Teaching with Gemini: Measuring the impact of Guided Learning on student mathematics progress in Sierra Leone

LearnLM Team, Google & Fab AI

We conducted a preregistered randomized controlled trial in Sierra Leone to evaluate the impact of the Guided Learning feature in Gemini on student learning outcomes in mathematics. The trial enrolled $N = 1763$ junior secondary school students across 48 mathematics classrooms. We asked teachers to incorporate Gemini into half of their weekly lessons for eight weeks, adding up to a requested total of 12 hours. Guided Learning significantly improved student learning outcomes, with an intent-to-treat effect of +0.258 SD ($p = 0.029$). Uptake exceeded our expectations, with 69.0% of students engaging with Guided Learning for at least 12 hours over the trial. Reaching this threshold yielded a significant treatment-on-the-treated effect of +0.380 SD ($p = 0.029$). In focus groups, teachers emphasized students' enthusiasm for Guided Learning sessions and credited Gemini with helping reveal new ways to teach familiar topics.

Keywords: education, efficacy, Gemini, artificial intelligence, randomized controlled trial, mathematics, Sierra Leone

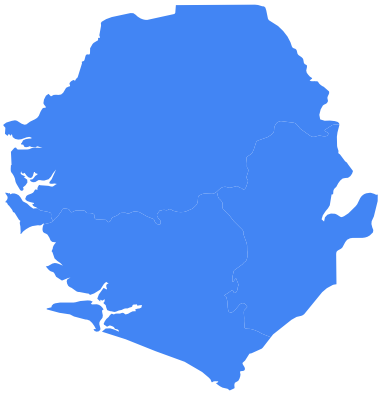
About this series

We are currently conducting a series of preregistered randomized controlled trials across multiple countries to understand the impact of the *Guided Learning feature in Gemini* on student learning. In these trials, Gemini forms a core part of structured classroom activities designed and led by teachers. This technical report describes the design and results from one of these trials. In the interest of advancing open science and disseminating timely insights, we will release brief technical reports as we progress through the trials. In the future, we also hope to write a holistic report synthesizing evidence across countries. For more information, see goo.gle/learnlm-impact.

Research question

How does the teacher-led integration of Gemini's Guided Learning feature into mathematics classes in Sierra Leone junior secondary schools affect student learning outcomes?

Country context



The education system in Sierra Leone has undergone substantial transformation in recent years. In 2018, the government began allocating over 20% of its annual budget to eliminate school fees for all students in the country [1]. This policy substantially expanded access to education and drove a dramatic surge in school enrollment [2, 3]. Access and enrollment, however, have not readily translated into improved learning outcomes. Qualified teachers remain scarce, particularly in rural areas [4]. Learning progress tends to fall short of intended benchmarks, and fewer than two-thirds of students ultimately complete junior secondary school [5, 6]. The government views technology as one means to close this gap and to support students and educators [7, 8]. Indeed, recent research suggests that when responsibly designed and deployed, tools like pedagogical artificial intelligence (AI) can provide effective, personalized support to students [9, 10]. With encouragement from the government, recent efforts have equipped

Corresponding author(s): learnlm-tech-report@google.com

hundreds of Sierra Leonean teachers with AI tools to assist lesson preparation and classroom instruction [11]. These initial inroads effectively position Sierra Leone to pioneer new research on pedagogical AI and the educational support it can provide—across Sub-Saharan Africa and beyond.

Design

We conducted a preregistered two-arm randomized controlled trial (RCT) across 12 government-supported junior secondary schools in Port Loko District, Sierra Leone. The trial enrolled $N = 1763$ students aged 13 or older in 48 grades 7 and 8 classrooms. We used clustered randomization for the trial with classrooms as clusters. That is, within each school and grade, we randomly assigned half of mathematics classrooms to use the Guided Learning feature in the Gemini app [12] and the other half to continue with standard instruction.

We trained all participating teachers for five to six hours over one day, covering device familiarization, an introduction to the Gemini app and to Guided Learning, guidance on preparing lessons using Guided Learning, and strategies for facilitating student use in the classroom. Control- and treatment-classroom teachers received the exact same training to avoid confounding effects from differential training. In particular, we trained teachers to deliver each Guided Learning lesson as a structured classroom activity, proceeding through four steps: the teacher introduces the learning objectives, assigns students to work in pairs with Gemini on the lesson content, leads a class discussion to consolidate their learning, and finally summarizes key takeaways. We also instructed teachers to prepare each lesson by defining learning objectives and crafting starter prompts and questions to scaffold students' interactions with Gemini.

We instructed treatment-classroom teachers to integrate Guided Learning (specifically using the Pro model) into half of their weekly mathematics lessons, making for a total of 12 hours of use over the course of the trial. In Sierra Leone, grades 7 and 8 classrooms hold four mathematics periods per week, so this amounted to two periods (90 minutes) per week for each classroom over eight weeks. Students accessed the Gemini app on tablets or desktop computers, sharing at a 2:1 student-to-device ratio. We instructed control-classroom teachers to continue with standard mathematics instruction (i.e., without incorporating the Gemini app into their lessons). To accommodate staggered starts across schools, the intervention itself ran for nine weeks total, from 6 October to 5 December 2025. A Gemini model update rolled out on 18 November [13], so in practice, students accessed Gemini 2.5 Pro [14] over the first six weeks of the trial and Gemini 3.0 Pro [15] over the final three weeks. In summary, we trained and equipped teachers with the Guided Learning feature in Gemini, with instructions to incorporate it into half of their math classroom time; the teachers then independently created and delivered their own Guided Learning lessons.

To insulate outcome measurement from potential bias [16], we contracted Oxford MeasurEd to develop and administer curriculum-aligned mathematics assessments at baseline and endline. Question difficulty inevitably varies across assessments, so raw test scores reflect both student learning progress and the difficulty of the specific questions included in the baseline and endline assessments [17]. Oxford MeasurEd used item response theory (IRT) to disentangle student performance from question difficulty (without knowledge of students' treatment assignment), producing scores on a common scale across baseline and endline. We also recorded and analyzed all student conversations with Gemini to assess the pedagogical quality of the Guided Learning sessions. To protect student privacy, we processed all conversational data via automated de-identification pipelines prior to analysis, removing personally identifiable information; the transcripts used in this analysis do not identify individual students. Finally, we carried out focus groups with the teachers who participated in the trial—and conducted inductive thematic analysis [18] with the resulting transcripts—to understand teachers' perspectives on Gemini, Guided Learning, the overall classroom intervention, and the trial itself.

Our trial protocol underwent independent ethical review and received approval from the Sierra Leone Ethics and Scientific Review Committee (No. 007/09/2025), as well as authorization from the Sierra Leone Ministry of Basic and Senior Secondary Education. We preregistered the trial protocol and analysis plan with the AEA RCT Registry (AEARCTR-0016651). Before beginning the trial, we collected parental consent, teacher consent, and student assent for all participating students.

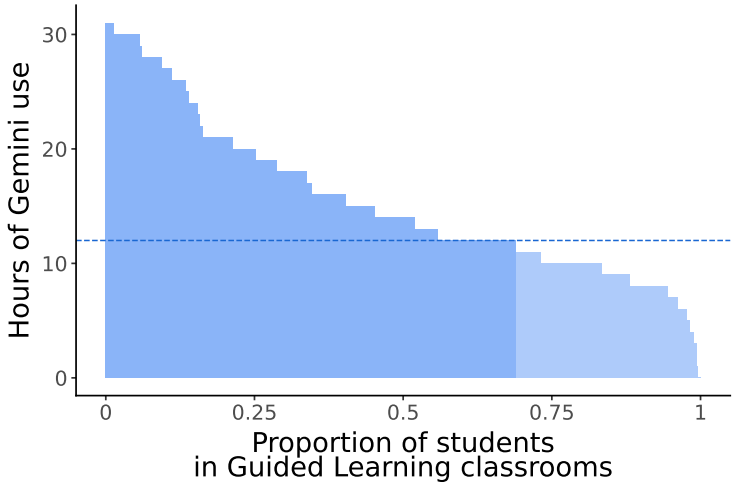


Figure 1. Uptake of Gemini exceeded our expectations. We asked teachers in Guided Learning classrooms to incorporate Guided Learning activities into half of their weekly classes, making for 12 hours of requested use total (dashed line). Remarkably, 69.0% of the 871 students in Guided Learning classrooms met this threshold, as indicated by the dark-blue shaded region. In fact, many teachers voluntarily integrated Gemini into their classes beyond the 12-hour mark, with some classrooms logging more than double that level of uptake.

Results

Uptake. We requested that classrooms use Guided Learning in half of their weekly lessons, for 12 hours of total use over the trial. Based on attendance records and corroborated by field notes, 69.0% of students met this threshold of Guided Learning use (Figure 1). In practice, treatment classrooms actually averaged approximately 15 hours of uptake over the trial (25% above the usage we requested).

Overall effect. Use of the Guided Learning feature in Gemini significantly improved students' learning outcomes, yielding a gain of +0.258 standard deviations across the Guided Learning classrooms (intent to treat; 95% confidence interval [0.027, 0.488], $p = 0.029$; Figure 2, left). To explain that result more concretely, a gain of 0.258 standard deviations is roughly equivalent to a student in the middle of their cohort—scoring at the 50th percentile on the baseline assessment—moving up to the 60th percentile on the endline assessment. Benchmarked against typical academic progress in low- and middle-income countries, that gain reflects roughly 1.2 to 1.7 years of extra learning progress [19]. This overall effect remains robust when adjusting for demographic covariates, with an estimated gain of +0.259 SD (95% CI [0.025, 0.494], $p = 0.031$).

Dosage. Overall, the more that students engaged with Guided Learning, the more they benefited. Each additional hour of Guided Learning use improved learning outcomes by +0.016 SD (treatment on the treated; 95% CI [0.002, 0.031], $p = 0.026$). Of course, not all classrooms reached the full 12 hours of use that we requested. When we account for variation in student uptake, completing 12 hours of Guided Learning activities (as intended) improved student learning by +0.380 SD (treatment on the treated; 95% CI [0.040, 0.719], $p = 0.029$). The per-hour and intended-dosage effects remain significant when adjusting for demographic covariates (per hour: +0.017 SD, 95% CI [0.002, 0.031], $p = 0.028$; intended dosage: +0.382 SD, 95% CI [0.035, 0.729], $p = 0.032$).

Heterogeneity. Students who entered the trial with stronger mathematics skills benefited more from Guided Learning. For each additional standard deviation of mathematics proficiency a student demonstrated at baseline, the treatment effect increased by +0.195 SD (95% CI [0.074, 0.315], $p = 0.002$; Figure 2, right). We analyze potential variation in the impact of Guided Learning across other student characteristics in the appendix.

