

Engagement Intensity as a Learner-Modeling Signal for Adaptive AI Ethics Instruction

YongKyung Oh¹, Lynn Talton¹ and Alex Bui^{1,*}

¹University of California, Los Angeles (UCLA), Los Angeles, CA, USA

Abstract

Adaptive AI ethics instruction in graduate research training benefits from intake measures that reflect differences in prior LLM experience. Prior coursework or workshop attendance is an obvious candidate, but it is not clear whether it is associated with pre-instruction ratings on key AI perception items. We compare three candidate intake features, self-reported usage frequency, self-rated LLM familiarity, and prior AI education, across five baseline perception outcomes in 93 bioscience graduate and postdoctoral trainees enrolled in a required research ethics course. Usage frequency shows Holm-corrected associations with all five outcomes, self-rated familiarity with three, and prior AI education with none. A threshold-like pattern at the lower end of the scale is most visible for training interest and accuracy trust rather than appearing as a uniform gradient across all five outcomes. In a short intake survey, reported LLM use is more consistently associated with these perceptions than prior coursework or workshops, with self-rated familiarity serving as a secondary indicator. These results suggest that simple pre-instruction behavioral signals can inform lightweight intake profiling for adaptive AI ethics education.

Keywords

AI ethics education, AI literacy, Learner modeling, Adaptive instruction, Engagement intensity

1. Introduction

Graduate trainees enter AI ethics instruction with different prior experiences and different views of what LLMs can do. LLMs introduce distinct challenges in higher education [1], including graduate research. Generative AI tools are already used for research and writing support [2, 3], and many trainees use them before formal instruction. This heterogeneity creates a design problem in higher-education settings already adapting to generative AI [4, 5]. A single AI ethics module may not serve learners with very different starting points equally well. AI ethics instruction for science and engineering trainees has already been implemented in graduate settings [6]. The open question is which intake features best differentiate learners before that instruction begins.

Established AI literacy frameworks focus on learner-centered instruction for non-expert users [7], and recent meta-reviews highlight the need for more rigorous empirical work on how learners engage with AI in educational settings [4]. Building on this perspective, we evaluate three practical intake measures that can be collected before instruction begins: self-reported usage frequency, self-rated LLM familiarity, and prior AI education. We compare their ability to differentiate learners at baseline and identify which measure provides the most informative signal for instructional personalization [8].

Prior coursework and workshop attendance are intuitive proxies for readiness, but whether these proxies capture the variation that actually matters for instructional design has not been tested in this context. Recent guidance argues that responsible research use of AI depends on user judgment, oversight, and disclosure [9, 10], but how to translate those principles into intake-level instructional design remains unclear. Existing AI ethics modules for graduate science and engineering trainees show measurable gains but treat learners as a relatively homogeneous group at intake [6, 11]. Therefore, this study examines pre-instruction survey data from 93 bioscience graduate and postdoctoral trainees in a required Responsible Conduct of Research (RCR) course.

This paper makes two contributions. (1) We compare three candidate intake features for intake-level segmentation, namely usage frequency, self-rated LLM familiarity, and prior AI education. (2) We find

AI for Education Day, SIGKDD 2026, August 12, 2026, Jeju, Korea

*Corresponding author.

✉ yongkyungoh@mednet.ucla.edu (Y. Oh); ltalton@mednet.ucla.edu (L. Talton); buia@mii.ucla.edu (A. Bui)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

that usage frequency is most consistently associated with baseline AI perceptions, self-rated familiarity is a weaker but meaningful secondary axis, and prior AI education does not survive family-wise correction, indicating that behavioral and self-perception signals are more informative intake features than education-history labels alone.

2. Related Work

Generative AI raises distinct pedagogical, methodological, and ethical challenges in higher education [12]. In doctoral populations, perceived usefulness and ease of use are also strong correlates of acceptance [13]. A pre-post study of undergraduates reports shifts in cognitive, affective, and behavioral attitudes after sustained exposure to generative AI [14]. Recent survey work also treats ChatGPT perceptions and usage as measurable constructs in higher-education and postgraduate contexts [15, 16]. On the educator side, instructors in U.S. higher education report uneven and often cautious trust in generative AI, indicating that AI perceptions vary across roles and not only with individual experience [17]. Calls to embed AI literacy and ethics into higher-education curricula have grown in parallel [8].

Prior Experience and AI Perception. Familiarity and perceived usefulness shape how people evaluate and adopt AI systems [18, 12]. A cross-national study of teacher trust in AI-based educational technology found that experiential variables such as AI self-efficacy and understanding carried more explanatory weight than demographic variables such as age and gender [19]. In higher-education samples, similar patterns emerge. Trust in AI-powered educational technology correlates with demographic and academic characteristics among higher-education students [20].

Trust Calibration in Human-AI Interaction. Trust calibration refers to the correspondence between a user’s subjective trust and a system’s objective capability [21]. Reviews of human trust in AI show that trust develops through system design and interaction context [22], while a recent review of human-AI interaction notes that the field still lacks a consistent definition of appropriate trust [23]. Users’ knowledge, skills, and abilities also shape how they interpret AI outputs [24].

Adaptive Instruction and Learner Modeling. In intelligent tutoring, the choice of initial learner variable matters for adaptive instruction. VanLehn’s review of intelligent-tutoring research shows that the match between learner starting point and instructional support shapes outcomes [25]. By analogy, we compare reported AI use, LLM familiarity, and prior AI education as alternative intake variables for the same cohort. Taken together, these studies point to experience-related differences, but they do not show whether reported AI use separates pre-instruction response patterns better than prior coursework or workshops in a graduate research setting.

3. Study Design

We use “engagement intensity” as a working label for the pair of intake features, reported use and self-rated familiarity, treated as complementary indicators of technology engagement in technology acceptance and AI literacy research [26, 18]. Throughout the paper, “adaptive” refers to intake-level learner grouping based on pre-instruction signals rather than differentiated instructional materials, adaptive sequencing, personalized feedback, or a full intelligent tutoring system [25, 8].

This study was approved by the University of California, Los Angeles (UCLA) Institutional Review Board (IRB-25-1106). Participation was voluntary, consent was implied by beginning the anonymous survey, respondents could skip any item, and no identifying information was stored. We surveyed 93 bioscience graduate and postdoctoral trainees enrolled in a required Responsible Conduct of Research course and analyze only pre-instruction responses. Required RCR training is now standard in U.S. research settings, although its content and delivery vary across programs [27, 28]. The descriptive intake distributions use the full cohort ($N = 93$). Two respondents left all five focal outcome items blank, yielding an outcome-valid analytic sample of $N_{\text{valid}} = 91$ for all association tests. Data were collected with an anonymous online Qualtrics survey (UCLA Health, about 10 to 15 minutes) administered around the RCR course in a pre/post design, of which this pilot analyzes the pre-class responses only.

The sample is predominantly early-stage doctoral students drawn from more than ten bioscience sub-disciplines. The baseline composition and disciplinary breakdown, the three intake-feature distributions, and reported LLM tools and purposes are detailed in Appendix A (Tables 4 and 5).

Measures Five perception items served as focal outcomes, each on a five-point Likert scale.

- Accuracy trust. “I believe LLMs provide accurate general scientific information.”
- Distinguishing capability. “I feel capable of distinguishing between factual and incorrect information produced by LLMs.”
- Complex-task trust. “I trust LLMs for complex ethical issues or nuanced scientific concepts.”
- Critical-thinking risk. “I believe over-reliance on LLMs might impair my critical thinking skills.”
- Training interest. “I am interested in receiving formal training to effectively utilize LLMs in my future research projects.”

Each item is a single five-point Likert statement (anchored from “Strongly disagree” to “Strongly agree”) and is analyzed as a distinct facet of pre-instruction AI perception rather than pooled into a composite. No internal-consistency or factor-structure statistic is reported, because the items are treated as distinct facets. Each item corresponds to a construct examined in prior work, namely trust in factual accuracy [21, 22], self-rated ability to evaluate AI output [7], trust calibration for complex tasks [23, 24], perceived over-reliance risk [29], and interest in formal training [11, 30].

Analysis Because Likert responses are ordinal rather than interval [31], usage frequency and self-rated LLM familiarity were tested as five-level ordinal predictors using Spearman correlations [32]. Prior AI education, a single categorical item covering training in AI, computational biology, or bioinformatics, was tested with Kruskal–Wallis H tests [33] under the three-group coding.

We additionally report a collapsed two-group coding (any prior versus none, with group sizes 51 and 40). Because the Kruskal–Wallis test reduces to the Mann–Whitney U test for two groups, the same test serves both codings, and Cliff’s δ summarizes the two-group effect size. Holm correction [34] was applied across the five outcomes within each predictor family, because each family was treated as a competing intake signal rather than pooled into a single 15-test omnibus family.

4. Results

Figure 1 summarizes the overall response distributions for the five focal outcomes. The profile shows high endorsement of critical-thinking risk and training interest, moderate confidence in distinguishing capability, and substantially more cautious trust for complex-task use than for general scientific accuracy.

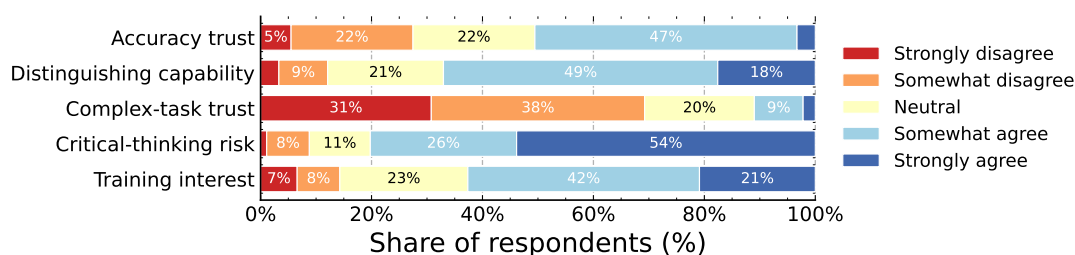


Figure 1: Response-profile distribution across the five focal perception outcomes. The stacked profile shows how agreement, neutrality, and disagreement are distributed within each outcome.

Tables 1, 2, and 3 summarize the three candidate intake features side by side. Usage frequency survives Holm correction on all five outcomes, self-rated familiarity on three, and prior AI education on none. Usage frequency shows the largest associations for complex-task trust ($\rho = .409$), accuracy trust ($\rho = .370$), and critical-thinking risk ($\rho = -.346$). Self-rated familiarity shows a weaker pattern and survives Holm correction for distinguishing capability ($\rho = .294$, $p_{\text{Holm}} = .023$), complex-task trust ($\rho = .278$, $p_{\text{Holm}} = .031$), and critical-thinking risk ($\rho = -.254$, $p_{\text{Holm}} = .045$). Accuracy trust shows a positive familiarity association but does not survive Holm correction ($p_{\text{Holm}} = .069$).

and training interest is not associated with familiarity. The familiarity pattern is bottom-anchored rather than smoothly graded. Separation is clearest at the lower end of the scale, while the medians for the “Somewhat familiar” and “Very familiar” groups are often identical. In this sample, the five pre-instruction items vary more by reported use and familiarity than by prior AI education.

Table 1

Usage-frequency associations with the five focal outcomes. For each outcome the table reports the median rating within every usage level from Never to Daily, together with the Spearman correlation ρ and its Holm-corrected p value computed across the five outcomes.

Outcome	Median by usage level					Test statistics	
	Never	Rarely	Occasionally	Frequently	Daily	ρ	p_{Holm}
Accuracy trust	2.0	3.0	3.0	4.0	4.0	.370	.001**
Distinguishing capability	3.0	4.0	4.0	4.0	4.0	.281	.014*
Complex-task trust	1.5	2.0	2.0	2.0	2.5	.409	<.001***
Critical-thinking risk	5.0	5.0	5.0	5.0	4.0	-.346	.002**
Training interest	2.0	4.0	4.0	4.0	4.0	.215	.041*
n ($N_{\text{valid}} = 91$)	10	13	21	27	20		

Table 2

Self-rated LLM-familiarity associations with the five focal outcomes. For each outcome the table reports the median rating within every familiarity level from Not at all to Very, together with the Spearman correlation ρ and its Holm-corrected p value computed across the five outcomes.

Outcome	Median by familiarity level					Test statistics	
	Not at all	Slightly	Neutral	Somewhat	Very	ρ	p_{Holm}
Accuracy trust	2.0	3.0	3.0	4.0	4.0	.222	.069
Distinguishing capability	3.0	3.0	4.0	4.0	4.0	.294	.023*
Complex-task trust	1.0	1.0	2.0	2.0	2.0	.278	.031*
Critical-thinking risk	5.0	5.0	5.0	5.0	4.0	-.254	.045*
Training interest	1.0	3.0	4.0	4.0	4.0	.042	.694
n ($N_{\text{valid}} = 91$)	1	13	5	47	25		

Table 3

Prior-education associations with the five focal outcomes under a three-group coding of formal, informal, and none and a collapsed two-group coding of any prior versus none. The table reports the group medians together with the Kruskal–Wallis H statistic, its Holm-corrected p value, and Cliff’s δ as the two-group effect size.

Outcome	Median (three-group)			Median (two-group)		Test statistics				
	Formal	Informal	None	Any prior	None	H_3	p_3	H_2	p_2	δ
Accuracy trust	4.0	4.0	3.0	4.0	3.0	2.38	.576	2.32	.524	.17
Distinguishing capability	4.0	4.0	4.0	4.0	4.0	3.80	.576	2.63	.524	.18
Complex-task trust	2.0	1.0	2.0	2.0	2.0	3.88	.576	1.12	.724	-.12
Critical-thinking risk	4.0	5.0	5.0	4.0	5.0	5.06	.397	1.37	.724	-.13
Training interest	4.0	4.0	4.0	4.0	4.0	3.56	.576	0.49	.724	.08
n ($N_{\text{valid}} = 91$)	32	19	40	51	40					

Table 3 isolates prior AI education, the one intake feature with no family-wise association. Under the three-group coding (formal / informal / none), the Kruskal–Wallis tests are non-significant for all five outcomes after Holm correction (largest $H = 5.06$), and the sample is dominated by the None category (44.1%). The null is robust to coding granularity. A collapsed two-group contrast (any prior versus none), reported in the same table, reaches the identical conclusion (minimum $p_{\text{Holm}} = .524$) with at most small effect sizes (Cliff’s $|\delta| < .19$). The largest two-group gap is a one-level median difference on accuracy trust and on critical-thinking risk, both non-significant. Prior coursework or workshop labels therefore separate baseline AI perceptions less sharply than reported use behavior or self-rated familiarity, an informative null that sharpens, rather than weakens, the intake-feature hierarchy.

5. Discussion

Usage frequency showed Holm-corrected associations on all five outcomes, self-rated familiarity on three, and prior AI education on none. Under this coarse three-category coding, workshop or course participation alone did not distinguish pre-instruction ratings as clearly as the two experience-related features. In Appendix A, Figure 2 shows this pattern most clearly. The early threshold-like pattern is strongest for training interest and accuracy trust, but it is not uniform across all five outcomes. A practical distinction in introductory design may therefore lie between learners who have not used AI tools and those with any sustained use, while still recognizing that some outcomes, such as critical-thinking risk, shift mainly among daily users.

The negative association between usage and critical-thinking risk ($\rho = -.346$) adds a calibration concern. Systematic review evidence suggests that over-reliance on AI dialogue systems can affect students' cognitive performance [29]. In the present sample, heavy users reported less worry about over-reliance, which could reflect appropriate confidence or reduced sensitivity to failure modes that grow harder to detect in more capable models [35]. The familiarity result points in the same direction. Learners who felt more familiar with LLMs also reported lower concern about critical-thinking risk ($\rho = -.254$). Taken together, these results raise a calibration concern among more engaged learners, who may need exercises that focus on when LLM outputs fail.

Why behavioral engagement tracks these perceptions is not established here, but two mechanisms are plausible. Repeated interaction may recalibrate accuracy trust as users encounter system failures [21, 22], and for the over-reliance item the negative association could instead reflect growing confidence or habituation rather than heightened concern [29, 35]. Hands-on use may also increase self-rated evaluation competence rather than measured skill. We treat these as candidate accounts to be tested longitudinally rather than as established pathways.

For intake profiling, the current data support using reported usage frequency as a first intake variable and self-rated familiarity as a secondary check. If a course uses simple routing, the present results are more consistent with a split between non-users and sustained users than with a split based on prior coursework or workshops. Heavy users may also need additional calibration-focused activities. Furthermore, we discuss alternative non-causal accounts of these associations in Appendix B.

6. Conclusion

Understanding what learners bring into AI ethics instruction has become a pressing concern as generative AI use spreads across graduate research. Among the three intake features examined, usage frequency shows the strongest association with the five pre-instruction perception items, self-rated familiarity shows a weaker pattern, and prior AI education does not reach significance after correction. A practical implication is that short pre-instruction surveys can use simple behavioral and self-perception signals to inform intake profiling, with reported LLM use indicating engagement level and self-rated familiarity serving as a secondary check on perceived readiness. In heterogeneous graduate cohorts, intake analytics could, in principle, help distinguish learners who need basic orientation from those who benefit from calibration-focused work on trust, over-reliance, and failure modes.

Limitations. This study is cross-sectional, associational, and drawn from a single institution. The single-respondent “Not at all familiar” group means the familiarity ladder should be interpreted descriptively and not overclaimed. Both predictors and outcomes are single-item self-report measures, so common-method variance cannot be excluded and multi-item reliability is unavailable. Other engagement dimensions remain outside the present analysis. Whether differentiated instruction based on these profiles improves learning outcomes requires an intervention study.

Future Work. Future work should test whether this hierarchy of intake features replicates across institutions, disciplines, and course formats. The engagement construct could also be broadened with behavioral measures such as session logs or task-level interaction data. Randomized studies could then evaluate whether engagement-based grouping improves calibration, learning, or transfer.

