

LLM Unplugged - Welcome

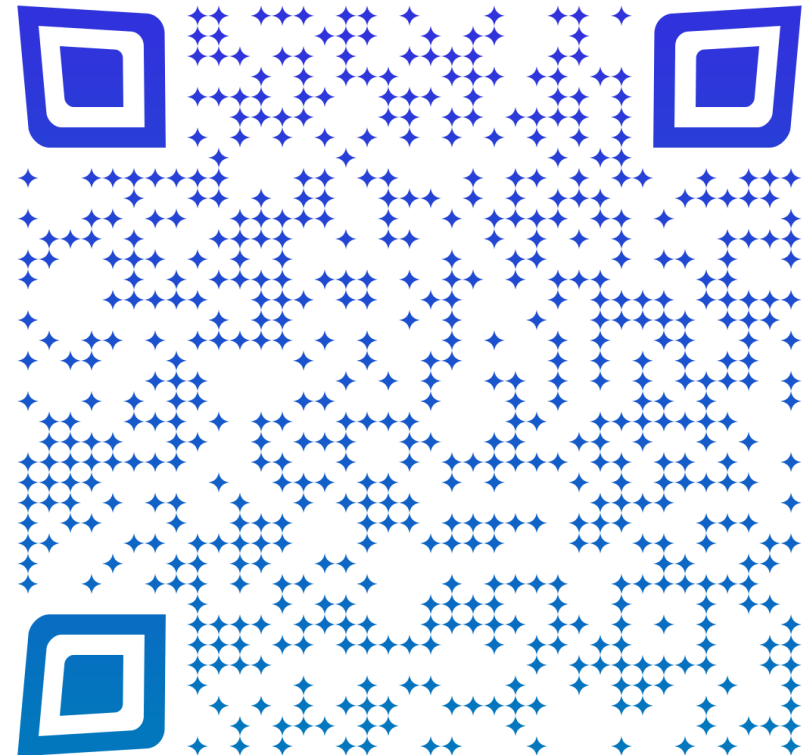
Prepare (if you haven't already):

- 1) Download **AnythingLLM** for your system
- 2) Install AnythingLLM (will do more heavy downloads)
- 3) Install **LM Studio** for your system
- 4) Download a model of your choice from inside LM Studio

Suggestion: LFM2.5-1.2B-Instruct-Q8_0

Exercise Files:

<https://gitlab.com/danilo.biella/llm-unplugged> -> agile



The Slide Pusher



Danilo Biella

Msc Csc 1999

Certificates

Communities

Always sharing #givingisgiving

All slides 

Created with alternative

SKIPADDI

Check-in: Expectations

Accuracy

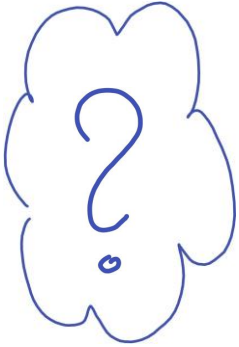
Fairness

Copyright

Speed

Transparency

Security



Availability

Sovereignty

Non-bias

Sustainability

Privacy



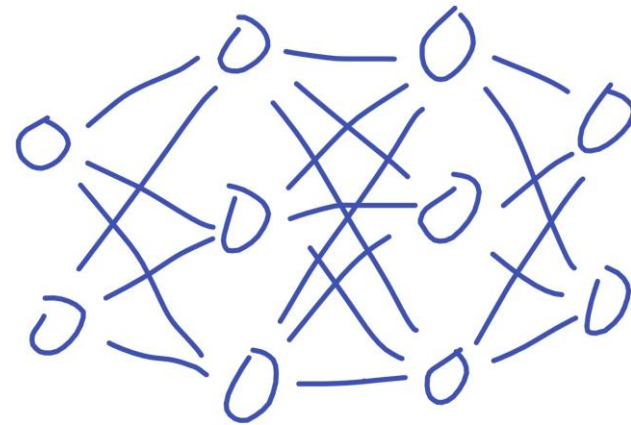


Quote of the Day

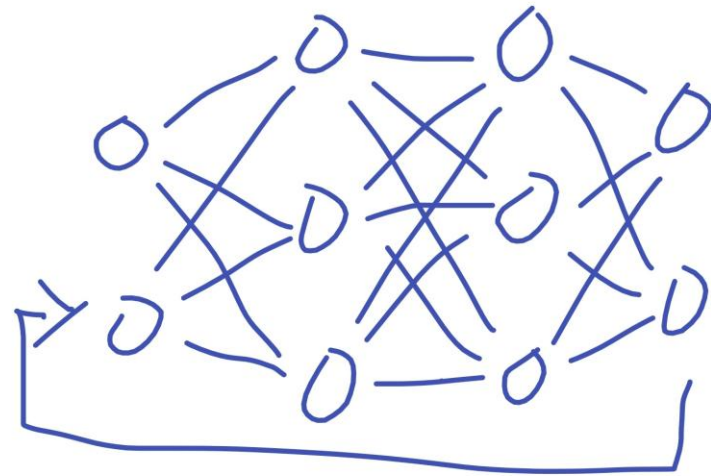
Arguing that you don't care about data privacy because you have nothing to hide is no different than saying you don't care about freedom of speech because you have nothing to say.

