Open WebUI @ MAX IV

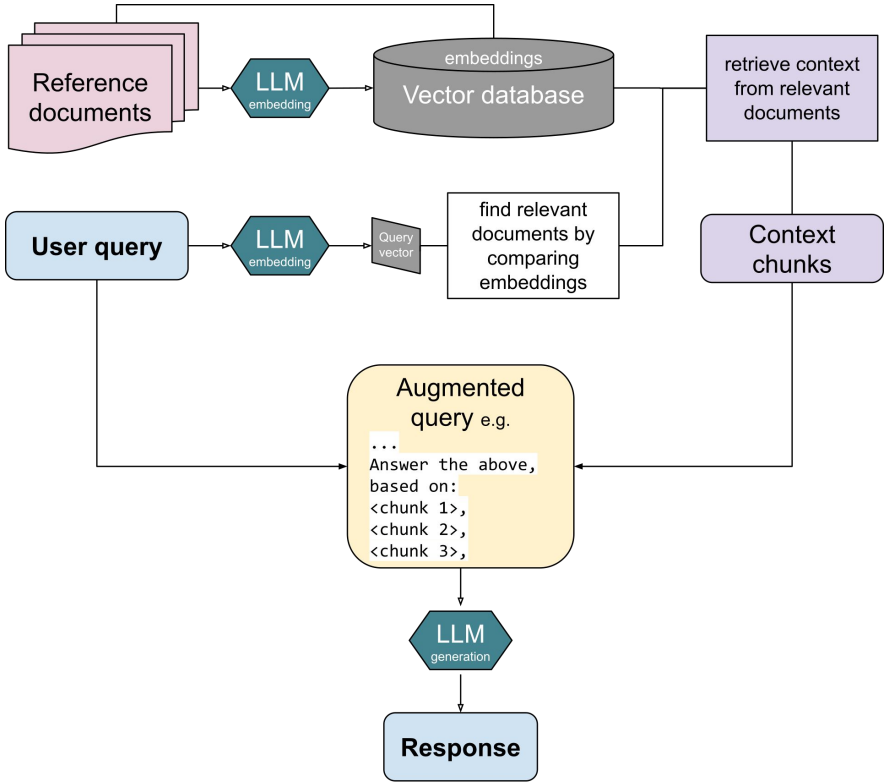
to find a right needle in a haystack of documentation

Dmitrii Ermakov, Andrii Salnikov
IT Infrastructure

Challenges of finding the MAX IV-specific information

- Information is scattered around ***different sources***, ***not well-structured***:
  - Elogy
  - Wiki
  - Gitlab repos / pages
  - Gitlab Issues
  - RT Tickets
  - …
- Hard to search for specific subject or issue
- Complex **system dependencies**
  - Software vs Hardware vs IT Infrastructure part
  - Usually knowledge is scoped to expertise area

Retrieval-augmented generation (RAG)



*https://en.wikipedia.org/wiki/Retrieval-augmented_generation

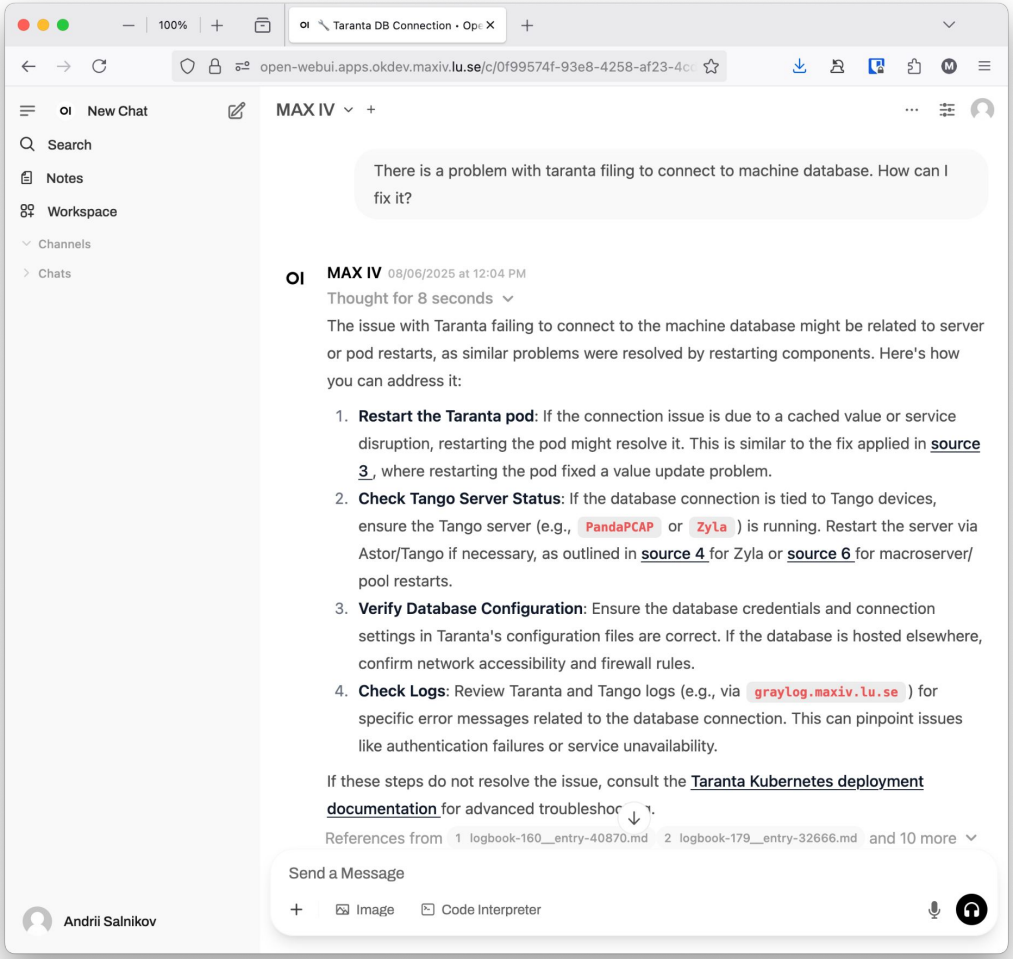
- Documentation ➪ ***LLM embeddings***
- No need to retrain the LLM
  - supplement information
- **Include sources URLs** in the responses via metadata
  - human can evaluate

