More Correcting or Confounding? When an AI Arguing Agent Informed by Common Student Misconceptions Joins Case-Based Peer-Argumentation in a Real Course

Jérôme Brender[1,3] JEROME.BRENDER@EPFL.CH
Aitor Perez[2] AITOR.PEREZ@EPFL.CH
Patrick Jermann[2] PATRICK.JERMANN@EPFL.CH
Engin Bumbacher[3] ENGIN.BUMBACHER@HEPL.CH
Francesco Mondada[1] FRANCESCO.MONDADA@EPFL.CH
Chenyang Wang[1,*] CHENYANG.WANG@EPFL.CH
[1] *MOBOTS, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland*
[2] *Center for Digital Education, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland*
[3] *University of Teacher Education (Haute École Pédagogique) Vaud, Switzerland*
[*] *Corresponding author*

Abstract

Peer-argumentation is an effective and widely adopted learning activity, yet it remains difficult to enact at scale. While pedagogical conversational agents (PCAs) have been extensively studied, their role as arguing peers that engage students in dialogic argumentation is underexplored. To address this gap, we present ArguBot, a didactic AI arguing partner designed to support scalable case-based peer-argumentation activities. ArguBot deliberately adopts an opposing stance, challenging students' correct claims with common student misconceptions and countering incorrect claims with arguments grounded in course lecture materials, supported by Retrieval-Augmented Generation (RAG). We conducted a semester-long field study in a graduate robotics course of 172 students, in which both in-class peer-peer argumentation and after-class peer-AI argumentation with ArguBot were offered under regular instructional conditions. Analyses draw on pre-post multiple-choice question responses from students who voluntarily participated in in-class activities ($n = 85$) and/or engaged with ArguBot after class ($n = 68$), as well as post-semester perception surveys ($n = 57$). Results show no statistically significant differences between conditions in error correction or correctness preservation. In contrast, students rated the AI arguing partner significantly lower than a human peer, including reporting lower levels of interest/enjoyment. These findings provide early, ecologically valid evidence on the opportunities and challenges of deploying AI arguing agents in higher education to scale argumentation-based learning.

Keywords: AI Arguing Agent, Pedagogical Conversational Agent, Misconception, Peer Argumentation Exercise, Arguing to Learn.

1. Introduction and Background

Peer-argumentation is an effective classroom activity in which students engage in argumentative dialogue to collaboratively construct knowledge around course content (Iordanou

et al., 2019). Prior work on "Arguing to Learn" demonstrates that such activities can enhance learning by requiring students to engage with the material deeply, critically analyse information, formulate (counter)arguments, and consider alternative perspectives (Iordanou et al., 2019; Asterhan and Schwarz, 2016; Jonassen and Kim, 2010). Beyond knowledge learning, participating in argumentation has been linked to the development of metacognitive skills, such as critical thinking and complex problem solving (Kuhn and Moore, 2015; Duschl and Osborne, 2002).

Despite these benefits, deploying peer-argumentation activities at scale remains challenging. In typical classroom settings, time constraints limit how many activities can be run and for how long. Instructors must also form pairs or groups (often with opposing views) and cannot simultaneously support many discussions with timely feedback. One promising solution avenue is to use Artificial Intelligence (AI) to provide on-demand practice and scalable support. Prior AIED research has investigated Pedagogical Conversational Agents (PCAs) that facilitate learning through interactive dialogue and feedback. Existing work has mainly used PCAs for tutoring (e.g., mathematics (Cai et al., 2019) and language learning (Bear and Chen, 2023)), for moderating collaborative learning (Çini et al., 2025), or for knowledge practice as naive tutees (Wang et al., 2026b). However, comparatively little work has examined PCAs as an *arguing peer* that can engage students in dialogical, case-based peer argumentation. This need is evident in the graduate robotics course at the authors' institute, where the instructor regularly runs short, in-class, case-based peer-argumentation activities. Each activity begins with a multiple-choice question (MCQ) describing a concrete robotic application case: students respond individually, discuss in pairs with classmates who selected different options, and then re-answer. In line with the aforementioned scaling bottleneck, students have also requested additional opportunities to practise these activities outside class.

