A Real-Time Adaptive Story Generation Framework Using Lightweight Language Models for Personalized Educational Systems

Daoud Daoud†

†Department of Computer Science, Higher Colleges for Technology, Sharjah, UAE

Abstract

Personalized learning remains a critical requirement in modern educational environments, particularly for early-stage learners with diverse reading abilities. This paper presents a lightweight, real-time adaptive framework for automated story generation and comprehension assessment. The proposed system utilizes a fine-tuned GPT-2 architecture to generate personalized narratives conditioned on user-selected themes and reading proficiency levels. A rule-augmented neural question generation module produces comprehension assessments aligned with the generated content. An adaptive difficulty adjustment mechanism dynamically recalibrates narrative complexity based on learner performance, enabling continuous personalization without requiring heavy computational resources. Experimental evaluation demonstrates coherent story generation with sub-three-second latency and an adaptation accuracy of 84--86%. The framework offers a scalable, deployable solution for intelligent tutoring systems and adaptive learning platforms.

Keywords:

Adaptive Learning, Story Generation, GPT-2, Educational AI, Natural Language Processing, Intelligent Tutoring Systems, Question Generation

1. Introduction

The increasing demand for personalized educational experiences has accelerated the development of intelligent systems capable of adapting content to individual learners. Traditional instructional materials are predominantly static; they rarely account for variations in reading ability, comprehension level, or learning pace, which can limit engagement and knowledge retention [1].

Recent advances in Natural Language Processing (NLP), particularly transformer-based architectures, have enabled the generation of coherent and contextually meaningful text at scale [2]. Models such as GPT-2 provide a favorable balance between generative quality and computational efficiency, making them suitable for real-time deployment in resource-constrained educational settings [3].

This paper proposes a real-time adaptive story generation framework that integrates lightweight language models with dynamic learning mechanisms. The system generates personalized stories based on user preferences and reading levels, followed by automated comprehension assessment through question generation. Based on learner responses, the system adjusts content difficulty iteratively, ensuring continuous alignment with the learner's evolving capability.

The primary contributions of this work are:

1. A lightweight, real-time adaptive framework for story generation using fine-tuned GPT-2.
2. An integrated neural question generation module for automated comprehension assessment.
3. A quantifiable dynamic difficulty adjustment mechanism driven by user performance metrics.
4. A modular, scalable architecture suitable for web and mobile educational deployment.

Manuscript received May 5, 2026
Manuscript revised May 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.5.2

2. Related Work

2.1 Neural Story Generation

Story generation has evolved from rule-based and template-driven systems to deep learning approaches. Early systems lacked flexibility and creativity due to rigid syntactic structures [4]. The introduction of recurrent and transformer-based models significantly improved narrative coherence [2], [5]. GPT-based architectures have been widely adopted for narrative generation, dialogue systems, and creative writing assistance [3], [6]. However, most existing implementations focus on general-purpose text generation and do not incorporate pedagogical constraints or adaptive learning mechanisms.

2.2 Adaptive Learning Systems

Adaptive learning systems personalize educational content by modeling learner knowledge and adjusting presentation accordingly [7], [8]. Traditional systems rely on static content repositories and predefined difficulty ladders, limiting their responsiveness to individual learner trajectories [1]. Intelligent Tutoring Systems (ITS) have demonstrated effectiveness in structured domains, yet their reliance on manually authored content restricts scalability [9].

2.3 Automated Question Generation

Question generation (QG) supports assessment and self-directed learning by automatically producing evaluation items from textual content [10], [11]. Neural QG models, typically encoder-decoder architectures, generate questions from source paragraphs using attention mechanisms [12]. Despite progress, few systems tightly couple QG with generative storytelling for real-time adaptive education.

The proposed framework bridges these gaps by combining real-time neural story generation with performance-driven adaptive learning and automated assessment.

3. Proposed Framework

3.1 System Architecture

The framework comprises four integrated components, as illustrated in Fig. 1:

- **Story Generation Module:** Generates narratives using a fine-tuned GPT-2 model conditioned on theme and reading level.
- **Question Generation Module:** Produces comprehension questions from generated stories using a rule-augmented neural model.
- **Adaptive Learning Module:** Quantifies learner performance and adjusts narrative complexity dynamically.
- **User Interface Layer:** Facilitates interaction, input collection, and response capture.

