

Universität
Münster

Building UniGPT: A Customizable On-Premise LLM-Solution for Universities

Jonathan Radas

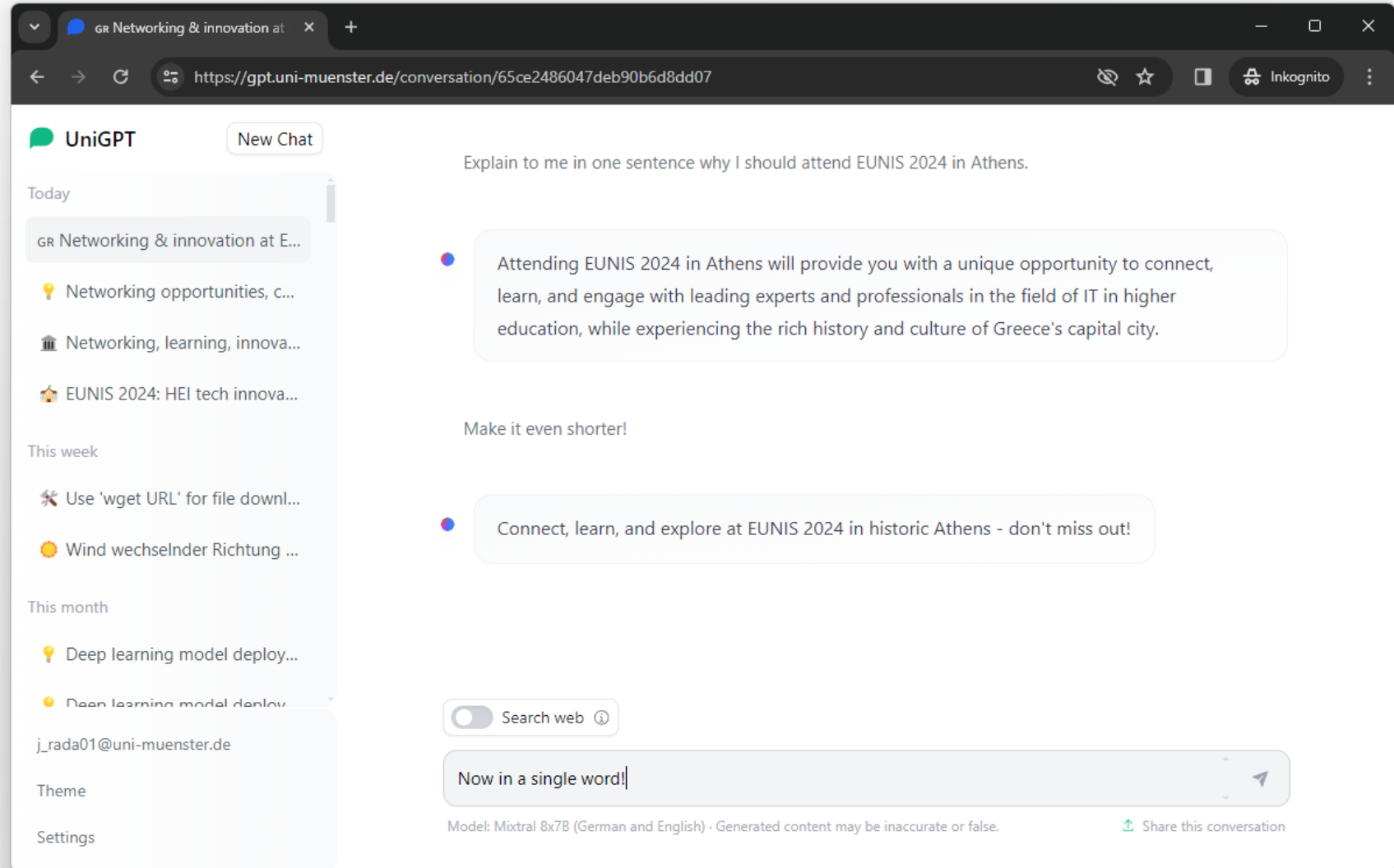
living.knowledge



Why deploy an on-premise large language model (LLM)?

