Public Disclosure Authorized

Public Disclosure Authorized

Policy Research Working Paper 11125

From Chalkboards to Chatbots

Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria

Martín De Simone
Federico Tiberti
Maria Barron Rodriguez
Federico Manolio
Wuraola Mosuro
Eliot Jolomi Dikoru

WORLD BANK GROUP
Education Global Department
May 2025

Reproducible Research Repository

A verified reproducibility package for this paper is available at http://reproducibility.worldbank.org, click **here** for direct access.

Policy Research Working Paper 11125

Abstract

This study evaluates the impact of a program leveraging large language models for virtual tutoring in secondary education in Nigeria. Using a randomized controlled trial, the program deployed Microsoft Copilot (powered by GPT-4) to support first-year senior secondary students in English language learning over six weeks. The intervention demonstrated a significant improvement of 0.31 standard deviation on an assessment that included English topics aligned with the Nigerian curriculum, knowledge of artificial intelligence and digital skills. The effect on English, the main outcome of interest, was of 0.23 standard deviations. Cost-effectiveness analysis revealed substantial learning gains, equating to 1.5 to 2 years of 'business-as-usual' schooling, situating the intervention among some of the most cost-effective programs to improve learning outcomes. An analysis of heterogeneous effects shows that while the program benefits students across the baseline ability distribution, the largest effects are for female students, and those with higher initial academic performance. The findings highlight that artificial intelligence-powered tutoring, when designed and used properly, can have transformative impacts in the education sector in low-resource settings.

This paper is a product of the Education Global Department. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at desimone@worldbank.org, ftiberti@worldbank.org, mbarronrodriguez@worldbank.org, fmanolio@worldbank.org, wmosuro@worldbank.org, and edikoru@worldbank.org. A verified reproducibility package for this paper is available at http://reproducibility.worldbank.org, click **here** for direct access.



The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Produced by the Research Support Team

From Chalkboards to Chatbots: Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria*

Martín De Simone, Federico Tiberti, Maria Barron Rodriguez, Federico Manolio, Wuraola Mosuro, Eliot Jolomi Dikoru†

Originally published in the Policy Research Working Paper Series in *May 2025*. This version is updated in *December 2025*.
To obtain the originally published version, please email prwp@worldbank.org.

Keywords: large-language models, adaptive learning, artificial intelligence, education technology, secondary education, teaching at the right level.
JEL Classification: C93, I21, J24, O15, O33.

*The team would like to thank Scherezad Latif and Halil Dundar, Education Practice Managers, World Bank. The team extends its appreciation to Dr. Joan Osa Oviawe and Jennifer Aisuan, for their collaboration throughout the implementation of the pilot, as well as Alex Twinomugisha, Robert Hawkins, and Cristóbal Cobo for their support with the intervention. The team thanks those who provided comments to a previous version of this paper, including David Evans, Halsey Rogers, Carolina Lopez, Francisco Haimovich, Daniel Rodriguez-Segura, Noah Yarrow, Juan Barón, and Lucas Gortazar. The team acknowledges the financial support received from the Mastercard Foundation.

†De Simone: The World Bank. E-mail: mdesimone@worldbank.org. Tiberti: The World Bank. E-mail: ftiberti@worldbank.org. Barron: The World Bank. E-mail: mbarronrodriguez@worldbank.org. Manolio: The World Bank. E-mail: fmanolio@worldbank.org. Mosuro: The World Bank. E-mail: wmosuro@worldbank.org. Dikoru: The World Bank. E-mail: edikoru@worldbank.org.

1 Introduction

The global education sector is grappling with a learning crisis. According to the Learning Poverty Index, approximately 70% of 10-year-olds in low- and middle-income countries cannot read and understand an age-appropriate text (World Bank, 2022). These deficits in learning accumulate and become particularly acute at the secondary school level, as evidenced by numerous international, regional, and national assessments.

In his seminal 1984 study, Bloom demonstrated that students receiving one-on-one tutoring outperformed their peers in traditional classroom settings by an average of two standard deviations (Bloom, 1984). Subsequent studies have consistently confirmed the significant benefits of one-on-one tutoring (Nickow et al., 2020). The challenge, however, is that implementing one-on-one tutoring at scale is costly and unaffordable for most education systems. Bloom referred to this challenge as the "two-sigma problem": how to replicate the gains of personalized tutoring at scale in a cost-effective manner.

This paper examines whether generative artificial intelligence, specifically large language models (LLMs), can help solve that problem. We evaluate a six-week after-school tutoring program in Nigeria that used a publicly available LLM (ChatGPT-4) to support students in learning English. First-year secondary students from nine public schools in Benin City were invited to participate; from this pool, 52% of eligible students expressed interest, and participants were randomly selected from among them. Those assigned to the intervention attended twelve 90-minute sessions in computer labs, engaging in curriculum-aligned activities guided by teachers. We use a randomized controlled trial (RCT) design to estimate the causal impact of the program on learning outcomes.

We present three main sets of results. First, we show that students selected to participate in the program score 0.31 standard deviation higher in the final assessment that was delivered at the end of the intervention. We find strong statistically-significant intent-to-treat (ITT) effects on all sections of that assessment: English skills (which included the majority of questions, 0.24 σ), digital skills (0.14 σ), AI skills (0.31 σ) and an Item Response Theory (IRT) composite score of each student's exam (0.26 σ). We also show that the intervention yielded strong positive results on the regular English curricular exam of the third term. This result is important because the content evaluated in that exam was broader than the one covered during the six weeks of the intervention and included the content of the entire year. We calculate an ITT effect of being selected for the program on the performance in the third-term exam of 0.21 standard deviations.

Second, we examine heterogeneity of the effects by certain pre-treatment characteristics.

Treatment effects were positive and statistically significant across all levels of baseline performance, but stronger among students with better prior performance. Similarly, treatment effects were positive and statistically significant over the entire distribution of a proxy for socioeconomic status, but stronger among students with a higher one. Lastly, treatment effects were stronger among female students, compensating for a deficit in their baseline performance.

Third, we conduct dose-response analysis. We estimate Local Average Treatment Effect (LATE) estimates, focusing on the impact of actual attendance to the intervention sessions, which averaged 72% among the treatment group. Using attendance data, we estimate a dose-response relationship, finding a strong linear association between days attended and improved learning outcomes, with an effect size of approximately 0.031 standard deviation per additional day of attendance. Further analysis predicts substantial gains with extended program duration, estimating an increase of between 1.2 and 2.2 standard deviations for a full academic year of participation, depending on attendance rates.

The findings, combined with a cost analysis, seem to indicate that the program was highly cost-effective. The six-week pilot generated learning gains that take between 1.5 and 2 years in a business-as-usual scenario. The program achieved 3.2 equivalent years of schooling (EYOS) per $100 invested, surpassing many comparable interventions. Using Learning-adjusted years of schooling (LAYS) as the metric for the analysis, the program generates up to 0.9 year of high-performance education. When benchmarked against evidence from both low- and middle-income countries, the pilot program ranks among the most cost-effective solutions for addressing learning crises.

Our study contributes to different strands of the literature that aim to identify the effect of programs that try to customize instruction to the level of students, both with and without technology. Efforts to address this challenge have included the development of the "Teaching at the Right Level" (TaRL) approach, which has shown to improve learning outcomes in contexts such as India, Kenya, Ghana, and Zambia (Banerjee et al., 2016). Implementation modalities of TaRL have varied, ranging from pulling students out of class (Banerjee et al., 2007), tracking classrooms (Duflo et al., 2011), providing extra instructional time outside of school (Banerjee et al., 2016), and employing volunteers instead of teachers (Banerjee et al., 2008). However, scaling TaRL programs remains difficult due to their labor-intensive nature. This challenge is particularly pronounced given the global shortage of teachers, which is particularly pronounced in Sub-Saharan Africa. Recent estimates suggest that by 2040, countries in the region will need 21% more secondary school teachers per year (Evans and Mendez Acosta, forthcoming). Teacher shortages are

further compounded by high attrition rates, and the need for specialized knowledge at the secondary level makes TaRL programs even more difficult to implement.

In recent years, adaptive learning software has emerged as a potential solution to the scalability of tutoring programs by using technology to mimic one-on-one tutoring. Evidence suggests that computer-adaptive learning systems can improve learning outcomes. For example, a study of personalized, technology-aided after-school instruction for middle school students in India reported gains of 0.37 standard deviation in math and 0.23 standard deviation in Hindi over a 4.5-month period (Muralidharan et al., 2019). A study in Cambodia that focused on math instruction for primary school students found impacts on cognitive skills due to students' increased learning productivity per hour (Ito et al., 2021). In El Salvador, the use of software for adaptive learning proved effective in an environment with heterogeneous classes and poorly qualified teachers (Büchel et al., 2022). Experiments in China have also found positive effects on standardized math scores (Lai et al., 2015a) and on Mandarin (Lai et al., 2015b), including when implemented during regular school hours (Mo et al., 2014). In Ecuador, the possibility to use an adaptive-learning software for 4 months led to large positive impact on standardized test scores in math (Angel-Urdinola et al., 2023). Other studies that do not follow experimental approaches have also estimated positive effects of similar software programs, such as a program in Uruguay that showed gains of 0.2 standard deviation on mathematics test scores (Perera and Aboal, 2019).

Despite these successes, adaptive learning programs face several challenges. First, most are not deployed in the world's most challenging educational contexts, particularly in Sub-Saharan Africa, raising questions about external validity. Second, these programs often rely on proprietary software, which typically involves both fixed and per-student costs, making them difficult to scale in resource-constrained environments.

Some adaptive-learning options are developed using artificial intelligence (AI) to adjust to the level of the students, but they primarily rely on pattern recognition and predictive algorithms, to provide students with exercises adjusted to their level based on a pool of thousands of items. The recent advances in generative artificial intelligence offer a promising avenue to use software to teach students while maintaining a more human-like interaction with students through the use of natural language.

Most of the studies that have examined generative AI in education have been conducted in developed countries and lab settings, assessing the short-term effects of brief interactions (Kumar et al., 2023). In Italy, studies have found positive effects of Large Language Models (LLMs) on learning outcomes through homework support (Vanzo et al., 2024). In

the United States, a human-AI approach with expert guidance through language models supports tutors instead of providing direct help to students, and found that students working on mathematics with tutors randomly assigned to have access to a tutor co-pilot are 4 percentage points more likely to master topics (Wang et al., 2024). A study carried out among undergraduate students at Harvard University showed that those who benefited from an AI-powered tutor at home performed better than those exposed only to active learning classes (Kestin et al., 2024).

Only a few studies evaluate the effect of generative AI to support students through tutoring. In Ghana, students who were given access to a phone for one hour a week and were allowed to use an AI-powered math tutor via a messaging app to independently study math improved their scores much more than those without access, with an effect size of 0.36 (Henkel et al., 2024). A recent study in Türkiye of an intervention that included only four sessions showed that while LLMs can improve mathematics learning outcomes, they can also be detrimental to learning in the long term if they are used as "crutches" rather than as tutors (Bastani et al., 2024). A similar effect was found for coding classes in a lab setting (Lehmann et al., 2024). This study showed more positive impacts with the LLM used with prompts to safeguard learning (Bastani et al., 2024).

Thus, this paper contributes to this recent literature by examining the impact of one of the first programs to leverage LLMs for educational purposes in a developing country context using a real experimental design in Sub-Saharan Africa. It also aims to address some of the challenges identified in recent reviews of emerging studies on the effect of LLMs on learning: the lack of objective measures to complement subjective assessments of impact, weaknesses in the definition of the control and treatment groups (Weidlich et al., 2025), and the lack of power analysis to determine adequate sample sizes (Deng et al., 2024). Furthermore, the intervention used a free, off-the-shelf model, requiring minimal customization and no pre-built question banks, which might facilitate its scalability.

The findings of this intervention underscore several critical policy implications for addressing the learning crisis in developing countries, particularly in Sub-Saharan Africa. The program demonstrated significant impacts on learning outcomes, even amid challenges such as internet disruptions and power outages, highlighting its potential in contexts with severe teacher shortages and resource constraints. AI-powered tutoring programs using LLMs can complement traditional teaching by enhancing teacher productivity and delivering personalized learning experiences, particularly when paired with guided prompts, teacher oversight, and alignment with the curriculum. The intervention's cost-effectiveness and scalability are promising, leveraging local staff and free tools

to minimize costs while eliminating the need for extensive question banks required by traditional adaptive software. However, policymakers must address potential inequities arising from disparities in digital literacy and access to technology. Investments in infrastructure, teacher training, and inclusive digital education are essential to ensure equitable access and mitigate the risk of exacerbating inequalities. Given the nascent application of LLMs in education, numerous questions remain unanswered, underscoring the importance of replicating this study, including with small variations.

