OpenLLM Day Agenda

May 30th, 2024

Registration: 08:00 AM to 09:00 AM

LLM Introduction, Benchmark, & Evaluations: 09:00 AM to 11:00 AM

- Introduction to Language Models (30 mins; by Pranav)
    - What are they?
    - An overview of the evolution and tech details
        - From RNNs to Transformers: The evolution of language models
        - Introduction to transformer architecture, attention mechanisms, KV cache.
        - How does text generation work during inference?
- Evaluating NL-SQL (90 mins; by Jeyaraj)
    - Different use cases LLMs - NL-to-NL, NL-to-Code, NL-to-SQL
    - Why we chose NL-to-SQL
    - Techniques and Benchmarks: A comprehensive report.
    - Discussion on the benchmarking process, focusing on the Bird dataset including metrics such as accuracy, latency, model size, and cost.
    - Insights into overall learning.

Break: 11:00 AM to 11:30 AM

Retrieval-Augmented Generation (RAG): 11:30 AM to 01:30 PM - by Jeyaraj

- Understanding RAG
    - An exploration of what RAG is and its applications.
    - Discussion on the effectiveness of RAG: scenarios where it excels and its limitations.
- Exploring RAG Techniques
    - Dive into various strategies such as HyDE, Re-Ranking, and Self-RAG.
- Interactive RAG Workshop
    - Participants will be divided into groups to engage in hands-on practice with different RAG techniques using tools like Google Colab, sample datasets, and access to a free MongoDB vector cluster.
    - Review of the RAG evaluation report from Petavue.

Lunch Break: 01:30 PM to 02:30 PM

Fine-Tuning Strategies: 02:30 PM to 05:00 PM - by Pranav Reddy

- Basics and Essentials of fine-tuning
  - Introduction to LLM fine-tuning
  - RAG vs Fine-tuning. What to use, When, Why fine-tuning is necessary.
  - Why and when Overview of various techniques including LoRA and PEFT.
- Hands-on fine-tuning session
  - LoRA fine-tuning and hyperparameters explained
  - Live demo of fine-tuning Mistral 7B for SQL generation using PEFT, followed by model evaluation
- Data collection & preparation strategies
  - Tips on how to collect and prepare data for fine-tuning, synthetic data generation, etc

Networking: 05:00 PM to 05:30 PM

An opportunity for attendees to network, discuss the day's learnings, and explore potential collaborations.