Not Faster, but Better: Generative AI and Teacher Skill in Peru

Carla Z. Glave Carolina Lopez Ezequiel Molina Román Andrés Zárate*

November 2026

Abstract

Generative artificial intelligence is being deployed across the public sectors of developing countries on the premise that it makes workers faster. This paper tests this in a randomized trial with 1,270 sixth-grade teachers across 390 public schools in Lima, Peru. A low-cost program—a one-day training, an artificial intelligence license, and eight months of light-touch follow-up—raised adoption sharply yet saved teachers no time on any of their core tasks, with precisely estimated nulls throughout. The gains lay elsewhere: treated teachers wrote markedly better prompts, including on a rubric designed independently of the training, and were sharper at catching weak artificial intelligence outputs. These gains were largest for older teachers, who began furthest behind, narrowing rather than widening skill gaps. Taken together, the results indicate that artificial intelligence training primarily develops the skills needed to generate and evaluate high-quality artificial intelligence outputs, rather than reducing the time teachers spend on their work.

JEL Codes: I21, J24, M53, O33, C93.

Keywords: Artificial Intelligence, ICTs, Edtech, Skills Development & Training, Primary Education

*We thank Marcos Saúl Tupayachi Cárdenas, Director of the Dirección Regional de Educación de Lima Metropolitana (DRELM), for allowing us to conduct this research. We are grateful to Alexandra Gutiérrez Traverso, María Andreína Cantele, and Jostin Kitmang for outstanding research assistance. The study was funded by the Human Capital Project of the World Bank and the Microsoft AI Economy Institute. The research protocol was approved by the University of Toronto (IRB approval number 48010) and registered in the AEA RCT registry (Lopez et al., 2025). The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent. All errors are our own. Glave: UW-Madison, carla.glave@wisc.edu, Lopez: World Bank, carolina_lopez@worldbank.org, Molina: World Bank, molina@worldbank.org, Zárate: University of Toronto, ra.zarate@utoronto.ca.

1 Introduction

A rapidly growing experimental literature documents large productivity gains from generative AI: 15 percent more calls resolved per hour by customer support agents (Brynjolfsson et al., 2025), 40 percent faster professional writing (Noy and Zhang, 2023), and 40 percent higher-quality consulting deliverables (Dell'Acqua et al., 2026). This evidence comes almost entirely from private sector professionals in high-income countries, performing tractable tasks assigned and measured by the experimenter. Whether these gains extend to complex, self-directed work in public sector settings in developing countries—where AI training programs are being actively scaled—has not been directly tested.

Whether generative AI raises the productivity of a given worker depends on two things: whether the worker uses it and how well. We study both in a school-level randomized controlled trial covering 1,270 sixth-grade teachers across the 390 public schools in Metropolitan Lima that operate at this grade level. Treatment schools (196 schools, 660 teachers) received a one-day, in-person training built around a general framework for using large language models in instruction, along with a Microsoft Copilot license and eight months of WhatsApp-based follow-up. Control schools (194 schools, 610 teachers) received no intervention but retained access to publicly available AI tools. Randomization was stratified by administrative district and by whether the school's baseline math performance was above or below the median. We tracked all teachers for eight months through administrative platform data and a blind-graded endline survey, with no differential attrition between treatment and control.

Training did not make teachers faster, but it made them substantially more competent users of AI. Activation of the Copilot license rose by 20.5 percentage points (p.p.) over a control mean of 2.8 percent, and the share of teachers naming Copilot their most-used AI tool rose by 29.3 p.p. over a control mean of 9.6 percent. The time savings that define the existing AI-productivity literature, however, did not appear. Across seven core professional tasks—lesson planning, differentiating materials, designing assessments, grading, collaborative work, administrative tasks, and parent communication—treated teachers saved no time, with precisely estimated null effects throughout. Higher adoption did not translate into a faster workday.

The training, however, changed the quality of teachers' interaction with AI. Blind-graded prompts written by treated teachers scored 0.29 points higher on a 5-point rubric built from the training content, and 0.38 points higher on a rubric designed independently of the training. Treated teachers were also better at judging AI output: they were 10 p.p. less likely to misjudge the quality of a weak prompt and 6.7 p.p. less likely to accept an inadequate AI response as good. That the effect on the independent rubric is no smaller than the effect on the training-based one weighs against a teaching-to-the-test interpretation: the training built a transferable skill for working with AI rather than proficiency with a fixed set of prompts. The intervention also shifted the composition of use toward professional work: treated teachers were 17 p.p. less likely than controls to report using AI more for personal than for professional purposes, indicating that the program built specifically professional human capital rather than general familiarity with the technology.

These gains were not concentrated among the teachers best positioned to capture them. The effect on prompt quality was largest for teachers above the median age of 54—those who scored lowest in the

absence of training—so that the program narrowed, rather than widened, the age gap in AI skill. To our knowledge, this is the first randomized evidence that the quality of a worker's interaction with generative AI—rather than access or speed—is the margin that responds to training, and that this margin can be built at scale, and most for those furthest behind, in a developing-country public sector.

These results speak to two strands of the literature. The experimental literature on AI and worker productivity has documented time savings on tractable, assigned tasks in high-income private-sector settings (Brynjolfsson et al., 2025; Noy and Zhang, 2023; Dell'Acqua et al., 2026; Cui et al., 2026), with only one prior experiment in a developing country—varying access alone among Kenyan entrepreneurs, and finding no average effect (Otis et al., 2024). We identify a distinct channel that dominates when the underlying task is open-ended: not the speed of a fixed task but the quality of what the worker can get the tool to produce, a form of human capital that structured training can build (Acemoglu and Restrepo, 2019). In parallel, two decades of experiments on educational technology in developing countries have established that making a technology available to schools rarely changes practice on its own (Cristia et al., 2017; Malamud and Pop-Eleches, 2011; Rodriguez-Segura, 2022). We show that the same lesson reappears with generative AI, but that a low-cost package—one day of training, a dedicated license, and light-touch follow-up over the school year—supplies the skill that access alone does not. The design follows what is known about teacher behavior change: that integration-specific confidence predicts classroom use more than access (McGehee, 2024; Gomez et al., 2022), and that structured programs with sustained support outperform one-shot workshops (Kraft et al., 2018; Darling-Hammond et al., 2017). Whether these gains in teachers' AI fluency translate into measurable improvements in student learning is an open question we address in ongoing work.

The remainder of the paper is organized as follows. Section 2 describes the educational context in Peru and the structure of the training program. Section 3 presents the experimental design, the sample, and the data. Section 4 outlines the empirical strategy. Section 5 reports the main results on adoption, time use, and the quality of teachers' interaction with AI. Section 6 concludes.

2 Context and Intervention

2.1 Educational Context in Peru

Peru provides a useful setting for studying whether structured training can build teachers' capacity to use generative AI for professional work. The Peruvian education system combines a nationally defined curriculum and discretion at the classroom level. The Ministry of Education defines the *Curriculo Nacional de la Educacion Basica* (CNEB), the national curriculum framework for basic education. The CNEB specifies the competencies students are expected to develop, their progression across grade levels, the expected standards by cycle and modality, and guidance for formative assessment and curricular diversification (MINEDU, 2016). In practice, this means that teachers are expected not only to cover grade-level content, but also to translate broad national competencies into lesson plans, learning activities, assessments, rubrics, feedback, and differentiated materials.

