Nova Science Press

Journal of Psychology & Education

Vol. 1, No. 4 (2026)

Confucius' "Teaching in Accordance with Students' Aptitude" Realized Through AI:A Case of Philosophical Academic Writing Course Reform

Siyan Yu[1]*

1 Department of Philosophy, Guizhou University, Guiyang, 550025, China

*Corresponding author: Siyan Yu; Email: yusy@gzu.edu.cn

Funding statement: This work was supported by the Guizhou University Teaching Reform Project "AI-Empowered Academic Writing and Research Ethics in Philosophy" (Project No. XJG2025065).

Journal of Psychology & Education • Vol. 1, No. 4 (2026)

DOI: https://doi.org/10.66581/m9260b32

Received 19 May 2026 • Accepted: 26 May 2026 • Published 31 May 2026

CITATION

Yu, S. (2026). Confucius' "Teaching in Accordance with Students' Aptitude" Realized Through AI: A Case of Philosophical Academic Writing Course Reform. *Journal of Psychology & Education*, *1*(4), 31. https://doi.org/10.66581/m9260b32

Published by Nova Science Press, HK

Abstract

Large philosophy classes make it difficult to realize Confucius' ideal of teaching in accordance with students' aptitude. Conventional assessments poorly capture growth in higher-order writing skills. This article reports a 16-week reform of an undergraduate philosophical academic writing course that integrated an AI "learning companion," structured "arguing-with-AI" activities, and peer-supported clinical tutoring. Drawing on work in intelligent tutoring, stealth assessment, and emerging research on generative AI in philosophy teaching, we designed an 8-week conceptual module plus an 8-week clinic-style supervision module. Using a quasi-experimental historical cohort design (N = 79), we compared rubric-based writing scores, process data from AI interaction logs, and student self-reports across cohorts. The AI-supported design was associated with greater increases in argumentative coherence, evidence use, and originality in this sample than a traditional design, and it expanded individualized feedback coverage. We discuss how generative AI can extend—but not replace—instructors' capacity to enact individualized instruction at scale in a philosophy context, and outline implications for AI governance and academic integrity.

Key words: individualized instruction; generative AI; philosophy writing; higher education; stealth assessment

1.Introduction

In Analects 11.22, Confucius offers different answers to different students regarding the same question., illustrating his pedagogical maxim of teaching "in accordance with the student's aptitude". The underlying idea—that instruction should

be responsive to individual learners' needs rather than uniformly delivered—is widely shared in contemporary educational psychology and pedagogy. Research on differentiated instruction, formative assessment, and adaptive technologies all converges on the view that learning improves when teaching connects with learners' prior knowledge, interests, and difficulties.

Yet in mass higher education, particularly in writing-intensive humanities courses, this ideal remains hard to realize. In many institutions, academic writing courses serve large cohorts and are taught by a small number of instructors with limited time for individual supervision. Students may receive only one or two rounds of feedback on a term paper, and that feedback often focuses on surface-level mechanics or broad comments rather than the fine-grained reasoning and textual engagement that philosophical writing demands. As a result, some students complete such courses having learned the formalities of citation and structure but still lacking confidence in constructing and defending a philosophical argument.

1.1 Challenges in philosophical academic writing

Philosophical writing poses distinctive challenges. Beyond form, students must learn to:

Formulate clear, focused questions in a dense conceptual landscape;

Map existing positions and arguments with textual accuracy;

Make the inferential structure of their own arguments explicit;

Anticipate and respond to objections;

Balance fidelity to texts with critical originality.

These abilities are inherently developmental and deeply personal: different students struggle with different parts of the process. Some can generate ideas but fail to structure them; others can recite texts but hesitate to take a position; some write fluently but loosely, with hidden assumptions and gaps in reasoning.

Traditional assessment formats—one major paper plus a handful of assignments—are ill-suited to tracking such development. They capture a snapshot of performance but reveal little about how students arrived at it, which strategies they used, and where they improved or regressed. Furthermore, feedback in such settings is often generic ("clarify your argument") rather than specific to the student's evolving pattern of strengths and weaknesses.