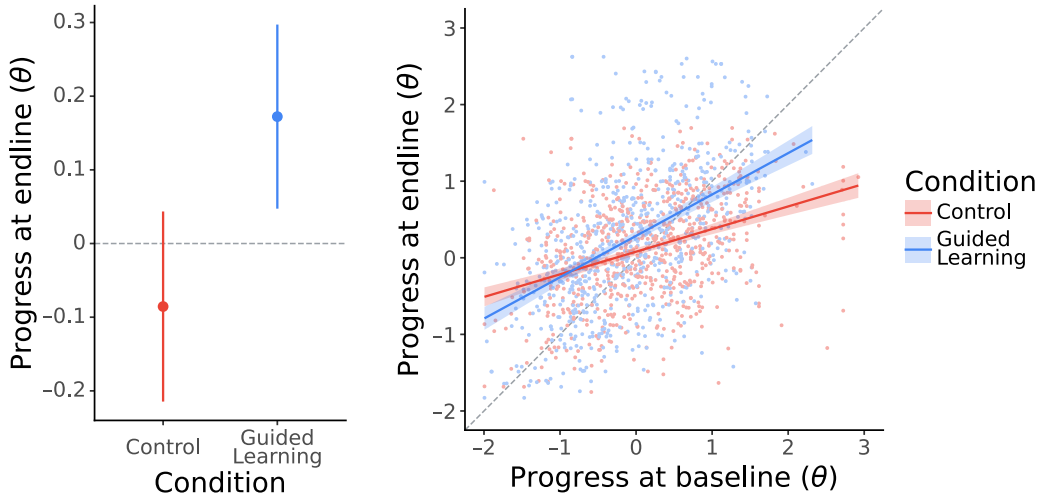


Figure 2. Studying with the Guided Learning feature in Gemini significantly improved learning outcomes for students, with an intent-to-treat effect of +0.258 SD ($p = 0.029$). We use item response theory to map student scores on the baseline and endline assessments onto a common scale of mathematics progress (θ). On the left, points and error bars indicate estimated marginal means and 95% confidence intervals, respectively. On the right, the scatterplot shows individual performance at baseline and endline. The solid lines reflect the relationships that we estimate, with the error bands indicating their 95% confidence intervals. The dashed line represents the identity line (i.e., no change from baseline to endline).

Student interactions. Conversations with Gemini demonstrated high pedagogical quality. In total, students and Gemini exchanged 113,344 messages over 7,421 conversations throughout the trial. These interactions remained remarkably on-task, with on-topic conversations accounting for 97.4% of all messages. Looking at student behavior in these interactions, we see that students seldom focused on seeking direct answers (5.0% of conversations by volume), and predominantly directed their efforts toward developing their understanding and skills (91.4%). In turn, Gemini supported students by posing scaffolding questions (76.4% of its messages) far more frequently than providing direct solutions (2.1%).

Teacher perspectives. In focus groups following the trial, teachers voiced largely positive experiences with Gemini and their Guided Learning lessons. Two themes particularly stood out: students' excitement for Guided Learning, and the value of Gemini for teachers' professional growth. Across every focus group, teachers discussed their classes' enthusiasm for Guided Learning sessions, reporting that students would rush to class and press for additional time with Gemini. Teachers also consistently credited Gemini with expanding their own knowledge and helping them discover new ways to teach familiar topics. At the same time, teachers flagged several classroom challenges (including gaps in traditional and digital literacy), and requested additional time and access to help students fully benefit from the tool. We provide the full set of themes and insights from our analysis in the appendix.

Reflections

This report represents the first dispatch from a series of randomized controlled trials investigating the efficacy of the Guided Learning feature in Gemini. For these reports, we would like to offer brief, honest reflections—surprises, lessons learned, areas to improve—and defer any extensive interpretation until we build a broader base of evidence.

We were quite surprised to see such high levels of uptake across the classrooms using Guided Learning. A central challenge when developing interventions—not only for education, but also for health, finance, and agriculture—is simply encouraging adoption among the intended users. Educational technologists and development economists alike frequently find that voluntary take-up of new tools or programs stubbornly hovers around five percent [20, 21]. So we entered this trial with relatively modest expectations, and did not

anticipate that 69.0% of students would ultimately adhere to our guidelines for sessions with Guided Learning. We similarly did not anticipate that several classrooms would go on to use Gemini for several hours more per week than we requested. To us, these indicators of engagement carry just as much promise as the intervention's significant and positive effects on assessment scores.

We conducted this trial on a much faster timescale than typical impact research (cf. [22, 23]). Educators, school administrators, and policymakers currently operate with very limited evidence to help guide their decisions on AI. Should classrooms adopt AI tools? If so, how can they most effectively integrate them? Which opportunities and risks most deserve teachers' attention? Shorter trials can provide timely answers to these questions and begin charting evidence-based paths through the wild west of AI and education. In the spirit of producing practical evidence, we also chose not to narrowly focus this trial on assessing the capabilities of a specific generation of Gemini. As Gemini improved, students and teachers simply used the latest version available, just like any classroom would. We believe that this sort of rapid, realistic trial will ultimately complement and strengthen the longer-term research needed to study sustained effects of AI on education.

Crucially, the intervention in this trial is not a technological fix [24]. It is an integrated approach [25] that weaves together pedagogical structure, teacher design and direction, peer interactions between students, and pedagogical AI. The teachers played an especially key role. They designed and aligned lessons with the curriculum, adapted activities as their students' needs shifted, and provided the relational foundation needed to foster learning. These contributions were as essential to students' engagement and improvement as the opportunities introduced by Gemini. Looking forward, we expect educators to play just as vital a role in any efforts to scale or build upon these results.

In this trial, studying with Gemini conferred the greatest benefits to students who started with strong mathematics skills. This is a common pattern in educational technology: new tools frequently widen achievement gaps rather than close them [26]. But our ambition is the opposite. We hope to develop pedagogical AI that helps all students, and that delivers particularly strong gains for students who start behind. The latter will not come for free from improving AI capabilities like accuracy, explanation, or personalization. Addressing this challenge will require deliberate work on the pedagogical approach, activity design, and relational aspects of our classroom interventions—guided by specific theories of change for reaching students who start behind. As a starting point, we suspect that motivation, digital literacy, and reading literacy will be key barriers to address. To make progress here, we intend to move our future trials beyond aggregate analysis and link specific tutoring moves and student experiences directly to learning outcomes.

Overall, our goal with this research is to identify the most promising opportunities and the most important limitations of pedagogical AI, so that it can *meaningfully support students and educators*. We wish to answer practical questions. If a teacher wants to incorporate pedagogical AI into their classes to materially help their students, what should they do? If a school wants to support its teachers, what training and support should it provide? If a policymaker wants to ensure that they pedagogical AI helps the students who need it most, what frameworks should they put in place? Collaboration between educators, learning scientists, field researchers, AI experts, and local communities will be essential to discovering new pathways to improve learning outcomes for all students.

References

[1] Desmond Bermingham, Myra Harrison, Florence Malinga, and Susy Ndaruhutse. Report of the Provisional Independent Technical Advisory Panel (ITAP) Assessment of Enabling Factors: Sierra Leone. Technical report, Global Partnership for Education, July 2022. URL https://www.globalpartnership.org/library/assessment-enabling-factors-sierra-leone-july-2022.

[2] GPE Secretariat. Sierra Leone: More efficient and equitable education financing. Global Partnership for Education, April 2025. URL https://www.globalpartnership.org/blog/sierra-leone-more-efficient-and-equitable-education-financing.

[3] Jack Rossiter and Might Kojo Abreh. Sierra Leone has made a big bet on free education for poor children—so long as they can pass the exams. Center for Global Development, August 2022. URL

https://www.cgdev.org/blog/sierra-leone-has-made-big-bet-free-education-poor-children-so-long-they-can-pass-exams.

[4] Emma Cameron, Marcela Gutierrez Bernal, Mari Shojo, Namrata Raman Tognatta, Md. Mokhlesur Rahman, Ali Ansari, and Afra Chowdhury. Sierra Leone – Data-driven approach to teacher deployment. Working Paper 203151, World Bank Group, July 2024. URL http://documents.worldbank.org/curated/en/099452307142536307.

[5] Roberta V. Gatti, Paul Andres Corral Rodas, Nicola Anna Pascale Dehnen, Ritika Dsouza, Juan Elias Mejalenko, and Steven Michael Pennings. The human capital index 2020 update: Human capital in the time of COVID-19. Working Paper 152967, World Bank Group, January 2021. URL http://documents.worldbank.org/curated/en/456901600111156873.

[6] Lidiya Tefera and Quentin Wodon. Sierra Leone: Trends in completion rates by education level and gender. Data Brief 2025-33, UNESCO International Institute for Capacity Building in Africa (IICBA), December 2025.

[7] Ministry of Basic and Senior Secondary Education (MBSSE). Sierra Leone Education Sector Plan 2022–2026: Transforming Learning for All. Technical report, Government of Sierra Leone, September 2022.

[8] World Bank. National digital learning strategy for the Sierra Leone Ministry of Basic and Senior Secondary Education and Teaching Service Commission 2025-2030. Report 204306, World Bank Group, August 2025. URL http://documents.worldbank.org/curated/en/099601408042527956.

[9] Martin Elias De Simone, Federico Hernan Tiberti, Maria Rebeca Barron Rodriguez, Federico Alfredo Manolio, Wuraola Mosuro, and Eliot Jolomi Dikoru. From chalkboards to chatbots: Evaluating the impact of generative AI on learning outcomes in Nigeria. Policy Research Working Paper 11125, The World Bank, 2025.

[10] LearnLM Team, Google and Eedi. AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms. *arXiv preprint arXiv:2512.23633*, 2025.

[11] Kabiru Mansaray, Foday Kalokoh, Miriam Mason-Sesay, and Abby Couralis. How AI is transforming education in Sierra Leone. EducAid Sierra Leone, March 2025. URL https://www.educaid.org.uk/how-ai-is-transforming-education-in-sierra-leone/.

[12] Maureen Heymans. Guided Learning in Gemini: From answers to understanding. The Keyword, August 2025. URL https://blog.google/products-and-platforms/products/education/guided-learning/.

[13] Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. A new era of intelligence with Gemini 3. Google Blog, November 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-3/.

[14] Google DeepMind. Gemini 2.5 Pro model card. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Pro-Model-Card.pdf.

[15] Google DeepMind. Gemini 3 Pro model card. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

[16] Alan C. K. Cheung and Robert E. Slavin. How methodological features affect effect sizes in education. *Educational Researcher*, 45(5):283–292, 2016.

[17] Ronald K. Hambleton, Hariharan Swaminathan, and H. Jane Rogers. *Fundamentals of item response theory*, volume 2. SAGE, 1991.

[18] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[19] David K. Evans and Fei Yuan. How big are effect sizes in international education studies? *Educational Evaluation and Policy Analysis*, 44(3):532–540, 2022.

[20] Laurence Holt. The 5 percent problem. *Education Next*, 24(4):26–31, 2024.

