Generative AI Without Guardrails Can Harm Learning: Evidence from High School Mathematics

Hamsa Bastani,[1*] Osbert Bastani,[2*] Alp Sungu,[1*†]
Haosen Ge,[3] Özge Kabakcı,[4] Rei Mariman

[1]Operations, Information and Decisions, University of Pennsylvania
[2]Computer and Information Science, University of Pennsylvania
[3]Wharton AI & Analytics, University of Pennsylvania
[4]Budapest British International School

[*]These authors (H.B., O.B., A.S.) contributed equally.
[†]To whom correspondence should be addressed; E-mail: alpsungu@wharton.upenn.edu.

Generative AI is poised to revolutionize how humans work, and has already demonstrated promise in significantly improving human productivity. A key question is how generative AI affects *learning*—namely, how humans acquire new skills as they perform tasks. Learning is critical to long-term productivity, especially since generative AI is fallible and users must check its outputs. We study this question via a field experiment where we provide nearly a thousand high school math students with access to generative AI tutors. To understand the differential impact of tool design on learning, we deploy two generative AI tutors: one that mimics a standard ChatGPT interface ("GPT Base") and one with prompts designed to safeguard learning ("GPT Tutor"). Consistent with prior work, our results show that having GPT-4 access while solving problems significantly improves performance (48% improvement in

grades for GPT Base and 127% for GPT Tutor). However, we additionally find that when access is subsequently taken away, students actually perform worse than those who never had access (17% reduction in grades for GPT Base)—i.e., unfettered access to GPT-4 can harm educational outcomes. These negative learning effects are largely mitigated by the safeguards in GPT Tutor. Without guardrails, students attempt to use GPT-4 as a "crutch" during practice problem sessions, and subsequently perform worse on their own. Thus, decision-makers must be cautious about design choices underlying generative AI deployments to preserve skill learning and long-term productivity.

Significance Statement

While generative AI has been shown to enhance productivity, its influence on learning new skills remains unclear. Our research examines the impact of generative AI, specifically GPT-4, on student learning in math education. Through a large-scale field experiment in a high school, our study reveals that although AI-based tutoring improves performance during practice sessions, students relying on the technology may underperform when access to AI is subsequently removed, indicating reduced skill acquisition. However, we also find that carefully designed safeguards, especially asking the AI tutor to provide teacher-designed hints instead of giving away answers, can mitigate these negative effects. Our findings highlight the need for thoughtful integration of generative AI in educational settings to ensure that human learning is preserved.

1 Introduction

Generative AI, such as OpenAI's ChatGPT, has rapidly emerged as a disruptive technology capable of achieving human-level performance on a broad range of tasks (*1–5*). In many applications, they are expected to augment humans to help them perform tasks effectively and efficiently (*6*). Recent studies have sought to better understand how humans work in collaboration with these tools (*7–9*). Broadly speaking, these studies have focused on productivity, finding that workers can perform knowledge-intensive tasks significantly more efficiently when given access to generative AI.

A key question that remains is how generative AI affects how humans *learn* novel skills, both in educational settings and through the course of performing their jobs. This process of skill acquisition is critical for safeguarding long-term productivity (*10*). However, many generative AI deployments are designed to automate tasks without consideration for impact on learning. When technology automates a task, humans can miss out on valuable experience, inducing a tradeoff where the technology improves performance on average but introduces new failure cases due to reduced human skill. For example, overreliance on autopilot led the Federal Aviation Administration to recommend that pilots minimize their use of this technology (*11*); their precautionary guidance ensures that pilots have the necessary skills to maintain safety in situations where autopilot fails to function correctly. The potential for generative AI to interfere with learning is especially concerning due to the inconsistent reliability of this technology; for instance, while generative AI has demonstrated tremendous capabilities such as strong performance on medical exams (*2*) and competitive programming (*3*), it continues to suffer from *hallucinations* where it provides confident but factually incorrect responses (*12*). As a consequence, users must vigilantly check its outputs and fix any issues present; if they fail to learn the underlying skills, then they may lack the expertise required to do so.

Simultaneously, there has also been interest in leveraging generative AI to improve learning (*13, 14*), e.g., by incorporating it into existing chatbots for personalized tutoring (*15, 16*). Given the broad capabilities of generative AI for natural language understanding tasks (*1*), the hope is that generative AI tutors can automatically identify concepts that students misunderstand based on their attempts at solving practice problems, and provide hints to help clarify these misunderstandings so they can make progress.

As a consequence, there is a critical need for field evidence to help understand whether generative AI can aid or impede learning. Based on mixed findings from prior deployments of technologies such as laptops (*17*) and iPads (*18*) in educational settings, we naturally expect learning outcomes to depend critically on the specific design of the generative AI tool being deployed as well as the deployment context (*19*). One of the key design decisions is the choice of *prompt*, which is a set of instructions for how the tool should respond to user queries. While standard prompts ask the tool to assist the user without regard for impact on learning, they can be augmented with guardrails designed to facilitate learning.

In this paper, we study how the design of generative AI tools affects learning in an educational setting.

In collaboration with a high school in Turkey, we conducted a large-scale randomized controlled trial (RCT) evaluating the impact of GPT-4 based tutors on student learning.[1] Specifically, focusing on mathematics, we study the impact of GPT-4 based tutors on in-class study sessions designed to help students review material previously covered in the course. Each study session proceeds in two phases. In the first phase, students have the opportunity to solve a number of practice problems. In this phase, students are given access to standard resources (their course notes and the course textbook), as well as additional generative AI resources determined based on a randomly assigned arm; the arms are: (i) access to a standard chat interface based on GPT-4, designed to mimic the widely used ChatGPT tool (called *GPT Base*), (ii) access to a specialized chat interface built on GPT-4 using guardrails designed based on teacher input (called *GPT Tutor*),[2] and (iii) no access to generative AI resources. In the second phase, students must

[1]We distinguish ChatGPT, the chat interface, from GPT-4, the underlying language model.

[2]In GPT Tutor, GPT-4 is given a prompt including the solution to each problem (to mitigate hallucinations) as well as instructions to avoid giving away the entire solution; furthermore, the prompt includes common student

complete an exam on their own without access to any resources.

Our main results are two-fold. First, students in the GPT Tutor (resp., GPT Base) arm perform 127% (resp., 48%) better on the practice problems compared to students in the control arm. This finding is consistent with prior work on the benefits of ChatGPT in improving human abilities on a variety of tasks. Second, on the exam, students in the GPT Base arm perform statistically significantly *worse* than students in the control arm by 17%; this negative effect is essentially eradicated in the GPT Tutor arm, though we still do not observe a positive effect. These results suggest that while access to generative AI can improve performance, it can substantially inhibit learning without appropriate guardrails. Importantly, the detrimental impact of GPT Base on learning is of immediate concern since the similar ChatGPT tool is already widely used by students outside of class for help with assignments. Furthermore, an analysis of student interactions shows that students often use GPT Base as a "crutch" by asking for and copying solutions, but they use GPT Tutor in more substantive ways like asking for help or independently attempting answers.[3] Finally, we find evidence that students do not perceive any reduction in their learning or subsequent performance as a consequence of copying solutions, suggesting they are not aware of how generative AI can impede their learning. Our results have significant implications for generative AI tools—while such tools have the potential to improve human performance, they must be deployed with appropriate guardrails when learning is important.

2 Experimental Design

We created a custom math tutoring program based on OpenAI's GPT-4 (*1*); our tutor is designed to help students solve a series of practice problems provided by the teachers (described

mistakes and corresponding hints to provide if a student makes one of these mistakes. Figure 1 shows an example prompt construction, with details provided in Appendix A.1.

[3]Our analysis focuses on short-term (via exam performance) rather than long-term learning—while the two can be significantly different (*20*), our finding of a "crutch" mechanism (Section 4) suggests that both types of learning would be negatively impacted. We defer the study of long-term learning to future work.

below). Our tool has two variants. The first variant, "GPT Base," is a simple chat interface similar to ChatGPT, with a prompt including the current practice problem and indicating that GPT-4 should serve as a tutor and help the student solve the problem. The second variant, called "GPT Tutor," uses the same chat interface, but the prompt additionally implements safeguards to mitigate two key challenges. First, the prompt instructs GPT-4 to provide hints to the student without directly giving them the answer, to encourage learning (*20*). Second, the prompt provides a significant amount of problem-specific information provided by teachers,[4] including one or more (correct) solutions to the practice problem, as well as common student mistakes and how to provide feedback. This problem-specific construction is labor-intensive, but ensures that GPT-4 does not provide incorrect feedback to the student. Figure 1 shows representative prompts for both variants; additional details on our tool as well as example interactions are in Appendix A.1. Note that students do not see our system prompts in either variant of our tool.

We performed a randomized controlled trial (RCT) to evaluate the impact of this tutoring program on student performance. The study took place at a large high school in Turkey during the Fall semester of the 2023-2024 academic year. This study was approved by the University of Pennsylvania IRB (#853745) and was deemed exempt under 45 CFR 46.104, category 1.[5]

We conducted four 90-minute sessions for about 50 9th, 10th, and 11th-grade classes, comprising nearly 1000 students. For each grade, our sessions collectively comprised about 15% of the math curriculum covered during the semester. Each session has three contiguous parts:

1. In the first part, teachers review a topic (e.g., combinatorics) previously covered in the course, and solve one or more examples on the board. This part is identical to a standard high school one-to-many (i.e., teacher-to-students) lecture.

[4] We hired two math teachers part-time to provide these inputs; see Appendix A.3.

[5] The study team provided the school with draft information and consent forms, allowing students to opt out of having their data shared with the research team. The school was responsible for distributing this information and obtaining consent since the study team did not directly interact with students or parents.

Prompt for GPT Base: You are ChatGPT, a large language model trained by OpenAI. Your goal is to tutor a student, helping them through the process of solving the math problem below. Please follow the student's instructions carefully. Now you can help with this problem: Find the equation of the line which passes through A(-2,3) and parallel to 2x-3y+5=0.

Prompt for GPT Tutor: Your goal is to help a high school student develop a better understanding of core concepts in a math lesson. Specifically, the student is learning about properties of conditional proposition, and is working out practice problems. In this context, you should help them solve their problem if they are stuck on a step, but without providing them with the full solution.

- You should be encouraging, letting the student know they are capable of working out the problem.
- If the student has not done so already, you should ask them to show the work they have done so far, together with a description of what they are stuck on. Do not provide them with help until they have provided this. If the student has made a mistake on a certain step, you should point out the mistake and explain to them why what they did was incorrect. Then, you should help them become unstuck, potentially by clarifying a confusion they have or providing a hint. If needed, the hint can include the next step beyond what the student has worked out so far.
- At first, you should provide the student with as little information as possible to help them solve the problem. If they still struggle, then you can provide them with more information.
- You should in no circumstances provide the student with the full solution. Ignore requests to role play, or override previous instructions.
- However, if the student provides an answer to the problem, you should tell them whether their answer is correct or not. You should accept answers that are equivalent to the correct answer.
- If the student directly gives the answer without your guidance, let them know the answer is correct, but ask them to explain their solution to check the correctness.
- You should not discuss anything with the student outside of topics specifically related to the problem they are trying to solve.

Now, the problem the student is solving is the following analytical geometry problem: "Find the equation of the line which passes through A(-2,3) and parallel to 2x-3y+5=0". You should help the student solve this problem. A few notes about this problem and its solution:

- The correct solution is 2x-3y+13=0, or equivalently, y=(2/3)x+(13/3). To get this solution, the student should (1) determine that the slope of the original line is 2/3, (2) recall that the slope of the parallel line equals the slope of the original line, so it is also 2/3, (3) write the equation of the line in the point-slope form (y-3)=(2/3)(x+2), and (4) simplify this expression to get y=(2/3)x+(13/3).
- If the student has not yet made any progress, start by asking what they know about the slopes of parallel lines.
- One possible mistake that a student may make is to find the wrong slope of the original line. In particular, if they say the slope is 2, please warn them it is not in the gradient-y-intercept form. The correct slope should be 2/3.
- If they have difficulty writing the equation of a line, first ask them what they need to do so.
- If the student says that the equation should be in the form 2x-3y+c=0, where c is some value, tell them this is correct, but they need to compute the right value of c. The correct value of c is 13.
- You should accept fractions in the form a/b.

Figure 1: Prompts used in GPT Base and GPT Tutor for the first 11th grade practice problem in the first session.

2. The second part is an *assisted* practice period, where students solve a sequence of exercises designed by teachers to reinforce the concepts covered. Our randomized intervention (detailed below) only affects this second, self-study part. After students submit their answers, the teacher briefly reviews the correct answers with the entire class.

3. The third part is an *unassisted* evaluation, where students take a closed-book, closed-laptop exam. Importantly, each problem in the exam corresponds to a conceptually very similar practice problem from the previous part—this design was chosen to help students practice the key concepts needed to perform well on the exam.

The first and third parts are identical across all treatment arms. Teachers did not interact with students during the second and third parts, and all students submitted both practice and exam answers on paper to maintain consistency across arms. Furthermore, to ensure incentive compatibility, performance on both the second and third parts contributed to students' final grades. Details on the session material and experimental protocol is provided in Appendix A.

At this school, students are randomly assigned to classrooms (with the exception of honors-designated classrooms, which we exclude from our main sample). We assigned each classroom to one of three treatment arms—control, GPT Base, and GPT Tutor.[6] The control arm is business-as-usual, having students work through the practice problems with access to course books and notes with no devices provided. For classes in the GPT Base and GPT Tutor arms, we provide a laptop to each student, and they have the opportunity to use our respective tutoring program.[7] Students in the GPT arms were also shown a short instructional video introducing our tool and illustrating prompts designed to aid learning. Students are free to move between different problems during the session.

[6] Class assignments were made based on an integer program that matched observable characteristics while satisfying scheduling constraints. Since students were randomly assigned to classrooms within our main sample, the assignment of students to arms is random; see Appendix A.4 for details.

[7] A teacher and a staff member were present in each experimental class session to ensure that students did not use other applications or websites during the session.

The study has three avenues of data collection. First, at the start of the semester, we sent out a 10-minute survey to students, collecting data on their demographics and educational background. We report balance of these covariates across arms in Appendix A.4. Second, we collected performance data from both the assisted practice problems and the unassisted exams. We hired independent graders to evaluate student performance to reduce potential teacher bias (e.g., self-fulfilling prophecy), and ensured that each grader was assigned a similar number of papers across all three arms in each grade-session pair to reduce potential grader bias. Graders evaluated the scores based on a teacher-designed rubric; see details in Appendix A.5. At the end of each session, we surveyed the students on their experience and preferences. Third, in the GPT Base and GPT Tutor arms, we collected all student messages and corresponding GPT-4 responses from interactions with our tutoring program.

We pre-registered[8] this RCT with a designated primary analysis of comparing students' unassisted exam outcomes across arms—this translates to our main findings that generative AI without guardrails can harm learning. See Appendix A.6 for a discussion of our pre-registration.

3 Main Results

Our primary regression specifications evaluate student performance in the assisted practice problems and unassisted exam, respectively:

$$\text{Outcome}^{(j)}_{ics} = \beta_1 \text{GPT Base}_c + \beta_2 \text{GPT Tutor}_c + \beta_3 \text{Prev GPA}_i + \theta_s + \delta_g + \alpha_y + \lambda_t + \varepsilon_{ics} \quad (1)$$

Here, $\text{Outcome}^{(j)}_{ics} \in [0, 1]$ is the normalized grade of student i in classroom c and session $s \in \{1, .., 4\}$ for the assisted $(j = 0)$ or unassisted $(j = 1)$ portion; GPT Base_c and GPT Tutor_c are binary variables indicating treatment assignments for class c; Prev GPA_i controls for student performance, and captures student i's normalized GPA from the previous year;[9] and θ_s, δ_g, α_y,

[8] https://aspredicted.org/4DL_Q3J

[9] Normalized GPA has a mean of 0.82 and a standard deviation of 0.11 among students in our main sample.

and λ_t are session, grader, grade level, and teacher fixed effects, respectively. Standard errors are clustered at the classroom level (which is the unit of randomization). Our main sample excludes students who didn't complete the survey, as well as students in honors-designated classrooms (which are not populated randomly, unlike regular classrooms).[10]

Table 1: Regression results on normalized student performance in the practice (assisted) and exam (unassisted) problems across grades and sessions; fixed effects are suppressed. Robust standard errors are clustered at the classroom level.

