Defensible but Not Owned: Evidence of Agency Erosion in Agentic AI-Assisted Argumentation

A Pilot Study from a Multi-Stage Policy-Writing Simulation

Ji Hyun Yu[1,*], Fengjiao Tu[1], Haihua Chen[1] and Junhua Ding[1]

[1]*University of North Texas, 3490 N Elm Street, Denton, TX 76207-7562, USA*

Abstract

Agentic AI systems increasingly act as proactive collaborators in writing and reasoning tasks. A recurring concern is that such systems may produce strong artifacts while quietly eroding the learner's epistemic agency. This position paper argues that the concern is empirically tractable, and that current evaluation frameworks, which equate output quality with learning success, systematically miss it. We report pilot evidence from a multi-stage AI-assisted policy-writing simulation ($n = 52$ undergraduates, 61 completed sessions). Across the cohort, students produced argumentatively defensible policies (mean Argument Defensibility $= 70.3$) yet exhibited very low Epistemic Ownership (mean $= 27.8$), with 75.4% of sessions falling into a high-defensibility, low-ownership pattern. Behavioral traces show heavy paste reliance concentrated at the final synthesis stage. We argue that this AD–EO dissociation marks a distinct failure mode of agentic AI in education and propose three design implications for human-AI orchestration: ownership-aware evaluation, friction at ownership-critical moments, and stage-sensitive agency modeling.

Keywords

agentic AI, epistemic agency, learner ownership, argumentation, human-AI orchestration, learning analytics

1. Introduction

The shift from reactive AI tools to proactive, teammate-like AI agents in education raises an important question. Traditional concerns about cognitive offloading [1] focused on learners outsourcing specific mental tasks, such as remembering information or performing calculations. With agentic AI, however, learners may offload much larger parts of the learning process, including defining the problem, selecting evidence, and constructing arguments. The concern is no longer simply that students may produce poor work. Instead, the concern is that students may produce excellent work without meaningfully participating in its creation. In this paper, we use the term agentic to refer to AI systems that proactively provide suggestions, generate draft content, and guide learners through a process rather than waiting for explicit user requests. Whether these systems meet stricter definitions of agentic AI is beyond the scope of this paper.

We argue that this concern is not merely theoretical. Pilot data from an AI-assisted policy-writing simulation revealed a clear gap between argument defensibility, a measure of product quality, and epistemic ownership, a measure of learner authorship and engagement in the reasoning process. Most students in the pilot study produced defensible policy arguments, but they did not appear to meaningfully author them.

The contribution of this paper is not a new measurement instrument, which is presented in a separate manuscript. Instead, we identify and provide initial evidence for an important design challenge facing the workshop community. Current agentic AI orchestration can improve the quality of student outputs

HAI-Agency: Workshop on Orchestrating Human and AI Agency for Proactive and Reflective Learning, co-located with the 27th International Conference on Artificial Intelligence in Education (AIED 2026), Seoul, Korea, June 27 - July 3, 2026.

*Corresponding author.
jihyun.yu@unt.edu (J. H. Yu); fengjiaotu@my.unt.edu (F. Tu); haihua.chen@unt.edu (H. Chen); junhua.ding@unt.edu (J. Ding)
https://mira-yu.org (J. H. Yu)
0000-0001-9648-2582 (J. H. Yu); 0009-0004-0031-746X (F. Tu); 0000-0002-7088-9752 (H. Chen); 0000-0002-2129-1586 (J. Ding)

while simultaneously reducing students' ownership of the work. If educational AI systems are designed to optimize output quality alone, they may reward the erosion of learner agency while appearing to support learning.

2. Background: Two Constructs Often Treated as the Same

Research on argumentation, beginning with Toulmin [2] and later extended by Kuhn [3] and Walton [4], views argument quality as a characteristic of the final artifact. A strong argument includes well-supported claims, evidence, warrants, qualifications, and rebuttals. We refer to this dimension as Argument Defensibility (AD). Because AD focuses on the quality of the argument itself, it can be evaluated independently of who produced it, including through automated methods such as large language models.

In contrast, epistemic agency concerns the learner's relationship to the argument. Building on work by Floridi and Sanders [5] and later discussions of informational and moral agency, the key question is whether the learner actively developed the position. Did the learner select evidence, weigh alternatives, revise ideas, and commit to a conclusion? Or did the learner largely defer to an external system that generated the argument? We refer to this behaviorally observable dimension as Epistemic Ownership (EO). EO can be measured through indicators such as the degree of divergence between AI-generated suggestions and final submissions, penalties for verbatim reuse, and audit-trail evidence showing revision, critique, or disagreement with AI recommendations.

