Can Intelligent User Interfaces Engage in Philosophical Discussions?: A Longitudinal Mixed-Methods Study of Philosophers' Judgments

Yibo Meng[1,*,†], Bingyi Liu[2,†], Lyumanshan Ye[3,†] and Yan Guan[1]

[1]Tsinghua University, Beijing, China
[2]University of Michigan, Ann Arbor, United States
[3]Shanghai Jiao Tong University, Shanghai, China

Abstract

This study examines how philosophy scholars' judgments of generative AI–based Intelligent User Interfaces (IUIs) in philosophical discourse evolve over time. We conducted a three-year (2023–2025) mixed-methods longitudinal study across three annual study waves, collecting data through repeated surveys, blind evaluation tasks, and semi-structured interviews. with 16 philosophy scholars and students. Qualitative interviews reveal a three-stage trajectory, moving from initial resistance and unfamiliarity, through instrumental acceptance of the IUI as a tool, to principled skepticism regarding its capacity for genuine philosophical thought. Quantitative data from blind assessments complicate this trajectory: when evaluating anonymized philosophical responses produced by both humans and an IUI, participants increasingly rated the IUI's outputs higher over time and, by 2025, above those of human experts across multiple dimensions of philosophical text quality, including originality of thought, despite persistent normative reservations. Taken together, these findings reveal a persistent divergence between participants' normative beliefs about what philosophical agents ought to be and their empirical evaluations of what AI systems can do, highlighting a failure of expectation calibration in expert judgments under conditions of uncertainty.

Keywords

Generative AI, Large Language Models (LLMs), Expert Judgment, Normative Evaluation, Blind Assessment, Longitudinal Study, Human–AI Evaluation, Design-Informed Research,

1. Introduction

Recent advances in generative AI have enabled systems to produce structured and fluent responses to questions traditionally addressed within philosophy, such as those concerning fairness, responsibility, and reasoning[1, 2, 3, 4]. As a result, generative AI is no longer only a topic of philosophical debate, but is increasingly encountered in philosophical pedagogy, writing, and evaluative contexts[5]. This development raises fundamental questions about how philosophical work should be evaluated when artificial systems are capable of producing its apparent form.

Philosophical responses to generative AI remain contested, with a prominent strand of work arguing that generative systems lack genuine understanding or normative commitment and should not be treated as intentional agents [6, 7, 8]. Yet this debate has largely centered on what AI should count as, leaving empirical investigations of how philosophers evaluate AI-generated philosophical content in practice relatively underexplored and largely cross-sectional.

Evaluation in philosophy is not a neutral assessment of clarity or correctness alone, but is embedded in disciplinary values, intellectual identity, and conceptions of philosophical responsibility[9]. Therefore, encounters with generative AI may generate tensions between the' stated beliefs of philosophers about the limits of artificial systems and their operational judgments when assessing philosophical arguments.

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus
*Corresponding author.
†These authors contributed equally.
mengyb22@tsinghua.org.cn (Y. Meng); bingyi@umich.edu (B. Liu); yuanlsy@mail.dlut.edu.cn (L. Ye); guany@tsinghua.edu.cn (Y. Guan)
0000-0002-0877-7063 (Y. Meng); 0009-0000-4312-153X (B. Liu); 0009-0007-1864-5957 (L. Ye); 0009-0004-5107-5050 (Y. Guan)

These tensions are unlikely to be resolved in a single encounter; instead, they may persist, shift, or be reinterpreted over time as both the technology and philosophers' familiarity with it develop [10].

To investigate these dynamics, we present a three-year longitudinal mixed-method empirical study (2023–2025) that examines how philosophers evaluate philosophical responses produced by generative AI over time. Rather than assessing the objective quality or superiority of AI-generated philosophy, this study focuses on how evaluative judgments are formed, challenged, and negotiated in repeated encounters with increasingly capable generative systems. The study combines repeated evaluative tasks with in-depth interviews to capture both observable judgment patterns and the evolving interpretations of those judgments by participants.

Specifically, we ask: (RQ1) how do philosophers' evaluative attitudes toward generative AI–produced philosophical responses change over time? and (RQ2) how do philosophers interpret and attempt to reconcile divergences between their evaluative judgments and stated commitments as system capability evolves and remains difficult to fix? By adopting a longitudinal and interpretive approach, this study provides empirical insight into how philosophical judgment is reshaped under conditions of ongoing technological uncertainty.

2. Related Works

Philosophical discussions of generative AI have primarily addressed normative questions concerning understanding, agency, and moral or epistemic responsibility [11, 12]. A prominent strand of work argues that large language models lack genuine understanding or normative commitment and therefore should not be regarded as philosophical or moral agents[13, 11]. While these accounts establish important conceptual boundaries, they focus on what AI ought to count as, rather than how philosophers evaluate AI-generated philosophical content under empirical, comparative, or blind assessment conditions.

In parallel, empirical studies in HCI and AI have examined the quality, creativity, and reasoning capabilities of generative AI systems[14, 15]. Such work typically evaluates AI outputs using researcher-defined criteria and is often conducted in single-session or short-term settings[16, 17]. Although these studies document rapid performance improvements, they offer limited insight into how expert judgments grounded in disciplinary values are formed or revised through repeated encounters with AI systems.

Prior research has emphasized that domain-specific norms, values, and professional identities shape evaluation. However, longitudinal investigations of how disciplinary judgment structures evolve under blind assessment conditions remain scarce. The present study addresses this gap by examining philosophers' evaluations of AI-generated philosophical responses across three years, combining repeated blind assessments with in-depth qualitative interviews.

3. Methods

We conducted a three-year longitudinal mixed-methods study (2023–2025) examining philosophers' engagement with generative AI-based Intelligent User Interfaces (IUIs). Each cycle combined surveys, semi-structured interviews, and a blind evaluation task to capture both stated beliefs and operational judgments, with the IUI presented as a conversational system responding to philosophical prompts.

3.1. Participants

We recruited participants through online academic and social media platforms. 16 participants with formal training in philosophy participated in the study. All participants provided informed consent, and the study received institutional ethical approval. Participants were assigned to one of the two groups described below.

Group 1 (Evaluators). Group 1 consisted of philosophy students and scholars who served as evaluators throughout the longitudinal study. Their role was to assess anonymized philosophical

Table 1
Study Constructs and Measurement Instruments

Construct	Items	Focus
Familiarity with Generative AI	A1–A6	Conceptual understanding, engagement, use, and self-perceived competence.
Attitudes toward IUI Participation	B1–B9	Perceived philosophical capabilities (e.g., explanation, argumentation, synthesis).
	B10–B14	Overall and meta-level judgments (e.g., usefulness, human-likeness).
Philosophical Text Quality	6-dim. rubric	Participant-designed questionnaire, blind evaluation across accuracy, clarity, coherence, rigor, depth, and originality.

responses and to reflect on the role of generative AI in philosophical discussion.

Group 2 (Human Contributors). Group 2 consisted of philosophy PhDs and professors who provided human-authored responses to philosophical questions. These responses were anonymized and mixed with AI-generated outputs for blind evaluation by Group 1.

3.2. Procedure

Each annual study cycle followed the same three-stage procedure (T1–T3).

T1 (Attitude and Familiarity Assessment). Participants in Group 1 completed a scale assessing their familiarity with generative AI (Appendix A) and an attitude scale regarding IUI participation in philosophical discussion (Appendix B). They also took part in a semi-structured qualitative interview probing their views on the role and limits of AI in philosophical practice (Appendix C).

T2 (Question Generation and Response Collection). Each member of Group 1 submitted at least two philosophical questions they believed could best distinguish between human-authored and IUI-generated responses. Submitted questions were collected, anonymized, and peer-reviewed by Group 1 for suitability. After revision and removal of duplicates, a final questionnaire was constructed (Appendix D).

Each member of Group 2 randomly answered one question from this questionnaire. In parallel, a large language model generated responses to all questions. Different models were used across study cycles (GPT-3.5 in 2023, GPT-4 in 2024, and GPT-4.5 in 2025). To ensure independent outputs, the AI session was refreshed after each prompt.

T3 (Blind Evaluation). Responses from the AI and Group 2 were anonymized, mixed, and presented to Group 1 for blind evaluation using a rubric they had collectively designed (Appendix E).

3.3. Instruments and Measures

We employed a mixed set of survey scales, interview protocols, and evaluation rubrics to assess participants' familiarity with generative AI, attitudes toward IUI participation in philosophical discussions, and judgments of philosophical response quality. Familiarity and attitudes were measured using custom Likert-scale instruments (Appendices A and B), complemented by semi-structured interviews probing participants' reasoning and value judgments (Appendix C). Philosophical performance was assessed through a participant-designed questionnaire and a six-dimensional evaluation rubric capturing core dimensions of philosophical text quality (Appendices D and E).

3.4. Data Analysis

Qualitative Analysis. All interviews were recorded with participant consent and ranged from 34 to 63 minutes in length. The recordings were transcribed within 24 hours and each transcript was independently reviewed and corrected by at least two researchers. Interview data were then analyzed

using thematic analysis to identify recurring patterns in participants' evaluative criteria, attitudes toward generative AI, and reflections on philosophical responsibility across study cycles.

Quantitative Analysis. Quantitative data from survey scales and blind evaluation ratings were analyzed using descriptive and inferential statistical methods to examine trends and changes over the three-year period.

4. Results

4.1. Qualitative Results

We conducted longitudinal interviews with 16 philosophy scholars over three years (2023–2025) to examine how their attitudes toward generative AI in philosophical discussion evolved over time. The analysis revealed a clear three-stage trajectory: initial resistance and unfamiliarity, followed by instrumental acceptance, and ultimately a shift toward principled philosophical critique.

Longitudinal Shift in Basic Attitudes At the beginning of the study in 2023, most of the participants expressed unfamiliarity and resistance to generative AI. This resistance was often rooted in disciplinary concerns about philosophical values, rather than in sustained engagement with technology. As one participant noted, *"I was initially dismissive of ChatGPT. I saw it as a kind of advanced parroting—noise rather than something relevant to philosophy's pursuit of truth"* (P7, 2023).

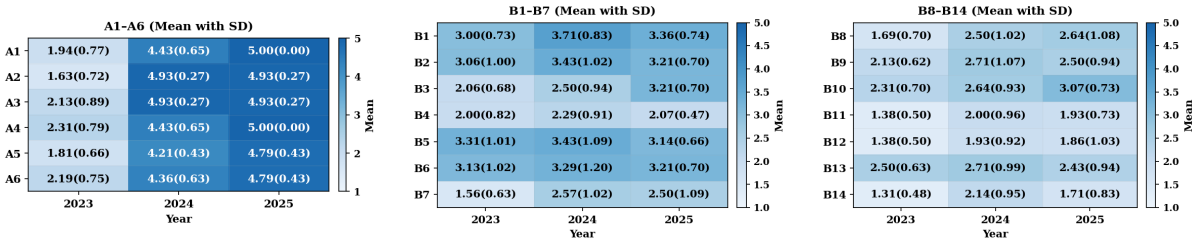
By 2024, after hands-on use and observation of AI-generated outputs, participants' attitudes shifted toward instrumental acceptance. Many began to frame generative AI as a useful auxiliary tool, while explicitly delimiting its role. Typical uses included information retrieval, summarization, and organization of preliminary arguments. One participant described this position succinctly: *"I treat it as a literature assistant or a starter. It gives me materials, not ideas"* (P4, 2024).

In 2025, participants' evaluations further evolved into a more principled, philosophical form of critique. At this stage, discussions moved beyond questions of usefulness to address whether generative AI possesses the fundamental capacities required for philosophical thinking, such as understanding, embodiment, and subjectivity. As one participant reflected, *"The issue is no longer whether it can analyze concepts, but whether it genuinely understands them. Philosophy is grounded in lived experience, and AI does not have that world"* (P8, 2025).

