DISCUSSION PAPER SERIES

DP21577
(v. 2)

THE GENERATIVE AI LEARNING PENALTY: EVIDENCE FROM CHINESE SECONDARY EDUCATION

David Strömberg, Victor Lei and Yanhui Wu

DEVELOPMENT ECONOMICS

ISSN 0265-8003

THE GENERATIVE AI LEARNING PENALTY: EVIDENCE FROM CHINESE SECONDARY EDUCATION

David Strömberg, Victor Lei and Yanhui Wu

Discussion Paper DP21577
First Published 02 June 2026
This Revision 02 June 2026

Centre for Economic Policy Research
187 boulevard Saint-Germain, 75007 Paris, France
2 Coldbath Square, London EC1R 5HL
Tel: +44 (0)20 7183 8801
www.cepr.org

THE GENERATIVE AI LEARNING PENALTY: EVIDENCE FROM CHINESE SECONDARY EDUCATION

Abstract

Using 30 months of panel data on 26,811 Chinese students in grades 7--12, we study how generative AI affects homework productivity and learning. The data combine monthly closed-book exams, high-school and college entrance exams, and homework scores and completion time across nine subjects. We exploit staggered AI adoption in a difference-in-differences design. AI adoption raises homework scores by 18% and reduces completion time by 30%, but lowers monthly exam scores by 20% within six months. High-stakes entrance-exam scores fall by 18 and 24%, with the full penalty emerging only after about two years. The losses are largest in social science subjects, followed by STEM and languages, and are especially large for junior students, high-achieving students, and boys. The learning losses are concentrated among roughly 80% of AI users whose behavior is consistent with homework outsourcing, as indicated by exceptionally short homework completion time coupled with high homework scores. AI users who maintain similar homework completion time as non-AI users experience small learning losses.

JEL Classification: O15, I20, O33

Keywords: Generative AI, Education, Human capital accumulation, China shock

David Strömberg - david.stromberg@su.se
Stockholm University

Victor Lei - victorlei.econ@gmail.com
University of Hong Kong

Yanhui Wu - yanhuiwu@hku.hk
University of Hong Kong and CEPR

The Generative AI Learning Penalty: Evidence from Chinese Secondary Education*

David Strömberg† Victor Lei‡ Yanhui Wu§

June 2026

Abstract

Using 30 months of panel data on 26,811 Chinese students in grades 7–12, we study how generative AI affects homework productivity and learning. The data combine monthly closed-book exams, high-school and college entrance exams, and homework scores and completion time across nine subjects. We exploit staggered AI adoption in a difference-in-differences design. AI adoption raises homework scores by 18% and reduces completion time by 30%, but lowers monthly exam scores by 20% within six months. High-stakes entrance-exam scores fall by 18 and 24%, with the full penalty emerging only after about two years. The losses are largest in social science subjects, followed by STEM and languages, and are especially large for junior students, high-achieving students, and boys. The learning losses are concentrated among roughly 80% of AI users whose behavior is consistent with homework outsourcing, as indicated by exceptionally short homework completion time coupled with high homework scores. AI users who maintain similar homework completion time as non-AI users experience small learning losses.

*The authors have no sources of research support, financial relationships, or other potential conflicts of interest to disclose. The data collection and usage of this project were approved by the Human Research Ethics Committee of the University of Hong Kong with the following reference number: EA260129. We thank Jing Cai, Ruixue Jia, Ginger Zhe Jin, Sendhil Mullainathan, Torsten Persson, Abhijeet Singh, Jonas Vlachos, David Yanagizawa-Drott and participants at the Stockholm Uppsala Education Economics Workshop, University of Maryland, and University of Virginia for useful comments and suggestions.

†Stockholm University, email: david.stromberg@su.se

‡University of Hong Kong, email: victorlei.econ@gmail.com

§University of Hong Kong, email: yanhuiwu@hku.hk

1 Introduction

A growing body of evidence shows that generative AI enables workers to complete knowledge-intensive tasks much more efficiently (Achiam et al., 2023; Dell'Acqua et al., 2023; Noy and Zhang, 2023; Brynjolfsson et al., 2025). While these efficiency gains could create substantial economic value, AI use may also crowd out the effort through which people acquire skills. This raises concerns about individual cognitive atrophy (Kosmyna et al., 2025), failures of workplace training (Garicano and Rayo, 2025), and broader declines in human knowledge (Acemoglu et al., 2026).

The possibility of an AI learning penalty is especially important for young students, who perform a wide range of cognitive tasks in school. For students, completing these tasks efficiently is not the goal; learning from them is. Hence, the rapid diffusion of generative AI tools among students in recent years has created widespread concerns about their learning.[1] Yet systematic evidence on the learning losses due to generative AI use in natural school settings is still lacking. Generative AI may improve learning by acting as a personalized tutor. However, it may also undermine learning by allowing students to engage in *homework outsourcing* in which students use generative AI to complete assignments and bypass the cognitive effort that produces learning.

Experimental evidence paints a mixed picture, with estimated learning effects varying widely by context.[2] In the setting of Turkish high-school students, Bastani *et al.* (2025) find that student access to an AI tutor that can provide direct answers to math homework questions improves their performance in practice problems but reduces subsequent closed-book exam performance. Conversely, Kestin *et al.* (2025) show that an AI tutor designed with pedagogical principles significantly improves college students' learning in an introductory physics course. These studies provide useful evidence on how specific AI tools affect short-run learning in narrowly defined curricular settings. However, much less is known about a broader question: how does self-directed use of generative AI affect cumulative learning over time in ordinary school settings?

To answer this question, we exploit rich administrative data and staggered AI adoption in a 30-month panel of more than 25,000 students in Chinese secondary education (grades 7–12). We measure task productivity using students' homework scores and completion time. We distinguish between two types of learning outcomes: short-run learning, measured by students' performance on monthly closed-book school exams, and long-run learning, measured by performance on high school and college entrance exams. To identify causal effects, we leverage staggered adoption of generative AI in a difference-in-differences (DID) design. Our setting is well suited to this design. Reported AI use rose from nearly zero in 2022 to around 80 percent by June 2025, with adoption spreading gradually across students.

Our setting has three advantages over the randomized controlled trials that dominate the existing literature. First, students choose their own AI tools and usage

[1] In the United States, 84 percent of upper-secondary students reported using generative AI for schoolwork in 2025 (Adair et al., 2025). In a recent survey conducted by the Center for Democracy and Technology, more than 70 percent of parents and teachers reported concern that students' use of AI may weaken their academic skills (Laird et al., 2025).

[2] See Lo et al. (2024) and Wang and Fan (2025) for meta-studies and OECD (2026) for a discussion.

patterns, allowing us to measure effects under realistic conditions rather than under mandated tool use. Second, we observe outcomes over months and years, a much longer period than in most existing studies. Third, we observe both short-run learning of recently taught material (monthly exams) and long-run accumulation of general skills (entrance exams) across all major secondary-school subjects. Our main findings are as follows.

First, we find a sharp divergence between task productivity (homework completion) and learning outcomes (exam scores). Generative AI dramatically boosts homework performance: after six months of use, homework scores rise by 18 percent of the baseline mean while completion time per assignment falls from 64 to 45 minutes. However, students' scores in monthly school exams fall by 20 percent of the baseline mean, which is equivalent to 1.4 standard deviations (SD for short).

Second, the negative effect of generative AI on long-run learning (as measured by high-stakes entrance exams) is equally large, but accumulates more slowly. While regular-exam scores deteriorate fully within six months after initial AI adoption, it takes two years for the negative effect on entrance-exam scores to reach its full magnitude of 18 and 24 percent of the baseline mean (1.3 and 1.5 SD) for the college entrance exam (Gaokao) and high-school entrance exam (Zhongkao), respectively. This dynamic pattern implies that the short-duration studies that dominate the existing literature systematically underestimate the long-run learning cost of generative AI.

Third, the AI learning penalty appears to be driven primarily by the majority (81 percent) of AI users whose behavior is consistent with homework outsourcing. These students complete homework assignments in substantially less time than the fastest non-AI students, receive homework scores consistent with generative AI's known accuracy on similar problems, and perform poorly in closed-book exams.

By contrast, AI users who spend as much time on homework as non-users achieve similar exam scores, even though their higher homework scores indicate that they use AI. These students are not differentially selected on prior achievement, suggesting that generative AI does not reduce their learning efficiency.

Fourth, the AI learning penalty differs across subjects. Social science subjects (particularly Politics and Geography) are most negatively affected (-27 percent on average), followed by STEM subjects (-22 percent), and then languages (English, -17 percent, Chinese, -9 percent). This result is illuminating given that most RCTs study the effects of AI use on mathematics, programming, and foreign language skills.

Fifth, we characterize students with higher learning penalties. The exam scores of lower secondary students are more negatively affected than upper secondary students. Similarly, boys and high-achieving students suffer more from generative AI use.

Finally, we find suggestive evidence that the learning losses are mitigated after a certain period. Examining the effect of five months of AI exposure over calendar time, the estimated AI learning penalty falls from around 25 percent in early 2023 to around 16 percent by June 2025. The result holds for a fixed sample of early adopters. These findings indicate that students (or teachers) are gradually adapting to generative AI, but there remain substantial barriers to eliminating the persistent learning losses associated with AI use.

Our study contributes to the emerging research on the impact of generative AI on education and human capital accumulation. Despite substantial differences in samples and methods, our estimated effects are similar in magnitude to existing estimates in the literature when expressed as a percentage of the outcome mean. In standard-deviation units, our effects appear large: 1.4 SD for monthly average exam scores and 1.3–1.5 SD for entrance exams. This reflects the fact that averaging across 7–9 subjects compresses the monthly-average standard deviation to roughly half the within-subject standard deviation. Benchmarked against within-subject standard deviations, a more common denominator in the literature, our effects are correspondingly about half as large, though still substantial.

Our study reveals a number of new insights regarding the impact of generative AI use on learning. Our results are consistent with Bastani et al. (2025), who find large positive effects on homework scores and negative effects on exam scores in an RCT with short sessions among 1,000 Turkish high school students. However, they estimate an immediate (within the same 90-minute session) negative effect of -17 percent on mathematics scores, whereas we find a small effect in the first month that grows in magnitude to -22 percent after six months. One likely explanation is that their students used a streamlined AI system pre-programmed with the homework questions, making immediate outsourcing straightforward. In contrast, students in our setting use general-purpose AI tools, and it takes time to learn these tools effectively to outsource homework tasks.

Our results may appear to contrast with Kestin et al. (2025), who find that Harvard first-year physics students learn more in less time with a specialized AI tutor than with in-class active learning. However, their result is not inconsistent with ours. They demonstrate that, with a well-designed tool and highly motivated students, AI can enhance learning efficiency. We show that, under realistic conditions in which teenagers choose their own tools, the dominant pattern is not tutoring but homework outsourcing and a large AI learning penalty.

Our finding of pervasive homework outsourcing among students speaks to the literature on learning processes and skill dynamics. Kosmyna et al. (2025) find that writing essays with ChatGPT impairs users' ability to recall their own arguments and document reduced neural connectivity in users' brains during the writing process. Relatedly, cognitive psychologists show that practice testing, including homework-like exercises, improves retention and induces more cognitive effort than passive study (Roediger and Karpicke, 2006; Dunlosky et al., 2013). Our results suggest that, through homework outsourcing, students forgo this beneficial practice effect.

Moreover, our finding that exam scores deteriorate gradually over time suggests that homework outsourcing may undermine not only the retention of current material but also the absorption of future material. This is a form of dynamic complementarity in skill formation (Heckman and Corbin, 2016), where skills beget skills. It is also consistent with recent evidence on "cognitive endurance," the capacity to sustain effortful attention in improving cognitive ability (Brown et al., 2025).

More broadly, our study relates to research on digital technology in K–12 education. In a series of RCTs in rural China during 2010-2014 (Mo et al., 2014, 2015, 2020; Ma et al., 2024), researchers find that computer-assisted learning tools often have significant positive effects on students' academic achievement in math and language, but the effects are strong only when implementers follow the protocol and closely monitor technology use. In India, Muralidharan et al. (2019) and Muralidharan and Singh (2025) find that a well-designed computer-based personalized adaptive learning software strongly improves middle-school students' learning of math and Hindi in both the short run and long run. They also emphasize the importance of student learning time and the monitoring of implementation. In contrast to these studies, our findings show that generative AI, which is likely to become a prevalent technology for education, has a substantial negative impact on student learning. At the same time, our findings echo this line of research by highlighting the importance of incentives and implementation.

2 Background and Data

We collected data from a county in central China. The county has a population exceeding 1 million and GDP per capita slightly below 6,000 US dollars. It is representative of the vast majority of Chinese counties outside the well-developed coastal areas.

Throughout the six years of secondary education, students follow a national curriculum consisting of three core subjects (Chinese, Mathematics, and English) alongside several secondary subjects (Physics, Chemistry, Biology, History, Politics, and Geography). In our sample, class sizes range from 50 to 60 students. A central goal of secondary schooling is preparation for two high-stakes exams: the Senior High School Entrance Examination (Zhongkao) and the College Entrance Examination (Gaokao). The first exam sorts students into high schools of different quality and into vocational schools, and the second exam is the sole determinant of student admission to colleges.

2.1 Survey Design and Sample

We use data from a survey regarding students' use of generative AI tools that was conducted in late June 2025. The survey covered 26,811 students in grades 7–12 across five junior high schools and four senior high schools, accounting for 90% of all students enrolled for secondary education within the county. Students not covered by the survey mainly live 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) commonly used for research 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 check the registration dates in the AI tools they had used.[3] This was done to ensure the accuracy of students' self-reported timing of initial use of AI. We discuss potential measurement error in this key variable and its implications for our estimates in Section 4. The survey also asked

[3] Although some AI tools (e.g., DeepSeek, Doubao) allow use without registration, this only applies to the web version which provides partial services (e.g., no memory of past conversations). Most students use the App version of AI tools, which allows for a full function but requires registration of an account.

how many hours per week the student used generative AI. The final valid response rate exceeded 96%.

2.2 Exam and Homework Data

Data on students' academic performance were provided by the local educational bureau, which extracted the monthly exam results from a digital grade-management system. Monthly exams were designed and implemented by school teachers and conducted for all subjects (7 in lower secondary and 9 in upper secondary). These exams are graded through a unified digital platform, where teachers score students' work directly in the system.

In addition to the monthly exams, we obtained three sets of student performance data. The first is a 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 grades 7 and 9. The second is the Zhongkao, taken by grade 9 students near the end of junior high school. The third is the Gaokao, taken by grade 12 students each June. Both the 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 these exams can be found in the appendix.

The local educational bureau also provided scores for each student's weekly homework assignments in each subject. Homework assignments were submitted and graded through a digital platform. Importantly, the digital system records timestamps for homework access and submission. We define homework completion time as the interval between retrieving an assignment and submitting the answers on the digital platform. The average homework completion time is 58 minutes. Although this variable may not be a perfect measure of the actual time spent on each homework assignment, it is a reasonable proxy for the time students spend on the weekend homework assignments on average.[4]

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

In sum, our main outcomes are weekly homework scores and completion time as well as the scores of the monthly regular and county-wide exams and the entrance exams. We rescale all these scores so that they are reported as percentage of baseline mean, defined as the mean across all student-month observations in which students have not yet adopted generative AI. A score of 100 means that the student has the same score as the average among non-AI students. In our monthly panel, we use averages of

[4] While our data consist of the homework completion time, we also observe the log-in and submission timestamps for a subsample. Students in this subsample work on their homework mainly from Friday to Sunday. In this sample, 99 percent of students work on the assignments sequentially, logging in to retrieve one assignment, working on it, and submitting it before retrieving the next homework.

these variables across subjects (e.g., exam score average) within month as dependent variables. Summary statistics are shown in Table 1.

2.3 Demographic data

Student demographic data come from digital enrollment records collected by local educational authorities. The data include information on student gender, urban/rural residence, parents' occupation, and monthly family income. The full list of demographic variables is presented in Table A4.

2.4 Trends in generative AI use

At the start of our sample period (September 2022), widely available generative AI systems had more limited capabilities than later models. At this point, only a small fraction of students in our sample were using generative AI (see Figure 1).[5] The use of generative AI remained low throughout 2023. In September 2024 and January 2025, we see an acceleration in the uptake of generative AI among students. This acceleration coincided with the release of DeepSeek V2.5 and DeepSeek R1, respectively. Another popular generative AI tool, Doubao, was launched in August 2023, significantly upgraded in August 2024 and again in January 2025.

As shown in Figure 1, at the time of our survey (June 2025), around 80 percent of the students report using generative AI. The most popular tools were Doubao, DeepSeek, ChatGLM, Ernie Bot, and Qwen. Generative AI was reported to be most frequently used in mathematics (66 percent) and English (55 percent), while between 30 and 40 percent of the students report AI use in other subjects.

Chinese generative AI tools were relatively competent at answering questions from Chinese K–12 exams even in the early stage of our sample period. For example, Hou et al. (2024) find that, in zero-shot settings, Qwen-72B (Alibaba) and Ernie Bot 4.0 (Baidu) correctly answer around 90 percent of questions from various homework, practice questions, and mock exams freely available on the internet. 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 them.

3 Empirical Strategy

Our object of interest is the effect of having started using generative AI on our outcomes. 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

[5] GPT was banned in mainland China, but some Chinese students obtained access via VPN.

sample period, while the classification of students as *AI students* or *non-AI students* varies over time.

We estimate these effects using the staggered DID regression adjustment 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. This way, we estimate the effect 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 students who never adopt AI during the entire sample period. Standard errors are clustered at the class level (524 classes in the full sample).

The key identifying assumption of the above estimator is parallel trends: 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 the causal effects of AI adoption. For a detailed formulation, see the Appendix.

