Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Method for Reasoning Attribution Identification and Instructional Early Warning in Medical Microbiology Education: A Randomized Controlled Trial

Yue Shi
Changchun University of Chinese Medicine
Silin Zhang
Changchun University of Chinese Medicine
Lu Hou
Changchun University of Chinese Medicine
Fang Wang
13514487032@163.com

Changchun University of Chinese Medicine

Research Article

Keywords: medical microbiology education, generative artificial intelligence, reasoning attribution identification, counterfactual premise perturbation, knowledge space theory

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11116981/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Method for Reasoning Attribution Identification and Instructional Early Warning in Medical Microbiology Education: A Randomized Controlled Trial

Yue Shi[1], Silin Zhang[1], Lu Hou[1], Fang Wang[1,*]

[1] School of Basic Medical Science, Changchun University of Chinese Medicine, Jilin 130117, PR China

* Correspondence: Fang Wang, 13514487032@163.com

Email addresses:
Yue Shi: octoberpearl@126.com
Silin Zhang: zhangchuchu1103@163.com
Lu Hou: houlu@ccucm.edu.cn
Fang Wang: 13514487032@163.com
ORCID: Yue Shi, https://orcid.org/0000-0002-1112-4645

Abstract

Background: The widespread adoption of generative artificial intelligence in medical education has made the phenomenon of "AI borrowing" in the classroom a direct threat to the evaluation of learning authenticity. Pathogen reasoning tasks in medical microbiology inherently have a strict dependency structure. However, existing AIGC text detectors remain confined to surface features, neural cognitive diagnosis systems stop at knowledge-component-level scores, and educational early warning tools provide only outcome labels. None of these approaches fills the methodological gap between node-level attribution identification and actions that teachers can execute.

Methods: This study proposes EduAttr, which introduces three principled innovations. The reasoning commitment consistency test based on counterfactual premise perturbation introduces the intervention semantics of causal inference into attribution identification. The identifiable path signature based on the dependency partial order violation rate uses the downward closure property of knowledge space theory and forms a complementary coupling with the counterfactual probe. The instructional early warning node ranking mechanism integrates dual-channel signals with dependency depth weights and outputs a Top-K list of follow-up questions that teachers can execute. The system was evaluated in a two-arm randomized controlled trial at two medical colleges in China between September 2024 and June 2025, with 240 second-year clinical medicine students in eight intact classes allocated 1:1 to supervised writing without generative AI or to AI-permitted writing whose interaction transcripts anchored node-level attribution labels. The primary outcome was the node-level F_1 score.

Results: Across 1,200 reports and 22,362 annotated nodes, EduAttr achieves a node-level F_1 of 0.843, exceeding the strongest surface feature baseline Ghostbuster by 13.0 percentage points. Under severe rewriting attacks, this gap further expands to 29.5 percentage points. Agreement with the adjudicated labels reaches a Cohen's kappa of

0.762, which is statistically indistinguishable from the interteacher agreement of 0.741, the teacher adoption rate is 87.5%, and grading time is reduced by 42.7%.

Conclusions: EduAttr advances identification accuracy, robustness, and instructional operability, providing a technical pathway that can be embedded into classroom follow-up questioning workflows in medical education.

Trial registration: Chinese Clinical Trial Registry, ChiCTR2600118749. Registered 2 October 2026 (retrospectively registered).

Keywords: medical microbiology education; generative artificial intelligence; reasoning attribution identification; counterfactual premise perturbation; knowledge space theory

Introduction

In recent years, generative artificial intelligence technology has penetrated higher education at an exceptionally rapid pace, and medical education, a field with the most stringent requirements for rigorous reasoning chains, has been among the first to be affected. As a bridge discipline between basic medicine and clinical medicine, medical microbiology requires each pathogen case analysis by a student to proceed sequentially through specimen description, morphological observation, mechanistic hypothesis, and clinical conclusion, with explicit causal dependencies among the corresponding nodes. The widespread availability of large language models such as ChatGPT and GPT-4o enables students to obtain apparently complete reasoning reports within seconds, while instructors have difficulty determining, within limited grading time, whether each reasoning step was completed independently by the student or generated with AI assistance. This phenomenon of "AI borrowing" not only threatens academic integrity but also fundamentally undermines the pathway centered on case analysis for cultivating pathogen reasoning competence, depriving instructional evaluation of a reliable means of assessing learning authenticity.

Research addressing this challenge has developed along three main lines. Large language models have reached or even surpassed the performance of typical medical students in medical multitask question answering and clinical reasoning tasks [1], and their capacity to encode clinical knowledge has been demonstrated across diagnosis, treatment, and mechanistic explanation [2]. The medical education community has systematically summarized the pathways and risks associated with integrating generative AI into curricula and assessment [3], while review evidence has also indicated rapid growth in this research area but generally weak methods for evaluating learning authenticity [4]. Analyses in educational technology have identified the double-edged effects of large language models in the classroom. They can serve as personalized tutoring resources, but they can also weaken students' critical thinking and writing practice [5]. The benefits, boundaries, and risks of GPT-4 in medical question answering have been systematically discussed [6], whereas publicly available AI text detection tools in commercial and academic settings commonly exhibit substantial declines in accuracy across domains, languages, and scenarios involving disguise through rewriting [7]. Evaluations of clinical decision-making further show that large language models retain systematic limitations in causal consistency and structural errors, and these defects are difficult to detect directly from surface outputs [8]. Research on the interpretability of diagnostic reasoning prompts further highlights deficiencies in reasoning transparency at the model level [9]. Collectively, these studies reveal the same methodological gap. Existing detection methods treat student reports as flat text, ignore the dependency structure inherent in pathogen reasoning, and cannot translate identification results into classroom follow-up questioning actions that teachers can execute directly.

To address this gap, this study proposes EduAttr, a pathogen reasoning attribution identification and instructional early warning system designed for classroom teaching scenarios, and introduces principled innovations in three dimensions. The reasoning commitment consistency test based on Counterfactual Premise Perturbation, CPPC, introduces the intervention semantics of causal inference into AI borrowing

identification. It replaces static consistency measurement with node-level counterfactual responses, thereby elevating the criterion from "whether the text is internally consistent" to "whether the reasoning remains consistent when the premise is changed." The identifiable Path Signature based on the dependency Partial Order Violation Rate, POVR-Sig, constructs a principled criterion using the downward closure property of knowledge space theory. It extracts the structural misalignment inherent in AI-generated content, in which "a downstream node appears correct while an upstream node is actually incorrect," as a computable signature and forms a complementary coupling with CPPC that is difficult for the same generation strategy to evade simultaneously. The Teaching Alert Node Ranking mechanism, TANR, integrates the dual-channel identification signals with dependency depth weights and outputs a Top-K list of follow-up questions that teachers can execute directly, thereby closing the long-standing operational gap between "identifying AI borrowing" and "generating classroom follow-up questioning actions." The overall research framework is shown in Figure 1. All innovations are systematically validated on data from a two-arm randomized controlled trial in medical microbiology education, and their effectiveness is evaluated in terms of identification accuracy, robustness, and alignment with teacher judgments.

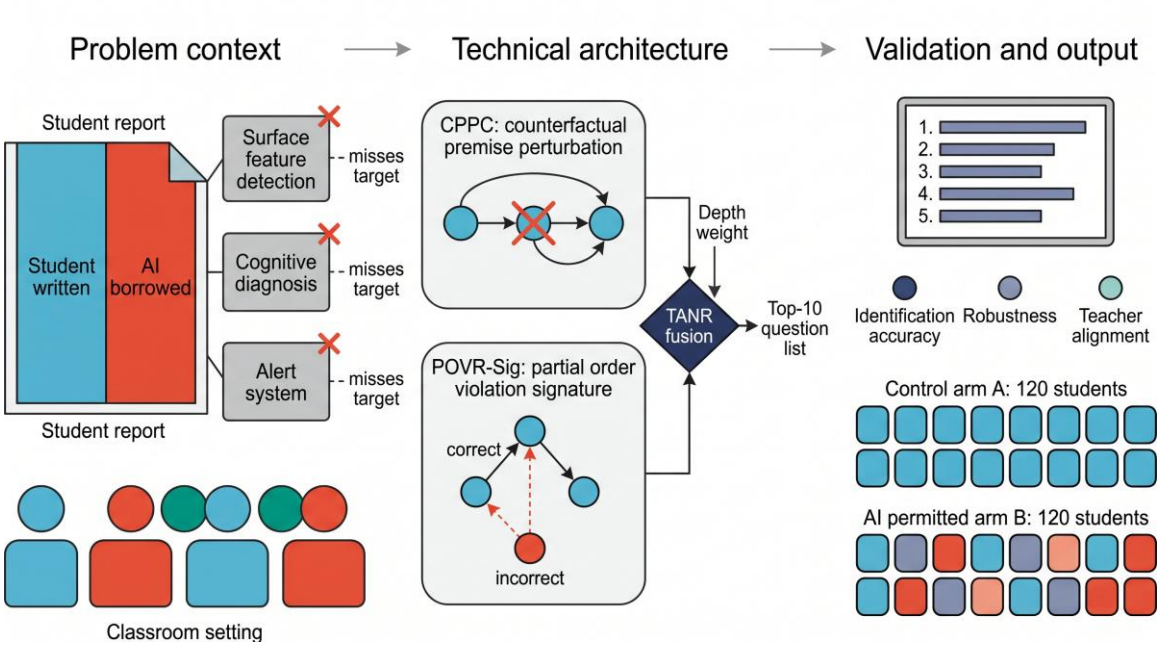


Figure 1 Overall research framework of this study

Related Work

Generative Content Detection Methods and the Bottleneck of Surface Linguistic Features

Mainstream approaches to identifying AI borrowing place their criteria on statistical features at the textual surface and perform binary classification using word probability distributions, perplexity curvature, or feature selection based on weak model ensembles. Recent systematic analyses have reached relatively cautious conclusions regarding this pathway. Current detection methods are excessively sensitive to text length, writing style,

and successive generations of generative models, and the stability of their criteria drifts significantly as scenarios change [10]. In educational contexts, the social cost of this drift is more pronounced. Mainstream commercial detectors exhibit systematic bias against texts written by nonnative English authors, misclassifying more than half of TOEFL samples as AI-generated while achieving nearly perfect accuracy on samples from native-speaking students [11]. Students in medical education classrooms likewise display linguistic characteristics similar to those of nonnative authors, including high terminology density, formulaic sentence patterns, and convergent writing styles, which place them directly within the error range of these detectors. Independent evaluations in academic settings further show that mainstream commercial tools generally lack discriminative ability across models and rewriting scenarios, and their false positive rates rise rapidly as the number of rewrites increases [12]. Collectively, this evidence indicates that the fundamental limitation of the surface feature pathway does not lie in implementation details. Instead, it lies in the basic mismatch of treating "whether the wording was written by AI" as equivalent to "whether the reasoning was completed independently by the student." Consequently, this pathway cannot provide a credible attribution criterion for classroom settings such as medical microbiology, where the rigor of the reasoning chain is central.

Dependency Structured Reasoning Modeling and the Bottleneck in Characterizing Causal Consistency

Another research stream attempts to move beyond lexical statistics and measure consistency at the structural level of reasoning chains. Common approaches include chain-of-thought self-consistency, factual consistency testing, and dependency graph alignment. This direction confirms that exploitable criterion signals do exist at the reasoning-chain level, but it remains insufficient with respect to the fundamental question of what type of consistency can genuinely indicate reasoning attribution. Recent mechanistic research shows that the reasoning commitments of large language models are systematically unreliable. Semantic entropy has been used as an unsupervised signal for hallucination detection, indicating a stable inconsistency within models between "what I said" and "whether I believe what I said" [13]. In broader open domains, factuality challenges have been shown to possess compound, cross-topic, and cross-source complexity, making it difficult for conventional fact-checking workflows to cover the full risk surface [14]. Reasoning assessment in medical education is even more vulnerable. Review evidence indicates that consistency testing for large language models in clinical case analysis often relies on post-hoc judgment and fails to characterize the intervention semantics of whether downstream reasoning updates when an upstream premise changes. It is therefore difficult to distinguish reports that are superficially consistent but substantively borrowed [15]. The common blind spot of this work is that self-consistency, factuality, and dependency chain structural alignment all remain at the level of post-hoc scoring. Premise repairability has not been used as a criterion for distinguishing authentic reasoning from borrowed reports, even though it is precisely the central gap addressed by the CPPC channel in this study.