1.2 AI-supported individualized instruction and stealth assessment

Research on artificial intelligence in education (AIED) has long explored ways to deliver individualized support at scale. Intelligent tutoring systems (ITSs) adapt problem sequences, hints, and feedback based on estimated learner states. Meta-analytic work suggests that some ITSs can rival, and occasionally match, the effectiveness of human tutoring in procedural domains such as mathematics and physics (VanLehn, 2011). Koedinger et al. (1997) showed that ITSs can be integrated into ordinary school settings and still produce meaningful gains.

A complementary strand, stealth assessment, embeds unobtrusive measures of competence within ongoing activity rather than relying solely on separate tests (Shute & Ventura, 2013). Process data—such as sequences of actions, time on task, and choice patterns—are used to infer underlying skills and update learner models. While most stealth assessment work has focused on games and STEM learning environments, the basic logic—using rich process traces to understand and support learning—can, in principle, extend to writing.

1.3 Generative AI in academic and philosophical writing

The emergence of large language models (LLMs) such as ChatGPT adds a new dimension to AIED. Unlike domain-specific tutors, LLMs are general text generators that respond to natural language prompts. They can brainstorm ideas, suggest outlines, rephrase sentences, and even mimic argumentative styles. Early empirical work indicates that students, especially second-language writers, use LLMs to generate ideas, improve organization, and refine language (Cheng et al., 2025). At the same time, LLMs are prone to "hallucinations," producing plausible-looking but false information and fabricated references (Oertner, 2024), and can encourage a form of superficial engagement where students accept fluent prose without critically evaluating its content.

Within philosophy education, scholars have begun to explore more reflexive uses of AI. Mouser (2024) suggests that ChatGPT can be treated as a "student" to be taught: by attempting to explain philosophical concepts and arguments to the model and then critiquing its paraphrases, human students are pushed to clarify their own understanding.

Lemanek (2025) proposes using ChatGPT to generate "artificial referee reports" on student drafts, which students then must answer, thereby rehearsing the skills involved in responding to peer review. These designs align with a view of AI as a fallible interlocutor whose outputs are raw material for critical thinking rather than authoritative answers.

1.4 Gaps and research questions

Despite this emerging body of work, three gaps remain relevant to philosophical writing instruction:

Domain specificity. Most AI-in-writing studies focus on general composition or language learning, not on the particular epistemic and argumentative norms of philosophy.

Course-level integration. AI is often used for single tasks (e.g., generating a draft) rather than integrated across an entire course with aligned pedagogy and assessment.

Integrity and agency. There is limited empirical evidence on how AI-mediated support interacts with academic integrity norms and students' sense of authorship in credit-bearing, high-stakes contexts.

The present study addresses these gaps by examining a reform of an undergraduate philosophical academic writing course at a large Chinese university. Rather than testing an isolated intervention under laboratory conditions, we analyze how an AI "learning companion," structured "argue-with-AI" tasks, and a peer-supported clinical tutoring

model functioned in a real institutional context.

We asked:

To what extent was the AI-supported design associated with increased individualized feedback and more differentiated writing support in a large philosophy class?

How were key philosophical writing competencies — argumentative coherence, evidence use, originality — related to participation in the AI-supported design, compared with a previous, non-AI cohort?

How did students and tutors perceive the role of AI in relation to academic integrity and authorship?

Given the quasi-experimental, historical cohort design, we treat our analyses as providing evidence of associations rather than strong causal claims. Our aim is to offer a theoretically informed and empirically grounded case study that can inform, and be refined by, future work on AI-supported philosophy writing instruction.

2. Methods

2.1 Design and participants

We used a quasi-experimental historical cohort design. Participants were third-year undergraduate philosophy majors at Guizhou University enrolled in the compulsory course "Thesis Writing and Academic Norms."

Reform cohort (AI-supported condition). 2025–2026 academic year, 41 students

(22 female, 19 male). The course used the AI-supported design described below.

Comparison cohort (traditional condition). 2024–2025 academic year, 38 students (21 female, 17 male). The course followed a traditional design, emphasizing lectures on formats and norms, one term paper, and limited written feedback.