These cohort-specific effects are then aggregated across adoption cohorts and event time to obtain the summary measures reported below. We focus on three types of aggregation. First, 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 post adoption. These estimates trace out how the effect of AI adoption evolves with time since the first use.

Second, since our effects grow initially and then stabilize, we report the average effects within a maturation window as a compact summary of the full effect of AI adoption. Matching the empirical results, we use a window of 6-10 months after adoption for the regular exams and 24 months or more after adoption for the entrance exams. Third, 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.

The proposed estimator is well suited to our setting because it provides consistent estimates even when treatment effects vary across adoption cohorts and time. Such heterogeneity is plausible here. For instance, 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.

A substantial literature has shown that standard two-way fixed effects regressions may be biased when treatment effects vary over time or by adoption cohort (e.g., Cengiz et al., 2019; de Chaisemartin and D'Haultfœuille, 2020; Sun and Abraham, 2021; Goodman-Bacon, 2021; Callaway and Sant'Anna, 2021; Borusyak, Jaravel, and Spiess, 2024). For comparison, we also estimate effects using the extended two-way fixed-effects model outlined in Wooldridge (2025) and a standard two-way fixed-effects regression controlling for student and month fixed effects.

Means plot To visualize the effect of AI adoption, we plot mean outcomes by event time for ever-AI students alongside a reweighted comparison series of never-AI students.

At each event time, the reweighting matches the control group to the treated group by calendar month, grade, and school. The gap between the lines therefore does not reflect seasonal, grade, or school-composition differences. Details are in Appendix 9.1. Although identification comes from the DID estimator, the means plot is useful because it shows pre-trends, baseline levels, and post-treatment divergence in one figure.

Sample balance Although our identification does not rely on random assignment, the parallel-trends assumption is more plausible when ever-adopters and never-adopters are similar. In our data, ever-adopters and never-adopters are well balanced on all the demographic variables as well as students' initial exam scores. Figure A1 and Table A1 show that the standardized difference in means between ever- and never-AI students is smaller than 0.1 in absolute magnitude for all characteristics, except for whether a student's parents are in the knowledge-intensive sector, which produces a difference of 0.103. This is well below the 0.25 rule-of-thumb threshold discussed by Imbens and Wooldridge (2009).[6]

Characterizing early adopters Balance on observables does not imply that AI adoption is unpredictable. In fact, given our large sample, we can significantly predict AI adoption using the demographic variables. Figure A2 shows the coefficients from separate regressions of an indicator variable for ever-AI students on each demographic variable. Several patterns are worth noting. First, the relationship between AI adoption and pre-AI (September 2022) exam scores is nonlinear. Compared to students in the first tercile, students in the middle tercile are 3 percentage points (pp) less likely to adopt generative AI while students in the highest tercile are 3 pp more likely to adopt generative AI.

Second, students in grades 9 and 12 are significantly less likely to adopt generative AI, probably because they are in their graduation year and overloaded with preparation for the entrance exams. Third, boys are 2.5 pp more likely than girls to adopt AI. This aligns with the general pattern of AI adoption by gender (Bick et al., 2024). Fourth, there is a strong but non-monotonic relationship between generative AI adoption and household income. Students from families with monthly household incomes ranging from 5,000 to 10,000 Yuan (RMB) are 11 pp more likely to adopt generative AI than those from households with extremely low income (below 200 Yuan), while those with over 30,000 Yuan per month are only 7 pp more likely. This is consistent with the findings from other countries that income is not a linear predictor of AI adoption (e.g., McClain et al., 2026; Chatterji et al., 2025). Fifth, students whose parents work in knowledge-intensive occupations are 4 pp more likely to use generative AI, consistent with greater exposure to generative AI in these households.

We also characterize the early adopters among the ever-AI adopters. Specifically, we regress the number of months since the first AI adoption as of June 2025 separately on each demographic variable. Students in the highest pre-AI exam score tercile, who have adopted generative AI by June 2025, have on average done so 2-3 months earlier

[6] Imbens and Wooldridge (2009) suggest that with a normalized difference exceeding 0.25, linear regression methods tend to be sensitive to the specification.

than students in the lowest tercile. Other demographic characteristics do not predict early generative AI use. Bick et al. (2024) and Chatterji et al. (2025) show that in a highly heterogeneous sample, younger, more educated, and higher paid individuals tend to adopt generative AI for work earlier, but the prediction is fairly noisy. Our sample is much more homogeneous and the difficulty of predicting early AI adopters is perhaps not surprising.

The balance-test results and the predictability results are not contradictory. Table A1 shows that differences between ever-AI and never-AI students are small in magnitude, relative to the natural variation in the data. Figure A2 shows that adoption is statistically predictable because of the large sample size. For example, a 2.5 pp gender gap in AI adoption is highly significant but corresponds to a standardized difference of only about 0.08, well below the conventional imbalance thresholds. More importantly, identification relies on parallel trends rather than random assignment, and we will investigate pre-trends to support this assumption.

4 Main Results

In this section, we first present the effects of generative AI adoption on homework productivity (homework scores and completion time) and next on short-run learning (regular exams) with a number of robustness checks. We then examine effects on long-run learning (high-school and college entrance exams) and analyze the effect dynamics.

Figure 2 displays the results with regard to students' homework productivity and their performance in the regular exams. The left-hand figures show histograms of the outcomes for three groups: the never-AI students and the ever-AI students split by pre- and post-adoption periods. The right-hand figures plot the event-time means of these outcomes for the ever-AI and never-AI students from 10 months before to 10 months after a student's first AI use. The numbers inside the graphs report the average effect 6–10 months after adoption and the overall average treatment effect on the treated. They are computed from the Callaway and Sant'Anna (2021) regression adjustment estimator, with standard errors clustered by class in parentheses.

As shown in the histograms in Figure 2, during the pre-AI period, the outcome distributions of the ever-AI students are almost indistinguishable from those of the never-AI students. In the event-time plots, the means of the ever-AI and never-AI students are virtually on top of each other in the pre-period, for all outcomes. The mean outcomes for the two groups exhibit parallel trends before AI adoption. Moreover, they are very similar in mean levels. Figure A3 zooms in on the pre-period. During this period, ever-adopters have statistically higher mean exam scores, but the difference is only about 0.2 percent. Similarly, the right panels of Figure A4 (event-study graphs) show a negligible difference in trends between the treatment and control groups during the pre-period.

Because the pre-adoption differences in outcome means and distributions are very small, post-adoption raw differences closely track the DID estimates.[7] Hence, the gap

[7]The difference-in-differences estimator computes average differences in outcomes pre- and post

in raw means and the shift in distributions between pre- and post-adoption provide a useful visualization of the estimated effects of generative AI adoption.

4.1 Homework Productivity

Panels A and B of Figure 2 show the effects of AI use on homework productivity. The ever-AI students' homework productivity improves dramatically over the first six months after AI adoption. Their average homework scores rise by 18 percent after five months, while the never-AI students' average scores remain at the baseline mean. After six months, the difference between ever-adopters and never-adopters remains roughly stable at around 18 percent. This pattern is consistent with the statistically significant DID estimate of 18 percent (1.9 SD) for the average effect 6-10 months after adoption. The overall ATT is a statistically significant increase of 11 percent, smaller than the 6–10 month estimate because it includes the initial ramp-up period.

At the same time, the ever-AI students' average homework completion time declines from 64 minutes to around 45 minutes while the never-AI students remain at the 64-minute level. The DID estimate of the average effect 6-10 months after adoption is -19 minutes.

A notable feature of the histograms is that the AI students' homework time displays a secondary hump at higher levels (and similarly for low homework scores). These humps consist of students who recently started using generative AI and likely have not yet learned to use it fully. The humps disappear for students who have used generative AI for more than five months; see Figure A5.

4.2 Short-run Learning: Regular Exams

Simultaneously, the ever-AI students experience an equally dramatic decline in exam performance, measured by their scores on the monthly closed-book in-class exams. As shown in Panel C of Figure 2, their exam scores fall by 20 percent after five months, while the never-AI students' average scores remain close to the baseline mean. The DID estimates reveal the same pattern; see Figure A4. The ever-AI and never-AI students follow parallel trends before the month in which the ever-AI students start using generative AI. After that point, the estimated negative effect grows for about six months and then stabilizes. The average estimated effect for months 6–10 is 20 percent and highly significant. This effect amounts to a substantial fall of 1.4 standard deviations. The overall average effect, including all post-adoption months, is a statistically significant fall of 13 percent.

The effects build up gradually over approximately six months before stabilizing. Two forces can account for this pattern. First, students become more proficient at

values for treatment and control:

$$\begin{aligned}(\overline{y}^T_{post} - \overline{y}^T_{pre}) - (\overline{y}^C_{post} - \overline{y}^C_{pre}) &= (\overline{y}^T_{post} - \overline{y}^C_{post}) - (\overline{y}^T_{pre} - \overline{y}^C_{pre}) \\ &\approx \overline{y}^T_{post} - \overline{y}^C_{post}, \text{ since } \overline{y}^T_{pre} \approx \overline{y}^C_{pre}.\end{aligned}$$

using generative AI, as reflected in steadily declining homework completion time. Second, learning deficits accumulate as larger fractions of course material are processed via generative AI. The growing ratio of the negative exam effect to the reduction in homework time suggests that both mechanisms contribute.

Gradual proficiency in AI usage helps reconcile our findings of gradual ramp-up with the immediate effects found in the aforementioned experimental studies. They report large immediate effects because students are assigned highly accessible AI tools with explicit instructions and thus bypass the learning curve. In real-world settings such as ours, students must learn through trial and error how to apply general-purpose AI, delaying the impact on homework productivity and exam performance.

The stark patterns in Figure 2 can be explained by several factors. First, our sample is considerably larger than those in existing studies, which improves precision. Second, as we have shown above, students who adopt generative AI early are not strongly selected on either pre-treatment outcomes (parallel trends and similar levels in the pre-adoption period) or observable characteristics (balance tests in Section 3). Third, generative AI is a transformative technology, so its effects can be large relative to the natural variation in exam scores. Our effect size is comparable to that in Bastani et al. (2025), who find that a generative AI tool providing direct answers to homework questions immediately reduces closed-book math exam scores by 17 percent. In our data, six months after generative AI adoption, students' average exam score across all subjects falls by 20 percent. Finally, natural variation over time in our outcomes is small relative to the estimated AI effect. It is extremely rare for student exam scores to drop by 20 percent over a six-month period for reasons other than generative AI adoption.[8]

4.3 Robustness

This subsection presents a number of robustness tests for the main estimated effects reported above and further assesses the validity of a causal interpretation.

Alternative specifications. We re-estimate the average treatment effect of the treated using several alternative econometric specifications. First, we include students who have not yet started using generative AI in the control group. Second, we include school-by-grade fixed effects to account for school-by-grade specific trends. Third, we estimate effects using the extended two-way fixed-effects model outlined in Wooldridge (2025) and a standard two-way fixed-effects regression controlling for student and month fixed effects only. As shown in Table A2, the estimated average effect 6–10 months after adoption is very similar across specifications. The total average effect is smaller in absolute magnitude for the standard two-way fixed-effects model, perhaps because it is not robust to heterogeneous effects.

[8]To show this, we compute the change between the 6 to 10 month forward average and the once lagged value, including all student-month observations not affected by generative AI adoption (all ever-AI students 11 months before generative AI adoption and all never-AI students). Figure A6 shows that our estimated effects are extreme outliers in these distributions.

Self-reported AI adoption. 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. Misreporting could go in two directions. First, students could report an adoption month that is later than their actual adoption month. This would produce a pre-trend, with outcomes diverging before the reported start month. The absence of pre-trends makes this type of error unlikely. Second, students could report a starting date that is too early—for example, by reporting the date they registered an account rather than when they began using it actively. This would appear as a slower initial effect and could partly explain the six-month ramp-up, but it would not affect the estimated magnitude of the full effect or the finding (reported below) that entrance exam scores deteriorate more slowly than regular exam scores.

Contemporaneous shocks. A student who suddenly falls sick, becomes less engaged, or is more time-constrained may begin using AI and simultaneously reduce effort. If shocks of this magnitude were common, similar outcome changes should also appear among non-AI students. As mentioned previously, it is extremely rare that student exam scores drop by more than 20 percent of the baseline mean (the mean estimated effect for ever-AI students) over a six-month period for any student who does not use generative AI. In fact, this only happens to five non-AI students in our sample. This is especially telling in 2022 and 2023, when AI adoption in our sample is still very low. If large adverse shocks are common, we should still observe them among the large majority of students who remained non-AI in those years. Figure A6 plots the natural (non-AI-related) variation in our outcomes and shows that, for all our main outcome variables, changes as large as our estimated effect are rare among non-AI students. Hence, contemporaneous shocks are unlikely to explain our results.

Spillover effects. Another potential concern is spillovers from AI students to non-AI students. In response to falling student achievement, teachers may write simpler exams or exams that discriminate less strictly across student ability levels. A related concern is that AI students may have more time to socialize and discourage their classmates from studying. More generally, if never-AI students are affected by such spillovers, the parallel-trends assumption may be violated.

To address the concern that teachers adapt the exams to the share of AI students in the class, we study student performance on the county-wide exams, where all students in a grade, subject and month, although from different schools, answer the same questions. There are typically two county-wide exams per semester, and we observe this outcome only for students in grades 9 and 12. Despite the small number of observations, the estimated effects are very similar to our main results. Exam scores fall for the first six months after students start using generative AI, and the negative effect stabilizes at around -23 percent (1.3 SD); see Figure A7.

To test for general spillovers, we investigate whether the performance of the never-AI (control) students is correlated with the share of AI students in their class. We do this in a two-way fixed-effects model, regressing the exam scores of never-AI students on the share of AI students in their class, controlling for student and month fixed effects. Table A3 shows that the exam scores of the control group are not significantly correlated with the share of students who have adopted generative AI in their class (column 1) or in their school and grade (column 4).

We also estimate a specification inspired by Butts (2021) designed to capture both the direct treatment effect and the spillover effect. This specification includes an AI-student dummy and the interaction between a non-AI-student dummy and the share of AI students in the class (or school and grade). The estimated spillover effect is not significant, and the estimated treatment effect is virtually unaffected. Hence, we find no evidence of noticeable spillovers.

4.4 Long-run Learning: Entrance Exams

We now examine the impact of using generative AI on the longer-run learning outcomes, the Zhongkao and Gaokao. These exams differ from regular exams in several ways. First, the stakes are much higher because the entrance-exam scores determine students' admission into different high schools and colleges. While some students may not study hard for regular exams, this is unlikely for the entrance exams. Second, the entrance exams are much more challenging, covering material from multiple years. 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 and require strong integration across topics. For example, a math problem may combine algebra, geometry, and functional reasoning without signaling which method to use. Finally, the entrance exams are written and graded by external expert committees, whereas regular exams are written and graded by classroom teachers.

We observe the entrance-exam scores only for a subset of students: Zhongkao scores in 2022 and Gaokao scores in 2025 for the 12th-graders and Zhongkao scores in 2025 for the 9th-graders. Figure A8 plots the histograms for these students split into ever- and never-AI students. Students who use generative AI have lower average entrance exam scores in 2025, but the gap is considerably smaller than the corresponding gap for regular exams.

The fact that we observe the entrance exam outcomes only once or twice raises challenges in identifying the causal effect of generative AI adoption. We pursue two strategies to overcome some of the challenges. First, we run a standard regression controlling for the 2022 Zhongkao score (for the Gaokao outcome) and the September 2022 regular exam score (for the Zhongkao outcome). Second, we use the regular exam score one month before the treated students' first use of generative AI as the before-AI performance and apply the DID approach used in the previous estimation. Note that we cannot show a pre-trend in this case, because we do not observe students who start using generative AI after June 2025.

We report the estimated total ATT effects using these two strategies. Table 2 reports the results from the first strategy. Compared with non-AI students, AI students score around 5-6 percent lower on the Zhongkao and the Gaokao (columns 1 and 4). The estimated effects are robust to the inclusion of various controls. Column 7 provides a placebo test, showing that the treatment and control groups display no difference in exam performance in June 2022, before generative AI was available.

Figure 3 presents the results from the second strategy. The DID estimator yields a statistically significant total ATT of −7 percent for both the Zhongkao and Gaokao. For comparison, we also estimate the average effect of generative AI on regular exam scores in the same month (June 2025) within the same sample. The entrance-exam effects are significantly smaller in magnitude than the corresponding regular-exam effects in June 2025, which are −14 and −11 percent for 9th- and 12th-graders, respectively.

Although our identification strategies for the entrance exams are not as strong as those for the monthly regular exams, a causal interpretation remains plausible, as the results are consistent with evidence from settings where identification is stronger. First, as shown previously, comparing outcomes in the post-period yielded estimates very similar to our DID estimates for the monthly exams. The reason is that, prior to AI adoption, the ever-AI students have very similar outcomes to the never-AI students, both in mean and distribution. This is also true for the entrance-exam scores.[9] Similarly, the sample balance result documented in Section 3 applies with equal force here. Second, as we will show in Section 6, the ordering of heterogeneous effects is nearly identical across regular and entrance exams, even though these exam types differ substantially in scope, stakes, grading authority, and timing. This pattern is difficult to reconcile with confounders, which would have to replicate, for example, the precise subject-level structure of effects that we establish causally for the regular exams.

