

Real-world Impact of a GenAI Pedagogical Agent and Child-AI Discourse Analysis in K12 Math Learning in the Middle East

Xin Miao¹

xinmiaouae2018@gmail.com

Alef Education Consultancy

Pawan Kumar Mishra

Alef Education Consultancy

Qi Zhou

University College London

Research Article

Keywords: Generative AI tutor, Pedagogical Agent, Intelligent tutoring system, Quasi-experimental evaluation, Child-AI discourse analysis, K12 Math education

Posted Date: August 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10455194/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Real-world Impact of a GenAI Pedagogical Agent and Child-AI Discourse Analysis in
K12 Math Learning in the Middle East

Xin Miao^{1*}, Pawan Kumar Mishra¹, Qi Zhou²

¹ Alef Education Consultancy L.L.C., Al Muntazah, Ministries Complex, P.O.
Box 769719, Abu Dhabi, United Arab Emirates

² University College London (UCL), Institute of Education, UCL Knowledge
Lab, 20 Bedford Way, London, WC1H 0AL, United Kingdom

*Corresponding author: Xin Miao, email: xinmiaouae2018@gmail.com, ORCID
(0000-0002-9915-8275)

Large-scale Real-world Impact of a GenAI Tutoring Agent and Child-AI Discourse Analysis in K12 Math Learning in the Middle East

Abstract. This study investigates the integration of a Generative AI tutoring agent with a K12 math intelligent tutoring system to support mathematics learning among students from Grade 5 to Grade 8 in the United Arab Emirates. Conducted over an academic year with 3,771 students from 40 public schools, the research evaluates the effectiveness of the GenAI tutor in improving learning outcomes and explores how students initiate math-related discourse with the system. Results indicate that students perform better on GenAI Tutor integrated ITS than students using ITS alone, and this is particularly true for below-grade-level students. Analysis of 6,794 student-AI interactive utterances reveals diverse discourse patterns, ranging from procedural math queries to conceptual understanding, highlighting how students utilize the GenAI tutor to solve math problems. Furthermore, the study shows that students engaged with more on-task math discourse (i.e. math talk group) perform better than students engaged with fewer on-task discourse over the entire academic year, which implies consistent on-task use of GenAI tutor could lead to better performance. The study also shed light on GenAI powered K12 pedagogical agent design features and its implications on student self-regulated learning within ITS. This research underscores the potential of carefully designed GenAI K12 learning tools to address educational challenges in real-world contexts and emphasizes the importance of thoughtful research and design to optimize their impact.

Keywords: Generative AI tutor, Pedagogical Agent, Intelligent tutoring system, Quasi-experimental evaluation, Child-AI discourse analysis, K12 Math education

1. Introduction

Mathematics is a foundational discipline critical to students' academic and professional success, fostering problem-solving, logical reasoning, and analytical skills essential to thrive in a knowledge-driven economy. PISA examines the degree to which students at the age of 15, have developed essential skills required to participate in modern societies, as they conclude K12 education. According to PISA 2022 Mathematics report¹, 36 countries (with a total of 81 participating countries) scored statistically significantly below the OECD benchmark, including 3 lower Gulf states (i.e. Saudi Arabia, United Arab Emirates, Qatar). The United Arab Emirates performed the best among the Gulf countries, yet still below OECD benchmark. Researchers Miao et al.² investigated predictive factors for Grade 5 to Grade 9 students' approach to math learning in UAE (United Arab Emirates) public schools, and found out that students lag behind international peers and the average grade gap is observed with a growing trend moving from Grade 5 to Grade 9, and the average grade gap in Grade 9 is reaching -3.366, which means that on average, the actual math performance level of an average Grade 9 student equals that of a fifth grader reaching the seventh month of the academic year. Additionally, English language literacy is found to be the No. 1 factor affecting math achievement, and key self-system factors (e.g. self-efficacy and math anxiety), metacognitive regulatory strategies (e.g. memorization strategies, problem-solving strategies) are top predictive factors too. These learning challenges are faced with children in the UAE, and likely other countries in the Middle East and other regions too. Policy reform is far from tackling them, which underscores an urgent need for innovative intervention to solve these challenges early, before they reach 15 years old.

Artificial Intelligence (AI) has been employed to support mathematics learning through tools like intelligent tutoring system (ITS)³ and dialog-based ITS (e.g. chatbot)⁴. ITSs have been the subject of research during the past decades, with many commercial products available worldwide. While such technologies have shown promises in providing personalized instruction at scale and addressing skill gaps, existing literature lacks nuanced

analysis of how students interact with these tools ⁵. As AI technologies advance to GenAI, new forms of learner-AI interaction emerges, which would lead to more understanding of how K12 children interact with Generative AI.

This study evaluates the impact of a GenAI-powered tutoring agent on student math performance through a math ITS system implemented across 3771 students in UAE K12 schools over the 2023-2024 academic year. Additionally, the study also investigates student math discourse interaction with the GenAI tutor, and how these discourse interactions affect learning outcomes.

2. Literature Review

2.1 ITS: Mechanics and Impact on K12 Learning

ITS (Intelligent Tutoring System) is one of the most widely used AI technologies in K12 mathematics education. Aleven et al.⁶ summarized that ITS system architecture consists of four main components: a. domain model; b. student model; c. pedagogical model; d. problem-solving environment. The domain model represents expert knowledge within a particular task area. It includes concepts, skills, strategies, constraints, and defines how knowledge components relate to each other through prerequisite relationships and how they connect to practice problems and their constituent steps. The student model captures the current mastery level of knowledge and skills targeted in the instruction, and might capture other aspects of learning (e.g. affective or motivational state). According to Miao et al.⁷, the pedagogical model uses information from the student model to determine the domain knowledge presented to the learner (e.g. identify topics a learner needs to master, provide feedback based on learner mastery), while problem-solving environment, refers to the interactive workspace where individual students engage in specific tasks. For instance, a key design consideration is learner agency and effective task engagement. The four models work in tandem, touching on learner cognitive skill development, but also affective, motivational and self-regulatory engagement. Létourneau et al.⁸ conducted a systematic review to identify the impact of ITSs on K12 students' learning and performance. The review included 28 studies that focused on middle and high school students, predominantly examined STEM subjects, and utilized quasi-experimental designs with different intervention time frames. Overall, the findings show that ITSs have a positive impact on K12 student learning outcomes but the results are mitigated when compared to non-intelligent tutoring systems. The authors highlighted some limitations: a. intervention duration tends to be short, and effect size might be due to novelty effect; b. sample size tends to be small and suggests larger and more diverse sample size is needed for future studies; c. AI ethical considerations are largely overlooked in the reviewed studies. Most of the evaluation studies shed light on ITS impact on academic performance, yet the nuances of how ITS are designed to engage K12 learners, how learners interact with AI on a task level, how learner self-system factors and metacognitive skills are incentivized during cognitive tasks, are critical for effective learning but lacking in existing literature.

