Stanford Course Fall'23: Creating Production Ready LLMs

arxiv-search

Sharat Sripada

Introduction - *arxiv search*

- Leverage pre-trained Large Language Models (LLM) for Information Retrieval (IR) tasks
- Using techniques like *fine-tuning* - Low Ranking Adaptation/LoRA, Efficient fine tuning of quantized LLMs - qLoRA, Parameter efficient Fine Tuning/PEFT etc., *prompt engineering* and *Retrieval Augmented Generation (RAG)* fit custom data and build a personalized Question-Answer (QA)/Hybrid-search engine

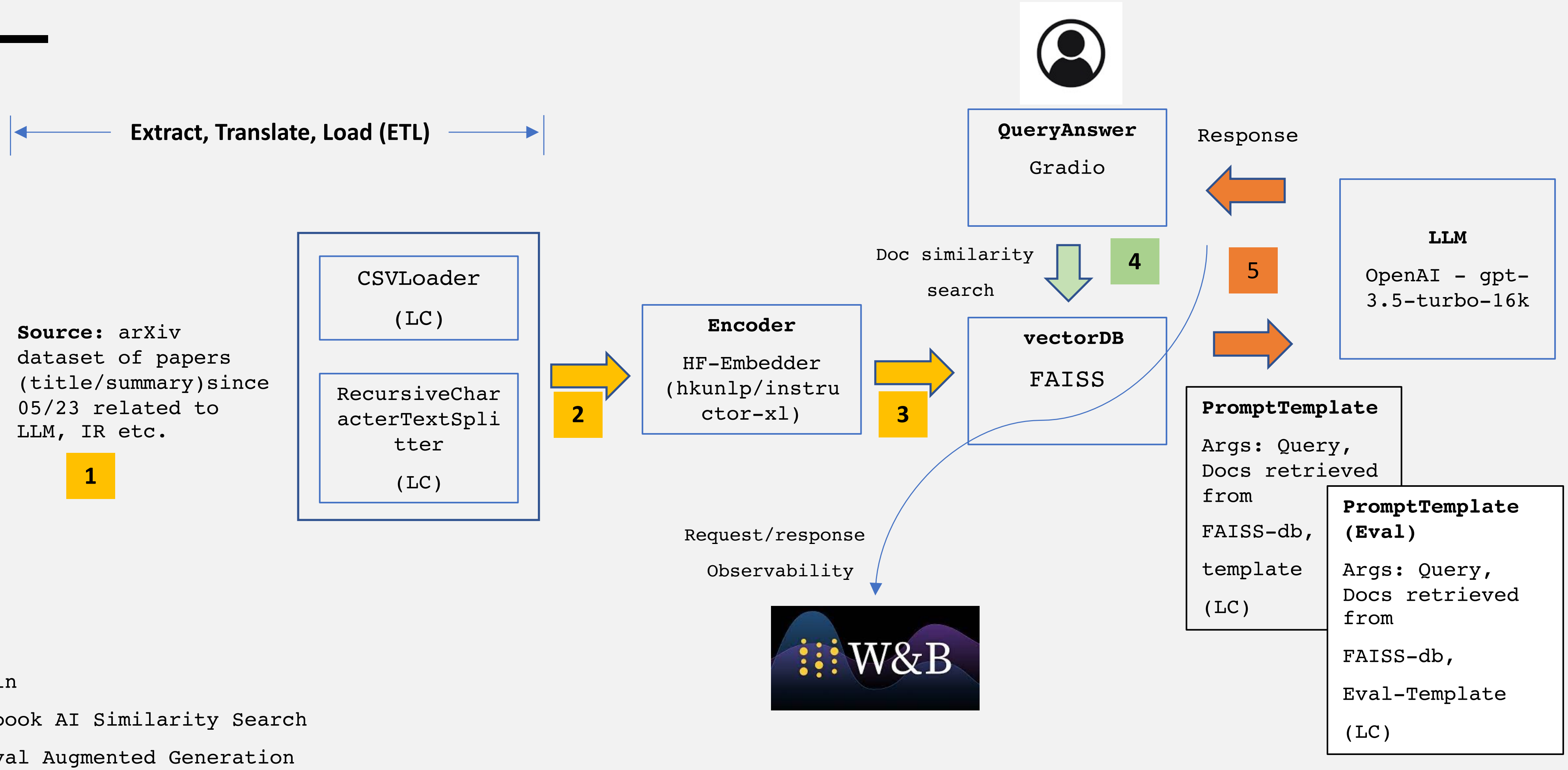
Problem Statement

The space of LLMs/IR is evolving at a rapid pace and research papers is one way to stay up-to-date, learn about problems being solved and innovate. Using papers from arxiv.org for past six-months and a pre-trained LLM (OpenAI - ChatGPT-3.5)* the objective is to build a QA model and seek answers to questions like:

