Contribution ID: **8** Type: **On-going projects**

Deploying Open Source LLMs for on site usage

LLMs have become a tool used by many researchers in a wide variety of tasks and several libraries are available to facilitate access to the most common LLMs. At the same time many workstations used by researchers don't have the capacity to run llms locally and at the same time researchers are hesitant to feed potentially sensitive data to models hosted on external webservices like Azure or OpenAI. We have set up a local llm deployment, based on kubernetes, FastAPI and llama.cpp. The deployment provides several models along with a self-service checkout for researchers to set up their own API keys. While the service, currently is not intended for high throughput inference, it can serve as a testing ground and can easily be extended.

Primary author: PFAU, Thomas (Aalto University)

Presenter: PFAU, Thomas (Aalto University)

Track Classification: Track 1