

# AI in eArchiving, with or without the supercomputer “Hippu”

Anssi Jääskeläinen

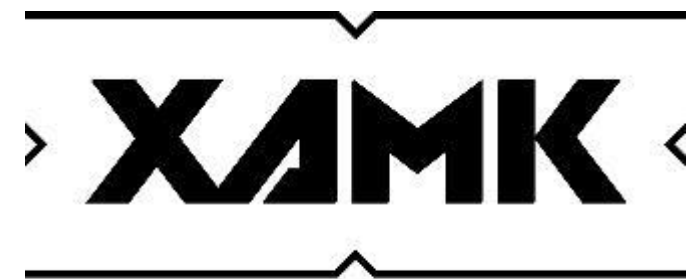
Research manager Xamk/Digitalia

[Anssi.jaaskelainen@xamk.fi](mailto:Anssi.jaaskelainen@xamk.fi)

<https://memorylab.fi/>



South-Eastern Finland  
University of Applied Sciences



**Memory Lab**

**Digitalia**



*"AI is everywhere, even now; in this very room, you can see it when you look out your window or when you turn on your television, you can feel it when you go to work."*

# Agenda

---

- Short intro
- AI in generally
  - Foundation models
  - Transformers
  - LLMs
  - Generative AI
- Hippu supercomputer
  - Technological overview
  - Software stack
- Practical examples & RDI
  - Utilization of existing models
  - Training of new models

1. Introduction

Zoom Meeting, Oct 10, 2020



# Intro

1. Introduction  
2. Overview of the course  
3. Course objectives  
4. Course structure  
5. Course materials  
6. Course evaluation  
7. Course contact information



# Xamk / Digitalia / Memory Lab

- South-Eastern Finland University of Applied Sciences
  - <https://www.xamk.fi/en/frontpage/>
- Memory Lab
  - AI specialized ~680 000€ technical environment
  - Operational June 23.
  - <https://memorylab.fi>
- Digitalia – Research Center on Digital Information Management
  - Founded 2015
  - Usability of digital materials
  - Automated things
  - Visualization
  - [Digitalia.fi](https://digitalia.fi)
  - More things:  
<https://digitalia.xamk.fi/>



# AI – The first things to remember

---

- Not an all-mighty thing that solves every possible problem
  - Although best LLMs might seem like it
- Often designed and trained to solve one specific thing
  - page empty or not
  - female or male
  - Cat or dog
- Current hype in generative AI
  - Includes LLMs and image generators
- Training an AI takes A LOT of time
  - Data collection & annotation ~80-90% of the time
  - Altering training variables & training 10-20%
- Can hallucinate, never blindly trust your AI
  - It is just predicting and calculating probabilities



# About hallucination...

## **what can you tell me about ELKA that is located at Mikkeli**

- I'm sorry, but I don't have access to current information. However, ELKA is a Finnish company that produces electronic components and systems, and they have a factory in Mikkeli.

## **What can you tell me about Central Archives for Finnish Business Records**

- The Central Archives for Finnish Business Records is a national archive in Finland. It's located in the city of Mikkeli and it's responsible for storing and preserving business records from all over the country. It's a very important resource for researchers and historians who want to learn more about the history of Finnish businesses.

## **Who is the director of Central Archives for Finnish Business Records**

- I'm sorry, but I don't have access to current information. However, you can visit the official website of the Central Archives for Finnish Business Records to find out more about their current leadership team.

## **can you link me to the official website**

- Sure! The official website of the Central Archives for Finnish Business Records is <https://www.arkisto.fi/en/>.



A green wireframe cube is centered in the frame. Inside the cube, there are several smaller, semi-transparent green shapes, including a pyramid and some irregular polygons. The background is dark with a pattern of small white dots, resembling a starry sky or a digital space. The text "AI in generally" is written in white, sans-serif font across the middle of the cube.