The rest of this paper is organized as follows. Section 2 describes the intervention and the experimental design, including the data used. Section 3 presents our main results, including a heterogeneity and dosage analysis, as well as a robustness analysis. Section 4 discusses cost effectiveness, proposes future research directions, and presents policy implications.

2 Intervention and Study Design

2.1 The Intervention

The study analyzes the effects of an after-school program in which students interacted with a large language model twice per week to improve their English skills, following the national curriculum. The intervention was implemented in Benin City, Nigeria, using Copilot, an LLM powered by the GPT-4 model at the time of implementation.[1] The program was implemented over a six-week period between June and July 2024, targeting first-year senior secondary school students, who are typically 15 years old.[2] The intervention aimed to improve learning outcomes in English language classes using an AI chatbot as a virtual tutor. The selected tool was Microsoft Copilot, powered by ChatGPT-4, which was freely available and required only student registration. The program was conducted in nine schools and the students were grouped according to the number of computers in each school lab, with an average of 30 students per session. Each student was allowed to participate in a maximum of two 1.5-hour after-school sessions per week.

The selection of schools was based on the availability of computer labs. These labs varied in the types of devices they used, ranging from laptops to desktop computers. Internet access, essential for real-time interaction with the LLM, was provided through routers

[1] GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10 percent of test takers (Achiam et al., 2023).

[2] A detailed implementation timeline can be found in Table 14.

and mobile telephone signals. However, internet disruptions and power outages were common challenges faced during the intervention. Despite these issues, students were able to interact with the chatbot for the majority of the sessions.

All students' guardians signed consent forms, agreeing to their children's participation in the pilot program. Students worked in pairs, sharing a computer, and engaged in dialogue with the AI tool to enhance their learning. Teachers, who played a critical role in guiding the students but did not provide direct instruction, participated in a single three-day training program with one cohort. This training introduced teachers to the functionalities of the LLM and equipped them with pedagogical techniques to ensure their responsible use and supervise students during the sessions. It also made them aware of potential risks, such as hallucinations and biases, that the LLM could have.

In the first session, teachers familiarized students with Microsoft Copilot, emphasizing both its educational benefits and potential risks, such as over-reliance on the model and the possibility of hallucinations and biased outputs. The goal was to foster responsible usage, encouraging students to complement their learning with the AI tool while retaining critical thinking skills.

Each subsequent session focused on a topic from the first-year English language curriculum, aligned with the material that students covered during their regular classes. The sessions began with a teacher-provided prompt, followed by free interaction between the student pairs and the AI tool. Teachers circulated the classroom, ensuring students' interactions remained relevant and on task. Each teacher was provided with a three-part implementation toolkit which included: a) curated online learning resources on the use of Copilot and LLMs; b) a handbook focused on AI literacy and potential risks and benefits; and c) session guidelines, including suggested initial prompts and potential follow-up questions to assist students if needed. Teachers were also provided with contacts in case they faced any problems with the program implementation, and a group-chat was created to streamline communications. The students also had a customized guide, which included the initial prompts.

The lesson guides and their prompts were carefully crafted to position the LLM as a tutor, focusing on facilitating learning rather than simply providing direct answers. These prompts were informed by principles from the science of learning and were tailored to the cultural context of southern Nigeria, incorporating familiar names and customs to resonate with students.[3] Some of the prompt structures were derived from Mollick and

[3]One of the strategies employed to enhance learning through prompting was to encourage the LLM to leverage "desirable difficulties" rather than simply providing direct answers. These are conditions that,

Mollick (2023a). This design aimed to encourage the LLM to adapt to each student's individual learning level, providing pedagogical support through contextually relevant examples and diverse teaching techniques. Students interacted with the LLM by asking questions, completing exercises, and receiving personalized feedback. At the end of each session, the students were encouraged to reflect and discuss lessons learned and challenges encountered during session to facilitate knowledge sharing among the group.

To ensure the fidelity of program implementation, monitors were first trained, provided with monitoring guidelines, and then assigned to track student attendance and gather information about each session using Kobo Toolbox.[4] This system allowed for real-time data collection, ensuring that the intervention was carried out as intended in each school and offered the opportunity to respond promptly to any challenges.[5]

2.2 Sample and Randomization

The randomization for the pilot program was conducted at the student level in the nine selected schools. All first-year senior secondary school students in these schools were informed about the program through information sessions and given a window of ten days to express their interest in participating. Only students who voluntarily expressed interest within this period were included in the randomization pool.

To assess whether students who expressed interest in the after-school program differed systematically from those who did not, we compare pre-program exam scores between students who were eligible for the lottery (i.e., those who later expressed interest) and those who were not. Table 12 reports estimates from regressions of baseline academic outcomes on eligibility status. In the first term, students who would later express interest scored 0.085 standard deviations higher than their peers (p ¡ 0.1) (see Figure 6). However, by the second term—still prior to the lottery—this relationship reverses: students who

while seemingly challenging, foster more durable and flexible learning (Bjork, 1994). For example, the initial and suggested prompts incorporated evidence-based principles such as retrieval practice—shown to be effective for upper secondary students when implemented through multiple-choice and short-answer quizzes (McDermott et al., 2014)—elaborative interrogation (Dunlosky et al., 2013), and the use of concrete examples (Weinstein et al., 2018). However, we believe there is significant potential for future iterations of the intervention to more fully exploit evidence-based strategies for improving learning outcomes. For instance, while in our program, each session was dedicated to a single curriculum topic, future programs could experiment with variations, such as incorporating interleaving (Weinstein et al., 2018) and spacing practices (Kang, 2016). These approaches would allow for the coverage of multiple topics within a single session, revisiting and reinforcing them over time to enhance long-term retention and understanding.

[4]For details on this tool, see Das (2024).

[5]The monitoring data included teacher and student attendance, punctuality, power and internet conditions, as well as participants' engagement, among other factors

did not express interest scored 0.147 standard deviations higher (p ¡ 0.01) (see Figure 7). The absence of a consistent directional pattern across terms suggests that selection into the program was not strongly or systematically correlated with academic performance. While our analysis focuses on treatment effects among those who expressed interest, the lack of clear academic selection implies that results may generalize beyond this group. Nevertheless, we lack demographic data on non-interested students, which limits our ability to assess representativeness along other dimensions.

Once the period to express interest closed, the randomization was carried out using simple random sampling without replacement[6] among interested students to assign them either to the treatment group, which participated in the program, or to the control group, which did not receive any intervention but continued their regular learning in the classroom. The students completed a baseline survey and an end-line survey with sociodemographic information. Initially, 657 students were assigned to the treatment group and 671 to the control group. However, only 422 students in the treatment group and 337 in the control group completed the final assessment, which constitutes the final sample used for the analysis.

Table 1 provides summary statistics and balance tests for key observable characteristics of the two groups. Demographic variables include gender, age, and a socio-economic status (SES) index. This index was derived from a principal components analysis of household characteristics, such as access to goods (computers, phones), services (internet connection), study spaces, and parental education.[7] The SES index, as well as other variables such as the proportion of female students and age, shows that the sample is balanced across the treatment and control groups, with differences that are small and not statistically significant. These results confirm that the randomization process achieved balance in key characteristics, supporting the validity of subsequent comparisons between the treatment and control groups.

In addition to sociodemographic information, baseline academic performance was measured using scores from the First and Second Term Exams conducted prior to the intervention. The difference between treatment and control group students in mean baseline scores for the First Term Exam is 0.131 (SE = 0.073), and for the Second Term Exam, it

[6]Randomization was conducted without stratification using a computerized system. Although the randomization process did not incorporate a fixed random seed, the allocation results were documented and saved, ensuring the assignments are reproducible and transparent, as recommended by Bruhn and McKenzie (2009).

[7]For a discussion on using principal component analysis to build a SES index, see Vyas and Kumaranayake (2006).

is 0.096 (SE = 0.073). These differences are also statistically insignificant, indicating that students in both groups had comparable academic performance before the program.

2.3 Learning Data used as the Dependent Variable

At the end of the six-week intervention, participating and non-participating students completed a standardized assessment designed to measure three key outcomes: (a) English language proficiency aligned with the Nigerian curriculum for the corresponding period (our main outcome of interest), (b) knowledge of AI, and (c) understanding of basic digital concepts (from now on referred as "digital skills" for convenience). The majority of the questions aimed to assess English language. To minimize the risk of cheating, multiple versions of the assessment were created, each with a randomized order of questions. Additionally, monitors were stationed at the schools to oversee the administration of the assessments and ensure compliance with testing protocols. The assessment was administered in a traditional pencil-and-paper format and consisted of multiple-choice questions designed by experts based on the Nigerian curriculum.

For each student, a simple score was generated based on the percentage of correct answers across all topics, as well as separate scores for each of the three domains (English language, AI knowledge, and digital skills). In addition to these unweighted scores, weighted scores were calculated for each domain and for the overall assessment. The weights were based on the ex-ante difficulty of each test item, which was determined by the test designers prior to the administration.

An additional score for proficiency in each theme was calculated using Item Response Theory (IRT).[8] This method allowed for comparability across students by placing their performance on a common scale, taking into account the actual difficulty of each question as observed in the sample. The IRT-derived score provided a more nuanced measurement of English language ability -as well as knowledge of AI and digital skills- by considering both the students' responses and the varying difficulty levels of the assessment items.

In addition to the intervention-specific assessment, an additional dependent variable was derived from the student's final English exam scores. This exam, which was conducted independently by the school, covered the entire term's content, which extended beyond the six-week period of the after-school program.

[8]For a more detailed explanation of IRT models, see Van der Linden and Hambleton (2015). For a discussion on the importance of using IRT in education RCTs, see Muralidharan (2017).

3 Results

3.1 Model and Main Results

We estimate the intent-to-treat (ITT) effects of the intervention using the following regression:

$$Y_{iks} = \alpha + \gamma_s \cdot Baseline_{ik} + \beta_s \cdot \text{Treatment}_i + \phi_k + \epsilon_{iks} \quad (1)$$

The dependent variable Y_{iks} represents student i's test score in school k and subject s. The specification includes a coefficient γ_s for the student's pre-treatment exam score from the second term ($Baseline_{ik}$), and a term β_s for our primary variable of interest—an indicator for whether the student won the treatment lottery. To account for unobserved school-level heterogeneity, we include school fixed effects (ϕ_k).

Although the difference in academic performance before the intervention between the treatment and the control group was not statistically significant, it was still positive in favor of the treatment group. Therefore, we followed a conservative approach to ensure the robustness of our results, including the performance in the second term exam as a control variable. This approach is also recommended by Muralidharan (2017).

Table 2 reports the intent-to-treat (ITT) effects of the intervention on three main outcomes: the total score of the final assessment (weighted and IRT scaled) and the score from the third term exam. The coefficient for the second-term exam is significant across all models, reflecting the predictive validity of prior performance. School-fixed effects are included in all models, and the number of observations ranges between 636 and 654, depending on the outcome.

The treatment effect on the total score (weighted) is 0.31 standard deviation (SE = 0.068), and the effect remains positive and significant at 0.263 standard deviation (SE = 0.068) when scaled using Item Response Theory (IRT). These results indicate that the intervention led to substantial improvements in students' performance on the assessment directly tied to the program. Importantly, treatment also had a positive and significant impact on the third term exam score, with an effect size of 0.206 standard deviation (SE = 0.067), although this exam was not limited to the intervention's specific content. This suggests that the intervention may have fostered generalizable skills or improved learning outcomes beyond the targeted content.

Table 3 provides a more granular analysis by disaggregating the total score into English skills, digital skills, and AI skills. The results reveal that the intervention had the largest effect on AI knowledge, with a coefficient of 0.309 standard deviations (SE = 0.077), followed by English skills (0.238 σ, SE = 0.068) and digital skills (0.139 σ, SE = 0.076). Effects on English skills (our main outcome of interest) and on AI skills are statistically significant at the 1% level. The effects on digital skills are statistically significant at the 10 % level. The positive and significant effects on AI and digital skills further indicate potential spillovers to other skill areas, even though these were not the primary targets of the program. As with Table 2, the second term exam scores are strong predictors of outcomes, and school fixed effects are included to account for unobserved factors at the school level.