Students in both cohorts had similar admission scores and prior GPA; institutional records showed no statistically significant differences on these indicators. No students withdrew from either course, and all completed major assessments.

The choice of a historical comparison was driven by ethical and practical constraints: the department preferred not to randomize students within the same cohort to substantially different experiences in a compulsory course. We thus interpret findings conservatively as associations conditioned on cohort-level differences.

2.2 Course intervention

2.2.1 Overall structure and design principles

The reformed course spanned 16 weeks (32 contact hours), divided into two phases:

Weeks 1 – 8: Conceptual and norms module. Students developed core skills in problem formulation, literature review, argument structure, citation, and AI literacy.

Weeks 9–16: Clinical supervision module. Students iteratively drafted, critiqued, and revised their papers, engaging in "argue-with-AI" activities, peer review, and tutor consultations, culminating in a full paper and mock defense.

Four design principles guided the intervention:

Cognitive ecology. AI, knowledge graphs, rubrics, and peer feedback were treated as cognitive tools that, together with teachers and students, form an extended reasoning system.

Problem-driven learning. Each student developed a single research question into a full philosophical paper through staged tasks (topic proposal, literature review, argument outline, draft, revision).

Adaptive plus stealth assessment. While we did not implement a full ITS, we used AI and simple analytics to support differentiated tasks and used process data (prompt logs, reference verification tables, revision histories) to infer aspects of competence.

Social collaboration and governance. A tutor team and heterogeneous peer groups provided multi-source feedback, while explicit AI governance measures addressed hallucinations and authorship concerns.

2.2.2 AI learning companion and governance

Students in the reform cohort used a university-hosted large language model configured as an "AI learning companion." The system logged all prompts and responses with timestamps and anonymized identifiers. In designing AI use, we aimed to strike a balance: we wanted students to benefit from AI's generative and diagnostic capacities, but we also wanted to avoid dependence and preserve academic integrity.

Three governance mechanisms were implemented:

Prompt templates. We provided structured templates for three functions: retrieval,

generation, and critique. For example, retrieval prompts required AI to list only references that could be found in CNKI or Web of Science and to provide essential metadata (author, year, journal). Generation prompts asked AI to synthesize arguments based only on a given list of verified sources. Critique prompts instructed AI to identify potential logical gaps and counterexamples in student arguments.

Reference verification. For any AI-suggested reference, students had to perform manual database searches and record outcomes in a verification table. Unverifiable items or sources that did not match AI's descriptions were to be discarded or replaced.

AI use statement. With the final paper, each student submitted a statement detailing when, how, and why AI was used, what kinds of suggestions were accepted or rejected, and how hallucinations were detected. These statements, along with prompt logs, were graded and also used for academic integrity checks.

We presented AI to students as a fallible partner rather than an authority: helpful but not trustworthy by default.

2.2.3 Weekly activities (overview)

Weeks 1–8: Conceptual and norms module

Week 1: Orientation and baseline.

The instructor introduced course goals, the writing rubric, philosophical writing expectations, and AI governance principles. Students completed a baseline 800-word essay on a philosophical prompt and a self-efficacy survey. After a brief demo of the AI

companion, they practiced inputting simple prompts and reflecting on output quality.

Week 2: Topic narrowing and knowledge graphs.

Using CiteSpace maps of selected philosophical debates (e.g., moral responsibility, personal identity), students identified potential research niches and drafted three candidate paper titles. Class discussions linked topic coherence to logical laws. Homework involved AI-assisted refinement of one topic and a 300-word rationale.

Week 3: Literature search and verification.

Sessions covered advanced search techniques, journal hierarchies, and typical AI citation errors. In groups, students collected AI-recommended references and attempted to verify them. Homework required submitting a verification table and a corrected reference list.

Week 4: Argument structure and counterexamples.

The instructor modeled how to reconstruct arguments from classic texts and design counterexamples. Students wrote short arguments for their theses and used AI as a critic. They then revised their arguments based on AI and peer feedback.

Week 5: Prompt engineering.