This institutional context is especially relevant for the use of generative AI. Unlike a textbook or a scripted lesson plan, a general-purpose AI tool does not provide a fixed instructional sequence. Its output

depends on the quality of the instructions supplied by the user. For teachers, effective use therefore requires the ability to describe the student audience, specify the learning objective, request an appropriate format, and assess whether the output is pedagogically sound and aligned with the national curriculum and with their class context. These are professional skills, not simply access conditions. The same tool can produce a useful lesson plan, a vague worksheet, or an inappropriate explanation depending on the teacher's prompt and subsequent interaction with the response.

Our study takes place in Metropolitan Lima, the largest urban public education system in the country. Public schools in Lima are overseen by the Direccion Regional de Educacion de Lima Metropolitana (DRELM), which operates through seven local education management units (UGELs). Schools are also organized into educational networks through which principals and teachers participate in coordination meetings, professional development activities, and administrative communication. These networks constitute an important part of the institutional environment in which new practices circulate. Teachers and principals in the same network may exchange information about training opportunities, instructional materials, and the usefulness of new tools, making the network a natural unit of social and professional exchange within the system.

Metropolitan Lima is also a policy-relevant environment because teachers are not operating in a setting of complete digital exclusion. Evidence from Peru's national teacher survey, the Encuesta Nacional a Docentes (ENDO), shows that in 2021 most public school teachers reported access to basic digital devices and services: 99 percent had a cellphone, 81 percent had a laptop or tablet, 37 percent had a computer, 88 percent had a mobile data plan, and 63 percent had internet at home (MINEDU, 2021). Further, recent evidence from the Ministry's national assessment (ENLA) in 2025 indicates that teachers nationwide are already using AI, as roughly 70 percent of teachers in grades 4, 8 and 11 report using AI to create math teaching materials and 40 percent report using it for this purpose for half of their classes or more. These numbers suggest that many teachers have the minimum infrastructure required to access online tools, and that they are already using new technologies like artificial intelligence. However, access to devices and connectivity does not imply the ability to use a new technology productively for instructional work.

The need for effective instructional support is also salient given Peru's persistent learning challenges. In PISA 2022, Peru continued to exhibit low levels of mathematics proficiency and substantial inequality in performance. Socioeconomically advantaged students outperformed disadvantaged students by 86 score points in mathematics, and socioeconomic status accounted for 17 percent of the variation in mathematics performance (OECD, 2023). These patterns imply that teachers face classrooms with considerable heterogeneity in student preparation. In such settings, tools that can help generate differentiated materials, explanations, examples, and feedback are valuable. However, the potential value of these tools depends on whether teachers can use them in ways that respond to students' needs rather than simply producing generic materials.

The setting therefore illustrates a broader challenge for education technology policy. Governments and education systems increasingly face pressure to provide access to digital tools, including AI platforms. But if the binding constraint is the human capital required to use the technology well, then access alone is unlikely to change professional practice. This is particularly true for generative AI, where the

quality of the output is jointly produced by the model and the user. The intervention we evaluate was designed around this premise: the relevant policy question is not only whether teachers can access AI, but whether structured training can build the skills required to use AI for instructional work.

2.2 The Intervention

The intervention was designed to test whether a low-cost package of access, structured training, and low-cost follow-up could build teachers' capacity to use generative AI for professional tasks. Teachers assigned to treatment received access to Microsoft Copilot through institutional education accounts provisioned in partnership with DRELM. The platform could be accessed from a phone or computer with an internet connection and did not require schools to purchase additional hardware. Control schools did not receive the licenses or the training, although teachers retained access to publicly available AI tools.

The central component of the intervention was a one-day, in-person training workshop conducted at the beginning of the 2025 school year. The workshop introduced teachers to the basic logic of large language models, hands-on examples of the types of teaching tasks for which they could be used, prompt templates, and the limitations of AI-generated outputs. Rather than presenting AI as a general-purpose productivity tool, the training focused on concrete tasks teachers routinely perform: lesson planning, generating practice exercises, adapting materials to students' learning levels, preparing assessments, designing rubrics, giving feedback, and drafting communications to families. The training materials were co-designed with 12 teachers who had won an innovation contest from DRELM in 2024.

A core feature of the workshop was a structured prompting framework organized around four elements: *who*, *what*, *how*, and *CNEB*. Teachers were trained to specify the intended student audience or student profile (*who*); the content, topic, or learning objective (*what*); the desired format, pedagogical approach, or constraints of the output (*how*); and the relevant competencies or curricular expectations from the national curriculum (*CNEB*). The framework was intended to reduce the cognitive cost of using AI by converting an unfamiliar interaction with a chatbot into a repeatable professional routine. Teachers practiced applying the framework to their own upcoming lessons and received examples that could be adapted to common classroom tasks.

The training also emphasized critical evaluation. Teachers were encouraged not to treat AI outputs as finished products, but as drafts requiring professional judgment. This component is important because generative AI can produce fluent but incomplete, inaccurate, or pedagogically weak responses. For teachers, the relevant skill is therefore not only prompt construction but also output assessment: identifying whether an AI-generated lesson is aligned with the learning objective, whether an explanation is age-appropriate, whether a worksheet matches students' level, and how much a response should be revised before being used in class.

After the in-person workshop, treatment teachers received weekly WhatsApp messages. These messages provided prompt examples, reminders, and practical suggestions for applying the framework to common teaching tasks. The follow-up was designed to support sustained use after the initial training and to reduce the likelihood that teachers would treat AI as a one-time novelty. The WhatsApp component also created variation in two dimensions. First, the messenger varied across arms, comparing messages sent by an education specialist from DRELM's Office of Pedagogical Administration with

messages sent by another teacher. Second, the content varied across arms, comparing general AI-use support with messages that emphasized adapting instruction to students' learning levels, consistent with teaching-at-the-right-level principles. These cross-randomized components allow us to examine whether sustained skill formation depends more on institutional authority, peer encouragement, or pedagogical framing.

3 Experimental Design

3.1 Sample and Randomization

The initial sample comprises all schools in Lima Metropolitana that offer sixth grade of primary education. The Dirección Regional de Educación de Lima Metropolitana (DRELM) organizes schools into administrative networks in which principals and teachers frequently interact and share information. Given the potential for spillovers within networks, the experimental design incorporates two stages of randomization.

In the first stage, we randomized eligibility for the intervention at the network level. There are 143 school networks in total, with an average of six schools per network. We randomly selected 84 of the 143 networks to be eligible for the intervention. To ensure balance between treated and control networks, we implemented the randomization procedure 1,000 times and selected the assignment that maximized the minimum p-value across a set of baseline characteristics, including school size, instructional capacity, network size, participation in the national standardized assessment, average fourth-grade academic performance (math, reading, and socioemotional indicators), dispersion in these outcomes, and baseline diagnostic test scores for sixth-grade students in math and reading.

In the second stage, we randomized treatment assignment at the school level within the networks selected for eligibility. This randomization was stratified by UGEL (the local school district) and by whether the school's baseline average math performance was above or below the median. As in the first stage, we implemented the randomization 1,000 times and selected the allocation that maximized the minimum p-value across the same set of baseline characteristics described above. The final experimental sample consists of 390 schools, of which 196 were assigned to the treatment group and 194 to the control group.

In addition to the main treatment, the experimental design cross-randomized two additional interventions. The first subtreatment varies the identity of the person responsible for monitoring progress and guiding the application of the tools introduced during the in-person training session. In the *authority* arm, messages were sent by an education specialist from the Office of Pedagogical Management of DRELM. In the *peer networks* arm, messages were sent by a teacher selected based on their participation during the in-person training session and their active use of the AI license provided as part of the program.