- More control over the models
- Less dependency on OpenAI
- Privacy and copyright concerns with OpenAI
- Starting point for further research

Live Demo



The screenshot shows a web browser window with the UniGPT interface. The browser's address bar shows the URL `https://gpt.uni-muenster.de/conversation/65ce2486047deb90b6d8dd07`. The UniGPT interface includes a sidebar on the left with a "New Chat" button and a list of recent conversations. The main chat area shows a user prompt: "Explain to me in one sentence why I should attend EUNIS 2024 in Athens." The AI response is: "Attending EUNIS 2024 in Athens will provide you with a unique opportunity to connect, learn, and engage with leading experts and professionals in the field of IT in higher education, while experiencing the rich history and culture of Greece's capital city." A follow-up prompt asks to "Make it even shorter!" and the AI response is: "Connect, learn, and explore at EUNIS 2024 in historic Athens - don't miss out!". At the bottom, the user has typed "Now in a single word!". The interface also features a "Search web" toggle, a model identifier "Model: Mixtral 8x7B (German and English)", and a "Share this conversation" link.

gr Networking & innovation at ...

UniGPT New Chat

Today

- gr Networking & innovation at E...
- 💡 Networking opportunities, c...
- 🏛️ Networking, learning, innova...
- 🏆 EUNIS 2024: HEI tech innova...

This week

- 🔧 Use 'wget URL' for file downl...
- 🌞 Wind wechselnder Richtung ...

This month

- 💡 Deep learning model deploy...
- 🟡 Deep learning model deplov...

j_rada01@uni-muenster.de

Theme

Settings

Explain to me in one sentence why I should attend EUNIS 2024 in Athens.

- Attending EUNIS 2024 in Athens will provide you with a unique opportunity to connect, learn, and engage with leading experts and professionals in the field of IT in higher education, while experiencing the rich history and culture of Greece's capital city.

Make it even shorter!

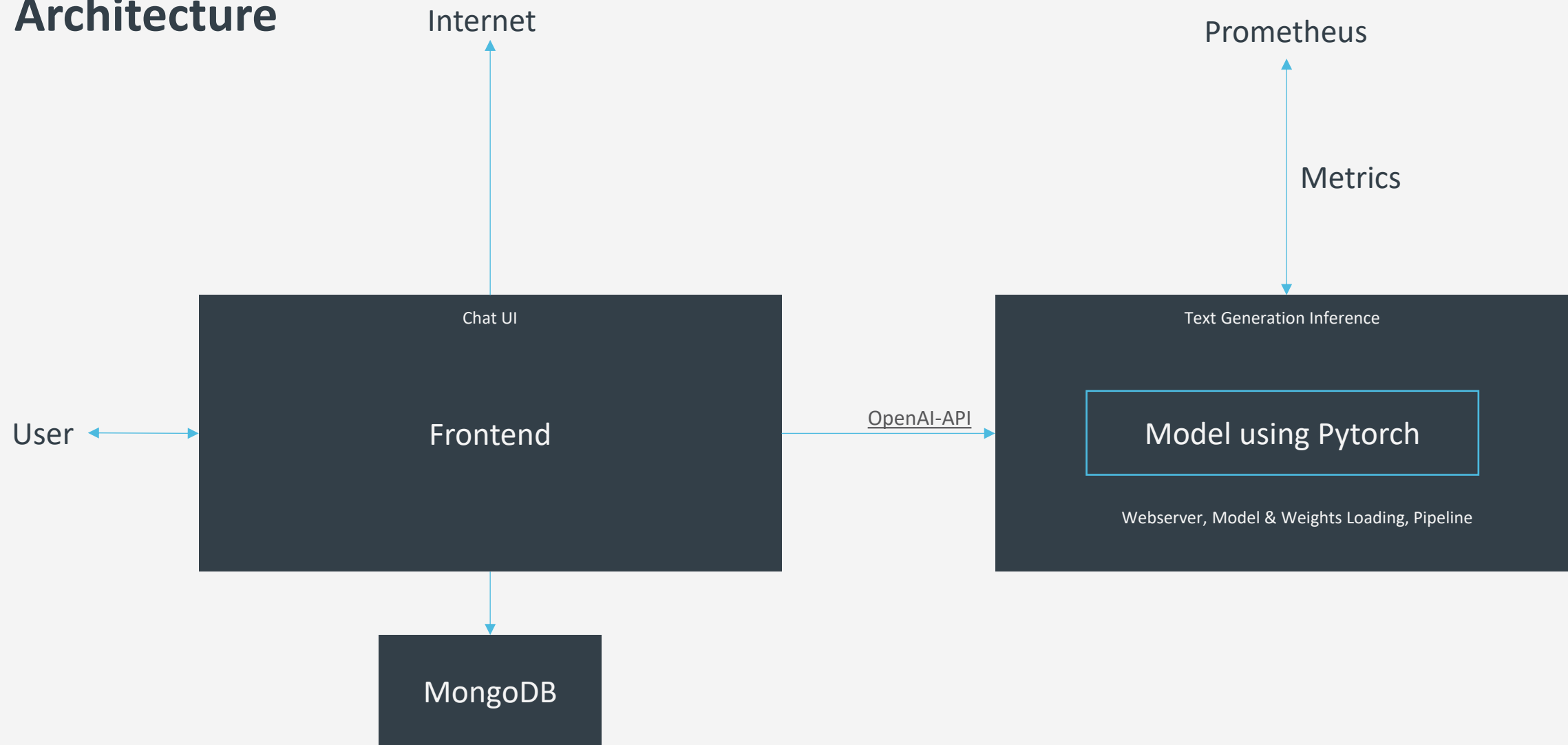
- Connect, learn, and explore at EUNIS 2024 in historic Athens - don't miss out!

Search web ⓘ

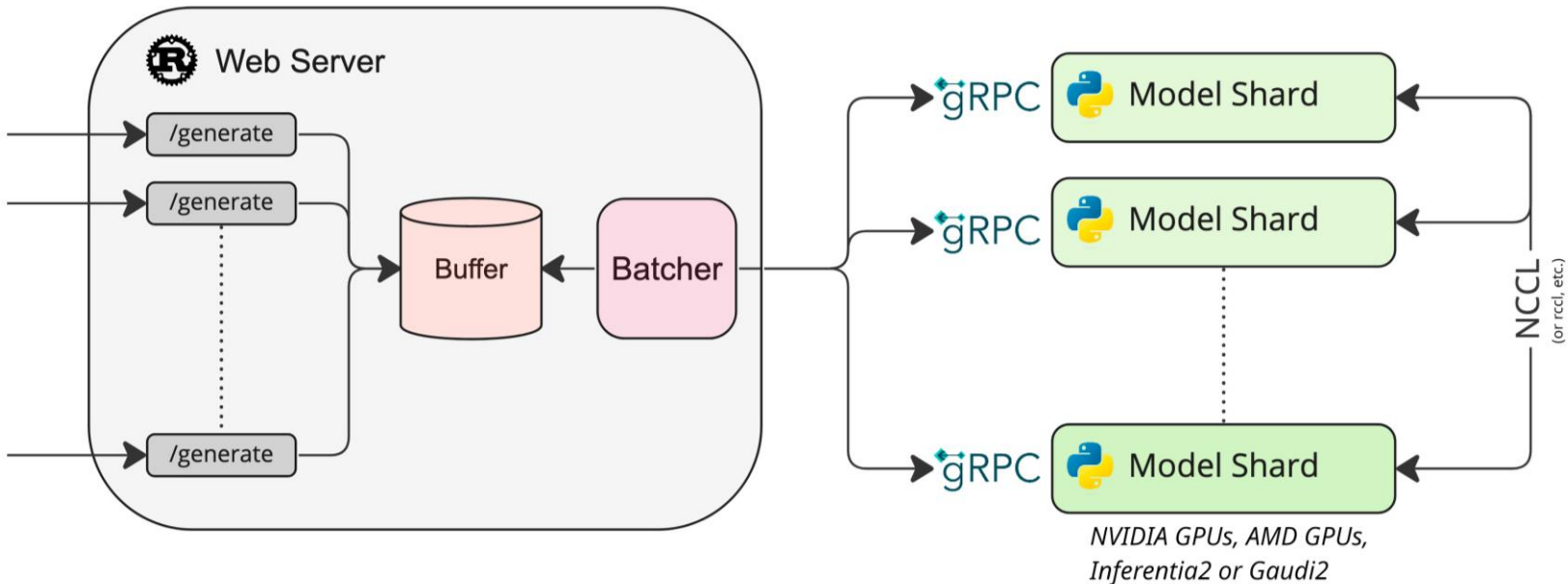
Now in a single word!