-- Edward Snowden

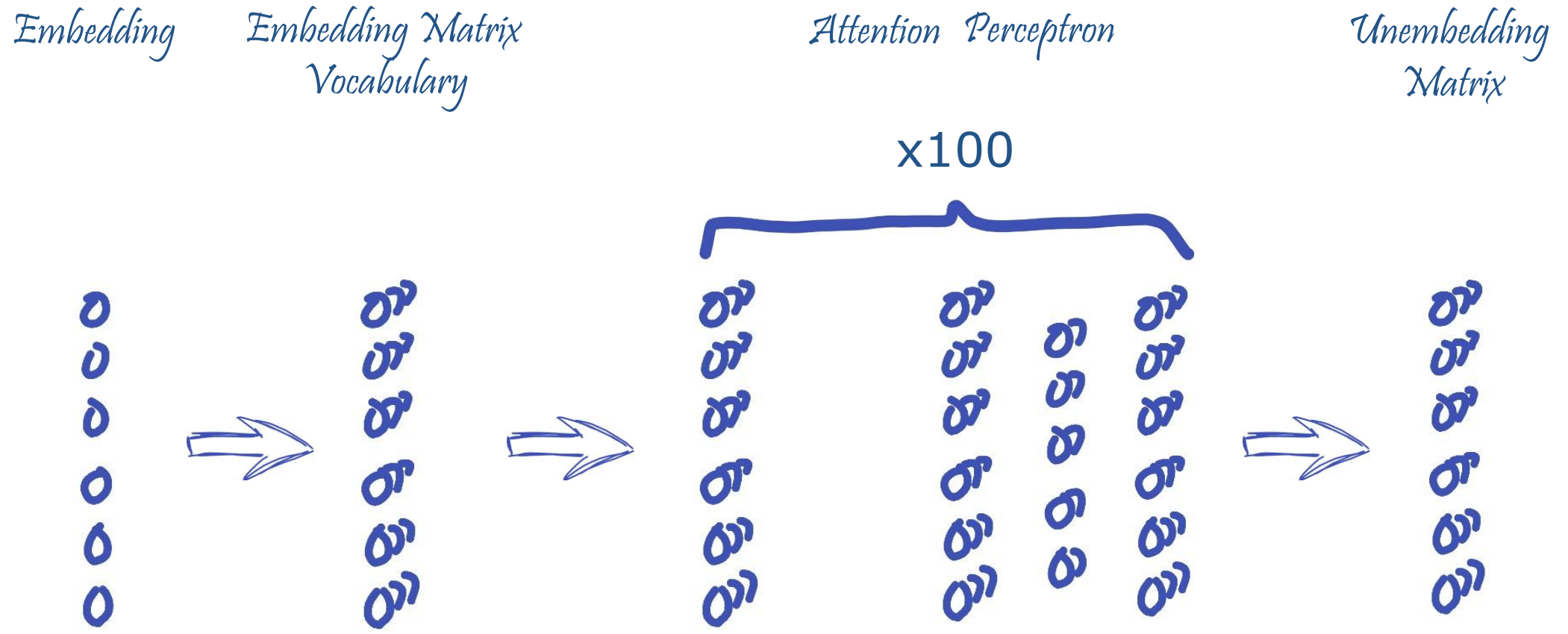
Basic Functionality



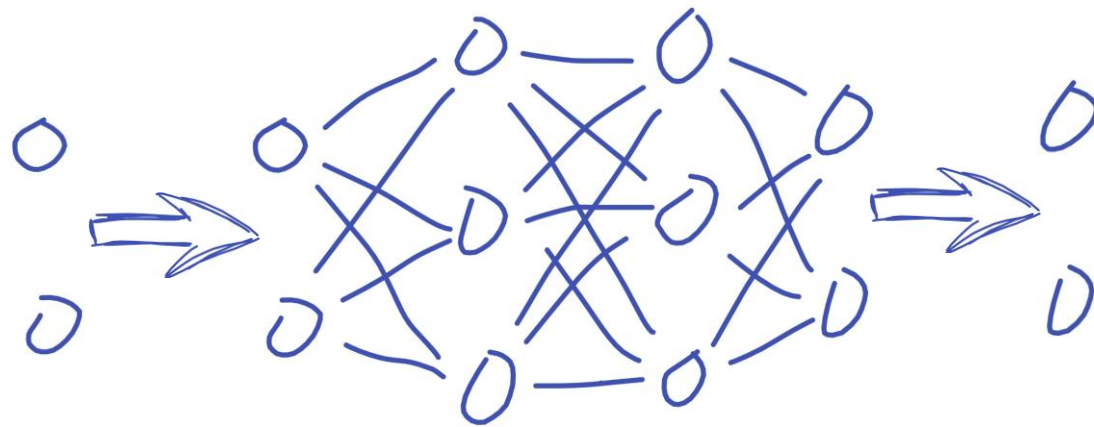
Basic Functionality



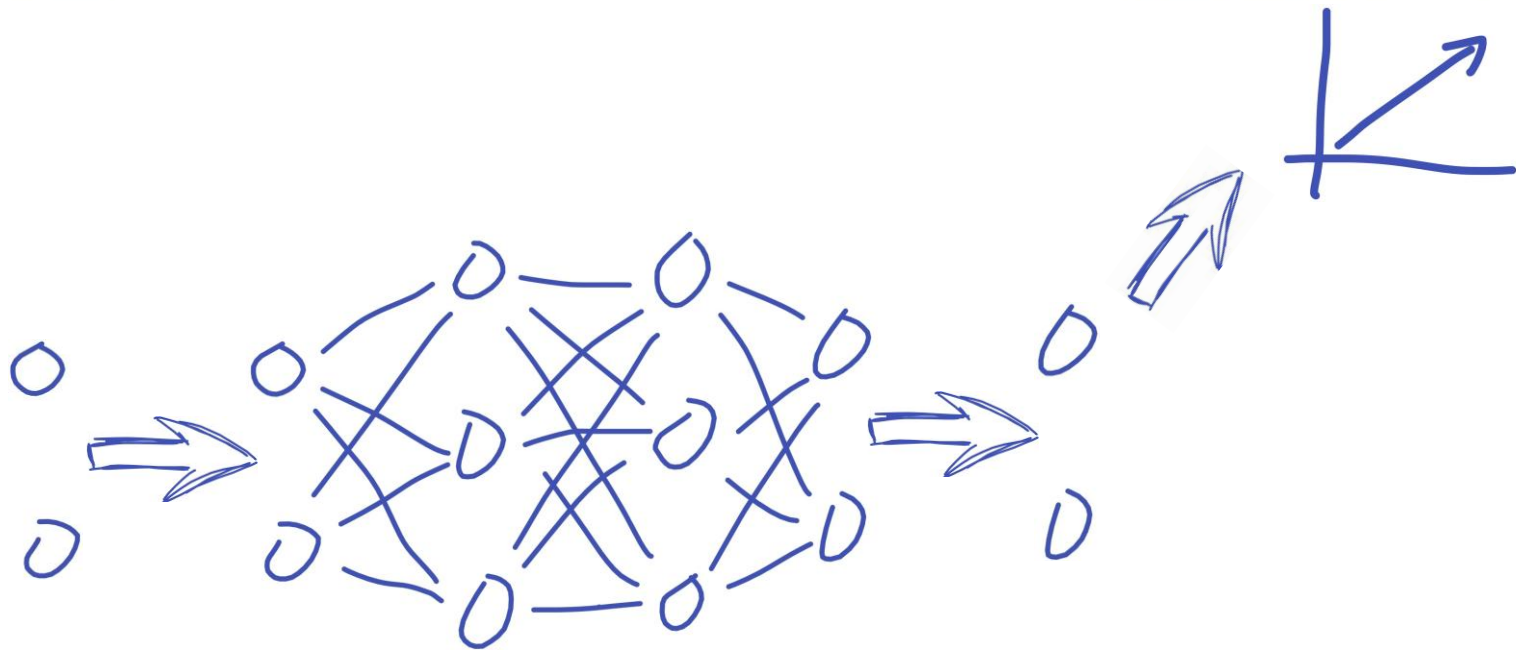
Transformer



