在第二份作业中用去七次提示，其中两次明确要求模型核验格律，然而颔联与颈联的不对仗自始至终未获纠正，第五句更在修订过程中新增“橘”字出律。核验轮次的增加并不等于可靠性的提高——当唯一的验证者与生成者同源时，轮次只是提供了让同一错误被反复确认、并以日益自信的措辞加以陈述的机会。以思维回路观之，此即“反思解释”环节被生成者占据的结果：回路仍在转动，却不再引入回路之外的判准[8]。

这也需要更谨慎地理解“过度依赖”。学生 A 的风险来自低投入与整体接受；学生 B 的风险则来自知识尚未稳固时，对模型技术性解释的过度信任。Lee 等人在知识工作场景发现，对 GenAI 的信心越高，自报的批判性思维投入越低，而任务自信越高则相反[6]；本研究的两例与这一关系具有启发性的对应，但学生案例与知识工作者调查的对象和情境不同，不能直接等同。更合理的结论不是“懂得越多越容易被骗”，而是局部知识与积极参与并不足以自动形成可靠判断；学习者仍需把直觉转化为可说明、可核验的规则。

据此，学生 B 的困境不宜简化为“信任过度”。她确实产生了有效的错误信号——察觉诗句“读来别扭”、指出首句与颈尾“生硬不自然”——但直觉只能报告异常的存在，不能定位异常，更不能给出判定异常的理由；模型以“拗救”“宽对”提供的形式化解释，恰好补上了她无法自行完成的定位与举证环节，于是低分辨率的正确怀疑被高分辨率的错误结论取代。这提示 MTA 各维度并非齐头并进：任务目标觉察与评价标准觉察可以显著分离，而后者才是审核 AI 输出的关键。表 6 依 MTA 的各维度并列两案例的表现及其后果。

Table 6. Manifestations and consequences of MTA dimensions in the two cases
表 6. 元任务觉察各维度在两案例中的表现与后果

MTA 维度[8]	案例 1 (学生 A)	案例 2 (学生 B)	对纠错的后果
任务目标觉察	薄弱：提示用于凑足次数，转向赏析与重述创作意图	较强：自订诗题、主动调整意象与诗眼	决定是否持续推进任务，但不足以保证合律
评价标准觉察	缺失：未提出对仗要求，亦未核验	缺失：察觉异常却无法定位与举证	两案例的共同缺口，为纠错失败的关键所在
提示建构	单轮倾倒全部要求	分轮递进，但曾委托模型自行界定术语	提示只能传递已被觉察的要求；未觉察者必被遗漏
反思解释	让渡：全盘接受模型判定	让渡：接受“拗救”“宽对”之说	思维回路的反思环节由生成者占据，回路失去外部参照
轨迹意识	缺乏：各轮次彼此孤立	部分具备：后续提示承接前轮修订	提升投入品质，仍不能替代评价标准

提示次数限制的价值主要是研究性，而非教学效果本身。本团队前一阶段研究未设置此类限制[9] [10]；本次将第二项任务限为七次，使提示成为有限资源，从而较清楚地显现学生如何安排创作、核验与修订。学生 A 出现填充性提示，学生 B 则较集中地推进诗作。由于样本只有两人，本文只能说该设计“有助于显现互动策略差异”，尚不能断言它能提高学习成效或准确测量学习态度。

就教学应用而言，LLM 较适合被定位为诊断性“陪练”，而非格律答案的终审者。教师可要求学生标出疑点、引用规则、提出替代方案，并比较模型修订前后的得失；如此，模型的不可靠性才可能转化为练习批判性判断的材料。前提是先提供足够的基础知识、格律表与可复核工具。学生也不宜把自己尚未理解的术语直接交给模型后照单全收，而应说明术语的适用条件并要求模型给出逐字、逐句的证据。这样的任务设计把关键责任留在学习者与教师一侧，也回应了认知外包研究对保留认知投入的提醒[5]-[7]。