[21] Gabriel Lara Ibarra, David McKenzie, and Claudia Ruiz-Ortega. Estimating treatment effects with big data when take-up is low: An application to financial education. *The World Bank Economic Review*, 35(2): 348–375, 2021.

[22] Katharine M. Conn. Identifying effective education interventions in sub-Saharan Africa: A meta-analysis of impact evaluations. *Review of Educational Research*, 87(5):863–898, 2017.

[23] Patrick J. McEwan. Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research*, 85(3):353–394, 2015.

[24] Henrik Skaug Sætra. *Technology and sustainable development: The promise and pitfalls of techno-solutionism*. Taylor & Francis, 2023.

[25] Fred E. Emery and Eric L. Trist. Socio-technical systems. *Management Science, Models and Techniques*, 2: 83–97, 1960.

[26] Andrew A. Tawfik, Todd D. Reeves, and Amy Stich. Intended and unintended consequences of educational technology on social inequality. *TechTrends*, 60(6):598–605, 2016.

[27] David Hayes. Cascade training and teachers' professional development. *ELT Journal*, 54(2):135–145, 2000.

[28] Anna Popova, David K. Evans, Mary E. Breeding, and Violeta Arancibia. Teacher professional development around the world: The gap between evidence and practice. *The World Bank Research Observer*, 37(1): 107–136, 2022.

[29] Elizabeth G. Cohen. Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research*, 64(1):1–35, 1994.

[30] Education Endowment Foundation. Collaborative learning approaches. Education Endowment Foundation, July 2021. URL https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit/collaborative-learning-approaches.

[31] Jill Denner, Linda Werner, Shannon Campe, and Eloy Ortiz. Pair programming: Under what conditions is it advantageous for middle school students? *Journal of Research on Technology in Education*, 46(3): 277–296, 2014.

[32] Google DeepMind. Gemini 3.1 Flash Lite model card. Technical report, Google DeepMind, 2026. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Flash-Lite-Model-Card.pdf.

[33] Allan Wigfield and Jenna Cambria. Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, 30(1):1–35, 2010.

[34] Michelene T. H. Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G. Hausmann. Learning from human tutoring. *Cognitive Science*, 25(4):471–533, 2001.

[35] Greg Guest, Kathleen M. MacQueen, and Emily E. Namey. *Applied thematic analysis*. SAGE, 2011.

Contributions and Acknowledgments

Core Contributors

The following individuals contributed to the work described in this report. These lists are ordered alphabetically, and do not indicate ranking of contributions.

On the Google team, the following individuals made core contributions:

Albert Wang, Andrea Huber, Asha Johnson, Brian Veprek, Daniel Gillick, Eleni Sgouritsa, Garrett Honke, Irina Jurenka, Kaiz Alarakyia, Katie Stasaski, Kevin R. McKee, Khyati Jain, Markus Kunesch, Srinithi Gopinath, Yilmazcan Ozyurt, and Yusra Ibrahim.

On the Fab AI team, the following individuals made core contributions:

Ignacio Roman Muñoz, Jared Lee, María José Ogando Portela, Natalia Valdes Aspillaga, Paul Atherton, Tomáš Koutecký, and Usman Khawar.

As local implementation, data collection, and assessment development partners for this trial, Erin Northley, Ibrahim Alhaji Fortune, Lansana Bakarr and Miriam Mason (at EducAid), Ayumi Uchiyama, Ishmail Kamara and Kelsey Julianna Hunt (at Laterite), and Paulina Valenzuela and Rachel Outhred (at Oxford MeasurEd) also made core contributions in collaboration with the Fab AI team.

Kevin R. McKee led the preparation of this report.

Acknowledgements

This work represents a close collaboration between Google and Fab AI.

For Fab AI: We would like to thank Ana Paola Ramirez and Orla Humphries from the Fab AI team for their contributions to data quality assurance. We are also grateful to EducAid for fielding a dedicated team of 12 monitors who provided on-site support throughout the implementation: Abu Bakarr Bangura, Abu Bakarr Turay, Abubakarr Jalloh, Amadu Conteh, Betty Caulker, Idrissa Abass Bangura, Isatu Salia, Kadijah M. Turay, Mohamed Ansumana, Mohamed F. Bangura, Queen Kaday Mambu, and Yambom S. Kanu. We would also like to extend our sincere thanks to Ahmad Jawad Asghar and Benjamin Piper at the Gates Foundation for their support and sponsorship of this work.

For Google: We completed this work as part of the LearnLM effort—a cross-Google project, with members from Google DeepMind, Google Research, Google LearnX, and more. This tech report represents only a small part of the wider effort, and only lists team members who made direct contributions to this report. The dedication and efforts of numerous teams make our work possible. We would like to acknowledge support from Aakash Kaku, Aliya Rysbek, Amy Sims, Antonia Mould, Avishkar Bhoopchand, Darryl Wright, Dave Messer, Diana Carreno, Filip Bar, Gabi Hill, HJ Choe, Jennifer Shen, Jonathan Caton, Josh Caldwell, Julia Wilkowski, Kristen Morea, Liam McCafferty, Mana Jabbour, Maritza Handal, Matt Dawson, Miriam Schneider, Miruna Pîslar, Nupur Jain, Rachel Hashimshoni, Rishabh Trivedi, and Shanice Onike. Furthermore, we would like to thank the Gemini team, the Gemini app team, and Google LearnX. Finally, we would like to acknowledge the support provided by our leads and sponsors, who made this project possible: our genuine thanks go to Benedict Gomes, Chris Phillips, Lila Ibrahim, Shakir Mohamed, and Zoubin Ghahramani.

We gratefully acknowledge Google.org for their support, which made this work possible.

This work was supported, in whole or in part, by the Gates Foundation (INV-078363). The conclusions and opinions expressed in this work are those of the author(s) alone and shall not be attributed to the Foundation. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. Please note works submitted as a preprint have not undergone a peer review process.

A. Partners

Google DeepMind (https://deepmind.google/) is a world-leading AI research lab committed to building AI responsibly for the benefit of humanity, and spearheaded the overall research collaboration.

Fab AI (https://www.fab-ai.org/) works to shape the world's best technologies for those learning the least, and led the overall impact evaluation for this trial.

EducAid (https://www.educaid.org.uk/) is a UK-registered charity with over 30 years of experience running free schools and education programmes in Sierra Leone, and led local implementation for this trial.

Laterite (https://www.laterite.com/) is an Africa-based research consultancy specializing in rigorous, tech-enabled data collection and social impact research, and led data collection for this trial.

Oxford MeasurEd (https://www.oxfordmeasured.co.uk/) is an education research consultancy specializing in psychometrically rigorous, context-sensitive learning assessment design, and led assessment development for this trial.

B. Design

We conducted the trial in 12 government-supported junior secondary schools in Port Loko District, Sierra Leone during the second term of the 2025 academic year. The trial enrolled $N = 1763$ students aged 13 years or older across 16 grade 7 classrooms and 32 grade 8 classrooms. The intervention ran for nine weeks, from 6 October to 5 December 2025. We preregistered this trial protocol with the AEA RCT Registry (AEARCTR-0016651).

In Sierra Leone, the mathematics curriculum for grade 7 covers numbers and numeration (number patterns, fractions, place value), everyday arithmetic (operations, ratios, percentages), geometry and measurement (angles, area, perimeter), and introductory algebra (expressions and simple equations). The mathematics curriculum for grade 8 revisits the same topics and subdomains, albeit with greater depth.

Trial implementation began with teacher training. We adopted a cascade approach to training [27, 28]. That is, we delivered an initial five-hour training session to a cohort of teacher trainers recruited by EducAid, preparing them to deliver the trial's technical and instructional program. The teacher trainers then trained all participating mathematics teachers. Each participating teacher went through a single-day session, comprising roughly five to six hours of material. These sessions familiarized teachers with the trial devices, introduced them to the Gemini app and Guided Learning, coached them on lesson preparation, and worked through approaches to facilitating student use in the classroom. Because we conducted training before classroom randomization, in effect, both treatment- and control-classroom teachers attended the same training.

After training the participating teachers, we conducted randomization. Then, within each school and grade, we randomly assigned half of the participating classrooms to use Guided Learning (treatment classrooms) and the other half to continue with standard instruction (control classrooms).[1] This resulted in an allocation of nine grade 7 and 15 grade 8 classrooms to the treatment arm, and seven grade 7 and 17 grade 8 to the control arm. At this point, we informed teachers of the arms to which we had assigned their classrooms. Due to the classroom-level randomization, some teachers taught only treatment classrooms, some taught only control classrooms, and many taught classrooms in both arms. Randomizing at the classroom level rather than the school level improves statistical power, though at the cost of a higher risk of spillovers from treatment to control classrooms. In addition to the steps we took to minimize contamination (described in the following paragraphs), we note that any spillovers would actually narrow the observed differences between arms, rather than inflate them.

EducAid recruited and trained local staff to act as field monitors, responsible for supporting operations, ensuring implementation fidelity, and guarding against spillover throughout the trial. We stationed one field monitor at each of the 12 schools for the duration of the trial. To support classroom operations, field monitors charged and prepared devices before lessons and assisted teachers with technical issues. They tracked implementation integrity and spillover by logging incidents such as teacher absences, lesson cancellations, use of Gemini in control classrooms, as well as by submitting reports tracking delivery of Guided Learning lessons. Field monitors also managed the consent process by distributing parental consent information, collecting completed consent forms, and confirming that only students whose parents provided consent participated in trial activities. To mitigate observer effects, we instructed monitors to generally stay outside of classrooms, entering only to handle technical issues and leaving once resolved.

We worked closely with the field monitors to track the overall progress of trial implementation. Throughout the trial, the monitors escalated any urgent problems directly to us so that we could help address them. We also held two scheduled reviews with the full cohort of monitors to check in and address any open questions on their minds. Outside of the (infrequent) escalations, our protocol included no other intervention for teachers (e.g., no ongoing coaching) to ensure the trial reflected a realistic, teacher-led implementation.

Overall, with this trial, we sought to evaluate the efficacy of the Guided Learning feature with the Pro model in the Gemini app. To reflect realistic deployment conditions, we maintained the standard update schedule for the Gemini app during the trial. Following this schedule, the Gemini app routed students to Gemini 2.5 Pro [14] for the first six weeks of the trial and Gemini 3.0 Pro [15] for the final three weeks. For this trial, we configured all student accounts to operate exclusively with the Guided Learning feature enabled.

[1] For this trial, we restricted Gemini use to students aged 13 or older. During Guided Learning sessions in treatment classrooms, we planned to move any students under the age of 13 to other mathematics classrooms. Consequently, where schools lacked meaningful capacity to reassign students, we excluded classrooms with high proportions of younger students from the trial.

