The Generative AI Learning Penalty
Evidence from Chinese K12 Education

Victor Lei* David Strömberg† Yanhui Wu‡

March 2026

Preliminary and incomplete. Please do not circulate.

Abstract

We study the effects of generative AI on student learning using 30 months of panel data on 26,811 Chinese students in grades 7–12. Our data include monthly closed-book exams, high school and college entrance exams, and homework scores and completion times across 7–9 subjects. We estimate effects using a staggered difference-in-differences design exploiting variation in the timing of students' adoption of generative AI. Generative AI dramatically increases homework productivity: scores rise by 13 percentage points (1.9 SD) while completion times fall by 19 hours per month. However, scores on closed-book classroom exams fall by 12 percentage points (1.4 SD), indicating large reductions in the learning of recently taught material. Scores on college and high school entrance exams—which test the long-term accumulation of general skills—fall by 2.5–3 percentage points (0.3–0.4 SD). The effects on entrance exams accumulate more slowly, reaching their full magnitude only after 20 months of AI use, compared to 6 months for regular exams. This dynamic pattern implies that short-duration studies systematically underestimate the long-run learning cost of generative AI. The mechanism seems to be reduced study effort, not reduced learning efficiency. More than half of AI-using students outsource homework to generative AI, completing homework far quicker while receiving high scores. Students who spend the same time studying learn equally well with or without generative AI. In other words, generative AI does not seem to make studying less productive—it makes students study less.

*University of Hong Kong
†Stockholm University
‡University of Hong Kong

1 Introduction

Generative AI has become increasingly capable of performing knowledge-intensive tasks at human-level proficiency (Achiam et al., 2023). A growing body of evidence shows that it allows workers to complete such tasks much more efficiently (Brynjolfsson et al., 2025; Noy and Zhang, 2023). However, using AI to complete a task means that workers may not learn from the task because they are not actually doing it. This issue is particularly important for young students who are accumulating general human capital through practicing a wide range of cognitive tasks in school. A key open question is whether generative AI improves learning by acting as a personalized tutor, or undermines learning by enabling humans to outsource tasks and bypass the cognitive effort that produces learning.

We study this question using 30 months of panel data on 26,811 Chinese students in grades 7–12. Our data include monthly closed-book classroom exams, high school and college entrance exams, and homework scores and completion times, all by subject. We exploit staggered variation in the timing of students' adoption of generative AI in a difference-in-differences design (Callaway and Sant'Anna, 2021). Our setting is well suited for this design: students' reported use of generative AI expanded rapidly during our sample period, from marginal in 2022 to around 80 percent by June 2025, with adoption staggered across students. Already early in our sample period, Chinese generative AI tools such as Qwen and ERNIE Bot were quite capable, providing correct responses to around 90 percent of homework and exam questions in "zero-shot" queries (Hou et al., 2024).

This setting offers three advantages over the randomized trials that dominate the existing literature.[1] One, students choose their own AI tools and usage patterns, so we capture how generative AI affects learning under realistic demand conditions rather than mandated tool use. Two, we observe learning outcomes over months and years, considerably longer than existing studies. Three, we measure both short-run learning of recent material (monthly exams) and long-run accumulation of general skills (entrance exams), across multiple subjects. Thus, our study has the advantage of addressing the poorly understood question of how AI usage affects human capital accumulation. Our main findings are as follows.

First, we find a sharp divergence between task completion and learning. Generative AI dramatically increases homework productivity: after six months of use, homework scores rise from 73 to 86 percent correct while completion time falls from 64 to 45 hours per month. However, scores on closed-book classroom exams fall from 60 to 48 percent correct—a decline of 12 percentage points (1.4 SD). Scores on the high-stakes college and high school entrance exams fall by 2.5–3 percentage points (0.3–0.4 SD).

Second, the effects on regular and entrance exams differ not in magnitude but in speed. Regular exam scores, which are based on recently taught material, deteriorate fully within six months of AI adoption and then stabilize. Entrance exam scores, which are based on skills accumulated over years, deteriorate more slowly and reach their full negative magnitude only after about 20 months. Because most students in

[1] For meta studies, see Lo et al. (2024), Wang and Fan (2025) and OECD (2026) for a discussion.

our sample adopted generative AI relatively recently, the average effect on the June 2025 entrance exams understates the long-run impact. For long-term users, the negative effects on entrance exams are as large as those on regular exams. This dynamic pattern implies that short-duration studies that dominate the existing literature systematically underestimate the long-run learning cost of generative AI.

Third, we find that these negative effects are largely driven by students' decreased learning effort instead of reduced learning efficiency. We document evidence on this mechanism through the joint distribution of homework time, homework scores, and exam scores. More than half of AI-using students complete their homework assignments in substantially less time on homework than even the fastest non-AI student, receive homework scores consistent with generative AI's known accuracy on similar problems (ca 90 percent), and score very poorly on exams. This pattern suggests a story of outsourcing: students ask their AI tool for direct answers rather than working through the material themselves.

Critically, AI-using students whose efforts on homework are comparable to those of non-AI students achieve similar exam scores despite also using generative AI for homework (as evidenced by their increased homework scores). These students are not differentially selected on prior achievement. This indicates that generative AI reduces the time students spend learning but does not reduce learning efficiency conditional on time. In other words, generative AI does not make studying less productive, but it makes students study less.

Fourth, we estimate the effects across subjects. Social science subjects are most negatively affected (particularly politics and geography, at -14 percentage points), followed by STEM (-10 pp) and languages (English, -7 pp, Chinese, -5 pp). This is perhaps surprising, given that students most commonly report using AI for mathematics and English, and that most RCTs study effects on mathematics, programming and foreign language skills.

Our paper contributes to a growing literature on the effects of AI on student learning. The most closely related study is Bastani et al. (2024), who study 1,000 Turkish high school students in an RCT. Students using a standard LLM scored lower on a post-exercise closed-book exam, while those using a specially designed "GPT tutor" scored comparably to non-AI students. Their estimated negative effect on math scores of around 17 percent is similar to ours. However, the estimates capture different objects: Bastani et al. measure the immediate effect of a mandated tool over 90 minutes with fixed time, while we estimate the cumulative effect of student-chosen AI tools and use over months and years. Conditional on study time, we find no effect of AI on learning—the negative effect operates through reduced time.

Kestin et al. (2025) find that Harvard physics students learn more in less time using a specialized AI tutor compared with in-class active learning. Their result is not inconsistent with ours: with a well-designed tool and highly motivated students, AI can enhance learning efficiency. Our contribution is to show that, under realistic conditions—where typical teenagers choose their own tools—the dominant pattern is outsourcing, not tutoring.

Much of the existing literature has focused on supply-side technical solutions: designing AI tutoring tools that scaffold learning and prevent direct answers. Some of

these designs have proven effective (e.g., Kestin et al., 2025). Our findings suggest that equal attention should be given to the demand side. Well-designed AI tutors exist, but the teenagers in our sample overwhelmingly choose generic tools that provide direct answers, spend far less time studying, and learn substantially less. Addressing the learning costs of generative AI may require changing student incentives—not just improving AI tools.

2 Data

2.1 Survey design and sample

We use data from a survey regarding students' use of generative AI tools that was conducted in a county located in a central province of China in late June, 2025. The survey covered approximately 26,811 students from Grade 7 to Grade 12 across five junior high schools and four senior high schools, accounting for 90% of all students enrolled for secondary education within the county. The students beyond our survey coverage are those living in remote rural areas and are difficult to track.

The survey was designed and administered by the local Educational Bureau and distributed via a professional online platform (https://www.wjx.cn), which is commonly used for running research-oriented surveys in China. Local teachers sent the survey link to their students through a class-based WeChat group used for day-to-day communication with students. Before distributing the survey link, teachers instructed students to verify their actual registration dates with the AI tools they used. This was done to ensure the accuracy of students' self-reported timing of initial usage of AI. The final valid response rate exceeded 96%.

2.2 Exam and homework data

Data regarding students' academic performance were provided by the local educational bureau, which extracted the monthly exam results from a digital grade-management system. Monthly exams are given every month in each subject (7 subjects in lower secondary, 9 subjects in upper secondary school). The exams are processed through a unified digital grading platform where teachers grade students' work directly within the system. This platform records not only the aggregate scores but also granular, item-level data for each sub-question and every student using a student ID number. Similarly, weekly homework assignments in each subject are submitted and graded via a digital platform, which in addition to student responses and grades also logs timestamps for when the homework assignment was released by the teacher and the answer submitted by the student. We use the time it takes for students to complete and hand in homework assignments, defined as the time between when the homework assignment was released and when the student submitted the answers on the digital platform. This is an imperfect measure of the actual time spent working on the home-

work.[2] Nevertheless, we believe that this variable is a reasonable proxy for the time students spend on assignments on average.

We have past exam and homework data for each student in their school at the time of the survey (June 2025). This implies that for students in 9th and 12th grade in 2025, we have three years of data, for students in 8th and 11th grade we have two years, and for students in 7th and 10th grade we have one year of data. In robustness analyses, we will focus on the 4,067 9th graders and 4,399 12th graders for which we have three years of outcome data.

The local educational bureau also provided three sets of student performance data beyond the exams administered by each school. The first is the Joint Exam administered at the county level to assess students' learning outcomes using the same exam papers across schools. This was provided for students in Grade 7 and Grade 9. The second is Zhongkao, the high-school entrance exam for Grade 9 students towards the end of their graduation. The third is Gaokao, the college entrance exam for Grade 12 students taking place in June every year. Both Zhongkao and Gaokao are administered at the provincial level, and students in the same province are required to take the same exams. Details about the administration and grading of Zhongkao and Gaokao can be found in the appendix.

Our main outcomes, monthly regular and county-wide exams, homework scores, homework completion time, and entrance exams are all at the subject by month level for each student. We rescale all exam scores (regular, county-wide, entrance) so that they are measured as percent correct on a 0–100 scale. A decline of 11 percentage points therefore means that students answer 11 fewer questions correctly out of every 100. In our monthly panel, we use averages of these variables across subjects (e.g., exam score average) as dependent variables.

2.3 Demographic data

Student demographic data were obtained from digital records collected by local educational authorities at the point of student enrollment, based on standardized protocols implemented across all schools. Furthermore, we obtained from schools information regarding a student's leadership role.

2.4 Trends in generative AI use

At the start of our sample period, in September 2022, OpenAI's GPT-3 was the state-of-the-art model with moderate math, science and reasoning capabilities and ability to write short essays. At this point, only a small fraction of our sample students were using generative AI (see Figure 1). Use of generative AI remained low throughout 2023. From September 2024 and again in January 2025, we see an acceleration in the uptake of generative AI among students. This acceleration coincided with the release of DeepSeek V2.5 in September 2024 and DeepSeek R1 in January 2025. Another popular

[2] For example, students may not submit immediately after completing a task and students may complete assignments before the platform opens up for accepting submissions, truncating completion times from below.

generative AI tool, Doubao, was launched in August 2023, significantly upgraded in August 2024 and again in January 2025.

At the time of our survey, in June 2025, around 80 percent of the students report using generative AI. The most popular tools were Doubao, DeepSeek, ChatGLM, ERNIE and Qwen, see Figure 1. Generative AI was most frequently used in mathematics (66 percent) and English (55 percent), while between 30 and 40 percent of the students report that they use it in other subjects.

Chinese generative AI tools were relatively competent in answering questions on Chinese exams already early in our sample period. For example, Hou et al. (2024) gather questions from free, regional homework, practice questions, and mock exams available on the internet. They find that QWEN-72B (Alibaba) and ERNIE Bot 4.0 (Baidu) correctly answer around 90 percent of these questions in a "zero-shot" query. Similarly, Zhang et al. (2023) collect questions from the more complex Gaokao exams and find that GPT-4 correctly answers 72 percent of the questions and ERNIE Bot correctly answers 57 percent of the questions.

3 Specification

We define treatment by the month in which a student first adopts generative AI. Students who adopt by the end of the sample period, June 2025, are *ever-AI students*, while students who have not adopted by then are *never-AI students*. We refer to students who have adopted generative AI by a given month as *AI students* in that month, and to students who have not yet adopted by that month as *non-AI students*. Thus, the classification of students as ever-AI or never-AI students is fixed over the sample period, while the classification of students as AI or non-AI students varies over time.

