Overcoming Context Length Limitations in LLM’s Integrating LSTM, Retrieval-Augmented Generation, and Agentic Frameworks for Enhanced Business Data Analysis

Dhruvansh Gandhi[1], Aryan Giri[2], Prof. Swati Uparkar[3]

[1,2]Student, Artificial Intelligence & Data Science, Shah and Anchor Kutchhi Engineering College
[3]Professor, Artificial Intelligence & Data Science, Shah and Anchor Kutchhi Engineering College

Abstract

Large Language Models (LLMs) such as GPT and BERT demonstrate remarkable capabilities in various natural language processing (NLP) tasks. However, their performance is constrained by context length limitations, leading to inefficiencies in processing extended text sequences. This paper explores the challenges posed by context length limitations and proposes innovative solutions combining Long Short-Term Memory (LSTM), Retrieval-Augmented Generation (RAG), and Agentic Framework. We present an AI-powered solution tailored for businesses, enabling efficient data processing, analysis, and visualization. The solution integrates actionable insights, streamlines operations, and drives growth through automated data integration, AI-powered analytics, and interactive visualizations.

Keywords: Large Language Models (LLMs), GPT, Natural Language Processing (NLP), Long Short-Term Memory (LSTM), Retrieval-Augmented Generation (RAG), Agentic Framework, AI-powered Data Analytics

1. Introduction

Large Language Models (LLMs) have transformed NLP by enabling machines to understand and generate human-like text. Despite their transformative capabilities, LLMs face inherent context length limitations, which hinder their ability to process long documents, books, or extended conversations. This limitation poses challenges in domains requiring comprehensive data analysis or extended text understanding.

The paper proposes leveraging LSTM’s sequential learning capabilities, RAG’s external knowledge retrieval, and an Agentic framework to address these limitations. By enhancing LLM context handling, the proposed solution ensures seamless data processing and advanced analytics for businesses.

2. Problem Statement

- Context length limitations in LLMs constrain their ability to:
- Handle long-form text comprehension and generation.
- Retain coherence over extended interactions.
- Integrate external, domain-specific knowledge efficiently.

These challenges negatively impact applications such as legal document analysis, multi-turn customer support, and academic research, where extensive context is crucial.

3. Related Work

Numerous strategies have been explored to address LLM context limitations, including:

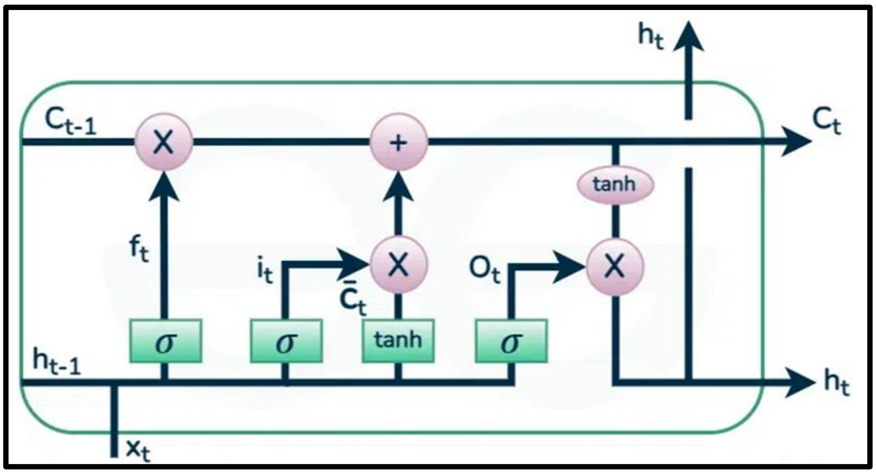
- Chunking and hierarchical modeling.
- Attention mechanisms with memory extensions.
- Hybrid approaches integrating neural networks and retrieval-based systems.

While effective, these approaches often fail to balance computational efficiency with scalability.