The results of this pilot study are particularly notable given that several factors likely attenuated the estimated treatment effects. First, the randomization was conducted at the student level rather than at the school level, a design feature that may have led to spillover effects, as students in the control group may have interacted with peers in the treatment group during regular school hours, potentially diffusing the impact of the intervention. Second, monitoring and evaluation data indicate that some control group students inadvertently gained access to the after-school sessions, due to the lack of willingness of some teachers to enforce the distinction, especially during the first weeks. Additionally, significant implementation challenges occurred during the first weeks of the program, with many students encountering difficulties creating accounts and engaging with the LLM. Despite these challenges, the intervention yielded positive and significant results, suggesting that the observed effects should be interpreted as conservative estimates of the intervention's impact.

A recent review of randomized controlled trials in pre-primary, primary, and secondary education in low and middle-income countries found a median effect of 0.10 standard deviations in overall test scores and 0.14 in reading (Evans and Yuan, 2022). Thus, the results found in this study are situated at least at the 80th percentile of all RCTs, even though effect sizes in secondary education tend to be lower than in primary. Even when considering only RCTs that had between 500 and 1000 participants, the results are still higher than 80% of the other studies. When considering only the effects on language outcomes, the results are near the 70th percentile of all studies.

3.2 Heterogeneity

As a first step, we conducted quantile regressions to examine the treatment effect at different points in the outcome distribution. The analysis indicates that the treatment has a positive and statistically significant effect across all quantiles, suggesting broad benefits for students regardless of their initial performance levels. Table 4 examines the heterogeneity in treatment effects by gender, SES, and baseline academic performance. Column (1) explores heterogeneity by gender, using an interaction term between the treatment indicator and a female dummy variable. While the main effect of treatment is not statistically significant for this specification, the interaction term for treatment and being female is positive and significant (0.420) at the 5% level, indicating that the intervention had a larger positive effect on female students compared to their male counterparts. This result should be interpreted with caution, as it appears to be influenced by the inclusion of a girls-only school that performed worse than others in the sample prior to the intervention.

Column (2) considers heterogeneity by baseline academic performance, as measured by the second-term exam score. The interaction term between treatment and the second-term exam score is positive and significant at the 5% level, indicating that students with higher prior academic performance benefited more from the intervention. Column (3) examines heterogeneity by socio-economic status, using an interaction term between treatment and the SES index. The interaction term is positive (0.113) and significant, also at the 5% level, suggesting that students from higher SES backgrounds experienced larger treatment effects. This finding aligns with anecdotal evidence indicating that, for students from poorer households, this was often one of their first experiences using computers. While these students still experienced significant gains relative to the counterfactual of not participating, the initial unfamiliarity with technology may have moderated the magnitude of the intervention's impact.

3.3 Dosage Effects

All the results presented thus far are ITT estimates, which are based on an average attendance rate of approximately 72% among participants in the treatment group. In this section, we present LATE and OLS estimates, which measure the impact of actually attending the sessions. These estimates leverage attendance data collected as part of the program's monitoring and evaluation efforts (see Figure 5). Additionally, under further assumptions, we provide predicted treatment effects at varying levels of program expo-

sure. We estimate the dose-response relationship between days of attendance and value-added using the following model:

$$Y_{iks} = \alpha + \gamma_{ik} \cdot \text{Baseline}_{ik} + \mu \cdot \text{Attendance}_{ik} + \epsilon_{iks} \tag{2}$$

The dependent variable Y_{is} represents student i's test score in subject s in school k. The specification includes a coefficient γ_{ik} for the student's second-term pre-treatment exam score ($Baseline_{ik}$), and a term μ for our primary variable of interest—the number of days the student attended intervention sessions, which takes a value of zero for the control group (Attendance_{ik}).

Table 5 shows that higher attendance is strongly associated with improved learning outcomes, providing evidence of a dose-response effect, with an estimated effect size of approximately d=0.033 for each additional day of attendance. This finding underscores the importance of consistent participation, as greater exposure to the program results in meaningful improvements in student outcomes.

These estimates capture the average causal response (ACR) of the treatment, which represents a weighted average of the causal effects of a one-unit change in treatment (in this case, an additional day of attendance) for individuals whose treatment status is influenced by the instrument (Angrist and Imbens, 1995). However, using these Instrumental Variable (IV) estimates to predict the effects of different levels of attendance requires additional assumptions, as stated in Muralidharan et al. (2019): (i) assumptions about how treatment effects vary across students, as the ACR is identified only for a subset of compliers (those who attend at least one session) and not the entire sample, and (ii) assumptions about the functional form of the relationship between days attended and treatment effects, since the ACR reflects an average across varying levels of treatment intensity.

For the first assumption, we cannot assume that the effect on the non-compliers would have been the same as the compliers. This is because we have evidence showing that low performing students tend to benefit less from the program (Table 4), and that better performance in previous examinations correlates positively with attendance (Table 9). Therefore, we follow a conservative approach and assume that the non-compliers -those who are assigned to the treatment group but do not attend any sessions- would not benefit at all had they attended the program. Under that assumption, and given that the non-compliers represent only 3.5% of those assigned to the treatment group, the estimated effect size of each additional day of attendance is of 0.031.

We follow this conservative approach despite the fact that, following Muralidharan et al.

(2019), two other pieces of evidence would be consistent with expecting the ACR to apply even to the non-compliers. First, we cannot reject the equality of the IV estimates of equation (3) and the ordinary least squares (OLS) estimates using a value-added (VA) specification, which suggests that the average treatment effects and local average treatment effects (ATE and LATE) may be similar. Second, the constant term in the OLS VA specifications (corresponding to 0 attendance) is similar when estimated using the full sample and when estimated using only the data in the treatment group. This suggests equality of potential outcomes across students with different compliance rates.

Regarding the functional form of the relationship between days attended and the treatment effect, a graphic representation suggests a linear relationship (Figure 4). Moreover, while the value added of the scores is strongly correlated with the number of days attended in a linear specification, adding a quadratic term does not improve the fit, and the quadratic term is not significant, as shown on Table 7. From a more theoretical point of view, the linear effect is also expected given the adaptive nature of LLMs. Thus, it seems fair to assume that, while effect is not the same across students, the effect for each student does not show diminishing returns to program exposure.

Qualitative evidence from the intervention aligns with these results, suggesting that the first days of the program had little to no measurable impact on learning outcomes. This lag may reflect the time required for students to familiarize themselves with the technology and adjust to the new instructional format. Beyond this initial acclimation period, the effects of additional days of attendance remained consistently positive, with no evidence of plateauing. This suggests that extending the implementation period beyond the six weeks of the intervention could have further amplified learning gains. It also suggests that the LATE estimates might be underestimated, given the null effects of the first few days of participation.

Under the assumptions of constant treatment effects diminished by the lack of effect on non-compliers and a linear dose-response relationship—both of which appear reasonable in this context— and following the methodology used in (Muralidharan et al., 2019), our instrumental variables analysis predicts that attending the program for 36 weeks (the equivalent of a school year) would lead to gains of approximately 2.23 standard deviations. A more conservative estimate based on an attendance rate of 72% -the empirical value in our sample-, predicts that attending the program for 21 weeks (over a theoretical total of 36) would lead to gains in the range of 1.55 standard deviations. Under a more pessimistic scenario of only 50% attendance rates, the estimate would still be 1.2 standard deviations. These findings highlight the potential of sustained implementation to yield

transformative impacts on learning outcomes.

3.4 Robustness

To test the robustness of the results, we performed several checks, progressively adjusting the baseline model to improve robustness and account for potential sources of bias. First, for all the models presented, we employed robust standard errors to address heteroskedasticity in the data. Second, we included school-fixed effects in all the models presented, with robust standard errors, to control for unobservable school-level characteristics that might influence student outcomes. Third, we incorporated students' performance in the second-term exam, conducted prior to the intervention, as a control variable. Although the difference in performance between the treatment and control groups was not statistically significant, the treatment students exhibited slightly higher scores. As mentioned above, adding this variable allows us to adopt a conservative approach, mitigating the risk of any potential, albeit unlikely, selection bias. The models with all these specifications are the basis for our main results, presented in Table 2.

Furthermore, we estimated the models using alternative specifications of the dependent variable to assess the robustness of our findings. For each outcome of interest, total score, English language proficiency, digital skills, and AI knowledge, we analyzed specifications with and without the experts' weighted difficulty of questions. Additionally, we employed IRT as an alternative approach to scoring the assessments. This approach ensures that the estimated outcomes are not unduly influenced by the design or composition of the test. We presented the main results for the total score weighted (Column (1) of Table 2), which are 0.31 standard deviation, and the IRT scaled results (Column (2) of Table 2), of 0.263 standard deviation.

Table 8 presents a sensitivity analysis to test the robustness of the treatment effects by iteratively excluding one school at a time from the sample. The dependent variable shown is the total score (weighted) from the final learning assessment, but similar results are found with other specifications. The estimated treatment effects range from 0.156 (SE = 0.085) when excluding Idia College to 0.360 (SE = 0.072) when excluding Imaguero College. While the effect remains statistically significant at the 1% level in most cases, the exclusion of Idia College reduces the effect size and significance to the 10% level. This is, however, understandable, given the large size of Idia College (219 of 657, or 33% of the sample), whose exclusion reduces the power of the calculation. For all other schools, the estimated coefficients are stable around 0.30 and maintaining statistical significance at

the 1% level. Overall, the results demonstrate that the treatment effects remain consistent and statistically significant across most specifications, indicating that the findings are not driven by the influence of any single school.

Finally, since the difference in attrition between the treatment and control groups is significant, we first provide Lee-bounds estimates of ITT effects on the outcome variables. These bounds indicate the range of the ITT effects estimated using Lee (2005) bounding approach. This method accounts for possible bias due to attrition or selection into the program. The analysis shows that the treatment effects are always positive and significant, even following this conservative approach (Table 10). In addition, we assess the robustness of our findings to attrition by modeling the likelihood of participation in the endline based on observed characteristics. We then calculate inverse probability weighted treatment effects, finding that the estimated ITT effects remain largely unchanged (Table 11). Thus, our main conclusions hold even in the presence of non-random attrition at endline.

4 Discussion

4.1 Cost-Effectiveness

This section provides a cost-effectiveness analysis of the pilot, drawing comparisons with other high-dosage programs, and discusses some challenges and opportunities for scalability. We measure the program's nominal cost using planning and budget data, and estimate its implicit costs on pro rata basis (Valdivia Teixeira, 2019). Implementing the 6-week pilot for 657 students, had a per-pupil cost of approximately $48, and the marginal cost is estimated at $9. Furthermore, extending the pilot over the four academic quarters would cost $124 per pupil, which can be particularly useful to frame policy discussions about the return on investment of a longer intervention even without further improvement in learning outcomes (a conservative approach considering our dosage effects results). Table 13 provides a breakdown of the pilot's costs and our estimation for a four-quarter program. As the pilot was implemented, fixed costs accounted for 43% of overall costs (39% in a hypothetical four-quarter program), which provides an idea of the potential cost reduction in a second round of the program.[9]

To analyze the effectiveness of the pilot, we follow the methodology used in Evans and

[9]This is particularly relevant considering that content development accounted for 72% of all fixed costs, and would be 80% for yearly implementation.

Yuan (2019). To estimate effectiveness, size effects are translated into EYOS, which expresses the years of 'business-as-usual' schooling in terms of learning outcomes delivered by a given intervention. Our ITT effect of 0.238 standard deviation in English[10] is equivalent to increasing 1.5 years of 'business-as-usual' schooling in Nigeria, and the total score gain of 0.31 standard deviation is equivalent to 2 additional years of schooling in the country. This is more than twice the effect of some of the most effective interventions in education, such as structured pedagogy, which is typically implemented for the entire school year. For the remainder of this section, we use effects on English given that this is our main outcome of interest. The intervention's cost-effectiveness, measured in EYOS per $100 (theoretically) invested per participant, is estimated to yield 3.2 EYOS under the assumption of constant returns.

