IMPLEMENTATION OF A GENERATIVE AI CHATBOT FOR ELEMENTARY MATHEMATICS LEARNING: EVIDENCE FROM A QUASI-EXPERIMENTAL STUDY IN UZBEKISTAN

Kosimova Nafosat Shoydulla kizi
International Joint Degree Program, Termez State University and Faculty of Mathematics and Physics Education, Universitas Pendidikan Indonesia

Abstract: *Generative artificial intelligence (AI) chatbots are increasingly proposed as tools for personalising instruction, yet empirical evidence from primary mathematics classrooms in Central Asia remains scarce. This study examined the effect of a generative AI-based chatbot on the mathematics achievement, motivation, and classroom engagement of elementary school students in Uzbekistan. Using a quasi-experimental pre-test/post-test design, 60 students in Grades 4–6 were divided into an experimental group (n = 30), which used the chatbot as a supplementary learning tool, and a control group (n = 30), which received conventional teacher-led instruction over an eight-week intervention. Data were collected through achievement tests, a motivation questionnaire, classroom observation, and chatbot interaction logs. The experimental group's mean score rose from 61.40 to 79.20 (gain = 17.80), compared with the control group's rise from 60.80 to 70.10 (gain = 9.30). Students who used the chatbot reported high perceived usefulness (M = 4.32/5), motivation to practise (M = 4.25/5), and comfort asking questions (M = 4.18/5). Observation data indicated stronger participation and independent problem-solving in the experimental group, and interaction logs showed that more frequent chatbot use was associated with larger gains. These findings suggest that generative AI chatbots can meaningfully support elementary mathematics instruction when integrated as a supplement to, rather than a replacement for, teacher guidance. Implications for curriculum alignment, teacher training, and infrastructure policy in Uzbekistan and comparable developing education systems are discussed, alongside limitations related to sample size, study duration, and the absence of inferential statistical testing.*

Keywords*:* *generative artificial intelligence; educational chatbot; elementary mathematics; learning motivation; Uzbekistan; AI in education*

1. INTRODUCTION

Artificial intelligence has reshaped instructional delivery worldwide, and generative AI chatbots — systems capable of producing human-like, context-sensitive dialogue — are now positioned as tools for personalised tutoring at scale (Holmes et al., 2021; Kasneci et al., 2023). Because mathematics learning depends heavily on iterative practice, timely correction, and step-by-step reasoning, it is often cited as a subject well suited to chatbot-based support (Kumar et al., 2020; Chorianopoulou & Grigoriadou, 2021). International assessments continue to document persistent weaknesses in primary-level arithmetic, fraction reasoning, and word-problem interpretation (OECD, 2021), and researchers argue that these gaps often stem from insufficient individualised feedback and constrained teacher time (Hattie & Timperley, 2007; Shute, 2008).

Large-scale AI tutoring initiatives in China, Singapore, South Korea, and several Western countries have reported gains in accuracy, motivation, and homework completion attributable to AI-supported mathematics instruction (Chung et al., 2024; Holmes et al., 2021). Underlying this progress is the emergence of transformer-based large language models, which allow chatbots to generate novel, context-aware explanations rather than retrieving pre-scripted responses (Vaswani et al., 2017; Brown et al., 2020). These systems can decompose problems into sequential steps, adapt explanations to a learner's apparent level of understanding, and sustain multi-turn dialogue — capabilities that align with established learning-science principles such as scaffolding and social learning (Bandura, 1997; Bruner, 1996).

In Uzbekistan, the "Digital Uzbekistan 2030" programme has made digital transformation a national education priority (Ministry of Digital Technologies of Uzbekistan, 2023), yet systematic use of AI for personalised primary-level learning remains limited (Mirzaev, 2022). National assessment data show recurring difficulties in arithmetic, fractions, and word problems (State Inspectorate for Education Quality, 2022), while classroom practice remains largely teacher-centred, restricting opportunities for differentiated support (Sattarov, 2023). Despite the country's growing digital infrastructure, no identified empirical study has examined the effect of a generative AI mathematics chatbot within Uzbek primary classrooms, nor how Uzbek-speaking children respond to AI-mediated instruction. This gap is significant both academically — given the near-total absence of Central Asian evidence in the broader AI-in-education literature (Kasneci et al., 2023) — and practically, given the country's stated policy commitment to digital learning reform.

