Evaluating the Impact of AI-Assisted Learning on Students' Understanding of Economics and Business Concept[†]

Sining Wang, Scott Shane, Yufan Dong, Heyu Xiong, Jie Ning, Leonardo Madureira, Richard Boyatzis

Draft version: February 2026

Abstract

This paper reports causal evidence from a randomized field experiment embedded in introductory accounting and economics courses that tests how different modes of generative AI assistance affect learning and perceptions of learning. Students were randomized at the student–topic level within session blocks to one of four conditions: no AI, AI that solves problems for them, AI that clarifies mistakes, and interactive AI learning. Each experimental session included a practice exercise in which students had access to AI-assisted tools depending on their treatment status, followed by a quiz exercise in which all students reverted to a level playing field, and a post-activity survey recording their perceptions of learning. We find no statistically detectable increase in quiz performance across AI modes. Two AI modes—solution provision and interactive assistance—significantly increase practice performance. Solution-providing AI additionally reduces student perceptions of learning, whereas interactive AI does not meaningfully change perceptions. AI that clarifies mistakes produces no detectable effects on quiz performance, practice performance, or perceptions of learning. Overall, the findings show that how AI is used matters more than whether it is used, and that the effects of AI assistance are consistent with changes in the timing and structure of cognitive work rather than changes in student effort.

Keywords: Generative AI, AI-assisted learning, field experiment, randomized controlled trial, higher education

[†] All authors are at the Weatherhead School of Management, Case Western Reserve University.

Introduction

Generative AI tools have diffused rapidly into undergraduate economics and business classrooms (Kasneci et al., 2023; Zhai, 2022). Institutions and instructors are now making consequential decisions about whether and how to integrate these tools into routine instruction, often in the absence of credible evidence on their effects on learning (Brynjolfsson et al., 2017; Bulman & Fairlie, 2016). Recent work on generative AI in education emphasizes that learning effects depend not only on model capability but also on how AI is integrated into the learning process, including whether systems provide final answers, interactive guidance, or post-hoc feedback (Dwivedi et al., 2023; Kasneci et al., 2023). Students can use AI in substantively different ways, ranging from delegating problem solving to AI, to using AI as feedback that clarifies mistakes, to engaging in more interactive AI-supported learning. These modes of AI assistance plausibly affect short-run task performance, assessed learning, and students' perceptions in different ways (Kane & Staiger, 2012; Soderstrom & Bjork, 2015) .

From an economic perspective, these distinctions matter because different AI usage modes alter the production of observable performance in different ways, potentially decoupling short-run task completion from the acquisition of knowledge that persists beyond the supported environment (Becker, 1964; Todd & Wolpin, 2003). Empirically, however, most available evidence does not distinguish among AI usage modes or isolate their causal effects in real classroom environments.

Prior studies in this area face several recurring limitations (Holmes et al., 2022; Luckin et al., 2016). First, many rely on observational variation in AI adoption or self-reported usage, making it difficult to separate the effects of AI from differences in student ability, motivation, or instructor practices. Second, a large share of the evidence comes from laboratory or highly controlled settings that may not generalize to regular course instruction, where time constraints, grading incentives, and heterogeneous instructional contexts shape student behavior. Third, much of the existing work emphasizes student perceptions or satisfaction, which may not align with objective measures of learning and may respond differently to AI assistance than performance does (Carrell & West, 2010; Uttl et al., 2017). As a result, there remains limited causal field-experimental evidence on whether AI-assisted gains in observed performance reflect genuine learning or instead reflect changes in how answers are produced under support.

This paper addresses these gaps using a randomized field experiment embedded within regular instruction in two undergraduate courses at a highly selective private university: introductory accounting and introductory microeconomics. Students are randomized at the student–topic level within session blocks to one of four conditions: a no-AI control, AI that

solves the problems, AI that clarifies mistakes, and interactive AI-supported learning. This design isolates the causal effects of AI assignment within the same topic, class section, and instructor. We measure multiple outcomes at the student–topic level, including practice performance, quiz performance as an objective measure of learning, and a post-session survey capturing students' perceptions of learning. Students had access to AI-assisted tools, depending on treatment status, during the practice activity. By holding instructional content, assessment, and timing constant within session blocks, the design allows a clean comparison between supported performance during practice and performance under independent evaluation.

Our study is contributing to a growing literature that uses field experiments to evaluate AI-assisted learning in educational settings. A recent example is Bastani et al. (2025), who study AI tutoring in high school mathematics and show that while access to GPT improves short-run performance, it can reduce performance once AI support is removed(Bastani et al., 2025). Rather than focusing on the presence or withdrawal of AI per se, our study shifts attention to the *structure of AI use*. We examine how distinct modes of AI assistance—ranging from solution provision to interactive learning—shape effort allocation, short-run performance, and subsequent learning in undergraduate accounting and economics courses, explicitly separating supported practice from unassisted assessment.

The contribution of the study is threefold. First, it provides causal evidence from a real classroom setting on how distinct modes of AI assistance affect objective learning outcomes. Second, it examines how the alignment and potential divergence between short-run performance during practice and subsequent assessed learning on quizzes depend on the mode of AI assistance. By explicitly accounting for how AI shapes students' effort allocation across practice and unassisted assessment, we provide evidence consistent with a channel through which interactive AI improves learning efficiency without reducing total effort.

The third contribution is that this study simultaneously documents how students perceive the learning value of different AI modes. Perceived learning is not treated here as a proxy for objective knowledge acquisition, but as a distinct outcome that shapes student engagement, satisfaction, and acceptance of instructional technologies. Divergences between perceived and objective learning are therefore economically relevant, as they influence adoption, compliance, and welfare assessments of educational interventions.

Overall, by shifting attention from the presence of AI to the structure of AI use, the paper reframes the evaluation of AI in education as a problem of input design and outcome measurement rather than adoption per se.

Conceptual Framework

In economic terms, students produce observable performance during learning activities by combining ability, effort, and instructional inputs subject to the constraints of the learning environment (Becker, 1964; Fairlie & Robinson, 2013; Todd & Wolpin, 2003). Importantly, different educational outputs may be produced under different constraints even when they are intended to reflect the same underlying knowledge. As a result, changes in instructional inputs can alter observed performance without necessarily altering the stock of knowledge that persists beyond the learning activity (Soderstrom & Bjork, 2015). Related experimental evidence on instructional inputs and student outcomes includes work showing null effects on performance measures in online settings under quasi-experimental designs(Bettinger et al., 2017).

In the instructional settings studied here, students produce two distinct outputs for each topic. During practice, performance is produced in a supported environment in which feedback, structure, or guidance may be available contemporaneously with task completion. During quizzes, performance is produced under uniform, independent constraints that require students to apply what they have internalized without external assistance. From a production perspective, practice performance reflects output generated under assistance, while quiz performance reflects output generated under unsupported conditions (Todd & Wolpin, 2003).

Because these outputs are produced under different constraints, they need not respond similarly to changes in instructional inputs, even when intended to measure the same underlying knowledge. Other large-scale experimental work in undergraduate settings has identified null effects on traditional academic outcomes despite intervention engagement (Oreopoulos et al., 2020).

Generative AI alters the conditions under which practice performance is produced, but it does so in substantively different ways depending on how it is used. The three AI modes studied here correspond to distinct modifications of the production environment, with different implications for effort, information, and the cost of producing correct answers.

These considerations yield a set of testable hypotheses linking AI usage modes to observable outcomes across performance, learning, perceptions, and time use. The hypotheses are structured hierarchically. We first examine how different modes of AI assistance affect supported practice performance (H1), then assess whether such changes translate into learning under unsupported assessment and into alignment between practice and quiz performance (H2–H3). We next examine how these modes shape students' perceived learning value (H4), and finally use time-use patterns to distinguish effort-based from cognitive mechanisms underlying the observed effects (H5).

