Enhancing Critical Thinking with Multimodal Generative AI

Alanoud Qutaim Alqahtani, Dalya Abdulaziz Alghofaily, Reem Faisal Alsahli, Haya Buayyan Alwizrah and Norah Saleh Alghamdi

Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

Abstract. The increasing popularity of Artificial Intelligence (AI) in education has transformed learning methodologies by enabling personalized and interactive experiences. However, over-reliance on generative AI technologies has raised concerns about the decline of critical thinking skills, particularly among elementary school students in mathematics. To address this issue, an AI-based educational system is proposed to enhance critical thinking and cognitive reasoning abilities among sixth-grade students. The system employs storytelling and Socratic questioning to promote engagement and cognitive development. It is built on Large Language Models (LLMs) for text and image generation, incorporating external knowledge to improve accuracy and reduce errors. Through the integration of mathematical story generation and contextual visual representations, the system aims to foster deeper learning and problem-solving skills among young learners.

Keywords: Artificial Intelligence (AI), Education, Critical Thinking, Elementary School Students, Mathematics, Socratic Questioning, Large Language Models (LLMs), Text Generation, Image Generation, Mathematical Storytelling, AI-based Learning Systems.

1 Introduction

The rapid adoption of Artificial Intelligence (AI) in education has enabled personalized learning and scalable instructional support. While AI-assisted tools offer clear benefits, concerns have emerged regarding their impact on students' analytical and critical thinking skills, particularly in mathematics education [1]. Generative AI systems, when used for direct answer generation, risk promoting surface-level learning rather than reasoning-based understanding, especially among elementary school learners. Addressing this challenge requires educational approaches that prioritize reasoning and cognitive engagement over solution delivery. Storytelling and Socratic questioning have been shown to support structured reasoning by encouraging reflection, problem decomposition, and active participation in learning [2]. This work investigates the use of AI-driven methods that integrate these pedagogical strategies to support critical thinking and cognitive development in elementary mathematics.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_36

2 Related Work

Recent advances in large language models (LLMs) have expanded their use in educational settings, particularly for mathematical reasoning, explanation generation, and dialog-based learning. Early research has largely focused on English-language benchmarks and models.

To improve reasoning consistency in Math Word Problem (MWP) solving, [3] proposed a Retrieval-Augmented Generation (RAG) approach starts by classifying the user prompt to identify requested domain and retrieves relevant context along with the solution steps in this approach domain-specific examples and solution steps to guide Large Language Model Meta AI (LLaMA). Evaluation using an LLM-as-a-judge framework and a proposed reasoning score demonstrated improved explanation coherence and reasoning reliability.

Socratic questioning has also been explored to promote deeper reasoning. The Socratic Playground for Learning (SPL) integrates Generative Pre-trained Transformers (GPT) with Chain-of-Thought prompting refers to a technique in which language models are guided to generate a coherent sequence of intermediate reasoning steps that lead to a final answer [4] within an Intelligent Tutoring System to support multi-turn instructional dialogues, reporting improvements in critical thinking while noting limitations in personalization.

Story generation has been investigated as a pedagogical strategy for learning support. In [5], an expert-validated LLM pipeline was proposed for generating instructional stories and questions, with GPT-4 outperforming earlier models. However, the study identified reasoning errors and continued reliance on human oversight.

Arabic-centric LLM development has recently gained attention. The Arabic large language model (ALLaM) [6] employs a balanced Arabic–English training corpus (45/45), demonstrating improved Arabic language understanding through large-scale training and combined automatic and human evaluation.

For Arabic educational content, Arabic LLM (AraLLaMa) was fine-tuned for story generation using synthetic and translated datasets [5]. Automatic and human evaluations revealed discrepancies between GPT-4–based metrics and human judgment, highlighting challenges in evaluating Arabic narrative generation.

Image generation has been explored as a complementary modality for educational engagement. In [7], a framework combining GPT-4 for story generation and DALLE-3 for character-consistent image creation demonstrated increased engagement, while revealing challenges in narrative–visual alignment.

