TC Document

I. Basic Information for TC

▪ Country/Region:	PANAMA
▪ TC Name:	AI-Tutoring Is All You Need: An RCT on the Educational Promises of Large Language Models
▪ TC Number:	PN-T1359
▪ Team Leader/Members:	Duenas Herrera, Ximena (SCL/EDU) Team Leader; Lopez Gelb Loren Viviana (SCL/EDU); Orellana, Miguel Angel (VPC/FMP); Corriols Diaz, Leonor Odilie (VPC/FMP); Bonilla Merino Arturo Francisco (LEG/SGO); Zarate Martinez, Jose Daniel (ITE/IPS); Maya Iglesias, Viviana Mariela (LEG/SGO); Duarte Rodriguez Jorge Leonardo (SCL/EDU); Moreno, Michelle Leonor (ITE/IPS); Parra Alvarez, Juliana (CID/CPN) Corriols Diaz, Leonor Odilie (VPC/FMP); Bonilla Merino Arturo Francisco (LEG/SGO); Zarate Martinez, Jose Daniel (ITE/IPS); Maya Iglesias, Viviana Mariela (LEG/SGO); Duarte Rodriguez Jorge Leonardo (SCL/EDU); Moreno, Michelle Leonor (ITE/IPS); Parra Alvarez, Juliana (CID/CPN)
▪ Taxonomy:	Research and Dissemination
▪ Operation Supported by the TC:	.
▪ Date of TC Abstract authorization:	02 Sep 2024.
▪ Beneficiary:	Ministerio de Educación, Republic of Panamá
▪ Executing Agency and contact name:	Inter-American Development Bank
▪ Donors providing funding:	OC SDP Window 2 - Economic Growth(W2F); OC SDP Window 2 - Social Development(W2E)
▪ IDB Funding Requested:	OC SDP Window 2 - Social Development (W2E): US$187,500.00 OC SDP Window 2 - Economic Growth (W2F): US$187,500.00 Total: US$375,000.00
▪ Local counterpart funding, if any:	US$0
▪ Disbursement period (which includes Execution period):	30 months
▪ Required start date:	October 31, 2024
▪ Types of consultants:	Individuals and Firms
▪ Prepared by Unit:	SCL/EDU-Education
▪ Unit of Disbursement Responsibility:	CID/CPN-Country Office Panama
▪ TC included in Country Strategy (y/n):	Yes
▪ TC included in CPD (y/n):	No
▪ Alignment to the Update to the Institutional Strategy 2024-2030:	Gender equality; Institutional capacity and rule of law; Productivity and innovation; Social inclusion and equality

II. Objectives and Justification of the TC

2.1 **Objective**. The objective of this TC is to validate the hypothesis that Artificial Intelligence (AI) tutoring positively impacts Panamanian students' reading skills and

Panamanian teachers' *Aprendamos Todos a Leer* (ATAL)[1] adoption and AI tools. We operationalize this objective through two main treatments: a personalized AI Tutor for students and a personalized AI Tutor for teachers. One of the few ways to unequivocally determine the impact of an intervention is through a randomized controlled trial through an experimental design that allows us to measure the impact of each treatment, the combined effect, and complementary or competing treatments. The design spans through two school academic years, such that in the first year we can capture data and in the second year, train our own models.

2.2 **Justification**. The landscape of education is evolving with the advent of advanced technologies. Among these, Large Language Models (LLMs), like GPT-4, have shown potential in transforming educational methodologies. The integration of LLMs in education holds promise for revolutionizing how we teach and learn. By providing personalized, engaging, and accessible educational experiences, LLMs can address many of the challenges faced by traditional education systems. Research is beginning to explore how teachers can leverage generative AI to create personalized learning experiences for students that transform teaching and learning (Mollick, Ethan and Mollick, Lilach, 2024).