When AI provides complete solutions to practice problems, it directly substitutes for student problem-solving effort by reducing the need for independent reasoning. This mode substantially lowers the cost of generating correct answers during practice, making high observed performance possible even with limited cognitive engagement. As a result, improvements in practice performance may reflect answer substitution rather than genuine concept acquisition(Bastani et al., 2025; Bjork et al., 2011; Kumar et al., 2025; Soderstrom & Bjork, 2015).

When AI clarifies mistakes after students complete independent practice, it alters the informational environment while preserving the requirement that students generate answers themselves. This mode does not substitute for problem-solving effort but instead modifies the feedback students receive conditional on their own inputs. Because effort and independent reasoning are preserved, this form of assistance is unlikely to mechanically inflate observed practice performance or to decouple practice performance from learning (Hattie & Timperley, 2007).

> **H1a:** AI that clarifies mistakes after students complete practice independently will produce little or no change in practice performance relative to no AI.

Interactive AI assistance occupies an intermediate position between direct solution provision and post-hoc feedback. By responding dynamically to student inputs, it can guide problem solving while preserving the requirement that students actively generate answers. This mode may reduce unproductive friction during practice without fully substituting for cognitive effort, potentially leading to modest improvements in observed practice performance relative to no AI. Whether such improvements translate into learning depends on whether the assistance complements, rather than replaces, the cognitive work required for concept internalization (Graesser et al., 2011; Lira et al., 2025).

> **H1b:** Interactive AI assistance will produce a positive increase in practice performance relative to no AI.

Quiz performance, by contrast, is produced under uniform and unsupported conditions. Improvements in supported practice performance therefore need not translate into higher quiz scores if those gains reflect changes in answer production rather than changes in underlying knowledge. Economic models of skill formation predict that inputs which primarily reduce effort during production will have limited effects on performance once that support is removed (Soderstrom & Bjork, 2015; Todd & Wolpin, 2003)

> **H2:** Relative to no AI, none of the three AI-assisted conditions is expected to exhibit meaningful short-horizon gains in quiz performance once assistance is removed.

The divergence between supported and independent performance is most directly captured by the difference between quiz and practice outcomes. If AI raises practice performance by lowering the cost of producing correct answers without increasing learning, the gap between quiz and practice performance will widen, implying a decline in quiz-minus-practice performance (Soderstrom & Bjork, 2015).

> **H3a:** AI that provides complete solutions will generate a decline in quiz-minus-practice performance.
>
> **H3b:** AI that clarifies mistakes will produce no decline in quiz-minus-practice performance.
>
> **H3c:** Interactive AI assistance will generate a smaller decline in quiz-minus-practice performance than AI that provides complete solutions.

Beyond objective performance under supported and unsupported conditions, AI assistance may also shape how students perceive the learning value of the instructional experience. Perceived learning reflects students' subjective assessments of effort, autonomy, and skill acquisition during practice, which need not align with objective performance outcomes (Eom et al., 2006; Uttl et al., 2017). In particular, when AI substantially lowers the cost of producing correct answers, students may infer reduced learning value from the experience even if observed practice performance improves.

> **H4a**: AI that provides complete solutions will reduce perceived learning relative to no AI.
>
> **H4b**: AI that clarifies mistakes will not reduce perceived learning relative to no AI.
>
> **H4c**: Interactive AI assistance will not reduce perceived learning relative to no AI.

A central distinction in interpreting the effects of AI assistance concerns whether observed performance changes arise from increased effort or from changes in the cognitive production process itself. An effort-based mechanism implies that improved performance is driven by greater time investment or sustained exertion across tasks. A cognitive mechanism, by contrast, implies that AI alters the efficiency, timing, or nature of cognitive work (such as reducing friction, reallocating effort across stages, or substituting for certain forms of reasoning) without necessarily increasing total effort.

This distinction has sharp empirical implications. Under an effort-based mechanism, improvements in performance should be accompanied by increases in total time-on-task or uniform changes in time spent across stages. Under a cognitive mechanism, performance may improve even when total time remains unchanged, and time-use patterns may diverge

across practice and quiz activities in ways that reflect reallocation or substitution of cognitive work rather than increased effort.

Time-use outcomes provide a useful way to distinguish effort-based from cognitive mechanisms. An effort-based explanation predicts uniform changes in time spent across activities, while a cognitive mechanism predicts stage-specific shifts in time allocation and, potentially, unchanged total time.

> **H5a:** AI that provides complete solutions will not increase total time-on-task and will reallocate time away from practice toward quizzes, consistent with cognitive substitution rather than increased effort.

> **H5b:** Interactive AI assistance will shift time investment toward the practice stage and reduce time spent on quizzes, consistent with front-loaded cognitive engagement rather than effort substitution.

In sum, these arguments generate mode-specific predictions that link AI usage to distinct changes in production constraints and to observable differences across practice performance, quiz performance, their alignment, and student perceptions. The empirical analysis evaluates whether observed outcomes conform to these theoretically grounded predictions.

Experimental Design

This study was conducted as a field experiment embedded within regular instruction in two undergraduate courses at a private U.S. university: introductory accounting and introductory microeconomics. In both courses, instruction is organized around discrete instructional topics (e.g., supply and demand; job costing), which serve as the primary learning units for the course. The experiment was implemented within these naturally occurring instructional units, allowing students to complete course-required practice and assessment activities under different instructional environments without altering the structure or pacing of the courses. The study protocol was reviewed and approved by the university's Institutional Review Board (IRB), and all participating students provided informed consent prior to participation.

Randomization Strategy

The unit of randomization was the student–topic level. Each course included four experimental sessions over the semester, each corresponding to a distinct instructional topic. We implemented a balanced Latin square design such that, across the four experimental sessions, each student was assigned to each of four experimental conditions

exactly once. Treatment order was fully balanced, with each condition preceding and following every other condition exactly once across students. Randomization was conducted within session blocks defined by topic, class section, instructor, and timing. This ensured that treatment order effects were fully balanced across students.

Implementation

Each experimental session followed a standardized structure embedded within regular class time. Students first attended the scheduled lecture for the instructional topic, which lasted approximately 40 minutes and was delivered by the course instructor as usual. After the lecture concluded, the instructor left the classroom and a member of the research team took over to administer the experimental session.

Students then completed a practice stage consisting of ten topic-specific practice questions. During this stage, students were assigned to one of four experimental conditions that differed only in the form of AI assistance available. In the no-AI control condition (T0), students completed the practice problems independently without any AI support. In the AI solves the problem condition (T1), students could submit practice questions to an AI system that generated final answers directly, without explanation. In the AI clarifies mistakes condition (T2), students were instructed to first complete the practice problems independently and receive grades and could then consult an AI system to ask targeted questions or clarify specific mistakes. In the interactive AI learning condition (T3), students worked through the practice problems while interacting with an AI tutor designed to provide step-by-step guidance and prompts, without supplying final answers. At the end of the practice stage, students received a grade on their practice performance. The AI-generated solutions were provided for structurally similar problems rather than for the exact practice questions themselves, ensuring that students still needed to map solutions to the assigned problem context.

After completing the practice stage, students proceeded to a quiz stage consisting of a separate set of ten questions aligned with the same instructional topic. No AI assistance was available during the quiz stage in any condition. Students did not receive feedback or see their quiz grades at the end of the quiz stage. Finally, students completed a short post-activity survey measuring their perceptions of learning and the instructional experience (See Figure A1. Experimental Implementation Flowchart).

Each experimental session lasted approximately 30–45 minutes. Sessions were monitored by at least one member of the research team and two research assistants, none of whom were the course instructor, to ensure compliance with assigned conditions. In addition, each AI tool was implemented with built-in constraints that prevented students from using the system outside its intended mode (e.g., the interactive AI tutor could not generate final

answers even when explicitly prompted). Quiz performance contributed to students' course grades, providing incentives for full participation. More details about AI tool configuration can be found in Appendix A.