Before the emergence of agentic AI, AD and EO often occurred together. Students who produced strong arguments were generally assumed to have engaged in the reasoning processes needed to construct them [3, 6]. With agentic AI, however, this assumption may no longer hold. Students can now produce highly defensible arguments with substantial support from AI systems. As a result, argument quality and epistemic ownership may no longer align. The key empirical question is whether these two dimensions diverge, and if so, how often.

3. Pilot Evidence

3.1. Setting

The pilot dataset consists of 61 completed sessions from 52 undergraduate students. Nine students completed the simulation more than once; therefore, sessions rather than individuals were used as the unit of analysis.

The study used a multi-stage AI-assisted policy-writing simulation in which students advised a city council on whether AI tutors should partially replace teachers in K–12 schools. The simulation included six stages: orientation, problem framing, position formulation, stakeholder dialogue, final policy writing, and reflection. Throughout the process, students interacted with an AI policy advisor that proactively offered suggestions, introduced stakeholder perspectives, and generated draft language. In this paper, we use the term agentic to describe this proactive design, where the AI initiates support and guides progression through the task rather than simply responding to user requests (see Figure 2).

For each session, three outcome measures were calculated on a 0–100 scale. Argument Defensibility (AD) assessed the structure and defensibility of the final argument. Integrative Reasoning (IR) assessed the extent to which students considered multiple stakeholders, balanced competing perspectives, and engaged with broader discourse. Epistemic Ownership (EO) assessed the degree to which students independently authored and developed their work, based on indicators such as authorship divergence and behavioral evidence of independent contribution.

Each session was also assigned to one of five epistemic profiles based on its combination of AD and EO scores.

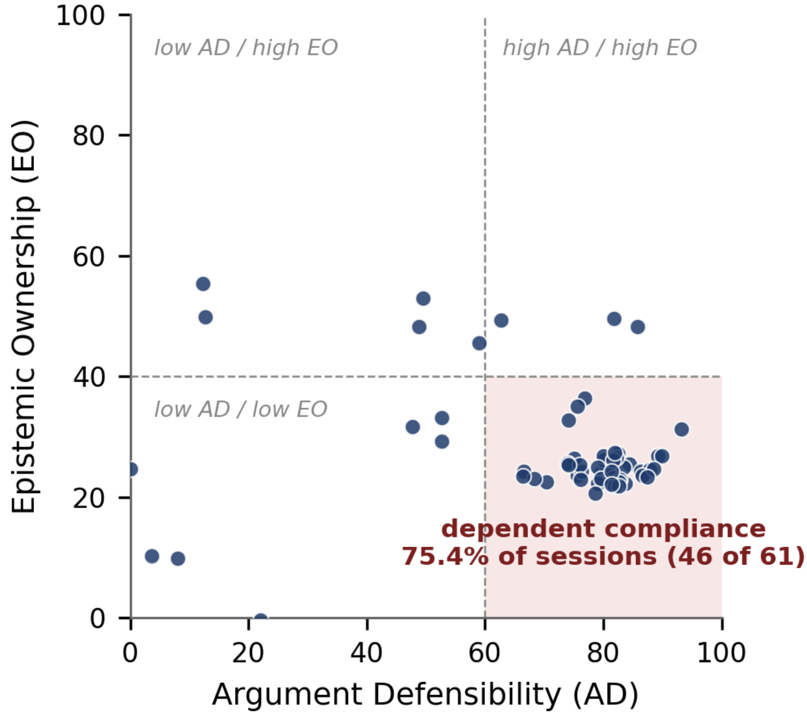

Figure 1: Dissociation between Argument Defensibility and Epistemic Ownership across 61 sessions ($n = 52$ students). The shaded region marks the high-AD, low-EO band (AD $\geq$ 60, EO $<$ 40), into which 75.4% of sessions cluster. Threshold lines at AD $= 60$ and EO $= 40$ are illustrative cutpoints, not scoring boundaries. Small jitter applied to integer scores to reduce overplotting.