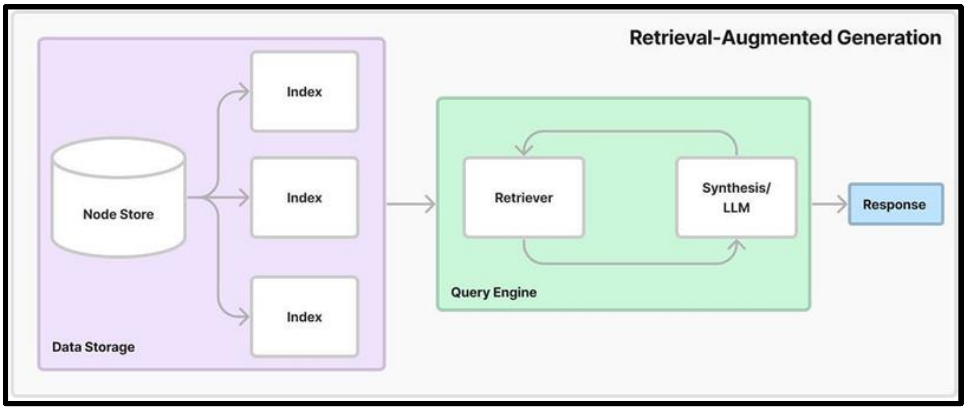
4. Proposed Solution

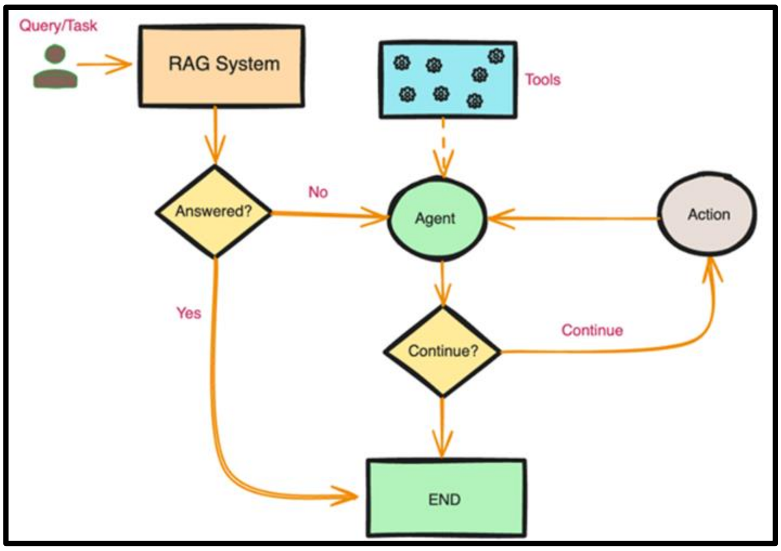
The proposed framework combines LSTM, RAG, and Agentic approaches:

4.1 LSTM for Sequential Memory LSTM models excel in retaining long-term dependencies, making them ideal for maintaining coherence across extended text sequences. By integrating LSTM with LLMs, context length limitations can be mitigated, allowing for sequential processing of large datasets.



4.2 Retrieval-Augmented Generation (RAG) - RAG incorporates external knowledge retrieval into LLM workflows. By querying a structured knowledge base or database, RAG retrieves relevant information to supplement LLM's limited context. This ensures accurate, context-aware responses without exceeding token limits.





4.3 Agentic Approach The Agentic approach involves designing AI systems capable of:

- Proactively retrieving and integrating external knowledge.
- Iteratively analyzing and summarizing extended text.
- Orchestrating multiple LLMs or models for task-specific processing.

The agent acts as a mediator, dynamically managing resources to optimize performance and scalability.

5. AI-Powered Solution for Businesses

The solution is designed to process, analyze, and visualize business data efficiently. Key features include.

- **Automated Data Integration:** Seamlessly aggregates data from multiple sources, ensuring consistency and accuracy.
- **AI-Powered Analytics:** Applies advanced analytics to uncover actionable insights, optimize operations, and forecast trends.
- **Interactive Visualizations:** Transforms insights into intuitive visual representations, aiding decision-making.