Measurement

Objective learning outcomes are drawn from the practice and quiz components of each experimental session. Practice performance is measured as the score on the practice assignment, recorded on a 0–10 scale. Quiz performance is measured as the score on the quiz assessment for the same instructional topic, also on a 0–10 scale. In addition, we construct a quiz-practice difference measure defined as the quiz score minus the practice score, capturing the alignment between supported practice performance and subsequent independent quiz performance within the same session. Time spent on the practice and quiz stages is measured using system-generated timestamps that record the duration (in seconds) of student activity in each stage.

Subjective learning outcomes are measured using a post-activity survey administered immediately following the quiz stage. The survey captures students' perceived learning and satisfaction with the instructional experience using ten items adapted from established and validated instruments in the learning environments literature (Eom et al., 2006; Hwang & Kim, 2022; Walker & Fraser, 2005). The items assess perceived understanding, application to real-world problems, critical thinking, skill acquisition, depth of learning, engagement, overall satisfaction, willingness to recommend the learning approach, alignment with expectations, and confidence.

All survey items are measured on a five-point Likert scale ranging from strongly disagree (1) to strongly agree (5), with higher values indicating more favorable perceptions. In the primary analysis, the items are aggregated into an unweighted composite perception index. As a robustness check, the composite index is decomposed into two subscales capturing perceived skill acquisition and satisfaction with the learning experience.

Estimation Strategy

We estimate intent-to-treat (ITT) effects of three modes of AI assistance—AI solves the problem, AI clarifies mistakes, and interactive AI learning—relative to a no-AI control condition. For each outcome, we estimate linear regressions of the form:

$$Y_{is} = \alpha + \beta_1 T1_{is} + \beta_2 T2_{is} + \beta_3 T3_{is} + \gamma_s + \varepsilon_{is},$$

where Y_{is} denotes the outcome for student i in session s; $T1_{is}$, $T2_{is}$, and $T3_{is}$ are indicator variables for the three AI treatment conditions; and the omitted category is the no-AI

control condition. All specifications include session fixed effects (γ_s), which absorb instructional content, class section, instructor, course field, and timing within each experimental session.

Standard errors are clustered at the student level to account for within-student correlation in outcomes across multiple topics and, for a subset of students, across courses. Because treatment assignment is randomized at the student–topic level within session blocks, coefficient estimates are identified from randomized variation in AI assignment within the same instructional context. As a robustness check, we additionally estimate specifications that include student fixed effects; these models rely solely on within-student variation across topics.

In addition to performance and perception outcomes, we examine how AI assistance affects the allocation of time across stages of the learning activity and the efficiency with which students complete the quiz assessment. For each experimental session, we observe the time a student spends during the practice stage and during the quiz stage, measured in seconds. The unit of observation is the student–topic pair; references to student–session observations are used only to describe the timing and sequencing of practice and assessment within the experimental design. From these measures, we construct three time-use outcomes: practice time, quiz time, and total time-on-task, defined as the sum of practice and quiz time.

We estimate intent-to-treat effects of AI assignment on each time-use outcome using the same specification as above, replacing Y_{is} with the relevant time-use measure.

To relate time use to assessment performance, we construct a measure of quiz fluency defined as the ratio of quiz performance to time spent on the quiz:

$$\text{Fluency}_{is} = \frac{\text{QuizScore}_{is}}{\text{QuizTime}_{is}/60}$$

where QuizScore_{is} is the quiz score (0–10) and QuizTime_{is} is quiz time measured in seconds, so that fluency is expressed as points per minute. Quiz fluency captures the efficiency with which students apply knowledge under independent assessment. Holding scores constant, higher fluency reflects lower cognitive friction during problem solving rather than greater output alone. We estimate intent-to-treat effects of AI assignment on quiz fluency using the same regression specification, including session fixed effects and clustering standard errors at the student level. Student fixed effects specifications are again reported as robustness checks.

Results

The study includes a total of 411 unique student participants. Of these, 257 were enrolled in introductory accounting and 241 in introductory microeconomics, with 87 students enrolled in both courses during the semester. Across both courses, there were six instructors and four sections per course.

The resulting analysis sample consists of 1,821 student–topic level observations, including 909 observations from introductory accounting and 912 observations from introductory microeconomics. Table 1 reports the distribution of student–topic level observations across courses and experimental conditions. Table 2 reports the distribution of students across experimental sessions. Additional descriptive statistics for practice and quiz performance are reported in Table B1 in Appendix B.

We begin by examining effects on practice performance (see Table 3, column 1). Assignment to AI that solves the problem produces a large and statistically significant increase in practice scores of approximately 1.33 points *($\beta = 1.33$, $p < 0.01$)*. Assignment to interactive AI learning also increases practice scores, though by a smaller magnitude of approximately 0.58 points *($\beta = 0.58$, $p < 0.01$)*. These effects correspond to improvements of roughly 16 percent and 7 percent of the average practice score for AI solves the problem and interactive AI learning, respectively. In contrast, assignment to AI that clarifies mistakes does not produce a statistically significant change in practice performance relative to no AI. Thus, two of the three AI modes substantially improve students' immediate performance during practice activities.

We next turn to quiz performance, which serves as our primary measure of learning (see Table 3, column 2). Across all three AI modes, we find no statistically significant differences in quiz scores relative to the no-AI condition. The estimated coefficients are small in magnitude and imprecisely estimated, with point estimates close to zero for AI solves the problem, AI clarifies mistakes, and interactive AI learning. Despite sizable improvements in practice performance for two AI modes, these gains do not translate into detectable increases in quiz scores.

We then examine student perceptions, measured using a composite survey index that averages the ten items capturing perceived learning, engagement, confidence, and satisfaction (see Table 3, column 3). Relative to no AI, assignment to AI that solves the problem leads to a statistically significant decline in survey scores of approximately 0.37 points *($\beta = -0.37$, $p < 0.01$)*—or an 11% reduction based on the average survey score. In contrast, assignment to AI that clarifies mistakes or to interactive AI learning does not produce statistically significant changes in student perceptions. Thus, only the AI mode that directly solves problems for students is associated with lower perceived learning.

Finally, to make the divergence between practice and quiz outcomes explicit, we construct a quiz-practice difference measure defined as quiz score minus practice score (see Table 3, column 4). Relative to no AI, assignment to AI that solves the problem is associated with a statistically significant decline in the quiz-practice difference of approximately 1.37 points ($\beta = -1.37$, $p < 0.01$). Assignment to interactive AI learning is also associated with a statistically significant decline in the quiz-practice difference of approximately 0.64 points ($\beta = -0.64$, $p < 0.01$). Assignment to AI that clarifies mistakes does not produce a statistically significant change in the quiz-practice difference. Those results indicate that for AI modes that raise practice performance, gains observed during practice do not translate into improved quiz performance.

Taken together, the results reveal a consistent pattern across outcomes. AI that solves problems for students and AI that provides interactive assistance both improve short-term practice performance, but neither increases learning as measured by quiz performance, and both are associated with a deterioration in the quiz-practice difference. Only AI that solves the problem negatively affects students' perceptions of learning and satisfaction, while interactive AI learning leaves perceptions unchanged. AI that clarifies mistakes produces no detectable effects on practice, quizzes, perceptions, or improvement, and behaves similarly to the no-AI condition across all outcomes.

To assess the robustness of the main findings, we conduct a series of additional exercises for each primary outcome—practice performance, quiz performance, student perceptions, and the quiz-practice difference.

First, we added student fixed effects as a robustness check (Table 4). For quiz performance, treatment coefficients remained small and statistically indistinguishable from zero, with magnitudes similar to the baseline specification. For practice performance, the strong positive effects for AI solves the problem and interactive learning were essentially unchanged. For student perceptions, the negative effect of AI solves the problem remained large and statistically significant, while the other AI modes remained statistically indistinguishable from no AI. For quiz-practice difference, the negative effects for AI solves the problem and interactive learning remained large and statistically significant with similar magnitudes. Across outcomes, adding student fixed effects does not alter the qualitative pattern of results, indicating that the estimates are not driven by compositional differences across students.