4.5 Effect Dynamics

We now compare the effect dynamics for the regular and entrance exams, using the DID strategy that applies to both exam types. Figure 3 shows that the full negative effects on regular exam scores are realized after six months, consistent with our earlier findings obtained from a longer sample (recall Figure 2). However, the negative effects on entrance exam scores accumulate more slowly, reaching their full magnitude only after two years. This pattern is consistent with the cumulative nature of the entrance exams. A student who adopted generative AI in January 2025 experienced its effects on learning only for material taught in the spring semester of 2025, while learning in previous years was unaffected. Since entrance exams cover material from both before and after adoption—with earlier material receiving greater weight for earlier adopters—the erosion of the entrance exam performance is naturally gradual. In contrast, regular exams test recently covered material, so most tested content has been learned post adoption, and students' performance is affected more quickly.

Crucially, the smaller average effects on entrance exams in June 2025 are a consequence of slower accumulation. Because many generative AI users in June 2025 adopted relatively recently, the full negative effect on entrance exam scores had not yet materialized in the average. For students who adopted generative AI two or more years earlier, the estimated effects are −24 percent (1.5 SD) for the Zhongkao and −18 percent (1.3 SD) for the Gaokao—comparable to the corresponding effects on June 2025 regular exams (−21 percent for 9th-graders and −14 percent for 12th-graders).

[9] The bottom pane of Figure A8 shows that the Zhongkao exam score distributions of the ever- and never-AI students in 2022 (before generative AI) are very similar. Column 7 of Table 2 displays a precise null effect for the placebo outcome of Zhongkao 2022 exam scores. Hence, controlling for previous exam scores does not affect estimates much.

For comparison, we also estimate dynamic effects using the regression specification used to produce Table 2. Specifically, we regress the entrance exam scores on an indicator for the number of months since a student started using generative AI, controlling for demographic characteristics and class-fixed effects. The estimated dynamic effects are similar to the corresponding DID estimates, as shown in Figure A10. In particular, for students who adopted generative AI two or more years earlier, the estimated effects are -23 percent for the Zhongkao and -14 percent for the Gaokao, close to the estimates using other approaches.

4.6 Effect size over time and adaptation

We first investigate how the effects of generative AI have evolved over time. Several factors may explain changes in the AI learning penalty. The capabilities of generative AI tools have improved, the composition of AI students has changed, and teachers and students may have adapted to generative AI.

To study this, we plot the effects five months after generative AI adoption. For each month, the sample consists of students who started using generative AI five months earlier. We use the five-month effect to maximize the time window over which we can estimate effects.[10] As shown in the left pane of Figure 4, these effects were around -25 percent in early 2023 and then fell gradually in absolute terms to -16 percent in June 2025. This could be because the estimation sample changes over time.

To assess the issue of sample composition, we examine the effect in a fixed sample of early users, namely those who started using generative AI in October 2022. Holding the estimation sample constant, the effects over time for this group of early adopters (the right pane of Figure 4) are roughly the same as the effects in the left panel. This indicates that the falling absolute effect size is not driven by the composition of affected students. Instead, it is driven by the changing effects of generative AI, perhaps due to students or teachers adapting to the use of generative AI.

Given the magnitude of the AI learning penalty we document, why do students, teachers, school administrators, or parents not respond more strongly to further mitigate the learning losses? Several factors may contribute.

School administrators may react slowly because aggregate effects emerged only recently and were initially modest. We compute the aggregate effect of generative AI on the county-level average exam scores in a given month by multiplying the average treatment effect on the treated by the share of students treated. Our estimates imply that in the two school years from September 2022 to June 2024, generative AI caused the average exam score across all students in the county to fall by only -3.4 percent. The total induced fall grows to -9.9 percent by June 2025, primarily because the share of students adopting generative AI rises rapidly to 80 percent. The aggregate effect is not larger at this point because of the large share of recent adopters for which the full negative effect has not yet been realized.

[10] The full effects are developed after 5 months. Using the effect at five months, we can estimate effects from February 2023 (5 months after September 2022).

Teachers may not react because they typically observe student performance in only one subject. As noted above, it is extremely uncommon for a non-AI student's average score across all subjects to fall over six months by as much as our estimated effect. By contrast, it is much more common for a student's score in a single subject to fall by this amount. This occurs in roughly 4 percent of student-month observations among non-AI students. From the perspective of an individual subject teacher, the decline is large but not extraordinary.

Students and parents observe the full drop in exam performance across all subjects. However, students who fall behind may not recognize that the reason is their use of generative AI. The gradual deterioration of exam scores over six months may conceal the relationship. Studies of secondary education have documented a disconnect between students' actual learning and their perception of learning. Most importantly, when students experience the increased cognitive effort associated with active learning, for example, struggling with their homework assignments, they initially interpret that effort as a signal of poorer learning (Deslauriers et al., 2019). Relatedly, although students who use generative AI perform worse on exams, they do not perceive that they perform worse or learn less (Bastani et al., 2025).

5 Homework Outsourcing with Generative AI

Students may use generative AI in very different ways, from homework outsourcing to tutoring. In this section, we show that the majority of AI students exhibit behavior consistent with homework outsourcing, spending little time completing homework assignments, and experiencing learning losses. The remaining AI students spend as much time completing homework as non-AI students, receive higher homework scores and similar exam scores. In other words, they are learning as efficiently as the non-AI students. We then discuss mechanisms that may explain the AI learning penalty.

5.1 Outsourcing vs. Non-outsourcing Students

Outsourcing students In traditional educational settings, homework performance reliably predicts exam performance, reflecting the practice and consolidation effects of completing learning tasks. Figure A11 illustrates this relationship for non-AI students. However, among AI students with homework scores above the baseline mean, higher homework scores are associated with lower exam scores. In this range, average homework time also falls: from 62 minutes at the baseline mean score of 100 to 43 minutes at 120 percent of the baseline mean. This suggests that high homework scores among AI students reflect greater reliance on AI assistance rather than better learning, consistent with the "homework outsourcing," as opposed to "tutoring," pattern of AI use.

This interpretation is further supported when we examine the joint distribution of homework time, homework scores, and exam scores. Figure 5 plots the median and interquartile ranges of homework and exam scores as functions of homework completion time, separately for AI students (red solid curve) and non-AI students (blue dotted curve). For non-AI students, homework completion generally takes at least 50 minutes, and both their homework and exam scores are monotonically increasing in homework completion time. This pattern is consistent with homework time proxying for student effort.

For AI-using students, the pattern is quite different. More than half of AI students spend 20–50 minutes completing homework, less time than even the fastest non-AI students. This group receives very high homework scores. Their median score 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. This pattern—low time on task, high homework scores, and low exam scores—is entirely consistent with homework outsourcing.

Among AI students, in the range of homework completion time between 20-45 minutes, both the homework score and exam score are roughly invariant in the homework completion time. However, within the 45–50 minute range, homework scores fall while exam scores rise rapidly. One possibility is that an increasing share of students in this group have adopted generative AI but do not use it for homework. However, because students in this group complete homework substantially faster than even the fastest non-AI students, this interpretation seems less likely. The more plausible interpretation is that students in this group are only partly engaging in homework outsourcing. This increases the time that students work on the homework, as well as their learning effectiveness as measured by exam scores. Notably, it also reduces the students' homework scores.

The majority of AI students appear to engage in homework outsourcing. We define an AI student as engaging in homework outsourcing if the student completes homework in less than 50 minutes, and as engaging in full homework outsourcing if the student completes it in less than 45 minutes. In our data, 58 percent of AI students engage in homework outsourcing and 34 percent engage in full homework outsourcing. After more than five months of generative AI use, these shares rise to 81 percent and 50 percent, respectively.

Non-outsourcing students At the other end of the distribution, AI students who spend more than 65 minutes on their homework receive homework and exam scores similar to those of non-AI students, suggesting that these students do not use generative AI for homework assignments. However, this group consists entirely of students who adopted generative AI no more than five months. Six months after adoption, no AI student spends more than 65 minutes completing their homework (see Figure A5). This is consistent with the gradual process of learning how to use AI tools. It also suggests that AI crowds out the highest level of effort.

Interestingly, in the range of 50–65 minutes, the median and the interquartile range of exam scores of AI and non-AI students are similar. This implies that, in the range where AI students and non-AI students have overlapping homework times, students who spend the same amount of time completing 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.[11] The AI students in this range

[11]Suppose that a share s of AI students at a given completion time use AI for homework while

receive significantly higher homework scores than non-AI students, indicating that they do employ generative AI for homework assignments. Importantly, they are not positively selected on pre-AI ability: their average pre-AI exam score is close to the baseline mean (Figure A12). This makes it unlikely that similar exam scores at equal homework time simply reflect positive selection into the high-time AI group.

This pattern shows that students who spend the same amount of time on homework learn similarly, with or without generative AI. In other words, generative AI reduces time spent learning for the majority of AI students but not learning efficiency for those who spend the same time studying as the non-AI students.

5.2 Discussion

The patterns above suggest that learning losses are concentrated among the roughly 80 percent of AI students who spend substantially less time on homework than non-AI students. The relationship between homework completion time and exam scores should not be interpreted as causal. Homework completion time after AI adoption allows us to characterize how the group of students whose learning is negatively affected use generative AI. Still, other factors that are correlated with our measured homework completion time may also contribute to the reduced learning. For example, students who engage in homework outsourcing of their weekend homework assignments likely similarly engage in outsourcing other school tasks, such as smaller weekday homework and in-school assignments.

The importance of homework time in learning It seems likely that such other factors contribute to the learning loss, since the effect on homework completion time is moderate relative to the total time students spend studying. Our estimated one-third reduction in homework completion time amounts to 2.2–2.8 fewer hours per week, or a 5–6 percent reduction in total weekly study time.[12] This is small compared to the 20 percent reduction in exam scores.

Nevertheless, homework time is particularly conducive to learning. Dunlosky et al. (2013) review various strands of evidence and conclude that practice testing, such as homework assignments, is among the most effective learning techniques, improving performance substantially more than spending the same amount of time studying. In addition, students who engage in homework outsourcing may also learn less during class time. Specifically, experimental studies have found that spending time on practice tests enhance students' ability to organize and integrate new material (Roediger and

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 $\overline{y}_{AI} = \overline{y}_{No-AI}$.

[12] Students in our sample spend roughly 42–44 hours per week on learning: 35 hours in class and 7–9 hours completing the main homework assignments. We estimate that homework completion time falls by approximately 20 minutes per homework assignment, or roughly one-third of total homework time.

Karpicke, 2006). The total effect of homework on learning may be substantially larger than its partial direct effect of learning homework content.

Policy implications The finding that AI students who spend the same amount of time on homework as non-AI students achieve similar exam scores does not necessarily imply that mandating longer homework time would restore learning for students engaged in homework outsourcing. Some AI-adopting students continue to spend more than 50 minutes per homework assignment. This could reflect several factors, such as stronger intrinsic motivation, closer parental monitoring, better information about the learning costs of homework outsourcing, or greater skill in using generative AI as a learning aid.

Among these explanations, stronger intrinsic motivation seems less likely because students engaged in homework outsourcing have ex ante exam scores similar to those who do not. If higher homework time reflects stronger monitoring, then mandating longer homework time, monitored by parents or teachers, may increase learning. However, if higher homework time reflects better knowledge of how to learn with AI (e.g., if students engaged in homework outsourcing do not know how to use generative AI productively), then simply requiring them to spend more time on homework with access to generative AI may not restore their learning. In this case, an effective treatment may combine mandated longer homework time with other interventions, such as information on how to use generative AI effectively for learning, or information about the AI learning penalty associated with homework outsourcing.

An implication of our findings is that AI weakens the informational value of homework. Among AI students with above-median homework scores, higher homework scores are associated with lower exam scores (see Figure A11). This creates a monitoring problem for teachers and parents. Better homework performance may coexist with weaker learning, and the divergence may not become visible until closed-book exams or high-stakes assessments.

6 Heterogeneity

6.1 What subjects are most affected?

A priori, it is difficult to predict which subjects should be most affected by generative AI. The RCT literature focuses heavily on mathematics, programming, and foreign languages. Meta-studies that compare effects across studies (e.g., Wang and Fan 2025) find no significant differences in (positive) learning effects between the studies of STEM-related courses and those of language learning and academic writing. However, such comparisons are made across very different samples in different grade levels and countries, and the estimates are typically based on small samples.

We estimate the effects across nine subjects in a unified setting with a focus on the fully developed effects (6–10 months post-adoption for the regular exam scores and 24 months post-adoption for the entrance exam scores). Figure 6 reports the average effect from the Callaway and Sant'Anna DID estimator by subject. All subject-level effects are statistically significant, and the results for regular and entrance exams are strikingly similar across subjects.[13] Because regular and entrance exams differ in scope, stakes, grading authority, and timing, such consistency across exam types provides a cross-validation for the causal interpretation of the entrance exam estimates.

Across subjects, the largest losses occur in social sciences with negative effects of −27 percent of the baseline mean on average. Next are STEM subjects with average negative effects of around −22 percent. Least affected are English (−17 percent) and Chinese (−9 percent). For comparison with previous studies, the effect on mathematics equals −22 percent of the baseline mean and 0.84 standard deviations. Note that the effect on subject-level exam scores measured in standard deviations is smaller than the corresponding effect on the monthly average exam scores, simply because the standard deviation of individual subject exam scores is larger than that of the mean monthly scores.

6.2 Which students are most affected?

We next investigate whether effects differ by educational level, weekly reported AI time (dose-response), gender, and initial exam score (pre-AI skill). To this end, we estimate effects in samples split by each of these characteristics. Again, we focus on the fully developed effects; see Figure 7. Figure A14 shows the corresponding event-study graphs as well as the results for homework scores and completion time.

Educational level The younger students in lower secondary education are more negatively affected than those in upper secondary education, as shown in the upper left pane of Figure 7. The differences are relatively large, roughly 40 percent in both regular and entrance exams. For example, the lower-secondary students experience a drop of -24 percent in their regular exam scores, compared with -17 percent among the upper secondary students. This result is consistent with insights from our interviews with school teachers who stated that high-school students were more loaded with supervised course work and more restrictive in the use of generative AI.

Dose response We find a clear dose response in the sense that students who report using generative AI more suffer more in terms of learning. As indicated in the upper right pane of Figure 7, the average estimated effect 6–10 months after adoption for students who report using generative AI 0–1 hours per week is −5 percent, compared to −30 percent for students who report using generative AI five hours or more.

Gender We study whether the use of generative AI widens the female-male gender gap in academic performance. The literature has documented that female students typically receive higher grades than male students, perhaps due to differences in noncognitive skills and behaviors (Cornwell, Mustard, and Van Parys, 2013) or teacher-related reasons (e.g., Dee, 2007; Lavy and Schlosser, 2011). In our setting, we also find that girls outperform boys in exam scores.

[13] The exception is that regular exam scores in biology are less negatively affected than the corresponding entrance exam scores, and scores in mathematics are marginally less negatively affected.

The lower left pane of Figure 7 plots the estimated effects for the samples of female and male students, respectively. The average effect 6–10 months after adoption is -18.4 percent for girls, compared to -21.6 percent for boys. The difference of 3.2 percent of the baseline mean exam score implies that the negative effect of generative AI on the exam scores is 17 percent greater for boys than for girls. A gender difference of similar magnitude also appears for entrance exam scores.

The differential effect on regular exam scores can be explained mainly by the higher reported intensity of generative AI use among boys. Multiplying the share of boys who report different weekly hours of generative AI use with the corresponding estimated effects in the upper right graph of Figure 7 produces a weighted-average effect of -21.3 percent for boys. The corresponding weighted-average effect for girls is -18.6 percent. Thus, the reported intensity of generative AI use explains 2.7 percent of the 3.2 percent difference in the estimated effect between boys and girls.

The gender gap could also have widened if more male students use generative AI at the extensive margin. In our June 2025 survey, 80 percent of girls had adopted generative AI compared to 82 percent of boys, demonstrating a small gender difference in generative AI adoption. This is similar to other surveys of generative AI use in China and the US, whereas significantly more males use AI in the UK.[14]

Another potential channel is that boys use generative AI differently. In particular, boys may be more likely to engage in homework outsourcing. However, we find little evidence supporting this channel. It is true that homework completion time falls more and homework scores increase more for boys than for girls, but the differences are very small, as shown in Figure A14c.

Pre-AI student skill We present results on the distributional consequences of generative AI for student learning by their pre-AI ability (the initial average exam-score of each student in September 2022). Specifically, we split students into terciles based on these scores, from lowest to highest so that the third tercile contains the one third of students with the highest grades. We then estimate effects separately by the pre-AI exam-score tercile. The results are shown in the lower right pane of Figure 7. The negative learning effects are larger for students with higher initial achievement. The differences in the estimated full (6–10 month average) effects are substantial, with a 50% gap between the most negative effect (-24 percent) for the highest tercile and the least negative (-16 percent) for the lowest tercile. The difference of 8 percent of the baseline mean amounts to around 0.5 SD, so the generative AI effect on the exam score distribution is considerable. The differences are similar for entrance exams, with effects of -18 percent for the top tercile compared to -11 percent for the bottom tercile.

This finding points to a compression of the skill distribution, but through a different mechanism than that identified in previous work. Earlier studies show that generative AI disproportionately improves task performance among lower-skilled individuals (e.g., Noy and Zhang, 2023; Brynjolfsson et al., 2025; Cui et al., 2026). In contrast, the

[14] See https://xinwen.bjd.com.cn/content/s69c46a21e4b0cd719e9f86bd.html for evidence from China, Faverio and Sidoti (2025) for evidence from the US, and Ofcom (2024) for evidence from the UK.

compression effect in our setting arises because the AI learning penalty is largest among high-performing students.

7 Conclusion and Discussion

Using large-scale evidence from a natural school setting, we find that generative AI use substantially reduces learning among secondary-school students. The central pattern is stark: AI raises homework scores and reduces homework time, but lowers performance on closed-book exams. Regular monthly exam scores fall by 20 percent, while high-stakes entrance-exam scores fall by 18 and 24 percent. The negative learning effects fully materialize after six months for regular exam scores and after two years for entrance exams.

