Does AI Help or Hurt Learning? *

Catalina Franco† Natalie Irmert‡ Siri Isaksson§

March 26, 2026

Abstract

AI is transforming how students learn, raising concerns about whether it expands educational opportunities or widens existing gaps. We examine this question in a pre-registered lab experiment (N=572) in which students study a novel topic under one of three conditions: browsing only (control), AI-assisted, or AI-guided, and then complete an exam without AI access. We find no overall effect of AI access on learning outcomes. However, this average zero effect masks substantial heterogeneity. High-GPA women appear to benefit the most from AI-guided access, while the effects on men and low-GPA students are weaker and in some cases negative. We also find that students with AI access attempt fewer practice questions during the study phase. This suggests that studying with AI crowds out other learning activities, but does not lead to an overall change in exam performance. Finally, exploratory analyses of prompt data provide suggestive evidence on why some students benefit more than others. More delegative AI use (measured by copy-pasting practice questions into the chatbot) is associated with attempting more questions but performing worse on the final exam. High-GPA women rely on this strategy the least and perform the best. Overall, AI appears to crowd out some traditional study effort without reducing learning on average, but because its benefits are concentrated among already advantaged students, it may reinforce existing educational inequalities.

KEYWORDS: Generative AI; learning outcomes; study behavior; educational inequality; randomized experiment; human-AI interaction

JEL CODES: C90, D83, I23, J16, O33

*We are grateful for funding from the Jan Wallanders och Tom Hedelius stiftelse samt Tore Browaldhs stiftelse (grant P24-0189). The funding body had no role in the study design, data collection, management, analysis or interpretation of the data or preparation of the manuscript. We thank Iver Finne, Jakob Svensson, Adam Zylbersztejn, and the seminar audiences at the 2025 ASSA Meeting in San Francisco, IFN Stockholm, Umea University, NTNU Trondheim, 2025 ESA-Asia, ISER Osaka Workshop, 2025 AFSEE Meeting, 2025 SEET Meeting in Dijon, 2025 CEN in Aarhus, the 2025 Stockholm Uppsala Education Economics Workshop, RABE Workshop at Gate Lyon 2, LEMNA Lab at University of Nantes, and the 2025 CESifo / ifo Junior Workshop on the Economics of Education. The study received ethical approval from the Nottingham School of Economics and the Norwegian School of Economics. We pre-registered the study with a pre-analysis plan at the AEA RCT registry: AEARCTR-0014930.

†Center for Applied Research (SNF) at NHH – Norwegian School of Economics.

‡Lund University **Corresponding author.** Address: Lund University, Department of Economics, P.O. Box 7080, 220 07 Lund, Sweden. Email: natalie.irmert@nek.lu.se

§Burgundy School of Business Groupe ESC Dijon Bourgogne - CEREN.

1 Introduction

The rapid spread of generative artificial intelligence (AI) has sparked a growing debate about its consequences for education (Bengio et al., 2026; Mishra et al., 2025). Optimistic views emphasize that AI can expand access to explanations, feedback, and personalized support, potentially helping students learn more effectively. More skeptical perspectives argue that these tools may instead weaken learning by allowing students to bypass the effort required for understanding and skill formation. Because learning depends not only on access to information but also on active engagement, a central question is whether AI complements students' effort or substitutes for it. The answer matters not only for average learning outcomes, but also for inequality: if some students are able to use AI more productively than others, the technology may widen or reduce existing gaps in educational achievement.

Despite the intensity of this debate, credible causal evidence on how AI affects learning remains limited. A central challenge is that AI use is difficult to observe and even harder to restrict in field settings, making it challenging to compare students with and without access to the technology. Causal studies are still few and often use small samples or face important identification challenges, including differences in teachers, classroom environments, and compliance with treatment assignment (Bastani et al., 2025; Lehmann et al., 2024; Tan and Rajaratnam, 2024). In this study, we set out to fill this empirical gap.

To the best of our knowledge, we report on the first pre-registered laboratory experiment designed to study the effects of generative AI on learning. The laboratory setting allows us to cleanly isolate access to AI, eliminate spillovers across students, and hold constant the topic, materials, and study environment. While such a design is naturally better suited to measuring short-run learning than long-run skill accumulation, it offers strong internal validity and allows us to examine not only average effects, but also distributional consequences across student groups and the behavioral mechanisms through which AI may affect learning. In particular, the laboratory setting allows us to observe and analyze students' prompts in a way that is typically not feasible in real-world educational settings, providing a unique window into how students use AI during the learning process.

Our experiment consists of three between-subject treatment conditions. University stu-

dents are asked to study a topic that is new to them—Esperanto—by reviewing lesson materials and working through practice questions during a fixed study period.[1] In the control condition, students study with access to Google Search but without AI, approximating the pre-AI status quo. In the AI-assisted condition, students have access to a generative AI chatbot but receive no formal instruction on how to use it beyond basic operational guidance. In the AI-guided condition, students also have access to AI, but are additionally given structured guidance on how to use the tool in ways intended to support their own learning.[2] After the study period, all students complete the same exam without access to AI or other external aids, allowing us to identify the causal effect of AI access during study on learning.

Our experiment yields three key findings on the impact of AI tools on learning outcomes. First, we find no evidence that AI affects learning on average. Contrary to common concerns that AI may undermine learning, the estimated effects on both exam performance and self-assessed learning are small and statistically insignificant. AI access neither improves nor worsens performance on average: students in all treatment conditions score about 8 out of 15 on the final Esperanto exam. Self-assessed learning is also high across groups, with about 80% of students across treatment conditions reporting that they agree or somewhat agree that they learned a lot while working through the practice questions.

Second, this average zero effect masks substantial heterogeneity across students. The subgroup that appears to benefit most from AI access is high-GPA women, for whom the AI-guided treatment increases the likelihood of reaching the top of the score distribution. This suggests that AI may reinforce existing inequalities in educational achievement, particularly if students who are already high-performing are better able to translate AI access into learning gains. Given that women have outperformed men in educational attainment for decades in most countries (e.g., Voyer and Voyer, 2014), these results also raise the possibility that AI could widen existing gender gaps in the grade distribution.

Third, AI access changes study behavior via study time allocation between chatting with

[1] Esperanto is a constructed international language designed for easy learning and communication. Its simplicity and unfamiliarity to our sample make it well suited to the experiment, since students can learn meaningful material in a short period while starting from a common baseline.

[2] For example, we provided advice such as "Rather than asking for direct answers, request explanations or steps."

the bot and active problem solving. Students in the AI-assisted and AI-guided treatments attempt about 4.0 and 5.6 fewer practice questions, respectively, relative to a control-group mean of 33.8 questions for women and 35.6 for men. Thus, AI crowds out conventional practice during the study phase, but this reduction does not produce an average decline in exam performance. The subgroup results suggest that high-GPA women reduce attempted practice less than other students under AI-guided access, consistent with a less substitutive use of AI.

In addition, for students in the AI treatments, we conduct an exploratory analysis of prompts to shed light on a plausible mechanism behind why high-GPA women are the group that benefits most. Specifically, we match our experimental data to the text data of participant's prompts to the AI chatbot and define a measure of delegation to AI based on the extent to which participants copy-paste, i.e., insert practice questions verbatim into their prompts. There are three key takeaways from this analysis. First, more prompting is associated with more attempted practice questions. However, this does not translate into learning: there is no correlation between the number of prompts and final exam score. Second, copy-pasting is negatively associated with final exam scores: participants with the highest copy-paste index, score on average 6.7 on the final exam, a significant difference of 16% from the control mean equal to 8. Third, high-GPA women are the group least likely to use copy-pasting as a learning strategy. Taken together, these patterns suggest that students who primarily use AI to delegate work tend to have lower exam performance. This may help explain why high-GPA women benefit more from AI-guided access.

Through this paper, we contribute to three strands of literature. First,we contribute to the literature on how students use generative AI and how such use relates to academic performance. To date, evidence in this area remains limited and is largely descriptive. For instance, Wecks et al. (2024) show that students identified as AI users perform worse on exams on average, while Pallant et al. (2025) find that learning outcomes differ markedly depending on whether students use AI to construct knowledge or simply to complete tasks procedurally. Our paper advances this literature by moving beyond correlations and experimentally varying AI access. This approach allows us to identify the causal effect of AI availability sep-

arately from selection into AI use, given that nearly 90% of students assigned to AI access actively engage with the tool.

We also contribute to the growing experimental literature on AI and learning. Recent work has shown that AI can both help and hinder skill acquisition depending on how it is deployed. In a large field experiment in high-school mathematics, Bastani et al. (2025) show that unrestricted access to GPT-4 improves performance during practice but can reduce subsequent performance once AI is removed, whereas tutor-like safeguards mitigate these negative effects. Similarly, Lehmann et al. (2024) find that Large Language Models (LLMs) can support learning when students use them as tutors, but harm learning when students rely on them to solve exercises for them. Related work, summarized in Han et al. (2025), points to substantial heterogeneity across contexts, subjects, and designs. We contribute to this literature by providing, to our knowledge, the first pre-registered laboratory experiment in economics on AI and learning. Our design allows to circumvent potential pitfalls of field experiments by offering tight control over AI access, common learning materials, and the study environment. It also allows us to document substantial heterogeneity in who benefits from AI, a distributional dimension that remains largely unexplored. Finally, our data allow us to open the black box of student-AI interaction by analyzing prompt behavior directly.

Our third contribution is to the literature on guided and curated AI tutoring. A central insight in this work is that the educational value of AI depends not only on access, but also on how interaction with the system is structured. Studies of customized AI tutors and personalized AI feedback generally find positive effects of AI tutors and stronger learning effects when the technology is designed to scaffold students' reasoning (Pardos and Bhandari, 2024; Kestin et al., 2025; Dai et al., 2025). We complement this literature by comparing standard AI access to an AI-guided condition that provides brief, scalable instructions on how to use the tool to support learning. Put differently, rather than studying highly curated AI tutoring systems designed to optimize learning, we focus on the kind of everyday use of off-the-shelf AI tools that students engage in when given access. Also, unlike highly curated tutor systems, our AI guided intervention is much closer to the kind of lightweight guidance that schools and universities could realistically implement at scale. This comparison there-

fore speaks not only to whether simple guidance can make AI use more productive, but also to an important practical issue in less structured settings: time spent interacting with AI may crowd out time that would otherwise be devoted to conventional study activities.

Taken together, our findings suggest that AI may change not only how much students learn, but also how they learn. We provide suggestive evidence of how using AI as a substitute rather than complement to own effort while studying hurts learning. We also find evidence of winners from AI assisted studying: women with High-GPA are expected to benefit the most, despite being the group with the lowest adoption rates as shown in previous research (Carvajal et al., 2024). These patterns raise a broader concern that AI may reinforce existing inequalities in educational achievement if some students are able to use it more productively than others. This possibility is especially salient in light of growing concern over the "boy crisis" in education in many OECD countries (Autor and Wasserman, 2013; Autor et al., 2023).

2 Experimental Design

We conducted the experiment in the CedEX Lab at the University of Nottingham in December 2024. Participants (N=572) were undergraduate and graduate students from programs in all fields of study across the university.

The experiment was designed to mimic the stages of a typical university course, where students first learn about the new topic, then deepen and practice the content through practice questions, and finally are tested through an exam. To test how generative AI affects learning, we experimentally varied students' access to AI tools during the second stage of the experiment (completion of practice questions). In particular, participants were randomly assigned to one of three treatments, which are explained in more details below: T1: no access to ChatGPT (Control), T2: access to ChatGPT (AI-assisted), and T3: guided access to ChatGPT (AI-guided). For students in the AI treatments, we provided paid ChatGPT accounts that were already logged in on the lab computers to which students were randomly assigned. To preserve treatment compliance, AI websites were blocked on all machines, students were

not allowed to use phones inside the lab, and they could not access their own AI accounts.

2.1 Learning task

All participants, independently of treatment status, learned about the same topic: Esperanto—a constructed international language designed for easy learning and communication. Its simplicity and unfamiliarity to the sample made it ideal for measuring how well students learn new material, rather than relying on their prior knowledge.[3] Holding the topic constant across treatments is important for identification, as it ensures that any differences in learning outcomes can be attributed to differences in access to AI rather than to differences in subject matter or teacher delivery.

Learning a new language provides a useful setting for studying learning because it requires students to absorb new vocabulary and rules and then apply them in practice. This makes it well suited to our setting, where the goal is to measure how students learn genuinely new material in a short period of time. Esperanto is particularly useful for this purpose because its grammar and word formation are highly regular and transparent, allowing students to make measurable progress quickly without reducing the task to simple memorization (Koutny, 2015). Although broader claims about general cognitive benefits from language learning remain debated, the literature strongly supports the view that learning a language involves understanding and applying explicit rules, which is central to the kind of short-run learning we study here (DeKeyser, 2003). Thus, while knowledge of Esperanto is not useful in itself, it is particularly well suited for our study because learning it requires the acquisition of new rules and vocabulary while starting from a common baseline of unfamiliarity among students.

2.2 Session procedures

Each session adhered to the same procedure:

[3]This is a key design choice in our study because previous knowledge of a topic may discourage students to use AI since they can answer the questions themselves. We would then not be able to estimate the effect of AI on learning.

1. **Stage 1**: Students learn Esperanto for 15 minutes using written learning materials provided by the researchers.[4]

2. **Stage 2**: Students complete practice questions with access to different learning aids depending on treatment for about 20 minutes. This is the stage where the treatment variations take place.

3. **Stage 3**: Students take an exam testing their learning without access to any notes or learning aids.

4. **Stage 4**: Students take a post-study survey asking them about demographics as well as their learning experience during the session and attitudes and beliefs about AI.[5]

The entire experiment was conducted in Qualtrics. The learning materials and practice questions were presented across different pages, which allowed participants to move back and forth. More questions than possible to answer in the allotted time were included to accommodate different learning speeds. When the time for the lesson and practice questions was up, the screen moved automatically to a wait screen so that everybody would start the exam at the same time. There was no time limit on the exam but it was calibrated to last about 10 minutes on average. The total duration of each session was about one hour.

2.3 *T1: No access to ChatGPT*

T1 was designed to mirror learning before the advent of generative AI and to work as a baseline against which we can measure the impact of (guided) AI access on learning outcomes. As students had access to a web browser and online learning before the advancement of generative AI, we provided participants in this group access to a web browser during stage 2. Websites providing access to AI chatbot sites, however, were blocked in T1. See instructions for students in Figure C.1.

[4]Although 15 minutes may seem brief for learning, Esperanto is a highly regular and easy-to-decode language, with simple grammatical rules and consistent morphology. In that time, the lesson covered personal pronouns, verbs in the present and past tense, negation, nouns, adjectives, adverbs, and plural formation.

[5]Basic questions such as gender and experience with AI were asked at the beginning of the study, before stage 1.

2.4 *T2: Access to ChatGPT*

T2 was designed to mirror learning when ChatGPT is available to students without clear guidelines, structure and training—a situation in which it is up to each student whether and how she or he wants to use AI in the learning process. This mirrors the situation which many universities face as they are grappling with the important task of adjusting their policies to the advent of AI.