Our object of interest is the effect of having started using generative AI. Specifically, we estimate the cohort- and event-time-specific average treatment effect on the treated: the effect, at a given event time, of having adopted generative AI for students who first adopt in a given month, relative to the outcome path they would have followed had they not adopted. We proxy this counterfactual outcome path using never-adopters.

We estimate these effects using the staggered difference-in-differences estimator of Callaway and Sant'Anna (2021). For students who first adopt in month g, the estimator compares the change in outcomes between month $g-1$ and month t to the corresponding change for never-adopters.

These cohort-specific effects are then aggregated across adoption cohorts and event times to obtain the summary measures reported below. We focus on three such aggregations. First, we report the overall average treatment effect on the treated (ATT), which averages the group-time effects across all treated cohorts and all post-adoption periods for which they are observed. This summarizes the average post-adoption effect for ever-AI students. Second, we report dynamic effects by event time. For each exposure length l, we average the group-time effects across all adoption cohorts observed l months after adoption. These estimates trace out how the effect of AI adoption evolves with time since first use. Third, we report the average effect 6 to 10 months

after adoption. We focus on this window because, for our main monthly outcomes, the effect appears to reach its full magnitude after about six months. Averaging over event times 6 through 10 therefore provides a compact summary of the full effect of AI adoption.

This framework is well suited to our setting because it allows treatment effects to vary across adoption cohorts. Such heterogeneity is plausible here: generative AI systems improved rapidly over the sample period, so students who adopted earlier may have faced substantially different tools than those who adopted later. The effect of adoption may therefore differ by cohort.

The key identifying assumption is a parallel-trends assumption: absent adoption, ever-AI and never-AI students would have experienced the same evolution in outcomes. Under this assumption, differences in outcome changes between adopters and never-adopters can be interpreted as causal effects of adoption. For further details, see the Appendix and Callaway and Sant'Anna (2021). Standard errors are clustered at the class level (there are 283 classes in our sample).

We also plot event-time means of each outcome from 10 months before to 10 months after first AI use. For ever-AI students, each point is the average outcome at a given event time relative to adoption. For never-AI students, we construct a reweighted comparison series so that, at each event time, the control observations have the same distribution over month, grade, and school as the corresponding treated observations. The appendix provides details.

3.1 Sample balance

Although our identification relies on the parallel trends assumption and not random assignment, parallel trends is a priori more likely to hold if the ever- and never-AI students are more similar. In fact, the ever- and never-AI students are quite balanced on all our demographic variables as well as initial exam scores. The left graph in Figure A1 and Table A1 show that the standardized difference in means between ever- and never-AI students lies within 0.1 for all characteristics, except for whether the parents are in knowledge-intensive or general work, which is at 0.103.

We also characterize the early adopters, relevant in particular for the analysis using not-yet-adopting students as controls. We do this by regressing the number of months since first AI adoption as of June 2025 on our demographic variables. Conditional on grade, our demographic characteristics do not predict early generative AI use. However, students in higher grades are more likely to be early adopters.

4 Results

We begin by presenting our main results, analyzing outcomes in the monthly student panel, which allows us to trace the effects of AI adoption on homework scores, homework completion time, and regular exam scores. For each of these outcomes, we first show the distribution of scores in a histogram and plot the means, separately for AI

and non-AI students. We then present our difference-in-differences estimates, including event-study graphs, the average treatment effect and the effect after 6-10 months.

In the remainder of this section, we examine the mechanism underlying these effects, turn to high-stakes entrance exams as a measure of longer-run learning, and finally consider heterogeneity across subjects. Table 1 contains summary statistics for the key variables in our analysis.

4.1 Homework

We first study how generative AI affects students' productivity in completing homework assignments.

4.1.1 Homework Score

The distribution of homework scores for AI and non-AI students is shown in the left pane in panel A of Figure 2. The generative AI users clearly have higher homework scores. The secondary hump at lower scores consists of students who recently started using generative AI and have likely not yet learned to exploit it fully. The hump disappears for students who have used generative AI more than five months, see Figure A2.

Of course, the scores of AI-students could be higher than those of non-AI students because of selection—if, for example, students with high homework scores are more likely to adopt generative AI early. However, the average homework scores are very similar for ever- and never-AI students before the former start using generative AI, both in the level and the trend; see the center pane in panel A of Figure 2. The ever-AI students' average homework scores rise from around 73 percent correct the month before they start using generative AI to around 86 percent correct after five months, while the control students' average scores remain at 73 percent.

The right pane in panel A of Figure 2 shows our difference-in-differences estimates. The event study graph displays no significant pre-trend. The effect grows for six months and then stabilizes. The average treatment effect and the effect 6-10 months after generative AI adoption are reported below the graph. There is a statistically significant ($p < 0.01$) average effect of 13 percentage points (1.9 SD) between six and ten months after adoption. The overall ATT is a significant 8 percentage point increase, smaller than the 6–10 month estimate because it includes the initial ramp-up period.

The results demonstrate that students use generative AI to significantly boost their homework assignment scores. The effect size after six months is quite large—but the resulting average of 86 percent correct is entirely consistent with the known accuracy of existing generative AI tools.

The key identifying assumption is that, absent generative AI use, ever- and never-AI students would have had parallel trends in homework scores. This assumption is not testable, but the fact that the trends in the pre-period are not statistically different is strongly supportive. The ever- and never-AI students also have very similar average completion times before adoption, indicating that faster students are not differentially selecting into generative AI use.

4.1.2 Homework time

We next study the students' homework completion time. The histogram in panel B of Figure 2 plots the homework completion time for AI students and non-AI students. Completion times are clearly lower when students use generative AI. Again, this could reflect selection: perhaps students who complete assignments quickly are also more likely to adopt generative AI early. However, homework completion times are very similar for ever- and never-AI students before the former start using generative AI; see the right pane in panel B of Figure 2. After the ever-AI students start using generative AI their average homework completion times decline rapidly, from 64 hours to around 45 hours while the never-AI students remain at 64 hours.

The right pane presents results from our difference-in-differences estimator. The event study graph shows that the ever- and never-AI students are on parallel trends prior to the month the former started using generative AI. For six months, the homework completion time falls monotonically. After 6-10 months, the average estimated effect is a reduction of the homework completion time by a statistically significant ($p < 0.01$) 19 hours.

Note that we ask students when they started using generative AI, independent of purpose. In principle, the drop in homework completion times could be because students use generative AI to provide answers to the homework questions, or because students use generative AI for entertainment leaving less time to study. However, the dramatic increase in homework scores is a clear indication that students use it for help with the homework. Note also that the effects on homework time are fully realized after around six months. This indicates that it takes students a while to learn how to use generative AI effectively for homework assignments.

Students using generative AI receive high grades while quickly completing their homework. In this sense, they are much more productive in completing homework assignments than non-AI students.

4.2 Regular exams: short-run learning of novel skills

The next question is how students' use of generative AI affects their learning of novel skills. Our first measure of student learning is their scores on the monthly closed-book exams that students take in class. These exams test whether students have learned recently taught material. Their content closely tracks that of recent homework assignments.

The left pane in panel C of Figure 2 plots the distribution of exam scores for AI and non-AI students. Exam scores are clearly lower for students who use generative AI, with, in particular, a much thicker negative tail. The average exam scores are very similar for ever- and never-AI students before the former start using generative AI; see the center pane of Figure 2, panel C. After the ever-AI students start using generative AI, their average grades decline rapidly from 60 percent correct to around 48 percent correct after five months.

The results from the difference-in-differences estimator in the right pane tell a very similar story. The ever-AI and never-AI students are on parallel trends prior to the

month when the ever-AI students start using generative AI. After this point, the size of the estimated negative effect increases for six months to around -12 percentage points. The average estimated effect for months 6-10 is 12 percentage points ($p < 0.01$). This is a very large effect, amounting to a fall of 1.4 standard deviations or 20 percent of the baseline mean of 60 percent correct answers. The total average effect, including all months, is a statistically significant fall of 8 percentage points.

A potential concern is that the exams are constructed and graded by different classroom teachers and on specific material. Some teachers may write exam questions that are more difficult or that discriminate more strictly based on student ability. To address this concern, we also study student scores on county-wide exams, where all students in a grade, subject and month answer the same questions. There are two county-wide exams per semester and we only have this outcome for 9th- and 12th-graders, so we have fewer observations for this outcome. However, the estimated effects are very similar. Exam scores fall the first six months after students start using generative AI and the negative effect then stabilizes at around -13 percentage points (1.3 SD, $p < 0.01$), see Figure A3.

We finish the presentation of our main results by discussing some identification concerns. The main identification assumption is that of parallel trends. This could be violated if, for example, students whose grades are already declining turn to generative AI to complete their homework. We cannot test this directly, but the ever- and never-AI students have parallel trends in the pre-period in all outcomes, and even the mean levels of these outcomes are very similar. So there is no evidence that these students lose interest or have declining grades in the month before they start using generative AI.

We use the never-AI students as controls—the 20 percent of students who had not started using generative AI by June 2025. Although we find no significant differences in observables (see Figure A1), these students could still be different in ways we do not observe. For robustness, we also include students who have not yet started to use generative AI in the current month of analysis in the control group. Figure A4 presents results for such estimated effects, analogous to the right panel of Figure 2. The results are very similar.

Our main treatment variable, the month that each student started using generative AI, is self-reported and retrospective. Teachers instructed students to verify registration dates, but we cannot observe whether they actually did so. We can, however, rule out that a significant share of students report an adoption month that is later than their actual adoption month. The reason is that this would result in a pre-trend, where outcomes diverge in the months before the reported AI start month. There is clearly no pre-trend in any of the outcomes, so we can rule out this type of error. It could be that all misreporting students report a starting date that is too early. For example, it could be that students register a generative AI account, but only start using it actively with some lag. This would be quite close to our previous interpretation that it takes some time for students to learn how to use their AI tools for homework. It would appear in our data as a slower effect. We measure the effect on our monthly outcomes as fully developed after six months and part of the delay could be caused by this type

of reporting error. Note that this type of error would increase the delay time of the full effect, but not impact the estimated size of the fully developed (after six months) effect. This type of error would also not affect the result that the general skills tested in the entrance exams deteriorate more slowly than those in the regular exams (see below).

4.3 Mechanism

In this section, we examine the joint distribution of homework time, homework scores, and exam scores to identify the mechanism behind the negative effect on learning.

4.3.1 Outsourcing homework to generative AI

In traditional educational settings, homework performance reliably predicts exam performance, reflecting the practice and consolidation effects of completing assignments. Figure A5 shows that this relationship remains strongly positive for non-AI users but becomes *negative* for AI users above the median homework score. This suggests that high homework scores among AI users signal greater reliance on AI assistance rather than greater learning, consistent with the "outsourcing" rather than "tutoring" pattern of AI use.

This view is confirmed when we investigate how exam and homework scores are related to homework time. Figure 3 plots the median and interquartile ranges of homework and exam scores separately for AI and non-AI students. Non-AI students spend at least 50 hours completing their homework. Both homework and exam scores are monotonically increasing in this time, which makes perfect sense given that homework time is a proxy for student effort.

The pattern is very different for students who use generative AI. Slightly more than half of AI students spend only 20–50 hours completing homework, less than *any* non-AI student. This group receives extremely high homework scores. Their median score of around 87 percent is consistent with the performance of generative AI tools on similar problems (Hou et al., 2024). At the same time, they receive extremely low exam scores (around 45 percent correct). This pattern—low time on task, high homework scores, and low exam scores—is entirely consistent with outsourcing, and characterizes more than half of AI-using students.

4.3.2 Learning time and learning efficiency

At the other end, AI students who spend more than 65 hours on their homework receive similar homework and exam scores as non-AI students, indicating that these students do not use generative AI for homework assignments. However, this group consists entirely of students who adopted generative AI within five months. Six months after adoption, no AI student spends more than 65 hours completing their homework (see Figure A2).

For the AI students in the range 45-50 hours, homework scores are falling and exam scores are rapidly rising. It could be that the share of students who are not

using generative AI for homework is rising in this group (while the share outsourcing is falling). However, the fact that all students in this group complete their homework in substantially less time than even the fastest non-AI student makes this interpretation seem less likely. The more likely interpretation is that students in this group are using generative AI more as tutors providing hints and information, as opposed to providing direct answers (outsourcing). This increases the time that students work on the homework, as well as their learning as measured by exam scores. Notably, it also reduces the students' homework scores. This implies that students who only care about their homework scores should avoid using generative AI as tutors, or if they do, also ask for direct homework answers to submit.