- what is recent research in RAG
    - Explain PRAG
        - What is the title of the paper for PRAG
    - Explain KGR
        - What is the title of the paper for KGR
- are there any references to local language models
- what is local different from PEFT? (ask generic questions)

**trained on data till Jan'23*

Information Retrieval using RAG



LC – LangChain

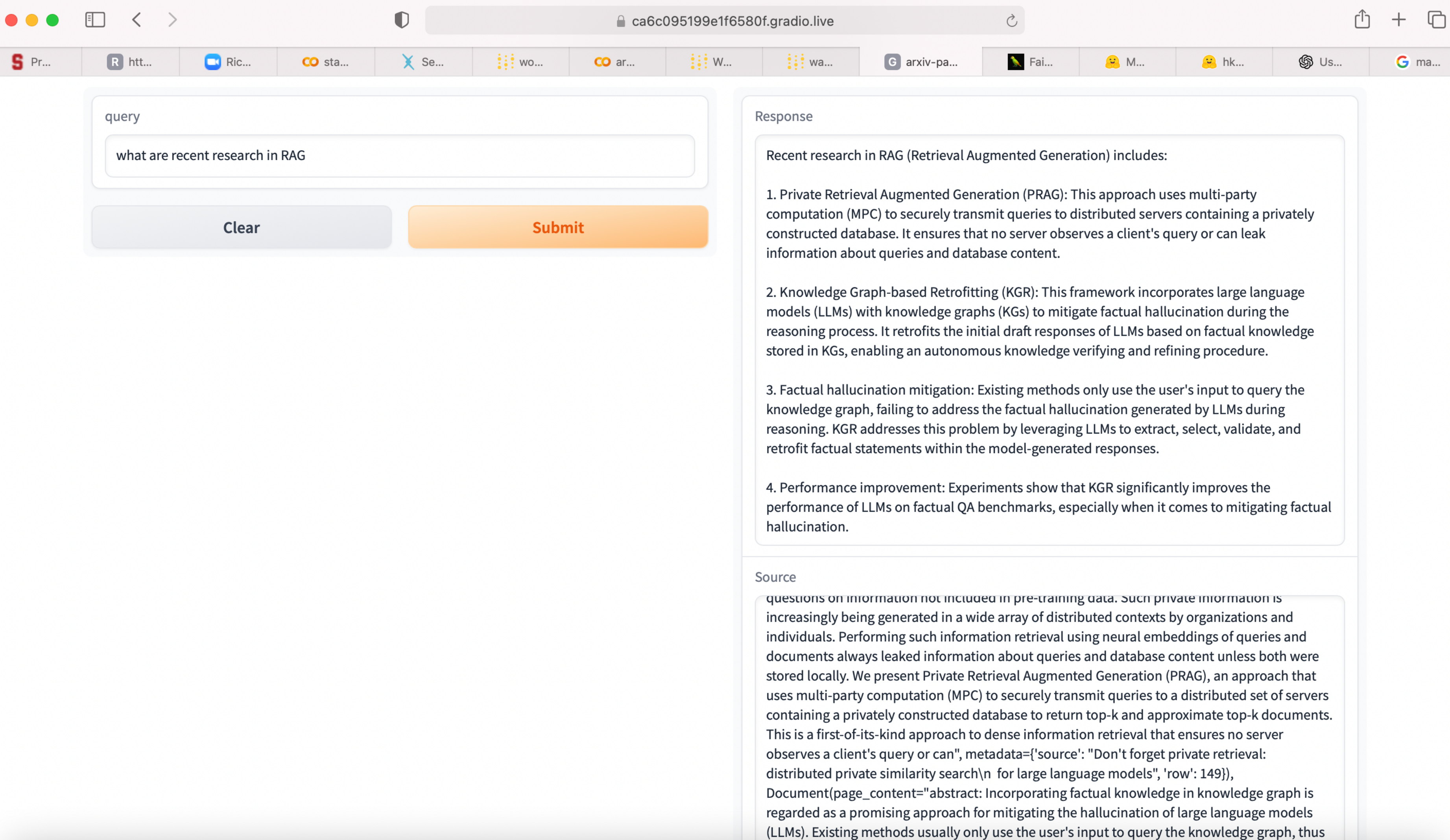
FAISS – Facebook AI Similarity Search