Learner Modeling and Knowledge Tracing and the Bottleneck Caused by Failure of the Independence Assumption

Learner modeling and knowledge tracing originate from intelligent education and provide a basis for personalized tutoring by mapping students' response sequences to mastery of knowledge components. A recent systematic review of this field traces the full development from Bayesian knowledge tracing to deep knowledge tracing and also notes that these models generally rely on two core assumptions, namely conditional independence among knowledge components and a single latent variable for mastery. Both assumptions fail once generative AI is introduced [16]. When students use AI to complete reasoning, a structural misalignment emerges between observed scores and actual mastery. A conventional model output such as "a score of 0.8 for this knowledge component" cannot distinguish how much derives from the student's own reasoning and how much is borrowed from AI scaffolding. Recent work on Q matrix constrained neural networks attempts to improve diagnostic interpretability by introducing prior knowledge. Although this approach shows some advantages in small-sample scenarios, it does not address the central problem created when prerequisite dependencies are erased. The structural misalignment in which "a downstream node appears correct while an upstream node is actually incorrect" remains unidentifiable along the dependency chain [17]. The common limitation of these studies is that modeling granularity remains at the knowledge component level. They neither treat node-level AI borrowing as an identifiable object nor use the partial order structure of the disciplinary task itself as a principled criterion. These two points constitute the theoretical basis of the POVR-Sig channel proposed in this study.

Educational Early Warning Systems and the Bottleneck in Translating Attribution Results into Instructional Operations

Teacher-oriented instructional early warning workflows deliver identification results to instructors through risk levels, suspicious passage annotations, or knowledge concept profiles. A recent controlled experiment involving teachers showed that teachers themselves achieved less than 40% accuracy in identifying AI-generated text and were more likely to misclassify formal and terminology-dense student responses as AI-generated. This finding indicates that teachers do not need an alert stating "whether it is AI," but rather an actionable output indicating "which node should be questioned first" [18]. Another survey from the teacher perspective shows that teachers' trust in AI-collaborative tools depends on whether the output can directly enter classroom actions. Systems lacking this level of output significantly weaken teachers' willingness to adopt them, meaning that even highly accurate identification results cannot complete the instructional loop [19]. The shared problem in this line of work is a structural mismatch between output granularity and the teacher workflow. Existing alert systems provide outcome labels, whereas teachers require ranked instructions specifying what should be done next. This is precisely the operational gap addressed by the TANR ranking mechanism in this study.

The four categories of bottlenecks described above point from different levels to the same methodological gap. Existing work remains confined to the textual surface, post-hoc scoring, knowledge-concept-level independence assumptions, or alert formats based on outcome labels. None provides an integrated pathway combining node-level attribution identification with ranked instructional actions. EduAttr begins with a

principled analysis of these four bottlenecks and simultaneously introduces intervention semantics, downward closure identification, and instructional action generation at the methodological level, thereby completing the full loop from "identifying AI borrowing" to "generating classroom questioning actions."

EduAttr Method Design

Problem Formalization: From Document-Level Detection to Node-Level Attribution Identification

Conventional AIGC text detection treats a student submitted report R as a flat string and outputs a single binary probability $p(\mathrm{AI}\mid R)$. Its criteria can only rely on surface features such as word frequency, perplexity, and syntactic statistics [20], and the robustness limitations of such methods under cross-model rewriting have been repeatedly discussed in recent evaluations [21, 22]. In pathogen reasoning tasks for medical microbiology, student responses naturally contain parsable dependency relations. From "specimen description" to "morphological observation," and then to "pathogen hypothesis" and "clinical conclusion," each step has an explicit causal dependency on upstream nodes. EduAttr exploits this structural fact and redefines attribution identification as a node-level discrimination problem.

Given a student report R, a domain dependency parser π is used to construct a reasoning dependency graph

$$G=(V,E),\quad V=\{v_1,v_2,\ldots,v_N\},\quad E\subseteq V\times V \tag{1}$$

where V is the set of reasoning nodes appearing in the report, N is the number of nodes, and E is the set of directed dependency edges among nodes. Semantically, $(u,v)\in E$ indicates that node u is a causal prerequisite of node v. Each node v carries a reasoning assertion claim(v) and an evidence citation evidence(v) supporting that assertion. The attribution identification task is formalized as a mapping

$$A:V\to[0,1],\quad A(v)=\Pr(\text{AI borrowing}\mid v,G,R) \tag{2}$$

The closer the mapping value is to 1, the more likely node v originates from AI borrowing. The closer it is to 0, the more likely it reflects independent student reasoning. Compared with a document-level binary classification probability, the node-level mapping A preserves the information that is critical for classroom questioning, namely which step was completed by the student and which step was borrowed.

The overall architecture of EduAttr is shown in Figure 2. Beginning with the original report, the system sequentially performs dependency graph construction, dual-channel attribution signal computation, and instructional early warning node ranking. It ultimately produces a ranked list that can be directly embedded into a classroom questioning workflow.

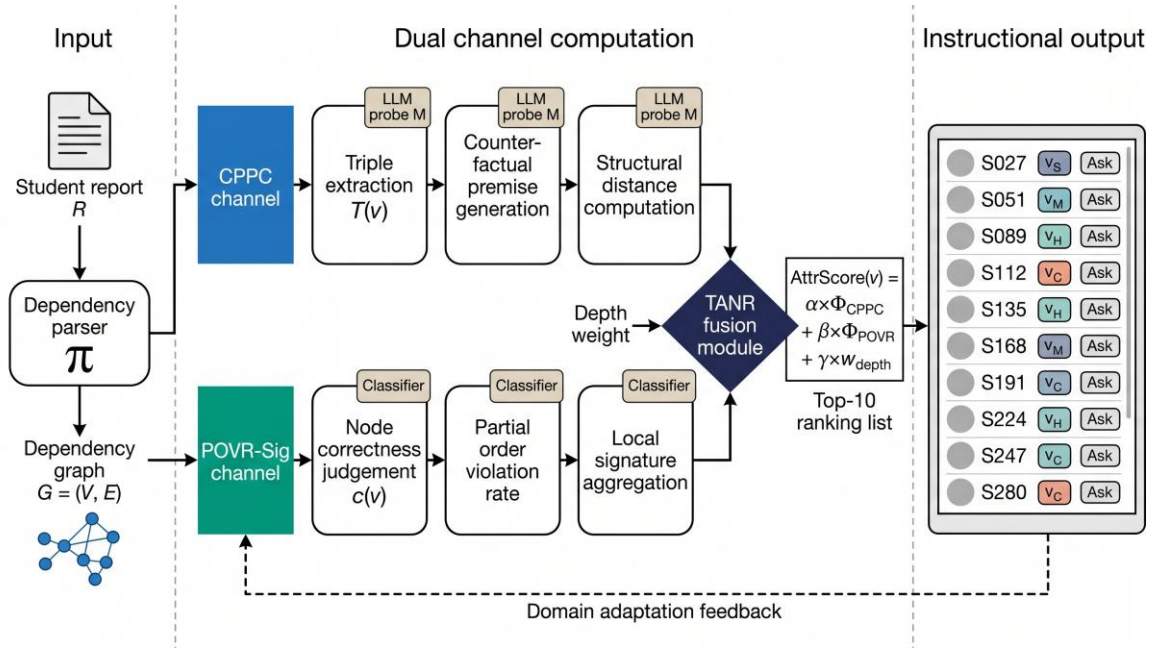


Figure 2 Overall architecture of the EduAttr system

Reasoning Commitment Consistency Test Based on Counterfactual Premise Perturbation

The central insight of CPPC is that the causal commitments of authentic reasoning are repairable. When a student is told that an upstream mechanism does not hold, a student who truly understands the pathogenic mechanism can rederive the downstream assertion along the reasoning chain. By contrast, a report generated with AI assistance is fixed at the level of textual templates and reveals fractures when confronted with a counterfactual premise. Self-consistency measures for chain-of-thought reasoning [23, 24] and NLI criteria both remain static measures of whether the text is internally consistent. They do not use whether reasoning remains consistent after the premise changes as a criterion. Chain-of-thought reasoning without an intervention mechanism has also been shown to exhibit systematic unfaithfulness [25]. CPPC introduces the intervention semantics of causal inference into AIGC attribution identification for the first time, and its rationale is corroborated by findings from counterfactual explanation methodologies for graph neural networks [26, 27].

The directionality of this distance reflects a structural asymmetry between authentic and borrowed reasoning. When a student has internalized the pathogenic mechanism, the claim at a node is anchored primarily in the evidence citation evidence(v); the probing model, forced to accept the negated upstream premise but still constrained by evidence(v), tends to produce a response that remains close in substance to the student report. When the report has been copied from an AI template, however, the claim is tightly entangled with the original premise and lacks grounding in evidence(v); negating the premise therefore propagates to a counterfactual response that diverges from the student report. The distance d is thus small when the student truly understands the mechanism and large when the report has been borrowed. Taking Gram staining as a concrete example, a report that genuinely explains the pathogenic mechanism of

Staphylococcus aureus by referring to coagulase-positive and catalase-positive cocci in grape-like clusters produces a counterfactual response that preserves these evidence-anchored claims even when the probe is instructed to assume Gram-negative bacilli, whereas a template-borrowed report whose claim is lexically bound to Gram-positive morphology produces a sharply divergent counterfactual.

The node-level reasoning commitment triplet is extracted as follows

$$T(v) = \left(v, \text{claim}(v), \text{evidence}(v)\right), \quad v \in V \tag{3}$$

For each upstream node $u \in \text{Pa}(v)$, a probing large language model M is used to generate a counterfactual premise

$$\neg p(u) = M\left(p(u), \text{prompt}_{\text{cf}}\right) \tag{4}$$

where $p(u)$ is the original premise proposition carried by node u, and $\text{prompt}_{\text{cf}}$ is a template that guides M to produce an output that is structurally valid but semantically opposite. Taking Gram staining as an example, if $p(u)$ asserts that "the bacterium is a Gram positive coccus," then $\neg p(u)$ is induced as "assume that the bacterium is actually a Gram negative bacillus." After the counterfactual premise is substituted, the downstream node v is reasoned again by the probing model to obtain a counterfactual response

$$r\left(v|\ \neg p(u), \text{evidence}(v)\right) = M\left(\neg p(u), \text{evidence}(v)\right) \tag{5}$$

CPPC is not concerned with the surface similarity between r and the original text, but with their structural distance on the dependency graph. The distance measure is defined as

$$d\left(r_1, r_2\right) = \lambda_1 \cdot d_{\text{sem}}(r_1, r_2) + \lambda_2 \cdot d_{\text{struct}}(r_1, r_2) \tag{6}$$

where d_{sem} is the cosine distance between domain fine-tuned BERT embeddings, and d_{struct} is defined as the graph kernel distance on the dependency subgraph. It measures structural differences between two reasoning processes in causal connectives, evidence citations, and dependency edge relations. λ_1 and λ_2 are weights with default values of 0.4 and 0.6. A higher weight is assigned to the structural term because text-level similarity can be easily disguised by AI through colloquial rewriting, whereas similarity in dependency structure is difficult to fabricate simultaneously using the same strategy. Before weighting, both components were min-max normalized to the interval [0, 1] using statistics computed on the training set. After integrating all upstream nodes, the CPPC score of node v is

$$\phi_{\mathrm{CPPC}}(v)=\frac{1}{|\mathrm{Pa}(v)|}\sum_{u\in\mathrm{Pa}(v)} d\Big(r\big(v|\ \neg p(u)\big), r_{\mathrm{obs}}(v)\Big) \tag{7}$$

When a student genuinely understands the reasoning mechanism, $r_{\mathrm{obs}}(v)$ and the counterfactual reasoning converge, and $\phi_{\mathrm{CPPC}}(v)$ is small. When a report borrows an AI template, $r_{\mathrm{obs}}(v)$ remains fixed in the original template, and $\phi_{\mathrm{CPPC}}(v)$ increases substantially. This score is equivalent to performing a structural auscultation of causal coupling at every node in the reasoning chain and represents an intervention semantic signal that no static consistency measure can capture.

The operating mechanism of CPPC is illustrated in Figure 3. The upper portion of the figure presents the path of authentic student reasoning, whereas the lower portion presents the path of AI borrowing. The separation between their response trajectories in the dependency graph space after counterfactual premise substitution is clearly observable.