Interestingly, in the range 50-65 hours, the median and the interquartile range of exam scores of AI and non-AI students are similar. This implies that, in the entire region where the homework completion times of AI and non-AI students overlap, students who work the same hours on their homework on average receive similar exam scores. In other words, the two groups have roughly equal learning productivity (score per homework completion time) in this range.[3] The AI students in this range receive significantly higher homework scores than non-AI students, indicating that they do employ generative AI for the homework assignments. They are not differentially selected based on their pre-AI grade: their average is around 60, close to the sample mean, see Figure A6. This indicates that students who spend the same number of hours learn as much with or without generative AI, in other words, that generative AI reduces time spent learning but not learning efficiency.

4.4 Entrance exams: long-run learning of general skills

Our next measure of student learning is their scores on college and high-school entrance exams. These entrance exams differ from regular exams in several ways. First, the stakes are much higher in the entrance exams, since they determine the students' ranking in high-school and college admissions. Some students may not study hard for the regular exam, and perhaps even less so if they are distracted by generative AI applications. This is not likely the case for the entrance exams, which have a large impact on the students' future careers. Second, the type of question is different. The regular exams test whether recent material was learned, often mirroring homework, worksheets or textbook examples. The Zhongkao and Gaokao have a much broader cumulative scope, covering course material from multiple years. There is also a strong integration across topics, requiring the students to choose which method to use before applying it. For example, a math problem may combine algebra, geometry, and functional reasoning, without signaling which tool to use. Finally, the entrance exams are

[3] Suppose that a share s of AI students at a given completion time use AI for homework while a share $1-s$ do not. Let $\overline{y}_{AI}$ be the mean exam score for students who use AI for homework and $\overline{y}_{No-AI}$ the mean exam score for those who don't (all non-AI students and a share $1-s$ of the AI students). The graph shows that the mean exam scores for AI and non-AI students are equal. This means that

$$\overline{y}_{No-AI} = s\overline{y}_{AI} + (1-s)\overline{y}_{No-AI},$$

which is only true if $\overline{y}_{AI} = \overline{y}_{No-AI}$.

written and graded by external expert committees, in contrast to the regular exams, which are written and graded by classroom teachers.

We observe entrance exam scores at most twice for every student. For the students in 12th grade, we observe their Zhongkao scores in 2022 and Gaokao scores in 2025. For the 9th-graders, we observe the Zhongkao scores in 2025. The students who use generative AI have worse average scores on the Zhongkao and Gaokao, but the gap is smaller than for regular exam scores; see the histogram in Figure A8. More importantly, because we only observe the outcomes once or twice, it is more difficult to address selection concerns and convincingly identify the causal effect of generative AI use on these outcomes.

We pursue two strategies to partially address this concern. First, we condition on the regular exam score one month before the treated students started using generative AI and apply our difference-in-differences approach, relating grade changes to the number of months since adoption. Note that, in this case, we cannot show a pre-trend, since we do not observe generative AI start months after June 2025. Second, we run a standard regression, controlling for the Zhongkao score in 2022 for the Gaokao outcome and the September 2022 regular exam score for the Zhongkao outcome.

The results are presented in Figure A9. The difference-in-differences estimator yields a statistically significant ($p < 0.01$) estimated effect of -3.4 percentage points for both the Zhongkao and Gaokao scores. Note that we are now estimating a different object than before, namely the average effect in June 2025, across student exposure lengths to generative AI in that month. For comparison, we similarly estimate the average effect of generative AI on regular exam scores in June 2025 for 9th graders (also taking the Zhongkao) and 12th graders (also taking the Gaokao). The estimated effects for the Zhongkao and Gaokao scores are significantly smaller than the estimated effects on the regular exam scores in the same month (June 2025), which are -8.6 and -6.5 percentage points (1.0 and 0.7 SD).

4.4.1 How quickly do skills erode?

To analyze this difference in effect sizes across regular and entrance exams more closely, we consider the dynamic response for these outcomes. Figure 4 shows that the full negative effects on the June 2025 regular exam scores are realized after six months, similar to what we saw before. However, the negative effects of generative AI on the June 2025 entrance exams accumulate more slowly and are only fully realized after two years. This makes perfect sense. The entrance exams test skills accumulated over several years, so these are only gradually affected. The regular exams test novel skills and are more rapidly affected.

The difference in average effects in June 2025 is not because the effects on the entrance exams are smaller but because they are slower. There are many relatively recent users of generative AI in June 2025. The full negative effect of generative AI on the entrance exam scores was not yet realized in the average effect at that time.

Next, we employ our regression control strategy and regress the entrance exam scores on a dummy for whether the student used generative AI in 2025. The results are shown in Table 2. Students using generative AI score on average 2.5 to 3 percentage

points lower on the exams than those who don't (0.3 and 0.4 SD, respectively; $p < 0.01$ for both). This holds, controlling for class-fixed effects and demographic controls as well as the students' initial exam scores (their 2022 Zhongkao scores for 12th graders and their September 2022 regular exam grades for 9th-graders). Exam scores are highly persistent: the coefficients on exam scores lagged three years are around 0.8 and highly significant. As a placebo, we regress students' 2022 Zhongkao score on their June 2025 use of generative AI. The coefficient is insignificant and small, see column (7) of Table 2. We finally compare estimated effect dynamics to the previous approach. The effect dynamics are very similar, see Figure A10.

4.5 Heterogeneity

We finally analyze how the estimated effects vary over time, across subjects and by student characteristics.

4.5.1 Have effects changed over time?

There are several reasons why the effects of generative AI may change over time. The capabilities of generative AI tools have improved and the selection of adopting students has changed. In addition, teachers and students may have adapted to generative AI, changing its effect.

We now investigate how effect sizes change over time. Since the full effect on exam scores appears after five months, we first plot the effects of five months of generative AI exposure, by month (see the left pane of Figure 5). The effects after five months of student AI use are decreasing over time. In early 2023, these effects were around -15 percentage points but the effects fell in absolute terms to -10 percentage points in June 2025. This could be because the estimation sample changes over time (for each month, students who started using generative AI five months ago are included in the sample).

To see whether selection is affecting the results, we study the effect of generative AI on a fixed sample of early users over time. This way, we hold the selection of treated students constant. The effects over time for the group of early adopters are plotted in the right pane of Figure 5. The pattern is roughly the same as the effects after five months. This indicates that the falling absolute effect size is driven by changes in the effects of generative AI, rather than the composition of affected students. The falling effect size could imply that students or teachers are adapting, diminishing the negative effects of generative AI on learning.

4.5.2 What subjects are most affected?

A priori, it is difficult to guess for what subjects generative AI tools would affect student learning more. Empirically, each individual study of the effects of generative AI typically studies one topic in one subject. There is a strong focus on topics in mathematics, programming and foreign languages in the RCT literature on the effect of AI tools on learning. Meta-studies that compare effects across studies, for example,

Wang and Fan (2025), find no significant differences in (positive) learning effects in studies of STEM and related courses relative to those found in studies of language learning and academic writing. However, this compares estimates from very different samples, including different countries and grade levels, and estimates typically based on quite small samples.

We study effects across nine subjects in our setting. Figure 6 shows our estimated average effects of generative AI on exam scores by subject. Overall, social science subjects are most negatively affected, in particular politics and geography with negative effects of 14 percentage points. Next are STEM subjects with negative effects of around 8 to 10 percentage points. Least affected are English (-7 percentage points) and Chinese (-5 percentage points). All subject-level effects are statistically significant at the one-percent level. The figure also contains the effect size relative to the sample mean (blue dots, right axis). For example, the negative effect of generative AI on math exam scores is around 17 percent of the mean score.

4.5.3 What students are most affected?

We next investigate whether effects differ by gender, initial exam score (ability) and weekly reported AI time (dose-response). To this end, we estimate effects in samples split by each of these characteristics. We find a strong heterogeneity by the weekly time that students report using generative AI. The average (ATT) effect for students who report using generative AI one hour per week or less is -3 percentage points, compared to -13 percentage points for students who report using generative AI five hours or more, see Figure A11. There thus appears to be a clear dose response. Students who use AI more learn less. This is not driven by students who have adopted generative AI within five months and who don't use it much. The difference between these groups is even larger for the average effect 6-10 months after adoption (13 percentage points).

The total average (ATT) effect on girls is -7.3 percentage points compared to -8.8 percentage points for boys. The difference of 1.5 percentage points is small. Part of this difference is likely explained by the higher reported generative AI use among boys.

Finally, we split students into three groups by their September 2022 exam score tercile. Students in the highest tercile are most negatively affected, with an average treatment effect of 11 percentage points, compared to 7 percentage points for students in the lowest exam score tercile.

We also investigate heterogeneous effects by initial exam-score tercile for the entrance exams, which are the most important exams for student careers. For the 2025 Zhongkao, the "initial grades" are each student's June 2022 exam scores. For the 2025 Gaokao, they are each student's 2022 Zhongkao scores. We interact the treatment (post-AI) dummy with dummy variables for whether the initial exam scores are in the top, medium or bottom tertile. The results are shown in columns (3) and (6) of Table 2. For the Zhongkao, the top tertile is most negatively affected by generative AI. For the Gaokao, there is no significant difference in treatment effects across terciles.

5 Conclusion and discussion

We find that while generative AI significantly improves homework, it also causes large reductions in learning. Regular monthly exam scores, testing novel skills, fall by 12 percentage points (1.3 SD), while scores on entrance exams, testing general skills and long-term learning, fall by 3 percentage points (0.4 SD). The full negative effects are realized after six months for regular exam scores and after twenty months for entrance exams. This dynamic pattern explains the difference in average effects across the two exam types. Many students have recently adopted generative AI and the full negative effect on entrance exams had not been realized in June 2025.

The negative effects on learning are mostly driven by the majority of AI-using students who spend less time on homework than even the fastest non-AI student, receive high homework scores consistent with measures of generative AI capabilities, and very low exam scores – a pattern consistent with outsourcing the homework assignments to generative AI. Generative AI students who spend as much time studying as non-AI students receive very similar grades. Hence, generative AI seems mainly to influence the time spent learning, not learning efficiency.

The literature on the effects of generative AI on learning has largely focused on technical solutions on the *supply* side: how can we best design generative AI tools to facilitate scaffolding and hinder outsourcing? These designs have been evaluated in RCTs where the use of a specific tool in a specific way has been mandated.

Our findings highlight the importance of the *demand* side and student *incentives*. In China (and elsewhere), tutoring generative AI tools exist, but are not chosen. Instead, students outsource homework to generic LLMs that provide them with answers. Perhaps some young teenagers will choose to spend hours with a generative AI tool that avoids providing answers and instead gives hints and tries to lead them on the correct path, rather than one that just gives the right answer. But most students are unlikely to choose a tutoring tool that withholds answers when a generic LLM provides them instantly. How do we incentivize students to choose tutoring LLMs in an environment with general LLMs that will give direct answers? Given our results on time, how do we incentivize students to spend as much time studying as they did before generative AI? One step is to inform students of the negative learning effects of generative AI. This may help since several studies find that students who use generative AI are overconfident about their skills.

Acknowledgements. The authors have no sources of research support, financial relationships, or other potential conflicts of interest to disclose.

6 References

References

[1] Achiam, Josh, et al. "Gpt-4 technical report." arXiv preprint arXiv:2303.08774 (2023).

[2] Bastani, Hamsa, et al. "Generative AI without guardrails can harm learning: Evidence from high school mathematics." Proceedings of the National Academy of Sciences 122.26 (2025): e2422633122.

[3] Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond. "Generative AI at work." The Quarterly Journal of Economics 140.2 (2025): 889-942.

[4] Hou, Jinchang, et al. "E-eval: a comprehensive Chinese k-12 education evaluation benchmark for large language models." arXiv preprint arXiv:2401.15927 (2024).

[5] Callaway, Brantly, and Pedro HC Sant'Anna. "Difference-in-differences with multiple time periods." Journal of econometrics 225.2 (2021): 200-230.

[6] Kestin, Greg, et al. "AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting." Scientific Reports 15.1 (2025): 17458.

[7] Lo, Chung Kwan, Khe Foon Hew, and Morris Siu-yung Jong. "The influence of ChatGPT on student engagement: A systematic review and future research agenda." Computers & Education 219 (2024): 105100.

