



Računarski fakultet

Review on the topic of

Retrieval-Augmented Generation (RAG)

Course: Data mining

Advisor:

prof. Nemanja Ilić

Student:

Vanja Kovinić

Belgrade, January 2025.

Contents

1	Introduction	1
2	Large Language Models: Background and Limitations	2
2.1	Understanding Language Models	2
2.2	Limitations of Traditional LLMs	2
2.2.1	The Knowledge Cutoff Problem	2
2.2.2	Fine-tuning Challenges	3
2.3	Prompt Engineering as a Partial Solution	3
2.3.1	Benefits of Prompt Engineering	3
2.3.2	Limitations of Prompt Engineering	4
2.4	The Need for Alternative Approaches	5

1 Introduction

Recent advances in **Large Language Models (LLMs)** have revolutionized natural language processing, demonstrating remarkable capabilities in tasks ranging from text generation to complex reasoning. Models like **GPT-4** [1] and **LLaMA** [2] have achieved good performance in understanding and generating human-like text. However, these models face significant limitations that impact their practical deployment and reliability in real-world applications.

One of the fundamental challenges with traditional LLMs is their reliance on static knowledge encoded in model parameters during training. This approach presents several critical issues.

- First, the knowledge contained within these models becomes outdated as new information emerges, requiring costly and time-consuming re-training or fine-tuning processes.
- Second, the models are constrained by their parameter count, with even the largest models unable to capture all potentially relevant information.
- Third, fine-tuning these models on new data can lead to forgetting, where the model loses previously acquired knowledge while adapting to new information.

To address these limitations, researchers have developed **Retrieval-Augmented Generation (RAG)** [3], a novel architecture that combines the generative capabilities of LLMs with dynamic access to external knowledge sources. RAG represents a paradigm shift in how we approach language models, moving from purely parametric knowledge to a hybrid system that can actively retrieve and incorporate relevant information during inference.

The core innovation of RAG lies in its ability to decompose the knowledge acquisition and generation processes. Instead of relying solely on information encoded in model parameters, RAG systems maintain a separate knowledge base that can be easily updated without modifying the underlying language

model. This separation provides several advantages: it allows for real-time knowledge updates, reduces the memory requirements for storing information, and maintains consistent access to accurate, verifiable information.

2 Large Language Models: Background and Limitations

2.1 Understanding Language Models

LLMs are probabilistic systems trained to predict the next token in a sequence based on the provided context. During training, these models process vast amounts of text data, learning to capture patterns, relationships, and factual information within their parameters. The knowledge acquired during training is encoded in billions of parameters - for instance, **GPT-3** [4] contains 175 billion parameters, while models like **LLaMA 2** [5] can range from 7 billion to 70 billion parameters.

The training process involves exposing the model to a diverse corpus of text data, allowing it to learn language patterns and accumulate knowledge about various topics. This knowledge, however, is static and bounded by the training cutoff date, creating what is known as the knowledge cutoff problem. Any information or events occurring after the training period are unknown to the model, regardless of their significance.

2.2 Limitations of Traditional LLMs

2.2.1 The Knowledge Cutoff Problem

The knowledge cutoff represents a fundamental limitation of LLMs. For example, a model trained with data up to 2022 would have no inherent knowledge of significant events or technological developments that occurred in 2023. This limitation becomes increasingly problematic as time passes, making the model's knowledge progressively outdated.

2.2.2 Fine-tuning Challenges

Organizations facing the knowledge cutoff problem might consider fine-tuning as a solution. However, fine-tuning presents several significant challenges:

- **Cost Implications:** Fine-tuning large models requires substantial computational resources and expertise, making it prohibitively expensive for many organizations.
- **Parameter Limitations:** The number of parameters in a model may not be sufficient to capture all desired knowledge. Even large models like **LLaMA 2** (70B) have finite capacity for storing information.
- **Forgetting:** Fine-tuning can lead to the model forgetting previously learned information. For instance, a model heavily fine-tuned on recent events might lose its ability to reason about historical contexts.