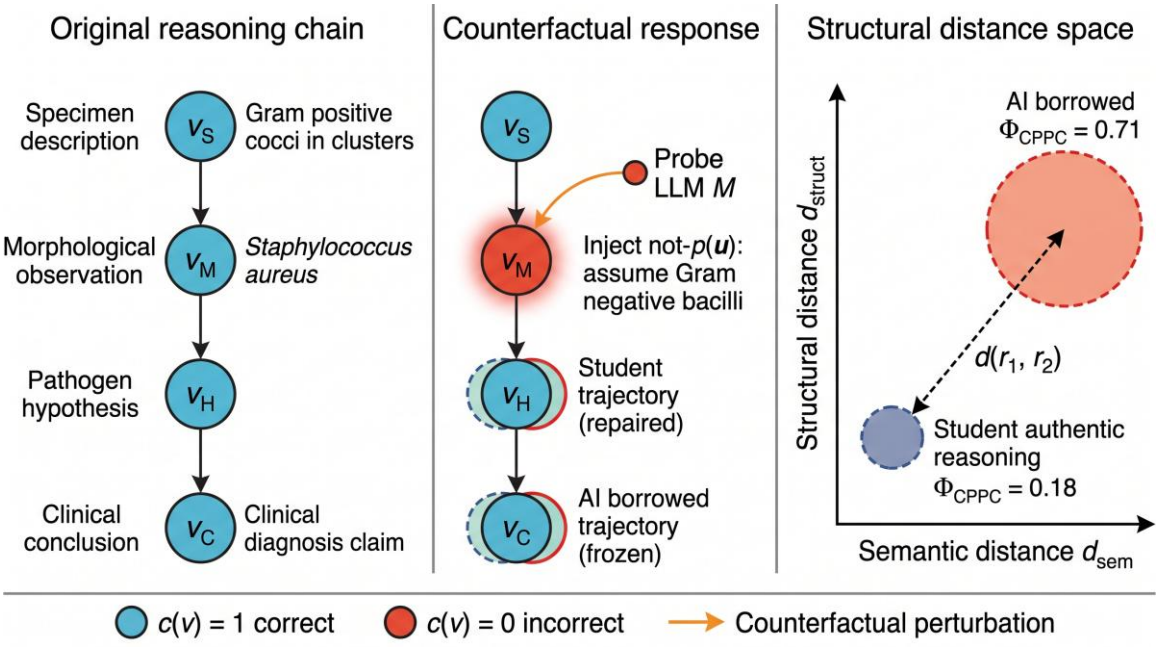


Figure 3 Schematic illustration of the CPPC counterfactual premise perturbation mechanism

The complete computation procedure of the CPPC channel is presented in Algorithm 1.

Algorithm 1 CPPC Counterfactual Response Generation and Distance Computation

Input: student report R, dependency graph $G=(V,E)$, probing model M, domain BERT embedder ϕ_{BERT}
Output: CPPC score for each node $\Phi=\{\phi_{\mathrm{CPPC}}(v)\}_{v\in V}$
1: for each $v\in V$ do
2: Extract the triplet $T(v)=(v,\mathrm{claim}(v),\mathrm{evidence}(v))$
3: Initialize accumulator $s\leftarrow 0$

4: for each $u \in \mathrm{Pa}(v)$ do
5: Generate the counterfactual premise according to $\mathrm{prompt}_{\mathrm{cf}}$: $\neg p(u) \leftarrow M(p(u), \mathrm{prompt}_{\mathrm{cf}})$
6: Invoke M to generate the counterfactual response $r \leftarrow M(\neg p(u), \mathrm{evidence}(v))$
7: Compute the semantic distance $d_{\mathrm{sem}} \leftarrow 1-\cos(\phi_{\mathrm{BERT}}(r), \phi_{\mathrm{BERT}}(r_{\mathrm{obs}}(v)))$
8: Compute the structural distance $d_{\mathrm{struct}} \leftarrow \mathrm{GraphKernel}(r, r_{\mathrm{obs}}(v))$
9: $s \leftarrow s+\lambda_1 d_{\mathrm{sem}}+\lambda_2 d_{\mathrm{struct}}$
10: end for
11: $\phi_{\mathrm{CPPC}}(v) \leftarrow s / \max(\lvert \mathrm{Pa}(v) \rvert, 1)$
12: end for
13: return Φ.

Identifiable Path Signature Based on the Dependency Partial Order Violation Rate

POVR-Sig relies on a more rigid structural signature. A classical conclusion long established in knowledge space theory states that an authentic cognitive state must be downward-closed with respect to the competence partial order. In other words, genuinely understanding a downstream node necessarily entails understanding all of its prerequisite nodes. This downward closure no longer holds in the presence of generative AI. An autoregressive model can make a downstream node appear fluent and its conclusion apparently correct while the corresponding upstream nodes are entirely incorrect at the factual level. Existing work in knowledge tracing and cognitive diagnosis generally models mastery at the item-level [28, 29] and has not yet used node-level partial order relations as a principled criterion for attribution identification. POVR-Sig extracts this structural misalignment as a computable path signature, exposing for the first time at a quantifiable level a problem that static assessment previously concealed.

Given a node v and a lightweight correctness classifier $c: V \rightarrow \{0,1\}$, the criterion $c(v)=1$ indicates that the statement at the node is factually correct, whereas $c(v)=0$ indicates that it is factually incorrect. In this study, c was implemented by fine-tuning a Chinese medical BERT on 4,200 nodes with factual-correctness labels annotated by the panel of teachers independently of the attribution labels. On a held-out validation set of 1,200 nodes, the classifier achieved an accuracy of 0.918 and an F1 of 0.908, which is adequate for the subsequent partial order violation computation. On a dependency edge $(u,v) \in E$, a partial order violation is defined as a case in which a downstream node appears correct while an upstream node is actually incorrect. The global POVR is

$$\Phi_{POVR}=\frac{1}{|E|}\sum_{(u,v)\in E} 1\,[c(v)=1 \wedge c(u)=0] \quad (8)$$

where $1[\cdot]$ is the indicator function. The node-level POVR considers only the set of incoming edges terminating at v

$$\Phi_{POVR}^{local}(v)=\frac{1}{\max(|Pa(v)|,1)}\sum_{u\in Pa(v)} 1\,[c(v)=1 \wedge c(u)=0] \quad (9)$$

The guarantee provided by downward closure theory can be written more rigorously as follows. If $K \subseteq V$ is the subset representing the learner's authentic cognitive state, then under the dependency partial order $\preceq$ it must hold that

$$v \in K \wedge u \preceq v \Rightarrow u \in K \quad (10)$$

An authentic reasoning path lies within the downward-closed lattice, and $\Phi_{\mathrm{POVR}}^{\mathrm{local}}$ approaches 0. A report involving AI borrowing violates downward closure with nonzero probability, and $\Phi_{\mathrm{POVR}}^{\mathrm{local}}$ becomes significantly greater than 0. POVR-Sig and CPPC possess a natural complementary coupling. If AI forces every node to be correct in order to evade POVR, it loses the capacity to respond to counterfactual premises and is exposed by CPPC. If AI permits errors in upstream nodes to preserve a natural reasoning style, POVR immediately produces an alert. In principle, the two signals cannot be evaded simultaneously by the same generation strategy. This property is central to the ability of EduAttr to resist multiple borrowing strategies.

The node type semantics used by POVR-Sig are presented in Table 1. Nodes in the dependency graph are divided into four categories, and errors in each category are assigned different weights according to their downstream effects. The same categorization is reused during instructional early warning ranking.

Table 1 Node Types and Semantic Definitions in the Pathogen Reasoning Dependency Graph

Node type	Notation	Semantic description	Typical position
Specimen description node	V_S	Description of specimen source, sample type, and collection conditions	Beginning of the report
Morphological observation node	V_M	Record of microscopic morphology, staining reactions, and culture characteristics	Early to middle section
Mechanistic hypothesis node	V_H	Inference regarding pathogen identity hypotheses and pathogenic mechanisms	Middle to late section
Clinical conclusion node	V_C	Diagnostic conclusions and recommendations regarding treatment direction	End of the report

Node type information is incorporated into the hierarchically weighted version of POVR-Sig. Figure 4 visualizes the partial order lattice and violation detection.

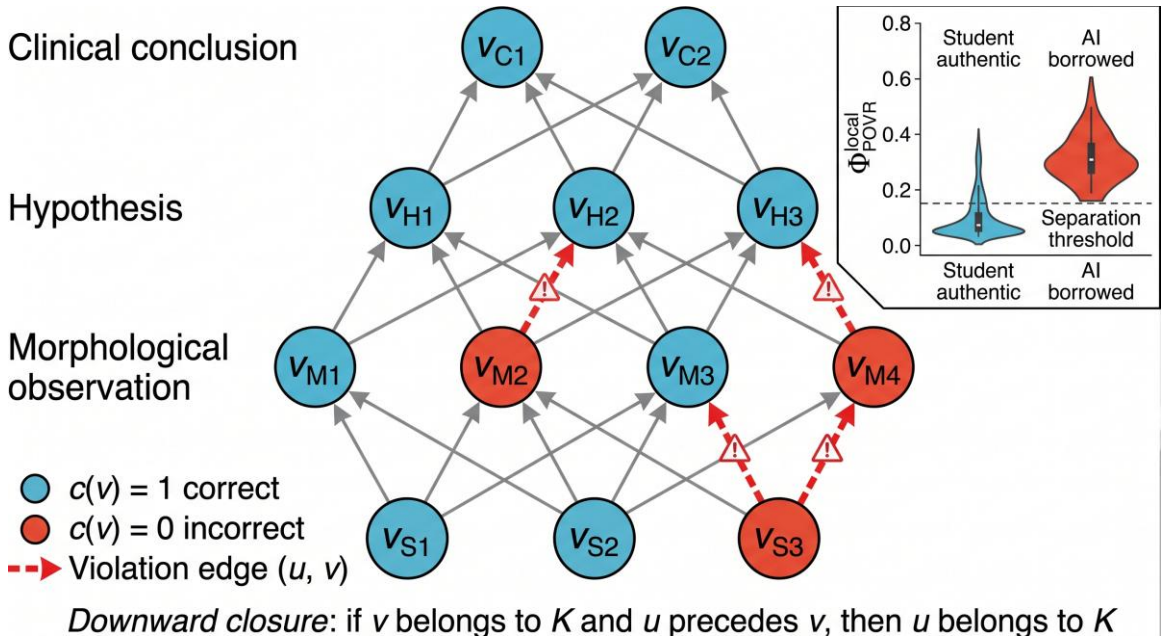


Figure 4 Schematic illustration of the dependency partial order lattice and POVR violation detection

The detailed computation procedures for the discriminator and the partial order violation rate are presented in Algorithm 2.

Algorithm 2 POVR-Sig Partial Order Violation Rate Computation

Input: dependency graph $G=(V,E)$, correctness classifier c
Output: global Φ_{POVR} and node-level $\{\Phi^{\mathrm{local}}_{\mathrm{POVR}}(v)\}_{v\in V}$
1: Initialize the global violation count $n_g \leftarrow 0$; violation edge set $E_{\mathrm{vio}} \leftarrow \varnothing$
2: for each $(u,v)\in E$ do
3: if $c(v)=1$ and $c(u)=0$ then
4: $E_{\mathrm{vio}} \leftarrow E_{\mathrm{vio}} \cup \{(u,v)\}$; $n_g \leftarrow n_g+1$
5: end if
6: end for
7: $\Phi_{\mathrm{POVR}} \leftarrow n_g / \max(\lvert E\rvert,1)$
8: for each $v\in V$ do
9: Count the violating incoming edges terminating at v: $n_v \leftarrow \lvert\{u\in \mathrm{Pa}(v):(u,v)\in E_{\mathrm{vio}}\}\rvert$
10: $\Phi^{\mathrm{local}}_{\mathrm{POVR}}(v) \leftarrow n_v / \max(\lvert \mathrm{Pa}(v)\rvert,1)$
11: end for
12: return Φ_{POVR}, $\{\Phi^{\mathrm{local}}_{\mathrm{POVR}}(v)\}$, E_{vio}.

Attribution Signal Fusion and Instructional Early Warning Node Ranking

The preceding two channels provide node-level scores from the perspectives of intervention semantics and partial order structure, respectively. The task of TANR is to translate these two heterogeneous signals into follow-up questioning actions that teachers can execute directly. Simply adding the two scores would overlook a critical instructional fact. The closer a node is to the clinical conclusion in the dependency chain, the greater

the potential instructional harm caused by borrowing at that node and the higher the marginal value of teacher follow-up questioning. TANR introduces depth weights to explicitly model this hierarchical difference, thereby upgrading attribution identification from a scoring module to an instructional action generator.

The depth of node v in the dependency graph is defined as

$$\mathrm{Depth}(v) = \max_{P \in \mathrm{Path}(v_0 \to v)} |P| \quad (11)$$

where v_0 is the root node at the specimen-description level, $\mathrm{Path}(v_0 \to v)$ is the set of all directed paths from v_0 to v, and |P| is the path length. To prevent the depth weight from being dominated by extremely long paths, sigmoid normalization is introduced

$$w_{\mathrm{depth}}(v) = \frac{1}{1 + \exp\left(-(\mathrm{Depth}(v) - \mu_D)\right)} \quad (12)$$

where μ_D is the mean depth of the dependency graph for the sample. It serves as a baseline calibration, ensuring that the depth weight is distributed between 0 and 1 and is insensitive to graph size. The attribution fusion score is defined as Before fusion, the two channel signals were z-standardized using the training-set mean and standard deviation so that the fusion coefficients operate on comparable scales.

$$\mathrm{AttrScore}(v) = \alpha \cdot \phi_{\mathrm{CPPC}}(v) + \beta \cdot \Phi^{\mathrm{local}}_{\mathrm{POVR}}(v) + \gamma \cdot w_{\mathrm{depth}}(v) \quad (13)$$

where α, β, and γ have default values of 0.4, 0.4, and 0.2, respectively, and can be adjusted by teachers according to disciplinary priorities in actual teaching. The instructional early warning list is obtained through Top-K ranking

$$\mathrm{T}_K = \text{arg-topK}_{(s,v) \in \mathrm{S} \times V(G_s)} \mathrm{AttrScore}_s(v) \quad (14)$$

where S is the set of students in the class and G_s is the dependency graph of student s. The output list is arranged as "student, node" pairs, with each entry corresponding to a specific focal point for classroom follow-up questioning. This output granularity is unavailable in existing AIGC detectors and neural cognitive diagnosis systems, and it constitutes the key interface through which EduAttr moves from identification to instructional action. Figure 5 presents a visual interface for the TANR ranking output.