AI in generally



# What is a foundation model?

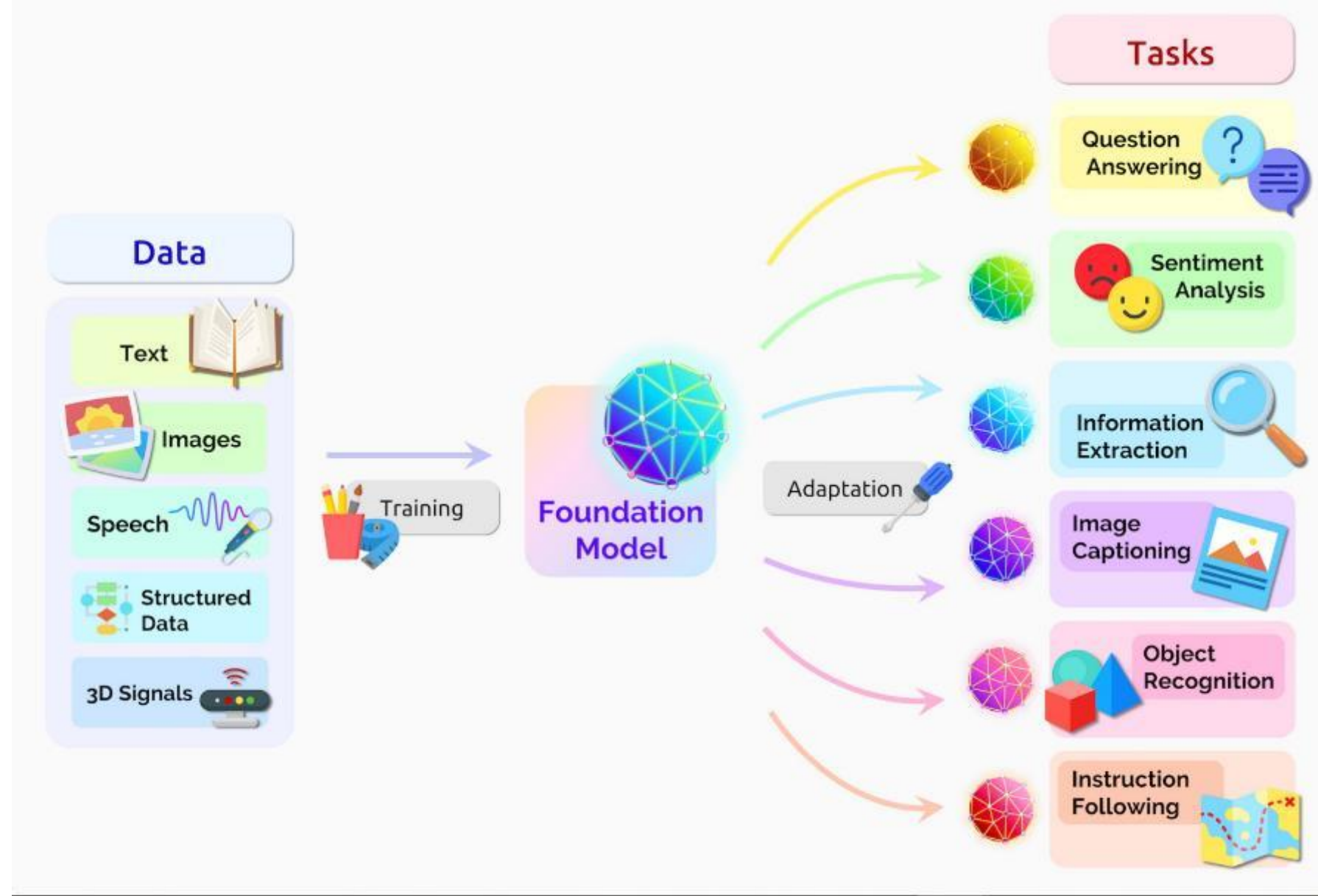
---

- Shortly: Pre-trained model, ready to be utilized
- Instead of training from the beginning → Take the foundation model and fine tune it to suit your needs
  - Lot more effective
  - ~550 000 foundation models freely available
  - Most known example
    - GPT Base model
    - ChatGPT, PDFGPT



# Transformer models & techniques

transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data



# The most known transformer model = GPT

- General Pre-trained Transformer
- Transformer architecture landmark paper “[Attention is All You Need](#)”
- All familiar with chatGPT
  - A finetuned GPT model for Q&A tasks
  - Broke the Turing test:  
<https://www.nature.com/articles/d41586-023-02361-7>
    - Based on GPT-4 (free GPT 3.5)

“transformer architecture” by Stable Diffusion

*Steps: 100, Sampler: DPM++ 2M Karras, CFG scale: 7, Seed: 34, Face restoration: CodeFormer, Size: 512x512, Model hash: f6a8a4e67d, Model: brixIAMustInYour\_v5EndOfTheLine, Version: v1.7.0*