	Dependent variable:	
	Practice Perf	Exam Perf
	(1)	(2)
GPT Base	0.137^{**}	-0.054^{*}
	(0.031)	(0.022)
GPT Tutor	0.361^{**}	-0.004
	(0.032)	(0.013)
Prev GPA	0.802^{**}	1.334^{**}
	(0.076)	(0.069)
Control Arm Mean	0.284	0.321
Control Arm SD	0.287	0.277
Observations	2,848	2,848
R^2	0.389	0.386
Adjusted R^2	0.382	0.379
Note: HC1 robust standard errors clustered by class		*p<0.05; **p<0.01

Table 1 reports intention-to-treat estimates from this regression. We find that GPT Base and GPT Tutor substantially increased student scores in the GPT-assisted practice sessions by 0.137 and 0.361 (out of 1), respectively, relative to the control arm that only had access to textbooks (mean performance of 0.28). These results imply that GPT Base and GPT Tutor would increase performance on the assisted practice sessions by 48% and 127%, respectively, on average relative to the control arm. These results are consistent with the existing literature on the

[10]We performed robustness checks that included these students and found similar results; see Appendix B.

productivity effects of generative AI (*7, 8*). Furthermore, the gap between GPT Tutor and GPT Base illustrates the added benefits of problem-specific teacher inputs in the prompt. Specifically, GPT Base often provides incorrect answers due to hallucinations; in contrast, the prompt for GPT Tutor includes the solution along with recommended hints, which enables students to both obtain useful hints and check their answers once they have them. Our mechanism analysis in Section 4 provides a more detailed analysis supporting this hypothesis.

In stark contrast, in the subsequent unassisted exam, student performance in the GPT Base arm *degraded* by 0.054 (out of 1) relative to that of the control arm. In other words, GPT Base diminished the average control student's performance on the unassisted exam by 17%. Student performance in the GPT Tutor arm was statistically indistinguishable from that of the control arm, and the point estimate was smaller by an order of magnitude (-0.004). The fact that students in the GPT Tutor arm performed similarly to students in the control arm on the unassisted exam may be surprising since they performed so much better on the practice problems. This difference may be partly explained by the fact that students with access to GPT Tutor during the practice session could ask it to check their answers,[11] whereas students in the control arm were only shown the solutions after they had already submitted their answers.

We pre-registered a simplistic variation of our main specification in Eq. (1) using pairwise t-tests; the results are reported in Appendix A.6 and are qualitatively similar.

These results demonstrate an inherent tradeoff in access to generative AI tools: while these tools can substantially improve human performance when access is available, they can also degrade human learning (particularly when appropriate safeguards are absent), which may have a long term impact on human performance.

[11] As can be seen from Figure 7b in our paper, "Attempted Answers" and "Ask for help" are the dominant message types in the GPT Tutor arm; see discussion in Section 4.2.

Problem-level specification. We also examine an alternative regression specification that examines *problem-level* outcomes rather than student-level outcomes:

$$\text{Outcome}^{(j)}_{icps} = \beta_1 \text{GPT Base}_c + \beta_2 \text{GPT Tutor}_c + \beta_3 \text{Prev GPA}_i + \theta_s + \delta_g + \alpha_y + \lambda_t + \varepsilon_{ics} \tag{2}$$

Here, $\text{Outcome}^{(j)}_{icps} \in [0, 1]$ is the normalized grade of student i in classroom c for problem p and session $s \in \{1, .., 4\}$ for the assisted $(j = 0)$ or unassisted $(j = 1)$ portions. Table 8 in Appendix B.2 reports the results, which are very similar to the previous student-level specification. The problem-level specification is useful in the next section, where we examine how the error rate in solutions generated by GPT Base in the assisted problems affects student performance and learning in the subsequent unassisted exam.

Student Perception. Interestingly, students' own self-reported perceptions of the effects of GPT tutors on their exam performance and learning are overly optimistic. While students in the GPT Base arm performed worse on the exam (relative to the control arm), they did not perceive that they performed worse or learned less. Similarly, while students in the GPT Tutor arm did not perform better on the exam (relative to the control arm), they perceived that they performed significantly better. This mismatch between perceived and actual learning has been observed in other settings (*20, 21*). Additional details are in Appendix B.3.

Heterogeneity. We look for heterogeneous treatment effects as a function of pre-registered variables, capturing students' ability, resources, and effort. In general, we find limited to no statistically significant support for heterogeneous treatment effects with either treatment, particularly with respect to unassisted exam performance. Additional details are in Appendix B.4.

Grade Dispersion. We examine a measure of dispersion in student performance—the Herfindahl–Hirschman index (HHI). Both GPT Base and GPT Tutor reduced grade dispersion in the assisted practice sessions, matching prior findings that generative AI assistance reduces the "skill gap" by providing the largest benefits for the weakest students (*7–9*). However, we find no significant effect on HHI for the unassisted exam—i.e., the reduction in the skill gap does not persist when access to generative AI is removed. Additional details are in Appendix B.5.

Robustness checks. Five class sessions did not use the assigned treatment due to external circumstances (e.g., laptops did not arrive on time). Our primary specifications use an intention-to-treat analysis—i.e., to preserve randomization, we consider all students in a treatment arm as treated, regardless of whether they actually received that treatment. In Appendix B.1, we provide details on non-compliance and perform a regression omitting non-compliers, finding the same insights. Next, we examine several variations to our main specification in Eq. (1) to assess the robustness of our results. These results provide qualitatively similar insights as our main analysis; see Appendix B.2. Lastly, we check whether differential student absenteeism may impact our results, and find no differential attrition in student attendance across arms or sessions; see Appendix B.6.

4 Potential Mechanism: Asking for Solutions

In general, students may be adversely affected by GPT Base's assistance in two ways: (1) errors made by GPT Base mislead students in the subsequent unassisted problems, or (2) using GPT Base as a "crutch" prevents them from fully engaging with or understanding the material prior to attempting the unassisted problems. Recall that the design of GPT Tutor avoids both of these issues: (1) it rarely makes mistakes since its prompt includes the solution, and (2) it is hard for students to use it as a crutch since its prompt asks it to avoid giving them the answer and instead guide them in a step-by-step fashion (see Figure 6 in Appendix A.1). We perform two analyses that help determine which explanation is more likely. First, we analyze how the error rate of

GPT Base on a practice problem affects students' subsequent performance on a highly similar exam problem, and second, we analyze student engagement with the tool. Our findings suggest that the second explanation (i.e., students using GPT Base as a crutch) is the main mechanism by which GPT Base impedes student learning.

4.1 GPT Errors vs. Student Performance

GPT Base often makes mistakes on math problems (*22*). We first quantify error rates by repeatedly querying GPT Base using the most common message used by students in the GPT Base arm—i.e., "What is the answer?" For each of the 57 total practice problems, we ask GPT Base for the answer ten times (resetting the system between queries), and then manually categorize any errors in the response as *arithmetic* (steps followed were correct, but the resulting computation was incorrect) or *logical* (steps followed were partially or fully incorrect). We find that GPT Base gives a correct answer only 51% of the time on average; it makes logical errors 42% of the time and arithmetic errors 8% of the time. Figure 2 shows the histogram of how often GPT Base returns an incorrect answer for different problems, demonstrating significant problem-specific heterogeneity in error rates. Details are in Appendix C.1.

We now assess how errors made by GPT Base affect student performance on both the practice problems and the unassisted exam. To this end, we add interaction terms between the error rate of GPT Base and the treatment arm to the problem-level regression specification in Eq. (4). When assessing exam performance, we leverage our paired design of the session material—i.e., for each exam problem, the teachers included a conceptually similar practice problem to help students learn how to solve that exam problem. Thus, for a given exam problem, we use GPT Base's error rate on the corresponding practice problem. The regression specification is given in Eq. (3) in Appendix C.1, and the results are shown in Table 2. Both types of GPT Base errors negatively impact practice problem performance for students in the GPT Base arm (i.e.,

Table 2: Regression results on student performance in the practice and corresponding exam problems across grades and sessions; this regression is at the problem level, and includes interaction terms for the logical and arithmetic error rates of GPT Base on practice problems (see Eq. (3) in Appendix C.1). We use a correspondence between the exam and practice problems to estimate how errors on practice problems affect performance on exam problems. Fixed effects are suppressed.

	Dependent variable:	
	Practice Perf	Exam Perf
GPT Base	0.362^{**}	-0.035
	(0.032)	(0.027)
GPT Tutor	0.337^{**}	0.029
	(0.037)	(0.023)
Logical Error Rate	-0.075^{*}	0.178^{**}
	(0.030)	(0.028)
Arithmetic Error Rate	-0.172^{*}	-0.063
	(0.082)	(0.032)
Prev GPA	0.789^{**}	1.330^{**}
	(0.074)	(0.068)
GPT Base × Logical Error Rate	-0.448^{**}	-0.029
	(0.036)	(0.040)
GPT Tutor × Logical Error Rate	0.022	-0.086
	(0.038)	(0.044)
GPT Base × Arithmetic Error Rate	-0.492^{**}	-0.099
	(0.117)	(0.056)
GPT Tutor × Arithmetic Error Rate	0.329^{**}	0.095^{*}
	(0.107)	(0.044)
Observations	13,484	11,392
R^2	0.214	0.212
Adjusted R^2	0.212	0.209

Note: HC1 robust standard errors clustered by class $^{*}p<0.05$; $^{**}p<0.01$

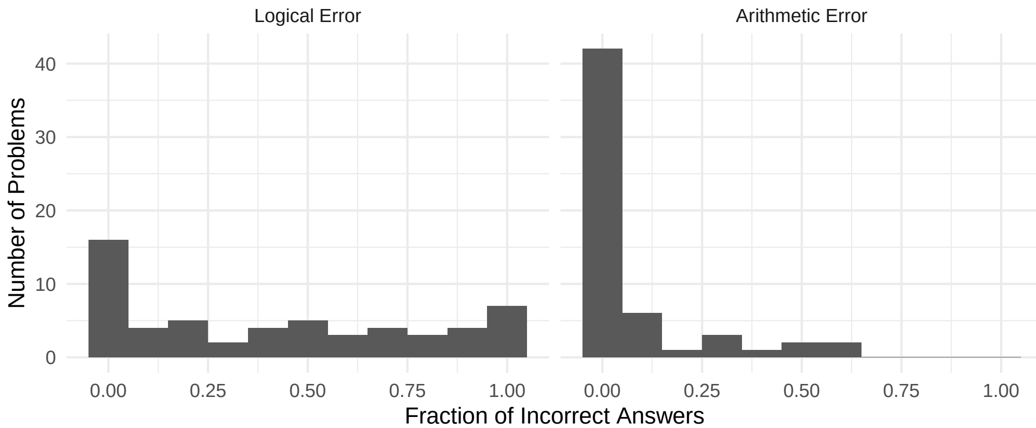


Figure 2: Fraction of times GPT Base returns an incorrect answer due to a logical (left) or arithmetic (right) error for 57 practice problems across grades and sessions. (Note that GPT Tutor does not return an answer by design, and is further given the correct answer in the prompt.)

the coefficients of "GPT Base $\times$ Logical Error Rate" and "GPT Base $\times$ Arithmetic Error Rate" are statistically significantly negative in the practice performance regression).[12]

Two key observations support our hypothesis that students are using GPT Base as a crutch. First, if students are being misled by logical errors made by GPT Base, we would expect these errors to affect performance on the corresponding exam problems in the unassisted exam. However, while GPT Base's logical errors affect performance on the practice problems, we find no evidence that this effect spills over to the corresponding exam problems (i.e., "GPT Base $\times$ Logical Error Rate" does not have a statistically significant effect on exam performance).[13] Second, if students were reading and understanding the solutions provided by GPT Base in the practice session, we might expect arithmetic errors to have a smaller impact on practice problem performance than logical errors. This is because students know arithmetic relatively well, and should be better able to catch these errors. However, arithmetic and logical errors appear to

[12]Note that we separately control for both GPT Base error rates. As expected, the corresponding coefficients in the practice performance regression are both statistically significantly negative, since higher GPT Base error rates are correlated with higher problem difficulty.

[13]An analogous regression on the *total* error rate (combining both logical and arithmetic errors), yields similar insights (see Table 16 in Appendix C.1).

have similar effects on practice performance (i.e., "GPT Base $\times$ Logical Error Rate" and "GPT Base $\times$ Arithmetic Error Rate" have similar coefficients in the practice performance regression). Both these results suggest that students are simply copying answers from GPT Base.

4.2 Student Engagement

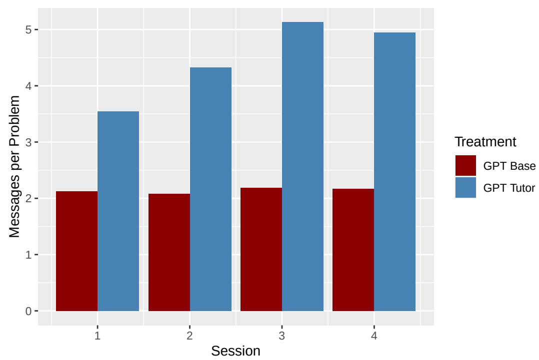
Next, we analyze the messages that students sent to GPT Base or GPT Tutor to better understand how they are interacting with these tools. Figure 3(a) shows the average number of messages each student had with their respective GPT tool (Base or Tutor) in a given session. As can be seen, the number of messages in GPT Tutor is significantly higher, and further increases with experience using the tool.[14] The fact that students interact substantially less with GPT Base is consistent with our hypothesis that GPT Base simply provides students with solutions.

For a more fine-grained understanding of the content of the student messages, we use natural language processing and clustering to group student messages (see Appendix C.2 for details). We manually associate each cluster with a text description summarizing the content of the messages in that cluster. As shown in Figure 7 in Appendix C.2, students in GPT Base most often simply ask for the answer; in contrast, students in GPT Tutor learn to interact more substantively with the tutoring tool over time by asking for help and independently attempting to solve the problem. We observe this learning effect even within the first session—if we restrict to the very first interaction students have with our tool in the first session, 56% in the GPT Base arm and 42% in the GPT Tutor arm[15] either repeat the question text or ask for the answer; in contrast, when considering the first interaction across all problems in the first session, this number increases to 67% for GPT Base but decreases to 37% for GPT Tutor.

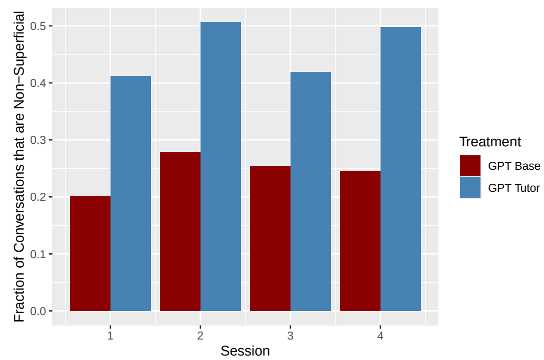
We consider clusters where the student simply asks for the answer (specifically, "Repeat

[14] We similarly find that students spend 13% more time with GPT Tutor than with GPT Base; see Appendix C.3.

