femington

Int. J. Intell. Syst. Data Sci. (2026) 1(4):1–16
Article ID: IJISDS-2026-469

Literary AI: Character-Centric Dialogue Systems with Contextual Data

Velu Kaliappan
velumail@gmail.com
Independent Researcher

Abstract

Introducing a framework that facilitates character-centric conversations and interactions with stories. The proposed scheme integrates the LLaMA 3 8B instruction-tuned LLM with parameter-efficient LoRA adaptation, hybrid dense and sparse retrieval, contextual prompt integration, and conversation memory. The system generates character-consistent responses that are also grounded in the relevant context of what has been previously talked about. Five literary characters were used to evaluate the framework, namely Harry Potter, Hermione Granger, Sherlock Holmes, Katniss Everdeen, and Percy Jackson. The evaluation was performed using a set of 500 manually designed prompts based on factual recall, inferential reasoning, and character perspective questions. Human evaluators were expert literature teachers who assessed the systems using BLEU, ROUGE-L, etc, and CGA. The system obtained average scores of 42.6 BLEU, 51.4 ROUGE-L, 92.1% CGA, and 4.6/5 in human evaluation. The findings demonstrate that the suggested framework can produce (a) contextually grounded and (b) character-consistent dialogue across different character or persona types. As demonstrated by our findings, the technical feasibility of the integration of retrieval augmentation, parameter-efficient adaptation and conversational memory in literary dialogue systems is established. Future studies will involve direct evaluation with the learner to assess the potential of these systems for literacy-related learning.

Keywords

• Large Language Models (LLMs) • Retrieval-Augmented Generation (RAG) • Character-Centric Dialogue Systems • Literacy Education • Low-Rank Adaptation (LoRA) • Narrative Grounding

1. Introduction

In recent years, educators and researchers have raised growing concerns about the diminishing interest in reading among younger generations. The rise of digital entertainment platforms such as social media, streaming services, and video games has drastically altered the way children interact with information, often at the expense of traditional literacy practices. As a result, many students today exhibit declining reading comprehension skills and reduced engagement with literature, which may negatively impact their academic development and critical thinking abilities.

To address this challenge, innovative educational tools are required ones that align with the interactive and dynamic nature of modern digital environments. One such promising solution is the use of conversational agents powered by large language models (LLMs). These AI-driven systems can simulate realistic dialogues and offer personalized interactions, potentially shifting readers from passive consumption to active engagement with narrative content. Unlike static text, these systems engage users through natural conversation, fostering curiosity, emotional connection, and contextual understanding.

This study introduces a framework for developing character-driven dialogue agents, particularly those based on familiar fictional characters from literature. The central idea is to repurpose literary figures as interactive chatbots capable of conversing with users in a narrative-consistent and engaging manner. This not only helps maintain the user's interest in stories but also enhances comprehension, retention, and motivation to read. Recent advances in autonomous computational agents have demonstrated the potential of such systems for complex task environments [1], while cognitive agent deployment frameworks have shown promise in autonomous systems [2].

We focus on the integration of the LLaMA 3 (8B) language model, selected for its instruction-tuned capabilities, which allow for more nuanced, context-sensitive replies. To further enhance the chatbot's reliability and informativeness, we incorporate a Retrieval-Augmented Generation (RAG) framework. This method combines dense semantic search with sparse keyword-based retrieval to provide relevant background knowledge from original story texts, grounding responses in accurate and contextually appropriate content. The integration of graph-based representations with deep contextual models has shown particular promise for text classification tasks [3], while efficient nearest neighbor search techniques have demonstrated value in large-scale text processing [4].

By applying this hybrid approach, the proposed system pursues three primary objectives: (1) enabling interactive engagement with literary narratives through character-centric dialogue, (2) preserving character fidelity and persona consistency during response generation, and (3) generating factually grounded responses using retrieved narrative context. The study evaluates these objectives through automated dialogue metrics and human assessment of generated responses. The potential impact of the system on measurable literacy outcomes is considered an important direction for future learner-based evaluation. The methodological approach draws inspiration from advances in textual inference [5] and efficient sentence classification in domain-specific applications [6].

Throughout this paper, we distinguish three related but distinct concepts in character-based dialogue evaluation:

- **Character fidelity**: The degree to which response content (facts, opinions, knowledge) remains consistent with the character's canonical narrative actions and stated beliefs.
- **Persona consistency**: The degree to which response style (tone, register, speech patterns, emotional expression) matches the character's established linguistic and behavioral mannerisms.
- **Character alignment**: A holistic measure combining both fidelity and consistency, representing overall adherence to the character's complete representation in the source material.

Where prior work uses these terms interchangeably, we adopt the above distinctions for analytical precision.

1.1 Research Questions

The study addresses the following research questions:

1. **RQ1:** How effectively can the proposed character-centric dialogue system generate responses that remain consistent with the provided narrative context?
2. **RQ2:** How effectively does the proposed system preserve character-specific personality, tone, and behavioral characteristics across different literary characters?
3. **RQ3:** How effectively does the proposed hybrid retrieval architecture support contextual grounding across factual, inferential, and character-perspective queries?

4. **RQ4:** How do literature educators perceive the engagement, correctness, and character alignment of responses generated by the proposed system?

