NET WORK LISER

# Discussion Paper Series

IZA DP No. 19072

November 2026

# From Cheap Answers to Lasting Learning: AI Tutoring and Educational Inequality

**Hai-Anh H. Dang**
World Bank, IZA@LISER, GLO,
Indiana University, LSE and UEH

**Jere Behrman**
University of Pennsylvania

The IZA Discussion Paper Series (ISSN: 2365-9793) ("Series") is the primary platform for disseminating research produced within the framework of the IZA@LISER Network, an unincorporated international network of labour economists coordinated by the Luxembourg Institute of Socio-Economic Research (LISER). The Series is operated by LISER, a Luxembourg public establishment (établissement public) registered with the Luxembourg Business Registers under number J57, with its registered office at 11, Porte des Sciences, 4366 Esch-sur-Alzette, Grand Duchy of Luxembourg.

Any opinions expressed in this Series are solely those of the author(s). LISER accepts no responsibility or liability for the content of the contributions published herein. LISER adheres to the European Code of Conduct for Research Integrity. Contributions published in this Series present preliminary work intended to foster academic debate. They may be revised, are not definitive, and should be cited accordingly. Copyright remains with the author(s) unless otherwise indicated.

IZA DP No. 19072 November 2026

# From Cheap Answers to Lasting Learning: AI Tutoring and Educational Inequality*


## Abstract

Nearly half of 15-year-old students in OECD countries, and more in some poorer countries, use AI weekly for learning. Generative AI makes schoolwork help cheap for connected students, but cheaper help need not yield lasting learning. We link the economics of supplementary education with learning science to examine how effort, entry costs, public provision and assessment shape AI tutoring's benefits. Randomized trials find gains and losses varying by setting, use and outcome, while lasting, equitable gains remain uncertain. We distinguish program effects from generative AI's contribution, compare full costs with alternatives, and propose decision rules and a trial design.



## JEL classification

I21, I24, I28, H52, O33

## Keywords

artificial intelligence, tutoring, learning, educational inequality, public investment, randomized trials



## Corresponding author

Hai-Anh H. Dang
hdang@worldbank.org


* We would like to thank Gero Carletto and various colleagues at and outside the World Bank for helpful discussions. We would like to thank Minh Do for helpful research assistance.

## 1. Introduction

Generative artificial intelligence (AI) is changing how students get help with their schoolwork. In 2025, on average across Organisation for Economic Co-operation and Development (OECD) countries, 45.5% of 15-year-old students reported using AI at least weekly to help them learn, as did 68.6% in Viet Nam and 45.7% in Kenya (OECD 2026). For students who are online, a chatbot can explain concepts, check solutions or draft essays at a money price close to zero. Help that families have long bought from private tutors, often at high prices and mostly in better-off households (Bray 2006; Dang and Rogers 2008; Supplementary Table S1, Panel B), is now available on demand.

Cheaper help does not by itself mean cheaper skills. AI assistance can raise performance on tasks at hand while leaving students less able to work without it (Bastani *et al.* 2025; Yan *et al.* 2025). Yet supported AI programs have also raised test scores in schools in Nigeria and Sierra Leone (De Simone *et al.* 2025; LearnLM Team and Fab AI 2026). Whether AI narrows or widens educational inequality therefore depends on more than access. It depends on which students use the tools in ways that build skills, on what the tools replace, and on whether public provision of AI reaches the students whom private tutoring has left out.

These questions are pressing because governments are deciding about AI in school now, and some have changed course. New York City blocked ChatGPT on school devices in 2023, soon eased the block, and paused student-facing AI through grade 8 again in 2026 (Banks 2023; Elsen-Rooney 2023; K-12 Dive 2026). Estonia and the United Arab Emirates, by contrast, expanded AI in their schools (Estonian Ministry of Education and Research 2025; Gulf News 2025). Such choices commit public money and students' time, yet the causal evidence behind them is thin, even

in high-income countries. Of 818 studies relevant to AI in school education collected by October 2025, only 20 met a standard for strong causal evidence (Fesler *et al.* 2026).

We use "AI tutoring" for student-facing generative AI used for learning: general-purpose chatbots used for schoolwork and tutors built on large language models (LLMs), with or without guardrails or human support. Earlier adaptive software serves as a benchmark. Teacher-facing tools and AI in assessment are connected applications. We discuss them where they change what students learn or how learning is measured, but do not review them in the same depth.

Here we argue that public provision can widen access to AI tutoring without narrowing learning gaps. We connect learning evidence with household responses and full costs, monetary and nonmonetary, to identify when that provision is worthwhile, and for whom. Three questions organize our Perspective: (1) Does a program build lasting skills, not only assisted performance? (2) Which students benefit, once entry costs and household responses are counted? (3) Is the program preferable to feasible alternatives at full cost?

Recent reviews cover generative AI from the perspectives of learning science and human–computer interaction (Yan *et al.* 2024). Others synthesize experiments on technology in education (Escueta *et al.* 2020; Burneo *et al.* 2026), on tutoring (Nickow *et al.* 2024) and on AI in schools (Fesler *et al.* 2026). We add an economic and institutional treatment, starting from the supply-and-demand model of private tutoring in Dang and Rogers (2008). It covers the nonmonetary price of help, entry costs, public provision and household responses to it, the instruction that AI replaces, and assessment incentives. We also read recent trials by what was randomized, in line with audits of the field (Weidlich *et al.* 2025; Lawson *et al.* 2026).

Section 2 separates access to AI, its use and its learning returns, and Section 3 develops the framework. Section 4 reviews the trials, which bear on the first two questions, and Section 5 turns

to full costs and equity, which bear on the second and third. Section 6 proposes conditional decisions for governments and a comparative trial, and Section 7 concludes.

## 2. Who uses AI for learning, and why

Three inequalities are often merged. Unequal access concerns devices, connectivity and electricity. Unequal use concerns how often students use AI and for what. Unequal learning returns concern how much a given use teaches. Each requires different evidence and different policies.

The OECD's Programme for International Student Assessment (PISA) 2025 offers the first internationally comparable measure of use (Supplementary Table S1). Weekly use exceeded 50% in Singapore, the United Arab Emirates, Estonia and (South) Korea, and in several middle-income systems, including Viet Nam, Colombia, Peru and Indonesia (OECD 2026). PISA samples 15-year-olds enrolled in school, so it describes enrolled students and says little about those out of school.

Use also differs within countries. Across OECD countries, advantaged students (those in the top quarter of PISA's socio-economic index in each system) used AI for schoolwork more than disadvantaged students, on average by about a fifth of a standard deviation on the PISA index of such use. Two-thirds of advantaged students, against 58% of disadvantaged students, also reported assessing AI-generated information in their lessons. Both gaps are positive and statistically significant in most systems with data and significantly negative in none (Supplementary Table S1; OECD 2026). In the United States, 54% of teenagers have used chatbots for help with schoolwork. Doing all or most schoolwork with that help is reported by 20% of teenagers in households earning less than US$30,000 a year, against 7% in households earning US$75,000 or more (Pew Research Center 2026). Such reliance measures use, not learning; what matters for learning is whether it replaces students' own work (Section 3).

Cross-sectional associations between use and achievement do not show effects. In PISA 2025, students who used chatbots for schoolwork such as summarizing, preliminary research or drafting scored about 20 points lower in science than students who did not. Students who used chatbots to help them learn once or twice a week scored about the same as non-users, on average across the OECD (Schleicher 2026). Both patterns could reflect selection, for example if weaker students rely on AI more often.

Access sets the outer limit. In low-income countries, 23% of people use the internet, against 94% in high-income countries (ITU 2025). Worldwide, about 655 million people still lack electricity (Tracking SDG7 2026). PISA focuses on students' access mostly in middle- and high-income countries and does not show how much of the gap in use between countries reflects infrastructure.

Demand for AI help plausibly rises with examination pressure, with teacher shortages, estimated at 44 million teachers by 2030 (UNESCO 2024), and with learning poverty—the inability to read and understand a simple text by age 10—which affected an estimated 70% of 10-year-olds in low- and middle-income countries (World Bank 2022). Lower income need not mean that families value education less. Differences in guidance, school resources, language support and the opportunity cost of time are more plausible reasons for unequal use. Supply comes from free tiers, public procurement and a few global platforms, and restrictions on private tutoring may shift demand toward AI. Sales of learning tablets in China reached 6.3 million in 2025, four years after curbs on for-profit tutoring (Caixin Global 2026), although the timing alone does not show that the curbs caused the rise. Viet Nam restricted paid tutoring from February 2025 (Ministry of Education and Training of Viet Nam 2024), and the effect on AI use is not yet known.

## 3. A framework: what AI changes in the market for supplementary learning

We start from a framework depicting the market for supplementary learning (Figure 1a), adapted from Dang and Rogers (2008). The vertical axis is the full marginal price of a unit of learning to households, in money, time and effort, and the horizontal axis is the quantity of learning. Public schooling ($S_1$) supplies learning at a low price to households up to its capacity, point A, beyond which it cannot offer students more learning in the short run. Because public schooling is subsidized, households pay less than the cost of this schooling to society, whereas they pay the full marginal cost of tutoring in a competitive market (Supplementary Note 1). Households therefore use public schooling first, and private tutoring adds a supply curve ($S_2$) that starts at A and relaxes the capacity constraint. $S_2$ is steeper than the lower part of $S_1$: beyond A, tutoring costs households more than public schooling would if schools could expand.

Households with a low demand for education ($D_1$) are satisfied by public schooling and consume $Q_1$. These low-demand households do not buy tutoring, but high-demand households ($D_2$) buy tutoring beyond what schools provide and consume $Q_2^*$. Because demand, which reflects the ability to pay as well as preferences, rises with income, better-off households buy more tutoring.

AI tutoring can add a flatter supply curve ($S_3$), but only under certain conditions. Its marginal money cost is close to zero, yet cheap assistance does not by itself establish a lower price per unit of learned skill. $S_3$ lies below $S_2$ only if use turns assistance into learning. Where $S_3$ lies below $S_2$, a high-demand household consumes more learning, moving from $Q_2^*$ to $Q_2^{**}$. The surplus gain over private tutoring ($\Delta$ in Figure 1a) is conditional. With capacity at A, public schooling and its costs are the same in both cases and cancel out, but $\Delta$ is measured before the entry and public-provision costs of AI (Section 5). Households keep all of it only if AI providers charge the near-zero marginal cost (Supplementary Note 1).

Entry costs decide who takes up AI tutoring (Figure 1b). A device (and connection to a local server) requires mostly one-off costs, whereas data and electricity recur and add to the full marginal price. A household adopts when the additional surplus from AI tutoring, over its best existing option, covers the additional entry cost. School-provided devices and offline tools lower that cost, partly by shifting it to the public budget, and can bring in households with moderate demand, whom private tutoring never reached.

Public provision also changes what households do. In a randomized after-school tutoring program in rural China, home tutoring fell for children living with a parent, who also gained less in mathematics from the program than children left in the care of grandparents, who provided less home tutoring before the program (Behrman *et al.* 2024). Program effects are thus net of household responses, which can differ by family background. The same may hold for AI provided at school, although none of the AI trials we review measured such responses.

Assistance is not necessarily learning (Figure 1c). Assistance can improve immediate performance without improving subsequent unaided performance. Learning measured without the tool depends on whether use preserves productive cognitive activity, such as generating explanations, retrieving from memory and attempting problems before receiving hints (Chi and Wylie 2014; Yan *et al.* 2025). Assistance that removes unproductive work can save time and still improve learning. In a US physics course, students who studied with an AI tutor spent a median of 49 minutes, against about an hour in class, and scored higher on a post-test the same day (Kestin *et al.* 2025). Assistance that replaces productive work can lower learning below its level without AI, as unrestricted GPT-4 did in Türkiye (Bastani *et al.* 2025).

By effort we mean productive cognitive activity, not time spent or difficulty. Defined this way, the mechanism can be tested rather than assumed. Evaluations can record attempts before

hints, students' explanations, retrieval and the fading of support, and can specify these features before outcomes are observed. Assessment incentives act on the same margin. When Mexican high schools rewarded students for their mathematics scores, students studied more and scores rose, but copying also became more common (Behrman *et al.* 2015). AI lowers the cost of such shortcuts. Where assessments cannot separate assisted from unassisted work, measured performance will overstate learning.

The analogy with private tutoring breaks in three places. First, school-provided and teacher-facing tools can substitute for existing instruction, so their counterfactual is the instruction they replace rather than no tutoring. Second, the price of AI help is mostly nonmonetary. Third, supply is concentrated in a few global platforms, so previous concerns about corrupt teachers who tutor their own students for money become a question about providers' prices, defaults and use of data. AI could also raise public capacity (from A to A′ in Figure 1a), through teacher-facing tools or by reducing demand for the private tutoring that teachers sell, which can divert their effort from class (Jayachandran 2014; Supplementary Note 1). Teacher-facing tools can, however, lower the quality of teaching (Sungu *et al.* 2026).