We convert gains in test scores into increased wages (Evans and Yuan, 2019). We estimate that improvements in English would result in an increase of 14% in wages.[11] The additional annual income ranges between $392 and $630.[12] Over their work-life, each participant has a present discounted value of income gains of $7,767 to $12,517[13] When considering the long-term wage effects and the cost of our pilot, the benefit-cost ratio of our pilot program is 161 to 260. As a reference point, we calculate that running the pilot for one year without further improvements in learning outcomes would still yield a benefit-cost ratio as high as 62 to 100.

For comparison purposes, the return on investment of the program is very high compared to recent evidence from high-dosage personalized tutoring programs with and without technology in the United States. Guryan et al. (2023), Bhatt et al. (2024), Fryer and Howard-Noveck (2020) yield benefit-cost ratios that range from 2.4 to 8. However, evidence from low-income countries (LICs) and lower-middle-income countries (LMICs) shows benefit-cost ratios that vary between 8 and 156 (Glewwe et al. (2010), Duflo et al. (2011), Banerjee et al. (2007), Evans and Yuan (2019)). In line with our results, a recent review of 150 interventions in global education (Angrist et al., 2023) found that programs teaching at the right level with a technology component generate the largest benefit-cost

[10]This section focuses on English outcomes since most of the literature that uses EYOS and LAYS focuses on language or math skills. If instead we used the overall scores, the estimates would be larger.

[11]Following Evans and Yuan (2019), since we do not have data on returns to learning in Nigeria, we use that of Kenya, the country with the closest income per capita to Nigeria (($6,200 versus $6,020 in PPP, current 2023).

[12]These values depend on what labor-share of income is applied. The World Penn Table estimates it at 0.465. (Feenstra et al., 2015), and the ILO estimation is 0.748. While we prefer the ILO methodology because it adjusts for self-employed income, we provide both values.

[13]We use a 3% discount rate, assume our representative agent will enter the labor market at 20 years old, and have a 40-work life. We also assume that wage returns to skills are constant across one's working life.

ratios across LICs and LMICs, with a weighted average of 65.

Like in any other lower-middle income country, the productivity of a 'business as usual' day in a Nigerian school is lower than in a high-performing country. To facilitate cross-country comparisons, we calculate the LAYS (Angrist et al., 2025; Filmer et al., 2020), which adjusts the learning gains in years of schooling of our program by the quality of learning in Nigeria. Using the learning gain in English skills (0.24 σ), we estimate the LAYS under two scenarios. If the effects last only one year, the intervention produces 0.3 LAYS. Conversely, if the impacts last for the remaining school life expectancy, the program generates an additional 0.9 year of high-quality schooling for each participant.[14] In other words, our Nigerian participants gained (on average) up to 0.9 years of a top-performing country education. Finally, our program yields between 0.6 and 1.9 LAYS per $100. If instead of the estimates based on the 6-week program we used the estimates calculated in the dosage effect section for English, the LAYS for one year of the program -accounting for the observed attendance rate- would be 1.25.

Beyond comparing small quantitative differences, it is important to assess results as ordinal rather than cardinal, given the varying underlying assumptions, imprecision in estimations, and contextual conditions of each study. Through this lens, analysis can inform broad trade-offs in policy making, budget allocation, and program design. Consequently, the pilot program's benefit-cost ratios and other cost-effectiveness metrics are situated at the upper end or above benchmarks, highlighting its potential as a cost-effective solution for addressing learning crises in low-resource settings.

4.2 Future Research Directions

Several potential avenues for future research emerge from the findings of this study. First, extending the duration of the program beyond six weeks could provide insight into whether a longer intervention leads to more substantial or sustained improvements in learning outcomes and what the shape of the learning curve is as time progresses. A more extended program might also allow for more complex interactions with the chatbot, further enhancing its educational benefits. These extended studies should be complemented with a more qualitative assessment of the interaction between the students and the AI tool, in order to understand the causal mechanism driving the improvements in learning, and the specific ways that students benefit from the virtual tutoring.

[14] Considering that our participants are in Senior Secondary 1, we use $t = 3$ as the remaining expected years in school.

Second, expanding the study to include a more diverse set of schools, particularly those in rural areas, would improve the external validity of the findings. By investigating the effectiveness of the program in various educational settings, it could be possible to assess its scalability and adaptability to different contexts. Both a longer duration of the intervention and a broader set of schools could show insights into what Rodriguez-Segura (2022) calls "general equilibrium" effects after the roll-out of educational technology interventions, which could include potential changes in teacher attitudes, effort, and behavior, even as part of the regular instruction.

Another promising avenue for exploration is the addition of an additional treatment arm that offers one-on-one tutoring by teachers without the use of technology. This would allow for a direct comparison between the effectiveness of LLM-powered tutoring and traditional teacher-led tutoring, providing valuable insights into cost-effectiveness and pedagogical efficacy, and helping to calculate the productivity enhancement effect that technology might have on teachers. Similarly, additional arms can help disentangle the multiple causal mechanisms that might be driving the effect, including additional instructional time and the interaction with the chatbot with teacher support.

Understanding the long-term impacts of the intervention is also crucial. Future studies should investigate whether the positive effects observed in the short term persist over time, contributing to lasting improvements in students' academic trajectories. Similarly, from a policy perspective, it would be valuable to assess whether an after-school program like this leads to a long-term shift in effort or time away from productive in-school activities, and whether an in-school program might offer even greater effectiveness as an alternative.[15]

Finally, further research could explore whether students are transferring -without explicit instruction- their skills in using AI tools from one subject area to another. For example, future studies could examine whether familiarity with AI tools in English language lessons enhances student performance in other subjects, such as mathematics or science. This cross-curricular application of AI would provide insights into the broader academic potential of LLMs in education.

[15]For a discussion on the advantages and disadvantages of after-school programs related to computer-assisted adaptive learning, see Mo et al. (2014).

4.3 Policy Implications

The findings of this pilot intervention highlight several promising policy implications for addressing the learning crisis in developing countries, particularly in Sub-Saharan Africa.

First, the intervention demonstrated substantial impacts on learning outcomes, despite some implementation challenges, such as internet disruptions and power outages. This is particularly encouraging for countries grappling with severe teacher shortages, high population growth, and increasing teacher attrition rates. One key takeaway for policymakers is that investing in AI-powered tutoring programs powered by LLMs, could significantly increase teacher productivity, which is consistent with recent qualitative evidence (Keppler et al., 2024). By supplementing traditional classroom instruction with AI-based support, education systems can deliver personalized learning experiences, particularly in contexts where human resources are stretched thin.

Second, the program shows that LLMs can improve learning when used properly. A recent debate in the literature seems to show that LLMs can harm learning when used as shortcuts, in other words, when used to facilitate responses to students' questions without encouraging them to think. These are results found in studies such as Bastani et al. (2024). Some studies have also shown that the use of LLMs can lead to lower-quality reasoning and argumentation when students use them to search for information (Stadler et al., 2024). Contrarily, the intervention we assessed seems to suggest that LLMs can improve learning when used specifically as tutors adapted to specific use cases and contexts via prompting. Thus, the findings are consistent with the idea emphasized in Gerlich (2025) that educational strategies should promote critical engagement with AI technologies in order to avoid cognitive offloading, which might reduce critical thinking skills. The intervention evaluated in this paper leveraged third key mechanisms to achieve effective tutoring. First, the prompts were intentionally crafted to guide the LLM to provide explanations and support grounded in the principles of the science of learning, rather than simply supplying direct answers. Second, teachers played an essential role in monitoring and guiding students' use of the LLM to ensure it was used appropriately and productively. Third, the content of each session was aligned with the official curriculum.[16]

In other words, we interpret that the intervention as a whole -which includes the interaction with the LLM and teacher guidance with specific prompts- is driving the results. We have reasons to believe that the effects are not driven solely by the additional time

[16] In this context, while our intervention differs in its use of LLMs, it aligns more closely with "computer-assisted instruction"—integrated into teachers' instruction and curriculum—than with "computer-assisted learning," which operates independently. This distinction is outlined by Bai et al. (2016).

with teachers, given that the impact of human tutoring tends to be very low when is not one-on-one or in small groups (Nickow et al., 2020; Kraft and Lovison, 2024).[17] This interpretation suggests that there might be complementarities between teachers and technology and the way technology is used and deployed is critical to understand its impact. Following Muralidharan et al. (2019), our results can also be interpreted as showing the extent to which using technology -and in particular, LLMs- in education can raise the productivity of an instructor.[18]

Third, although this intervention was conducted on a pilot scale, its cost-effectiveness makes it a promising candidate for large-scale implementation. While the effect sizes of interventions typically decrease as sample sizes grow (Evans and Yuan, 2022), the rapid advancements in LLMs and the potential for improving the implementation process suggest that future iterations of the program could be even more impactful. Moreover, the fact that the intervention was implemented with local staff (including teachers and monitors), might help with its scalability. Similarly, the use of a free tool, rather than traditional subscription-based computer-adaptive software, can significantly lower marginal costs. Moreover, LLMs offer a distinct advantage for adaptive learning: they eliminate the need to develop extensive question banks with varying difficulty levels to accurately categorize students into performance tiers, a requirement that Rodriguez-Segura (2022) emphasizes is essential for traditional adaptive software. This potential for scalability is particularly important for policy makers seeking affordable, efficient ways to address learning gaps in resource-constrained environments.

Fourth, while AI interventions have the potential to reduce learning gaps, policy makers must be vigilant about areas where such programs could inadvertently widen inequities. Although the intervention may offer Pareto-optimal benefits, disparities in digital literacy and access to technology could exacerbate existing inequalities. A prerequisite to ensure that all students can benefit from AI-powered tutors, digital skills and AI literacy programs should be introduced early in the curriculum in a practical and inclusive manner and teachers should receive training to leverage digital skills to improve their pedagogical practice and support students to be digital and AI literate. Additionally, significant investments in infrastructure and equipment are needed to provide equitable access to technology across regions. Policy makers must ensure that efforts to integrate AI into

[17] Furthermore, Rodriguez-Segura (2022) compares Büchel et al. (2022), a study from El Salvador, and Ma et al. (2024), a study from China, to suggest that in traditional computer-assisted adaptive learning, additional instructional time is unlikely to be the primary driver of improved outcomes in countries with relatively low state capacity, such as Nigeria.

[18] This aligns with the perspective presented by Mollick and Mollick (2023b), of AI as a "force multiplier" for instructors if implemented cautiously and thoughtfully in service of evidence-based teaching practices.

education are accompanied by initiatives that address the digital divide, particularly in low-income and rural areas. This may require a realignment of priorities across sectors, as education budgets are often heavily weighted toward recurrent expenditures such as salaries.

Finally, the rapid development of generative AI presents a unique opportunity to tackle the global learning crisis. By harnessing the responsible use of AI to provide personalized, adaptive learning at scale, governments can take decisive steps toward improving learning outcomes in contexts that have traditionally faced significant educational challenges.

References

ACHIAM, J., S. ADLER, S. AGARWAL, L. AHMAD, I. AKKAYA, F. L. ALEMAN, D. ALMEIDA, J. ALTENSCHMIDT, S. ALTMAN, S. ANADKAT, ET AL. (2023): "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*.

ANGEL-URDINOLA, D., C. AVITABILE, AND M. CHINEN (2023): *Can Digital Personalized Learning for Mathematics Remediation Level the Playing Field in Higher Education? Experimental Evidence from Ecuador*, The World Bank.

ANGRIST, J. D. AND G. W. IMBENS (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

ANGRIST, N., E. AURINO, H. A. PATRINOS, G. PSACHAROPOULOS, E. VEGAS, R. NORDJO, AND B. WONG (2023): "Improving Learning in Low- and Lower-Middle-Income Countries," *Journal of Benefit-Cost Analysis*, 14, 55–80.

ANGRIST, N., D. K. EVANS, D. FILMER, R. GLENNERSTER, H. ROGERS, AND S. SABARWAL (2025): "How to Improve Education Outcomes Most Efficiently? A Review of the Evidence Using a Unified Metric," *Journal of Development Economics*, 172, 103382.

BAI, Y., D. MO, L. ZHANG, M. BOSWELL, AND S. ROZELLE (2016): "The Impact of Integrating ICT with Teaching: Evidence from a Randomized Controlled Trial in Rural Schools in China," *Computers & Education*, 96, 1–14.

BANERJEE, A., R. BANERJI, J. BERRY, E. DUFLO, H. KANNAN, S. MUKHERJI, M. SHOTLAND, AND M. WALTON (2016): "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India," Tech. rep., National Bureau of Economic Research.