Students accessed ChatGPT on their assigned computer browser with a already logged in premium ChatGPT account. The account was provided by the researchers and was cleared of any memory. The prompting history was archived and saved by researchers for analysis of prompting. See instructions for students in Figure C.2.

2.5 *T3: Access to Guided ChatGPT*

This treatment is the exact same as T2, except that students now also receive guidance (written by the researchers) on how to use ChatGPT. The idea behind this treatment is to explore the potential of providing not just equal access but also equal training and encouragement in AI use for students. See instructions for students in Figure C.3.

2.6 Incentives

We aimed to mimic the incentives provided in typical college courses in the lab. The incentives in most real courses around the world are to assign a smaller percentage of the final grade to assignments and practice questions and to have a larger weight on a final exam or midterms and a final exam. The idea of the assignments with lower percentages is to incentivize students to cover the materials and prepare them for the exam(s). In the same spirit, we offered incentives to reward correct responses and help students deepen their learning in the Esperanto lesson, while offering higher incentives for every correct answer in the exam. The structure of incentives is as follows:

- GBP 5 for completing the study

- GBP 7 for correctly answering at least 20 out of 92 practice questions[6]
- GBP 1 per correct answer in the test (maximum amount GBP 15)

The maximum payment they could earn is GBP 27.

From the beginning of the session students were informed that the most important part of the session (and where they can earn the most) is the exam. This is emphasized throughout the session. The intention is that they focus their efforts of trying to learn as much as possible form all the resources they have available, just as they would in a real-life course.

3 Empirical Strategy

3.1 Outcomes of Interest

We focus on three primary measures of learning. First, our main variable of interest is exam performance, captured through both an objective and a subjective metric. The objective measure is the total number of correct answers on the final exam (ranging from 0 to 15). The subjective measure is students' self-assessed learning, recorded as an indicator equal to one if they slightly or strongly agreed with the statement "I feel like I learned a lot from solving practice questions" (on a four-point scale), and zero otherwise. For both measures, we investigate heterogeneous effects by background characteristics such as gender and GPA.

Second, we examine distributional outcomes by estimating the likelihood of students scoring in the top or bottom of the test score distribution. A top scorer was prespecified in the pre-analysis plan (see Section E) as having more than 10 correct answers, while a low scorer was prespecified as solving less than 5.

Finally, we assess students' exam preparedness by analyzing two measures: the number of practice questions solved correctly and the amount of time spent completing the final exam.

[6]We set this low threshold on purpose so that everybody was able to earn the GBP 7.

3.2 Econometric Specifications

The main specification regresses the outcomes presented above on indicators for the two AI treatments, keeping the control as the excluded indicator:

$$y_i = \alpha_0 + \alpha_1 \text{AI-assisted}_i + \alpha_2 \text{AI-guided}_i + \varepsilon_i \tag{1}$$

The main specification adds baseline covariates to control for different levels of AI adoption and students characteristics.[7]

We asses whether there are any gender differences in responses to the treatment using the specification:

$$\begin{aligned} y_i = \beta_0 + \beta_1 \text{AI-assisted}_i + \beta_2 \text{AI-guided}_i + \beta_3 \text{Female}_i + \beta_4 \text{AI-assisted}_i \times \text{Female}_i \\ + \beta_5 \text{AI-guided}_i \times \text{Female}_i + \varepsilon_i \end{aligned} \tag{2}$$

Where the coefficients β_1 and β_2 measure the effects of the treatments among men, β_3 measures any gender gaps in the control group, and β_4 and β_5 measure the differential effects of the treatments for women.

A similar specification replaces the Female variable with high-GPA, a measure of academic skill equal to 1 if students are in first-class honours (GPA 70% or above) and 0 otherwise.

A final specification involves the interaction effects between treatments, gender and GPA. Following the previous literature, this specification may allow us to see differential effects by gender and GPA simultaneously. For example, if the treatments affect high-GPA female students differentially.

[7]The covariates include: Age 21 or younger, undergraduate student, indicators for field of study, use AI occasionally or all the time, has a paid subscription to an AI provider. In regressions without interactions with gender and high-GPA, we add these variables as covariates.

$$
\begin{aligned}
y_i = \gamma_0 + \gamma_1 \text{AI-assisted}_i + \gamma_2 \text{AI-guided}_i + \gamma_3 \text{Female}_i + \gamma_4 \text{high-GPA}_i + \\
\gamma_5 \text{AI-assisted}_i \times \text{Female}_i + \gamma_6 \text{AI-assisted}_i \times \text{high-GPA}_i + \gamma_7 \text{AI-guided}_i \times \text{Female}_i + \\
\gamma_8 \text{AI-guided}_i \times \text{high-GPA}_i + \gamma_9 \text{AI-assisted}_i \times \text{Female}_i \times \text{high-GPA}_i + \\
\gamma_{10} \text{AI-guided}_i \times \text{Female}_i \times \text{high-GPA}_i + \varepsilon_i
\end{aligned} \tag{3}
$$

The triple interactions measure whether top female students are differentially affected by the treatments. The double interactions with high-GPA will measure if top male students are differentially affected by the treatments.

3.3 Balance across Treatments and Descriptive Statistics

Table 1 reports baseline characteristics by treatment status. Overall, the randomization produced groups that are similar along most observed dimensions. The shares of women, younger students, undergraduates, field of study, prior AI use, and paid AI subscriptions are close across the control, AI-assisted, and AI-guided groups. One variable is not balanced: the share of high-GPA students is lower in the AI-guided group than in the control group (22.6% versus 36.8%). In all our results, we control for baseline covariates.

The experimental sample consists of 604 students. In terms of composition, about 61% of participants are women, 57% are aged 21 or younger, and 64% are undergraduate students. Academic background is mixed across fields of study, and about 34% of participants are classified as high-GPA students. Prior exposure to AI tools is common: roughly 63% report using AI occasionally or all the time, while 7% report having a paid AI subscription. Because 22 students left the survey while working on the exam despite being instructed not to do so, we do not have outcomes for them and we classify them as attritors (see last row of Table 1). Table A1 shows that the sample remains balanced across treatment arms when these attritors are excluded. We further drop 10 students who do not report a binary gender. Thus, the analytical sample consists of 572 students.

Figures 1 and 2 provide a non-parametric view of the exam performance outcome dis-

tributions. Figure 1 shows substantial overlap in exam-score distributions across the three treatment arms in the full sample, consistent with the absence of large average treatment effects. Figure 2 suggests, however, that the distributions differ somewhat across subgroups. In particular, the score distribution for women and for high-GPA students appears somewhat more right-shifted under AI access, whereas the distributions for men and low-GPA students show less favorable patterns relative to the control group. The left tails of the distributions are typically longer in the control group, in particular for men (Figure 2b) and high-GPA students (Figure 2c). In the distribution for men, it also appears that not many achieve top scores in the control group, but more men are able to achieve high scores under the AI treatments (Figure 2b). The distribution of high-GPA students under the AI-guided treatment seems more compressed (i.e., not many obtaining low scores or high scores) relative to both the control and AI-assisted distributions (Figure 2c).

When we look at average exam performance, we do not see any differences as students in each treatment condition obtain a score of 8 (out of 15) on average (Figure 3). However, the descriptive differences in the density plots suggest that the average effects hide substantial heterogeneity and foreshadow the heterogeneous treatment effects discussed in the next section.

4 Results

4.1 No Overall Impact of AI on Learning

We begin by examining whether access to generative AI affects overall learning. This is a central outcome in the policy debate over whether AI tools should be restricted or incorporated into educational settings, given concerns that they may reduce students' own learning.

Table 2 reports the effects of the two AI treatments on our two learning outcomes: exam performance in a setting without access to external aids, and students' self-assessed learning. Across specifications, we find no evidence of an overall effect of AI access on learning. In columns (1) and (5), where we estimate specifications based on Equation 1, assignment to the two AI treatments have small effects and none of the estimates is statistically distin-

guishable from zero. Relative to a control-group mean of around 8 correct answers in exam performance and between 81% and 83% of students indicating that they somewhat agree or agree with the statement "I feel like I learned a lot working through the questions.", these effects are economically small.

Columns (2), (3), (6), and (7) of Table 2 report the pre-registered heterogeneity analyses by gender and GPA following Equation 2. These specifications do not show broad, systematic differences in the effects of AI on exam performance or self-assessed learning when heterogeneity is examined along only one dimension at a time. Taken together, they support the conclusion that AI access has no average effect on learning and provide little evidence of heterogeneity by gender or GPA separately.

Columns (4) and (8) estimate a fully saturated model with treatment effects allowed to vary jointly by gender and GPA following Equation 3. While the triple interactions are not statistically significant, their magnitudes are economically meaningful. Most notably, the interaction between AI-guided access, female, and high-GPA is associated with 1.4 additional correct answers on the exam, about 17% of the relevant control-group mean. For self-assessed learning, the corresponding triple interaction is -19.9 pp. This estimate should be interpreted relative to the positive coefficient for high-GPA men in the same specification (20.6 pp), implying a reversal in sign: while high-GPA men report higher perceived learning under AI-guided access, high-GPA women report lower perceived learning. Although we are underpowered to estimate these triple interactions precisely, the pattern suggests that high-GPA women may benefit disproportionately from AI-guided access in terms of measured performance while perceiving that they learn less, in sharp contrast to high-GPA men.

Figure 4 complements these regression results by plotting subgroup-specific treatment coefficients for exam performance. These subgroup coefficients are useful for visualizing how the estimates vary across cells defined jointly by gender and GPA, but they were not separately pre-registered; our pre-analysis plan pre-specified the interaction analyses reported in the tables. The figure reinforces the main result of no overall treatment effect, while also suggesting that heterogeneity may be concentrated in more specific subgroups rather than along gender or GPA alone. In particular, point estimates under AI-guided access are more

favorable for high-GPA women, but more negative for low-GPA women and high-GPA men (Figure 4b). Under AI-assisted access (Figure 4a), the corresponding subgroup patterns are smaller and less systematic. The confidence intervals are wide because we are not powered enough to identify small effects in the subsamples, so these subgroup differences should be interpreted cautiously.

We find a similar absence of average effects for students' self-assessed learning. Figure 5 shows that point estimates are generally close to zero and imprecisely estimated across subgroups in both AI treatments. Overall, the evidence in Table 2 and Figure 4 indicates that providing AI access neither improves nor worsens learning on average in our setting, even though some subgroup-specific patterns emerge in the estimates.

4.2 Heterogeneous Effects Emerge in the Tails of the Score Distribution

The absence of average effects on exam performance masks meaningful heterogeneity in where students end up in the score distribution. To examine this more directly, Table 3 reports treatment effects on two pre-registered binary outcomes: obtaining a top score (more than 10 correct answers) and obtaining a low score (fewer than 5 correct answers). Figure 6 complements these estimates by plotting subgroup-specific treatment coefficients for the same outcomes.

At the aggregate level, neither treatment changes the probability of reaching the top or the bottom of the score distribution. In column (1), the effects on obtaining a top score are close to zero for both AI-assisted and AI-guided. Likewise, in column (5), the effects on obtaining a low score are economically small and statistically insignificant in the full sample. Thus, the null average effects in test scores are not driven by uniform shifts in performance across all students.

The heterogeneity results point instead to differential effects across student groups, particularly under AI-guided access. The clearest pattern appears for low scores. In column (7), AI-guided increases the probability of scoring in the bottom tail by 4.6 pp, while the interaction with high-GPA is negative and statistically significant (-12.3 pp), implying that the adverse effect is concentrated among lower-GPA students. A similar, though less precisely

estimated, pattern appears for AI-assisted access. These results suggest that AI access can increase the risk of very poor performance among academically weaker students, even if average scores remain unchanged.

For top scores, the evidence is less clear. The interaction specifications in columns (2)–(4) do not indicate broad average gains in the probability of reaching the top tail, but they do suggest meaningful differences once gender and GPA are considered jointly. In particular, the fully interacted specification for AI-guided access in column (4) implies a larger representation among top scorers for high-GPA women and a lower representation for high-GPA men. Figure 6 makes this pattern even more clear: under AI guidance, the point estimate for obtaining a top score is most positive for high-GPA women, whereas the estimates are weaker or negative for low-GPA women and high-GPA men. At the same time, the bottom-tail results in the figure show that low-GPA students and, in particular, low-GPA men are the group most likely to experience an increase in very poor performance under both AI treatments.

Taken together, Table 3 and Figure 6 suggest that AI does not shift the entire score distribution upward or downward. Instead, its effects are concentrated in the tails and differ sharply across subgroups. In our setting, AI-guided access appears to help some students reach high performance while increasing the risk that other students, especially lower-GPA students, end up at the bottom of the distribution.

4.3 AI Changes Study Behavior: Fewer Practice Questions Attempted

A natural question is whether access to AI changes how students allocate study time. On the one hand, AI may substitute for active engagement with traditional course material if students rely on the chatbot for answers rather than working through practice problems themselves. On the other hand, AI could complement conventional study tools by helping students ask follow-up questions or explore material in greater depth. Understanding how AI affects study behavior is therefore important, particularly if time spent with AI crowds out traditional practice.

Table 4 and Figure 7 examine two measures of study behavior. The first outcome is time

spent on the final exam, a setting in which no student had access to AI and no time limit was imposed. The second is the number of practice questions attempted during the study session, which took place under the assigned treatment condition and within a fixed 37-minute window.[8] Because time is fixed during the study period, the number of attempted practice questions provides a useful measure of how students allocate their effort across AI use and conventional practice.

Columns (1)–(4) of Table 4 show no meaningful effect of either treatment on time spent on the final exam. Across all specifications, the estimated effects of AI-assisted and AI-guided access are small and statistically insignificant relative to an average of 10.8 minutes for women and 10.4 minutes for men in the control group. This is unsurprising, since AI was not available during the exam itself.

By contrast, columns (5)–(8) show that AI access substantially reduces the number of practice questions attempted during the study session. In the control group, students attempt between 33.8 and 35.6 practice questions on average, for men and for women respectively. Assignment to the AI-assisted treatment lowers this by 4.17 questions, while assignment to the AI-guided treatment lowers it by 5.77 questions, both highly statistically significant. These correspond to reductions of about 12 and 17%, respectively for each treatment, relative to the control-group mean. Thus, while AI does not change how long students spend on the final test, it clearly changes how they study beforehand, leading students to attempt fewer practice questions.

The pre-registered heterogeneity specifications in columns (6) and (7) do not show broad, systematic differences by gender or GPA considered separately. Figure 7 is nevertheless informative about how these effects vary across narrower subgroups. The figure suggests that the decline in attempted practice questions is concentrated among men, especially high-GPA men, whereas high-GPA women appear much less affected. In the fully interacted specification in column (8), the estimated reduction under AI-guided access is about 5.2 questions for the baseline group, but the positive triple interaction for high-GPA women is economically large, implying a much smaller decline for that subgroup. These

[8]The 37 minutes were split into about 15 minutes to study an Esperanto lesson and the remaining time for working on practice questions.

subgroup differences are not estimated precisely enough to support strong claims, but they are consistent with the broader pattern in the paper that high-GPA women appear to use AI in a way that is less substitutive of traditional study effort.