Overall, Prior work on mathematical reasoning, Socratic dialogue, story generation, and multimodal content has mostly been isolated and English-focused. Limited research targets AI systems for Arabic learners. This work integrates story-driven explanations, Socratic dialogue, and multimodal content to promote guided reasoning and sustained cognitive engagement rather than answer accuracy alone.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_36

3 Methodology

The proposed system is organized as a structured, multi-stage interaction that guides students from topic selection to reflective problem-solving through narrative grounding, Socratic dialogue, and multimodal reinforcement. At a high level, a selected mathematical domain triggers retrieval of curriculum-aligned context, which is used to generate an open-ended mathematical story in Arabic. This story serves as a shared reference for both visual generation and interactive reasoning, culminating in a Socratic dialogue that encourages step-by-step analysis rather than solution delivery. The overall system workflow and model interactions are illustrated in Fig. 1, The following sections detail how data design, model specialization, fine-tuning experiments, and generation control mechanisms are combined to realize this end-to-end reasoning workflow.

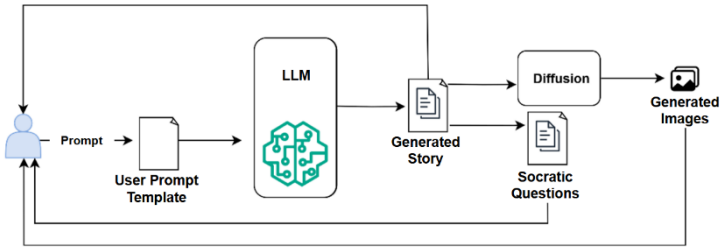

Fig. 1. High-level overview of the proposed reasoning-centered AI pipeline.

3.1 Data And Model Selection

The methodology relies on reasoning-supervised and retrieval-oriented datasets to support distinct stages of the reasoning-centered workflow. The Grade School Math 8K dataset (GSM8K) [8] provides grade-school mathematical word problems annotated with intermediate Socratic-style reasoning steps, enabling supervision of step-by-step reasoning. In contrast, Academia Sinica Diverse MWP Dataset (ASDiv) [9] consists of annotated problems labelled by grade level and type, supporting curriculum-aligned retrieval and contextual grounding during generation. To ensure linguistic and instructional consistency for the target learner population, both datasets are translated from English to Arabic using ALLaM. For ASDiv, Named Entity Recognition is applied to localize foreign names, followed by manual validation to preserve coherence. GSM8K samples are reformatted into structured User–Assistant dialogue templates to align the data with the intended Socratic interaction.

Model selection follows a task-driven specialization strategy. Narrative generation is assigned to GPT-4o due to its ability to produce coherent, open-ended stories, while ALLaM handles mathematical reasoning and Socratic questioning, leveraging its alignment with Arabic educational content. LLaMA 3 is employed for scene decomposition and visual prompt preparation, while a diffusion-based model performs image synthesis.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_36

3.2 System Workflow and Reasoning Control

To ensure that mathematical reasoning is expressed through guided inquiry rather than answer delivery, ALLaM is adapted for Socratic questioning using Parameter-Efficient Fine-Tuning (PEFT) a strategy designed to make task-specific adjustment without changing the full model's parameters [10] In particular, Low-Rank Adaptation (LoRA) introduces low-rank matrices into the model's architecture, to capture task-specific information without modifying the full weight matrices [11]. Fine-tuning is treated as an experimental variable rather than a fixed configuration. Four variants are evaluated: a fully frozen base model (LoRA-only training) and models with the final 3, 4, or 7 transformer layers unfrozen. This controlled setup enables systematic analysis of the trade-off between representational flexibility and stability, isolating the effect of partial unfreezing on step-by-step Socratic reasoning. All configurations are trained under a unified setup to ensure comparability.

To control generation quality and ensure curriculum alignment, the system integrates RAG as the entry point of the workflow [12]. As shown in Fig. 2, the student first selects a mathematical domain, triggering retrieval of topic-relevant examples from a vector store populated with ASDiv problems. Fig. 3 illustrates the learner-facing interface for mathematical topic selection, which serves as the entry point for the retrieval-augmented generation pipeline. Retrieved contexts ground the generation process and are passed to GPT-4o, which generates an open-ended mathematical story in Arabic. This story establishes a shared semantic context that anchors subsequent reasoning and multimodal stages.

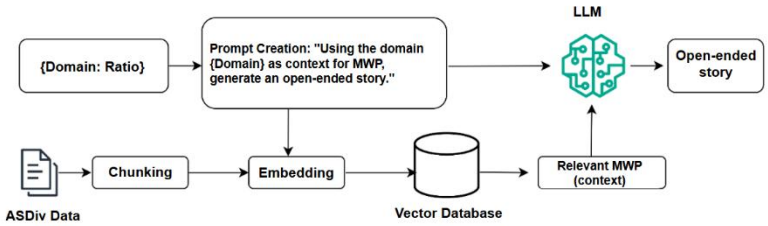

Fig. 2. The retrieval-augmented generation workflow used for story creation.