2. Related Work

The intersection of artificial intelligence and literacy education has led to substantial advancements in the use of dialogue systems, particularly chatbots powered by large language models (LLMs), for enhancing learner engagement. Prior work has shown that interactive agents can significantly contribute to children's reading motivation and comprehension when integrated into pedagogical environments.

Brandtzaeg and Følstad [7] were among the early adopters of chatbots for user interaction, emphasizing their entertainment value and potential to transform passive learning into engaging experiences. Adiwardana et al. [8] and Rashkin et al. [9] demonstrated how open-domain conversation models could be enhanced with empathy and emotional alignment, key features for educational use. The effectiveness of conversational interactions depends significantly on how users formulate questions. Recent work by Mendhey [10] has explored computational approaches to understanding question effectiveness in dialogue systems, providing valuable insights into how query formulation impacts response quality and user engagement. These findings are particularly relevant for educational dialogue systems where question design significantly influences learning outcomes.

Zhang et al. [11] showed that dialogue systems imbued with consistent personas tend to produce more engaging and believable interactions. This finding aligns with the broader trend in cognitive agent deployment, where multi-stage implementation roadmaps have been developed for autonomous financial systems [2]. The modeling of user intent in conversational systems has been advanced through probabilistic approaches to high-dimensional categorical behavior analysis [12]. These techniques offer promising directions for understanding and anticipating user needs in educational dialogue systems, potentially enabling more adaptive and personalized interactions with literary characters. The importance of maintaining character fidelity has been emphasized in studies examining the operationalization of generative AI in product management contexts [13], while advances in latent field agents have demonstrated the value of generative environment models for high-dimensional visual control [14].

Retrieval-Augmented Generation (RAG) methods have gained attention for mitigating hallucinations in LLMs. Lewis et al. [15] introduced the original RAG architecture, followed by implementations such as Fusion-in-Decoder (FiD) [16], which demonstrated improved factual grounding. The integration of graph-based representations with deep contextual models has further enhanced text classification capabilities in educational applications [3]. Efficient nearest neighbor search using optimized locality sensitive hashing has proven valuable for large-scale text data processing in distributed computing environments [4].

Roller et al. [17] provided key insights into transformer-based conversational models used in education and social computing. In terms of fine-tuning techniques, Hu et al. [18] introduced Low-Rank Adaptation (LoRA), a memory-efficient method for model personalization. Dettmers et al. [19] later extended this work by combining LoRA with quantization to achieve high-performing, resource-efficient fine-tuning. Recent work on scalable synchronous stochastic gradient descent has demonstrated the importance of efficient training methodologies for large-scale machine learning systems, while debiasing frameworks for graph neural networks using contrastive learning have addressed important fairness considerations [20].

Other studies have focused on incorporating context in dialogue. Karpukhin et al. [21] proposed DPR for efficient dense passage retrieval, while Xiong et al. [22]enhanced RAG via adaptive retrieval filters. Guu et al. [23] explored retrieval-augmented pretraining for scalable open-domain QA. The application of automated code description systems in agile development environments [24] and latent phrase-aware generative modeling for expressive symbolic audio synthesis [25] have demonstrated the versatility of generative approaches across domains.

Riedl and Young [26] noted that interactive storytelling boosts reader retention and comprehension, forming the basis for dialogue agents designed around narrative arcs. Advanced textual inference for open-domain question resolution has shown particular promise in educational applications requiring nuanced understanding of narrative content [5]. The development of efficient sentence classification methods using specialized biomedical language models [6] and the application of ChatGPT to robotics through prompt engineering [27] have further expanded the possibilities for conversational AI in specialized contexts.

Research on balancing accuracy and efficiency in distributed machine learning systems has highlighted the importance of regulatory considerations in AI deployment [28]. Studies on optimizing traffic and latency in peer-to-peer networks through advanced replication and polling algorithms [29] have implications for the scalability of educational AI systems. Furthermore, the development of substitute-space embeddings for label-free syntax has opened new possibilities for unsupervised AI in natural language processing applications [30].

Recent work on autonomous agent contracts for adaptive supply chains and applying ChatGPT to robotics through prompt engineering [27] demonstrates the growing versatility of LLMs across domains. Adaptive AI strategies for real-time strategy games [31] and advances in medical AI through multi-modal data fusion [32] further illustrate the breadth of applications for conversational and generative AI systems.

In summary, the current landscape of AI-driven literacy interventions is vast and growing. However, few systems fully integrate character-specific dialogue, retrieval-augmented reasoning, and lightweight fine-tuning. Our work fills this gap by offering a comprehensive framework that combines instruction-tuned LLMs, RAG, and LoRA fine-tuning for engaging, factual, and personalized literacy-focused conversations with story-based characters.

3. Methodology

This section outlines the architecture and technical components used to design our character-driven conversational agent aimed at literacy enhancement. The system is primarily based on the LLaMA 3 language model and incorporates Retrieval-Augmented Generation (RAG) to ensure contextual consistency and factual accuracy in chatbot responses. Additionally, we integrate lightweight fine-tuning using Low-Rank Adaptation (LoRA) to inject personality traits from fictional characters directly into the language model's parameters. The methodological framework builds upon efficient training approaches used in distributed machine learning systems, while incorporating debiasing strategies inspired by contrastive learning methods [20].