The framework yields four hypotheses that organize Sections 4 and 5. First, unassisted gains from AI depend on whether AI use preserves productive activity and on the instruction it replaces. Second, AI take-up and benefits depend on entry costs, public provision and household responses. Third, AI-assisted performance overstates learning where assessment cannot separate the two. Fourth, AI providers' pricing affects how much of the gain households keep.

## 4. What the trials show

We organize the evidence by what was randomized (Figure 2; Supplementary Table S2; Supplementary Note 2). But several challenges exist. Differences across unrelated trials mix age,

subject, baseline skill, dose, comparator and implementation, so they do not clearly show which design works (Weidlich *et al.* 2025). Each trial also evaluates an offer of a particular implementation, not AI as a technology. In high-income countries, much of the evidence comes from short experiments: of the 14 strong causal studies of student-facing tools identified by Fesler *et al.* (2026), seven were university experiments in the United States and Germany. In general, across these studies, AI tended to raise performance while students had access, with mixed effects once access was removed.

**What the randomized comparisons establish.** Programs compared with usual instruction show what an AI-enabled package can achieve (Figure 2a). Several encouraging school trials come from poorer countries. In Sierra Leone, mathematics lessons supported by Gemini, held twice a week for eight weeks, raised scores by 0.26 standard deviations (SD) [95% confidence interval (CI) 0.03, 0.49] (LearnLM Team and Fab AI 2026). In Nigeria, a six-week after-school program raised English scores by 0.24 SD; 43% of students missed the endline test, but the estimates remain positive under bounds and reweighting for attrition (De Simone *et al.* 2025). Ghana's gain with a WhatsApp tutor was reported without an interval (Henkel *et al.* 2024). In Tennessee middle schools, a Khanmigo package had an estimated effect of 0.06 SD a year [−0.01, 0.13] over two years (Oreopoulos and Low 2026).

These programs combined AI with devices, facilitators, timetables and sometimes additional time. Their effects do not isolate what generative AI adds over the same provision without it. Trials that add AI to an otherwise unchanged setting come closer to identifying that contribution, but their effects vary by setting and outcome (Figure 2b). In Türkiye, unrestricted GPT-4 in practice sessions lowered closed-book exam scores by 0.19 SD [−0.35, −0.04], whereas a hint-only tutor showed no detectable difference from practice without AI (Bastani *et al.* 2025).

Practice scores rose with both forms of access (Box 1). By contrast, US college students allowed generative AI alongside traditional resources while learning a new topic scored 0.27 SD higher on a proctored test a week later (Contractor and Reyes 2026). Adding an AI tutor to mastery-based practice in Tennessee made small and imprecise differences a week later (Oreopoulos *et al.* 2026). At a Maryland university, offering a course chatbot had an estimated effect of −0.29 SD [−0.60, 0.03] on final grades, which may include AI-assisted work. The estimate covers all students offered the chatbot, of whom about 15% used it; the authors conjecture that its presence changed AI use in the course more widely (Liu *et al.* 2026).

Comparisons with other learning methods ask whether AI improves on an alternative use of students' time (Figure 2c). Results are mixed and mostly short-run. In a US university physics course, students scored higher after studying with an AI tutor than after an active-learning class (Kestin *et al.* 2025). The authors' estimate of 0.63 SD adjusts for time on task, although the AI tutor itself changed that time, so it is not a total effect and is not plotted. In a German laboratory study, unrestricted access to a textbook-grounded tutor raised test scores by 0.34 SD over textbook study alone (Fischer *et al.* 2025). Yet in English secondary schools, note-taking outperformed LLM use by 0.38 SD on comprehension and 0.44 SD on retention three days later (Kreijkes *et al.* 2026).

Randomizing how the same tool is used gives more direct evidence on design (Figure 2d). In Türkiye, the hint-only tutor scored about 0.18 SD above unrestricted access on the closed-book exam (Bastani *et al.* 2025). In Chile, guidance on learning-oriented use before the final examination raised grades by 0.22 SD [0.02, 0.41], while a separate encouragement to adopt the course chatbot before the midterm had no detectable effect on midterm grades (Gallegos 2026). These randomizations preceded different examinations and are not a head-to-head test. Nor does

greater restriction necessarily improve learning: in the German laboratory study, unrestricted access scored 0.21 SD [−0.01, 0.42] above access delayed by ten minutes of reading (Fischer *et al.* 2025). Design can change outcomes with the same model, but the evidence does not yet establish which features work across settings.

**Persistence and the strength of the evidence.** The comparison is only part of the interpretation; the outcome and its timing also matter (Box 1). The short follow-ups discussed above provide limited evidence on retention and transfer. An endline gain does not establish persistence after support ends, and whether Khanmigo's first-year gains survived the summer cannot be determined (Oreopoulos and Low 2026).

Pooled estimates summarize the evidence but do not resolve these differences. Across eight comparisons in which AI tutors came with human support or guardrails, the pooled gain is 0.21 SD [0.05, 0.36]. Separate analyses excluding the two trials rated at high risk of bias or retaining only school trials each give 0.16 SD, with intervals that include zero (Dang and Do 2026). Burneo *et al.* (2026) find an average of 0.125 SD across 14 randomized evaluations of adaptive and AI-enabled software. Their estimated difference between generative and earlier tools, 0.022 SD (not statistically significant), compares studies, not randomized arms. These syntheses overlap with the trials discussed here and with each other (Supplementary Note 2). For scale, tutoring raised learning by 0.29 SD across 89 trials (Nickow *et al.* 2024; Figure 2e), and the median effect in education trials is about 0.10 SD (Kraft 2020). These are benchmarks, not a ranking of alternatives.

Inclusion criteria and outcome definitions also help explain why syntheses differ. Education-journal meta-analyses report large average effects on performance, such as 0.71 SD across 51 effect sizes (Deng *et al.* 2025); one reporting 0.87 SD was retracted in April 2026 (Wang and Fan 2025). Such syntheses can apply lenient inclusion criteria and treat performance with the

tool as learning (Bauer *et al.* 2025). Audits find few clean comparisons (Weidlich *et al.* 2025; Lawson *et al.* 2026), and ratings on pedagogical benchmarks are not learning outcomes (Jurenka *et al.* 2025).

**Who benefits within trials.** Average effects leave the second question of Section 1 unresolved: which students gain? Within-trial interactions are more informative about equity than comparisons across trials, but few studies report effects by gender, age or family background. Their measures also differ, so we retain each trial's definitions rather than compare subgroup estimates across studies (Supplementary Table S3). In Sierra Leone, each additional SD of baseline mathematics score added about 0.2 SD to the effect (LearnLM Team and Fab AI 2026). In Nigeria, effects on a composite score rose with baseline exam scores and a household socioeconomic index; girls gained more than boys, although one girls-only school may drive that difference (De Simone *et al.* 2025). In Maryland, the estimated effect on final grades was more negative for first-generation students (Liu *et al.* 2026). German laboratory experiments found that ChatGPT access widened the gap by prior knowledge (Lehmann *et al.* 2025).

Baseline achievement, first-generation status and household socioeconomic status are distinct dimensions. In these selected student-facing trials, estimates favor students who start ahead on each, but the latter two dimensions each rest on one trial. These patterns concern the outcomes measured in particular settings, not a general conclusion that public provision must widen gaps. As Section 5 discusses, population-wide effects also depend on coverage, take-up and household responses.

**Teacher-facing tools.** AI can also affect students through the support given to their teachers and tutors. Real-time AI guidance for tutors raised topic mastery by 4 percentage points on average and by 9 points for students of lower-rated tutors (Wang *et al.* 2026). This is a

difference by tutor quality, not by students' baseline skill or family background. In England, ChatGPT and an AI lesson assistant cut preparation time by 31% and 24%. The first trial found no detectable difference in sampled resource quality and the second found it non-inferior, but neither measured students' learning (Roy *et al.* 2024, 2026). In Türkiye, a GPT-4o teaching assistant lowered students' intrinsic motivation by 0.11 SD, with no detectable effect on average grades (Sungu *et al.* 2026). Copilot training in Peru saved no detectable time but improved teachers' prompts (Glave *et al.* 2026). Teachers' workload, the quality of their work and students' learning are therefore separate outcomes.

Taken together, the trials speak most clearly to the first and third hypotheses of Section 3: how AI is used can change outcomes, and better assisted performance need not imply better unassisted learning. Both supported programs and less-restricted access can yield gains, while some uses cause harm. Evidence on the second hypothesis is incomplete: optional take-up was low in Maryland, but no included trial varied entry costs or measured household responses. The fourth hypothesis, on provider pricing, remains open. Section 5 therefore asks how these effects, their distribution and full costs compare with feasible alternatives, informing the conditional decisions and trial design in Section 6.

## 5. Costs and equity

Cost evidence is scarce and incomplete (Table 1, Panel A). Of the trials pooled by Burneo *et al.* (2026), only two have comparable per-student costs, and neither uses generative AI. The Nigerian trial, which that review excludes for attrition, adds a third. Reported figures also mix cost concepts: about US$5 per student (marginal) in Ghana, US$48 per student for the six-week Nigerian program (US$9 marginal, with electricity and infrastructure excluded), about US$15 per

student a year for a Khanmigo license, and none in Sierra Leone. None of these studies includes all of devices, connectivity, electricity and staff time, so these figures cannot rank programs.

Rather than rank programs by their reported prices, policymakers should ask whether an AI program offers better value than a feasible alternative, $B$. Compare both on the same learning outcome, from the same starting point and over the same period. Let $\Delta C = C_{AI} - C_B$ be the difference in full cost per student and $\Delta g = g_{AI} - g_B$ the difference in learning. The break-even rule is $\Delta C \leq \lambda \Delta g$, where $\lambda$ is what one more unit of comparable learning would cost through the best alternative use of funds locally, at the scale being considered. Put simply, AI's extra cost should be no greater than the cost of achieving its extra learning through other means. For a whole program, $B$ may be usual instruction; for AI added to an existing platform, $B$ is that platform without AI. The latter comparison measures what AI itself adds, which few trials estimate (Section 4). A cheaper program can also pass the test with a small learning loss if its savings exceed $\lambda$ times the learning lost, subject to the safeguards in Section 6.

Figure 1a illustrates the same idea. The height of $S_2$ at $Q_2^*$ marks $\lambda$: the cost of obtaining one more unit of learning through private tutoring in this example. The lower $S_3$ curve shows AI's potential to reduce that cost. From society's perspective, the comparison must also include devices, staff and other costs omitted from the curve. Students' time is already included. When AI increases learning, $\Delta C/\Delta g$ is its full additional cost per unit gained and must be no greater than $\lambda$ to pass the test. The shaded area $\Delta$ represents a potential gain before additional entry and public-provision costs. An alternative's average cost per SD can stand in for $\lambda$ only if the cost of further learning stays constant as provision expands, that is, if its supply curve is flat. Even with a common $\lambda$, the relevant alternative $B$ differs across households: those on $D_2$ already buy tutoring, whereas those on $D_1$ do not. The same program can therefore offer good value for one group but not the other.

Table 1, Panel B illustrates the resulting ceilings with three non-generative anchors, treating each anchor's average cost per SD as $\lambda$, as if its curve were flat. Phone tutoring in Botswana cost about US$157 per SD of numeracy (Angrist *et al.* 2022), and adaptive software in India about US$830 per SD of mathematics (Muralidharan *et al.* 2019; Burneo *et al.* 2026). Technology-supported human tutoring in the United States cost about US$13,000 per SD of a composite of learning outcomes (Bhatt *et al.* 2024; Burneo *et al.* 2026). For a gain of 0.2 SD, the ceilings are respectively about US$31, US$170 and US$2,600 per student for the three countries.

But these figures are for illustration only. In an actual comparison, costs should be put on the anchor's price base and short-run effects should not be annualized, but matching price bases does not make numeracy, mathematics and composite outcomes interchangeable. Price bases still matter. The Nigerian program looks cheap at market exchange rates, but in 2022 purchasing power parity (PPP) dollars Burneo *et al.* (2026) put it at about US$1,530 per SD of learning, against about US$1,024 for Mindspark on a composite of mathematics and Hindi. Actual choices also depend on local feasibility, affordability, uncertainty and distributional safeguards.

Two rules keep such comparisons consistent. First, count differences in resource use between the alternatives, including resources already owned. Identical use cancels out, but owning a resource does not make its use free. Added device wear, scarce computer time, training, supervision and students' (and caregivers') time are costs even without a new invoice. Second, avoid counting the same foregone learning twice. Instruction displaced by a program is already reflected in the effect, but other opportunity costs should still be recorded. Costs should also be labeled by perspective, to the school system or to society, which also counts students' time and donated resources (Institute of Education Sciences 2020). Both $\Delta C$ and $\lambda$ must share one perspective. Procurement decisions should include recurrent costs, such as licenses, device

replacement, connectivity and continued training (Yarrow *et al.* 2026). They should also plan for the end of donated or subsidized access at scale, which shifts costs to public budgets or households. Table 1, Panel C lists the items to cost under two scenarios, with devices and connectivity already in place or newly supplied.