The second subtreatment varies the content of the messages. While all teachers received general messages with examples of prompts that could be used in the classroom, this subtreatment assigned messages that either focused on general instructional support or emphasized *teaching at the right level* (TaRL) practices, encouraging teachers to adapt instruction to the heterogeneity in students' achievement levels.

3.2 Data Collection

This study draws on five complementary data sources collected throughout the intervention period. First, we use administrative school and teacher records provided by DRELM and school principals prior to the training. These data define the study sample and provide baseline information on teacher and school characteristics used to assess balance and improve the precision of treatment effect estimates.

Second, we use attendance records from the March training sessions, which document participation in the intervention and serve as our primary measure of program take-up. These records allow us to examine whether teachers assigned to the treatment group attended the training sessions offered by the program.

Third, we use an immediate post-training survey administered at the conclusion of the training sessions. This survey was used primarily to verify and consolidate the roster of participating teachers and to collect information on time-invariant teacher sociodemographic characteristics that are included as baseline covariates in our analysis.

Fourth, we use administrative data provided by DRELM on the use of Microsoft Copilot licenses assigned through the program. These weekly usage reports allow us to track activation and engagement with the AI tool over time and provide an objective measure of technology adoption among participating teachers.

Fifth, we use endline survey data collected near the end of the school year, in December. The endline survey gathered information on teachers' use of AI, teaching practices, attitudes toward technology, and other outcomes of interest. All teachers in both the treatment and control groups were invited to complete the survey during in-person sessions that also included additional AI training. After approximately 300 teachers attended these sessions, an online version of the survey was distributed to the remaining teachers to maximize response rates. These survey data constitute our primary source for measuring the program's main outcomes.

4 Empirical Strategy

As the intervention was randomly assigned, we estimate the causal effects of the program using an intention-to-treat (ITT) framework, comparing outcomes between individuals in schools assigned to treatment and those in schools assigned to control. Our main specification is:

$$y_{is} = \alpha + \beta T_s + \Gamma' X_{is} + \gamma_{\mu(s)} + \varepsilon_{is}, \tag{1}$$

where y_{is} denotes the outcome of teacher or student i in school s, and T_s is an indicator equal to one if school s was assigned to the treatment group. The vector X_{is} includes a set of baseline covariates chosen using the double post-lasso procedure to improve precision, $\gamma_{\mu(s)}$ are fixed effects for the randomization strata $\mu(s)$ (defined by UGEL and whether baseline school math performance is above or below the median), and ε_{is} is an idiosyncratic error term. The coefficient of interest, β, captures the ITT effect of assignment to the program. In all specifications, standard errors are clustered at the school level, the unit of treatment assignment.

5 Results

5.1 Balance, Program Take-up, and AI usage

We begin by assessing baseline balance for the main treatment and the two subtreatment arms. Table 1 reports the results of these balance tests, based on estimates of the parameter β in equation 1, using baseline teacher and school characteristics as outcome variables. Overall, the evidence suggests that treatment assignment is well balanced across these characteristics, with no statistically significant differences between treated and control schools. Likewise, Appendix Tables A1 and A2 report balance tests for the two subtreatments. These tables show that the subtreatment arms are balanced relative to both the control group and each other.

We next describe the treatment effects on teacher outcomes. Table 2 presents the effects on program take-up. The results show that the treatment increases the probability of attending the in-person training session by 48.1 p.p. (s.e. 3.5 p.p.), relative to a control-group participation rate of 12.8%. Some teachers in control schools also attended the training session because the official invitation documents issued by DRELM mistakenly included a subset of control schools.

In addition to participation in the training session, we examine whether teachers activated the Copilot licenses they received as part of the program. Column 2 of Table 2 shows that treatment increased the likelihood of activating a Copilot license by 20.4 p.. (s.e. = 2.1 p.p.). This corresponds to an average increase of 3.15 active weeks on the platform (s.e. = 0.34) over the 28-week program period (column 3). Column 4 shows that treatment did not affect the likelihood of completing the endline survey: the estimated effect is 2.8 p.p. and it is not statistically significant (s.e. = 3.1 p.p.). Treated participants were, however, more likely to complete the survey in person rather than online. We therefore control for survey modality in all survey-based analyses. The results are robust to alternative specifications that either include or exclude this control.

We next turn to AI-use outcomes from the survey data. Column 5 of Table 2 reports the effect on the likelihood that teachers identify Copilot as their most frequently used AI tool. We find an increase of 29.3 p.p. (s.e. 3.5 p.p.), relative to a control-group usage rate of 9.6%. Moreover, the treatment increases overall AI use. Column 6 shows that the general index of AI use rises by 0.214σ relative to the control mean. Finally, Column 7 indicates that teachers in treated schools are less likely than teachers in control schools to report using AI more for personal purposes than for teaching purposes. The treatment reduces this likelihood by 17.2 p.p. (s.e. 4.9 p.p.), from a control mean of 61.9%.

Panels B and C of Table 2 report first-stage results by subtreatment arm. Overall, the results in Panel B show larger effects for the authority treatment than for the peer treatment on some outcomes, such as being an active user and the number of active weeks. However, there are no statistically significant differences between the two treatments in training attendance or survey completion. Finally, the results in Panel C show no statistically significant differences between the TaRL and non-TaRL subtreatments for any of the first-stage outcomes.

5.2 Time Allocation, Prompt Quality, and Equity

Next, we examine how the treatment affected teachers' time allocation across different activities. Panel A of Table 3 reports treatment effects on whether teachers indicate in the survey that they performed each activity, while Panel B reports effects on the number of minutes spent on each activity. For the time-use measures, we assign zero minutes when a teacher reports not engaging in the corresponding activity.

Overall, we find no evidence that the treatment affected teachers' time allocation. Treated and control teachers report similar rates of engaging in lesson planning, differentiating instruction based on student achievement, designing assessments, grading, collaborating with other teachers, performing administrative tasks, and communicating with parents. For most outcomes, we cannot reject the null of no effect on either the extensive margin (whether the activity was performed) or the intensive margin (time spent on the activity). If anything, treated teachers are somewhat less likely to report differentiating materials by students' achievement levels. However, even for this activity, we do not observe significant differences in the amount of time spent. Appendix Table A3 reports time-use results by subtreatment arm and shows no meaningful differences across subtreatments.

While the treatment did not change time allocation and had only modest effects on general AI use, we find significant improvements in prompt quality. Table 4 reports treatment effects on prompt quality based on survey responses. In the survey, teachers were asked to complete a set of tasks designed to capture both prompt generation and evaluation skills. First, they were asked to write the full prompt they would give to an AI system to design a lesson on improper fractions for sixth-grade students. Specifically, teachers were instructed to write the complete instruction they would provide to the AI for this task. Second, teachers were asked to assess the quality of a simple prompt ("Give me exercises on improper fractions for 6th grade") and to improve this basic prompt by writing a more effective and detailed version.

Panel A of Table 4 presents the results on prompt quality using a rubric derived from the program's training materials. Specifically, the rubric evaluates whether the prompt clearly specifies a *who* (column 2), referring to the intended audience for the material (e.g., grade level, student ability, or specific learning needs); a *what* (column 3), referring to the task or objective the teacher wants the AI to perform (e.g., generate a lesson plan, create assessment questions, or explain a concept); and a *how* (column 4), capturing the extent to which the teacher provides guidance on how the output should be structured or delivered (e.g., level of detail, format, tone, or pedagogical approach). The rubric also captures whether the prompt explicitly connects the task to the Peruvian national curriculum (column 5). Column 1 reports a general assessment based on these four measures.