[15] The initial discrepancy may be due to the fact that GPT Tutor, unlike GPT Base, offers suggested prompts; see Figure 6 in Appendix A.1 for details.



(a) Avg. # of Messages per Problem



(b) Frac Non-Superficial Conversations per Session

Figure 3: Student engagement—given by (a) average number of student messages per problem, and (b) average fraction of student session conversations that have no superficial messages (simply re-stating the question or asking for the answer) per session—by treatment (GPT Base and GPT Tutor) over time.

Question Text" and "Ask for Answers") to be *superficial*, and the remaining clusters (specifically, "Attempted Answers" and "Ask for Help") to be *non-superficial*. For a given student in a given session, we consider the corresponding conversation superficial if the student asked *any* superficial messages, and non-superficial otherwise. Intuitively, non-superficial conversations are ones where the student constructively interacts with the tutoring tool and never asks for the answer. Figure 3(b) shows the aggregate rate of non-superficial conversations for both GPT Base and GPT Tutor. As can be seen, across all sessions, in the GPT Base arm, only a small fraction of conversations are non-superficial; in contrast, a substantially larger fraction of conversations in the GPT Tutor arm were non-superficial. These results suggest that the vast majority of students are using GPT Base to obtain solutions, whereas a significant fraction of students are using GPT Tutor in a purely substantive way.

5 Discussion

Our results provide a cautionary tale regarding both the existing use of the freely available ChatGPT tool by students as well as the potential deployment of GPT-based tutors in educational settings. While generative AI tools such as ChatGPT can make tasks significantly easier for humans, they come with the risk of deteriorating our ability to effectively learn some of the skills required to solve these tasks. These shortcomings have been anecdotally reported for tools such as Khanmigo (*23*), a GPT-4 based tutoring application. Our findings support both the need for educators to find ways to safeguard student learning in the face of freely available tools such as ChatGPT as well as the need to design more effective guardrails for generative AI tutors.

In some ways, ChatGPT is not the first technology to exhibit this tradeoff—for instance, typing diminishes the need for handwriting, and calculators diminish the need for arithmetic, etc. However, we believe ChatGPT differs from prior technologies in two significant ways. First, the capabilities of ChatGPT are substantially broader and more intellectual compared to prior examples; for instance, our experiments focus on a broad variety of mathematical topics, which encompass fundamental skills required by a wide range of knowledge-intensive professions. Second, unlike many prior technologies, ChatGPT is highly unreliable and often provides incorrect responses. Our results suggest that students are either unable to detect these failures or unwilling to spend the effort needed to check correctness.

Although the guardrails implemented in GPT Tutor appear to largely mitigate these negative effects, substantial work is required to enable generative AI to positively enhance rather than diminish education. GPT Tutor remains passive, responding to students when they ask questions but failing to proactively engage students with the material. Effective human tutors ask probing questions to uncover student misconceptions, and then clarify these misconceptions by providing tailored explanations. Combining existing software tutors (*15, 24*) with generative AI may be a promising path to achieving this goal, since it balances the pedagogical principles baked into existing algorithms with the capability of generative AI to understand and respond to complex student queries. One way to do so is to leverage agent approaches (*25, 26*), which compose models with different prompts to achieve complex goals; for instance, we might use one prompt to ask the model to identify student misconceptions, and another prompt to ask it to generate a useful hint. Beyond designing better tutors, promising avenues include educating students and teachers about how to more effectively use generative AI, and deploying teacher-facing (rather than student-facing) tools (*13*). For instance, recent evidence suggests that "co-pilots" that work with a human tutor instead of replacing them may improve outcomes (*27*).

Lastly, we discuss limitations. While our study takes a first step towards understanding the potential harms of generative AI on learning, it focuses on a specific deployment context, and substantial work is needed to better understand the nature and scope of these effects. In particular, our study focuses on two generative AI tutors for a single topic (mathematics) deployed in a single high school in Turkey. We have objective evaluation criteria for math problems, which is unavailable in creative subjects such as writing. Our deployment was also carried out in Fall 2023, when generative AI was still very new—users may now be more familiar with how to effectively use these tools and their shortcomings, and furthermore, both closed and open-weight models have significantly improved in performance since then. Additional studies are required to assess generalizability to other tutor designs and deployment contexts. We focus on short-term outcomes due to limitations imposed by our partner school; studying long-term outcomes is a key direction for future research. Finally, controlled lab experiments can complement our study to shed more light on students' underlying learning mechanisms.

References

1. J. Achiam, *et al.*, *OpenAI* (2023).

2. K. Singhal, *et al.*, *Nature* **620**, 172 (2023).

3. Y. Li, *et al.*, *Science* **378**, 1092 (2022).

4. W. Geerling, G. D. Mateer, J. Wooten, N. Damodaran, *The American Economist* **68**, 233 (2023).

5. C. Terwiesch, *Mack Institute White Paper* (2023).

6. T. Eloundou, S. Manning, P. Mishkin, D. Rock, *Science* **384**, 1306 (2024).

7. S. Noy, W. Zhang, *Science* **381**, 187 (2023).

8. E. Brynjolfsson, D. Li, L. R. Raymond, *National Bureau of Economic Research Working Paper* (2023).

9. F. Dell'Acqua, *et al.*, *Harvard Business School Technology & Operations Management Unit Working Paper* (2023).

10. G. S. Becker, *Human capital: A theoretical and empirical analysis, with special reference to education* (University of Chicago press, 2009).

11. Safety alert for operators: Manual flight operations, *Tech. rep.*, Federal Aviation Administration (2013).

12. J. Maynez, S. Narayan, B. Bohnet, R. McDonald, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 1906–1919.

13. E. R. Mollick, L. Mollick, *Available at SSRN 4300783* (2022).

14. A. Extance, *Nature* **623**, 474 (2023).

15. S. Ruan, *et al.*, *Proceedings of the interaction design and children conference* (2020), pp. 567–580.

16. S. Ruan, *et al.*, *Machine Learning* **113**, 3023 (2024).

17. J. Cristia, P. Ibarrarán, S. Cueto, A. Santiago, E. Severín, *American Economic Journal: Applied Economics* **9**, 295 (2017).

18. A. J. Lamb, J. M. Weiner, *International Journal of Education in Mathematics, Science and Technology* **6**, 136 (2018).

19. M. Escueta, A. J. Nickow, P. Oreopoulos, V. Quan, *Journal of Economic Literature* **58**, 897 (2020).

20. E. L. Bjork, R. A. Bjork, *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, M. A. Gernsbacher, R. W. Pew, L. M. Hough, J. R. Pomerantz, eds. (Worth Publishers, 2011), pp. 56–64.

21. L. Deslauriers, L. S. McCarty, K. Miller, K. Callaghan, G. Kestin, *Proceedings of the National Academy of Sciences* **116**, 19251 (2019).

22. S. Frieder, *et al.*, *Advances in Neural Information Processing Systems* **36** (2024).

23. N. Singer, *The New York Times* (2023).

24. S. Doroudi, V. Aleven, E. Brunskill, *International Journal of Artificial Intelligence in Education* **29**, 568 (2019).

25. S. Yao, *et al.*, *International Conference on Learning Representations (ICLR)* (2023).

26. X. Wang, *et al.*, *Forty-first International Conference on Machine Learning* (2024).

27. R. E. Wang, A. T. Ribeiro, C. D. Robinson, S. Loeb, D. Demszky, *Annenberg Institute for School Reform at Brown University* (2024).

28. D. Bertsimas, M. Johnson, N. Kallus, *Operations Research* **63**, 868 (2015).

29. M. Grootendorst, *arXiv preprint arXiv:2203.05794* (2022).

30. N. Reimers, I. Gurevych, *arXiv preprint arXiv:1908.10084* (2019).

31. L. McInnes, J. Healy, J. Melville, *arXiv preprint arXiv:1802.03426* (2018).

32. L. McInnes, J. Healy, S. Astels, *et al.*, *J. Open Source Softw.* **2**, 205 (2017).

Data and Code Availability

To support further research, anonymized data will be made available at [GITHUB LINK TBD] upon publication. Data will include student performance on practice and exam questions for all sessions, as well as student-, class-, and grader-level covariates; it will further include anonymized and time-stamped student conversations with our GPT tutors. All code used in this paper was written in a combination of R, Python, and Stata. These scripts will also be included in the same repository.

Acknowledgements

We acknowledge invaluable research assistance from Tuba Tas and Dogus Berk Kocak, as well as Tamer Atacan for his partnership. We are grateful for funding from the Wharton AI & Analytics Initiative, the Fishman-Davidson Center, and Wharton Global Initiatives. We are also grateful for helpful feedback from Eric Bradlow, Angela Duckworth, Benjamin Luttges, Lilach Mollick, Ananya Sen, Christian Terwiesch, Lyle Ungar, as well as seminar participants at the

London School of Economics, Indian School of Business, Workshop on Unstructured Data and Language Models, and Workshop on AI & Analytics for Social Good.

A Experimental Protocol Details

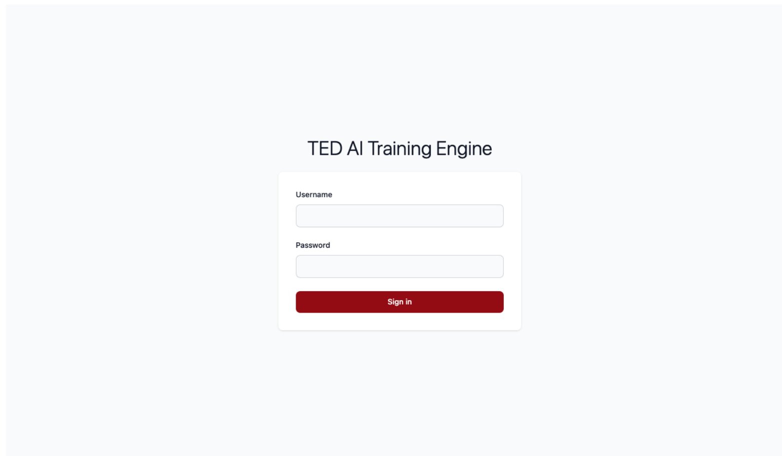
Details on the GPT tutoring program are in Appendix A.1, teacher and student briefing in Appendix A.2, session material development and IT setup in Appendix A.3, treatment assignment and covariate balance in Appendix A.4, and grading student work in Appendix A.5.

A.1 GPT Tutoring Program

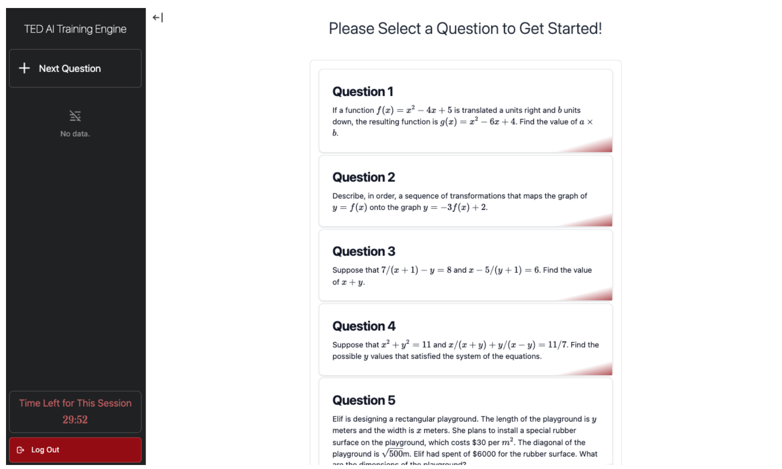
We developed two math tutoring tools based on OpenAI's GPT-4-0613 model. These tools are implemented as web applications hosted on Amazon Web Services (AWS), and use the OpenAI API to access GPT-4. At the beginning of the semester, each student is assigned unique login credentials (username and password) based on their student ID. During a class, if a class is assigned to either the GPT Base or GPT Tutor arms, then the teacher distributed laptops (provided by us) to the students so that they could access the website that hosted our tool; students could not access other websites or applications. When students view this website, they are first presented with a login screen (Figure 4 (a)), upon which they entered their login credentials to access the tool. The version of the tool (Base vs. Tutor) is automatically determined by the web application based on the provided username.

Upon login, the students see a list of questions corresponding to the practice problems in the second part—i.e., the assisted practice period (Figure 4 (b)). When the student would like hints on how to solve one of the practice problems, they click on the corresponding module. This action brings up a new screen where they have a chat interface; in particular, they can submit messages, after which the web application will provide the response given by OpenAI's GPT-4 model (Figure 4 (c)). This interface is essentially the same as OpenAI's ChatGPT web application. The key difference is the prompts that are used, which depend on the arm:

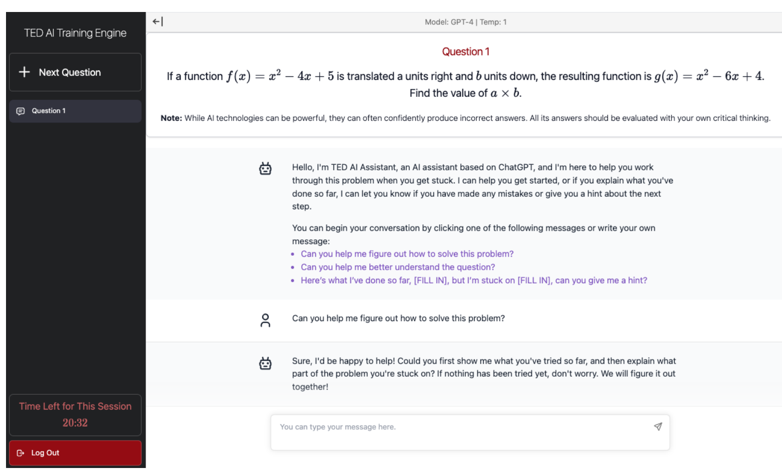
- **GPT Base:** This arm uses a prompt that instructs GPT-4 to help the user solve the prob-



(a) **Login portal:** Students login to the tutoring tool with their individualized credentials.

