Natural Language Processing
Large Language Models Usage and Evaluation Patterns

Felipe Bravo-Marquez

November 22, 2023

Introduction

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

Source:

`https://www.masayume.it/img/masayume/Large-Language-Models.jpg`

Recap: What is an LLM

- An autoregressive language model trained with a Transformer neural network on a large corpus (hundreds of bullions of tokens) and a large parameter space (billions) to predict the next word.
- It is usually later aligned to work as a user assistant using techniques such as Reinforcement Learning From Human Feedback [**?**] or supervised fine-tuning.
- Some are private (access via API or web browser): Google Bard, ChatGPT, etc.
- Others are open (model's weights can be downloaded): Llama, LLama2, Falcon, etc.

Talk Overview

Usage Patterns

1. Fixed-knowledge Assistant
2. knowledge-augmented Assistant
3. Applications with LLMs in the middle
4. Agents

Evaluation Patterns

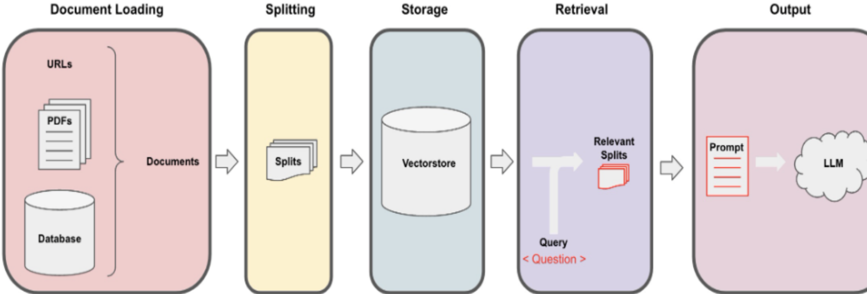
- MTBench
- LLM Arena

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 [**?**].
- `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

Datasets for Instruction Fine-Tuning

- Standford Alpaca Dataset (Vicuna)
- ShareGPT (Alpaca)
- Dolly-15K
- Orca Dataset

Parameter Efficient Fine Tuning

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

Token-Incrementation

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

LLMBench and LLm Arena

- MT-bench (categories)
- HuggingFace Open LLM Leaderboard
- LLM Arena

LangChain and Agents

- Bla

Questions?

Thanks for your Attention!

References I