The negative effects on learning outcomes appear to be mostly driven by the 81 percent of AI-using students, who spend less time on homework than even the fastest non-AI student, receive high homework scores matching the capability of generative AI tools they are using, and yet very low exam scores. By contrast, AI students who spend as much time on homework as non-AI students achieve similar exam scores and higher homework scores.

Our findings highlight the importance of the demand side and student incentives. Much of the existing RCT literature focuses on the supply-side question of how to design AI tools that scaffold learning while limiting outsourcing. In China, as in many other countries, such tutoring tools already exist and are often available at low or zero cost. Yet most students in our setting do not choose them. Instead, they rely on general-purpose AI tools that provide quick answers and facilitate homework outsourcing.

How can we incentivize students to choose tutoring AI tools in an environment with general-purpose AI tools that give direct answers? Our findings suggest several possible interventions. First, providing credible information about the long-run learning costs of homework outsourcing may alter student behavior. Second, increasing the weight placed on closed-book, in-person assessments may help restore the connection between effort and reward. Third, parents and teachers may be more effective if they monitor inputs, such as homework time and study effort, rather than outputs such as homework scores.

8 References

References

[1] Acemoglu, Daron, Asuman Ozdaglar, and Dingwen Kong. 2026. "AI, Human Cognition and Knowledge Collapse." NBER Working Paper.

[2] Achiam, Josh, Steven Adler, Sandhini Agarwal, et al. 2023. "GPT-4 Technical Report."arXiv preprint arXiv:2303.08774.

[3] Adair, Alexandra, Jessica Howell, Amanda Jacklin, and Alexandria Walton Radford. 2025. "U.S. High School Students' Use of Generative Artificial Intelligence: New Evidence from High School Students, Parents, and Educators." College Board Research.

[4] Bastani, Hamsa, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, & R. Mariman, "Generative AI without guardrails can harm learning: Evidence from high school mathematics." Proceedings of the National Academy of Sciences 122.26 (2025): e2422633122.

[5] Bick, Alexander, Adam Blandin, and David J. Deming. 2024. "The Rapid Adoption of Generative AI." NBER Working Paper No. 32966.

[6] Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event-Study Designs: Robust and Efficient Estimation," Review of Economic Studies, 91 (2024), 3253–3285.

[7] Brown, Christina, Supreet Kaur, Geeta Kingdon, Heather Schofield, Cognitive Endurance as Human Capital, The Quarterly Journal of Economics, Volume 140, Issue 2, May 2025, Pages 943–1002.

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

[9] Butts, Kyle. "Difference-in-differences estimation with spatial spillovers." arXiv preprint arXiv:2105.03737 (2021).

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

[11] Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer, "The Effect of Minimum Wages on Low-Wage Jobs," Quarterly Journal of Economics, 134 (2019), 1405–1454.

[12] Chatterji, Aaron, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman, "How People Use ChatGPT," NBER Working Paper 34255 (2025), https://doi.org/10.3386/w34255.

[13] Cornwell, C., Mustard, D. B., & Van Parys, J. (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. Journal of Human Resources, 48(1), 236-264.

[14] Cui, K. Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., & Salz, T. (2026). The effects of generative AI on high-skilled work: Evidence from three field experiments with software developers. Management Science.

[15] de Chaisemartin, Clément, and Xavier D'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," American Economic Review, 110 (2020), 2964–2996.

[16] Dee, Thomas S. "Teachers and the gender gaps in student achievement." Journal of Human Resources 42.3 (2007): 528-554.

[17] Dell'Acqua, Fabrizio, Ethan McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, and Karim R. Lakhani. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Harvard business school technology & operations mgt. Unit working paper, (24-013).

[18] Deslauriers, Louis, et al. "Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom." Proceedings of the National Academy of Sciences 116.39 (2019): 19251-19257.

[19] Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public Interest, 14(1), 4-58.

[20] Faverio, Michelle, and Olivia Sidoti. "Teens, social media and AI chatbots 2025." Pew Research Center (2025).

[21] Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." Journal of Econometrics 225 (2): 254–77.

[22] Heckman, James J., and Chase O. Corbin. 2016. "Capabilities and Skills." Journal of Human Development and Capabilities 17 (3): 342–59.

[23] Hou, Jinchang, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, and Min Yang. "E-eval: a comprehensive Chinese k-12 education evaluation benchmark for large language models." arXiv:2401.15927 (2024).

[24] Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." Journal of Economic Literature 47 (1): 5–86.

[25] Kestin, Greg, Miller, Kelly, Klales, Anna, Milbourne, Timothy, and Ponti, Gregorio. "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.

[26] Kosmyna, Nataliya, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. "Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task." Working paper, MIT Media Lab.

[27] Laird, Elizabeth, and Maddy Dwyer. 2025. "Up in the Air: Educators Juggling the Potential of Generative AI with Detection, Discipline, and Distrust." Center for Democracy and Technology Report.

[28] Lavy, Victor, and Analia Schlosser. "Mechanisms and impacts of gender peer effects at school." American Economic Journal: Applied Economics 3.2 (2011): 1-33.

[29] 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.

[30] Ma, Yue, Robert W. Fairlie, Prashant Loyalka, and Scott Rozelle, 2024. "Isolating the "Tech" from EdTech: Experimental Evidence on Computer Assisted Learning in China," Economic Development and Cultural Change, Vol. 72, No. 4.

[31] McClain, Colleen, Monica Anderson, Olivia Sidoti, and William Bishop. 2026. "How Teens Use and View AI." Pew Research Center.

[32] Mo, Di, Linxiu Zhang, Renfu Luo, Qinghe Qu, Weiming Huang, Jiafu Wang, Yajie Qiao, Matthew Boswell, and Scott Rozelle. 2014. "Integrating Computer-Assisted Learning into a Regular Curriculum: Evidence from a Randomised Experiment in Rural Schools in Shaanxi." Journal of Development Effectiveness 6 (3): 300–323.

[33] Mo, Di, Weiming Huang, Yaojiang Shi, Linxiu Zhang, Matthew Boswell, Scott Rozelle, Computer technology in education: Evidence from a pooled study of computer assisted learning programs among rural students in China, China Economic Review, Volume 36, 2015, Pages 131-145.

[34] Mo, Di, Yu Bai, Yaojiang Shi, Cody Abbey, Linxiu Zhang, Scott Rozelle, Prashant Loyalka, Institutions, implementation, and program effectiveness: Evidence from a randomized evaluation of computer-assisted learning in rural China, Journal of Development Economics, Volume 146, 2020.

[35] Muralidharan, K., A. Singh, and A. J. Ganimian (2019): "Disrupting education? Experimental evidence on technology-aided instruction in India," American Economic Review, 109, 1426–60.

[36] Muralidharan, Karthik and Abhijeet Singh, "Adapting for scale: Experimental Evidence on Technology-aided Instruction in India," NBER Working Paper 34205 (2025), https://doi.org/10.3386/w34205.

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

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

[39] Ofcom, "Online Nation 2024 Report", (2024).

[40] Roediger, Henry L., III, and Jeffrey D. Karpicke. 2006. "Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention." Psychological Science 17 (3): 249–55.

[41] Sun, Liyang, and Sarah Abraham. 2021. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." Journal of Econometrics 225 (2): 175–99.

[42] 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.

[43] Wooldridge, Jeffrey M. "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators." Empirical Economics 69.5 (2025): 2545-2587.

[44] Zhang, Xudong, Chunyang Li, Yiming Zong, Zhe Ying, Liang He, and Xipeng Qiu. 2023. "Evaluating the Performance of Large Language Models on Gaokao Benchmark." arXiv preprint arXiv:2305.12474.

9 Appendix

9.1 Mean Plots

In Figure 2, we plot mean outcomes by event time for ever-AI students alongside a reweighted comparison series of never-AI students in an interval of 10 months before and after the month students start using generative AI. At each event time, the reweighting matches the control group to the treated group by calendar month, grade, and school. This appendix describes the process in detail.

Let s be an indicator for each unique combination of generative AI adoption month, g, grade and school. This indicator partitions the data; each student belongs to one and only one s. Let m be the current month. We define the lag $l = m - g$ which is the number of months since the student started using AI. For all treated individuals i observed in lag l, we plot the average outcome

$$\overline{y}_l^T = \frac{1}{n_l^T} \sum y_{i(g_i+l)}^{s_i} = \sum_s \rho_{sl}^T \overline{y}_{sl}^T$$

where n_l^T is the number of treated students observed with lag l or, equivalently, observed in month $m = g + l$. We can rewrite this as

$$\overline{y}_l^T = \sum_s \rho_{sl}^T \, \overline{y}_{sl}^T,$$

where ρ_{sl}^T is the share of treated individuals with lag l that are in group s.

For comparison, we also plot the corresponding mean for never-AI students who have not yet started using generative AI in June 2025. For each s, we compute a representative matched control outcome, $\overline{y}_s^C$ which is the mean outcome in the control group in s. We then compute and plot the average control outcome for each lag,

$$\overline{y}_l^C = \sum_s \rho_{sl}^T \, \overline{y}_s^C .$$

9.2 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 $G_i \in \{1, \dots, T, \infty\}$ 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 to capture

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

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.

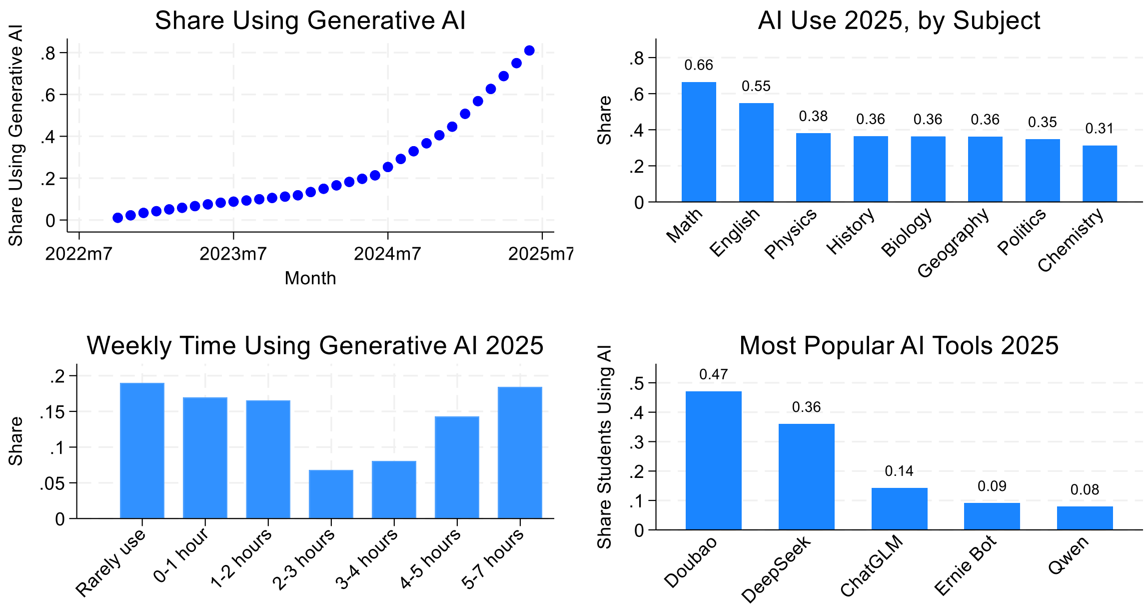
9.3 Question design and grading of the high-school and college entrance exams

For the high-school and college entrance exams, the entire process from question design to grading is centralized at the provincial level, bound by strict disciplinary and legal constraints. All personnel involved in exam design and grading sign confidentiality agreements. Any violations of the protocols will result in criminal sanctions. Through standardized measures such as personnel screening, physical isolation, anonymous segmentation, double-evaluation arbitration, and real-time monitoring, the fairness and non-manipulability of the exams are institutionally guaranteed.

The examination questions are designed by an expert panel that has been strictly screened. Members are required to have no immediate family members taking the exam in the current session and no experience working in private tutoring institutions. During the exam period, starting from question design, the panelists are required to reside in a fully enclosed facility, without any external communication until the exams conclude. The panel sets precise rules regarding the allocation of the weight of knowledge points and the proportion of difficulty across the entire exam paper. They also collect and compare all teaching and reference materials available in the market to avoid similarity with existing ones. Before finalization, a group of teachers not involved in the question-design process conducts on-site trial tests to verify answer accuracy, difficulty coefficients, and completion time.

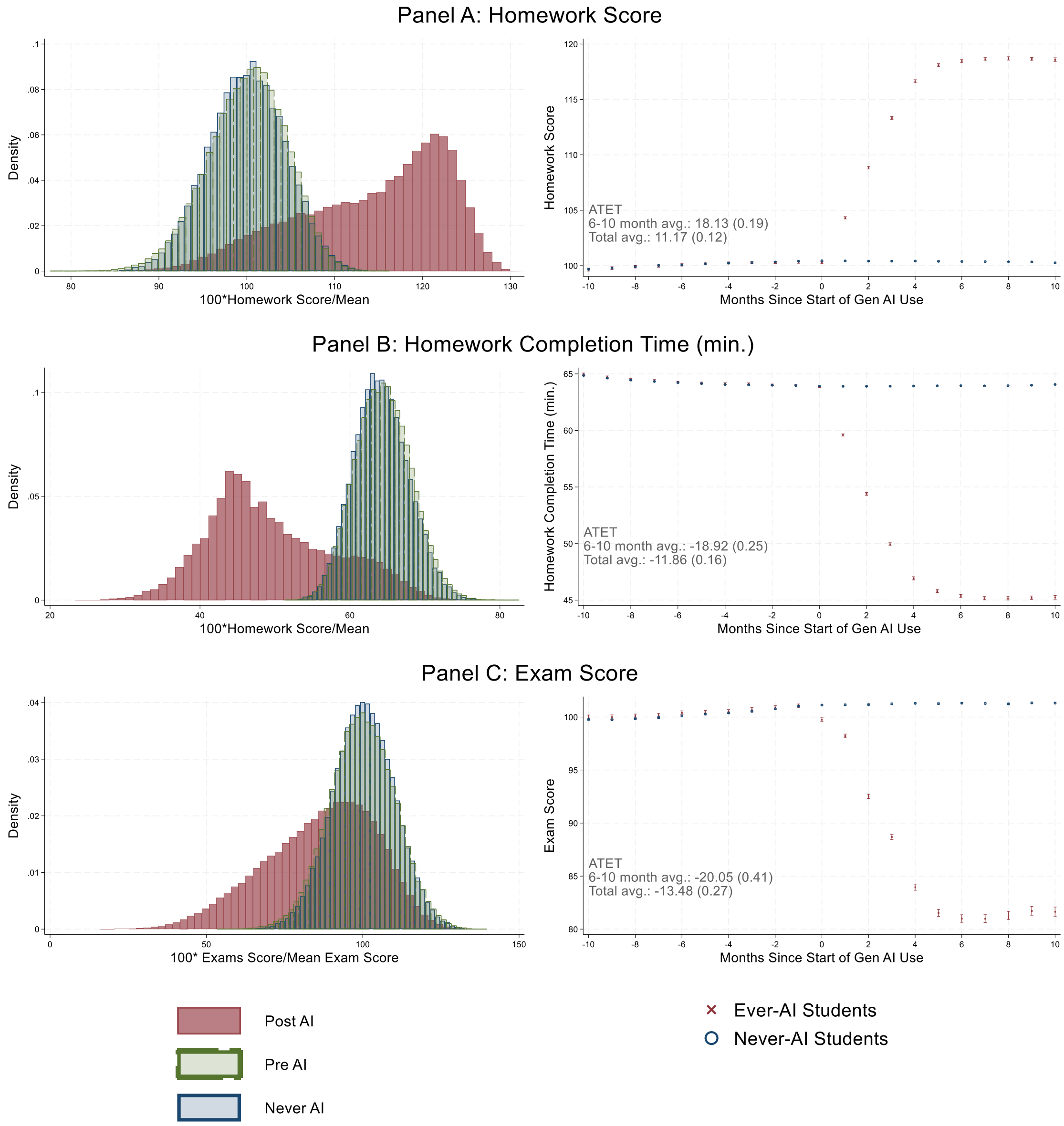
The grading is conducted by a panel of expert teachers selected from various schools in a designated location under close monitoring. Before formal grading begins, the expert panel develops detailed scoring guidelines through trial evaluations of sample papers. The grading follows a "double-blind, double-evaluation" mechanism. For the double-blind part, the process relies on an "electronic grading" system to ensure full anonymity and monitoring throughout. This applies not only to the multiple-choice questions but also to the essay or analytical questions. After answer sheets are scanned and stored, these non-multiple-choice questions are divided into independent image fragments and randomly distributed to different graders. To implement the double-evaluation principle, each question is independently scored by two graders. If the score difference exceeds a preset threshold, the system automatically triggers arbitration by a third grader or review by a team leader to ensure objective scoring.