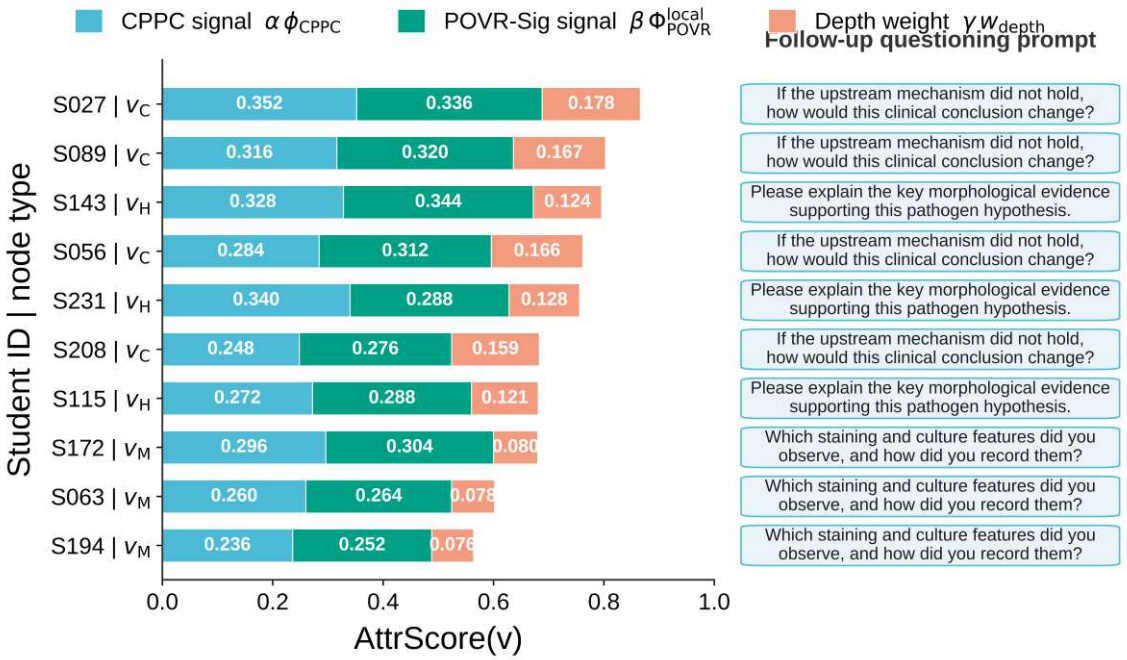


Figure 5 Schematic illustration of the TANR instructional early warning node ranking output

The TANR ranking and list generation procedure is presented in Algorithm 3.

Algorithm 3 TANR Attribution Fusion and Top-K Instructional Early Warning Ranking

Input: class student set S, individual dependency graph G_s, CPPC scores $\{\phi_{\mathrm{CPPC}}\}$, POVR scores $\{\Phi^{\mathrm{local}}_{\mathrm{POVR}}\}$
Output: instructional early warning list T_K and questioning prompt field
1: Initialize candidate set $\mathrm{C} \leftarrow \varnothing$
2: for each $s \in \mathrm{S}$ do
3: Compute $\mu_D \leftarrow \mathrm{mean}_{v \in V(G_s)} \mathrm{Depth}(v)$
4: for each $v \in V(G_s)$ do
5: Compute $\mathrm{Depth}(v)$ and $w_{\mathrm{depth}}(v)$
6: $\mathrm{AttrScore}(v) \leftarrow \alpha\phi_{\mathrm{CPPC}}(v) + \beta\Phi^{\mathrm{local}}_{\mathrm{POVR}}(v) + \gamma w_{\mathrm{depth}}(v)$
7: $\mathrm{C} \leftarrow \mathrm{C} \cup \{(s, v, \mathrm{AttrScore}(v), \mathrm{type}(v))\}$
8: end for
9: end for
10: Sort C in descending order by $\mathrm{AttrScore}$
11: $\mathrm{T}_K \leftarrow \mathrm{C}[1:K]$
12: for each $(s, v, \cdot, \tau) \in \mathrm{T}_K$ do
13: Match the questioning prompt template according to node type $\tau \in \{V_S, V_M, V_H, V_C\}$
14: end for
15: return T_K

System Integration and Embedding in the Teacher Workflow

At the system level, the three innovative components are assembled into a complete data flow. Student reports are converted into dependency graphs by the domain dependency parser π. CPPC and POVR-Sig are computed in parallel through the two channels, and TANR performs node-level fusion and class-level ranking. The resulting follow-up question list is then presented to the instructor through a workbench. Neural cognitive diagnosis systems stop at "outputting a score for a knowledge component" [30], while existing educational early warning tools stop at "marking a suspicious passage." Neither addresses the instructional-action-level gap of determining which specific node should be questioned first. Through the dual outputs of node-level attribution and a ranked list, EduAttr simplifies preclass grading from reading all reports in full to focusing on Top-K nodes and transforms follow-up questioning from experience-based judgment into data-driven action.

From the perspective of data flow, follow-up questioning outcomes and students' subsequent responses return as supervisory signals for the next iteration. For each question in the Top-K list asked by the teacher in class, if the student's in-class response is logically valid, the attribution score of the node is calibrated downward. If the response still shows traces of borrowing, the sample is added to the domain retraining set for the probing model M and the classifier c. This closed-loop process makes EduAttr more than a one-time identification system. It becomes a continuously evolving mechanism embedded in the classroom follow-up questioning workflow, consistent with the human-AI collaborative assessment pathway repeatedly emphasized in recent medical education informatization research [3].

Figure 6 presents the overall system data flow and its embedding in the teacher workflow. The key hyperparameters and their default values are summarized in Table 2 to facilitate replication and comparison in the subsequent experimental chapter.

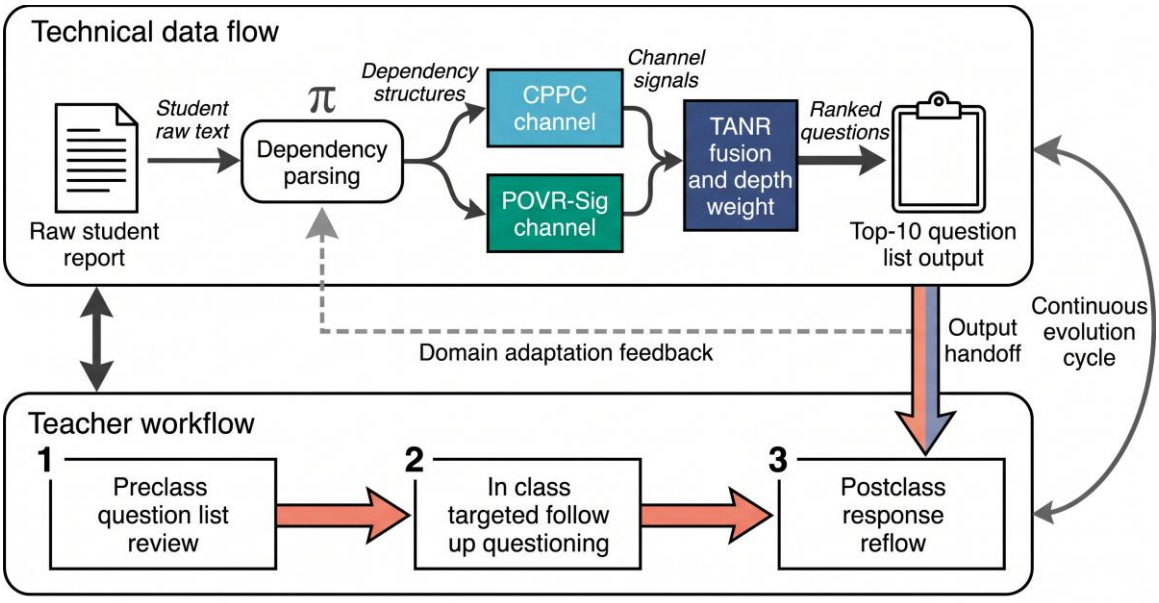


Figure 6 Schematic illustration of the overall EduAttr system data flow and its embedding in the teacher workflow

Table 2 Default Values and Meanings of Key EduAttr Hyperparameters

Parameter	Default value	Module	Semantic description
λ_1, λ_2	0.4, 0.6	CPPC distance measurement	Fusion weights for semantic distance and structural distance
α, β, γ	0.4, 0.4, 0.2	TANR attribution fusion	Fusion coefficients for CPPC, POVR, and depth weights
K	10	Top-K ranking	Number of questioning entries output for each class session
μ_D	Data adaptive	Depth normalization	Mean dependency graph depth baseline
$\text{prompt}_{\text{cf}}$	Domain template library	Counterfactual premise generation	Template guiding M to generate counterfactual premises

At this point, the three methodological innovations and the system integration approach have been fully presented. CPPC departs in principle from the detection paradigm based on surface features. POVR-Sig provides an identifiability guarantee through the downward closure property of knowledge space theory. TANR completes the translation from identification to questioning at the instructional-action level. Together, the three components form a complete loop from text to structure and then to instructional action. Section 4 quantitatively validates the identification performance, robustness, and teacher adoption of this methodological system on data from a two-arm randomized controlled trial.

Experiments and Results

Trial Design, Dataset Construction, and Real Classroom Annotation

This study was designed as a two-arm, parallel-group randomized controlled trial with a 1:1 allocation ratio and is reported in accordance with the CONSORT guidelines [31]; the completed CONSORT checklist is provided as Additional file 1. The trial was retrospectively registered in the Chinese Clinical Trial Registry on 2 October 2026 (ChiCTR2600118749). The study protocol was approved by the ethics committee of the leading institution before recruitment and is available from the corresponding author on reasonable request.

The dataset was collected from second-year students majoring in clinical medicine at two medical colleges in China and covers the autumn and winter semesters of the 2024 to 2025 academic year. Recruitment started on September 10, 2024 and the last report was submitted on June 20, 2025. Students were eligible if they were enrolled in the medical microbiology course in one of the participating classes, were at least 18 years old, and provided written informed consent; no further exclusion criteria were applied. The study protocol was approved by the ethics committee of the leading institution, and all participants signed informed consent forms. Enrolled students were randomly assigned to control arm A and experimental arm B. An investigator not involved in teaching, grading, or annotation generated the allocation sequence using a computer-generated random number list stratified by institution with four classes allocated per institution, and the allocation was concealed from course teachers and annotators until all classes had been enrolled and the baseline pretest had been completed. Each arm contained 4 intact classes with 120 students, yielding a total of 240 participants. Because the eight participating

classes constituted the full available pool of eligible intact classes at the two institutions during the study period, the sample size was determined pragmatically; a post hoc precision analysis confirmed that the achieved sample provided a 95% confidence interval half-width of 0.014 for the primary outcome, which exceeded the pre-specified target precision of 0.03. Students in the control arm prepared their reports in supervised classroom sessions in which access to external networks was disabled, so that no generative AI tool could be used. Students in the experimental arm were permitted and encouraged to use large language models such as GPT-4o and ChatGPT to assist their reasoning, and were required to export and submit the complete interaction transcript together with each report. These transcripts provide the objective evidence from which node-level attribution labels were derived, which is what distinguishes the present annotation setting from the unaided teacher judgment condition reported in [18]. The assignments focused on core chapters of medical microbiology. The five pathogen cases were Staphylococcus aureus, Mycobacterium tuberculosis, Salmonella Typhi, Mycoplasma pneumoniae, and Treponema pallidum. They covered three major morphological categories, namely cocci, bacilli, and spirochetes, as well as three metabolic ecological types, namely aerobic, anaerobic, and intracellular parasitic organisms. Apart from the writing condition, students in both arms completed the same five case assignments using the same four-part reporting structure of specimen description, morphological observation, mechanistic hypothesis, and clinical conclusion, received the same teaching materials, and were graded by instructors who were not members of the annotation panel. Each student completed 5 reports, producing 1,200 original texts in total.