Students learned and practiced structured prompts for retrieval, generation, and critique. Emphasis was placed on specifying premises, constraints, and expected output format. Homework documented prompt – response iterations leading to improved paragraphs.

Week 6: Academic norms and degree regulations.

Discussions centered on plagiarism, patch-writing, self-plagiarism, and emerging norms on AI-assisted writing. Case studies illustrated AI hallucinations and misattributions. Students drafted a literature review section with fully verified citations.

Week 7: Peer review training.

Students anonymously reviewed two peers' literature reviews using the rubric and a peer-feedback template. In-class activities analyzed constructive and unhelpful feedback examples. Students then revised their work.

Week 8: Midterm presentations.

In small groups, students presented their questions, key literature, and preliminary argument outlines. A tutor panel offered suggestions, focusing on feasibility and depth.

Weeks 9–16: Clinical supervision module

Weeks 9–10: Individual consultations.

Each student had at least one 20-minute one-to-one consultation with a tutor, reviewing drafts and AI logs. Tutors asked students to explain selected prompts and justify which AI suggestions they followed or ignored.

Week 11: Public "argue-with-AI" workshop.

Students demonstrated AI-generated critiques of their arguments and and formulated counterarguments in front of peers. The class and tutors then offered

additional challenges, approximating a low-stakes defense.

Week 12: Ethics and AI authorship.

Sessions focused on authorship norms, responsibility for content, and equity concerns. Students drafted AI use statements.

Week 13: Full draft submission and blind review.

Students submitted full papers, AI logs, verification tables, and draft AI use statements. Three raters blind-evaluated the papers.

Week 14: Mock defenses.

Students presented their papers and answered questions from a mixed panel of peers and tutors. Oral feedback targeted clarity, structure, and engagement with objections.

Week 15: Final revision and submission.

Students revised papers based on feedback and submitted final versions, including polished AI use statements and short reflective essays.

Week 16: Posttest and portfolios.

Students completed a posttest essay on a new prompt and a post-course survey. They assembled portfolios documenting progress in argumentation, literature use, and AI literacy.

2.3 Measures

2.3.1 Writing rubric

Our four-dimension rubric (0 – 20 per dimension, total 80) was adapted from existing work on analytical writing and philosophical argumentation, refined through local pilot testing:

Argumentative coherence. Criteria evaluated whether the paper articulated a clear thesis, explicitly laid out supporting reasons, avoided major logical fallacies, and engaged with at least one objection.

Conceptual and textual accuracy. Items assessed the correctness and consistency of key concepts and the fidelity of textual interpretations.

Evidence use and citation integrity. Items examined the relevance of sources, integration of textual evidence into argument, and correctness of citations (including verification against databases).

Originality and critical depth. Items captured the extent to which the student articulated a defensible position, identified tensions among views, and offered non-trivial reasoning rather than mere summary.

Each dimension comprised 4 – 5 behavioral indicators rated on a 5-point scale; scores were summed and rescaled. Cronbach's alpha for the total scale at posttest was .86.

2.3.2 Process indicators

For the reform cohort, we derived:

Prompt iteration count: Number of distinct prompts issued for key assignments (topic selection, literature review, argument drafting, revision).

Hallucinated reference rate: Proportion of AI-suggested references marked as unverifiable in students' verification tables.

Revision depth: Revision depth: Approximated by the number of tracked changes between draft versions and the qualitative categories of those changes (local edits vs. structural revisions).

Peer feedback quality: Ratings on a short rubric assessing specificity (e.g., references to particular sentences), constructiveness (suggestions, not just criticism), and alignment with rubric criteria.

2.3.3 Surveys and interviews

Writing self-efficacy was measured by a 6-item scale adapted from existing writing research (e.g., "I am confident that I can construct a clear philosophical argument in writing"), rated from 1 (strongly disagree) to 5 (strongly agree). Cronbach's alpha at pretest was .79 and .83 at posttest.

Attitudes toward AI were measured using items on perceived usefulness, perceived risk (e.g., "AI might make me lazy in my thinking"), and perceived control ("I feel I can decide when and how much to rely on AI"), also rated on a 5-point scale.