We offer some further reflections. First, more equal access, positive gains for poorer students and a narrower achievement gap are distinct outcomes, and a program can achieve one without the others. On a common outcome scale, a gap narrows only if the population-average gain is larger for disadvantaged students. That average depends on coverage, take-up and the net learning effect among users (and any effects on non-users), including household responses, so positive effects among connected participants cannot establish it. Access comes first. The entry cost in Figure 1b bites hardest where connectivity is lowest, and only 40% of primary schools worldwide are connected to the internet (UNESCO 2023a). Shared school devices lower the entry cost per student. Whether they reach students without devices or connectivity at home should be measured, not assumed.

Second, regarding AI use, PISA results show that advantaged students use AI for schoolwork more and more often learn to assess its output, whereas in the United States heavy reliance on chatbots for schoolwork is more common in lower-income households. Frequency of use and reliance are different margins, and neither gradient shows a difference in learning.

Third, regarding learning returns, for student-facing tools, the few within-trial interactions in Section 4 favor students who start ahead. A public program can still narrow gaps if households with more support at home reduce their own tutoring in response, as in the Chinese tutoring trial (Behrman *et al.* 2024), but whether AI provision triggers similar responses is unknown.

Finally, language is a further margin. None of the trials we include tested a tutor in a low-resource language, and randomized evidence on language learning itself is thin. Nigeria's English gain and an imprecise Italian homework trial (0.25 SD [−0.24, 0.74]; Vanzo *et al.* 2025) are the only such generative AI outcomes in our set, and meta-analyses of chatbots for language learning draw largely on studies from before large language models (Lyu *et al.* 2025). Tokenization can cost up to 15 times more in some languages than in others, and models perform far worse on many tasks in African languages (Petrov *et al.* 2023; Ojo *et al.* 2025). These are reasons to test tutors in local curricula and languages, not evidence that every learner gets a worse tutor.

Yet, several questions from previous discussions of private tutoring remain open. Does free AI help weaken parents' pressure to improve public schools? Does AI raise or lower public capacity? Does AI-produced work erode the signal in grades? On the third, in a 2023–24 survey of more than 95,000 students at 20 US public research universities, 37% used generative AI at least monthly, and an indirect estimate suggests that about 9% had used it to cheat (Chirikov *et al.* 2026). And how much of the surplus do providers keep through prices, defaults, data and lock-in? Provider involvement also matters for the evidence itself. Evaluations run or co-authored by providers need disclosure and independent replication, not automatic dismissal.

**6. Conditional decisions for governments**

Public policies seem to have moved ahead of research evidence and can generally be sorted into three groups (Figure 3; Supplementary Table S4). Some restrict or scale back AI use, as New York City, China and Korea have done (China Daily 2025; Rest of World 2025; K-12 Dive 2026). Others wait, or guide or regulate it. England and the United States have issued guidance (Department for Education 2025; US Department of Education 2025), and the European Union's AI Act treats admission, the evaluation of learning and proctoring as high-risk uses (European

Union 2024, 2026). Still others encourage it, through AI curricula, tools for students and partnerships with providers (Associated Press 2025; Estonian Ministry of Education and Research 2025; Gulf News 2025; OpenAI 2025; Ministry of Education and Training of Viet Nam 2026). These measures are not exclusive. China restricts independent use by young students while Beijing requires AI education, and New York City runs vetted pilots in high schools during its pause.

A reversal is not proof of failure. Korea reclassified its AI textbooks after political opposition and low adoption (Rest of World 2025). The episode shows the fiscal and political risk of scaling before evaluating, not that the textbooks failed to teach. Teaching about AI and teaching with an AI tutor also need different evidence. The first is a curriculum choice, and Fesler *et al.* (2026) found no high-quality causal study of AI-literacy instruction. The second is an instructional intervention whose effects on learning can be tested.

Table 2 therefore organizes conditional decisions by use case, school level and delivery capacity rather than by country. For each use, it states when a bounded pilot is reasonable, what evidence would justify expansion, and what findings should prompt redesign or withdrawal. We propose five principles that run across these decisions.

**Expand on evidence of learning or of savings.** Expand when credible evidence, relative to feasible alternatives, shows better learning at a full cost that passes the break-even test, or rules out an educationally important loss while delivering substantial savings. Expansion also requires safeguards and benefits for disadvantaged learners; costs alone do not justify it. A cheaper substitute can be worthwhile without outperforming a strong alternative. The acceptable loss must be specified in advance, however, because failing to reject a difference is not evidence that learning is preserved. For teacher-facing tools, a lighter workload can be a benefit in itself, even when the time saved is not all spent on teaching.

**Price the AI layer against its increment.** When AI is added to an existing platform, the relevant gain is the increment over the same platform, schedule and staffing, not the effect of the full package. Governments can ask providers for evidence on that increment and its full cost, or fund trials that estimate it (Section 5; Box 2).

**Treat imprecise results as information.** A precise finding that rules out an important loss differs from an imprecise estimate and from a meaningful harm. An imprecise pilot should lead to further evaluation, not to automatic expansion or withdrawal. Redesign is warranted when a program fails the break-even or equity conditions, for example when it raises assisted practice but neither improves unassisted learning nor saves resources. Withdrawal is warranted when unassisted learning falls by more than the prespecified acceptable loss, with adequate precision, or when safeguards fail.

**Lower entry costs for the unconnected.** School provision through shared devices and low-bandwidth or offline tools could reverse the equity pattern of household-financed tutoring, provided that the program builds skills for the students it reaches.

**Set conditions on assessment, procurement and measurement.** Assessment reform, such as invigilated or oral components and disclosure of AI use, can be combined with narrowly targeted restrictions. Procurement can require independent assessment of learning, child-appropriate safeguards, data minimization, accessibility, teacher oversight, a complaints channel and exit arrangements. Of 164 education technology products endorsed by 49 governments during the Covid-19 pandemic, 89% risked or infringed children's rights, for example by tracking children across the internet or sharing their data with advertising companies (Human Rights Watch 2022). National assessments and surveys can also measure AI use and opportunities to learn about AI, because provider data cannot show who learns.

Box 2 proposes a comparative trial to guide public investment. It randomizes a generative tutor and human facilitation in a 2×2 design on a common non-generative platform, so it can separate what AI adds, what facilitation adds and whether the two are complements. It also measures delayed unassisted learning, household responses and full incremental costs, so that its results can be set against the break-even rule in Section 5. The box ends with short- and long-term research priorities.

## 7. Conclusion

AI has made answers cheap, but lasting learning still requires productive effort, support and time. The evidence reviewed here supports cautious, evaluated use. Randomized trials show that both supported programs and less-restricted access can improve learning in some settings, while other uses can impair it. The policy question is whether a specified program builds lasting skills, or preserves learning at lower full cost, relative to a feasible alternative, and who benefits. In the few student-facing trials that report subgroup results, students who start ahead fared better. Public provision can widen access without narrowing learning gaps, because who benefits depends on how students use the tools, what the tools replace and how households respond.

The comparative trial in Box 2 is designed to inform all three questions of Section 1; access and assessment arrangements need evaluations of their own.

Two decades ago, private tutoring spread as a supplement that families bought, and better-off households bought the most. Private use of AI is spreading in the same way. Publicly provided AI tutoring differs in one respect that matters for policy: governments can design, procure and evaluate it before scaling it up. Whether AI narrows educational inequality will depend in part on whether they do.

## Table 1. Cost evidence for AI tutoring programs and benchmarks: reported costs, break-even ceilings and items to cost

**Panel A. Reported costs**

| No. | Program (comparator) | Reported cost and type | Not included | Effect (SD); duration |
|---|---|---|---|---|
| 1 | Ghana: WhatsApp mathematics tutor (Rori), grades 3–8 (regular instruction) | About US$5 per student; marginal cost as reported; currency year not stated | Phones or devices, mobile data, study-hall supervision, electricity | 0.36 (no CI); 8 months |
| 2 | Nigeria: after-school AI program (Copilot), senior secondary (no program) | US$48 per student (US$9 marginal); program cost from budget data | Electricity and school infrastructure (excluded by the source) | English 0.24 [0.10, 0.37]; composite 0.31; 6 weeks |
| 3 | Sierra Leone: Gemini in mathematics lessons, grades 7–8 (usual lessons) | Not reported | Shared devices, connectivity, electricity, accounts, implementation support | 0.26 [0.03, 0.49]; 8 weeks |
| 4 | United States: Khanmigo in remedial blocks, grades 6–8 (usual remediation) | About US$15 per student a year; license only | Devices and staff time (used in both arms); the comparison software is not netted out | 0.06 a year [−0.01, 0.13]; 2 years |
| 5 | India: Mindspark adaptive software, after school (no access, by lottery); benchmark | US$307 per student over 4.5 months (2022 PPP dollars), computed in a synthesis | Not itemized in the synthesis | Mathematics 0.37 [0.24, 0.50]; Hindi 0.23; 4.5 months |
| 6 | Botswana: phone calls plus text messages (no program); benchmark | US$19 per child (US$5 for text messages only), as reported | Not itemized | 0.12 [0.03, 0.21]; 8 weeks |
| 7 | United States: Saga technology-supported human tutoring (usual instruction); benchmark | US$2,488 per student over 9 months (2022 PPP dollars), computed in a synthesis | Not itemized in the synthesis | About 0.19 (implied by the synthesis); 9 months |

**Panel B. Illustrative break-even ceilings, $\Delta C \leq \lambda \Delta g$ (US$ per student)**

| Additional learning, $\Delta g$ (SD) | Phone tutoring, Botswana ($\lambda \approx$ US$157 per SD, numeracy) | Adaptive software, India ($\lambda \approx$ US$830 per SD, mathematics) | Tutoring with technology, United States ($\lambda \approx$ US$13,000 per SD, composite) |
|---|---|---|---|
| 0.10 | 16 | 83 | 1,300 |
| 0.20 | 31 | 170 | 2,600 |
| 0.30 | 47 | 250 | 3,900 |

**Panel C. Items to cost, by scenario (template)**

| Item | Scenario 1: devices and connectivity already in schools | Scenario 2: devices and connectivity newly supplied |
|---|---|---|
| Devices and their replacement | Added use (wear, computer time), even if owned, in $\Delta C$ | Annualized purchase and replacement, in $\Delta C$ |
| Connectivity, data and electricity | Additional use only, in $\Delta C$ | Full cost, including backup power, in $\Delta C$ |
| Licenses and content | Recurrent fees, in $\Delta C$ | Recurrent fees, in $\Delta C$ |
| Facilitators' and teachers' time | Additional time, in $\Delta C$ | Additional time, in $\Delta C$ |
| Training and implementation support | Initial and recurrent, in $\Delta C$ | Initial and recurrent, in $\Delta C$ |
| Students' and caregivers' time and own devices | Societal perspective only | Societal perspective only |
| Instruction displaced by the program | Not charged again; the loss is in the effect | Not charged again; the loss is in the effect |

*Source:* Panel A, by row: 1, Henkel *et al.* (2024) and World Bank (2026); 2, De Simone *et al.* (2025); 3, LearnLM Team and Fab AI (2026); 4, Oreopoulos and Low (2026); 5, Muralidharan *et al.* (2019) and Burneo *et al.* (2026, Table A16); 6, Angrist *et al.* (2022); 7, Bhatt *et al.* (2024) and Burneo *et al.* (2026, Table A16). Panel C, Institute of Education Sciences (2020) and Yarrow *et al.* (2026).

*Note:* Costs are as reported, on different price bases, so Panel A does not rank programs. Panel B is an illustration, not a ranking or a procurement threshold: it treats each anchor's average cost per SD as $\lambda$, on the anchor's price base (Supplementary Note 2), and rounds ceilings to two significant figures.