Panel B, in contrast, reports treatment effects using a more general rubric that is not directly tied to the training materials. In particular, we consider four dimensions of prompt quality. First, clarity (column 2) captures whether the prompt is written in a way that is easy to understand and free of ambiguity. Second, specificity (column 3) measures the extent to which the prompt provides precise and concrete instructions—such as clearly defined objectives, constraints, or outputs—rather than vague or open-ended requests. Third, context provision (column 4) evaluates whether the prompt includes relevant background information (e.g., subject matter, classroom setting, or learning goals) that helps guide the AI's response. Finally, well-specified components (column 5) capture whether the prompt integrates

multiple relevant elements (such as audience, task, and structure) in a coherent way. As in Panel A, column 1 reports an aggregate quality index combining these four components.

Overall, the treatment substantially improved the quality of teachers' AI prompts. Across the different quality measures in Panels A and B, we find positive and statistically significant effects ranging from 0.18 to 0.50 points, corresponding to increases of 4.4% to 17.2% relative to the control mean. The aggregate indices in Column 1 show effects of 29.3 points in Panel A and 37.5 points in Panel B, which correspond to increases of 9.2% and 11.1% relative to the control mean, respectively. Appendix Table A4 shows no statistically significant differences in prompt quality across subtreatments.

In addition to improving the quality of their prompts, teachers are also more likely to identify how to improve the response given by the AI. In particular, during the survey we asked teachers to evaluate and refine an AI-generated output in a realistic classroom scenario. Teachers were asked to imagine that a sixth-grade teacher requested an AI system to generate a worksheet on multiplying decimals for a math class and received a draft response. They were then asked whether they would use the worksheet as is, or whether it required improvements in formatting or content. Teachers were also asked how they would improve the worksheet, whether by revising it themselves, asking the AI to improve it, or combining both approaches, and to write the specific instruction they would give to the AI to enhance the material.

Table 5 reports the results on these measures. Column 1 reports whether the teacher would use AI to improve the lesson plan, Column 2 whether the teacher gave an incorrect answer to the prompt assessment, Column 3 whether the teacher considered the AI response to be good, and Columns 4 and 5 whether the teacher considered the response inadequate due to content or formatting, respectively.

The results show that teachers in treated schools are more likely to identify mistakes in AI-generated responses and how to improve them. Column 1 shows an increase of 9.4 p.p. (s.e. 3.9 p.p.) in the likelihood that teachers would use AI to improve the response, relative to a baseline of 68.7% in control schools. Moreover, teachers in treated schools are 10.0 p.p. less likely (s.e. 4.7 p.p.) to incorrectly assess the quality of the prompt and 6.7 p.p. less likely (s.e. 2.5 p.p.) to consider the AI response to be good, as shown in Columns 2 and 3, respectively. While the estimates in Columns 4 and 5 are not statistically significant, the pattern suggests that these teachers are more likely to attribute deficiencies in the AI output to content-related issues rather than formatting concerns. As with the results on prompt quality, Appendix Table A5 shows no statistically significant differences across subtreatments in teachers' identification of errors or their proposed responses in these prompt scenarios.

Finally, as complementary evidence we assess teachers' beliefs about equity, including which groups of teachers benefit most from AI use and their willingness to pay. In particular, we ask teachers to identify which groups of teachers would benefit the most from AI, whether they believe AI could reduce disparities, and whether—and how much—they would be willing to pay for AI.

Appendix Table A6 reports the ITT estimates on these beliefs. Column 1 reports whether teachers think AI benefits a particular group of teachers, Column 2 whether they believe it benefits everyone, and Column 3 whether they think AI will reduce disparities among teachers. Columns 4–6 report different measures of willingness to pay: whether they would pay for AI (Column 4), whether they would pay the market price or more (Column 5), and the price bracket they would be willing to pay (Column 6).

Overall, the results show that teachers in treated schools are more likely to believe that AI use would

benefit a particular group of teachers rather than everyone, with no statistically significant effects on their beliefs about disparities. Teachers in treated schools are 7.4 p.p. more likely (s.e. 3.9 p.p.) to report that AI benefits a particular group (Column 1) and 8.2 p.p. less likely (s.e. 4.6 p.p.) to report that it benefits everyone (Column 2). However, we do not find statistically significant differences in their perceptions of whether AI can reduce disparities among teachers. Finally, Columns 5 and 6 show that willingness to pay is lower among teachers in treated schools, likely reflecting anchoring effects from the free provision of AI as part of the program. Panels B and C of Appendix Table A6 report treatment effects by subtreatment. The only statistically significant differences are that non-TaRL teachers are more likely to believe that AI reduces disparities, and teachers in the peer treatment are less likely than those in the authority treatment to pay the market price or more for the Copilot license.

5.3 Heterogeneity

Finally, we also examine whether treatment effects differ across school and teacher characteristics. Specifically, we test for heterogeneity by school performance level, teacher gender, and teacher age, estimating interaction models between teachers characteristics and the treatment. Overall, the treatment effects are mostly consistent across subgroups, as the gains in take-up, active engagement, and prompt quality are present for both low- and high-performing schools and male and female teachers.

However, we find substantial heterogeneity in treatment effects by teacher age, with older teachers benefiting more from the intervention than their younger counterparts. Appendix Table A7 shows that teachers above the median age of 54 experience significantly larger gains across most dimensions of both the training-based and general prompt-quality rubrics. Although older teachers perform worse in the control group, the interaction between treatment assignment and age is positive and statistically significant for most outcomes, suggesting that the program substantially reduces age-related gaps in AI prompting skills.

6 Conclusion

We studied whether a low-cost training program can build public school teachers' capacity to use generative AI for professional work, in a setting where access to the technology was no longer the binding constraint. The program raised adoption sharply but saved teachers no time on their core tasks. Instead, it changed instead was the quality of teachers' interaction with AI: treated teachers wrote better prompts and were better at judging when an AI output was inadequate. Because the gains appeared on a rubric designed independently of the training, and not only on one drawn from its content, they reflect a transferable skill for working with AI rather than familiarity with a fixed set of prompts.

This pattern clarifies what generative AI does for open-ended professional work. The early experimental evidence, drawn largely from tractable tasks in high-income private-sector settings, framed AI as a tool that makes a given task faster. In a teacher's job—where the task is to translate a broad curricular objective into a lesson, an assessment, or a differentiated worksheet—the margin that moves is not speed but the quality of what the worker can get the tool to produce, and that quality is itself a form of human capital that training can build. The result also extends a long-standing lesson from educational

technology in developing countries: providing a technology rarely changes practice on its own, but a modest package of training and light-touch support can supply the skill that access alone does not. That the largest gains accrued to the teachers who started furthest behind suggests such programs can narrow, rather than widen, skill gaps within the workforce.

These findings come with a clear limit. We measure teachers' AI fluency, not student learning, and higher-quality prompts need not translate into better outcomes in the classroom. Whether the skills documented here improve instruction and student achievement is the question we take up in ongoing work. What this study establishes is the necessary first link: that the capacity to use generative AI productively can be built among in-service teachers, at low cost and at scale, in exactly the kind of public sector setting where these tools are now being deployed.