Semi-structured interviews with 12 reform cohort students and 4 tutors probed:

How they used AI at different stages;

How their views of AI changed;

How they understood authorship and responsibility in AI-assisted writing.

2.4 Data analysis

Quantitatively, we used paired-sample t tests to assess within-cohort pre/post changes in rubric scores and self-efficacy. Independent-samples t tests compared final rubric scores across cohorts. To adjust for potential cohort differences, we ran multiple regression models with final total scores as outcomes and cohort (reform vs. comparison), baseline scores, admission scores, and prior GPA as predictors. For robustness, we conducted propensity score matching (PSM) to create matched subgroups based on baseline writing scores and GPA, then re-estimated cohort differences.

We reported Cohen ' s d and 95% confidence intervals along with p values. Statistical analyses were performed in R. Qualitative data were thematically coded by two researchers; disagreements were resolved through discussion. We used qualitative findings to interpret quantitative patterns and to illustrate mechanisms, not as independent outcome measures.

3. Results

3.1 Individualized feedback and process engagement

One aim of the reform was to increase individualized feedback coverage. In the comparison cohort, course records showed that only 3 of 38 students (7.9%) received

more than two substantial, paper-specific feedback episodes beyond generic comments (e.g., a margin note such as "needs clearer structure"). Such episodes typically involved a face-to-face meeting or detailed written comments.

In contrast, in the reform cohort, all 41 students received at least four distinct feedback episodes targeting their own work, combining:Clinical consultations with tutors;Structured peer reviews.

The mean number of feedback episodes per student was 4.8 (SD = 1.2). AI logs indicated that 100% of students engaged the AI in at least one retrieval, one generation, and one critique function across the term. The median number of prompt–response rounds related to the term paper was 16 (IQR = 12–21), suggesting sustained, rather than incidental, AI use.

Students' qualitative accounts corroborated these patterns. Several noted that in the traditional course they "barely spoke" with the instructor about their actual arguments, whereas in the reformed course they had opportunities to "test ideas with both AI and tutors" and felt "seen" as individual learners.

3.2 Writing outcomes

Descriptive statistics for writing rubric scores are summarized in Table 1 (not shown here).

In the reform cohort, total rubric scores increased from M = 47.3 (SD = 8.1) at baseline to M = 61.9 (SD = 7.4) on the final paper, $t(40) = 9.15$, $p < .001$, $d = 1.43$, 95%

CI [0.99, 1.86]. Gains were observed on all four dimensions, with Cohen's d ranging from 0.62 (originality and critical depth) to 1.13 (argumentative coherence).

In the comparison cohort, total scores improved from M = 46.0 (SD = 7.9) to M = 52.1 (SD = 8.0), $t(37) = 4.32$, $p < .001$, $d = 0.70$, 95% CI [0.37, 1.02]. Dimension-level gains were smaller, particularly in evidence use and originality.

An independent-samples t test on final scores showed a significant reform cohort advantage, $t(77) = 3.95$, $p < .001$, $d = 0.89$, 95% CI [0.44, 1.33]. In regression models controlling for admission score, prior GPA, and baseline writing performance, cohort (reform vs. comparison) remained a significant predictor ($\beta \approx 4.8$, $p < .01$). PSM analyses yielded similar estimates (average treatment effect on the treated $\approx$ 4.2 points, 95% CI excluding zero), suggesting that baseline differences alone are unlikely to account for the observed performance gap.

3.3 AI governance and reference integrity

We next examined how the AI governance measures related to reference integrity. In the first substantial writing assignment (early literature review drafts), 137 AI-suggested references were recorded in students' verification tables. Of these, 20 (14.6%) were classified as unverifiable—either not found in CNKI or Web of Science, or present but with mismatched metadata (e.g., wrong journal or year). Common patterns included plausible-sounding but nonexistent journals and authors with no publications on the given topic.

After repeated verification practice and explicit classroom discussions about hallucinations, the number of unverifiable AI-suggested references decreased substantially. In later drafts (Weeks 10–12), across 124 AI-suggested references, only 4 (3.2%) were deemed unverifiable. No unverifiable references remained in final papers. Instead, when AI suggested questionable sources, students increasingly replaced them with verified ones or omitted them altogether.