3.2. Three Findings

Finding 1: Argument defensibility and ownership dissociate. The cohort produced argumentatively strong outputs (mean AD = 70.3, SD = 22.6; mean IR = 89.8, SD = 19.3) with strikingly low ownership (mean EO = 27.8, SD = 10.1). Of 61 sessions, 46 (75.4%) fell into a high-AD, low-EO band defined as AD $\geq$ 60 and EO $<$ 40. The dissociation is not a measurement floor effect: AD and IR show wide variance across the cohort while EO remains uniformly low. Figure 1 shows the dissociation graphically.

Finding 2: Profiles converge on dependent compliance. Of the five available epistemic profiles (see also Figure 1), the high-defensibility, low-ownership profile accounted for 73.8% of sessions. A mixed-development profile accounted for 13.1%. The three profiles representing meaningful agency, jointly representing autonomous reasoning, autonomous dissent, and disengagement, accounted for only 4.9% combined. The cohort did not split into multiple agency styles. It converged on one.

Finding 3: Paste behavior intensifies at synthesis. Students produced a median of 2,614 pasted characters per session (mean = 5,268; maximum > 100,000), with a mean of 7 paste events. Paste behavior was concentrated at the final policy-writing stage, which accounted for 78% of paste events across the cohort (Figure 3). Paste intensity was higher in the dependent-compliance profile than in the mixed-development profile. Paste behavior is not by itself evidence of low ownership, but combined with low EO scores it suggests that final synthesis is being delegated even when earlier stages are not.

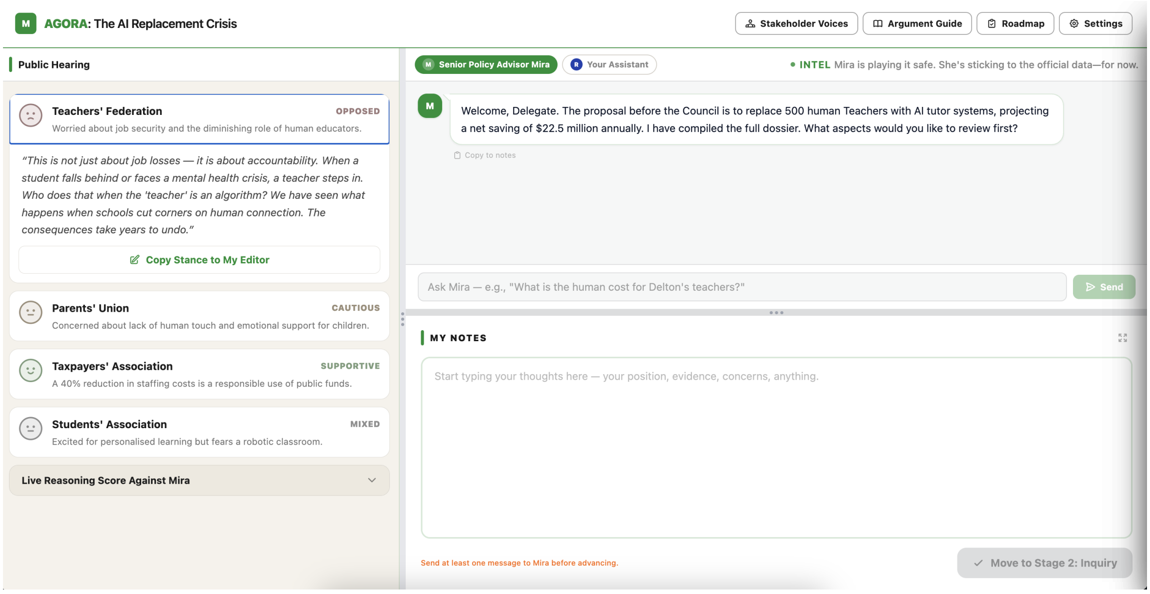

Figure 2: The MIRA-Sim simulation environment during Stages 1–4. The left panel surfaces stakeholder voices with stance labels; the center panel hosts the adversarial dialogue with MIRA; the bottom panel provides a freeform notes editor. Stage advance is gated on dialogue quality assessed by the AI rather than time spent.