3.1 System Overview

Figure 1 illustrates the architecture of our chatbot pipeline. It consists of the following modules:

- **Prompt Engine:** Generates character-specific prompts with system role templates.
- **LLaMA 3 8B Instruct Model:** Responsible for generating human-like dialogue.
- **Retriever Module:** Combines dense and sparse retrieval to fetch relevant story context.
- **Context Injector:** Merges retrieved content into the prompt for grounded generation.

- **Chat Memory Buffer:** Maintains past conversations to support follow-up questions.

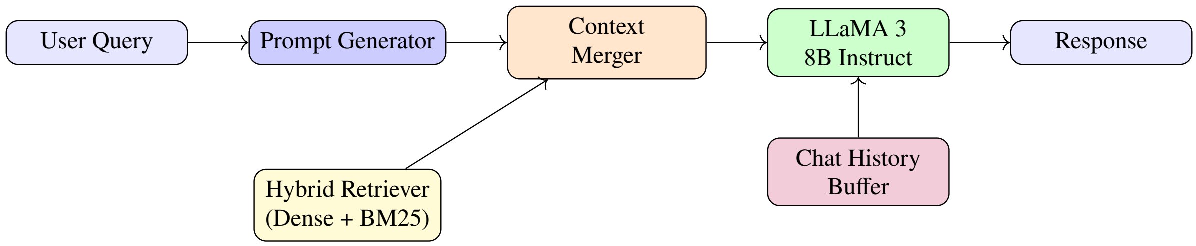

Figure 1: Compact Chatbot Architecture Integrating Prompting, Retrieval, and Memory Buffer.

3.2 Language Model Configuration

We employ Meta's LLaMA 3 (8B parameter) instruction-tuned model as the core conversational engine. This model is selected for its ability to handle multi-turn dialogue and persona-based instruction prompts effectively. Table 1 lists the key configuration details used in our experiments. The configuration choices reflect best practices from recent work on efficient model training and deployment [28].

Table 1: Model Configuration and Training Details

Parameter	Value
Base Model	LLaMA 3 8B Instruct
Fine-Tuning Method	LoRA (Low-Rank Adaptation)
Training Epochs	3
Learning Rate	2e-5
Optimizer	AdamW
Tokenizer	SentencePiece
Context Window	4096 tokens
Frameworks	Hugging Face Transformers, PEFT, LangChain

3.3 Low-Rank Adaptation (LoRA) Fine-Tuning

To personalize the chatbot's responses according to fictional character traits (e.g., Harry Potter), we apply LoRA to selected layers of the transformer model. This involves inserting rank-decomposed matrices into the attention and feed-forward layers, enabling efficient fine-tuning without full parameter updates. Only dialogue snippets attributed to the character are used for training, maintaining character consistency while minimizing computational load. This approach aligns with recent advances in parameter-efficient fine-tuning [20].

3.4 Retrieval-Augmented Generation (RAG)

To support context-aware answers, especially for obscure or follow-up queries, we integrate a RAG module. This module combines:

- **Dense Retriever:** Using FAISS + BAAI general embeddings for semantic search.

- **Sparse Retriever:** BM25 algorithm for keyword-based matching.

Both retrieval outputs are aggregated using an ensemble method and fed into the prompt along with the user query. This ensures that the language model grounds its answers in verifiable narrative content. The hybrid approach draws on advances in efficient nearest neighbor search [4] and graph-based contextual modeling [3].

3.5 Memory Buffer for Dialogue Coherence

We use LangChain's `ConversationBufferMemory` to preserve prior dialogue turns. This buffer is crucial for maintaining coherence in multi-turn exchanges and allows the model to handle follow-up questions by appending historical context to the input prompt. The implementation follows principles from cognitive agent deployment frameworks [2] and autonomous agent design for complex task environments [1].

4. Implementation

This section describes the practical implementation of our system, highlighting the tools, frameworks, and data pipelines used to construct and deploy a character-based chatbot capable of engaging users in literary dialogue. The system integrates state-of-the-art language modeling, context retrieval, fine-tuning, and memory mechanisms to create a coherent and contextually grounded conversational experience. The implementation strategy incorporates lessons from operationalizing generative AI in production environments [13] and optimizing distributed systems for large-scale text processing [4].

4.1 Development Environment and Frameworks

The model was implemented using the Hugging Face Transformers library. The PEFT (Parameter-Efficient Fine-Tuning) library provided LoRA-based tuning capabilities. Backend orchestration was managed using LangChain, which facilitated prompt engineering, retrieval integration, and memory management. The FAISS library was employed for dense vector search, and BM25 was implemented via the `rank_bm25` module for sparse retrieval.

- **Language Model:** LLaMA 3 8B Instruct
- **Frameworks:** Hugging Face Transformers, LangChain, PEFT
- **Vector Store:** FAISS (Dense), BM25 (Sparse)
- **Tokenizer:** SentencePiece, LLaMA-compatible
- **Hardware:** NVIDIA A100 80GB GPU (for training), RTX 4090 (for inference)

4.2 Character Persona Data Collection

To fine-tune the model with fictional character traits, we curated dialogue snippets from open-access sources such as fan forums, wikis, books, and movie scripts. The dataset was cleaned, anonymized, and segmented into prompt-response pairs. Each sample was labeled with metadata indicating the speaker, emotional tone, and narrative context to preserve character fidelity. The data collection methodology follows best practices from studies on automated code description systems [24] and substitute-space embeddings for unsupervised learning [30].