上述建议之所以可能有效，其理由可由前文的系统分析导出，而非仅凭经验直观。第一，以格律表或程序化工具进行核验，作用在于为系统重新引入一条与生成过程相互独立的检错通道；平仄与对仗的判定可还原为字词的声调归部及词性、结构的位置比对，属于可外部裁决的形式规则，因而不必交由生成者自评。此点亦与文献相互印证：形式约束的符合度可通过规则化的评价与反馈获得改善[1]，而经专门微调的较小模型在特定形式任务上可优于通用大模型[3]——两者共同说明，格律核验所需的是明确的判准与针对性的知识，而非更强的通用生成能力。第二，要求学生在提示中自行给出术语的操作定义，作用在于消除定义空缺；如 4.2 所示，空缺一旦存在，模型便倾向以自订标准填补，核验的基准反而由被核验者设定。第三，将评分重心由“作品是否优美”移向“能否找出并论证模型的错误”，作用在于改变任务的可代劳性：前者模型能够代为完成，后者则必须由学习者保有判准方能完成，认知投入因而由可选项转为任务的必要条件。换言之，这三项设计都在把 MTA 由内在倾向转为外显要求：先陈述判准，再据判准审读输出，正是 MTA 可教、可练的具体形式[8]。

须强调第三项的前提。本研究恰恰显示，当学生与模型同时缺乏判准时，“找错”任务无从成立——两名学生皆未发现对仗问题。因此“诊断性陪练”是评价标准觉察达到可判定水平之后的教学策略，而非替代基础教学的捷径。就跨领域推广而言，可迁移者并非近体诗教学法本身，而是上述三项机制所依赖的条件；其适用范围与主要挑战概述于表 7。

Table 7. Conditions and challenges for cross-domain transfer
表 7. 跨领域推广的条件与挑战

项目	内容
可迁移的机制	形式规范的可外部判定性；术语判准的可操作化；评分对象由产出转向检错
适用领域举例	数学证明的步骤合法性、程序设计的语法与单元测试、法律文书的条文引用、化学方程式的配平、外语的语法与韵律训练
挑战一：判准的情境性	文学诠释、伦理论证、设计审美等领域缺乏共识判准，独立核验通道无从锚定，此一设计难以直接移植
挑战二：术语标准化程度	工程与医学多有规范定义；人文领域常因流派而异(如本研究“四不犯”“三不破”并无统一说法)，对学生操作化能力要求更高
挑战三：门槛依赖	方案对已具基础者更为有利，可能扩大差距；正规教育对认知外包具缓冲作用[5]，故教学支架的分配须审慎设计
挑战四：教师端成本	依赖专家逐句批注(本研究仅一名专家)，大班教学需可替代的规则工具或评分量表

本研究有四项限制。第一，分析仅涵盖两名学生的四份作业，案例具有示例性，不能推论总体分布。第二，仅由一名专家核验，尚缺少多评者一致性与三角互证。第三，只考察单一模型与单一写作任务，结果可能受版本、提示方式及课程情境影响。第四，本文依据对话文本解释知识建构与批判性思维，未使用前后测或独立量表评估学习成效。后续研究应扩充样本、预先制定编码本、报告评者一致性，并比较不同模型、提示条件与教学支架。

综合而言，本文结论的适用范围可界定如下。本研究能够支持的主张是：在所观察的案例中存在一条可辨识的失败路径——未明示的体裁规则被遗漏、自检因同源而失效、学生因缺少判准而无法拦截，三者叠加使错误留存至定稿。本研究不能支持的主张至少有四类：其一，不能断定该路径在同类学生中的发生比例，两例的选取旨在呈现机制而非估计分布；其二，不能断定模型版本更新使风险整体上升或下降，本文并未设置跨版本对照；其三，不能断定学生 B 的失败由模型单独造成——本次设计刻意排除了教师即时介入与规则工具，若保留这些支架，她的审美直觉本有转为规则学习的可能；其四，不能由对话文本推断学习成效的变化，本文未使用前后测或独立量表。上述界限并不削弱案例的价值：机制性案例的作用在于提出可检验的路径假设，而非估计效应量。