Next, we interacted treatment assignment with course field to test for field-specific artifacts (Table 5). These models estimate the treatment effect in Economics and the differential effect in Accounting. This robustness check is intended to assess whether the main treatment effects are driven by differential course composition or institutional features across fields, rather than by the mode of AI assistance itself. For quiz performance, neither

the Economics main effects nor the Accounting differentials show systematic departures from the overall null pattern. For practice performance, AI solves the problem remains strongly positive in Economics and is larger in Accounting, while interactive learning remains positive in Economics with no statistically significant Accounting differential. For student perceptions, the negative effect of AI solves the problem is present in Economics and stronger in Accounting. For quiz-practice difference, the negative effect of AI solves the problem is present in Economics and is larger in Accounting, while interactive learning remains negative in Economics with no statistically significant Accounting differential. Overall, these interaction results do not suggest that the core findings are driven by one field, while at the same time implying that the effects are exacerbated for Accounting when compared to Economics courses.[‡]

Although some treatment effects are larger in Accounting than in Economics, these interaction results are not the focus of the analysis and are not interpreted as evidence of discipline-specific mechanisms. The primary contribution of the paper lies in identifying how different modes of AI assistance affect learning-related outcomes across instructional contexts, rather than in explaining cross-field differences that may reflect bundled institutional or curricular factors.

We also evaluated whether treatment effects vary systematically by topic or by instructor by estimating models that interact treatment assignment with topic indicators and, separately, with instructor indicators, while maintaining session fixed effects and student-clustered standard errors. We assessed heterogeneity using joint F-tests of whether each set of interaction terms is collectively zero and examined the implied topic- and instructor-specific treatment effects. We find no evidence that the main results are driven by topic- or instructor-specific artifacts, allowing us to treat the estimated effects as robust across instructional contexts.

The main analysis measures student perceptions using a composite index constructed from all ten survey items. As a robustness check, we decompose this index into its two conceptual components—perceived skills learned (items capturing understanding, skill acquisition, and depth of learning) and satisfaction with learning (items capturing engagement, confidence, and overall satisfaction). Both subscales exhibit high internal consistency (Cronbach's $\alpha > 0.90$), and yield qualitatively similar results to the composite index (see Appendix B Table B2).

[‡] Average scores for both practices and quizzes are lower for Accounting vs Economics courses across all modes, suggesting students face more difficulties with the Accounting exercises. The higher improvement in practice scores and the extra reduction in perception of learning for Accounting when AI solves the problem is consistent with the perceived higher level of difficulty of these exercises.

The results are substantively unchanged when perceptions are disaggregated. AI that solves the problem is associated with a large and statistically significant decline in both perceived learning and satisfaction, while AI that clarifies mistakes and interactive AI learning have no statistically significant effects on either construct. Thus, the negative perception effects observed in the main analysis are not driven solely by satisfaction or affective responses; they are equally present in students' self-assessments of learning. These findings confirm that the combined perception index used in the main analysis does not mask offsetting effects across distinct perceptual dimensions.

We next examine how AI assistance affects time allocation across stages of the learning activity. Table 6 reports intent-to-treat estimates of time spent (in seconds) during the practice stage, the quiz stage, and in total.

Table 6 shows that assignment to AI that solves the problem (T1) reduces time spent during practice by approximately 21–24 seconds ($p < 0.01$), while increasing quiz time by approximately 19–20 seconds ($p < 0.05$), depending on whether student fixed effects are included. Total time-on-task is statistically indistinguishable from the no-AI condition. This pattern indicates a reallocation of time from the practice stage to the quiz stage rather than a reduction in overall time investment.

In contrast, assignment to interactive AI learning (T3) produces the opposite pattern. T3 increases practice time by approximately 117–121 seconds ($p < 0.01$), reduces quiz time by approximately 23–24 seconds ($p < 0.01$), and increases total time-on-task by approximately 94–97 seconds ($p < 0.01$). These effects are robust to the inclusion of student fixed effects. Assignment to AI that clarifies mistakes (T2) produces no statistically significant changes in practice time, quiz time, or total time-on-task.

Taken together, these results indicate that different AI modes reshape how students allocate time across learning stages rather than uniformly increasing or decreasing overall time investment.

Next, we examine treatment effects on quiz fluency, which captures how efficiently students apply knowledge on the quiz, with higher values indicating less time required to achieve the same score. We define quiz fluency as the number of points earned per minute spent on the quiz, capturing the rate at which students translate time into assessed performance. Table 7 reports the corresponding estimates. Assignment to interactive AI learning (T3) increases quiz fluency by approximately 0.12–0.13 points per minute, an effect that is statistically significant and robust to the inclusion of student fixed effects. In contrast, AI that solves the problem (T1) produces a small and statistically insignificant decline in fluency, while AI that clarifies mistakes (T2) has no detectable effect.

The magnitude of the T3 effect is economically meaningful. A 0.12-point increase per minute implies that, over a ten-minute quiz, students assigned to interactive AI effectively answer about 1.2 additional points' worth of questions correctly, out of 10 points, relative to the no-AI condition.

Building on the time-allocation results, these fluency patterns indicate that different AI modes operate through distinct cognitive mechanisms rather than simple changes in effort. When AI provides complete solutions (T1), students spend less time during practice and more time during the quiz, with no change in total time-on-task and no improvement in quiz fluency. This pattern is consistent with cognitive substitution: AI replaces independent reasoning during practice, inflating short-run performance while deferring cognitive work to the assessment stage, where students must re-engage in problem solving without efficiency gains.

In contrast, interactive AI learning (T3) increases time spent during practice, reduces time spent during the quiz, raises total time-on-task, and improves quiz fluency. This configuration is difficult to reconcile with effort-saving behavior and instead reflects cognitive scaffolding, whereby AI induces earlier and deeper engagement during practice that lowers cognitive friction under independent assessment. Although neither treatment increases average quiz scores, the opposing reallocations of time and the presence of fluency gains under T3 reveal that AI primarily reshapes the timing and structure of cognitive work rather than aggregate performance levels.

Discussion

This study evaluates how distinct modes of generative AI assistance affect short-run performance, assessed learning, time allocation, and student perceptions in a real classroom environment. By randomizing AI usage modes within tightly controlled instructional contexts, the design isolates causal effects of *how* AI is used, allowing interpretation of results through the lens of how AI alters the production of performance and learning rather than through broad claims about AI adoption in education. Because the randomized intervention varies only the mode of AI assistance while holding the underlying problems and instructional setting fixed, the experiment isolates the causal effects of AI usage modality rather than differences in generic problem-solving pedagogy.

Several limitations of the study should be noted. First, the experimental design measures learning outcomes over a short horizon, with quizzes administered immediately following the practice period. As a result, the estimated effects speak to near-term performance under different modes of AI assistance and do not directly capture the durability of learning over

longer time horizons. Second, the close sequencing of practice and assessment reflects the instructional context in which the intervention was deployed, but it necessarily limits the ability to disentangle longer-run knowledge accumulation from short-run performance effects. Third, while the study carefully distinguishes between different modes of AI use—such as solution provision, post-hoc clarification, and interactive assistance—the analysis does not attempt to evaluate the underlying quality or frontier status of the AI system itself. The AI tool should therefore be interpreted as representative of technologies plausibly available to students at the time of the study, rather than as a benchmark for state-of-the-art performance. Relatedly, although safeguards were implemented to ensure reasonable and consistent AI outputs, the paper does not claim that the implementation is optimal, nor that the results would generalize mechanically to all AI systems or instructional settings. Finally, the study abstracts from detailed process measures of how individual students engage with the interactive AI condition, focusing on reduced-form treatment effects. A more granular analysis of engagement dynamics is left for future work.