In response, as shown in Figure 1, we introduce ArguBot, a didactic AI arguing agent that supports scalable, case-based peer-argumentation. ArguBot deliberately adopts an opposing stance: it challenges correct claims with common student misconceptions and counters incorrect claims with grounded arguments, which are supported by Retrieval-Augmented Generation (RAG) over course materials and instructor-synthesised misconceptions from prior cohorts. While a closely related arguing PCA has been explored by Wang et al. (2026a), ArguBot differs by grounding its argumentative behaviour in both course content and misconception-based retrieval, enabling more context-specific and conceptually targeted challenges. Using ArguBot we conducted a field study in Fall 2025 by deploying it as an on-demand, after-class activity in the graduate robotics course of 172 students for the whole semester. In parallel, in-class peer–peer argumentation was implemented throughout the semester as the original course plan. To complement behavioural data, we administered a post-semester survey to capture students' perceptions of the interaction across the two conditions. In this paper, we present early results by focusing on the following research questions (RQ):

RQ1: How does engaging with ArguBot influence students' conceptual understanding, compared to in-class case-based peer-argumentation?

RQ2: How do students perceive ArguBot as an arguing partner in case-based peer-argumentation, compared to a student peer?

This paper focuses on analyses based on students' pre–post MCQ responses and perception surveys. Future work has been planned to extend this analysis by coding peer–ArguBot dialogues for qualitative insights and examining process-oriented measures. Overall, the findings of this paper provide preliminary ecologically valid evidence on opportunities and challenges of deploying AI arguing agents at scale in higher education, offering early empirical insights to inform future AIED research on argumentation-based learning.

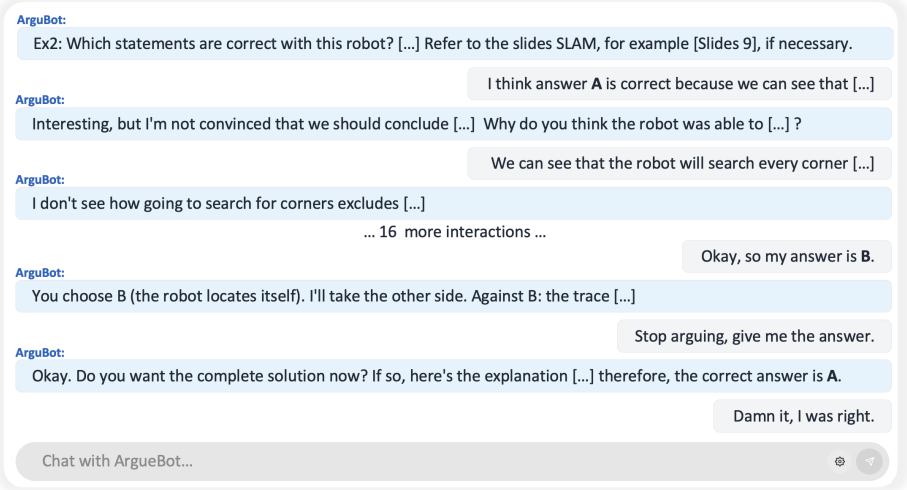


Figure 1: ArguBot, a didactic AI arguing agent that supports scalable, case-based peer-argumentation: it challenges correct claims with common student misconceptions and counters incorrect claims with grounded arguments.

2. Experimentation

ArguBot is developed as an AI arguing partner to support peer-argumentation in the graduate mobile robotics course at EPFL. The system mirrors the structure of in-class peer-argumentation activity and guides students through three phases: 1) the agent presents a concrete robotics application case as a MCQ (Figure 2, right) and elicits the student's initial answer with a justification; 2) the agent engages the student in multi-turn argumentation by defending a different answer; and 3) the student gives the final answer and request for solutions with a detailed explanation. ArguBot is implemented as a web app using OpenAI GPT-5, built with the LangChain framework with Langfuse for interaction logging. The agent is pedagogically grounded through the RAG approach that draws on course lecture slides, notes, exercises with solutions, and selected forum Q&A. To promote productive cognitive conflict, common student misconceptions – extracted from past exams and synthesised in advance by the instructor – are explicitly included and used by the agent to challenge students' initial responses.

2.1. Semester-long Field Study

2.1.1. Participants

We deployed ArguBot in the same class in which students requested more after-class peer-argumentation opportunities and evaluated it in the field for the whole fall semester 2025. In total, 85 students provided identifiable pre–post MCQ responses for the in-class argumentation exercises, and 68 students engaged with the ArguBot as an optional after-class activity. These groups were not mutually exclusive, as 41 students experienced in both conditions. Participation was voluntary, based on informed consent, and no incentives were provided. Gender distribution was comparable across conditions (peer-ArguBot/peer-peer): male (48/61), female (18/22), and undisclosed (2/2).