This study addresses that gap by evaluating the implementation of a generative AI chatbot as a supplementary tool for elementary mathematics instruction in an Uzbek primary school. Specifically, it asks:

1. Does chatbot-supported learning produce higher mathematics test gains than traditional instruction alone?
2. How do students perceive the usability, comfort, and motivational value of chatbot-assisted learning?
3. What benefits and limitations emerge when a generative AI chatbot is integrated into classroom mathematics teaching?

By combining achievement data, perception measures, classroom observation, and system usage logs, the study offers a multi-source empirical account of chatbot-assisted mathematics learning in a context that has received little prior research attention, while also identifying practical conditions — teacher supervision, curriculum alignment, and infrastructure support — under which such tools are likely to be effective.

2. Literature Review

2.1 AI chatbots in education: global evidence

Chatbots in education evolved from rule-based dialogue systems with fixed response scripts (Fryer & Carpenter, 2006) toward generative, transformer-based systems capable of dynamic, learner-adapted dialogue (Dale, 2016; Kasneci et al., 2023). This evolution matters pedagogically because static instructional tools impose uniform pacing, whereas generative chatbots can diagnose misconceptions and adjust explanations in real time (VanLehn, 2011;

Johnson et al., 2020). A substantial body of evidence links AI tutoring to improved outcomes across educational levels and subjects, including higher-education computer science instruction (VanLehn et al., 2005) and language learning (Kerly, Hall, & Bull, 2007), with meta-analytic reviews generally reporting moderate-to-large positive effects relative to conventional instruction (Kulik & Fletcher, 2016).

Within mathematics specifically, chatbot interventions have been associated with improved procedural fluency, reduced error frequency, and greater persistence with difficult problems (Table 1). Kumar et al. (2020) reported significant gains in arithmetic accuracy among Indian primary pupils using a step-by-step chatbot, while Chorianopoulou and Grigoriadou (2021) found that adaptive chatbot feedback improved algebraic reasoning among Greek primary students. Motivational effects have also been widely documented: Shin and Pierre (2021) observed reduced mathematics anxiety among South Korean elementary pupils using an AI tutor, and Sundararajan and Ma (2020) reported stronger engagement and quiz participation among Singaporean primary learners. Biswas and Sengupta (2024) extended this evidence to a resource-constrained setting, finding meaningful multiplication and division gains among rural Bangladeshi students — a finding of direct relevance to developing-country contexts such as Uzbekistan.

Table 1. Selected global evidence on chatbot-supported mathematics learning

Study	Setting	Design	Focus	Key finding
Kumar et al. (2020)	India, primary	Experimental (n = 52)	Arithmetic practice	Improved addition/subtraction accuracy and enjoyment
Chorianopoulou & Grigoriadou (2021)	Greece, primary	Quasi-experimental (n = 89)	Algebra support	Enhanced reasoning, reduced error frequency
Shin & Pierre (2021)	South Korea, elementary	Mixed methods (n = 60)	Motivation	Reduced anxiety, greater persistence
Alamri et al. (2019)	Saudi Arabia, secondary	Pre–post (n = 70)	Q&A chatbot	Improved procedural fluency
Ruan et al. (2022)	China, secondary	Longitudinal (n = 84)	Personalised hints	Increased homework completion
Sundararajan & Ma (2020)	Singapore, primary	Case study	Motivation-centred chatbot	Stronger engagement, higher confidence
Biswas & Sengupta (2024)	Bangladesh, primary	Field trial (n = 122)	Adaptive chatbot	Gains in multiplication/division; effective for rural students

Note. Adapted from the source thesis (Qosimova, 2025), based on the cited primary studies.

Despite this evidence, limitations are consistently reported. Generative models can produce fluent but factually incorrect output — a phenomenon termed "hallucination" — which is particularly consequential in a domain such as mathematics where a single incorrect step can mislead a learner (Ji et al., 2023). Scholars therefore emphasise that chatbot effectiveness is contingent on curriculum alignment, teacher facilitation, and human oversight rather than on the technology operating autonomously (Winkler & Söllner, 2018; Luckin et al., 2016).

