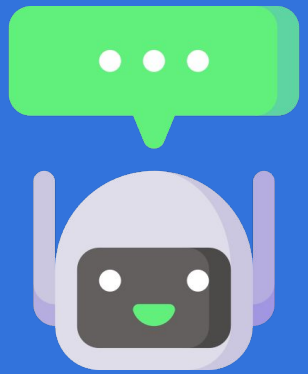
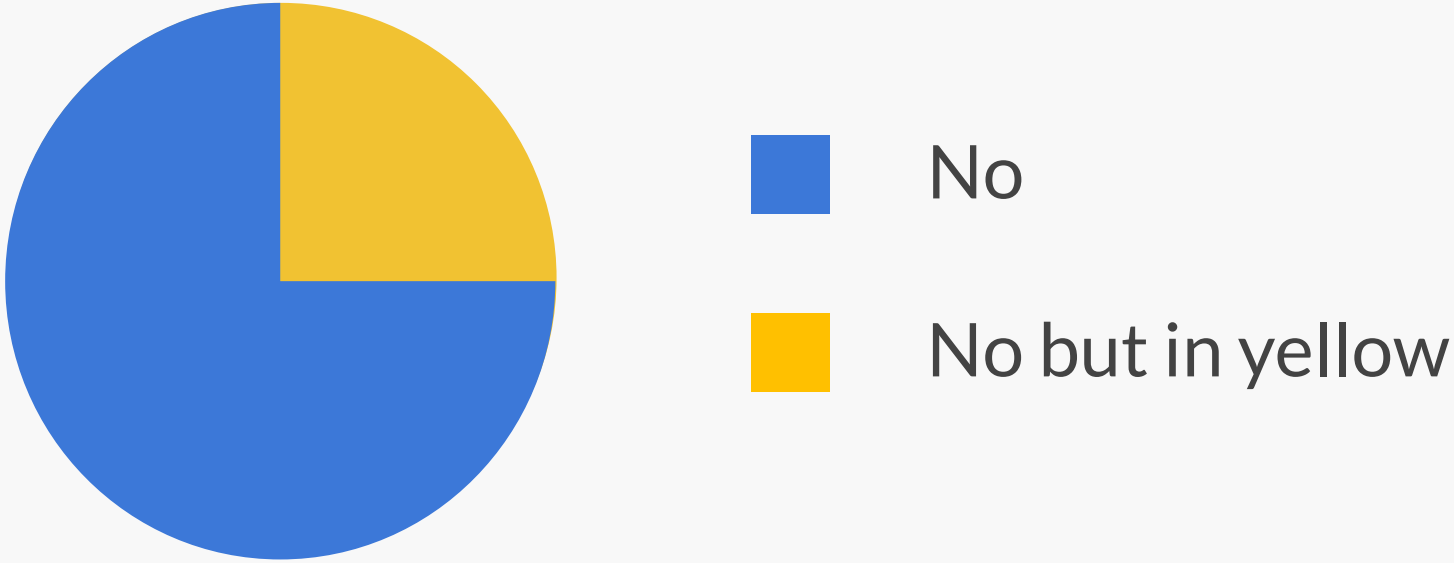
A Practical Guide to Fast and Private LLMs



Do I Even Care?



Why?

It’s **MY** fucking data!

Ali Farhat
Posted on 1 Aug 2025 • Edited on 1 Jan

🩸ChatGPT Privacy Leak: Thousands of Conversations Now Publicly Indexed by Google

#ai #privacy #gdpr #chatgpt

Google has indexed thousands of ChatGPT conversations — exposing sensitive prompts, private data, and company strategies. Here's what happened, why it matters, and how you can protect your AI workflows.