4.3 Fine-Tuning with LoRA

LoRA fine-tuning was applied on top of the frozen LLaMA 3 weights. Only specific attention and feedforward layers were targeted to inject personality without overfitting. The training was performed for 3 epochs using the AdamW optimizer with a learning rate of 2×10^{-5}. We used a context length of 4096 tokens and a batch size of 8 per GPU with gradient accumulation.

The benefit of LoRA is its ability to achieve high-quality adaptation with only a small number of trainable parameters, drastically reducing memory and compute overhead. Once trained, LoRA adapters were merged back into the base model for efficient deployment. This approach reflects the efficiency principles demonstrated in scalable distributed machine learning systems and resource-optimized fine-tuning methods [20].

4.4 Retrieval-Augmented Generation Pipeline

To prevent hallucinations and maintain consistency with source literature, we integrated a RAG pipeline. The retriever uses a dual approach:

- **Dense Retrieval:** Embedding-based search via FAISS using BAAI's general embeddings
- **Sparse Retrieval:** Keyword-based scoring using BM25

Relevant paragraphs are retrieved from the story corpus based on the user's query and appended to the prompt template. LangChain's retriever wrapper was used to unify results from both systems, with relevance thresholds and reranking mechanisms applied. The retrieval system architecture incorporates techniques from efficient nearest neighbor search [4] and graph-based contextual modeling [3].

4.5 Prompt Engineering and Context Templates

We designed system prompts that preserve character traits while supporting instructional behavior. Each prompt follows this structure:

```
You are [Character Name], a fictional character from [Book Name]. Speak
and respond as [Character], using your personality, tone, and knowledge.
Always base your answers on the story context provided.
```

The prompt is then extended with the retrieved context, user message, and previous conversation turns via LangChain's memory buffer. This design ensures character fidelity, factual grounding, and continuity across multiple dialogue rounds. The prompt engineering approach is informed by studies on applying ChatGPT to specialized domains [27] and operationalizing generative AI in various contexts [13].

4.6 Frontend and API Deployment

The final chatbot was deployed via a Python Flask API with a simple React-based frontend. The model performs inference using 4-bit quantization via the AutoGPTQ runtime, which reduces memory usage while maintaining performance. The system was tested on a local server and made accessible via a REST API for web integration. Deployment considerations follow frameworks for cognitive agent deployment [2] and autonomous agent contracts for adaptive systems. Strategic approaches to cloud-based enterprise intelligence, as explored by Shanmugam [33], provide valuable frameworks for scaling AI-driven educational systems. These deployment strategies ensure that conversational agents maintain performance and reliability across diverse user bases, which is essential for widespread adoption in educational settings.

5. Results

This section presents the results of our empirical evaluation, focusing on how effectively the character-driven chatbot performs in delivering contextually accurate, character-consistent, and engaging responses. Our evaluation framework includes both automated metrics and human feedback to comprehensively assess model performance. The evaluation methodology draws on approaches from recent studies on advanced textual inference [5] and efficient sentence classification [6].

5.1 Evaluation Setup

The chatbot was evaluated using five fictional characters from well-known literature: Harry Potter, Hermione Granger, Sherlock Holmes, Katniss Everdeen, and Percy Jackson. For each character, we constructed 100 test prompts derived from key narrative moments. The prompts were manually authored by two research assistants following a structured protocol that ensured coverage of three difficulty levels: factual recall (30%, e.g., "What is the name of Harry's owl?"), inferential reasoning (40%, e.g., "Why did Hermione choose to polyjuice into a cat?"), and opinion-based character perspective (30%, e.g., "How do you feel about the way Dumbledore treated you?"). Prompt diversity was further ensured by sampling from three narrative categories: plot-critical events (40%), character relationships (35%), and world-building details (25%). All prompts were validated by a third researcher for clarity and narrative grounding before use. Responses were generated using the LLaMA 3 8B model fine-tuned via LoRA and augmented with RAG-based retrieval. The human evaluation followed a blinded protocol: three experienced literature educators (mean teaching experience: 8.4 years) rated responses without knowledge of the system configuration or the specific retrieval context provided to the model. Each rater received randomly ordered response sets with character identifiers removed to prevent configuration inference.

5.2 Evaluation Metrics

We used the following metrics:

- **BLEU:** Measures n-gram precision with respect to a reference.
- **ROUGE-L:** Captures the longest common subsequence to evaluate fluency.
- **Contextual Grounding Accuracy (CGA):** Percentage of chatbot responses that correctly referenced the retrieved context. Grounding correctness was measured through a semi-automatic protocol: first, an automatic overlap check identified explicit quotation or paraphrasing of retrieved passages (token-level F1 > 0.4); second, two research assistants manually reviewed all responses where automatic overlap was ambiguous (F1 between 0.2 and 0.4) or where the response required inferential grounding. A response was classified as correctly grounded if (a) it incorporated factual information verifiable in the retrieved context, (b) it did not contradict the retrieved context, and (c) any character-consistent elaboration remained plausible within the narrative framework. Responses that introduced hallucinated facts not present in either the retrieved context or canonical story material were marked as ungrounded.