2.2 Granular Learning Interaction through Integrated ITS with Dialog-based ITS

Nuances of learner-AI interaction are critical to examine, including cognitive interactions, as well as affective, motivational and self-regulatory nuances. On the cognitive front, VanLehn, K.⁹ introduced a key hypothesis "interaction plateau" indicating that the granularity of interaction decreases (e.g. from answer-based to step-based and sub-step based), the effectiveness increases to a point. After reaching the plateau, further decreases in granularity do not lead to significant increases in effectiveness. Chatbots bring forward the potential to allow one-

to-one interaction, particularly step-based or sub-step learning. Cheng et al. (2024)⁴ studied the integration of a rule-based chatbot into an online math ITS (ASSISTments), as an alternative to traditional hint systems. The design of the chatbot incorporates scaffolding strategies and social emotional considerations to allow students to construct new understanding while interacting with a chatbot that is designed to motivate learners and foster resilience. The sampled population consists of 85 middle and high school students in the United States. No immediate positive impact on learning outcomes is found, due to short intervention duration. However, student feedback revealed that traditional hints do not give examples of how to solve problems, and are difficult for slow learners or visual learners to process, suggesting the chatbot's potential for one to one adaptive learning support. Even though the chatbot studied did not use GenAI technologies, this study is an attempt to improve interaction granularity between individual learner and ITS system, and such granularity is expanding from cognitive scaffolding, to integrating learner social emotional needs (e.g. motivation, resilience). Martínez-Téllez and Camacho-Zuñiga¹⁰ examined the use of an AI-powered chatbot in a flipped mathematics learning environment, finding that students achieved lower-than-expected math problem-solving scores. The authors attributed this to students' unfamiliarity with AI tools.

In short, large-scale, long-term impact study of GenAI-powered chatbots implemented in K12 is extremely scarce. Attention is needed not only on the impact on academic learning outcomes, but also the pedagogical interactions between GenAI chatbot and K12 learner, considering student motivation, engagement, and avoiding deskilling learners in productive thinking processes. This research attempts to bridge the gap.

2.3 Animated K12 Pedagogical Agent

Apart from cognitive interaction, AI pedagogical agents bring forward the possibility of using child-friendly AI agent designs to engage K12 learners socially and emotionally.

“A pedagogical agent (PA) can be defined as a virtual character or physical robot that seeks to promote learning, enhance motivation and provide support to engage in an educational activity. PAs are typically considered as part of the user interface and designed in the context of system features intended to support learning, such as problem solving with natural interaction through speech, gesture, emotions and action¹¹”.

Substantial effort has gone into investigating the design, roles and impact of pedagogical agents during the past three decades. Baylor and Kim¹¹ investigated design aspects of pedagogical agents, including ethnicity and gender, and found that agents that challenged stereotypes lead to greater transfer of learning and Schroeder and Adescope¹¹ found that gender had no impact on cognitive or affective outcomes in their studies. Finkelstein et al.¹¹ found that a peer agent that used a culturally congruent dialect produced improved science communication and science reasoning by learners.

Researcher Lane, H. Chad¹¹ introduced a few key points as roadmap for PA: 1) support for emotional self-regulatory skills, 2) support for interest development (e.g. personalized motivational techniques that reflect personal relevance), 3) identify learner assets and broaden adaptivity of PAs (e.g. PAs learn about learner interests, strengths, possible goals that would influence the pedagogical choices such as culturally aligned learning activities and feedback policy. VanLehn, K.⁹ also highlighted the importance of developing self-improving tutoring systems that monitor their own effectiveness and adjust teaching strategies based on learner performance. This is partly possible now given GenAI Agent advancement in memory. For example, GenAI agent “episodic memory” defined as “stored sequences of agent’s past memory” according to LangChain¹², could allow AI agents

to remember and learn its interactive sequences with learners. Assuming AI has a good understanding of each learner, including personality, interests, behavioral patterns, thinking patterns etc., it is possible for an AI Agent to customize how to interact with each individual student more human-like, with social emotional engagement.

2.4 Learner-AI Math Dialogic Interaction

Mathematical communication is widely considered an essential tool for expressing mathematical ideas, explaining misconceptions, and demonstrating mathematical connections¹³. It enables students to understand and analyse mathematical concepts while fostering the development of mathematical thought¹⁴. Numerous empirical studies have examined the impact of mathematical communication on learning outcomes and mathematical skills. For example, Klibanoff and colleagues¹⁶ conducted a study with 198 preschool students from 26 classrooms in 13 preschools to investigate the effect of math talk. Their results showed a significant relationship between the amount of math talk and students' growth in mathematical knowledge.

Wang et al.¹⁵ conducted a computational analysis of dialogic interactions in online one-on-one mathematics tutoring, using data from a Singapore-based platform serving K-12 students. The platform pairs students with a suitable human tutor, and through pedagogical dialog the human tutor assists the student in resolving mathematical queries. The researchers developed a dialog coding system for online tutoring sessions and leveraged advanced AI techniques, including a RoBERTa-based automatic annotator augmented with ChatGPT-generated training data, to analyze tutoring conversations. Results show that, 1) human tutors often applied both didactic and other effective teaching strategies though dominated conversations following a predominantly "provide-and-explain" model; 2) off-task chatting accounts for a significant proportion of the tutoring session; 3) high school students exhibited greater engagement and cognitive abilities than primary and middle school students through their active and superior reasoning skills; 4) High-level constructive expressions are less frequent compared to lower level claims, which could impede cognitive processes and learning gains.

In short, existing literature highlighted gaps in 1) lack of large-scale, real-world study of the impact of GenAI tutor on K12 student academic learning; 2) lack of literature on adding GenAI tutoring agent to ITS platforms to enhance interaction granularity; and 3) lack of nuances on how K12 students interact with GenAI tutor. This study attempts to bridge these gaps in real-life, large-scale applications.