We instructed treatment-classroom teachers to incorporate Guided Learning into their instruction for two of their four weekly mathematics periods—totaling to 90 minutes of Guided Learning activities per week. In terms of pedagogical design, teachers learned during their training to deliver Guided Learning lessons through a structured four-part framework: first, a teacher-led introduction to establish the learning objectives and check students' prior knowledge; second, a main activity where students worked in pairs with Gemini; third, a consolidation phase where the class discussed and reviewed what they had learned; and fourth, a plenary to summarize key takeaways. The training taught teachers that each of these Guided Learning lessons required preparation in advance. In particular, teachers needed to define clear learning objectives, write the starter prompts that students would type into Gemini, and draft question stems to write on the classroom chalkboard as scaffolding to facilitate interactions between students and Gemini. The training also showed teachers how to embed educational and regional context into prompts (for example, specifying the student's grade level and the setting of Sierra Leone) to help Gemini ground and appropriately level its responses.

Our pedagogical design required students to work in pairs, each sharing a single tablet or desktop computer [29, 30]. The training instructed teachers to assign "driver" and "navigator" roles [31] to facilitate learning within each pair of students, with roles swapping each lesson. The driver typed, while the navigator took notes and helped think through questions. Device scarcity or logistical constraints occasionally prevented pairing. In these rare instances, the teacher would group three students to share a device—with two students acting as navigators, and the remaining as the driver. After assigning roles and getting students started on the lesson, teachers circulated among the classroom to check that students remained on task and to support pairs struggling to formulate effective questions. When technical issues arose, teachers called in field monitors to resolve them.

We explicitly instructed teachers not to use Gemini or any Gemini outputs outside the designated mathematics lessons. Control-classroom teachers continued with standard mathematics instruction, without incorporating the Gemini app. To assess spillover risks, we reviewed field monitor reports for evidence of spillover across multiple channels, including students in control classrooms using Gemini, teachers bringing treatment-classroom approaches into control classrooms, and students passing content from Guided Learning lessons across arms. Field monitors documented no evidence of spillover throughout the trial, aside from two isolated incidents in which a student from a control classroom sat in on part of a Guided Learning lesson.

We contracted Oxford MeasurEd to develop and administer curriculum-aligned mathematics assessments at baseline and endline. The parallel structure of the grade 7 and grade 8 curricula—the same topics at two levels of complexity—allowed Oxford MeasurEd to design shared assessment instruments for both grades. Oxford MeasurEd designed the assessment materials so that half of the assessment items covered curricular content from the intervention period, and the remaining half covered broader concepts from the curricula for grades 7 and 8, along with foundational knowledge from earlier grades. This approach balanced curriculum alignment with broader measurement of mathematical skills. Oxford MeasurEd designed an additional assessment measuring English reading comprehension to administer at baseline. The scores from this assessment allowed us to investigate the interplay of student reading proficiency with the effectiveness of Guided Learning.

Oxford MeasurEd scored the assessments using item response theory (IRT), a standard psychometric method that estimates student mastery independently of the difficulty of the specific questions on an assessment [17]. For context, the raw score from an assessment (e.g., the percentage of questions a student answered correctly) inherently reflects not just student mastery, but also the difficulty of the specific items on the assessment. On an incredibly difficult assessment, a well-prepared student might only answer half of the questions correctly, whereas an easy assessment might allow a poorly prepared student to receive a perfect score. This prevents meaningful comparisons of raw scores from one assessment to another. In simple terms, IRT resolves this challenge by placing both students and test questions on a shared interval scale (i.e., a scale in which a one-unit change represents a consistent difference in mastery, regardless of the exact starting or ending points). On this scale, IRT locates items according to their difficulty and places students according to their response patterns. The interval nature of this scale provides the mathematical foundation required to link different assessments together and enable meaningful comparisons of student progress over time. To prepare the trial data for our analysis, Oxford MeasurEd applied IRT to the pooled data from all students, without access to treatment assignment.

To better understand the dynamics underlying the Guided Learning sessions, we analyzed de-identified student transcripts from the trial. As an early observation, we noticed that students occasionally continued

existing chat threads (across multiple days) rather than starting a new thread for each Guided Learning session. As a consequence, we split thread transcripts into separate conversations by a 20-minute inactivity gap. We then conducted the following analyses using Gemini 3.1 Flash-Lite [32]. We classified messages as on-topic if they addressed the mathematics topics defined in the curriculum for the trial period, and off-topic otherwise. We then applied two taxonomies to assess pedagogical quality. To understand the way that students engaged with Guided Learning sessions, we classified student conversations by the different goal orientations they reflected (i.e., the overarching aims they bring to the interaction [33]). Then, to understand the way that Gemini responded to students, we classified Gemini's responses into different tutoring moves (i.e., instructional actions by a tutor [34]). We report the results of each classification analysis as a percentage of messages exchanged or sent, such that they reflect the prevalence of these patterns across conversations. Both because students shared tablets in pairs and because different pairs used the same device across lessons, we could not link transcripts to individual students' assessment outcomes.

Finally, after the trial concluded, we invited teachers from treatment classrooms to participate in focus groups to discuss their experiences with lesson planning, classroom supervision, and Gemini itself. We held four of these virtual focus groups, lasting 60 to 90 minutes with four to five teachers each. We recorded and transcribed these discussions, then conducted an iterative, inductive thematic analysis [18, 35] of the transcripts. We completed the entire analysis without AI assistance, with multiple team members serving as coders. The coders first skimmed every transcript to familiarize themselves with the full corpus. Each coder then independently generated codes on a single shared transcript. We discussed this transcript and the group's codes before consolidating the latter into themes and developing a formal codebook. We then independently applied the codebook to all transcripts, with at least two coders assigned to code each transcript. After finishing this coding, each pair of coders met to reconcile any disagreements and produce unified coding for their transcript. We iterated on the codebook throughout this stage, refining definitions and incorporating new patterns as the coders worked through the full corpus of transcripts.

C. Results

We preregistered our analysis plan with the AEA RCT Registry (AEARCTR-0016651).

C.1. Study sample

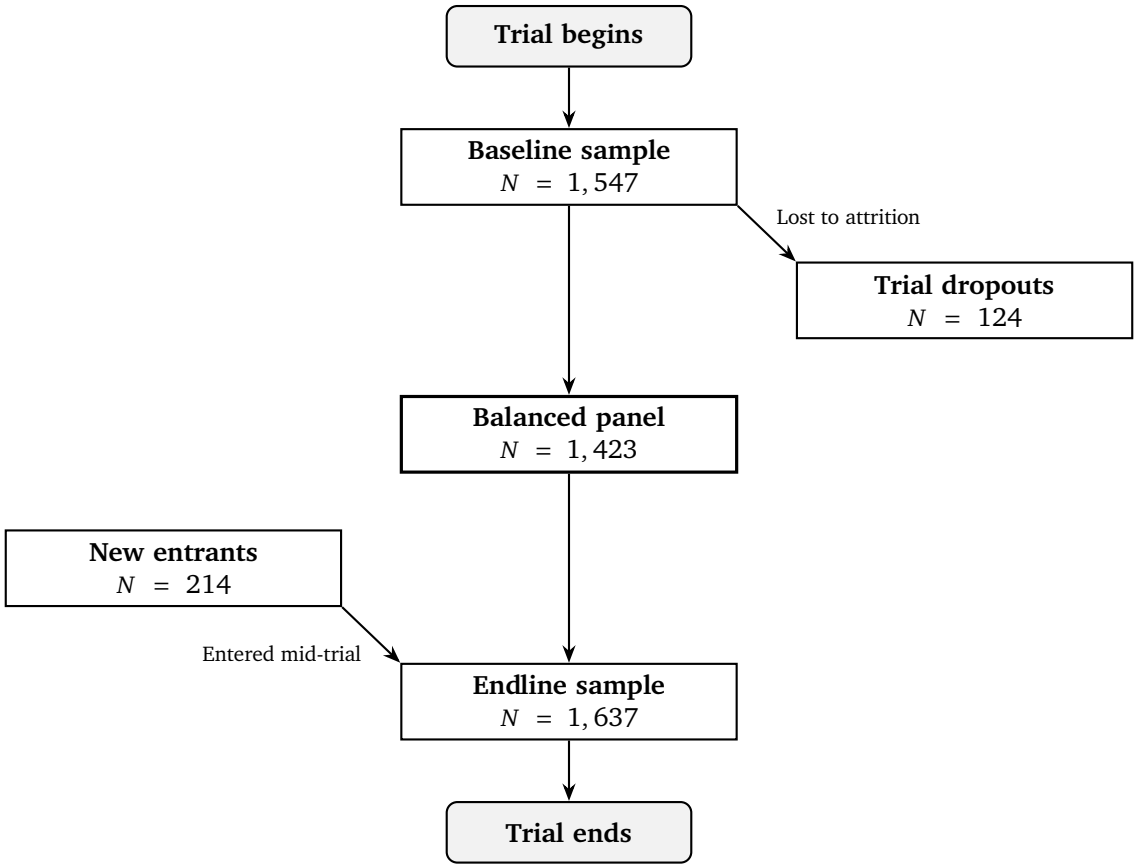


Figure C.1. Our overall sample for the trial—comprising students assessed in at least one round—included $N = 1763$ students across 48 classrooms. As the trial progressed, background levels of student mobility caused changes to our sample. Some students left school after our baseline assessment, and others joined classes later in the term. Most of our statistical analyses focus on the balanced panel, composed of students present at both baseline and endline.

Table C.1. Balance of student characteristics at baseline

	Control	Treatment
Number of students	784	763
Baseline mathematics score (SD)	0.086 (0.860)	−0.081 (0.837)
Baseline reading score (SD)	−1.372 (0.840)	−1.351 (0.788)

Values represent means, with standard deviations in parentheses.

Table C.2. Student retention at follow-up

	Control	Treatment
Students entering at baseline	784	763
Students retained at endline	711	712
Follow-up rate	0.907	0.933

The follow-up rate represents the proportion of the baseline sample size that remained in the study at endline.