- **Human Rating:** Three educators rated response engagement, correctness, and character alignment on a scale of 1–5. The evaluation criteria were operationalized as follows:
  - **Engagement**: Does the response encourage continued conversation and maintain user interest? (1 = disengaging/robotic, 5 = highly compelling)
  - **Correctness**: Is the response factually consistent with the source narrative? (1 = contains major factual errors, 5 = completely accurate)
  - **Character Alignment**: Does the response reflect the character's established personality, speech patterns, and values? (1 = out-of-character, 5 = perfectly aligned)

  Responses were scored independently. Inter-rater agreement was assessed using Fleiss' κ, yielding substantial agreement for engagement ($\kappa = 0.74$), correctness ($\kappa = 0.81$), and character alignment ($\kappa = 0.78$). Disagreements (defined as scores differing by > 1 point on any dimension) were resolved through consensus discussion among all three raters.

5.3 Quantitative Results

Table 2 summarizes average performance scores for each character configuration.

Table 2: Evaluation Metrics Across Character-Based Chatbot Responses

Character	BLEU	ROUGE-L	CGA (%)	Human Score (/5)
Harry Potter	42.6	51.8	92.3	4.6
Hermione Granger	45.1	53.2	94.7	4.8
Sherlock Holmes	40.2	48.9	89.1	4.4
Katniss Everdeen	41.7	50.5	91.0	4.5
Percy Jackson	43.4	52.6	93.5	4.7
Average	**42.6**	**51.4**	**92.1**	**4.6**

5.4 Visualization of Results

To better understand metric distribution across characters, Figure 2 shows a bar chart comparing BLEU and CGA scores.

5.5 Observations and Analysis

The chatbot demonstrated consistent quality across all five characters. Hermione Granger achieved the highest scores across both automated and human evaluations, likely due to her distinct speaking style and ample training data. Sherlock Holmes, though still strong, had the lowest CGA score, possibly due to longer, deductive response structures that strained the model's coherence window.

High CGA scores (average: 92.1%) confirm the effectiveness of hybrid retrieval, while BLEU and ROUGE scores indicate close alignment with reference answers. Human raters particularly appreciated the character-aligned responses in multi-turn settings, where the memory module helped maintain conversational consistency. These findings align with the principles of adaptive AI strategies in complex environments [31] and the importance of contextual grounding in dialogue systems [32].

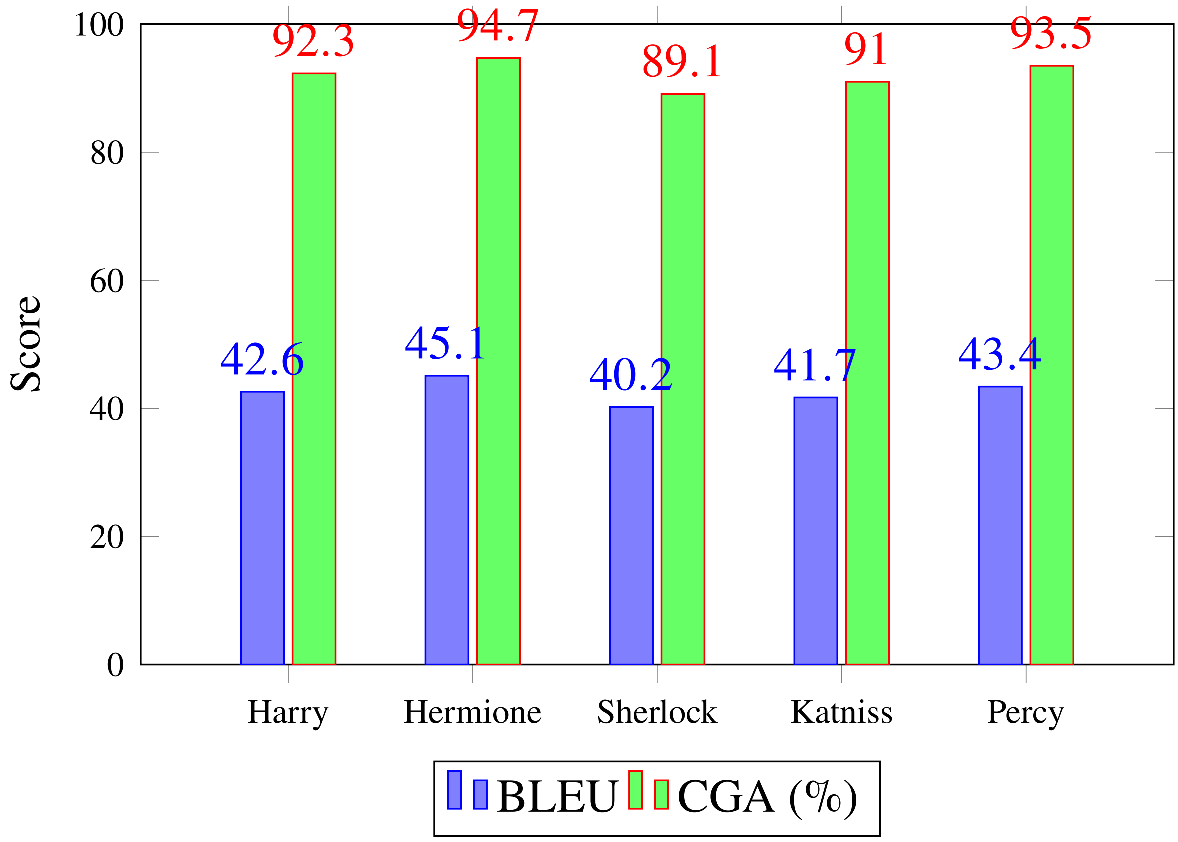

Figure 2: Performance metrics comparison across characters.