2.3 The true capabilities of LLMs are constantly evolving, generating human-like text based on the data they have been trained on. LLMs allow for the structuring of unstructured data (building databases from human language) and enable agentic capabilities such as Contextual Understanding, Decision-Making, Task Execution, and Personalization. Frameworks like LangChain and LlamaIndex, which allow for programming with LLMs and connecting them with data, have seen rapid growth in the last two years. These capabilities translate into cost reductions, especially when showing positive impacts. Scalability is particularly beneficial for education systems with limited resources. LLMs can scale up educational interventions, making it possible to reach a larger number of students without a proportional increase in costs.

2.4 For these reasons, the experiments propose a "serverless" application or "software as a service (SaaS)". This allows for a reduction in costs because it works as a pay--as--you-go system. It is possible to create, adapt or replicate code, adjust the service API keys, and deploy an application instantly. Broadly, the code is divided into two parts. First, the agentic logic (Yao, Shunyu, et al, 2023; Madaan, Aman, et al, 2023; Ridnik, Tal; Kredo, Dedy and Friedman, Itamar. 2024): the application uses LLMs to create agents with specific roles: item creator, item quality inspector, exam evaluator, educational tutor, disseminator, and orchestrator. These roles run on code hosted in the cloud and are accompanied by a vector store (which stores, in a language understood by the LLMs, the ATAL material, categorized theoretical exam, and user interactions), and in a database (which contains user authentication and object storage paths). The second part of the code is the data-based model. This is code with weights:

[1] Aprendamos todos a leer is a program that promotes the development of precursors in early childhood education, and the acquisition, development and consolidation of fundamental reading and writing skills in the early grades. It is an explicit, systematic, structured teaching program with formative assessment support. The development of this educational material was made possible thanks to the support of the resources provided by the Program Improving the Efficiency and Quality of the Education Sector (PN-L1143). Loan Agreement No. 4357/OC-PN with the Inter-American Development Bank (IDB), through the Comprehensive and Continuous Pedagogical Support component.

parameters that encapsulate the regularities found in the data. This will allow, in the second stage of the experiment, not to depend on the generic capabilities of foundational models, but to have a specific model for understanding students' challenges. Both parts of the code comprise the complete application, but they can be of interest separately. The agent logic can be adapted for other disciplines or sectors, and the data-based model can be used with other logic.

2.5 **Theory of change**. There is ample evidence of the potential positive effect of technology-based personalized tutoring (VanLehn, Kurt, 2011). LLMs can significantly lower the barriers to creating effective, engaging simulations, opening new possibilities for experiential learning at scale (Mollick, Ethan, et al. 2024). The mechanism by which our specific IA-Tutor for students contributes to the objective operates in two ways. First, the time students spend on the application is learning time (practice based on the students' interests, real-time correction, and feedback). Second, the students' personalized assessment report serves as a resource for teachers to make informed decisions about the best use of classroom time. Similarly, the mechanism by which the IA-Tutor for teachers contributes to the objective also operates in two ways. The application facilitates access to the ATAL program materials and frees up class preparation time. Both mechanisms from both interventions could help extend and make classroom time more efficient.

2.6 **Methodology**. Loan PN-L1143 financed the distribution of *Aprendamos Todos a Leer* (ATAL) material to all students and teachers from first to third grades in Panama in 2024. At the start of the academic year, teachers also received training on this program's pedagogical proposal. Because the materials were distributed to all schools, a randomly selected sample of schools can be selected to receive the treatment and control. Our study is based on the premise that all students and teachers from grades 1 to 3 have access to the ATAL materials distributed through Loan PN-L1143. While some teachers may not be aware of or choose not to implement the materials, randomization mitigates potential selection bias by evenly distributing such variability across treatment and control groups. Any differences in ATAL adoption across regions become additional covariates for exploration rather than confounding factors. This approach does not affect the validity of the experiment; instead, it provides an opportunity to analyze whether our treatments have varying effects depending on existing levels of ATAL adoption, potentially indicating that the interventions primarily facilitate the adoption of ATAL. The experiment will be conducted at the school level to prevent spillovers, and treatment will begin in second grade during the first year and cover second and third grade during the second year of the experiment so the treated students will be followed for two years, and second grade teachers will receive the treatment for two years.