Table C.3. Predictors of student retention at follow-up

	Joint estimation
Arm: Treatment	0.032*
	(0.013)
Age (years)	−0.006
	(0.006)
Gender: Female	0.022
	(0.017)
Baseline mathematics score (SD)	0.021*
	(0.009)
Arm: Treatment × Baseline mathematics score (SD)	−0.025
	(0.017)
Observations	1,543
R-squared (R^2)	0.048
Adjusted R-squared	0.031

This table presents estimated coefficients from the specified model (with standard errors in parentheses), as well as model summary statistics. The model predicts a binary indicator for completing the endline mathematics assessment. We restrict the sample to students with complete baseline covariates. The model includes block fixed effects and clusters standard errors at the classroom level. The reference level for "Arm" is the control arm, and for "Gender" is male students. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

C.2. Overall effect

Table C.4. Intent-to-treat effects on endline mathematics scores

	(1) Unadjusted	(2) Baseline adjusted	(3) Fully adjusted
Arm: Treatment	0.216	0.258*	0.259*
	(0.137)	(0.115)	(0.116)
Baseline mathematics score (SD)	–	0.270**	0.268**
		(0.044)	(0.044)
Baseline reading score (SD)	–	0.035	0.031
		(0.022)	(0.021)
Gender: Female	–	–	–0.273**
			(0.044)
Age (years)	–	–	–0.016
			(0.014)
Observations	1,637	1,423	1,421
R-squared (R^2)	0.307	0.347	0.370
Adjusted R-squared	0.297	0.335	0.358

This table presents estimated coefficients from several model specifications (with standard errors in parentheses), as well as model summary statistics. The models predict endline mathematics scores in standard deviation units. All models cluster standard errors at the classroom level and include block fixed effects. Models (2) and (3) use an ANCOVA framework, controlling for baseline test scores. The reference level for "Arm" is the control arm, and for "Gender" is male students. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table C.5. Difference-in-differences intent-to-treat effects on endline mathematics scores

	(1) Block fixed effects	(2) Student fixed effects
Arm: Treatment × Timepoint: Endline	0.317* (0.146)	0.317* (0.146)
Timepoint: Endline	0.007 (0.091)	0.007 (0.090)
Arm: Treatment	–0.122 (0.093)	–
Observations (students · timepoints)	2,846	2,846
R-squared (R^2)	0.279	0.720
Adjusted R-squared	0.272	0.439

This table presents estimated coefficients from several model specifications (with standard errors in parentheses), as well as model summary statistics. The models predict mathematics scores in standard deviation units. We estimate all models on the balanced panel and cluster standard errors at the classroom level. The reference level for "Arm" is the control arm, and for "Timepoint" is the baseline assessment.
† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Our pre-analysis plan specified the estimation of treatment and dosage effects on five mathematics subdomains (numbers and numeration, everyday arithmetic, geometry and measurement, algebra, and statistics and probability). We ultimately did not conduct these analyses; in retrospect, the assessment contained too few items per subdomain for IRT scoring to produce stable subdomain-level estimates.

C.3. Dosage

Table C.6. Treatment-on-the-treated effects on endline mathematics scores (per hour)

	(1) Unadjusted	(2) Baseline adjusted	(3) Fully adjusted
Treatment dosage (hours)	0.013 (0.008)	0.016* (0.007)	0.017* (0.007)
Baseline mathematics score (SD)	–	0.272** (0.043)	0.270** (0.043)
Baseline reading score (SD)	–	0.032 (0.022)	0.028 (0.022)
Gender: Female	–	–	–0.272** (0.044)
Age (years)	–	–	–0.017 (0.014)
Observations	1,637	1,423	1,421
R-squared (R^2)	0.310	0.349	0.373
Adjusted R-squared	0.297	0.335	0.358

This table presents estimated coefficients from several model specifications (with standard errors in parentheses), as well as model summary statistics. We used treatment assignment as an instrument for total dosage hours. The models predict endline mathematics scores in standard deviation units. All models include block fixed effects and cluster standard errors at the classroom level. Models (2) and (3) use an ANCOVA framework, controlling for baseline test scores. The reference level for "Gender" is male students. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table C.7. Treatment-on-the-treated effects on endline mathematics scores (intended dosage)

	(1) Unadjusted	(2) Baseline adjusted	(3) Fully adjusted
Dosage: Completed requested hours	0.292 (0.186)	0.380* (0.169)	0.382* (0.172)
Baseline mathematics score (SD)	–	0.273** (0.043)	0.271** (0.043)
Baseline reading score (SD)	–	0.030 (0.022)	0.025 (0.022)
Gender: Female	–	–	–0.275** (0.043)
Age (years)	–	–	–0.015 (0.015)
Observations	1,637	1,423	1,421
R-squared (R^2)	0.307	0.347	0.370
Adjusted R-squared	0.297	0.335	0.358

This table presents estimated coefficients from several model specifications (with standard errors in parentheses), as well as model summary statistics. We used treatment assignment as an instrument for completing the requested hours. The models predict endline mathematics scores in standard deviation units. All models include block fixed effects and cluster standard errors at the classroom level. Models (2) and (3) use an ANCOVA framework, controlling for baseline test scores. The reference level for "Dosage" is students who did not reach the requested hours, and for "Gender" is male students. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table C.8. Difference-in-differences treatment-on-the-treated effects on endline mathematics scores (per hour)

	Block fixed effects
Treatment dosage (hours) × Timepoint: Endline	0.013†
	(0.007)
Timepoint: Endline	0.062
	(0.075)
Observations (students · timepoints)	2,846
R-squared (R^2)	0.276
Adjusted R-squared	0.270

This table presents estimated coefficients from the specified model (with standard errors in parentheses), as well as model summary statistics. We used treatment assignment as an instrument for total dosage hours. The model predicts mathematics scores in standard deviation units. We estimate the model on the balanced panel, include block fixed effects, and cluster standard errors at the classroom level. The reference level for "Timepoint" is the baseline assessment. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table C.9. Difference-in-differences treatment-on-the-treated effects on endline mathematics scores (intended dosage)

	Block fixed effects
Dosage: Completed requested hours × Timepoint: Endline	0.293†
	(0.161)
Timepoint: Endline	0.060
	(0.076)
Observations (students · timepoints)	2,846
R-squared (R^2)	0.276
Adjusted R-squared	0.270

This table presents estimated coefficients from the specified model (with standard errors in parentheses), as well as model summary statistics. We used treatment assignment as an instrument for completing the requested hours. The model predicts mathematics scores in standard deviation units. We estimate the model on the balanced panel, include block fixed effects, and cluster standard errors at the classroom level. The reference level for "Dosage" is students who did not reach the requested hours, and for "Timepoint" is the baseline assessment. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

C.4. Heterogeneity

Table C.10. Intent-to-treat effects on endline mathematics scores by baseline mathematics score

	Baseline adjusted
Arm: Treatment	0.250*
	(0.111)
Arm: Treatment × Baseline mathematics score (SD)	0.195**
	(0.060)
Baseline mathematics score (SD)	0.187**
	(0.047)
Observations	1,423
R-squared (R^2)	0.354
Adjusted R-squared	0.343

This table presents estimated coefficients from the specified model (with standard errors in parentheses), as well as model summary statistics. The model predicts endline mathematics scores in standard deviation units. We estimate the model on the balanced panel, include block fixed effects, and cluster standard errors at the classroom level. The reference level for "Arm" is the control arm. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table C.11. Intent-to-treat effects on endline mathematics scores by baseline reading score

	Baseline adjusted
Arm: Treatment	0.221*
	(0.101)
Arm: Treatment × Baseline reading score (SD)	–0.027
	(0.061)
Baseline reading score (SD)	0.048
	(0.033)
Baseline mathematics score (SD)	0.270**
	(0.044)
Observations	1,423
R-squared (R^2)	0.347
Adjusted R-squared	0.335

This table presents estimated coefficients from the specified model (with standard errors in parentheses), as well as model summary statistics. The model predicts endline mathematics scores in standard deviation units. We estimate the model on the balanced panel, include block fixed effects, and cluster standard errors at the classroom level. The reference level for "Arm" is the control arm. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table C.12. Intent-to-treat effects on endline mathematics scores by gender

	Baseline adjusted
Arm: Treatment	0.225*
	(0.103)
Arm: Treatment × Gender: Female	0.083
	(0.093)
Gender: Female	–0.315**
	(0.062)
Baseline mathematics score (SD)	0.276**
	(0.043)
Observations	1,423
R-squared (R^2)	0.370
Adjusted R-squared	0.358

This table presents estimated coefficients from the specified model (with standard errors in parentheses), as well as model summary statistics. The model predicts endline mathematics scores in standard deviation units. We estimate the model on the balanced panel, include block fixed effects, and cluster standard errors at the classroom level. The reference level for "Arm" is the control arm, and for "Gender" is male students. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table C.13. Intent-to-treat effects on endline mathematics scores by age

	Baseline adjusted
Arm: Treatment	0.243*
	(0.121)
Arm: Treatment × Age: Above expected for grade	0.041
	(0.073)
Age: Above expected for grade	–0.049
	(0.039)
Baseline mathematics score (SD)	0.273**
	(0.044)
Observations	1,421
R-squared (R^2)	0.346
Adjusted R-squared	0.334

This table presents estimated coefficients from the specified model (with standard errors in parentheses), as well as model summary statistics. The models predict endline mathematics scores in standard deviation units. All models cluster standard errors at the classroom level and include block fixed effects. We classified students as above expected age for grade if they were 14 years or older in Grade 7, or 15 years or older in Grade 8. The reference level for "Arm" is the control arm, and for "Age" is students at or below the expected age for their grade. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table C.14. Intent-to-treat effects on endline mathematics scores by grade

	Baseline adjusted
Arm: Treatment	–0.078*
	(0.037)
Arm: Treatment × Grade: Grade 8	0.429**
	(0.142)
Baseline mathematics score (SD)	0.274**
	(0.042)
Observations	1,423
R-squared (R^2)	0.355
Adjusted R^2	0.344

This table presents estimated coefficients from a single model specification (with standard errors in parentheses), as well as model summary statistics. The model predicts endline mathematics scores in standard deviation units. We estimate the model on the balanced panel, include block fixed effects, and cluster standard errors at the classroom level. For this analysis, our pre-analysis plan specified estimating treatment effects for each grade using separate regressions. We instead compute a single model with a grade interaction to directly test the difference in effectiveness between grades. The reference level for "Arm" is the control arm, and for "Grade" is Grade 7. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Table C.15. Treatment-on-the-treated effects on endline mathematics scores by baseline reading level (intended dosage)

	(1) Unadjusted	(2) Baseline adjusted	(3) Fully adjusted
Dosage: Completed requested hours	0.339**	0.407**	0.412**
	(0.159)	(0.158)	(0.156)
Dosage: Completed requested hours × Baseline reading level: pre-foundational	−0.014	−0.033	−0.038
	(0.149)	(0.131)	(0.134)
Baseline reading level: pre-foundational	−0.118*	−0.055	−0.049
	(0.062)	(0.053)	(0.052)
Baseline mathematics score (SD)	–	0.271**	0.269**
		(0.043)	(0.042)
Gender: Female	–	–	−0.276**
			(0.043)
Age (years)	–	–	−0.014
			(0.014)
Observations	1,423	1,423	1,421
R-squared (R^2)	0.294	0.347	0.371
Adjusted R-squared	0.281	0.335	0.358

This table presents estimated coefficients from several model specifications (with standard errors in parentheses), as well as model summary statistics. We used treatment assignment and its interaction with baseline reading level as instruments for the corresponding dosage terms. The models predict endline mathematics scores in standard deviation units. We classified students as pre-foundational if they scored at level 0 on the baseline reading assessment. All models include block fixed effects and cluster standard errors at the classroom level. The reference level for "Dosage" is students who did not reach the requested hours, for "Baseline reading" is students above the pre-foundational level, and for "Gender" is male students. † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

C.5. Student interactions

Table C.16. Interactions over time

Period	Number of conversations	Number of messages	On-topic (prop.)
6 – 12 Oct	458	5,241	0.981
13 – 19 Oct	1,186	11,834	0.902
20 – 26 Oct	1,092	13,518	0.976
27 Oct – 2 Nov	880	10,524	0.966
3 – 9 Nov	997	16,419	0.984
10 – 16 Nov	856	16,785	0.991
17 – 23 Nov	891	16,932	0.978
24 – 30 Nov	679	15,270	0.993
1 – 5 Dec	382	6,821	0.980
Overall	7,421	113,344	0.974

This table presents descriptive statistics from de-identified student transcripts in the treatment arm. We classified on-topicness at the conversation level, then weighted conversations by length to calculate these proportions (i.e., they reflect the relative prevalence of staying on topic).