*Time taken: 26.2 sec.*

*A: 8.18 GB, R: 9.23 GB, Sys: 9.8/79.1533 GB (12.4%)*



# LLM – Large Language Model

- Trained on massive datasets
  - Can generate false, offensive, harmful, and biased content
- Best open source options (at the moment)
  - LLaMA 2, BLOOM, BERT, Falcon, OPT
  - Test Falcon-180B Chat  
<https://huggingface.co/spaces/tiiuae/falcon-180b-demo>
  - Was able to solve this: “You are in a room that has three switches and a closed door. The switches control three light bulbs on the other side of the door. Once you open the door, you may never touch the switches again. How can you definitively tell which switch is connected to each of the light bulbs?”
    - Can you?
- Pick the right one for your purposes
  - Not a simple tasks but luckily switching is simple
    - pipe = pipeline("summarization",  
model="facebook/bart-large-cnn")





# LLM efficiency – tests

- [https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard)
  - ARC, HellaSwag, MMLU, TruthfulQN, Winogrande and GSM8K
- HellaSwag
  - Dataset for evaluating commonsense
    - <https://github.com/rowanz/hellaswag/tree/master/data>
  - OpenAI GPT4 = 95,3 % (1760B parameters)
  - Humans ~95%
  - Best OS model (currently) saltlux/luxia-21.4B-alignment-v1.0 ~91,2%
  - Tiiuae/falcon180B = 88,89%
  - LLaMA2 70B = 85,3%

# LLM - general challenges

- Picking a model
  - Wrong foundation model / model trained with incorrect parameters
- Running a model
  - Output is too long or short
    - Not specified before the use
  - Generation mode does not match the need
    - Creative tasks vs. translations
  - Prompting
    - Proper content / role for the AI hasn't been defined

The image shows a dark-themed user interface for configuring LLM training. It features several sections with sliders and input fields for different parameters.

- LoRA Rank:** A slider set to 32. Description: "Also called dimension count. Higher values = larger file, more content control. Smaller values = smaller file, less control. Use 4 or 8 for style, 128 or 256 to teach, 1024+ for fine-detail on big data. More VRAM is needed for higher ranks."
- LoRA Alpha:** A slider set to 64. Description: "This divided by the rank becomes the scaling of the LoRA. Higher means stronger. A good standard value is twice your Rank."
- Batch Size:** A slider set to 128. Description: "Global batch size. The two batch sizes together determine gradient accumulation (gradientAccum = batch / microBatch). Higher gradient accum values lead to better quality training."
- Micro Batch Size:** A slider set to 4. Description: "Per-device batch size (NOTE: multiple devices not yet implemented). Increasing this will increase VRAM usage."
- Cutoff Length:** A slider set to 256. Description: "Cutoff length for text input. Essentially, how long of a line of text to feed in at a time. Higher values require drastically more VRAM."
- Save every n steps:** An input field with the value 0. Description: "If above 0, a checkpoint of the LoRA will be saved every time this many steps are completed."
- Epochs:** An input field with the value 3. Description: "Number of times every entry in the dataset should be fed into training. feed each item in once, 5 means feed it in five times, etc."
- Learning Rate:** An input field with the value 3e-4. Description: "In scientific notation. 3e-4 is a good starting base point. 1e-2 is extremely high. 1e-5 is extremely low."
- LR Scheduler:** A dropdown menu showing "linear". Description: "Learning rate scheduler - defines how the learning rate changes over time. 'constant' means never change, 'linear' means to go in a straight line from the learning rate to 0, cosine follows a curve, etc."
- Advanced Options:**
  - LoRA Dropout:** A slider set to 0,05. Description: "Percentage probability for dropout of LoRA layers. This can help reduce overfitting. Most users should leave at default."
  - Stop at loss:** A slider set to 0. Description: "The process will automatically stop once the desired loss value is reached. (reasonable numbers are 1.5-1.8)"
  - Optimizer:** A dropdown menu showing "adamw\_torch". Description: "Different optimizer implementation options, for advanced users. Effects of different options are not well documented yet."
- Warmup Steps:** An input field with the value 100. Description: "For this many steps at the start, the learning rate will be lower than normal. This helps the trainer prepare the model and precompute statistics to improve the quality of training after the start."
- Train Only After:** An input field with the value "### Response:". Description: "Only consider text \*after\* this string in any given chunk for training. For datasets, use '### Response:' to only train the response and ignore the rest of the chunk."
- Add EOS token:** An unchecked checkbox. Description: "Adds EOS token for each dataset item. In case of raw text, the EOS will be added at the end of the line. If checked, changes Rank/Alpha slider above to go much higher. This will require a datacenter-class GPU."
- Enable higher ranks:** An unchecked checkbox. Description: "If checked, changes Rank/Alpha slider above to go much higher. This will require a datacenter-class GPU."