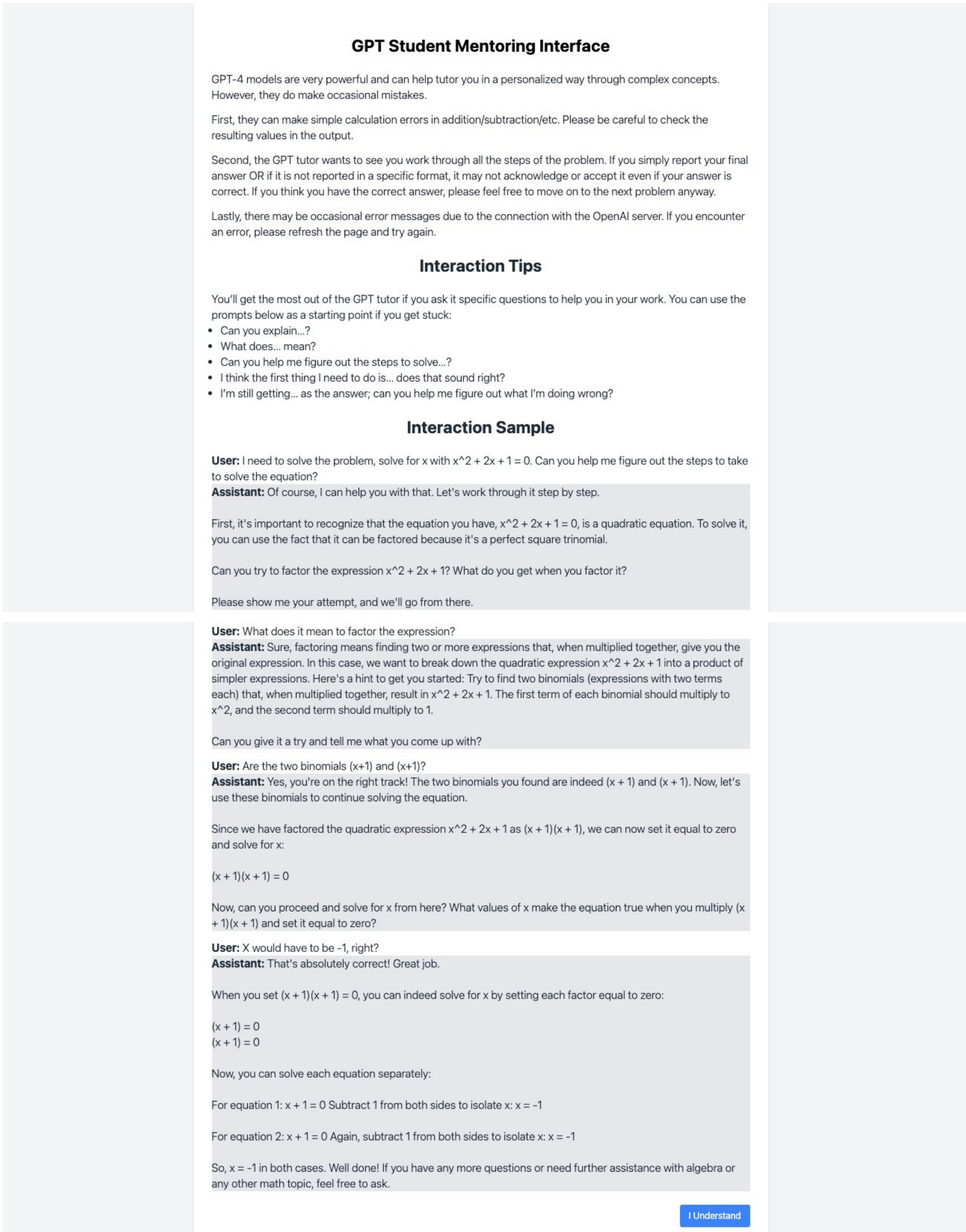


(b) **Question selection:** Students select the question (for their session and grade) corresponding to the practice problem they want to solve.



(c) **Chat interface:** After selecting a question, students interact with GPT-4 by typing messages and receiving responses. The prompt differs between GPT Base and Tutor.

Figure 4: Screen shot of different steps in using our GPT Base and GPT Tutor tools. The tools are identical except for the prompt used in the backend and the welcome message in the chat interface for GPT Tutor.



GPT Student Mentoring Interface

GPT-4 models are very powerful and can help tutor you in a personalized way through complex concepts. However, they do make occasional mistakes.

First, they can make simple calculation errors in addition/subtraction/etc. Please be careful to check the resulting values in the output.

Second, the GPT tutor wants to see you work through all the steps of the problem. If you simply report your final answer OR if it is not reported in a specific format, it may not acknowledge or accept it even if your answer is correct. If you think you have the correct answer, please feel free to move on to the next problem anyway.

Lastly, there may be occasional error messages due to the connection with the OpenAI server. If you encounter an error, please refresh the page and try again.

Interaction Tips

You'll get the most out of the GPT tutor if you ask it specific questions to help you in your work. You can use the prompts below as a starting point if you get stuck:

- Can you explain...?
- What does... mean?
- Can you help me figure out the steps to solve...?
- I think the first thing I need to do is... does that sound right?
- I'm still getting... as the answer; can you help me figure out what I'm doing wrong?

Interaction Sample

User: I need to solve the problem, solve for x with x^2 + 2x + 1 = 0. Can you help me figure out the steps to take to solve the equation?

Assistant: Of course, I can help you with that. Let's work through it step by step.

First, it's important to recognize that the equation you have, x^2 + 2x + 1 = 0, is a quadratic equation. To solve it, you can use the fact that it can be factored because it's a perfect square trinomial.

Can you try to factor the expression x^2 + 2x + 1? What do you get when you factor it?

Please show me your attempt, and we'll go from there.

User: What does it mean to factor the expression?

Assistant: Sure, factoring means finding two or more expressions that, when multiplied together, give you the original expression. In this case, we want to break down the quadratic expression x^2 + 2x + 1 into a product of simpler expressions. Here's a hint to get you started: Try to find two binomials (expressions with two terms each) that, when multiplied together, result in x^2 + 2x + 1. The first term of each binomial should multiply to x^2, and the second term should multiply to 1.

Can you give it a try and tell me what you come up with?

User: Are the two binomials (x+1) and (x+1)?

Assistant: Yes, you're on the right track! The two binomials you found are indeed (x + 1) and (x + 1). Now, let's use these binomials to continue solving the equation.

Since we have factored the quadratic expression x^2 + 2x + 1 as (x + 1)(x + 1), we can now set it equal to zero and solve for x:

(x + 1)(x + 1) = 0

Now, can you proceed and solve for x from here? What values of x make the equation true when you multiply (x + 1)(x + 1) and set it equal to zero?

User: X would have to be -1, right?

Assistant: That's absolutely correct! Great job.

When you set (x + 1)(x + 1) = 0, you can indeed solve for x by setting each factor equal to zero:

(x + 1) = 0
(x + 1) = 0

Now, you can solve each equation separately:

For equation 1: x + 1 = 0 Subtract 1 from both sides to isolate x: x = -1

For equation 2: x + 1 = 0 Again, subtract 1 from both sides to isolate x: x = -1

So, x = -1 in both cases. Well done! If you have any more questions or need further assistance with algebra or any other math topic, feel free to ask.

I Understand

Figure 5: The landing page for the GPT Base and GPT Tutor arms. Upon logging into their accounts, each user will see this page and must click "I understand" to access the practice problems.

lem; it includes the problem the user is currently working on.

- **GPT Tutor:** The prompt used by GPT Tutor includes the instructions and information provided in the GPT Base prompt. There are two significant additional aspects. First, it includes detailed instructions asking GPT-4 to avoid giving away the full solution, and instead provide incremental hints to help the student solve the problem. Second, it includes one or more solutions, as well as a list of common mistakes and corresponding hints to provide to the student should they make one of these mistakes.

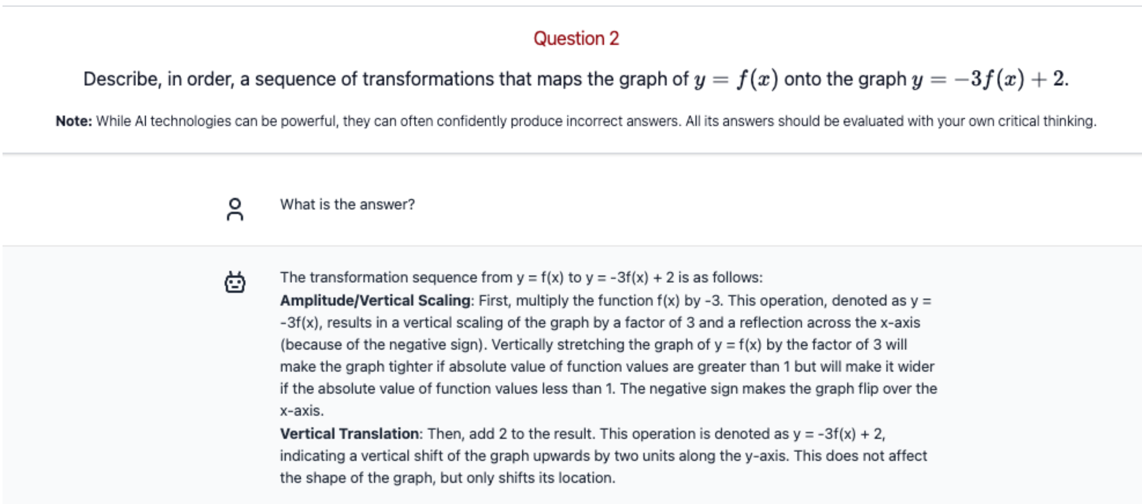
Besides the different prompts, one additional feature of GPT Tutor is that it provides a welcome message to the student, providing them with a description of its capabilities as well as three suggested prompts that they can use (see Figure 6 for an illustration).

The problems, solutions, and common mistakes are developed by the teachers (see Appendix A.3). We designed the prompt templates, and incorporated the material provided by the teachers into the prompts. These prompts were then uploaded into the web application prior to each session. Representative prompts for both GPT Base and Tutor for a question on linear systems (from the first 11th grade practice session) are shown in Figure 1 in the main paper.

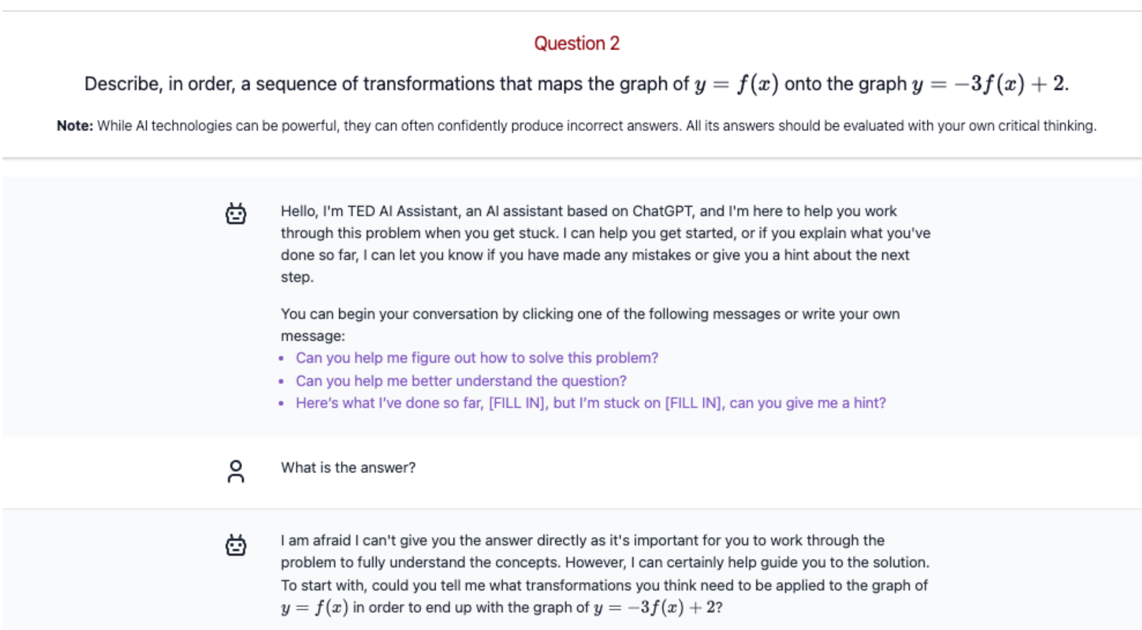
Finally, Figure 6 illustrates the impact of the prompt on students' interactions with GPT-4 in the chat interface. In particular, Figure 6a shows a hypothetical interaction with GPT Base, and Figure 6b shows the same hypothetical interaction with GPT Tutor. As can be seen, GPT Base gives away the solution, whereas GPT Tutor asks the student to provide their progress.

A.2 Teacher and Student Briefing

Before the start of the study, our team held a meeting with all participating teachers to introduce the experiment protocol and to demo both versions of our tool. Teachers were not provided any training on the best usage of generative AI or our tool. (Note that they did not interact with students during the assisted practice periods, or instruct students on generative AI usage.)



(a) GPT Base



(b) GPT Tutor

Figure 6: Response to a message asking for the answer in (a) GPT Base, and (b) the GPT Tutor. For GPT Base, the GPT-4 will typically return the solution to the problem. In contrast, for GPT Tutor, GPT-4 will typically avoid returning the solution and will instead ask the student to explain what they are stuck on.

Prior to the first session, teachers informed students about our experiment, including its general structure (i.e., there would be four 90 minute sessions with 3 stages, etc.) and its goals, but they were not informed of the specific topics that would be covered. Students were also told that their performance on both the practice and exam questions in these sessions would be factored into their final grades, and that the exam questions would cover the same topics as the practice questions.

Students in the GPT arms were shown short videos (< 5 minutes) introducing our tool and sample non-superficial interactions.[16] The landing page of our tool (identical for GPT Base and Tutor) also includes some tips on best interacting with the tool as well as sample interactions (see Figure 5). No further training was provided to students on specific prompting styles or ways in which they could best use the mentoring interface to facilitate their learning.

A.3 Material Development and IT Setup

One of the co-authors, Özge Kabakcı, a high school math teacher and former department chair of the math department at our partner Turkish high school, led the development of all session materials. Additionally, we employed a math teacher, Tuğba Taş, who was formerly at our partner high school as a part-time research assistant to aid in material development. Notably, these materials comprise a substantial portion (about 15%) of the curriculum in each of three grades. Our content creation was guided by two primary sources. First, we adhered strictly to the syllabus prescribed by the Ministry of Education in Turkey.[17] Second, we incorporated supplementary materials from two extra books for each grade level, developed by the partner high school. These books, not publicly accessible, have been reviewed by the educational board of the Ministry of Education in Turkey, ensuring that our materials align with official pedagogical

[16] GPT Base classes were shown `https://www.youtube.com/watch?v=klas2MZbG6g` and GPT Tutor classes were shown `https://www.youtube.com/watch?v=Mwtb--FyHhs`.

[17] Available at https://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=343

guidelines. The questions developed for each session are documented in Appendix D. All material in this study was consistent with the school's existing curriculum and teaching methods—no adjustments were made to the sequence of topics or instructional practices.

We purchased 52 laptops, as the largest class size was 26, ensuring sufficient IT resources to cover two classes simultaneously. We hired two IT experts as research assistants to provide full-time IT support during all sessions. The partner high school has Wi-Fi coverage in classrooms, with each classroom equipped with a computer and often utilizing the internet as a teaching resource. While we primarily relied on the school's existing Wi-Fi infrastructure, we acquired four portable Wi-Fi dongles, two for each session, to guarantee reliable internet connectivity at all times. In case of connectivity issues within the school's network, we used these dongles to supplement Wi-Fi access. All computers were charged overnight, and we used batteries during the sessions. Additionally, we carried extension cables to address any battery drainage. The laptops were stored in a secure location within the school (the IT support team's office). Teachers submitted requests before reaching the topics covered in our sessions. Our IT RA team transported the laptops and set up the IT equipment during the break before the start of each lecture, ensuring that the IT setup did not encroach upon session time, with one exception noted in Appendix B.1. Once a student opens the laptop lid, they are directly presented with our GPT interface. Students were explicitly forbidden from accessing other websites or opening additional applications. To enforce this protocol, both the teacher and the IT research assistant actively monitored the classroom, ensuring strict adherence to the guidelines.

A.4 Treatment Assignment and Covariate Balance

We assign treatment arms (Control, GPT Base, and GPT Tutor) at the class level—i.e., all students in a single class are assigned to the same arm across all four sessions. One challenge is that there are overlapping classes but a limited number of laptops. Therefore, we randomize classes to treatment arms while accounting for scheduling constraints using an integer program, based on the strategy proposed in (*28*). We also include additional constraints to help balance the number of students in each arm. Note that, except for honors students (who are excluded from our main analysis), the school randomly assigns students to classes, so the treatment assignment at the student level is random. In our main regression analysis, we cluster the standard error at the class level to account for the fact that treatment is assigned at the class level.

Table 3 presents sample descriptive statistics by trial arm. We separately report balance across covariates for our main sample (which excludes students in special honors-designated classes), and the full survey sample.

A.5 Grading

Official solutions and grading rubrics were created as part of material development for both practice and exam questions. To mitigate potential grading biases stemming from teachers' personal connections and expectations of their students, we engaged independent graders to carry out a blinded grading process. We hired one lead grader to oversee the entire process, responsible for collecting the papers from the school, distributing them to other graders, training them on the grading rubric, and returning the graded papers to the teachers. Each grader evaluated a similar number of papers from all three arms for each grade-session pair. Furthermore, we note that the submitted papers did not include any information about the assigned arm, so graders could not be influenced by the treatment arm during grading. Initially, we hired two additional graders for the first session. For the subsequent three sessions, we employed a different team of five graders. All graders were either master's or Ph.D. students from the engineering or mathematics departments of two top Turkish universities. They entered all the grades into an Excel spreadsheet, each associated with a unique student ID. Graders were assigned to different trial arms to incorporate grader fixed effects, with grader assignments made at the class level.