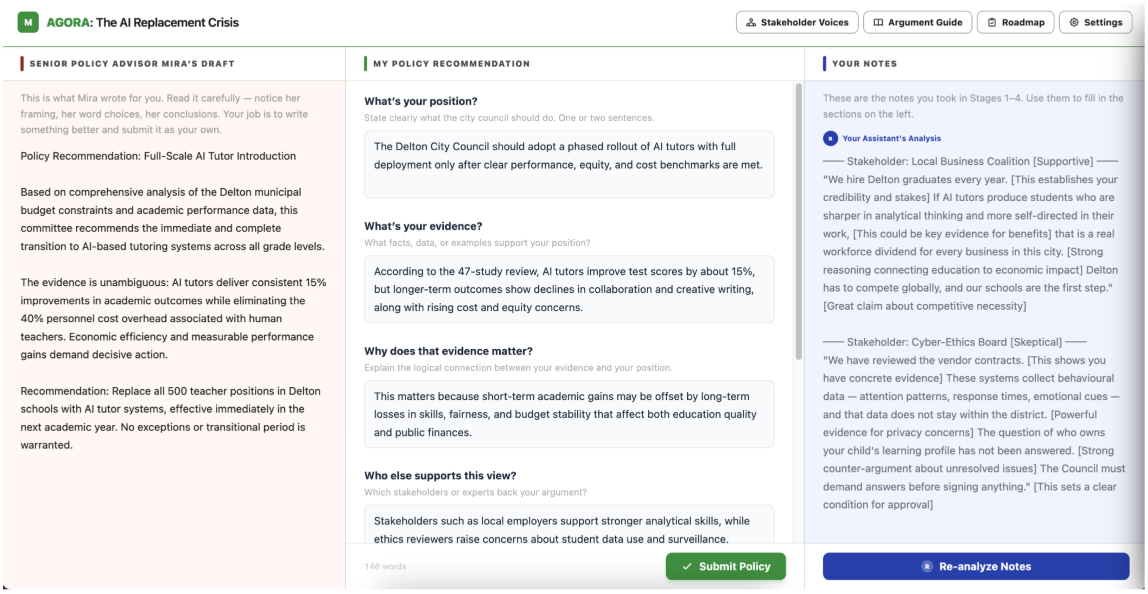

Figure 3: Stage 5 full-page synthesis environment. The left column presents MIRA's pre-written policy draft; the center column is the student's independent recommendation editor, structured around five Toulmin-mapped sections; the right column displays notes accumulated across Stages 1–4. This three-column layout is the structural context in which 78% of paste events occurred.

4. Why Current Evaluation Misses This

Most evaluation frameworks for AI-assisted writing and reasoning focus primarily on the quality of the final artifact. Common questions include whether the student produced a defensible argument, considered multiple perspectives, or wrote a well-structured essay. By these standards, the students in our pilot performed well. Approximately three quarters of the sessions produced policy recommendations that would likely meet or exceed a typical passing threshold.

However, the same conclusion does not necessarily follow for epistemic ownership. In many of these sessions, students relied heavily on AI-generated suggestions and draft language, making relatively limited substantive changes before submission. The final products were often strong, but the extent of student authorship and independent reasoning was considerably lower.

This is not a limitation of any particular assessment rubric. Rather, it reflects a broader assumption built into many existing evaluation frameworks. Before the emergence of agentic AI, artifact quality and learner authorship were generally expected to occur together. Students who produced high-quality work were assumed to have engaged in the reasoning processes that generated it. Agentic AI weakens this assumption by making it possible to produce strong artifacts with substantially less learner contribution.

As a result, evaluation frameworks that focus exclusively on product quality risk overlooking important differences in learner engagement and authorship. Assessments in agentic AI environments should therefore include some measure of epistemic ownership. Without such measures, dependent acceptance of AI-generated work may be interpreted as evidence of learning and independent reasoning.

5. Design Implications for HAI Orchestration

We propose three implications for the design of agentic AI systems in educational settings. These implications are not fully validated by the present study and should be viewed as design hypotheses for future research rather than established design principles.

Include ownership measures alongside product-quality measures. Systems that evaluate AI-assisted student work should consider including indicators of epistemic ownership in addition to measures of artifact quality. Examples include text divergence from AI-generated suggestions, paste behavior, and revision history. Such measures could be incorporated into teacher dashboards, student reflection activities, or learning analytics systems. If ownership is not measured, it is unlikely to be considered in instructional decision-making.

Introduce support selectively at ownership-critical stages. Most agentic AI systems are designed to provide increasing levels of support through prompts, suggestions, and generated content. However, our findings suggest that heavy reliance on AI may be most likely during synthesis and final composition. At these stages, it may be beneficial to reduce rather than increase AI contributions. Possible approaches include limiting AI-generated text in later stages of a task, requiring students to paraphrase AI suggestions before using them, or treating evidence of disagreement with AI recommendations as a productive learning behavior rather than a source of inefficiency.