Importantly, attempting fewer questions does not necessarily imply lower efficiency. Appendix Table A2 shows that students in both AI treatment groups solve a larger share of attempted practice questions correctly relative to the control group, with little evidence of subgroup differences. The results therefore suggest a shift in study strategy: students with AI attempt fewer questions, but do so with a higher success rate. That is, studying with AI takes longer but does not meaningfully affect performance in the final exam where no AI tools were allowed.

In sum, these results suggest that high-GPA women are the subgroup that benefits most from AI-guided access. Unlike most other students, they show only a modest decline in attempted practice questions relative to their control counterparts and are simultaneously more represented among top scorers, which is consistent with a less substitutive and potentially more effective use of AI during study.

5 Who is hurt and who is helped by AI? Descriptive Evidence from Prompting Analysis

In this section, we use exploratory analyses to unpack how prompting affects learning. We create a merged dataset in which experimental outcomes, such as the score on the final exam, are matched with the prompts that each student typed. The matched dataset includes 376 students in either AI treatment, who together generated 4,452 prompts. An example prompt is seen in Figure B.1.

On average, students write 11 prompts in a session.[9] In general, the number of prompts students write does not differ by demographic characteristics or treatment assignment (AI-guided vs. AI assisted), as seen in Figure B.2. We also find no significant correlations between baseline characteristics and either the likelihood of writing zero prompts or the likelihood

[9]This average includes 33 students who wrote zero prompts.

of writing a very large number of prompts (more than 41, the 90th percentile of the prompt distribution). In particular, women are no more likely than men to write zero prompts, suggesting that the positive effects observed for high-GPA women are unlikely to be driven simply by lower AI use.

5.1 Methodology

The exploratory analysis is based on correlating the total number of prompts with two behavioral outcomes that are central for our study: the number of practice questions attempted at the study stage, and final exam score.

We also aim to provide a measure capturing the content of the prompts. As this analysis is exploratory, we do not have predefined measures for types of prompts. We did, however, pre-specify that we wanted to differentiate between complement and substitute learning style. One clear, quantifiable, and objective candidate measure for this, which has been used in at least one recent related paper (see Lehmann et al., 2024) is the degree to which subjects use copy-paste to solve practice questions.

We create a copy-paste index constructed using the ChatGPT conversation logs. For each prompt, we detect whether the message contains verbatim text from any of the 92 practice questions used in the study. Matching is based on case-insensitive string overlap between the prompt and a predefined list of phrases and words corresponding to the practice questions, for instance "La knabo _____ Esperanton". We then compute, for each conversation, the share of prompts that contain such matches. This share is mapped onto the five-point copy-paste index as follows:

- 1 - Original: No prompts contain copied practice-question text
- 2 - Mostly Original: Less than 20% of prompts contain copied text
- 3 - Mixed: 20–49% of prompts contain copied text
- 4 - Mostly Copied: 50–79% of prompts contain copied text
- 5 - Verbatim Copy: 80–100% of prompts contain copied text

Higher values therefore indicate a greater student reliance on directly inserting practice questions into the AI system rather than actively learning and formulating questions and prompts on their own.

5.2 Correlation between number of prompts and behavioral outcomes

Figure 8a presents a scatterplot showing a strong and positive relationship between the number of practice questions attempted and the number of prompts. Figure 8b shows no relationship between the scores in the final exam and the number of prompts. The correlation between number of prompts and exam score is informative on whether using AI while studying actually helps students learn as it is measured at a point in time when AI access is removed. The lack of a positive relationship suggests that on average, participants use AI to solve practice questions but that does not translate into learning. In other words, more prompts allow students to attempt more study questions without any real learning taking place.

5.3 Correlation between copy-pasting and behavioral outcomes

One hypothesis for why more prompting does not correlate with learning is that students could just use AI to copy-paste practice questions and get the answers. We explore the correlations between copy-pasting and both attempted practice questions and final exam scores in Figure 9.

We find substantial and significant relationships between copy-pasting, study behavior, and learning outcomes. First, students who do not engage in copy-pasting attempt significantly fewer practice questions than those that engage in some degree of copy-pasting (Figure 9a). This pattern is consistent with the idea that formulating original prompts and engaging more directly with the material is more time-intensive, leaving less time to move through a large number of practice questions. Second, learning outcomes are substantially worse among students who rely most heavily on copy-pasting (Figure 9b). In particular, students in the highest copy-pasting category—those who copy-paste in 80 to 100% of prompts—score 6.7 correct answers on average on the final exam, about 16% below the

control-group mean of 8.0. The boxplot in Figure B.3 underscores how weak performance is in this group: even students at the 75th percentile of its score distribution perform only around the control-group mean. These patterns suggest that students who use AI in a more delegative way tend to learn less than those who use it in a more original and engaged way.

5.4 Do high-GPA women prompt differently?

Finally, we turn to the subgroup that appears to benefit most from AI-guided access in our study: high-GPA women. As discussed above, this group is more likely to appear in the top of the exam-score distribution under the AI-guided treatment and is also the only subgroup for which we see a lower decline in the number of attempted practice questions relative to the control group. We next examine their prompting behavior to better understand why they may be more likely to benefit from AI.

Figure 10 plots the share of students who primarily delegate effort to copy-paste—this is defined as having at least 50% of prompts categorized as copy-pasted. High-GPA women exhibit the lowest rate of high copy-pasting of any subgroup, at about 15%. This share is substantially and significantly lower than the corresponding share among high-GPA men, 36.5%. Thus, the subgroup that appears to benefit most from AI-guided access is also the subgroup least likely to rely on a highly delegative prompting strategy.

One possible interpretation, which remains speculative, is that women—and especially high-GPA women—may be more likely to adhere to guidance about appropriate AI use. We cannot test this mechanism directly in our experiment. However, the pattern is consistent with the design of the AI-guided treatment, in which students were explicitly told not to ask for direct answers and instead to use AI to obtain explanations, ask follow-up questions, and receive feedback. In fact, we see that students in AI-guided are less likely to copy-paste than those in AI-assisted (Figure B.4).[10] It is also consistent with evidence from Carvajal et al. (2024), where women, and particularly top-performing women, are more responsive than men to whether instructors allow or forbid AI use. This may help explain why high-GPA

[10]Figure B.4 also reveals additional patterns in copy-pasting behavior. Younger students and undergraduates are less likely to rely on copy-pasting, whereas students who use AI more intensively in their daily lives and those who have a paid AI subscription are more likely to do so.

women in our setting are less likely to engage in prompting that delegates work to AI.

6 Discussion

Our findings provide new evidence on the heterogeneous effects of AI on learning. The subgroup that appears to benefit most from AI-guided access is high-GPA women. They are more likely to reach the top of the exam-score distribution and, unlike other groups, do not show as large an average decline in the number of attempted practice questions relative to the control group. Our prompting analysis suggests one possible reason: high-GPA women are the subgroup least likely to rely on copy-pasting during the study phase. These patterns suggest that they use AI in a way that is less substitutive of their own effort and therefore more conducive to learning. This is particularly striking in light of prior evidence that high-ability women have relatively lower AI adoption rates (Carvajal et al., 2024).

An important finding is that studying with AI appears to take time away from other activities. Students in the AI treatments attempt significantly fewer practice questions during the fixed study period, indicating that time spent interacting with AI crowds out other uses of time. This point is central for thinking about AI in education. In our setting, this crowding out does not reduce exam performance on average, but in real educational environments the displaced activity could matter greatly. If AI crowds out study time in other school subjects, the broader effect on learning could be negative even when performance in the focal subject does not decline. If instead it crowds out leisure, then AI could impact students' overall welfare. More generally, our results suggest that the educational impact of AI cannot be understood only through test scores. It also depends on what activities AI replaces.

Our findings also raise broader concerns about inequality in education. If students who are relatively advantaged to begin with are better able than others to use AI productively, AI may reinforce existing gaps in academic achievement rather than reduce them. This possibility is especially relevant in light of the so-called "boy crisis" in education, which refers to the growing evidence that boys are falling behind girls on a range of educational outcomes with important later-life consequences (Autor and Wasserman, 2013; Autor et al., 2023). In

most OECD countries, boys are more likely than girls to fall below baseline proficiency in core subjects such as mathematics, reading, and science, and they also drop out of school at higher rates (OECD, 2015). Our results suggest that the spread of AI in educational settings could interact with these pre-existing inequalities.

At the same time, several limitations of this study should be kept in mind. First, our study measures short-run learning in a controlled laboratory setting and does not speak directly to long-run retention. In particular, we do not observe whether students would still remember the Esperanto material some time after the experiment, an important dimension of learning that logistical constraints prevented us from measuring through a later follow-up session. This limitation is important because human capital accumulates over repeated learning experiences. Precisely for that reason, the short-run losses we observe for some students are concerning: if substituting one's own effort with AI already reduces learning in a single session, repeated exposure to this pattern could plausibly generate meaningful longer-run losses in skill accumulation. Second, the laboratory setting abstracts from many features of real educational environments, including repeated exposure to AI over time, teacher responses, peer interactions, and variation in assessment design. Third, while our prompting analysis provides useful insight into how students use AI, it remains exploratory and does not identify the causal effect of specific prompting strategies. These limitations notwithstanding, the experiment provides clean evidence that AI changes study behavior even in the short run, and that its effects on learning crucially depend on who uses it and how.

References

Autor, David and Melanie Wasserman, "Wayward sons: The emerging gender gap in labor markets and education," *Third Way Report*, 2013, *20013*.

_ , David Figlio, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman, "Males at the tails: How socioeconomic status shapes the gender gap," *The Economic Journal*, 2023, *133* (656), 3136–3152.

Bastani, Hamsa, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman, "Generative AI without guardrails can harm learning: Evidence from high school mathematics," *Proceedings of the National Academy of Sciences*, 2025, *122* (26), e2422633122.

Bengio, Yoshua, Stephen Clare, Carina Prunkl, Maksym Andriushchenko, Ben Bucknall, Malcolm Murray, Rishi Bommasani, Stephen Casper, Tom Davidson, Raymond Douglas et al., "International ai safety report 2026," *arXiv preprint arXiv:2602.21012*, 2026.

Carvajal, Daniel, Catalina Franco, and Siri Isaksson, "Will Artificial Intelligence Get in the Way of Achieving Gender Equality?," *NHH Dept. of Economics Discussion Paper*, 2024, (03).

Dai, Xusheng, Zhaochun Wen, Jianxiao Jiang, Huiqin Liu, and Yu Zhang, "How students use AI feedback matters: Experimental evidence on physics achievement and autonomy," *arXiv preprint arXiv:2505.08672*, 2025.

DeKeyser, Robert, "Implicit and explicit learning," *The handbook of second language acquisition*, 2003, pp. 312–348.

Han, Xiaoli, Hongchao Peng, and Mingzhuo Liu, "The impact of GenAI on learning outcomes: A systematic review and meta-analysis of experimental studies," *Educational Research Review*, 2025, p. 100714.

Humlum, Anders and Emilie Vestergaard, "The Adoption of ChatGPT," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, 2024, (2024-50).

Kestin, Greg, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti, "AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting," *Scientific Reports*, 2025, *15* (1), 17458.

Koutny, Ilona, "A typological description of Esperanto as a natural language," *Język. Komunikacja. Informacja*, 2015, *10*, 43–62.

Lehmann, Matthias, Philipp B Cornelius, and Fabian J Sting, "AI Meets the Classroom: When Does ChatGPT Harm Learning?," *arXiv preprint arXiv:2409.09047*, 2024.

Mishra, Punya, Danah Henriksen, Lauren J Woo, and Nicole Oster, "Control vs. agency: Exploring the history of AI in education," *TechTrends*, 2025, *69* (2), 247–253.

OECD, "The ABC of Gender Equality in Education. Aptitude, Behavior, Confidence," 2015.

Pallant, Jessica L, Janneke Blijlevens, Alexander Campbell, and Ryan Jopp, "Mastering knowledge: The impact of generative AI on student learning outcomes," *Studies in Higher Education*, 2025, pp. 1–22.

Pardos, Zachary A and Shreya Bhandari, "ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills," *Plos one*, 2024, *19* (5), e0304013.

Tan, Steffi and Vaikunthan Rajaratnam, "Critique of Generative AI Can Harm Learning Study Design," *Available at SSRN 4898213*, 2024.

Voyer, Daniel and Susan D Voyer, "Gender differences in scholastic achievement: a meta-analysis.," *Psychological bulletin*, 2014, *140* (4), 1174.

Wecks, Janik Ole, Johannes Voshaar, Benedikt Jost Plate, and Jochen Zimmermann, "Generative AI usage and exam performance," *arXiv preprint arXiv:2404.19699*, 2024.

AI statement

In preparing this manuscript, the authors used ChatGPT to check spelling, grammar, and wording. All content was subsequently reviewed and edited by the authors, who take full responsibility for the final publication.

7 Tables

Table 1: Treatment balance

	(1) Control (T1)	(2) AI-assisted (T2)	(3) AI-guided (T3)	(4) Diff. (T1)-(T2)	(5) Diff. (T1)-(T3)
Female	0.618	0.613	0.602	0.004	0.016
	(0.487)	(0.488)	(0.491)	(0.049)	(0.049)
21 years old or younger	0.584	0.520	0.613	0.063	-0.029
	(0.494)	(0.501)	(0.488)	(0.049)	(0.049)
Undergraduate student	0.641	0.602	0.683	0.039	-0.042
	(0.481)	(0.491)	(0.466)	(0.048)	(0.047)
Arts	0.124	0.112	0.116	0.012	0.009
	(0.331)	(0.316)	(0.321)	(0.032)	(0.032)
Economics or Business School	0.201	0.224	0.241	-0.024	-0.040
	(0.402)	(0.418)	(0.429)	(0.041)	(0.041)
Engineering	0.163	0.163	0.131	-0.001	0.032
	(0.370)	(0.371)	(0.338)	(0.037)	(0.035)
Medicine and Health Sciences	0.163	0.153	0.146	0.010	0.017
	(0.370)	(0.361)	(0.354)	(0.036)	(0.036)
Social Sciences	0.153	0.199	0.176	-0.046	-0.023
	(0.361)	(0.400)	(0.382)	(0.038)	(0.037)
Other	0.196	0.148	0.191	0.048	0.005
	(0.398)	(0.356)	(0.394)	(0.037)	(0.039)
High GPA	0.368	0.429	0.226	-0.060	0.142***
	(0.484)	(0.496)	(0.419)	(0.049)	(0.045)
Use AI occasionally or all the time	0.627	0.612	0.648	0.015	-0.021
	(0.485)	(0.488)	(0.479)	(0.048)	(0.048)
Paid AI subscription	0.057	0.056	0.090	0.001	-0.033
	(0.233)	(0.231)	(0.288)	(0.023)	(0.026)
Clicked outside of the test page	0.014	0.066	0.030	-0.052***	-0.016
	(0.119)	(0.249)	(0.171)	(0.020)	(0.015)
Observations	209	196	199	405	408