2.7 The experiment is divided into two stages, corresponding to two school academic years. The first stage is the data capture stage, utilizing foundational models. Here, an AI-Tutor for students that administers an adaptive test, provides real-time feedback. It functions as an exam in that it reveals the student's ability, but for the student, it is an interactive exercise. Parents can accompany their children during this exercise for up to 25 minutes each day throughout the intervention period. Each parent in the treatment group will decide whether to use it, and whether to share the results with the teachers. For the AI-Tutor for teachers, in this first stage, we propose a chatbot that delivers all the ATAL information (materials, courses, recorded conferences), based

on teachers' requests. Both applications operate on WhatsApp to maximize ease of use.

2.8 The second stage involves the development of models that incorporate data from the experiments' first stage. In this stage, the incorporated data from the first stage will reveal common patterns among all students. For instance, the data will help determine if the intensity, rhythm, volume, and sentiment analysis of the voice relate to learning stages and specific recommendations. Thus, in the first stage, personalization is based on each user's history and on theoretical learning path derived from ATAL's formative assessment skills, but in the second stage, the guidance for each student considers the patterns of all students. The same applies to the intervention for teachers. After understanding the teachers' needs, the model can transition from a chatbot to a monitoring and supervision system. The aim is for the AI-Tutor for teachers to save administrative time. Teachers spend time designing lessons, thinking about exams, grading them, and making reports for their superiors and parents. We want to identify the most important tasks to help teachers have more time for what they consider most important.

2.9 Four treatments will be deployed in each of the two academic years so that some groups receive all four interventions, others receive several, and one group receives none (pure control group). In the second school year, improvements will be implemented in the treatments, technologies, and experimental design. The treatments are the AI tutor for teachers, the AI tutor for students, a monthly newsletter for teachers reinforcing the use of ATAL, and a monthly newsletter for teachers reinforcing the use of the AI tutor for teachers (and AI in general). This way, four independent interventions can be conducted to observe both isolated and combined effects. The newsletters' objective is to serve as a distractor, as alternate hypotheses: perhaps teachers need constant reminders of the benefits of ATAL materials rather than real-time interactions with artificial intelligence (or maybe the effects are cumulative).

2.10 In the study, AI refers to Large Language Models (LLMs) that act as personalized tutors for teachers and students. LLMs simulate having multiple individuals who can answer teachers' questions and assess students based on ATAL materials. This approach is based on cost reduction and scalability premises, enabling each teacher and student to have their own personalized tutor—an innovation that was unthinkable just a few years ago. In the second stage, we plan to develop smaller machine learning models using data from the first stage. By analyzing voices, grades, emotions, and error types, we aim to improve the data capture process with shorter assessments and provide better diagnostics, either through our own models or by fine-tuning LLMs to suit our specific needs.

2.11 The randomized controlled trial involves four interventions—AI Tutors for students and teachers, plus two newsletters—and divides schools into all 16 possible combinations of these interventions. This design allows both isolated and combined treatments effects. Randomization will be conducted at the school level to minimize potential spillover effects between treatment and control groups. By assigning entire schools, the likelihood of interaction between participants in different treatment groups will be reduced. The interventions are delivered via WhatsApp, linked to specific phone numbers, ensuring only designated groups receive them. While communication between teachers at different schools cannot be prevented, minimal spillover effects are anticipated, and we plan to account for any that occur by analyzing data based on school proximity or communication networks.

2.12 The second stage will have a similar design, but specific interventions will remain open as the rapid advancement in computational capacity and the increasing capabilities and decreasing costs of LLMs make us optimistic about what can be achieved. For example, in the first stage of DEIF, GPT-4o ("omni") by OpenAI had not been released, which is half the price of GPT-4 and includes vision capabilities (facilitating the grading of written exams). Although GPT-4o real-time audio capabilities have not been released to the public, they will change our strategy. Audio will not need to first be transcribed to be processed by an LLM; instead, a single technology will recognize voice characteristics in real-time. Instead of an exchange of audios, the exam could be a video call. In the first stage, we propose using OpenAI technologies, but we should test other technologies in the second one. If impacts from the first stage are clear, we could replace the informational bulletins (or even the control group) with other interventions.