[8] Noy, Shakked, and Whitney Zhang. "Experimental evidence on the productivity effects of generative artificial intelligence." Science 381.6654 (2023): 187-192.

[9] OECD. *Digital Education Outlook 2026*. Paris: OECD Publishing, 2026.

[10] Zhang, X., Li, C., Zong, Y., Ying, Z., He, L., & Qiu, X. (2023). Evaluating the performance of large language models on gaokao benchmark. arXiv preprint arXiv:2305.12474.

[11] Wang, Jin, and Wenxiang Fan. "The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis." Humanities and Social Sciences Communications 12.1 (2025): 1-21.

7 Appendix

7.1 Mean Plots

In the mean (center) plots in Figure 2, we plot the mean outcomes of ever-AI students who start using generative AI in our sample period in an interval of 10 months before and after the month students start using generative AI. Let s be the month in which a student started using generative AI and m be the current month. We then define the lag $l = m - s$ which is the number of months since the student started using AI. We plot

$$\overline{y}_l^T = \frac{1}{n_l^T} \sum_i y_{i(s+l)}^s,$$

where n_l^T are the number of treated students observed with lag l or equivalently, observed in month $m = s + l$.

As comparison, we also plot the corresponding mean for never-AI students who have not yet started using generative AI in June 2025. For each treated student, we compute a representative matched control outcome, $\overline{y}_{i(s+l)}^C$ which is the mean outcome in the control group in the same month, grade and school. We then compute the average control outcome for each lag,

$$\overline{y}_l^C = \frac{1}{n_l^T} \sum_i \overline{y}_{i(s+l)}^C.$$

7.2 Empirical specification

We define treatment by the month in which a student first adopts generative AI. Students who adopt by the end of the sample period, June 2025, are *ever-adopters*, while students who have not adopted by then are *never-adopters*. We refer to students who have adopted generative AI by a given month as *AI students* in that month, and to students who have not yet adopted by that month as *non-AI students*. Thus, the classification of students as ever-adopters or never-adopters is fixed over the sample period, while the classification of students as AI or non-AI students varies over time.

Our object of interest is the effect of entering the post-adoption state. Specifically, we estimate the cohort- and event-time-specific average treatment effect on the treated: the effect, at a given event time, of having adopted generative AI for students who first adopt in a given month, relative to the outcome path they would have followed had they not adopted. We proxy this counterfactual outcome path using never-adopters.

Let i index students and t months. Let $G_i \in \{1, \ldots, T, \infty\}$ denote student i's adoption month, where $G_i = \infty$ for never-adopters. A student is treated in month t if $t \geq G_i$. For each adoption cohort g and calendar month t, we define the group-time average treatment effect as

$$ATT(g, t) = \mathbb{E}\left[Y_{it}(g) - Y_{it}(\infty) \mid G_i = g\right],$$

where $Y_{it}(g)$ denotes the potential outcome in month t if student i first adopted in

month g, and $Y_{it}(\infty)$ denotes the potential outcome in month t if the student never adopted during the sample period.

We estimate these effects using the staggered difference-in-differences estimator of Callaway and Sant'Anna (2021). For students who first adopt in month g, the estimator compares the change in outcomes between month $g-1$ and month t to the corresponding change for never-adopters:

$$ATT(g,t) = \mathbb{E}[Y_{it} - Y_{i,g-1} \mid G_i = g] - \mathbb{E}[Y_{it} - Y_{i,g-1} \mid G_i = \infty].$$

These cohort-specific effects are then aggregated across adoption cohorts and event times to obtain the event-study estimates reported below.

This framework is well suited to our setting because it allows treatment effects to vary across adoption cohorts. Such heterogeneity is plausible here: generative AI systems improved rapidly over the sample period, so students who adopted earlier may have faced substantially different tools than those who adopted later. The effect of adoption may therefore differ by cohort.

The key identifying assumption is a parallel-trends assumption: absent adoption, ever-adopters and never-adopters would have experienced the same evolution in outcomes. Formally, for each adoption cohort g and post-treatment month $t \geq g$,

$$\mathbb{E}\left[Y_{it}(\infty) - Y_{i,g-1}(\infty) \mid G_i = g\right] = \mathbb{E}\left[Y_{it}(\infty) - Y_{i,g-1}(\infty) \mid G_i = \infty\right].$$

Under this assumption, differences in outcome changes between adopters and never-adopters can be interpreted as causal effects of adoption. Standard errors are clustered at the class level.

7.3 Robustness

7.3.1 Parallel trends

The main identification assumption is that of parallel trends. This could be violated, for example, if some students that lose interest in school or for other reasons have declining grades, start using generative AI to complete their homework. We cannot test this directly, but the ever- and never-AI students have parallel trends in the pre-period, and even the mean levels of the outcomes are very similar. So there is no evidence that these students lose interest or have declining grades in the month before they start using generative AI. It could still be that there is a shock to student motivation in the same month. However, the effects are so dramatic and affect a majority of students. It seems highly unlikely that 80 percent of students receive huge unobserved motivation or other exam score shocks in the month they start using generative AI.

7.3.2 Reported generative AI adoption month

Our main treatment variable, the month that each student started using generative AI, is self-reported and retrospective. Teachers instructed students to verify registration dates, but we cannot observe whether they actually did so. Under some assumption on

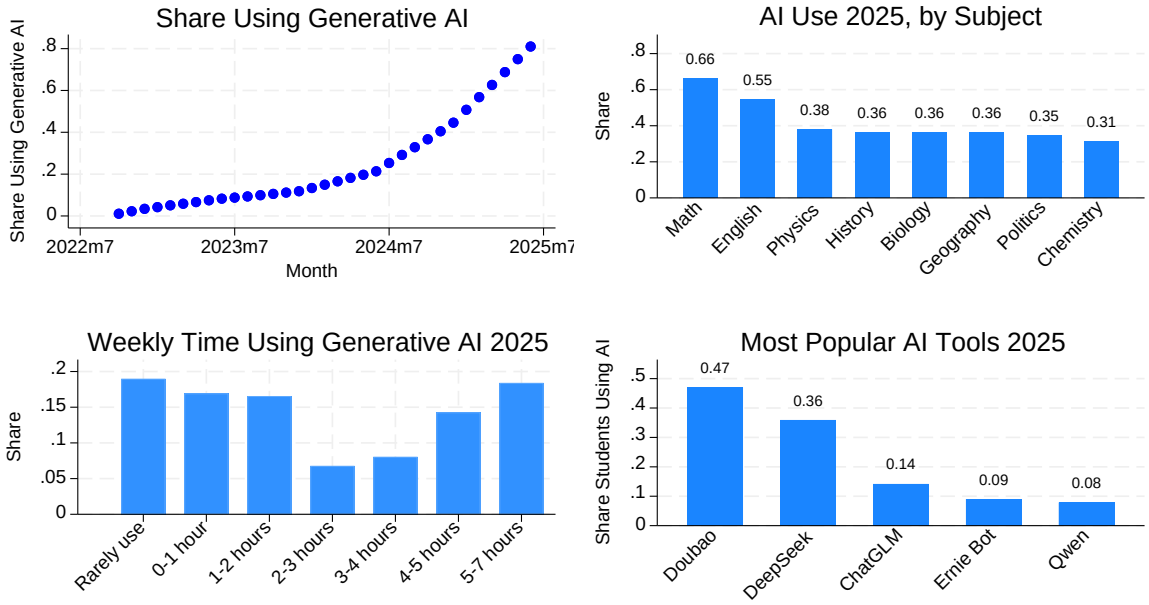
the type of error, we argue that the error is likely to be small. This assumption is that if students make mistakes, then these are sometimes positive and sometimes negative.

We can rule out that students make positive mistakes, reporting an adoption month that is later than their actual adoption month. The reason is that this would result in a pre-trend, where outcomes diverge in the months before the reported AI start month. There is clearly no pre-trend in any of the outcomes, so we can rule out that any significant share of students made this type of error.

It could be that all reporting mistakes are negative, so that misreporting students always report a too early starting date. For example, it could be that students register an account at a generative AI service, but only start using it actively with some lag. This would be quite close to our previous interpretation that it takes some time for students to learn how to use their AI tools for homework. It would appear in our data as a slower effect. We measure the effect on our monthly outcomes as fully developed after six months and part of the delay could be caused by this type of reporting error. Note that this type of error would increase the delay time of the full effect, but not impact the estimated size of the fully developed (after six months) effect. This type of error would also not affect the result that the general skills tested in the entrance exams deteriorate more slowly than those in the regular exams (see below).

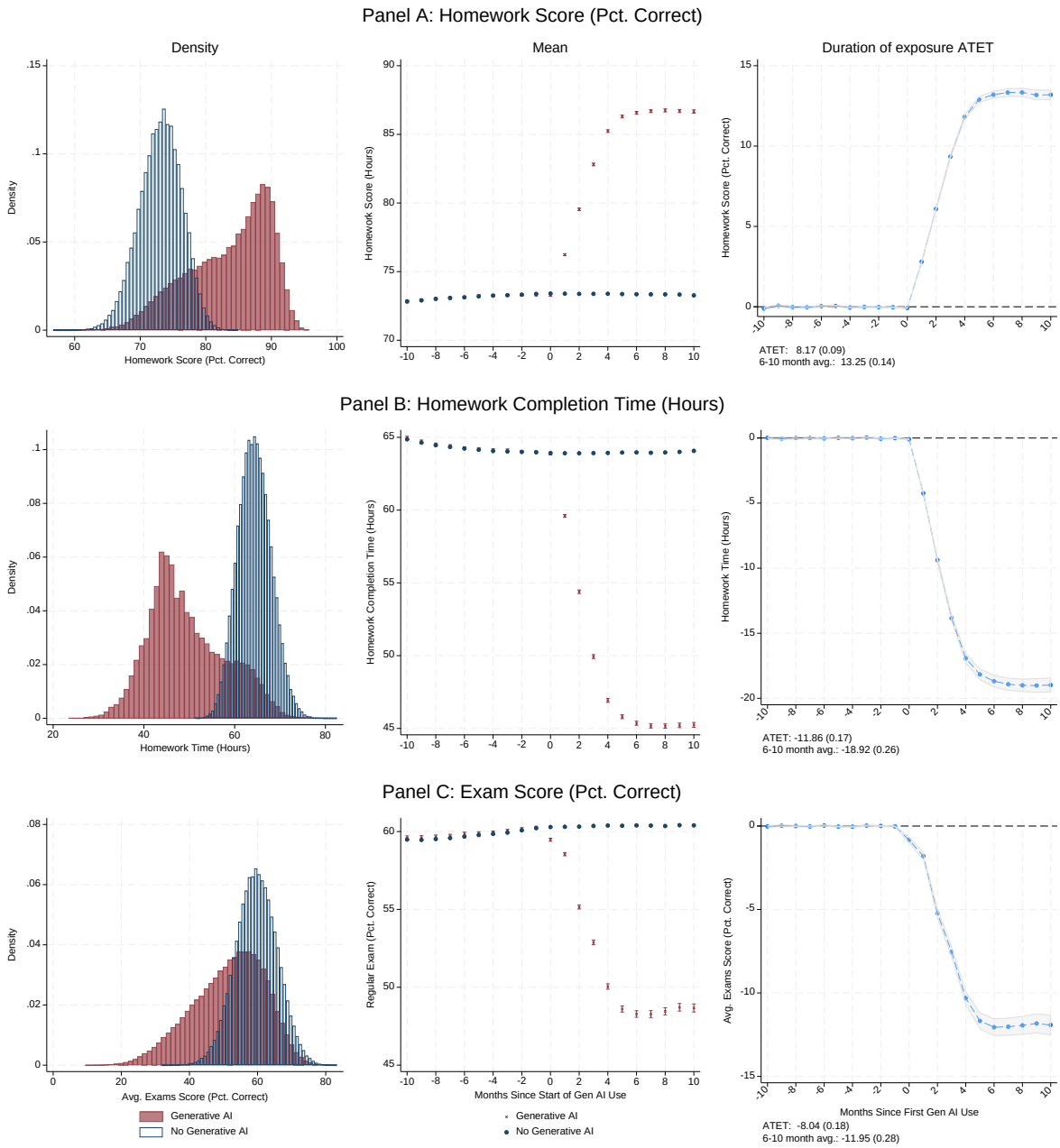
Given that there is no pre-trend, indicating no significant positive misreporting errors, we believe that the reported starting dates are most likely quite accurate. However, we cannot fully rule out the existence of negative reporting errors and hence are a bit more guarded regarding the interpretation of the six months to full effect result.