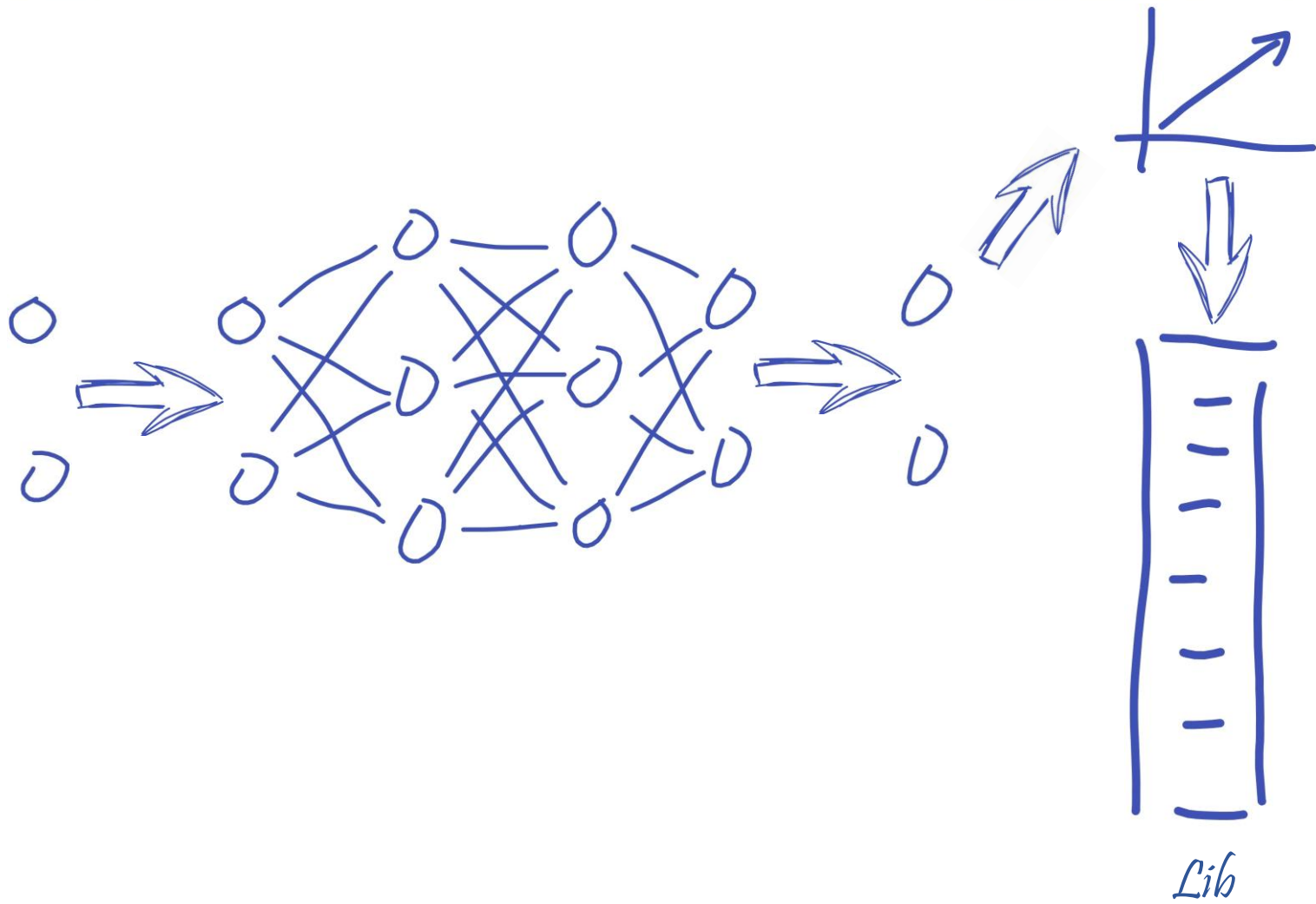
Basic Functionality



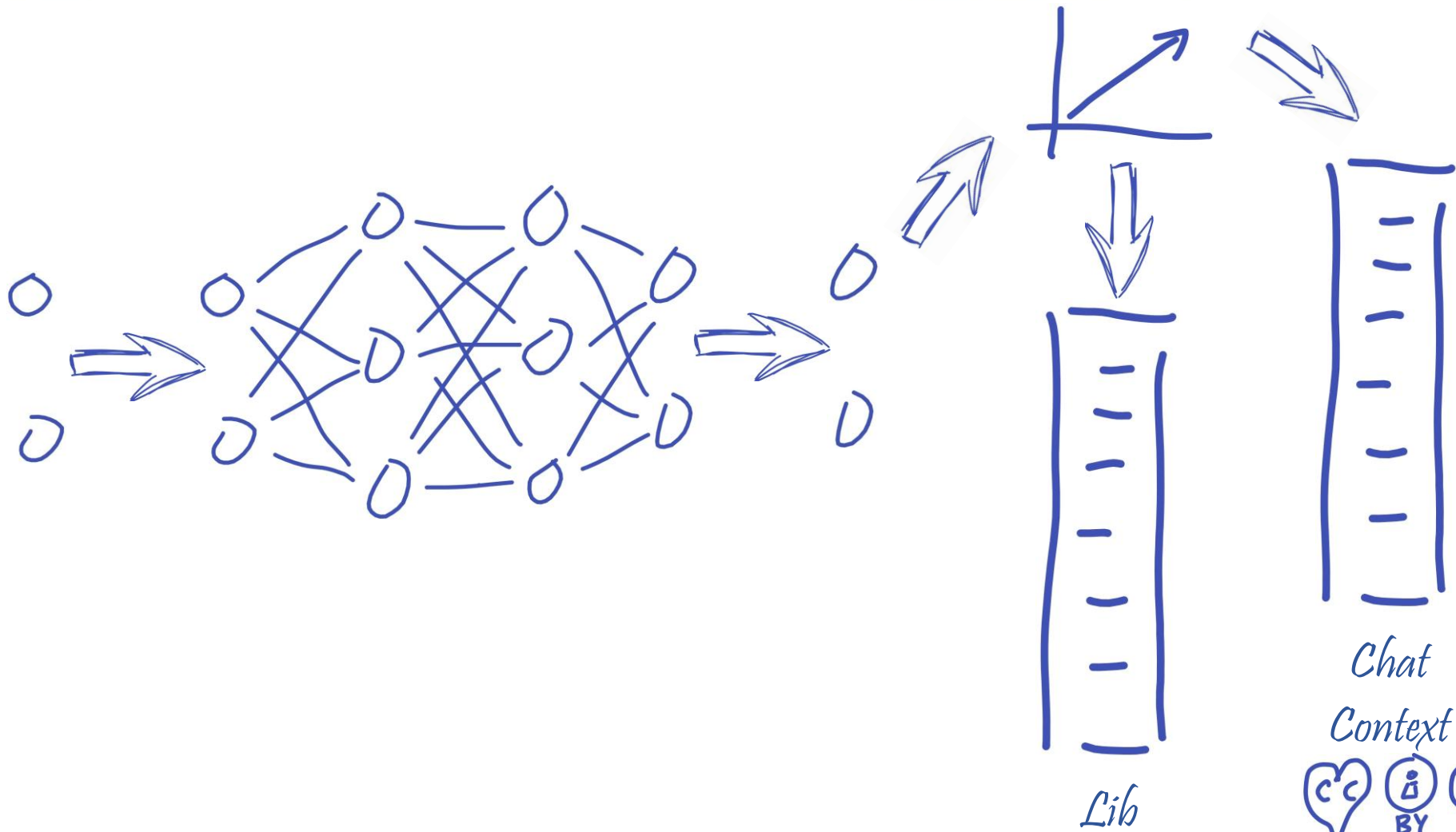
Basic Functionality



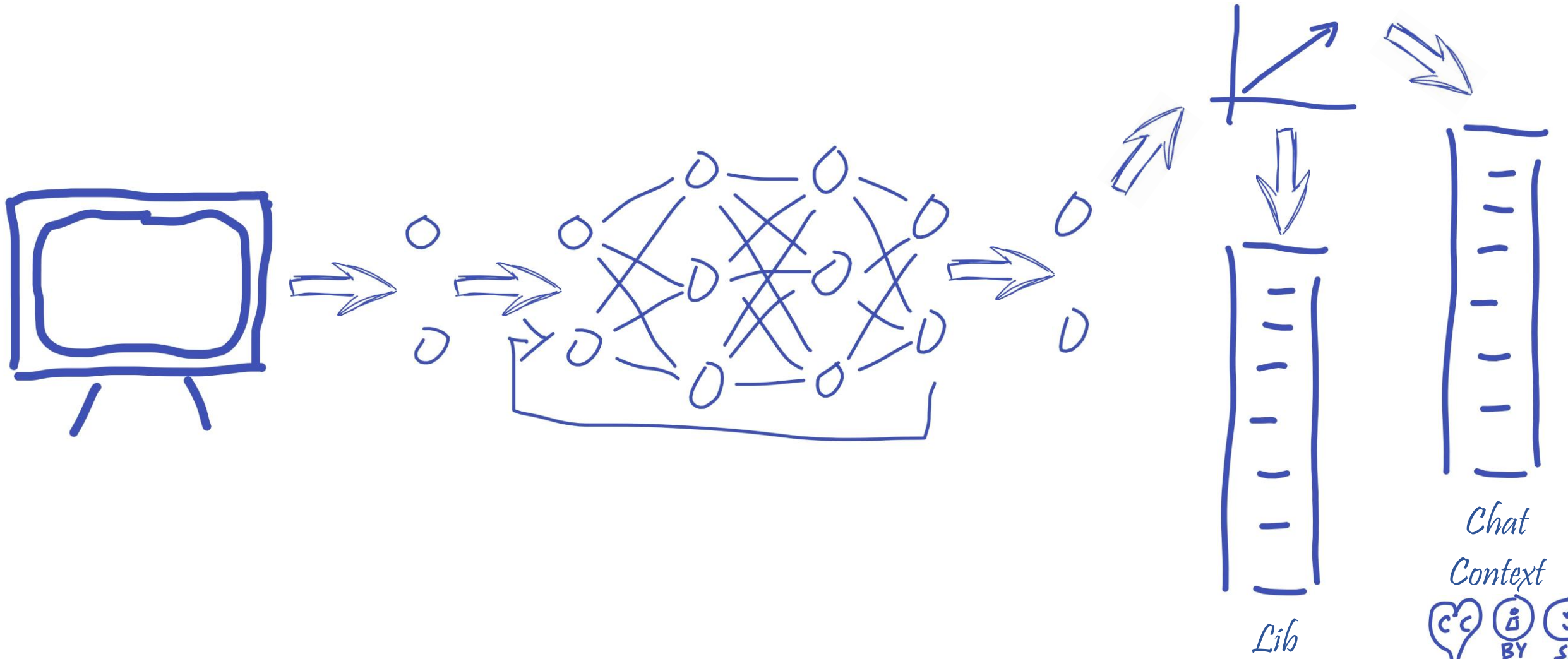
Basic Functionality



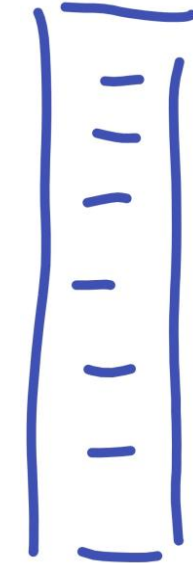
