Agenda

- AI Today: The Generative Landscape
- **Serving a LLM On-Prem: Open WebUI**
- *Coffee break*
- RAG – Knowledge Bases and Linking a SQL Database
- How Agents Talk to Tools: Towards MCP
- Closing Remarks

About me

Alexander Sternfeld

- Associate researcher
  @ Reliable Information Lab, HES-SO
- Safety and Security @ Apertus Team
- Former intern
  @ Cyber-defence Campus, armasuisse
- Data science graduate
  @ EPFL

How to choose your LLM?

Think carefully about your requirements

- What is the task at hand?
- Do I run locally or do I use an inference provider?
- How important is inference speed?
- What is my budget?

Compare Models

- Relevant Benchmarks?
- AI Sheets ?
- Model training data includes relevant data?
- Toy example test?

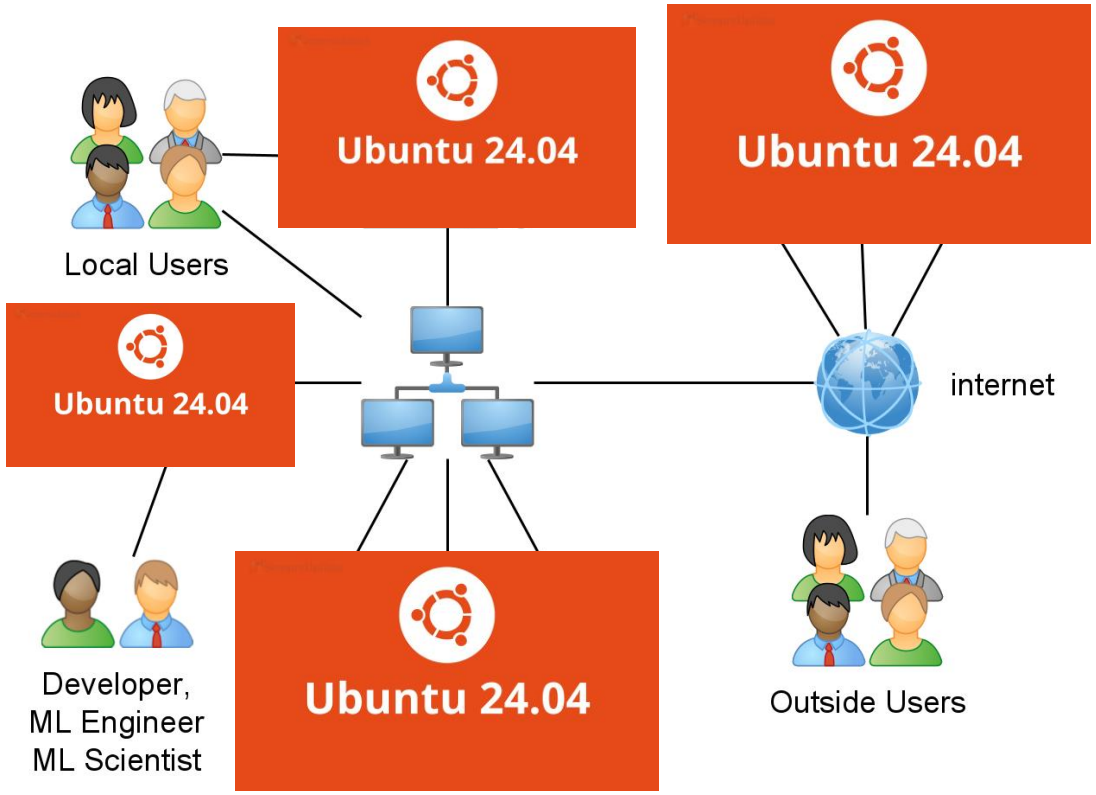
A Few Considerations on Model Size Scaling

- Non-quantized models:
  - **N** B parms
    => **N** x 2 GB of vRAM + ~1GB
  - **70** B params => H200
  - **32** B params => H100
  - **8** B params => 18 GB vRAM
- Quantized:
  - F16: N x GB vRAM + ~1GB
  - INT8/F8: N x ½ GB vRAM + ~1 GB

- Attention:
  - Those estimations are for small "chat" context windows. Long context increases memory need.