Acknowledgments

This study was approved by the University of California, Los Angeles (UCLA) Institutional Review Board (IRB-25-1106). We thank the trainees who participated in the Spring 2025 Responsible Conduct of Research (RCR) course survey.

Declaration on Generative AI

During the preparation of this work, the authors used generative AI tools for grammar and spelling checking and writing-style improvement. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of this publication.

References

- [1] S. Milano, J. A. McGrane, S. Leonelli, Large language models challenge the future of higher education, *Nature Machine Intelligence* 5 (2023) 333–334. doi:10.1038/s42256-023-00644-2.
- [2] J. Roberts, M. Baker, J. Andrew, Artificial intelligence and qualitative research: The promise and perils of large language model (LLM) ‘assistance’, *Critical Perspectives on Accounting* 99 (2024) 102722. doi:10.1016/j.cpa.2024.102722.
- [3] J. S. Barrot, Using ChatGPT for second language writing: Pitfalls and potentials, *Assessing Writing* 57 (2023) 100745. doi:10.1016/j.asw.2023.100745.
- [4] M. Bond, H. Khosravi, M. De Laat, N. Bergdahl, V. Negrea, E. Oxley, P. Pham, S. W. Chong, G. Siemens, A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour, *International Journal of Educational Technology in Higher Education* 21 (2024) 4. doi:10.1186/s41239-023-00436-z.
- [5] N. McDonald, A. Johri, A. Ali, A. H. Collier, Generative artificial intelligence in higher education: Evidence from an analysis of institutional policies and guidelines, *Computers in Human Behavior: Artificial Humans* 3 (2025) 100121. doi:10.1016/j.chbah.2025.100121.
- [6] M. Usher, M. Barak, Unpacking the role of AI ethics online education for science and engineering students, *International Journal of STEM Education* 11 (2024) 35. doi:10.1186/s40594-024-00493-4.
- [7] D. Long, B. Magerko, What is AI Literacy? Competencies and Design Considerations, in: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–16. doi:10.1145/3313831.3376727.
- [8] W. Xing, N. Nixon, S. Crossley, P. Denny, A. Lan, J. Stamper, Z. Yu, The Use of Large Language Models in Education, *International Journal of Artificial Intelligence in Education* 35 (2025) 439–443. doi:10.1007/s40593-025-00457-x.
- [9] D. B. Resnik, M. Hosseini, The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool, *AI and Ethics* 5 (2025) 1499–1521. doi:10.1007/s43681-024-00493-8.
- [10] M. Hosseini, D. B. Resnik, K. Holmes, The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts, *Research Ethics* 19 (2023) 449–465. doi:10.1177/17470161231180449.
- [11] J. Borenstein, A. Howard, Emerging challenges in AI and the need for AI ethics education, *AI and Ethics* 1 (2021) 61–65. doi:10.1007/s43681-020-00002-7.
- [12] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günemann, E. Hüllermeier, et al., ChatGPT for good? On opportunities and challenges of large language models for education, *Learning and Individual Differences* 103 (2023) 102274. doi:10.1016/j.lindif.2023.102274.
- [13] M. Zou, L. Huang, To use or not to use? Understanding doctoral students’ acceptance of ChatGPT in writing through technology acceptance model, *Frontiers in Psychology Volume 14 - 2023* (2023). doi:10.3389/fpsyg.2023.1259531.
- [14] W. Suh, Generative AI integration in higher education shifts students’ attitudes from tool use to innovation, *Discover Education* 4 (2025) 508. doi:10.1007/s44217-025-00914-8.
- [15] D. Ravšelj, D. Keržič, N. Tomažević, L. Umek, N. Brezovar, N. A. Iahad, A. A. Abdulla, A. Akopyan, M. W. Aldana Segura, J. AlHumaid, et al., Higher education students’ perceptions of ChatGPT: A global study of early reactions, *PLOS ONE* 20 (2025) e0315011. doi:10.1371/journal.pone.0315011.
- [16] M. Nemt-allah, W. Khalifa, M. Badawy, Y. Elbably, A. Ibrahim, Validating the ChatGPT Usage Scale: psychometric properties and factor structures among postgraduate students, *BMC Psychology* 12 (2024) 497. doi:10.1186/s40359-024-01983-4.