Dependency parsing was performed using a Chinese spaCy pipeline fine-tuned on a specialized medical microbiology corpus. On a reserved sample of 100 reports annotated independently by two microbiology instructors, the parser achieved a node extraction F1 of 0.903 and a dependency edge F1 of 0.878. Each report initially yielded an average of 19.4 reasoning nodes, resulting in 23,246 candidate nodes across the entire dataset. Node-level attribution annotation was completed by 5 teachers with associate professor rank and more than 8 years of teaching experience. Each node was first annotated independently by 3 teachers who read only the student report. Because of the nature of the intervention, neither the students nor their course teachers could be blinded to allocation, whereas annotators in this first stage worked from anonymized reports without access to the interaction transcripts or to arm labels, and the analyst who computed the final evaluation metrics was blinded to study arm during method development. Agreement at this stage was quantified as the mean of the pairwise Cohen's kappa values across the three rater pairs and reached 0.741, which falls within the range of substantial agreement. The interaction transcripts of the experimental arm were then released to the panel, and the final gold label of every node was fixed by aligning the assertion and the evidence citation carried by that node against the transcript, with residual disputes resolved through collective consultation. Control arm nodes were annotated under the same protocol and no borrowing was identified in that arm. After removing 3.8% of nodes that could not be parsed or whose conflicting labels did not reach consensus, 22,362 valid nodes remained, with an average of 18.6 analyzable nodes per report. Every experimental arm report containing at least one borrowed node, 517 reports in total, was further assigned a dominant borrowing strategy label from the five categories listed in

Table 9. The dataset was divided into training, validation, and test sets at a ratio of 70/15/15. Stratification was performed by student rather than by report to prevent writing style features from the same student leaking across sets. Baseline characteristics of the participants and the principal statistics of the dataset are presented in Table 3.

Table 3 Baseline Characteristics of Participants and Core Statistics of the EduAttr Dataset

Item	Control arm A	Experimental arm B	Total
Number of students	120	120	240
Number of classes	4	4	8
Age (years), mean (SD)	19.4 (0.7)	19.5 (0.6)	19.4 (0.7)
Female, n (%)	68 (56.7)	71 (59.2)	139 (57.9)
Microbiology pretest (0 to 100), mean (SD)	61.2 (8.4)	60.5 (8.7)	60.9 (8.5)
Prior anatomy and physiology grade (0 to 100), mean (SD)	80.4 (5.8)	79.9 (6.1)	80.1 (5.9)
Number of reports	600	600	1,200
Number of valid nodes	11,528	10,834	22,362
Mean nodes per report	19.2	18.1	18.6
Proportion of AI borrowing nodes	0	34.7%	16.8%
Mean dependency depth	3.42	3.36	3.39
Mean three-rater kappa	0.752	0.729	0.741

All 240 allocated students completed the five assignments and all 1,200 reports were included in the analysis, so no participant was lost to follow-up after allocation. No serious adverse events were observed; in the first week of classroom deployment two students in the experimental arm expressed mild discomfort during follow-up questioning, and no further concerns were raised after a protocol clarification emphasizing the formative, non-disciplinary purpose of the questioning. The flow of participants through the trial and the construction of the analysis dataset are shown in Figure 7.

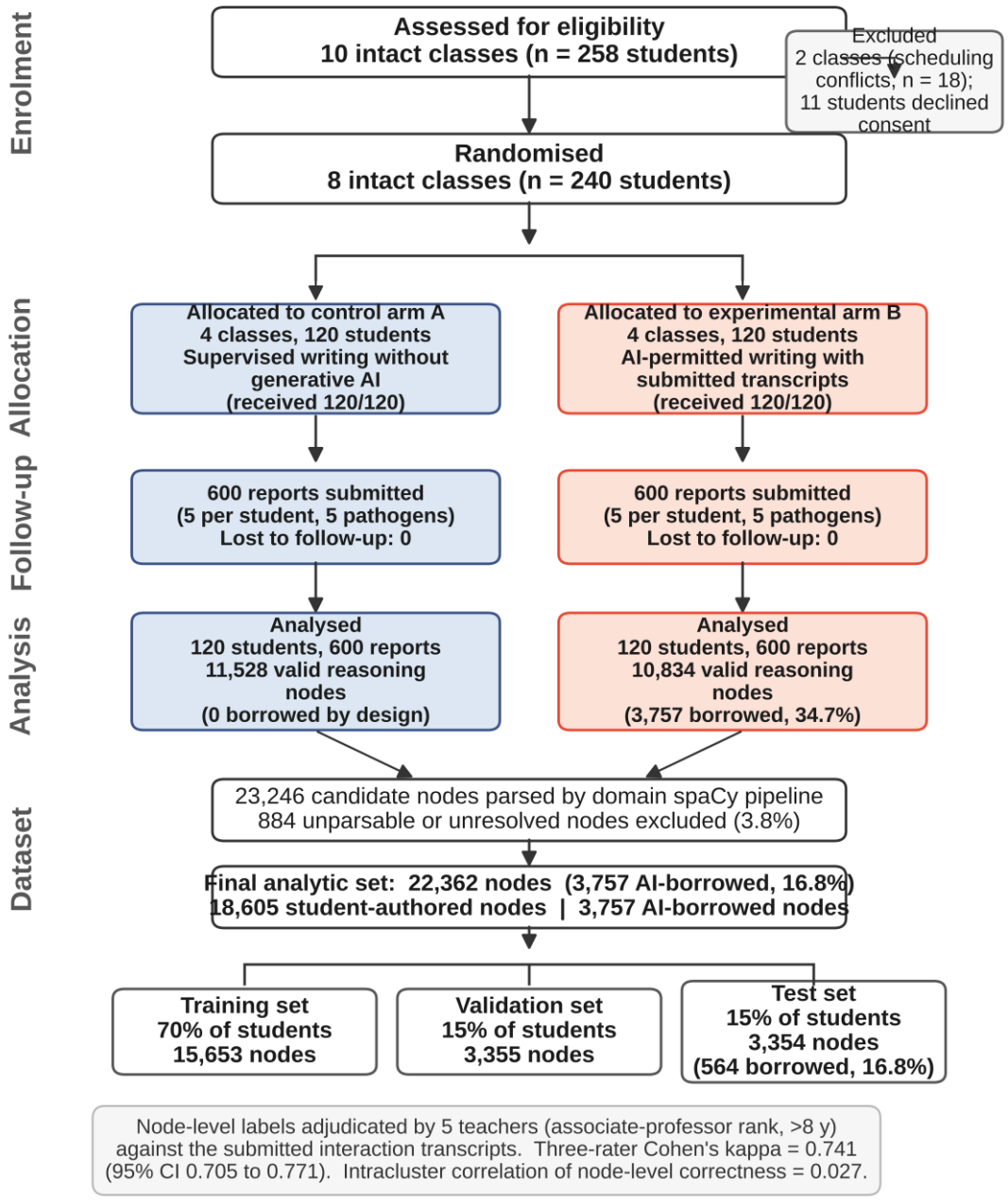


Figure 7 CONSORT flow diagram of participants and construction of the EduAttr dataset. The diagram reports, in successive stages, assessment for eligibility, allocation of eight intact classes to two arms with 120 students per arm, completion of five assignments by all 240 students, parsing of 23,246 candidate nodes, exclusion of 884 unparsable or unresolved nodes (3.8%), formation of the final set of 22,362 analyzable nodes (11,528 in the control arm and 10,834 in the experimental arm), adjudication of node-level labels against the interaction transcripts for the experimental arm (3,757 borrowed nodes, 16.8% overall prevalence), and the student-level 70/15/15 split into training, validation, and test sets

Baseline Methods, Evaluation Metrics, and Experimental Settings

Seven representative baselines were selected to cover four mainstream pathways. The textual surface feature pathway includes GLTR [32], DetectGPT [20], and Ghostbuster [21], all of which use criteria based on word probability distributions or perplexity. The supervised classification pathway is represented by a RoBERTa detector [33] fine-tuned on the training set. The learner modeling pathway includes NeuralCD-adapted [30] and KTM-adapted [34], which are adapted by reformulating node attribution as a knowledge concept mastery prediction task. The reasoning consistency pathway is represented by a chain-of-thought self-consistency measure that examines the logical closure within each node. Detailed baseline configurations are presented in Table 4.

Table 4 Key Configurations of the Baseline Methods

Method	Category	Key component	Training mode
GLTR	Surface features	Word probability quantile histogram	Zero-shot
DetectGPT	Surface features	Chinese GPT-2 medium curvature score + 100 perturbations	Zero-shot
Ghostbuster	Surface features	Chinese weak language model ensemble + structured feature search	Feature selection + LR
RoBERTa-det	Supervised classification	RoBERTa-large-zh dual classification head	Supervised fine-tuning
NeuralCD-adp	Learner modeling	Node embedding + interaction function network	Supervised fine-tuning
KTM-adapted	Learner modeling	Factorization machine + factor interaction	Supervised fine-tuning
CoT-consist	Reasoning consistency	Sampling of 20 reasoning chains + majority voting	Zero-shot

The primary outcome was the node-level F_1 score on the held-out test set, and the remaining metrics described below served as secondary outcomes. The evaluation metrics cover three levels. Node-level metrics include precision P, recall R, F_1, and AUROC. Report-level performance is evaluated using F_1. Follow-up question node localization is measured using Precision@10 and MRR. The former measures the proportion of genuine AI borrowing nodes in the Top-10 list, while the latter measures the reciprocal rank of the first correctly identified follow-up question node. Teacher alignment is quantified using Cohen's kappa between system predictions and the adjudicated gold labels, with the interteacher agreement of the annotation panel serving as the reference band. The three practicality indicators were collected during actual classroom deployment in the experimental arm, covering 20 sessions, that is, 4 classes across 5 pathogen cases, each session receiving one Top-10 list and giving 200 entries in total. Deployment was confined to the experimental arm because the follow-up questioning success indicator is defined only where borrowing is present, and the control arm contains no borrowed node. Because attribution performance can only be estimated where both borrowed and non-borrowed nodes are present, the attribution outcomes were estimated on the pooled test set, to which the control arm contributed only non-borrowed nodes and the experimental arm contributed both borrowed and non-borrowed nodes. All experiments were implemented in PyTorch 2.2 on four NVIDIA A100 GPUs. The probing large language model M in EduAttr was GPT-4o (version gpt-4o-2024-08-06, accessed via the OpenAI API between October 2024 and June 2025 at temperature 0.0 for deterministic probing; a manual audit of 100 randomly sampled counterfactual premises showed that 94 were

judged by two microbiology instructors to be structurally valid and semantically opposite to the original premise, and the counterfactual premise template is provided in Additional file 3), the domain embedder was a Chinese biomedical BERT further pretrained on a medical microbiology corpus, and the node-level classifier c was obtained by fine-tuning a Chinese medical BERT model for binary classification. All reported results are the means of 5 independent runs together with 95% confidence intervals obtained from 1,000 bootstrap resamples of the test set. Bootstrap resamples were drawn at the student level to reflect the clustering of nodes within reports and reports within students, so that confidence intervals and paired tests account for the nested data structure. Pairwise comparisons between EduAttr and each baseline on the primary outcome used paired bootstrap tests on the same resamples, and p values were adjusted using the Holm procedure. The intraclass correlation coefficient of node-level correctness at the class level was 0.027, consistent with a limited influence of class-level clustering on the primary outcome. Analyses were performed in Python 3.10 with scikit-learn 1.3 and SciPy 1.11. Comparisons between EduAttr and each baseline were assessed with paired bootstrap tests, and the resulting p values were adjusted for the seven comparisons using the Holm procedure. The complete implementation of EduAttr, covering the counterfactual probing module, the partial order violation signature, the node ranking mechanism, and all evaluation scripts, is openly available at https://github.com/a18907052542-lang/EduAttr.git.

Attribution Identification Performance and Accuracy of Follow-Up Question Node Localization

The main experimental results are presented in Table 5. EduAttr achieves a node-level F_1 of 0.843, representing improvements of 12.1 percentage points over the strongest supervised baseline, RoBERTa-det, and 13.0 percentage points over the strongest surface feature baseline, Ghostbuster. Both differences remain significant after Holm adjustment according to paired bootstrap tests. EduAttr achieves AUROC and Doc-F_1 values of 0.979 and 0.874, respectively, exceeding the strongest baselines by 4.1 and 9.1 percentage points. The AUROC margin is compressed relative to the margin in node-level F1 because all methods operate at a low false positive rate on a test set in which only 16.8% of nodes are borrowed, which is the regime where ranking metrics separate methods more sharply than threshold-free area measures. The area under the precision-recall curve, which is more informative under the 16.8% positive class prevalence, was 0.924 for EduAttr, 0.805 for RoBERTa-det, 0.796 for Ghostbuster, 0.752 for CoT-consist, 0.749 for DetectGPT, 0.665 for NeuralCD-adp, 0.664 for KTM-adapted, and 0.648 for GLTR, with an EduAttr margin of 11.9 percentage points over the strongest baseline. Analyses restricted to the experimental arm showed that the node-level F1 of EduAttr was 0.828 (95% CI 0.812 to 0.843), 14.5 percentage points higher than that of RoBERTa-det (0.683, 95% CI 0.661 to 0.705) and 15.7 points higher than that of Ghostbuster (0.671, 95% CI 0.650 to 0.692); the within-arm margin was larger than the overall between-method gap, indicating that the identification advantage derives from the dependency structure of reasoning rather than from differences in writing conditions between arms. The method flagged 49 of 1,729 control-arm test nodes (2.8%) as borrowed, consistent with the overall false positive rate, and no control-arm student

appeared more than once in any Top-10 list, limiting the practical risk of false alerts in the arm with no true borrowing.

