Optimizing AI Requests in ServiceNow: RAG vs. Prompt Engineering

DT ADVISORY

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

To achieve optimal performance of large language models (LLMs) for specific tasks, one can utilize one of three advanced techniques: prompt engineering, fine-tuning, or Retrieval-Augmented Generation (RAG).

These methods are designed to enhance the effectiveness of LLMs by tailoring their capabilities to meet particular requirements. However, in this article, we will not focus on fine-tuning, as it is often a complex and resource-intensive process. Instead, we will concentrate on prompt engineering and RAG, which offer more accessible and cost-effective approaches for optimizing LLM performance.

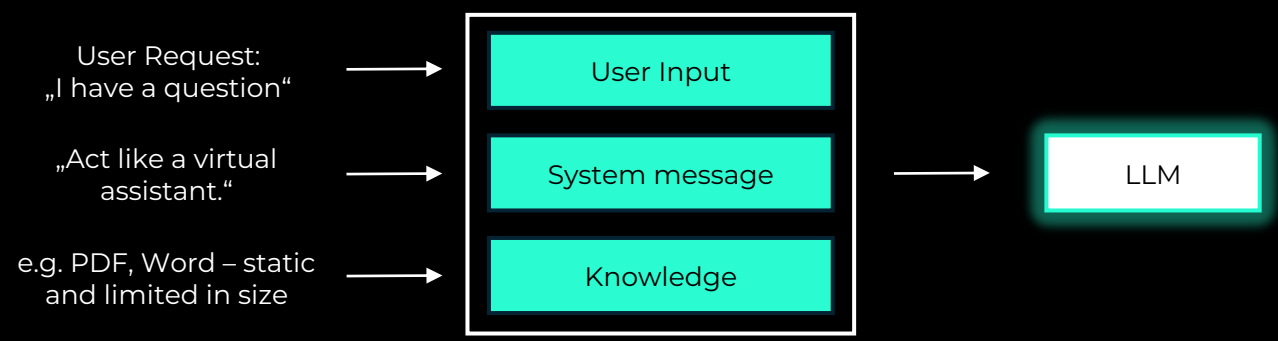
Main Reasons for including personal data in AI

Enhanced User Experiences
- ✓ Personal information enables more accurate, relevant AI responses.

Effective Task Automation
- ✓ AI automates tasks better with detailed user profiles.

Proactive Support
- ✓ Predictive analytics prevent issues using personal data patterns.

Prompt Engineering



Description

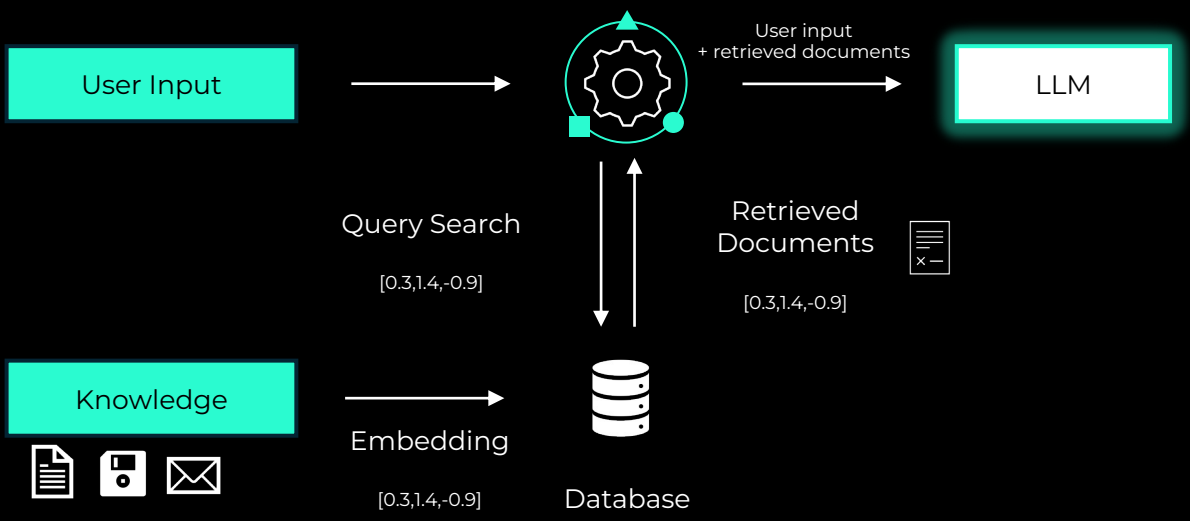
Prompt engineering involves creating precise, detailed **but static** instructions to guide LLMs in generating accurate and relevant outputs. This includes three main components:

1. The user message provides specific input or questions
2. The system message sets the context and rules
3. Additionally, Knowledge like PDFs can be added to the prompt, but the size is limited to the LLM constraints

All this information is passed into the prompt to steer the AI's response.

- ☑ Low implementation Effort
- ☑ Enough for simple standardized tasks
- ☒ Limited Prompt length/context
- ☒ No data access control

RAG



Description

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

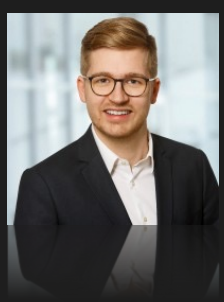
1. Retrieving documents or passages from provided knowledge (PDFs, Documents, etc.) based on an input query. This query is generated using the user input.
2. Using this information to generate more factual, informative, and grounded outputs through a language model by combining the retrieved information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

- ☑ No additional training efforts
- ☑ Supports domain knowledge
- ☑ Can handle large amount of data
- ☒ Need to build retrieval logic
- ☒ Slower execution due to additional
- ☒ Retrieval logic

ServiceNow Implementation

Prompt Engineering and RAG can be implemented using custom tailored solutions.

Contact our Expert



Sören Maucher
ServiceNow Architect
Soeren.Maucher@dt-advisory.ch

Visit DT Advisory