BANERJEE, A., R. BANERJI, E. DUFLO, R. GLENNERSTER, AND S. KHEMANI (2008): "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India," Working Paper 14311, National Bureau of Economic Research.

BANERJEE, A. V., S. COLE, E. DUFLO, AND L. LINDEN (2007): "Remedying Education: Evidence from Two Randomized Experiments in India," *The Quarterly Journal of Economics*, 122, 1235–1264.

BASTANI, H., O. BASTANI, A. SUNGU, H. GE, O. KABAKCI, AND R. MARIMAN (2024): "Generative AI Can Harm Learning," *Available at SSRN*, 4895486.

BHATT, M. P., J. GURYAN, S. A. KHAN, M. LAFOREST-TUCKER, AND B. MISHRA (2024): "Can Technology Facilitate Scale? Evidence from a Randomized Evaluation of High Dosage Tutoring," Working Paper 32510, National Bureau of Economic Research.

BJORK, R. A. (1994): "Memory and metamemory considerations in the training of human beings," .

BLOOM, B. S. (1984): "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring," *Educational Researcher*, 13, 4–16.

BRUHN, M. AND D. MCKENZIE (2009): "In pursuit of balance: Randomization in practice in development field experiments," *American economic journal: applied economics*, 1, 200–232.

BÜCHEL, K., M. JAKOB, C. KÜHNHANSS, D. STEFFEN, AND A. BRUNETTI (2022): "The Relative Effectiveness of Teachers and Learning Software: Evidence from a Field Experiment in El Salvador," *Journal of Labor Economics*, 40, 737–777.

DAS, A. S. (2024): "Chapter 4 - KoboToolbox," in *Open Electronic Data Capture Tools for Medical and Biomedical Research and Medical Allied Professionals*, ed. by A. Pundhir, A. K. Mehto, and A. Jaiswal, Academic Press, 241–329.

DENG, R., M. JIANG, X. YU, Y. LU, AND S. LIU (2024): "Does ChatGPT Enhance Student Learning? A Systematic Review and Meta-Analysis of Experimental Studies," *Computers & Education*, 105224.

DUFLO, E., P. DUPAS, AND M. KREMER (2011): "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 101, 1739–1774.

DUNLOSKY, J., K. A. RAWSON, E. J. MARSH, M. J. NATHAN, AND D. T. WILLINGHAM (2013): "Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology," *Psychological Science in the Public interest*, 14, 4–58.

EVANS, D. AND F. YUAN (2019): "Equivalent Years of Schooling: A Metric to Communicate Learning Gains in Concrete Terms," *World Bank Policy Research Working Paper*.

EVANS, D. K. AND A. MENDEZ ACOSTA (forthcoming): "Teacher Demand in African Countries: Trends and Challenges," CGD Working Paper, Center for Global Development.

EVANS, D. K. AND F. YUAN (2022): "How Big Are Effect Sizes in International Education Studies?" *Educational Evaluation and Policy Analysis*, 44, 532–540.

FEENSTRA, R. C., R. INKLAAR, AND M. P. TIMMER (2015): "The Next Generation of the Penn World Table," *American Economic Review*, 105, 3150–3182.

FILMER, D., H. ROGERS, N. ANGRIST, AND S. SABARWAL (2020): "Learning-Adjusted Years of Schooling (LAYS): Defining a New Macro Measure of Education," *Economics of Education Review*, 77, 101971.

FRYER, R. G. AND M. HOWARD-NOVECK (2020): "High-Dosage Tutoring and Reading Achievement: Evidence from New York City," *Journal of Labor Economics*, 38, 421–452.

GERLICH, M. (2025): "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking," *Societies*, 15, 6.

GLEWWE, P., N. ILIAS, AND M. KREMER (2010): "Teacher Incentives," *American Economic Journal: Applied Economics*, 2, 205–227.

GURYAN, J., J. LUDWIG, M. P. BHATT, P. J. COOK, J. M. V. DAVIS, K. DODGE, G. FARKAS, J. FRYER, ROLAND G., S. MAYER, H. POLLACK, L. STEINBERG, AND G. STODDARD (2023): "Not Too Late: Improving Academic Outcomes among Adolescents," *American Economic Review*, 113, 738–765.

HENKEL, O., H. HORNE-ROBINSON, N. KOZHAKHMETOVA, AND A. LEE (2024): "Effective and Scalable Math Support: Experimental Evidence on the Impact of an AI-Math Tutor in Ghana," in *International Conference on Artificial Intelligence in Education*, Springer, 373–381.

ITO, H., K. KASAI, H. NISHIUCHI, AND M. NAKAMURO (2021): "Does Computer-Aided Instruction Improve Children's Cognitive and Noncognitive Skills?" *Asian Development Review*, 38, 98–118.

KANG, S. H. (2016): "Spaced repetition promotes efficient and effective learning: Policy implications for instruction," *Policy Insights from the Behavioral and Brain Sciences*, 3, 12–19.

KEPPLER, S., W. P. SINCHAISRI, AND C. SNYDER (2024): "Backwards Planning with Generative AI: Case Study Evidence from US K12 Teachers," *Available at SSRN*.

KESTIN, G., K. MILLER, A. KLALES, T. MILBOURNE, AND G. PONTI (2024): "AI Tutoring Outperforms Active Learning," *Research Square*.

KRAFT, M. A. AND V. S. LOVISON (2024): "The Effect of Student-Tutor Ratios: Experimental Evidence from a Pilot Online Math Tutoring Program. EdWorkingPaper No. 24-976," *Annenberg Institute for School Reform at Brown University*.

KUMAR, H., D. M. ROTHSCHILD, D. G. GOLDSTEIN, AND J. M. HOFMAN (2023): "Math Education with Large Language Models: Peril or Promise?" *Available at SSRN*.

LAI, F., R. LUO, L. ZHANG, X. HUANG, AND S. ROZELLE (2015a): "Does Computer-Assisted Learning Improve Learning Outcomes? Evidence from a Randomized Experiment in Migrant Schools in Beijing," *Economics of Education Review*, 47, 34–48.

LAI, F., L. ZHANG, Q. QU, X. HU, Y. SHI, M. BOSWELL, AND S. ROZELLE (2015b): "Teaching the Language of Wider Communication, Minority Students, and Overall Educational Performance: Evidence from a Randomized Experiment in Qinghai Province, China," *Economic Development and Cultural Change*, 63, 753–776.

LEE, D. S. (2005): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," .

LEHMANN, M., P. B. CORNELIUS, AND F. J. STING (2024): "AI Meets the Classroom: When Does ChatGPT Harm Learning?" *arXiv preprint arXiv:2409.09047*.

MA, Y., R. FAIRLIE, P. LOYALKA, AND S. ROZELLE (2024): "Isolating the "Tech" from Edtech: Experimental Evidence on Computer-Assisted Learning in China," *Economic Development and Cultural Change*, 72, 1923–1962.

MCDERMOTT, K. B., P. K. AGARWAL, L. D'ANTONIO, H. L. ROEDIGER III, AND M. A. MCDANIEL (2014): "Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes." *Journal of Experimental Psychology: Applied*, 20, 3.

MO, D., L. ZHANG, R. LUO, Q. QU, W. HUANG, J. WANG, Y. QIAO, M. BOSWELL, AND S. ROZELLE (2014): "Integrating computer-assisted learning into a regular curriculum: Evidence from a randomised experiment in rural schools in Shaanxi," *Journal of Development Effectiveness*, 6, 300–323.

MOLLICK, E. AND L. MOLLICK (2023a): "Assigning AI: Seven Approaches for Students, with Prompts," *arXiv preprint arXiv:2306.10052*.

MOLLICK, E. R. AND L. MOLLICK (2023b): "Using AI to Implement Effective Teaching Strategies in Classrooms: Five Strategies, Including Prompts," *The Wharton School Research Paper*.

MURALIDHARAN, K. (2017): "Field experiments in education in developing countries," in *Handbook of Economic Field Experiments*, Elsevier, vol. 2, 323–385.

MURALIDHARAN, K., A. SINGH, AND A. J. GANIMIAN (2019): "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India," *American Economic Review*, 109, 1426–1460.

NICKOW, A., P. OREOPOULOS, AND V. QUAN (2020): "The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence," Working paper, National Bureau of Economic Research.

PERERA, M. AND D. ABOAL (2019): "The Impact of a Mathematics Computer-Assisted Learning Platform on Students' Mathematics Test Scores," MERIT Working Papers 2019-007, United Nations University - Maastricht Economic and Social Research Institute on Innovation and Technology (MERIT).

RODRIGUEZ-SEGURA, D. (2022): "EdTech in Developing Countries: A Review of the Evidence," *The World Bank Research Observer*, 37, 171–203.

STADLER, M., M. BANNERT, AND M. SAILER (2024): "Cognitive Ease at a Cost: LLMs Reduce Mental Effort but Compromise Depth in Student Scientific Inquiry," *Computers in Human Behavior*, 160, 108386.

VALDIVIA TEIXEIRA, S. (2019): "Capturing Cost Data," https://thedocs.worldbank.org/en/doc/994671553617734574-0090022019/original/CapturingCostData190314.pdf, accessed: 2025-01-24.

VAN DER LINDEN, W. J. AND R. K. HAMBLETON (2015): *Handbook of Item Response Theory*, CRC Press.

VANZO, A., S. P. CHOWDHURY, AND M. SACHAN (2024): "GPT-4 as a Homework Tutor Can Improve Student Engagement and Learning Outcomes," *arXiv preprint arXiv:2409.15981*.

VYAS, S. AND L. KUMARANAYAKE (2006): "Constructing socio-economic status indices: how to use principal components analysis," *Health Policy and Planning*, 21, 459–468.

WANG, R. E., A. T. RIBEIRO, C. D. ROBINSON, S. LOEB, AND D. DEMSZKY (2024): "Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise," *ArXiv preprint arXiv:2410.03017*.

WEIDLICH, J., D. GAŠEVIĆ, AND P. A. KIRSCHNER (2025): "ChatGPT in education: An effect in search of a cause," .

WEINSTEIN, Y., C. R. MADAN, AND M. A. SUMERACKI (2018): "Teaching the science of learning," *Cognitive research: principles and implications*, 3, 1–17.

WORLD BANK (2022): "The State of Global Learning Poverty Report: 2022 Update," Tech. rep., World Bank, accessed: 2024-09-20.

Table 1: Sample descriptive characteristics and balance on observables

	Mean (treatment)	Mean (control)	Difference	SE	95% CI
Demographic characteristics					
Female	0.773	0.795	-0.023	0.03	[-0.082,0.036]
Age	15.19	15.12	0.07	0.08	[-0.088,0.222]
SES index	0.059	-0.075	0.133	0.105	[-0.073,0.339]
Baseline test scores					
First Term Exam	0.062	-0.069	0.131	0.073	[-0.013,0.275]
Second Term Exam	0.045	-0.05	0.096	0.073	[-0.048,0.24]
School					
School A	0.031	0.018	0.013	0.011	[-0.009,0.035]
School B	0.045	0.071	-0.026	0.017	[-0.06,0.008]
School C	0.052	0.059	-0.007	0.017	[-0.04,0.026]
School D	0.242	0.347	-0.105	0.033	[-0.17,-0.04]
School E	0.052	0.033	0.019	0.015	[-0.009,0.047]
School F	0.095	0.134	-0.039	0.023	[-0.085,0.007]
School G	0.159	0.148	0.01	0.026	[-0.042,0.062]
School H	0.23	0.128	0.102	0.027	[0.048,0.156]
School I	0.095	0.062	0.032	0.019	[-0.006,0.07]

Note: Treatment and control groups refer to students randomly assigned to attend the Copilot sessions. Demographic variables used to assess covariate balance in this table were measured in a baseline survey. The SES index was estimated using the first factor from a Principal Components Analysis consisting of indicators of access to certain goods (computers, phones), services (internet connection), study spaces at home, and parents' educational attainment. Baseline test scores were measured by observing students' performance in regular curricular school exams from the two terms prior to the intervention.

