**SUMARY**

**Model**

Finetuning Qwen3-1.7B with LoRA PEFT. Using the following dataset:

Link: <https://huggingface.co/datasets/axiong/pmc_llama_instructions>

Size: 513,999

Method: Split the training dataset with the possibility of randomly taking 10% samples from the entire dataset for each epoch.

* We have: 51399 data for each epoch.
* Solve the problem of VRAM shortage when we encounter hardware limitations during training.

Limitations: Due to the model training method not being truly optimal, overfitting still occurs.

**Application**

**Architecture:** LLMAPI -> BackEnd -> FrontEnd

Dockerize: LLMAPI -> BackEnd -> FrontEnd

Database(MariaDB)

**Technology:**

* **LLMAPI :** FastAPI
* **BackEnd:** FastAPI, MariaDB
* **FrontEnd:** NextJS, Zuststand, Tailwindcss

**Features: Login – Register, Chatbox**

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a chatbot

AI-generated content may be incorrect.A screenshot of a chatbot

AI-generated content may be incorrect.

A screenshot of a chatbot

AI-generated content may be incorrect.