Students' AI use statements and interviews further illustrated changes in behavior. Early in the semester, some students described a kind of epistemic deference to AI:

"At the beginning, I just believed whatever reference AI listed. It looked so formal that I didn't think to question it. I assumed the problem was with my search, not with the AI." (Student 7)"

As the course progressed, many students reported that the mandatory verification procedures had reshaped their habits:

"The verification table was annoying at first, but it forced me to slow down and actually check. Now I feel I have a kind of 'second sense' for suspicious references. I'm less easily impressed by a long list of sources." (Student 19)"

"Because I knew I had to explain my AI use to the tutor, I became more careful. I started to ask myself: if I can't find this article, should I trust any of AI's other claims here?" (Student 3)"

At the same time, several students pointed out that reference verification added

workload and occasionally diverted attention from other aspects of writing:

"Sometimes I felt I spent more time chasing after bad references than improving my argument. It helped my research skills, but there were weeks when the technical checking felt heavier than the philosophical thinking." (Student 12)"

These mixed reflections suggest that governance mechanisms can indeed foster more critical and individualized use of AI — students learn to distinguish between reliable and unreliable outputs and to adjust their behavior accordingly—but they also introduce cognitive and temporal costs that need to be balanced against other instructional priorities.

3.4 Student perceptions

Reform cohort students showed significant gains in writing self-efficacy (pre: M = 2.9, SD = 0.7; post: M = 3.7, SD = 0.6 on a 5-point scale), $t(40) = 6.28$, $p < .001$, $d = 0.98$. Changes were especially pronounced on items related to structuring arguments and responding to objections. Students who initially reported that they "didn't know where to start" in philosophical writing often described, by the end of the course, a clearer sense of how to break a broad question into manageable moves.

Interviews highlighted that many students experienced the AI-supported design as a form of individualized instruction that had previously been unavailable in large classes. One student commented:

"In other courses, I feel like the teacher only sees the final version of my work.

Here, the AI and the tutors saw my messy thinking step by step. The AI wouldn't get tired of my questions, and the tutor would focus on the parts where I seemed really stuck. It felt closer to Confucius' idea of teaching different students differently." (Student 5)"

Another drew a contrast with more traditional courses:

"Usually, I just get one set of general comments at the end. This time, I could ask AI very specific questions about my own topic and get an immediate response. Then in the consultation, the tutor reacted to my prompts and drafts, not just to a generic problem set. I felt the course was built around my learning trajectory, not around a fixed script." (Student 21)"

At the same time, not all perceptions were unambiguously positive. Some students expressed ambivalence about the psychological effects of easy access to AI:

"AI is like a calculator for arguments. It's great because it shows me patterns and counterarguments I didn't think of. But I also worry that if I use it too much, I might stop trying to think things through myself." (Student 9)"

Others noted that the presence of AI sometimes made it harder to judge what "counts" as their own work:

"When I read my draft, sometimes I honestly couldn't remember which sentences started from me and which came from AI suggestions I had edited. That made me uncomfortable. I don't want my final paper to feel like it belongs to the machine."

(Student 14)"

A few students also pointed to subtle equity concerns:

"Students who already have strong logic skills seem to use AI to go even deeper—they know how to push it. Some of us who are weaker can become too passive, just accepting what it gives us. So AI can widen gaps if we are not careful." (Student 2)"

These comments resonate with our quantitative findings and with broader concerns in the literature. On the one hand, students experienced AI as a means of receiving more tailored, immediate support than a single instructor could provide, especially when combined with clinical tutoring and peer review. On the other hand, they recognized risks of over-reliance, blurred authorship, and the possibility that those who are already more skilled may benefit disproportionately from AI tools.

Taken together, student perceptions underline the central claim of this study: generative AI can contribute to realizing individualized instruction at scale in philosophical writing, but only when embedded within a carefully governed ecology that attends both to opportunities (rich, personalized feedback) and to emergent risks (dependency, inequity, and erosion of authorship).