# Hippu supercomputer

Leverage from  
the EU  
2014–2020



European Union  
European Regional  
Development Fund



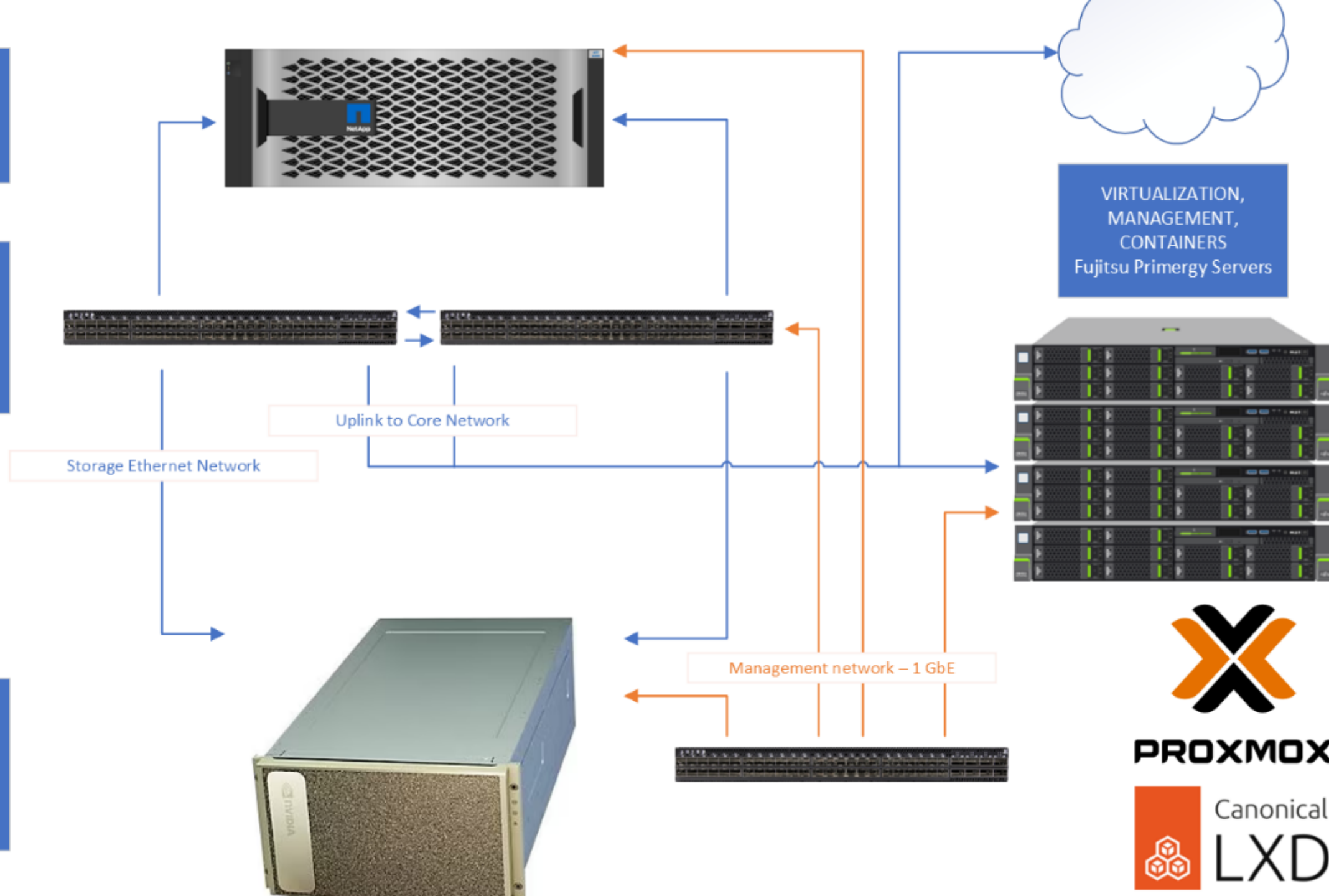
# Memory Lab: “Hippu” DGX A100

- [Among 200 the most powerfull](#) in a world and ranks 3-5 in Finland.
  - 157 = BioHive-1 - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband
- Theoretical peak performance 5 Petaflops (6,5kW)
  - LUMI 531 petaflops (7100kW)
  - Mahti 9,5 petaflops (1070kW)
  - Puhti+Puhti AI 4,5 petaflops (585kW)
- Comparable to about 20 000 basic laptop computers.
- Based on Nvidia A100 cards, each has 80Gt HBM2 memory
  - Total VRAM 640 GBs
  - <https://www.nvidia.com/en-us/data-center/a100/>