Notes. This table reports means and standard deviations (in parentheses) for the covariates listed in the left-hand column for the control group (Column 1), the AI-assisted treatment (Column 2), and the AI-guided treatment (Column 3). Columns 4 and 5 report the mean differences between each treatment group and the control group, with standard errors in parentheses. The treatment was assigned within each laboratory section by designating a third of the available computers to each treatment condition and randomly allocating participants to computers. A small fraction of students left the survey page during the final exam even though they were instructed not to do so (last row in the table). They were not allowed to finish the survey and are thus considered as attritors. Table A1 shows that balance remains when excluding attritors. Statistical significance of the differences in Columns 4 and 5 is indicated by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Treatment effects on exam performance and self-assessed learning

	Test score				Self-assesment of learning			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI-assisted	0.078	-0.224	-0.046	-0.464	0.004	-0.049	-0.014	-0.038
	(0.192)	(0.295)	(0.236)	(0.383)	(0.040)	(0.066)	(0.052)	(0.088)
AI-guided	-0.234	-0.211	-0.318	-0.066	-0.041	-0.069	-0.068	-0.119
	(0.189)	(0.293)	(0.217)	(0.332)	(0.041)	(0.067)	(0.048)	(0.083)
Female		-0.047		0.110		-0.025		-0.001
		(0.262)		(0.287)		(0.056)		(0.069)
AI-assisted × Female		0.498		0.667		0.087		0.039
		(0.386)		(0.485)		(0.083)		(0.108)
AI-guided × Female		-0.039		-0.403		0.045		0.081
		(0.390)		(0.440)		(0.085)		(0.102)
High GPA			0.006	0.288			-0.049	-0.012
			(0.290)	(0.441)			(0.059)	(0.093)
AI-assisted × High GPA			0.310	0.471			0.047	-0.021
			(0.397)	(0.616)			(0.083)	(0.135)
AI-guided × High GPA			0.258	-0.554			0.095	0.206
			(0.437)	(0.651)			(0.094)	(0.135)
Female × High GPA				-0.452				-0.059
				(0.593)				(0.120)
AI-assisted × Female × High GPA				-0.269				0.116
				(0.815)				(0.170)
AI-guided × Female × High GPA				1.362				-0.199
				(0.878)				(0.187)
Constant	7.448***	7.546***	7.514***	7.539***	0.828***	0.854***	0.843***	0.857***
	(0.311)	(0.322)	(0.311)	(0.318)	(0.067)	(0.069)	(0.068)	(0.075)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean women	7.968	7.968	7.968	7.968	0.808	0.808	0.808	0.808
Mean men	8.184	8.184	8.184	8.184	0.829	0.829	0.829	0.829
Observations	572	572	572	572	572	572	572	572

Notes. This table reports estimates from Equation 3 (pre-registered) for two primary outcomes. Exam score is measured on a 0–15 scale. Self-assessed learning is an indicator equal to one if the respondent selected "somewhat agree" or "agree" to the statement, "I feel like I learned a lot working through the questions," and zero otherwise ("disagree" or "somewhat disagree"). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Treatment effects on scoring at the top or bottom

	Top score				Low score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI-assisted	0.009	0.016	0.034	0.057	-0.003	0.040	0.029	0.073*
	(0.028)	(0.041)	(0.035)	(0.048)	(0.020)	(0.029)	(0.024)	(0.043)
AI-guided	-0.002	0.026	0.010	0.078*	0.009	0.040	0.046**	0.073**
	(0.026)	(0.038)	(0.029)	(0.043)	(0.020)	(0.030)	(0.023)	(0.036)
Female		0.043		0.065*		0.029		0.015
		(0.035)		(0.038)		(0.025)		(0.020)
AI-assisted × Female		-0.011		-0.038		-0.070*		-0.071
		(0.056)		(0.068)		(0.039)		(0.051)
AI-guided × Female		-0.047		-0.109*		-0.050		-0.044
		(0.053)		(0.058)		(0.043)		(0.048)
High GPA			0.057	0.097			0.056*	0.031
			(0.042)	(0.062)			(0.032)	(0.035)
AI-assisted × High GPA			-0.061	-0.104			-0.081*	-0.078
			(0.059)	(0.090)			(0.043)	(0.063)
AI-guided × High GPA			-0.036	-0.175**			-0.123***	-0.117**
			(0.062)	(0.070)			(0.039)	(0.052)
Female × High GPA				-0.065				0.038
				(0.083)				(0.060)
AI-assisted × Female × High GPA				0.069				-0.006
				(0.121)				(0.086)
AI-guided × Female × High GPA				0.236**				-0.007
				(0.116)				(0.076)
Constant	0.022	0.012	0.010	-0.015	0.103***	0.079**	0.081**	0.064**
	(0.042)	(0.042)	(0.043)	(0.042)	(0.038)	(0.034)	(0.036)	(0.031)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean women	0.088	0.088	0.088	0.088	0.056	0.056	0.056	0.056
Mean men	0.053	0.053	0.053	0.053	0.013	0.013	0.013	0.013
Observations	572	572	572	572	572	572	572	572

Notes. This table reports estimates from the pre-registered specification in Equation 3 for two binary outcomes. The thresholds for these outcomes were pre-specified as follows: Top scorer equals one if the participant answered more than 10 questions correctly, and low scorer equals one if the participant answered fewer than 5 questions correctly. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

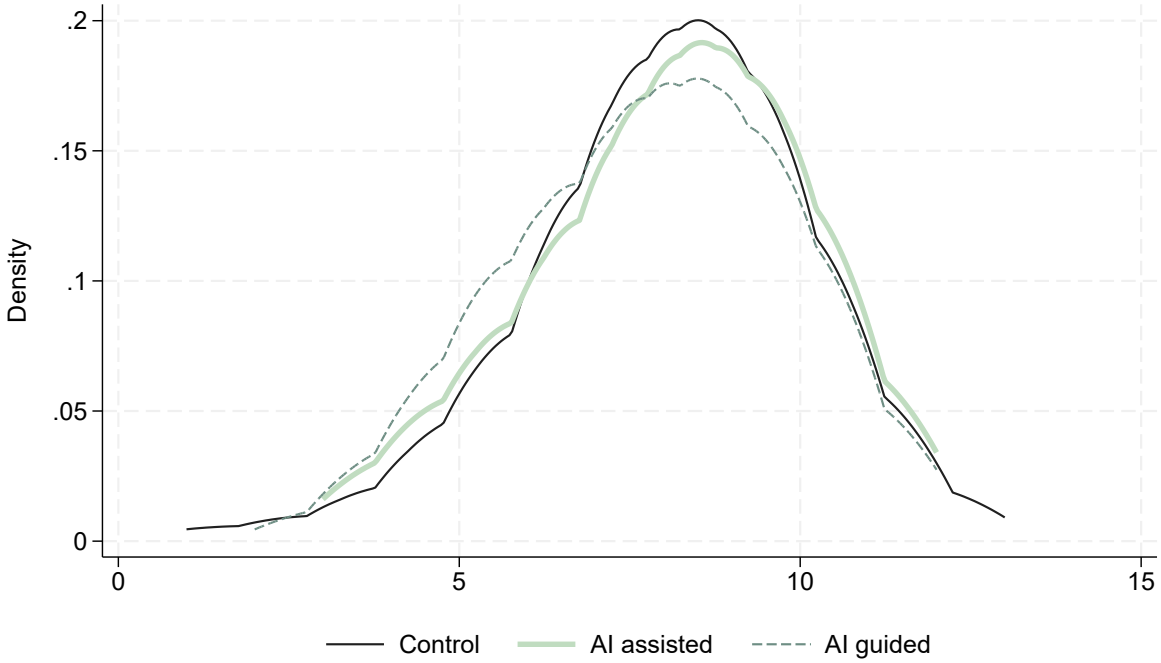
Table 4: Treatment effects on number of practice questions attempted

	Time spent on test (minutes)				Number of practice questions attempted			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI-assisted	-0.132	0.162	-0.056	0.346	-4.171***	-6.255***	-4.337***	-5.022**
	(0.248)	(0.420)	(0.315)	(0.550)	(1.094)	(1.710)	(1.351)	(2.121)
AI-guided	0.149	0.225	0.090	0.426	-5.773***	-6.412***	-6.166***	-5.243***
	(0.243)	(0.371)	(0.290)	(0.441)	(1.125)	(1.751)	(1.305)	(1.852)
Female		0.479		0.904**		-1.865		0.085
		(0.377)		(0.455)		(1.538)		(1.691)
AI-assisted × Female		-0.484		-0.649		3.379		1.037
		(0.518)		(0.668)		(2.227)		(2.760)
AI-guided × Female		-0.124		-0.540		1.048		-1.469
		(0.494)		(0.583)		(2.333)		(2.580)
High GPA			0.279	1.004*			1.107	4.415
			(0.385)	(0.609)			(1.695)	(2.884)
AI-assisted × High GPA			-0.174	-0.560			0.460	-3.237
			(0.518)	(0.865)			(2.320)	(3.724)
AI-guided × High GPA			0.266	-0.435			1.396	-3.327
			(0.541)	(0.815)			(2.567)	(3.998)
Female × High GPA				-1.169				-5.357
				(0.783)				(3.484)
AI-assisted × Female × High GPA				0.555				6.197
				(1.072)				(4.731)
AI-guided × Female × High GPA				1.129				7.767
				(1.092)				(5.175)
Constant	10.112***	9.989***	10.109***	9.758***	30.538***	31.447***	30.718***	30.525***
	(0.435)	(0.472)	(0.446)	(0.510)	(1.991)	(2.025)	(1.946)	(2.072)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean women	10.819	10.819	10.819	10.819	33.857	33.857	33.857	33.857
Mean men	10.447	10.447	10.447	10.447	35.628	35.628	35.628	35.628
Observations	572	572	572	572	594	594	594	594

Notes. This table presents results from the pre-registered specification in Equation 3 for two outcomes. Time spent is measured as the total minutes spent on the exam. Practice questions solved is the total number of practice questions attempted. There were 90 practice questions in total. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

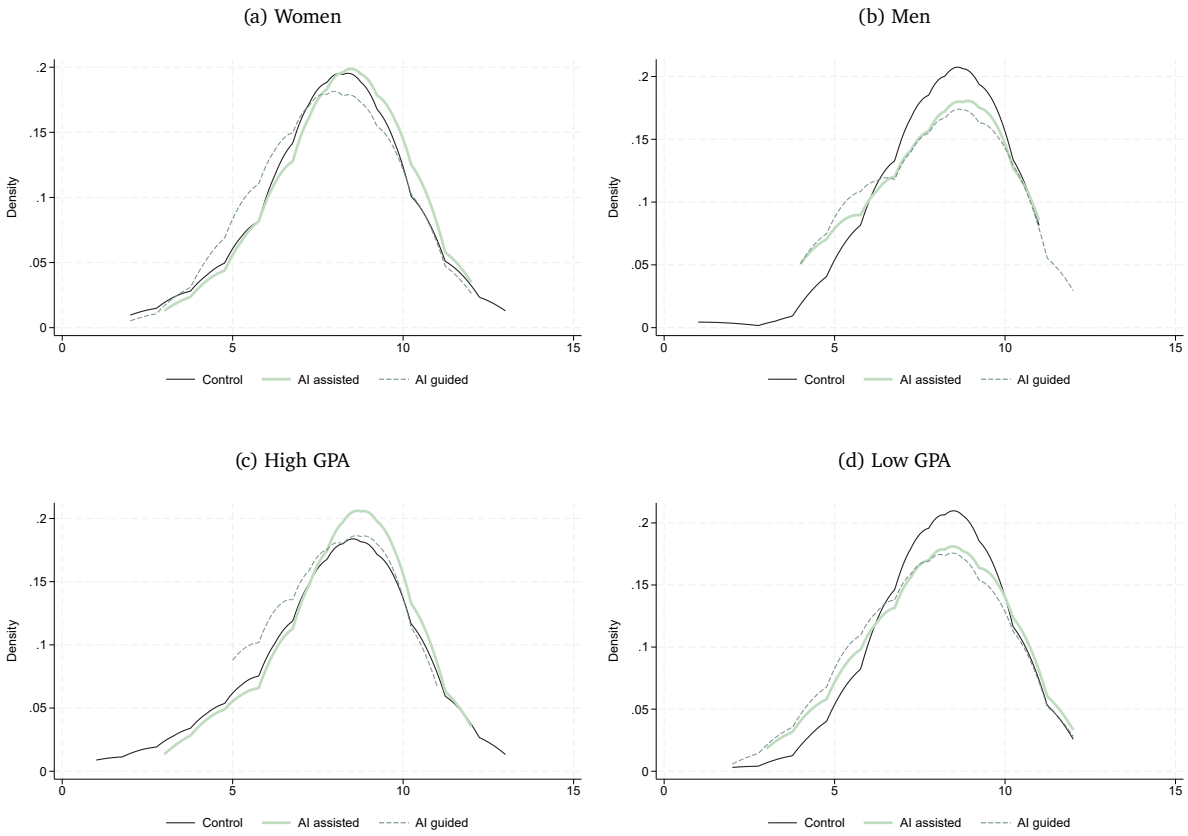
8 Figures

Figure 1: Density of exam performance



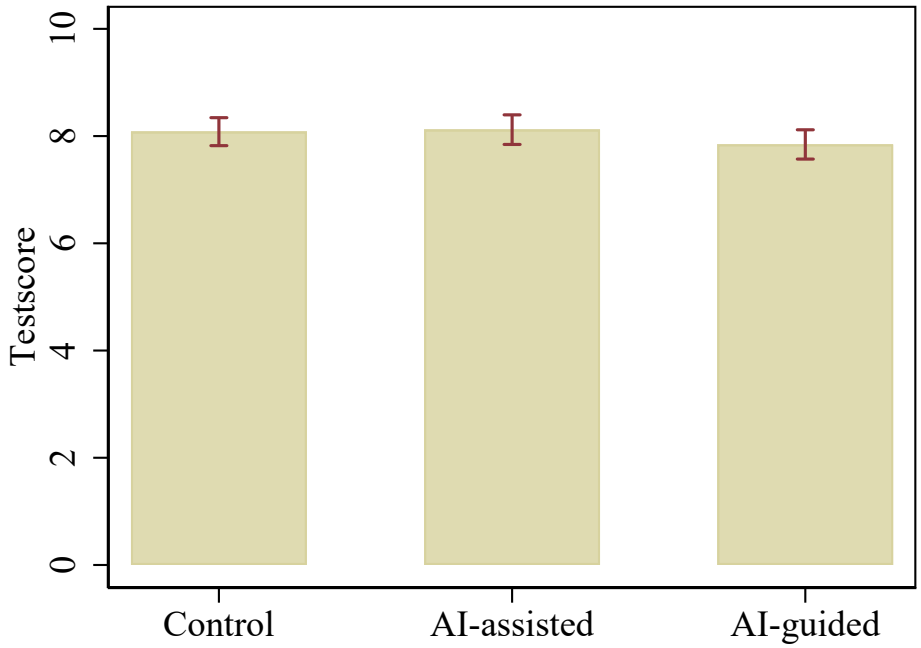
Notes. This figure presents kernel density estimates using the Epanechnikov kernel function with the width of the density window around each point equal to 1, separately for each treatment arm.