2.2 Mathematics learning difficulties in primary education

Difficulties in primary mathematics are well documented and multifactorial. A persistent gap exists between procedural performance and conceptual understanding, with many children able to execute familiar algorithms but unable to transfer reasoning to novel problems (Hiebert & Grouws, 2007; Rittle-Johnson, Siegler, & Alibali, 2001). Word-problem interpretation is a further common obstacle, as students often extract numbers and apply operations mechanically rather than modelling the underlying situation linguistically (Verschaffel, Greer, & De Corte, 2000). Mathematics anxiety compounds these difficulties: negative affective responses to mathematics reduce working-memory availability and persistence, creating a self-reinforcing cycle of avoidance and underachievement (Ashcraft & Krause, 2007; Dowker, Sarkar, & Looi, 2016).

Teacher-related and systemic factors also contribute. Primary teachers are frequently generalists with variable confidence in mathematical content knowledge (Ball, Thames, & Phelps, 2008), and large class sizes limit the individualised, timely feedback that formative-assessment research identifies as central to improvement (Black & Wiliam, 1998; Hattie & Timperley, 2007). These constraints are especially salient in the Uzbek context, where classroom observation research indicates a predominantly teacher-centred instructional style with limited scope for differentiated practice (Sattarov, 2023).

2.3 Generative AI: theoretical and technological foundations

Generative AI models synthesise novel outputs by learning statistical regularities from large training corpora, in contrast to earlier retrieval- or rule-based systems (Goodfellow et al., 2014). The transformer architecture (Vaswani et al., 2017), which uses self-attention to model relationships across an entire input sequence, underlies contemporary large language models and has substantially improved the coherence and contextual relevance of generated explanations. Pre-training on broad corpora followed by task-specific fine-tuning (Brown et al., 2020) allows such models to be adapted for domain-specific applications, including mathematics tutoring, while reinforcement learning from human feedback helps align model outputs with pedagogical and safety expectations (Christiano et al., 2017).

These capabilities map onto established learning theories. Cognitive load theory suggests that instructional content should be sequenced to avoid overwhelming working memory (Sweller, 2010), a property that step-by-step chatbot explanations can support directly. Social learning and constructivist perspectives emphasise the value of dialogic, learner-driven interaction for building understanding (Bandura, 1997; Bruner, 1996) — conditions that conversational AI is structurally suited to provide. At the same time, the probabilistic nature of generative text production introduces the risk of confidently stated inaccuracies (Ji et al., 2023), underscoring the need for teacher oversight when such tools are deployed with children.

2.4 The Uzbek context and research gap

Uzbekistan's national digital transformation strategy has prioritised ICT expansion and digital literacy in schools (Khodjayev, 2021; Ministry of Digital Technologies of Uzbekistan, 2023), in line with UNESCO recommendations for Central Asia (UNESCO, 2022). However, adoption of AI specifically for personalised learning remains at an early and uneven stage, with rural–urban disparities in connectivity and teacher digital readiness well documented (Sattarov & Yuldashev, 2023; Yakubov, 2023). Existing Uzbek-focused scholarship addresses ICT infrastructure, teacher preparedness, and digital-textbook development (Nazarov & Soliev, 2023), but no identified study has empirically tested a generative AI mathematics chatbot with primary-age learners, nor examined Uzbek-language chatbot interaction specifically. This absence is consistent with a broader pattern in the international literature, which is dominated by North American, European, and East Asian evidence, leaving the transferability of chatbot effectiveness claims to Central Asian, multilingual, and lower-resource contexts largely untested (Kasneci et al., 2023). The present study addresses this gap directly by evaluating a generative AI mathematics chatbot with Uzbek elementary students under controlled classroom conditions.

3. Materials and Methods

3.1 Research design

This study used a quasi-experimental pre-test/post-test control-group design (Campbell & Stanley, 1966), comparing an experimental group that used a generative AI chatbot as a supplementary mathematics learning tool with a control group that received conventional teacher-led instruction only. Both groups were taught by the same mathematics teacher and followed an identical curriculum sequence over an eight-week period, isolating chatbot use as the primary instructional variable of interest. The design combined quantitative achievement and perception data with qualitative classroom observation, following a triangulated mixed-data approach intended to strengthen interpretive validity (Tashakkori & Teddlie, 2010).

3.2 Participants and sampling

Participants were 60 students in Grades 4–6 (ages approximately 9–12) at a public primary school in Uzbekistan, an age range considered developmentally appropriate for chatbot interaction given adequate reading comprehension and prior exposure to the arithmetic, fraction, and word-problem content targeted by the study (Sarama & Clements, 2009; Siegler et al., 2012). The school was purposively selected for its digital-device access, internet connectivity, and administrative willingness to participate (Etikan, Musa, & Alkassim, 2016); students were then randomly assigned to the experimental (n = 30) or control (n = 30) group to reduce selection bias. Random assignment and the use of a single teacher for both groups were intended to control for instructor- and ability-related confounds (Campbell & Stanley, 1966). Participation required parental consent and basic Uzbek literacy; students with severe cognitive or behavioural support needs, or with attendance patterns likely to disrupt continuity, were excluded. Ethical safeguards included anonymised student identifiers, confidential data handling, the right to withdraw at any time, and post-study chatbot access for the control group.