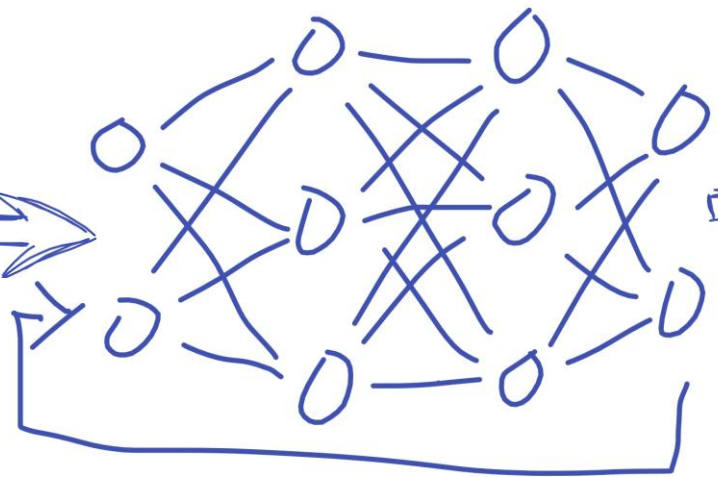
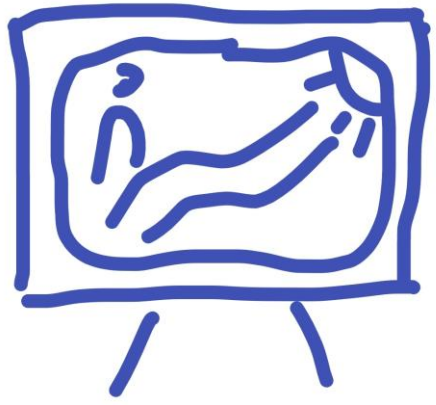
Basic Functionality



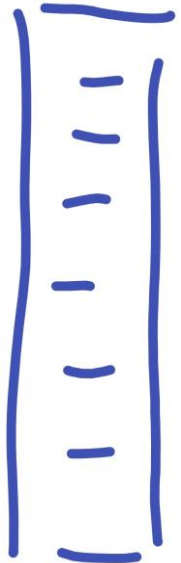
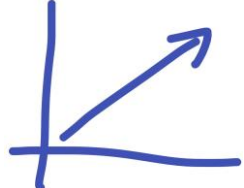
AI-Imaging