Figure 1: Survey of Generative AI Adoption



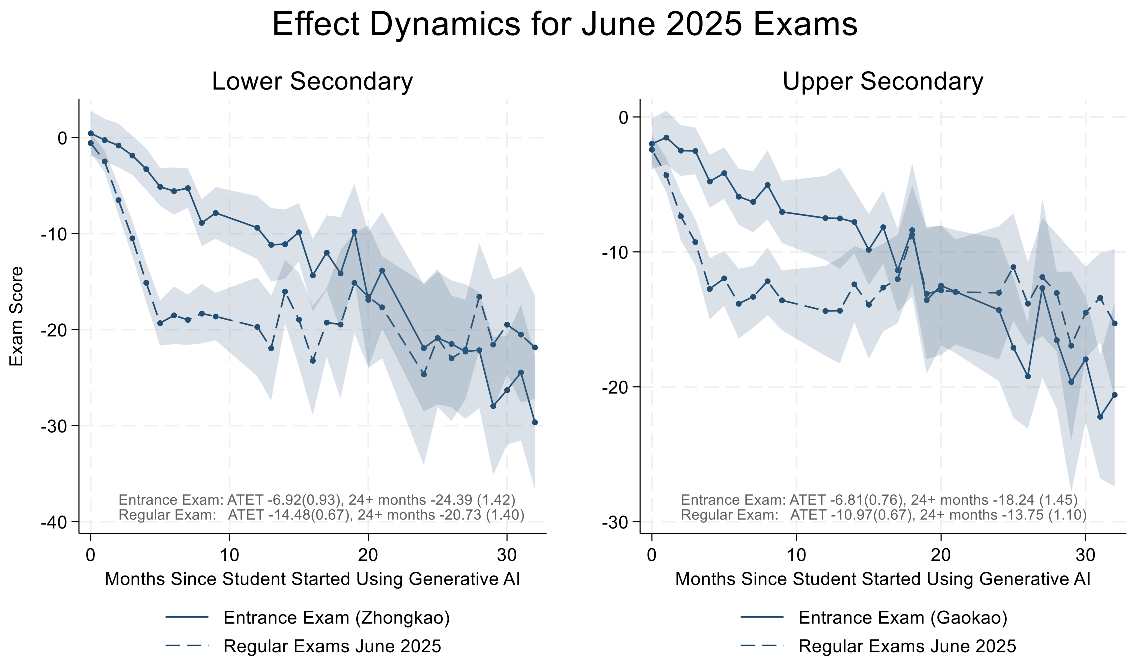
Note: The figures are based on responses to our survey of 26,811 students in grades 7-12. The first figure plots the share of students who adopted generative AI by each month. The following figures present the share of students in June 2025 who report using generative AI for different subjects, how many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

Figure 2. Main Results



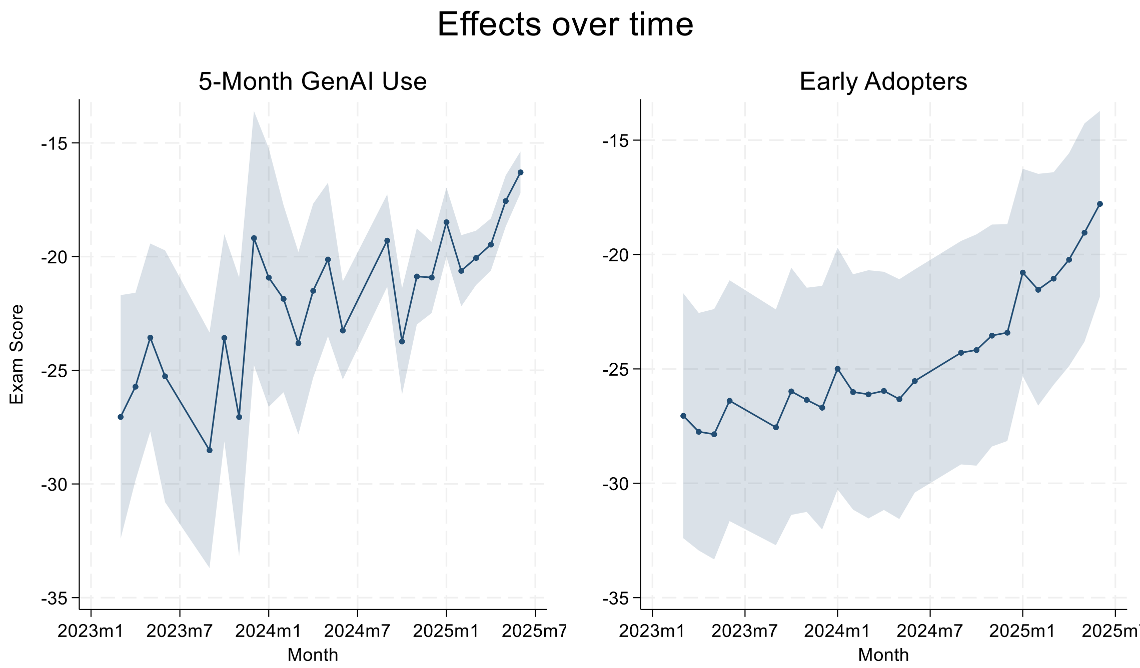
Note: The figures to the left plot the distribution of our outcomes, split by student-month observations for students who eventually adopt generative AI before adoption (Pre AI), after adoption (Post AI) and for student-month observations associated with students who have no adopted generative AI by June 2025 (Never AI). The figures to the right plot the mean outcomes by event time for students who eventually adopt generative AI before adoption (Ever-AI Students). For the Never-AI Students, we construct a reweighted mean comparison series so that, at each event time, the control observations have the same distribution over month, grade, and school as the corresponding Ever-AI observations. The data is at the student-month level and the samples are shown in Table 1. In panel A, the outcome is the average score (percent relative to baseline average) on the homework assignments. In panel B, the outcome is the average homework completion time in minutes. In Panel C, the outcome is the average score (percent relative to baseline average) on the closed-book exams in class.

Figure 3. Learning Deterioration Speed for Regular and Entrance exams



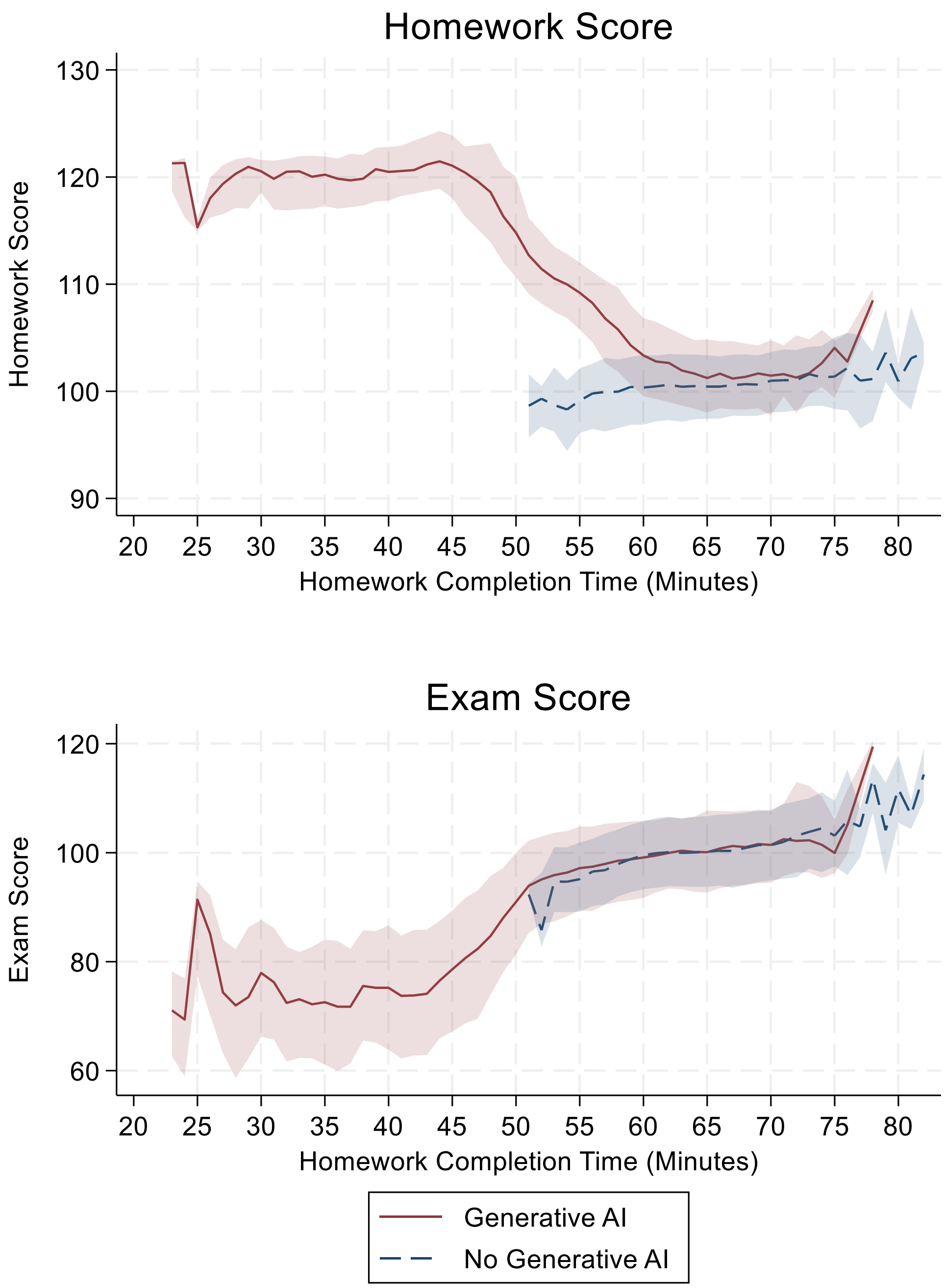
Note: The figure plots the average treatment effects by event time (and the corresponding 95% confidence intervals) of generative AI adoption on the June 2025 regular exam scores and entrance exam scores, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator. The sample consists of students in grades 9 and 12 in June 2025. For the entrance exams, the 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 4. Effects over Time



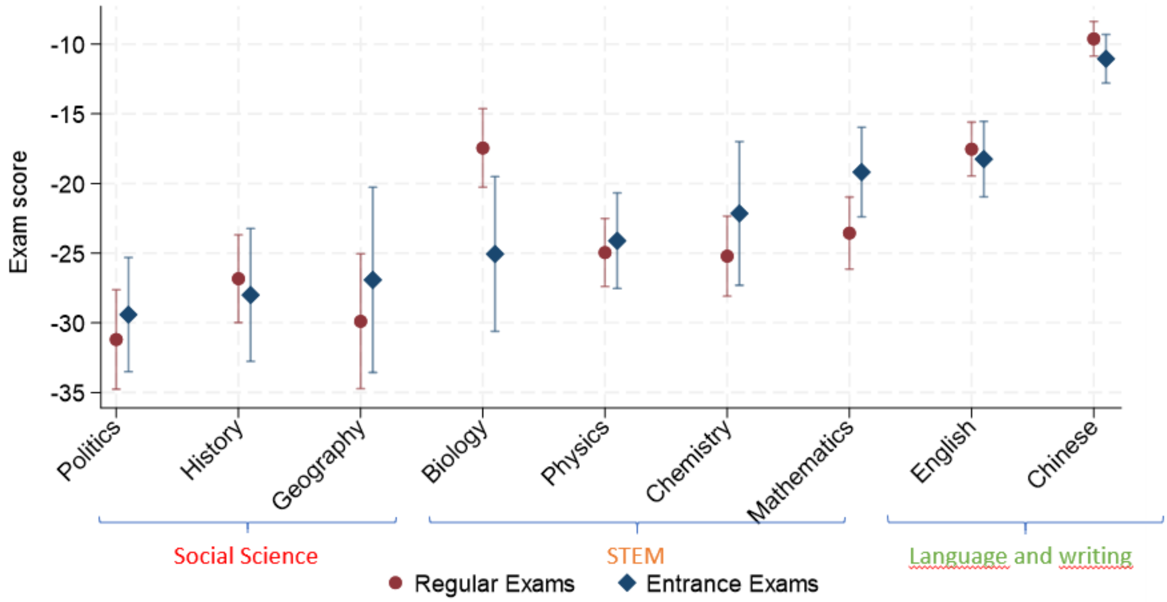
Note: The graph to the left plots the estimated average treatment effects (and 95% confidence intervals) of having adopted generative AI for exactly five months on regular exam scores, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator, by month. In this graph, the estimation 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 5. Homework and Exam Scores by Homework Completion Time



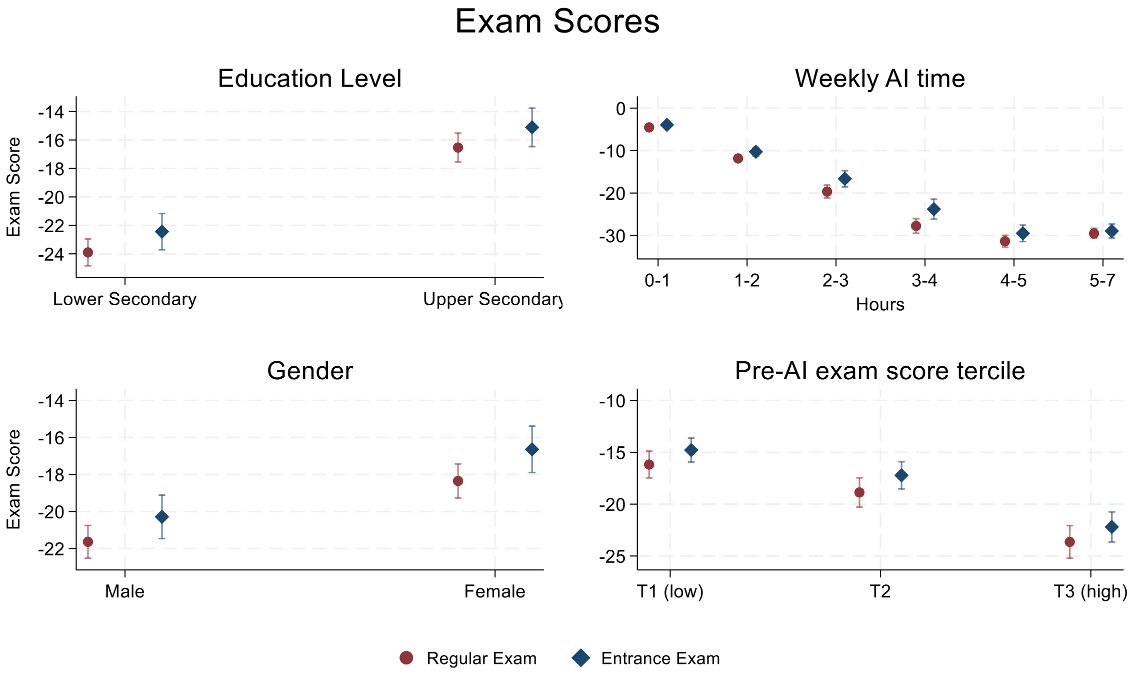
Note: The figure plots the median and interquartile ranges of homework and exam scores by homework completion time, separately for AI students and non-AI students. The data is at the student-month level and the sample is the one shown in Panel A of Table 1 with non-missing observations for homework time.

Figure 6. Heterogeneous Effects by Subject



Note: The figure plots, by subject, the average treatment effect 6-10 months post generative AI adoption and the 95% confidence intervals on regular and entrance exam scores, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator. The data is at the student-by-month-by subject level. Standard errors are clustered by class id.

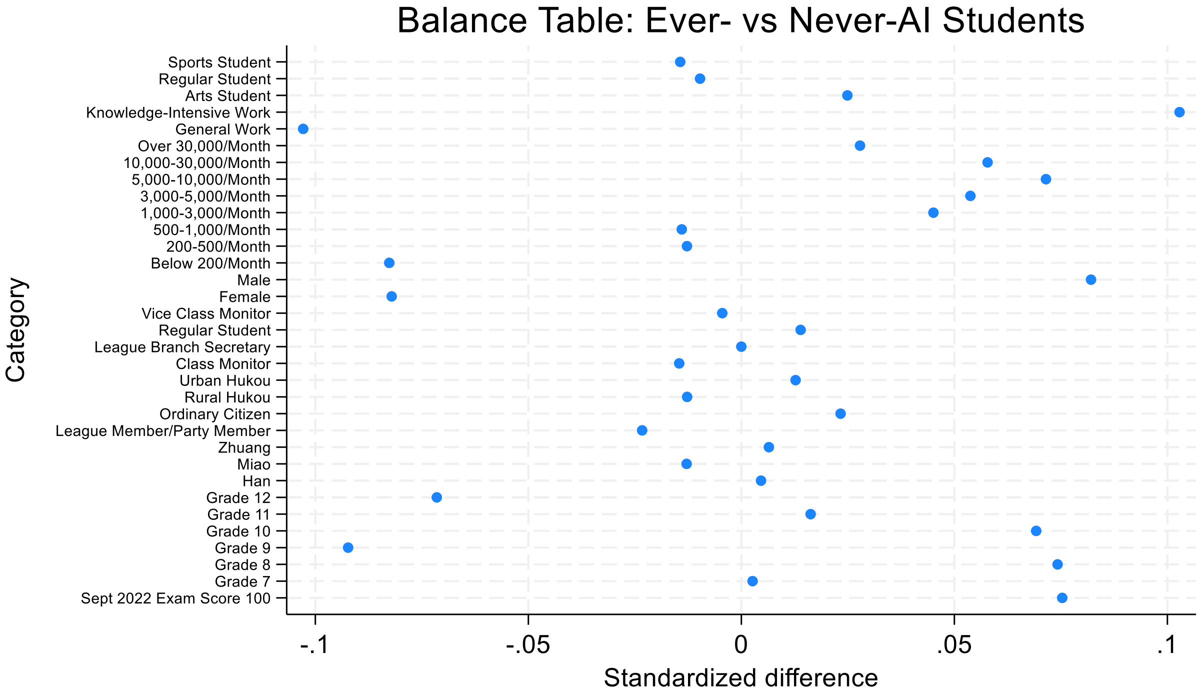
Figure 7. Heterogeneous Effects on Regular Exam Scores by Student Characteristics



Note: The figure plots the estimated average treatment effects on the treated (and 95% confidence intervals) of generative AI adoption on regular exam scores (above) and entrance exam scores (below), computed using the Callaway and Sant'Anna (2021) regression adjustment estimator. For effects for regular exam scores averaged 6-10 months post adoption, while for entrance exams the effects are averages for 24 months since adoption or more. Estimation samples are split by student characteristics. Standard errors are clustered by class id.