Figure 2: Density of exam performance by gender and GPA



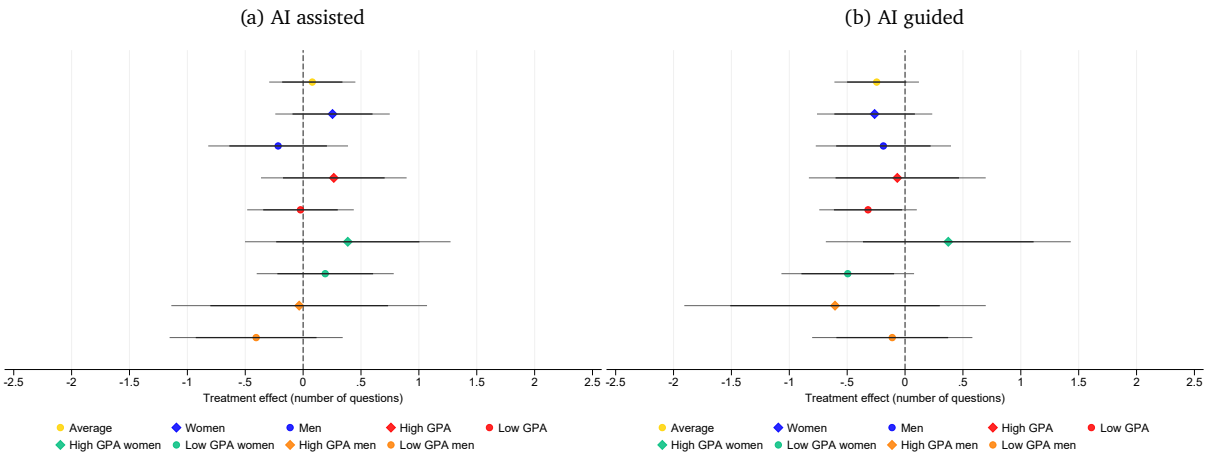
Notes. These figure presents kernel density estimates using the Epanechnikov kernel function with the width of the density window around each point equal to 1, separately for each treatment arm. The panels define the sample split on which the estimates are generated: gender and GPA.

Figure 3: Exam performance by treatment



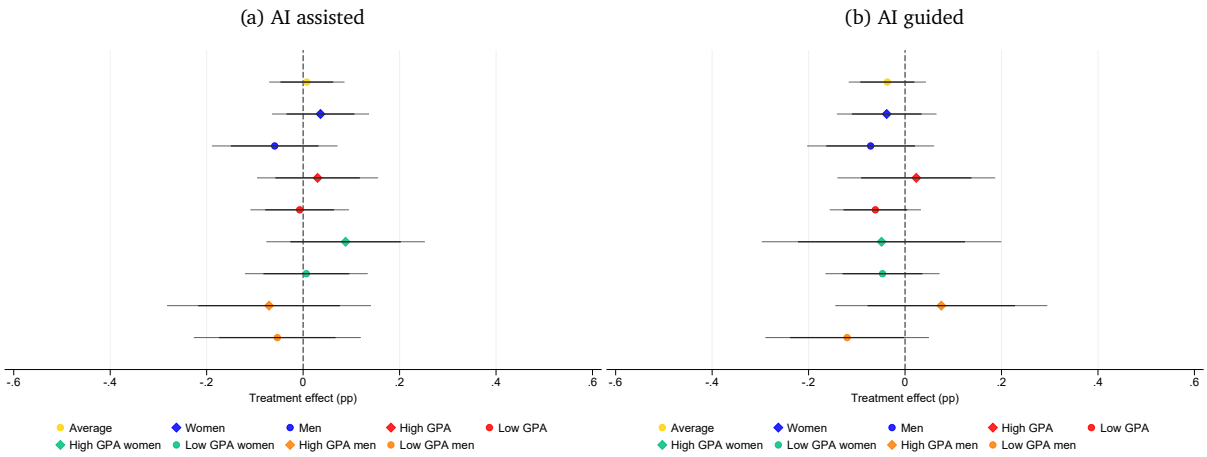
Notes. Each bar plots the raw mean score in the Esperanto exam for students each treatment condition, along with 95% confidence bars.

Figure 4: Treatment effects on exam performance by subgroups



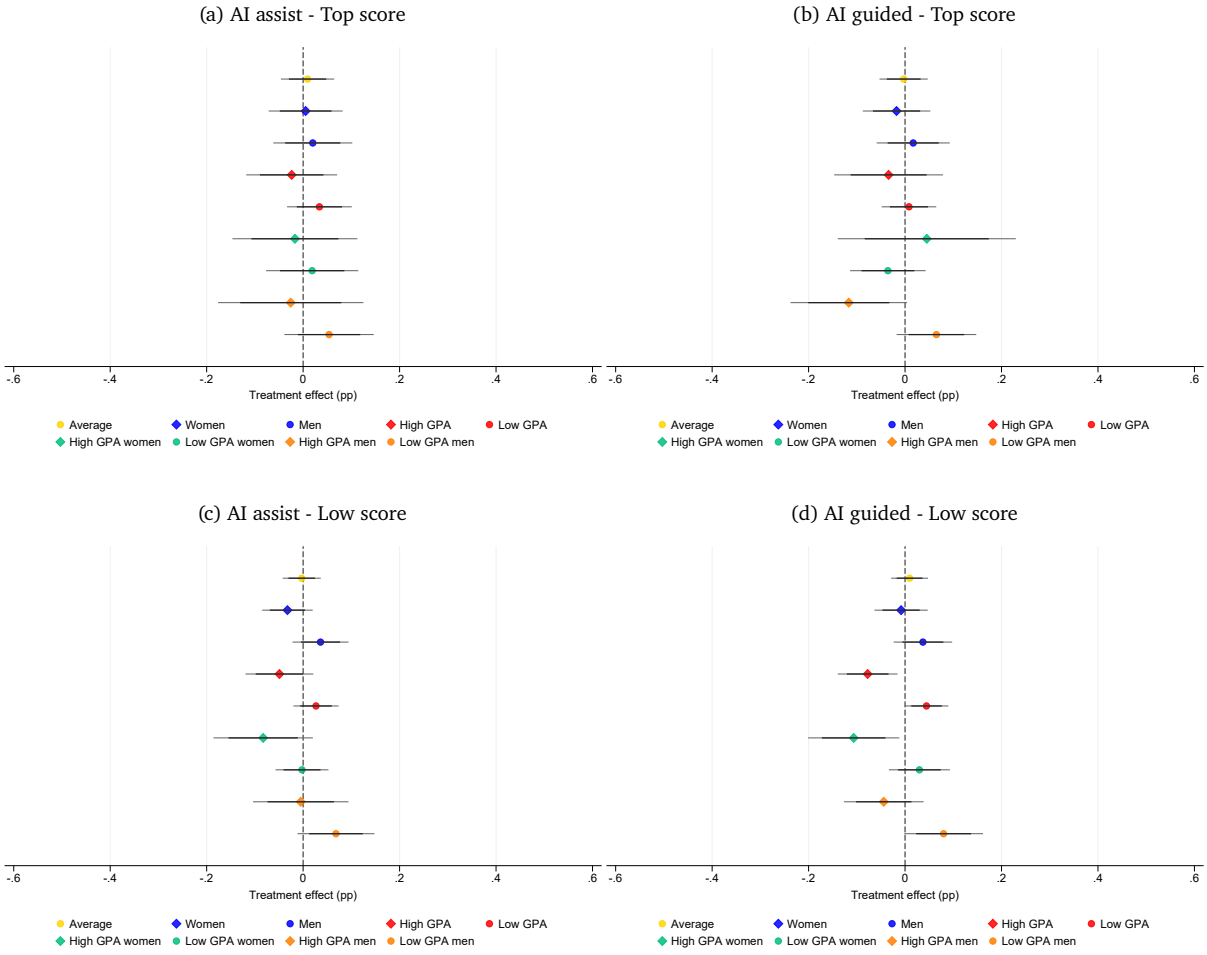
Notes. The plots report point estimates along with 95% confidence intervals of treatment effects on exam performance for the full sample (top coefficient) and for the subsamples listed below. 83% confidence intervals are added to facilitate visual comparisons of differences across subgroups. Subsample estimates are obtained by comparing treated students within a given subgroup to control-group students in the same subgroup. Panel (a) reports the effect of assignment to the AI-assisted treatment relative to the control group, and Panel (b) reports the effect of assignment to the AI-guided treatment relative to the control group. All regressions from with the coefficients are obtained control for: Age 21 or younger, undergraduate student, indicators for field of study, use AI occasionally or all the time, has a paid subscription to an AI provider. In regressions without interactions with gender and high GPA, we add these variables as covariates.

Figure 5: Treatment effects on self-assessed learning by subgroups



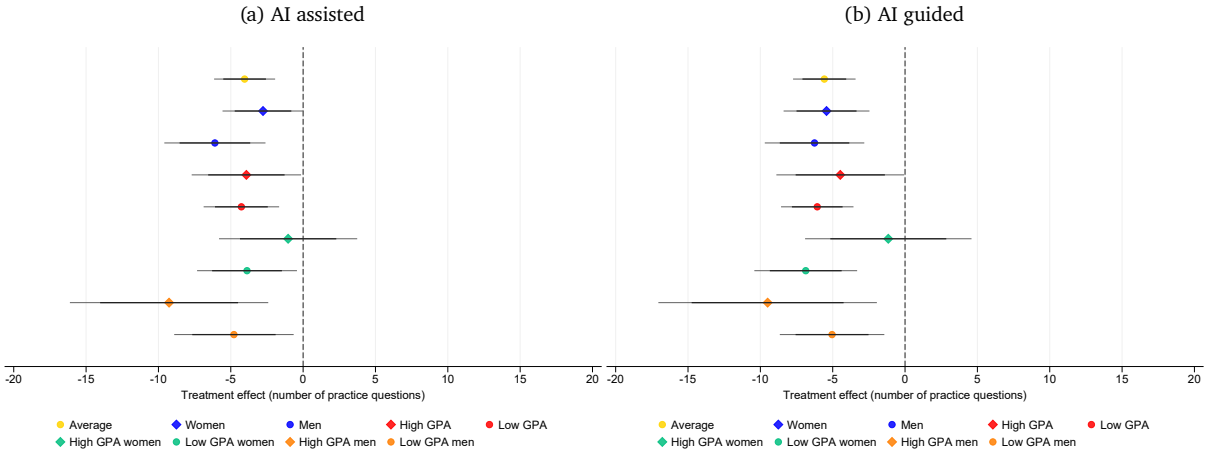
Notes. The plots report point estimates along with 95% confidence intervals of treatment effects on self assessed learning for the full sample (top coefficient) and for the subsamples listed below. 83% confidence intervals are added to facilitate visual comparisons of differences across subgroups. Subsample estimates are obtained by comparing treated students within a given subgroup to control-group students in the same subgroup. Panel (a) reports the effect of assignment to the AI-assisted treatment relative to the control group, and Panel (b) reports the effect of assignment to the AI-guided treatment relative to the control group. All regressions from with the coefficients are obtained control for: Age 21 or younger, undergraduate student, indicators for field of study, use AI occasionally or all the time, has a paid subscription to an AI provider. In regressions without interactions with gender and high GPA, we add these variables as covariates.

Figure 6: Treatment effects on scoring at the top or bottom



Notes. The plots report point estimates along with 95% confidence intervals of treatment effects on obtaining a top or low exam performance for the full sample (top coefficient) and for the subsamples listed below. 83% confidence intervals are added to facilitate visual comparisons of differences across subgroups. Subsample estimates are obtained by comparing treated students within a given subgroup to control-group students in the same subgroup. Panels (a) and (b) report the effect on obtaining a score higher than 10 in the exam, and Panels (c) and (d) report the effect on obtaining a score lower than 5. Both score thresholds were pre-registered. All regressions from with the coefficients are obtained control for: Age 21 or younger, undergraduate student, indicators for field of study, use AI occasionally or all the time, has a paid subscription to an AI provider. In regressions without interactions with gender and high GPA, we add these variables as covariates.

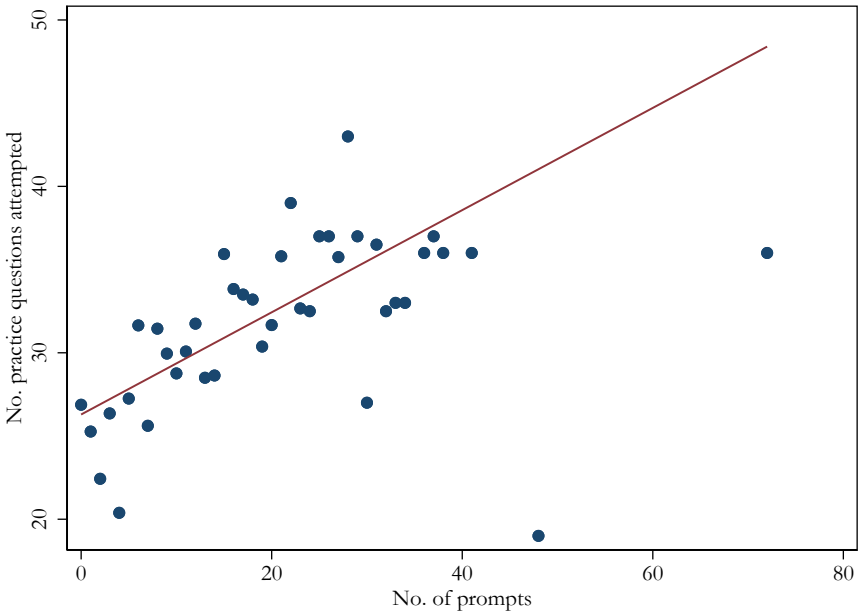
Figure 7: Treatment effects on number of practice questions attempted



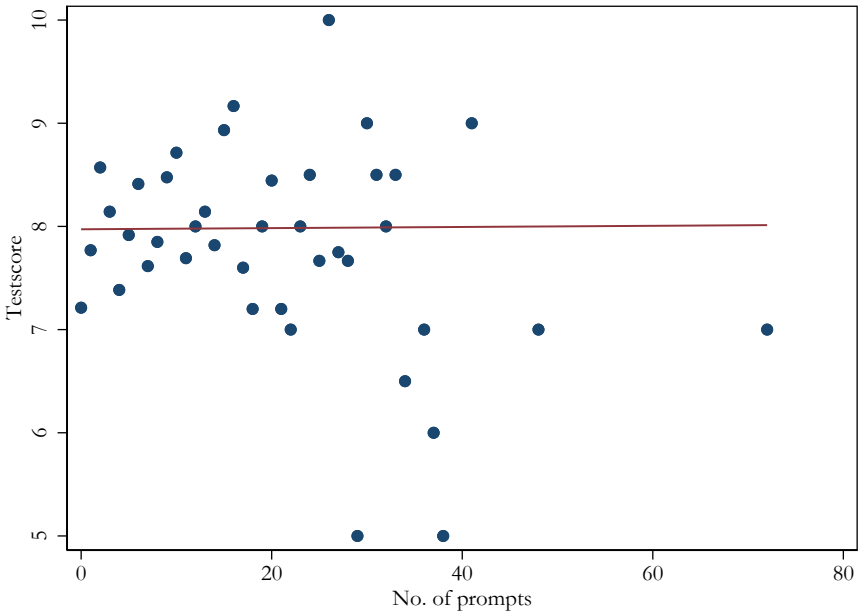
Notes. The plots report point estimates along with 95% confidence intervals of treatment effects on number of practice questions attempted for the full sample (top coefficient) and for the subsamples listed below. 83% confidence intervals are added to facilitate visual comparisons of differences across subgroups. Subsample estimates are obtained by comparing treated students within a given subgroup to control-group students in the same subgroup. Panel (a) reports the effect of assignment to the AI-assisted treatment relative to the control group, and Panel (b) reports the effect of assignment to the AI-guided treatment relative to the control group. There were 90 practice questions in total. All regressions from with the coefficients are obtained control for: Age 21 or younger, undergraduate student, indicators for field of study, use AI occasionally or all the time, has a paid subscription to an AI provider. In regressions without interactions with gender and high GPA, we add these variables as covariates.