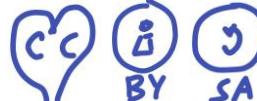
AI-Imaging



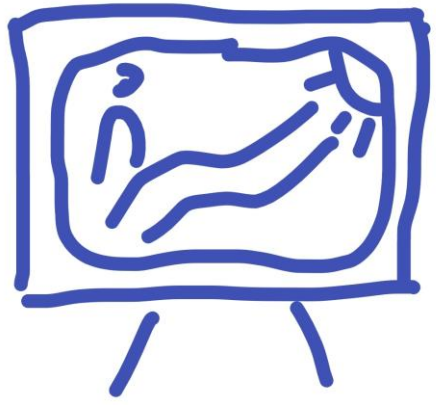
Lib

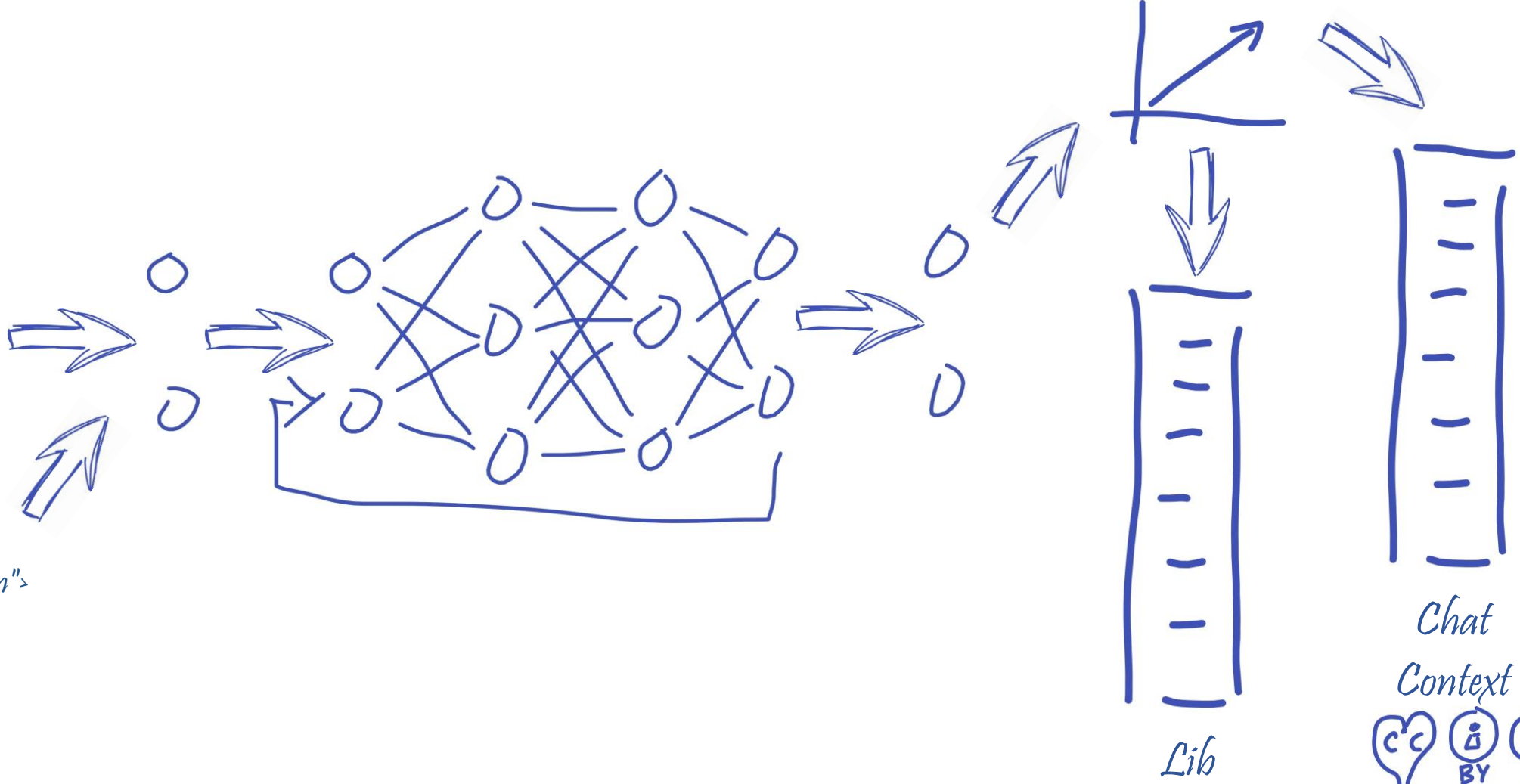


Chat
Context

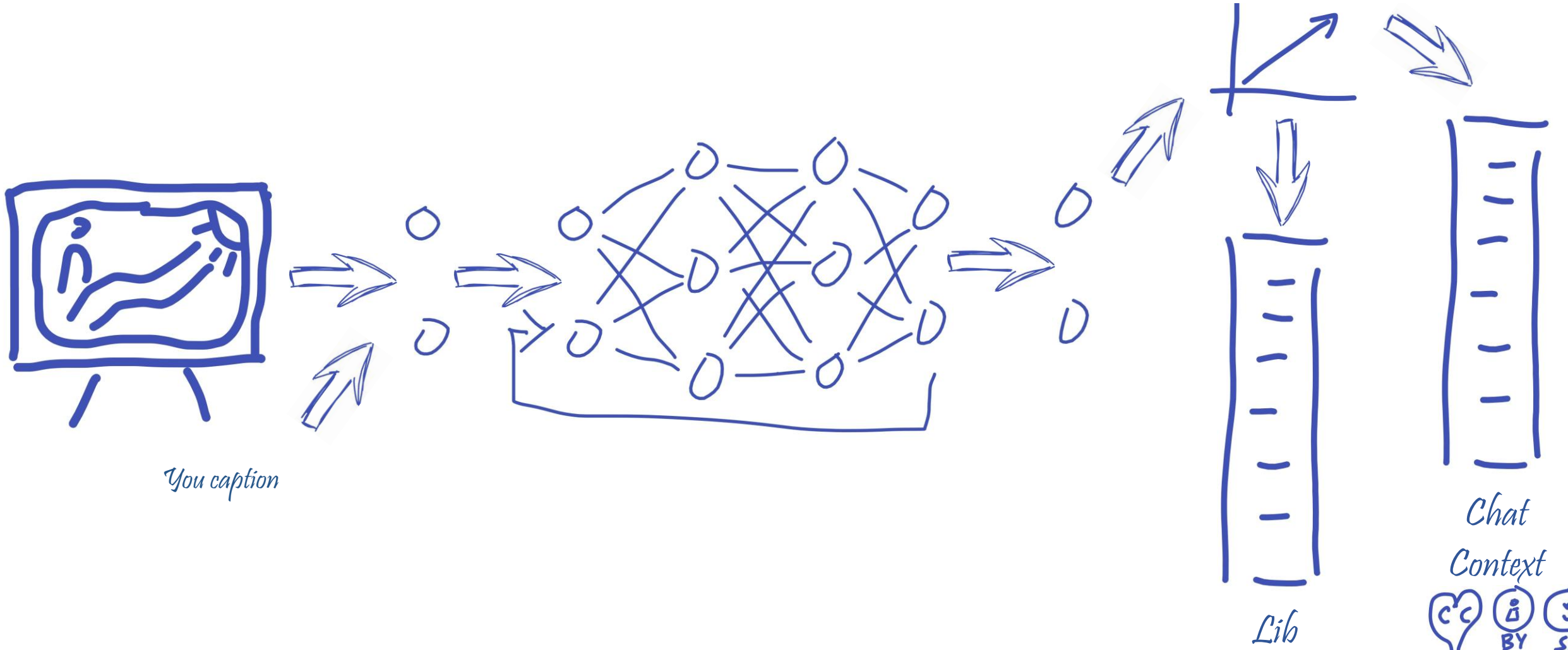


Stable Diffusion

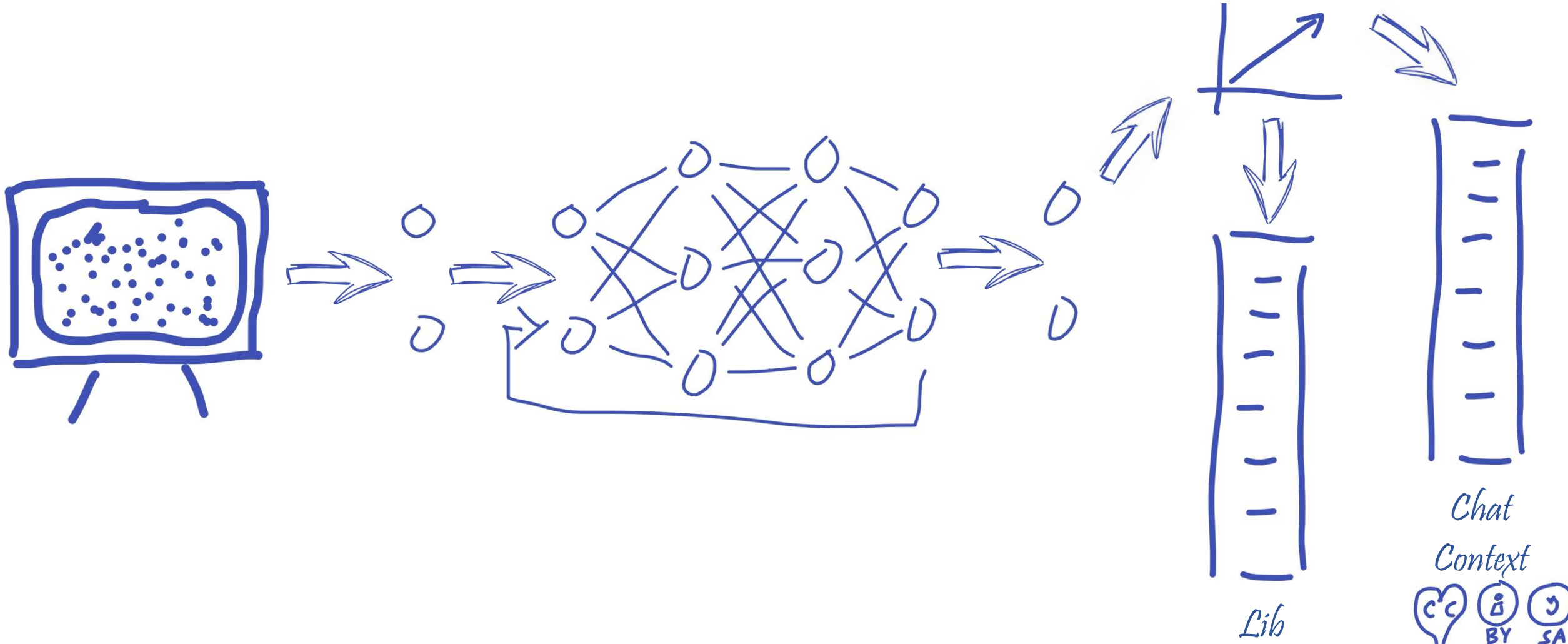




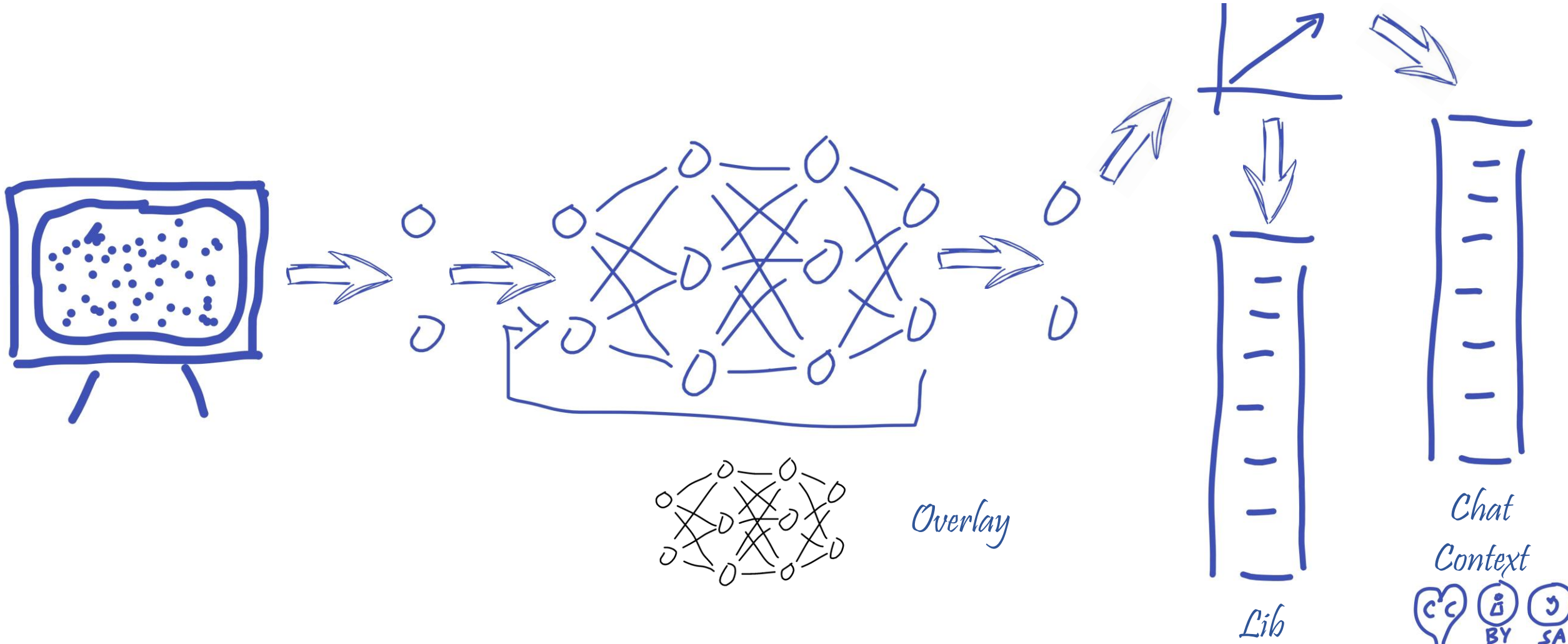
Stable Diffusion



Stable Diffusion



Stable Diffusion



Ways to Shrink Large Models

Distill

Quantize

SLM & RAG

Train dedicated - not all purpose

Refine, e.g. LORA

Main Tasks Libraries Languages Licenses Other

Models 2,803,017 Filter by name

Full-text search Inference Available Sort: Trending

Tasks

- Text Generation Any-to-Any Image-Text-to-Text Image-to-Text Image-to-Image Text-to-Image Text-to-Video Text-to-Speech + 44