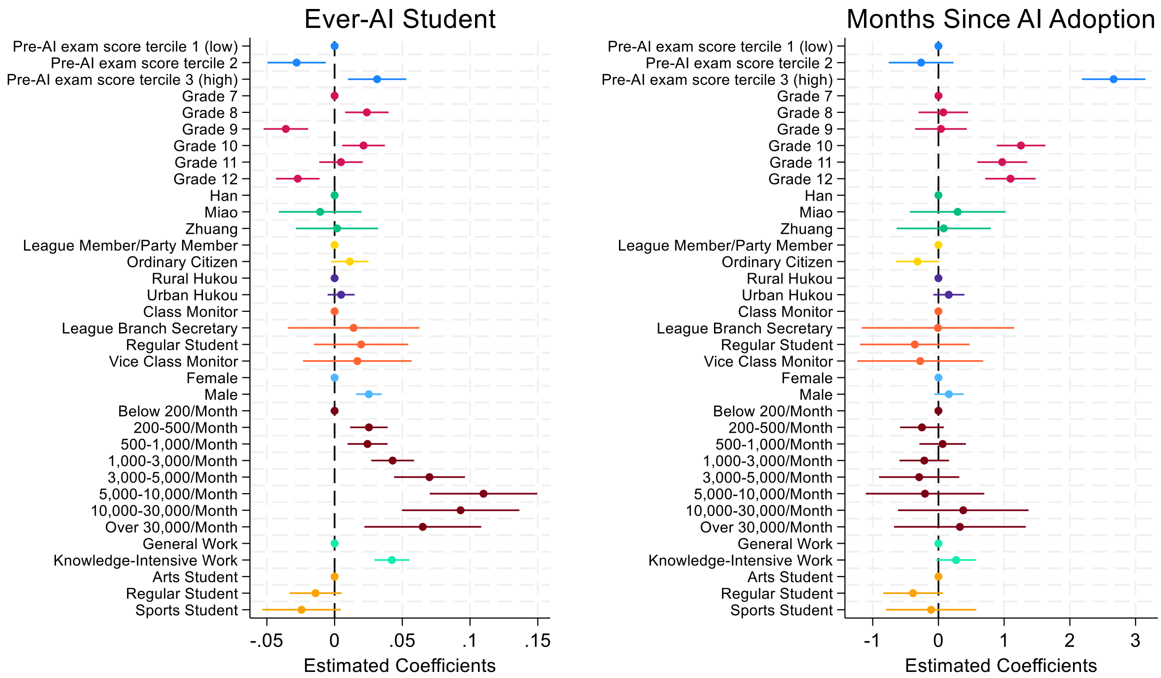
Appendix

Figure A1: Balance Table



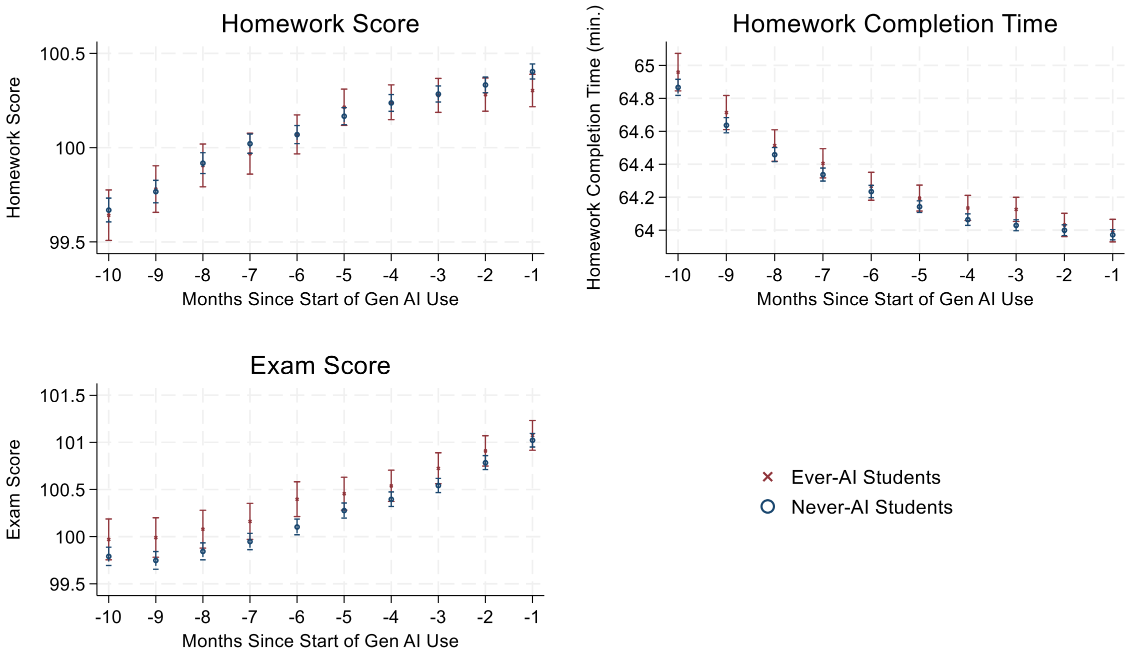
Note: The graph shows the standardized difference in means between students who have started using generative AI in June 2025 (Ever-AI Students) and those who have not (Never-AI Students). The data is at the student level and includes all 26,811 students, except for the last entry, which is based on the 8,466 students who are in 12th or 9th grade in 2025 and for whom we have exam scores for September 2022.

Figure A2. Bivariate Regressions on Student Characteristics



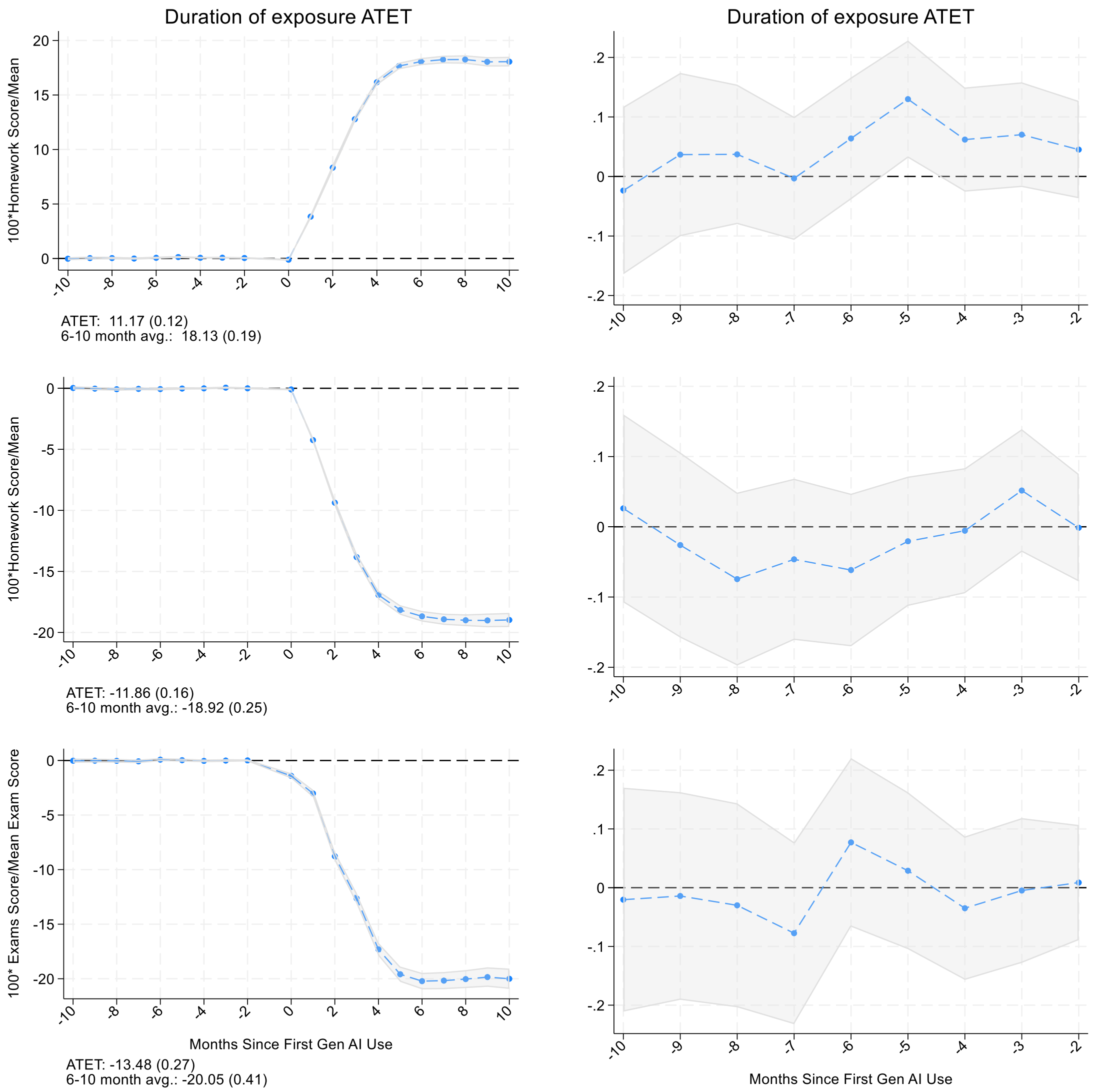
Note: The table shows the estimated coefficients and 95% confidence intervals from OLS-regressions. The dependent variable in the left graph is an indicator for an ever-AI student. The dependent variable in the graph to the right is the number of months since AI adoption in June 2025 for the sample of ever-AI students. Regressions were run separately for each categorical variable (color coded) and the plots include the base category at zero. The data is at the student level and includes all 26,811 students, except for the indicators for the tercile in pre-AI exam score distribution (September 2022), 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.

Figure A3. Pre-period Outcome Means



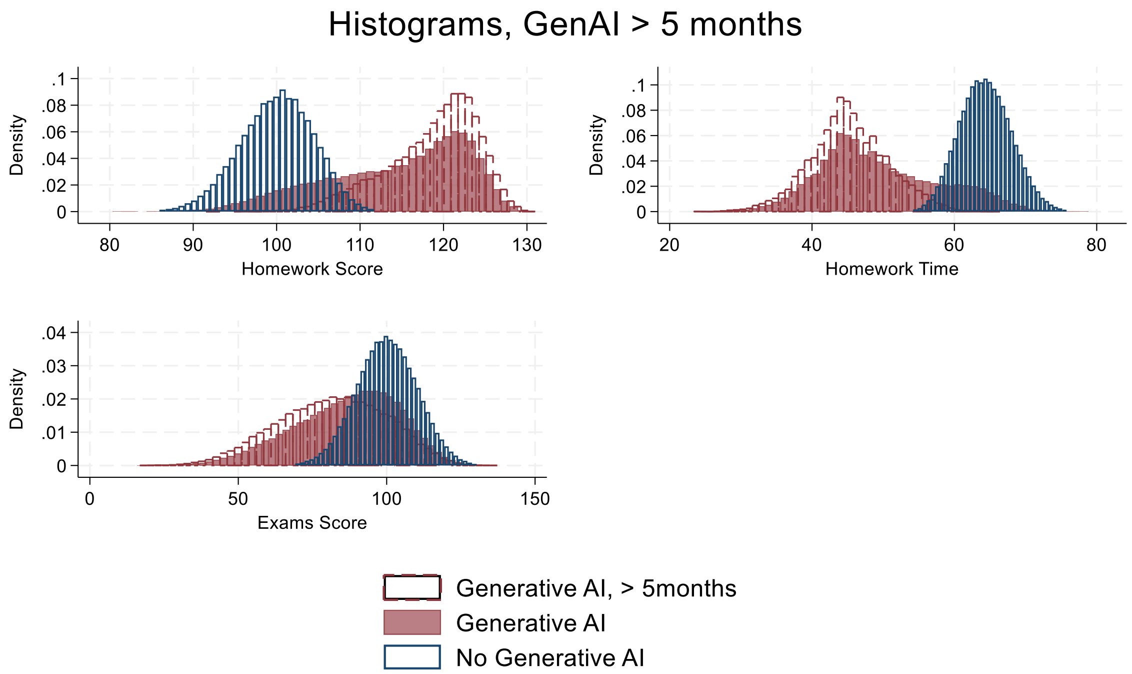
Note: The figures plot the mean outcomes in the pre-adoption period by event time for students who eventually adopt generative AI before adoption (Ever-AI Students). For the Never-AI Students, we construct a reweighted mean comparison series so that, at each event time, the control observations have the same distribution over month, grade, and school as the corresponding Ever-AI observations. The data is at the student-month level and the samples are shown in Table 1.

Figure A4. Event-study Graphs, Main Outcomes



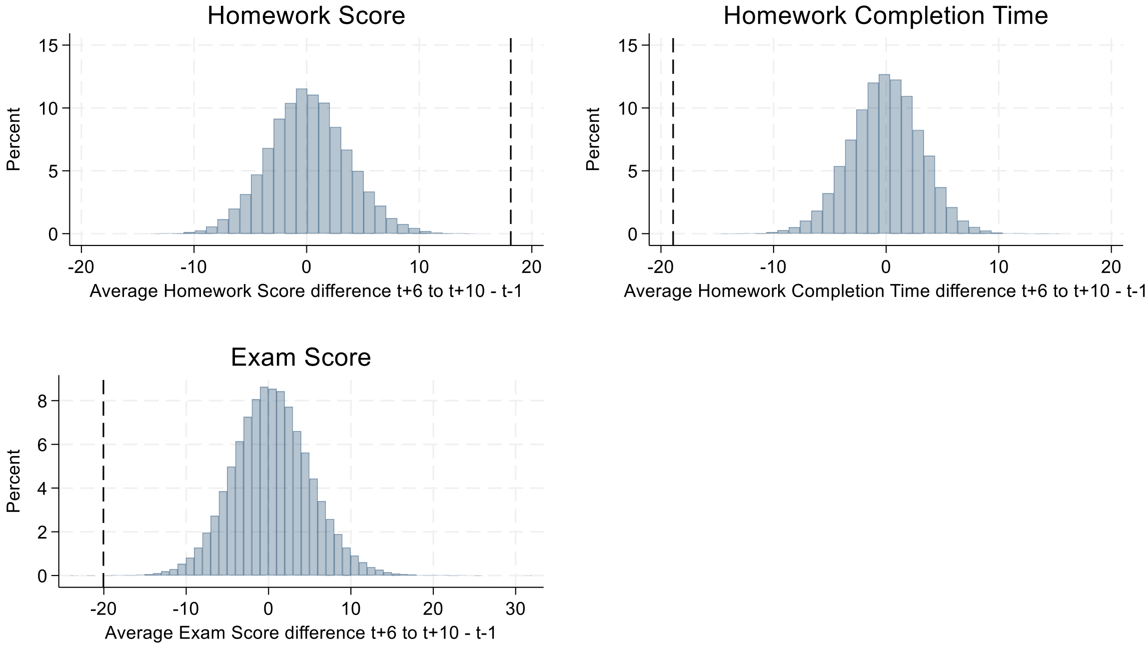
Note: The figure plots the average treatment effects by event time (and their 95% confidence intervals) of generative AI adoption on the outcome variables listed in the panel headings, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator and with never-AI students as the control group. The graphs to the right focus on the pre-period or the graph to the right in the same row. 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 is reported below the corresponding graphs. Standard errors are clustered by class id.

Figure A5. Histograms for >5 months GAI Use



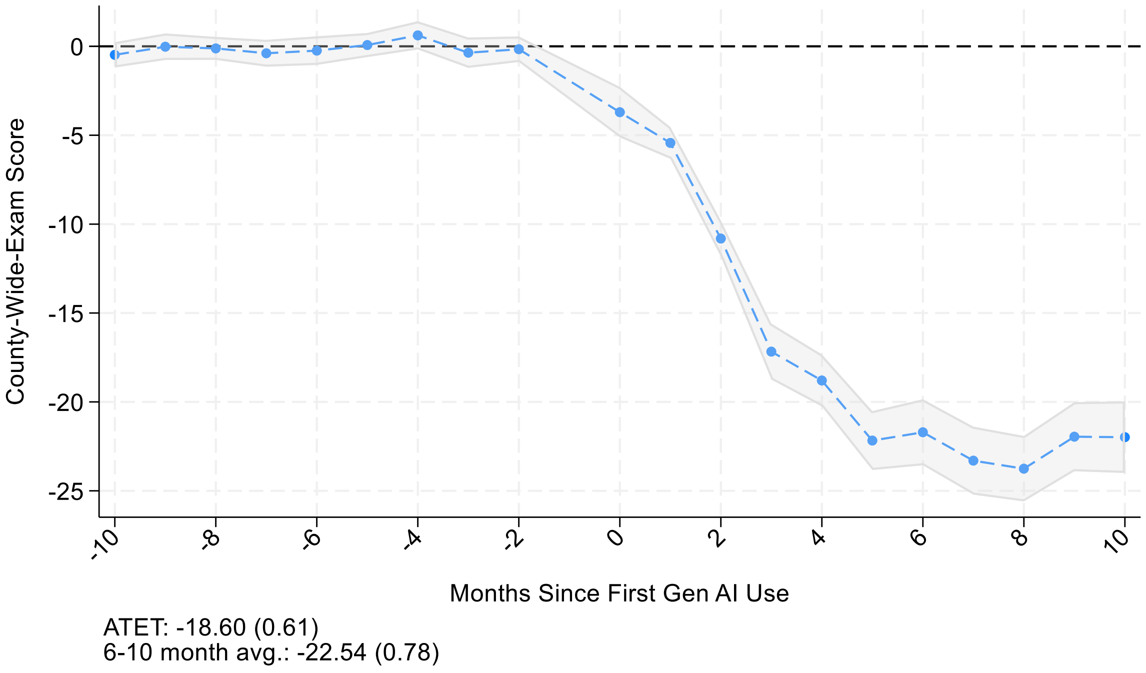
Note: The figure plots the distributions of homework score, homework completion time and regular exam scores, for three groups of students: (i) those who have adopted generative AI, (ii) those who adopted generative AI more than 5 months ago, and (iii) those who have not adopted generative AI. The data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

Figure A6. Natural Variation in Difference between Outcomes t-1 and Average Outcomes t+6 to t+10.