Table 2: Intent To Treat (ITT) Effects on Main Outcomes

	Dependent variable:		
	(1)	(2)	(3)
	Total Score (Weighted)	Total Score (IRT Scaled)	Third Term Exam
Treatment	0.310***	0.263***	0.206***
	(0.068)	(0.068)	(0.067)
Second Term Exam	0.470***	0.435***	0.504***
	(0.038)	(0.038)	(0.042)
Constant	0.274	0.232	−0.334
	(0.223)	(0.220)	(0.159)
School Fixed Effects	✓	✓	✓
Observations	654	654	636

Note: *p<0.1; **p<0.05; ***p<0.01. Heteroskedasticity-robust standard errors in parenthesis. Treatment is a dummy variable indicating whether a student is assigned to attend the Copilot sessions. Outcome in model 1 is the total score in the final learning assessment in the intervention, as described in Section 4.1. Outcome in model 2 is the same assessment score, but scaled using Item Response Theory models. Outcome in model 3 is the score obtained in the regular curricular exam in the third school term, which took place after the intervention and which content was unrelated to the intervention's material. All outcomes are standardized to have a mean of zero and a standard deviation of one.

Table 3: Intent To Treat (ITT) Effects on Specific Areas

	Dependent variable:		
	(1)	(2)	(3)
	English Skills Score	Digital Skills Score	AI Skills Score
Treatment	0.238***	0.139*	0.309***
	(0.068)	(0.076)	(0.077)
Second Term Exam	0.401***	0.324***	0.344***
	(0.038)	(0.041)	(0.042)
Constant	0.153	0.228	0.350
	(0.238)	(0.159)	(0.224)
School Fixed Effects	✓	✓	✓
Observations	654	654	654

Note: *p<0.1; **p<0.05; ***p<0.01. Heteroskedasticity-robust standard errors in parenthesis. Treatment is a dummy variable indicating whether a student is assigned to attend the Copilot sessions. Outcome in model 1 is the total score in the final learning assessment in the intervention, as described in Section 4.1. Outcome in model 2 is the same assessment score, but scaled using Item Response Theory models. Outcome in model 3 is the score obtained in the regular curricular exam in the third school term, which took place after the intervention and which content was unrelated to the intervention's material. All outcomes are standardized to have a mean of zero and a standard deviation of one.

Table 4: Heterogeneity in treatment effect by gender, socio-economic status and previous students' performance

	Interaction Term Variable		
	(1)	(2)	(3)
	Female	Second Term Exam Score	SES Index
Treatment	−0.039	0.311***	0.305***
	(0.172)	(0.068)	(0.071)
Second Term Exam	0.477***	0.393***	0.480***
	(0.038)	(0.048)	(0.039)
Treatment * Second Term Exam		0.151**	
		(0.072)	
Female	−0.293		
	(0.294)		
Treatment * Female	0.420**		
	(0.188)		
SES Index			−0.054
			(0.033)
Treatment * SES Index			0.113**
			(0.047)
Constant	0.521	0.257	0.257
	(0.362)	(0.229)	(0.222)
School Fixed Effects	✓	✓	✓
Observations	654	654	617

Note: *p<0.1; **p<0.05; ***p<0.01. Heteroskedasticity-robust standard errors in parenthesis. All models use as dependent variable the total score in the final learning assessment in the intervention, as described in Section 4.1. The interaction term in model 1 includes a dummy variable for female students. The interaction term in model 2 includes the score obtained by students in their regular curricular school exams from the second school term. The interaction term in model 3 includes the SES index, which was estimated using the first factor from a Principal Components Analysis consisting of indicators of access to certain goods (computers, phones), services (internet connection), study spaces at home, and parents' educational attainment.

Table 5: Dose-response analysis of attendance to the program's sessions: IV estimates

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Total Score (Weighted)	Total Score (IRT Scaled)	Third Term Exam	Digital Skills Score	AI Skills Score	English Skills Score
Days of attendance	0.033***	0.028***	0.022***	0.015*	0.033***	0.025***
	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)	(0.007)
Constant	0.215	0.183	−0.372**	0.202	0.291	0.108
	(0.224)	(0.223)	(0.163)	(0.166)	(0.227)	(0.239)
School Fixed Effects	✓	✓	✓	✓	✓	✓
Observations	654	654	636	654	654	654
R^2	0.262	0.233	0.318	0.133	0.140	0.204

Note: *p<0.1; **p<0.05; ***p<0.01. Heteroskedasticity-robust standard errors in parentheses. All models are two-stage least square estimates in which the main independent variable, the number of days of attendance to the Copilot workshop sessions, is instrumented for with a dummy variable indicated having been randomly assigned to attend them. Models 1, 2, 4, 5 and 6 use as dependent variable the total score in the final learning assessment in the intervention, as described in Section 4.1, either the total score (models 1 and 2) or scores in specific sections of the exam (models 4, 5 and 6). Model 6 uses as dependent variable the score in students' third term regular curricular exam.

Table 6: Dose-response analysis of attendance to the program's sessions: OLS estimates

	Dependent variable:					
	OLS VA (full sample)			*OLS VA (Treatment group)*		
	(1)	(2)	(3)	(4)	(5)	(6)
	Total Score (Weighted)	Total Score (IRT Scaled)	Third Term Exam	Total Score (Weighted)	Total Score (IRT Scaled)	Third Term Exam
Days of attendance	0.034***	0.029***	0.022***	0.041***	0.034**	0.050***
	(0.007)	(0.007)	(0.007)	(0.016)	(0.015)	(0.017)
Second Term Exam	0.418***	0.399***	0.498***	0.498***	0.502***	0.454***
	(0.036)	(0.036)	(0.037)	(0.054)	(0.053)	(0.052)
Constant	−0.229***	−0.212***	−0.117**	−0.305**	−0.272*	−0.395**
	(0.046)	(0.047)	(0.047)	(0.155)	(0.144)	(0.169)
Observations	654	654	636	344	344	336
R^2	0.225	0.198	0.268	0.260	0.250	0.233

Note: *p<0.1; **p<0.05; ***p<0.01. Heteroskedasticity-robust standard errors in parentheses. All models are ordinary least square estimates in which the main independent variable is the number of days of attendance to the Copilot workshop sessions. Models 1, 2 and 3 analyze the entire sample, while models 4, 5 and 6 replicate the analysis in a sample restricted to the treatment group. Models 1, 2, 4 and 5 use as dependent variable the total score in the final learning assessment in the intervention, as described in Section 4.1. Models 3 and 6 use as dependent variable the score in students' third term regular curricular exam.

Table 7: Dose-response analysis of attendance to the program's sessions: OLS estimates

	Dependent variable:					
	OLS VA (full sample)			*OLS VA (Treatment group)*		
	(1)	(2)	(3)	(4)	(5)	(6)
	Total Score (Weighted)	Total Score (IRT Scaled)	Third Term Exam	Total Score (Weighted)	Total Score (IRT Scaled)	Third Term Exam
Days of attendance	0.013	0.018	−0.021	−0.006	0.020	0.025
	(0.030)	(0.029)	(0.029)	(0.057)	(0.053)	(0.061)
Days of attendance (squared)	0.002	0.001	0.004	0.003	0.001	0.002
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)
Second Term Exam	0.418***	0.399***	0.498***	0.500***	0.503***	0.456***
	(0.036)	(0.036)	(0.037)	(0.054)	(0.053)	(0.053)
Constant	−0.221***	−0.208***	−0.100**	−0.147	−0.223	−0.313
	(0.047)	(0.048)	(0.048)	(0.224)	(0.197)	(0.256)
Observations	654	654	636	344	344	336
R^2	0.226	0.198	0.270	0.261	0.250	0.233

Note: *p<0.1; **p<0.05; ***p<0.01. Heteroskedasticity-robust standard errors in parentheses. All models are ordinary least square estimates in which the main independent variable is the number of days of attendance to the Copilot workshop sessions. Models 1, 2 and 3 analyze the entire sample, while models 4, 5 and 6 replicate the analysis in a sample restricted to the treatment group. Models 1, 2, 4 and 5 use as dependent variable the total score in the final learning assessment in the intervention, as described in Section 4.1. Models 3 and 6 use as dependent variable the score in students' third term regular curricular exam.

Table 8: Sensitivity analysis: results excluding iteratively each school from the sample

	Dependent variable:
	Total Score (weighted)
Excluding School A	0.305***
	(0.068)
Excluding School B	0.351***
	(0.070)
Excluding School C	0.314***
	(0.068)
Excluding School D	0.156*
	(0.085)
Excluding School E	0.319***
	(0.070)
Excluding School F	0.360***
	(0.072)
Excluding School G	0.291***
	(0.074)
Excluding School H	0.346***
	(0.074)
Excluding School I	0.346***
	(0.070)

Note: *p<0.1; **p<0.05; ***p<0.01. Heteroskedasticity-robust standard errors in parenthesis. Each row shows the coefficient for the treatment dummy variable of the baseline model, where the dependent variable is the total score in the final learning assessment in the intervention, as described in Section 4.1. All models include control for second term exam performance and school fixed effects.

Table 9: Correlates of Attendance

	Dependent variable:		
	Days of Attendance		
	(1)	(2)	(3)
Female	-1.600^{***}	-1.946^{***}	-2.394
	(0.492)	(0.547)	(1.913)
SES Index	0.348^{***}	0.257^{*}	0.243^{*}
	(0.134)	(0.145)	(0.146)
Second Term Exam Score		0.397^{*}	0.351
		(0.207)	(0.227)
Constant	6.532^{***}	6.625^{***}	10.481^{***}
	(0.443)	(0.497)	(2.347)
School Fixed Effects	N	N	✓
Observations	712	615	615

Note: *p<0.1; **p<0.05; ***p<0.01. Heteroskedasticity-robust standard errors in parenthesis. All models use as dependent variable the total number of days each student attended the Copilot workshop.

Table 10: Lee bounds estimates of ITT effects

	Dependent variable:			
	(1)	(2)	(3)	(4)
	Total Score (Weighted)	English Skills Score	Digital Skills Score	AI Skills Score
Lower	0.255	0.202	0.138	0.232
	(0.072)	(0.074)	(0.08)	(0.079)
Upper	0.327	0.238	0.145	0.351
	(0.069)	(0.074)	(0.077)	(0.071)
95% CI	[0.115, 0.462]	[0.058, 0.384]	[-0.018, 0.297]	[0.078, 0.489]

Note: Analytic standard errors in parentheses. This table presents Lee (2009) bounds on the ITT effects of being selected to participate in the after-school program, on different outcome variables. The models are equivalent to those in Tables 2 and 3, for easier interpretability of the results. The bounds are tightened using dummy variables for each school.

Table 11: ITT estimates with inverse probability weighting for attrition

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
	Total Score (Weighted)	Total Score (IRT Scaled)	Third Term Exam	English Skills Score	Digital Skills Score	AI Skills Score
Treatment	0.272***	0.227***	0.134**	0.194***	0.102	0.301***
	(0.067)	(0.067)	(0.059)	(0.068)	(0.075)	(0.074)
Second Term Exam	0.456***	0.408***	0.459***	0.387***	0.309***	0.331***
	(0.040)	(0.038)	(0.040)	(0.038)	(0.045)	(0.043)
Constant	0.313	0.261	−0.206	0.201	0.227	0.347
	(0.229)	(0.222)	(0.167)	(0.242)	(0.160)	(0.223)
School Fixed Effects	✓	✓	✓	✓	✓	✓
Observations	700	700	858	700	700	700
R^2	0.246	0.210	0.329	0.186	0.124	0.140

Note: *p<0.1; **p<0.05; ***p<0.01. Heteroskedasticity-robust standard errors in parentheses. Treatment is a dummy variable indicating random assignment to the treatment. Results in this table are weighted by the inverse of the predicted probability of being present in the measurement of the outcome. The probability is predicted using a probit model with second-term exam, sex of the student and dummies for individual schools as predictors. Models 1, 2, 4, 5 and 6 use as dependent variable the total score in the final learning assessment in the intervention, as described in Section 4.1, either the total score (models 1 and 2) or scores in specific sections of the exam (models 4, 5 and 6). Model 3 uses as dependent variable the score in students' third term regular curricular exam.

Table 12: Analysis of selection into lottery: results of eligible v. ineligible students

	Dependent variable:	
	First Term Exam	Second Term Exam
	(1)	(2)
Not eligible for Lottery	-0.085^{*}	0.147^{***}
	(0.049)	(0.050)
Constant	0.332^{**}	0.222^{*}
	(0.138)	(0.126)
Observations	1,564	1,586
R^2	0.149	0.138

Note: *p<0.1; **p<0.05; ***p<0.01.
Heteroskedasticity-robust standard errors in parentheses.