3. Research Questions

Therefore, the following research questions are investigated:

- 1) Does integrating a math ITS with a GenAI tutoring agent lead to learning gains compared to using a math ITS alone?
- 2) What types of math discourse are initiated by students to get learning support from the GenAI Tutoring agent?
- 3) Do students who engage in more on-task math discourse with GenAI Tutoring agent show greater learning improvements compared to those who engage in fewer on-task discourse?

4. Background & Educational Context

4.1 Context Statistics

This study is conducted from the 2023 to 2024 academic year, with a total number of 3771 students onboarded to use the GenAI tutoring agent on a math ITS, from Grades 5 to Grade 8 across 40 K12 schools in the United Arab Emirates. The distribution of students by grade is as follows: 36.72% in Grade 5, 29.51% in Grade 6, 25.67% in Grade 7, and 8.09% in Grade 8. During the academic year, students can use the AI tutoring system anytime through their own devices. By design, the AI system incentivizes learners to use it as much as they should. Below is a detailed description of the entire AI system, including the ITS and integrated GenAI tutoring agent.

4.2 Designed AI System: GenAI Dialog-based Tutoring Agent Integrated into K12 Math ITS

In this section, details of the AI tutoring system are explained. A GenAI dialog-based tutoring agent is integrated into the K12 math ITS that consist of the following key features:

1) Adaptive diagnostic tests for domain subject (i.e. Math) and instructional language subject (i.e. English) are developed and localized to identify learning gaps, as mentioned earlier, English literacy has been identified as the number one factor contributing to students' math learning in the UAE. The math diagnostic test is developed based on the Common Core State Standards (CCSS). MetaMetrics' Quantile measures are used to identify students' current skill levels, which could be used to benchmark against international peers who follow CCSS. The English reading test follows CEFR level.

2) Based on diagnostic insights, personalized remediation pathways are designed to address individual missing skills and concepts through bite-sized, localized digital instructional content. The Math ITS remediation journey consists of four domains (i.e. geometry, measurement data and statistics, algebra and algebraic thinking, numbers and operations), and each domain consists of levels that include lessons and activities. For example, A Grade 8 student could be recommended with Grade 7 Geometry level lessons and Grade 6 Algebra level lessons. Content design focuses on using simple, relatable texts, localized child-friendly characters, playful and positive learning designs to engage K12 learners.

3) On-demand one-to-one tutoring support from a GenAI agent is provided, as the ITS is a student self-regulated learning system and learners are assumed to encounter questions, misconceptions at any point of the remediation journey. Fig. 1 presents illustrations of how the GenAI tutoring agent is integrated into the math ITS. When students are in the personalized remediation journey, they can access the tutoring function when needed. For instance, children can click on the GenAI tutor icon in green on the right top corner of the ITS system (e.g. Ask Me, or Hide Me). Additionally, children also have the option to ask questions from their teachers, if they prefer. The "Ask Your Teacher" tab is placed next to the GenAI tutor icon. Details of the GenAI tutoring agent are explained below.

4) Learning analytics dashboards are designed to reflect learner progress, skill mastery, engagement real-time, and each student has a gamified profile page to reflect progress made, skills mastered.

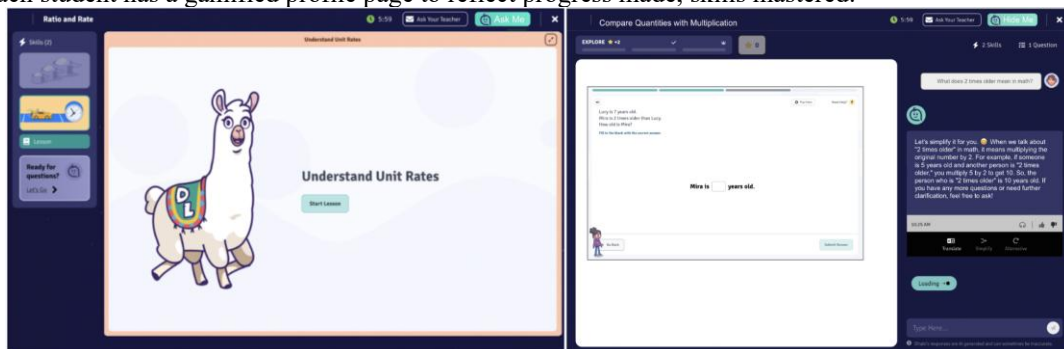


Fig. 1 GenAI tutoring agent integrated into the K12 Math ITS

4.3 GenAI Tutoring Agent

The GenAI tutor is powered by OpenAI's GPT 4.0 model, Fig. 2 presents its technical architecture. Tutor AI leverages its privately hosted endpoints via secure cloud platform, ensuring that no user data is shared or stored for external training purposes. The model is further trained using in-house data including math domain-specific content, anonymised metadata from the ITS, and student performance data. Continuous data feedback loops are in place to monitor and improve response quality.

The main pedagogical support are conceptual understanding, step-by-step scaffolding, personalized instructions based learner proficiency (e.g. step-by-step explanations with simplified language for struggling learners), and on-demand practice recommendation based on skills learners previously struggled with, powered by the question generation agent. Additionally, cognitive load support features (i.e. multilanguage translation, text simplification, and putting math in real-life situations) are provided on-demand, based on field research experience of local student characteristics and learning challenges. During the entire Q&A and scaffolding process, learners are not provided with final answers or next steps automatically, but guided to propose next steps. When learners engage with Tutor AI for off-task chats, Tutor AI guides the chat to be for math only.