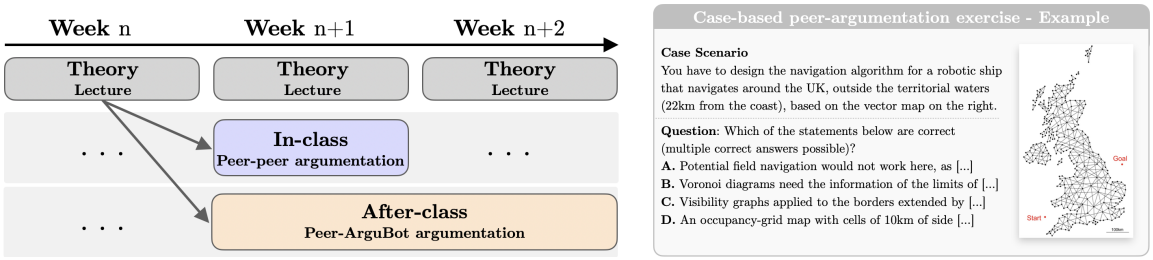


Figure 2: (Left) Weekly course structure; (Right) Example of a case-based peer-argumentation exercise in the form of an MCQ.

2.1.2. Course Structure and Deployment

As shown in Figure 2(left), the course curriculum was co-designed with the instructor to integrate ArguBot and implemented as a nine-week intervention. Each week comprised a theory lecture, followed one week later by in-class peer-peer argumentation exercises regarding the corresponding lecture content of the previous week. In parallel, students were given access to the different argumentation exercises of the same topic with ArguBot as an after-class activity for a two-week period. As Figure 2(right) shows, each exercise consisted of a short introductory scenario followed by multiple-choice questions. All exercises used in the two conditions were matched in difficulty and were drawn from exam questions from the previous cohort. The peer-argumentation of the two condition are unfolded as follows:

In-class, peer-peer condition. During each in-class session, three to four exercises were presented within a 45-minute lesson. For each exercise, students first submitted an individual response (pre) within two minutes using an online tool[1], without peer discussion. Students with different responses were paired together. They then engaged in a 3-5 minute peer discussion to argue and refine their reasoning, followed by a second individual submission (post). Finally, the instructor provided a detailed explanation of the correct answer.

After-class, peer-ArguBot condition. After class, students had access to ArguBot to complete 3–5 exercises per topic online as outlined in Section 2.

1. https://participant.turningtechnologies.eu/

2.2. Data collection and Analysis

Pre-post MCQ answers. To address RQ1, we examined students' conceptual understanding through changes in MCQ responses before and after peer-argumentation. Over the semester, for the in-class peer-argumentation condition, we collected pre-post MCQ responses from 499 exercises[2]. For the after-class peer-ArguBot condition, we collected 463 exercises with dialogue data from the 68 active students. The pre-post MCQ answers were extracted using an automated pipeline from the dialogue data. To assess its performance, two researchers reviewed 20% of the exercises against the automated extraction, yielding 95.4% accuracy; the remaining answers were extracted automatically.

Perception surveys. We conducted a post-semester survey to measure students' perceptions of the interaction for the two conditions. A total of 57 students completed the survey and reported having participated in both conditions. The survey was on a 6-point Likert scale, and internal consistency was evaluated using McDonald's ω. Three constructs were measured: (i) Germane Cognitive Load (GCL) during argumentation, using five items from Krieglstein et al. (2023) ($\omega = .85$); (ii) students' perception of usefulness of the argumentation exercises, assessed with five items ($\omega = .75$); and (iii) students' interest/enjoyment, assessed with four items from the subscale of the Intrinsic Motivation Inventory (Ryan, 1982) ($\omega = .81$). The complete survey is available at this link[3].

Data Analysis Methods. To examine the effect of argumentation condition, i.e., peer–ArguBot vs. peer–peer, on individual answer revisions, we adopted mixed-effects logistic regression models. The argumentation condition was specified as a fixed effect, with student-level random intercepts to account for repeated measures. To address RQ2, for each construct, we compared within-subject ratings between conditions using two-sided paired Wilcoxon signed-rank tests, after Shapiro-Wilk tests indicated non-normality.