A central pattern emerging from the results is a divergence between practice performance and assessed learning. Two AI modes—AI that solves the problem and interactive AI learning—substantially raise performance during practice, yet neither produces gains in quiz performance, and both are associated with declines in the quiz-practice difference. This divergence highlights that higher short-run task performance under AI assistance does not necessarily correspond to greater learning under independent assessment, a distinction long emphasized in the economics of education between observed performance and durable human capital formation (Angrist & Lavy, 2002; Becker, 1964).

This pattern reflects differences in how AI alters the cognitive production of answers during practice. When AI provides complete solutions, it directly lowers the cost of generating correct responses, allowing students to achieve higher observed performance without engaging in independent problem solving. The resulting practice gains therefore reflect changes in answer production rather than changes in the underlying stock of knowledge that persists to assessment. Interactive AI learning, while less extreme, similarly modifies the practice environment in ways that improve immediate task completion without yielding detectable gains in subsequent quiz performance. By contrast, AI that clarifies mistakes preserves the core production process and correspondingly produces no detectable effects on either performance or learning. This pattern echoes prior field evidence showing that instructional inputs that improve task completion do not always translate into learning gains (Banerjee et al., 2007).

The divergence between practice and quiz outcomes underscores the importance of distinguishing between performance and learning as separate outputs produced under different constraints. Practice occurs in a supported environment, whereas quizzes require

independent application under uniform conditions. Improvements in the former need not translate into gains in the latter. The quiz-practice difference makes this divergence explicit, showing that for AI modes that raise practice scores, assessment performance does not keep pace.

Student perceptions provide a further dimension on which AI modes differ. AI that solves the problem reduces perceived learning, while interactive AI learning leaves perceptions unchanged, despite both modes failing to increase quiz performance. These findings demonstrate that subjective evaluations capture a distinct outcome dimension shaped by features of the instructional experience—such as autonomy and engagement—rather than serving as reliable proxies for objective learning (Eom et al., 2006; Walker & Fraser, 2005).

The time-use and fluency results clarify the mechanisms underlying these patterns. Rather than uniformly increasing or decreasing effort, different AI modes induce sharply different reallocations of cognitive work across stages of the learning activity. When AI provides complete solutions, students spend less time during practice and more time during the quiz, with no change in total time-on-task and no improvement in quiz fluency. This configuration is consistent with cognitive substitution: AI replaces independent reasoning during practice, inflating short-run performance while deferring cognitive work to the assessment stage, where assistance is unavailable.

Interactive AI learning produces the opposite configuration. Students spend substantially more time during practice, less time during the quiz, and more time overall, while also exhibiting higher quiz fluency. This pattern is difficult to reconcile with shortcut-taking or reduced effort. Instead, it reflects cognitive scaffolding, whereby interactive AI induces earlier and deeper engagement during practice that lowers cognitive friction under independent assessment. Although this reallocation does not translate into higher quiz scores, it fundamentally alters how students arrive at those scores.

Taken together, the evidence shows that generative AI does not merely change how much students work, but when and how cognitive work is performed. The central risk of AI in education is therefore not reduced effort per se, but the outsourcing or restructuring of key cognitive processes during learning. Evaluations that rely solely on performance levels, time spent, or student satisfaction risk missing these dynamics. By jointly analyzing performance, time allocation, fluency, and perceptions, this study demonstrates that the educational effects of AI hinge critically on the structure of AI use and on how learning outcomes are measured.

Conclusion

This paper provides causal field-experimental evidence on how different modes of generative AI assistance affect performance, learning, time allocation, and student perceptions in undergraduate economics and accounting courses. By randomizing AI access at the student–topic level within tightly controlled instructional contexts, the study isolates the effects of *how* AI is used rather than the presence of AI itself. The results demonstrate that AI can substantially alter observed performance during practice without improving learning as measured under independent assessment conditions, and that subjective perceptions of learning need not align with objective outcomes.

The central contribution of the paper is to show that how AI is used matters more than whether it is used. AI deployed as a solution-providing input increases short-run task performance but does not generate learning gains and is associated with a widening gap between practice and quiz outcomes. Interactive AI assistance, while less disruptive to students' perceptions of learning, likewise fails to improve assessed learning. In contrast, AI that clarifies mistakes without reducing the requirement for independent problem solving produces no detectable effects on either performance or learning. Together, these findings underscore the importance of distinguishing between performance under support and learning under independent conditions when evaluating AI-assisted instruction.

The results have direct implications for education policy, instructional design, and emerging approaches to AI governance in educational settings. Institutions evaluating AI adoption should avoid equating higher observed performance during AI-assisted activities with learning gains and should be cautious about deploying AI primarily as a substitute for students' cognitive work. From a design perspective, AI tools that preserve students' responsibility for generating solutions are less likely to distort perceptions of learning, even if they do not yield immediate performance gains. More broadly, the findings suggest that effective governance frameworks should focus not only on access to AI, but on the structure of AI use within instructional environments.

By shifting attention from the existence of AI to the economic consequences of different AI usage modes, this study contributes evidence that can inform both future research and institutional decision-making. Because the experiments are conducted within a single institution, the results should be interpreted as internally valid estimates of how AI usage modes affect learning under a common instructional environment, rather than as population-wide treatment effects. The design nonetheless captures economically relevant variation in instructional structure that is common across many selective universities. As generative AI continues to diffuse rapidly into classrooms, careful evaluation of how these

tools reshape the production of learning will be essential for aligning technological innovation with educational objectives.

Materials and Methods

The study involved human participants and was reviewed and approved by the Institutional Review Board of Case Western Reserve University. All participants provided informed consent prior to participation. Students were informed that participation involved the use of instructional tools as part of regular course activities, that their responses would be used for research purposes in de-identified form, and that participation was voluntary and would not affect their course standing.

Acknowledgments

The authors acknowledge the financial support of Cloudforce LLC in undertaking this research. We are grateful to the instructors who generously opened their classrooms to allow us to implement the experimental interventions underlying this study. Their willingness to integrate the research design into their regular teaching made this project possible. We thank Jonathan Ernest**,** Jenny Hawkins**,** Abraham Joseph, and Piotr Skomra for their cooperation, flexibility, and support throughout the semester. We are grateful for the helpful comments from Anthony Bucaro, Jeffrey Capadona, Timothy Fogarty, Susan Helper, Mark Schweitzer, and Daniel Shoag.

References

Angrist, J., & Lavy, V. (2002). New evidence on classroom computers and pupil learning. *The Economic Journal*, *112*(482), 735–765. https://academic.oup.com/ej/article/112/482/735/5079771

Banerjee, A. V, Cole, S., Duflo, E., & Linden, L. (2007). Remedying educaiton: evidence from two randomized experiments in India. *The Quarterly Journal of Economics*, *122*(3), 1235–1246. https://academic.oup.com/qje/article/122/3/1235/1879525

Bastani, H., Bastani, O., Sungu, A., Ge, H. I., Kabakcı, Ö., Mariman, R., & Brunskill, E. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, *122*(26). https://doi.org/10.1073/pnas

Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education, First Edition*. University of Chicago Press.

Bettinger, E., Doss, C., Loeb, S., Rogers, A., & Taylor, E. (2017). The effects of class size in online college courses: Experimental evidence. *Economics of Education Review*, *58*, 68–85. https://doi.org/10.1016/j.econedurev.2017.03.006

Bjork, E. L., Bjork, R., Roediger, H. L., Mcdermott, K. B., & Mcdaniel, M. A. (2011). Learning Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning. In *Psychology and the real world: Essays illustrating fundamental contributions to society,* (pp. 59–68).

Brynjolfsson, E., Rock, D., Syverson, C., Abrams, E., Agrawal, A., Autor, D., Benzell, S., Gans, J., Goldfarb, A., Goolsbee, A., Saint-Jacques, G., Meyer, A., & Tratjenberg, M. (2017). *Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics* (24001). http://www.nber.org/papers/w24001

Bulman, G., & Fairlie, R. W. (2016). Technology and Education: Computers, Software, and the Internet. *Handbook of the Economics of Education*, *5*, 239–280. http://one.laptop.org/about/countries.