References

Acemoglu, Daron and Pascual Restrepo, "Automation and New Tasks: How Technology Displaces and Reinstates Labor," *Journal of Economic Perspectives*, 2019, *33* (2), 3–30.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond, "Generative AI at Work," *The Quarterly Journal of Economics*, 2025, *140* (2), 889–942.

Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Eugenio Severín, "Technology and Child Development: Evidence from the One Laptop per Child Program," *American Economic Journal: Applied Economics*, 2017, *9* (3), 295–320.

Cui, Zheyuan Kevin, Mert Demirer, Sonia Jaffe, Leon Musolff, Sida Peng, and Tobias Salz, "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers," *Management Science*, 2026. Forthcoming. https://doi.org/10.1287/mnsc.2025.00535.

Darling-Hammond, Linda, Maria E. Hyler, and Madelyn Gardner, "Effective Teacher Professional Development," Technical Report, Learning Policy Institute, Palo Alto, CA 2017.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality," *Organization Science*, 2026, *37* (2), 403–423. Working Paper 24-013, Harvard Business School, 2023.

Gomez, Frank C., Jesus Trespalacios, Yu-Chang Hsu, and Dazhi Yang, "Exploring Teachers' Technology Integration Self-Efficacy through the 2017 ISTE Standards," *TechTrends*, 2022, *66* (2), 159–171.

Kraft, Matthew A., David Blazar, and Dylan Hogan, "The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence," *Review of Educational Research*, 2018, *88* (4), 547–588.

Lopez, Carolina, Ezequiel Molina, and Roman Andres Zarate, "AI in the Classroom: Evaluating the Impact of Teacher Training on Practices and Student Outcomes," AEA RCT Registry July 2025. AEARCTR-0016336.

Malamud, Ofer and Cristian Pop-Eleches, "Home Computer Use and the Development of Human Capital," *The Quarterly Journal of Economics*, 2011, *126* (2), 987–1027.

McGehee, Nikolas, "Breaking Barriers: A Meta-Analysis of Educator Acceptance of AI Technology in Education," Technical Report, Michigan Virtual Learning Research Institute, Lansing, MI 2024.

MINEDU, "Curriculo Nacional de la Educacion Basica," 2016. Lima: Ministerio de Educacion.

___ , "Encuesta Nacional a Docentes de Instituciones Educativas Publicas de Educacion Basica Regular: ENDO Remota 2021," 2021. Lima: Ministerio de Educacion.

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, 2023, *381* (6654), 187–192.

OECD, "PISA 2022 Results: Country Note – Peru," 2023. OECD Publishing.

Otis, Nicholas G., Rowan Philip Clarke, Solène Delecourt, David Holtz, and Rembrand Koning, "The Uneven Impact of Generative AI on Entrepreneurial Performance: Evidence from a Field Experiment in Kenya," *Working Paper*, 2024. Available at SSRN: https://ssrn.com/abstract=4671369.

Rodriguez-Segura, Daniel, "EdTech in Developing Countries: A Review of the Evidence," *The World Bank Research Observer*, 2022, *37* (2), 171–203.

Appendices

TABLE 1: Descriptive Statistics and Balance

	Control Mean (1)	Treatment Mean (2)	Difference (3)	Complier Mean (4)
A. Teacher Characteristics				
Age	52.227	52.327	0.053	52.143
	[0.407]	[0.362]	(0.562)	(0.593)
Gender	0.802	0.792	-0.012	0.809
	[0.017]	[0.016]	(0.023)	(0.026)
Has university degree	0.785	0.746	-0.036	0.778
	[0.019]	[0.019]	(0.026)	(0.028)
Teaching experience (years)	24.115	24.334	0.135	23.980
	[0.490]	[0.418]	(0.675)	(0.664)
Teach. exp. current school (years)	12.930	12.896	-0.036	11.873
	[0.554]	[0.507]	(0.796)	(0.861)
6th grade teach. exp. (years)	6.198	6.395	0.183	6.241
	[0.321]	[0.300]	(0.438)	(0.481)
No additional employment	0.771	0.755	-0.015	0.729
	[0.019]	[0.019]	(0.028)	(0.032)
Responded to endline	0.470	0.486	0.021	0.502
	[0.020]	[0.019]	(0.032)	(0.042)
B. School Characteristics				
Total students	91.402	90.224	-1.288	
	[3.201]	[3.183]	(4.442)	
Total classes	3.082	3.092	0.011	
	[0.098]	[0.100]	(0.141)	
Network size	5.593	5.556	-0.041	
	[0.153]	[0.153]	(0.212)	
Mean math score	0.127	0.132	0.007	
	[0.031]	[0.034]	(0.045)	
Mean language score	0.142	0.103	-0.037	
	[0.029]	[0.038]	(0.046)	
Share of tech devices	0.614	0.624	0.011	
	[0.016]	[0.016]	(0.022)	

Notes: Difference column reports coefficients from regressions of each variable on treatment indicator with strata fixed effects, clustering at school level. Complier mean (Panel A only) from instrumental variable regression on the interaction of each covariate with compliance regressed on compliance, which in turn uses treatment as instrument.

TABLE 2: First Stage: Take-up of AI Training Program

	Attended training	Active user	Active weeks	Responded endline	Copilot most used	AI Tasks Index	Personal $\geq$ Prof
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Main treatment effect							
Treatment	0.488***	0.204***	3.154***	0.019	0.293***	0.214**	-0.172***
	(0.035)	(0.021)	(0.341)	(0.031)	(0.035)	(0.096)	(0.049)
Control mean	0.128	0.028	0.337	0.470	0.096	-0.000	0.619
N	1,269	1,269	1,269	1,269	531	531	448
B. Authority cross-randomization							
Peers effect	0.497***	0.164***	2.427***	-0.011	0.254***	0.267**	-0.116*
	(0.047)	(0.027)	(0.424)	(0.040)	(0.047)	(0.128)	(0.062)
Authority effect	0.514***	0.265***	4.234***	-0.008	0.368***	0.205*	-0.173***
	(0.042)	(0.031)	(0.517)	(0.040)	(0.047)	(0.109)	(0.058)
p-value Auth − Peers	0.759	0.011	0.005	0.954	0.049	0.623	0.397
Control mean	0.128	0.028	0.337	0.470	0.096	-0.000	0.619
N	1,269	1,269	1,269	1,269	531	531	448
C. TaRL cross-randomization							
Non-TaRL effect	0.497***	0.209***	3.368***	-0.032	0.349***	0.254**	-0.123*
	(0.045)	(0.028)	(0.457)	(0.039)	(0.045)	(0.117)	(0.064)
TaRL effect	0.513***	0.220***	3.303***	0.012	0.286***	0.217*	-0.164***
	(0.045)	(0.030)	(0.507)	(0.040)	(0.049)	(0.119)	(0.057)
p-value TaRL − No-TaRL	0.779	0.776	0.923	0.343	0.290	0.769	0.550
Control mean	0.128	0.028	0.337	0.470	0.096	-0.000	0.619
N	1,269	1,269	1,269	1,269	531	531	448

Notes: All panels report treatment effects from double post-lasso regressions with baseline covariates, missingness indicators and strata fixed effects, clustering at school level. Columns (5)–(7) additionally control for whether teachers responded to the endline survey online. Panel A reports the main treatment effect. Panel B reports effects by authority cross-randomization arm: peers message (T1) and authority message (T2), each relative to control; the p-value row reports the p-value of the T2 − T1 difference from a `lincom` test. Panel C reports effects by TaRL cross-randomization arm: non-TaRL (T1) and TaRL (T2), each relative to control; the p-value row reports the p-value of the T2 − T1 difference. * p<0.10, ** p<0.05, *** p<0.01.