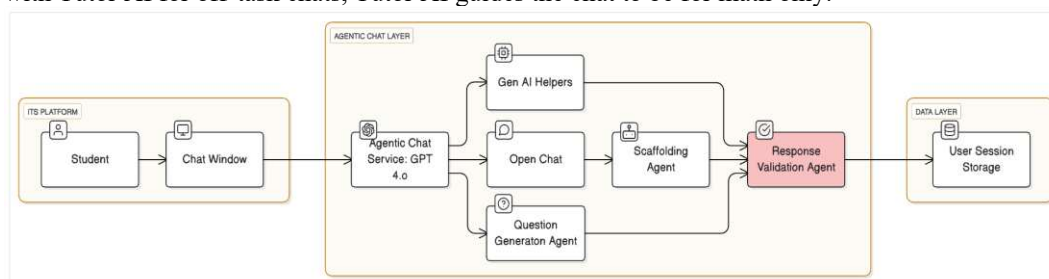


Fig. 2 GenAI tutoring agent technical architecture

The design of the GenAI tutoring agent are localized, child-friendly characters designed to engage with local students. The visual look of the tutoring agent reflects the task that the agent is taking on while interacting with students in the entire ITS journey, not only within the chat function. The purpose of the agent is also to guide students in the entire self-regulated learning journey within the ITS. The non-cognitive skill intended to foster is self-regulated learning. Additionally, when students first onboard the product, a list of questions are sent to students from the GenAI tutoring agent for two purposes: 1) students get familiar with how to interact with a dialog-based AI tutor; 2) capture key learner characteristics (e.g. resilience, interest, etc.) to improve personalized conversation in the future.

5 Methodology

5.1 Data Description

Adaptive diagnostic tests were administered to students at the beginning and end of the academic year. Student test scores at the beginning of the year serve as baseline measurements, while learning outcome was measured by the end-of-year diagnostic test results. To ensure the validity of the analysis, we excluded data from students who completed both diagnostic tests in less than 15 minutes.

Based on baseline diagnostic test results using MetaMetrics' Quantile Grade Level Charts (<https://hub.lexile.com/quantile-grade-level-charts/>), students were segmented into four performance groups:

- Below grade (baseline diagnosed grade level < 1 actual grade level)

- Approaching grade (baseline diagnosed grade level is approaching actual grade level, within 1 grade level)
- At grade (baseline diagnosed grade level equals actual grade level)
- Above grade (baseline diagnosed grade level ≥ 1 actual grade level)

Learner characteristic data, including baseline diagnostic test scores, grade level, gender, and school organization data (i.e. school district information) were used to create a control group, using PSM (Propensity Score Matching). More details on propensity score matching, control and treatment group are discussed in the following section.

To address research questions two and three, log data from the GenAI tutor was collected to investigate how students have interacted with the GenAI tutoring agent. In total, 6794 utterances were recorded. We applied the MPnet V2 model straight on the raw chat logs, as it easily handles typos and colloquial words, so we could preserve the real flavour of students' conversations for clustering.

5.2 Quasi-experimental design

To answer research question one, a quasi-experimental design was employed to measure the impact of the GenAI tutor on students' math learning outcomes. This approach was chosen because it provides causal evidence to feasibly evaluate EdTech solutions in real-life implementation settings. Specifically, the study utilized Propensity Score Matching (PSM) techniques to construct a control group closely resembling the treatment group in characteristics.

A student was considered treated if, at any point during the academic year, a) launched at least one chat session with GenAI tutor and b) exchanged three or more conversational exchanges within that session. Students who have never met conditions a) and b) constitute the eligible control pool. If a treated student later stopped using the GenAI tutor, they remained in the treatment group, consistent with an intention-to-treat perspective. Therefore, the treatment group consists of students engaged with the GenAI Tutor (based on conditions a and b), while the control group comprises students who were not engaged with the GenAI tutor under the defined conditions.

Propensity score matching is widely regarded as an effective approach to measure the impact of intervention programs and is highly recommended by researchers¹⁷. We have conducted this analysis with the use of the causallib library provided by IBM in python²¹. First, a set of background characteristics were selected for each student, namely grade, gender, and baseline diagnostic test scores and school organization information, to account for pre-existing student differences and ensure balanced covariate distributions between treatment and control groups in the propensity score model. Second, logistic regression was used to train a model where the binary outcome was treatment status (yes/no), and their background characteristics, along with their baseline diagnostic test scores were used as predictor variables. This process resulted in a propensity score for each student, representing their probability of being assigned to the tutor program based on their background characteristics.

Next, a nearest-neighbour matching algorithm was applied to identify the nearest control group matches for treatment students with similar propensity scores. The threshold for matching was set at 0.1; this means we only paired students whose propensity scores differed by less than 0.1 in absolute terms. After applying PSM, the distributions of background characteristics between the treatment and control groups were compared to ensure balance. The distribution of propensity scores is shown in Fig. 3. In Fig. 3, the orange 'unweighted' markers depict covariate differences before matching, whereas the blue 'weighted' markers show the balance achieved after applying propensity-score weights. The impact of the GenAI tutor was assessed by comparing the end of year diagnostic test scores between the treatment group and their matched pseudo-control group. To estimate the confidence intervals for the results, 100 bootstrap simulations were conducted, generating 95% confidence

intervals. As shown in Fig.3 below, there is a very small absolute mean difference in the attributes of two groups, indicating the treatment and control groups are the same.

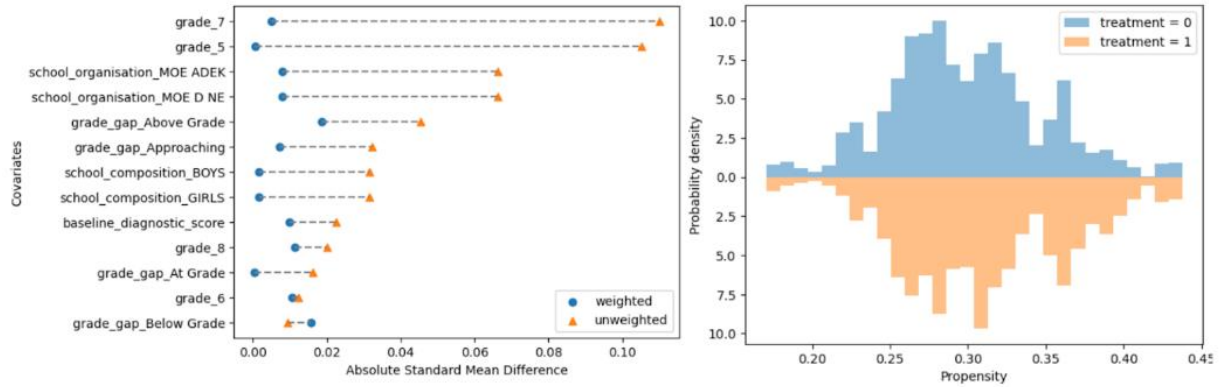


Fig. 3. Differences between control group and treatment group after applying PSM

Following matching and data cleaning, we created two comparable groups of 814 students each: one treatment group and one control group, drawn from the total 3,771 students originally onboarded to use the GenAI tutor on the math ITS.