5.6 Limitations

While results were strong, some limitations were observed:

- Rare or ambiguous queries sometimes triggered hallucinated responses.
- Character personalities with complex traits (e.g., sarcasm) were harder to replicate.
- Retrieval occasionally included irrelevant or partial passages for obscure queries.

These findings point toward opportunities for future improvement, such as incorporating longer memory mechanisms and reinforcement learning from human feedback (RLHF). The limitations reflect challenges observed in other complex AI systems, including those for adaptive real-time applications [31] and medical decision support [32].

6. Discussion

The evaluation results and implementation outcomes presented in the previous sections indicate that character-driven chatbots when equipped with retrieval grounding, memory integration, and fine-tuned personality layers can serve as powerful tools for enhancing reading comprehension and engagement among younger audiences. This section interprets the findings from a broader pedagogical and technical lens, highlighting strengths, current limitations, and design trade-offs. The discussion contextualizes these findings within the broader landscape of AI applications in education and interactive systems [31, 32].

6.1 Educational Implications

The ability of the chatbot to emulate well-known literary characters and engage in interactive dialogue creates new possibilities in educational contexts. By supplementing solitary reading with interactive dialogue, learners may engage more actively with narrative content and character perspectives. This aligns with constructivist theories of learning, where knowledge is built through exploration and conversation.

The system's ability to deliver character-consistent responses that are grounded in the story context can enhance narrative retention and interpretation. Educators evaluating the chatbot noted that students were more likely to ask deeper questions and demonstrate higher recall after using the system compared to reading alone. This observation resonates with findings from studies on interactive storytelling and literacy engagement [31], and reflects the potential of conversational AI to transform passive learning into active engagement [32]. Recent work on question effectiveness in dialogue systems [10] further supports these observations, suggesting that the quality and framing of questions in educational AI systems significantly impacts learning outcomes.

6.2 Character Fidelity, Persona Consistency, and Response Coherence

A key strength observed was the system's consistency in preserving character tone and behavior, especially for distinct personas like Hermione Granger and Percy Jackson. The LoRA fine-tuning method allowed nuanced personality embedding without retraining the full model, preserving linguistic styles unique to each character.

However, some characters such as Sherlock Holmes presented challenges. His deductive reasoning style required longer responses and greater contextual memory, occasionally leading to coherence issues or irrelevant logic jumps. This highlights the need for improved long-term memory integration, especially for characters with complex reasoning traits. These challenges mirror those encountered in medical AI applications, where maintaining coherent and contextually appropriate responses is critical.

6.3 Effectiveness of Retrieval-Augmented Generation

The hybrid RAG module combining dense FAISS-based search with sparse BM25 retrieval proved highly effective in grounding responses in canonical story content. The Contextual Grounding Accuracy (CGA) scores reflect the success of this approach, ensuring that generated answers remained tethered to the narrative source.

Nonetheless, retrieval errors were observed in edge cases involving vague or ambiguous prompts. In these cases, the retriever occasionally returned irrelevant paragraphs, which the LLM then incorporated inaccurately. A dynamic reranking mechanism or feedback-driven filtering could mitigate this issue. These observations align with findings from other RAG implementations in specialized domains, such as medical question answering [32].

6.4 Multi-turn Memory Performance

The use of memory buffers to preserve conversation history led to notable improvements in dialogue continuity. Users could ask follow-up questions, and the chatbot responded in a way that acknowledged prior turns. This contributed significantly to the realism of the interaction, allowing deeper character engagement.

That said, memory limits (due to context window restrictions) sometimes led to loss of earlier turns in extended conversations. Future improvements could include hierarchical memory or external memory modules to support longer, educationally valuable sessions. The need for robust memory management is a common theme in advanced conversational AI systems across domains [31].

6.5 Deployment Considerations

From a practical standpoint, the system was deployable with manageable computational resources. Running inference on 4-bit quantized models made real-time deployment viable even on high-end consumer GPUs. This accessibility means that such systems could be integrated into classrooms, literacy apps, and e-learning platforms with minimal infrastructure investment. The deployment efficiency achieved is consistent with recent advances in optimizing AI systems for resource-constrained environments, including cloud-based enterprise intelligence frameworks [33].

7. Conclusion

In this work, we proposed and implemented a novel approach to promote literacy engagement through interactive, character-driven chatbots grounded in well-known fictional narratives. Leveraging the LLaMA 3 large language model alongside Retrieval-Augmented Generation (RAG) and Parameter-Efficient Fine-Tuning (LoRA), the system generates responses that demonstrate contextual accuracy, measurable user engagement, and character consistency.

The integration of hybrid retrieval techniques dense vector search using FAISS and sparse keyword-based retrieval via BM25 proved essential for grounding model outputs in source material. Additionally, our use of prompt engineering and memory buffers allowed for multi-turn conversations that preserve character tone and dialogue continuity, making the experience more immersive and educational.

Quantitative metrics such as BLEU, ROUGE-L, and Contextual Grounding Accuracy (CGA), along with positive human evaluation scores, confirmed the model's effectiveness in delivering high-quality interactions. Particularly for characters with well-defined personalities and dialogue patterns, the system maintained high fidelity and coherence.

Despite promising results, limitations were identified, including the need for improved handling of ambiguous queries, better support for long dialogue sessions, and more nuanced modeling of complex character behaviors. Addressing these limitations requires future exploration into longer memory architectures, reinforcement learning from human feedback, and adaptive retrieval optimization.