Table C.17. Student goal orientation over time

Period	Skill-seeking (prop.)	Solution-seeking (prop.)	Info-seeking (prop.)	Other (prop.)
6 – 12 Oct	0.677	0.251	0.056	0.016
13 – 19 Oct	0.808	0.101	0.076	0.015
20 – 26 Oct	0.925	0.031	0.031	0.013
27 Oct – 2 Nov	0.897	0.052	0.036	0.015
3 – 9 Nov	0.907	0.057	0.028	0.009
10 – 16 Nov	0.953	0.028	0.012	0.007
17 – 23 Nov	0.963	0.018	0.011	0.008
24 – 30 Nov	0.980	0.010	0.006	0.015
1 – 5 Dec	0.917	0.063	0.012	0.007
Overall	0.914	0.050	0.026	0.010

This table presents descriptive statistics from de-identified student transcripts in the treatment arm. We classified goal orientation at the conversation level, then weighted conversations by length to calculate these proportions (i.e., they reflect the relative prevalence of each goal orientation). "Other" represents interactions that included either insufficient or ambiguous signal to identify a goal orientation (typically because the conversation was very short).

Table C.18. Tutor moves over time

Period	Scaffolding questions (prop.)	Direct solutions (prop.)	Other (prop.)
6 – 12 Oct	0.642	0.070	0.288
13 – 19 Oct	0.739	0.048	0.213
20 – 26 Oct	0.765	0.014	0.220
27 Oct – 2 Nov	0.740	0.028	0.231
3 – 9 Nov	0.745	0.023	0.231
10 – 16 Nov	0.792	0.013	0.195
17 – 23 Nov	0.799	0.009	0.193
24 – 30 Nov	0.809	0.004	0.187
1 – 5 Dec	0.742	0.022	0.236
Overall	0.764	0.021	0.214

This table presents descriptive statistics from de-identified student transcripts in the treatment arm. We classified tutor moves at the message level, then directly calculated these proportions (i.e., they reflect the relative prevalence of each tutor move). "Other" represents messages that we could not classify as scaffolding questions or direct solutions (anecdotally, often greetings or clarifications).

C.6. Teacher perspectives

After independently coding and then jointly discussing one focus group transcript, we developed an initial codebook composed of seven themes and 30 codes. We then refined the codebook by applying it to the full set of transcripts and iteratively discussing our notes. Our final codebook comprised seven themes and 43 codes. We present the final codebook on the following pages, with an interpretive summary for each theme and definitions, example markers, boundary notes, and illustrative quotes for each code.

To validate the quality and coherence of our initial codebook, we investigated the interrater reliability of the decisions from our initial, independent coding of each transcript. Thematic coding is not typically amenable to reliability analysis [35]: reliability measures assume coders label a set of predefined, fixed units, whereas thematic analysis asks coders to dynamically unitize the source material (that is, to identify and label excerpts of any length). In this case, we observed *post hoc* that each coded excerpt happened to fall within a single speaker turn, and that no coder applied the same code more than once per turn. This allowed us to treat speaker turns as defined, fixed units for coding, and thus to simulate interrater reliability (that is, coders' agreement on whether a given code applied to a given speaker turn). This yielded a Krippendorff's alpha of $\alpha = 0.69$, showing that the coders applied the initial codebook in relatively similar ways, even while coding independently.

Theme: Students' excitement for Gemini

If we were to identify one singular takeaway from our focus groups, it would be the level of students' enthusiasm for Gemini. Teachers from every school emphasized just how much Gemini excited their students. They shared stories of Guided Learning classes captivating student attention, and voiced strong requests to provide broader access to Gemini, beyond the bounds of the present trial.

Code: Request to expand student access to Gemini

Identified in remarks from 11 different teachers.

Definition: Teachers discuss requests to share Gemini with students beyond the current classrooms with access.

Example markers: Mention of control-classroom students pressuring teachers; or request to extend the trial to more students.

Boundary notes: Distinct from *Need for one-on-one use of Gemini* and *Request to expand teacher access to Gemini*.

Quotes:

> "I have to talk to them and say, 'Patience, patience. This is the first phase. Maybe by second term, all of you will come inside... Go and exercise patience.' So please... we appreciate if we can extend this opportunity to all of the classes in our schools. And not just only in our schools. We know there are other schools in that need this opportunity."

Code: Excitement to use AI and other technology

Identified in remarks from eight different teachers.

Definition: Teachers describe student enthusiasm for AI and other technology.

Example markers: Discussion of student excitement for Gemini; mention of students rushing to class; mention of students asking to go to the computer lab; discussion of students wanting to use tablets; mention of increase in student attendance; or reference to student enjoyment of Guided Learning lessons.

Boundary notes: Distinct from *Gemini supporting active learning*.

Quotes:

> "But the introduction of AI—I mean, let me confess, I've seen children rushing to attend classes."

> "The children, they have more interest in this app. Yesterday we were supposed to have classes, but due to the meeting that we are supposed to have yesterday... one or two issues comes up that did not make us to held the meeting. Some of them, most of them met me outside. 'So we are supposed to go to the lab. We need the lab. We want to go and play with the lab.'"

Theme: Teachers' use of Gemini

In each focus group, teachers volunteered detailed accounts of the specific approaches they developed to apply Gemini to lesson preparation. Strikingly, teachers consistently raised the topic of professional learning in these accounts. They recounted moments when Gemini helped them discover new ways to teach familiar topics—fractions, ratios, place value—and credited Gemini with expanding their own subject knowledge alongside their students'.

Code: Use of Gemini for lesson preparation

Identified in remarks from 13 different teachers.

Definition: Teachers describe applying Gemini to plan and generate content for lessons.

Example markers: Discussion of preparing daily lesson notes with Gemini; description of chatting with Gemini to create instructional materials; or mention of using Gemini to generate the full content for a lesson.

Boundary notes: Distinct from *Professional learning from Gemini use*.

Quotes:

"In fact, almost every blessed day, I must make sure that I use Gemini to prepare my lesson notes. And I've seen a lot of differences. Before, I was struggling a lot in preparing my lesson notes. But since the introduction of this app, it has created big differences in my life."

"Then I will just use it as this, to prepare well for my kids to teach them. Almost all the time, I use it for me."

"I'm using AI for research purpose. Especially if I happen to come across a word that is new for the very first time, I will use the AI for research purpose."

Code: Professional learning from Gemini use

Identified in remarks from 12 different teachers.

Definition: Teachers make reference to their experiences with Gemini expanding their own knowledge.

Example markers: Discussion of learning new vocabulary from Gemini; or reference to personal use of Gemini to learn.

Boundary notes: Distinct from *Use of Gemini for lesson preparation*.

Quotes:

"So the use of AI, especially when we talk to Gemini, is so much enriching. It helps us as teachers to build our knowledge, to add more to our knowledge, especially with certain facts which are not within our reach. Gemini explores those facts and reveals it to us."

"You know, Gemini really explains this to me in particular, because I use it in teaching the children. So with that, I learned a new vocabulary called 'placeholder,' which previously is just for me to say, 'Okay, you put a zero.' Now that zero has a name. We call it 'placeholder.' So I learned something very interesting from Gemini."

"For me, teaching mathematics and agricultural science, and even using it now in various subjects—researching various subjects which I'm not even teaching. Sometime people call me now, they said, 'You know everything.' When you bring on a subject, I will most find out—I say, 'Give me a time.' I will research through it and brush with it and respond. 'Say, how do you know about this subject?' Not knowing that I'm using this AI now, this particular one."

Code: Comparison of AI workflows for lesson preparation

Identified in remarks from four different teachers.

Definition: Teachers draw a contrast between lesson preparation with one AI tool to lesson preparation with another.

Example markers: Expression of a preference for Gemini without Guided Learning in teacher workflows; or comparison of Gemini with other AI systems.

Quotes:

"So we discuss, actually. We chat. Like the other one gives—the other apps I've been using just give me direct answers... But Gemini, with certain concepts, it try to dig in to know deeply what is my thought and what do I think I want to know much better. So it guides me depending on the mood I use."

"So when I go home, I will use the Gemini to at least prepare myself, and then get ready for the lessons. So I use that one... it is that Flash mode. Use in the Flash mode to at least prepare."

Code: Request to expand teacher access to Gemini

Identified in remarks from four different teachers.

Definition: Teachers discuss requests to share Gemini with additional colleagues.

Example markers: Request to extend Gemini access with other teachers.

Boundary notes: Distinct from *Requests to expand student access to Gemini*.

Quotes:

"Maybe it will aid us with the other subjects, because the other teachers are now asking—the other teachers are now asking, 'Is it just mathematics?'"

Code: Less time required to prepare lessons with Gemini

Identified in remarks from one teacher.

Definition: Teachers describe lesson planning with Gemini as less time consuming.

Example markers: Reference to Gemini decreasing preparation time; or mention of Gemini enabling the rapid generation of lesson materials.

Boundary notes: Distinct from *Comparison of AI workflows for lesson preparation*.

Quotes:

"Before this time actually, it takes a lot of waste of time, a lot of time chatting different books. When we are using textbooks, sometime the textbook is not available, you have to meet a friend to lend you to see how you can—and your friend will be panting behind you, 'Can I get my book back?' And sometime even it brings conflict between you. Maybe you fell in love with that book so much that you... You don't want to return this book again. So really with this Gemini, I don't need to really—I just need few resources. Let me always have my data and do my research. So it's really lessen my time and it's even giving me a comfortable time. Anytime I can do my research. Midnight, I don't need to think worried about it during the day. At any time when I have data and when I'm ready, I can do my research. So it's really less time for now."

"The time is less, it's less using the Gemini."

Code: More time required to prepare lessons with Gemini

Identified in remarks from one teacher.

Definition: Teachers describe lesson planning with Gemini as more time consuming.

Example markers: Reference to Gemini increasing preparation time; or discussion of the time needed to work through different options.

Boundary notes: Distinct from *Comparison of AI workflows for lesson preparation*.