RAG – Retrieval Augmented Generation

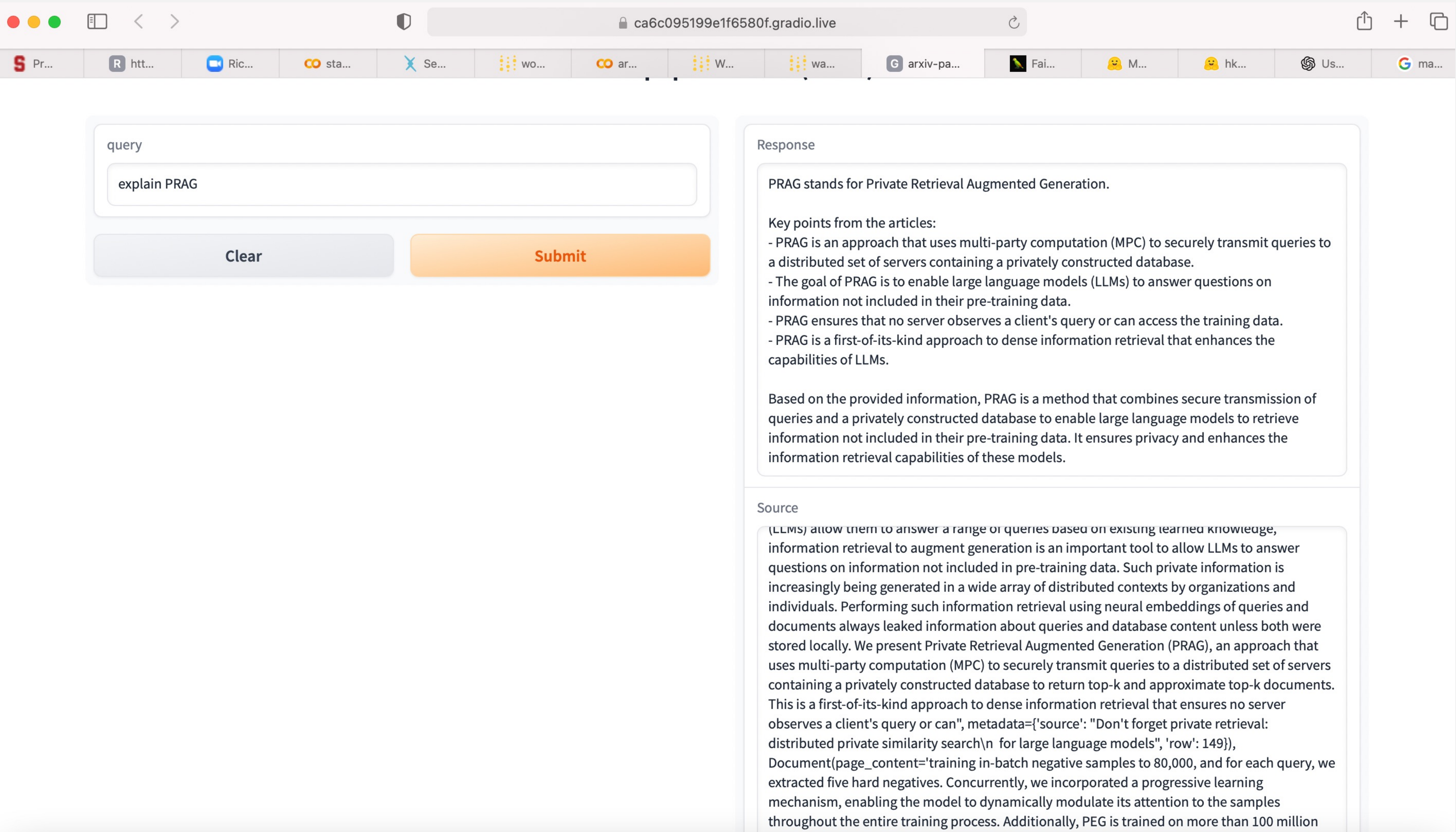
W&B - Weights and Biases

Demo

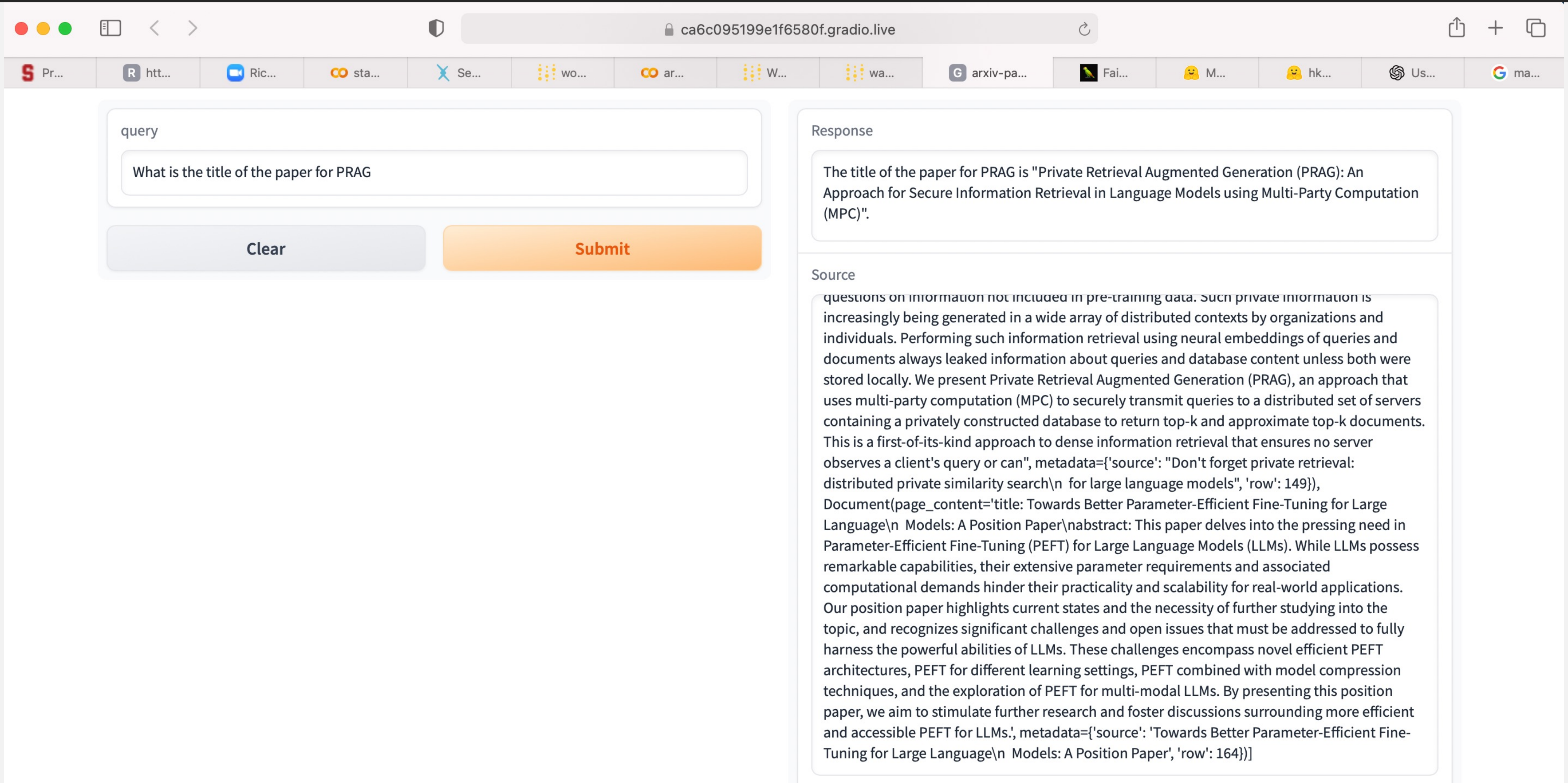
arxiv-search - *"What's research community thinking about RAG"*