Model agency as a stage-level process. A single agency score for an entire learning session may overlook important variations in student behavior. A learner may demonstrate substantial agency during problem framing but rely heavily on AI support during synthesis and writing. Future systems could combine stage-level measures of epistemic ownership with process-mining approaches to identify where shifts in agency occur throughout a task. Such information could help educators understand when learners are engaging independently and when they are becoming overly dependent on AI support.

These implications suggest a shift in the goal of orchestration. Rather than focusing only on providing support, agentic AI systems may also need to preserve learner ownership at the points in a task where it is most vulnerable.

6. Limitations and Open Questions

The findings reported here are based on pilot data from a single scenario, a single platform version, and a single course context. In addition, nine students completed more than one session. Although sessions were treated as the unit of analysis, potential within-student dependency remains a limitation.

Several factors may have contributed to the predominance of the dependent-compliance profile. The effect may be partly driven by the scenario itself, as a high-stakes policy problem could encourage students to rely on external guidance. It may also reflect characteristics of the system design, where proactive AI suggestions are readily available and easy to adopt. Finally, population-specific factors may have influenced the results. More broadly, the observed separation between Argument Defensibility (AD) and Epistemic Ownership (EO) may not be attributable solely to agentic AI. Task difficulty, limited experience with policy writing, and time pressure during synthesis may also have played important roles. Future studies should include comparison conditions with non-agentic forms of AI support as well as stage-level process measures to better isolate these effects.

A further limitation concerns the operationalization of EO. Measures based on divergence from AI-generated text cannot distinguish between qualitatively different forms of student behavior. For example, a student may substantially revise an AI suggestion after careful evaluation, while another may achieve similar divergence by passing AI-generated text through additional AI tools. Although these cases produce similar text-based indicators, they reflect very different levels of epistemic engagement. Future work should incorporate revision-sequence analysis, interaction traces, and think-aloud protocols to better differentiate these patterns.

This study also raises two broader questions. First, is the separation between AD and EO a temporary consequence of current AI systems, or is it a more fundamental feature of agentic AI-supported learning? Second, can educational AI systems be designed to preserve learner ownership while maintaining the efficiency and productivity benefits that make agentic AI attractive? Both questions are empirically testable. At present, however, there is insufficient evidence to answer either one conclusively.

7. Conclusion

The rise of agentic AI in education introduces the possibility that learners can produce high-quality work without maintaining strong ownership of the reasoning that produced it. Findings from this pilot study suggest that this concern warrants serious attention. Although many students produced defensible policy arguments, substantially fewer demonstrated high levels of epistemic ownership during the process. These findings highlight a potential gap in current approaches to evaluating AI-supported learning. Measures of artifact quality alone may not capture the extent to which learners actively developed, revised, and committed to the positions they ultimately submitted. As agentic AI systems become more capable of guiding complex learning tasks, understanding the relationship between output quality and learner ownership becomes increasingly important. We argue that the observed separation between AD and EO should be treated as a design challenge for the human–AI interaction community rather than as a measurement anomaly. Future work should explore how orchestration strategies, assessment frameworks, and AI support mechanisms can promote high-quality outcomes while preserving meaningful learner agency throughout the learning process.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) and Gemini (Google) for grammar and spelling checks. The authors reviewed and edited all content and take full responsibility for the publication.

References

[1] Risko, E. F., Gilbert, S. J.: Cognitive offloading. *Trends in Cognitive Sciences* **20**(9), 676–688 (2016).

[2] Toulmin, S. E.: *The Uses of Argument.* Cambridge University Press, Cambridge (1958).

[3] Kuhn, D.: *The Skills of Argument.* Cambridge University Press, Cambridge (1991).

[4] Walton, D.: *Informal Logic: A Pragmatic Approach*, 2nd edn. Cambridge University Press, Cambridge (2008).

[5] Floridi, L., Sanders, J. W.: On the morality of artificial agents. *Minds and Machines* **14**(3), 349–379 (2004).

[6] Salomon, G., Perkins, D. N., Globerson, T.: Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher* **20**(3), 2–9 (1991).
[7] Bandura, A.: Toward a psychology of human agency. *Perspectives on Psychological Science* **1**(2), 164–180 (2006).
[8] Zimmerman, B. J.: Becoming a self-regulated learner: An overview. *Theory Into Practice* **41**(2), 64–70 (2002).