Model: Mixtral 8x7B (German and English) · Generated content may be inaccurate or false. [Share this conversation](#)

Architecture

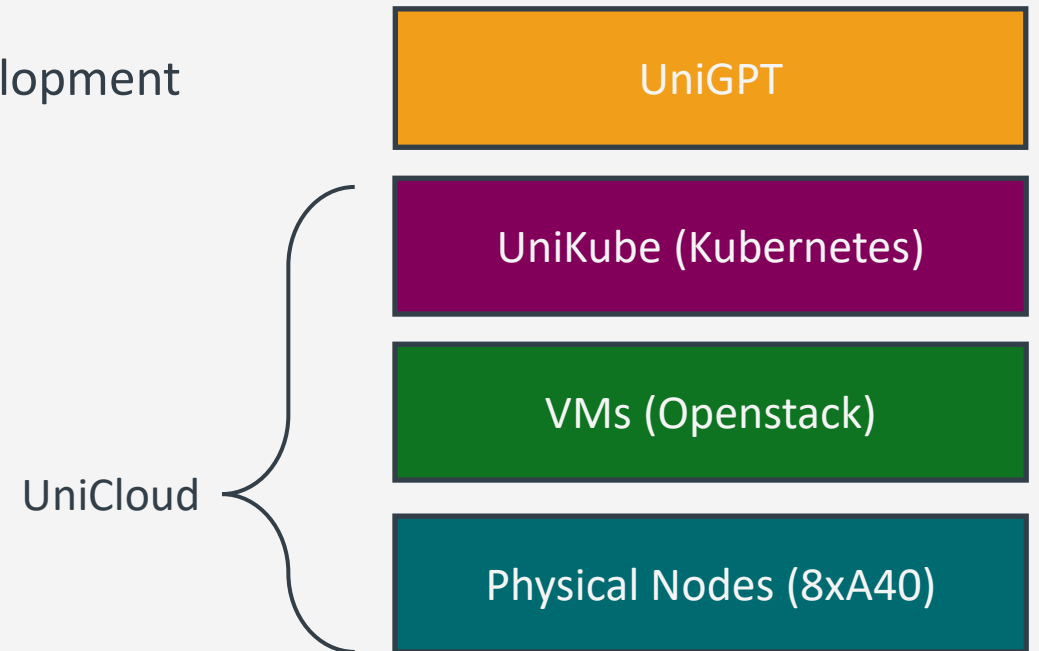


“The Backend” – Text Generation Inference (TGI)



Soft and Hardware Stack

- 8 Nvidia A40 with 48 GB of memory
- 2 GPUs for each model on production and 1 GPU for development
- Multiple GPUs mainly for availability not throughput
- Running on a Kubernetes cluster based on OpenStack