Figure 8: Correlation between prompting behavior, and study behavior and exam performance

(a) Practice questions attempted



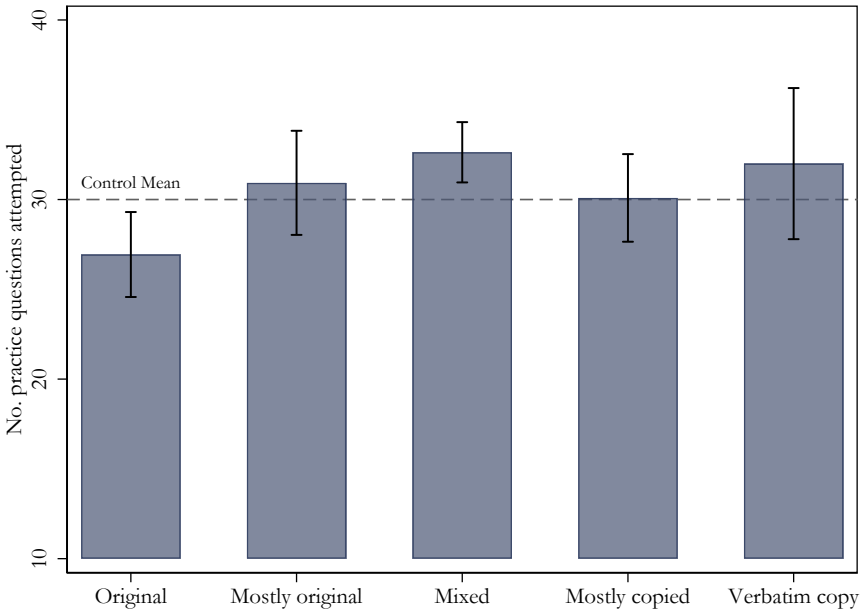
(b) Exam performance



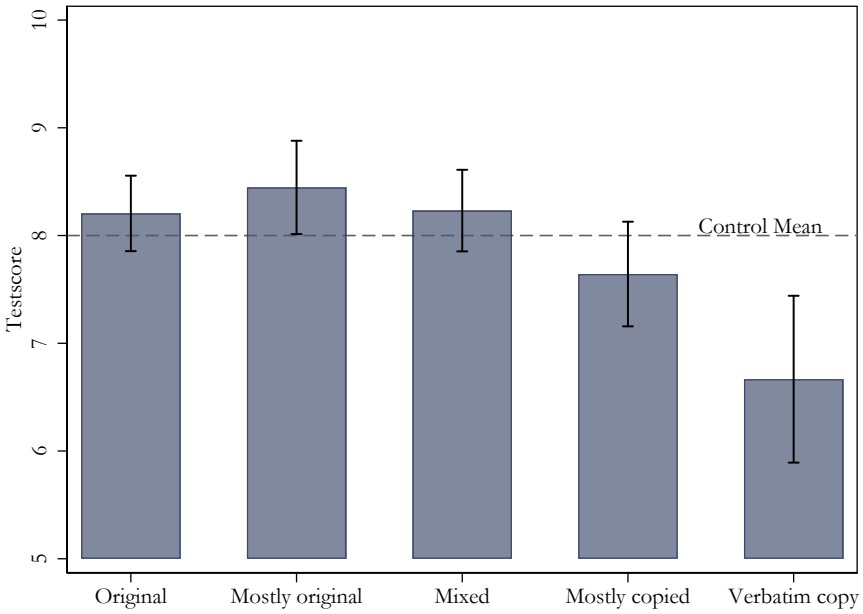
Notes. The figures report scatterplots correlating the number of prompts each student submitted with the number of practice questions they attempted (Panel (a)) and their score in the final Esperanto exam (Panel (b)). In total, 395 students were given AI access and almost 90% of them wrote at least one prompt.

Figure 9: Correlation between copy-pasting, and study behavior and exam performance

(a) Practice questions attempted

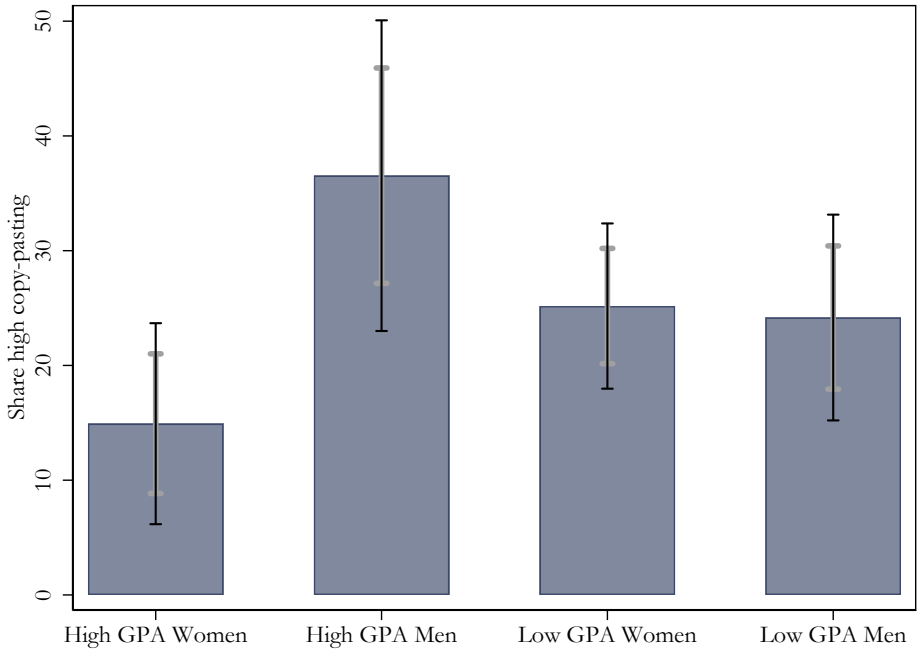


(b) Exam performance



Notes. The plots report means and confidence intervals for the outcomes in the panel headings on a categorization of copy-pasting behavior observed in the prompts. The copy-pasting categories indicate the following prompting behavior: Original = no copied practice-question text, Mostly original = less than 20% of prompts copy-pasted, Mixed = 20% to less than 50% copy-pasted, Mostly copied = 50% to less than 80% copy-pasted, Varbatim copy = 80% or more copy-pasted. The horizontal dotted lines represent the average value for the outcome in the control group.

Figure 10: Copy-pasting by subgroups



Notes. The plot reports the means and 95% confidence bars for copy-pasting behavior by subgroups. High copy-pasting is a dummy variable with 1 indicating 50% or more of prompts of the participant being copy-pasted, and 0 indicating less than 50% copy-pasted. The means illustrate the share of participants who are of the high-copy-paste type among all participants in their subgroup. 83% confidence bars (in grey) are added to facilitate visual comparisons of differences across subgroups.

A Appendix Tables

Table A1: Treatment balance with final analytical sample

	(1) Control (T1)	(2) AI-assisted (T2)	(3) AI-guided (T3)	(4) Diff. (T1)-(T2)	(5) Diff. (T1)-(T3)
Female	0.622	0.597	0.605	0.025	0.017
	(0.486)	(0.492)	(0.490)	(0.050)	(0.049)
21 years old or younger	0.577	0.541	0.616	0.036	-0.039
	(0.495)	(0.500)	(0.488)	(0.051)	(0.050)
Undergraduate student	0.637	0.619	0.684	0.018	-0.047
	(0.482)	(0.487)	(0.466)	(0.050)	(0.048)
Arts	0.119	0.110	0.121	0.009	-0.002
	(0.325)	(0.314)	(0.327)	(0.033)	(0.033)
Economics or Business School	0.204	0.210	0.237	-0.006	-0.033
	(0.404)	(0.408)	(0.426)	(0.042)	(0.042)
Engineering	0.164	0.166	0.137	-0.002	0.027
	(0.371)	(0.373)	(0.345)	(0.038)	(0.036)
Medicine and Health Sciences	0.164	0.160	0.153	0.004	0.012
	(0.371)	(0.368)	(0.361)	(0.038)	(0.037)
Social Sciences	0.149	0.193	0.163	-0.044	-0.014
	(0.357)	(0.396)	(0.370)	(0.039)	(0.037)
Other	0.199	0.160	0.189	0.039	0.010
	(0.400)	(0.368)	(0.393)	(0.039)	(0.040)
High GPA	0.373	0.436	0.216	-0.063	0.157***
	(0.485)	(0.497)	(0.412)	(0.050)	(0.045)
Use AI occasionally or all the time	0.642	0.619	0.647	0.023	-0.006
	(0.481)	(0.487)	(0.479)	(0.050)	(0.049)
Paid AI subscription	0.060	0.050	0.095	0.010	-0.035
	(0.238)	(0.218)	(0.294)	(0.023)	(0.027)
Observations	201	181	190	382	391

Notes. This table reports means and standard deviations (in parentheses) for the covariates listed in the left-hand column for the control group (Column 1), the AI-assisted treatment (Column 2), and the AI-guided treatment (Column 3). Columns 4 and 5 report the mean differences between each treatment group and the control group, with standard errors in parentheses. The treatment was assigned within each laboratory section by designating a third of the available computers to each treatment condition and randomly allocating participants to computers. A small fraction of students left the survey page during the final exam even though they were instructed not to do so. They were not allowed to finish the survey and are thus considered as attritors. This table reports the balance of covariates using the final analytical sample after excluding attritors (see Table 1). Statistical significance of the differences in Columns 4 and 5 is indicated by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

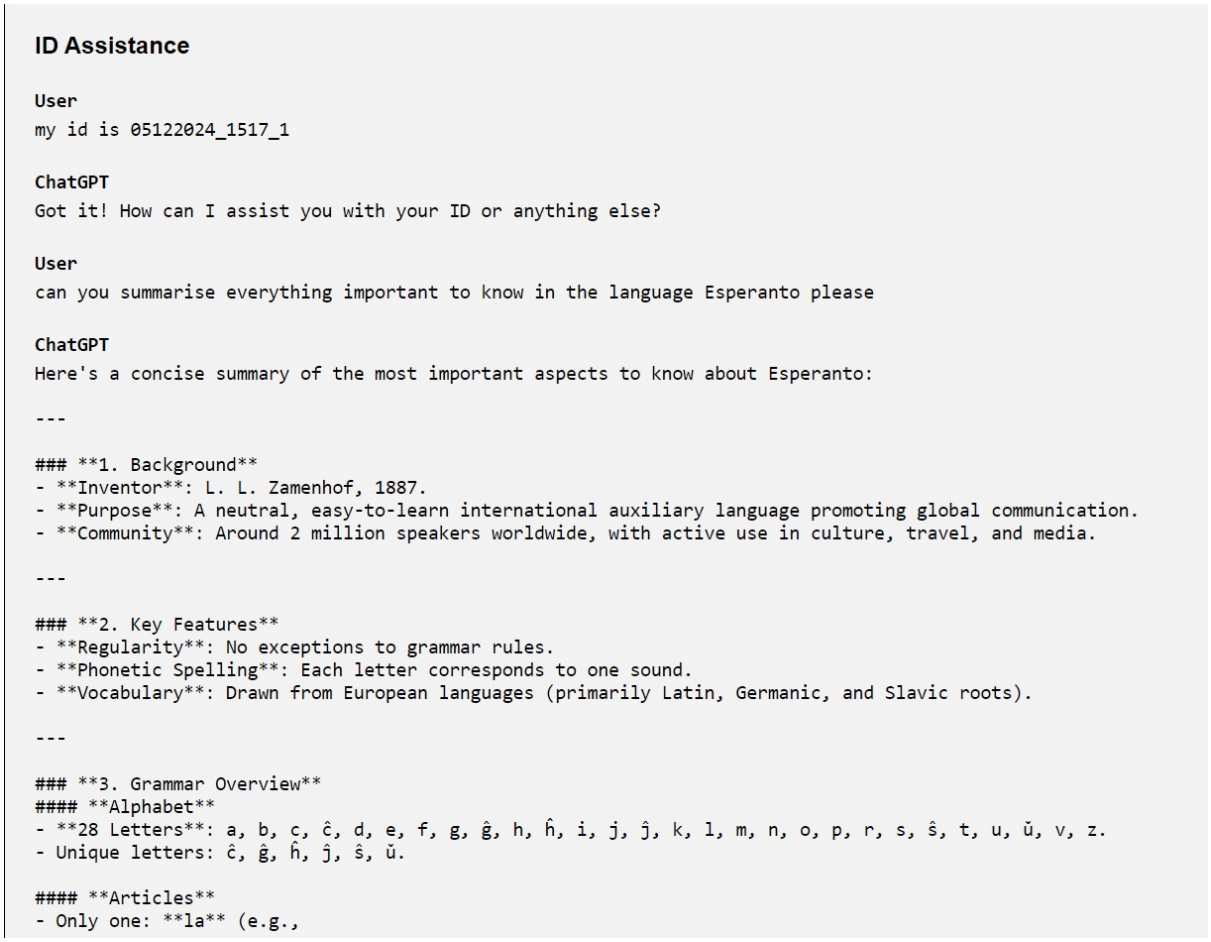
Table A2: Share of correctly solved practice questions

	Success rate in attempted practice questions			
	(1)	(2)	(3)	(4)
AI-assisted	0.055***	0.016	0.038*	-0.004
	(0.017)	(0.028)	(0.022)	(0.035)
AI-guided	0.062***	0.027	0.056***	0.021
	(0.017)	(0.026)	(0.020)	(0.031)
Female		-0.051**		-0.048*
		(0.023)		(0.029)
AI-assisted × Female		0.063*		0.064
		(0.035)		(0.045)
AI-guided × Female		0.057		0.056
		(0.035)		(0.041)
High GPA			0.008	0.008
			(0.023)	(0.037)
AI-assisted × High GPA			0.042	0.044
			(0.034)	(0.056)
AI-guided × High GPA			0.012	0.015
			(0.036)	(0.057)
Female × High GPA				-0.006
				(0.047)
AI-assisted × Female × High GPA				0.002
				(0.070)
AI-guided × Female × High GPA				-0.000
				(0.073)
Constant	0.685***	0.712***	0.692***	0.719***
	(0.028)	(0.030)	(0.028)	(0.031)
Controls	Yes	Yes	Yes	Yes
Mean women	0.678	0.678	0.678	0.678
Mean men	0.736	0.736	0.736	0.736
Observations	595	585	595	585

Notes: The outcomes are at the top of the columns. Estimates based on the specifications and heterogeneity reported in our pre-analysis plan. $* \ p < 0.10$, $** \ p < 0.05$, $*** \ p < 0.01$.

