Limitations of Retrieval-Augmented Generation (RAG)

Md. Sopon Abdullah

Department of Computer Science
University Chittagong

October 30, 2025

Overview

1. Introduction
2. Core Limitations
3. Summary

What is RAG?

- **Retrieval-Augmented Generation (RAG)** combines:
  - **Retrieval:** Fetch relevant documents from an external knowledge base.
  - **Generation:** Use a language model (LLM) to generate responses using retrieved context.
- Enhances factual accuracy and domain specificity.
- Commonly used in *domain-specific chatbots, academic assistants, and knowledge QA systems.*

1. Dependence on Retrieval Quality

- The model's output is only as good as the retrieved documents.
- Poorly retrieved or irrelevant documents lead to incorrect or misleading answers.
- Retrieval model (e.g., BM25, DPR, ColBERT) limits accuracy if not domain-tuned.

2. Context Length Limitation

- LLMs have fixed input context windows.
- Only a few retrieved passages can be fed at once.
- Important information may be truncated or ignored.

3. Knowledge Inconsistency

- Conflicts between:
  - Retrieved facts and model's internal knowledge.
  - Multiple retrieved sources with contradicting information.
- Leads to confusing or incoherent responses.

4. Lack of Dynamic Reasoning

- RAG systems retrieve static text.
- They cannot perform reasoning over multiple retrieved pieces effectively.
- Often fail in multi-hop question answering.

5. Domain Adaptation Challenges

- Requires domain-specific tuning of both retriever and generator.
- LLMs may not understand specialized terminology or structure of documents.
- Costly to curate, clean, and index large domain datasets.

6. Latency and Computational Cost

- RAG involves multiple stages:
  1. Document retrieval
  2. Encoding
  3. Response generation
- Increases response time and computational overhead.

7. Evaluation Difficulty

- Difficult to measure factual accuracy and citation correctness.
- No standardized metrics for RAG-specific evaluation.
- Human evaluation often required.

Summary

Main Limitations

- Dependence on retrieval quality
- Limited context window
- Knowledge inconsistency
- Weak multi-hop reasoning
- Domain adaptation cost
- High latency
- Evaluation challenges

Summary

Main Limitations

- Dependence on retrieval quality
- Limited context window
- Knowledge inconsistency
- Weak multi-hop reasoning
- Domain adaptation cost
- High latency
- Evaluation challenges

Future Directions

- Hybrid retriever models
- Adaptive context compression
- Retrieval-grounded reasoning models

Thank You!

Questions?