4. Discussion

4.1 Interpreting the findings

Within the limits of a quasi-experimental historical comparison, our findings suggest that the AI-supported design was associated with both expanded individualized

feedback and larger gains in several dimensions of philosophical writing than those observed in a prior cohort under a traditional model. Students in the reform cohort received more frequent, more varied, and more targeted feedback, and their final papers, on average, exhibited clearer argument structures, more appropriate use of evidence, and somewhat greater originality.

From a theoretical standpoint, these patterns align with ideas from AIED and stealth assessment. Rather than implementing a full ITS, we created a hybrid ecology in which AI, tutors, and peers each played specific roles. The AI companion supported brainstorming, organization, and critique; tutors provided higher-order guidance and ethical framing; peers offered perspective-taking and audience awareness. Process data—prompt logs, verification tables, revision histories—functioned as both learning artifacts and assessment evidence, consistent with stealth assessment's emphasis on continuous, embedded measurement (Shute & Ventura, 2013).

Our results also speak to current debates about generative AI in writing. They suggest that harmful patterns—such as uncritical acceptance of hallucinated references or overreliance on AI-generated prose—are not inevitable. Under structured conditions, with explicit verification routines and reflective AI use statements, students can learn to treat AI outputs as starting points and objects of critique rather than as authoritative sources. This is in line with Mouser's (2024) and Lemanek's (2025) calls to position AI as a "student to be taught" or an "artificial reviewer" whose contributions must be interrogated.

4.2 Limitations

Several limitations qualify these conclusions. First, the historical cohort design lacks randomization; as such, we cannot definitively attribute cohort differences to the AI-supported design. Unmeasured cohort-specific factors (e.g., broader departmental culture, external stressors) may have played a role, even though we controlled for admission scores, GPA, and baseline writing performance and used PSM.

Second, the sample was limited to philosophy majors at a single Chinese university. Cultural factors may influence both students' receptivity to AI and their interpretations of Confucian notions like individualized instruction. The generalizability of our findings to other disciplines, institutions, or countries is therefore uncertain.

Third, our measures capture short-term outcomes within one course. We do not yet know whether the observed gains in writing and self-efficacy persist over time or transfer to other contexts, such as capstone theses or professional writing. Longitudinal research is needed.

Fourth, although our process indicators are richer than those in many writing studies, they remain relatively coarse. For instance, we counted prompt iterations but did not systematically code prompt quality or characterize AI responses beyond reference hallucinations. More fine-grained computational linguistics or protocol analyses could yield deeper insights into how specific patterns of AI use relate to learning.

Finally, while many students reported becoming more critical in their AI use, some

still described strong temptations to "let AI think for me," especially when under time pressure. Our design mitigated but did not eliminate this risk, highlighting the need for ongoing attention to students' epistemic habits and moral agency.

4.3 Implications and future directions

Despite these limitations, the study offers several implications for research and practice.

For instructors in philosophy and related disciplines, the findings suggest that generative AI can be integrated into writing pedagogy in ways that support, rather than undermine, core disciplinary values. Key design choices include:

- Framing AI as a fallible interlocutor whose outputs must be checked and debated;
- Embedding AI use within a broader system of human tutoring and peer review;
- Requiring students to log and reflect on their AI interactions, thereby making the "invisible work" of thinking more visible.

For educational researchers, our work illustrates the value of combining product-based measures (blind-rated essays) with process-based evidence when studying AI-mediated learning. Future research could investigate how different prompt strategies, degrees of AI reliance, or patterns of verification correlate with learning gains or with changes in epistemic beliefs.

Two avenues of future work seem particularly important. One is comparative: implementing similar designs in other humanities and social science courses (e.g.,

history, law, sociology) to identify which elements are discipline-specific and which are transferable. The other is longitudinal: following cohorts into their capstone projects and early careers to see whether AI-supported writing instruction has sustained effects on reasoning and communication practices.