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy*, *118*(3). https://doi.org/https://doi.org/10.1086/653808

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Eom, S. B., Wen, H. J., & Ashill, N. (2006). The Determinants of Students' Perceived Learning Outcomes and Satisfaction in University Online Education: An Empirical Investigation*. *Decision Sciences Journal of Innovative Education*, *4*(2), 215–235. https://doi.org/10.1111/j.1540-4609.2006.00114.x

Fairlie, R. W., & Robinson, J. (2013). *Experimental Evidence On the Effects of Home Computers on Academic Achievement Among Schoolchildren* (NBER Working Paper 19060). http://one.laptop.org/about/countries.

Graesser, A. C., Mello, S. D. ', & Cade, W. (2011). Instruction based on tutoring. *Handbook of Research on Learning and Instruction*, 422–440.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Holmes, W., Bialik, M., & Fadel, C. (2022). Artificial intelligence in education. In *Data ethics : building trust : how digital technologies can serve humanity* (pp. 621–653). Globethics Publications. https://doi.org/10.58863/20.500.12424/4276068

Hwang, S., & Kim, H. K. (2022). Development and validation of the e-learning satisfaction scale (eLSS). *Teaching and Learning in Nursing*, *17*(4), 403–409. https://doi.org/10.1016/j.teln.2022.02.004

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching Combining High-Quality Observations with Student Surveys and Achievement Gains*.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*. https://doi.org/10.1016/j.lindif.2023.102274

Kumar, H., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. (2025). Artificial Intelligence in Education. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial Intelligence in Education. AIED 2025* (Vol. 15880, pp. 60–75). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-98459-4

Lira, B., Rogers, T., Goldstein, D. G., Ungar, L., & Duckworth, A. L. (2025). *Coach not crutch: AI assistance can enhance rather than hinder skill development* (2502.02880. 2025). http://arxiv.org/abs/2502.02880

Luckin, R., Holmes, W., Griffiths, M., & Pearson, L. B. F. (2016). *Intelligence Unleashed An argument for AI in Education*.

Oreopoulos, P., Petronijevic, U., Logel, C., & Beattie, G. (2020). Improving non-academic student outcomes using online and text-message coaching. *Journal of Economic Behavior and Organization*, *171*, 342–360. https://doi.org/10.1016/j.jebo.2020.01.009

Soderstrom, N. C., & Bjork, R. A. (2015). Learning Versus Performance: An Integrative Review. *Perspectives on Psychological Science*, *10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Todd, P. E., & Wolpin, K. I. (2003). On The Specification and Estimation of The Production Function for Cognitive Achievement. *The Economic Journal*, *485*(485), 3–33. https://academic.oup.com/ej/article/113/485/F3/5079558

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Walker, S. L., & Fraser, B. J. (2005). Development and validation of an instrument for assessing distance education learning environments in higher education: The Distance Education Learning Environments Survey (DELES). *Phenomenology and the Cognitive Sciences*, *4*(3), 289–308. https://doi.org/10.1007/s10984-005-1568-3

Zhai, X. (2022). *ChatGPT User Experience: Implications for Education*. https://orcid.org/0000-0003-4519-1931

Figures and Tables

Table 1. Student-Topic Observations per Experiment Condition

Course	**No AI**	**AI Solves the Problem**	**AI Clarifies Mistakes**	**Interactive AI Learning**	**Total**
Accounting	226	225	230	228	909
Microeconomics	230	235	224	223	912
Total	456	460	454	451	1,821

Table 2. Distribution of number of students per experimental session

Course	# of Sessions	Average	Min	Median	Max
Accounting	16	56.81	47	57.5	66
Microeconomics	16	57.00	45	57.5	65
Total	32	56.91	45	57.5	66

Table 3. Effects of AI Assistance on Learning Outcomes

Dependent Variables	**Practice (1)**	**Quiz (2)**	**Perceptions (3)**	**Quiz − Practice (4)**
AI solves the problem (T1)	1.334***	−0.032	−0.371***	−1.365***
	(0.102)	(0.112)	(0.049)	(0.146)
AI clarifies mistakes (T2)	−0.037	0.108	0.029	0.143
	(0.102)	(0.115)	(0.044)	(0.137)
Interactive AI (T3)	0.582***	−0.056	0.047	−0.637***
	(0.101)	(0.115)	(0.047)	(0.139)
Session FE	Yes	Yes	Yes	Yes
Student FE	No	No	No	No
Clustered SE (student)	Yes	Yes	Yes	Yes
Adjusted R^2	0.281	0.104	0.071	0.186
No. of Observations	1,821	1,817	1,627	1,817
Students	411	411	407	411

Notes: *All coefficients are intent-to-treat estimates from linear regressions. The omitted reference category is no AI (T0). Outcomes are measured on their respective scales (practice and quiz scores: 0–10; perceptions: 1–5). All specifications include session fixed effects. Standard errors are clustered at the student level and reported in parentheses. Statistical significance is denoted as: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.*

Table 4. Student Fixed Effect Robustness Check

Dependent Variables	**Practice (1)**	**Quiz (2)**	**Perceptions (3)**	**Quiz − Practice (4)**
AI solves the problem (T1)	1.318*** (0.103)	−0.063 (0.111)	−0.378*** (0.050)	−1.379*** (0.146)
AI clarifies mistakes (T2)	−0.039 (0.102)	0.058 (0.114)	0.030 (0.044)	0.096 (0.136)
Interactive AI (T3)	0.572*** (0.100)	−0.113 (0.111)	0.044 (0.047)	−0.683*** (0.139)
Session FE	Yes	Yes	Yes	Yes
Student FE	Yes	Yes	Yes	Yes
Clustered SE (student)	Yes	Yes	Yes	Yes
Adjusted R^2	0.416	0.390	0.441	0.237
No. of Observations	1,813	1,808	1,603	1,808
Students	403	402	383	402

Notes: *All coefficients are intent-to-treat estimates from linear regressions. The omitted reference category is no AI (T0). Outcomes are measured on their respective scales (practice and quiz scores: 0–10; perceptions: 1–5). All specifications include session fixed effects and student fixed effects, so identification comes from within-student variation across topics and sessions. Standard errors are clustered at the student level and reported in parentheses. Statistical significance is denoted as: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.*

Table 5. Treatment × Accounting Robustness Checks

Dependent Variables	**Practice (1)**	**Quiz (2)**	**Perceptions (3)**	**Quiz − Practice (4)**
AI solves the problem (T1)	1.137***	0.062	−0.231***	−1.071***
	(0.117)	(0.148)	(0.065)	(0.182)
AI clarifies mistakes (T2)	0.051	0.263*	−0.003	0.212
	(0.123)	(0.141)	(0.060)	(0.171)
Interactive AI (T3)	0.542***	−0.019	0.057	−0.559***
	(0.129)	(0.155)	(0.062)	(0.185)
T1 × Accounting	0.403**	−0.190	−0.292***	−0.597**
	(0.196)	(0.225)	(0.089)	(0.267)
T2 × Accounting	−0.174	−0.311	0.056	−0.140
	(0.198)	(0.233)	(0.080)	(0.263)
T3 × Accounting	0.080	−0.076	−0.022	−0.157
	(0.192)	(0.227)	(0.084)	(0.257)
Session FE	Yes	Yes	Yes	Yes
Student FE	No	No	No	No
Clustered SE (student)	Yes	Yes	Yes	Yes
Adjusted R^2	0.283	0.103	0.076	0.187
No. of Observations	1,821	1,817	1,627	1,817
Students	411	411	407	411

Notes: *All coefficients are intent-to-treat estimates from linear regressions. The omitted reference category is no AI (T0). Outcomes are measured on their respective scales (practice and quiz scores: 0–10; perceptions: 1–5). All specifications include session fixed effects. Main treatment coefficients correspond to effects in* ***Microeconomics****, while interaction terms report the* ***differential effect in Accounting****. Standard errors are clustered at the student level and reported in parentheses. Statistical significance is denoted as: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.*