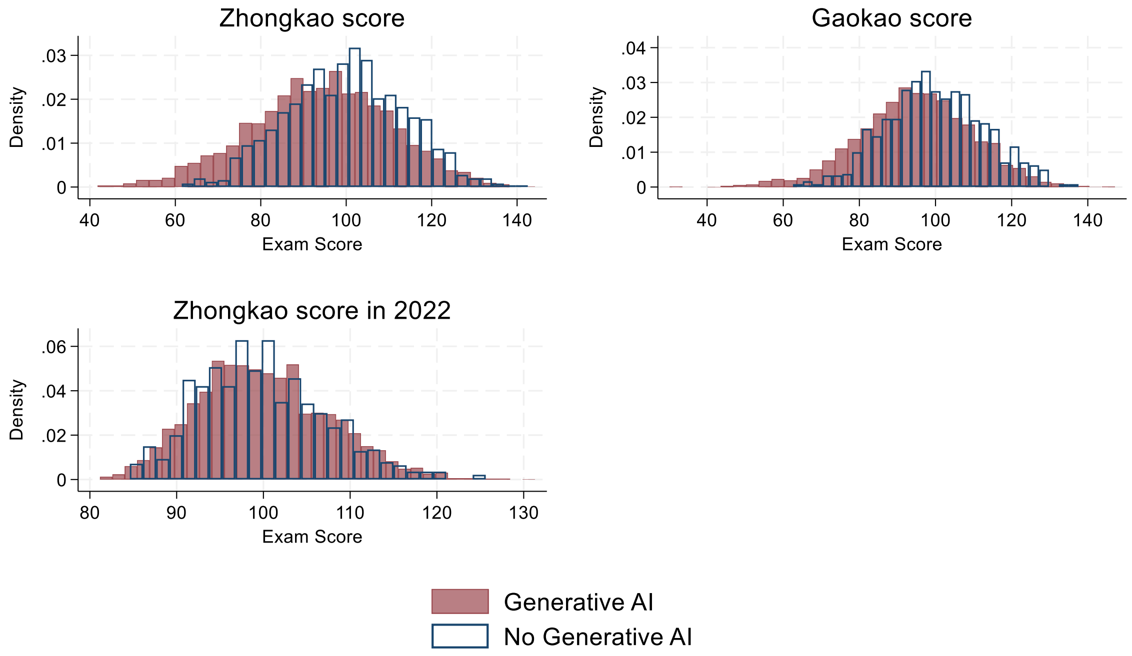
Note: Our main outcome is the difference between the outcome in t-1 and the average outcome in months t+6 to t+10, where t is the month the student adopted generative AI. We now compute the same statistic for every month t including all student-month observations not affected by generative AI adoption (including all month observations for ever-AI students and all months up until 11 months before generative AI adoption for ever-AI students). These differences are then plotted to show the natural variation in our outcomes. This natural variation is very small compared to effect sizes. For example, it is extremely rare that this variable is -12 or below for exam scores (4/212,370). The plots also include the estimated effects (dotted lines).

Figure A7. Effects on County-wide Exams by Event Time



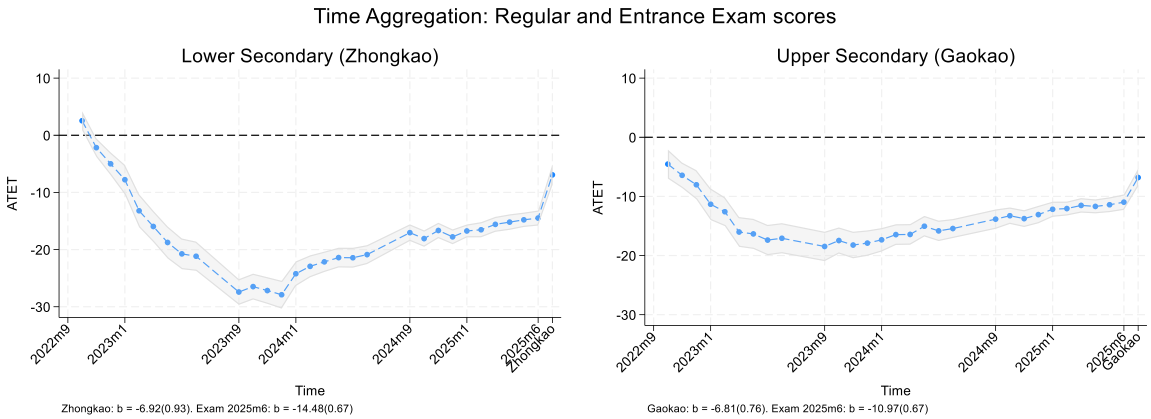
Note: The figure plots the average treatment effects by event time (and their 95% confidence intervals) of generative AI adoption on the county-wide exam score, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator. The total ATT and the average effects 6–10 months post-treatment 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 A8. Entrance Exam Histograms



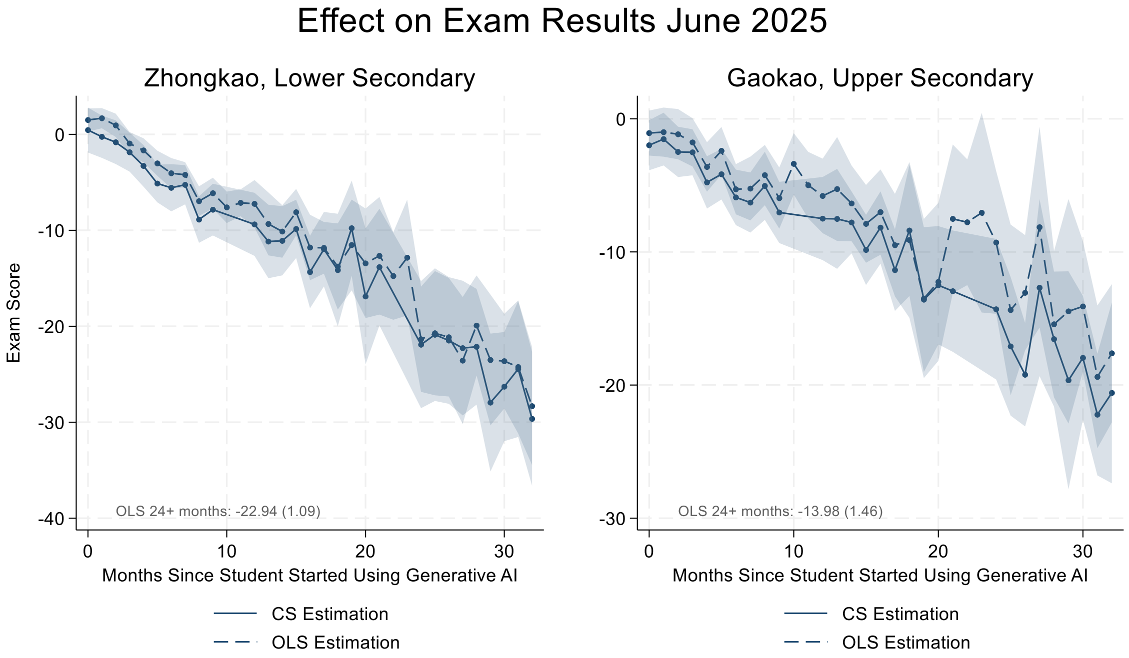
Note: The figure plots the distribution of entrance exam scores, separately for AI students and non-AI students. The Zhongkao score in 2022 is included to show that AI and non-AI students are not selected based on pre-AI entrance exam scores. The data is at the student level and the sample summary statistics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

Figure A9. Time Aggregation: Regular and Entrance Exams



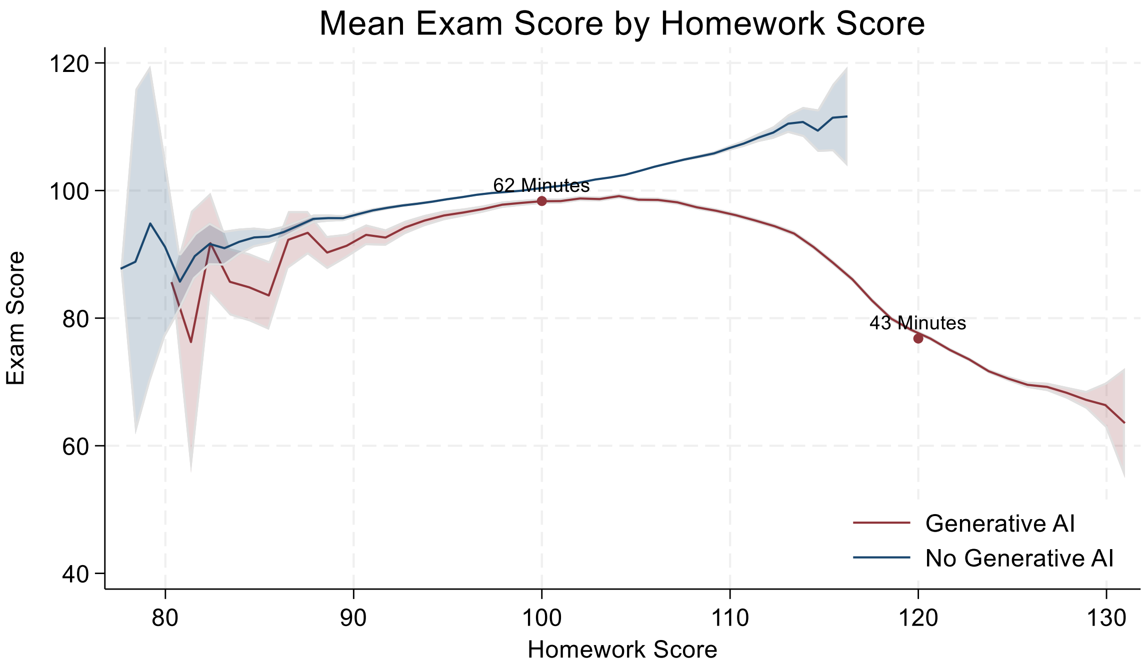
Note: The figure plots the estimated average treatment effects on the treated and 95% confidence intervals of generative AI adoption on regular exam scores and entrance exam scores in June 2025, computed using the Callaway and Sant'Anna (2021) regression adjustment estimator. The total ATT and the average 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



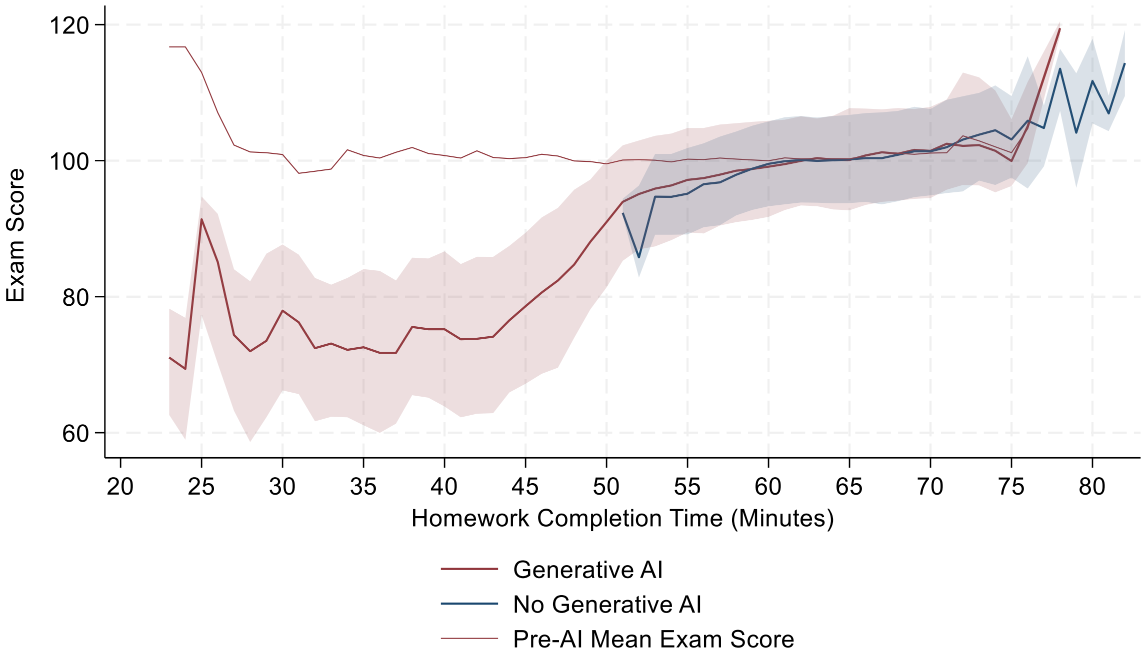
Note. The figure shows the the estimated average treatment effects and the 95% confidence intervals by event time computed using the Callaway and Sant'Anna estimator (CS Estimation) and using an OLS regression, controlling for class id and our demographic controls.

Figure A11. Homework and Exam Scores



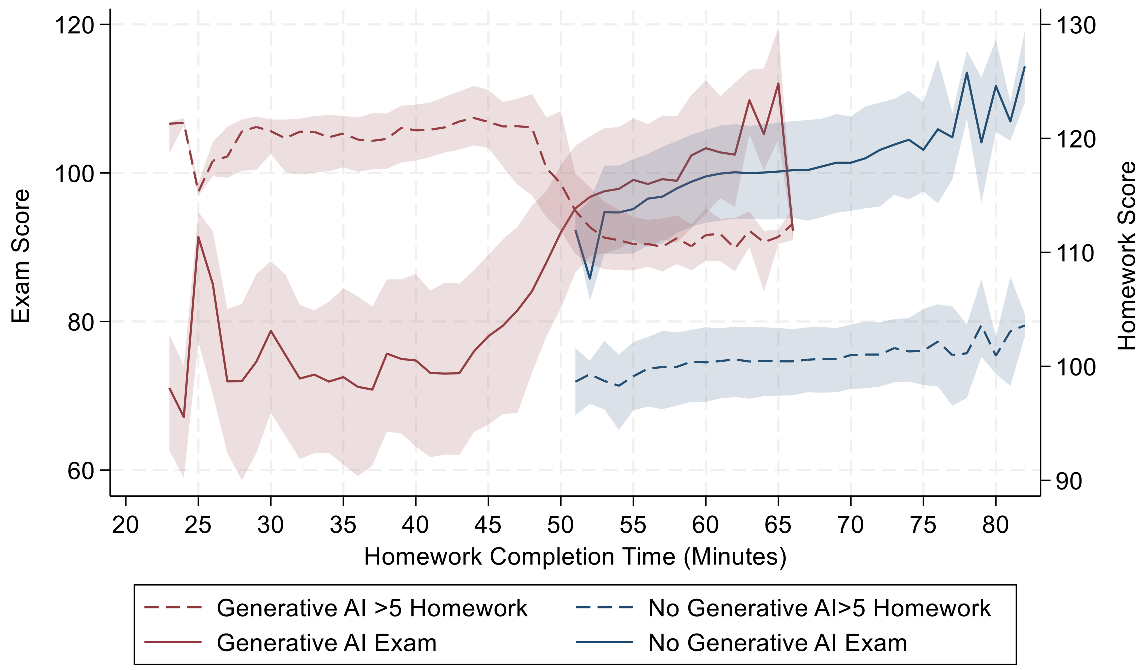
Note: The figure plots the smoothed mean and its 95% confidence intervals, by homework score, separately for students who have adopted generative AI and those who have not. The data also shows the average homework completion time for AI students with homework scores of 70 and 90. The data is at the student-month level and the sample summary statistics are shown in Panel A of Table 1.

Figure A12. Median and interquartile range of exam scores by homework completion time.