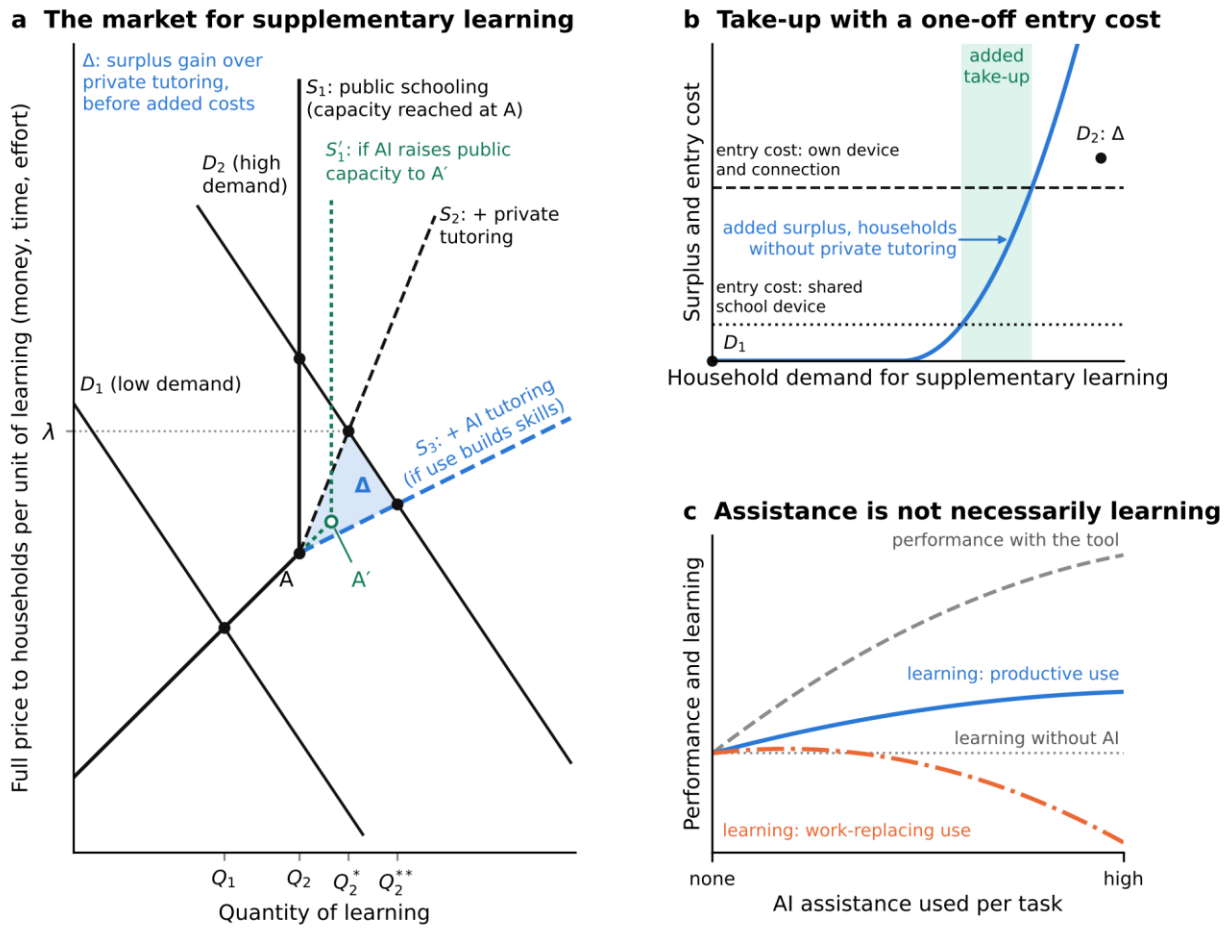
## Table 2. Conditional decisions for AI in education, by use case, school level and delivery capacity

| No. | Use case | Allow a bounded pilot when | Expand when evidence shows | Redesign or withdraw if | Illustrative cases |
|---|---|---|---|---|---|
| 1 | Teaching about AI (AI literacy), all levels | Curriculum time and trained teachers are available without cutting foundational literacy and numeracy | Students can judge and use AI output on assessed tasks | Core-subject results fall by more than the prespecified margin | United Arab Emirates; Beijing; Viet Nam |
| 2 | AI tutor, primary school | Use is teacher-led, content is curated, open-ended chat is off and safeguards fit young children | Unassisted foundational skills improve, or an educationally important loss is ruled out at substantially lower cost, by the end of the year | Unassisted skills fall by more than the prespecified margin, or safeguards for young children fail | China; New York City |
| 3 | AI tutor, secondary school | The design keeps productive work with the student, and a timetable, a facilitator and devices are in place | Delayed unassisted gains, or no educationally important loss at substantial savings, under routine public delivery | Assisted practice rises without unassisted gains or savings (redesign); unassisted scores fall by more than the prespecified margin (withdraw) | Sierra Leone; Nigeria; Estonia |
| 4 | AI tools, tertiary education | Disclosure rules and guidance on learning-oriented use exist | Guided use improves outcomes measured under secure assessment, or preserves them at substantially lower cost | Graded work stops measuring learning | Chile; US public universities |
| 5 | Teacher-facing tools | Teachers keep oversight and data use is minimized | Workload falls with no educationally important loss in student outcomes, or better materials raise them | Motivation or results fall by more than the prespecified margin (withdraw); neither workload nor outcomes improve (redesign) | England; Türkiye; Peru |
| 6 | AI in assessment and admission | Human review, appeal and validation by group are in place | Validity and fairness hold across groups | Errors or bias by group appear | European Union |
| 7 | Low delivery capacity (unreliable devices, power or connectivity) | Shared devices in school, low-bandwidth or offline tools, teacher mediation | Learning improves or is preserved, and full costs, including devices and power, stay below the break-even | Costs exceed the break-even, or use concentrates among already connected students | Sierra Leone |
| **All uses: conditions for expansion.** Better learning at a full cost below the break-even, or an educationally important loss ruled out at substantially lower full cost, relative to a feasible alternative; benefits for disadvantaged learners; safeguards in place. Imprecise results call for further evaluation, not for expansion or withdrawal. | | | | | |
| **All uses: procurement conditions.** Independent assessment of learning; recurrent and replacement costs in the budget; child-appropriate safeguards; data minimization; accessibility (low bandwidth, local languages, disability); teacher oversight; a complaints channel; exit arrangements (data portability, no lock-in). | | | | | |

*Source:* Cases by row: 1, Gulf News (2025), Australian Government Department of Education (2025) and Ministry of Education and Training of Viet Nam (2026); 2, China Daily (2025) and K-12 Dive (2026); 3, De Simone *et al.* (2025), Estonian Ministry of Education and Research (2025) and LearnLM Team and Fab AI (2026); 4, Gallegos (2026) and Chirikov *et al.* (2026); 5, Roy *et al.* (2024, 2026), Sungu *et al.* (2026) and Glave *et al.* (2026); 6, European Union (2024, 2026); 7, LearnLM Team and Fab AI (2026).

*Note:* Proposals for discussion, not findings. Acceptable losses are set before the evaluation; "preserved" and "no educationally important loss" mean that such a loss is ruled out, not merely undetected.

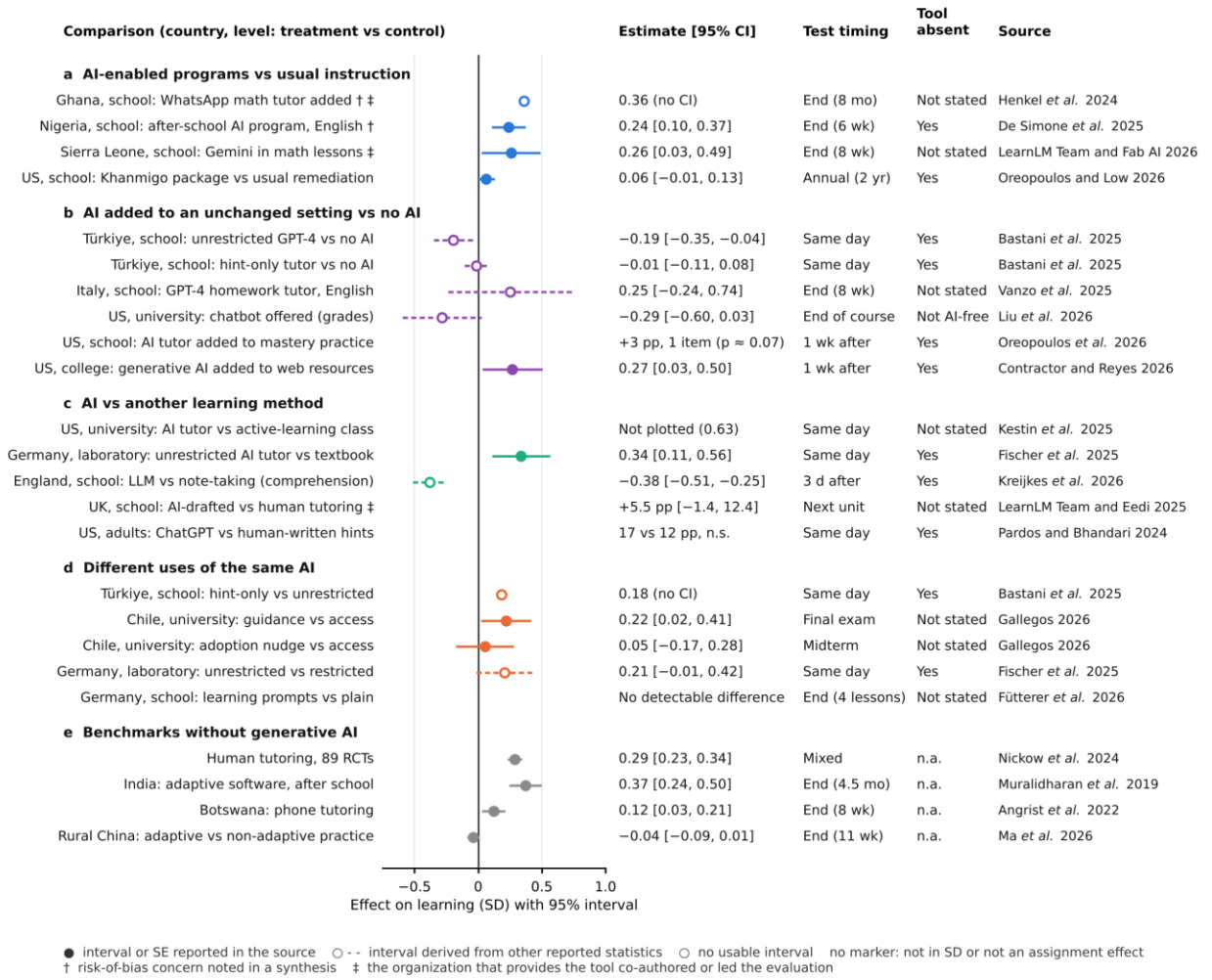
**Figure 1. AI tutoring in the market for supplementary learning**




*Source:* Authors' illustration, adapted and extended from Dang and Rogers (2008).

*Note:* **a**, Public schooling ($S_1$) is capped at A. Private tutoring ($S_2$) and AI tutoring ($S_3$) supply learning beyond A, $S_3$ only if use turns assistance into learning. A high-demand household ($D_2$) consumes $Q_2$ with public schooling only, $Q_2^*$ with private tutoring and $Q_2^{**}$ with AI tutoring; a low-demand household ($D_1$) buys neither. Δ is a conditional surplus gain over private tutoring, before entry and public-provision costs. $\lambda$, the height of $S_2$ at $Q_2^*$, is the cost of further learning through the best alternative (Section 5). $S_1'$ and A′ show public supply and capacity if AI raised capacity (a hypothesis); A′ is not where $S_3$ meets $D_2$. **b**, A household adopts AI tutoring if its added surplus over its best existing option covers the entry cost. The blue curve is for households that private tutoring does not reach; a household on $D_2$ already buys tutoring, so it gains at most Δ if tutors earn no producer surplus. Shared school devices lower the entry cost, bringing in households in the shaded band. **c**, Hypothesized paths, not estimates: learning without the tool rises when use keeps productive work with students and can fall below its level without AI when use replaces it.

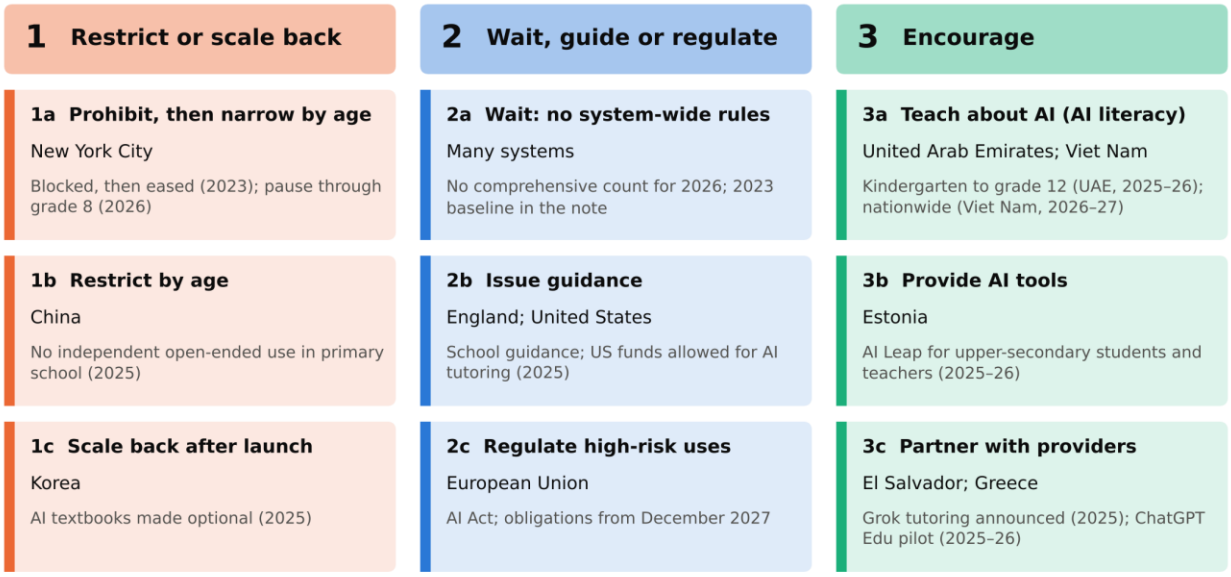
**Figure 2. Effects of AI on learning, organized by what was randomized**




*Source:* Published estimates; Supplementary Table S2 gives each source, the table or page used and its version.