Table 6. Effect of treatment on time spent (seconds) in practice and quiz stage

Dependent Variables	Practice time spent		Quiz time spent		Total time spent	
	(1)	(2)	(3)	(4)	(5)	(6)
AI Solves the Problem (T1)	-21.259***	-23.901***	18.959**	20.439***	-2.252	-3.669
	(7.284)	(7.087)	(7.444)	(7.439)	(10.070)	(9.825)
AI clarifies mistake (T2)	-6.605	-9.735	-0.952	0.293	-7.773	-9.061
	(6.640)	(6.717)	(7.839)	(7.743)	(10.904)	(10.741)
Interactive AI (T3)	121.294***	117.190***	-24.251***	-23.091***	96.635***	93.890***
	(8.857)	(8.648)	(7.179)	(7.096)	(11.759)	(11.533)
Session FE	Yes	Yes	Yes	Yes	Yes	Yes
Student FE	No	Yes	No	Yes	No	Yes
Clustered SE (Student)	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.323	0.501	0.323	0.539	0.372	0.638
No. of observations	1,821	1,813	1,817	1,808	1,817	1,808
Students	411	403	411	402	411	402

*Notes: All coefficients are intent-to-treat estimates from linear regressions. The omitted reference category is no AI (T0). Outcomes, i.e., time spent, are measured in seconds. Robust standard errors are clustered at the student level and reported in parentheses. Statistical significance is denoted as: *** p < 0.01, ** p < 0.05, * p < 0.10.*

Table 7. Treatment effect on student fluency in quiz

Dependent Variable	**Fluency (quiz score/minutes spent in quiz)**	
	(1)	**(2)**
AI Solves The Problem (T1)	-0.058 (0.062)	-0.068 (0.058)
AI clarifies mistake (T2)	0.058 (0.066)	0.034 (0.067)
Interactive AI (T3)	0.131** (0.059)	0.124** (0.058)
Session FE	Yes	Yes
Student FE	No	Yes
Clustered SE (Student)	Yes	Yes
Adjusted R-squared	0.296	0.444
No. of observations	1,817	1,808
Students	411	402

Notes: *All coefficients are intent-to-treat estimates from linear regressions. The omitted reference category is no AI (T0). Outcomes, i.e., quiz fluency, is defined as the quiz score divided by the time (in minutes) a student spends in quiz. Standard errors are clustered at the student level and reported in parentheses. Statistical significance is denoted as: *** p < 0.01, ** p < 0.05, * p < 0.10.*

Appendix A. AI Tool Configuration

This appendix documents the configuration, constraints, and usage of the AI tools employed in the experiment. The purpose of this appendix is to clarify how AI assistance differed across experimental conditions, to provide transparency about the implementation, and to facilitate replication, while abstracting from implementation details that are not central to the study's identification strategy.

A.1 Underlying Model and System Configuration

For the purposes of experimental control, data collection, and implementation consistency, we developed customized AI tools for this study rather than relying on off-the-shelf commercial products such as ChatGPT.

All AI tools used in the experiment were built on the same underlying large language model, GPT-4o, which remained fixed throughout the duration of the study. No model updates, parameter changes, or architectural modifications occurred during the experimental period.

To ensure comparability across experimental conditions, the same base model was used for all AI-assisted treatments. Differences across treatments arose solely from constraints on permitted interactions and outputs, rather than from differences in model capability.

Sampling parameters affecting response stochasticity (e.g., temperature, top-p) were held constant across all AI modes and experimental sessions. These parameters were selected prior to the start of the experiment and were not adjusted over time.

- Temperature = 0.5
- Top-p = 0.95
- Frequency penalty = 0

By fixing sampling parameters, the study ensures that differences in student outcomes across AI conditions reflect differences in the structure of AI assistance rather than differences in response randomness or variability.

A.2 Scope and Representativeness of the AI Tool

The AI system used in this study is intended to be representative of generative AI tools plausibly available to undergraduate students at the time of the experiment, rather than a benchmark of frontier or optimal AI performance. The study does not attempt to evaluate the absolute quality, accuracy, or pedagogical optimality of the AI model itself.

Accordingly, the results should be interpreted as estimates of how different *modes of AI use* affect learning outcomes within a common instructional environment, not as claims about the educational

potential of state-of-the-art AI systems per se. This distinction is important for interpreting the findings as evidence about AI usage modality rather than AI capability.

A.3 Treatment-Specific AI Behavior and Constraints

System-level constraints were implemented to enforce the intended behavior of each AI mode and to prevent students from circumventing the assigned treatment structure. These guardrails were implemented through system prompts and response filters that restricted the types of outputs the AI could generate in each condition.

A.3.1 AI Solves the Problem (T1)

In the *AI Solves The Problem* condition (T1), students could submit practice questions to the AI system, which generated **final answers directly**, without explanation, hints, or intermediate reasoning steps. The system was explicitly restricted from providing guided solution paths, conceptual explanations, or partial reasoning even when explicitly prompted by students.

Importantly, AI-generated solutions were provided for **structurally similar problems**, rather than for the exact practice questions themselves. This design ensured that students still needed to map AI outputs to the specific problem context they were assigned, while substantially lowering the cost of producing correct answers during practice.

Representative example (T1):

> ***Student Inquiry:***
> "Which market likely has the most elastic short-run supply?
>
> a. 3D-printed phone cases
> b. Rental apartments
> c. Electricity generation tomorrow
> d. Stadium seats for tonight of a game"
>
> ***AI response*:**
> "a. 3D-printed phone cases"

A.3.2 AI Clarifies Mistakes (T2)

In the *AI clarifies mistakes* condition (T2), students were instructed to first complete the practice problems independently and submit their answers. After receiving their practice grades, students could consult the AI system to ask specific concepts aimed at clarifying specific mistakes or misunderstandings.

Representative example (T2):

Student's inquiry
"Are price elasticity and elasticity different?"

AI response
"Price Elasticity and Elasticity are related concepts, but they refer to different scopes depending on their context. Here's a brief overview to clarify:

1. Price Elasticity: - Specifically refers to the responsiveness of the quantity demanded or supplied of a good to changes in its price. - The most common type is Price Elasticity of Demand (PED), which measures how much the quantity demanded changes in response to a price change. - Similarly, there is Price Elasticity of Supply (PES), which measures the responsiveness of the quantity supplied to price changes.

2. Elasticity: - This is a broader term that describes how one variable responds to changes in another variable. - Besides price elasticity, there are other types of elasticity such as: - Income Elasticity of Demand: Measures how the quantity demanded changes as consumer income changes. - Cross-Price Elasticity of Demand: Measures the responsiveness of the demand for one good to a change in the price of another good.

In summary, while price elasticity is a specific type of elasticity focused on price changes, elasticity, in general, can refer to responsiveness related to various economic variables."

A.3.3 Interactive AI Learning (T3)

In the *interactive AI learning* condition (T3), students worked through practice problems while interacting with an AI tutor designed to provide step-by-step guidance, prompts, and hints. The system responded dynamically to student inputs but was explicitly prohibited from supplying final answers.

When students attempted to solicit complete solutions, the system redirected them toward reasoning steps, conceptual cues, or diagnostic questions. These constraints were designed to preserve active problem solving while reducing unproductive friction during practice.

Representative example (T3):

Student Inquiry:
"can you help me understand the concept of equilibrium?"

AI response:
"Absolutely, I'd be happy to help you understand the concept of "equilibrium" in microeconomics! Let's break it down step by step.

Basic Understanding

> Equilibrium in economics refers to a situation where supply and demand are balanced. This means that the quantity of a good or service that consumers are willing to buy equals the quantity that producers are willing to sell at a given price.

Let's start with a question:

What do you think happens to the price of a product if there is more demand than supply?"