5.3 GenAI-Child Discourse Analysis

According to City et al.¹⁸ Task predicts performance and skill development comes from interactive tasks learners carry out. The tasks learners engage with GenAI Tutor are conversational. Investigating types of conversations initiated by learners is key to understanding a) without prompting guide or training, how UAE K12 students are using GenAI; b) what kind of help they are requesting from a GenAI agent and how that relates to learning outcomes.

Data cleaning was intentionally avoided prior to applying MPnet V2 model on student's interaction data as it helps to preserve the natural structure and nuances of conversations. The MPnet V2 model is robust to noisy and informal text and it manages to capture semantic meaning effectively even with typos or colloquial language. Having raw data ensured the clustering process to reflect authentic interaction patterns.

In order to analyse the conversation between students and the GenAI tutor, clustering techniques have been applied. Prior to applying clustering techniques, we first generated sentence embeddings for messages (numeric representations of text in matrix form) that enabled clustering methods to identify the underlying semantic patterns within each cluster. To be more specific, a multilingual transformer model SBert MPnet v2: the paraphrase-multilingual-mpnet-base-v2 model provided by sentence BERT was used to generate sentence embeddings for the chat interaction data, resulting in an embedding matrix 6794x768. This model was chosen for its strong semantic representation capabilities across more than 50 languages, which is particularly important in the UAE context where student utterances may reflect multilingual influences. Its ability to generate high-quality, context-rich sentence embeddings makes it well-suited for short, conversational data typically seen in K12 learner-AI interactions. The model also enables efficient comparison of sentence-level meaning, supporting meaningful clustering based on semantic similarity rather than surface-level features.

K-means clustering was applied to cluster the embeddings. The number of the identified clusters was decided based on the avg. within centroid distance (1.47) and the elbow plot. At last, six clusters have been identified from 6794 utterances. Then, authors of this paper worked together to analyse the examples from each cluster and discussed types of conversations conducted between Grade 5 to Grade 8 learners and the GenAI tutoring agent.

6. Results

6.1 Comparison between control group and treatment group

T-tests were conducted to compare the learning outcomes between the control group and the treatment group. Growth is measured as a percentage change in results from baseline (i.e. begin-of-year diagnostic test) to end-of-year diagnostic for the treatment group measured relative to the control group's improvement, reflecting the overall improvement effect of the treatment. After matching, the mean baseline diagnostic test score is 430 for both treatment and control groups.

$$\text{Performance Growth} = \frac{\text{End-of-year diagnostic Score} - \text{Begin-of-year diagnostic Score}}{\text{Begin-of-year diagnostic Score}}$$

Table 1 presents an overall comparison of the two groups. The results indicate that the treatment group (Mean = 501.37, SD = 144.01) exhibited significantly higher performance on the end-of-year diagnostic test compared to the control group (Mean = 481.23, SD = 142.25), $p = 0.004$, Cohen's $d = 0.141$. Additionally, the treatment group (Mean = 70.72, SD = 130.01) demonstrated significantly higher performance growth than the control group (Mean = 50.36, SD = 131.2), $p = 0.002$, Cohen's $d = 0.156$.

Table 1. Comparison between the treatment group and the control group

	Treatment group			Control Group			P value	Cohen's d
	n	Mean	SD	n	Mean	SD		
End-of-year Test Score	814	501.37	144.01	814	481.23	142.25	<.01	.14
Performance Growth	814	70.72	130.01	814	50.36	131.2	<.01	.16

A comparison was also conducted at four performance segmentation levels using a t-test. As Table 2 shows, there was a significant difference between the treatment group and the control group at the below-grade segmentation level. At this level, students in the treatment group (Mean = 485.31, SD = 141.8) achieved significantly higher end-of-year diagnostic scores than students in the control group (Mean = 468.11, SD = 140.4). However, no significant differences were observed at the other three levels, which might be due to smaller sample sizes in other performance segmentation groups.

Table 2. Comparison between the treatment group and the control group, by performance segmentation groups

	Treatment Group			Control Group			P value	Cohen's d
	n	Mean	SD	n	Mean	SD		
Performance Seg								
Below Grade	666	485.31	141.8	666	468.11	140.4	.02	.12
Approaching	91	544.03	109.99	91	511.57	115.17	.05	.29

At Grade	27	618.37	148.89	27	587.52	136.44	.43	.22
Above Grade	30	623.33	152.26	30	584.63	170.38	.36	.24

6.2 Clustering result of student-GenAI conversation

In order to explore what conversation has been conducted between students and the GenAI tutor, K-means clustering was conducted. Based on the Elbow plot, six clusters have been identified. The average centroid distance is 1.47. Figure.4 shows the number of questions raised by students in each cluster.

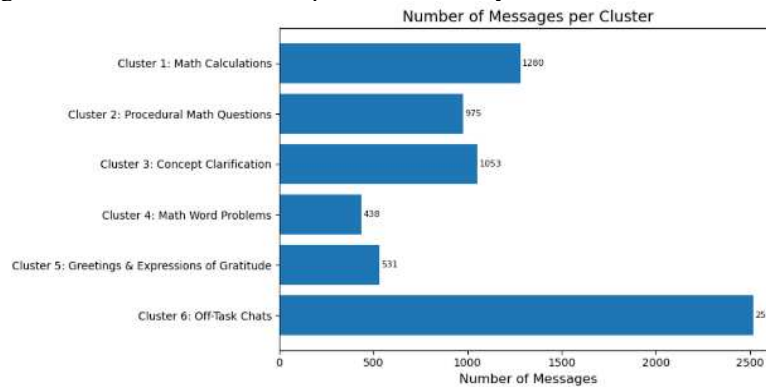


Fig. 4. An overall distribution of conversations from different clusters

Math Calculations. In this cluster, students ask the AI tutor to provide results for math calculations. For example: “ $56 \div 4$ ”, “ 7×49 ”, and “ $2 \times 5 \times 6$ ”. In total, 1280 messages from 436 students have been clustered. By design, the AI tutor avoids providing final answers but guides students to solve the problem step by step.