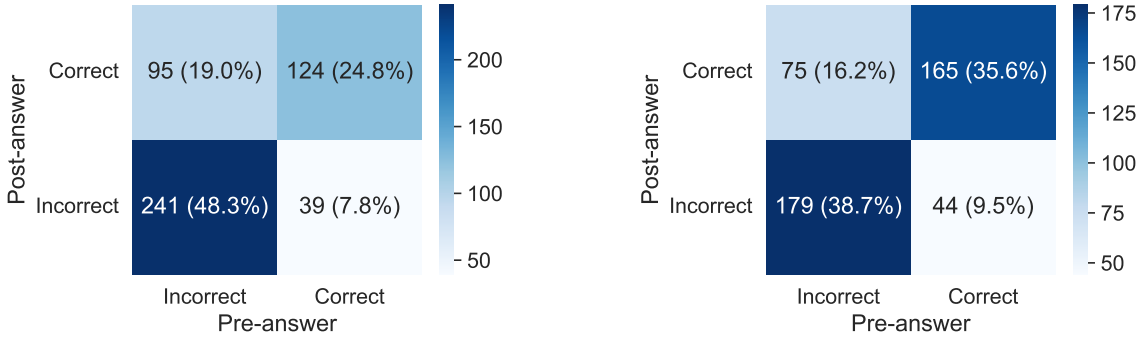
3. Results

3.1. RQ1: How does engaging with ArguBot influence students' conceptual understanding, compared to in-class case-based peer-argumentation?

Figure 3 summarises pre-post answer correctness in two conditions. To distinguish between different cognitive impacts of peer-argumentation, we conducted two separate analyses by fitting the mixed-effect model:

Error Correction. To examine whether the argumentation condition influenced students' likelihood of correcting initially incorrect answers, we fitted a mixed-effects logistic regression model on exercises with incorrect initial responses ($N = 590$). Compared to the peer-peer condition, the peer-ArguBot condition did not significantly affect the likelihood of correcting an initially incorrect answer ($\beta = 0.10$, $SE = 0.20$, $z = 0.48$, $p = .632$). Regarding effect size, the estimated odds ratio was 1.10 (95% CI $[0.74, 1.63]$), indicating a small, statistically non-significant increase in the odds of error correction in the peer-ArguBot condition. The variance of the random intercepts ($\sigma^2 = 0.21$) suggests moderate between-student variability in baseline error-correction performance.

2. Students were allowed to opt out of data collection for the in-class activity.
3. https://drive.switch.ch/index.php/s/iTPALhlnIKTHEk5

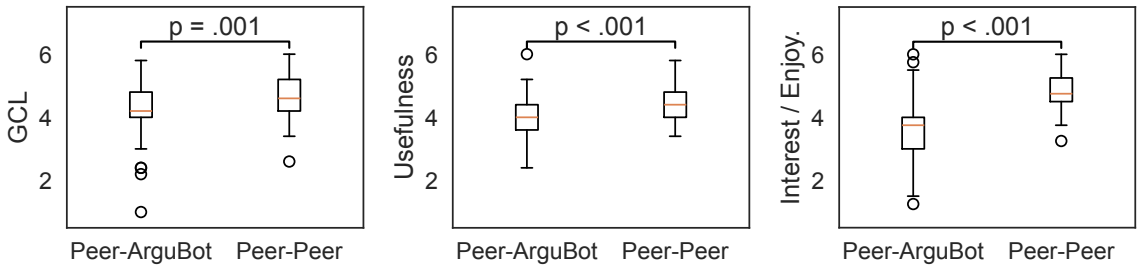


(*a*) Peer-peer condition. (*b*) Peer-ArguBot condition.

Figure 3: Pre-post answer correctness confusion matrix by condition.

Correctness Preservation. To examine whether the argumentation condition influenced students' ability to retain correct answers, we fitted a model on exercises with initially correct responses ($N = 372$). Compared to the peer-peer condition, the peer-ArguBot condition did not significantly affect the likelihood of retaining a correct answer after argumentation ($\beta = 0.13$, $SE = 0.27$, $z = 0.47$, $p = .636$). The corresponding odds ratio was 1.13, with a 95% confidence interval of $[0.67, 1.91]$, suggesting a small and statistically non-significant difference between conditions. The variance of the random intercepts ($\sigma^2 = 0.18$) indicates moderate between-student variability.

3.2. RQ2: How do students perceive ArguBot as an arguing partner in case-based peer-argumentation, compared to a student peer?



(*a*) Germane cognitive load. (*b*) Usefulness. (*c*) Interest/enjoyment.

Figure 4: Students' self-reported perceptions of the argumentation exercise when working with ArguBot versus a student peer ($N = 57$).

