LLM Apps

The new OpenAI API as alternative to Langchain

Intro

- One of the reasons LangChain became popular initially is that it was seen as a simpler and more plural alternative to OpenAI's API.

- This trend might change due to three factors:
  - Until now, OpenAI's ChatGPT model is by far the most used, so at this moment the value of LangChain as a multi-model framework virtually makes no sense.

  - With the adoption of the new LCEL language, LangChain is becoming more complex.

  - At the DevDay in November 2023, OpenAI has launched a simpler and more versatile version of its API.

Caution

- Just because the OpenAI API can do something does not mean it is the best way to do it:
  - Using the new functionalities of OpenAI can be very expensive.
  - Using the new functionalities of OpenAI implies having a lesser degree of knowledge and control over the settings, given that OpenAI's API is opaque in many senses.

Analysis if the New OpenAI in three lessons

- In this lesson, we will present the main changes that OpenAI introduces in its API since the DevDay of November 2023, as well as their impact for developers of LLM applications.

- In the lesson chapter, we will do a brief review of OpenAI's API.

- And we will dedicate a third lesson to analyze in more detail the most interesting functionality of the OpenAI API: the OpenAI functions.

Main changes introduced in the OpenAI API

- Context Window de 128.000 tokens (modelo GPT-4 Turbo)
- Multiple function calling.
- JSON mode.
- Reproducible outputs (seed parameter).
- Assistants.
- RAG Assistant.
- Vision.
- Text to Speech.

GPT-4 Turbo with 128k context

- RAG is not necessary for simple cases.
- You can now upload one full book (up to 300 pages) and make questions about it.
  - But be careful: this is a very expensive functionality.
- GPT-3.5 Turbo has 16k context now.

Multiple function calling

- Before it, you had to call one function at a time. This simplifies the process and saves time.
- Now you do not pass in functions, but tools. The type of the tool is function.

JSON mode

- Before it, chatGPT responded with text (strings) and we had to use an Output Parser to convert it to JSON.

Assistants

- Agent-like experience.
- Can call models and tools to perform tasks.
- Code-interpreter, retrieval, function calling.
- Persistent and infinitely long conversation threads. ChatGPT keeps the conversation memory for us.
- Use cases:
  - Coding assistant
  - Vacation planner
  - Voice-controlled DJ
  - Visual canvas

- See Assistants at work in the OpenAI playground.

Seed parameter: reproducible outputs

- The seed parameter ensures that the output (response) of the model to the same input is going to be the same (or very similar). This does not work 100% of times, but most of the time. It is good for testing.

Multi-modality: Vision

- Ask about the content of an image
- Text-to-image with Dall-E

Multi-modality: Audio

- Before we had just speech-to-text with Whisper.
- Now we also have text-to-speech.