On-Premises:

EXOSCALE



- We simulate a “clean machine”
- Our “clean machine” is on virtual private cloud
  - Not exposed to internet otherwise
- The virtual private cloud is hosted by Exoscale
  - Geneva (A30); Vienna (A5000)
  - CH-based Cloud, Host, Compute
  - Compliance (GDPR) & Usability

Tooling for This Workshop

- Website: https://llm-on-prem-amld2026.github.io

- Private servers (Exoscale)
  - Full set-up of a clean machine
  - From a minimal PoC
  - To a minimal personal prod.
- Re-do the workshop yourself:
  - With your hardware
  - Or on the same Private Servers

LLM Chat GUI

- Many varieties
- Different functionalities, different (dis)advantages
- We choose Open WebUI
  - High performance
  - API integration
  - User-friendly
  - Ongoing community-based development
  - Positive Experience with it

Technical set-up

- Open the website: https://llm-on-prem-amld2026.github.io
- Make sure you have gotten a paper with an **(IP, password)** combination

Go through the sections ***Connecting to your machine*** and ***Setting up your own Open WebUI, carefully follow the steps.***

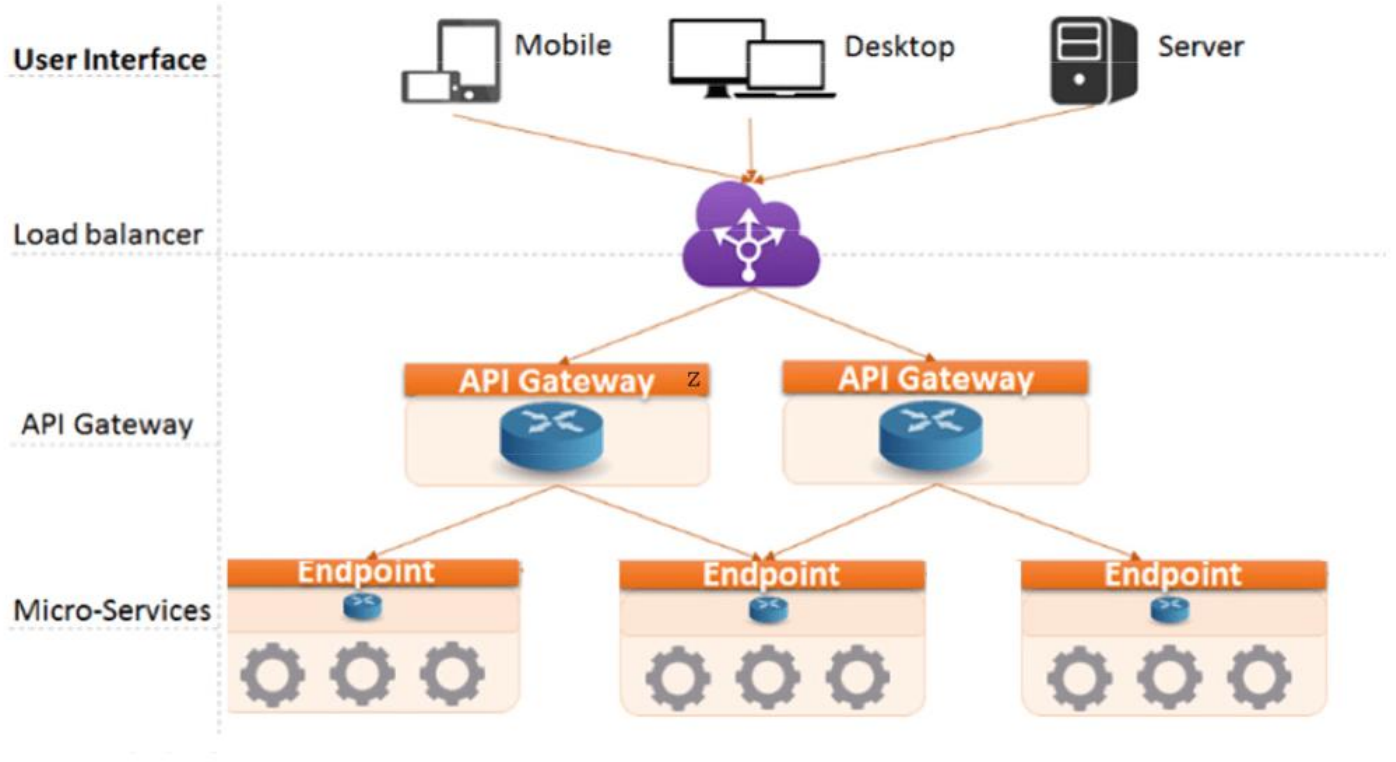
- In case you have any problems, please raise your hand, and do not hesitate to collaborate with your neighbors ☺

Now you have your own Chat UI!

- **Limitation:** computational power
- Can only use **cpu-only** models…

- How did you like the performance of the cpu-only model?

Solution: usage of an API



Technical set-up - continuation

Follow the instructions on the page ***Adding the API key*** to add the two endpoints to your Open WebUI instance

- Password for the encrypted zip file: **amld2026llmonprem**

→ You now unlocked: