Journal for Educators, Teachers and Trainers
The PROFESIOLab electronic, peer-reviewed, open-access Magazine

ISSN 1989-9572
DOI: 10.47750/jett.2026.17.01.08

A Replicable 15-Minute LLM Tutoring Exercise for Quantitative Problem Solving in a Flipped Classroom

Robert J. McKeown

Journal for Educators, Teachers and Trainers, Vol.17 (1)

https://jett.labosfor.com/

Date of reception: 02 February2026

Date of revision: 25 February2026

Date of acceptance: 05 March 2026

Robert J. McKeown (2026). A Replicable 15-Minute LLM Tutoring Exercise for Quantitative Problem Solving in a Flipped Classroom. ***Journal for Educators, Teachers and Trainers, Vol.17 (1) 132-146***

Journal for Educators, Teachers and Trainers
The LabOSfor electronic, peer-reviewed, open-access Magazine

A Replicable 15-Minute LLM Tutoring Exercise for Quantitative Problem Solving in a Flipped Classroom

Robert J. McKeown, Associate Professor, Teaching Stream
Department of Economics, York University

Email: mckeow99@yorku.ca

Abstract

Mathematical under-preparedness is a persistent barrier for many university students in quantitative disciplines, particularly in settings where instructional time is limited and individual support is scarce. This paper evaluates a brief, reusable classroom exercise that introduces students to using a large language model (LLM) as an on-demand, interactive tutor within a flipped classroom at a university or high school level. The activity takes approximately fifteen minutes and combines small-group problem solving, individual LLM-guided coaching on the same problem, a class discussion and evaluating feedback quality. We collect student evidence during and after the activity using an in-class worksheet and survey items on engagement, perceived learning, and the relative value of an LLM help compared with peers and existing course software. Students report high engagement and rate LLM assistance as at least as useful as in-class peer support, while describing the LLM as complementary to the course's learning platform. Reported learning experiences are more positive among students using more advanced LLM versions. These findings suggest that a short, structured in-class exercise can help students learn how to use LLMs productively for quantitative problem solving, and provides a practical template for instructors seeking to integrate generative AI into quantitative gateway courses.

Keywords: Economics education, Large language models, Generative AI, Socratic tutoring, Flipped classroom, Mathematics education

1. Introduction

Large, heterogeneous first-year quantitative courses face a persistent constraint: students need timely, individualized feedback during problem-solving, but instructor attention is scarce and unevenly distributed across learners. This feedback bottleneck motivates scalable supports that can prompt students to articulate steps, diagnose errors, and revise reasoning during practice. Large language models (LLMs), including ChatGPT, can generate worked solutions and explanations in accessible language, and

their conversational interface can approximate one-to-one tutoring by prompting students to justify steps, check assumptions, and iterate after errors. At the same time, the pedagogical challenge is not merely access to LLMs but also competent use, especially when the goal is learning and mastery of a subject. Students may underuse these tools, use them primarily for getting answers, or naively accept outputs without verification. In mathematics, where errors can be subtle and confidence is fragile, uncritical use may impede learning even when the final answer is correct. Beyond immediate course performance, LLM competence is increasingly a form of general-purpose tool literacy. Graduates are likely to encounter AI-mediated workflows in analysis, communication, and routine problem solving. Effectiveness depends on prompting and validating outputs. An introductory quantitative course is a natural setting to teach these skills because verification is feasible: solutions can be checked against definitions, algebraic identities, and boundary conditions. The instructional objective of this study and exercise is twofold: to help students learn course content, and to help them learn how to learn with an LLM in a way that transfers to future academic and professional tasks.

This paper addresses a practical gap in the emerging literature on generative AI in education: instructors need brief, replicable, in-class protocols that teach students how to use LLMs as learning tools rather than as answer engines, particularly in quantitative gateway courses. We introduce and evaluate a concise, instructor-supervised exercise designed to familiarize students with using an LLM as a Socratic-style mathematics tutor. The intervention takes approximately 15 minutes and consists of three components: (i) small-group attempts on a targeted calculus problem, (ii) an individual LLM dialogue in which students request hints, explanations, and error checks while documenting the interaction, and (iii) a whole-class debrief that surfaces common misconceptions and models verification practices (e.g., checking intermediate steps, testing special cases, and comparing against alternative solution paths). The exercise is designed to be reusable with many mathematics, statistics, or economics problems. It complements existing supports such as peer collaboration and other education technology such as an adaptive learning system (ALS).