Storage  
NetApp AFF A400

Ethernet Network  
NVIDIA Mellanox  
100/25GbE  
NFS, VMs, Containers

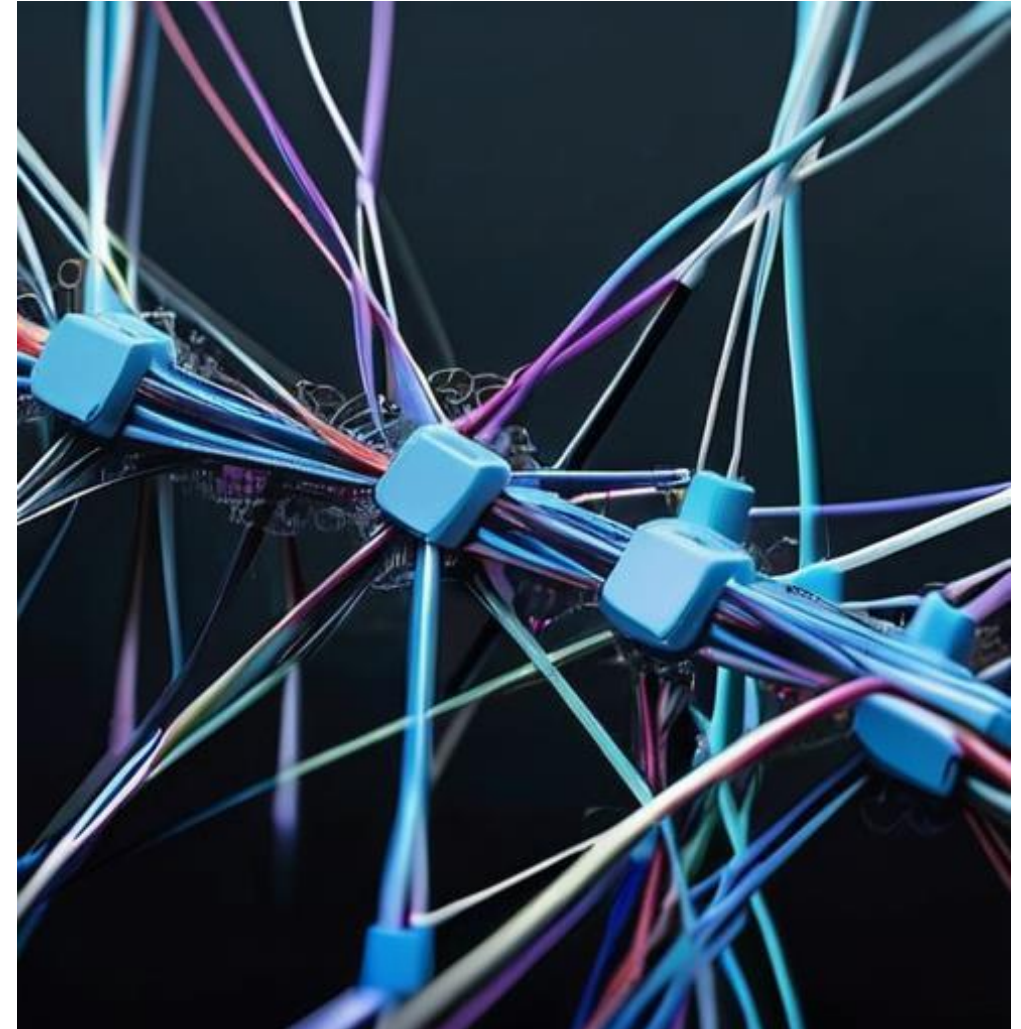
COMPUTE  
NVIDIA DGX A100  
8 x A100 80GB GPU



# Software stack == “Wasting” performance

---

- DGX unit has been virtualized
  - Two VM:s with 2 GPU:s
  - One VM with 4 GPU:s
- HPC is commonly run with Slurm or something similar
- Proxmox as a hypervisor, main reasoning
  - Utilizing existing VMs
  - Open Source
  - Previous experience
  - High availability
  - Scheduled backups
  - Live migration
- → We decided to go with Proxmox
  - <https://www.proxmox.com/en/>
  - KVM/QEMU-based





A robotic hand with a grey, segmented design is shown interacting with a large, horizontal digital tablet. The tablet displays a complex interface with various data visualizations, including a bar chart on the left, a line graph in the center, and a grid of small data points on the right. The robotic hand is positioned over the tablet, with its fingers extended as if it is about to touch or is touching the screen. The background is a dark, textured surface, possibly a desk or a wall, which adds to the futuristic and technological feel of the image.

# Practical examples & RDI

# Similary, summarizer, tagger, tone analyzer

- Try it out yourself: <https://memorylab.fi/AIKO/cluetail-demo/>
  - Summarizer: [facebook/bart-large-cnn](https://huggingface.co/facebook/bart-large-cnn)
  - Tagger: [facebook/bart-large-mnli](https://huggingface.co/facebook/bart-large-mnli)
  - Tone analyzer: [cardiffnlp/twitter-roberta-base-sentiment-latest](https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest)
  - +[Scikit Learn – cosine similarity](#)
- UI with Gradio, <https://www.gradio.app/>
  - pip install gradio
- Code not yet published, will be no later than 30<sup>th</sup> of April 24
  - Xamkfi github

# Automated translations



- Same thing that google offers for websites
- Only with freely usable AI models
  - Practically any language pair is possible
  - Demo from Finnish to English
  - Uses: [Helsinki-NLP/opus-mt-fi-en](https://helsinki-nlp.fi/opus-mt-fi-en)

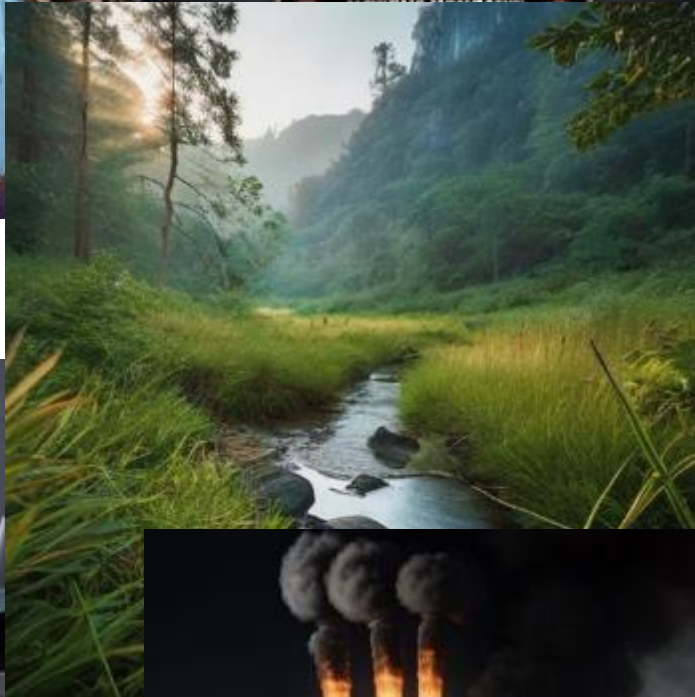
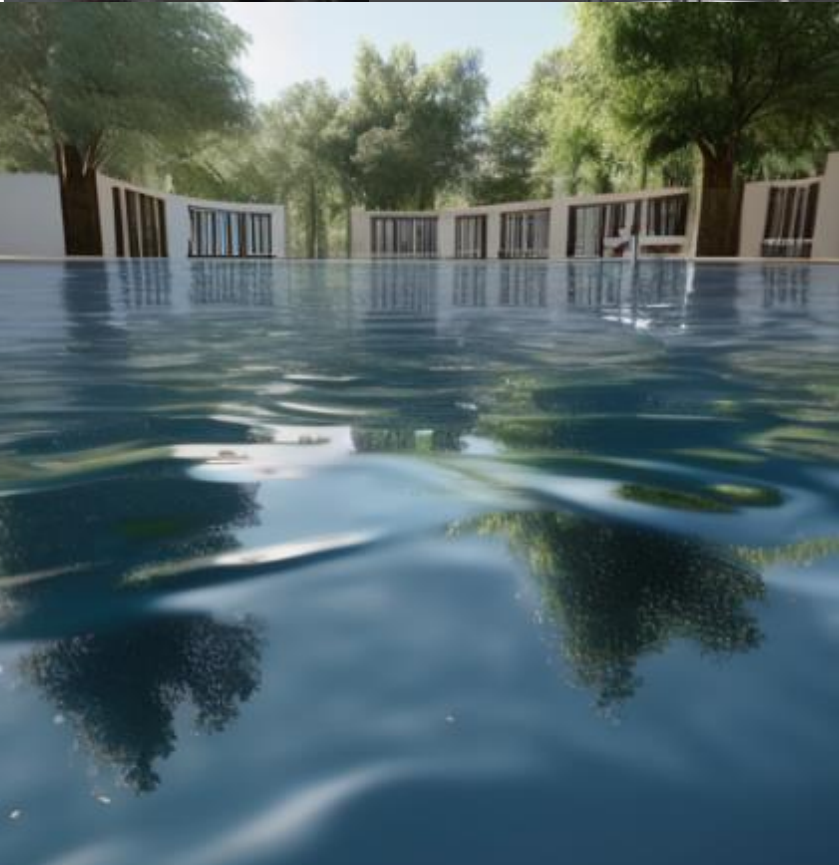
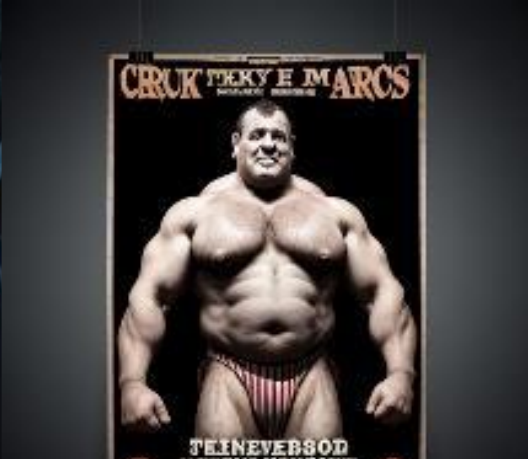