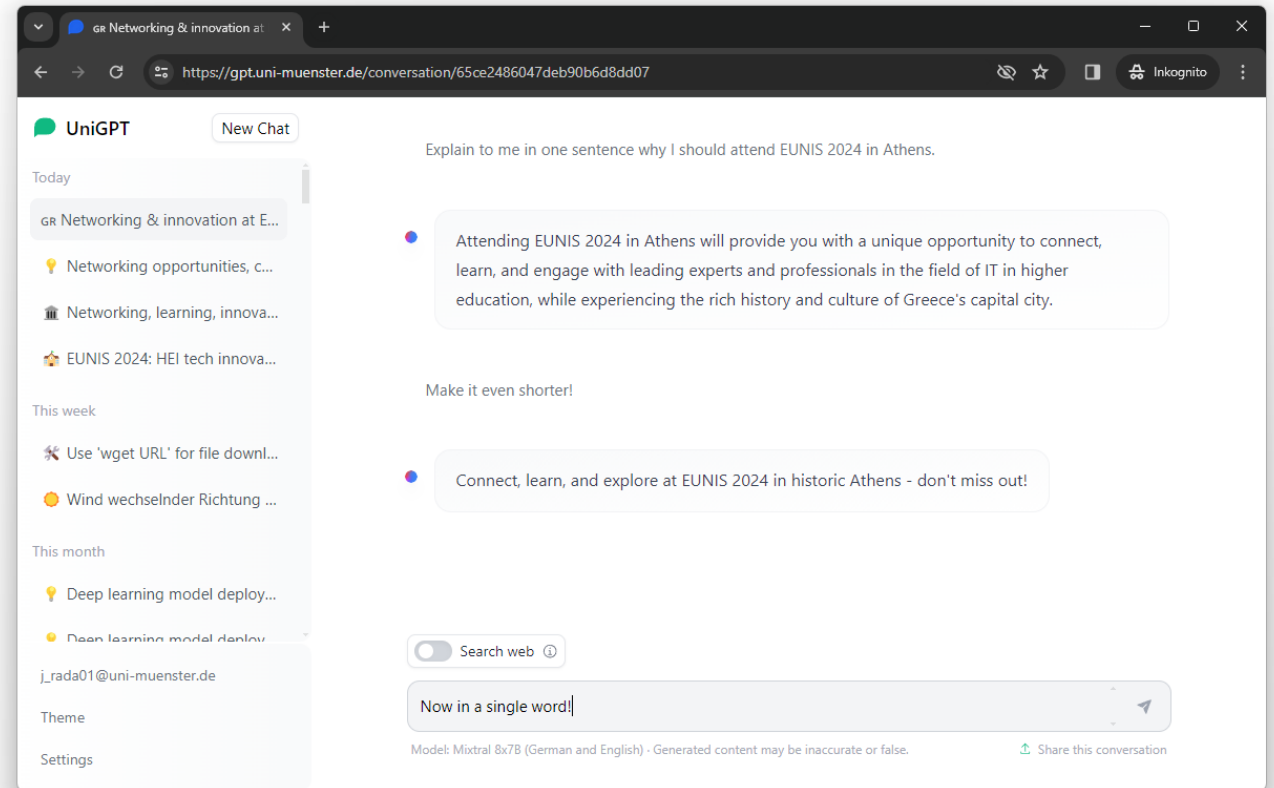


Models

Model	Creator	Size in Parameters	License	Deployed
Llama	Meta AI	7-80B (400B)	Own License	Yes
Mixtral	Mistral AI	8x7B	Apache	Yes
Falcon	TII	140B	Modified Apache	No
Gemma 2	Google	9-27B	Own License	No (but planned)
GPT-3/4/4o	OpenAI	Unknown	Proprietary	Yes (through API)

The Frontend

- ChatUI by Hugging Face
- text-generation-webui by oobabooga
- LobeChat by LobeHub
- LM Studio by Element Labs

