Better LLM Responses in ServiceNow: RAG vs. Prompt Engineering ✨

DT ADVISORY

In This Article

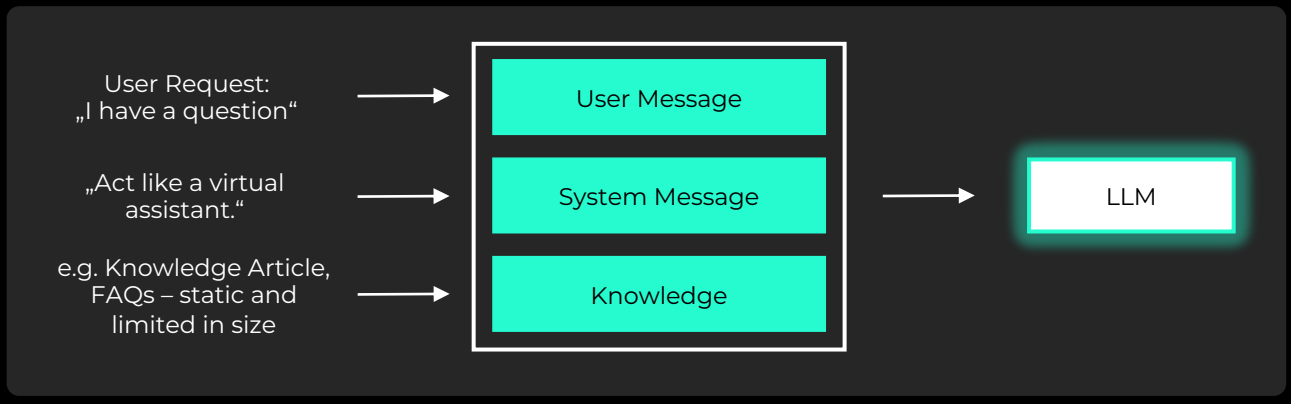
- ✓ Prompt Engineering
- ✓ Retrieval-Augmented Generation (RAG)
- ✓ Implementation in ServiceNow

Bring Your Own Data: Effective LLM Optimization with Prompt Engineering and RAG

To get the best performance from large language models (LLMs) for complex tasks, you can use three advanced techniques: prompt engineering, fine-tuning, or Retrieval-Augmented Generation (RAG).

These methods enhance the effectiveness of LLMs by tailoring their capabilities to meet custom requirements. In this article, we will concentrate on prompt engineering and Retrieval-Augmented Generation (RAG), as these offer more accessible and cost-effective approaches for optimizing LLM performance compared to fine-tuning.

Prompt Engineering



Description

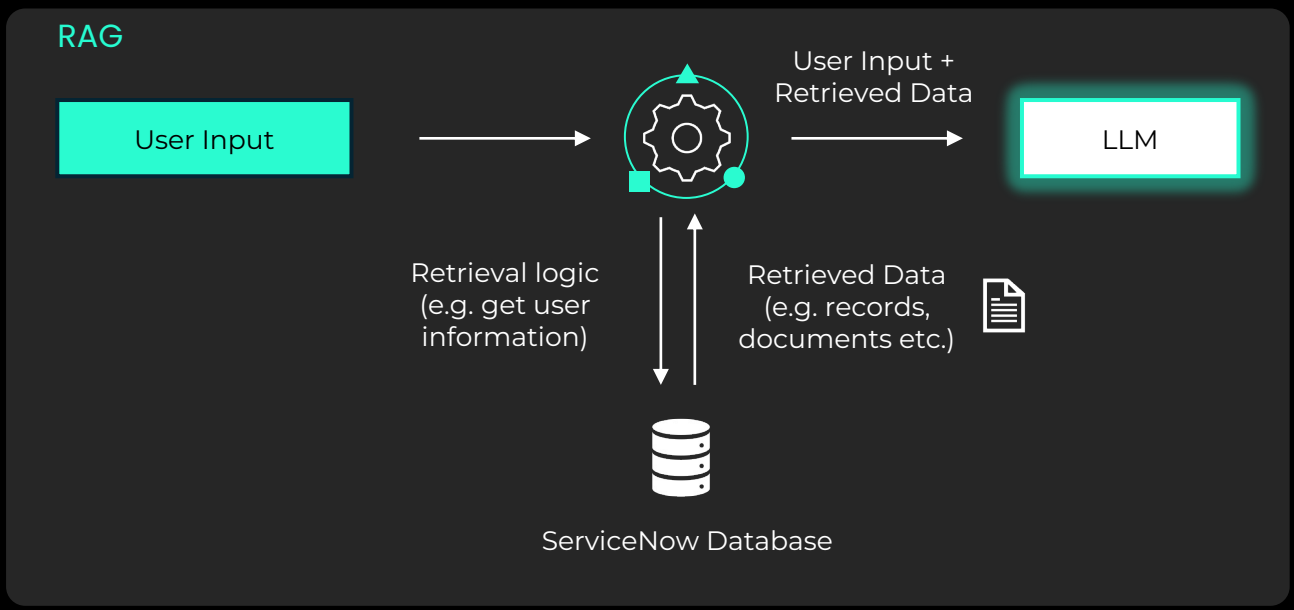
Prompt engineering involves creating precise, detailed, but static instructions to guide LLMs in generating accurate and relevant outputs. This includes three main components:

1. User message: Provides specific input or questions.
2. System message: Sets the context and rules.
3. Knowledge: Adds domain knowledge by providing context specific knowledge articles or FAQs.

These elements are integrated into the prompt to shape the AI's output.

Benefits and Limitations

- ☑ Low implementation Effort
- ☑ Sufficient for simple standard tasks
- ☒ Limited prompt length/context
- ☒ No data access control



Description

Retrieval-Augmented Generation (RAG) enhances language models by retrieving and incorporating relevant external data. It involves two steps:

1. Retrieving documents or passages from provided knowledge (PDFs, Documents, etc.) based on an input query. This query is generated using the user input.
2. Using this information to generate more factual, informative, and grounded outputs through a language model by combining the retrieved information with the user input and sending it to an LLM.

Benefits and Limitations

- ☑ No additional training efforts
- ☑ Supports domain knowledge
- ☑ Can handle large amount of data
- ☒ Retrieval logic needs to be defined and built
- ☒ Slower execution due to retrieval logic

ServiceNow Implementation

Prompt Engineering and Retrieval-Augmented Generation (RAG) can be implemented using custom-tailored solutions. Each AI step can be seamlessly integrated into the ServiceNow Flow Designer, allowing AI calls to be executed within a flow step. Depending on the architecture, a MID server can be used to access a Large Language Model (LLM), or the query can be made directly via the LLM's API. Additionally, external databases such as ChromaDB can be connected. For structured data, ServiceNow's Access Control Lists (ACLs) can be utilized to directly access and enrich AI queries with data from ServiceNow itself.

Ready to elevate your ServiceNow Platform?

Discover the full potential of your ServiceNow platform with our help. Prompt engineering is just the start. Our experts can guide you through advanced AI concepts like Retrieval-Augmented Generation (RAG) and fine-tuning specifically for ServiceNow. Have questions or need personalized advice? We're here to help you leverage Now Assist capabilities and support your AI strategy. Contact DT Advisory today and let's unlock the full potential of AI for your business together.

Sören Maucher
ServiceNow Solution Architect
soeren.maucher@dt-advisory.ch

DT Advisory AG
Richtiarkade 4, 8304 Wallisellen
www.dt-advisory.ch