Table 5 Main Results for Attribution Identification (Test Set, Mean ± Half Width of the 95% Confidence Interval)

Method	Node-P	Node-R	Node-F_1	AUROC	Doc-F_1	P@10	MRR
GLTR	0.612±0.023	0.583±0.026	0.597±0.021	0.878±0.014	0.678±0.024	0.421±0.031	0.487
DetectGPT	0.681±0.020	0.653±0.024	0.667±0.019	0.914±0.011	0.744±0.021	0.487±0.028	0.559
Ghostbuster	0.724±0.017	0.702±0.020	0.713±0.015	0.935±0.009	0.783±0.017	0.534±0.025	0.612
RoBERTa-det	0.716±0.018	0.729±0.019	0.722±0.016	0.938±0.009	0.779±0.018	0.519±0.026	0.601
NeuralCD-adp	0.593±0.025	0.641±0.023	0.616±0.022	0.886±0.013	0.649±0.025	0.446±0.029	0.508
KTM-adapted	0.611±0.024	0.628±0.022	0.619±0.021	0.889±0.013	0.657±0.023	0.462±0.028	0.516
CoT-consist	0.702±0.019	0.671±0.022	0.686±0.017	0.923±0.010	0.755±0.020	0.502±0.027	0.581
EduAttr	0.856±0.014	0.831±0.016	0.843±0.013	0.979±0.005	0.874±0.014	0.847±0.019	0.892

Follow-up question node localization is the dimension in which the advantage of the proposed method is most pronounced. EduAttr achieves a Precision@10 of 0.847, indicating that, on average, 8.47 of the nodes in the output Top-10 follow-up question list are genuine AI borrowing points according to the adjudicated labels. This P@10 value is 31.3 percentage points higher than that of the strongest baseline, Ghostbuster. The MRR reaches 0.892, indicating that the top-ranked follow-up question node output by the system is the most valuable focal point in the vast majority of samples. This gap confirms the principled advantage of the dual-channel signals proposed in Section 3 for node-level ranking.

Figures 8, 9, and 10 compare the methods from different perspectives. The operating points marked on Figures 8 and 9 correspond to the precision and recall values reported in Table 5.

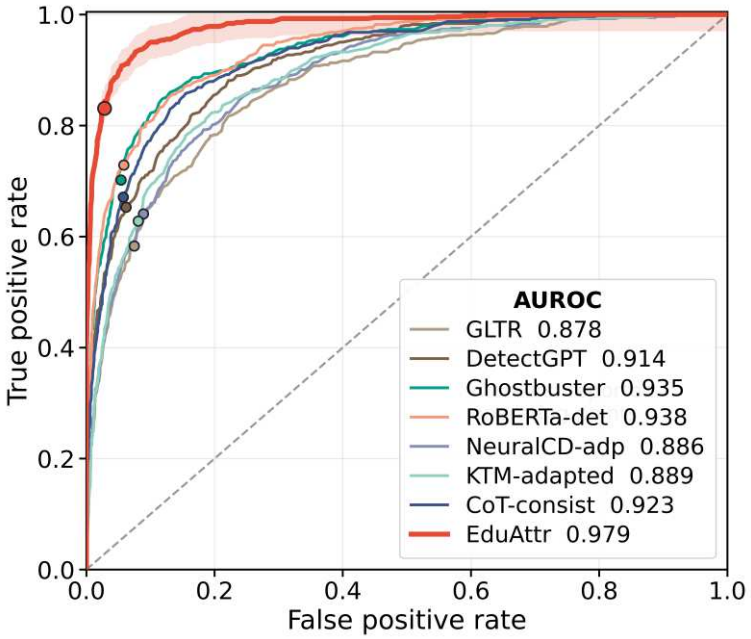


Figure 8 Node level ROC curves of the eight methods

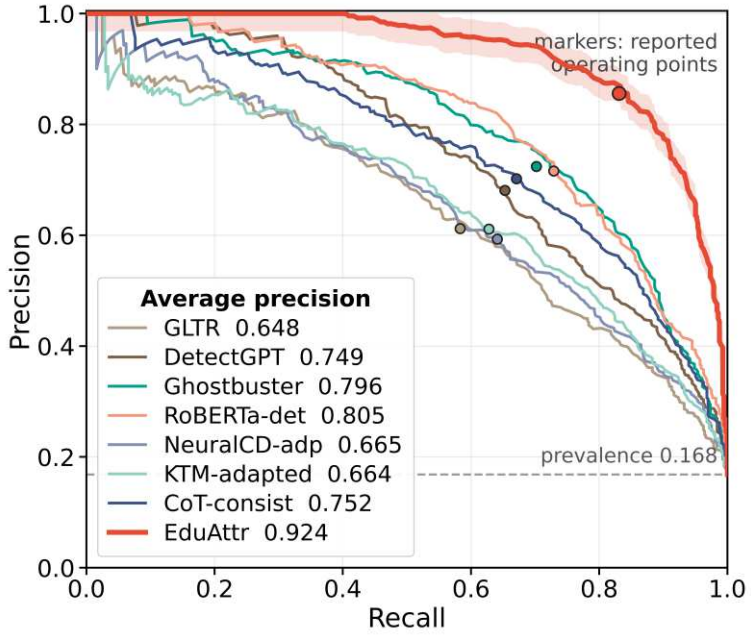


Figure 9 Node level Precision-Recall curves of the eight methods

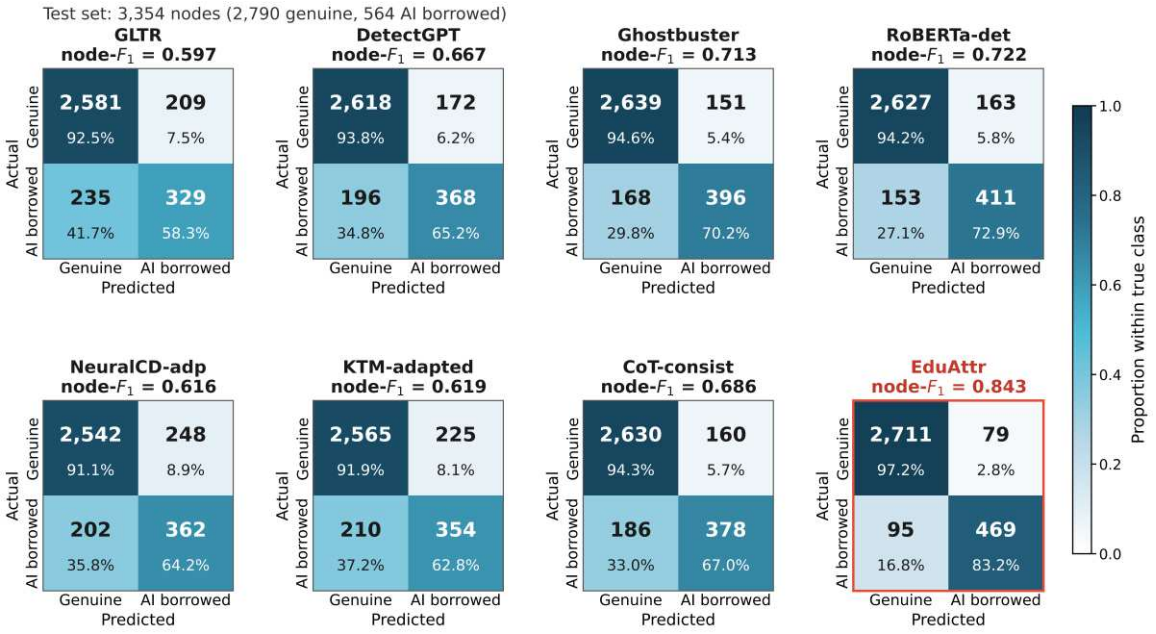


Figure 10 Node level confusion matrices of the eight methods on the test set

Ablation Analysis, Robustness Testing, and Alignment with Teacher Judgments

The ablation experiments isolate the independent contributions of CPPC, POVR-Sig, and the depth weight. The results are presented in Table 6. Removing CPPC decreases F_1 by 7.2 percentage points, while removing POVR-Sig decreases F_1 by 8.7 percentage points. The magnitudes are comparable, and both declines are substantially greater than the 1.6 percentage point decrease caused by removing the depth weight. This structural evidence supports the claim made in Section 3 that each channel contributes a nonredundant criterion dimension. Because the probing model and the model used by students in the experimental arm belong to the same family, the counterfactual channel was also rerun with two probing models from different families. Node-level F1 moved to 0.831 and 0.828, so the advantage of the method does not depend on matching the generator. When CPPC or POVR-Sig is used alone, F_1 decreases to 0.712 and 0.694, respectively, both substantially below the 0.843 achieved by the complete method. This result confirms the complementary coupling between the two signals rather than a simple redundant addition. The two variants that omit the depth weight differ in how the remaining channels are combined. In the w/o depth weight variant the two channel scores are still standardized separately and their fusion coefficients are retuned on the validation set, giving 0.827, whereas the simple sum variant adds the two raw channel scores with equal coefficients and no per-channel calibration, giving 0.798. The 2.9 percentage point interval between them therefore isolates the contribution of score calibration, and the further 1.6 percentage point interval up to the complete method isolates the contribution of the depth weight itself. As shown in Figure 11, the complete method dominates every variant on all six dimensions, and the spread across variants is widest on P@10 and MRR, the two dimensions that determine the quality of the ranked questioning list. The shapes therefore provide an intuitive representation of the complementary roles of the components.

Table 6 Ablation Experiment Results

Variant	Node-F_1	ΔF_1	AUROC	P@10	Doc-F1	MRR	Robust-F1
EduAttr (full)	0.843	--	0.979	0.847	0.874	0.892	0.723
w/o CPPC	0.771	-0.072	0.957	0.723	0.812	0.784	0.664
w/o POVR-Sig	0.756	-0.087	0.951	0.694	0.801	0.759	0.612
w/o depth weight	0.827	-0.016	0.974	0.796	0.861	0.843	0.708
CPPC only	0.712	-0.131	0.934	0.641	0.762	0.706	0.581
POVR-Sig only	0.694	-0.149	0.926	0.617	0.748	0.688	0.573
Simple sum	0.798	-0.045	0.966	0.752	0.833	0.801	0.679

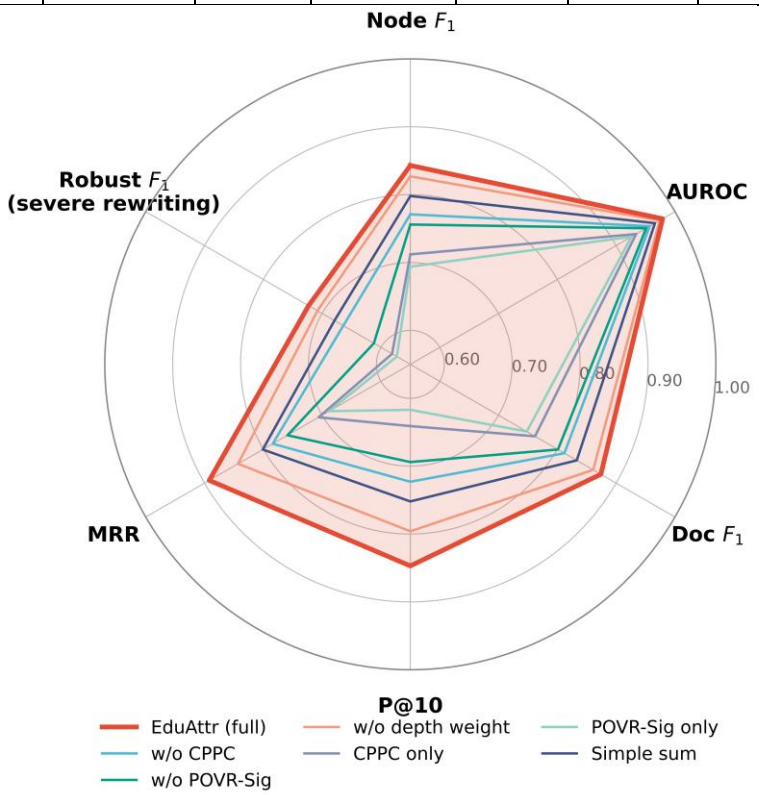


Figure 11 Radar comparison of EduAttr and its ablation variants across six metrics

The robustness test simulates three levels of disguise attacks, corresponding to mild rewriting through synonym substitution, moderate rewriting through sentence reordering and connective replacement, and severe rewriting through paragraph-level paraphrasing and structural reorganization. The results are presented in Table 7. Baselines based on textual surface features and supervised classification both deteriorate severely under severe rewriting. The F_1 values of DetectGPT and RoBERTa-det fall to 0.362 and 0.401, respectively, indicating that their criteria depend heavily on surface statistical features or writing patterns closely associated with the training distribution. Under the same conditions, EduAttr maintains an F_1 of 0.723, a decline of only 12.0 percentage points from the original condition. This robustness arises from the principled design of the two signals. CPPC operates through intervention semantic responses in the reasoning chain, while POVR-Sig operates through the downward closure structure of the dependency partial order. Neither depends on lexical statistics, making both naturally immune to text-level disguise. As shown in Figure 12, the EduAttr row remains on the cool side of the

scale at every attack level. Across the five cases its degradation from the original condition to severe rewriting decreases monotonically as mean dependency depth increases: 14.2 percentage points for Mycoplasma pneumoniae at depth 3.05, 13.7 for Staphylococcus aureus at 3.11, 12.6 for Salmonella Typhi at 3.28, 9.8 for Mycobacterium tuberculosis at 3.74, and 9.7 for Treponema pallidum at 3.77. The baselines show no comparable pattern and in fact perform slightly worse on the two deep chain cases. This pattern reflects a structural amplification effect of the proposed method in reasoning tasks with greater dependency depth.