- [17] W. Lyu, S. Zhang, T. Chung, Y. Sun, Y. Zhang, Understanding the practices, perceptions, and (dis)trust of generative AI among instructors: A mixed-methods study in the U.S. higher education, *Computers and Education: Artificial Intelligence* 8 (2025) 100383. doi:10.1016/j.caeai.2025.100383.
- [18] S. Kelly, S.-A. Kaye, O. Oviedo-Trespalacios, What factors contribute to the acceptance of artificial intelligence? A systematic review, *Telematics and Informatics* 77 (2023) 101925. doi:10.1016/j.tele.2022.101925.
- [19] O. Viberg, M. Cukurova, Y. Feldman-Maggor, G. Alexandron, S. Shirai, S. Kanemune, B. Wasson, C. Tømte, D. Spikol, M. Milrad, et al., What Explains Teachers' Trust in AI in Education Across Six Countries?, *International Journal of Artificial Intelligence in Education* 35 (2025) 1288–1316. doi:10.1007/s40593-024-00433-x.
- [20] T. Nazaretsky, P. Mejia-Domenzain, V. Swamy, J. Frej, T. Käser, The critical role of trust in adopting AI-powered educational technology for learning: An instrument for measuring student perceptions, *Computers and Education: Artificial Intelligence* 8 (2025) 100368. doi:10.1016/j.caeai.2025.100368.
- [21] J. D. Lee, K. A. See, Trust in Automation: Designing for Appropriate Reliance, *Human Factors* 46 (2004) 50–80. doi:10.1518/hfes.46.1.50_30392.
- [22] E. Glikson, A. W. Woolley, Human Trust in Artificial Intelligence: Review of Empirical Research, *Academy of Management Annals* 14 (2020) 627–660. doi:10.5465/annals.2018.0057.
- [23] S. Mehrotra, C. Degachi, O. Vereschak, C. M. Jonker, M. L. Tielman, A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges, *ACM Journal on Responsible Computing* 1 (2024) 26:1–26:45. doi:10.1145/3696449.
- [24] G. M. Alarcon, A. Capiola, Explicating the trust process for effective human interaction with artificial intelligence and machine learning systems, *Frontiers in Computer Science Volume 7 - 2025* (2025). doi:10.3389/fcomp.2025.1662185.
- [25] K. VanLEHN, The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems, *Educational Psychologist* 46 (2011) 197–221. doi:10.1080/00461520.2011.611369.
- [26] V. Venkatesh, M. G. Morris, G. B. Davis, F. D. Davis, User Acceptance of Information Technology: Toward A Unified View, *Management Information Systems Quarterly* 27 (2003) 425–478. doi:10.2307/30036540.
- [27] N. H. Steneck, R. E. Bulger, The History, Purpose, and Future of Instruction in the Responsible Conduct of Research, *Academic Medicine* 82 (2007) 829–834. doi:10.1097/ACM.0b013e31812f7d4d.
- [28] M. S. Anderson, A. S. Horn, K. R. Risbey, E. A. Ronning, R. De Vries, B. C. Martinson, What Do Mentoring and Training in the Responsible Conduct of Research Have To Do with Scientists' Misbehavior? Findings from a National Survey of NIH-Funded Scientists, *Academic Medicine* 82 (2007) 853–860. doi:10.1097/ACM.0b013e31812f764c.
- [29] C. Zhai, S. Wibowo, L. D. Li, The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review, *Smart Learning Environments* 11 (2024) 28. doi:10.1186/s40561-024-00316-7.
- [30] L. McCoy, N. Ganesan, V. Rajagopalan, D. McKell, D. F. Niño, M. C. Swaim, A Training Needs Analysis for AI and Generative AI in Medical Education: Perspectives of Faculty and Students, *Journal of Medical Education and Curricular Development* 12 (2025) 23821205251339226. doi:10.1177/23821205251339226.
- [31] S. Jamieson, Likert scales: how to (ab)use them, *Medical Education* 38 (2004) 1217–1218. doi:10.1111/j.1365-2929.2004.02012.x.
- [32] C. Spearman, The Proof and Measurement of Association between Two Things, *The American Journal of Psychology* 15 (1904) 72. doi:10.2307/1412159.
- [33] W. H. Kruskal, W. A. Wallis, Use of Ranks in One-Criterion Variance Analysis, *Journal of the American Statistical Association* 47 (1952) 583–621. doi:10.1080/01621459.1952.10483441.
- [34] S. Holm, A Simple Sequentially Rejective Multiple Test Procedure, *Scandinavian Journal of Statistics* 6 (1979) 65–70.
- [35] L. Zhou, W. Schellaert, F. Martínez-Plumed, Y. Moros-Daval, C. Ferri, J. Hernández-Orallo, Larger and more instructable language models become less reliable, *Nature* 634 (2024) 61–68. doi:10.1038/s41586-024-07930-y.
- [36] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, N. P. Podsakoff, Common method biases in behavioral research: A critical review of the literature and recommended remedies., *Journal of Applied Psychology* 88 (2003) 879–903. doi:10.1037/0021-9010.88.5.879.

