Natural Language Processing
Large Languade Models Usage Patterns

Felipe Bravo-Marquez

November 10, 2023

Introduction

- Since the inception of Large Language Models, different usage patterns of this technology have emerged.
- In this talk we try to organize these patterns and give a general overview of them.

What is an LLM

- An autoregressive language model trained with a Transformer neural network on a large corpus and a large parameter space.
- It is usually later aligned to work as a user assistant.
- Some are private: Google Bard, ChatGPT
- Others are open: LLama2, Falcon
- `https://ai.meta.com/llama/get-started/?trk=feed_main-feed-card_reshare_feed-article-content`.

LLMs Usage Patterns

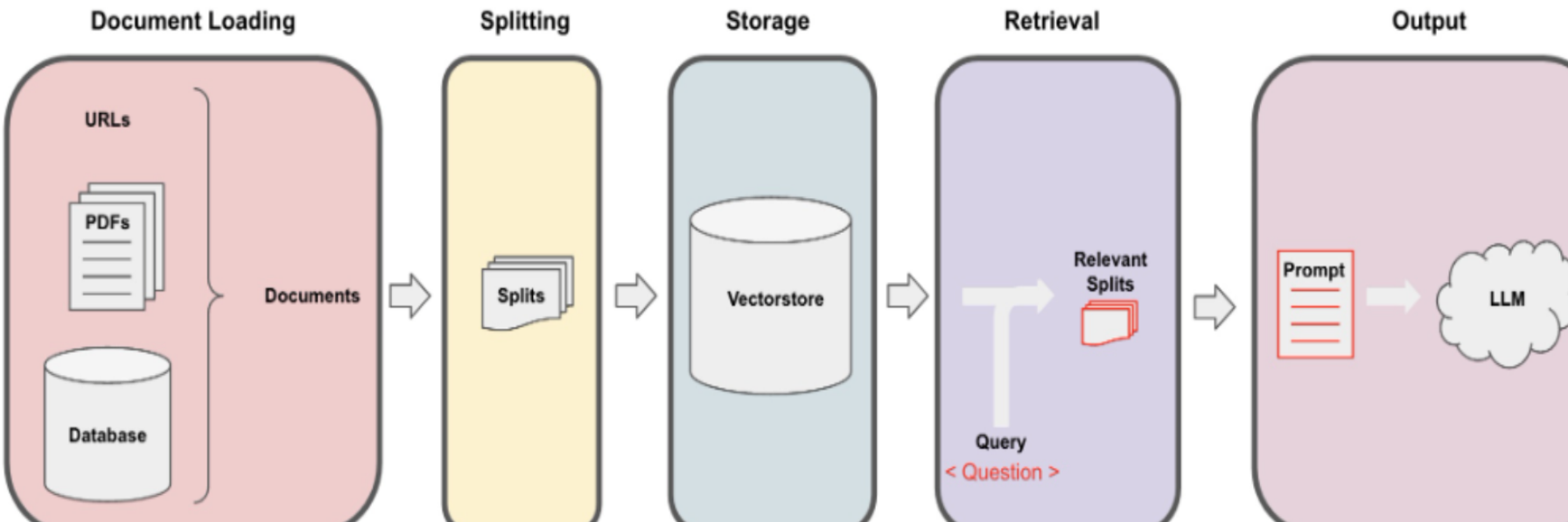
- Prompting
- Vector Databases
- Fine-Tuning
- Evaluation
- Agents

Prompting

- Prompt Engineering
- Chain of thought Prompting

Vector Databases

- Idea incorporate domain-scpefific knowledge not included during training.
- Rely on a Vector Database embed queries, retrieve relevant documents, append them into the prompt [Lewis et al., 2021].
- https://www.infoworld.com/article/3709912/vector-databases-in-llms-and-search.html
- https://learn.deeplearning.ai/vector-databases-embeddings-applications/lesson/1/introduction
- https://stackoverflow.blog/2023/10/09/from-prototype-to-production-vector-databases-in-generative-ai-ap



Instruction Fine-Tuning

- Paid Fine-Tuning (GPT-4??)
- Alpaca, Vicuna, Llama, Llama2
- https://blog.gopenai.com/paper-review-qlora-efficient-finetuning-of-quantized-llms-a3c857cd0cca

Parameter Efficient Fine Tuning

- Lora, QLora
- https://blog.gopenai.com/paper-review-qlora-efficient-finetuning-of-quantized-llms-a3c857cd0cca

LLMBench and LLm Arena

- Bla

LangChain and Agents

- Bla

Questions?

Thanks for your Attention!

References I


Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2021).
Retrieval-augmented generation for knowledge-intensive nlp tasks.