Figure 1: Survey of Generative AI Adoption



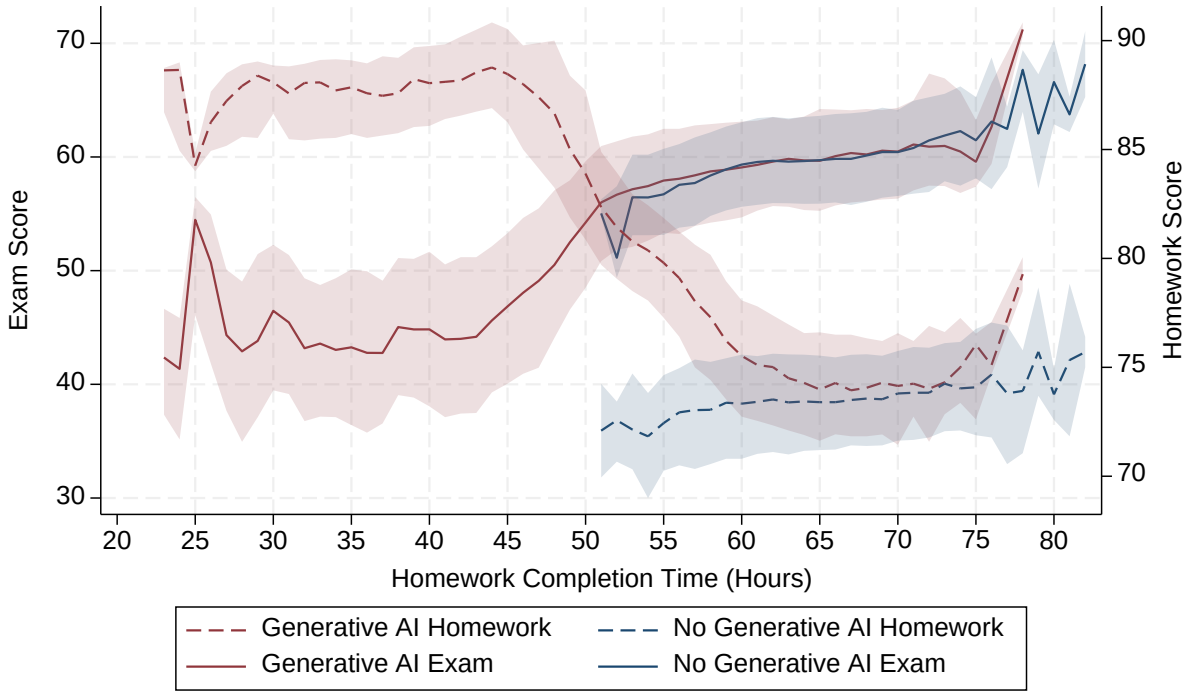
Note: The figures are based on responses to our survey of 26,811 students in grades 7-12. The first plot the share of students who adopted generative AI by each month. The following figures present the share of students in June 2025 who report using generative AI for different subjects, how many hours they use generative AI per week, and what tools are most used.

Figure 2. Main results



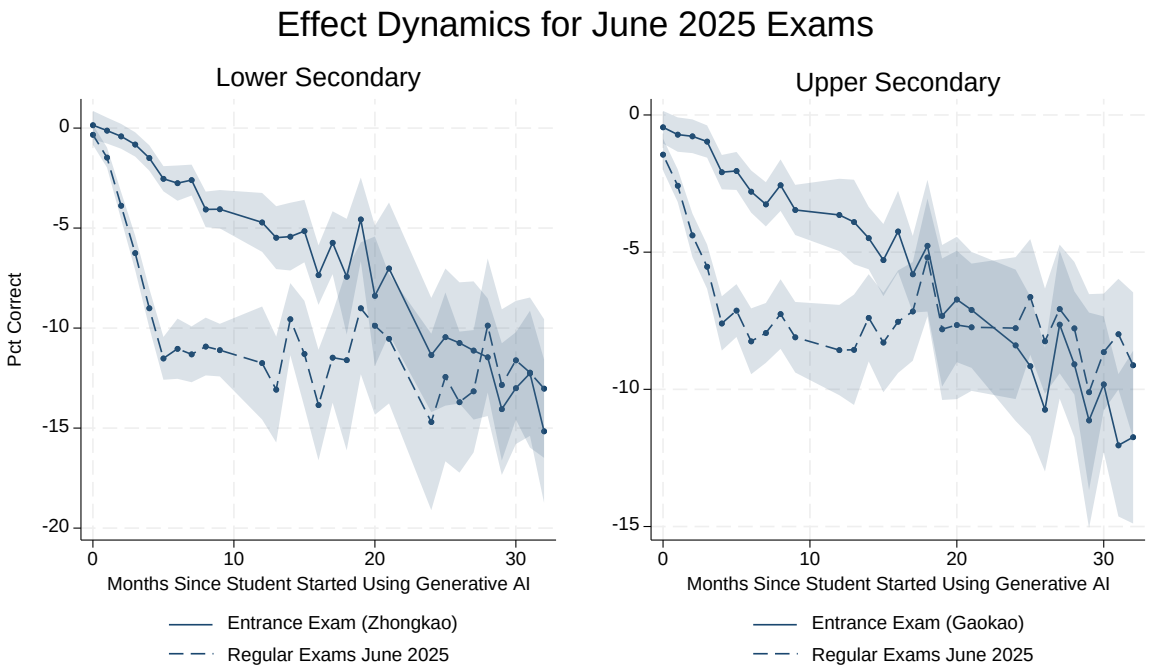
Note: The figure shows our main results. The data is at the student-by-month level and the sample summary statistics are presented in panels A and B of Table 1. The histograms to the left plot the outcome variables in the panel headings separately for students who have started using generative AI and those who have not. The middle graphs plot the mean outcome variables and 95 percent confidence bands for the same two groups. The graphs to the right plot the average treatment effects by event time (and their 95% confidence intervals) of generative AI adoption on the outcome variables listed in the panel headings, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator. The total ATT and the average effects 6–10 months post-treatment are reported below the corresponding graphs. Standard errors are clustered by class id.

Figure 3. Homework and exam scores (medians and interquartile ranges)
by homework completion time



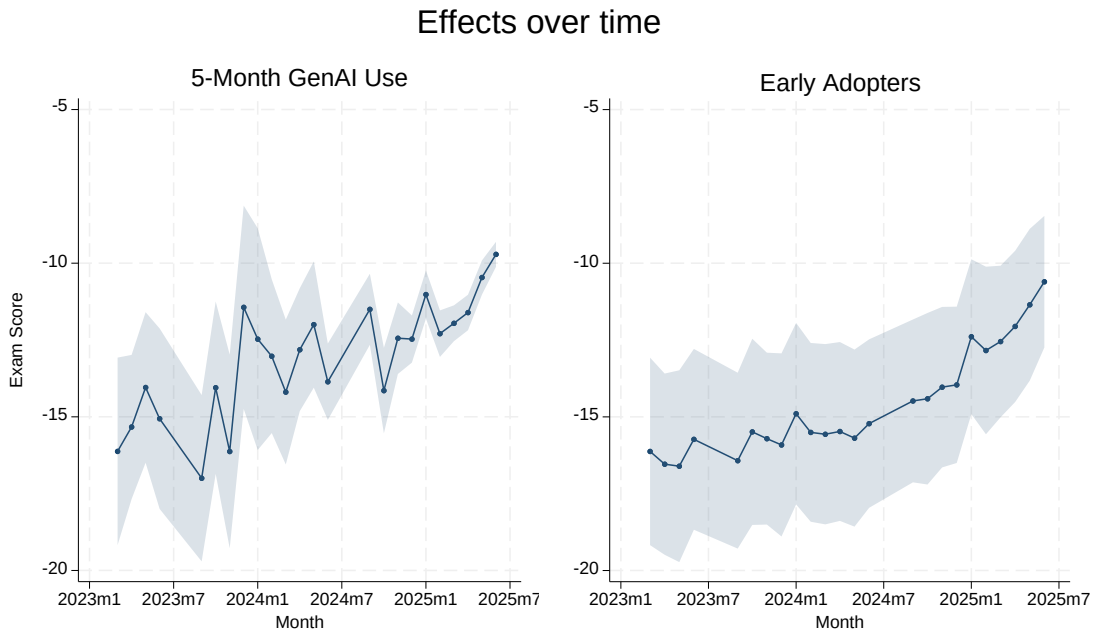
Note: The figure plots the median and interquartile ranges of homework and exam scores by homework completion time, separately for AI students and non-AI students. The data is at the student-month level and the sample is the one shown in Panel A of Table 1 with non-missing observations for homework time.

Figure 4. Learning deterioration for regular and entrance exams



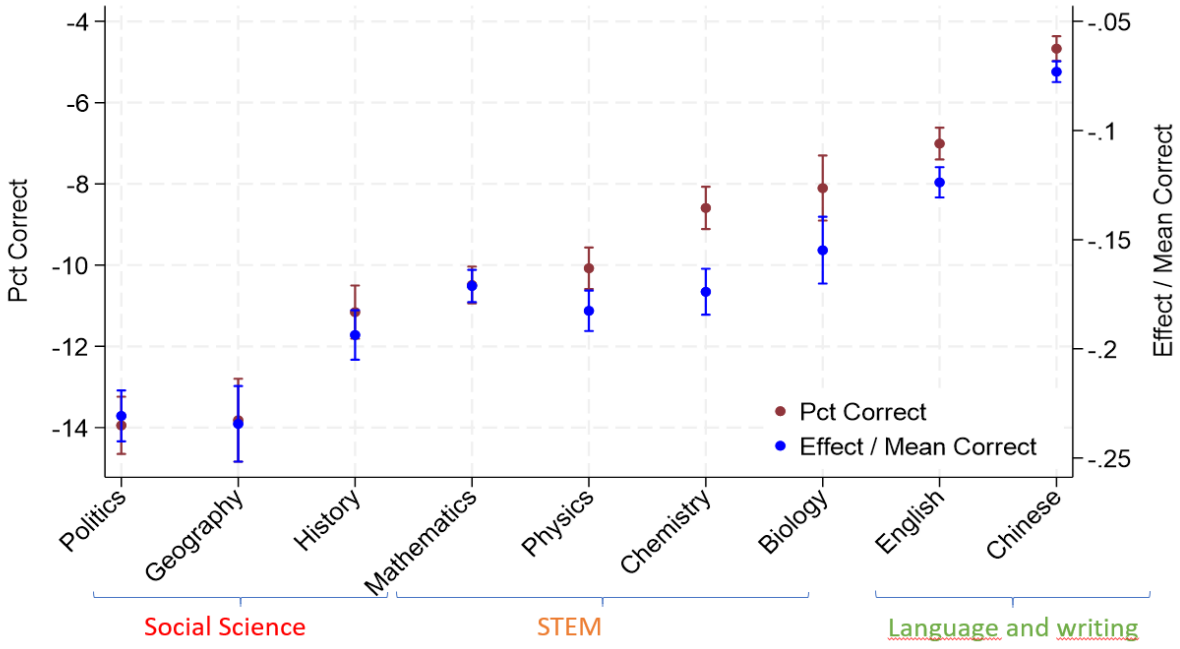
Note: The figure plots the average treatment effects by event time (and their 95% confidence intervals) of generative AI adoption on June 2025 regular exam scores and entrance exam scores, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator. The sample consists of students in grade 9 and 12 in June 2025. For the entrance exams, differences are computed based on regular exam scores the month before generative AI adoption. OLS regressions of generative AI adoption on entrance exam scores, controlling for class id and our demographic variables, produce very similar results; see Figure A10.

Figure 5. Effects over time



Note: The graph to the left plots the estimated average treatment effects (and 95% confidence intervals) of having adopted generative AI for exactly five months on regular exam scores, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator, by month. In this graph, the estimation sample in each month consists of students who started using generative AI five months ago. The graph to the right plots the estimated average treatment effects (and 95% confidence intervals) of having adopted generative AI on regular exam scores, for the cohort that started using generative AI in October 2022, by month. In this graph, the sample of students is the same and the length of exposure is increasing over time. Standard errors are clustered by class id.