*Note:* Effects in standard deviations (SD) of the outcome, mostly test scores, with 95% intervals as published or derived from published statistics. **a**, AI-enabled programs versus usual instruction; **b**, AI added to an otherwise unchanged setting versus no AI; **c**, AI versus another learning method; **d**, different uses of the same AI; **e**, benchmarks without generative AI. Pooled estimates of AI effects are not plotted (Supplementary Note 2). Test timing: "End" is the end of the program (duration in parentheses); later tests are dated from the end of support; "Next unit" is the first question on a later topic, answered the same day. Tool absent: whether students could not use the tool during the test ("Not stated": the source does not say); "Not AI-free": course grades that may include AI-assisted work. The US physics estimate is adjusted for time on task, which the tutor changed, so it is not plotted; the UK interval is a 95% credible interval. Risk-of-bias concerns (†): Ghana, 11 schools with student-level inference; Nigeria, 43% did not take the endline. SE, standard error; RCTs, randomized controlled trials; pp, percentage points; n.s., not significant; n.a., not applicable; d, days; wk, weeks; mo, months; yr, years.

**Figure 3. Measures governments have taken on generative AI in schools: three groups of three variants**




*Source:* Supplementary Table S4, which gives each measure with its date, its status in October 2026, whether it concerns teaching about AI or with AI, and its sources.

*Note:* Each variant except 2a is illustrated by one or two systems, and none is a census. For variant 2a, fewer than 10% of more than 450 schools and universities polled by UNESCO in 2023 had formal guidance on generative AI, a historical baseline only (UNESCO 2023b).

**Box 1. Four outcomes that are easy to confuse**

Evaluations of AI in education report four different outcomes. Mixing them is a common source of contradictory claims.

***Assisted performance*** is how well students do with the tool available. It rises in most trials, including the Turkish one, where practice grades rose with both unrestricted and hint-only access (Bastani *et al.* 2025; Fesler *et al.* 2026). It shows that the tool helps with the task, not that students learned (Soderstrom and Bjork 2015).

***Unassisted learning*** is what students can do without the tool soon after using it. In the same Turkish trial, unrestricted access lowered closed-book exam scores by about a fifth of an SD, and the hint-only tutor showed no detectable difference from practice without AI.

***Retention and transfer*** need separate tests. Retention is what remains days or months after support ends; transfer is the ability to apply what was learned to new problems. Evidence on both is thin. A gain appeared on a test a week later at a US college (Contractor and Reyes 2026), note-taking beat LLM use three days later in England (Kreijkes *et al.* 2026), and in Tennessee an AI tutor showed a suggestive gain on one practiced item a week later ($p \approx 0.07$), with no detectable gain on unpracticed items (Oreopoulos *et al.* 2026). Transfer requires applying learning to new tasks and can be assessed at endline or after a delay.

***Competent AI-assisted work*** is the skill of using the tool well, for example by judging and correcting its output. It is valuable in itself: Peruvian teachers trained to use Copilot wrote better prompts and spotted weak output (Glave *et al.* 2026). But it is not evidence of unaided mastery.

Four questions classify any reported gain. Was the tool absent at the test? How long after support ended was the test taken? Was the test external to the program? And did it repeat practiced material or require applying learning to new tasks?

## Box 2. A comparative trial to guide public investment

***Design.*** Randomize schools or classrooms to four arms on a common non-generative adaptive platform, in a 2×2 design: the platform alone; the platform with a generative tutor; the platform with human facilitation; and the platform with both. A fifth arm with usual instruction is optional when the question is whether to upgrade an existing platform, but needed to value the whole package where that platform is not yet in use.

***What it identifies.*** Comparisons with and without the tutor identify the AI increment, with and without facilitation. Comparisons with and without facilitation identify its value, with and without AI. Whether the AI increment is larger with facilitation than without it shows whether the two are complements; a design that adds facilitation only to the AI arm cannot tell. Detecting that interaction needs a larger sample than either main effect, so the trial should state whether complementarity is a primary question.

***Conditions.*** Holding devices and instructional time equal across arms estimates the value of components under common conditions. Evaluating the package as governments would routinely deliver it answers a separate policy question, and the trial should state which question it answers.

***Outcomes to prespecify.*** Unassisted learning at the end of the school year and a year later; transfer to new problems; assisted competence; displacement of other learning time; household responses, such as changes in home tutoring, spending and caregivers' time (Behrman *et al.* 2024); actual take-up and dosage; and full incremental costs. A small set of equity interactions, by baseline skill, socioeconomic status, home connectivity and gender, should be prespecified with their uncertainty.

***Inference.*** Name the primary comparisons and power the study for them. Prespecify clustering at the school or classroom level and the handling of attrition. Record students' use of AI outside the program and the versions of the models used. Fix the date on which program support ends and record continued access to AI, so that the one-year test can be read in light of continued use. A randomized withdrawal of access, for example in a subsample, would separate persistence from the effect of continued use. Household responses help explain effects but are not identified mediators simply because AI was randomized.

***Short-term priorities (next two years).*** For every trial, report learning without the tool, its timing, whether the tool was absent, whether the test was external and used new tasks, and full ingredient costs. Keep an auditable, living map of the evidence. Add items on AI use and AI literacy to national assessments and household surveys.

***Long-term priorities (next five years).*** Run government-led trials at scale in low-income countries and in low-resource languages. Study effects on public-school capacity, on teachers and on parents' demands on schools. Track long-run outcomes: attainment, earnings, the signal carried by credentials, and inequality.

**References**


Angrist, N., Bergman, P., and Matsheng, M. (2022). Experimental evidence on learning using low-tech when school is out. *Nature Human Behaviour*, 6, 941–950.

Associated Press. (2025). El Salvador teams up with Elon Musk's xAI to bring AI to 5,000 public schools. December 11. baynews9.com

Australian Government Department of Education. (2025). *China's Approach to AI Education in Schools (Year 1–12)*. Canberra, information as at September 2025. education.gov.au

Banks, D. C. (2023). ChatGPT caught NYC schools off guard. Now, we're determined to embrace its potential. Chalkbeat, May 18. chalkbeat.org/newyork/2023/5/18/23727942/chatgpt-nyc-schools-david-banks

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122.

Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., and Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review*, 37, 45.

Behrman, J. R., Parker, S. W., Todd, P. E., and Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools. *Journal of Political Economy*, 123(2), 325–364.

Behrman, J. R., Fan, C. S., Guo, N., Wei, X., Zhang, H., and Zhang, J. (2024). Tutoring efficacy, household substitution, and student achievement: Experimental evidence from an after-school tutoring program in rural China. *International Economic Review*, 65(1), 149–189.

Bhatt, M. P., Guryan, J., Khan, S. A., LaForest-Tucker, M., and Mishra, B. (2024). Can technology facilitate scale? Evidence from a randomized evaluation of high dosage tutoring. NBER Working Paper 32510.

Bray, M. (2006). Private supplementary tutoring: Comparative perspectives on patterns and implications. *Compare: A Journal of Comparative and International Education*, 36(4), 515–530.

Bray, M. (2025). Taking private tutoring out of the shadows. NORRAG blog, October 30. norrageducation.org/taking-private-tutoring-out-of-the-shadows

Burneo, A., Dinarte-Diaz, L., Lopez, C., and Molina, E. (2026). Can EdTech close learning gaps? Global evidence from digital interventions. Background paper for the *World Development Report 2026*. World Bank. thedocs.worldbank.org

Byun, S.-Y., Chung, H. J., and Baker, D. P. (2018). Global patterns of the use of shadow education: Student, family, and national influences. *Research in the Sociology of Education*, 20, 71–105.

Caixin Global. (2026). From cram schools to smart tablets: AI fuels China's new edtech boom. June 11. caixinglobal.com

Chi, M. T. H., and Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243.

China Daily. (2025). Guideline to regulate use of artificial intelligence in schools. May 16. global.chinadaily.com.cn/a/202505/16/WS6826951fa310a04af22bfc2e.html

Chirikov, I., Smirnov, I., and Kizilcec, R. F. (2026). Generative AI use and misuse call for assessment reform in higher education. *Science*, 392, 818–820.

Contractor, Z., and Reyes, G. (2026). Experimental evidence on the learning impact of generative AI. IZA Discussion Paper 18792.

Dang, H.-A. H., and Rogers, F. H. (2008). The growing phenomenon of private tutoring: Does it deepen human capital, widen inequalities, or waste resources? *World Bank Research Observer*, 23(2), 161–200.

Dang, H.-A. H., and Glewwe, P. W. (2018). Well begun, but aiming higher: A review of Vietnam's education trends in the past 20 years and emerging challenges. *Journal of Development Studies*, 54(7), 1171–1195.

Dang, H.-A. H., and Do, M. N. N. (2026). Can AI improve household welfare in poorer countries? Evidence and measurement challenges. IZA Discussion Paper 18984.

De Simone, M., Tiberti, F., Barron Rodriguez, M., Manolio, F., Mosuro, W., and Dikoru, E. J. (2025). From chalkboards to chatbots: Evaluating the impact of generative AI on learning outcomes in Nigeria. World Bank Policy Research Working Paper 11125, revised December 2025.

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

Department for Education. (2025). *Generative Artificial Intelligence (AI) in Education*. Guidance, updated August 12. London. gov.uk

Elsen-Rooney, M. (2023). NYC education department blocks ChatGPT on school devices, networks. Chalkbeat, January 3. chalkbeat.org/newyork/2023/1/3/23537987/nyc-schools-ban-chatgpt-writing-artificial-intelligence

Escueta, M., Nickow, A. J., Oreopoulos, P., and Quan, V. (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature*, 58(4), 897–996.

Estonian Ministry of Education and Research. (2025). Estonia announces a groundbreaking national initiative: AI Leap programme to bring AI tools to all schools. February 25. hm.ee

European Union. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). eur-lex.europa.eu/eli/reg/2024/1689/oj

European Union. (2026). Regulation (EU) 2026/1744 of 8 July 2026 (Digital Omnibus on AI). eur-lex.europa.eu/eli/reg/2026/1744/oj

Executive Office of the President. (2025). Executive Order 14277: Advancing artificial intelligence education for American youth. April 23. presidency.ucsb.edu

Fesler, L., Martinez, J., Agnew, C., and Loeb, S. (2026). *The Evidence Base on AI in K-12: A 2026 Review*. AI Hub for Education, SCALE Initiative, Stanford University. scale.stanford.edu/research-in-action/understanding-evidence-base-ai-k12-education

Fischer, M., Rau, H. A., and Rilke, R. M. (2025). AI tutoring enhances student learning without crowding out reading effort. IZA Discussion Paper 18338.

Fütterer, T., Bardach, L., Kuhn, J., Keller, S. D., and Gerjets, P. (2026). Enhancing school students' self-regulated learning through generative AI support: A randomized controlled trial. *Educational Psychology Review*, 38, 42.

Gallegos, S. (2026). Guidance over adoption: Experimental evidence on AI-assisted learning. IZA Discussion Paper 18513.

Glave, C. Z., Lopez, C., Molina, E., and Zárate, R. A. (2026). Not faster, but better: Generative AI and teacher skill in Peru. World Bank Policy Research Working Paper 11476. documents.worldbank.org/en/publication/documents-reports/documentdetail/099309210012625906

Gulf News. (2025). New school year begins with focus on student readiness and new AI curriculum. August 25. gulfnews.com

Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., and Lee, A. (2024). Effective and scalable math support: Experimental evidence on the impact of an AI math tutor in Ghana. In *Artificial Intelligence in Education: Posters and Late Breaking Results* (Communications in Computer and Information Science 2150, pp. 373–381). Springer.

Human Rights Watch. (2022). *"How Dare They Peep into My Private Life?" Children's Rights Violations by Governments that Endorsed Online Learning during the Covid-19 Pandemic*. May 25. hrw.org/news/2022/05/25/governments-harm-childrens-rights-online-learning

Institute of Education Sciences. (2020). *Cost Analysis: A Starter Kit* (IES 2020-001). US Department of Education, Washington, DC. ies.ed.gov/ies/2025/01/cost-analysis-starter-kit

ITU. (2025). *Facts and Figures 2025*. International Telecommunication Union, Geneva. itu.int/itu-d/reports/statistics/2025/10/15/ff25-internet-use

Jayachandran, S. (2014). Incentives to teach badly: After-school tutoring in developing countries. *Journal of Development Economics*, 108, 190–205.

Jurenka, I., *et al.* (2025). Towards responsible development of generative AI for education: An evaluation-driven approach. arXiv:2407.12687 (version 4).

K-12 Dive. (2026). New York City pauses AI use for elementary, middle school students. September 2. k12dive.com/news/new-york-city-pauses-ai-use-for-elementary-middle-school-students/829496

Kestin, G., Miller, K., Klales, A., Milbourne, T., and Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15, 17458.