Parameters



Libraries

- PyTorch TensorFlow JAX Transformers Diffusers sentence-transformers Safetensors ONNX GGUF Transformers.js MLX MLX + 42

Apps

Qwen/Qwen3.6-35B-A3B Image-Text-to-Text · 36B · Updated 5 days ago · 335k · 977

tencent/HY-Embodied-0.5 Image-Text-to-Text · 4B · Updated 6 days ago · 1.66k · 880

unsloth/Qwen3.6-35B-A3B-GGUF Image-Text-to-Text · 35B · Updated about 3 hours ago · 816k · 525

baidu/ERNIE-Image Text-to-Image · Updated 3 days ago · 4.14k · 483

tencent/HY-World-2.0 Image-to-3D · Updated 4 days ago · 476

MiniMaxAI/MiniMax-M2.7 Text Generation · 229B · Updated about 5 hours ago · 314k · 988

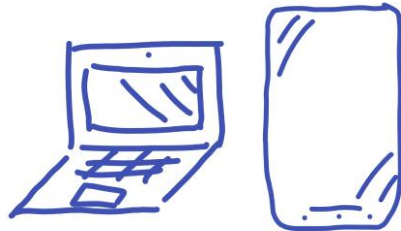
OBLITERATUS/gemma-4-E4B-it-OBLITERATED Text Generation · 8B · Updated about 12 hours ago · 50.7k · 363

baidu/ERNIE-Image-Turbo Text-to-Image · Updated 3 days ago · 5.2k · 227

Experimented on...



New Intel Laptop, no CUDA
Apple Mac Pro M1



Old Intel Laptop, no CUDA
New, Performant Android



Poor, Old Android (2018)

Minstral-3-3B-Instruct-2512-Q4_K_M (2.99 GB)

Phi-4-mini-instruct-Q4_K_M (2.48 GB)

q4_0-orca-mini-3b (1.98 GB)

Llama-3.2-3B-Instruct-Q4_K_S(1.93 GB)

LFM2.5-1.2B-Instruct-Q8_0 (1.25 GB)

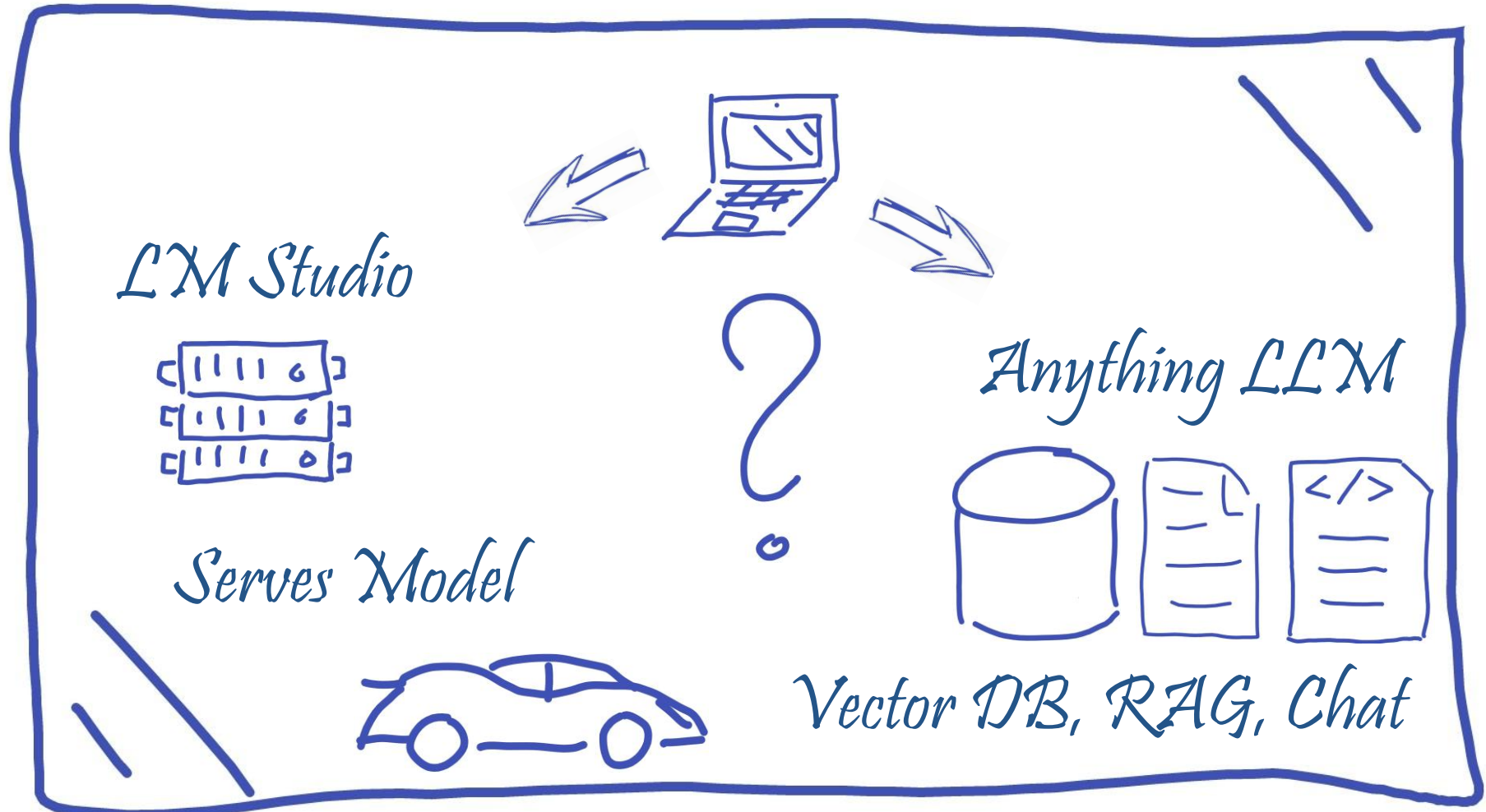
gemma-4-26b-a4b-it (19 GB)

gpt-oss-20b (12 GB)