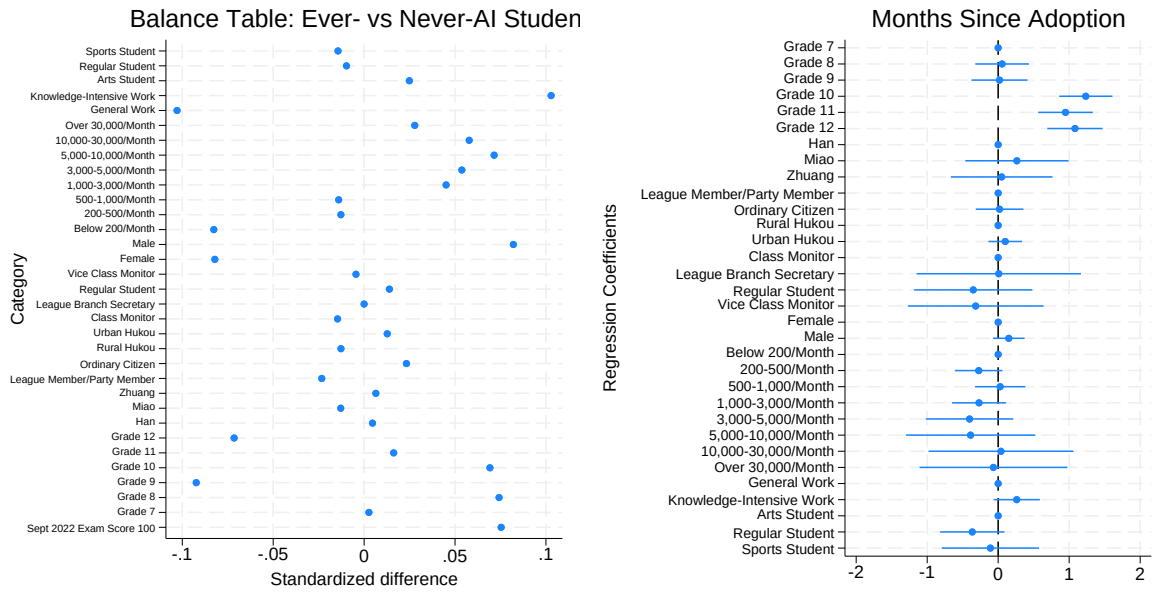
Figure 6. Effects by subject



Note: The figure plot the plots, by subject, the total average treatment effect and 95% confidence intervals of having adopted generative AI on regular exam scores, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator. Results are presented both in terms of percent correct answers (left axis) and relative to the mean share correct answers (right axis). The data is at the student-by-month-by subject level. Standard errors are clustered by class id.

Appendix

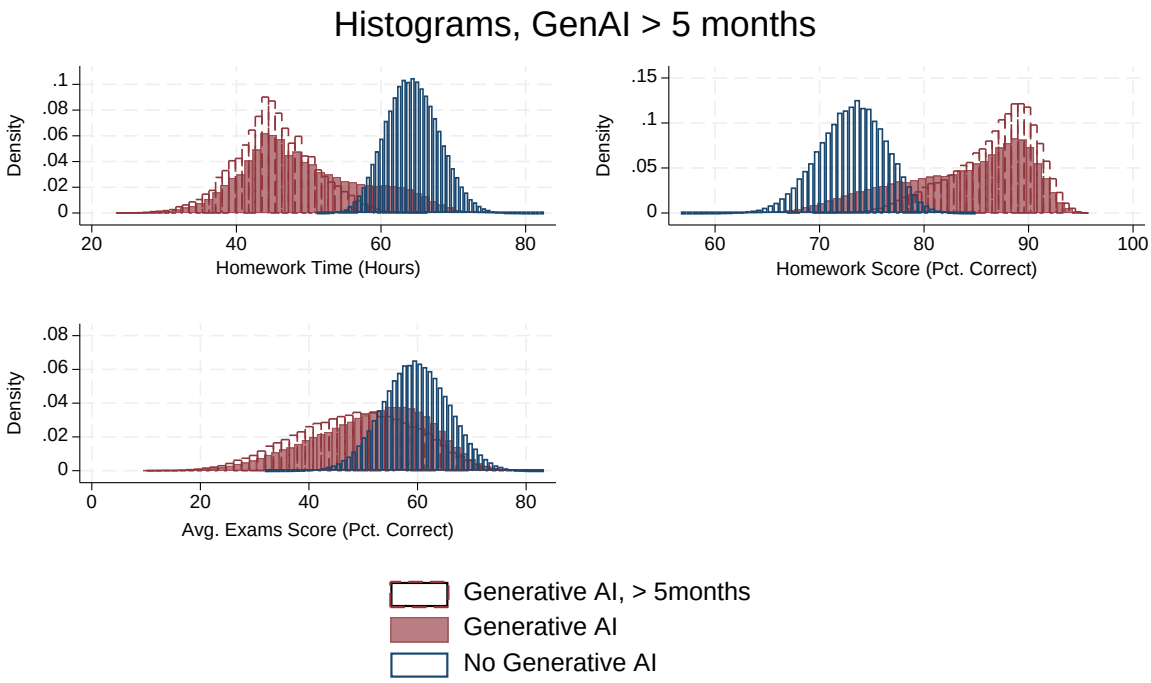
Figure A1a: Selection



Pre-scores base on 8,466 students with Sept 2022 grade data.

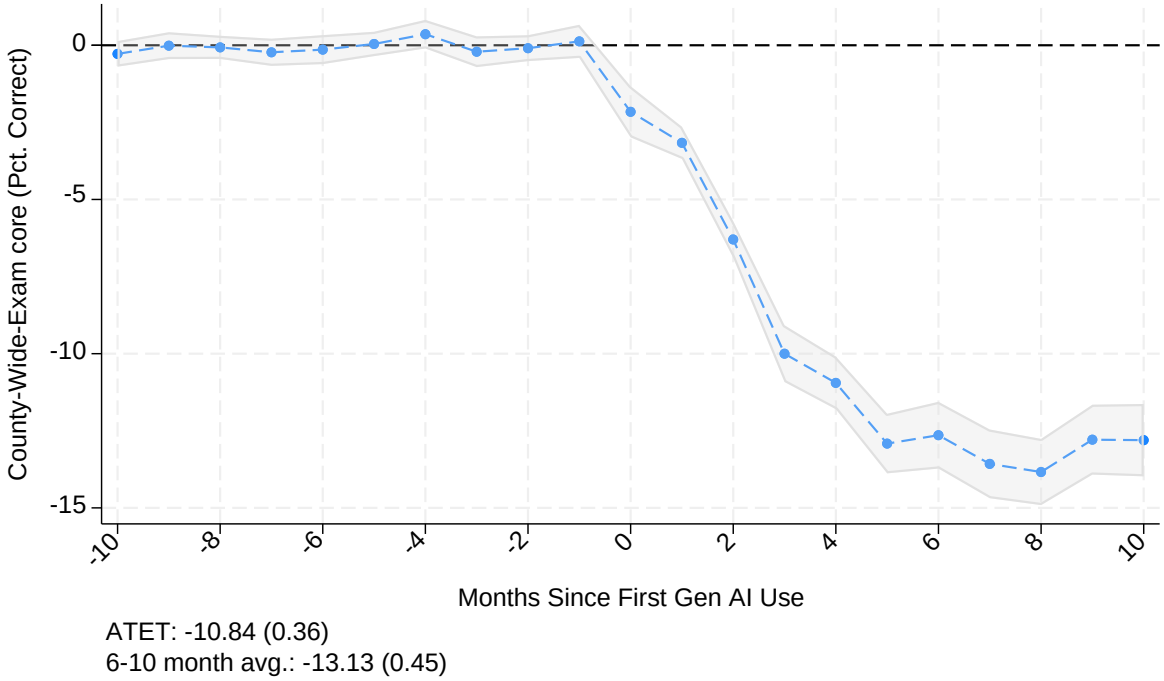
Note: The graph to the left shows the standardized difference in means between students who have started using generative AI in June 2025 (Ever-AI Students) and those who have not (Never-AI Students). The data is at the student level and includes all 26,811 students, except for the last entry, which is based on the 8,466 students who are in 12th or 9th grade in 2025 and for whom we have exam scores for September 2022. The graph to the right shows the coefficient estimates from an OLS regression of the number of months since the student adopted generative AI on our demographic variables. The sample for this regression is students who had adopted generative AI by June 2025.

Figure A2. Histogram including >5 months GAI use



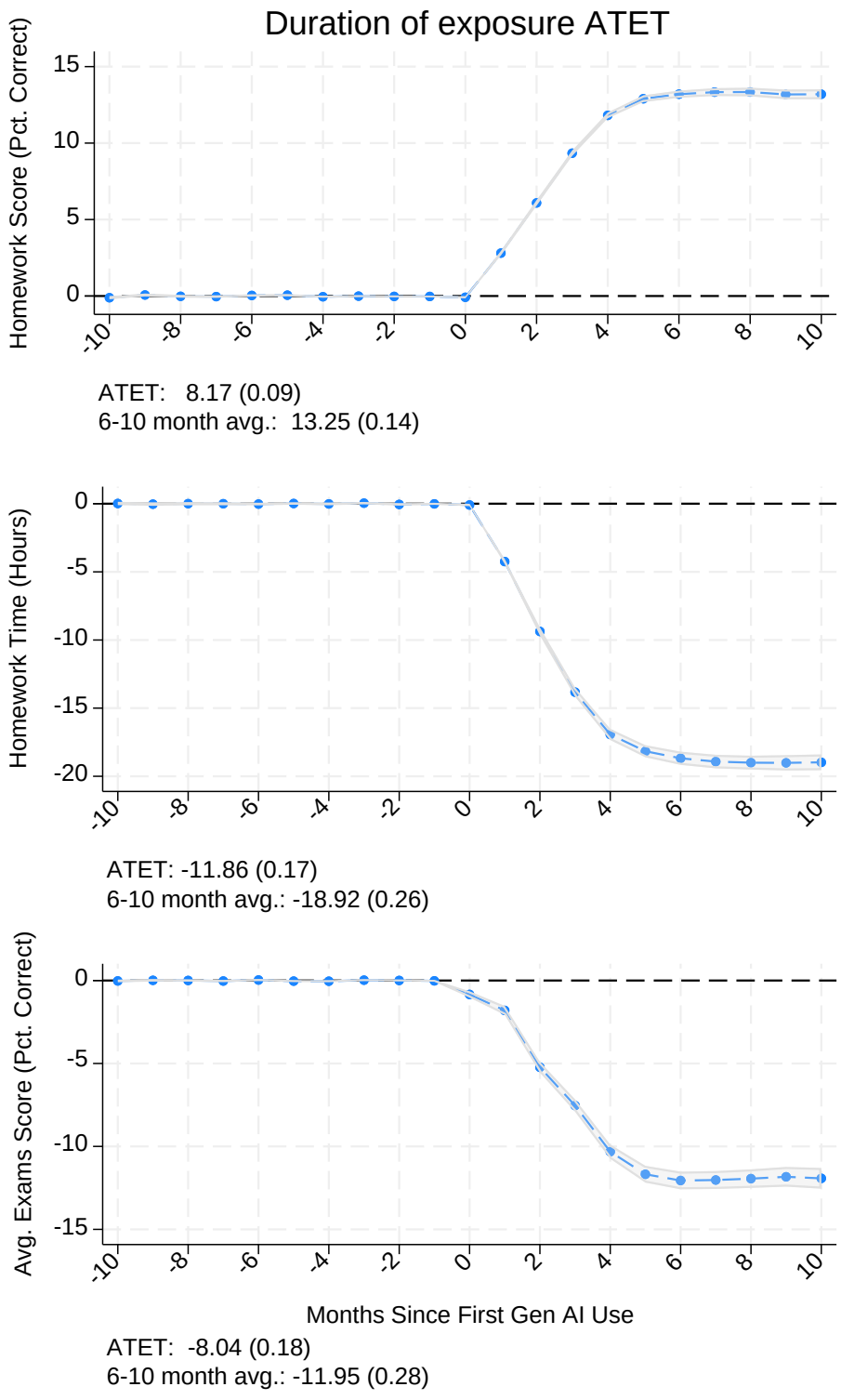
Note: The figure plots the distributions of homework score, homework completion time and regular exam scores, for three groups of students: (i) those who have adopted generative AI, (ii) those who adopted generative AI more than 5 months ago, and (iii) those who have not adopted generative AI. The data is at the student-by-month level and the samples are those summarized in Table 1 for the homework and regular exam data.