Procedural Math Questions. The second cluster presents questions about procedural math, which requires students to describe the operations of the number in sentences. For example: “*what is the sum of 4 and 10*”, “*Write a division sentence with a remainder for 25 divided by 4*”, and “*Compare the fractions $\frac{1}{3}$ and $\frac{2}{5}$ using a benchmark of $\frac{1}{2}$* ”. In total, this cluster includes 975 messages from 346 students. The AI tutor avoids providing final answers to procedural math questions. For example, when a learner asked “*what is the sum of 4 and 10*”, AI tutor responded,

“*Let's solve the addition problem step by step! 😊*”

1: To find the sum of 4 and 10, simply add the two¹ numbers together: $4 + 10$.

Give it a try and let me know what you get! 🤔”

Concept Clarification. The third cluster includes students’ requests to clarify the mathematical concepts or provide examples for the concepts. This cluster includes questions phrased by students themselves to enquire various levels of conceptual understanding to get the necessary help needed. In comparison, other clusters might have questions cited from other resources, despite the fact that the AI tutor avoids providing a final answer. For example: “*What are some examples of quadrilaterals?*”, “*Can you identify the unit rate in the given graph?*”, and “*why is importantbt to graph equation*”. In total, 388 students asked 1053 questions for these concept

¹ Note: Quoted italicized sentences (e.g. $56 \div 4$, *thank you my kind brother*) in section 6.2 Clustering Result of Student-AI Conversation are direct quotes of student responses with AI, including emojis (e.g. 🐾 😊) and grammatical errors (e.g. *importantbt*). The purpose is to present K12 students real interaction with GenAI tutor.

clarifications. For example, when students ask, “*why is it important to graph equations*”, the AI tutor provides benefits of graphing and provides practice challenges for the learner to try graphing.

Math Word Problems. In this cluster, students asked more complex math word problems. For example, “*If a recipe calls for 2 cups of flour and 3 eggs to make 12 cookies, how much flour and how many eggs would you need to make 36 cookies*”. These questions require students to comprehend and extract math concepts first, then solve the problem and apply math in real-life situations. In total, 438 messages from 185 students were clustered. When students raise this type of question, the first challenge learners face normally is English comprehension and extracting math concepts. AI Tutor breaks the word problem into smaller steps for learners to solve, instead of providing all the scaffolded steps in one go.

Apart from the questions related to maths, there were also some questions without maths communication.

Greetings & Expressions of Gratitude. This cluster consists of greetings, farewells, and polite acknowledgments from students. For example: “*Hiiii!! i am XXXX nice to meet you*”, and “*thank you my kind brother*”. In total 266 students sent 531 messages in this cluster.

Off-task Chats. The last cluster contains random chats from students, which is not relevant to the learning tasks. Students might send this type of message because of their curiosity. They tried to explore the boundary of the AI tutor. For example: “*what animal you love*”, “*let chat a bit*”, and “*No i just want you to be my friend*”. In total, 2517 messages from 407 students have been clustered. For example, when a student asks, “*what animal you love*”, the AI tutor responds by having a direct response on shared love for animals, and then guides the conversation back to math questions, “*It sounds like you're curious about animals! 🐾 While I love all animals, I'm here to help you with math questions. If you have any math-related queries or need assistance with a math problem, feel free to ask! 😊*”

6.3 Comparison between students with different math conversations

The third research question concerns whether students initiated different math discourse with the AI tutor and achieved different learning outcomes. To be more specific, 814 students were divided into two groups based on their chat history with the AI tutor. Students who asked more math-related questions(i.e. sum of frequencies of Cluster 1, 2, 3, and Cluster 4) than non-related questions(i.e. sum of frequencies of Cluster 5, and Cluster 6) were put in the first group, named math talk group. Otherwise, students who asked more non-related questions were put in the off-task group. T-test was conducted to compare the scores of the end-of-year diagnostic test between these two groups. Table.3 showed the result of the t-test. Math Talk Group (Mean = 514.05, SD = 138.39) showed significantly higher learning gains in the end-of-year diagnostic test than the Off Task Group (Mean = 473.11, SD = 152.35), $p < 0.001$, Cohen’s $d = 0.287$. It means that students who asked more math-related questions performed significantly better than students who asked more non-related questions during their one year learning journey.

Table 3. Comparison between the math talk group and the off task group

	Math Talk Group			Off Task Group			P Value	Cohen’s d
	n	Mean	SD	n	Mean	SD		
Performance	562	514.05	138.39	252	473.11	152.35	<.001	.287

7. Discussions, Future Work, AI Ethics for Child-AI Interactions

As mentioned earlier, this study attempts to bridge the gap in 1) lack of real-world, large scale, long-term causal impact study of GenAI tutor on K12 student learning, particularly integrating GenAI tutor to enhance one-to-one interaction granularity on ITS; 2) lack of literature on the nuanced interaction between K12 students and dialog-based GenAI tutor, and how such interactions affect learning outcomes.

The first research question addresses the real-world, large-scale, one-year impact of the treatment (i.e. one-year engaged use of GenAI Tutor on a K12 math ITS) by comparing student performance and learning gains between treatment and control groups. The results showed that engaged use of GenAI tutor improves student math performance on ITS, compared to students who are not engaged with the tutor on ITS. Dialog-based tutoring agent, by design, allows individual learners to initiate interactive Q&A that are related to pinpoints of the specific learner real-time, which assumably could enable learner agency (e.g. voluntary self-assessment and adjustment of learning strategies) and interaction granularity (e.g. individual control over what and how AI feedback is shared and utilized per each learner) as previous literature highlighted. The GenAI tutor feature on ITS functions on-demand, which indicates student self-selection into help-seeking real time with AI. Assumably, students who are actively seeking help, is an indicator of motivation and active problem-solving behavior. Additionally, the results revealed that the difference between the treatment and control groups was significant for below-grade-level students but not for approaching, at grade and above grade students. We assume this might be due to substantial smaller sample sizes observed in approaching, at grade and above grade student groups. This finding is still very important within the studied educational context, because the actual performance of the majority of the student population is below grade level, and GenAI tutor is proven helpful to improve learning outcomes for this segment of the student population.