Quotes:

"In terms of time, Gemini consumes time, because... I mean, it exposes you to various options that you have to choose the appropriate one. Whereas the previous one, as I told you, is a matter of rote learning, verbatim, word for word."

Theme: Teachers' general attitudes toward AI

While focus group discussions surfaced occasional misconceptions and apprehensions concerning AI, teachers far more frequently voiced positive attitudes toward both AI in general and Gemini in particular. For many teachers, the trial provided their first direct exposure to AI. The hands-on experience appeared to make a strong impression, with some teachers pressing for adoption among their peers and across other subjects.

Code: Positive feelings towards Gemini

Identified in remarks from 10 different teachers.

Definition: Teachers express general positive sentiment towards Gemini. Child code under *Positive feelings towards AI*.

Example markers: Description of Gemini as friendly; or mention of frequent use of Gemini.

Boundary notes: Distinct from *Positive feelings towards AI*.

Quotes:

"Gemini now is my friend. I rely on this particular one."

"Gemini, it's a friendly app for me."

Code: Positive feelings towards AI

Identified in remarks from eight different teachers.

Definition: Teachers express general positive sentiment towards AI.

Example markers: Mention of frequent use of AI tools; reference to the world improving with AI; or expression of positive sentiment for AI in education.

Boundary notes: Distinct from *Excitement to use AI and other technology* and *Positive feelings towards Gemini*.

Quotes:

"I came to realize that the world is improving with AI. And with AI, I'm sure everybody will understand exactly what they want to understand."

Code: General misconceptions of AI

Identified in remarks from six different teachers.

Definition: Teachers make statements about AI capabilities that reveal a misunderstanding of the underlying technology.

Example markers: Description of an AI system thinking on its own; or reference to an AI system consulting sources independently.

Boundary notes: Distinct from *Comparison of AI workflows for lesson preparation*.

Quotes:

"And I came to understand that ChatGPT is the only AI."

"What I knew about AI before this time—I was just thinking that AI is just used for us to search for materials only."

Code: First experience with AI

Identified in remarks from two different teachers.

Definition: Teachers describe this trial as their first exposure to AI.

Example markers: Reference to the trial as a first experience with AI; or mention of no prior exposure to AI tools.

Boundary notes: Distinct from *Positive feelings towards AI*.

Quotes:

"This is the first time for me to use and have access with the use of AI. You know, this is my first time—when the EducAid people come, and I started using this AI to search for notes and do other things."

Code: Negative feelings towards AI

Identified in remarks from one teacher.

Definition: Teachers express general negative sentiment towards AI.

Example markers: Description of AI as controversial; reference to reluctance to use AI; or mention of data privacy concerns.

Boundary notes: Distinct from *General misconceptions of AI*.

Quotes:

> "I'm not too happy... I've been listening greatly to the international media. Some of these discussion that data are not protected, they hack out data from people through this AI, they know much about people using AI. So there is this lot of conspiracy theory about AI. People saying negative things about AI."

Theme: Student learning with Gemini

Teachers devoted considerable attention to Gemini's effects on the learning process itself. When asked to compare Guided Learning sessions with their regular lessons, they frequently emphasized the way Gemini encouraged active learning among students, as well as its tendency to coach students step by step through their questions, rather than providing direct solutions. Multiple teachers reported already seeing improvements in their students' learning. Sustained use of Gemini also appeared to build students' comfort and familiarity interacting with technology.

Code: Comparisons of Gemini's instructional approach with traditional teaching

Identified in remarks from nine different teachers.

Definition: Teachers draw an explicit contrast between student learning in Gemini lessons and their regular classes.

Example markers: Expression of direct comparisons between Guided Learning lessons and standard lessons.

Quotes:

> "So if they have problem, I will tell them that they will call for them to call me. And with so doing, they will interact with the app. When they have issue, they call on me and I will explain certain concept to that particular set of group. From there, any other people that have issue, they will also call me. So by so doing, I'm sure most of them—most of them, they can collect the concept very faster than I do the talking."

Code: Gemini supporting active learning

Identified in remarks from nine different teachers.

Definition: Teachers mention students participating more actively during Guided Learning lessons.

Example markers: Mention of students engaging more with Guided Learning lessons; emphasis of students responding to Gemini; mention of students interacting with the app; or discussion of active learning during Guided Learning lessons.

Boundary notes: Distinct from *Excitement to use AI and other technology*.

Quotes:

> "But Gemini—most of the students participate, they respond. Gemini asks them, they respond. So, if we are to compare it to our previous one, I believe Gemini is more better than our normal teaching. Thank you."

Code: Gemini offering guided scaffolding

Identified in remarks from seven different teachers.

Definition: Teachers discuss Gemini working step by step with students, rather than simply giving away an answer.

Example markers: Discussion of Gemini responding socratically to students; mention of students and Gemini working side-by-side until arriving at an answer; or mention of Gemini responding to student "I don't know" messages by providing directions.

Quotes:

> "Gemini is so much encouraging to me because it helps me a lot. When we talk to Gemini, it directs us what to do so as to arrive at the appropriate answer. In other words, Gemini is not solving the problem for you all by itself. We work side by side: you ask questions, he directs you what to do. And following those instruction, we arrive at a point that is fruitful, especially for the children in our custody."

Code: Gemini improving student learning

Identified in remarks from three different teachers.

Definition: Teachers reference Gemini enhancing the speed or depth of students' academic progress.

Example markers: Mention of students needing less time to understand concepts during Guided Learning lessons; discussion of students grasping concepts more quickly with Gemini; or reference to Gemini improving student performance on assessments.

Quotes:

> "I'm sure most of them—most of them, they can collect the concept very faster than I do the talking."

Code: Gemini use improving student skills with technology

Identified in remarks from two different teachers.

Definition: Teachers discuss Guided Learning lessons improving students' abilities to use technology.

Example markers: Mention of Gemini use improving student typing skills; or reference to Gemini use helping to familiarize students with software in general.

Boundary notes: Distinct from *Gemini improving student learning*.

Quotes:

> "Due to the Gemini app—this app makes them to be more familiar with the instruments, with the facility... Now, they have the chance for them to go there and make use of the facility."

Code: Gemini engaging students in multiple languages

Identified in remarks from one teacher.

Definition: Teachers mention students interacting with Gemini in a language other than English.

Example markers: Reference to students prompting Gemini in a language other than English; or mention of Gemini switching languages.

Quotes:

> "One time we are treating about laws of indices. We notice, I think the Gemini pick it up... can I introduce Creo? 'Yes,' the child click on that—Creo. And the thing was the Gemini was teaching laws of indices in Creo. I even make a snapshot to that particular. I said—I call the field monitor. I said, so Creo is involved here."

Theme: Lesson delivery with Gemini

Teachers' discussions of lesson delivery frequently extended beyond Gemini itself to include the broader classroom environment. They described playing an active, hands-on role during Guided Learning sessions: circulating among pairs of students, supporting those who struggled, and fielding a steady stream of questions. Teachers also reflected on the peer interactions that shaped each Guided Learning session. Overall, their comments outlined a framework for instruction that depended on teacher leadership and pedagogical structure just as much as on Gemini itself.

Code: Teacher facilitating Gemini lessons

Identified in remarks from 14 different teachers.

Definition: Teachers describe shifting from a lecturer role to a facilitator role during Guided Learning lessons.

Example markers: Discussion of teachers walking around the class to facilitate lessons; mention of teachers asking if students are dealing with problems; mention of teachers explaining concepts to small groups of students; or reference to Gemini not cutting teachers out of the loop.

Quotes:

> "So we have to do thorough monitoring, moving from one seat to the other. Don't just sit or stand there talking, talking. No, no. That one will not enhance effective learning. You have to move from point to point or from table to table."

Code: Pairing students with differing levels of mastery

Identified in remarks from nine different teachers.

Definition: Teachers describe pairing strongly-prepared students with students needing support during Guided Learning lessons.

Example markers: Reference to pairing one stronger and one weaker student together for Gemini lessons; discussion of having stronger students assist weaker students; or reference to rotating pairs of students based on quiz performance.

Quotes:

> "Taking a good student and a weak one together. Pair them, it will enhance what? Good learning. And indeed, that's what we are doing. And again, in the area of writing, there are students that don't know even how to write... Let's say for example, I gave about two or three questions. Because they are in two, this one will write this one onto the answer, then the other will write the next question onto the answer."

Code: Gemini lesson protocols

Identified in remarks from six different teachers.

Definition: Teachers describe structured routines for conducting Gemini lessons.

Example markers: Mention of students starting a conversation with Gemini by typing their level, their serial number, or a topic; mention of teacher writing a question on the board to begin a Guided Learning lesson; or mention of teacher providing the class with an initial prompt to send to Gemini.

Quotes:

> "For me, I have already a set program. No sooner the tablet is given to the students or the pupils, they just go to the tablet and write their level already. If it's JSS1, they write the JSS1 serial number. If it is JSS2..."

Code: Teacher burden from student questions during Guided Learning lessons

Identified in remarks from two different teachers.

Definition: Teachers discuss becoming overloaded by students' queries while the class engages with Gemini.

Example markers: Discussion of teachers dealing with excess of student questions about Gemini; or mention of difficulty in managing classroom during Gemini lessons.

Boundary notes: Distinct from *Teacher facilitating Gemini lessons*.

Quotes:

> "When Gemini is introduced to them, it's difficult to manage those classes. Like I said earlier on, the timing of this program is not our friend. And when this Gemini sessions is going on, like I say, most of the kids are having issues with their reading and comprehension. So during the course of the Gemini session, you have a situation where kids will be calling you, 'Mister, please come and help me here. Mister!' They will be calling from all angles. So it was a very difficult situation to actually have the kids under control compared to our normal teaching, where everybody will be paying attention to the teacher."

Code: Teacher collaboration across trial arms

Identified in remarks from two different teachers.

Definition: Teachers describe teachers from control and treatment classrooms working together.

Example markers: Teachers collaborating closely across different study arms.

Quotes:

> "The man you have talked to just now is in charge of the intervention class. And I'm in charge of the control class. But we are working hand in gloves, meaning we are consulting each other on every basis, especially on topics that we want to teach. We... find appropriate methodology which to follow so that the children will grasp the content of each topic that we are about to teach."

Code: Request for additional teacher control over Guided Learning

Identified in remarks from one teacher.

Definition: Teachers discuss the need for additional control over Gemini's behavior.

Example markers: Request for teachers to manage Gemini's ability to generate videos; or preference for increased oversight over outputs from Gemini.

Boundary notes: Distinct from *Gemini introducing content from outside the planned class curriculum*.

Quotes:

> "The video should be requested by the instructor or the—or the person that is researching."

Code: Supplemental out-of-class support for students

Identified in remarks from one teacher.

Definition: Teachers reference the need for additional classes outside of the regular classroom schedule.

Example markers: Reference to out-of-school or extra sessions.