A. Sample Composition and Supporting Visualizations

Table 4 reports the baseline composition. The sample is predominantly early-stage doctoral students (77.4%), with a modal age of 25 to 29 (52.7%) and a slight female majority (53.8%). Participants span more than ten bioscience subdisciplines, with neuroscience and molecular biology the largest at 16.1% each and no field exceeding 17.0%, so no single subfield drives the results.

Table 4

Baseline sample composition for the full cohort of 93 trainees ($N = 93$).

Demographics	n (%)	Discipline	n (%)
<i>Academic stage</i>		Molecular Biology	15 (16.1)
PhD student, year 1–2	72 (77.4)	Neuroscience	15 (16.1)
PhD student, year 3 or above	7 (7.5)	Genetics & Genomics	11 (11.8)
Postdoctoral researcher	6 (6.5)	Cell & Developmental Biology	9 (9.7)
Dual-degree student (MD/PhD)	6 (6.5)	Physics & Biology in Medicine	8 (8.6)
Other trainee	2 (2.2)	Biochemistry	7 (7.5)
<i>Gender</i>		Microbiology & Immunology	7 (7.5)
Female	50 (53.8)	Pharmacology	7 (7.5)
Male	39 (41.9)	Bioinformatics	4 (4.3)
Prefer not to say	4 (4.3)	Physiology	4 (4.3)
<i>Age</i>		Biomathematics	1 (1.1)
18–24	32 (34.4)	Chemistry	1 (1.1)
25–29	49 (52.7)	Environmental & Molecular Toxicology	1 (1.1)
30–34	7 (7.5)	Medical Informatics	1 (1.1)
35–39	3 (3.2)	Psychology	1 (1.1)
40 or older	2 (2.2)	Structural Biology	1 (1.1)

Table 5 reports the generative-AI engagement profile. The three intake features differ in shape, which bears on their resolution as predictors. Reported usage spreads fairly evenly across the five levels, self-rated familiarity concentrates at the upper end (about 79.6% Somewhat or Very familiar, with a single Not-at-all respondent), and prior AI education is dominated by the None category (44.1%). Reported tool use is dominated by ChatGPT (86.0%), and the most common purposes are searching for scientific information (52.7%) and summarizing literature (40.9%).

Table 5

Generative-AI engagement profile ($N = 93$). Tools used and purposes permit multiple responses (percentages of $N = 93$). Usage, familiarity, and education are single-response.

LLM usage	n (%)	AI background	n (%)
<i>Tools used</i>		<i>Usage frequency</i>	
ChatGPT	80 (86.0)	Never	10 (10.8)
Google Gemini	21 (22.6)	Rarely	13 (14.0)
Microsoft Copilot	12 (12.9)	Occasionally	22 (23.7)
Perplexity	11 (11.8)	Frequently	27 (29.0)
Claude	3 (3.2)	Daily	21 (22.6)
Other	2 (2.2)	<i>LLM familiarity</i>	
<i>Purposes for use</i>		Not at all familiar	1 (1.1)
Search for scientific information	49 (52.7)	Slightly familiar	13 (14.0)
Summarize literature/readings	38 (40.9)	Neutral / Not sure	5 (5.4)
Understand complex concepts	36 (38.7)	Somewhat familiar	48 (51.6)
Assist with writing papers	24 (25.8)	Very familiar	26 (28.0)
Identify knowledge gaps	23 (24.7)	<i>Prior AI education</i>	
Translate technical texts	19 (20.4)	Formal	33 (35.5)
Generate ideas/questions	17 (18.3)	Informal	19 (20.4)
Create outlines/drafts	11 (11.8)	None	41 (44.1)

Figure 2 visualizes the gradients described in Section 4. Usage frequency shows the strongest pattern, particularly for training interest and accuracy trust. Familiarity shows weaker gradients for evaluation capability, complex-task trust, and critical-thinking risk, although the “Not at all familiar” group contains only one respondent. Prior AI education shows little separation under either coding scheme, consistent with its null result. Overall, the patterns distinguish non-users from learners with sustained engagement but do not support a uniform binary threshold across outcomes.