Capability Assessment: Formal Competence versus Philosophical Depth Across interviews, participants consistently acknowledged the strengths of generative AI in areas of formal competence, including logical structure and knowledge recall. For example, one participant observed that *"when it comes to constructing standard arguments or identifying formal fallacies, it can be more stable than some students"* (P2, 2025).

However, participants drew a sharp distinction between such formal performance and what they regarded as philosophical depth. In tasks requiring dialectical engagement, genuine revision of positions, or original insight, AI outputs were widely perceived as inadequate. Several participants noted that when challenged, AI responses tended to recycle existing formulations or juxtapose concepts without resolving underlying tensions. As one scholar explained, *"It looks like it is defending a position, but it is really just circling around the same claims. There is no sense of argumentative pressure that leads to a real shift in thinking"* (P9, 2025).

Value and Limits: Technical versus Principled Constraints Participants commonly differentiated between technical limitations of generative AI and what they viewed as principled constraints. Technical limitations—such as factual errors or contextual mismatches—were generally considered improvable through model iteration. In contrast, principled limitations were seen as intrinsic to AI's non-agential nature.



(a) Self-reported Familiarity with Generative AI (A1–A6)

(b) Perceived Philosophical Capabilities of IUI (B1–B7)

(c) Overall Attitudes toward IUI Participation (B8–B14)

Figure 1: Mean (SD) item scores for (a)familiarity (A1–A6) and (b,c) attitudes toward IUI participation (B1–B14) across 2023–2025. Values are mean (SD) on a 1–5 Likert scale; color encodes mean.

In terms of value, some participants noted that AI could occasionally prompt reflection, not by offering correct or profound answers, but by disrupting habitual ways of thinking. One participant remarked, *"Its value is not in giving the right answer, but in forcing you to reconsider assumptions you normally take for granted"* (P10, 2024). Nevertheless, such value was consistently framed as indirect and derivative rather than as evidence of genuine philosophical agency.

Overall Positioning of Generative AI in Philosophical Practice Across the three-year period, participants' overall stance toward generative AI became more cautious and clearly bounded. While its role as a powerful tool was broadly acknowledged, participants increasingly rejected the idea that generative AI could function as a philosophical agent. A recurring metaphor in later interviews characterized AI as reflective rather than generative in a strong sense. As one participant summarized, *"AI is more like a mirror. Through it, we are pushed to rethink what kinds of philosophical thinking can only be done by humans"* (P13, 2025).

4.2. Quantitative Results

We conducted a longitudinal quantitative analysis of participants' self-reported familiarity with generative AI (A1–A6) and attitudes toward IUI participation in philosophical contexts (B1–B14) across three study waves (2023–2025). All items were rated on a 1–5 Likert scale, and Figure 1 summarizes item-level means and standard deviations for each year. Familiarity ratings increased markedly over time, rising from uniformly low levels in 2023 to substantially higher and more stable levels in 2024 and 2025, with reduced variability on several items indicating growing consensus. Evaluations of the IUI's philosophical capabilities (B1–B7) followed a more moderate trajectory, improving from initial skepticism in 2023 to higher and stable levels in 2024 and 2025. In 2025, AI-generated responses received higher mean ratings than human-authored responses across accuracy, clarity, coherence, argumentative rigor, philosophical depth, and originality of thought, with the largest advantage observed in originality. In contrast, overall attitudes toward IUI participation in philosophical contexts (B8–B14) evolved more slowly and unevenly; although some items showed modest increases over time, evaluations remained variable, reflecting persistent ambivalence regarding the appropriateness and normative role of AI systems within philosophical practice.