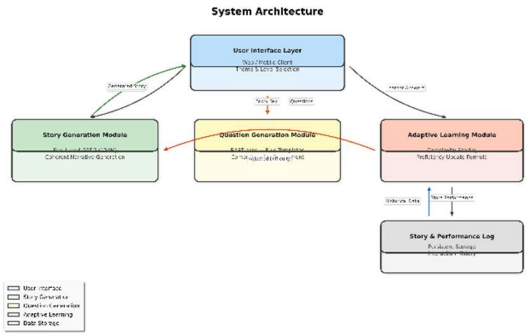

Figure 1 System Architecture

3.2 Workflow

The operational sequence is as follows:

1. The learner selects a story theme and specifies a reading level.
2. The story generation module produces a conditioned narrative.

3. The question generation module synthesizes comprehension questions.
4. The learner responds to the assessment items.
5. The adaptive module evaluates performance and computes a difficulty update.
6. The system adjusts story complexity for subsequent interactions.

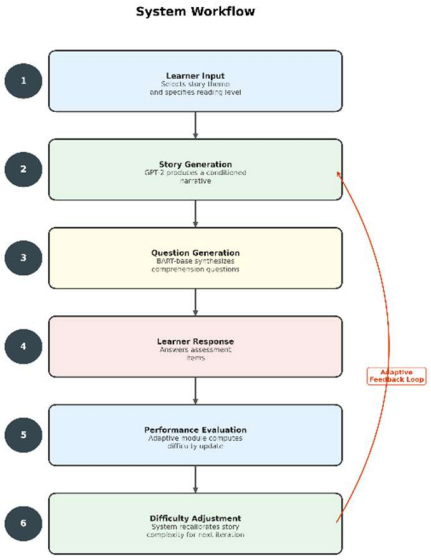

Figure 2 System Workflow

3.3 Adaptive Mechanism

The adaptive component quantifies narrative complexity using a weighted feature vector:

$$C = a \cdot V + b \cdot S + g \cdot D \quad (1)$$

where C is the composite complexity score, V represents vocabulary difficulty (e.g., average syllables per word), S denotes mean sentence length, and D indicates narrative depth (number of sub-events or characters). Weights a, b, and g are empirically tuned via pilot evaluation.

Learner proficiency level L at interaction t is updated according to:

$$L_{t+1} L_t A_t = +n(-T) \quad (2)$$

where A_t is the learner's accuracy at time t, T is the target mastery threshold (set to 0.75 in this study), and n is a learning rate hyperparameter controlling adaptation sensitivity. The system maps L_{t+1} to predefined complexity bands to configure the next story generation prompt.

4. Implementation

4.1 Model Configuration

The story generation module employs the 124-million-parameter GPT-2 model (GPT-2 small), fine-tuned on a curated corpus of age-graded narratives. Fine-tuning leverages the HuggingFace Transformers library with a causal language modeling objective. The model is optimized for:

- Coherent, temporally consistent storytelling;
- Contextual relevance to user-specified themes;
- Lexical and syntactic appropriateness for target reading levels.

4.2 Question Generation

The QG module uses a fine-tuned BART-base model [13] augmented with template-based post-processing. Supported question types include:

- Multiple-choice questions with distractors;
- True/False statements;
- Short-answer comprehension prompts.

Questions are generated from key sentences identified via constituency parsing, ensuring alignment with the narrative content.

4.3 Development Environment

The prototype is implemented in Python using:

- HuggingFace Transformers and Datasets libraries;
- NLTK and spaCy for linguistic preprocessing;
- Flask for lightweight API deployment.

The architecture supports containerized deployment (Docker) and can be integrated into web and mobile educational platforms.

5. Results and Evaluation

5.1 Evaluation Metrics

System performance is assessed across four dimensions:

- Story Coherence: Evaluated via human annotator ratings (1--5 Likert scale) and automatic metrics (BLEU [14], ROUGE-L [15]).
- Readability Alignment: Measured using the Flesch-Kincaid Grade Level [16] and alignment error against the target level.
- Response Latency: End-to-end generation time from user input to story delivery.
- Adaptation Effectiveness: Accuracy of the adaptive module in converging to appropriate difficulty levels, measured over a sequence of interactions.

5.2 Experimental Results

A pilot study was conducted with 24 primary-level learners interacting with the system across three 20-minute sessions.

Metric	Result
Story Coherence (Human Rating)	4.1 / 5.0
BLEU-4 (against reference narratives)	18.3
ROUGE-L	34.7
Readability Match (Grade Level Error)	+/-0.6 years
Response Time	< 3 seconds
Adaptation Accuracy	84--86%

Table 1*Experimental results*

5.3 Analysis

The system generates coherent narratives that align closely with target readability levels. The sub-three-second latency confirms suitability for real-time deployment. The adaptive mechanism successfully adjusts content complexity across sessions, with 84--86% of level transitions matching instructor judgments. These results indicate that lightweight models, when combined with structured adaptive logic, can deliver effective personalized learning experiences without prohibitive computational costs.