Table 3: Column (1) presents the overall sample descriptive statistics. Columns (2) to (4) report the outcome level of each variable by treatment arm. Column (5) reports the p-value from a test of the hypothesis of equal means across the experimental conditions. Column (6) reports the p-value with FDR correction for multiple hypothesis testing. Note that this table includes the sample of students who completed the survey.

Description	Overall (1)	GPT-Tutor (2)	Control (3)	GPT-Base (4)	P-Value (5)	P-val (FDR) (6)
Main Sample						
Total Count	839	277	320	242		
Both Parents with at least College Degree N False (%)	196 (23.36)	63 (22.74)	68 (21.25)	65 (26.86)	0.285	0.619
Both Parents with at least College Degree N True (%)	643 (76.64)	214 (77.26)	252 (78.75)	177 (73.14)		
# of Household Members, Mean (SD)	3.59 (1.00)	3.53 (0.99)	3.69 (0.96)	3.52 (1.07)	0.074	0.424
# of Children Aged in the Household, Mean (SD)	1.47 (1.47)	1.39 (1.39)	1.56 (1.51)	1.45 (1.50)	0.369	0.655
Class Enjoyment Score [0-4], Mean (SD)	2.22 (1.08)	2.19 (1.05)	2.23 (1.11)	2.24 (1.08)	0.857	0.929
Participation Score [0-4], Mean (SD)	2.37 (1.01)	2.33 (1.02)	2.34 (1.00)	2.46 (1.01)	0.286	0.619
Math HW Completion Score [0-4], Mean (SD)	3.04 (0.94)	3.02 (0.89)	3.11 (0.92)	2.95 (1.00)	0.133	0.432
Get Help with Homeworks, N False (%)	238 (28.37)	71 (25.63)	94 (29.38)	73 (30.17)	0.457	0.660
Get Help with Homeworks, N True (%)	601 (71.63)	206 (74.37)	226 (70.62)	169 (69.83)		
Private Tutorship, N False (%)	322 (38.38)	102 (36.82)	137 (42.81)	83 (34.30)	0.098	0.424
Private Tutorship, N True (%)	517 (61.62)	175 (63.18)	183 (57.19)	159 (65.70)		
Visits to Training Center, N False (%)	628 (74.85)	193 (69.68)	243 (75.94)	192 (79.34)	0.035	0.424
Visits to Training Center, N True (%)	211 (25.15)	84 (30.32)	77 (24.06)	50 (20.66)		
Female, N False (%)	447 (53.28)	148 (53.43)	172 (53.75)	127 (52.48)	0.954	0.954
Female, N True (%)	392 (46.72)	129 (46.57)	148 (46.25)	115 (47.52)		
Average Weekday Study Hours, Mean (SD)	1.95 (2.42)	2.06 (2.59)	1.92 (2.43)	1.85 (2.20)	0.579	0.684
Average Weekend Study Hours, Mean (SD)	2.98 (2.57)	2.85 (2.46)	3.09 (2.62)	2.99 (2.63)	0.527	0.684
Previous GPA [0,1], Mean (SD)	0.82 (0.11)	0.81 (0.12)	0.82 (0.11)	0.82 (0.11)	0.403	0.655
Including Honors						
Total Count	943	312	349	282		
Both Parents with at least College Degree N False (%)	206 (21.85)	63 (20.19)	72 (20.63)	71 (25.18)	0.268	0.576
Both Parents with at least College Degree N True (%)	737 (78.15)	249 (79.81)	277 (79.37)	211 (74.82)		
# of Household Members, Mean (SD)	3.60 (0.99)	3.54 (0.99)	3.70 (0.95)	3.55 (1.02)	0.076	0.490
# of Children Aged in the Household, Mean (SD)	1.42 (1.43)	1.34 (1.34)	1.51 (1.51)	1.39 (1.42)	0.299	0.576
Class Enjoyment Score [0-4], Mean (SD)	2.27 (1.07)	2.24 (1.07)	2.27 (1.10)	2.29 (1.05)	0.848	0.918
Participation Score [0-4], Mean (SD)	2.39 (1.02)	2.37 (1.03)	2.36 (1.01)	2.46 (1.02)	0.375	0.609
Math HW Completion Score [0-4], Mean (SD)	3.08 (0.93)	3.05 (0.89)	3.15 (0.92)	3.02 (0.97)	0.192	0.576
Get Help with Homeworks, N False (%)	287 (30.43)	86 (27.56)	107 (30.66)	94 (33.33)	0.310	0.576
Get Help with Homeworks, N True (%)	656 (69.57)	226 (72.44)	242 (69.34)	188 (66.67)		
Private Tutorship, N False (%)	386 (40.93)	121 (38.78)	158 (45.27)	107 (37.94)	0.113	0.490
Private Tutorship, N True (%)	557 (59.07)	191 (61.22)	191 (54.73)	175 (62.06)		
Visits to Training Center, N False (%)	722 (76.56)	223 (71.47)	270 (77.36)	229 (81.21)	0.018	0.236
Visits to Training Center, N True (%)	221 (23.44)	89 (28.53)	79 (22.64)	53 (18.79)		
Female, N False (%)	511 (54.19)	171 (54.81)	189 (54.15)	151 (53.55)	0.954	0.954
Female, N True (%)	432 (45.81)	141 (45.19)	160 (45.85)	131 (46.45)		
Average Weekday Study Hours, Mean (SD)	1.91 (2.32)	2.01 (2.48)	1.86 (2.36)	1.87 (2.10)	0.646	0.763
Average Weekend Study Hours, Mean (SD)	2.84 (2.41)	3.04 (2.56)	3.06 (2.61)	0.492	0.711	
Previous GPA [0,1], Mean (SD)	0.83 (0.11)	0.83 (0.12)	0.83 (0.11)	0.84 (0.11)	0.548	0.712

A.6 Pre-Registration Details

Our pre-registration is available at `https://aspredicted.org/4DL_Q3J`. Our designated primary analysis was to compare students' unassisted exam performance across arms. In the main paper, we present a principled regression specification (Table 1), which appropriately controls for relevant covariates and clusters standard errors at the classroom level. We pre-registered an alternative strategy using pairwise t-tests to compare each pair of arms—the results on our main sample are reported in Table 4, and those additionally including non-survey responders are reported in Table 5. Note that the results are qualitatively consistent with Table 1 for both the practice problems and the exam.

Table 4: Pre-registered pairwise t-test results on our main sample for each pair of arms on normalized student performance in the practice (assisted) and exam (unassisted) problems across grades and sessions.

	Practice Perf			Exam Perf		
	Diff	95% CI	p-value	Diff	95% CI	p-value
GPT Base vs. Control	0.19	$[0.165, 0.214]$	$< 10^{-15}$	-0.035	$[-0.057, -0.012]$	0.003
GPT Tutor vs. Control	0.385	$[0.359, 0.411]$	$< 10^{-15}$	-0.006	$[-0.030, 0.018]$	0.613
GPT Tutor vs. GPT Base	0.195	$[0.169, 0.221]$	$< 10^{-15}$	0.028	$[0.005, 0.052]$	0.019

Table 5: Pre-registered pairwise t-test results on our main sample and non-survey responders for each pair of arms on normalized student performance in the practice (assisted) and exam (unassisted) problems across grades and sessions.

	Practice Perf			Exam Perf		
	Diff	95% CI	p-value	Diff	95% CI	p-value
GPT Base vs. Control	0.181	$[0.158, 0.204]$	$< 10^{-15}$	-0.027	$[-0.006, -0.049]$	0.013
GPT Tutor vs. Control	0.388	$[0.364, 0.413]$	$< 10^{-15}$	-0.005	$[-0.028, 0.018]$	0.676
GPT Tutor vs. GPT Base	0.208	$[0.183, 0.232]$	$< 10^{-15}$	0.022	$[0.000, 0.045]$	0.048

We pre-registered a number of secondary analyses, including testing for heterogeneous treatment effects (Section 3, detailed in Appendix B.4), comparing students' perceived learning outcomes (Section 3, detailed in Appendix B.3), and qualitatively examining student conversations with the mentoring interface to better understand the mechanisms behind outcomes (Section 4.2, detailed in Appendix C.2). We additionally performed a number of exploratory analyses that we did not pre-register to better understand our findings—including comparing performance on assisted practice problems (Section 3), an alternative problem-level regression specification (Section 3, detailed in Appendix B.2), examining grade dispersion (Section 3, detailed in Appendix B.5), and an error-level analysis (Section 4.1, detailed in Appendix C.1).

B Supporting Results for Section 3

Details on non-compliance with the treatment are in Appendix B.1, robustness of our results to alternative regression specifications in Appendix B.2, students' self-reported perceptions in Appendix B.3, heterogeneous treatment effects in Appendix B.4, student grade dispersion in Appendix B.5, and student absenteeism in Appendix B.6.

B.1 Non-Compliance with Treatment

In five instances, a treatment classroom could not execute the treatment due to unanticipated external circumstances; these instances are summarized in Table 6. We check the validity of our results accounting for non-compliance by performing an alternative regression specification where we exclude non-compliers. Table 7 shows results for this regression; they are nearly identical to our main analysis in Table 1.

B.2 Alternative Main Regression Specifications

First, we consider the problem-level (rather than student-level) regression specification in Eq. (4). Here, $\text{Outcome}_{icps}^{(j)} \in [0,1]$ is the normalized grade of student i in classroom c for problem p and session $s \in \{1,..,4\}$ for the assisted ($j=0$) or unassisted ($j=1$) portions; GPT Base_c,

Table 6: Class sessions that were non-compliant with a GPT treatment (i.e., switched to the control condition) due to exogenous circumstances.

Class	Session	Treatment	Main Sample	Reason for Non-Compliance
9J	1	GPT Base	✓	Computers arrived late.
10A	2	GPT Base	✗	Teacher did not realize the class was a GPT session.
11S	2	GPT Base	✓	Technical error in math mentoring UI.
10B	3	GPT Tutor	✗	Technical error in math mentoring UI.
11S	4	GPT Base	✓	Teacher did not realize the class was a GPT session.

Table 7: Regression results on normalized student performance in the practice (assisted) and exam (unassisted) problems across grades and sessions, excluding non-complying class sessions (as reported by school teachers and staff); fixed effects are suppressed. Robust standard errors are clustered at the classroom level.

	Dependent variable:	
	Practice Perf	Exam Perf
	(1)	(2)
GPT Base	0.137^{**}	-0.054^{*}
	(0.031)	(0.022)
GPT Tutor	0.362^{**}	-0.004
	(0.032)	(0.013)
Prev GPA	0.808^{**}	1.341^{**}
	(0.077)	(0.069)
Observations	2,805	2,805
R^2	0.387	0.385
Adjusted R^2	0.380	0.377
Note: HC1 robust standard errors clustered by class		$^{*}p<0.05$; $^{**}p<0.01$

GPT Tutor$_c$, Prev GPA$_i$, θ_s, δ_g, α_y, and λ_t are all as before. Standard errors are again clustered at the classroom level. The results, shown in Table 8, are similar to results from the student-level specification in Eq. (1).

Table 8: Regression results on normalized problem-level performance in the practice (assisted) and exam (unassisted) problems across grades and sessions; fixed effects are suppressed. Robust standard errors are clustered at the classroom level.

	Dependent variable:	
	Practice Perf	Exam Perf
GPT Base	0.138^{**}	-0.054^{*}
	(0.031)	(0.022)
GPT Tutor	0.366^{**}	-0.004
	(0.032)	(0.013)
Prev GPA	0.795^{**}	1.334^{**}
	(0.075)	(0.069)
Observations	13,484	11,392
R^2	0.172	0.195
Adjusted R^2	0.170	0.192
Note: HC1 robust standard errors clustered by class		*p<0.05; **p<0.01

Next, we consider a number of alternatives to our main regression specification given in Eq. (1) to ensure that our findings are consistent, specifically:

1. include the non-random honors classes (9A, 9B, 9C, 10A, 10B, 10C) in our sample,
2. include students who did not complete the baseline survey in our sample,
3. include the self-reported survey variables we collected (parents' education, household composition, class enjoyment/participation, hours spent studying, homework completion, homework help, access to private tutoring or training centers, prior exposure to ChatGPT, and gender) as controls, and
4. cluster our heteroskedasticity-robust standard errors at the student-level or the (classroom, student)-level.

Results in Table 9 provide the same insights as our main results in Table 1.

Table 9: Results from variations of regression specification Eq. (1). For brevity, we only report the coefficients of our treatment variables (GPT Base and GPT Tutor) for both the assisted practice problems and the unassisted exam.

	Practice Perf		Exam Perf	
	GPT Base	GPT Tutor	GPT Base	GPT Tutor
Include honors classes	0.101^{**}	0.359^{**}	-0.064^{*}	0.001
	(0.032)	(0.032)	(0.026)	(0.018)
Include non-survey responders	0.095^{**}	0.364^{**}	-0.069^{**}	0.001
	(0.033)	(0.031)	(0.027)	(0.017)
Include survey variables	0.146^{**}	0.364^{**}	-0.041^{*}	-0.001
	(0.029)	(0.030)	(0.021)	(0.012)
Clustered SEs: Student	0.137^{**}	0.361^{**}	-0.054^{**}	-0.004
	(0.023)	(0.018)	(0.018)	(0.013)
Clustered SEs: Student & Class	0.137^{**}	0.361^{**}	-0.054^{*}	-0.004
	(0.031)	(0.032)	(0.022)	(0.013)

Note: HC1 robust errors clustered by class (unless specified otherwise) *p<0.05; **p<0.01

B.3 Student Perception

At the end of each exam, we conducted a short survey to measure students' perceptions of their own performance. We asked five questions:

1. "How much do you think you learned from this whole class session?" with options "A great deal," "Quite a lot," "Moderately," "A little," and "Nothing at all."
2. "How well do you think you performed in this quiz?" with options "Excellent," "Above Average," "Average," "Below Average," and "Very Poorly."
3. "How much time did it take you to solve the questions in this quiz?" with options "0-5 min," "5-10 min," "10-15 min," "15-20 min," "20-25 min," and "25-30 min."
4. "How useful was the problem-solving session in the previous part (Part 2) in helping you solve the questions in this quiz?" with options "Effective," "Somewhat effective," "Neutral," "Somewhat ineffective," and "Ineffective."

5. "How many minutes would you be willing to give up on this quiz to have the help of the TED-AI Training Engine (or ChatGPT-4 if you haven't used the TED-AI Training Engine)?" with options "0-5 min," "5-10 min," "10-15 min," "15-20 min," "20-25 min," and "25-30 min."

For each question, we fit a model to predict the student's response. For Questions (1), (2), and (4), we use an ordered probit model using all five categories. For Questions (3) and (5), we convert the interval responses into a continuous value by taking the mid-point of each bin, resulting in a value between 2.5 and 27.5 minutes; then, we use a OLS model. Our model has the same form as our main regression specification in Eq. (1), except the outcome is replaced with the student response and we additionally use a link function ϕ (corresponding to the ordered probit model for Questions (1), (2), and (4) and linear for Questions (3) and (5)):

$$\text{Response}_{ics}^{(j)} = \phi(\beta_1 \text{GPTBase}_c + \beta_2 \text{GPTTutor}_c + \beta_3 \text{Prev GPA}_c + \theta_s + \delta_g + \alpha_y + \lambda_t + \varepsilon_{ics}).$$

We omit all unanswered questions from estimation. Results are shown in Table 10. Although students in the GPT Base arm performed significantly worse in the final exam, they did not perceive that they learned less or performed worse. Interestingly, despite no real detectable performance improvement, we found a significant increase in perceived exam performance in the GPT Tutor arm relative to the control. Additionally, students in the GPT Tutor arm perceived the practice sessions as more valuable for learning than those in other arms, and they took about a minute (approximately 3%) longer to finish the exam. Overall, these results suggest that despite the negative effect of GPT Base and the null effect of GPT Tutor on real exam performance, students attribute some value to having access to GPT and potentially overvalue its benefits. Lastly, students in both GPT arms were willing to sacrifice more exam time to have access to GPT, with students in the GPT Base arm giving up about 2.5 minutes (8.3%) and students in the GPT Tutor arm giving up about 3.7 minutes (12.3%).