We evaluate the exercise using student feedback and performance on formative and summative assessments. The paper documents student perceptions of LLM usefulness after the activity, benchmarks those perceptions against peer help and the adaptive learning system (ALS), evaluates heterogeneity across cohorts and model quality, Fall 2023 and Winter 2024 versus Fall 2024, and tests for changes in performance on related formative and summative assessments. Across cohorts, students report strong engagement with the activity. Students reported that the LLM produced a correct solution to the assigned problem 91 percent of the time, and they rated LLM assistance as approximately as helpful as their peers for completing the worksheet. In 2023–2024, students reported the ALS as more beneficial to learning than the LLM, while student views of LLM use increased substantially in Fall 2024. Taken together, these results suggest that a short, structured, instructor-supervised protocol can help students develop more productive patterns of LLM use, while positioning the tool as complementary to peer learning and course software rather than as a replacement.

The paper proceeds as follows. Section 2 provides an in-depth literature of active learning in flipped classrooms and the potential LLM tutoring. Section 3 describes the course context and the design of the in-class exercise. Section 4 presents the data and evaluation approach, and Section 5 discusses implications for teaching quantitative gateway courses and for building transferable AI tool literacy. The appendix provides four reproducible handouts for students: instructions for accessing Copilot, practical prompting tips for beginners, a brief guide to LaTeX notation for mathematical expressions, and an example of an exercise including the tutoring prompt used in class.

2. Literature Review

Evidence from workplace settings suggests that LLM assistance can raise productivity and accelerate early skill acquisition, particularly for less-experienced users. Large field and experimental studies report that access to LLMs improves task performance and can disproportionately benefit novices, consistent with a mechanism in which on-demand guidance reduces fixed costs of getting started. This pattern has been documented in customer-support work and in controlled writing tasks, respectively (Brynjolfsson et al., 2025; Noy & Zhang, 2023). In education, the relevant question is not whether LLMs can supply answers, but whether they can be structured to preserve the cognitive work that produces durable learning. Decades of research on human tutoring show large gains relative to conventional instruction, although effects vary by implementation and domain (Bloom, 1984; VanLehn, 2011).

Flipped Classrooms and Active Learning

Active-learning and flipped classroom designs justify reserving class time for application, feedback, and error correction rather than first exposure. Meta-analytic evidence shows that active learning improves achievement and reduces failure rates across STEM settings relative to traditional lecture, consistent with reallocating in-class time toward problem solving with immediate support (Freeman et al., 2014). Flipped instruction similarly tends to yield modest gains in learning outcomes and student satisfaction when it meaningfully shifts lower-level content delivery outside of class and uses class meetings for guided practice (Lo & Hew, 2017). Within this logic, a brief, instructor-supervised routine in which students first attempt a quantitative problem independently, then revisit it with a Socratic dialogue, and then debrief as a class is a natural instantiation of in-class time for application and feedback. This adoption baseline strengthens the case for supervised in-class norm-setting: students learn how tools can be used for disciplined practice rather than defaulting to unstructured out-of-class use.

Active and flipped approaches are effective to the extent that they increase time-on-task, elicit productive struggle, and deliver timely feedback that helps learners repair misconceptions before they harden (Freeman et al., 2014; Lo & Hew, 2017). Theoretical accounts emphasise that learning improves when students generate solutions themselves, explain their reasoning, and update beliefs in response to feedback, rather than merely receiving worked solutions (Chi et al., 1994). The ICAP framework predicts that constructive, observable engagement, such as generating solutions step-by-step or explaining concepts, yields larger learning gains than passive reception. (Chi & Wylie,