5. Discussion

The qualitative findings reveal a longitudinal shift that is not merely attitudinal, but philosophical in nature. Over three years of sustained engagement, participants moved from initial resistance, through instrumental acceptance, to a principled critique of generative AI grounded in questions of understanding, embodiment, and responsibility[18]. Importantly, this trajectory does not reflect growing pessimism about the technology's performance. Rather, it reflects a deepening reflection on

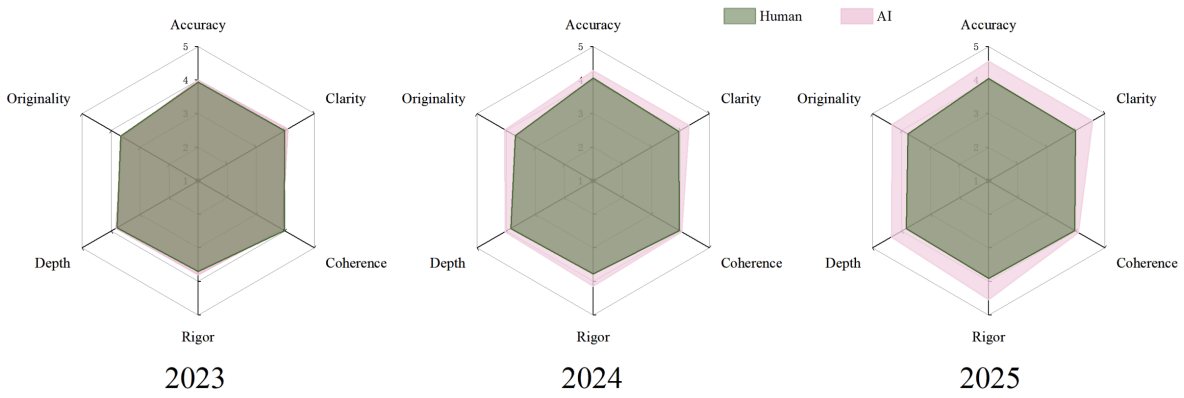


Figure 2: Mean ratings of human-authored and AI-generated philosophical responses across six evaluative dimensions for each study wave (2023–2025)

the nature of philosophical activity itself. As participants became more familiar with generative AI and increasingly recognized its formal competence, they simultaneously sharpened the boundaries they drew around what they regarded as genuinely philosophical thinking.At the same time, the quantitative findings complicate this trajectory. Although participants consistently denied generative AI philosophical standing in their interviews, blind evaluations indicate that they nonetheless attributed to AI forms of originality typically associated with human philosophical work. From a human–AI interaction perspective, this divergence reflects a failure of expectation calibration, whereby stable normative beliefs are not updated in line with evaluative performance as system capability improves.

Across interviews, participants consistently distinguished between formal adequacy and philosophical depth. While generative AI was acknowledged as capable of producing well-structured, fluent, and even impressive philosophical texts, it was repeatedly described as lacking the lived, situated, and normatively accountable engagement that participants viewed as central to philosophy. In this sense, participants' critiques were not primarily about correctness or usefulness, but about agency: who or what can bear responsibility for philosophical claims, experience the pressure of argument, and question the presuppositions within which reasoning takes place. This pattern suggests that, rather than interpreting the observed tension as an inconsistency in individual positions, it is better understood as a disciplinary evaluative structure that becomes visible under specific assessment conditions and shapes which qualities are recognized and valued in the evaluative process.

Seen in this light, generative AI was ultimately positioned by participants not as a philosophical agent, but as a reflective artifact. Several participants described AI as a "mirror" that makes visible patterns, assumptions, and limits within human philosophical practice itself. Engagement with AI-generated philosophy did not displace human philosophical authority; instead, it prompted renewed attention to those dimensions of philosophy that participants regarded as irreducibly human—such as embodied experience, value-laden judgment, historical situatedness, and the willingness to question one's own conceptual frameworks. The encounter with generative AI thus functioned less as a challenge to philosophy's legitimacy than as an occasion for its rearticulation.

Limitations Response length was not controlled in order to preserve naturalistic human and AI responses; this may have influenced perceived depth or originality in the blind evaluations.