Figure A3. County-wide exams effects by event time



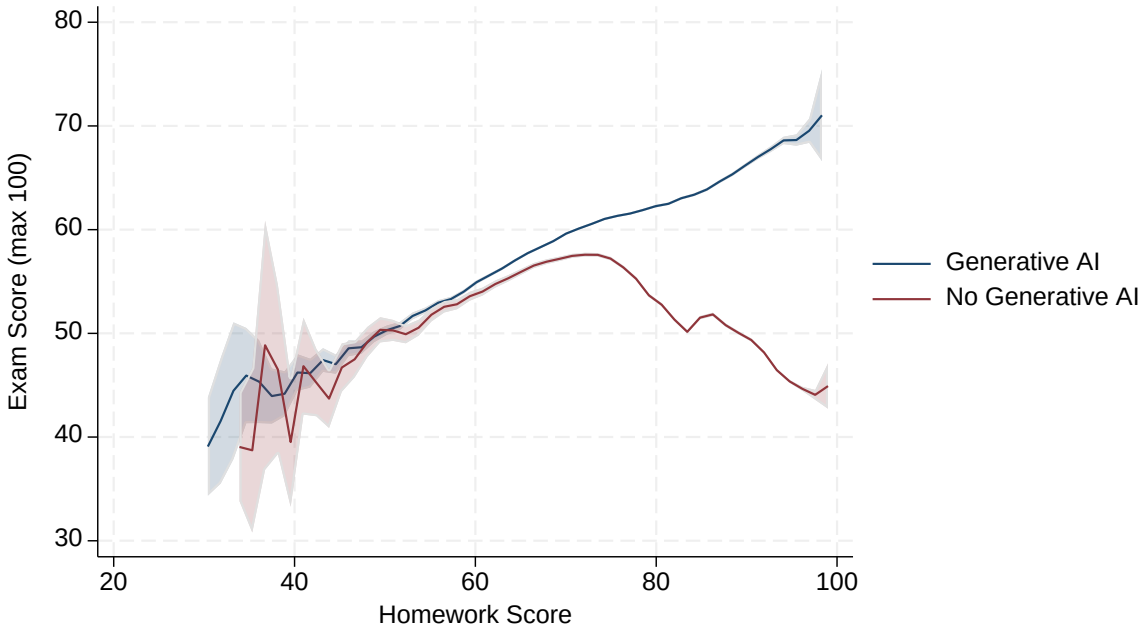
Note: The figure plots the average treatment effects by event time (and their 95% confidence intervals) of generative AI adoption on the county-wide exam score, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator. The total ATT and the average effects 6–10 months post-treatment is reported below the graph. The data is at the student by month level and the sample summary statistics are presented in Panel C of Table 1. Standard errors are clustered by class id.

Figure A4. Not yet treated included in control



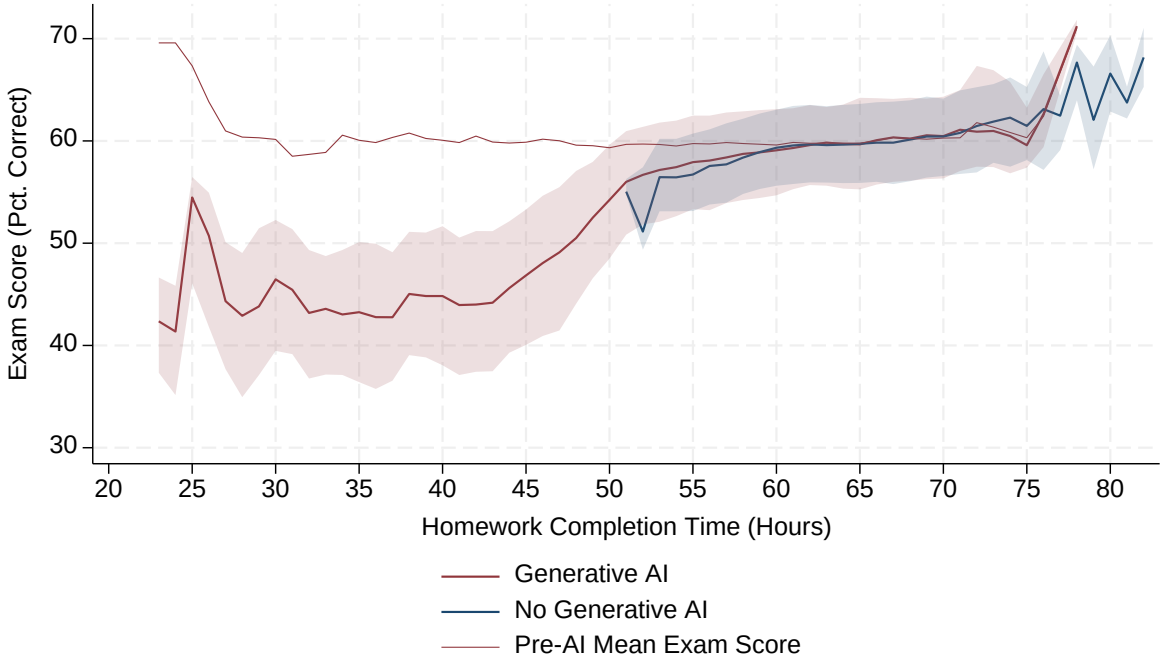
Note: The figure plots the average treatment effects by event time (and their 95% confidence intervals) of generative AI adoption on the outcome variables listed in the panel headings, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator and with not yet treated students included in the control group. The data is at the student-by-month level and the sample summary statistics are presented in panels A and B of Table 1. The total ATT and the average effects 6–10 months post treatment are reported below the corresponding graphs. Standard errors are clustered by class id.

Figure A5. Mean Exam Score by Homework Score



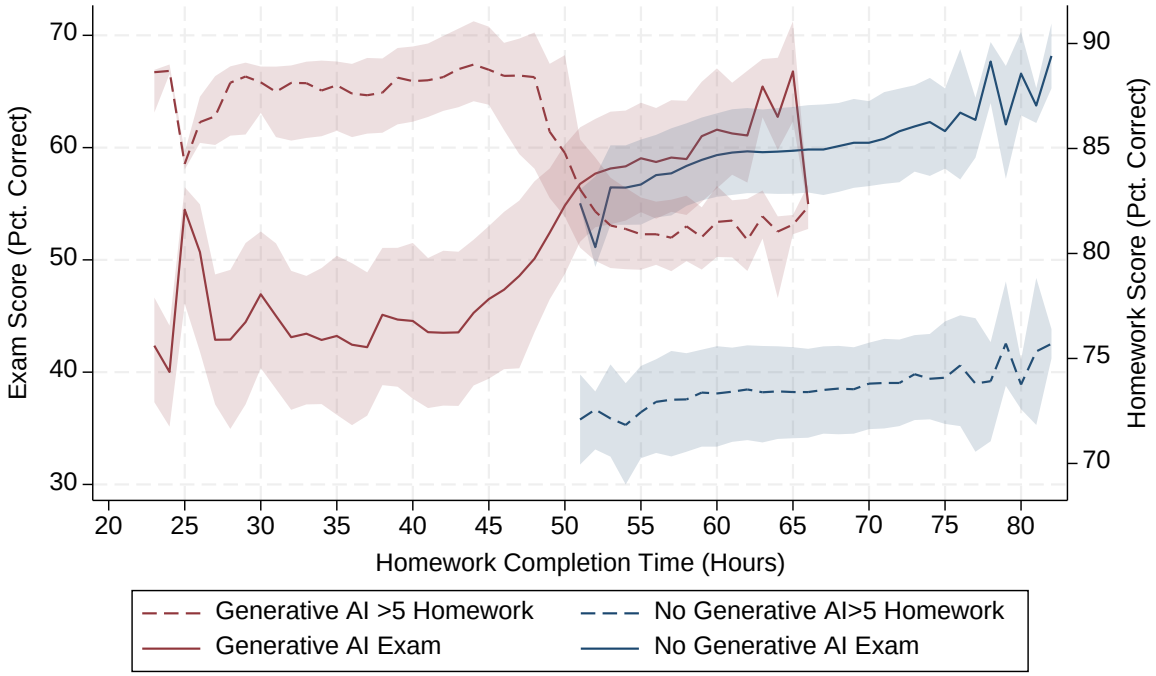
Note: The figure plots the smoothed mean and its 95% confidence intervals, by homework score, separately for students who have adopted generative AI and those who have not. The data is at the student-month level and the sample summary statistics are shown in Panel A of Table 1.

Figure A6. Median and interquartile range of exam scores by homework completion time.



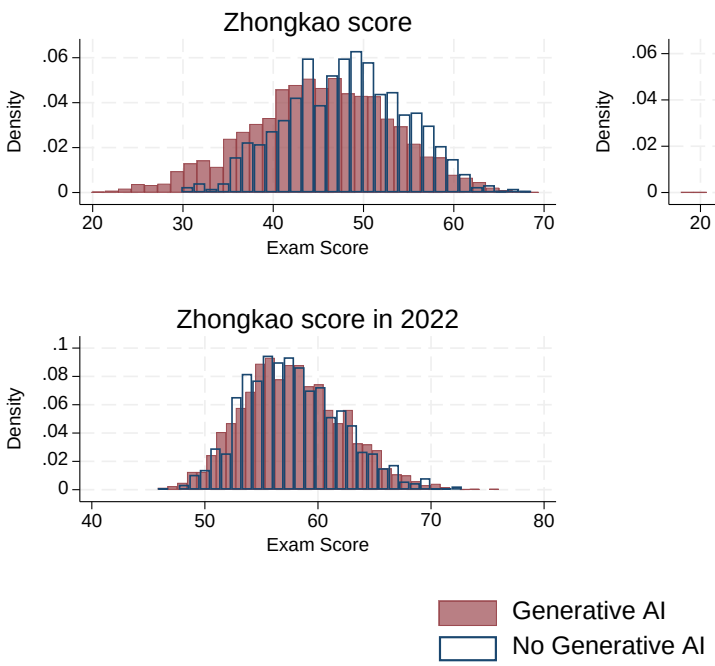
Note: The figure plots the median and interquartile ranges of exam scores by homework completion time, separately for AI students and non-AI students. It also plots the mean exam score for the AI students before the month they start using generative AI. The data are at the student-month level and the sample is that shown in Panel A of Table 1 with non-missing observations for homework time. There is little evidence of selection, except at the very low and high end, where students are positively selected.

Figure A7. Homework Completion Time and Score, > 5 months Gen AI Use



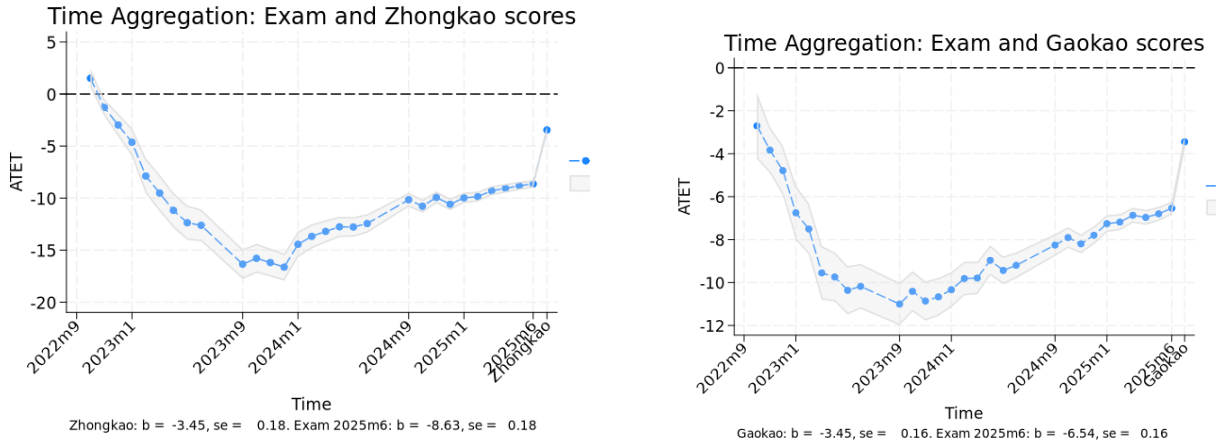
Note: The figure plots the median and interquartile ranges of exam score by homework completion time, separately for students who have adopted generative AI more than five months ago. The data is at the student-month level and the sample is that shown in Panel A of Table 1 with non-missing observations for homework time.