2014) LLM tutoring may matter primarily because it increases the likelihood that students persist with the worksheet, articulate intermediate reasoning, and obtain timely feedback at moments of confusion. Socratic tutoring is pedagogically consequential because it aims to preserve the causal mechanisms that make tutoring effective: eliciting reasoning, diagnosing gaps, and prompting learners to generate steps rather than copying answers. Foundational evidence suggests that tutoring can yield large improvements relative to conventional instruction (Bloom, 1984), and meta-analytic work indicates that the benefits of tutoring are strongly associated with features such as contingent questioning, feedback, and opportunities for students to explain their thinking (VanLehn, 2011). Using the Socratic method is therefore not an aesthetic preference but a mechanism-preserving design choice: prompting a tutor to avoid jumping to the final answer increases the probability that students engage in self-explanation and repair steps, a pathway supported by experimental work on self-explanation (Chi et al., 1994). Reviews linking Socratic dialogue to critical thinking likewise emphasise that disciplined questioning can shift learners from answer-getting to sense-making, especially when the dialogue targets assumptions and justifications rather than surface procedures (Mahoney et al., 2023). In a quantitative course, this matters because conceptual errors, such as interpreting derivatives or function behaviour, can be masked by correct algebra. Socratic prompts can surface the reasoning that differentiates durable understanding from fragile performance.

Technology-Aided Instruction, Generative AI, and Learning at Scale

Historically, intelligent tutoring systems (ITS) clarify which tutoring moves matter and which engineering constraints limited classroom scalability. Dialogue-based tutors such as AutoTutor were designed to sustain natural-language interaction that prompts explanation, supplies hints, and responds to student statements, demonstrating that conversational scaffolding can be engineered to elicit deeper processing (Graesser, 2016). Broader syntheses report that ITS can improve learning relative to business-as-usual instruction, with effects depending on how well systems implement adaptive feedback, step-level guidance, and opportunities for active responding (Kulik & Fletcher, 2016). From this perspective, the arrival of general-purpose LLMs relaxes engineering constraints such as the cost of authoring dialogue and anticipating student language. (Graesser, 2016; Kulik & Fletcher, 2016). A short, supervised in-class activity that frames an LLM explicitly as a Socratic tutor is thus consistent with the earlier ITS tradition. It foregrounds tutoring moves, such as asking questions, offering hints, and checking for understanding, rather than presenting the model as an omniscient authority.

Post-2022 research on LLMs in education is encouraging. LLMs can function effectively as tutors but results depend heavily on instructional design and the extent of student reliance. Systematic reviews and meta-analyses of experimental studies report positive average effects of LLM-supported instruction on learning, alongside substantial heterogeneity by implementation, discipline, and outcome measure (Deng et al., 2025; Schleicher, 2026). Recent classroom evidence also suggests that, under some conditions, an AI tutor can outperform well-executed active learning, as Kestin et al. (2025) found larger gains from a Gen AI tutoring condition than from an in-class active-

learning condition in a first-year physics course. This result strengthens the case for LLM-as-tutor designs, but it also implies that comparative effectiveness depends on the quality of the interaction, not simply answer acquisition. In mathematics, controlled evidence indicates that ChatGPT-generated hints can improve learning relative to no-hint conditions and can approach human-hint benchmarks, supporting structured hinting as a plausible mechanism (Pardos & Bhandari, 2024). At the same time, calculus evaluations document nontrivial error rates and occasional confident mistakes, particularly on tasks requiring careful symbolic reasoning or when prompts permit premature solution disclosure (Gandolfi, 2025). Together, these findings motivate an instructor-supervised routine in which students first attempt a problem unaided, then re-solve it using a Socratic dialogue prompt, then consolidate correct reasoning in a whole-class debrief.

A central instructional problem is improving learning per unit of class time given limited instructor attention. Evidence on technology-aided instruction suggests that gains hinge on design features that complement effortful cognition, especially attempt, feedback, and revision, rather than substituting for it. Generative AI intensifies this trade-off: LLMs can supply responsive explanations and dialogue at scale, but unguided use can short-circuit productive struggle and weaken durable skill development. Randomized evidence on personalized education technology (edtech) provides a benchmark for what well-designed, individualized support can achieve over short horizons. In a lottery-based evaluation of a technology-aided tutoring program, sizable achievement gains accrued over several months, with relatively larger benefits for initially weaker students, consistent with the view that targeted scaffolding can be equity-enhancing when it increases time-on-task at an appropriate difficulty level (Muralidharan et al., 2019). The relevant analogue for LLM-as-Socratic-tutor designs is not generic access to explanations, but structured interaction that pushes students to articulate steps and confront errors. Large-scale evidence on adapting a successful personalized adaptive learning model into routine schedules shows meaningful gains and highlights platform time as a low-cost proxy for fidelity, underscoring the value of observable engagement metrics in scale-ups (Muralidharan & Singh, 2025). For classroom LLM tutoring, comparable indicators include verified completion of an unaided first attempt, documented revisions after model questioning, and time spent in constrained dialogue tied to the targeted concept.