5.1. Conclusion

In conclusion, this study shows that increasing AI performance alone does not guarantee acceptance in complex knowledge domains such as philosophy. For HCI, the findings highlight the need to move beyond output quality toward designs that respect domain values, user autonomy, and interpretive authority. Rather than replacing human judgment, generative AI is better understood as an intellectual

tool and a reflective mirror that foregrounds the enduring human dimensions of philosophical inquiry.

Declaration on Generative AI

During manuscript preparation, large language models were used only for limited editorial refinement and consistency checks. All authors take full responsibility for the content, analysis, and interpretations, and all data are authentic with no AI-generated material included in the substantive work.

References

[1] OpenAI, Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[3] J. Jiao, S. Afroogh, A. Murali, et al., Llm ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large language models, Scientific Reports 15 (2025) 34642. doi:10.1038/s41598-025-18489-7.

[4] R. Bommasani, D. A. Hudson, E. Adeli, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).

[5] M. Bohlmann, A. M. Berger, Chatgpt and the writing of philosophical essays, Teaching Philosophy 47 (2024) 233–253. doi:10.5840/teachphil202451200.

[6] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[7] M. Zafar, Normativity and AI moral agency, AI & Ethics 5 (2025) 2605–2622. doi:10.1007/s43681-024-00566-8.

[8] M. Klenk, Ethics of generative AI and manipulation: A design-oriented research agenda, Ethics and Information Technology 26 (2024). URL: https://doi.org/10.1007/s10676-024-09745-x. doi:10.1007/s10676-024-09745-x, published online 3 February 2024.

[9] B. Williams, Philosophy as a humanistic discipline, Philosophy 75 (2000) 477–496.

[10] Y. Rogers, Discussion, in: HCI Theory: Classical, Modern, and Contemporary, Synthesis Lectures on Human-Centered Informatics, Springer, Cham, 2012. URL: https://doi.org/10.1007/978-3-031-02197-8_7. doi:10.1007/978-3-031-02197-8_7.

[11] M. Zafar, Normativity and ai moral agency, AI and Ethics 5 (2025) 2605–2622. doi:10.1007/s43681-024-00566-8.

[12] G. John-Stewart, Ethics of artificial intelligence, in: J. Fieser, B. Dowden (Eds.), Internet Encyclopedia of Philosophy, 2023. URL: https://iep.utm.edu/ethics-of-artificial-intelligence/, accessed: 2026-01-04.

[13] C. Véliz, Moral zombies: why algorithms are not moral agents, AI & Society 36 (2021) 487–497. doi:10.1007/s00146-021-01189-x.

[14] S. Grassini, M. Koivisto, Artificial creativity? evaluating AI against human performance in creative interpretation of visual stimuli, International Journal of Human–Computer Interaction 41 (2025) 4037–4048. doi:10.1080/10447318.2024.2345430.

[15] F. Gilardi, M. Alizadeh, M. Kubli, ChatGPT outperforms crowd workers for text-annotation tasks, Proceedings of the National Academy of Sciences of the United States of America 120 (2023) e2305016120. doi:10.1073/pnas.2305016120.

[16] H.-P. H. Lee, A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks, N. Wilson, The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706598.3713778. doi:10.1145/3706598.3713778.
[17] M. Fischer, C. Lanquillon, Evaluation ofnbsp;generative ai-assisted software design andnbsp;engineering: A user-centered approach, in: Artificial Intelligence in HCI: 5th International Conference, AI-HCI 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29 – July 4, 2024, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2024, p. 31–47. URL: https://doi.org/10.1007/978-3-031-60606-9_3. doi:10.1007/978-3-031-60606-9_3.
[18] H. L. Dreyfus, What Computers Still Can't Do: A Critique of Artificial Reason, MIT Press, Cambridge, MA, USA, 1992.

A. Appendix A. Familiarity with Generative AI Scale

This appendix presents the six-item scale used to assess participants' familiarity with generative AI, including conceptual understanding, engagement, and self-perceived competence. All items were rated on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree).