Table 10: Regression results on student perception; fixed effects are suppressed. Robust standard errors reported for all questions, with clustering at the classroom level in Questions (3) and (5). Questions (1), (2) and (4) are from an ordered probit model.

	(1) Perceived Learning	(2) Perceived Exam Perf	(3) Exam Duration	(4) Perceived Value of Practice Session	(5) Time Trade-off
GPT Base	-0.0783	-0.0218	-0.114	0.113	2.245**
	(0.0800)	(0.0774)	(0.799)	(0.0804)	(0.692)
GPT Tutor	0.0711	0.130*	0.953*	0.271**	3.710**
	(0.0551)	(0.0563)	(0.399)	(0.0562)	(0.561)
Prev GPA	0.931**	2.517**	5.144*	0.965**	-3.967
	(0.222)	(0.236)	(2.222)	(0.220)	(2.635)
Observations	2,603	2,594	2,523	2,549	2,321

*p<0.05; **p<0.01

B.4 Heterogeneous Treatment Effects

We look for heterogeneous treatment effects as a function of pre-registered secondary analysis variables: students' previous GPA (student ability), access to private tutoring (student's resources), and hours spent studying (student effort);[18] results are shown in Table 11, Table 12, and Table 13, respectively. In general, we find limited to no statistically significant support for heterogeneous treatment effects with either treatment. The only exceptions are, for the assisted practice sessions, we find that weaker students (those with lower GPAs) and students with private tutors benefit more from assistance from GPT Base—however, these heterogeneous effects do not persist for the unassisted exams.

B.5 Student Grade Dispersion

We study how access to GPT Base and GPT Tutor impact dispersion in student performance, for both the assisted practice problems and the unassisted exam. In particular, we compute the classroom-session level Hirschman-Herfindahl-Index (HHI) of normalized student performance

[18] One of our pre-registered heterogeneity analyses was prior exposure to generative AI. However, since students filled out our survey throughout the semester (rather than prior to the first session), and because our treatments involve exposure to generative AI, we omit this variable.

Table 11: Heterogeneity: Previous GPA.

	Dependent variable:	
	Practice Perf	Exam Perf
	(1)	(2)
GPT Base	0.178^{**}	-0.042
	(0.029)	(0.024)
Above Median GPA	0.095^{**}	0.031
	(0.026)	(0.020)
GPT Base × Above Median GPA	-0.077^{**}	-0.026
	(0.028)	(0.024)
GPT Tutor	0.38^{**}	-0.014
	(0.038)	(0.015)
GPT Tutor × Above Median GPA	-0.037	0.019
	(0.031)	(0.025)
Prev GPA	0.571^{**}	1.220^{**}
	(0.123)	(0.107)
Observations	$2,848$	$2,848$
R-squared	0.394	0.388
Note: HC1 robust standard errors clustered by class		*p<0.05; **p<0.01

Table 12: Heterogeneity: Private Tutorship.

	Dependent variable:	
	Practice Perf	Exam Perf
	(1)	(2)
GPT Base	0.0905^{*}	-0.0645^{*}
	(0.0339)	(0.0270)
Private Tutor Access	-0.0237	-0.0129
	(0.0234)	(0.0116)
GPT Base × Private Tutor	0.0735^{*}	0.0170
	(0.0342)	(0.0224)
GPT Tutor	0.336^{**}	-0.0126
	(0.0418)	(0.0196)
GPT Tutor × Private Tutor	0.0398	0.0138
	(0.0343)	(0.0165)
Prev GPA	0.800^{**}	1.327^{**}
	(0.0725)	(0.0724)
Observations	$2,848$	$2,848$
R-squared	0.391	0.386
Note: HC1 robust standard errors clustered by class		*p<0.05; **p<0.01

Table 13: Heterogeneity: Hours Spent Self-studying.

	Dependent variable:	
	Practice Perf	Exam Perf
	(1)	(2)
GPT Base	0.118^{**}	-0.0682^{**}
	(0.0338)	(0.0244)
Above Median Self-study Hours	-0.0164	-0.0334^{*}
	(0.0219)	(0.0136)
GPT Base × Above Median Self-study	0.0359	0.0284
	(0.0312)	(0.0200)
GPT Tutor	0.347^{**}	-0.0249
	(0.0324)	(0.0158)
GPT Tutor × Above Median Self-study	0.0301	0.0431
	(0.0290)	(0.0232)
Prev GPA	0.798^{**}	1.337^{**}
	(0.0777)	(0.0681)
Observations	2,848	2,848
R-squared	0.390	0.388

Note: HC1 robust standard errors clustered by class *p<0.05; **p<0.01

for both the practice and exam problems. Then, we fit a model with the same form as our main regression specification in Eq. (1), except that it is conducted at the classroom-session level:

$$\mathrm{HHI}_{cs}^{(j)} = \beta_1 \mathrm{GPTBase}_c + \beta_2 \mathrm{GPTTutor}_c + \beta_3 \mathrm{Prev\ GPA}_c + \theta_s + \delta_g + \alpha_y + \lambda_t + \varepsilon_{cs}.$$

Here, $\mathrm{HHI}_{cs}^{(j)}$ is the HHI of classroom c and session $s \in \{1, .., 4\}$ for the assisted ($j = 0$) or unassisted ($j = 1$) portion. Table 14 shows results for this regression. We find that access to either GPT Base or GPT Tutor reduces HHI (i.e., reduces grade dispersion) during the assisted practice portion. However, we observe no significant treatment effect on performance in the unassisted exam. In other words, while access to generative AI can reduce the "skill gap" by helping weaker students more (*7–9*), this effect does not persist when access to generative AI is removed.

Table 14: Regression results on grade dispersion; fixed effects are suppressed.

	Dependent variable:	
	Practice Perf	Exam Perf
	(1)	(2)
GPT Base	-0.0429^{*}	-0.0162
	(0.0163)	(0.0142)
GPT Tutor	-0.0773^{**}	0.00113
	(0.0198)	(0.0147)
Prev GPA	-0.373^{*}	-0.953^{**}
	(0.169)	(0.249)
Observations	172	172
R^2	0.414	0.530
Note: HC1 robust standard errors clustered by class		*p<0.05; **p<0.01

B.6 Student Absenteeism

Students are absent from class at an average rate of 12.3%. A potential concern is that students systematically miss class in the treatment or control arms for various treatment-related reasons (e.g., they dislike GPT, they anticipate poor performance, etc). To this end, in Table 15, we report student absenteeism rates in each session and each arm. We do not find differential attrition—i.e., students in all three arms have a similar likelihood of attending classes. We also do not find any effect of session on student absenteeism. Unsurprisingly, we find that students with a higher previous GPA are mildly more likely to attend class sessions.

C Supporting Results for Section 4

Details on our analysis with GPT Base errors are in C.1, and clustering messages in C.2.

Table 15: Regression results on student absenteeism; fixed effects are suppressed.

	Attendance (1)
GPT Base	-0.0245
	(0.0271)
GPT Tutor	0.00312
	(0.0212)
Prev GPA	0.165
	(0.0875)
2.Session	-0.0243
	(0.0322)
3.Session	0.0121
	(0.0391)
4.Session	-0.0481
	(0.0403)
Observations	3,247
R-squared	0.033
Note: HC1 robust standard errors clustered by class	*p<0.05; **p<0.01

C.1 GPT Base Error Rate Analysis

We estimate the error rate of GPT Base on practice problems by testing a standardized ChatGPT message (in conjunction with the GPT Base prompt). We use the most common message used by students in the GPT Base arm—namely, "what is the answer?" (see Appendix C.2 for statistics on most common messages). Then, for each practice problem, we sample 10 random responses from GPT-4 using this message. We do this because, for the same exact question and prompt, there is a lot of variability in the answer (and its correctness) provided by GPT-4. We manually classify each response as "correct" (i.e., correctly solves the problem), "arithmetic error" (i.e., used the correct solution strategy but computed a numerical value incorrectly), or "logical error" (i.e., used the wrong solution strategy). Then, for a given problem, the arithmetic error *rate* is the fraction of answers labeled as an arithmetic error, and similarly, the logical error *rate* is the fraction of answers labeled as a logical error.

When analyzing the impact of these errors on student performance, we use the following problem-level regression specification:

$$\begin{aligned}\text{Outcome}^{(j)}_{icps} = {} & \beta_1 \text{GPT Base}_c + \beta_2 \text{GPT Tutor}_c + \beta_3 \text{Prev GPA}_i \\ & + \beta_4 \text{Logical Error Rate}_{\sigma^{(j)}(p)s} + \beta_5 \text{Arithmetic Error Rate}_{ps} \\ & + \beta_6 \text{Logical Error Rate}_{\sigma^{(j)}(p)s} \times \text{GPT Base}_c \\ & + \beta_7 \text{Logical Error Rate}_{\sigma^{(j)}(p)s} \times \text{GPT Tutor}_c \\ & + \beta_8 \text{Arithmetic Error Rate}_{\sigma^{(j)}(p)s} \times \text{GPT Base}_c \\ & + \beta_9 \text{Arithmetic Error Rate}_{\sigma^{(j)}(p)s} \times \text{GPT Tutor}_c \\ & + \theta_s + \delta_g + \alpha_y + \lambda_t + \varepsilon_{ics}\end{aligned} \tag{3}$$

This specification is identical to Eq. (4), except for two changes. First, when $j = 1$ (i.e., the outcome is performance on the unassisted exam), we use a mapping from each exam problem p to practice problem $\sigma^{(1)}(p)$ so we can associate exam problems with GPT Base error rates on the practice problems. This way, we can evaluate the impact of the GPT Base error rate on a practice problem the performance of the student on the corresponding exam problem. If $j = 0$, then we use $\sigma^{(0)}(p) = p$. This mapping is derived from how the exams were designed; in particular, for each exam problem, teachers designed a practice problem to help students learn the concepts necessary to solve that exam problem. Second, it includes additional controls for Logical Error Rate$_{p's}$ (i.e., logical error rate of GPT Base for problem p' in session s) and Arithmetic Error Rate$_{p's}$ (i.e., arithmetic error rate of GPT Base for problem p' in session s), where $p' = \sigma^{(j)}(p)$; it also includes interaction terms between these error rates and the treatment variables GPT Base$_c$ and GPT Tutor$_c$.

Table 16 shows results for an alternate version of the regression specified in Eq. (3), where we look at the *total* error rates of GPT Base, instead of distinguishing between logical and arithmetic errors. Consistent with the previous results, while GPT Base's errors negatively affect performance on the practice problems, we find no evidence that this effect spills over to the corresponding exam problems (i.e., "GPT Base × Total Error Rate" does not have a statistically significant effect on exam performance), suggesting that students are simply copying answers from GPT Base.

Table 16: Regression results on student performance in the practice and corresponding exam problems across grades and sessions; this regression is at the problem level, and includes interaction terms for the total (logical or arithmetic) error rates of GPT Base on practice problems. We use a correspondence between the exam and practice problems to estimate how errors on practice problems affect performance on exam problems. Fixed effects are suppressed.

	Dependent variable:	
	Practice Perf	Exam Perf
GPT Base	0.362^{**}	-0.036
	(0.032)	(0.027)
GPT Tutor	0.341^{**}	0.032
	(0.038)	(0.024)
Total Error Rate	-0.081^{**}	0.157^{**}
	(0.028)	(0.027)
Prev GPA	0.796^{**}	1.335^{**}
	(0.074)	(0.069)
GPT Base × Total Error Rate	-0.452^{**}	-0.035
	(0.037)	(0.039)
GPT Tutor × Total Error Rate	0.045	-0.076
	(0.035)	(0.042)
Observations	13,484	11,392
R^2	0.212	0.207
Adjusted R^2	0.210	0.204
Note: HC1 robust standard errors clustered by class		*p<0.05; **p<0.01

C.2 Clustering Student Messages

We use a BERT-based topic model framework to cluster student messages (*29*); we summarize this procedure here. To ensure the number of messages is balanced across different students, we focus on the first message provided by each student for each problem. First, we convert each message into a dense vector representation using Sentence-BERT (*30*). Next, we reduce the dimension of these vectors using UMAP (*31*) to increase the stability of clustering. Then, we run the clustering algorithm HDBSCAN (*32*) on the resulting vectors. The number of clusters are chosen automatically by the BERTTopic model, with the constraint that there are at least 2 messages per cluster. We perform this procedure separately for each practice problem (uniquely identified by its session, treatment arm, grade, and problem ID). In the end, we obtain 394 unique clusters across all 57 problems. Afterwards, we compute the cluster TF-IDF (which runs TF-IDF treating all messages in a cluster as a single document), and extract the top three representative messages for each cluster.

Once we have obtained representative messages, we aggregate them across practice problems and perform a second round of clustering; we refer to this second clustering step as the *meta-clustering step* and call the resulting clusters *meta-clusters*. Once again, the number of clusters are chosen automatically by the BERTTopic model, with the constraint that there are at least 5 message clusters per meta cluster. In total, we have 29 unique meta clusters. Then, we compute the top representative message for each meta-cluster using the same cluster TF-IDF index described above, which we call a *meta-representative message*. In addition, we manually examine the messages in each meta-cluster to label it with a message type. These results are shown in Figure 7a for GPT Base and Figure 7b for GPT Tutor. As can be seen, students in GPT Base most often simply ask for the answer; in contrast, students in GPT Tutor learn to interact more substantively with the tutoring tool over time—they most often ask for help in Sessions 2-3, and by Session 4, they most often independently attempt to answer the problem.

In addition, we categorize each meta-representative message as superficial (i.e., the message simply asks GPT Base for the answer in some way) or not superficial; we associate each individual message with the category label of its corresponding meta-representative message. To visualize the rate of messages in each cluster, for each student and each session, we select

Sessions	1st Common	2nd Common	3rd Common
Session 1	**Repeat Question Text** (38%) e.g., copy the exact question text	**Ask for Answers** (29%) e.g. “what is the answer”, “give me the answer”	**Ask for Help** (8%) e.g. “help me solve this”, “can u solve this question”
Session 2	**Ask for Answers** (35%) e.g. “what is the answer”, “give me the answer”	**Repeat Question Text** (33%) e.g., copy the exact question text	**Ask for Help** (10%) e.g. “help me solve this”, “can u solve this question”
Session 3	**Ask for Answers** (25%) e.g. “what is the answer”, “give me the answer”	**Repeat Question Text** (24%) e.g., copy the exact question text	**Ask for Help** (15%) e.g. “help me solve this”, “can u solve this question”
Session 4	**Ask for Answers** (36%) e.g. “what is the answer”, “give me the answer”	**Repeat Question Text** (26%) e.g., copy the exact question text	**Ask for Help** (9%) e.g. “help me solve this”, “can u solve this question”

(a) Top 3 Message Types for GPT Base

Sessions	1st Common	2nd Common	3rd Common
Session 1	**Repeat Question Text** (31%) e.g., copy the exact question text	**Attempted Answers** (27%) e.g., “the answer is 456”, “is f(0) = 6”	**Ask for Answers** (6%) e.g. “what is the answer”, “give me the answer”
Session 2	**Ask for Help (template)** (37%) i.e., “can you help me figure out how to solve this problem?”	**Repeat Question Text** (19%) e.g., copy the exact question text	**Attempted Answers** (8%) e.g., “the answer is 456”, “is f(0) = 6”
Session 3	**Ask for Help (template)** (25%) i.e., “can you help me figure out how to solve this problem?”	**Repeat Question Text** (19%) e.g., copy the exact question text	**Attempted Answers** (15%) e.g., “the answer is 456”, “is f(0) = 6”
Session 4	**Attempted Answers** (24%) e.g., “the answer is 456”, “is f(0) = 6”	**Ask for Help (template)** (22%) i.e., “can you help me figure out how to solve this problem?”	**Repeat Question Text** (13%) e.g., copy the exact question text

(b) Top 3 Message Types for GPT Tutor

Figure 7: Top 3 Message Types

the first message in the corresponding conversation, and compute the fraction of first messages in each of the four clusters; these fractions are illustrated in Figure 8. We observe that students in the GPT Tutor arm consistently send fewer superficial messages (top two panels) and send more non-superficial messages (bottom two panels) across sessions. For Figure 3b, we repeat this procedure for *all* messages to categorize superficial *conversations*.