6. 结论

本研究以两名学生的四份人机对话为例，探讨 LLM 辅助五言律诗创作时的规则遵循、知识建构与批判性判断。案例显示：第一，模型较能处理提示中明确列出的平仄与押韵要求，却可能遗漏对仗等未明示规则，并在自检时产生术语化但不成立的解释；第二，低投入与知识不足会以不同方式削弱纠错，积极参与本身不能替代稳固的格律知识；第三，七次提示限制在本案例中有助于显现学生分配提示与维持任务主导权的差异。

因此，LLM 在近体诗教学中的价值不应以“能否快速写出一首诗”衡量，而应看它能否被纳入可核验的学习流程。较可行的路径是：先建立基本格律知识，再要求学生对模型输出逐项举证、质疑与修订，并由教师或规则工具进行复核。本文提出的是一个待检验的案例框架，而非关于学习成效的定论；其适用范围仍需通过扩大样本、多专家评定与跨模型比较加以检验。若以元任务觉察总结全文，三点发现可归为一句：模型能力的提升改变了错误的外观，却未改变纠错所需的条件——学习者对任务评价标准的觉察，以及思维回路中反思解释环节不被让渡[8]。在 AI 日趋主动的后提示时代，此一条件只会更为关键。

基金项目

本研究接受黄冈师范学院教学研究项目“教育大语言模型于现代教育技术的应用研究”(编号：0601202435)。

参考文献

[1] Qu, Z., Yuan, S. and Färber, M. (2026) PoeTone: A Framework for Constrained Generation of Structured Chinese Songci with LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, **40**, 32764-32772. https://doi.org/10.1609/aaai.v40i39.40555

[2] Ma, B., Yao, Y. and Haensch, A.C. (2025) Capabilities and Evaluation Biases of Large Language Models in Classical Chinese Poetry Generation: A Case Study on Tang Poetry. arXiv: 2510.15313. https://arxiv.org/abs/2510.15313

[3] Tang, Y., Li, Z., Yu, J. and Yang, L. (2025) A Simple Approach of Chinese Poetry Generation Using Pre-Trained LLMs. *Proceedings of the* 2025 7*th Asia Pacific Information Technology Conference*, 10-12 January 2025, 106-112. https://doi.org/10.1145/3726101.3726121

[4] Song, P., Han, P. and Goodman, N. (2026) Large Language Model Reasoning Failures. https://openreview.net/forum?id=vnX1WHMNmz

[5] Gerlich, M. (2025) AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, **15**, Article 6. https://doi.org/10.3390/soc15010006

[6] Lee, H.P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., et al. (2025) The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects from a Survey of Knowledge Workers. *Proceedings of the* 2025 *CHI Conference on Human Factors in Computing Systems*, Yokohama, 26 April-1 May 2025, 1-22. https://doi.org/10.1145/3706598.3713778

[7] Žnidarič, U., Štrumbelj, E. and Machidon, O. (2026) A Review of the Negative Effects of Digital Technology on Cognition. arXiv: 2603.10025. https://arxiv.org/abs/2603.10025

[8] Wong, L. (2025) Transforming Education with Generative AI: Designing for Post-Prompting Era. *Information and Technology in Education and Learning*, **5**, Inv-p003. https://doi.org/10.12937/itel.5.1.inv.p003

[9] Hsu, H.C. and Chen, H.C. (2025) Large Language Model-Assisted Regulated Verse Composition: Practice and Reflection Based on Human-AI Interaction Analysis of DeepSeek. *Proceedings of the* 29*th Global Chinese Conference on Computers in Education* (*GCCCE* 2025), *Workshop Proceedings*, Wuxi, 24-28 May 2025, 607-615.

[10] Hsu, H.C. and Chen, H.C. (2026) The Application of Large Language Models in Classical Poetry Teaching: A Case Study of DeepSeek's Prosodic Cognition in Qiyan Jueju. *Journal of Education Research*, **4**, 84-90.