3.3 Instruments

Four instruments were used to capture achievement, motivation, behavioural, and usage data: (1) a curriculum-aligned mathematics achievement test, administered in parallel pre- and post-test forms covering arithmetic, word problems, and conceptual items, reviewed by two primary mathematics teachers for content validity; (2) a Likert-scale motivation and perception questionnaire adapted from established motivation frameworks (Deci & Ryan, 2000), with additional chatbot-specific items for the experimental group; (3) a teacher-completed classroom observation checklist recording participation, attention, and problem-solving behaviour; and (4) automated chatbot interaction logs recording session frequency, question volume, topic focus, and duration for the experimental group. Table 2 summarises the intended validity and reliability procedures for each instrument.

Table 2. Instruments, validity, and reliability procedures

Instrument	Purpose	Validity procedure	Reliability target
Mathematics pre-/post-test	Measure achievement change	Expert review; curriculum alignment; piloting	Cronbach's $\alpha \geq .80$
Motivation & perception questionnaire	Assess attitudes and chatbot perception	Expert content review; translation check	Cronbach's $\alpha \geq .75$
Observation checklist	Document engagement behaviour	Expert validation of behavioural constructs	Inter-observer agreement ≥ 85%
Chatbot usage logs	Record interaction frequency and focus	System-level functional validation	Automated tracking accuracy check

3.4 Procedure

Following instrument piloting and ethical clearance, all students completed the mathematics pre-test under identical conditions, after which they were randomly allocated to groups. During the eight-week intervention, the control group received standard textbook-based, teacher-led instruction, while the experimental group used the chatbot during classroom practice periods and at home, under teacher supervision that ensured age-appropriate, curriculum-aligned use without providing direct solution steps. Weekly classroom observations were conducted in both groups to document engagement. Following the intervention, both groups completed the post-test under matched conditions, and the motivation and perception questionnaire was administered (with chatbot-specific items limited to the experimental group). All data — test scores, questionnaire responses, observation records, and usage logs — were anonymised and stored securely prior to analysis.

3.5 Data analysis

Test scores, motivation ratings, and usage-log data were entered into a coded spreadsheet by group and analysed descriptively (means, standard deviations, minimum/maximum values) to compare pre-test equivalence, post-test outcomes, and pre–post gain scores between groups. Motivation and perception items were summarised as mean Likert scores. Observation data were analysed qualitatively by indicator and group. Chatbot usage levels (high, moderate, low, based on session frequency) were cross-tabulated against individual gain scores to explore a possible usage–outcome relationship. Because inferential statistical tests (e.g., independent- and paired-samples *t*-tests) were not computed on the available dataset, the results reported below are descriptive; differences between groups are therefore interpreted as observed rather than statistically confirmed (see Limitations).

4. Results

4.1 Baseline equivalence

Pre-test scores were closely matched between groups, indicating comparable initial mathematics achievement (Table 3): the experimental group's mean score was 61.40 (*SD* = 8.75), and the control group's was 60.80 (*SD* = 8.92), a difference of 0.60 points.

Table 3. Pre-test descriptive statistics

Group	*n*	*M*	*SD*	Min	Max
Experimental	30	61.40	8.75	45	78
Control	30	60.80	8.92	44	77

4.2 Post-test outcomes and learning gains

After the eight-week intervention, both groups improved, but the experimental group showed a considerably larger increase (Table 4). The experimental group's mean score rose to 79.20 (*SD* = 7.10), and the control group's rose to 70.10 (*SD* = 8.05) — a post-test difference of 9.10 points. Mean gain scores were 17.80 points for the experimental group and 9.30 points for the control group, meaning the experimental group's improvement was approximately double that of the control group.