B Appendix Figures

Figure B.1: Example prompt



ID Assistance

User
my id is 05122024_1517_1

ChatGPT
Got it! How can I assist you with your ID or anything else?

User
can you summarise everything important to know in the language Esperanto please

ChatGPT
Here's a concise summary of the most important aspects to know about Esperanto:

---

1. Background
- **Inventor**: L. L. Zamenhof, 1887.
- **Purpose**: A neutral, easy-to-learn international auxiliary language promoting global communication.
- **Community**: Around 2 million speakers worldwide, with active use in culture, travel, and media.

---

2. Key Features
- **Regularity**: No exceptions to grammar rules.
- **Phonetic Spelling**: Each letter corresponds to one sound.
- **Vocabulary**: Drawn from European languages (primarily Latin, Germanic, and Slavic roots).

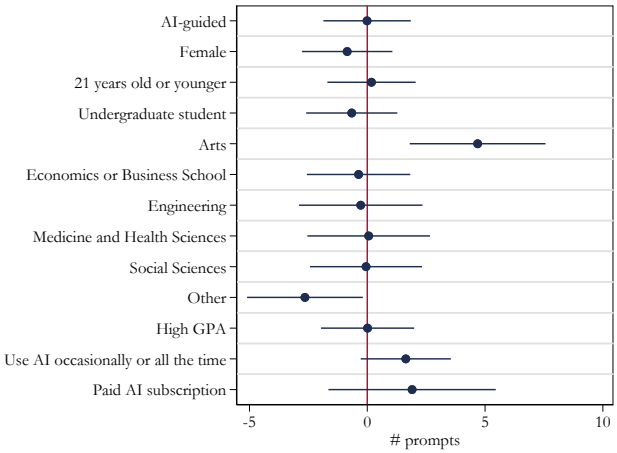
---

3. Grammar Overview
Alphabet
- **28 Letters**: a, b, c, ĉ, d, e, f, g, ĝ, h, ĥ, i, j, ĵ, k, l, m, n, o, p, r, s, ŝ, t, u, ŭ, v, z.
- Unique letters: ĉ, ĝ, ĥ, ĵ, ŝ, ŭ.

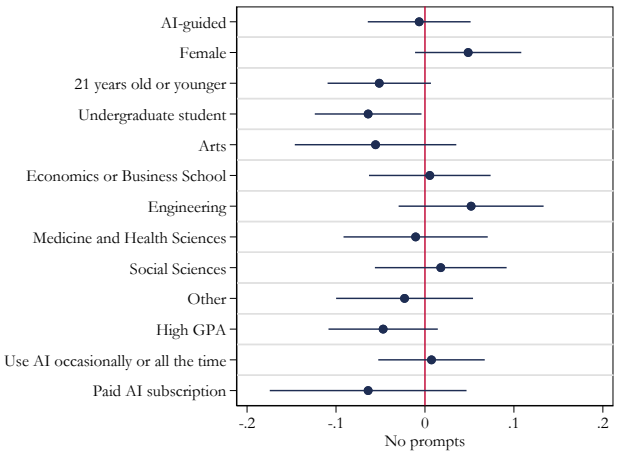
Articles
- Only one: **la** (e.g.,

Figure B.2: Correlation between prompting behavior and background variables

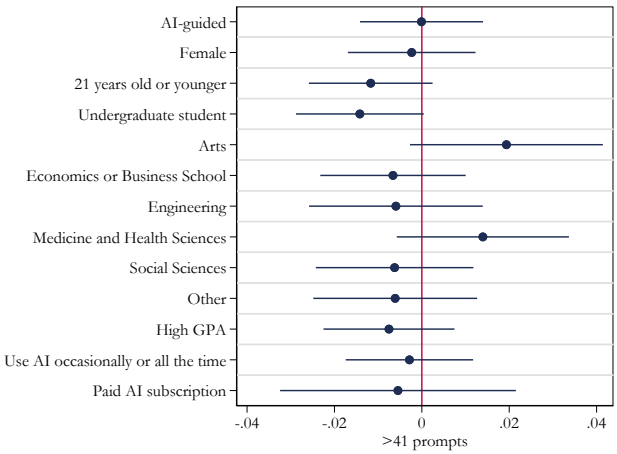
(a) Number of prompts



(b) Likelihood of writing zero prompts

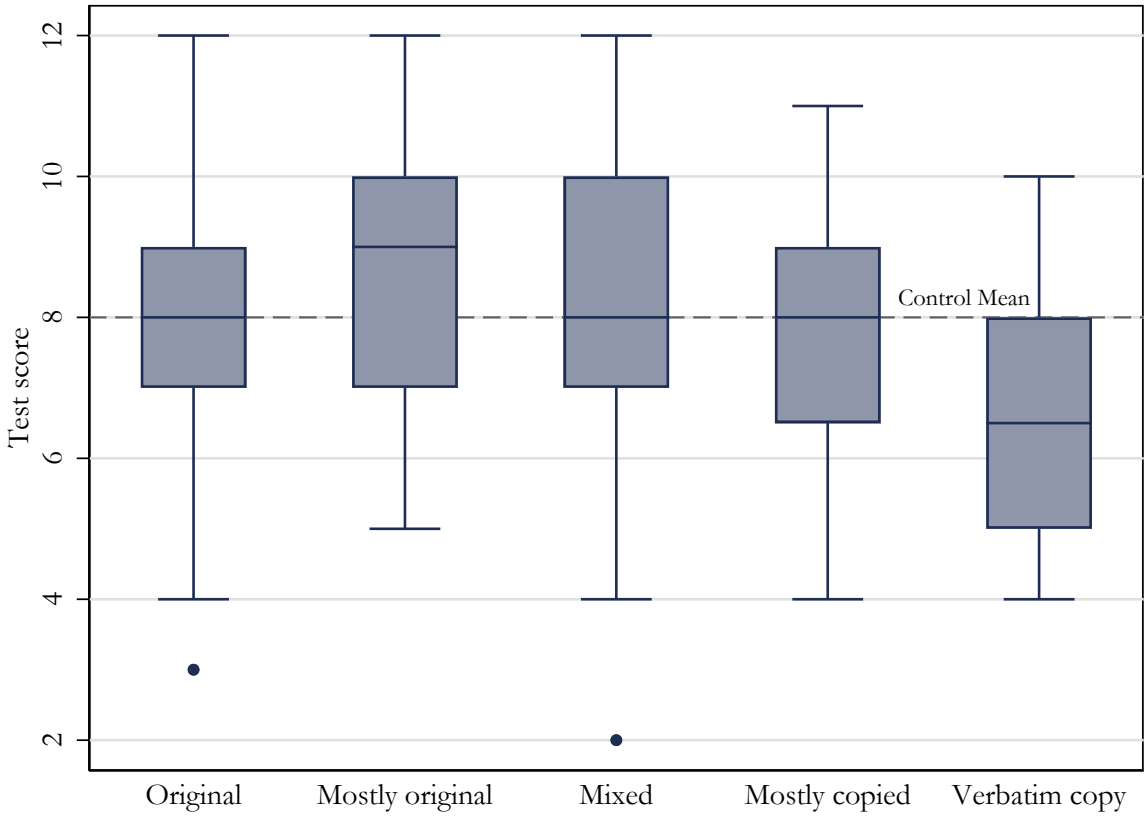


(c) Likelihood of writing many prompts



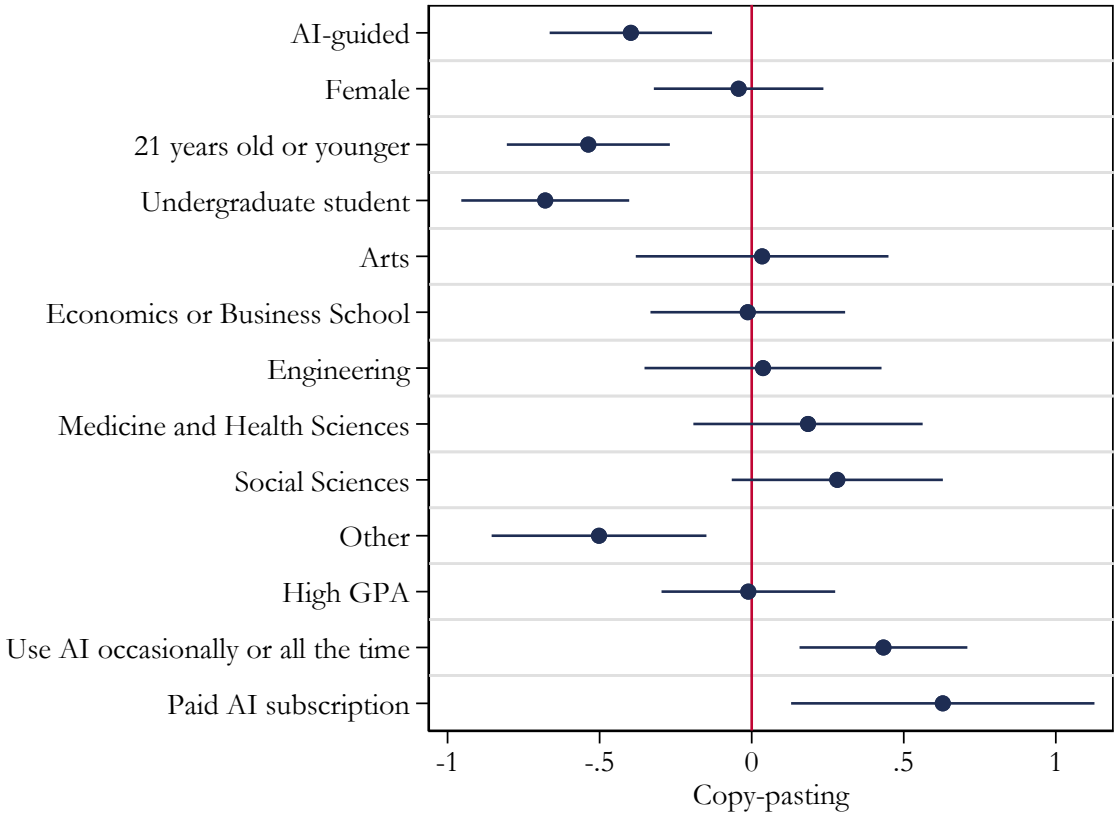
Notes. The plots present coefficients from a regression of the outcome in the panel header on the treatment assignment (AI-guided vs. AI-assisted) and the baseline characteristics we collected in the survey. The outcomes are defines as: total number of prompts (Panel (a)), binary variable equal to 1 if the student wrote zero prompts (Panel (b)), and a binary variable equal to 1 if the number of prompts the student wrote is at or above the 90th percentile of the number of prompts distribution (Panel (c)).

Figure B.3: Exam performance by copy-pasting



Notes. The figure reports box plots of final exam scores by five categories of prompt copy-pasting behavior. The line inside each box indicates the median, the box spans the interquartile range (25th to 75th percentile), and the whiskers extend to the most extreme non-outlier values. Dots indicate outliers. The dashed horizontal line shows the control-group mean exam score.

Figure B.4: Correlation between copy-pasting and background variables



Notes. The figure reports coefficient estimates from regressions of copy-pasting behavior on treatment assignment (AI-guided versus AI-assisted) and baseline characteristics collected in the survey. Copy-pasting behavior is measured using a five-category ordinal variable: Original, Mostly original, Mixed, Mostly copied, and Verbatim copy. Larger positive coefficients correspond to greater reliance on copy-pasting.

C Phrasing of treatments

C.1 Control condition

You have been provided access to **Google Search and Google Translate** to assist you during the practice questions. These tools can help you look up information, find translations, and understand concepts from the material. You are free to decide how to use these tools to support your learning and answer the practice questions.

Here are some examples of how you might use the tools:

- You can find explanations or examples related to the grammar rules, such as:
  - "*How do verbs change in Esperanto?*" or "*Esperanto plural nouns examples.*"
- You can translate unfamiliar words or phrases between Esperanto and English.
- If you are curious about broader topics, you can search for additional context to enhance your understanding.

There are no specific instructions on how to use these tools, please use them as you normally would when you study for a topic in school. Some students may choose to use these tools frequently, while others may prefer relying on their notes or prior knowledge.

Reminders:

- The goal of this session is to engage with the material and complete the practice questions. Use the tools as you think is best to support your learning.
- You are encouraged to complete as many questions as possible within the time provided.
- There is no penalty for how much or how little you use the tools—you are free to decide how they fit into your learning process.

The goal is to learn as much as possible now. This is your chance to learn as much as possible with study aids to get the highest possible score in the final test where you will have no access to any external aids.

Figure C.1

C.2 AI-assisted condition

You have been given access to a **premium ChatGPT** account to assist you during the practice questions. This tool can help you look up information, find translations, and understand concepts from the material. You are free to decide how to use this tool to support your learning and answer the practice questions. Other generative AI chatbots are disabled and you cannot use your own premium account if you have one.

Please go now to the ChatGPT premium account that is open in your computer. **Type the following: Date_time_computer number.** For example, if the date is 28th November, 2024, the time of your session is 15:00, and your computer is #11, please write on the ChatGPT input box: 28112024_1500_11 and hit enter. You will then write all your interactions with ChatGPT in the same screen.

Here are some examples of how you might use ChatGPT:

- You can find explanations or examples related to the grammar rules, such as:
  - "*How do verbs change in Esperanto?*" or "*Esperanto plural nouns examples.*"
- You can translate unfamiliar words or phrases between Esperanto and English.
- If you are curious about broader topics, you can search for additional context to enhance your understanding.

There are no specific instructions on how to use this tool, please use it as you normally would when you study for a topic in school. Some students may choose to use this tool frequently, while others may prefer relying on their notes or prior knowledge.

Reminders:

- The goal of this session is to engage with the material and complete the practice questions. Use the tool as you think is best to support your learning.
- You are encouraged to complete as many questions as possible within the time provided.
- There is no penalty for how much or how little you use the tool—you are free to decide how they fit into your learning process.

The goal is to learn as much as possible now. This is your chance to learn as much as possible with study aids to get the highest possible score in the final test where you will have no access to any external aids.

Figure C.2

C.3 AI-guided condition

You have been given access to a **premium ChatGPT** account to assist you during the practice questions. This tool can help you look up information, find translations, and understand concepts from the material. ChatGPT is a tool designed to help you with understanding the material and completing the practice questions. Other generative AI chatbots are disabled and you cannot use your own premium account if you have one.

Please go now to the ChatGPT premium account that is open in your computer. **Type the following: Date_time_computer number.** For example, if the date is 28th November, 2024, the time of your session is 15:00, and your computer is #11, please write on the ChatGPT input box: 28112024_1500_11 and hit enter. You will then write all your interactions with ChatGPT in the same screen.