2.13 Thanks to our experimental design, the identification of the causal effect is through a differences-in-differences estimator (without additional controls): the average difference between controls and treatments in the change between the pre-test and post-test outcome variables identify the impact in an experimental design (Duflo, E., Glennerster, R., and Kremer, M. 2006). As a first step, each experiment (both tutors and both bulletins) can be analyzed separately (each with its respective control) to understand the isolated total impact (Banerjee, A., et al. 2016). To understand the combined effect and effects according to students' skill levels, we will estimate several models using treatments and constructing additional controls (Duflo, E., Dupas, P., and Kremer, M. 2011; Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer, 2002).In addition to the causal results of the experiments, we will explore several correlational findings among the treated. Since the treatments themselves generate data, we can determine if usage intensity correlates with outcomes and understand both teachers' and students' interests, motivations, weaknesses, perceptions, and learning strategies.

2.14 **Sources of data**. The initial source is the list of schools in Panama Province that offer second and third grade, class rosters, teachers that teach those grade levels, schools' directors and subdirectors, and regional supervisors. Through school administrators' parents and caregivers' data will be obtained.

2.15 The study universe consists of 244 public schools from four districts in the Province of Panama, the four most populous districts in the province: Arraiján (35), La Chorrera (53), Panama (119), and San Miguelito (37). Based on historical records from MEDUCA for the years 2023 and 2024, we estimate that for the first cycle, second grade, there will be 838 teachers and 16,530 students - Arraiján (131; 3,191), La Chorrera (131; 3,066), Panama (419; 7,373), San Miguelito (157; 2,901). For the second cycle, third grade, we estimate there will be 816 teachers and 15,919 students - Arraiján (129; 2,884), La Chorrera (125; 2,940), Panama (413; 7,400), San Miguelito (149; 2,696).

2.16 The randomized controlled trials' results will be disseminated in two stages: one document for the first cycle and another for the entire cycle. Both documents should describe the methodology (both the experiment and the applications), the process, the results, cost-effectiveness, and lessons learned.

2.17 **A note on the data**: The chatbot interactions with teachers and students will generate data in different formats such as audio recordings, text messages and assessment results. . Given the collection of recordings, chat interactions, and assessment data the team will ensure compliance with the IDB's Personal Data Privacy Policy and its guidelines (document GN-3030). The student AI-Tutor application will produce perfect data for making predictions for non-intervened students. We will have audio recordings that can be analyzed acoustically (intonation, rhythm, intensity, volume). These recordings will be categorized by student skill/difficulty levels. All this information allows us to find patterns using machine learning techniques (to train models from scratch or to fine-tune existing models). These patterns will enable precise recommendations with short, low-cost instruments. The same logic applies to the data captured with the teacher AI-Tutor. Since there is no single way to train models, the data itself is a resource for investigating patterns. This action does not necessarily depend on the intervention's effectiveness.

2.18 **Main variables to construct**. The selected universe, regardless of treatment randomization, must be fully evaluated with two instruments. For students, a controlled assessment of reading skills (phonological awareness, alphabetic principle, fluency, vocabulary, and comprehension) by gender. For teachers, a perception survey about ATAL (usage, effectiveness, satisfaction) and about Artificial Intelligence in the education sector (challenges, fears, need for training). The primary outcome variables will be constructed from these assessments and surveys. The treatment itself acts as a data generator in two ways. First, capturing data on students' skills (audio recordings, labels, and their relation to achievements) by gender helps identify patterns (through machine learning and/or fine-tuning existing models) that could be used to create shorter assessments (predicting outcomes with fewer data) and utilize the assessment to provide personalized educational recommendations. Second, the application users will reveal their usage, frequency, and perceptions, which will reinforce the information about students' skills and teachers' questions. This information constitutes control variables for secondary outcomes.