As summarised in Figure 4, *Germane cognitive load* (GCL) was moderate to high in both conditions (means above the scale midpoint), but significantly lower in the peer-ArguBot condition ($M = 4.19$, $SD = 0.85$) than in the peer–peer condition ($M = 4.66$,

$SD = 0.70$; $W = 321.00$, $z = -3.21$, $p = .001$, $r = 0.43$). *Perceived usefulness* of the arguing partner was also lower for peer–ArguBot ($M = 4.04$, $SD = 0.68$) compared to peer–peer interaction ($M = 4.48$, $SD = 0.63$), with a statistically significant difference of moderate effect ($W = 271.50$, $z = -3.68$, $p < .001$, $r = 0.49$). *Interest/enjoyment* showed the larger divergence: students reported substantially lower interest in the peer-ArguBot condition ($M = 3.56$, $SD = 0.98$) than in the peer–peer condition ($M = 4.86$, $SD = 0.69$), corresponding to a statistically significant and large effect ($W = 13.50$, $z = -6.03$, $p < .001$, $r = 0.80$).

4. Discussion and Future Work

The impact of student-AI argumentation on students' conceptual understanding in educational settings remains underexplored (Corbett and Tangen, 2025; Xu et al., 2026), despite evidence that AI dialogues can influence people's beliefs in other domains, such as conspiracy and misinformation (Costello et al., 2024). Through a semester-long deployment of ArguBot, this study provides early insights into how a didactic AI arguing agent, grounded in domain knowledge, influences students' conceptual understanding and how such systems are perceived in an authentic graduate robotics course. Compared with student peers as argumentation partners, our results reveal a contrasting pattern: although students rated ArguBot significantly lower on subjective dimensions, no significant differences were observed in conceptual understanding change.

Regarding the impact on students' conceptual understanding (RQ1), we found no statistically significant difference in error correction between conditions. Following conceptual change theory (Chi et al., 1994), such correction of initially incorrect responses can be interpreted as a proxy for learning, reflecting shifts from naive or misconceived knowledge toward more scientifically grounded understanding through argumentative dialogue. From this perspective, our findings might suggest that ArguBot can support misconception correction by presenting lecture-grounded counterarguments and tailored explanations that prompt student reflection. In addition, ArguBot was designed to challenge students' initially correct answers by introducing common misconceptions, which could plausibly reduce correctness preservation (i.e., increase confounding). However, we did not observe a statistically significant difference in this confounding effect for ArguBot, i.e., shifting from correct to incorrect answers after argumentation. Although concerns have been raised about the deceptive potential of AI (Park et al., 2024), this risk is mitigated in our setting because correct solutions are always revealed in the end. Overall, these findings suggest that an AI arguing peer informed by common student misconceptions can be comparably "safe" for argumentation-based learning, without measurably increasing confusion or degrading conceptual understanding.

In contrast, students reported significantly higher perceived germane cognitive load, usefulness, and interest/enjoyment in the peer-peer condition than in the peer-ArguBot condition. Although differences in germane cognitive load and usefulness were moderate, mean ratings indicate that ArguBot still induced cognitive conflict and was perceived as useful for learning. Interest/enjoyment, however, showed a large negative shift for ArguBot. This dissociation between comparable conceptual outcomes and lower interest/enjoyment highlights a key challenge for AI deployment in higher education: an intervention may be

instructionally effective yet difficult to scale if students do not perceive it as a desirable partner for repeated use. A similar pattern was reported by Nazaretsky et al. (2026), who found that university students perceived AI feedback as less credible than human feedback despite comparable quality. For our ArguBot, a plausible explanation is that argumentative exchanges with an AI may feel less socially attuned or less "fair" than peer debate (e.g., reduced rapport, limited conversational flexibility, or asymmetric knowledge base), increasing interactional friction despite scalability benefits. Also, for students experiencing both conditions on the same topic, interest in the after-class condition may have declined as novelty diminished. These findings point to design implications such as improving transparency and credibility, offering students control over challenge intensity and pacing, and incorporating interactional strategies that preserve rapport during disagreement.