Table 4. Post-test descriptive statistics and gain scores

Group	*n*	Pre-test *M*	Post-test *M*	*SD* (post)	Mean gain
Experimental	30	61.40	79.20	7.10	17.80
Control	30	60.80	70.10	8.05	9.30

4.3 Motivation and perception

Students in the experimental group reported consistently positive perceptions of chatbot-assisted learning (Table 5). Perceived usefulness of chatbot explanations received the highest mean rating (*M* = 4.32), followed by overall satisfaction (*M* = 4.28) and motivation to practise more mathematics problems (*M* = 4.25). Students also reported relatively high comfort asking the chatbot questions (*M* = 4.18) and confidence in problem-solving (*M* = 4.05).

Table 5. Motivation and perception ratings (experimental group)

Indicator	*M* (of 5)	Interpretation
Interest in mathematics learning	4.21	High
Confidence in solving mathematics problems	4.05	High
Perceived usefulness of chatbot explanations	4.32	Very high
Comfort asking questions to chatbot	4.18	High

Motivation to practise more problems	4.25	Very high
Overall satisfaction with chatbot learning	4.28	Very high

4.4 Classroom observation

Teacher observation data indicated that the experimental group displayed higher levels of active participation, sustained attention, question-asking, independent problem-solving, task completion, and general learning enthusiasm than the control group, whose behaviour was more dependent on direct teacher prompting (Table 6). Peer discussion levels were similar (moderate) across both groups.

Table 6. Classroom observation summary

Indicator	Experimental group	Control group
Active participation	High	Moderate
Attention during activities	High	Moderate
Asking questions	High	Low–moderate
Independent problem-solving	High	Moderate
Peer discussion	Moderate	Moderate
Task completion	High	Moderate
Learning enthusiasm	High	Moderate

4.5 Chatbot usage patterns and their relationship to gains

Interaction logs showed that experimental-group students engaged regularly with the chatbot: the average number of sessions per student was 14.6, and the average number of questions asked was 38.2, with 83% of students classified as regular users. The most common topics were multiplication/division and word problems, and average session duration was 12–15 minutes. When students were grouped by usage intensity, those with high usage (n = 12) achieved the largest average gain score (21.40), followed by moderate users (n = 13, gain = 17.20) and low users (n = 5, gain = 11.60), suggesting a positive association between engagement frequency and learning improvement (Table 7).

Table 7. Chatbot usage level and average gain score

Usage level	*n*	Average gain score
High	12	21.40
Moderate	13	17.20
Low	5	11.60

5. Discussion

5.1 Chatbot use and mathematics achievement

The near-equivalent pre-test scores (0.60-point difference) support the assumption that both groups began the study with comparable mathematics ability, providing a reasonable basis for comparing post-intervention outcomes. The considerably larger post-test gain observed in the experimental group (17.80 vs. 9.30 points) is consistent with prior findings that AI-supported tutoring can produce learning gains through adaptive explanation and immediate feedback (VanLehn, 2011; Kulik & Fletcher, 2016). Formative-feedback theory suggests that feedback is most effective when timely, specific, and closely tied to the learner's current task (Shute, 2008); the chatbot's capacity to respond instantly to student questions may have reduced the persistence of misconceptions relative to the delayed feedback typical of whole-class instruction. The results also resonate with Vygotskian scaffolding theory, in that the chatbot appeared to function as a responsive support structure — offering hints and worked examples without directly supplying answers — within students' zone of proximal development.

5.2 Motivation, comfort, and affective outcomes

The consistently high motivation and perception ratings, particularly around perceived usefulness and comfort in asking questions, align with research indicating that non-judgemental, private interaction can reduce the reluctance many children feel about asking questions publicly (Dowker, Sarkar, & Looi, 2016; Shin & Pierre, 2021). This is a pedagogically meaningful finding for mathematics specifically, given the well-documented role of mathematics anxiety in suppressing participation and achievement (Ashcraft & Krause, 2007). The chatbot's conversational format may have provided a sense of social presence that supported sustained engagement, consistent with prior chatbot-motivation research (Fryer & Carpenter, 2006; Winkler & Söllner, 2018).

5.3 Engagement, autonomy, and usage intensity

Observation data showing higher participation and independent problem-solving in the experimental group, together with the positive relationship between usage frequency and gain scores, together suggest that the chatbot supported not only achievement but also self-directed learning behaviour, consistent with self-regulated learning theory (Zimmerman, 2002). The concentration of chatbot queries around multiplication, division, and word problems mirrors well-established difficulty areas in primary mathematics (Verschaffel, Greer, & De Corte, 2000; Siegler et al., 2012), indicating that students used the tool to target genuine points of struggle rather than for incidental interaction.