Table 2
Familiarity with Generative AI Scale

ID	Item
A1	I understand the basic concept of generative AI (e.g., ChatGPT).
A2	I can clearly distinguish between generative AI and a traditional search engine.
A3	I frequently follow news or discussions about the development of generative AI.
A4	I have used generative AI in my studies or research.
A5	I consider myself a skilled user of generative AI.
A6	I believe I understand the basic working principles of generative AI.

B. Appendix B. Attitudes Toward IUI Participation in Philosophical Discussions

This appendix reports the attitude scale used to assess participants' perceptions of generative AI participation in philosophical discussions, covering perceived capabilities, usefulness, and human-likeness. All items were rated on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree).

Table 3
Attitude Scale on IUI Participation in Philosophy

ID	Statement
B1	GenAI can provide accurate definitions for complex philosophical concepts.
B2	GenAI can organize a logically clear and well-structured philosophical argument.
B3	GenAI can effectively identify and avoid common logical fallacies.
B4	When challenged, GenAI can defend its argument or revise it reasonably.
B5	GenAI can clearly explain core theories of major philosophers.
B6	GenAI's summaries of classic philosophical texts are comprehensive and accurate.
B7	GenAI can synthesize philosophical views to propose innovative ideas.
B8	GenAI can apply philosophical theories to real-world moral dilemmas.
B9	GenAI can appropriately use philosophical thought experiments.
B10	Overall, GenAI's philosophical answers are information-rich and substantive.
B11	Overall, GenAI's answers are intellectually stimulating and provoke deeper thinking.
B12	Without knowing the author, I would find it difficult to distinguish AI-written philosophy from human-written philosophy.
B13	Overall, GenAI can provide helpful answers to philosophical questions.
B14	Overall, GenAI's thinking and responses are similar to a human's.

C. Appendix C. Semi-Structured Interview Protocol

This appendix outlines the semi-structured interview protocol used to probe participants' views on generative AI, philosophical judgment, and disciplinary values across study cycles.

Table 4
Interview Modules and Core Prompts

Module	Core Questions and Probes
Background	Academic background, philosophical focus, prior experience with generative AI.
Basic Attitude	Views on AI understanding; metaphors for AI (e.g., tool, mirror); impact on concepts such as intelligence or consciousness.
Capability Assessment	Open-ended prompts corresponding to items B1–B9 (e.g., AI's ability to construct arguments).
Value and Limitations	Distinction between technical vs. principled limits; value of AI as correct vs. unexpected; uniqueness of human philosophy.
Overall Attitude	Reflections corresponding to items B10–B14.
Closing	Additional comments or suggestions.

D. Appendix D. Philosophical Question Test Questionnaire

This appendix describes the structure of the philosophical question tasks used in the blind evaluation, designed to probe different dimensions of philosophical competence.

Table 5
Structure of Philosophical Question Tasks

Category	Assessment Goal	Sample Question
Definition & Elucidation	Conceptual understanding	Difference between Kant's categorical imperative and utilitarianism.
Argument Construction	Logical rigor	Construct an argument for free will.
Textual Analysis & Synthesis	Historical comparison	Wu wei vs. Stoic naturalism.
Practical Application	Applied ethics	Rawls and algorithmic justice.
Dialogue & Revision	Dialectical reasoning	Defending objective beauty.
Meta-philosophical Reflection	Limits of explanation	Phenomenology of pain.

E. Appendix E. Philosophical Text Quality Assessment Rubric

This appendix presents the six-dimensional rubric used by evaluators to assess the quality of philosophical responses under blind conditions.

Table 6
Evaluation Dimensions and Rating Criteria

Dimension	Description
Accuracy	Correctness and precision of philosophical concepts.
Clarity	Comprehensibility and lucidity of expression.
Coherence	Logical flow and structural organization.
Argumentative Rigor	Strength and consistency of reasoning.
Philosophical Depth	Engagement with underlying concepts and traditions.
Originality	Degree of synthesis, creativity, or conceptual contribution.