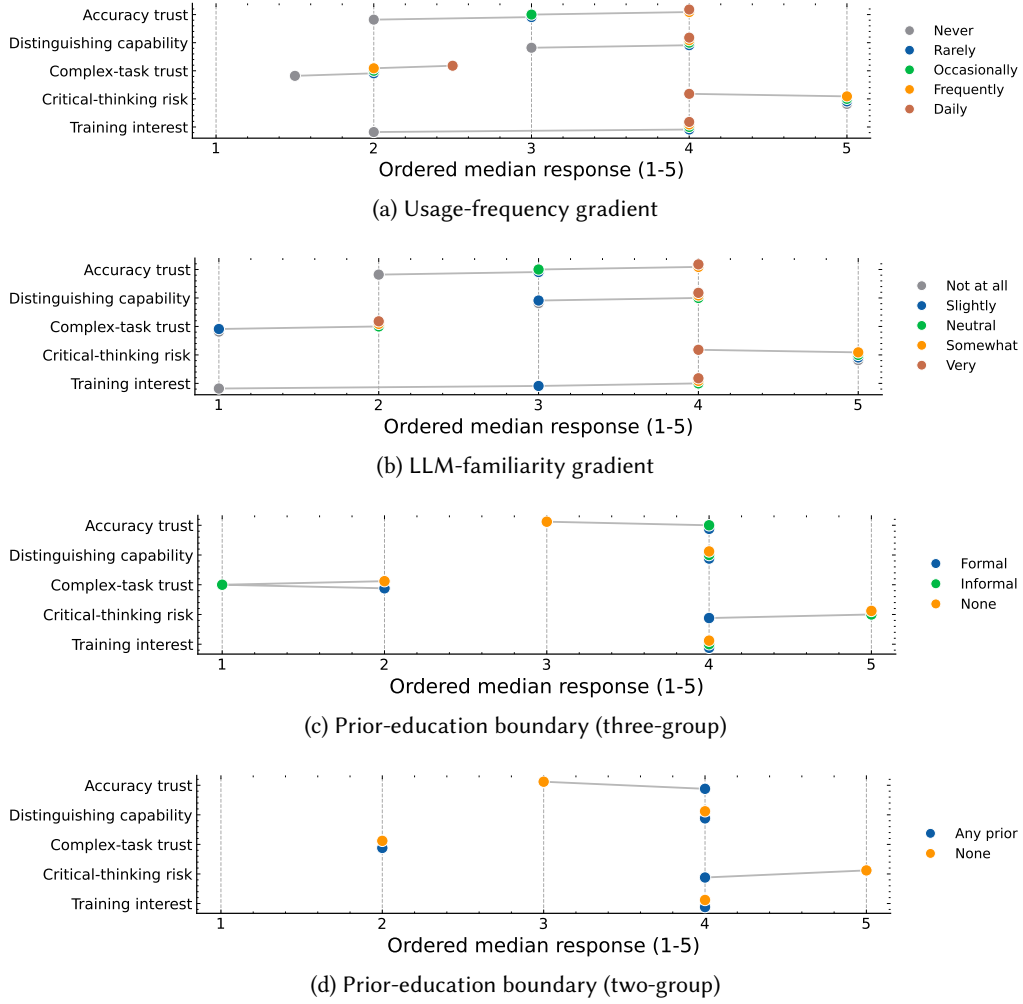


Figure 2: Ordered-median ladders for accuracy trust, distinguishing capability, complex-task trust, critical-thinking risk, and training interest across the candidate intake features. Panels (c) and (d) show prior AI education under the three-group (formal / informal / none) and two-group (any prior / none) codings.

B. Alternative Interpretations

The study design does not establish directionality. Engagement may influence perceptions, perceptions may influence engagement, or both may reflect upstream factors such as disciplinary norms or dispositional openness [26, 22]. The usage measure also captures self-reported frequency on a Never-to-Daily scale rather than observed behavior, making it an indirect measure of engagement [16].

Part of the observed associations may reflect general response styles, such as acquiescence or extremity bias, rather than differences in knowledge or readiness, because both predictors and outcomes were measured through self-report [36]. The cross-sectional design cannot distinguish between these possibilities. The five outcomes were treated as facets of distinct constructs documented in prior work on trust calibration and AI literacy, namely accuracy trust, evaluation capability, complex-task trust, over-reliance risk, and training interest, rather than as indicators of a single latent dimension [21].

The null result for prior AI education should be interpreted cautiously. AI literacy spans multiple competencies [7], so the three-category coding (Formal, Informal, None) cannot capture the depth, recency, or duration of training and is coarser than the five-level usage and familiarity scales. However, a higher-powered two-group comparison (any prior education versus none) produced the same null result, with at most small effects. This pattern suggests that the limitation lies in the measurement of prior AI education rather than in the construct itself. Future measures should capture the depth, recency, and duration of training.