Open WebUI

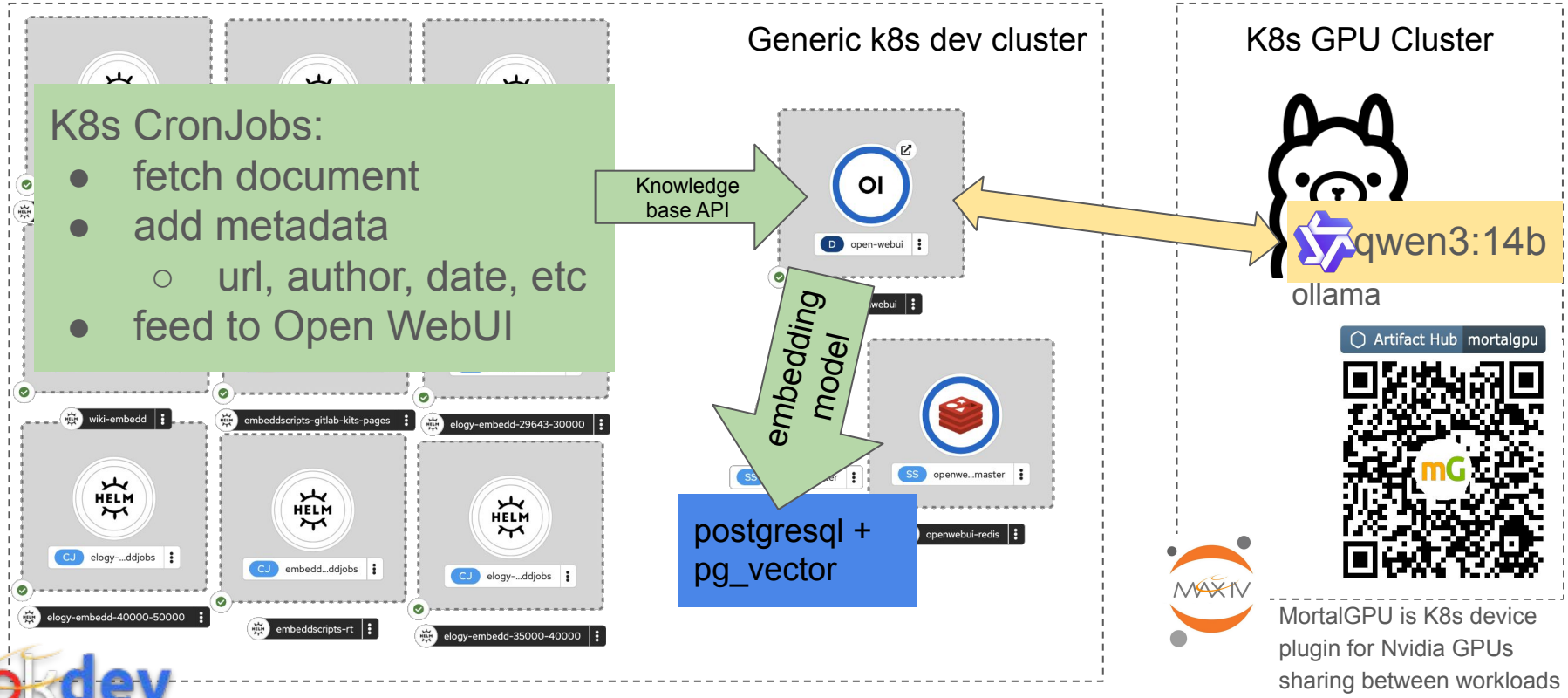
“Open WebUI is an extensible, feature-rich, and user-friendly self-hosted AI platform designed to operate entirely offline.

It supports various LLM runners like **Ollama** and **OpenAI-compatible APIs**, with **built-in inference engine** for RAG, making it a **powerful AI deployment solution**.”*

*https://docs.openwebui.com/



PoC Deployment



Going forward

- add more information sources
  - experimenting with **AI Agents** **(“Tools” in Open WebUI)** to query live data from ***Graylog***, ***Prometheus, etc***
- fine-tuning of the “**system prompt**”:
  - “Include URLs to the response”
- production deployment
  - dedicated resources
- more wide usage as **first-line support tool** in KITOS/On-call

Generated with AI · Microsoft Copilot


MAX IV
KUBERNETES
Thank you for attention!
Generated with AI · ChatGPT