Free to play

UN Trade and Development's experience with developing its own open-source RAG and fine-tuning LLM application

Daniel Hopp

UN Trade and Development

20.06.2024

1. LLMs
2. Uses
3. UNCTAD's application

1. LLMs
2. Uses
3. UNCTAD's application

What is an LLM?

- Large language models (LLMs) are artificial neural networks (ANNs) trained on huge amounts of natural language text scraped from the internet or other sources.
- They are able to predict the next word or series of words based on prior context.
- The end results are convincing and reasonable responses to prompts and queries.
- This combination of a vast knowledge base and flexible natural language input and output have made them very popular

What is RAG?

- Base LLMs' knowledge is limited to the data they were trained on, in both scope and time
- Retraining an LLM from scratch every week to make it aware of the latest information is prohibitively expensive
- Retrieval augmented generation (RAG) is an approach to make the LLM aware of new information without needing to retrain the model

What is fine-tuning?

- With RAG, the LLM is only aware of the retrieved chunks passed to the context window alongside the query
- Fine-tuning changes the underlying LLM by freezing the majority of its weights and only retraining a subsample
- This enables knowledge injection and task augmentation with much lower computational and time requirements

1. LLMs
2. Uses
3. UNCTAD's application

How is this useful to the work of NSOs and international organizations?

- Select a few research papers on a topic you're looking into and ask natural language questions about them, such as, "Explain [technical concept presented in the papers] for a non-technical person"
- Ask the LLM for help with coding, including for obscure libraries or programming functions, by passing alongside the documentation for that library or programming language
- Ask for summaries, key takeaways, and syntheses of multiple reports from different agencies

(cont.)

- Fine-tune a model on an organization's publications and get answers on new topics from that organization's perspective
- Incorporate LLM-generated responses into back-end workflows without the chat interface, e.g., pre-populating mission report fields
- Use RAG or fine-tune on your official statistics database API documentation to allow people to ask natural language questions about its contents, generate the correct query, retrieve the data, then comment on the retrieved data

1 LLMs

2 Uses

3 UNCTAD's application

Constituent libraries

- Document processing: nlp_pipeline
- LLM and RAG system: local_rag_llm
- LLM fine-tuning: local_llm_finetune
- Front end: streamlit_rag
- All libraries are Dockerized and the entire application can be run on any Nvidia GPU-enabled PC or Apple Silicon Mac, as well as on CPU at slower generation speeds

Why don't we just use ChatGPT?

- Cost itself, plus institutional overhead of getting new cost streams approved
- Data privacy and security
- Customization and control
- Fostering of institutional knowledge and capacities