Korea Herald. (2025). Private education spending hits record despite fall in student numbers. March 13. koreaherald.com/article/10440947

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253.

Kreijkes, P., Kewenig, V., Kuvalja, M., Lee, M., Hofman, J. M., Vitello, S., *et al.* (2026). Effects of LLM use and note-taking on reading comprehension and memory: A randomised experiment in secondary schools. *Computers & Education*, 243, 105514.

Lawson, A. P., Martella, A. M., Weidlich, J., Mulders, M., and Buchner, J. (2026). Color me confounded: A critical analysis of media comparisons on ChatGPT in education. *Computers & Education*, 254, 105701.

LearnLM Team and Eedi. (2025). *AI Tutoring Can Safely and Effectively Support Students: An Exploratory RCT in UK Classrooms*. Technical report, November 11. Google DeepMind. storage.googleapis.com/deepmind-media/LearnLM/learnLM_nov25.pdf

LearnLM Team and Fab AI. (2026). *Teaching with Gemini: Measuring the Impact of Guided Learning on Student Mathematics Progress in Sierra Leone*. Technical report, May 15. Google DeepMind. storage.googleapis.com/deepmind-media/LearnLM/learnLM_sierraleone_may26.pdf

Lehmann, M., Cornelius, P. B., and Sting, F. J. (2025). AI meets the classroom: When do large language models harm learning? arXiv:2409.09047 (version 2).

Liu, J., Sweet, T., Chen, M. H., Engelberg, J., Masters, M., Clark, M., Persaud, A., Lancaster, A., Hollingsworth, J., and King Rice, J. (2026). The effects of course-integrated AI tutoring on student performance and engagement: A randomized university trial. EdWorkingPaper 26-1598, Annenberg Institute, Brown University.

Lyu, B., Lai, C., and Guo, J. (2025). Effectiveness of chatbots in improving language learning: A meta-analysis of comparative studies. *International Journal of Applied Linguistics*, 35(2), 834–851.

Ma, Y., Feng, T., Fairlie, R. W., Liu, C., Loyalka, P., Rozelle, S., and Zhang, X. (2026). Putting AI to the test: Evidence from a large-scale RCT in rural China. CESifo Working Paper 12837.

Ministry of Education and Training of Viet Nam. (2024). Circular 29/2024/TT-BGDĐT of December 30, 2024 on extra teaching and learning (effective February 14, 2025). luatvietnam.vn

Ministry of Education and Training of Viet Nam. (2026). Decision 2422/QĐ-BGDĐT of August 18, 2026 on the content framework for AI education in general schools. baovanhoa.vn/doi-song/giao-duc-ai-dai-tra-tu-nam-hoc-20262027-257693.html

Muralidharan, K., Singh, A., and Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review*, 109(4), 1426–1460.

Nickow, A., Oreopoulos, P., and Quan, V. (2024). The promise of tutoring for preK–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational Research Journal*, 61(1), 74–107.

OECD. (2026). *PISA 2025 Results (Volume I): Future-Ready Students*. OECD Publishing, Paris. oecd.org/en/publications/2026/09/pisa-2025-results-volume-i_5265bfb1.html

Ojo, J., *et al.* (2025). AfroBench: How good are large language models on African languages? arXiv:2311.07978.

OpenAI. (2025). OpenAI for Greece. September 5. openai.com/global-affairs/openai-for-greece

Oreopoulos, P., and Low, N. (2026). One click away: AI tutoring with Khanmigo in a two-year school experiment. NBER Working Paper 35620.

Oreopoulos, P., Liut, M., Sungu, A., and Low, N. (2026). Making AI tutoring productive: Evidence from a mastery-based math practice experiment. NBER Working Paper 35621.

Pardos, Z. A., and Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLOS ONE*, 19(5), e0304013.

Petrov, A., La Malfa, E., Torr, P. H. S., and Bibi, A. (2023). Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36.

Pew Research Center. (2026). How teens use and view AI. February 24. pewresearch.org/internet/2026/02/24/how-teens-use-and-view-ai

Rest of World. (2025). AI-powered textbooks fail to make the grade in South Korea. October 15. restofworld.org/2025/south-korea-ai-textbook

Roy, P., Poet, H., Staunton, R., Aston, K., and Thomas, D. (2024). *ChatGPT in Lesson Preparation: A Teacher Choices Trial*. Evaluation report. Education Endowment Foundation. nfer.ac.uk/publications/chatgpt-in-lesson-preparation-a-teacher-choices-trial

Roy, P., Poet, H., Staunton, R., Aston, K., Thomas, D., and Styles, B. (2026). *Lesson Planning Using AI Lesson Assistant, Aila: A Teacher Choices Trial*. Evaluation report. Education Endowment Foundation. educationendowmentfoundation.org.uk/projects-and-evaluation/projects/aila-teacher-choices-trial

Schleicher, A. (2026). Generative AI and student learning: What the PISA data reveal. OECD Education and Skills Today, October 13. oecdedutoday.com/generative-ai-and-student-learning-what-the-pisa-data-reveal

Soderstrom, N. C., and Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2), 176–199.

Sungu, A., Lira, B., and Duckworth, A. L. (2026). Generative AI can harm teaching. SSRN Working Paper 7007339.

TI-Hüpe. (2026). AI Leap to expand to vocational education and lower secondary school teachers this fall. September 1. tihupe.ee

Tracking SDG7. (2026). *The Energy Progress Report 2026*. un.org

UNESCO. (2023a). *Global Education Monitoring Report 2023. Technology in Education: A Tool on Whose Terms?* UNESCO, Paris. unesco.org/gem-report/en/technology

UNESCO. (2023b). UNESCO survey: Less than 10% of schools and universities have formal guidance on AI. Press release, June 1. gcedclearinghouse.org/en/node/130058

UNESCO. (2024). *Global Report on Teachers: Addressing Teacher Shortages and Transforming the Profession*. UNESCO, Paris. teachertaskforce.org/node/2256

US Department of Education. (2025). Guidance on artificial intelligence use in schools. Press release, July 22. ed.gov

Vanzo, A., Pal Chowdhury, S., and Sachan, M. (2025). GPT-4 as a homework tutor can improve student engagement and learning outcomes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (pp. 31119–31136).

Wang, J., and Fan, W. (2025). 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, 621. Retracted April 22, 2026.

Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., and Demszky, D. (2026). Tutor CoPilot: A human-AI approach for scaling real-time expertise. arXiv:2410.03017 (version 3).

Weidlich, J., Gašević, D., Drachsler, H., and Kirschner, P. A. (2025). ChatGPT in education: An effect in search of a cause. *Journal of Computer Assisted Learning*, 41(5), e70105.

World Bank. (2022). *The State of Global Learning Poverty: 2022 Update*. World Bank, Washington, DC. worldbank.org/en/topic/education/publication/state-of-global-learning-poverty

World Bank. (2026). *World Development Report 2026: The Promise of Artificial Intelligence*. World Bank, Washington, DC. worldbank.org/en/publication/wdr2026

Yan, L., Greiff, S., Teuber, Z., and Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*, 8, 1839–1850.

Yan, L., Greiff, S., Lodge, J. M., and Gašević, D. (2025). Distinguishing performance gains from learning when using generative AI. *Nature Reviews Psychology*, 4(7), 435–436.

Yarrow, N., Trucano, M., Rodon, G., De Simone, M., and Bentil, E. (2026). Costing AI use in education in low- and middle-income countries. Education & Skills Working Paper 27. World Bank, Washington, DC. documents1.worldbank.org/curated/en/099317007132640857/pdf/IDU-1709a634-79e8-45b6-8f67-1e1c901644d7.pdf

## Supplementary Information

### *Supplementary Note 1. Private and social costs in Figure 1, and teachers who tutor their own students*

*Prices in Figure 1*. Figure 1a plots the full marginal price that households pay for a unit of learning. Households pay the full marginal cost of private tutoring sold in a competitive market, and of AI tutoring from providers that charge their near-zero marginal cost. Public schooling is subsidized, so households pay less than its cost to society. In a simple benchmark, the subsidy lowers the marginal price, choices are informed, and there are no externalities or other distortions. Households below capacity, such as those on $D_1$, then use more public schooling than the level at which its cost to society equals their marginal benefit, so the efficient level lies below the private equilibrium. This benchmark counts a dollar of benefit the same for every household. Demand reflects the ability to pay as well as preferences, not educational need. The benchmark therefore does not say how much schooling any group should receive.

For households at capacity, such as those on D2, capacity rather than price limits their use of public schooling. Beyond A, the private and social equilibria then coincide if tutoring and AI are priced at marginal cost, learning has no externalities, and these households would still fill public capacity at its full cost. Positive externalities make the social value of learning exceed private demand, so the efficient level beyond A then lies above the private equilibrium. Below A, it depends on whether the externality exceeds the subsidy. Tutors' market power or AI providers' markups would also keep the private equilibrium below the efficient level. Externalities, credit constraints and equity goals are the usual reasons for subsidizing schooling.

*What Δ includes*. With capacity at A, public schooling and its costs are the same with private tutoring and with AI tutoring, so they cancel out of Δ, the surplus gain from AI tutoring over private tutoring. Figure 1b treats Δ as the gain of a high-demand household when AI providers charge their near-zero marginal cost; providers that charge more take part of it. This assumes that tutors earn no producer surplus, for example because the rise in S2 reflects households' own time and effort. If tutors earn producer surplus, as in Dang and Rogers (2008), households gain more than Δ by what tutors lose, a transfer rather than a gain to society. The public budget enters through the costs of AI that schools provide, which Δ excludes. Households then pay less than the cost of AI to society, as with public schooling, and the break-even rule in Section 5 counts these costs by comparing programs at full cost.

*Teachers who tutor their own students*. Teachers commonly tutor their own pupils for pay in several countries, mainly where salaries are low, for example Cambodia, India, Lebanon and Nigeria (Bray 2006). Such tutoring may reduce teachers' time and energy in class, and some teachers leave part of the curriculum out of class and teach it in paid lessons (Dang and Rogers 2008). Jayachandran (2014) finds in Nepal that teachers teach less during the school day where schools offer for-profit tutoring to their own students, and suggests that lowering barriers to outside tutors could raise achievement. AI tutoring could be such an outside option, with low barriers. Where tutoring substitutes for classroom teaching, competition from AI could reduce teachers' incentive to withhold instruction in class, and so raise the capacity of public schooling (from A to A′ in Figure 1a). Teachers' market power in tutoring can, however, be greatest in remote rural areas, where access to AI is weakest. None of the trials we review tests this channel.

***Supplementary Note 2. How the evidence was assembled***

*Scope.* We include randomized evaluations of student-facing generative AI used for learning that report a learning outcome, together with a small set of non-generative benchmarks and the teacher-facing trials discussed in Section 4. Sources are journal articles, working papers and technical reports available by October 2026, found through recent syntheses (Burneo *et al.* 2026; Dang and Do 2026; Fesler *et al.* 2026), their references and working-paper series. The set is purposive: it illustrates each kind of randomized comparison and does not support a new meta-analysis.

*Extraction.* For each comparison we record the treatment and control, the outcome, its timing, whether students could use the tool during the test, the samples randomized and analyzed, and the estimate with its uncertainty, citing the table or page of the version used (Supplementary Table S2). Effects are in SD as standardized by the authors unless noted. Where a source reports a coefficient and standard error in raw units, we divide both by the reported control-group SD; where it reports only a p value, we derive the standard error from it; and where it reports a standardized effect with an interval only for the raw coefficient, we scale that interval by the ratio of the two. Derived intervals are marked in Figure 2, and Supplementary Table S2 records, for each, the coefficient, the standard deviation used and the test. "Not reported" means that the source does not give the information.

*Flags.* A dagger marks a risk-of-bias concern noted in a published synthesis, and a double dagger marks evaluations co-authored or led by the tool's provider. These are disclosures, not ratings. Interactions and subgroup effects estimated within trials are listed in Supplementary Table S3 with their measures and outcomes.

*Pooled estimates.* The syntheses quoted in Section 4 overlap with the trials in Figure 2 and with each other, so they are summaries, not independent replications. The pool of Dang and Do (2026) enters Nigeria's composite outcome (0.31 SD), whereas Figure 2 plots its English score (0.24 SD). Burneo *et al.* (2026) exclude the Nigerian trial for attrition, and their contrast between generative and earlier tools compares different studies rather than randomized arms.

*Costs.* Costs are taken as reported, with their perspective (marginal, license only or program cost) and the ingredients left out (Table 1, Panel A). The anchors for Botswana and India in Panel B divide a program's cost per student by its effect on numeracy or mathematics: US$19 / 0.121 SD for Botswana, in nominal dollars, and US$307 / 0.37 SD for India, in 2022 PPP dollars. For the United States, we use the US$13,093 per SD of a composite computed by Burneo *et al.* (2026), also in 2022 PPP dollars.

