Introduction to LLMs, LangChain, and LangGraph

Kevin Toh

June 7, 2024

Overview

Large Language Models

Definitions and Terminologies

- **Large Language Model**: A neural network utilizing the transformer architecture.
- **Word Embeddings**: Represent words as vectors in a multi-dimensional space.
- **Vectorization**: Conversion of sentences into numerical representations.
- **Embeddings**: Real-valued vectors encoding word meaning.
- **Inference Engine**: Hosted language model using specialized hardware.

Examples of LLMs

- GPT-3.5, GPT-4o by OpenAI
- Llama3 8b/13b/70b by MetaAI
- Claude 3.5 Sonnet by Anthrophic
- Gemma 2 9b/27b by Google

Inference Example

```
from langchain_groq import ChatGroq
mixtral8x7b = ChatGroq(model="mixtral-8x7b-32768")
```

Understanding LLM Features

Parameter Size

- Determines model complexity and performance.
- Examples: GPT-3.5 (175B), Llama 3 (8B, 13B, 70B)

Context Window

- Maximum tokens considered by the model.
- Examples: GPT-3.5 (2048 tokens), Llama 3 (8192 tokens)

LLM Settings

Temperature

- Controls randomness of model's output.
- Low temperature for factual QA, high for creative tasks.

Top P

- Nucleus sampling to adjust model determinism.
- Low top P for exact answers, high for diverse responses.

LangChain: LLM Application Framework

Definition

Framework for developing LLM applications.

Example

$\texttt{essay}_{generation}prompt_{template} = ChatPromptTemplate.from_{messages}([(\text{"}system\text{"}, GENERATOR_{PROMPT}), (\text{"}human\text{"}, \text{"}question\text{"}$

LangGraph: Multi-Agent Framework

Definition

Framework for building multi-agent LLM applications.

Benefits

Improves performance by structuring agents as a graph.