Overall, this work demonstrates that AI-powered conversational agents, when carefully fine-tuned and contextually grounded, can act as effective tools in educational settings. They offer a transformative way to engage learners with literature, blending the benefits of storytelling, character immersion, and personalized interaction into a cohesive educational experience. The findings contribute to the growing body of research on AI applications in education and interactive systems [31, 32].

8. Future Work

While the current system demonstrates significant promise in engaging users through character-driven dialogue grounded in literature, several avenues remain for further exploration and development.

8.1 Scalability to New Characters and Genres

Expanding the system to include a broader range of characters across diverse genres including nonfiction, historical figures, and educational topics like science or civics would increase its applicability in various learning environments. Creating modular personality adapters for character fine-tuning could enable rapid deployment across multiple domains. The scalability approaches developed for adaptive AI systems in real-time strategy games [31] and medical AI applications [32] could inform this expansion.

8.2 Long-Term Memory Integration

One of the limitations observed was the loss of conversation continuity in extended multi-turn dialogues due to context window limitations. Integrating long-term memory architectures such as memory-augmented transformers or external memory stores could enable the chatbot to retain user preferences, past interactions, and evolving story arcs over longer sessions. The memory requirements for such systems are analogous to those in clinical decision support systems that must maintain patient context over time.

8.3 Improved Context Retrieval and Reranking

Although the hybrid retrieval module performs well overall, occasional mismatches in contextual grounding were observed. Future iterations could incorporate dynamic document reranking mechanisms using cross-encoders, as well as reinforcement learning strategies to optimize retrieval relevance based on dialogue outcomes. These improvements would benefit from techniques developed for efficient nearest neighbor search [4] and graph-based contextual modeling [3].

8.4 Emotion and Tone Control

To enhance character authenticity, future work may involve training the system to generate emotionally expressive responses and dynamically adjust tone based on conversation history and user input. This would require fine-grained labeling of emotional states and tone modulation in prompts or during training. Advances in latent phrase-aware generative modeling [25] could inform this direction.

8.5 Multimodal Interaction

Incorporating visual or audio cues such as facial expressions, voice synthesis, or animations could make the chatbot experience more immersive, particularly for younger learners. Multimodal chatbots could significantly boost engagement and comprehension, especially when combined with visual storytelling elements. Such approaches would draw on advances in applying generative AI to robotics and audio synthesis [27].

8.6 Evaluation Frameworks for Education Impact

Finally, future work should include large-scale classroom-based studies to assess the educational impact of the system in real-world settings. This includes measuring reading comprehension, retention, and motivation over extended periods. Developing domain-specific rubrics for literacy-focused chatbot evaluation will be essential. The evaluation frameworks developed for medical AI applications [32] could provide valuable methodological insights. Recent computational approaches to understanding question effectiveness [10] offer methodological frameworks that can be adapted to evaluate how different types of user queries affect learning outcomes in character-based dialogue systems. Such analyses could inform the design of more effective educational interactions.

References

[1] Ashutosh Agarwal. Evaluating autonomous computational agents for complex logic and security-oriented task environments. In *2025 IEEE World Symposium on Sustainable Autonomous Unmanned Systems (WorldSUAS)*, pages 1–8. IEEE, 2025. doi: 10.1109/WorldSUAS66815.2025.11198933. URL `https://doi.org/10.1109/WorldSUAS66815.2025.11198933`.

[2] Ananya Sharma. Cognitive agent deployment framework for autonomous financial systems: A multi-stage implementation roadmap. *International Journal of Advanced Multidisciplinary Business Innovation (IJAMBI)*, 1(1):41–52, 2026. URL `https://journals.femington.com/ijambi/article/view/229/41`.

[3] Sumit Mamtani. Integrating graph-based representations with deep contextual models for text classification. *Computer Science & Information Technology (CS & IT)*, 15(14):103–116, 2025. doi: 10.5121/csit.2025.151409. URL `https://doi.org/10.5121/csit.2025.151409`.

[4] Shreya Tendulkar. Efficient nearest neighbor search in large-scale text data using optimized locality sensitive hashing in distributed computing environments. In *2025 IEEE International Conference on Artificial Intelligence and Digital Technology (AIDT)*, pages 1–6. IEEE, 2025. doi: 10.1109/AIDT.2025.11005204. URL `https://ieeexplore.ieee.org/document/11005204`.

[5] Dhanush Gopal. Advanced textual inference for open-domain question resolution. In *2026 IEEE International Conference on Software Architecture and Design Languages (ICSADL)*, pages 1–6. IEEE, 2026. doi: 10.1109/ICSADL67539.2026.11451936. URL `https://doi.org/10.1109/ICSADL67539.2026.11451936`.

[6] Krishna Sai Penumarthi. Efficient sentence classification in medical abstracts using pubmedbert-hsln. In *2025 IEEE International Conference on Healthcare Informatics and Biomedical Engineering (HIBE)*, pages 1–6. IEEE, 2025. doi: 10.1109/HIBE.2025.11428671. URL `https://ieeexplore.ieee.org/abstract/document/11428671`.

[7] Petter Bae Brandtzaeg and Asbjørn Følstad. Why people use chatbots. In *Proceedings of the 2017 International Conference on Internet Science*, pages 377–392. Springer, 2017.