2.3 Prompt Engineering as a Partial Solution

Prompt engineering has emerged as a technique to mitigate the knowledge cutoff problem by providing relevant context within the input prompt. This approach allows users to "teach" the model new information on the fly, enabling it to reason about and generate responses based on up-to-date information (*Figure 1*).

2.3.1 Benefits of Prompt Engineering

- **Dynamic Knowledge Integration:** Users can provide current information without modifying the model's parameters.
- **Flexibility:** The context can be tailored to specific queries or use cases.
- **Verifiability:** Since the context is explicitly provided, the source of information is transparent and can be verified.

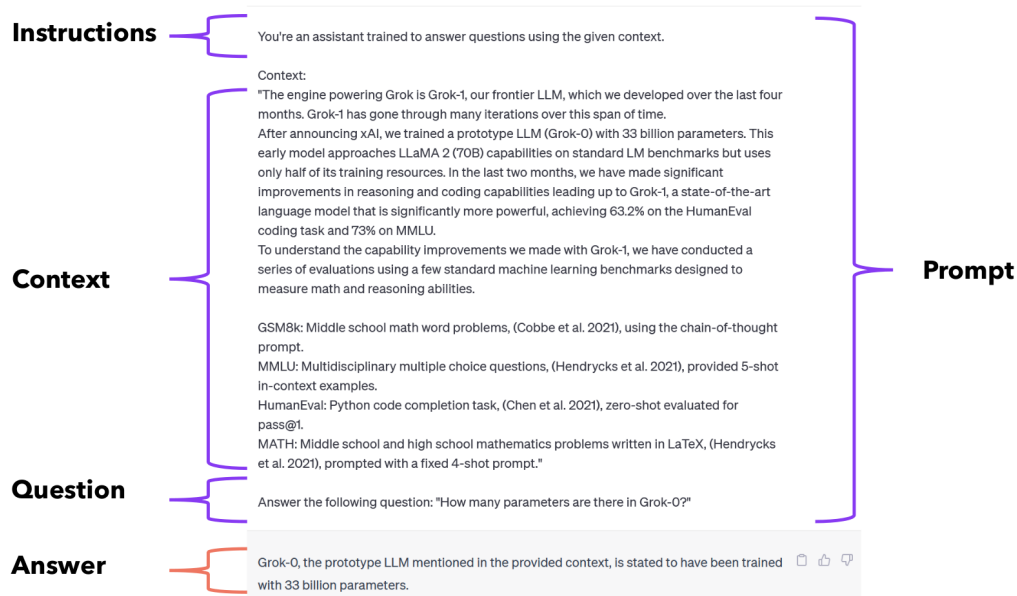


Figure 1: Example where user provides context with answer inside the prompt to enhance the LLMs response [6]

2.3.2 Limitations of Prompt Engineering

However, prompt engineering comes with significant drawbacks:

- **Context Window Constraints:** Every token used for context reduces the available space for the actual query and response. For example, if a model has a 4,096 token limit, providing 3,000 tokens of context leaves only 1,096 tokens for the query and response.
- **Manual Intervention:** Users must manually select and provide relevant context, which can be time-consuming and may require expertise in the subject matter.
- **Context Selection Challenge:** Determining which context is relevant for a given query can be difficult, especially for complex or ambiguous questions.

2.4 The Need for Alternative Approaches

While prompt engineering provides a workable solution for incorporating new knowledge into LLM interactions, its limitations make it suboptimal for many real-world applications. These limitations have motivated the development of more sophisticated approaches to augmenting LLM knowledge, leading to the emergence of Retrieval-Augmented Generation (RAG) as a promising solution. RAG addresses many of the limitations of both traditional LLMs and prompt engineering by providing a systematic way to incorporate external knowledge while maintaining efficiency and scalability.