For future research, large-scale, longitudinal investigation of GenAI tutor on student learning within the ITS will be investigated, especially including other performance segmentation groups (e.g. approaching, at grade and above grade) that represent the student population within the educational context. We assume the kind of AI support and interaction design would vary a lot depending on types of student segmentation (e.g. top performers vs. below grade performers). It is also important to track the performance changes of below grade students longitudinally to understand if there is continued active use of GenAI tutor and investigate student active problem-solving skills, if enhanced by GenAI tutor. Moreover, the studied math ITS offered students the option to opt for either asking questions from teacher or from the GenAI tutor (mentioned previously), the researchers will compare the usage of both, and explore if there are key differences in learner characteristics in the two types of behavioral choices.

The second research question delves into the nuances of the interactive task children have engaged with the GenAI tutor, namely, math discourses between GenAI tutor and learners. Through discourse analysis, six discourse clusters were identified. Over 37% of the discourse initiated by students was off-topic, and the authors further discovered that students from lower grade level tend to have higher percentage of off-task conversations (i.e. cluster 6): Grade 5 (44%), Grade 6 (33%), Grade 7 (24%), Grade 8 (24%). Surprisingly, this is similar to the literature highlighted by Wang, D. et al. (2025) through analyzing conversation between human tutor and students, “off task chatting accounted for a significant proportion of the tutoring session (over 30%) and high school students exhibited greater engagement and cognitive abilities than primary and middle school students through more active participation and superior reasoning skills”. This implies that age and corresponding cognitive or self-regulatory abilities might be predictors of K12 student engagement and effective use of GenAI. Additionally, if the nature of task interaction is conversational, either between human tutor and K12 learners, or between GenAI tutor and K12 learners, part of the interaction is casual chat. However, it is important to investigate longitudinally, if the chat interaction between GenAI tutor and learner is consistently off-task; if yes, it is important to find out

who these learners are, what are the reasons the engagement is consistently off task. Possibly, learners might not be interested in using GenAI for help, or might not be motivated learners, or simply do not know how to effectively use the tool without explicit task guidance due to age or cognitive and metacognitive abilities. Second, the on-task math discourse clusters (i.e. math calculation, procedural math, concept clarification and math world problem) show the type and level of math questions Grade 5 to Grade 8 students raised. Concept clarification cluster includes questions phrased or formulated by students, with much less likelihood that the questions are cited from elsewhere and implies more active engagement. But other on-task clusters could be questions children cited from elsewhere where, for instance, questions from the ITS system or elsewhere.

Then, the authors compared student performance between those that ask more math-related questions and those that raise less on-task questions throughout the entire academic year. The results showed that students engaged in more on-task mathematical discourse achieved significantly higher scores on the end-of-year diagnostic test. The above findings indicate that the support offered by GenAI tutor was most evident when students use the tool effectively and consistently. This aligns with prior research indicating that not all discourse effectively promotes mathematics learning¹⁰. Different types of mathematical discourse may have varying effects on learning outcomes. Future research should 1) investigate frameworks and best practices on K12 math discourse within the studied education context, and leveraged that into GenAI-learner interaction designs that guide students to better utilize GenAI tutor help; 2) investigate learners that consistently use the GenAI tutor on-task longitudinally, particularly finding out key learner factors that related to this behavior. These insights would be very useful to improve K12 student and GenAI interaction features for effective pedagogical designs.

Lastly, as mentioned earlier, the GenAI tutor was also designed as a child-friendly localized pedagogical agent that guides learners throughout the ITS, not only within chat. Hence, for future research, it is important to investigate the impact of the GenAI pedagogical agent on student self-regulated learning within the ITS, and how PA design features (including visual character, tasks, discourse style appear in different tasks, gamified incentives) impact on student engagement and learning outcomes. Additionally, as mentioned earlier, to onboard learners a list of questions were administered to capture important learner characteristics (e.g. resilience, interests, etc.), in addition to getting learners familiar with the GenAI tutor. The main purpose is to potentially personalize discourse interaction according to key learner characteristics. Hence, an important future R&D direction is to build self-improving K12 GenAI pedagogical agents that have important memory and understanding of each learner. That includes, key characteristics such as interest and resilience, interaction preference (e.g. texting or using natural language, preference over visual support or analytical text comprehension), existing behavioral pattern with GenAI agent (e.g. if the student is patient to follow through conversations with GenAI agent or skipping through tasks), etc.

When it comes to the above nuanced designs of GenAI for K12 learners, AI ethics need to be considered. According to Wang, G. et. al (2024)²², there is growing consensus on a set of high-level ethical AI principles; however, the AI-learner engagement ethical principles are limited, let alone how to effectively apply in practice. Given that GenAI will become more pervasive in children's lives now and in the future, the following principles are important considerations within this context:

- 1) The developmental aspect of children should be an important consideration, due to their diverse needs, age ranges, development stage, various skill acquisition needs, contextual background, and key characteristics;

2) Child-centered evaluations should consider children's best interests and rights. Research on 188 AI systems for children revealed that virtually all used only technical performance metrics for evaluation. Quantitative measures create significant assessment gaps, particularly concepts that technical metrics cannot adequately measure in terms of children's development and long term wellbeing.

Hence, when designing K12 GenAI pedagogical agents, especially nuanced interactions with individual learners, K12 children's developmental aspects, engagement, balanced evaluation metrics, data security, learner privacy, and should all be important considerations moving forward.

Limitations.

There are some limitations of this study that need to be noted. Firstly, PSM offers possibilities and advantages to explore the causal impact of an EdTech intervention in a real world implementation setting. However, one drawback is that PSM assumes all relevant confounders are observable and included in the model; if important variables are omitted, the estimated treatment effect may be biased¹⁹. Furthermore, PSM cannot account for unobservable differences between groups, which may lead to residual bias²⁰. Additionally, the matching process can reduce the sample size, which can limit the statistical power of the study. For instance, most of the students in this study were below grade level. More participants from other segmentations were needed for more reliable comparison. Lastly, conversation clustering analysis has a heuristic nature, as the purpose of this study is to explore as many types of question types as possible. However, there could be other clustering approaches, depending on research goals.

Ethics Statement. This study involved a retrospective secondary analysis of fully anonymized usage data generated through students' routine use of the Alef Education learning platform. No data were collected directly from participants, no interventions were carried out, and the research team did not access any personally identifiable information. The use of platform data for research purposes is governed by a formal data sharing agreement between Alef Education and the Abu Dhabi Department of Education and Knowledge (ADEK). As the study was limited to the analysis of pre-existing, de-identified data and was conducted in accordance with the terms of this agreement and applicable data protection regulations, formal ethics committee approval was not required under local regulations.