Table 13: Program Costs

	6-week Program Cost	4-quarter Program Cost
Fixed Cost		
Content Development	9,900	27,900
Equipment/ICT	1,400	1,400
Communications/PR	400	400
Travel	2,000	2,000
Total Fixed Cost	**13,740**	**31,740**
Variable Cost		
Data and Instructional Support	2,400	9,600
Equipment/ICT	2,321	9,286
Participant Fringe Benefits	4,844	16,559
Program Management	1,000	1,000
Tutor Stipends and Transportation	1,964	7,857
Tutor Training	5,400	5,400
Total Variable Cost	**17,929**	**49,702**
Grand Total	**31,669**	**81,442**
A. Program Size		
Participants	657	657
Tutors and Monitors	55	55
Students/Tutor	12	12
Schools	9	9
B. Costs		
Total Cost	31,669	81,442
Variable Cost	17,929	49,702
C. Average Total Cost		
Per Participant	48	124
D. Average Variable Cost		
Per Participant	27	76

Note: This table presents a detailed breakdown of the pilot program's costs. Costs are calculated using the 6-week program budget information and pro rata basis. Hardware costs are estimated considering the depreciation incurred during the pilot. Electricity and school infrastructure costs are not included. Cost estimates for a 4-quarter program are based on the 6-week program. All costs are in current US dollars, with conversions from Naira based on an exchange rate of USD/Naira = 1,400. "Tutors and Monitors" include back-up staff.

Table 14: Program Implementation Timeline: Selected Activities

Category/Activity	Start	Completion
Planning and Setup		
Project kick-off meeting	-	3/27/24
Discussion sessions: Ministry of Education and Directors	4/15/24	4/18/24
Schools and labs' assessment	4/15/24	4/22/24
Staff and student selection	4/15/24	5/27/24
Student enrollment/ expression of interest to participate	4/26/24	5/27/24
Training materials and curriculum development	4/9/24	5/10/24
Teachers and monitors selection	4/22/24	5/03/24
Procurement of goods and services	4/23/24	5/15/24
Randomization (control and treatment groups)	5/27/24	5/28/24
Baseline questionnaire	5/30/24	6/1/24
Setup for training sessions	5/10/24	5/23/24
Pilot Implementation		
Teacher training sessions	5/24/24	5/26/24
Monitoring training sessions	5/28/24	5/29/24
Pilot sessions	6/3/24	7/11/24
Pilot monitoring	6/3/24	7/11/24
Pilot Evaluation and Report		
Standardized assessment	7/11/24	7/12/24
Third term examination	7/12/24	7/12/24
Endline questionnaire	7/14/24	7/14/24
Evaluation analysis	7/15/24	8/30/24
Reporting and presentation	8/30/24	10/30/24

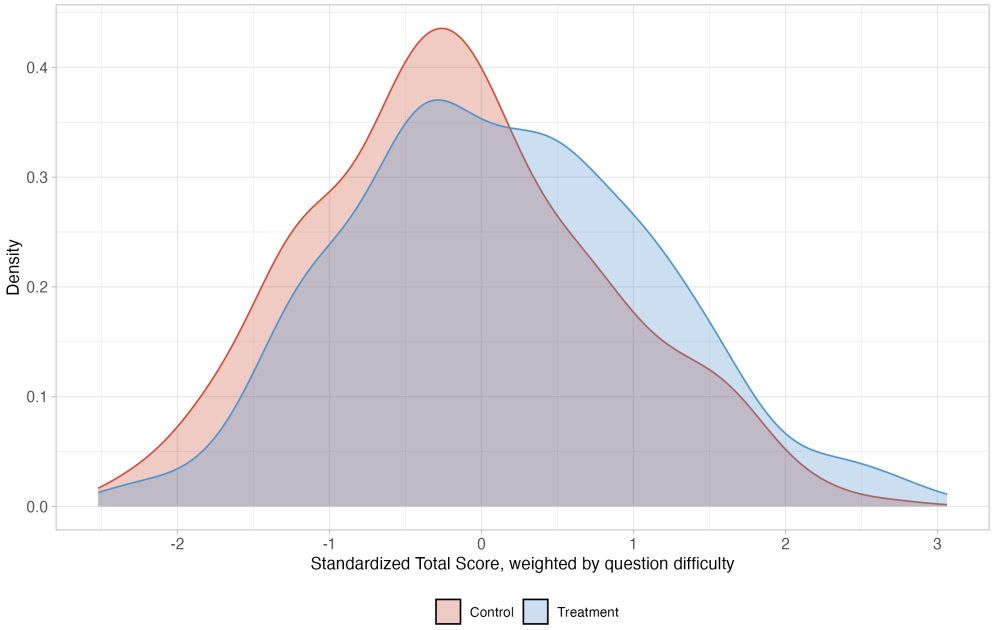

Figure 1: Distribution of Assessment Scores (combined) by Treatment Condition.

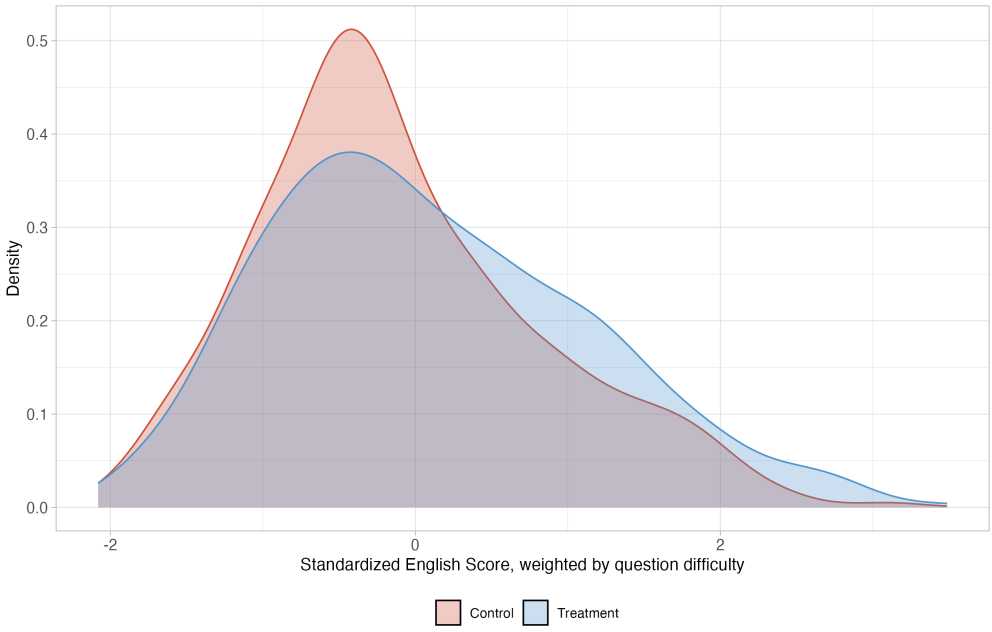

Figure 2: Distribution of English Scores by Treatment Condition.

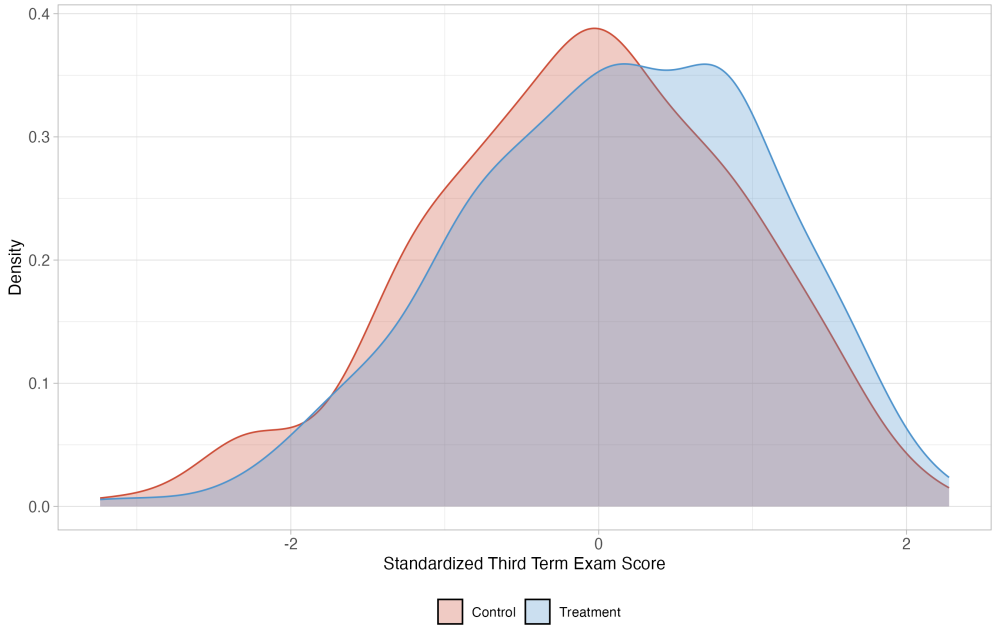

Figure 3: Distribution of Third Term Exam Scores by Treatment Condition.

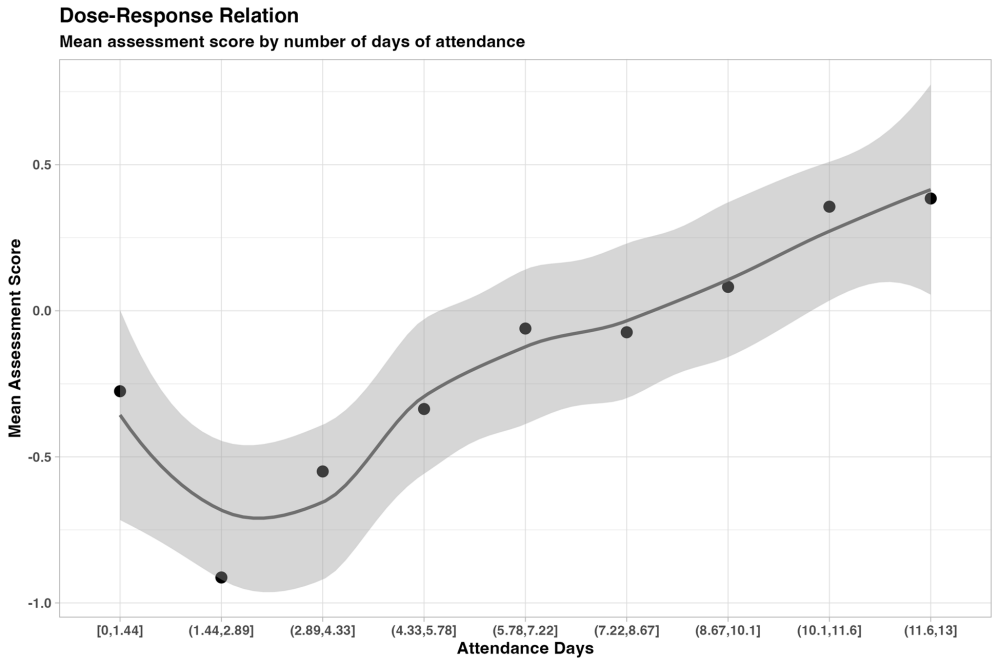

Figure 4: Dose-Response Relation.

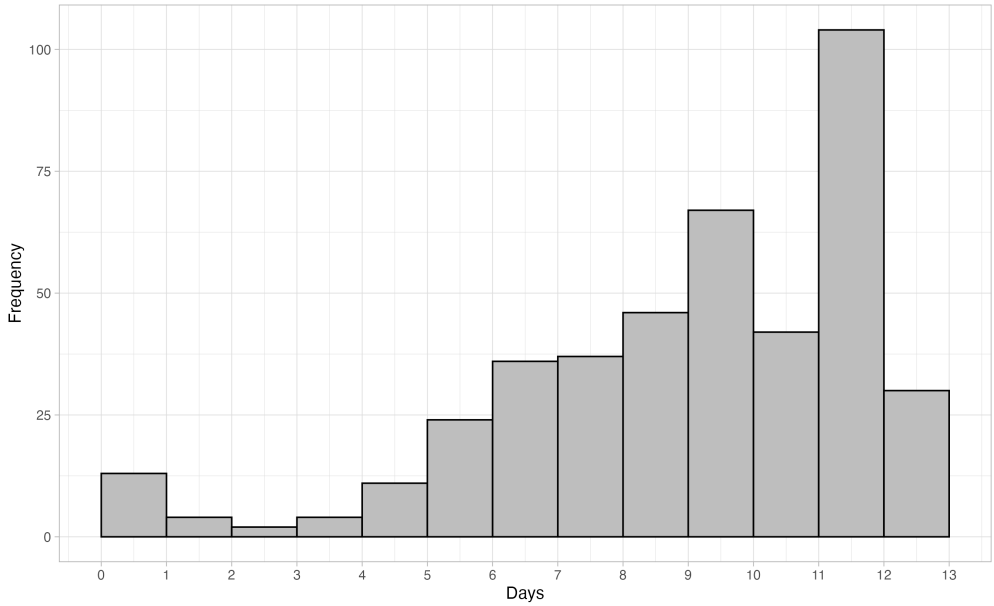

Figure 5: Distribution of the number of days of attendance to the program in the treatment group.