TABLE 3: Time Allocation

	Lesson Plan	Diff. Materials	Design Evaluation	Grading	Collab Work	Admin Work	Comm w/ Parents
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Reported last time activity was performed, endline							
Treatment	-0.004	-0.069**	-0.024	-0.017	-0.023	-0.046	-0.018
	(0.032)	(0.034)	(0.034)	(0.038)	(0.036)	(0.041)	(0.033)
Control mean	0.853	0.829	0.843	0.794	0.797	0.675	0.822
N	607	607	607	607	607	607	607
B. Time it took to perform last time they did, in minutes							
Treatment	1.653	-4.385	-1.737	0.541	4.738	-7.668	2.914
	(6.237)	(5.211)	(5.371)	(7.593)	(7.166)	(6.641)	(4.466)
Control mean	71.182	60.808	49.524	76.801	72.944	60.052	28.962
N	607	607	607	607	607	607	607

Notes: Both panels report effects from double post-lasso regressions with baseline covariates, missingness indicators, strata fixed effects and whether teachers responded to endline online, clustering at school level. Panel A reports effects for whether teachers who responded to endline reported a time in minutes for the activity. Panel B reports treatment effects for time reported, considering zero for teachers who did not report a time; * p<0.10, ** p<0.05, *** p<0.01.

TABLE 4: Prompt Quality

	Overall	Measure 1	Measure 2	Measure 3	Measure 4
	(1)	(2)	(3)	(4)	(5)
A. Training-based rubric					
Treatment	0.293***	0.248***	0.253***	0.358***	0.314***
	(0.071)	(0.073)	(0.063)	(0.078)	(0.083)
Control mean	3.197	3.155	3.596	2.955	3.083
N	562	562	562	562	562
B. General rubric					
Treatment	0.375***	0.179**	0.331***	0.502***	0.486***
	(0.086)	(0.077)	(0.082)	(0.102)	(0.102)
Control mean	3.374	3.891	3.528	2.970	3.106
N	562	562	562	562	562

Notes: Both panels report effects from double post-lasso regressions with baseline covariates, missingness indicators, strata fixed effects and whether teachers responded to endline online, clustering at school level. Panel A reports effects for a prompt rubric based on the program's training materials, where the measures are: (1) how well teacher specifies a who, (2) how well teacher specifies a what, (3) how well teacher specifies a how, and (4) how well teacher connects to peruvian national curriculum components. Panel B reports effects for a general rubric, not designed based-off the training materials, where the measures are: (1) clarity, (2) specificity, (3) context, and (4) well-specified components; * p<0.10, ** p<0.05, *** p<0.01.

TABLE 5: Prompt Scenarios

	Uses AI for Lesson Plan	Incorrect Prompt Assessment	Considers Good AI Response	Considers Bad AI Response Content	Considers Bad AI Response Format
	(1)	(2)	(3)	(4)	(5)
Treatment	0.094**	-0.099**	-0.067***	0.059	0.007
	(0.039)	(0.047)	(0.025)	(0.044)	(0.046)
Control mean	0.687	0.674	0.136	0.436	0.428
N	562	560	560	560	560

Notes: Table reports effects from double post-lasso regressions with baseline covariates, missingness indicators, strata fixed effects and whether teachers responded to endline online, clustering at school level; * p<0.10, ** p<0.05, *** p<0.01

TABLE A1: Balance: Authority Cross-Randomization

	Control Mean (1)	T1: Peer Mean (2)	T2: Authority Mean (3)	C vs T1 (4)	C vs T2 (5)	T1 vs T2 (6)
A. Teacher Characteristics						
Age	52.227	52.476	52.174	0.193	-0.175	-0.358
	[0.407]	[0.507]	[0.519]	(0.675)	(0.662)	(0.705)
Gender	0.802	0.802	0.781	0.000	-0.025	-0.022
	[0.017]	[0.023]	[0.024]	(0.026)	(0.029)	(0.032)
Has university degree	0.785	0.730	0.762	-0.052	-0.024	0.046
	[0.019]	[0.027]	[0.026]	(0.032)	(0.033)	(0.037)
Teaching experience (years)	24.115	24.661	23.996	0.453	-0.267	-0.781
	[0.490]	[0.590]	[0.591]	(0.822)	(0.775)	(0.855)
Teach. exp. current school (years)	12.930	13.438	12.336	0.352	-0.481	-0.963
	[0.554]	[0.708]	[0.727]	(0.930)	(0.967)	(1.003)
6th grade teach. exp. (years)	6.198	6.263	6.532	-0.059	0.396	0.326
	[0.321]	[0.390]	[0.458]	(0.509)	(0.542)	(0.565)
No additional employment	0.771	0.741	0.770	-0.029	0.001	0.017
	[0.019]	[0.027]	[0.026]	(0.035)	(0.033)	(0.040)
Responded to endline	0.495	0.451	0.473	-0.033	-0.016	0.016
	[0.020]	[0.028]	[0.028]	(0.040)	(0.040)	(0.046)
B. School Characteristics						
Total students	90.804	93.978	87.816	2.260	-3.217	-5.742
	[3.149]	[4.920]	[4.245]	(5.583)	(5.298)	(6.547)
Total classes	3.070	3.194	3.020	0.111	-0.056	-0.176
	[0.097]	[0.157]	[0.132]	(0.180)	(0.166)	(0.207)
Network size	5.598	5.645	5.459	0.031	-0.158	-0.172
	[0.150]	[0.223]	[0.218]	(0.270)	(0.260)	(0.314)
Mean math score	0.121	0.139	0.140	0.012	0.026	0.008
	[0.031]	[0.043]	[0.050]	(0.052)	(0.056)	(0.065)
Mean language score	0.135	0.111	0.110	-0.027	-0.023	0.008
	[0.029]	[0.045]	[0.060]	(0.052)	(0.062)	(0.072)
Share of tech devices	0.615	0.642	0.606	0.030	-0.008	-0.036
	[0.016]	[0.021]	[0.023]	(0.026)	(0.027)	(0.030)

Notes: Columns (4)–(6) report coefficients from regressions of each variable on group indicators with strata fixed effects, clustering at school level. T1 refers to the peer message treatment and T2 refers to the authority message treatment.