5.4 Interpretation in relation to the research questions

Regarding the first research question, the experimental group's larger post-test gain suggests that chatbot-supported learning was associated with greater improvement than traditional instruction alone, although — in the absence of inferential statistical testing on the available dataset — this should be read as a descriptively observed difference rather than a confirmed statistically significant effect (see Section 6). Regarding the second question, students reported favourable perceptions of usability, comfort, and motivation. Regarding the third, observation and usage-log data point to benefits in participation and independent practice, tempered by the well-established risk that generative systems can produce confidently incorrect explanations (Ji et al., 2023), which reinforces the need for continued teacher supervision rather than unmonitored AI use.

5.5 Comparison with prior literature

These findings broadly align with international evidence that AI tutoring systems and chatbots can enhance both achievement and engagement (VanLehn, 2011; Kulik & Fletcher, 2016; Winkler & Söllner, 2018), and extend this evidence to a previously unstudied context — Uzbek primary mathematics education. Consistent with cautions raised elsewhere in the literature (Kasneci et al., 2023), the results support a "teacher-plus-AI" model rather than AI as a substitute for instruction: the chatbot appeared most beneficial as a supplementary tool operating alongside, not instead of, teacher-led classroom instruction.

5.6 Practical implications

For Uzbekistan specifically, these findings suggest that generative AI chatbots could help address two persistent challenges — limited individualised feedback in large classes and uneven access to private tutoring — provided implementation is accompanied by curriculum alignment, Uzbek-language and culturally adapted content, teacher training in pedagogical (not merely technical) chatbot integration, and adequate digital infrastructure, particularly in rural areas where connectivity gaps remain (Sattarov & Yuldashev, 2023).

6. Limitations

This study has several limitations that should be considered when interpreting the findings. First, the sample size (N = 60) and single-school setting limit generalisability to the broader population of Uzbek elementary students, particularly given known urban–rural disparities in digital infrastructure. Second, the eight-week intervention period does not permit conclusions about the durability of learning gains or motivational effects over time. Third, the analysis reported here is descriptive: means, standard deviations, and gain scores are reported, but inferential statistical tests (e.g., independent- and paired-samples *t*-tests, effect sizes) were not available for this dataset; consequently, between-group differences should be interpreted as observed patterns rather than statistically confirmed effects, and future studies should report inferential statistics to support stronger causal claims. Fourth, the study addressed foundational arithmetic and word-problem content only, and findings should not be extended to more advanced mathematical domains without further testing. Fifth, as a generative system, the chatbot may occasionally have produced inaccurate or curriculum-misaligned explanations despite teacher supervision, a risk inherent to current large language model technology (Ji et al., 2023). Finally, variables such as home learning environment, socio-economic background, and prior digital experience were not systematically measured and may have influenced outcomes.

7. Conclusion

This study provides early empirical evidence, from a context with very limited prior research, that a generative AI chatbot used as a supplementary tool can be associated with greater mathematics achievement gains, more positive motivational outcomes, and stronger classroom engagement among elementary students than traditional instruction alone. Findings support a model in which generative AI functions as a scaffold for teacher-led instruction — extending feedback availability, reducing the social cost of asking questions, and encouraging independent practice — rather than as a substitute for teachers. For Uzbekistan and comparable developing education systems, these results offer a preliminary evidence base for integrating generative AI chatbots into primary mathematics instruction, contingent on continued teacher oversight, curriculum alignment, language localisation, and infrastructure investment. Larger, longer-term, and inferentially analysed studies across multiple schools and regions are needed to confirm and extend these findings.

REFERENCES:

Alamri, A., Alqahtani, M., & Hassan, A. (2019). Designing mathematics chatbots to support learner procedural fluency. *Journal of Computers in Mathematics and Science Teaching, 38*(3), 235–255.

Ashcraft, M. H., & Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review, 14*(2), 243–248.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389–407.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.

Biswas, K., & Sengupta, R. (2024). Mobile AI-chatbot tutoring for rural elementary mathematics learners: A field trial. *Asian Journal of Educational Research, 18*(2), 88–104.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Bruner, J. (1996). *The culture of education*. Harvard University Press.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Chorianopoulou, A., & Grigoriadou, M. (2021). Conversational mathematics tutoring through chatbot adaptive feedback. *Education and Information Technologies, 26*(5), 6079–6098. https://doi.org/10.1007/s10639-021-10615-2

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems, 30*, 4299–4307.