# Metadata Harvesting

- Tagging partially
- Dalai project –demo
  - <https://arkkiivi.fi/language/en/> (Project website)
  - <http://195.148.220.228/> (demo site, not https)
    - Metadata component description: [https://arkkiivi.fi/wp-content/uploads/2023/08/Metatiedot\\_en.pdf](https://arkkiivi.fi/wp-content/uploads/2023/08/Metatiedot_en.pdf)
  - All components freely available

## Components

Blank Page Detection	<input checked="" type="checkbox"/>
Faulty Image Detection	<input checked="" type="checkbox"/> 
Metadata Extraction	<input checked="" type="checkbox"/> 
Subject Indexing	<input checked="" type="checkbox"/>
Language	<input checked="" type="checkbox"/>
<b>Name entities</b>	
Person Names	<input checked="" type="checkbox"/>
Organizations	<input checked="" type="checkbox"/>
Finnish Business	<input checked="" type="checkbox"/>
Identity Code	<input checked="" type="checkbox"/>
Finnish Journal Number	<input checked="" type="checkbox"/>
Date	<input checked="" type="checkbox"/>
Locations	<input checked="" type="checkbox"/>
Geopolitical locations	<input checked="" type="checkbox"/>
Products	<input checked="" type="checkbox"/>
Events	<input checked="" type="checkbox"/>
Nationality, religious and political groups	<input checked="" type="checkbox"/>
Writing Type	<input checked="" type="checkbox"/>