TABLE A2: Balance: TaRL Cross-Randomization

	Control Mean (1)	T1: Non-TaRL Mean (2)	T2: TaRL Mean (3)	C vs T1 (4)	C vs T2 (5)	T1 vs T2 (6)
A. Teacher Characteristics						
Age	52.227	52.182	52.461	-0.205	0.288	0.367
	[0.407]	[0.515]	[0.510]	(0.667)	(0.657)	(0.694)
Gender	0.802	0.769	0.814	-0.036	0.010	0.047
	[0.017]	[0.024]	[0.022]	(0.028)	(0.028)	(0.031)
Has university degree	0.785	0.741	0.750	-0.044	-0.028	0.005
	[0.019]	[0.027]	[0.026]	(0.033)	(0.033)	(0.037)
Teaching experience (years)	24.115	24.205	24.454	-0.037	0.274	0.293
	[0.490]	[0.598]	[0.584]	(0.765)	(0.802)	(0.844)
Teach. exp. current school (years)	12.930	13.131	12.679	0.221	-0.349	-0.551
	[0.554]	[0.736]	[0.702]	(0.938)	(0.952)	(0.985)
6th grade teach. exp. (years)	6.198	6.201	6.575	-0.117	0.420	0.516
	[0.321]	[0.402]	[0.442]	(0.485)	(0.551)	(0.538)
No additional employment	0.771	0.757	0.754	-0.012	-0.015	0.004
	[0.019]	[0.027]	[0.026]	(0.035)	(0.034)	(0.039)
Responded to endline	0.495	0.440	0.483	-0.047	-0.007	0.047
	[0.020]	[0.028]	[0.028]	(0.042)	(0.039)	(0.044)
B. School Characteristics						
Total students	90.804	88.404	93.155	-2.795	1.845	5.075
	[3.149]	[4.460]	[4.695]	(5.466)	(5.590)	(6.702)
Total classes	3.070	2.989	3.216	-0.087	0.142	0.242
	[0.097]	[0.135]	[0.152]	(0.170)	(0.180)	(0.212)
Network size	5.598	5.383	5.711	-0.229	0.081	0.315
	[0.150]	[0.225]	[0.215]	(0.269)	(0.256)	(0.317)
Mean math score	0.121	0.121	0.157	-0.005	0.039	0.052
	[0.031]	[0.048]	[0.047]	(0.054)	(0.055)	(0.065)
Mean language score	0.135	0.096	0.125	-0.044	-0.008	0.047
	[0.029]	[0.044]	[0.062]	(0.050)	(0.065)	(0.073)
Share of tech devices	0.615	0.616	0.631	0.003	0.018	0.016
	[0.016]	[0.023]	[0.022]	(0.028)	(0.027)	(0.030)

Notes: Columns (4)–(6) report coefficients from regressions of each variable on group indicators with strata fixed effects, clustering at school level. T1 refers to the no-TaRL treatment arm and T2 refers to the TaRL treatment arm.

TABLE A3: Time Allocation by Cross-Randomized Treatment Arms

	Lesson Plan (1)	Diff. Materials (2)	Design Evaluation (3)	Grading (4)	Collab Work (5)	Admin Work (6)	Comm w/ Parents (7)
IA. Reported last time activity was performed – Authority cross-randomization							
Peers effect	-0.020	-0.089**	-0.040	-0.015	-0.017	-0.029	-0.032
	(0.041)	(0.044)	(0.043)	(0.047)	(0.045)	(0.052)	(0.042)
Authority effect	-0.001	-0.063	-0.003	-0.018	-0.043	-0.086	-0.015
	(0.034)	(0.039)	(0.039)	(0.044)	(0.045)	(0.053)	(0.038)
Auth − Peers (p-value)	0.622	0.562	0.382	0.945	0.602	0.326	0.687
Control mean	0.853	0.829	0.843	0.794	0.797	0.675	0.822
N	607	607	607	607	607	607	607
IB. Reported last time activity was performed – TaRL cross-randomization							
Non-TaRL effect	-0.040	-0.081*	-0.040	-0.057	-0.029	-0.046	-0.024
	(0.038)	(0.043)	(0.043)	(0.047)	(0.047)	(0.054)	(0.042)
TaRL effect	0.015	-0.071*	-0.005	0.019	-0.032	-0.068	-0.024
	(0.037)	(0.041)	(0.041)	(0.043)	(0.044)	(0.052)	(0.039)
TaRL − Non-TaRL (p-value)	0.153	0.827	0.439	0.126	0.960	0.707	0.998
Control mean	0.853	0.829	0.843	0.794	0.797	0.675	0.822
N	607	607	607	607	607	607	607
IIA. Time it took to perform last time they did, in minutes – Authority cross-randomization							
Peers effect	-3.265	-6.315	-6.200	0.329	6.002	-3.122	7.366
	(7.524)	(7.024)	(7.607)	(9.665)	(9.082)	(8.962)	(6.979)
Authority effect	2.890	-8.083	-1.346	-0.089	3.718	-11.143	-0.256
	(8.140)	(6.786)	(7.382)	(10.122)	(9.710)	(9.096)	(5.201)
Auth − Peers (p-value)	0.463	0.800	0.504	0.967	0.836	0.373	0.289
Control mean	71.182	60.808	49.524	76.801	72.944	60.052	28.962
N	607	607	607	607	607	607	607
IIB. Time it took to perform last time they did, in minutes – TaRL cross-randomization							
Non-TaRL effect	3.699	-6.027	-4.570	-5.179	8.459	-3.565	2.633
	(9.067)	(7.392)	(7.625)	(10.386)	(10.157)	(9.618)	(5.358)
TaRL effect	-3.464	-8.235	-3.019	4.693	1.720	-10.322	4.249
	(7.139)	(6.516)	(7.499)	(9.426)	(9.145)	(8.701)	(6.711)
TaRL − Non-TaRL (p-value)	0.441	0.757	0.837	0.332	0.566	0.470	0.819
Control mean	71.182	60.808	49.524	76.801	72.944	60.052	28.962
N	607	607	607	607	607	607	607

Notes: Both panels report treatment effects from double post-lasso regressions with baseline covariates, missingness indicators, strata fixed effects and whether teachers responded to endline online, clustering at school level. Panel A reports effects by authority cross-randomization arm: peers message (T1) and authority message (T2), each relative to control; the third row reports the difference between T2 and T1. Panel B reports effects by TaRL cross-randomization arm: non-TaRL (T1) and TaRL (T2), each relative to control; the third row reports the difference. * p<0.10, ** p<0.05, *** p<0.01.

TABLE A4: Prompt Quality by Cross-Randomized Treatment Arms

	Overall	Measure 1	Measure 2	Measure 3	Measure 4
	(1)	(2)	(3)	(4)	(5)
A. Training-based rubric – Authority cross-randomization					
Peers effect	0.338***	0.292***	0.308***	0.387***	0.363***
	(0.082)	(0.086)	(0.074)	(0.092)	(0.097)
Authority effect	0.313***	0.291***	0.222***	0.392***	0.349***
	(0.084)	(0.086)	(0.072)	(0.095)	(0.104)
Auth − Peers (p-value)	0.767	0.993	0.211	0.958	0.889
Control mean	3.197	3.155	3.596	2.955	3.083
N	562	562	562	562	562
A. Training-based rubric – TaRL cross-randomization					
No-TaRL effect	0.363***	0.313***	0.296***	0.452***	0.392***
	(0.087)	(0.090)	(0.072)	(0.098)	(0.107)
TaRL effect	0.292***	0.273***	0.236***	0.336***	0.324***
	(0.081)	(0.082)	(0.074)	(0.090)	(0.096)
TaRL − No-TaRL (p-value)	0.389	0.621	0.380	0.238	0.521
Control mean	3.197	3.155	3.596	2.955	3.083
N	562	562	562	562	562
B. General rubric – Authority cross-randomization					
Peers effect	0.463***	0.234**	0.418***	0.578***	0.624***
	(0.101)	(0.093)	(0.097)	(0.122)	(0.122)
Authority effect	0.436***	0.197**	0.374***	0.605***	0.569***
	(0.106)	(0.087)	(0.102)	(0.124)	(0.129)
Auth − Peers (p-value)	0.793	0.690	0.661	0.827	0.668
Control mean	3.374	3.891	3.528	2.970	3.106
N	562	562	562	562	562
B. General rubric – TaRL cross-randomization					
Non-TaRL effect	0.487***	0.250***	0.434***	0.597***	0.669***
	(0.111)	(0.096)	(0.105)	(0.130)	(0.136)
TaRL effect	0.416***	0.184**	0.362***	0.587***	0.532***
	(0.098)	(0.085)	(0.095)	(0.119)	(0.119)
TaRL − Non-TaRL (p-value)	0.507	0.471	0.487	0.941	0.302
Control mean	3.374	3.891	3.528	2.970	3.106
N	562	562	562	562	562

Notes: Both panels report treatment effects from double post-lasso regressions with baseline covariates, missingness indicators, strata fixed effects and whether teachers responded to endline online, clustering at school level. Panel A reports effects by authority cross-randomization arm: peers message (T1) and authority message (T2), each relative to control; the third row reports the difference between T2 and T1. Panel B reports effects by TaRL cross-randomization arm: non-TaRL (T1) and TaRL (T2), each relative to control; the third row reports the difference. * p<0.10, ** p<0.05, *** p<0.01.