Quotes:

> "So we have to go—especially myself who is in charge of the control class—I go the extra mile to organize classes for them outside the normal class system, so as to keep them abreast with the content of what it takes for them to go ahead to the next form."

Theme: Infrastructure for Gemini delivery

When we asked about the challenges teachers encountered during the trial, discussion centered primarily on practical constraints in their schools. Teachers most consistently voiced concerns and frustrations with internet

connectivity and power reliability. They spoke less frequently about devices, though several teachers expressed preferences for specific device types and suggested that classes needed more devices so that students could study individually.

Code: Internet and power reliability

Identified in remarks from nine different teachers.

Definition: Teachers identify losing connectivity or power as an obstacle during Guided Learning sessions.

Example markers: Reference to internet issues; suggestion to improve internet connectivity; mention of field monitors intervening to solve connection issues; or reference to power outages.

Quotes:

> "The internet, also. I will really appreciate if you can see how best we try to improve it more. At times, most often whilst we are using the app, we are having a lot of challenges in that area."

> "One of my challenges that I faced during the session is that—one of it is the internet. It normally goes off unless the team—the field monitor intervene to at least solve that particular issue."

Code: Implementation support from field monitors

Identified in remarks from five different teachers.

Definition: Teachers describe receiving support from field monitors.

Example markers: Mention of field monitors distributing devices; or mention of field monitors intervening with technical problems.

Quotes:

> "One of my challenges that I faced during the session is that—one of it is the internet. It normally goes off unless the team—the field monitor intervene to at least solve that particular issue."

Code: Need for one-on-one use of Gemini

Identified in remarks from four different teachers.

Definition: Teachers voice the need for students to be able to access Gemini individually.

Example markers: Reference to schools wanting more devices; request for one tablet per child; or suggestion to let each student take home a tablet.

Quotes:

> "My only advice is that you have to cater for a good number of tablets, so that each child will have a tablet."

Code: Preference for computers for Guided Learning lessons

Identified in remarks from three different teachers.

Definition: Teachers express a preference for conducting Gemini lessons with computers in the classroom.

Example markers: Reference to computer labs better supporting Guided Learning lessons; reference to reduced setup time for computers; or reference to reduced packing time for computers.

Quotes:

> "So the computer lab now, I think that will curtail this element of stealing the tablet. Or even getting it destroyed... For us to have computer labs... it will help equally in the teaching and learning of mathematics."

Code: Preference for tablets for Guided Learning lessons

Identified in remarks from three different teachers.

Definition: Teachers express a preference for conducting Gemini lessons with tablets in the classroom.

Example markers: Discussion of tablet portability better supporting Gemini use.

Quotes:

"A computer lab must be a special place where in all these equipments are to be installed. But for now, since we are just—we are young in the system, we prefer the tablets because the tablets are very easy to handle... So we prefer more tablets."

Code: Security concerns with tablets

Identified in remarks from two different teachers.

Definition: Teachers voice fears that tablets may be lost or damaged.

Example markers: Mention of theft risks with tablets; discussion of problems stemming from students taking tablets home; reference to anticipated loss if students take home tablets; or reference to anticipated misuse if students take home tablets.

Quotes:

"And there are times they are exposed to the risks of thieves, especially also in this part of the country now we are in the festive mood. People are looking out for greener pastures. So no sooner they see tablets, they will take it and then sell it for any amount so that they get on with their own social life. So the computer lab now, I think that will curtail this element of stealing the tablet."

Theme: Gemini usability for students

Teachers spoke relatively less about student challenges than on other aspects of their experiences, but discussions still touched on a number of factors that prevented students from fully engaging with Guided Learning lessons. Literacy gaps stood out as a particularly common barrier to students' interactions with Gemini. Teachers also flagged specific patterns in Gemini's behavior as areas for improvement, including inconsistent responses across devices and occasional departures from the planned lesson.

Code: Literacy gaps

Identified in remarks from 11 different teachers.

Definition: Teachers note student difficulties with reading and writing as a challenge to Guided Learning lessons.

Example markers: Reference to students' poor reading ability; mention of student reading comprehension impeding their interactions with Gemini; mention of teacher correcting students' spelling; mention of students using improper English in messages to Gemini; or mention of students transcribing questions incorrectly.

Boundary notes: Distinct from *Verbosity of content created by Gemini*.

Quotes:

"That is what I normally do to them because there are others like what [the other teacher] said, they have challenge of reading and understanding. To read—you have a few that can read well, and they can comprehend as well as the thing goes step by step onto the final answer."

"Especially in the area of not writing proper English. If you send a question and it's not structured, sometimes the answer you want will not come. That is one of the challenges only I face. But most of the kids because they have problem in their spellings... sometimes when they—even if you write on the board for them to transfer it—write it wrongly."

Code: Inconsistent responses from Gemini across the classroom

Identified in remarks from five different teachers.

Definition: Teachers discuss Gemini giving different answers across students which complicates whole-class guidance.

Example markers: Mention of student discussion about their tablets showing different examples; discussion of lack of consistent approach across conversations during the same lesson; or mention of teachers needing to manage discrepancies in Gemini responses across the class.

Boundary notes: Distinct from *Gemini introducing content from outside the intended lesson*.

Quotes:

"Most of them raise up their hands and started saying, 'That's not—that's not what I have on my tab.' So I found out that they're having different question."

Code: Insufficient time with Gemini

Identified in remarks from three different teachers.

Definition: Teachers say that the schedule for Guided Learning sessions did not give students enough time with Gemini to learn effectively.

Example markers: Discussion of lesson periods being too short for Guided Learning; mention of too little tutoring time during study; or expression of a desire for more frequent sessions with Gemini.

Quotes:

"But just 40 minutes or double period of 40 minutes—that is 1 hour, that's 80 minutes. To me, it is not... not too sufficient for them to understand the concept all. Especially the use of the typing on the tablet."

Code: Off-task distractions for students

Identified in remarks from three different teachers.

Definition: Teachers discuss students getting sidetracked by non-lesson content on their devices.

Example markers: Mention of students using other apps; mention of students browsing other online resources; or reference to monitoring to keep students focused.

Quotes:

"So their main focus is to watch for other things, instead of them focusing on what they are to do. So although there is only a single app that we have that we are using for the session, that is the Gemini app—but these kids are too smart enough. So, what I'm asking now is finding a ways and at least to mute all those things so that we are having one focus on them: the Gemini app instead of other apps."

Code: Verbosity of content created by Gemini

Identified in remarks from three different teachers.

Definition: Teachers mention that Gemini generates such long responses that students struggle to read and understand them.

Example markers: Reference to lengthy responses from Gemini; reference to the lengthy time needed for students to understand Gemini's responses; reference to students disengaging from a conversation after receiving long messages; or mention of pacing problems.

Boundary notes: Distinct from *Insufficient time with Gemini* and *Literacy gaps*.

Quotes:

"You know, at times it is difficult for them to grasp the content. Because it entails a lot of reading, and some

of them are poor readers. Because looking at our educational system nowadays in our country, Sierra Leone, children from the primary school to secondary school are not well grounded."

Code: Gemini introducing content from outside the intended lesson

Identified in remarks from two different teachers.

Definition: Teachers refer to Gemini broaching topics beyond the learning objectives scoped for a lesson.

Example markers: Discussion of Gemini bringing up material outside the current lesson; or mention of teachers helping students guide Gemini back to lesson topic.

Boundary notes: Distinct from *Inconsistent responses from Gemini across the classroom*.

Quotes:

"Gemini sometimes brings up materials to kids which actually they are not required."

"For example, if we are treating operation of world numbers, sometime it comes along with angles until—when you, the teacher, passing round, checking—you say, 'No, you are not supposed. Go back. Type back the topic we are on. So we all should learn the same thing. You should not be learning the different thing to that child.'"

Code: Guided Learning sessions disrupting the planned class curriculum

Identified in remarks from two different teachers.

Definition: Teachers discuss Guided Learning lessons delaying class progress through their scheme of work.

Example markers: Discussion of Gemini interrupting planned lesson curriculum; mention of schools setting scheme of work; or mention of Gemini providing information beyond school curriculum.

Boundary notes: Distinct from *Inconsistent responses from Gemini across the classroom* and *Decrease in lesson prep time with Gemini*.

Quotes:

"We have a scheme that each head of departments have prepared. And I'm one—that I've prepared my scheme. These are topics I need to cover at this specific time. With the intervention with Gemini, it's actually in that context—it somehow interrupts the way I am supposed to run that scheme, the way I was expecting it. Because there is a set time I have put that I must run this scheme, from this time to this time. That's a problem."

Code: Limited skills with technology

Identified in remarks from two different teachers.

Definition: Teachers describe students' lack of familiarity with technology as impeding their interactions with Gemini.

Example markers: Mention of students coming from backgrounds without access to technology; mention of students not typing properly; or reference to lack of student experience with devices like computers and tablets.

Boundary notes: Distinct from *Literacy gaps*.

Quotes:

"First of all, the challenges when we start this pilot phase, we face some of the challenges... when we start, most of the kids are not know how to use the tab and how to type—do other thing."

Code: Need for audio outputs from Gemini

Identified in remarks from two different teachers.

Definition: Teachers suggest that voice capabilities for Gemini would solve some student challenges.

Example markers: Mention of audio interaction as a solution student challenges with reading comprehension; request for voice output from Gemini; or discussion of the ways that audio capabilities would alleviate student spelling issues.

Quotes:

"And the sound—for me, the sound again is a challenge. I learned this again... we are using, for example, we are using the computers. If you allow all of them to use their speaker, as we are now on this meeting—we are using, we are listening to you from the speaker of the computer, we are getting you. Sometimes people list if they're listening, they can comprehend easily. And with the Gemini, we cannot allow all the kids to put on their speakers. But if they—if there is a microphone or phone for the kids to listen to what they want to listen..."

Code: Foundational knowledge gaps

Identified in remarks from one teacher.

Definition: Teachers mention that students lack prior knowledge needed for Guided Learning lessons.

Example markers: Reference to weakness with foundational skills; discussion of difficulty with background knowledge; or mention of gaps from earlier schooling.

Quotes:

"Children at times are really handicapped. Handicapped in the sense that even for—us teachers, we have discovered that some of them—this basic mathematical time table, they don't know it."

Code: Gemini responses fall short of cultural expectations

Identified in remarks from one teacher.

Definition: Teachers reference Gemini responses diverging from local norms.

Example markers: Mention of responses differing from students' expectations; or discussion of Gemini praise not matching local customs.

Quotes:

"One is the traditional language we have been using as teachers to at least enhance learning, or encourage—motivate other kids to answer. For example, to participate in class. For example, when you are teaching the child, you ask a question, a child answer, you say, 'Correct, can we clap for her? Can we clap for him?' So Gemini—the language Gemini have been using, sometimes the kids will say, 'What is it?' When for example, when you say... 'Exactly!' They will be, 'What is it?' I said the Gemini is congratulating you for what you have done."