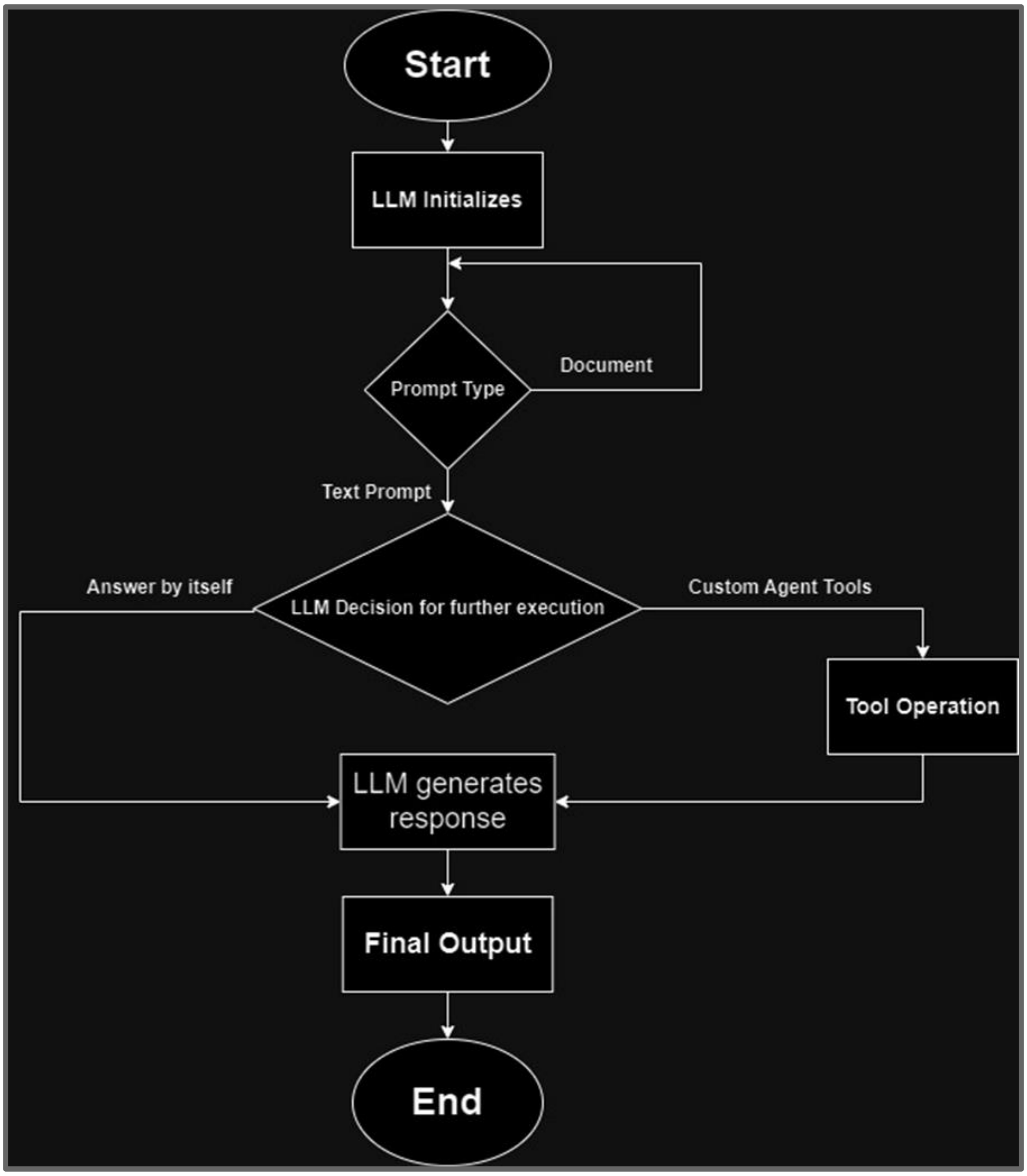
This integrated framework enhances productivity, decision-making, and business growth.

6. Implementation

6.1 Architecture The system comprises:

- An LSTM-augmented preprocessing layer.
- RAG-based knowledge retrieval module.
- Agentic framework coordinating LLM and external resources.

6.2 Workflow



1. Data is preprocessed using LSTM for sequential encoding.
2. Relevant knowledge is retrieved via RAG.
3. The Agentic system orchestrates LLM workflows, ensuring seamless integration and analysis.

6.3 Tools and Technologies

- TensorFlow/PyTorch for LSTM modeling.
- Elasticsearch for RAG knowledge retrieval.
- OpenAI API for LLM integration.
- Power BI/Tableau for data visualization.

7. Results and Evaluation

Preliminary experiments demonstrate:

- Enhanced context retention and coherence.

- Reduced computational overhead compared to hierarchical chunking.
- Improved business outcomes through actionable insights and data visualization.

8. Conclusion and Future Work

This research presents a novel framework combining LSTM, RAG, and Agentic approaches to overcome LLM context length limitations. Future work will focus on:

- Expanding knowledge base integration.
- Exploring real-time, multi-modal data processing.
- Enhancing scalability for enterprise-level deployments.

References

1. Vaswani, A., et al. (2017). Attention is All You Need.
2. Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
3. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory.
4. OpenAI. (2023). GPT-4 Technical Report.
5. .Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
6. Brown, T., et al. (2020). Language Models are Few-Shot Learners.
7. Raffel, C., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.
8. Dai, Z., et al. (2019). Transformer-XL: Attentive Language Models Beyond a FixedLength Context.
9. Karpukhin, V., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering.
10. Rajpurkar, P., et al. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text.
11. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.
12. Thoppilan, R., et al. (2022). LaMDA: Language Models for Dialog Applications.
13. Guu, K., et al. (2020). REALM: Retrieval-Augmented Language Model PreTraining.
14. Radford, A., et al. (2019). Language Models are Unsupervised Multitask Learners.
15. Beltagy, I., et al. (2020). Longformer: The Long-Document Transformer.