TABLE A5: Prompt Scenarios by Cross-Randomized Treatment Arms

	Uses AI for Lesson Plan (1)	Incorrect Prompt Assessment (2)	Considers Good AI Response (3)	Considers Bad AI Response Content (4)	Considers Bad AI Response Format (5)
A. Authority cross-randomization					
Peers effect	0.111**	-0.145**	-0.069**	0.105*	-0.036
	(0.052)	(0.059)	(0.032)	(0.060)	(0.056)
Authority effect	0.100**	-0.063	-0.115***	0.044	0.071
	(0.049)	(0.057)	(0.028)	(0.053)	(0.056)
Auth − Peers (p-value)	0.840	0.185	0.147	0.325	0.076
Control mean	0.687	0.674	0.136	0.436	0.428
N	562	560	560	560	560
B. TaRL cross-randomization					
Non-TaRL effect	0.059	-0.137**	-0.095***	0.120**	-0.025
	(0.055)	(0.058)	(0.032)	(0.056)	(0.057)
TaRL effect	0.145***	-0.073	-0.091***	0.033	0.058
	(0.045)	(0.058)	(0.029)	(0.056)	(0.056)
TaRL − Non-TaRL (p-value)	0.107	0.308	0.914	0.155	0.177
Control mean	0.687	0.674	0.136	0.436	0.428
N	562	560	560	560	560

Notes: Both panels report treatment effects from double post-lasso regressions with baseline covariates, missingness indicators, strata fixed effects and whether teachers responded to endline online, clustering at school level. Panel A reports effects by authority cross-randomization arm: peers message (T1) and authority message (T2), each relative to control; the third row reports the difference between T2 and T1. Panel B reports effects by TaRL cross-randomization arm: non-TaRL (T1) and TaRL (T2), each relative to control; the third row reports the difference. * p<0.10, ** p<0.05, *** p<0.01.

TABLE A6: Equity

	Benefits Particular Group	Benefits Everyone	Reduce Disparities	Would not Pay	Would Pay Market Price Or More	Price Bracket
	(1)	(2)	(3)	(4)	(5)	(6)
A. Main treatment effect						
Treatment	0.074*	-0.084*	0.050	-0.036	-0.078***	-5.558***
	(0.039)	(0.046)	(0.040)	(0.045)	(0.022)	(1.862)
Control mean	0.654	0.535	0.685	0.346	0.112	18.788
N	607	554	554	554	554	554
B. Authority cross-randomization						
Peers effect	0.106**	-0.138**	0.041	-0.047	-0.085***	-7.082***
	(0.046)	(0.056)	(0.053)	(0.056)	(0.023)	(2.100)
Authority effect	0.054	-0.046	0.082*	-0.062	-0.033	-2.569
	(0.048)	(0.055)	(0.049)	(0.052)	(0.028)	(2.262)
Auth − Peers (p-value)	0.282	0.130	0.442	0.810	0.051	0.044
Control mean	0.654	0.535	0.685	0.346	0.112	18.788
N	607	554	554	554	554	554
C. TaRL cross-randomization						
Non-TaRL effect	0.074	-0.079	0.115**	-0.053	-0.050*	-3.836*
	(0.049)	(0.056)	(0.051)	(0.056)	(0.028)	(2.324)
TaRL effect	0.084*	-0.100*	0.017	-0.056	-0.064***	-5.531***
	(0.046)	(0.057)	(0.050)	(0.053)	(0.025)	(2.129)
TaRL − Non-TaRL (p-value)	0.847	0.741	0.068	0.961	0.636	0.478
Control mean	0.654	0.535	0.685	0.346	0.112	18.788
N	607	554	554	554	554	554

Notes: All panels report treatment effects from double post-lasso regressions with baseline covariates, missingness indicators, strata fixed effects and whether teachers responded to endline online, clustering at school level. Panel A reports the main treatment effect. Panel B reports effects by authority cross-randomization arm: peers message (T1) and authority message (T2), each relative to control; the p-value row reports the p-value of the T2 − T1 difference from a `lincom` test. Panel C reports effects by TaRL cross-randomization arm: non-TaRL (T1) and TaRL (T2), each relative to control; the p-value row reports the p-value of the T2 − T1 difference. * p<0.10, ** p<0.05, *** p<0.01.

TABLE A7: Prompt Quality: Heterogeneity by Age

	Overall	Measure 1	Measure 2	Measure 3	Measure 4
	(1)	(2)	(3)	(4)	(5)
A. Training-based rubric					
Treatment (age $\leq$ median)	0.192**	0.138	0.135*	0.292***	0.208**
	(0.086)	(0.092)	(0.074)	(0.098)	(0.104)
Age above median	-0.452***	-0.500***	-0.414***	-0.386***	-0.493***
	(0.113)	(0.120)	(0.111)	(0.118)	(0.132)
Treatment $\times$ Age above median	0.341**	0.412***	0.283**	0.259*	0.391**
	(0.144)	(0.148)	(0.141)	(0.155)	(0.171)
Treatment (age $>$ median)	0.533	0.550	0.419	0.551	0.599
	(0.126)	(0.128)	(0.127)	(0.133)	(0.146)
Control mean (age $\leq$ median)	3.366	3.321	3.771	3.107	3.264
Control mean (age $>$ median)	2.929	2.838	3.365	2.730	2.784
N	562	562	562	562	562
B. General rubric					
Treatment (age $\leq$ median)	0.249**	0.090	0.164*	0.400***	0.348***
	(0.101)	(0.087)	(0.098)	(0.126)	(0.122)
Age above median	-0.514***	-0.381***	-0.477***	-0.576***	-0.603***
	(0.150)	(0.141)	(0.152)	(0.175)	(0.169)
Treatment $\times$ Age above median	0.405**	0.247	0.453**	0.404*	0.492**
	(0.181)	(0.168)	(0.180)	(0.212)	(0.212)
Treatment (age $>$ median)	0.655	0.337	0.617	0.804	0.840
	(0.158)	(0.147)	(0.155)	(0.183)	(0.187)
Control mean (age $\leq$ median)	3.568	4.050	3.729	3.179	3.314
Control mean (age $>$ median)	3.081	3.689	3.257	2.635	2.743
N	562	562	562	562	562

Notes: Both panels report treatment effects from double post-lasso regressions with baseline covariates, missingness indicators, strata fixed effects and whether teachers responded to endline online, clustering at school level. Age above median is a binary indicator equal to 1 if the teacher's age is above the sample median. reports the main treatment effect for teachers at or below the median age. reports the sum of the treatment effect and the interaction term (i.e. the treatment effect for teachers above the median age), estimated via `lincom`. Panel A uses the training-based prompt rubric; Panel B uses the general rubric. * p<0.10, ** p<0.05, *** p<0.01.