Student Inquiry:
"So there will be more people want to buy, but not enough stuff, so it will be more expensive?"

AI Response:
"Exactly! When demand exceeds supply, there's competition among buyers, which tends to drive the price up. Now, let's consider the opposite scenario where supply is greater than demand. What do you think happens to the price in this case?"

Student Inquiry:
"seller can't sell out all of their output so probably there will be a discount on things"

AI response:
"That's correct! When supply exceeds demand, sellers often lower prices to encourage more purchases. In a market, prices will adjust due to these forces until supply equals demand. This point is called the equilibrium price and the quantity is the equilibrium quantity. Let's think about this: Why do you think reaching equilibrium is beneficial for both buyers and sellers?"

All AI tools were implemented with built-in safeguards to enforce treatment-specific behavior. These safeguards included restrictions on disallowed output types (e.g., final answers in T3), monitoring of AI responses for compliance, and session supervision by members of the research team.

The AI tools interacted only with *practice activities* and were not accessible during the quiz stage in any condition. Quiz questions were distinct from, but aligned with, practice questions at the topic level and required independent problem solving under uniform conditions.

Representative practice and quiz questions are provided in Appendix C.

Appendix B. Additional Statistics and Analysis

This appendix presents additional descriptive statistics and robustness checks for practice performance, quiz performance, and subjective learning outcomes.

Table B1. Summary Statistics on Practice, Quiz, and Survey

Panel A. All Courses

	All		No AI (t0)	AI Solves the Problem (t1)	AI Clarifies Mistakes(t2)	Interactive AI Learning (t3)
Output	#	Mean	Mean	Mean	Mean	Mean
Practice	1821	8.23	7.75	9.10	7.70	8.35
Quiz	1817	6.43	6.44	6.39	6.54	6.37
Survey	1627	3.33	3.41	3.04	3.44	3.46

Panel B. Accounting Courses

	All		No AI (t0)	AI Solves the Problem (t1)	AI Clarifies Mistakes (t2)	Interactive AI Learning (t3)
Output	#	Mean	Mean	Mean	Mean	Mean
Practice	909	7.71	7.19	8.76	7.07	7.86
Quiz	908	5.88	5.96	5.79	5.93	5.83
Survey	794	3.25	3.37	2.83	3.42	3.39

Panel C. Microeconomic Courses

	All		No AI (t0)	AI Solves the Problem (t1)	AI Clarifies Mistakes(t2)	Interactive AI Learning (t3)
Output	#	Mean	Mean	Mean	Mean	Mean
Practice	912	8.74	8.30	9.43	8.34	8.85
Quiz	909	6.99	6.92	6.97	7.17	6.91
Survey	833	3.42	3.47	3.23	3.46	3.52

Table B2. Subjective Learning Outcome: Robustness Check

Dependent Variables	**(1)** **Perceived Skills Learned**	**(2)** **Satisfaction with Learning**
AI Solves the Problem (T1)	−0.399*** (0.053)	−0.345*** (0.049)
AI clarifies mistakes (T2)	-0.000 (0.047)	0.051 (0.047)
Interactive AI (T3)	0.021 (0.048)	0.062 (0.051)
Session FE	Yes	Yes
Student FE	No	No
Clustered SE (student)	Yes	Yes
Adjusted R^2	0.065	0.067
Observations	1,636	1,634
Students	407	407

Appendix C. Representative Practice and Quiz Questions

This appendix provides representative examples of practice and quiz questions used in the experiment. The purpose of these examples is to illustrate the structure, cognitive demands, and alignment between practice and quiz tasks, rather than to exhaustively document all instructional materials.

Across all sessions, practice and quiz questions were aligned at the topic level but were not identical. Practice questions were completed under varying AI-assisted conditions depending on treatment assignment, whereas quiz questions were completed independently without AI assistance in all conditions.

C.1 Microeconomics Example Questions

Practice Question (Demand and Supply)

In a competitive market, a **surplus** occurs when:
A. Price is below equilibrium and Qd > Qs
B. Price is above equilibrium and Qs > Qd
C. Price equals equilibrium and Qd = Qs
D. Government sets a binding price ceiling

Quiz Question (Demand and Supply)

A celebrity campaign boosts preferences for smartphones, mobile data plans (a complement) get cheaper, and a mild recession lowers incomes. If supply stays the same, what happens to the new equilibrium price and quantity (ceteris paribus)?
A. Price rises; quantity rises
B. Price falls; quantity falls
C. Price is ambiguous; quantity is ambiguous
D. Price rises; quantity is ambiguous

Practice Question (Market Structure)

Suppose several new firms enter a perfectly competitive market (ceteris paribus). What is the most likely outcome for market price and total quantity?
A. Price rises; total quantity falls
B. Price falls; total quantity rises
C. Price and total quantity both fall
D. Price and total quantity both rise

Quiz Question (Market Structure)

In long-run equilibrium for a perfectly competitive industry with free entry/exit, which set of conditions holds?

A. P = MC > ATC; positive profit
B. P = MC = min ATC; zero economic profit
C. P > MC = min ATC; positive profit
D. P < AVC; zero output

C.2 Accounting Example Questions

Practice Question (Inventory)

FIFO seldom coincides with the actual physical flow of inventory.
A. Work in process
B. Raw materials
C. Merchandise inventory
D. Finished goods

Quiz Question (Inventory)

Which of the following statements is correct with respect to inventories?

A. The FIFO method assumes that the costs of the earliest goods acquired are the last to be sold.
B. It is generally good business management to sell the most recently acquired goods first.
C. Under FIFO, the ending inventory is based on the latest units purchased.
D. FIFO seldom coincides with the actual physical flow of inventory

Please note that the questions shown here are not exhaustive; they are intended to be representative of the instructional tasks students encountered during the experiment.

Appendix D. Student Perceptions of Learning and Satisfaction Survey

This appendix documents the survey instrument used to measure students' perceived learning and satisfaction with the instructional techniques employed during the experimental sessions, including AI-assisted learning tools.

D.1 Survey Instrument

Students were asked to indicate their level of agreement with the following statements based on their experience during the experimental session. Responses were recorded on a five-point Likert scale:

1 = Strongly Disagree
2 = Disagree
3 = Neutral
4 = Agree
5 = Strongly Agree

Perceived Skills Learned

1. The techniques used in class today helped me better understand key concepts in this course.
2. Using the techniques improved my ability to apply knowledge to real-world problems.
3. The techniques enhanced my critical thinking skills.
4. I gained new skills that will be useful in future courses or professional settings.
5. The techniques helped me learn at a deeper level compared to other class sessions.

Satisfaction with Learning

6. I am satisfied with the overall learning experience provided by the techniques used today.
7. The techniques kept me engaged and motivated throughout the learning process.
8. I would recommend the use of these techniques to other students.
9. The techniques met my expectations for supporting my learning.
10. I feel confident in what I learned as a result of using these techniques.

All items were administered immediately following the quiz stage of each experimental session.

D.2 Validity and Reliability Foundations

The survey items were adapted from well-established instruments commonly used in research on learning environments, educational technology, and online instruction. These instruments have been extensively validated in prior work and are widely used to measure perceived learning outcomes and student satisfaction.

Specifically, the item content draws on the following sources:

- Distance Education Learning Environments Survey (DELES)
    - *Reliability*: Cronbach's alpha values across subscales range from 0.75 to 0.93.
    - *Validity*: Supported through exploratory and confirmatory factor analyses in multiple samples.
- Online Learner Satisfaction Questionnaire (OLSQ)
    - *Reliability*: Cronbach's alpha values exceeding 0.85 across all scales.
    - *Validity*: Supported via convergent validity and factor-analytic evidence.
- E-Learning Satisfaction Scale (eLSS)
    - *Reliability*: Cronbach's alpha = 0.89.
    - *Validity:* Demonstrated strong construct validity through confirmatory factor analysis and discriminant validity tests.