Table 7 Robustness Test Against Rewriting Attacks (Node-F_1)

Method	Original	Mild rewriting	Moderate rewriting	Severe rewriting
DetectGPT	0.667	0.601	0.489	0.362
Ghostbuster	0.713	0.658	0.552	0.428
RoBERTa-det	0.722	0.649	0.523	0.401
CoT-consist	0.686	0.647	0.578	0.484
EduAttr	0.843	0.812	0.774	0.723

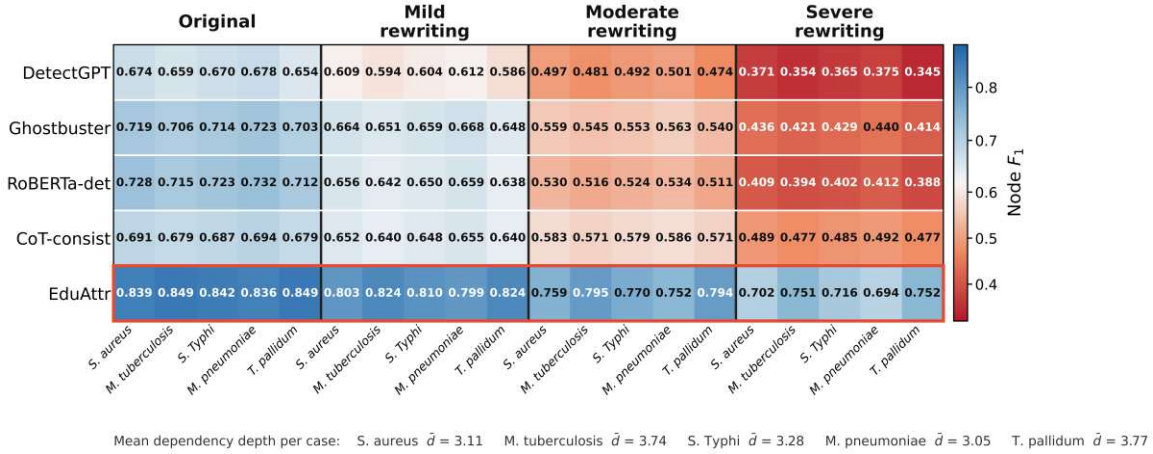


Figure 12 Node level F1 under multilevel rewriting attacks, by method, attack level, and pathogen case

Alignment with teacher judgments was quantified using Cohen's kappa and supplemented with three practical indicators, namely adoption rate, follow-up questioning success rate, and time saved. The results are presented in Table 8. Cohen's kappa between EduAttr predictions and the adjudicated labels reached 0.762 (95% CI 0.731 to 0.793). As a comparable reference, each annotator's independent first-stage labels made without access to the interaction transcripts were separately compared with the adjudicated labels, giving teacher-level kappa values that ranged from 0.698 to 0.768 with a mean of 0.738 (95% CI 0.705 to 0.771). The paired bootstrap difference between EduAttr's agreement and this teacher-level mean was 0.024 (95% CI -0.012 to 0.059, p = 0.19), supporting the use of EduAttr as a decision aid whose agreement with the final labels is at least as high as that of individual teachers reading reports without transcripts; the earlier interteacher value of 0.741 reported in Table 3 reflects the mutual reliability of unaided judgment and is retained for context. The teacher adoption rate examines how many entries in the Top-10 follow-up question list are judged by teachers to be directly executable classroom questions. EduAttr achieves 87.5%, that is, 175 of the 200 entries, substantially exceeding

the 64.0% of the supervised baseline. The classroom follow-up questioning success rate measures how many nodes, when actually used for questioning, cause students to acknowledge borrowing or become unable to respond. Of the 175 adopted entries, 162 were posed within class time, and 127 of them, or 78.4%, produced such a response, which is approximately 1.8 times the value obtained by RoBERTa-det. Teacher grading time decreases from 12.4 minutes to 7.1 minutes per report, a reduction of 42.7%, indicating that the ranked list successfully focuses teacher attention from reading the entire text to questioning key nodes. As shown in Figure 13, the kappa confidence interval of EduAttr contains the level of interteacher agreement and lies within the substantial agreement band. By contrast, the upper bounds of the confidence intervals for the other two methods remain below the interteacher reference line. Only EduAttr therefore reaches agreement with the gold labels that is statistically indistinguishable from the agreement observed among the expert annotators themselves, which supports its use as an additional rater in instructional decision-making rather than as a replacement for teacher judgment.

Table 8 Alignment with Teacher Judgments and Practicality Indicators

Indicator	EduAttr	Ghostbuster	RoBERTa-det	Interteacher agreement
Cohen's kappa	0.762	0.594	0.612	0.741
95% confidence interval	[0.731, 0.793]	[0.552, 0.636]	[0.573, 0.651]	[0.708, 0.774]
Teacher adoption rate	87.5%	61.5%	64.0%	--
Questioning success rate	78.4%	41.7%	44.2%	--
Time saved	42.7%	24.3%	26.1%	--

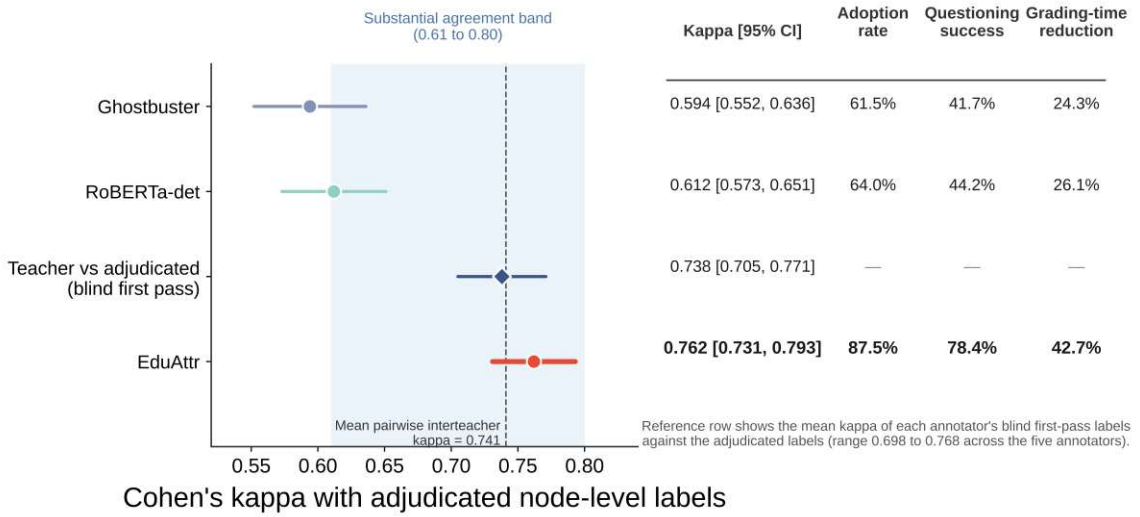


Figure 13 Forest plot of agreement with the adjudicated node level labels

The four sets of results from the main experiment, ablation analysis, robustness testing, and teacher alignment jointly support a central claim. By introducing intervention semantics and a partial order structural signature, EduAttr substantially surpasses existing baselines in node-level identification accuracy while also achieving principled robustness against text-level disguise and practical adoptability by teachers. It therefore fully closes the methodological gap between identifying AI borrowing and generating executable instructional actions. On the basis of this evidence, Section 5 discusses the theoretical significance, failure boundaries, and cross-disciplinary transfer pathways of the method.

Discussion

Mechanistic Complementarity of Counterfactual Probing and Partial Order Identification Signals

CPPC and POVR-Sig are grounded in two fundamentally different identification principles, and their mechanistic complementarity is not a post-hoc phenomenon but an inevitable consequence of the design. CPPC examines whether a reasoning commitment remains repairable when its premise is perturbed, and its criterion depends on interventional elicitation of model responses. POVR-Sig examines whether the dependency partial order satisfies the downward closure property of knowledge space theory, and its criterion depends on structural testing of node-level factual consistency. The two channels correspond to causal coupling and structural coupling, respectively, and are orthogonal in principle. A direct consequence of this orthogonality is that any AI borrowing strategy attempting to evade one channel necessarily leaves a trace in the other.

As shown in Table 9, CPPC and POVR-Sig exhibit substantially different response characteristics under five typical AI borrowing strategies. Borrowing an entire paragraph triggers both signals, and the joint operation of the two channels enables EduAttr to achieve a Node-F_1 of 0.912. Conclusion-level borrowing preserves authentic student reasoning but substitutes only the endpoint. CPPC fails because upstream counterfactual responses remain normal, whereas POVR-Sig precisely locates the violation in which a downstream node is correct while an upstream node is incorrect. Consequently, EduAttr exceeds Ghostbuster by 32.1 percentage points. Reasoning chain borrowing with a self-revised conclusion produces the opposite pattern. CPPC generates an alert because the middle section remains locked in a template, whereas the POVR-Sig signal weakens because the borrowed upstream chain is itself internally correct, so few pairs arise in which a downstream node is correct while its prerequisite is incorrect. The two channels therefore exchange roles. The ablation experiments further confirm this complementary coupling. Removing CPPC and POVR-Sig causes performance declines of 7.2 and 8.7 percentage points, respectively, showing comparable information contributions. The simple addition variant reaches only 0.798, substantially below the 0.843 of the complete version modulated by the depth weight, indicating that the two signals are not redundant but mutually calibrating. Because the strategy strata in Table 9 were sampled in balanced numbers rather than in proportion to their natural frequency, the harder strategies are deliberately over-represented and the absolute values in that table are not directly comparable with the pooled values in Table 5.

Table 9 Comparison of Dual-Channel Signal Responses and Identification Performance Under Typical AI Borrowing Strategies (Balanced Strategy Subset of Arm B, 60 Reports per Strategy)

Borrowing strategy	CPPC response	POVR-Sig response	EduAttr F_1	Ghostbuster F_1	F_1 gap	F_1 Share of arm B reports
Full paragraph borrowing	Strong	Strong	0.912	0.821	+0.091	28.6%

Borrowing strategy	CPPC response	POVR-Sig response	EduAttr F_1	Ghostbuster F_1	F_1 gap	F_1 Share of arm B reports
Conclusion-level borrowing (student reasoning retained)	Weak	Strong	0.837	0.516	+0.321	18.6%
Reasoning chain borrowing (conclusion self-revised)	Strong	Moderate	0.804	0.629	+0.175	17.6%
Full chain rewriting disguise	Moderate	Strong	0.756	0.418	+0.338	13.7%
Mixed borrowing (approximately 30% AI nodes)	Moderate	Moderate	0.821	0.593	+0.228	21.5%

Principled Advantages over AIGC Detectors and Failure Boundaries

The performance advantage of EduAttr over AIGC text detectors has a principled origin rather than resulting from engineering optimization. Criteria in the surface feature pathway focus on word frequency quantiles and perplexity curvature, and their identification target is whether literal statistical features resemble AI. The dual-channel criteria of EduAttr identify whether the reasoning chain resembles AI. The two types of criteria correspond to different phenomenological levels in principle. Under the original condition, EduAttr achieves a Node-F_1 of 0.843, exceeding the 0.713 of Ghostbuster by 13.0 percentage points. Under severe rewriting attacks, DetectGPT and Ghostbuster deteriorate by 30.5 and 28.5 percentage points, respectively, while EduAttr deteriorates by only 12.0 percentage points, expanding the gap between EduAttr and Ghostbuster to 29.5 percentage points. The substantial difference in deterioration arises from a principled fact. Literal disguise can conceal surface features, but it is difficult to disguise causal coupling and partial order structure simultaneously.