The GenAI tutoring literature further distinguishes short-run performance support from durable learning. Field evidence in secondary mathematics shows that access to a generative AI tutor can raise contemporaneous performance, while unrestricted chat-style access can reduce subsequent performance once the tool is removed; interface guardrails that slow solution delivery and preserve learning-relevant cognition mitigate these harms (Bastani et al., 2025). This finding directly motivates Socratic constraints in quantitative courses, which requires the student to generate an answer and reflect on immediate feedback. Related results from reskilling and workplace settings reinforce the same mechanism: GenAI can function as a productivity boost, improving output quality while available without proportional memory retention once removed (Wiles et al., 2024). Accordingly, evaluation of an LLM-as-tutor exercise should focus on post-intervention reasoning and student capability without AI. Designs that delay AI

assistance until after student commitment appear especially promising. Evidence from a large-scale deployment where generative AI activates only after students submit an answer and then supports debriefing and follow-up questions reports increased engagement and performance, with patterns consistent with longer-run skill development and larger gains for lower-performing students (Kim et al., 2025). This aligns closely with an instructor-supervised Socratic routine: attempt first, interrogate reasoning second, and close with verification and synthesis.

3. An Exercise with an LLM-tutor in a Flipped Classroom

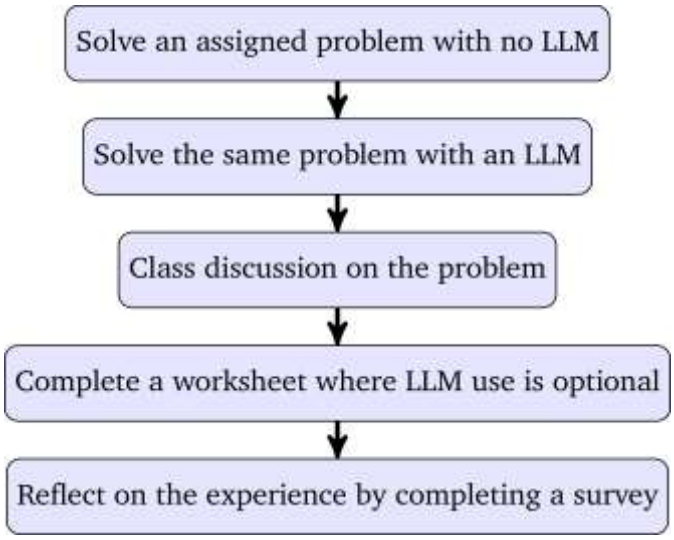
Exercise description

Before class, students are reminded to bring a laptop and are provided with short guidance on accessing an LLM, basic usage norms, and a reference sheet for entering mathematical notation using LaTeX. Appendix [App_A] and Appendix [App_B] provide the materials distributed to students. The in-class exercise follows the workflow summarized in Figure 1. First, students work in small groups to solve an assigned problem without an LLM and submit an initial answer through the learning management system (LMS). This initial attempt is intentional. It encourages students to engage in retrieval and problem representation, and it creates a concrete record of their reasoning that can be compared with subsequent feedback. It also reduces the likelihood that the LLM interaction becomes answer transcription by ensuring that students begin with a hypothesis, a partial solution path, or a specific point of confusion.

Second, students solve the same problem using an LLM. Students receive a written prompt from the instructor that commands the model to function as a Socratic style tutor rather than an answer generator. The prompt directs the LLM to elicit step-by-step reasoning, ask targeted follow up questions, and diagnose errors or missing justifications. Because the usefulness of LLM feedback is sensitive to the specificity of the input, students are given a prepared prompt that embeds the problem and establishes expectations for Socratic style guidance. An example prompt is provided in Appendix [App_C].

