Contribution ID: **3** Type: **not specified**

Helmholtz Blablador: An Inference Server for Scientific Large Language Models

Recent advances in large language models (LLMs) like chatGPT have demonstrated their potential for generating human-like text and reasoning about topics with natural language. However, applying these advanced LLMs requires significant compute resources and expertise that are out of reach for most academic researchers. To make scientific LLMs more accessible, we have developed Helmholtz Blablador, an open-source inference server optimized for serving predictions from customized scientific LLMs.

Blablador provides the serving infrastructure to make models accessible via a simple API without managing servers, firewalls, authentication or infrastructure. Researchers can add their pretrained LLMs to the central hub. Other scientists can then query the collective model catalog via web or using the popular OpenAI api to add LLM functionality in other tools, like programming IDEs.

This enables a collaborative ecosystem for scientific LLMs:

- Researchers train models using datasets and GPUs from their own lab. No need to set up production servers. They can even provide their models with inference happening on cpus, with the use of tools like llama.cpp.
- Models are contributed to the Blablador hub through a web UI or API call. Blablador handles loading models and publishing models for general use.
- Added models become available for querying by other researchers. A model catalog displays available LLMs from different labs and research areas.

Besides that, one can train, quantize, fine-tune and evaluate LLMs directly with Blablador.

The inference server is available at http://helmholtz-blablador.fz-juelich.de

Primary author: STRUBE, Alexandre (Heymholtz AI - Juelich Supercomputing Centre)

Presenter: STRUBE, Alexandre (Heymholtz AI - Juelich Supercomputing Centre)

Track Classification: Other