Fig. 3. Mathematical domain selection interface used for curriculum-aligned MWP retrieval.

For visual reinforcement, the generated story is segmented into scenes using LLaMA 3, which performs entity extraction and location inference before translating the Arabic narrative into structured English prompts compatible with the diffusion model. These

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_36

prompts are passed to a LoRA-enhanced Flux Diffusion model, specifically LoRA-enhanced Flux Diffusion model variant built upon FLUX.1-dev [13] to generate consistent, comic-style illustrations. An example of the generated visual narrative is shown in Fig. 4, demonstrating the transformation of the open-ended story into a comic-style illustration.

Fig. 4. Example of a generated visual scene derived from the open-ended mathematical story.

In parallel, the same story context is provided to the fine-tuned ALLaM model, which initiates a Socratic dialogue with the student. By grounding both visual synthesis and interactive questioning in a single narrative source, the system maintains coherence across modalities while guiding students through reflective, step-by-step mathematical reasoning.

To enforce structured reasoning throughout interaction, multiple interconnected prompts are employed as a control mechanism across all generative stages. Multi-instruction templates coordinate story generation, Socratic questioning, hint generation, and visual prompts within a unified reasoning flow. During dialogue, prompt chaining governs the Socratic flow: student responses are evaluated relative to expected reasoning steps, and follow-up prompts adapt accordingly to guide reflection without revealing answers. When misconceptions arise, targeted hints scaffold understanding while preserving learner agency.

4 Experimental Results and Evaluation

This section evaluates the proposed system across dataset translation, model fine-tuning, Socratic questioning, and image generation. Both automatic and subjective metrics are used to assess translation fidelity, reasoning quality, and multimodal alignment.

4.1 Translation Quality

Arabic translations generated using ALLaM were evaluated against Deep Translate references using Bilingual Evaluation Understudy (BLEU) [14] and Metric for Evaluation of Translation with Explicit Ordering (METEOR) [15]. As shown in Table 1, GSM8K achieved BLEU/METEOR scores of 0.70/0.66, while ASDiv achieved 0.76/0.70, indicating sufficient translation fidelity to support fine-tuning and retrieval-based generation.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_36

Table 1. Dataset translation performance using BLEU and METEOR

Dataset	BLEU Score	METEOR Score
GSM8K	0.70	0.66
ASDiv	0.76	0.70

4.2 Fine-Tuned Model Evaluation

ALLaM was fine-tuned under four configurations: a fully frozen base model (Model 1), and variants with the final 3, 4, and 7 transformer layers unfrozen (Models 2–4). Automatic evaluation using BERTScore which computes a similarity score for each token in the candidate sentence with each token in the reference sentence [16] produced consistent results across all configurations (F1 ≈ 0.83), indicating stable semantic similarity independent of the degree of unfreezing. In contrast, subjective evaluation using GPT-4o as an LLM-as-a-Judge revealed clearer differentiation. As shown in Table 2, Model 2 (last three layers unfrozen) achieved the highest average score (8.2/10). excelling in several key areas, particularly in Usefulness, Relevance, and Logical Coherence. It can be observed that Model 2 was the most effective in generating responses that were both coherent and relevant, providing content that was not only useful but also logically consistent, suggesting that limited unfreezing improves reasoning quality while avoiding overfitting.

Table 2. Subjective evaluation scores across ALLaM fine-tuning configurations

Model	Usefulness	Relevance	Clarity	Depth	Logical Coherence	Average
Model 1	7.60	7.60	7.47	6.60	8.93	7.84
Model 2	7.80	7.80	8.40	6.87	9.07	8.20
Model 3	7.67	7.67	8.20	6.80	9.07	8.11
Model 4	7.47	7.47	8.07	6.47	8.73	7.85

4.3 Socratic Questioning Evaluation

Chatbot interactions were evaluated across usefulness, Socratic style, relevance, clarity, and logical sequence. While relevance (7.4) and clarity (7.2) were strong, lower scores in Socratic style (4.6) indicate a tendency toward answer-oriented responses. Additionally, moderate scores in usefulness (6.4) and logical sequence (6.2) suggest occasional limitations in providing coherent and pedagogically effective guidance, highlighting an area for further refinement.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_36

4.4 Image Generation Evaluation

Image–text alignment was evaluated using BLIP [17], and CLIP [18]. Average scores were 0.43 and 0.33, respectively, reflecting moderate semantic and perceptual alignment. These results confirm the system's ability to generate visually relevant illustrations, with scope for improving prompt–image consistency.