Third, I facilitate a short whole class discussion that compares solution approaches, surfaces common misconceptions, and models verification practices. This debrief is a central component of the protocol because it reinforces the norm that LLM output is provisional and must be evaluated against mathematical definitions and intermediate steps. It also consolidates learning by resolving errors that may persist after the LLM interaction and by making effective validation strategies visible to students.

Figure 1: In Class Exercise Workflow



Following the exercise, students complete a worksheet as a formative assessment in the flipped classroom format. Students are permitted to use an LLM during this practice, but the worksheet is designed so that correct completion requires reasoning and verification rather than transcription of an answer. After submitting the worksheet responses, students complete a brief survey reflecting on the usefulness of the LLM, the quality of its feedback, and how the experience compared with other course supports.

Implementation and design decisions

I integrated the exercise into classes 6 through 12 for a total of six implementations.[1] In the first four weeks, the course emphasizes review of prerequisite material using an adaptive learning system (ALS) that provides feedback and worked solutions for precalculus but does not cover the calculus content introduced later in the term. I introduced the LLM exercise at the point where calculus topics begin, with the specific aim of providing a scalable source of step by step feedback during practice and of teaching students how to use an LLM as a learning tool. Several constraints guided the design. First, baseline familiarity with LLMs was limited at the start of the course, and many students had little experience using an LLM for structured learning rather than for casual querying. To reduce start-up costs and preserve instructional time, I provided a pre-written prompt that students could copy and paste into the LLM interface from the university learning management system (LMS). This standardization also reduced variation in student experiences that would otherwise arise from heterogeneous prompting skill, allowing class discussion to focus on mathematical reasoning and verification rather than on interface navigation.

Second, I selected problems that were difficult enough that students would benefit from scaffolding, but not so complex that the model would frequently fail, resort to opaque reasoning, or take too much class time. I also prioritized problems with objectively verifiable answers, which made it feasible to emphasize validation and error checking

[1] The course follows a 12 week schedule and includes an in class term test in week 10.

during the debrief. Third, I chose an LLM that students could access at no cost and with minimal friction. During the study period, Copilot, which provided access to the latest ChatGPT models and designed for deeper reasoning, offered a practical balance between performance and accessibility.[2] During each implementation, two teaching assistants were present to support students with technical issues and mathematical questions. After the structured exercise and debrief, students transitioned to the regular worksheet. This integration strategy allowed LLM use to be introduced without changing the curriculum or assessment structure, while making the use of the tool an explicit, teachable component of quantitative learning rather than an informal and unsupervised practice.

4. Data Collection and Results

Student feedback was collected at two levels: (i) brief exercise-level surveys administered after each in-class LLM activity where students recorded exercise observations, and (ii) an end-of-course survey. Exercise-level items measured (a) perceived correctness of the LLM solution: never correct; correct after one or more attempts; correct on the first attempt; did not use; no response, and (b) perceived helpfulness of the LLM relative to groupmates: less; equal; more. The end-of-course survey elicited software preferences comparing the course adaptive learning system (ALEKS) and LLMs, and asked whether the course should include ALS, LLMs, both, or neither.[3] Descriptive results are reported as proportions by academic year. The analyses are descriptive and associational, as students were not randomly assigned to LLM use.

Table 1. Correctness of the LLM

	2023-24		2024-25 Fall	
Correctness and attempts used	**Frequency**	**Percent**	**Frequency**	**Percent**
I did not use a LLM.	80	15.5	26	11.3
No Response	22	4.23	10	4.3
No, it never got the correct answer.	19	3.7	4	1.7
Yes, but it took one or more attempts.	148	28.7	32	13.4
Yes, on the first attempt.	246	47.8	159	68.8
Total	515	—	231	—

[2] Access conditions changed over time, including whether an account was required and the number of daily uses available.

[3] This study received ethics approval from York University's Office of Research Ethics, Human Participants Review Sub-Committee, for the project "Large Language Models as Personal Math Tutor in a Flipped Classroom" (Certificate e2023-221; initial approval July 13, 2023; renewed August 20, 2024 to August 20, 2025). Participation was voluntary, students provided informed consent, and responses were analyzed and reported in a de-identified, aggregate form.