# Importance of prompting

Husky

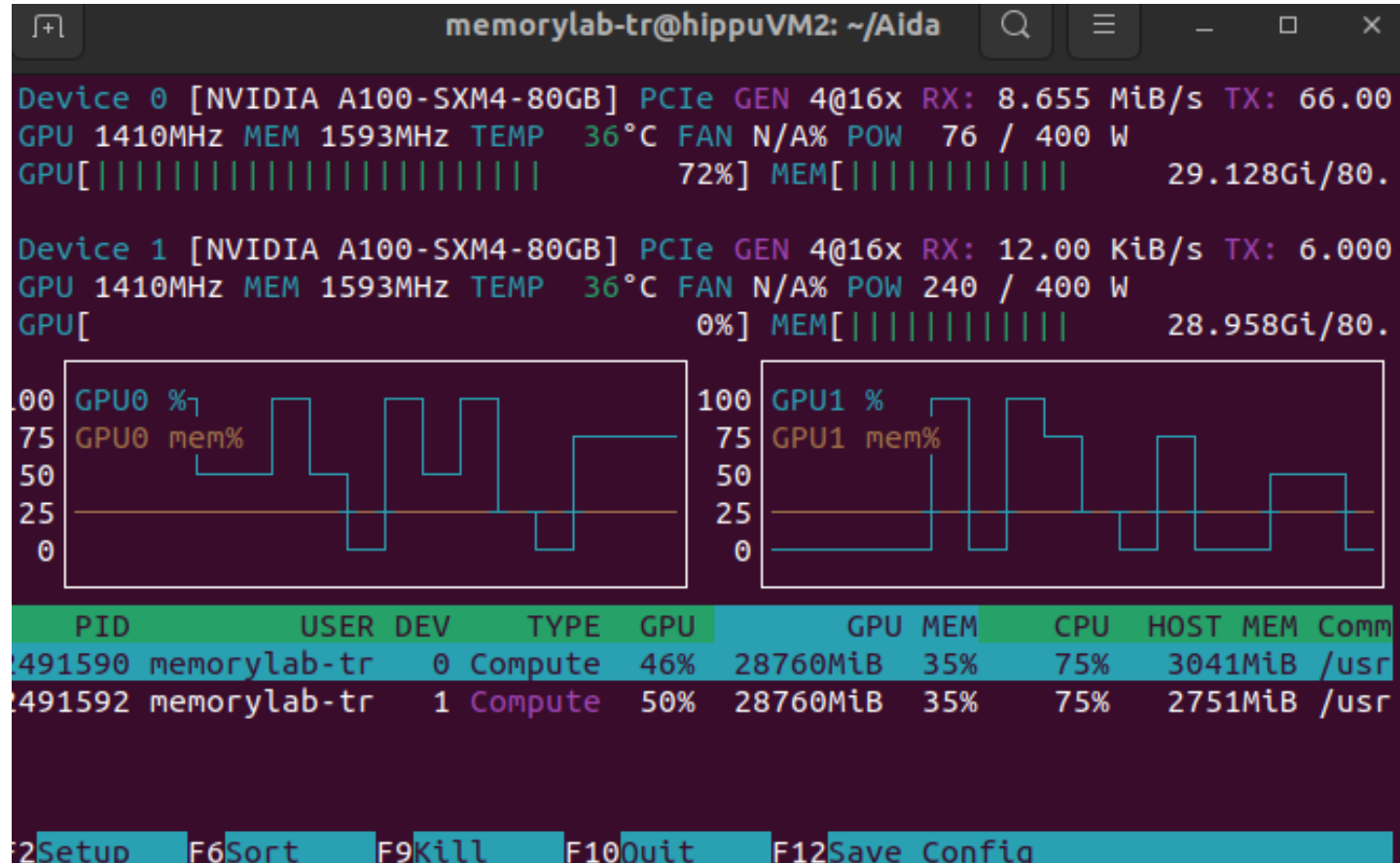
**wolf husky with a demonic glowing red eyes in front of a frozen lake staring directly to camera, Ice reflects the great mountains, photo taken with wide angle lens f=2.8. Ultra realistic fur with minor blood stains. fangs visible**





# Training AI models

- Training PaddleOCR
  - synthetic and annotated data
  - GPU usage 0-100%
  - MEM usage ~30Gt (gpu mem)
- Impossible to run on cpu / basic computer



```
eval model:: 100%| 102/102 [00:46<00:00, 5.74it/s]
eval model:: 100%| 102/102 [00:46<00:00, 2.18it/s]
[2024/03/15 12:08:31] ppocr INFO: cur metric, precision: 0.7981900452488688, recall: 0.922915940006976, hmean: 0.8560336460692334, fps: 7.8711479314500945
[2024/03/15 12:08:31] ppocr INFO: best metric, hmean: 0.9296162450524869, is_float16: False, precision: 0.9174592391304348, recall: 0.942099755842344, fps: 4.067021575
93601, best_epoch: 47
[2024/03/15 12:08:40] ppocr INFO: epoch: [108/500], global_step: 2808, lr: 0.000894, dila_dbloss_Student_Teacher: 0.435302, dila_dbloss_Student2_Teacher: 0.435951, los
s: 5.015853, dml_thrink_maps_0: 0.000849, db_Student_loss_shrink_maps: 1.293087, db_Student_loss_threshold_maps: 0.503031, db_Student_loss_binary_maps: 0.259231, db_St
udent_loss_cbn: 0.000000, db_Student2_loss_shrink_maps: 1.290825, db_Student2_loss_threshold_maps: 0.505025, db_Student2_loss_binary_maps: 0.258726, db_Student2_loss_c
bn: 0.000000, avg_reader_cost: 0.00017 s, avg_batch_cost: 0.80200 s, avg_samples: 12.0, ips: 14.96262 samples/s, eta: 8:28:53
[2024/03/15 12:08:41] ppocr INFO: save model in ./output/fi_PP-OCR_v3_seg_tr/latest
```

# With or without Hippu..?

- Honestly, only one presented activity truly requires Hippu / other powerful computational unit
- Others can also be run on plain CPU
  - But having a decent GPU speeds up processing
- So, to be able to start experimenting, you don't need highend computer



Questions, comments,  
criticism, worries, etc.?