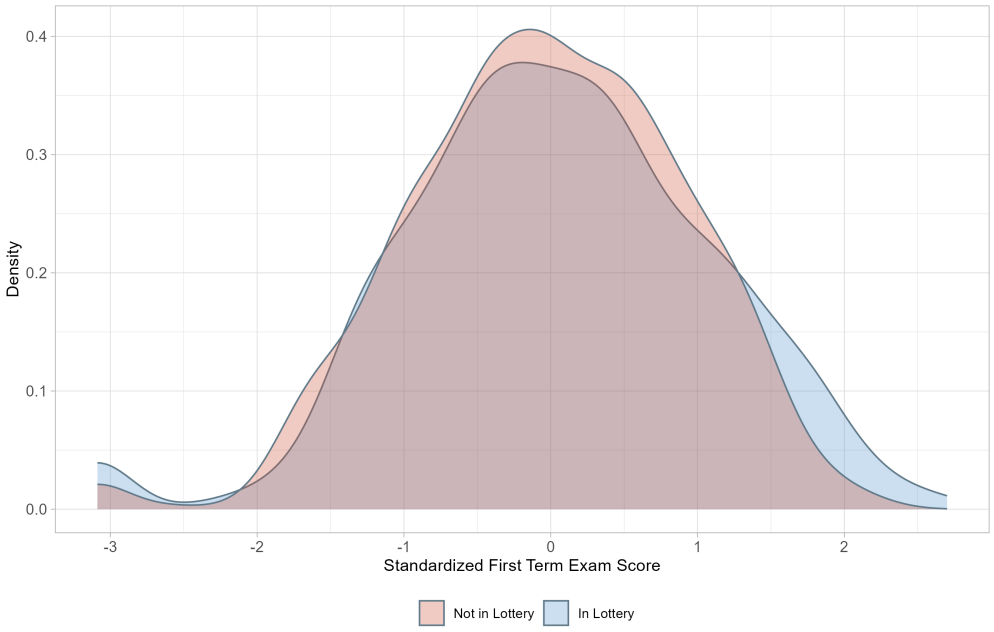

Figure 6: Distribution of first term exam scores by lottery eligibility.

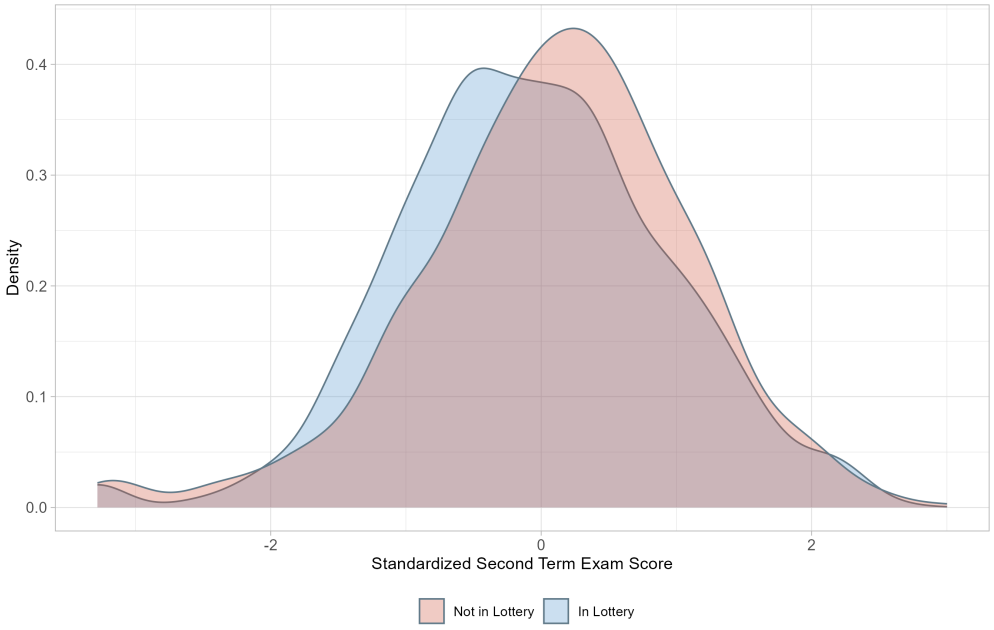

Figure 7: Distribution of second term exam scores by lottery eligibility.

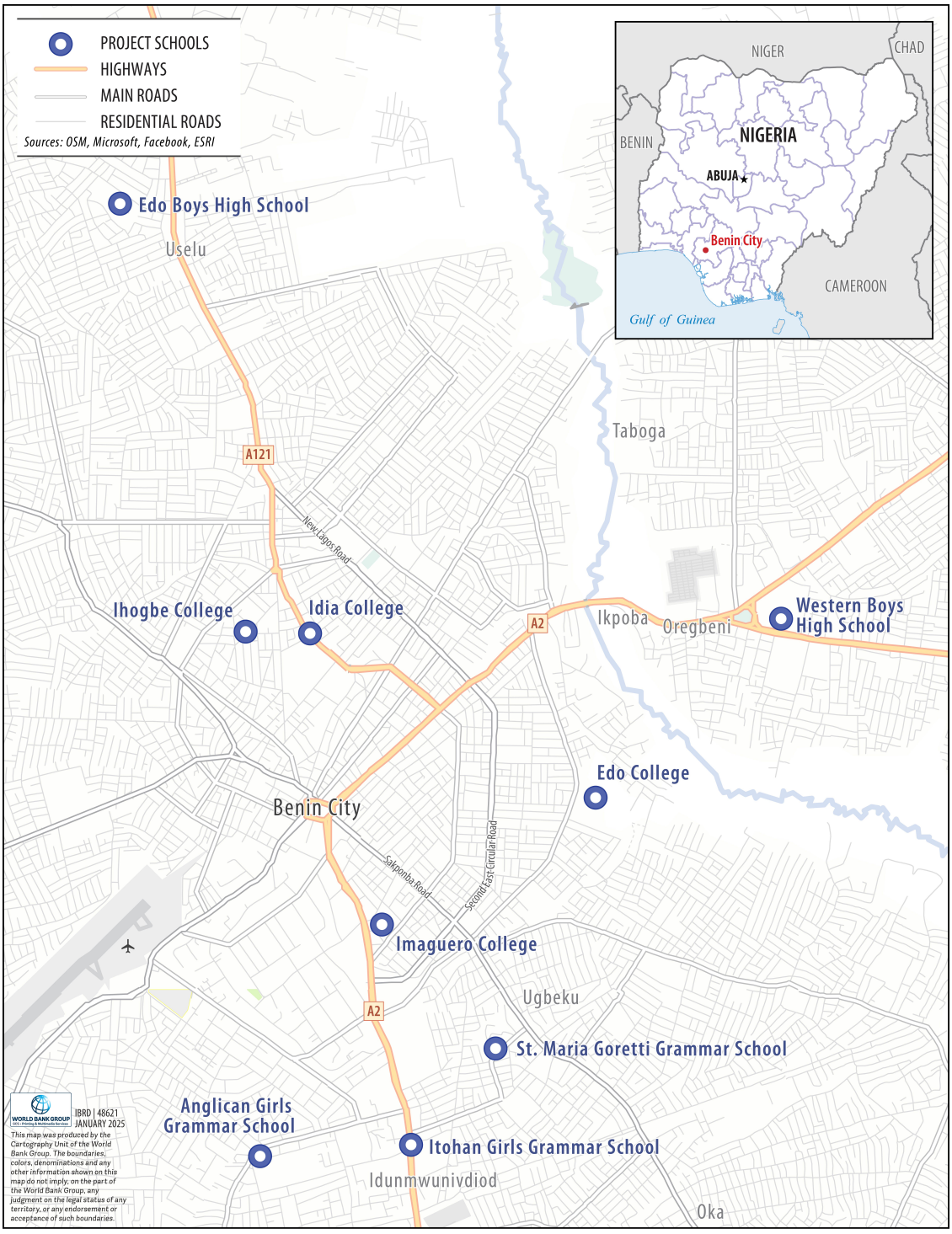

Figure 8: Locations of the Schools in the Sample.

A Appendix: Images from the after-school sessions and session guidelines



Figure 9: Students participate in an after-school session in Benin City, Nigeria.



Figure 10: Students participate in an after-school session in Benin City, Nigeria, in a female-only school.



Hello my friend!

We are **Eki** and **Odion**, your AI mascots, you will see us appear across this guide to give you some tips to have a fun learning experience using Generative Artificial Intelligence (GenAI).

We were developed using Copilot, a generative artificial intelligence (GenAI) tool, which you will be using weekly in the computer labs of your school.

Our first tip is that Copilot can produce both text and image according to your prompt (what you ask Gen AI to do when you type). We were developed with the following prompt: "*Create an image depicting a friendly mascot for a program to use AI in the classroom among senior secondary school students in Benin City, Nigeria. Provide male and female mascot options dressed in traditional clothing from the State of Edo and depict them holding books and/or a football.*" You can try modifying this prompt and developing your own mascot.

Let the fun begin!

1. What is Artificial Intelligence (AI)?

You are growing up in a world where artificial intelligence (AI) informs our daily lives, from photo filters, use of social media apps, creation of text and images with a single prompt, driving cars, and face recognition. But what is AI exactly?

In very simple terms, **Artificial Intelligence refers to programs or machines that simulate tasks that typically require human intelligence**.

8. **Congratulations! You've created an Outlook email account!**
9. Finally, the page will take you automatically Copilot again https://copilot.microsoft.com/ and you will be signed in with the user you created. You are all set! 😊 In case you are not automatically redirected you can enter Copilot again and log in with your credentials (email + password). Please always remember this information as it will help you for the rest of this program,.

Annex 3 : Prompt Library and Sample Questions

The program will take place between weeks 7 to 11 of the third term for first-year senior secondary students. Here you will find the collection of prompts that will be used during this period. As mentioned earlier, the sample questions are included to spark your imagination, feel free to ask different questions after introducing the starting prompt to Copilot. On week 11, teachers will indicate which prompt to use as it is a revision week.

Lesson Plan Week 7- Use of Clauses

Step 1: Introduce the starting prompt, after logging into your Copilot account https://copilot.microsoft.com. Set Copilot into creative mode.

Starting prompt

Good morning/afternoon I am {name/a student} of grade 1 of senior secondary education in Benin city, Nigeria. I would like to ask you to support me acting as a well-seasoned English grammar tutor to enhance my learning. My teacher {whose name is...} will be supervising this exercise, we are testing using Copilot to complement English lessons. We are currently studying how to use clauses. I will need you to reply to my questions in a motivational and engaging tone, your delivery should be clear and descriptive, yet to the point (using paragraphs and bullets). After providing an answer with links to sources of information and checking for understanding I will like you to follow up asking me a question or proposing an exercise on this topic. Wait for my answer, and if the answer is correct celebrate my progress, if it is incorrect provide encouraging words and provide a hint to arrive at the correct answer, which may include multiple options, if my reply is still incorrect you may provide the answer offering an explanation. After checking for my understanding continue with another question or exercise. Follow this model for the whole interaction. Please confirm that you have understood the task.

Step 2: You have the option to ask questions to Copilot alongside your computer partner, you can ask any question you'd like as long as they are related to the topic. Copilot will ask questions and pose exercises automatically but feel free to ask other questions, if you require more information or if you feel the questions are too simple or too difficult.

Some questions you may ask Copilot

- Can you create a sentence with an independent clause about the Benin Kingdom's history?

Figure 11: Samples of student guidelines.