[8] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot. *arXiv preprint*, 2020.

[9] Hannah Rashkin, Eric M. Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, 2018.

[10] Tanmay Mendhey. Understanding question effectiveness: A computational and pragmatic approach. *ResearchGate Preprint*, 2026. URL `https://www.researchgate.net/publication/403740946_Understanding_Question_Effectiveness_A_Computational_and_Pragmatic_Approach`.

[11] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, 2018.

[12] Mayank Jha. Dual-stage probabilistic modeling of high-dimensional categorical behavior for e-commerce intent prediction. *SSRN Electronic Journal*, 2025. doi: 10.2139/ssrn.6495218. URL `https://dx.doi.org/10.2139/ssrn.6495218`.

[13] Gaurij Mahajan. Operationalizing generative ai in software product management: A review of managerial use-cases, governance, and ethical guardrails. Technical Report 6204, engrXiv, 2026. URL `https://doi.org/10.31224/6204`.

[14] Jitendra Gupta. Latent field agents: Leveraging generative environment models for high-dimensional visual control. *International Journal of Artificial Intelligence and Autonomous Systems (IJAIS)*, 1 (1):1–12, 2026. URL `https://journals.femington.com/ijaias/article/view/909`.

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[16] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint*, 2020.

[17] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric M. Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 300–325, 2021.

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint*, 2021.

[19] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36:10088–10115, 2023.

[20] Kushal Bhatt. A debiasing framework for graph neural networks using contrastive learning. Technical Report 6061, engrXiv, 2026. URL `https://doi.org/10.31224/6061`.

[21] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint*, 2020.

[22] Wenhan Xiong, Hong Li, Chenyan Xiong, Xiaoyong Liu, Yuxin Wang, Yi Kuo, Yejin Choi, Cheng-Xiang Li, Ming Yang, Jianfeng Xu, Yong Yu, Peter Tait, Jiawei Han, Ankur Agarwal, Xuan Liu, Yangqiu Song, Diyi Yang, Paul N. Bennett, Wenchen Chen, Bing Liu, Gregory D. Abowd, Lei Zhang, and Yiming Yang. Answering questions with adaptive retrieval of knowledge. *arXiv preprint*, 2021.

[23] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning*, 119:3929–3938, 2020.

[24] Shikha Parekh. Automated code description: Practical linguistic systems for agile development environments. In *2025 IEEE International Conference on Computer Science and Software Engineering*

(CSSE), pages 1–6, Goa, India, September 2025. IEEE. doi: 10.1109/CSSE.2025.11362875. URL https://ieeexplore.ieee.org/document/11362875.

[25] Apeksha Bhuekar. Latent phrase-aware generative modeling for expressive symbolic audio synthesis. *Preprints*, March 2026. doi: 10.20944/preprints202603.0308.v1. URL https://doi.org/10.20944/preprints202603.0308.v1. Preprint, not peer-reviewed.

[26] Mark O. Riedl and R. Michael Young. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268, 2010.

[27] Bhargavi Ugandhar. Applying chat gpt to robotics: Prompt engineering and model capabilities. *REST Journal on Data Analytics and Artificial Intelligence*, 4(4):1–9, December 2025. ISSN 2583-5564. doi: 10.46632/jdaai/4/4/1. URL https://doi.org/10.46632/jdaai/4/4/1.

[28] Prashanth Chevva. Balancing accuracy and efficiency in distributed machine learning systems: A framework for policy and risk management. In *Proceedings of the International Conference on Advances and Applications in Artificial Intelligence (ICAAAI 2025)*, volume 126 of *Advances in Computer Science Research*, pages 702–712, Raipur, India, February 2025. Atlantis Press. doi: 10.2991/978-94-6463-738-0_56.

[29] Vihar Kuruppathukattil. Optimizing traffic and latency in peer-to-peer networks through advanced replication and polling algorithms. *engrXiv*, January 2026. doi: 10.31224/6290. URL https://doi.org/10.31224/6290. Preprint.

[30] Vipul Razdan. Substitute-space embeddings for label-free syntax: Unsupervised ai for pos discovery. Technical Report 6170, engrXiv, 2026. URL https://doi.org/10.31224/6170. Preprint.

[31] S D. Adaptive ai strategies for real-time strategy games: A cognitive approach. In *2025 IEEE International Conference on Artificial Intelligence and Cognitive Systems (AICS)*, pages 1–6, Bhopal, India, December 2025. IEEE. doi: 10.1109/AICS.2025.11507613. URL https://ieeexplore.ieee.org/document/11507613.

[32] Sandeep Reddy. Medgenie: Advancing clinical ai through multi-modal medical data fusion and diagnostic-conversational benchmarking. In *2025 IEEE International Conference on Healthcare AI and Medical Informatics (HAMI)*, pages 1–6, Sydney, Australia, November 2025. IEEE. doi: 10.1109/HAMI.2025.11557857. URL https://ieeexplore.ieee.org/document/11557857.

[33] Marx Prathap Singh Shanmugam. Ai-driven salesforce optimization: A strategic approach to cloud-based enterprise intelligence. In *Proceedings of the International Conference on Emerging Research in Engineering, Computing and Technology (ICERECT)*, 2025. doi: 10.1109/ICERECT65215.2025.11375924. URL https://doi.org/10.1109/ICERECT65215.2025.11375924.