Table 2. Helpfulness comparison by class and year

Compared to my groupmate(s), the LLM was ...	Class 1		Class 2		Class 3		Class 4	
	’24	’25	’24	’25	’24	’25	’24	’25
Equally helpful	38	13	34	10	35	15	37	14
Less helpful	14	5	15	4	5	2	8	3
More helpful	11	9	8	5	11	3	2	4
Total responses	63	27	57	19	51	20	47	21

	Class 5		Class 6				Total
	’24	’25	’24	’25	2023-24	2024-25	
Equally helpful	29	15	32	13	205	80	285
Less helpful	5	3	3	2	50	19	69
More helpful	4	3	6	3	42	27	69
Total responses	38	21	41	18	297	126	423

Note. Data presented were collected in Fall 2023, Winter 2024 and Fall 2024.

The LLM tutor was evaluated according to its correctness, usefulness compared with classmates and compared to the course software. The evaluation of the LLM’s success rate for solving assigned problems is presented in Table 1. This illustrates a notable level of effectiveness, with a significant portion of students successfully receiving solutions from an LLM on the first attempt or second attempt. The first-try success rate improved significantly from 47.8 to 68.8 percent in the second academic year, likely reflecting improved LLM capability. A minority of students chose not to participate in the LLM exercise, but few students tried and failed to use the LLM successfully. The summarized evaluations in Table 2 provide additional insights into students’ perceptions of the helpfulness of LLMs by comparing using an LLM with their groupmates during in-class worksheets. A majority of students found the LLM to be equally helpful as their groupmates, while a smaller proportion found it more or less helpful. Building on this impression, an end-of-course survey asks students which software better helped their

learning: the ALS or the LLM. Students preferred ALS by a modest majority in 2023-24. However, there is a slight preference for the LLM in 2024-25. See the right panel in Figure 2. Another question asked which software should stay in the course, a substantial majority preferred both. See the left panel in Figure 2.

Table 3. Which software had a greater impact on your overall learning in this course?

	2023-24		2024 Fall	
Software	**Frequency**	**Percent**	**Frequency**	**Percent**
ALS (ALEKS Math)	71	63.39	21	35.00
LLMs (ChatGPT or Bing AI)	41	36.61	39	65.00
Total	112	100.00	60	100.00

Figure 2: Which software had a greater impact on your overall learning

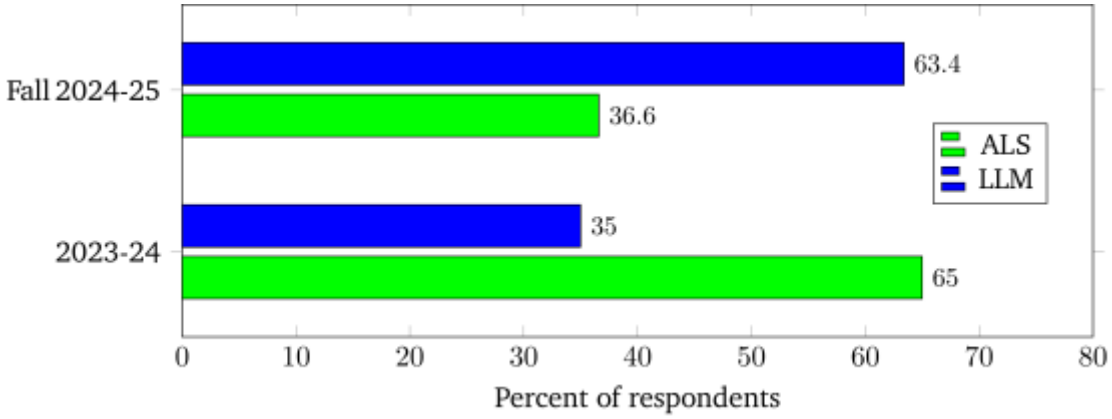
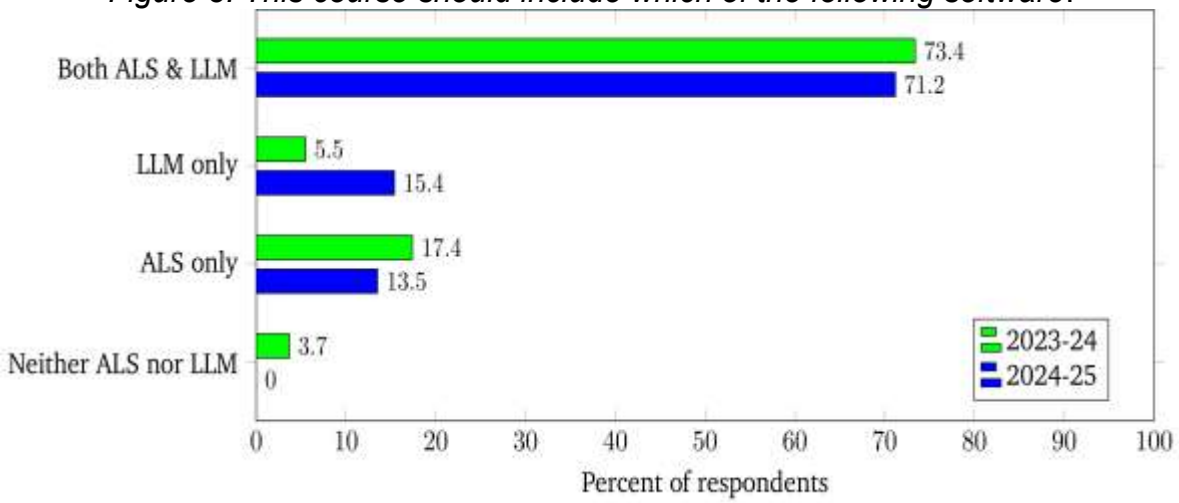