## *Supplementary Table S1. Use of AI for learning among 15-year-olds in school, gaps by socioeconomic status, and other measures with different populations*

**Panel A. PISA 2025, 15-year-olds enrolled in school**

| Country or economy | Use AI at least weekly to help them learn (%) | Learning to assess AI-generated information at school (%) | Learning to assess AI information, advantaged minus disadvantaged (pp) | Index of AI use for schoolwork, advantaged minus disadvantaged |
|---|---|---|---|---|
| Viet Nam | 68.6 | 70.0<br>(0.8) | 15.9<br>(2.3)[a] | 0.20<br>(0.04) |
| Singapore | 65.6 | 81.0<br>(0.5) | 6.5<br>(1.6) | 0.19<br>(0.04) |
| United Arab Emirates | 64.3 | 81.7<br>(0.3) | 8.1<br>(1.6) | 0.11<br>(0.03) |
| Estonia | 58.4 | 72.5<br>(0.8) | 6.2<br>(1.9) | −0.05<br>(0.04) |
| Colombia | 55.7 | 63.5<br>(0.8) | 17.4<br>(2.3)[a] | 0.18<br>(0.05) |
| Peru | 53.8 | 55.3<br>(1.1)[a] | 23.7<br>(2.6)[a] | 0.29<br>(0.05) |
| Indonesia | 53.5 | 73.2<br>(0.9) | 15.6<br>(2.0)[a] | 0.23<br>(0.04) |
| Korea | 50.7 | 66.5<br>(1.3) | 14.0<br>(2.3) | 0.40<br>(0.05) |
| Brazil | 47.8 | 51.2<br>(0.6)[a] | 13.7<br>(1.7)[a] | 0.31<br>(0.03) |
| Kenya | 45.7 | 55.5<br>(1.4)[a] | 31.8<br>(3.0)[a] | 0.40<br>(0.04) |
| Morocco | 44.9 | 52.7<br>(0.9)[a] | 23.0<br>(2.6)[a] | 0.25<br>(0.04) |
| United States | 37.3 | 56.3<br>(1.3)[a] | 14.5<br>(3.3) | 0.12<br>(0.07) |
| United Kingdom | 31.5 | 42.4<br>(0.9) | 6.3<br>(2.1) | 0.06<br>(0.04) |
| Japan | 27.4 | 31.2<br>(1.3) | 14.9<br>(2.2) | 0.22<br>(0.04) |
| Ireland | 25.9 | 47.2<br>(0.9) | 5.5<br>(2.3) | 0.05<br>(0.03) |
| OECD average | 45.5 | 62.6<br>(0.1) | 9.2<br>(0.3) | 0.19<br>(0.01) |

**Panel B. Other measures (not comparable with Panel A)**

| No. | Measure | Population and year | Estimate |
|---|---|---|---|
| 1 | Used chatbots for help with schoolwork (ever) | US teenagers aged 13–17, 2025 | 54% |
| 2 | Do all or most schoolwork with chatbot help | US teenagers aged 13–17, 2025, by household income | 20% below US$30,000; 7% at US$75,000 or more |
| 3 | Took part in private tutoring | Korea, school students, 2024 | 80% |
| 4 | Received private tutoring | Viet Nam, 2014, by school level | 37% primary; 53% lower secondary; 69% upper secondary |
| 5 | Received private tutoring | England, 2022 | 30% ever; 11% currently |
| 6 | Received private tutoring at some school levels | 23 countries, studies up to 2008 | 25% to 90% |
| 7 | Received private tutoring at secondary school level | 64 countries/economies participating in PISA 2012 | About one-third of all 15-year-old students enrolled in school |

*Source:* Panel A, OECD (2026), Table I.6 and annex Tables I.B1.4.53, I.B1.4.54 and I.B1.4.68 (stat.link/xmra2f). Panel B, by row: 1 and 2, Pew Research Center (2026); 3, Korea Herald (2025); 4, Dang and Glewwe (2018); 5, Bray (2025); 6, Dang and Rogers (2008); 7, Byun *et al.* (2018).

*Note:* Advantaged (disadvantaged) students are in the top (bottom) quarter of the PISA index of economic, social and cultural status in their system. The AI-use index combines how often students use AI chatbots to summarize texts, research new topics, draft writing and help them learn (OECD mean about 0, standard deviation about 1). pp, percentage points. Standard errors in parentheses; a, 50% to 75% of the population covered. The weekly share sums two response categories; its standard error would need their covariance, which the published tables do not report. Of the 84 systems with data, the two gaps are significantly positive (5% level) in 78 and 75, and significantly negative in none. United States: PISA sampling standards not met and limited questionnaire coverage; interpret with caution. Rwanda, Zambia and Cambodia took PISA 2025, but their AI items are missing.

## *Supplementary Table S2. Design, results and sources of the trials in Figure 2 and of the teacher-facing trials in Section 4*

| No. | Trial and sample | What was randomized | Design features | Outcome; timing; tool absent | Effect (SD) [95% CI]; how obtained | Location in the source; version | Status |
|---|---|---|---|---|---|---|---|
| **Panel A. Student-facing trials (Figure 2, panels a to d)** | | | | | | | |
| A1 | Türkiye, high school (grades 9–11); about 1,000 students randomized by classroom; 839 analyzed | Unrestricted GPT-4 or a hint-only GPT-4 tutor vs no AI, in teacher-led practice sessions | Hint-only arm: teacher-written hints and solutions in the prompt; teachers in all arms; no added time | Closed-book exam; same day; tool absent | Unrestricted −0.19 [−0.35, −0.04]; hint-only −0.01 [−0.11, 0.08]; hint-only vs unrestricted 0.18 (no CI). Derived: coefficients −0.054 (SE 0.022) and −0.004 (SE 0.013) on a 0–1 exam grade, with classroom-clustered SEs, divided by the control-arm SD of 0.277; normal approximation | Table 1; accepted manuscript (author-posted) | Peer reviewed |
| A2 | Ghana, grades 3–8 (analysis sample); 11 schools (unit of assignment); 477 analyzed | Math tutor on WhatsApp (Rori) added to regular instruction | Hints, then a worked solution; a teacher present for technical help; two 30-minute sessions a week for 8 months | Math growth scores; 8 months; not stated | 0.36; no SE or CI reported | Section 4.1 of the preprint (v2); conference version | Conference paper; developer co-authored; risk-of-bias concern |
| A3 | Nigeria (Benin City), first-year senior secondary; 1,328 assigned by lottery; 759 took the endline; 654 in the intention-to-treat regressions | After-school program with Microsoft Copilot (GPT-4) vs no program | Teachers' prompts on the English curriculum; teachers guide pairs; twelve 90-minute sessions over 6 weeks | Paper-based English test; end of the 6-week program; tool absent. Also the regular curricular exam a day after the sessions ended | English 0.24 [0.10, 0.37]; composite 0.31; curricular exam 0.21 (SE 0.07). Reported SEs; attrition checks in Tables 10 and 11 | Tables 2 to 4; working paper, revised December 2025 | Working paper; risk-of-bias concern (43% did not take the endline) |
| A4 | Sierra Leone, grades 7–8; 48 classrooms in 12 schools; 1,763 enrolled; 1,423 in the baseline-adjusted analysis (1,637 endline only) | Gemini Guided Learning in math lessons vs usual lessons (classroom level) | Restricted accounts; teacher-led lessons; two students per device; two of four weekly periods, no added time | External math test; 8 weeks (endline); not stated | 0.26 [0.03, 0.49]; interaction with baseline score in Supplementary Table S3. Reported CI | p. 3 and Tables C.4 and C.10; technical report, May 2026 | Provider report; preregistered |
| A5 | United States (Tennessee), grades 6–8; 53 grade-within-school clusters in 18 schools; 6,902 student-terms | Khan Academy with Khanmigo vs usual remedial instruction | Khanmigo set to guide rather than answer; groups of 4–8 with a teacher; no added time | District math test; annual, 2 years; tool absent | 0.06 a year [−0.01, 0.13] (0.02 in year 1; 0.08 in year 2). Reported SE; standardized by the population SD (0.07 [−0.02, 0.15] with the control-group SD, Table 7, Panel A) | Table 7, Panel B; working paper (August 2026) | Working paper |
| A6 | Italy, technical high school, English as a second language; 76 students in four classes, randomized within classes | GPT-4 homework tutor vs usual homework, 8 weeks | Prompted never to give the answer and to ask at least 10 questions per exercise; exercises set by the teacher; no live human support | Teacher-written 24-item test; 8 weeks (endline); not stated | 0.25 [−0.24, 0.74]. Derived: Cohen's d of 0.251 (authors' standardization), with the SE from the reported one-sided $p$ = 0.156; normal approximation | Section 4.4; conference proceedings | Conference paper |

| No. | Trial and sample | What was randomized | Design features | Outcome; timing; tool absent | Effect (SD) [95% CI]; how obtained | Location in the source; version | Status |
|---|---|---|---|---|---|---|---|
| A7 | United States (Maryland), university; 2,379 students, 30 instructors; 1,353 in the exact-match subsample | Course chatbot offered vs not (instructor level, within blocks) | Answers from course materials; tutor mode optional; about 15% of those offered used it | Final course grade; end of course; not an AI-free test | −0.29 [−0.60, 0.03]; exact match −0.37; first-generation differential in Supplementary Table S3. Derived: −2.88 points (SE 1.60) divided by the control-group SD of 10.10; normal approximation | Tables 3 and 4; working paper (October 2026) | Working paper |
| A8 | United States, university physics; 194 students; crossover | Custom GPT-4 tutor at home vs in-class active-learning lesson | Tutor designed by the authors; median 49 minutes with the tutor vs about 60 in class | Post-lesson quiz; same day; not stated | 0.63, adjusted for time on task and other controls; no SE or CI reported; not plotted. Medians 4.5 (n = 142) vs 3.5 (n = 174) student-lesson observations; rank-sum z = −5.6 | Results ('Learning gains: post-test scores' and 'Learning gains: linear regression model') and Supplementary Table S1; journal version | Peer reviewed |
| A9 | Germany, university students in a laboratory; 334 in three arms | Textbook-grounded GPT-4 tutor, unrestricted or after 10 minutes of reading, vs textbook only; 25-minute session | Tutor grounded in the textbook; no human support; equal study time | 25-item test; same day; tool absent | Unrestricted vs textbook 0.34 [0.11, 0.56] (reported SE); either AI arm vs textbook 0.23 [0.02, 0.43]; unrestricted vs restricted 0.21 [−0.01, 0.42]. Derived: difference of the reported effects, 0.337 − 0.132, in the authors' SD units, with the SE from the F-test $p$ = 0.066; normal approximation | Table 1, column 2; working paper (December 2025) | Working paper |
| A10 | United States, college; 204 students | Generative AI allowed alongside traditional resources vs traditional resources only, while learning an unfamiliar topic and writing an essay | Any generative AI tool allowed in the treatment arm; control limited to a search engine, Wikipedia and the library website | Proctored knowledge test; one week later; tool absent | 0.27 [0.03, 0.50]. Reported SE | Table 4, column 5; working paper (July 2026) | Working paper |
| A11 | England, secondary schools, ages 14–15; 405 randomized, 344 analyzed | LLM use vs note-taking vs both, while studying two passages | Instructions for active, comprehension-focused reading; no human support | Comprehension and retention tests; three days later; tool absent | Notes only vs LLM only: d = 0.38 (comprehension) and 0.44 (retention); LLM plus notes vs LLM only: 0.14 and 0.13. Reported d. Derived interval for Figure 2: the reported coefficient 0.95 [0.62, 1.28] scaled by d/coefficient (0.38/0.95) | Table 3; journal version | Peer reviewed; preregistered |
| A12 | United Kingdom, ages 13–15; 165 students (74 tutoring, 91 control) | AI-drafted tutor messages checked by expert tutors vs human tutors alone | Pedagogical model; expert tutors checked every message | First question of the next unit, attempted the same day as the tutoring; not stated | +5.5 percentage points (66.2% vs 60.7%); 95% credible interval −1.4 to 12.4; not in SD | Section 3 (pp. 4–5), Figure 3 and Tables F.5–F.6 (pp. 26–27); technical report, November 2025 | Provider report |
| A13 | United States, adults online; 274 participants | ChatGPT-generated worked-solution hints (screened) vs hints written by human tutors | Hints screened for errors; no human support | Post-test; same day; tool absent | Gains of 17.0 vs 11.6 percentage points; difference not significant ($p$ = 0.42); not in SD | Table 2 and p. 11; journal version | Peer reviewed |