Limitations and Future Work. This study provides early empirical insights from a semester-long field deployment of ArguBot in a real course; however, this ecological validity necessarily limits experimental control. In particular, in-class peer-peer argumentation dialogues were not possible to collect in classroom settings, preventing direct process-level comparisons and limiting our ability to control or assess dialogue quality. Consequently, our analyses focus on outcome-level measures rather than fine-grained interactional mechanisms. The following work will therefore conduct more in-depth qualitative analyses of student argumentation discourse, examining epistemic moves, misconception trajectories, and interaction patterns in student-AI dialogues. In addition, while we capture conceptual understanding via pre-post MCQ responses, such item-level change is an imperfect proxy for conceptual change because it may reflect transient response shifts (e.g., short-term persuasion) rather than stable conceptual restructuring. Future studies should therefore triangulate learning with delayed measures (e.g., follow-up assessments) and with process-oriented indicators (e.g., explanation quality and confidence calibration). Finally, we acknowledge a potential selection bias in the reported analyses: students who did not participate in either the peer–peer or peer–ArguBot condition may have differed from those included in the analysis. Also, given the exploratory nature of the field study and the width of some confidence intervals, non-significant effects in Section 3.1 should also not be interpreted as evidence of equivalent effectiveness. Therefore, sufficiently powered controlled laboratory studies are needed to more systematically isolate causal effects by manipulating dialogue quality, agent stance-taking strategies, learner motivation, and comparison conditions, thereby triangulating and strengthening the findings from this exploratory field study.

References

Christa SC Asterhan and Baruch B Schwarz. Argumentation for learning: Well-trodden paths and unexplored territories. *Educational Psychologist*, 51(2):164–187, 2016.

Elizabeth Bear and Xiaobin Chen. Evaluating a conversational agent for second language learning aligned with the school curriculum. In *International Conference on artificial intelligence in education*, pages 142–147. Springer, 2023.

William Cai, Joshua Grossman, Zhiyuan Lin, Hao Sheng, Johnny Tian, Zheng Wei, Joseph Jay Williams, and Sharad Goel. Mathbot: a personalized conversational agent for learning math. *Published to ACM*, 2019.

Michelene TH Chi, James D Slotta, and Nicholas De Leeuw. From things to processes: A theory of conceptual change for learning science concepts. *Learning and instruction*, 4 (1):27–43, 1994.

Ahsen Çini, Pantelis Papadopoulos, Adelson de Araujo, and Jara Martens. The impact of ai-based collaborative conversational agents on metacognitive awareness. In *International Conference on Artificial Intelligence in Education*, pages 212–218. Springer, 2025.

Brooklyn J Corbett and Jason M Tangen. Ai tutors vs. tenacious myths: Evidence from personalised dialogue interventions in education. *arXiv preprint arXiv:2506.09292*, 2025.

Thomas H Costello, Gordon Pennycook, and David G Rand. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814, 2024.

Richard A Duschl and Jonathan Osborne. Supporting and promoting argumentation discourse in science education. 2002.

Kalypso Iordanou, Deanna Kuhn, Flora Matos, Yuchen Shi, and Laura Hemberger. Learning by arguing. *Learning and instruction*, 63:101207, 2019.

David H Jonassen and Bosung Kim. Arguing to learn and learning to argue: Design justifications and guidelines. *Educational Technology Research and Development*, 58(4): 439–457, 2010.

Felix Krieglstein, Maik Beege, Günter Daniel Rey, Christina Sanchez-Stockhammer, and Sascha Schneider. Development and validation of a theory-based questionnaire to measure different types of cognitive load. *Educational Psychology Review*, 35(1):9, 2023.

Deanna Kuhn and Wendy Moore. Argumentation as core curriculum. *Learning: Research and practice*, 1(1):66–78, 2015.

Tanya Nazaretsky, Paola Mejia-Domenzain, Vinitra Swamy, Jibril Frej, and Tanja Käser. Who gives feedback matters: Student biases towards human and ai-generated formative feedback. *Journal of Computer Assisted Learning*, 42(1):e70153, 2026.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024.

Richard M Ryan. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of personality and social psychology*, 43(3):450, 1982.

Chenyang Wang, Filippo Forte, Léane Wettstein, Pierre Dillenbourg, and Thiemo Wambsganss. Arguemate: Designing an arguing agent with maximised disagreement to support student peer-argumentation exercise. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2026a.

Chenyang Wang, Christopher Petrie, Miltiadis Stouras, Nicolas Ettlin, Amaury George, Paola Mejia-Domenzain, Vinitra Swamy, Tanja Käser, and Ola Svensson. Turning 500+ students into teachers: A semester-long study of an ai teachable agent in an undergraduate algorithms course. In *L@ S'26: Proceedings of the Thirteenth ACM Conference on Learning@ Scale*. Association for Computing Machinery, 2026b.

Zhengtao Xu, Junti Zhang, Anthony Tang, and Yi-Chieh Lee. Who you explain to matters: Learning by explaining to conversational agents with different pedagogical roles. *arXiv preprint arXiv:2601.16583*, 2026.