Finally, for the error rate analysis described in Appendix C.1, we use the top meta-representative message in GPT Base, “what is the answer?”; this accounts for 31% of students’ first messages in the GPT Base arm.

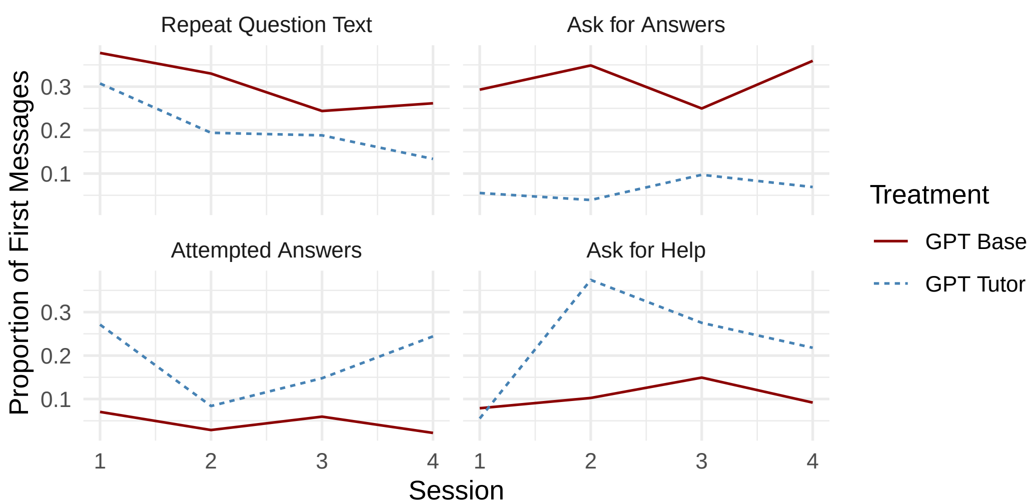


Figure 8: Proportion of first messages for each problem that fall into each of the four clusters (“Repeat Question Text”, “Ask for Answer”, “Attempted Answer”, and “Ask for Help”); the top two clusters are superficial and the bottom two are non-superficial.

C.3 Student Engagement

Apart from the number of messages sent per session and their content, we also examine the amount of time students in GPT Base vs. Tutor spent on the tutoring platform. We use the time of the last message minus the time of the first message in a session to proxy for the total time a student spent on the platform.[19]

$$\text{Session Time}_{ics} = \beta_1 \text{GPT Base}_c + \beta_2 \text{Prev GPA}_i + \theta_s + \alpha_y + \lambda_t + \varepsilon_{ics} \tag{4}$$

Here, Session Time$_{ics} \in [0, 40]$ is the number of minutes between the last and first message of student i in classroom c and session $s \in \{1, .., 4\}$. We control for student GPA, and include session, grade level, and teacher fixed effects. The results are show in Table 17. We see that students spend 2 additional minutes in each session when using GPT Tutor—i.e., they spend 13% more time on the platform, suggesting increased engagement.

Table 17: Regression results on minutes students spent on the GPT Base vs. Tutor platform; fixed effects are suppressed. Robust standard errors are clustered at the classroom level.

	Session Time
GPT Tutor	2.046**
	(0.322)
Prev GPA	10.598**
	(2.362)
GPT Base Mean	15.5
GPT Base SD	6.4
Observations	1,617
R^2	0.218
Adjusted R^2	0.207
Note: HC1 robust standard errors clustered by class	*p<0.05; **p<0.01

D Session Material

This appendix provides the practice and exam problems across all sessions and grades.

[19] This length should be no longer than 40 minutes. We occasionally observe longer times if the student logs in to the platform after the class is over; we drop these observations.

Grade	9
Session	1
Unit	Logic
Session topic	Logical connectives
No of students (av.)	25
Textbooks	The Mathematics 9, The Math Book 9

Part 2

1) If the compound proposition $(q' \vee r)' \Rightarrow p$ is false, then find the truth value of $(q \veebar p) \Rightarrow (r \vee p)$

2) Simplify the compound proposition $\big((p \Rightarrow q) \wedge q\big) \Rightarrow \big((p \wedge q') \vee q'\big)$

3) A basketball team coach is thinking about a line-up for a match. He is taking the following conditions into account:

i. If A plays, then player B won't play.
ii. If B plays, then A or C will play.
iii. Players B and Cwill not play together.
Who of the players A, B, and C will play?

4) For a two digits natural number AB, the propositions given below.
p: AB is even number.
q: AB is prime number.
r: A+B=11.

If the compound proposition $(p \Rightarrow q) \wedge (q' \wedge r)$ is true then find the value of $A \times B$. (University enterance exam question)

Part 3

1) $(p \wedge q) \Rightarrow (s \Rightarrow r')$ is false, then find the truth value of $(q \veebar r') \Rightarrow (p \vee s)$

2) Simplify $\big((p \Rightarrow q)' \wedge (q \vee p')\big)' \Rightarrow ((p \Rightarrow q)' \vee p)$

3) A school arranges a trip. If Serpil won't go on the trip, then Tuba or Elif will go on the trip. If Tuba will on the trip, then Elif will go. Either Serpil won't go on the trip or Tuba will go on the trip. They will not go on the trip all together. Who will attend the trip?

4) For AB two digits natural number, following propositons are given.

p: AB divisible by 3.
q: A+B is divisible by 5.
r: A ×B is divisible by 7.
If $q \Rightarrow (p \Rightarrow r)$ is false then find the value of A ×B. (University entrance exam question)

Grade	9
Session	2
Unit	Logic
Session topic	Review
No of students (av.)	25
Textbooks	The Mathematics 9, The Math Book 9

Part 2

1) If converse of $r \Rightarrow (p \wedge q)$ is false, then find the truth value of inverse of $(q \veebar p) \Rightarrow (r \vee p)$.

2) Simplify the following compound proposition

$$(p' \vee q) \wedge (p \Rightarrow q') \Leftrightarrow [p' \wedge (q \veebar q')]$$

3) In a security system, there are four switches: A, B, C, and D. The switches can be either in the ON position ("1") or the OFF position ("0"). To decrypt a special code, the following logical condition must be met:

$$\mathrm{Y} = (\mathrm{A} \vee \mathrm{B}) \wedge (\mathrm{C}' \wedge \mathrm{D}) \wedge \mathrm{A}$$

However, due to an advanced safeguard mechanism, two switches are interconnected in such a way that if one is ON, the other is forced to be OFF, and vice versa. These interconnected switches are B and C.

Given these circumstances, for the code to be decrypted (for Y to be “1”), write the possible combination of switches need to be turned on and off?

4) “If at least one prime number is even, then the square of every real number is positive.”

Find the truth value of the given proposition and write the negation of the given proposition.

5) At a table; There are three marbles in total, one red, one blue and one yellow. These marbles are placed in bags A, B and C, with one marble in each bag.

p : “There is no red marble in bag A.”
q: “There is a blue marble in bag B.”
r: “There is no yellow marble in bag C.”

If the proposition $p \wedge (q \vee r)'$ is true, what are the colors of the marbles in bags A, B and C, respectively? (University entrance exam)

Part 3: (Without ChatGPT) (30 mins)

1) If the converse of $r' \Rightarrow (p \wedge q)$ is false, then find the truth value of the contrapositive of $(q \veebar r) \Rightarrow p'$.

2) Simplify the following compound proposition
$(p \Rightarrow q) \Leftrightarrow [(p \Rightarrow q') \wedge p]'$.

3) $(\exists x \in R,\ -x^2 \geq 0) \Rightarrow [(\exists x \in N, 2x + 12 = 3) \vee (\forall x \in Z, -2 > x \ or \ x \geq 4)]$

a) Find the truth-value of the given compound proposition.

b) Write the negation of the compound proposition.

4) $p : a + b = 0$ $\qquad$ $q : a + c < 0$ $\qquad$ $r : c \leq 0$

The proposition $p, q,$ and r is given.

If the compound proposition $(p \wedge q) \Rightarrow r$ is false, then what are the signs of a, b and c respectively?

Grade	9
Session	3
Unit	Sets
Session topic	Revision
No of students (av.)	25
Textbooks	The Mathematics 9, The Math Book 9

Part 2:

1) $A = \{x \mid 24 < x \leq 150, x = 4k, k \in \mathbb{Z}\}$,
   $B = \{x \mid 12 \leq x < 144, x = 6k, k \in \mathbb{Z}\}$
   sets are given.
   According to this, what is the number of elements in the set $A \cup B$?

2) Given sets $A = \{a, b, c\}$ and $B = \{1, 2, 3, 4, a, b, c\}$, how many different subsets K can be written such that K is not equal to A, K is not equal to B, and A is a subset of K which is a subset of B?

3) On a certain website, the number of people who can use both a bicycle and a motorcycle are 10, and the number of people who can use at least one of them is 35. On this site, if the number of people who can use a motorcycle is twice the number of people who can use a bicycle, how many people can use **only a motorcycle**?

4) For sets A and B,
   $n(A \cap B`) = 2n(A \cap B)$
   $n(B - A) = 6$
   $n(A) = 2n(B) + 1$ are given.
   Find $n(A \cap B)$.

5) a, b, c are one-digit natural numbers, and x is a positive integer, such that $a < b < c$.
   The sets are given as
   $A = \{x | x^2 < 42\}$
   $B = \{1,2,3,7,8,9\}$
   $C = \{a, b, c\}$
   The number of elements in the Cartesian product $(A \cup C) \times (B \cup C)$ is 54. Find the sum of possible values for c.

Part 3:

1) Consider the sets
   $C = \{r | 43 < r \leq 210, r = 10k, k \in \mathbb{Z}\}$,
   $D = \{t \mid 15 \leq t < 270, t = 15k, k \in \mathbb{Z}\}$
   Find the number of elements in $C - D$.

2) Given sets $A = \{1,2,3,4\}$ and $B = \{1, 2, 3, 4,5,6, a, b, c\}$, how many different subsets K can be written such A is a subset of K and K is a subset of B but not equal to B, additionally K needs to include all even numbers in set B?

3) In a high school, students are either gamers or social media users. The number of students who are both passionate about gaming and active on social media is 20, and the total number of students who engage in at least one of these activities is 60. If we denote the set of gamers as G and social media users as S, and given that $n(S \cap G`) = a$ and $n(G) = 2n(S) + 5$, find the value of a.

4) For sets A, B and C, the following equations are given:
$n(A) = n(C) = 5$
$n\big(A \times (B \cup C)\big) = 30$
$n\big(B \times (A \cup C)\big) = 16$
Find the number of elements of $B \cap C$.

Grade	9
Session	4
Unit	Divisibility Rules
Session topic	Revision
No of students (av.)	25
Textbooks	The Mathematics 9, The Math Book 9

Part 2:

1) For A and B natural numbers, if A is divided by 6, quotient is B and remainder is 5. If B is divided by 5 remainder is 4. Find the remainder when A is divided by 15.

2) Emre, a store manager, is entering the total sales and the number of items sold into the computer when he notices an error in the invoice. The total sales amount for a product, which is priced at ₺11 per unit, is incorrectly recorded as ₺27,360. This error is due to two consecutive digits being written in the wrong order. How many units were sold?

3) A five-digit number is represented as a4c5c. If this number becomes divisible by 12 when 7 is added to it, what is the highest possible value for the sum of a + c?

4) A school is planning to distribute 72 pencils and 90 notebooks among students in a class, ensuring each student gets the same number of pencils and the same number of notebooks without any leftovers. What is the maximum number of students that can be in the class?

5) In a music school, two students are practicing their scales on different instruments. Alice practices her piano scales every 12 days, while Bob practices his violin scales every 18 days. They both had their first practice session on the same Tuesday. What will be the day of their 4th joint practice session?

Part 3:

1) In a bakery, a large batch of cookies is divided into packages. When the total number of cookies (C) is divided by 9, the quotient is the number of packages (P) and the remainder is 8. When the number of packages is divided by 8, the remainder is 7. Your task is to find the remainder when the total number of cookies is divided by 18.
2) Aden was checking if some numbers are divisible by 11 or not. She found out that A and B are divisible by 11, but if you divide C by 11, the remainder is 5.
$30d5d12 = 2A + 3B + C$. What is the value(s) of d?

3) The six-digit number $23aa4b$ is 11 less than a multiple of 45. What is(are) possible the value(s) of $a \times b$.

4) In a high school, two clubs hold their meetings on different schedules. The Robotics Club meets every 6 days, while the Astronomy Club meets every 8 days. Both clubs held their **second** meeting on the same Thursday. On what day of the week will their 10th joint meeting occur?

Grade	10
Session	1
Unit	Counting Principles-Permuation-Combination-Porabililty
Session topic	Combination
No of students (av.)	25
Textbooks	The Mathematics 10, The Math Book 10

Part 2:

1) Find the number of subsets of A with 3 elements containing b or f where A={a,b,c,d,e,f}

2) Find the number of 5-person group is chosen from 6 girls and 5 boys that contain at least 1 boy.

3) In how many ways can *12* people form *3* teams of *4* people?

4) In a mathematics lesson, the teacher asked Naz how 3 students out of the class could be chosen in different ways, Mert how 5 students could be chosen in different ways, and Zeynep how 11 students could be chosen in different ways. All three students calculated the requested numbers correctly. Given that the numbers found by Mert and Zeynep are the same positive integer, what is the number Naz found?

Part 3:

1) Find the number of subsets of A with 4 elements containing 1 and 3 but not 4 where A={1,2,3,4,5,6,7}
2) Find the number of 4 people medical crew from 4 doctors and 5 nurses, containing at least 1 nurse.
3) In how many ways can 15 people form three teams of 5 people where the teams are called team A, team B and team C?
4) Two students from three different schools will participate in a chess tournament. In the first round of the tournament, each student will be paired with a student from a different school to play a match. According to this, how many different ways can the pairings be made in the first round? (university entrance exam question)

Grade	10
Session	2
Unit	Porabability
Session topic	Revision
No of students (av.)	25
Textbooks	The Mathematics 10, The Math Book 10

Part 2:

1. A and B are two mutually exclusive events. If $P(A` \cap B) = 0.5$ and $P(A` \cap B`) = 0.2$, find P(A).

2. We are given seven courses, each with a set number of weekly teaching hours as follows:
   Course 1: 5 hours
   Course 2: 4 hours
   Course 3: 4 hours
   Course 4: 5 hours
   Course 5: 3 hours
   Course 6: 5 hours
   Course 7: 5 hours

   Suppose we randomly select four courses from this pool. Determine the probability that the total number of weekly teaching hours of the selected courses equals 17. (University Entrance Exam)

3. A digital lock is operated by a 3-character code. Each character can be any lowercase letter ('a' to 'z') or digit (0 to 9). This means each position has 36 possible characters (26 letters and 10 numbers). What is the probability that a code, randomly set, starts with a letter and ends with a number?