2.19 **Strategy Alignment**: The TC is consistent with the IDB Group Institutional Strategy: Transforming for Scale and Impact (CA-631), aligned with the objective of: (i) reducing poverty and inequality, by integrating teachers to the use of new technologies and; (ii) bolstering sustainable regional growth, by introducing tools for remote learning and developing inclusive multimedia content to promote the use of ATAL materials by families and involve them in their education processes. The TC is also aligned with Gender and Diversity Action Framework (GN-2800-13) as assessments will allow to monitor differences by gender and using newsletters, teachers will receive suggested strategies to address differences in results or gaps. The program is also aligned with the cross-cutting area of Institutional Capacity and Rule of Law, since MEDUCA's technological and administrative instruments will be strengthened. The TC is consistent with the Country Strategy with Panama (2021 2024) (GN-3055), in the priority of combating poverty and inequality, through education, science, technology and culture, specifically with the objectives of expanding access and quality of social protection, health, and water and sanitation services in vulnerable populations and to improve the quality and relevance of the educational system. It is also aligned with priority area 5 of inclusive social development financed with Ordinary Capital (GN-2819-14) OC SDP Window 2 - Social Development (W2E), specifically with the results of strengthening public institutions' efforts to become more effective and efficient in social programming, and social sector project execution; support clients to reduce inequality, gender equality, and diversity

through projects. Additionally, it is aligned with the sixth priority area of Inclusive Economic Growth financed with Ordinary Capital (GN-2819-14) OC SDP Window 2 – Economic Growth. Lastly, it is aligned with the Skills Development Sector Framework (GN 30123) related to the importance of developing relevant and culturally appropriate skills to respond to the specific needs of diverse populations such as migrants and refugees.

III. Description of activities/components and budget

3.1 **Component 1**. **Development of an IA-Tutor for students to measure literacy skills and provide feedback (US$230,000).** The specific objective is to promote the use of the IA-Tutor for students that offers personalized assessment and real-time feedback. Ultimately parents can share the student's results with teachers. This component will finance individual consultant for the, (i) deployment of the AI-Tutor for students in both school cycles design; an individual consultant for the (ii) creation (adaptation of ATAL's formative assessments) of a measurement instrument for literacy skills; a consulting firm for the (iii) application of an external literacy assessment; and an individual consultant for the (iv) dissemination of the results through two academic documents: one that presents the process and results of the first stage, and another that presents the complete research. From the experimental design, the expected results are to have a positive effect on students' reading proficiency measured with a literacy exam at the beginning and end of the school year. The exam measures the latent traits of students across various skills. Thus, reading learning translates into calibrated indices in the skills of phonological awareness, alphabetic principle, fluency, vocabulary, and comprehension (five independent indicators). Because the assessment is individual, differences in gender will be observed and outcomes will be used to inform teachers in the newsletters. Additionally, the intervention is focused on a periodical formative assessment that helps students and their caregivers monitor their reading progress, and as explained in the theory of change, caregivers can voluntarily share the results with teachers.

3.2 **Component 2**. **Development of an IA-Tutor for teachers to increase use and adoption of ATAL's materials (US$145,000).** The specific objective is to increase the use and appropriation of ATAL's materials by the Province of Panama teachers through their interaction with an IA tutor. This component will finance individual consultants for the, (i) design, deployment, continuous monitoring, and necessary changes (for the second school cycle) of the AI Tutor for teachers; (ii) development and application of a perception instrument about ATAL and the uses of artificial intelligence; and (iii) newsletters' design, which are directed at (randomly selected groups of) teachers. The expected results are to have a positive effect on teachers' perceptions of ATAL and the uses of generative IA tools. Based on the assessment results, the newsletters will include strategies to improve boy's reading proficiency in early grades. Based on the experimental design's different treatments, we want to identify if perceptions change in response to the treatment group and the cost and scalability implications of the responses.

3.3 **Budget**. This TC has a total budget of US$375,000, US$187,500 will be financed with resources from OC SDP Window 2 – Social Development (W2E) and US$187,500 from OC SDP Window 2 – Economic Growth (W2F). The project will span

30 months: 5 months before the intervention, 20 months during the intervention, and 5 months after the intervention. This timeline will approximately run from October 2024 to April 2027. No counterpart financing is expected.