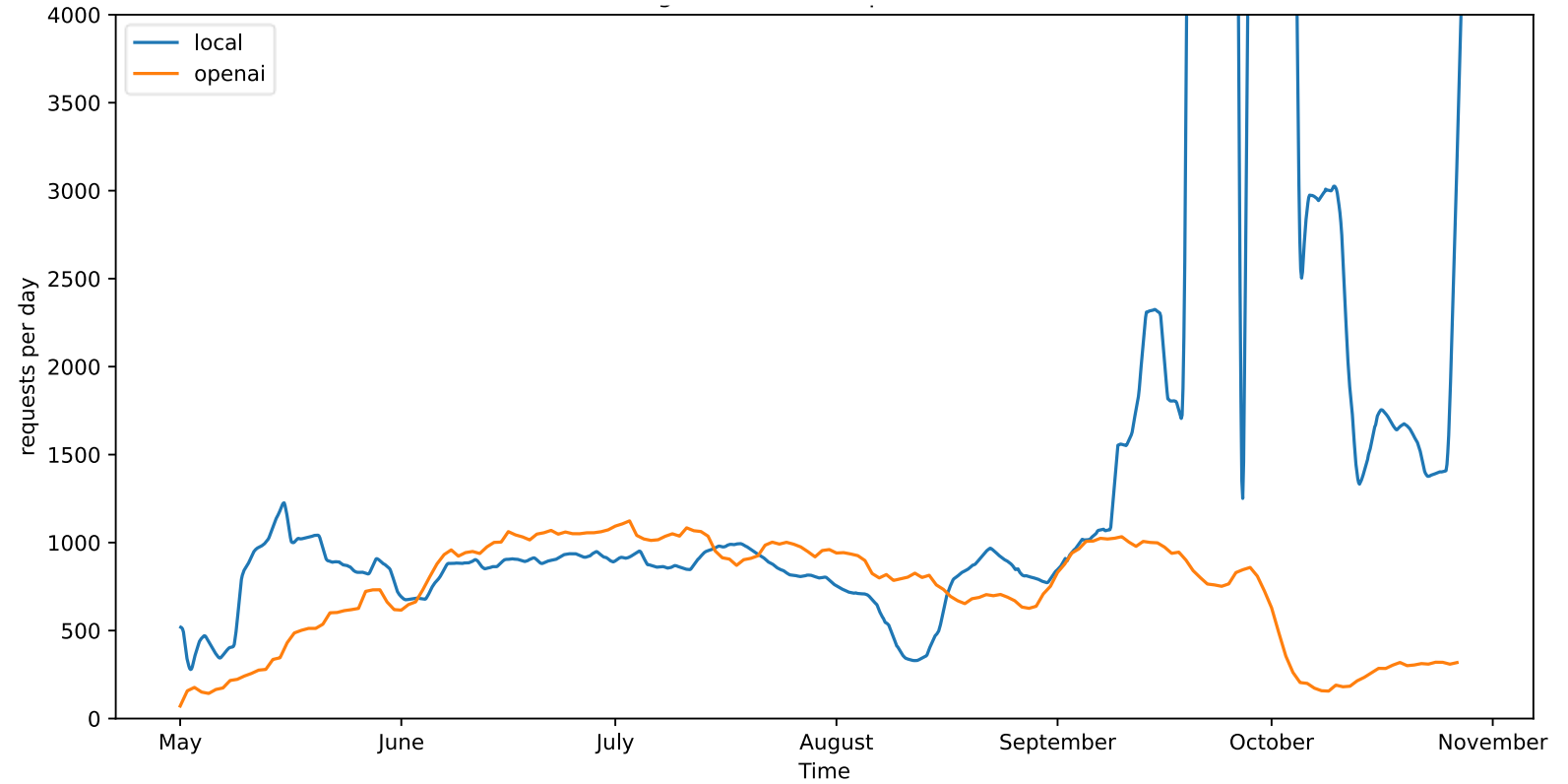


Organisational Challenges

- Maintaining and updating the models and infrastructure
 - New models arrive fast (very fast). Most models in their OpenLLM Leaderboard are less than 1 month old
 - New features arrive fast. Multimodality and function calling support arrived this year
- Terms of Use/Privacy
- People may criticize the model and the university because of toxic answers
 - Initial experiments show less safety in German than English