Table 4. This course should include:

	2023-24		2024 Fall	
Option	**Frequency**	**Percent**	**Frequency**	**Percent**
Both ALS and LLM	80	73.39	37	71.15
LLM only	6	5.50	8	15.38
ALS only	19	17.43	7	13.46
Neither ALS nor LLM	4	3.67	0	0.00
Total	109	100.00	52	100.00

Figure 3: This course should include which of the following software:



5. Discussion and Conclusion

This study evaluates a brief, replicable in class exercise routine that introduces first year economics students to using a large language model (LLM) as a Socratic style tutor. The central result is practical rather than technological. A time bounded, instructor supervised protocol can incorporate LLM tutoring into a standard active learning class without displacing core instructional time, while generating interpretable student feedback. Across cohorts, students reported that the LLM usually produced a correct solution to the assigned problem and that its help was approximately as useful as assistance from their groupmates. This LLM exercise matters because students will encounter Gen AI mediated workflows beyond the course, and the exercise teaches verification and questioning as transferable practices.

A second, equally important pattern is that students did not treat LLM tutoring and the ALS as substitutes. Students reported valuing both systems, and a majority indicated that both should remain in the course or that they were similarly useful for learning. This preference is consistent with a complementarity interpretation in which each tool addresses a different part of the feedback and practice problem faced by heterogeneous quantitative classes. The ALS provides structured practice, low friction input, immediate correctness feedback, and sequenced mastery oriented progression. These features are well matched to building fluency and persistence in prerequisite skills, particularly for novices. By contrast, LLM tutoring is flexible and conversational. It can generate alternative explanations, help students debug intermediate reasoning, and respond to specific questions that arise during problem solving, including for topics that fall outside the ALS content coverage. In this sense, the LLM is most valuable when it functions as a responsive scaffold that supports explanation, error diagnosis, and persistence during moments of confusion.

The year-to-year shift in perceived usefulness further underscores that the pedagogical value of LLM tutoring depends on the reliability and usability of the platform students encounter. In 2023–24, respondents more often credited the ALS than the LLM with improving their learning, whereas in Fall 2024–25 the pattern reversed. This change coincided with higher reported first attempt success on the LLM assisted problem. Because cohorts were not randomly assigned and multiple factors changed over time, including student composition and students' broader familiarity with LLMs, the shift should not be interpreted as evidence of a causal effect. Nonetheless, the pattern is consistent with a simple scope condition. When the tool is more reliable and easier to use, students are more likely to experience it as a credible source of feedback and explanation, and less likely to encounter errors that undermine trust or waste time.