| No. | Trial and sample | What was randomized | Design features | Outcome; timing; tool absent | Effect (SD) [95% CI]; how obtained | Location in the source; version | Status |
|---|---|---|---|---|---|---|---|
| A14 | Chile, university econometrics; 384 randomized; 289 and 303 analyzed | Two separate randomizations against access to the same course chatbot: guidance on learning-oriented use; encouragement to adopt | Guidance on tutor-style use, sent before the final exam; encouragement sent before the midterm; no added time | Final-exam grade (guidance); midterm grade (encouragement); not stated | Guidance 0.22 [0.02, 0.41]; encouragement 0.05 [−0.17, 0.28]. Reported SEs | Table 3, Panel B (p. 18) and Table 2, Panel B (p. 15); working paper | Working paper |
| A15 | United States (Tennessee), grades 6–8; 3,209 students in the mastery arm who took the delayed test | AI tutor inside mastery-based practice vs the same practice without AI | Mastery-based practice in both arms; no added time | Delayed test of practiced and unpracticed items; one week later; tool absent | Practiced item 1: +3.2 percentage points (SE 1.7; $p$ = 0.065); practiced item 2: +1.3 (1.7); unpracticed items 1 and 2: +0.2 (1.7) and +1.7 (1.6); not in SD | Section 5.5 and Table 9; working paper (August 2026) | Working paper |
| A16 | Germany, grades 7–9; 371 randomized, 241 at the posttest | Custom GPT with prompts for learning strategies or utility value vs standard ChatGPT | Prompts to support self-regulated learning; six 45-minute sessions (pretest, four learning sessions, posttest) | Domain knowledge test; after four learning sessions; not stated | No detectable difference (condition-by-time interaction, $\chi^2(1) = 1.28$, $p = 0.26$); effect size not reported | Results, 'Domain-specific knowledge' and Fig. 10; journal version | Peer reviewed |
| **Panel B. Benchmarks without generative AI (Figure 2, panel e)** | | | | | | | |
| B1 | Meta-analysis of 89 randomized trials, preK–12 | Tutoring by teachers, paraprofessionals, volunteers or parents vs no tutoring | Not applicable | Learning outcomes; mixed timing | 0.29 [0.23, 0.34]. Reported SE | Abstract; journal version | Peer reviewed |
| B2 | India (Delhi), middle school; 535 students (lottery) | Mindspark adaptive software plus group instruction, 90 minutes a day after school, vs no access | Adaptive practice at each student's level; teaching assistants lead group instruction | Math test; 4.5 months; not applicable | Math 0.37 [0.24, 0.50]; Hindi 0.23. Reported SE | Table 2, column 1; journal version (author-posted) | Peer reviewed |
| B3 | Botswana, grades 3–5; 2,815 students | Phone calls plus text messages vs no program, 8 weeks | Weekly phone tutorials while schools were closed | Numeracy; 8 weeks; not applicable | 0.12 [0.03, 0.21]. Reported CI | Table 1, Panel A, column 1; journal version | Peer reviewed |
| B4 | China (rural Sichuan), grades 4–6; 8,609 students | Commercial adaptive practice software (machine learning, not generative) vs non-adaptive practice software with the same timetable, or vs no computer-assisted learning, 11 weeks | Two weekly 40-minute sessions in both software arms, usually supervised by math teachers | Math test; 11 weeks; not applicable | Adaptive vs non-adaptive −0.04 [−0.09, 0.01]; adaptive vs none −0.04 [−0.09, 0.01]. Reported SEs | Table 1, column 2; working paper (July 2026) | Working paper |
| **Panel C. Teacher-facing trials discussed in Section 4** | | | | | | | |
| C1 | England, Year 7–8 science teachers; 259 teachers in 68 schools | ChatGPT with a guide vs usual lesson preparation, 10 weeks | Guide on using ChatGPT for lesson preparation | Preparation time; blind expert ratings of sampled resources | 31% less preparation time (about 25 minutes a week); no detectable difference in resource quality; student learning not measured | Tables 1 and 2 (pp. 4–5); evaluation report, December 2024 | Evaluation report |
| C2 | England, primary teachers (Years 3–6); 464 randomized in 108 schools, 416 analyzed | AI lesson assistant (Aila) vs usual lesson preparation | Aila used for lesson preparation in autumn 2025 | Preparation time; resource quality (non-inferiority test) | 24% less preparation time (about 49 minutes a week); resource quality non-inferior; student learning not measured | Tables 1 and 2 (pp. 6–7); evaluation report, October 2026 | Evaluation report |

| No. | Trial and sample | What was randomized | Design features | Outcome; timing; tool absent | Effect (SD) [95% CI]; how obtained | Location in the source; version | Status |
|---|---|---|---|---|---|---|---|
| C3 | United States, Title I schools; 900 tutors randomized, 783 tutors and 1,013 students analyzed | Real-time AI guidance for human tutors vs none (tutor level) | Suggestions based on expert reasoning; tutors deliver all instruction | Topic mastery (exit tickets); same session | +4 percentage points [1.1, 6.9]; +9 for students of lower-rated tutors; not in SD | Abstract, Table 1 (Panel A) and Figure 2a; preprint, version 3 (October 2026) | Working paper; preregistered |
| C4 | Türkiye, middle and high schools in a chain of 24 schools; 193 teachers and 2,816 students in 14 schools (analytic sample) | GPT-4o teaching assistant with training, with or without weekly reminders, vs business as usual (school–department level) | Curriculum materials of the Ministry of Education; one-hour training; one semester | Intrinsic motivation (endline survey) and standardized final exams; end of the semester | Intrinsic motivation −0.11 [−0.20, −0.02]; grades −0.02 [−0.09, 0.05]. Reported CI and SE | p. 6 and Tables S7 and S14 (pp. A5 and A10); working paper (June 2026) | Working paper; preregistered |
| C5 | Peru, sixth-grade teachers; 1,269 teachers in 390 schools | Copilot license and training vs none (school level) | One-day training; 8 months of follow-up by messaging | Teacher time, prompt quality and judgment of AI output; 8 months | No detectable time savings; better prompts and judgment of AI output | Tables 3 to 5; Policy Research Working Paper 11476 (October 2026) | Working paper; preregistered |

*Source:* By row: A1, Bastani *et al.* (2025); A2, Henkel *et al.* (2024); A3, De Simone *et al.* (2025); A4, LearnLM Team and Fab AI (2026); A5, Oreopoulos and Low (2026); A6, Vanzo *et al.* (2025); A7, Liu *et al.* (2026); A8, Kestin *et al.* (2025); A9, Fischer *et al.* (2025); A10, Contractor and Reyes (2026); A11, Kreijkes *et al.* (2026); A12, LearnLM Team and Eedi (2025); A13, Pardos and Bhandari (2024); A14, Gallegos (2026); A15, Oreopoulos *et al.* (2026); A16, Fütterer *et al.* (2026); B1, Nickow *et al.* (2024); B2, Muralidharan *et al.* (2019); B3, Angrist *et al.* (2022); B4, Ma *et al.* (2026); C1, Roy *et al.* (2024); C2, Roy *et al.* (2026); C3, Wang *et al.* (2026); C4, Sungu *et al.* (2026); C5, Glave *et al.* (2026).

*Note:* "Derived" marks intervals computed from other reported statistics (Supplementary Note 2). "Not stated": the source does not say whether students could use the tool during the test. Effects as plotted in Figure 2.

***Supplementary Table S3. Interactions and subgroup effects estimated within the selected trials***

| No. | Trial and sample | Measure | Outcome; timing | Estimate [95% CI] | Location in the source; version |
|---|---|---|---|---|---|
| 1 | Sierra Leone, grades 7–8; 1,423 students | Baseline mathematics score (per SD) | External mathematics test; end of the 8-week program | Interaction 0.195 [0.074, 0.315] SD per SD of baseline score | p. 3 and Table C.10; technical report, May 2026 |
| 2 | Nigeria, first-year senior secondary; 654 students | Baseline exam score (per SD) | Composite score (English, AI knowledge, digital skills); end of the 6-week program | Interaction 0.15 [0.01, 0.29] SD per SD of baseline score | Table 4; working paper, revised December 2025 |
| 3 | Nigeria, as in row 2 (617 students with the index) | Household socioeconomic index (parental education, household goods, internet access, study space) | As in row 2 | Interaction 0.11 [0.02, 0.21] SD per unit of the index | Table 4; working paper, revised December 2025 |
| 4 | Nigeria, as in row 2 | Female (vs male) | As in row 2 | Interaction 0.42 [0.05, 0.79] SD; the authors caution that one girls-only school may drive it | Table 4; working paper, revised December 2025 |
| 5 | United States (Maryland), university; 2,379 students | First-generation status | Final course grade, not an AI-free test; end of course | Effect for first-generation students −0.50 SD; difference from continuing-generation students −0.28 [−0.52, −0.05] (−2.87 points, SE 1.19, over a control-group SD of 10.10) | Table 4; working paper (October 2026) |
| 6 | Germany, university students in a laboratory; 176 in two experiments | Prior knowledge (pre-test below vs above the median) | Understanding (post-test, controlling for the learning phase); same session | Below the median: −0.95 points (SE 0.47; $p = 0.043$); interaction with above the median 1.61 (SE 0.75; $p = 0.034$); not in SD | Section 9.2 and Table 12 (pp. 22–23); working paper, version 2 (March 2025) |
| 7 | United States, Title I schools; 783 tutors and 1,013 students analyzed (teacher-facing tool) | Tutor rating (lower-rated tutors) | Topic mastery (exit tickets); same session | +9 percentage points for students of lower-rated tutors (56% to 65%), against +4 [1.1, 6.9] on average; not in SD | Abstract, Table 1 (Panel A) and Figure 2a; preprint, version 3 (October 2026) |

*Source:* By row: 1, LearnLM Team and Fab AI (2026); 2 to 4, De Simone *et al.* (2025); 5, Liu *et al.* (2026); 6, Lehmann *et al.* (2025); 7, Wang *et al.* (2026).

*Note:* Interactions are differences in the effect per unit of the measure; rows 5 and 6 also give subgroup effects, and row 7 gives only a subgroup effect. Row 6 is a laboratory study cited in Section 4 for this interaction; it is not among the estimates in Figure 2. Measures and outcomes differ, so the estimates are not comparable across trials. Intervals are the estimate ± 1.96 standard errors where the source reports only standard errors.

***Supplementary Table S4. Measures governments have taken on generative AI in schools, 2023 to 2026 (details for Figure 3)***

| No. | Variant | System | Measure (date) | Status, October 2026 | About AI or with AI |
|---|---|---|---|---|---|
| **1. Restrict or scale back** | | | | | |
| 1a | Prohibit, then narrow by age | New York City | ChatGPT blocked on school devices (January 2023); block eased (May 2023); one-year pause on student-facing AI for 2-K to grade 8, about 600,000 students (September 2026) | Pause in force for 2026–27; high schools run five vetted pilots | With |
| 1b | Restrict by age | China | Primary students may not use open-ended generative tools independently; approved-tool lists (May 2025); Beijing requires at least 8 class hours of AI education a year (from September 2025) | In force | Both |
| 1c | Scale back after launch | Korea | AI digital textbooks launched (March 2025); reclassified as optional supplementary materials (August 2025); adoption fell from 37% to 19% of schools | Optional; more than 1.2 trillion won spent by government | With |
| **2. Wait, guide or regulate** | | | | | |
| 2a | Wait: no system-wide rules | Many systems | No system-wide rules on generative AI in schools. Historical baseline: in a 2023 UNESCO poll of more than 450 schools and universities, fewer than 10% had formal guidance on generative AI | No comprehensive count for 2026; the 2023 poll cannot show 2026 positions | n.a. |
| 2b | Issue guidance | England; United States | School guidance (England, updated August 2025); Executive Order 14277 and federal guidance allowing funds for AI-enhanced tutoring (United States, 2025) | In force | Both |
| 2c | Regulate high-risk uses | European Union | AI Act lists education uses (admission, evaluation of learning, proctoring) as high risk; obligations postponed to December 2027 | In force, delayed | With |
| **3. Encourage** | | | | | |
| 3a | Teach about AI (AI literacy) | United Arab Emirates; Viet Nam | AI a subject from kindergarten to grade 12 (UAE, 2025–26); nationwide AI education from 2026–27, 12 periods a year (Viet Nam) | In force | About |
| 3b | Provide AI tools | Estonia | AI Leap: AI tools for about 20,000 upper-secondary students (grades 10–11) and their teachers (teacher training from August 2025; student access from January 2026) | Phase 2 under way | With |
| 3c | Partner with providers | El Salvador; Greece | Grok tutoring announced for more than 5,000 public schools (December 2025); ChatGPT Edu pilot starting with upper-secondary teachers (2025–26) | Announced; pilot | With |

*Source:* By row: 1a, Banks (2023), Elsen-Rooney (2023) and K-12 Dive (2026); 1b, China Daily (2025) and Australian Government Department of Education (2025); 1c, Rest of World (2025); 2a, UNESCO (2023b); 2b, Department for Education (2025), Executive Office of the President (2025) and US Department of Education (2025); 2c, European Union (2024, 2026); 3a, Gulf News (2025) and Ministry of Education and Training of Viet Nam (2026); 3b, Estonian Ministry of Education and Research (2025) and TI-Hüpe (2026); 3c, Associated Press (2025) and OpenAI (2025).

*Note:* Groups and variants as in Figure 3; systems often combine measures from more than one group. For variant 2a, the 2023 poll is a historical baseline only.