Finally, from a broader perspective, our case contributes to ongoing debates about the role of AI in education. Confucius' notion of individualized instruction rests on respect for students as persons with distinct capacities and trajectories. If designed and governed well, AI can extend teachers' ability to recognize and respond to these differences; if designed and governed poorly, it can encourage homogenization, passivity, and dependence. The challenge, and opportunity, for educators is to shape AI use in ways that are compatible with—and perhaps even deepen—the humanistic aims of higher education.

5. Conclusion

Confucius' dictum of teaching in accordance with students' aptitude has often been honored in rhetoric but frustrated in practice, particularly in large university classes where time and attention are scarce. In this article, we reported on a case in which generative AI was embedded in an undergraduate philosophical writing course in ways intended to support more individualized instruction without displacing human judgment.

Within the limits of a quasi-experimental historical comparison, the AI-supported design was associated with both expanded opportunities for individualized feedback

and larger gains in several key dimensions of philosophical writing than those observed in a prior cohort under a more traditional model. Process data and interview accounts suggest that many students came to treat AI less as an answer-providing oracle and more as a fallible interlocutor and diagnostic tool. At the same time, concerns about overreliance and hidden AI use remained, underscoring the importance of explicit governance and continued ethical reflection.

These findings do not demonstrate that generative AI, by itself, improves philosophical writing. Rather, they indicate that when AI is carefully configured—through prompt templates, verification routines, and AI use statements—and when it is combined with human clinical tutoring and peer review, it can contribute to an instructional ecology in which students' reasoning is made more visible and open to critique. In that sense, AI can extend instructors' capacity to approximate individualized instruction at scale, while still relying on teachers and peers for high-level judgment, ethical framing, and discipline-specific guidance.

Realizing the constructive potential of AI in philosophy education will require more rigorous designs, multi-site and longitudinal studies, and ongoing normative discussion about authorship, responsibility, and the aims of philosophical training. Our contribution is modest: a context-bound case study that illustrates one way in which Confucius' ideal of individualized instruction might be interpreted and partially instantiated in the age of generative AI, and a set of design principles that others may adapt, refine, or challenge.

References

Cheng, D., Li, M., & Lee, T. (2025). Leveraging ChatGPT for research writing: An exploration of ESL graduate students ' practices. Computers and Composition, 76, 102934. https://doi.org/10.1016/j.compcom.2025.102934

Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education, 8(1), 30–43.

Lemanek, K. (2025). Artificial reviewers: Teaching academic writing with ChatGPT. Teaching Philosophy, 48(3), 415–430. https://doi.org/10.5840/teachphil202548338

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence unleashed: An argument for AI in education. Pearson.

Mouser, R. (2024). Writing with ChatGPT. Teaching Philosophy, 47(2), 173 – 191. https://doi.org/10.5840/teachphil2024429197

Oertner, M. (2024). ChatGPT als Recherchetool? Fehlertypologie, technische Ursachenanalyse und hochschuldidaktische Implikationen. Bibliotheksdienst, 58(5), 259 – 297. https://doi.org/10.1515/bd-2024-0042

Pane, J. F., Steiner, E. D., Baird, M. D., & Hamilton, L. S. (2017). Informing progress: Insights on personalized learning implementation and effects (RR-2042-BMGF). RAND Corporation.

https://doi.org/10.7249/RR2042

Shute, V. J., & Ventura, M. (2013). Stealth assessment: Measuring and supporting learning in games. MIT Press. https://doi.org/10.7551/mitpress/9589.001.0001

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197 – 221. https://doi.org/10.1080/00461520.2011.611369

Ethics Approval

This study involved human participants (undergraduate students). All procedures were reviewed and approved by the Ethics Committee of the School of Philosophy, Guizhou University. Participation in research components (surveys, interviews, and use of de-identified course data for analysis) was voluntary. Students provided written informed consent, were informed that non-participation would not affect their course grades, and could withdraw at any time without penalty. All data were anonymized prior to analysis.

Research Transparency Statement

This study was not preregistered. All measures, conditions, and analytic decisions are reported. No observations were excluded after data collection. De-identified materials (rubric, sample prompts, survey items) and analysis code are available from the corresponding author on reasonable request, subject to institutional data-protection rules and local data-protection regulations.