These findings align with engagement based accounts of learning in quantitative courses. The exercise was designed to preserve productive struggle by requiring an initial attempt without AI, and to channel subsequent LLM use toward constructive and interactive engagement through a Socratic prompt and a whole class debrief. Under this design, LLM tutoring is not the learning outcome. It is a mechanism for increasing the likelihood that students persist, articulate intermediate reasoning, and receive feedback at the moment it is needed. The debrief then serves a second function that is central in an LLM mediated environment, establishing verification norms. Students observe how to evaluate a proposed solution against definitions, intermediate steps, and independent checks, and they see that model output is provisional rather than authoritative. Several limitations qualify the conclusions and point to next steps. Measures of LLM correctness and helpfulness are self reported and tied to the assigned problem used in the exercise rather than independently audited solution accuracy across many tasks. Students were not randomly assigned to different tools or prompting conditions, and platform choice and model quality likely varied across students. The intervention is also intentionally brief, so its most direct effects may be on process outcomes such as help seeking, calibration, persistence, and AI tool literacy, which are only imperfectly captured by standard course assessments. Future work could strengthen identification by randomly allocating sections or tasks to guided LLM use using a standardized Socratic prompt versus unguided LLM use, and by evaluating near transfer outcomes that capture the targeted process skills, including stepwise explanation and error diagnosis.

Despite these constraints, the evidence supports adopting short, supervised LLM as tutor activities in quantitative gateway courses as a supplement to existing supports and as an explicit introduction to effective LLM use. The design implications are straightforward: require an initial attempt without AI to preserve retrieval and productive struggle then offer a Socratic Method tutor that will require student engagement in a step-by-step process that forces the student to solve the problem. Pre-prepared Socratic prompts standardize the quality of the interaction and offer the student an example of effective prompting. Close the exercise with a whole class debrief that verifies the solution and makes evaluation practices explicit. Implemented this way, LLM tutoring functions less as an answer engine and more as a scaffold within active learning. It helps students practice how to ask precise questions, interpret feedback, and diagnose errors, while preserving instructor control over correctness and pacing. The broader payoff is not only improved support during problem solving, but the

development of transferable AI literacy that students can carry into future courses and professional settings.

References

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (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.2422633122

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16. https://doi.org/10.3102/0013189X013006004

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, *140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477. https://doi.org/10.1016/0364-0213(94)90016-7

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, *227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Gandolfi, A. (2025). GPT-4 in education: Evaluating aptness, reliability, and loss of coherence in solving calculus problems and grading submissions. *International Journal of Artificial Intelligence in Education*, *35*(1), 367–397. https://link.springer.com/article/10.1007/s40593-024-00403-3

Graesser, A. C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education*, *26*(1), 124–132. https://link.springer.com/article/10.1007/s40593-015-0086-4

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, *15*(1), 17458. https://www.nature.com/articles/s41598-025-97652-6

Kim, D., Mitrofanov, D., Wen, Q., & Xu, T. (2025). Generative AI Can Improve Performance and Engagement without Harming Learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5929576

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, *86*(1), 42–78. https://doi.org/10.3102/0034654315581420

Lo, C. K., & Hew, K. F. (2017). A critical review of flipped classroom challenges in k-12 education. *Educational Research Review*, *22*, 1–18. https://link.springer.com/article/10.1186/s41039-016-0044-2

Mahoney, B. B., Oostdam, R. R., Nieuwelink, H. H., & Schuitema, J. J. (2023). Learning to think critically through socratic dialogue: Evaluating a series of lessons designed for secondary vocational education. *Thinking Skills and Creativity*, *50*, 101422. https://www.sciencedirect.com/science/article/pii/S1871187123001906

Muralidharan, K., & Singh, A. (2025). *Adapting for scale: Experimental evidence on technology-aided instruction in india*. National Bureau of Economic Research Working Paper 34205. https://www.nber.org/papers/w34205

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in india. *American Economic Review*, *109*(4), 1426–1460. https://doi.org/10.1257/aer.20171112

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *Plos One*, *19*(5), e0304013. https://doi.org/10.1371/journal.pone.0304013

Schleicher, A. (2026, January 19). *How to effectively use Generative AI in education*. OECD. https://www.oecd.org/en/blogs/2026/01/how-to-effectively-use-generative-ai-in-education.html

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Wiles, E., Krayer, L., Abbadi, M., Awasthi, U., Kennedy, R., Mishkin, P., Sack, D., & Candelon, F. (2024). GenAI as an Exoskeleton: Experimental evidence on knowledge workers using GenAI on new skills. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4944588