Minstral-3-14B-Instruct-2512 (9GB)

Qwen3-1.7B-GGUF (1.7 GB)

Recommended Start Setup



Product range >

Product range > PC components > Graphics card > HPE NVIDIA H200 NVL 141GB PCIe Accelerator for

PC components

Air cooling

Barebones

Cases

Expansion cards

Graphics card

Motherboards

Optical disc drive

PC component accessories

Power Supply PC

Processors

RAM

RAM Model-specific

Storage

Water cooling





1 Image


105709.-
HPE NVIDIA H200 NVL 141GB PCIe Accelerator for


Ratings
★★★★★ 1

Brand
[More from HPE](#)

 Currently out of stock
 [Notify me when this product is available](#)

 Add to cart

 Compare

 Add to watch list

Free shipping

Product range >

Product range > PC components > Graphics card > HPE NVIDIA H200 NVL 141GB PCIe Accelerator for

PC components

Air cooling

Barebones

Cases

Expansion cards

Graphics card

Motherboards

Optical disc drive

PC component accessories

Power Supply PC

Processors

RAM

RAM Model-specific

Storage

Water cooling





1 Image


105709.-
HPE NVIDIA H200 NVL 141GB PCIe Accelerator for


Ratings
★★★★★ 1

Brand
[More from HPE](#)

 Currently out of stock
 [Notify me when this product is available](#)

 Add to cart

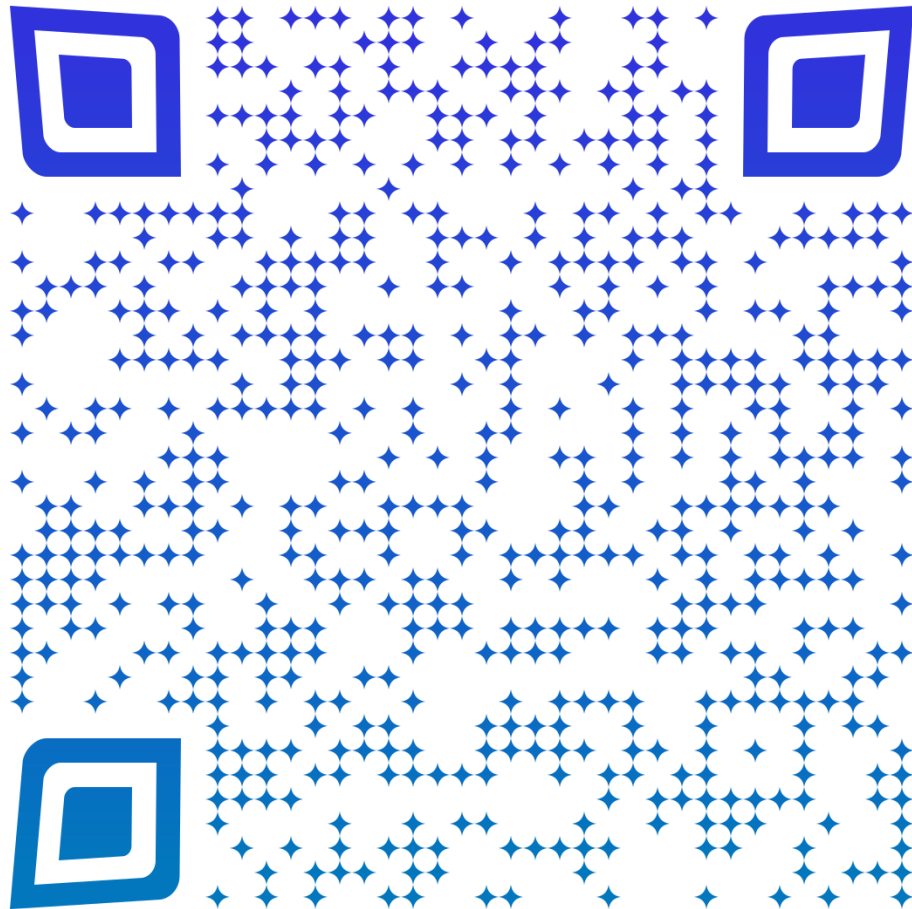
 Compare

 Add to watch list

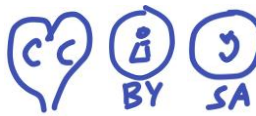
Free shipping

Link to downloads

<https://gitlab.com/danilo.biella/llm-unplugged> -> agile



- 2020-Scrum-Guide-US.pdf → RAG
- LunarBaseExercise.pdf → Exercises 2
- LunarBaseExercise.txt
- LunarBaseRequirements.pdf → RAG
- LunarBaseRequirements.txt
- ManifestoofAgileSoftwareDevelopr → RAG
- agilemanifesto.txt
- basic-prompts.txt → Exercises 1
- muloe-transcript.odt
- scrumguide.txt
- system-prompt.txt → Sys Prompt



Is there redundancy in the principles of the agile manifesto?

The Agile Manifesto values clarity, simplicity, and consistency. Upon reviewing the Agile Manifesto, it appears that the principles are not redundant. The values are:

1. Individuals and interactions
2. Working software
3. Customer collaboration over contract negotiation
4. Responding to change

These values are concise and distinct, emphasizing the importance of people, collaboration, and adaptability in software development.

[View Citations](#)

The principles, not the values.



Upon reviewing the Agile Manifesto principles, I notice that there are some similarities and overlap

Enter your prompt



Upon reviewing the Agile Manifesto principles, I notice that there are some similarities and overlap between them. Here are a few examples:

- Principle 1: Our highest priority is to satisfy the customer through early and continuous delivery of valuable software. This principle is similar to Principle 5: Business people and developers must work together daily throughout the project. Both emphasize the importance of collaboration and customer satisfaction.
- Principle 2: Welcome changing requirements, even late in development. This principle is similar to Principle 1: Our highest priority is to satisfy the customer through early and continuous delivery of valuable software. Both emphasize the importance of adapting to changing requirements.
- Principle 6: Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale. This principle is similar to Principle 3: Business people and developers must work together daily throughout the project. Both emphasize the importance of delivering working software in short iterations.



However, upon closer inspection, I don't see any significant redundancy between the principles.

Enter your prompt



shorter timescale. This principle is similar to Principle 3: Business people and developers must work together daily throughout the project. Both emphasize the importance of delivering working software in short iterations.

However, upon closer inspection, I don't see any significant redundancy between the principles. Each principle presents a unique idea or emphasis that complements the others.

That being said, some people might argue that Principle 5 and Principle 6 are related, as they both emphasize the importance of collaboration and delivering working software. However, Principle 5 focuses on the importance of collaboration in general, while Principle 6 focuses on the specific aspect of delivering working software frequently.

Overall, while there may be some similarities and overlap between the principles, I don't see any significant redundancy.

[View Citations](#)

Enter your prompt



Outro-Duction - Intro-Spection

AI Integration has become a priority
for nearly every business

-- Lize Raes