Indicative Budget

Activity/ Component	Description	IDB/ W2E	IDB/ W2F	Total Funding
Component 1	**Students Treatment**	**$132,500**	**$97,500**	**$230,000**
Activity 2.1	Census Evaluation of Student Literacy. Administered to the entire study population at the beginning and end of both academic cycles.		$90,000	$90,000
Activity 2.2	AI-Tutor for Students. Personal tutor providing assessments, feedback, and real-time reports. Randomly assigned to treatment groups.	$132,500		$132,500
Activity 2.3	Results of the Experiments. Technical documents describing the technology processes, identification of causal effects, and results.		$7,500	$7,500
Component 2	**Teachers Treatment**	**$55,000**	**$90,000**	**$145,000**
Activity 1.1	Perception Survey on ATAL and AI Tools for Teachers. Administered to the entire study population at the beginning and end of both academic cycles.		$45,000	$45,000
Activity 1.2	AI-Tutor for Teachers. Personal tutor providing information on ATAL classes and assessments. Randomly assigned to treatment groups.		$45,000	$45,000
Activity 1.3	Informative Newsletters for Teachers. Randomly distributed newsletters covering practical topics on ATAL and the use of generative AI.	$55,000		$55,000
TOTAL		**$187,500**	**$187,500**	**$375,000**

3.4 **Monitoring.** The development of all products will be closely coordinated by SCL/EDU who will provide guidance to ensure that the products will meet the needs and standards of the Bank. The Project Team will be responsible for the review of all

technical and financial reporting. The Team Leader will be responsible for monitoring activities in the field, and continuous progress meetings with the counterparts and consultants.

IV. Executing agency and execution structure

4.1 This TC will be executed by the Inter-American Development Bank (IDB) through the Education Division of the IDB's Social Sector (SCL/EDU). It will be executed according to TC guidelines (OP-619-4). In line with Appendix 10 of the Operational Guidelines for Technical Cooperation Products (OP-619-4), Bank execution of the TC is justified as contracting by the IDB enhances independence of key products to developed, namely the IA tutor for teachers and students. All disbursements will be executed through the Bank's systems and will require approval from SCL/EDU.

4.2 **Procurement plan.** All procurement to be executed under this Technical Cooperation have been included in the Procurement Plan (Annex IV) and will be hired in compliance with the applicable Bank policies and regulations as follows: (a) Hiring of individual consultants, as established in the regulation on Complementary Workforce (AM-650) and (b) Contracting of services provided by consulting firms in accordance with the Corporate procurement Policy (GN-2303-33) and its Guidelines.

4.3 All deliverables and any other material prepared under this TC are the sole and exclusive property of the Bank, and as such, the Bank has title, rights (including copyrights) and exclusive interests in the ownership of said products.

4.4 Given the delays in obtaining the letter from the liaison body, activities will not be initiated until the letter of no objection from the Ministry of Economy and Finance (liaison body in Panama) is sent.

V. Major issues

5.1 The activities of this TC are primarily based on collecting data from primary sources based on direct communication with teachers and students' caregivers for subsequent analysis. Because the sample will be collected at the school level, schools' directors and the regional supervisor of Panama Province will be critical to the implementation of the experiment. Given that the students' IA-tutor will be offered through their caregivers' mobile phones, we will include a disclaimer on the appropriate use of mobile phones by children. For caregivers and teachers, information on the use of their data will be shared and they must approve their participation in the RCT. Therefore, the associated risks can be considered low.

VI. Exceptions to Bank policy

6.1 None.

VII. Environmental and Social Aspects

7.1 This Technical Cooperation is not intended to finance pre-feasibility or feasibility studies of specific investment projects or environmental and social studies associated with them; therefore, this TC does not have applicable requirements of the Bank's Environmental and Social Policy Framework (ESPF).

Required Annexes:

Results Matrix_1750.pdf

Terms of Reference_56636.pdf

Procurement Plan_29415.pdf