Consent to Publish. Not applicable. This manuscript does not contain data from any individual person in identifiable form. All findings are derived from fully anonymized platform usage data and are presented in aggregate; therefore, consent for publication from individual participants was not required.

Consent to Participate. This study did not involve direct data collection from participants; all analyses were conducted on pre-existing, fully anonymized usage data from the Alef Education learning platform. Consent for the collection and processing of student data is obtained at the point of platform enrollment, in accordance with ADEK policies and applicable data protection regulations, and the use of data for research purposes is authorized under the data sharing agreement between Alef Education and ADEK. As the study involved only retrospective analysis of de-identified data, the requirement for individual informed consent for this specific research was not applicable.

Data availability. Data is available upon request.

Funding. This research did not receive any specific grant from funding agencies.

Author contributions. X.M. contributed conceptualisation, supervision, literature review, investigation, writing (original, review & editing), project administration, resources; P.K.M. contributed methodology, analysis, data curation, part of writing(original); Q.Z. contributed writing (original), literature review, investigation.

Competing interests. The authors declare no competing interests.

References

1. OECD. PISA 2022 Results (Vol. I): The State of Learning and Equity in Education (OECD Publishing, 2023). <https://doi.org/10.1787/53f23881-en>
2. Miao, X., Mishra, P.K. & Nadaf, A. Evidence and promises of AI predictions to understand student approaches to math learning in Abu Dhabi K–12 public schools. *Gulf Educ. Soc. Policy Rev.* **1**, 109–134 (2021).
3. Li, Y., Zhao, K. & Xu, W. Developing an intelligent tutoring system that has automatically generated hints and summarization for algebra and geometry. *Int. J. Inf. Commun. Technol. Educ.* **11**, 14–31 (2015).
4. Cheng, L., Croteau, E., Baral, S., Heffernan, C. & Heffernan, N. Facilitating student learning with a chatbot in an online math learning platform. *J. Educ. Comput. Res.* **62**, 907–937 (2024).
5. Hwang, G.-J. & Tu, Y.-F. Roles and research trends of artificial intelligence in mathematics education: A bibliometric mapping analysis and systematic review. *Mathematics* **9**, 584 (2021).
6. Aleven, V., Rowe, J., Huang, Y. & Mitrovic, A. Domain modeling for AIED systems with connections to modeling student knowledge: A review. In *Handbook of Artificial Intelligence in Education*, eds. du Boulay, B., Mitrovic, A. & Yacef, K., 127–170 (Edward Elgar Publishing, Cheltenham, 2023).
7. Miao, X., Brooker, R. & Monroe, S. Where generative AI fits within and in addition to existing AI K–12 education interactions: Industry and research perspectives. In *Machine Learning in Educational Sciences: Approaches, Applications and Advances*, ed. Khine, M.S., 359–384 (Springer Nature Singapore, 2024). doi:10.1007/978-981-99-9379-6_17
8. Létourneau, A., Deslandes Martineau, M. & Léger, P.M. A systematic review of AI-driven intelligent tutoring systems (ITS) in K–12 education. *npj Sci. Learn.* **10**, 29 (2025).
9. VanLehn, K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* **46**, 197–221 (2011).
10. Martínez-Téllez, R. & Camacho-Zuñiga, C. Enhancing mathematics education through AI chatbots in a flipped learning environment. In *Proceedings of the 2023 World Engineering Education Forum–Global Engineering Deans Council (WEEF-GEDC)*, 1–8 (IEEE, 2023).
11. Lane, H.C. Pedagogical agents for all: Designing virtual characters for inclusion and diversity in STEM. In *Handbook of Artificial Intelligence in Education*, eds. du Boulay, B., Mitrovic, A. & Yacef, K., 611–614 (Edward Elgar Publishing, Cheltenham, 2023).

12. LangChain. Memory of Agents. LangChain <https://blog.langchain.dev/memory-for-agents/> (2024).
13. Ballard, D. Discourse in math—don't just talk about it [White paper]. Consortium on Reaching Excellence in Education <https://www.corelearn.com/wp-content/uploads/2017/08/discourse-in-math-whitepaper.pdf> (2017).
14. Yang, E.F.Y., Chang, B., Cheng, H.N.H. & Chan, T.-W. Improving pupils' mathematical communication abilities through computer-supported reciprocal peer tutoring. *Educ. Technol. Soc.* **19**, 157–169 (2016).
15. Wang, D. *et al.* Investigating dialogic interaction in K–12 online one-on-one mathematics tutoring using AI and sequence mining techniques. *Educ. Inf. Technol.* **30**, 9215–9240 (2025).
16. Klibanoff, R.S., Levine, S.C., Huttenlocher, J., Vasilyeva, M. & Hedges, L.V. Preschool children's mathematical knowledge: The effect of teacher "math talk." *Dev. Psychol.* **42**, 59–69 (2006).
17. Andriillon, A., Pirracchio, R. & Chevret, S. Performance of propensity score matching to estimate causal effects in small samples. *Stat. Methods Med. Res.* **29**, 644–658 (2020).
18. City, E.A., Elmore, R.F., Fiarman, S.E. & Teitel, L. Instructional rounds in education: A network approach to improving teaching and learning. (Harvard Education Press, Cambridge, MA, 2009).
19. Caliendo, M. & Kopeinig, S. Some practical guidance for the implementation of propensity score matching. *J. Econ. Surv.* **22**, 31–72 (2008).
20. Dehejia, R.H. & Wahba, S. Propensity score-matching methods for nonexperimental causal studies. *Rev. Econ. Stat.* **84**, 151–161 (2002).
21. Shimoni, Y., Karavani, E., Ravid, S., Bak, P., Ng, T.H., Alford, S.H., Meade, D. & Goldschmidt, Y. An evaluation toolkit to guide model selection and cohort definition in causal inference. Preprint at <https://arXiv.org/abs/1906.00442> (2019).
22. Wang, G., Zhao, J., Van Kleek, M. & Shadbolt, N. Challenges and opportunities in translating ethical AI principles into practice for children. *Nat. Mach. Intell.* **6**, 265–270 (2024).