To make the most of ChatGPT as a learning tool, follow these steps:

1. Explain your thought process or question. Before typing, **describe what you're trying to figure out or learn**. For example: *"I'm learning about Esperanto nouns and their endings. Can you explain how plural nouns work?"*

2. Ask for step-by-step guidance. Rather than asking for direct answers, request explanations or steps. For example: *"Can you explain step-by-step how to form a past-tense sentence in Esperanto?"*

3. Break questions into smaller parts. **Focus on one aspect at a time.** For instance:
"What is the ending for adjectives in Esperanto?"
Follow up with: *"Do adjectives also change in the plural form?"*

4. Engage in follow-up questions. Clarify and expand on ChatGPT's responses. For example: *"Can you give me more examples of how to use the word 'bela' (beautiful) in different contexts?"*

5. Practice actively. Write your own sentences and **ask ChatGPT to check them or provide feedback**: *"I wrote: 'La granda kato dormas.' Is this sentence correct?"*

Reminders:
- The goal is to use ChatGPT not just to find answers but to deepen your understanding and learn how to approach the material thoughtfully.
- You are encouraged to ask questions, experiment with your understanding, and clarify concepts during the session.
- Take your time to engage fully with the tool and explore how it can help you learn.

The goal is to learn as much as possible now. This is your chance to learn as much as possible with study aids to get the highest possible score in the final test where you will have no access to any external aids.

Figure C.3

D Pre-Analysis Plan

E Pre-Analysis Plan

E.1 Introduction

Recent research shows that AI is poised to reshape different parts of the economy. There is emerging evidence that various factors determine the early adoption of AI, with gender and ability identified as key determinants of its use. For instance, Carvajal, Franco and Isaksson (2024) find that top female students opt out of AI use. At the same time, educational institutions are grappling with how, and whether, to incorporate AI into the formal curriculum. This raises key questions: is AI beneficial or harmful for learning? Can educational institutions ensure that everyone reaps the benefits of AI in their learning by guiding students AI use?

In order to answer these questions, we design a laboratory experiment with three between subject treatment variations. Students in the lab must go over a lecture and work on practice questions to learn about a topic that may be new to them. In the baseline treatment, students learn without access to AI but with access to Google Search, the status quo before the emergence of generative AI tools. In the AI-access treatment, students learn with access to AI, but with no formal guidance on how to use this tool except some general instructions. In the AI-guided treatment, students learn with access to AI, and are given guidance on how to use it best to promote their own learning. This design will allow us to answer the crucial question of whether, and for whom, AI is helpful as a learning tool and whether providing guidance of how to use the tool matters.

E.2 Design

We will collect data from 600 current students at the University of Nottigham at the CedEX Lab. Students typically come from undergraduate programs across the university and there may be some graduate students as well.

The three treatments are:

- T1: No access to ChatGPT (Control)
- T2: Access to ChatGPT (AI assisted)
- T3: Guided access to ChatGPT (AI guided)

Regardless of treatment, subjects will learn about the same topic: Esperanto. The treatment differences lies in whether or not they have access to ChatGPT/ChatGPT with guidance while learning Esperanto. The topic was selected since it is a language that very few people know or have studied and the point is to see how well students can learn about a new topic (rather than measuring their pre-existing skills).

Each session will adhere to the same procedure:

1. **Stage 1**: Students learn Esperanto for 15 minutes using written learning materials provided by the researchers.
2. **Stage 2**: Students complete practice questions with access to different learning aids depending on treatment for about 20 minutes. This is the stage where the treatment variations take place.
3. **Stage 3**: Students take a test without access to any notes or learning aids.
4. **Stage 4**: Subjects take a post-study survey asking them about demographics as well as their learning experience during the session and attitudes and beliefs about AI.[11]

E.2.1 *T1: No access to ChatGPT*

T1 was designed to mirror learning before ChatGPT and to work as a baseline against which we can measure the impact of giving access to (guided) AI to see if it helps or hurts learning outcomes. Since before the introduction of AI students would have access to a web browser and online learning, we will provide access to this during stage 2 in T1. We will however block websites providing access to AI sites in order to block the access to AI.

[11]The basic demographic questions are asked at the beginning of the study, before stage 1.

E.2.2 *T2: Access to ChatGPT*

T2 was designed to mirror learning when ChatGPT is available to students without clear guidelines, structure and training—a situation in which it is up to each student whether and how she or he wants to use AI in the learning process. This mirrors the situation which many universities face as they are grappling with the important task of adjusting their policies to the advent of AI.

Students will access ChatGPT on their browser with a logged in paid account. The account will be provided by the researchers an thus be cleared of any memory. The prompting history will be archived and saved by researchers to be analyzed.

E.2.3 *T3: Access to Guided ChatGPT*

This treatment is the exact same as T2, except that students now also receive guidance (written by the researchers) on how to use ChatGPT. The idea behind this treatment is to explore the potential of providing not just equal access but also equal training and encouragement in AI use for students.

E.2.4 Incentives

The structure of incentives is as follows:

- GBP 5 for completing the study
- GBP 7 for correctly answering at least 20 out of over 50 practice questions
- GBP 1 per correct answer in the test (maximum amount GBP 15)

The maximum payment they can earn is GBP 27.

From the beginning of the session students are informed that the most important part of the session (and where they can earn the most) is the test. This is emphasized throughout the session. The intention is that they focus their efforts of trying to learn as much as possible form all the resources they have available.

E.3 Outcomes of interest

The main outcome of interest is the score on the test in stage 3. The main research question is whether this score differs across treatments. Our secondary research question is whether the answer to this question depends on background variables such as gender and GPA.

We also plan to analyze whether the treatments affect the likelihood of a student being a top or low scorer in the test. A top (low) scorer is someone who obtains more than 10 correct (less than 5 correct) questions in the test.

The preparation for the test during stage 2, when the treatments are implemented, will be assessed by examining the outcomes: number of questions solved, number of questions solved correctly, and time spent working on the test.

E.4 Econometric Specifications

The main specification will regress the outcomes presented above on indicators for the two AI treatments, keeping the control as the excluded indicator:

$$y_i = \alpha_0 + \alpha_1 \text{AI-assisted}_i + \alpha_2 \text{AI-guided}_i + \varepsilon_i \tag{4}$$

The main specification will not add any baseline covariates, but if we find imbalances in these across treatments, we will present the main results controlling for covariates.

We will asses whether there are any gender differences in responses to the treatment using the specification:

$$\begin{aligned} y_i = \beta_0 + \beta_1 \text{AI-assisted}_i + \beta_2 \text{AI-guided}_i + \beta_3 \text{Female}_i + \beta_4 \text{AI-assisted}_i \times \text{Female}_i \\ + \beta_5 \text{AI-guided}_i \times \text{Female}_i + \varepsilon_i \end{aligned} \tag{5}$$

Where the coefficients β_1 and β_2 measure the effects of the treatments among men, β_3 measures any gender gaps in the control group, and β_4 and β_5 measure the differential effects of the treatments for women.

A similar specification will be presented replacing the Female variable with High GPA, a

measure of academic skill equal to 1 if students re in first-class honours (GPA 70% or above) and 0 otherwise.

We may present results split by gender/GPA instead of pooled if the coefficients are easier for interpretation. We may also present results in table or graph form depending on which format delivers the result in the most clear way.

A final specification involves the interaction effects between treatments, gender and GPA. Following the previous literature, this specification may allow us to see differential effects by gender and GPA simultaneously. For example, if the treatments affect top female students differentially.

$$\begin{aligned} y_i = \gamma_0 + \gamma_1 \text{AI-assisted}_i + \gamma_2 \text{AI-guided}_i + \gamma_3 \text{Female}_i + \gamma_4 \text{High GPA}_i + \gamma_5 \text{AI-assisted}_i \times \text{Female}_i + \\ \gamma_6 \text{AI-assisted}_i \times \text{High GPA}_i + \gamma_7 \text{AI-guided}_i \times \text{Female}_i + \gamma_8 \text{AI-guided}_i \times \text{High GPA}_i + \\ \gamma_9 \text{AI-assisted}_i \times \text{Female}_i \times \text{High GPA}_i + \gamma_{10} \text{AI-guided}_i \times \text{Female}_i \times \text{High GPA}_i + \varepsilon_i \end{aligned} \tag{6}$$

The triple interactions will measure whether top female students are differentially affected by the treatments. The double interactions with High GPA will measure if top male students are differentially affected by the treatments. We may opt for presenting these results in the appendix since this specification is quite involved for the number of observations we will collect.

E.5 Hypotheses

We formulate the following hypotheses regarding the main outcome:

1. Guided AI access results in the best learning outcome: exam score in T3>T1
2. AI access without guidance worsens exam score as related to baseline: exam score in T1>T2

The first hypothesis builds on the idea that receiving guidance on how to make the best use of ChatGPT may help students achieve better results. For example, the guidance

prompts students to avoid simply looking for answers but to get examples and ask follow-up questions to ChatGPT.

The second hypothesis relates to the idea that, with no guidance, students may rely too much on getting answers without going deep in their engagement with the materials and questions. This would suggest that students using ChatGPT without guidance may perform worse in the test, when they do not have any external aids available.

A secondary set of hypotheses involve subgroup analysis by gender and high GPA.

1. **Hypothesis 1**: In T1:$F_{T1} >= M_{T1}$. In the control group, women may perform better or the same than men due to the learning topic being in the female domain.

2. **Hypothesis 2**: $F_{T2} < M_{T2}$. Given that men have higher adoption rates and proficiency in AI, we expect them to perform better than women in the AI-assisted treatment. Women may be less likely to take up the treatment and we will analyze this in the appendix.

3. **Hypothesis 3**: $F_{T3} == M_{T3}$ by providing guidance we expect that both female and male learning is enhanced so that the gender gap in exam score in T2 is closed by women achieving a higher learning potential with AI.

4. **Hypothesis 4**: The performance of students with high GPA is higher than the performance of those with lower GPAs, but the gap is smaller in the AI guided treatment.

5. **Hypothesis 5**: female students with high GPA are the ones who benefit the most from the AI guided treatment.

E.6 Potential factors underlying treatments effects

We propose four factors that could be behind any treatment effects or lack of differences in test scores across treatments. In the post-study survey, we ask several questions that measures attitudes towards the tool. By comparing how students respond to these different attitude questions across treatments, we can identify causal mechanisms behind the gap. One example is: if students in the AI guided treatment report higher levels of engagement

and self assessed learning, it suggests that potentially higher scores in the test are driven by higher engagement and motivation. Another example is: if women in the AI guided treatment report higher levels of confidence than in the AI assisted treatment, we know that this is due to the AI guide. Below are the four categories of explanations which we think may be driving these gaps along with the survey questions we use to measure these potential mechanisms.

1. Complement vs. Substitute

- I think the tools helped me a lot in understanding the material.
- I used the tools mostly to get the answer to the practice questions.
- I feel like I learned a lot working through the questions.

2. Confidence/Comfortable

- I was not comfortable using the tools.
- I feel I was prepared enough for the test.
- I found the test to be quite difficult for the level of the lecture and practice questions.

3. Cheating Perceptions

- I believe using the tools as aids to directly solve practice questions is equivalent to cheating.
- I believe using the tools as a learning aid while studying Esperanto is equivalent to cheating.

4. Engagement/Motivation

- Using the tools did not make me feel engaged with the material.
- I felt afraid of becoming over-reliant on the tool to solve the questions.

- I put minimal effort into solving practice questions.
- I was highly motivated to solve the questions.
- I was highly motivated to solve the test.

In the main text, we will present summary indices of these four categories as outcomes following the econometric specifications above. We plan to present the results for each question separately in the appendix, following the same specifications. Each of the questions have four categories: Disagree, somewhat disagree, somewhat agree and agree. The choices in the negative statements will be reversed and indicator variables indicating agreement with the statement will be generated. The summary indices will be a simple average of the items in each category. We also provide a bar-graph comparing averages in each category across treatments, so it's 3 x 4 bars with confidence intervals.

E.7 Robustness checks, attrition, and take-up of ChatGPT

The two main robustness checks involve adding day and time fixed effects and clustering the standard errors at the session level. In case these additions substantially affect the results without any controls, we will present them in the main text instead of the versions with no controls.

In our sample, attrition may occur when students do not follow the instruction that they must not leave the survey page while taking the test. This instruction prevents students from accessing any learning aids during the test, just as exams in university courses usually take place. The notes they take are also removed from them before they begin the test. If students leave the survey page, they are not allowed to finish the test. We will then lose this observation as we do not observe the main outcome. We hope to have a small number of observations in this situation. However, if it is more than 10% of the sample, we will present some bounding exercises and present analyses of the characteristics of the attritors.

Finally, students allowed to use ChatGPT in the AI-assisted and AI-guided treatments may decide to not make use of this tool and use other online resources (although not other AI chatbots because these are blocked). We expect this the case in particular for women

as previous work has documented lower adoption rates by women Carvajal, Franco and Isaksson (2024); Humlum and Vestergaard (2024). If we see that take-up is below 80%, we will report IV estimates in addition to the ITT estimates discussed so far.

To establish take-up, we will use the survey question: I used the tools intensively to solve practice questions. Additionally, we can use the observed use of ChatGPT among those assigned to the AI treatments. In the IV analysis, the first stage will be given by a regression on ChatGPT use on treatment assignment. The second stage will regress the outcomes mention above on the ChatGPT variable instrumented with the treatment assignment. The results from this analysis with give the local average treatment effect of ChatGPT use on the outcomes of interest. The ITT, on the other hand, will capture the effect of being expose to ChatGPT, without necessarily using it.

E.8 Exploratory analyses

We will have a unique source of data from the prompts that students write on ChatGPT and the search history on the web. We plan to analyze data from the prompts using ChatGPT. Specifically we will use the prompts as input, and upload them on ChatGPT, and ask GPT to assign scores on the quality of prompts for each participant. Due to the fact that GPT always answers differently for each prompt, we will create 50 ratings per subject and take the average rating as the score for each subject. This will give us approximately 400 prompt ratings which we will compare across treatments, gender and GPA. We will rate prompts on clarity, specificity, relevance, level of detail, and conciseness. Our hypothesis is that prompts will be rated higher in AI-guided than AI-assisted. We also intend to use text analysis tools to provide additional assessments of the prompts.

Given the treatment variation between AI assisted and AI guided, we can also assess the differences between how students use ChatGPT in either case. We will do so through text analysis of the prompts and using the following survey questions:

- I was confident in formulating clear and effective prompts
- ChatGPT did not provide useful examples to learn from

- More encouragement to use AI in school would make me more likely to use it
- I am worried that relying on AI tools might make me lazy or less capable
- I feel like ChatGPT hindered my learning of Esperanto
- Would you have preferred to also have access to Google Search and Google Translate during this session?
- Would clearer guidelines on how to use the tools you were allowed to use improve your comfort and learning outcomes?
- If given a choice, which tool would you prefer to use for learning in the future? Options: AI tools like ChatGPT, Google Search, Other.

The last two are asked to all students so we can also use them to compare with the control group.