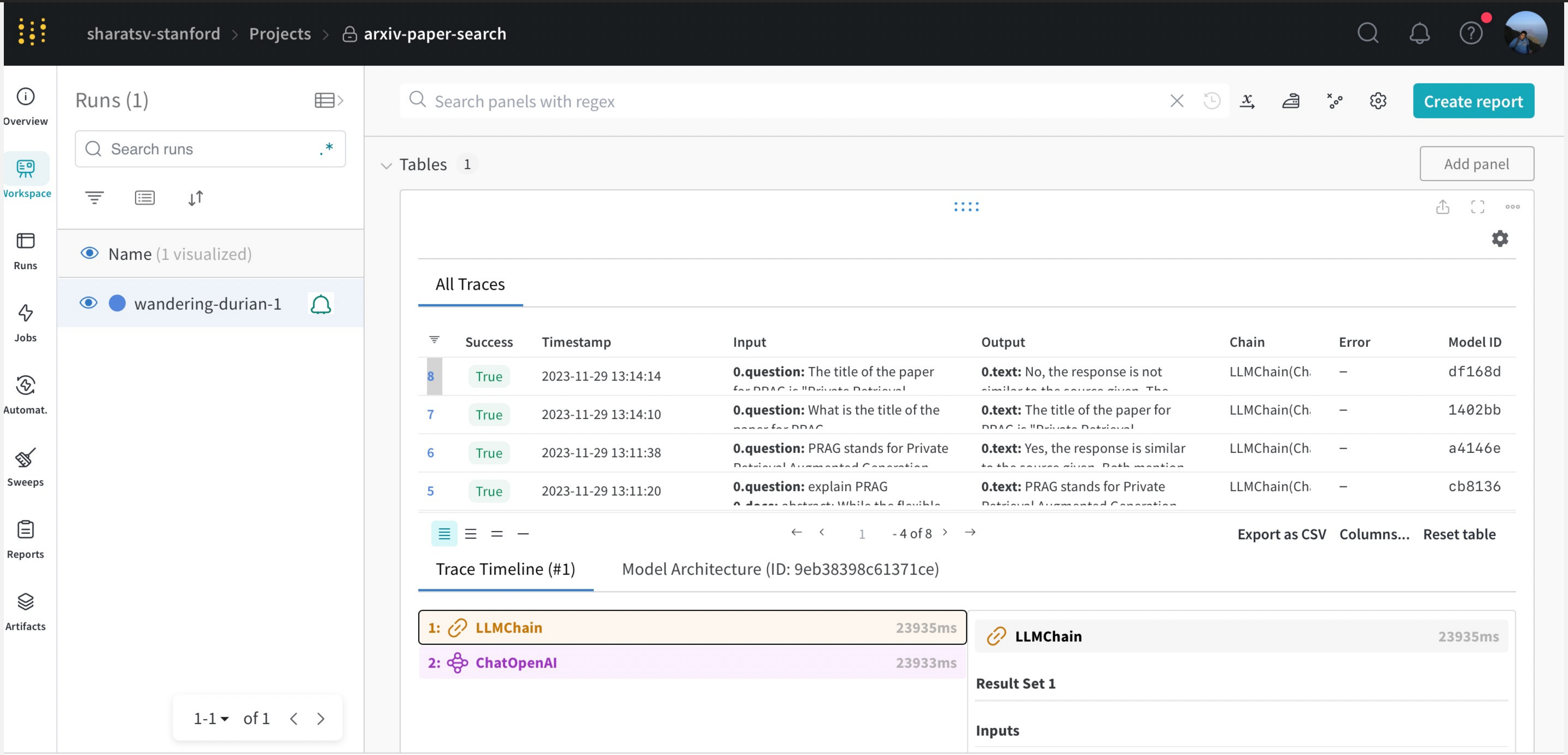
arxiv-search - *“PRAG, ok. Ask more about PRAG before I dive in”*



arxiv-search - *“I like PRAG..give me the reference title/paper”*



“Monitor infrastructure/GPU, query-response time..”



Future work

- Host LLM (example: LLaMA-cpp) locally on a small form-factor like Raspberry-Pi/Rock-Pi
- Experiment with Fine-tuning using LoRA/qLoRA versus RAG
- Improve query-response time (to < 1sec)
- Explore fetching summary of PDF/paper to QA model like:
    - what are recent research in RAG? (works)
        - what is the title of the paper? (works)
            - *fetch the corresponding paper and summarize it (TBD)*
- Feedback from Cohort