Open Models Track

A Summary

Generated by Gemini 1.5 Pro

September 2, 2024

Setting the Stage

- The Open Models track at AI Engineer World's Fair 2024 highlighted the rapid progress and exciting potential of open-source Large Language Models (LLMs).
- Researchers and developers from leading organizations showcased advancements in training, applications, and fostering a collaborative open-source AI community.
- This presentation summarizes the key takeaways and future directions discussed during this insightful event.

Meta AI: Democratizing Access

- Focus on making advanced LLMs widely accessible.
- In-depth explanation of the three LLM training stages:
  - Pre-training
  - Instruction Tuning
  - Learning from Human Feedback
- Introduction of Code Llama: open-source, code-generating LLM available on multiple platforms.

Txt: Taming LLMs with Structure

- Addressed the challenge of consistent structured output from LLMs.
- Introduced Outlines: Python library for structure generation (JSON, regular expressions, etc.).
- Benefits:
  - Reliable structured data generation.
  - Minimal inference overhead (sometimes even faster).
  - Improved sample efficiency.
  - Enhanced performance of open-source models.

Cohere: Developer-First RAG

- Focused on building developer-friendly tools for Retrieval Augmented Generation (RAG).
- Discussed challenges in RAG system development.
- Showcased Command R model family optimized for RAG, highlighting performance and cost-effectiveness.
- Open-sourced their UI toolkit to empower developers in building RAG applications.

UNSloth: Debugging and Optimization

- Highlighted the importance of bug identification and fixing in open-source LLMs (focus on Llama 3).
- Detailed common fine-tuning pitfalls and the need for careful configuration and pre-processing.
- Presented UNSloth's open-source platform:
  - Pre-built bug fixes.
  - Automatic chat template generation.
  - Support for long-context fine-tuning.

Liquid AI: Model Merging

- Provided an overview of fine-tuning techniques and strategic application for LLM performance.
- Discussed when to choose fine-tuning over prompt engineering.
- Explained various fine-tuning methods (full, LoRA, 4-bit quantization).
- Introduced Model Merging as a powerful technique for combining strengths of different fine-tuned models.
- Showcased merging techniques (Slurp, DeLoRA, Pass-through, Mixture of Experts) and real-world examples.

The Future of Open LLMs

- The Open Models track offered an inspiring look at the future of LLM development.
- It highlighted a collaborative ecosystem focused on groundbreaking research, accessibility, and responsible AI deployment.
- Open-source initiatives play a crucial role in shaping the future of LLMs, making them more powerful and beneficial for all.