Hands-On GenAI: LLMs, RAGs, and Agentic Systems for Beginners

Day 7

Adya Bhat and Tejas Venugopalan

Drawbacks of LLMs

source: https://aws.amazon.com/what-is/retrieval-augmented-generation/

providing:

- false information
- old/ generic responses
- responses from non-authoritative sources
- incorrect response due to terminology confusion

check this out for some examples of LLM hallucinations:
https://www.reddit.com/r/ChatGPT/comments/12fmrcd/examples_of_gpt4_hallucination/

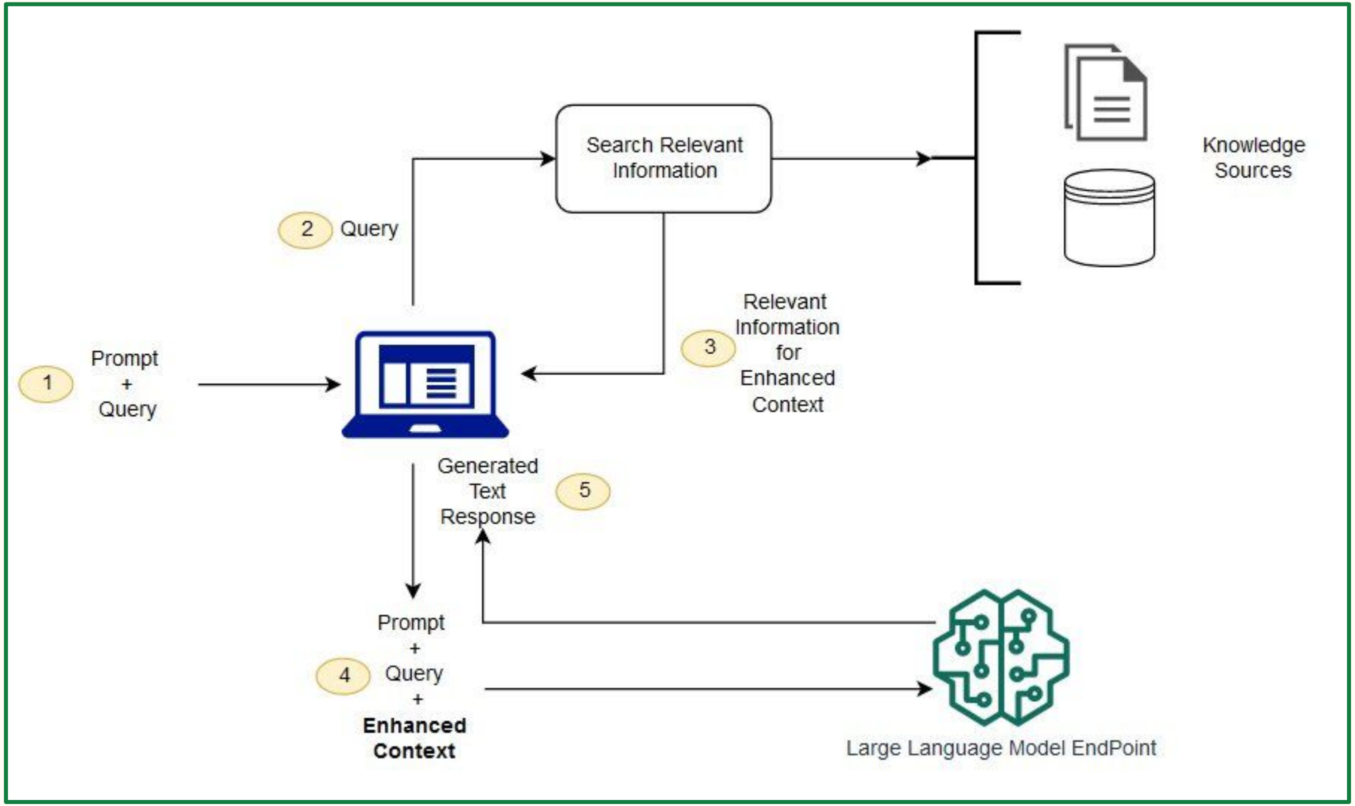
Steps in Natural Language Processing

- steps in an ML experiment for NLP:
  - **data:** text corpus
  - **data preprocessing:** tokenization, stopword removal, embedding generation (for tokens, sentences, topics, documents, etc.)
  - **task:** text generation, summarization, question and answer, translation etc.
  - **model:** text sequence, context modelling
  - **training**
  - **evaluation:** using metrics for accuracy of translation (e.g. BLEU), comparison of summarized text and reference (initial) text (e.g. ROUGE)

Enhancing Output of LLMs using RAG Architecture

- Retrieval- Augmentation- Generation architecture
- Components:
  - External data: text corpus converted to embeddings → stored in a vector database
  - Retrieval: user's question → vector representation → relevance search in vector database (similarity search)
  - Augmentation: user's question + retrieved information from vector database → converted to a prompt (retrieved information is added to augment the prompt to the LLM)
  - Generation: LLM generates an answer using the augmented prompt

RAG



Search Relevant Information

Knowledge Sources

2 Query

3 Relevant Information for Enhanced Context

1 Prompt + Query

Generated Text Response 5

4 Prompt + Query + **Enhanced Context**

Large Language Model EndPoint

source: https://aws.amazon.com/what-is/retrieval-augmented-generation/

Thank You!