5 Discussion and Conclusion

The evaluation results demonstrate the effectiveness of the system in promoting reasoning-centered mathematics learning through Socratic interaction and multimodal support. While BERTScore indicated comparable semantic similarity across fine-tuning configurations, subjective evaluation proved more informative, showing that selectively unfreezing a small number of transformer layers produced higher instructional quality in terms of usefulness, relevance, and logical coherence.

Limitations were observed in sustaining deep Socratic inquiry over longer interactions, primarily due to constraints in ALLaM's context window and model capacity. Nevertheless, relevance and clarity remained stable, indicating consistent grounding in the mathematical context. At the system level, integrating retrieval-augmented generation with diffusion-based image synthesis enhanced coherence. Overall, the results support the value of retrieval grounding, targeted fine-tuning, and multimodal generation for educational AI systems.

6 Future Directions

Although storytelling and Socratic questioning support structured reasoning and active learning [2], their effectiveness within AI-based systems remains underexplored. Future work will evaluate our proposed tool with real learners to assess its impact on critical thinking and learning outcomes.

Acknowledgments.
The authors would like to thank Eng. Muhammad Alagil and Nora Alagil for their support and generous grant in sponsoring the participation in this conference, and Princess Nourah bint Abdulrahman University for facilitating participation in this study.

Authors' Contributions.
D.A. and A.A. contributed to the conceptualization and methodology of the proposed system. A.A. and R.A. were responsible for system implementation, data preparation, and visualization. H.A. conducted experimental design and analysis. N.A. provided supervision, and project administration. D.A. and A.A. drafted the manuscript. All authors contributed to reviewing and editing the final paper.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_36

Disclosure of Interests.
The authors declare no conflict of interest.

References

1. Spector, J.M., Ma, S.: Inquiry and critical thinking skills for the next generation: from artificial intelligence back to human intelligence. Smart Learning Environments 6(1), 8 (2019). https://doi.org/10.1186/s40561-019-0088-z
2. Mena, A.: Critical thinking for civic life in elementary education: combining storytelling and thinking tools. Revista Educación 44, 23–43 (2020). https://doi.org/10.15517/revedu.v44i2.39699
3. Dixit, P., Oates, T.: SBI-RAG: Enhancing math word problem solving for students through schema-based instruction and retrieval-augmented generation. arXiv:2410.13293 (2024)
4. Zhang, L., Lin, J., Kuang, Z., Xu, S., Hu, X.: SPL: A Socratic playground for learning powered by large language models. arXiv:2406.13919 (2024)
5. El-Shangiti, A.O., Alwajih, F., Abdul-Mageed, M.: Arabic automatic story generation with large language models. arXiv:2407.07551 (2024)
6. Bari, M.S. et al.: ALLaM: Large language models for Arabic and English. arXiv:2407.15390 (2024)
7. Makridis, G., Oikonomou, A., Koukos, V.: FairyLandAI: Personalized fairy tales utilizing ChatGPT and DALL·E 3. arXiv:2407.09467 (2024)
8. Cobbe, K. et al.: Training verifiers to solve math word problems. arXiv:2110.14168 (2021)
9. Miao, S.-Y., Liang, C.-C., Su, K.-Y.: A diverse corpus for evaluating and developing English math word problem solvers. arXiv:2106.15772 (2021)
10. Xu, L., Xie, H., Qin, S.-Z.J., Tao, X., Wang, F.L.: Parameter-efficient fine-tuning methods for pretrained language models: a critical review and assessment. arXiv:2312.12148 (2023)
11. Hu, E.J. et al.: LoRA: Low-rank adaptation of large language models. arXiv:2106.09685 (2021)
12. Levonian, Z. et al.: Retrieval-augmented generation to improve math question answering: trade-offs between groundedness and human preference. arXiv:2310.03184 (2023)
13. Black Forest Labs: FLUX.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev, last accessed 22/04/2025
14. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: A method for automatic evaluation of machine translation. In: Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318. Association for Computational Linguistics, Philadelphia (2002). https://doi.org/10.3115/1073083.1073135
15. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein, J. et al. (eds.) Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72. Association for Computational Linguistics, Ann Arbor (2005)
16. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with BERT. arXiv:1904.09675 (2019)
17. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv:2201.12086 (2022)
18. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. arXiv:2104.08718 (2022)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_36