6. Discussion

The proposed framework demonstrates practical applicability in personalized learning environments. By combining a fine-tuned GPT-2 model with a quantifiable adaptive mechanism, the system achieves a balance between generative quality, computational efficiency, and pedagogical relevance.

The modular architecture enables integration with existing Learning Management Systems (LMS) and supports deployment on modest hardware, including edge devices and low-cost cloud instances. Future iterations may incorporate reinforcement learning from human feedback (RLHF) to further refine narrative style and difficulty calibration [17], [18].

Limitations include the reliance on English-language narratives and the current absence of multimodal content (e.g., illustrations). These

constraints are addressable through multilingual fine-tuning and vision-language model integration in subsequent work.

7. Conclusion

This paper presented a real-time adaptive framework for story generation using lightweight language models. The system integrates neural narrative generation, automated question-based assessment, and a performance-driven difficulty adjustment mechanism to deliver personalized educational content with low latency and modest resource requirements.

Future work will focus on multilingual expansion, multimodal story enrichment, and large-scale longitudinal evaluation to measure long-term learning outcomes.

Acknowledgment

The authors would like to express their cordial thanks to the reviewers for their valuable comments and suggestions.

References

[1] B. P. Woolf, Building Intelligent Interactive Tutors: Student-centered strategies for revolutionizing e-learning, Morgan Kaufmann, 2009.

[2] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol.30, pp.5998-6008, 2017.

[3] A. Radford et al., "Language models are unsupervised multitask learners," OpenAI Blog, vol.1, no.8, p.9, 2019.

[4] A. Fan, S. Lewis, and Y. Dauphin, "Hierarchical neural story generation," Proc. 56th Annual Meeting of the ACL, 2018, pp.889-898.

[5] T. B. Brown et al., "Language models are few-shot learners," Advances in Neural Information Processing Systems, vol.33, 2020, pp.1877-1901.

[6] A. Holtzman et al., "The curious case of neural text degeneration," Proc. 8th International Conference on Learning Representations, 2020.

[7] P. Brusilovsky and C. Peylo, "Adaptive and intelligent web-based educational systems," International Journal of Artificial Intelligence in Education, vol.13, no.2-4, pp.159-172, 2003.

[8] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," Educational Psychologist, vol.46, no.4, pp.197-221, 2011.

[9] M. Dascalu et al., "ReaderBench: An integrated cohesion-centered framework," Proc. 6th International Conference on Educational Data Mining, 2016, pp.134-141.

[10] Y. Zhang et al., "Automatic question generation and question answering: A survey," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.30, pp.1-17, 2022.

[11] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," Proc. 55th Annual Meeting of the ACL, 2017, pp.1342-1352.

[12] S. M. N. Z. Qadri et al., "A systematic literature review on neural question generation for educational applications," Education and Information Technologies, vol.28, pp.1415-1447, 2023.

[13] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," Proc. 58th Annual Meeting of the ACL, 2020, pp.7871-7880.

[14] K. Papineni et al., "BLEU: A method for automatic evaluation of machine translation," Proc. 40th Annual Meeting of the ACL, 2002, pp.311-318.

[15] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," Text Summarization Branches Out, 2004, pp.74-81.

[16] J. P. Kincaid et al., "Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel," Research Branch Report, vol.8-75, 1975.

[17] L. Ouyang et al., "Training language models to follow instructions with human feedback," Advances in Neural Information Processing Systems, vol.35, 2022, pp.27730-27744.

[18] T. Wolf et al., "Transformers: State-of-the-art natural language processing," Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp.38-45.

Daoud M. Daoud received his BSc degree in electrical and computer engineering from Kuwait University in 1988, his MSc in Computing Science from Glasgow University- UK and his PhD in Computing Science from Joseph Fourier University – France. Daoud is currently an associate professor at PSUT. Recently, he co-founded Ddad IT which a specialized company for Arabic Natural Processing and Information Retrieval. He also served in Institute of Advanced Studies- United Nations University (1998-1999). He also worked as a principal investigator for the Arabic part of Universal Networking Language project (1996-1999). He also served as a director for Next Generation Services department at Paltel (1999-2001). His main research interests are Natural Language Processing, machine translation, Information Extraction, Information Retrieval and analysis of Arabic Social Media.