The failure boundaries must also be discussed candidly. In scenarios involving extremely short reports, when the dependency graph contains fewer than 5 edges, the POVR-Sig path signature becomes unstable because of insufficient samples, and the advantage of EduAttr over the baselines narrows to 5 to 7 percentage points. A highly formulaic student writing style also weakens the identification capacity of CPPC because counterfactual responses have limited ability to distinguish formulaic reasoning from borrowed reasoning. A third boundary concerns the reference labels themselves. The control arm is treated as a clean negative reference on the strength of supervised writing conditions, and any residual undetected borrowing in that arm would inflate the measured precision of every method compared here. The transcript based gold standard that makes node-level labelling reliable is likewise available only in an instrumented study setting and cannot be assumed in routine coursework. In addition, the trial was conducted in a single course at two medical colleges, the reports were written in Chinese, and only five pathogen cases from the core syllabus were covered; the generalizability of the findings to other disciplines, languages, and institutions therefore remains to be confirmed in subsequent external evaluations. These boundaries are inherent limitations of the method under extreme scenarios. They do not affect its applicability to mainstream classroom

teaching, but they indicate that subsequent work should continue to advance along three pathways, namely dependency graph node expansion, style disentanglement, and label acquisition that does not depend on complete interaction transcripts.

Instructional Operability of Early Warning Outputs and Teacher Adoption

Instructional operability is the key dimension that distinguishes the proposed method from pure identification algorithms and is also the criterion determining whether the system can be genuinely embedded in classroom workflows. Cohen's kappa between EduAttr and teacher annotations reaches 0.762, with a 95% confidence interval of [0.731, 0.793]. The interval contains the interteacher agreement level of 0.741, so system outputs agree with the gold labels about as closely as the expert annotators agree with one another. A teacher adoption rate of 87.5% shows that nearly nine in ten of the Top-10 entries are considered by teachers to be suitable for direct classroom use. A follow-up questioning success rate of 78.4% further shows that nodes in the list actually elicit student acknowledgment of borrowing or an inability to respond during classroom questioning, thereby providing closed-loop validation of the operability of the system output.

The ranked output is accepted by teachers primarily because the list contains more than node labels. Each entry includes a follow-up question template automatically generated according to the node type. Teachers receive not an abstract risk rating but a question that is specific to instructional action, such as "Please explain the key morphological evidence supporting this pathogen hypothesis." This design is consistent with the principle of descriptive feedback long emphasized in the formative assessment literature and also addresses the prerequisite for adoption repeatedly reported in recent surveys from the teacher perspective, namely, that outcome labels are insufficient and a specific next action is required. Grading time decreases from 12.4 minutes to 7.1 minutes per report, a reduction of 42.7%. This finding confirms that the system successfully focuses teacher attention from reading the entire report to questioning key nodes, creating time for teachers to engage in higher-level reflective teaching.

Strengths and Limitations

This study has several strengths. Node-level attribution labels in the experimental arm were adjudicated against the complete interaction transcripts rather than inferred from text alone, which provides an objective reference standard that is seldom available in prior evaluations of AI-generated text detectors. Seven baselines spanning surface features, supervised classification, learner modeling, and reasoning consistency were compared on the same test set, and the data were split at the student level so that writing style from the same student could not leak across sets. The ranked output was tested in authentic classroom sessions rather than only in offline evaluation. Several limitations also warrant consideration. The trial recruited eight intact classes at two institutions in a single course, so transfer to other disciplines, languages, and institutional contexts requires further study. Students in the experimental arm used generative AI openly and submitted their transcripts, which enabled transcript-based adjudication but differs from routine coursework in which AI use is often concealed and transcripts are unavailable.

Control-arm labels relied on the supervised writing condition rather than on direct evidence of absence of borrowing. GPT-4o served both as a tool used by students in the experimental arm and as the probing model of EduAttr; rerunning the counterfactual channel with probing models from other model families produced results consistent with the main findings, but other generators that students might use in practice were not systematically tested. Finally, EduAttr depends on several components, namely the dependency parser, the correctness classifier, and the counterfactual generator, whose errors propagate to the final attribution score.

Conclusion

This study addresses the methodological gap in identifying AI borrowing and providing instructional early warning in medical microbiology classrooms. It systematically constructs the EduAttr framework and quantitatively validates it on data from a two-arm randomized controlled trial. The principal contributions are summarized as follows.

(1) At the theoretical level, attribution identification is reformulated from document-level binary classification as a node-level discrimination problem, aligning identification granularity with the classroom follow-up questioning workflow.

(2) The reasoning commitment consistency test based on Counterfactual Premise Perturbation, CPPC, is proposed. It introduces the intervention semantics of causal inference into AI borrowing identification and overcomes the inherent limitations of static consistency measurement.

(3) The identifiable Path Signature based on the dependency Partial Order Violation Rate, POVR-Sig, is proposed. It constructs a principled criterion using the downward closure property of knowledge space theory and forms a complementary coupling with CPPC that is difficult for the same generation strategy to evade simultaneously.

(4) The Teaching Alert Node Ranking mechanism, TANR, is proposed. It integrates dual-channel identification signals and dependency depth weights into a Top-K follow-up question list that teachers can execute, thereby closing the operational gap between identification and instructional action.

(5) Experiments involving 240 students and 22,362 reasoning nodes demonstrate that EduAttr exceeds the strongest surface feature baseline Ghostbuster by 13.0 percentage points in node-level F_1. Under severe rewriting attacks, this gap further expands to 29.5 percentage points, and Cohen's kappa with teacher judgments reaches 0.762.

The method still exhibits reduced identification capacity for extremely short reports and highly formulaic writing styles. The invocation cost of the probing large language model also imposes resource constraints on real-time deployment in large classes. Future work will continue in three directions, namely adaptive node expansion, style disentanglement, and lightweight probe design. The method will also be gradually

extended to other basic medical disciplines, including medical immunology and pathophysiology, to examine its transferability across different disciplinary dependency structures.

Ethics approval and consent to participate

All procedures involving human participants followed the principles of the Declaration of Helsinki. The study protocol was reviewed and approved by the Ethics Committee of Changchun University of Chinese Medicine, which acted as the leading institution, and the collaborating college conducted the study under that approval. The approval number was CUCM-SBMS-REC-2024-0710, dated September 15, 2024. All participants were adult students, participation was entirely voluntary, and written informed consent covering both participation and the analysis of the submitted reports and generative AI interaction transcripts was obtained from every student before enrolment. All materials were anonymized before annotation and analysis.

Consent for publication

Not applicable.

Availability of data and materials

The source code of EduAttr, covering the counterfactual probing module, the partial order violation signature, the node ranking mechanism, and the evaluation scripts, is openly available in the EduAttr repository at https://github.com/a18907052542-lang/EduAttr.git. The datasets generated and analysed during the current study are not publicly available because they consist of student reports and generative AI interaction transcripts whose consent terms do not permit open redistribution, but an anonymized subset is available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Key Project of the Jilin Provincial Higher Education Research Project sponsored by the Jilin Provincial Society of Higher Education (JGJX26B14), and by the Jilin Provincial Higher Education Research Projects JGJX25D0283 and JGJX25C064. The funding bodies had no role in the design of the study, in the collection, analysis, and interpretation of data, or in the writing of the manuscript.

Authors' contributions

YS conceived the study, designed the EduAttr framework, and drafted the manuscript. SZ implemented the counterfactual probing and partial order violation modules and carried out the experiments. LH organized the classroom data collection, coordinated the

annotation panel, and performed the statistical analysis. FW supervised the study, obtained the funding, and revised the manuscript critically for important intellectual content. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the teachers who served on the annotation panel and the students who took part in the study.

Additional files

Additional file 1: Completed CONSORT checklist.

Additional file 2: Study protocol.

Additional file 3: Counterfactual premise template, correctness classifier training labels summary, and parser evaluation details.

Reference

1. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180.

2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940.

3. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med. 2024;99(1):22-27.

4. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach. 2024;46(4):446-470.

5. Kasneci E, Sessler K, Kuchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ. 2023;103:102274.

6. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-1239.

7. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltynek T, Guerrero-Dib J, Popoola O, et al. Testing of detection tools for AI-generated text. Int J Educ Integr. 2023;19(1):26.

8. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(9):2613-2622.

9. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med. 2024;7(1):20.

10. Fraser KC, Dawkins H, Kiritchenko S. Detecting AI-generated text: factors influencing detectability with current methods. J Artif Intell Res. 2025;82:2233-2278.

11. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J. GPT detectors are biased against non-native English writers. Patterns. 2023;4(7):100779.

12. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. Int J Educ Integr. 2023;19(1):17.

13. Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. 2024;630(8017):625-630.

14. Augenstein I, Baldwin T, Cha M, Chakraborty T, Ciampaglia GL, Corney D, et al. Factuality challenges in the era of large language models and opportunities for fact-checking. Nat Mach Intell. 2024;6(8):852-863.

15. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. 2024;58(11):1276-1285.

16. Abdelrahman G, Wang Q, Nunes B. Knowledge tracing: a survey. ACM Comput Surv. 2023;55(11):1-37.

17. Tao J, Zhao W, Zhang Y, Ma Y, Zheng Q. Cognitive diagnostic assessment: a Q-matrix constraint-based neural network method. Behav Res Methods. 2024;56(7):6981-7004.

18. Fleckenstein J, Meyer J, Jansen T, Keller SD, Koller O, Moller J. Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. Comput Educ Artif Intell. 2024;6:100209.

19. Kim J. Leading teachers' perspective on teacher-AI collaboration in education. Educ Inf Technol. 2024;29(7):8693-8724.

20. Mitchell E, Lee Y, Khazatsky A, Manning CD, Finn C. DetectGPT: zero-shot machine-generated text detection using probability curvature. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR; 2023. p. 24950-24962.

21. Verma V, Fleisig E, Tomlin N, Klein D. Ghostbuster: detecting text ghostwritten by large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human

Language Technologies (Volume 1: Long Papers). Mexico City: Association for Computational Linguistics; 2024. p. 1702-1717.

22. Sadasivan VS, Kumar A, Balasubramanian S, Wang W, Feizi S. Can AI-generated text be reliably detected? Stress testing AI text detectors under various attacks. Trans Mach Learn Res. 2025.

23. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824-24837.

24. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. In: Proceedings of the 11th International Conference on Learning Representations. Kigali: OpenReview; 2023.

25. Turpin M, Michael J, Perez E, Bowman SR. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. Adv Neural Inf Process Syst. 2023;36:74952-74965.

26. Huang Z, Kosan M, Medya S, Ranu S, Singh A. Global counterfactual explainer for graph neural networks. In: Proceedings of the 16th ACM International Conference on Web Search and Data Mining. New York: ACM; 2023. p. 141-149.

27. Lucic A, ter Hoeve MA, Tolomei G, de Rijke M, Silvestri F. CF-GNNExplainer: counterfactual explanations for graph neural networks. In: Proceedings of the 25th International Conference on Artificial Intelligence and Statistics. Valencia: PMLR; 2022. p. 4499-4511.

28. Ghosh A, Heffernan N, Lan AS. Context-aware attentive knowledge tracing. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2020. p. 2330-2339.

29. Gao W, Liu Q, Huang Z, Yin Y, Bi H, Wang MC, et al. RCD: relation map driven cognitive diagnosis for intelligent education systems. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM; 2021. p. 501-510.

30. Wang F, Liu Q, Chen E, Huang Z, Chen Y, Yin Y, et al. Neural cognitive diagnosis for intelligent education systems. Proc AAAI Conf Artif Intell. 2020;34(4):6153-6161.

31. Hopewell S, Chan AW, Collins GS, Hrobjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. BMJ. 2025;389:e081123.

32. Gehrmann S, Strobelt H, Rush AM. GLTR: statistical detection and visualization of generated text. In: Proceedings of the 57th Annual Meeting of the Association for

Computational Linguistics: System Demonstrations. Florence: Association for Computational Linguistics; 2019. p. 111-116.

33. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019.

34. Vie JJ, Kashima H. Knowledge tracing machines: factorization machines for knowledge tracing. Proc AAAI Conf Artif Intell. 2019;33(1):750-757.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1CONSORTchecklist.docx
- Additionalfile2StudyProtocol.docx
- Additionalfile3TechnicalDetails.docx