Figure A8. Entrance exam histograms



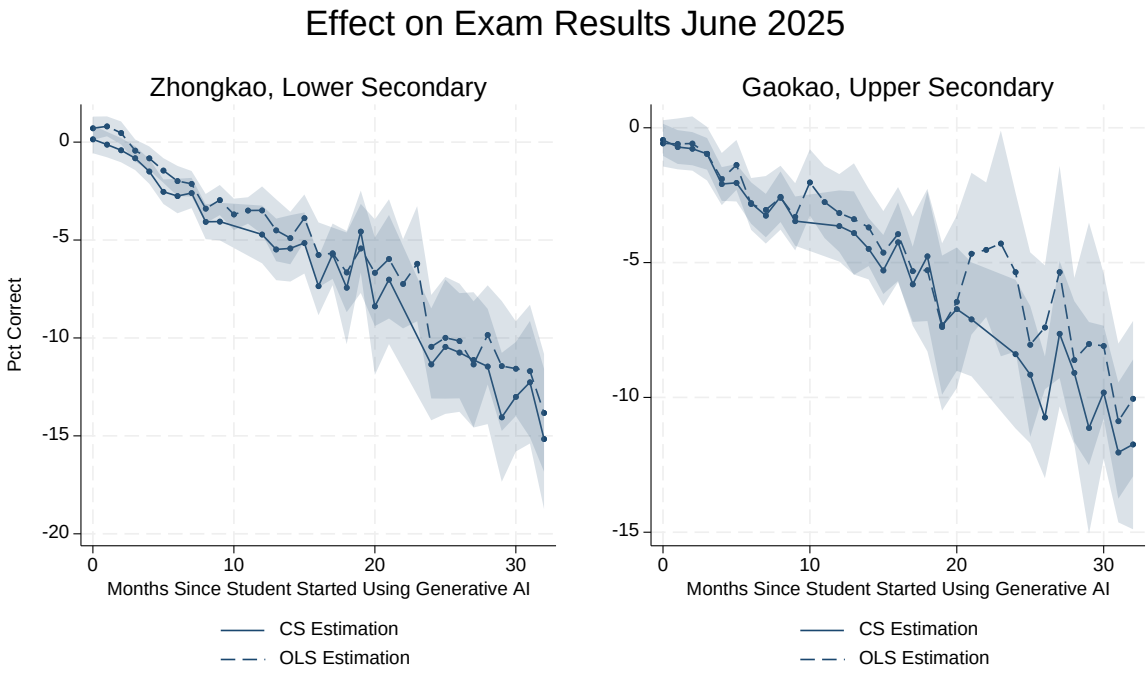
Note: The figure plots the distribution of entrance exam scores, separately for AI students and non-AI students. The Zhongkao score in 2022 is included to show that AI and non-AI students are not selected based on pre-AI entrance exam scores. The data is at the student level and the sample summary statistics are presented in Panel D of Table 1.

Figure A9. Time aggregation: Regular and entrance exams



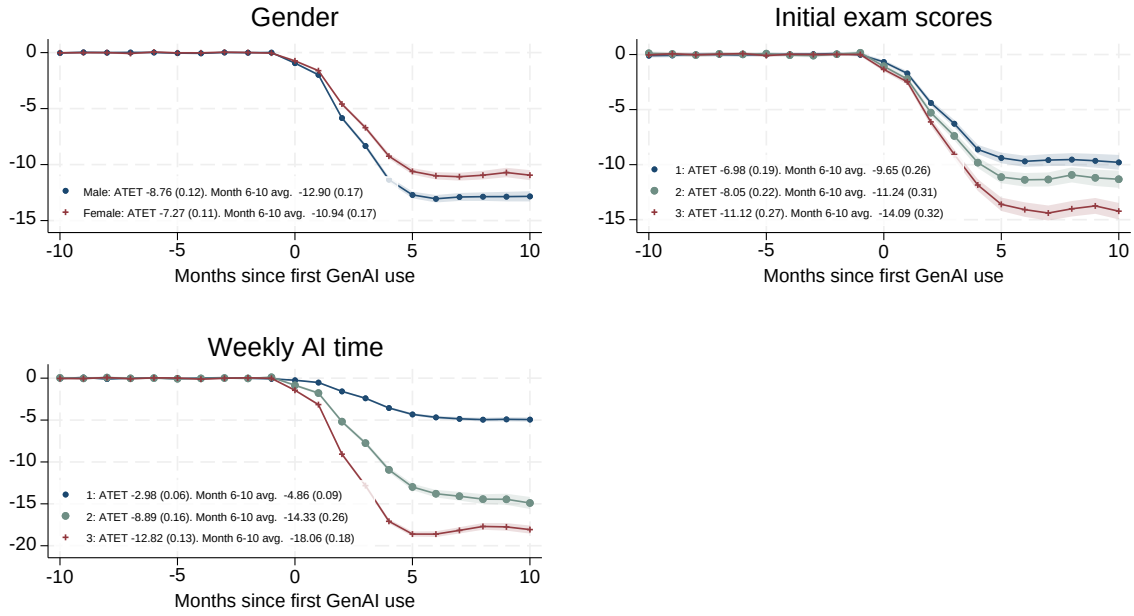
Note: The figure plots the estimated average treatment effects on the treated and 95% confidence intervals of generative AI adoption on regular exam scores and entrance exam scores in June 2025, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator. The total ATT and the average effects 6–10 months post treatment are reported below the corresponding graphs. The data is at the student-by-month level and the sample summary statistics are presented in panels A and B of Table 1. Standard errors are clustered by class id.

Figure A10. OLS and CS estimate comparison



Note. The figure shows the the estimated average treatment effects and the 95% confidence intervals by event time computed using the Callaway and Sant'Anna estimator (CS Estimation) and using an OLS regression, controlling for class id and our demographic controls.

Figure A11. Heterogeneous effects by student characteristics



Note: The figure plots the estimated average treatment effects on the treated (and 95% confidence intervals) of generative AI adoption on regular exam scores, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator, aggregated by event time for samples split by student characteristics. Standard errors are clustered by class id.

Tables

Table 1. Summary statistics

Variable	Obs.	Mean	SD	Min	Max
Panel A: Homework Data - 18,345 students, 20 months					
Homework Score (Pct. Correct)	272 140	77.52	7.01	56.72	95.71
Homework Time (Hours)	240 203	58.09	9.59	23.39	82.55
AI Student	272 140	0.42	0.49	0.00	1.00
Panel B: Regular Exam Data - 26,811 students, 30 months					
Regular Exams Score (Pct. Correct)	526 120	56.91	8.75	9.47	83.31
AI Student	526 120	0.33	0.47	0.00	1.00
Panel C: County-Wide Exam Data - 8,466 students, 18 months					
County-Wide-Exam Score (Pct. Correct)	89 059	55.64	9.73	5.43	85.74
AI Student	89 059	0.23	0.42	0.00	1.00
Panel D: Entrance Exam Data - 8,341 students, 1 month					
Zhongkao/Gaokao score (Pct. Correct)	8 341	50.11	8.91	17.22	81.17
AI Student	8 341	0.78	0.41	0.00	1.00

Notes: This table shows summary statistics for the samples used in our analysis.

Table 2. Entrance Exam Regressions

VARIABLES	(1) Zhongkao	(2) Zhongkao	(3) Zhongkao	(4) Gaokao	(5) Gaokao	(6) Gaokao	(7) Zhongkao 2022
AI Student	-2.946***	-2.525***	-2.240***	-2.498***	-3.002***	-3.018***	-0.013
	(0.593)	(0.184)	(0.218)	(0.385)	(0.336)	(0.387)	(0.207)
AI Student * Medium Initial Exam Score			0.355			0.112	
			(0.381)			(0.297)	
AI Student * High Initial Exam Score			-1.282***			-0.288	
			(0.382)			(0.460)	
Pre AI Exam Score		0.743***	0.703***		0.821***	0.338***	
		(0.018)	(0.033)		(0.033)	(0.037)	
Observations	4,034	4,034	4,034	4,307	4,260	4,260	4,260
R-squared	0.025	0.563	0.565	0.016	0.305	0.493	0.000
Mean Dep Var	45.864	45.864	45.864	54.080	54.080	54.080	57.798
SD Dep Var	7.766	7.766	7.766	8.035	8.035	8.035	4.436
Controls	No	Demographics and Pre-Score	Demographics and Pre-Score	No	Demographics and Pre-Score	Demographics and Pre-Score	No
Fixed Effects	No	Class ID	Class ID	No	Class ID	Class ID	No

Notes. This table presents the results of OLS regressions estimating the effect of student generative AI adoption on the percent correct answers in entrance exams (Zhongkao for high school entrance and Gaokao for university entrance). Columns (1) and (4) include no controls. Columns (2) and (5) control for pre-AI exam scores (September 2022 exam scores for the Zhongkao, June 2022 Zhongkao scores for the Gaokao), class fixed effects and fixed effects for our demographic variables (gender, ethnicity, political status, urban/rural, class cadre, monthly family income group, parents' occupation and special student). Columns (3) and (6) interact the AI Student dummy variable with the students' tercile in the distribution of pre-AI exam scores. Column (7) is a placebo replacing the dependent variable in the Gaokao regression with the June 2022 Zhongkao score. Standard errors clustered by class are in parentheses. $*p < .10$, $**p < .05$, $***p < .01$.

Table A1. Balance test

Variable	Control Mean	Control Std. Dev.	Treated Mean	Treated Std. Dev.	Standardized Difference
Sept 2022 Exam Score 100	58.743	5.870	59.222	6.805	0.075
Grade 7	0.174	0.379	0.175	0.380	0.003
Grade 8	0.144	0.351	0.171	0.376	0.074
Grade 9	0.179	0.384	0.145	0.352	-0.092
Grade 10	0.157	0.364	0.183	0.387	0.069
Grade 11	0.160	0.366	0.166	0.372	0.016
Grade 12	0.186	0.389	0.159	0.366	-0.071
Han	0.950	0.218	0.951	0.215	0.005
Miao	0.026	0.158	0.024	0.153	-0.013
Zhuang	0.024	0.154	0.025	0.155	0.006
League Member/Party Member	0.141	0.348	0.133	0.339	-0.023
Ordinary Citizen	0.859	0.348	0.867	0.339	0.023
Rural Hukou	0.336	0.473	0.330	0.470	-0.013
Urban Hukou	0.664	0.473	0.670	0.470	0.013
Class Monitor	0.020	0.141	0.018	0.133	-0.015
League Branch Secretary	0.019	0.137	0.019	0.135	0.000
Regular Student	0.907	0.290	0.911	0.285	0.014
Vice Class Monitor	0.053	0.224	0.052	0.223	-0.004
Female	0.531	0.499	0.490	0.500	-0.082
Male	0.469	0.499	0.510	0.500	0.082
Below 200/Month	0.200	0.400	0.168	0.374	-0.083
200-500/Month	0.333	0.471	0.327	0.469	-0.013
500-1,000/Month	0.246	0.431	0.240	0.427	-0.014
1,000-3,000/Month	0.163	0.370	0.180	0.384	0.045
3,000-5,000/Month	0.031	0.172	0.041	0.199	0.054
5,000-10,000/Month	0.009	0.093	0.017	0.128	0.072
10,000-30,000/Month	0.008	0.090	0.014	0.116	0.058
Over 30,000/Month	0.010	0.101	0.013	0.114	0.028
General Work	0.875	0.331	0.839	0.368	-0.103
Knowledge-Intensive Work	0.125	0.331	0.161	0.368	0.103
Arts Student	0.059	0.236	0.065	0.246	0.025
Regular Student	0.894	0.308	0.891	0.311	-0.010
Sports Student	0.047	0.212	0.044	0.206	-0.014

Note: The table shows the standardized differences in means between treated students who had started using generative AI by June 2025 and control students who had not, in our main sample of 26,811 students. The first entry is based on the 8,466 students who are in 12^{th} or 9^{th} grade in 2025 and for whom we have exam scores for September 2022.