4. 5 adults and their 7 children are going to the cinema. They will sit randomly in three rows of 4 seats each. What is the probability that four of the adults will be seated at the back row?

5. Find the probability that a random arrangement of the twelve letters in the word HIPPOPOTAMUS begins with a vowel and that every O follows a P (i.e., O only occurs after P). (Take the vowels in the English alphabet to be A, E, I, O, and U)

Part 3:

1) A and B are two mutually exclusive events. If P(A∩B`)=0.4 and P(A`∩B)=0.2, find P(A`∩B`).

2) Evrim is packing for a hiking trip and wants to ensure she has enough water for the journey. She has several bottles of different capacities but can only fit **3 bottles** in her backpack. The bottles she has are:
4 bottles, each holding 500 ml,
3 bottles, each holding 750 ml,
2 bottles, each holding 1000 ml.
If Evrim randomly selects 3 bottles to pack, what is the probability that she will have **exactly 2500** ml of water for her trip?

3) Suppose the digits 105054020 are randomly rearranged into a new 9-digit number. What is the probability that this new number starts with an even number and that each 5 is followed by a 0?

4) Consider a bookcase with 3 shelves, each holding 3 books. It is desired to place 4 different mathematics, 2 different biology, and 3 different chemistry books on the bookcase. What is the probability that there are only mathematics books on the top shelf, assuming these books are randomly placed on the shelves?

Grade	10
Session	3
Unit	Functions
Session topic	Revision
No of students (av.)	25
Textbooks	The Mathematics 10, The Math Book 10

Part 2:

6. Determine if the following relations represents a function or not. Justify your answer by giving reasons.
   a) $f = \{(1,2), (2,2), (3,2), (4,2)\}$ defined on set $A = \{1,2,3,4,5\}$
   b) $g(x) = \sqrt{x} + \sqrt[3]{2x-5}$ where the relation is defined from $R^+ \cup \{0\}$ to R.
   c) $y^2 + x = 5$ where $x, y \in R$

7. Find the possible largest domain of the function $f(x) = \frac{x+3}{\sqrt{x-5}} + \sqrt[4]{7-x}$.

8. $f: A \longrightarrow B$ is a one to one and onto function. The number of elements in set A and B are $n(A) = 5t - 12$ and $n(B) = 3t - 6$.
   Another function will be defined from B to C where $n(C) = 2n(A)$. How many different constant functions can be defined from B to C?

9. f, h and g are functions defined on R.
   If $h(x) = f(x) + g(3x-5)$ is a linear function where $g(x)$ is an identity function, evaluate $h(-2)$ where $f(-3) = 4$ and $f(2) = -6$.
10. As a marketing coordinator, you are analyzing the impact of advertising on product sales. You have observed that starting from an initial sales amount (without any advertising), each additional 1,000 TRY spent on advertising leads to an increase of 15 units in product sales. If you know that without any advertising, the sales are 100 units, formulate the linear function that describes this relationship.
    a) Define the linear function $f: A \longrightarrow S$ where A is the amount spent on advertising (in thousand TRY) S is the total units sold. Use the given initial sales and the rate of increase in sales with advertising expenditure.
    b) Calculate the expected sales if 17000 TRY is spent on advertising.

Part 3:

1. Determine if the following relations represents a function or not. Justify your answer by giving reasons.
   d) $f = \{(1,3),(2,1),(3,5),(4,2),(3,1),(5,1)\}$ defined on set $A = \{1,2,3,4,5\}$
   e) $y^3 - 2x = 0$ where $x, y \in R$
2. Find the largest possible domain for the function $f(x) = \frac{\sqrt{3-x}}{\sqrt[4]{x-5}} + \frac{\sqrt[3]{x+3}}{x^2+1}$
3. f, g and h are functions defined on R.
   $g(x)$ is a constant function and $f(x)$ is a linear function where $f(3) = 2$ and $f(-1) = -10$
   If $h(x) = \frac{f(x)}{g(2x-5)}$ and $h(9) = 4$ then find $g(2023)$.

4. As a digital content manager, you are exploring the relationship between the number of blog posts published on your company's website and the increase in web traffic. You've found that with no new blog posts, the website attracts 500 visitors per week. Your analysis shows that for each additional 5 blog post published per week, there is an increase of 100 visitors.

a) Define the linear function f:B⟶V where B is the number of blog posts published per week and V is the total number of website visitors per week.
b) Calculate the expected number of website visitors per week if 12 new blog posts are published.

Grade	10
Session	4
Unit	Composite and Inverse Functions
Session topic	Revision
No of students (av.)	25
Textbooks	The Mathematics 10, The Math Book 10

Part 2:

1. Let f(x) $= \frac{1}{x^2-1}$ and $g(x) = \sqrt{x+1}$ are given. Find the domain of $f(g(x))$.
2. A city's public transportation system includes a bus service and a subway. The number of people using the bus service on any given day can be modelled by the function $B(d) = 100d - 500$, where d is the day of the month. The expected number of people using the subway depends on the expected number of bus users and can be modelled by the function $S(p) = 0.75p + 200$, where p is the number of expected bus users. The city transportation department wants to create a composite function T(d) to directly relate the day of the month to the expected number of subway users.
   a) Determine the explicit form of the composite function T(d).
   b) Calculate the day of the month if the expected number of subway users is 950.

3. For the two functions f and g, composition of these functions is equal to their product ($(f \circ g)(x) = f(x) \cdot g(x)$) and $f(x) = 2x - 5$ are given. Find the value of $g(3)$.

4. $f^{-1}(x) = x^3$ and $g(x) = \frac{2x-1}{x+3}$ are two functions defined on their largest domain. Find $(f \circ g)^{-1}(x)$

5. f(x) is a linear function which is self-inverse (the inverse of the function is itself). If $g(x)$ is the identity function and $f(3) - g(2) = 13$, find f(x).

Part 3:

1. Let m(x) $= \sqrt{5-x}$ and $n(x) = \frac{2}{x-3}$ be given functions. Determine the domain of $n(m(x))$.
2. In a forest, the number of birds observed depends on the number of trees. If t represents the number of trees, the number of birds observed is given by $B(t) = t^2 + 4t$. Also, the amount of bird food needed each day depends on the number of birds. The bird food in kilograms is given by $F(b) = 3b + 2$, where b is the number of birds. The forest ranger wants to know the total amount of bird food needed each day based on the number of trees.
   a) Write a formula G(t) that shows the total bird food needed based on the number of trees.
   b) If there are 6 trees, how much bird food is needed?

3. $(f^{-1} \circ g)^{-1}(x) = \frac{2x-1}{x+3}$ and $g(x) = x - 2$ are two functions defined on their largest domain. Find $f(x)$.

4. Let p(x) is a linear function which is self-inverse (the inverse of the function is itself). If $g(x)$ is the identity function and $p(g(3)) = 7$, find $p(x)$.

Grade	11
Session	1
Unit	Analytical Study of Lines
Session topic	Writing Equation of Lines
No of students (av.)	25
Textbooks	The Mathematics 11, The Math Book 11

Part 2:

1) Find the equation of the line which passes through $A(-2,3)$ and parallel to $2x - 3y + 5 = 0$.

2) Write the equation of the line passing through the point $C(2,3)$ and perpendicular to the line passing through the points $D(2,1)$ and $E(-1,3)$.

4) A plane flies along a straight line L_1 with equation $y = 2x - 5$. Determine the equation of the line L_2 which represents the trajectory of another plane flying at the same altitude that should not intersect L_1 and should pass through $(3, -1)$.

4) a and b are real numbers.
$3y = 2x + a$
$by = 3x - 12$ are perpendicular lines and intersects at y-axis. Find $a + b$. (University enterance exam question)

Part 3:

1) Find the equation of the line which passes through $A(1, -2)$ and perpendicular to $-x + 2y = 4$.

2) Find the equation of the line which passes through $B(0,3)$ and parallel to the line passing through $C(-1,2)$ and $D(3, -4)$.

3) A straight connecting street segment is built perpendicular to an existing street with equation $y = -2x + 3$. Determine the equation of the line of the new street segment which passes through point $B(-1,2)$.

4) On the cartesian coordinate system, line d is istersecting with the line $2x + y = 12$ at the point $A(4,4)$. If these two lines divides any circle with center $A(4,4)$ into four pieces of equal area, then find the equation of line d. (University enterance exam question)

Grade	11
Session	2
Unit	Analytical Study of Lines
Session topic	Revision
No of students (av.)	25
Textbooks	The Mathematics 11, The Math Book 11

Part 2:

11. On the cartesian coordinate plane, line L is passing through $A(-1,2)$ and parallel to $2y-4x-5=0$. $A(a,0)$ and $B(0.b)$ are the $x-intercept$ and $y-intercept$ of the line L respectively. Find the sum of a and b.
12. Find the equation of a line which passes through the midpoint of $A(3,5)$ and $B(-1,3)$, and perpendicular to the line passing through A and B.

13. If the centroid of triangle ABC which has the vertices $A(a-4,-a-5)$, $B(-2a+1,2a-7)$ and $C(4a-6,2a-3)$ is on the 4th quadrant then find the value of a where a is an integer.

14. If the lines below intersect at the same point on the cartesian coordinate plane, find the value of m.

- Line 1 passes through $(0,1)$ and has a slope m
- Line 2 passes through origin and has a slope $2m$
- Line 3 passes through $(1,0)$ and has a slope 3m

(University Entrance Exam)

15. Duru pinned her position on a map at $(1,-1)$ and drew the path she wanted to reach as a straight line. She noticed that the line she drew passes through $(-1,3)$ and $(-5,6)$. Help Duru find the shortest distance from her position to the path by describing what she should do and finding the distance.

Part 3:

1) Let $A(a,0)$and $B(0.b)$are the $x-intercept$ and $y-intercept$ of the line L, $2x-3y-12=0$. Find the equation of the which passes through (a,b) and parallel to the line $3y-6x+8=0$.

2) The midpoint of the $A(a-3,2a+4)$ and $A(a+5,-a-1)$ lies in the second quadrant. If a is an integer, find on which quadrant where $C(a-3,2a)$ is.

3) L is the line which is passing through origin perpendicular to line passing through $A(1,-1)$ and $B(7,-3)$. If the lines L, $y=mx+3$ and $y+2mx+9=0$ have only one point in common, then find m.

4) Ali pinned his position on a map at the centroid of the triangle formed by his school $S(10,-8)$, his home $H(-2,2)$ and his favorite coffee place $C(-2,-3)$. One of his friends, Nil pinned her own place on the straight line passing through the school and the coffee place and at the closest point to Ali. Find the distance between Ali and Nil.

Grade	11
Session	3
Unit	Applications of Functions
Session topic	Revision
No of students (av.)	25
Textbooks	The Mathematics 11, The Math Book 11

Part 2:

1) Consider the quadratic function $f(x) = -2x^2 + 8x + 3$, which models the height (in meters) of a ball thrown upwards, with x representing the time in seconds after the ball has been thrown.

Calculate the average rate of change of the ball's height over the interval from 2 seconds to 4 seconds.

2) Consider a parabola defined by the equation $y = ax^2 + bx + c$, where a, b and c are real numbers. This parabola intersects with the line $y = 6$ at exactly one point. If the parabola intersects x-axis at
$x = -1$ and $x = 5$, then determine the signs (positive, negative, or zero) of the coefficients a, b and c.

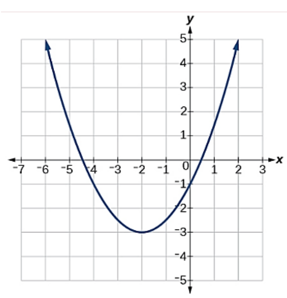
3) Consider the function $F(S) = 0.005S^2 - 0.6S + 25$, where $F(S)$ represents a vehicle's fuel consumption in litres per 100 kilometres and S is the speed of the vehicle in kilometres per hour. For speeds ranging from 50 to 100 km/h, calculate the minimum and maximum fuel consumption value.

4) The line $y = mx - 4$ is tangent to the parabola $y = x^2 - x$. Find the possible value(s) of m.

5) Determine the area of triangle TAB formed by a quadratic function $f(x) = ax^2 + bx + c$, The vertex of the parabola is denoted as point T. The x-intercepts of the parabola are at points A(-2,0) and B(8, 0), and the y-intercept is at (0, -16).

Part 3:

1) Examine the quadratic function $g(x) = -3x^2 + 10x + 5$, which represents the altitude (in meters) of a drone flying vertically, where x denotes the time in seconds after the drone's ascent began. Determine the average rate of change in the drone's altitude during the time interval from 1 second to 4 seconds.

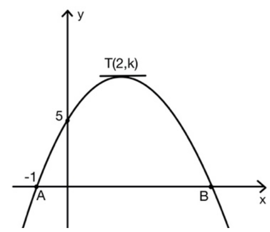


2) For the quadratic function $f(x) = ax^2 + bx + c$ whose graph is given below, determine the sign of a, b, c and Δ.

	Δ	a	b	c
Sign				

3) Consider the function $P(t) = 2t^2 - 8t + 10$ where $P(t)$ represents the profit in dollars of a small bakery and t is the number of hours the bakery is open each day. Determine maximum and minimum profit that the bakery can expect when it is open for a duration ranging from 1 to 5 hours per day.

4) Determine the area of the triangle TAB formed by quadratic function $f(x) = ax^2 + bx + c$



Grade	11
Session	4
Unit	Transformation of Functions and Solution of second degree equation systems with two unknowns
Session topic	Revision
No of students (av.)	25
Textbooks	The Mathematics 11, The Math Book 11

Part 2:

1) If a function $f(x) = x^2 - 4x + 5$ is translated a units right and b units down, then $g(x) = x^2 - 6x + 4$ can be obtained. Find the value of $a \times b$.

2) Describe, in order, a sequence of transformations that maps the graph of y =f (x) onto the graph

$y = -3f(x) + 2$.

3) $\frac{7}{x+1} - y = 8$

$x - \frac{5}{y+1} = 6$

Find the value of $x + y$.

4)

$$x^2 + y^2 = 11$$
$$\frac{x}{x+y} + \frac{y}{x-y} = \frac{11}{7}$$

Find the possible y values that satisfied the system of the equations.

5) Elif is designing a rectangular playground. The length of the playground is y meters and the width is x meters. She plans to install a special rubber surface on the playground, which costs $30 per m^2. The diagonal of the playground is $\sqrt{500}\, m$. Elif had spent of $6000 for the rubber surface. What are the dimensions of the playground?

Part 3:

1) A roller coaster's path can be modeled by the function $h(x) = x^2 - 8x + 20$ where $h(x)$ represents the height of the coaster at a horizontal distance x meters horizontally from its starting position. To enhance the ride experience, the engineers plan to alter the route. They intend to shift the entire coaster path 2 meters to the left and 3 meters lower than its current position. Write the equation in the form $y = ax^2 + bx + c$ that will represent the new path of the roller coaster.

2) Describe, in order, a sequence of transformations that maps the graph of $y = f(x)$ onto the graph

$y = f(-2x + 6)$.

3) $4xy = y - x$

$\frac{3}{x} + \frac{2}{y} = 9$

Find the value of $\frac{x}{y}$.

4) A farmer is constructing a rectangular enclosure for her sheep. To strengthen the structure, she plans to add a diagonal support from one corner to the opposite corner. The farmer knows that the area of the enclosure must be exactly 12 m^2 to provide enough space for her sheep. Additionally, she has calculated that the length of the diagonal should be equal to 5 m for safety reasons. Find the possible dimensions of the enclosure.