Usage of Local and OpenAI Models

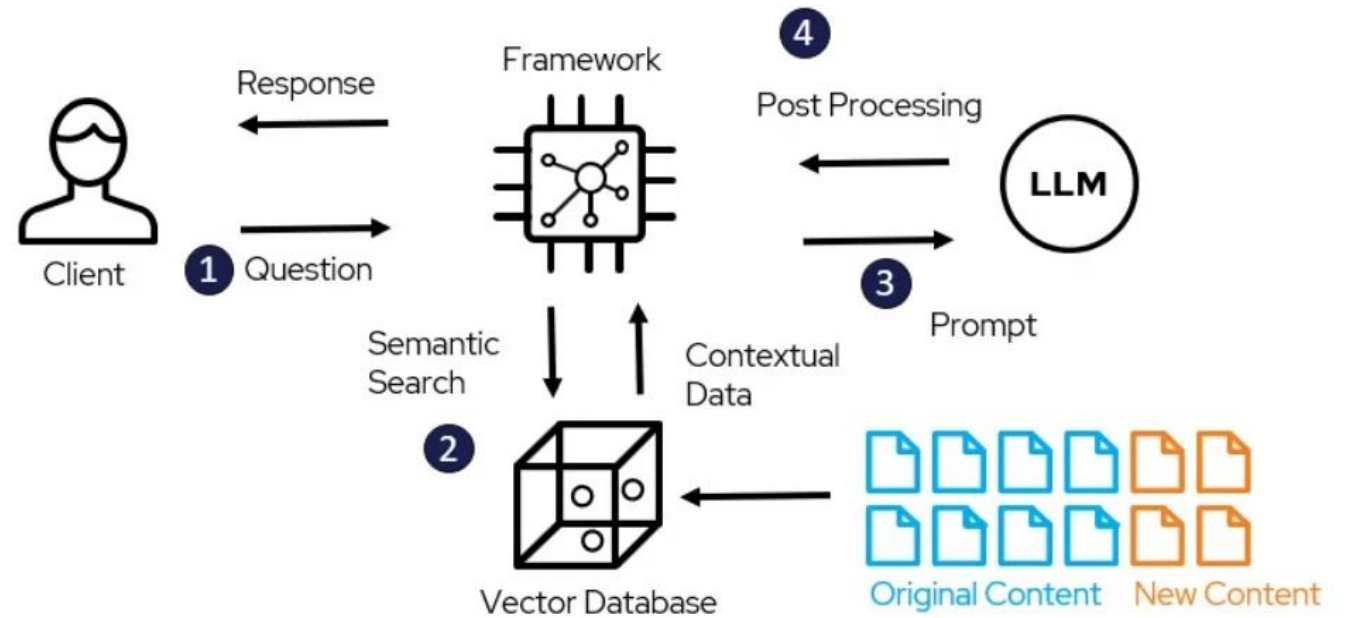
- 3000 Users in the first month
- Since May 5822 User (~12% of the university)
 - 2431 employees (~40%)
- Up 40M Token and 55k requests per day



RAG – Retrieval Augmented Generation

RAG is currently tested in two use cases at our university:

- IT-support chat bot
- course catalogue



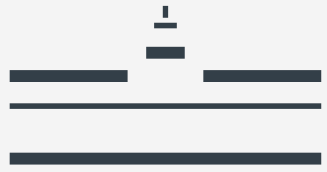
What we learned so far

- Quality of open source LLMs is catching up and in some areas even overtaking commercial LLMs
 - 50% of the users preferred Llama 3.1 Nemotron answers over OpenAI's latest flagship gpt-4o¹
- Privacy concerns are more important than initially expected
- Resistance and complaints among the students and professors are lower than initially thought
- The use of the API is significantly more in demand
 - We started offering an API with LiteLLM proxy

¹Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, & Ion Stoica. (2024). Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference.

Next steps (in increasing complexity)

- Enable tool access and multimodality
 - Deploy Llama-3.2-90B, once the A100 are installed
 - First demo of text-to-image based on flux in testing
- Hyperparameter Tuning
 - mainly sampling, e.g. temperature, top-k
- Finetune our own model, finetune heads



Universität
Münster

Thank You

Jonathan Radas

jonathan.radas@uni-muenster.de

living.knowledge