Chung, K., Park, S., & Lee, J. (2024). Conversational AI mentors in East Asian mathematics classrooms: A comparative review. *Computers & Education: Artificial Intelligence, 6*, 100201. *(Suggested addition — not in original thesis reference list.)*

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering, 22*(5), 811–817.

Deci, E. L., & Ryan, R. M. (2000). Self-determination theory and intrinsic motivation. *Contemporary Educational Psychology, 25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Dowker, A., Sarkar, A., & Looi, C. Y. (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in Psychology, 7*, 508. https://doi.org/10.3389/fpsyg.2016.00508

Etikan, I., Musa, S. A., & Alkassim, R. (2016). Comparison of convenience and purposive sampling. *American Journal of Theoretical and Applied Statistics, 5*(1), 1–4. https://doi.org/10.11648/j.ajtas.20160501.11

Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning & Technology, 10*(3), 8–14.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems, 27*, 2672–2680.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). Information Age Publishing.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2020). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education, 30*(2), 47–78. *(Suggested addition — updated citation for concept referenced in the original thesis.)*

Kasneci, E., Seifert, C., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kerly, A., Hall, P., & Bull, S. (2007). Bringing chatbots into education: Towards natural language negotiation of open learner models. *Knowledge-Based Systems, 20*(2), 177–185. https://doi.org/10.1016/j.knosys.2006.11.014

Khodjayev, S. (2021). Digital policy directions for Uzbekistan. *Public Administration Review of Uzbekistan, 7*(4), 101–119.

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42–78. https://doi.org/10.3102/0034654315581420

Kumar, V., Singh, A., & Gupta, R. (2020). AI-driven chatbot intervention for improving primary mathematics fluency: An experimental classroom trial. *International Journal of Educational Technology in Teaching and Learning, 17*(4), 221–236. https://doi.org/10.1504/IJETTL.2020.345672

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.

Ministry of Digital Technologies of Uzbekistan. (2023). *Digital Uzbekistan 2030: National program report*.

Mirzaev, N. (2022). Expansion of AI in Uzbekistan schooling. *Uzbekistan Education Research Journal, 5*(3), 20–33.

Nazarov, S., & Soliev, A. (2023). Localising AI learning platforms in Uzbekistan. *Journal of Central Asian Pedagogy, 3*(4), 112–129.

OECD. (2021). *PISA 2022 results: Mathematics performance overview*. OECD Publishing. *(Suggested addition — updated international assessment source.)*

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346–362.

Sarama, J., & Clements, D. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. Routledge. https://doi.org/10.4324/9780203881510

Sattarov, B. (2023). Teacher-centred pedagogy and instructional differentiation in Uzbek primary classrooms. *Central Asian Education Review, 10*(2), 45–61. *(Suggested addition — clarifies classroom-practice claim.)*

Sattarov, B., & Yuldashev, T. (2023). Digital divide in Uzbek schools: ICT infrastructure comparison. *Central Asian Education Review, 9*(1), 33–47.

Shin, D., & Pierre, M. (2021). Artificial intelligence chatbots and mathematics anxiety among elementary learners. *Computers & Education, 174*, 104311. https://doi.org/10.1016/j.compedu.2021.104311

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., Susperreguy, M. I., & Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science, 23*(7), 691–697.

State Inspectorate for Education Quality. (2022). *National student achievement report: Mathematics performance review*.

Sundararajan, S., & Ma, J. (2020). Using educational chatbots for mathematics practice: Engagement and motivation outcomes. *Technology, Knowledge and Learning, 25*(4), 1083–1102. https://doi.org/10.1007/s10758-019-09445-9

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*(2), 123–138.

Tashakkori, A., & Teddlie, C. (2010). *Mixed methodology: Combining qualitative and quantitative approaches*. SAGE Publications.

UNESCO. (2022). *AI in education: Policy recommendations for Central Asia*. UNESCO Publishing.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Five years of evaluations. *Artificial Intelligence in Education, 15*, 1–36.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems*. Swets & Zeitlinger.

Winkler, R., & Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. *Academy of Management Proceedings, 2018*(1), 15903. https://doi.org/10.5465/AMBPP.2018.15903abstract

Yakubov, F. (2023). Teacher preparedness for AI usage in Uzbek primary schools. *Uzbek Journal of Education Studies, 14*(2), 56–67.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2