Llama 3 Finetune Finder

Motivation

Small ML models can compete and even outperform larger and powerful ones when fine-tuned for specific tasks. This implies higher throughput, lower latency and lower costs. Additionally, it could enable deployment of ML models on edge devices.

How is it different from Chatbots Arena?

Unlike Chatbot Arena which compares LLMs in general benchmarks, Llama Finder’s goal is to help you pick the right Llama finetuned for your specific use-case.

Description

Llama 3 finetune finder takes as input prompt your specific needs. It then queries **HuggingFace** looking for specific keywords. It summarizes the available models with Llama 3 using **Together AI** API, and verifies if they are readily available in **Featherless.ai**.

It also uses **Brave API** to crawl for web resources. In the future, image search could be added and analyzed with Llama 3.2 vision capabilities. Alternatively, I intend to use **DSPy** to do multihop reasoning. At the moment "Llama Finder" is only able to summarize resources but it's not capable of doing a higher level comparison between the different Llama candidates. It also lists candidates based on alternative open source models.

The comparison table provides a summary of the model, whether it's finetuned to follow instructions, parameters and specific datasets used to fine-tune the model. It also listed resources crawled from the web at the bottom.