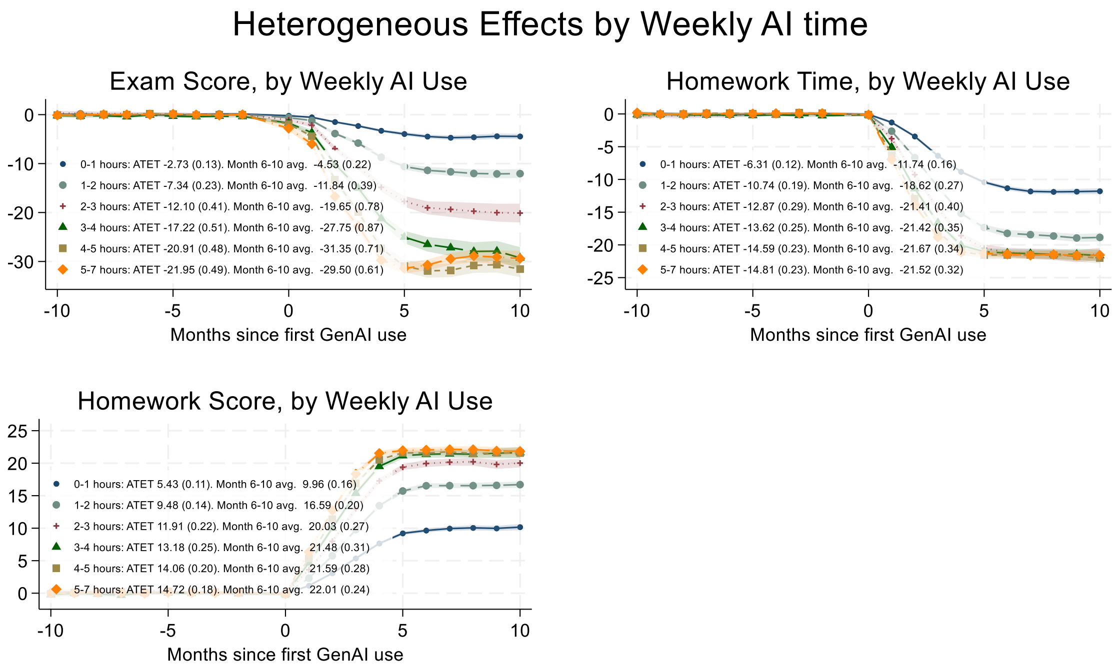
Note: The figure plots the median and interquartile ranges of exam scores by homework completion time, separately for AI students and non-AI students. It also plots the mean exam score for the AI students before the month they start using generative AI. The data are at the student-month level and the 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 A13. 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 A14a. Heterogeneous Effects by Reported Weekly AI-use Time



Note: The figure plots the estimated average treatment effects on the treated (and 95% confidence intervals) of generative AI adoption on regular exam scores, homework completion time and homework score, 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.

Figure A14b. Heterogeneous Effects by Educational Level

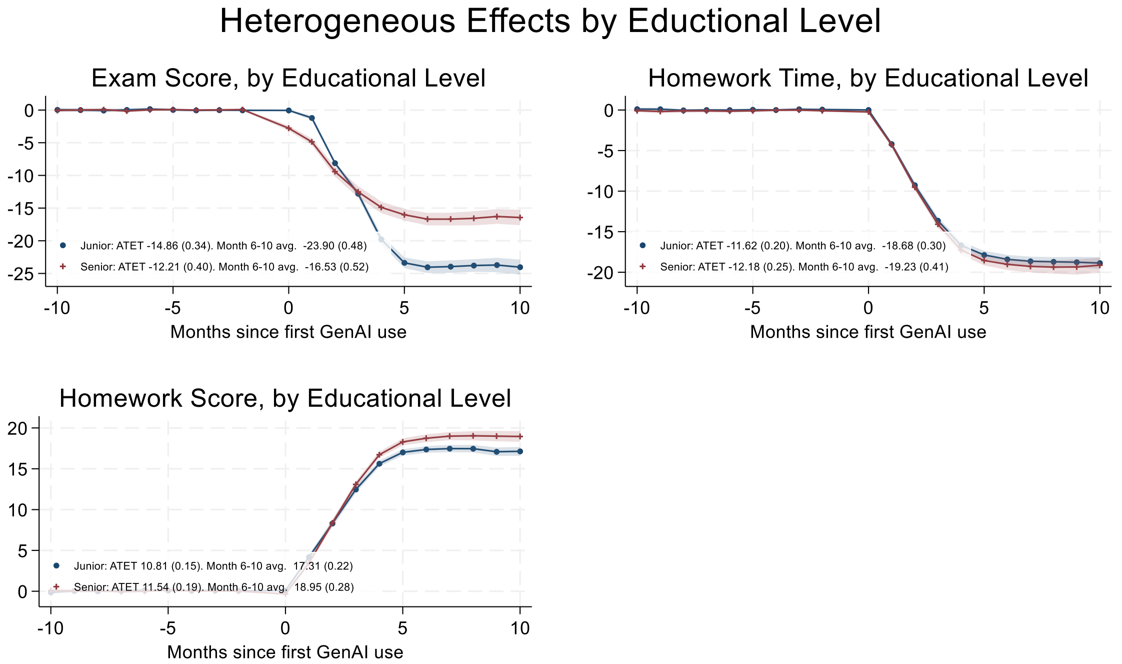
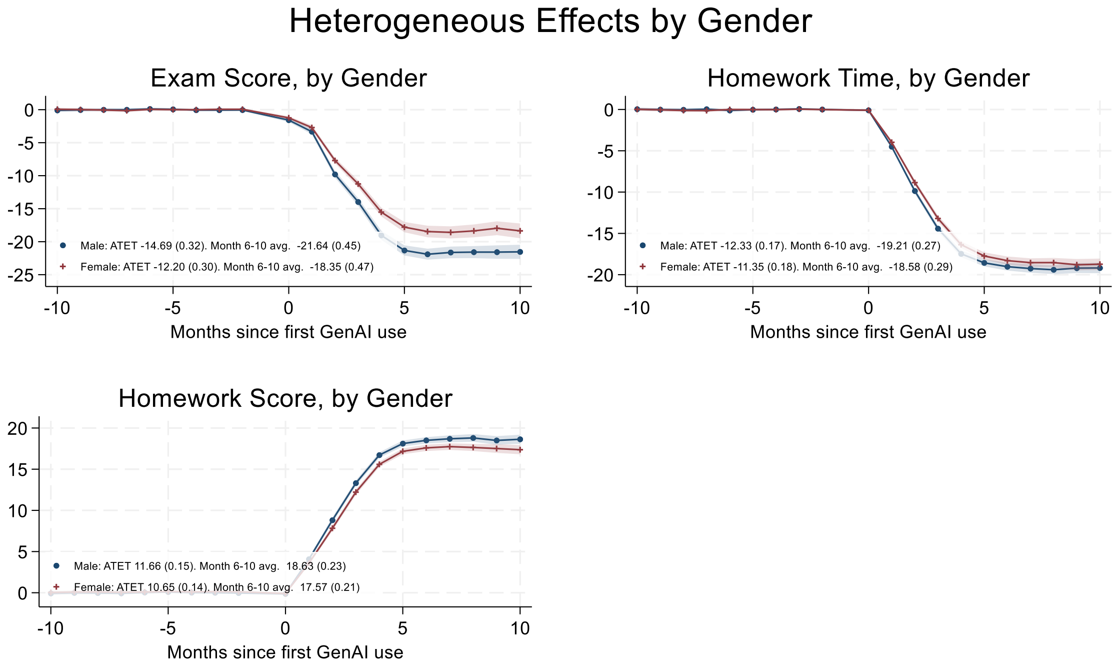
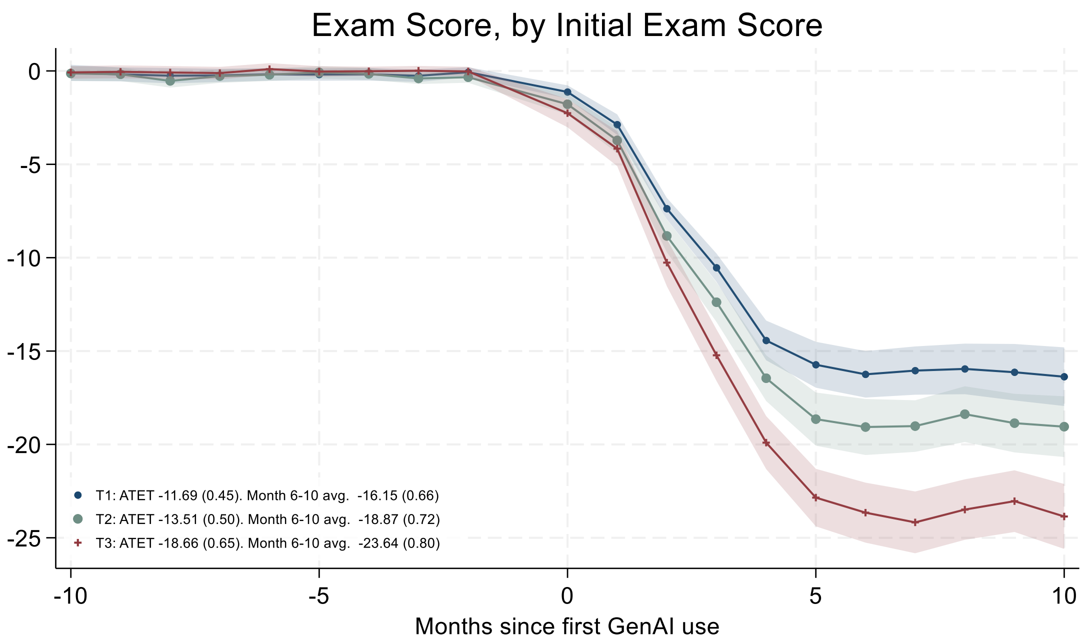


Figure A14c. Heterogeneous Effects by Gender



Note: The figure plots the estimated average treatment effects on the treated (and 95% confidence intervals) of generative AI adoption on regular exam scores, homework completion time and homework score, 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.

Figure A14d. Heterogeneous Effects by September 2022 Regular Exam Score Tercile



Note: The figure plots the estimated average treatment effects on the treated (and 95% confidence intervals) of generative AI adoption on regular exam scores, homework completion time and homework score, 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

Panel A: Homework Data - 18,345 students, 20 months					
Variable	**Obs.**	**Mean**	**SD**	**Min**	**Max**
Homework Score	272 140	106.08	9.59	77.62	130.98
Homework Time (Minutes)	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					
Variable	**Obs.**	**Mean**	**SD**	**Min**	**Max**
Regular Exams Score	526 120	95.47	14.68	15.88	139.75
AI Student	526 120	0.33	0.47	0.00	1.00

Panel C: County-Wide Exam Data - 8,466 students, 18 months					
Variable	**Obs.**	**Mean**	**SD**	**Min**	**Max**
County-Wide-Exam Score	89 059	95.52	16.70	9.32	147.20
AI Student	89 059	0.23	0.42	0.00	1.00

Panel D: Entrance Exam Data - 8,341 students, 1 month					
Variable	**Obs.**	**Mean**	**SD**	**Min**	**Max**
Zhongkao/Gaokao score	8 341	95.61	15.39	30.17	147.11
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	-6.070***	-5.198***	-4.521***	-5.214***	-5.309***	-5.521***	0.104
	(1.221)	(0.378)	(0.468)	(0.535)	(0.623)	(0.687)	(0.351)
AI Student * Medium Initial Exam Score			0.530			0.121	
			(0.804)			(0.557)	
AI Student * High Initial Exam Score			-2.705***			-0.007	
			(0.795)			(0.740)	
Pre AI Exam Score		0.920***	0.876***		0.894***	0.323***	
		(0.022)	(0.041)		(0.040)	(0.039)	
Observations	4,034	4,034	4,034	4,307	4,260	4,260	4,260
R-squared	0.025	0.563	0.565	0.021	0.256	0.473	0.000
Mean Dep Var	95.287	95.287	95.287	95.908	95.908	95.908	99.990
SD Dep Var	16.028	16.028	16.028	14.765	14.765	14.765	7.516
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 entrance exam scores (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	98.544	9.847	99.347	11.416	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
Athlete 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 12th or 9th grade in 2025 and for whom we have exam scores for September 2022.

Table A2 Robustness across Regression Specifications

Dependent variable	Estimator	Control Group	Controls	Total ATET	SE ATET	Average 6-10 month effect	SE 6-10
Exam Score	ols			-10.38	0.24	-20.06	0.40
Exam Score	ra	Never-AI Students	No	-13.48	0.27	-20.05	0.41
Exam Score	ra	Never-AI Students	Yes	-13.45	0.28	-19.99	0.41
Exam Score	ra	Not yet treated	No	-13.48	0.27	-20.05	0.41
Exam Score	ra	Not yet treated	Yes	-13.45	0.28	-19.99	0.41
Exam Score	twfe	Never-AI Students	No	-13.29	0.28	-19.81	0.41
Exam Score	twfe	Never-AI Students	Yes	-13.42	0.27	-19.91	0.40
Exam Score	twfe	Not yet treated	No	-13.29	0.28	-19.81	0.41
Exam Score	twfe	Not yet treated	Yes	-13.42	0.27	-19.91	0.40
Homework Score	ols			7.29	0.13	18.11	0.16
Homework Score	ra	Never-AI Students	No	11.17	0.12	18.13	0.19
Homework Score	ra	Never-AI Students	Yes	11.18	0.12	18.13	0.19
Homework Score	ra	Not yet treated	No	11.17	0.12	18.13	0.19
Homework Score	ra	Not yet treated	Yes	11.18	0.12	18.13	0.19
Homework Score	twfe	Never-AI Students	No	11.11	0.12	18.19	0.18
Homework Score	twfe	Never-AI Students	Yes	11.16	0.12	18.13	0.18
Homework Score	twfe	Not yet treated	No	11.11	0.12	18.19	0.18
Homework Score	twfe	Not yet treated	Yes	11.16	0.12	18.13	0.18
Homework Time	ols			-7.94	0.16	-18.76	0.23
Homework Time	ra	Never-AI Students	No	-11.86	0.16	-18.92	0.25
Homework Time	ra	Never-AI Students	Yes	-11.86	0.16	-18.91	0.25
Homework Time	ra	Not yet treated	No	-11.86	0.16	-18.92	0.25
Homework Time	ra	Not yet treated	Yes	-11.86	0.16	-18.91	0.25
Homework Time	twfe	Never-AI Students	No	-11.84	0.16	-18.93	0.25
Homework Time	twfe	Never-AI Students	Yes	-11.87	0.16	-18.88	0.25
Homework Time	twfe	Not yet treated	No	-11.84	0.16	-18.93	0.25
Homework Time	twfe	Not yet treated	Yes	-11.87	0.16	-18.88	0.25

Note: The table shows the estimated total and 6-10 month average treatment effects on the treated of generative AI adoption on regular exam scores, homework completion time and homework score. The estimates are computed using an OLS-regression controlling for student and month fixed effects (ols); the Callaway and Sant'Anna (2021) regression adjustment estimator (ra) and the extended two-way fixed-effects model outlined in Wooldride (2025). The included controls are school-by-grade fixed effects and gender for the ra estimator, and grade fixed effects for the twfe estimator (the model includes many interaction terms and could not be estimated with all controls in the ra specification). Standard errors are clustered by class id.

Table A3. Testing for Spillovers

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
Share of AI Students	-1.087			-0.903		
	(1.133)			(2.388)		
AI Student		-10.378***	-10.405***		-10.378***	-10.111***
		(0.236)	(0.339)		(0.236)	(0.354)
Non-AI Student x Share of AI Students			-0.066			0.603
			(0.482)			(0.500)
Observations	103,600	526,120	526,120	103,600	526,120	526,120
R-squared	0.840	0.763	0.763	0.840	0.763	0.763
Level of peer group	Class	Class	Class	School x Grade	School x Grade	School x Grade

Notes. This table presents the results of TWFE regressions, with student- and month-fixed effects. The data is at the student-by-month level and the sample summary statistics are presented in panel B of Table 1, except for columns (1) and (4) which contain only the subsample of never-AI students. The dependent variable is the regular exam score and the variable "Share of AI Students" is the share of students at the specified level (class in columns 1-3 and school-by-grade in columns 4-6) that have adopted generative AI. Standard errors clustered by class are in parentheses. ∗p < .10, ∗∗p < .05, ∗∗∗p < .01.

Table A4. Demographic Variables

Variable	**Code for Category**
Panel A: Student Demographic Variables	
Gender	0 = Female; 1 = Male
Ethnicity	0 = Han; 1 = Minority Ethnicity
Political Status	0 = Ordinary Citizen; 1 = League Member
Urban/Rural Classification	0 = Rural Hukou; 1 = Urban Hukou
Class Cadre	0 = Regular Student; 1 = Class Cadre
Parents' Occupation	0 = General Work; 1 = Knowledge-Intensive Work
Special Student	0 = Regular Student; 1 = Arts/Sports Student
Monthly Family Income	1 = Below 200; 2 = 200–500; 3 = 500–1,000; 4 = 1,000–3,000; 5 = 3,000–5,000; 6 = 5,000–10,000; 7 = 10,000–30,000; 8 = Over 30,000 (per month)
Panel B: Teacher Demographic Variables	
Gender	0 = Female; 1 = Male
Ethnicity	0 = Han; 1 = Miao / Zhuang (Minority)
Marital Status	0 = Unmarried; 1 = Married
Political Status	0 = Non-Party Member; 1 = Party Member
Education Level	0 = Bachelor's Degree; 1 = Master's Degree
Award History	0 = No Award History; 1 = Has Award History

Note: This table reports the coding scheme for the demographic variables used in the analysis. Panel A shows the student demographic variables; Panel B shows the teacher demographic variables.

Table A5. Summary Statistics for Subject-level Data

Regular Exam Scores

Subject	Obs.	Mean	SD	Min	Max
Biology	339 130	94.61	29.31	0.00	175.47
Chemistry	328 580	94.86	42.85	0.00	197.37
Chinese	526 120	97.81	20.60	42.28	149.54
English	526 120	96.33	24.62	8.61	160.65
Geography	268 060	92.50	29.33	0.00	165.19
History	342 730	94.35	31.61	0.00	162.44
Mathematics	526 120	94.84	26.23	0.00	157.19
Physics	442 110	95.33	30.20	0.00	172.00
Politics	347 940	93.70	28.30	0.00	157.30

Entrance Exam Scores

Subject	Obs.	Mean	SD	Min	Max
Biology	3 032	95.29	29.77	0.00	182.57
Chemistry	7 106	95.66	48.03	0.00	227.18
Chinese	8 341	96.72	21.87	23.93	157.63
English	8 341	96.64	24.91	4.72	181.15
Geography	1 677	91.75	28.13	0.00	154.42
History	4 772	94.11	32.34	0.00	177.59
Mathematics	8 341	94.91	28.72	0.00	165.57
Physics	7 603	96.55	36.32	0.00	203.50
Politics	4 867	94.32	28.88	0.00	172.76