r/OpenAI 2h • arstechnica.com

OpenAI slams court order to save all ChatGPT logs, including deleted chats

47 12

r/LocalLLaMA 1h • arstechnica.com

After court order, OpenAI is now preserving all ChatGPT and API logs

95 11

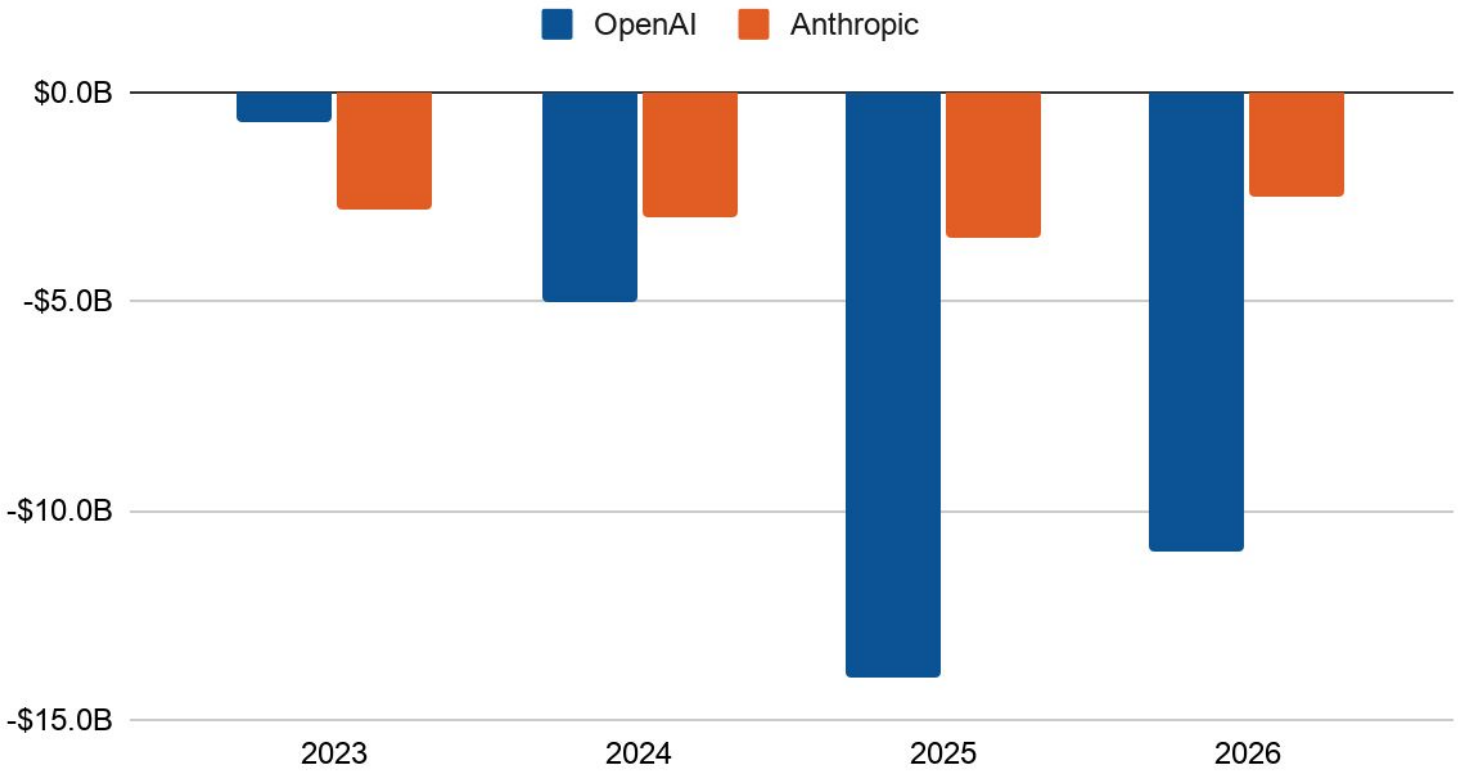
Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs

github.com/leak-llm/leak-llm.github.io

Why?

It’s **MY** fucking money!

Yearly Earnings



How?

It’s **MY** fucking LLM!

Public

ChatGPT / Claude / Mistral

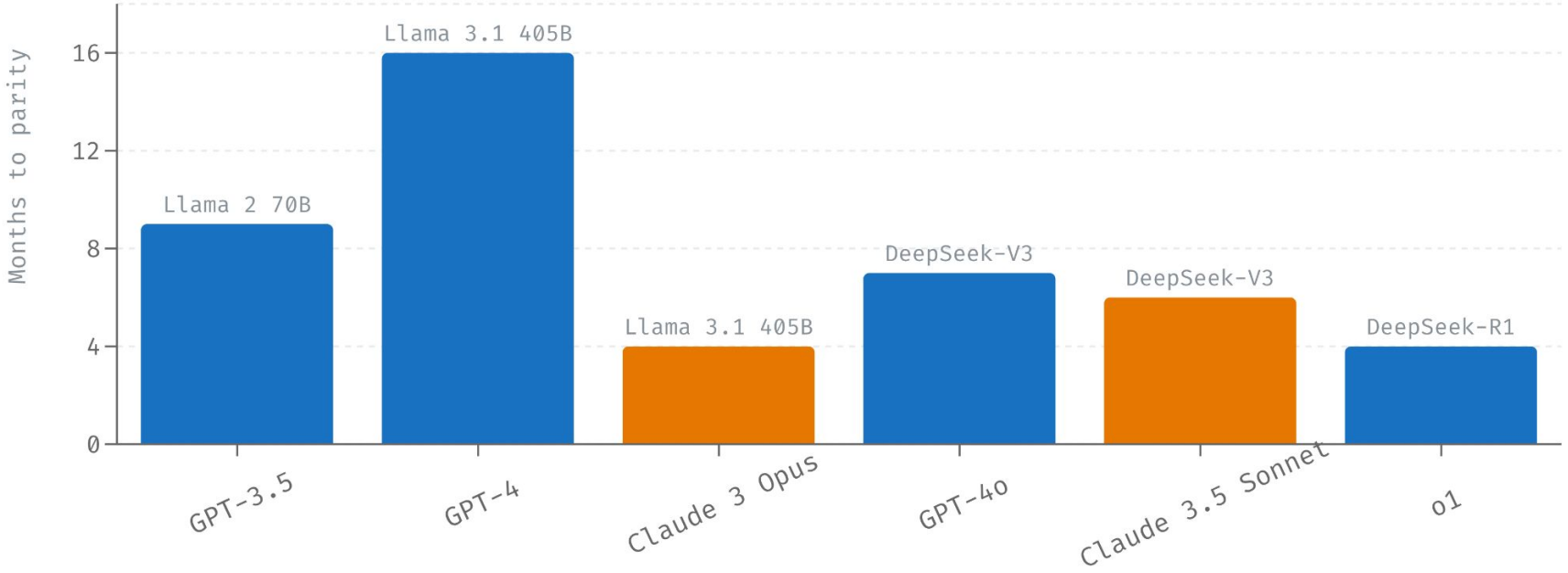
Hosted

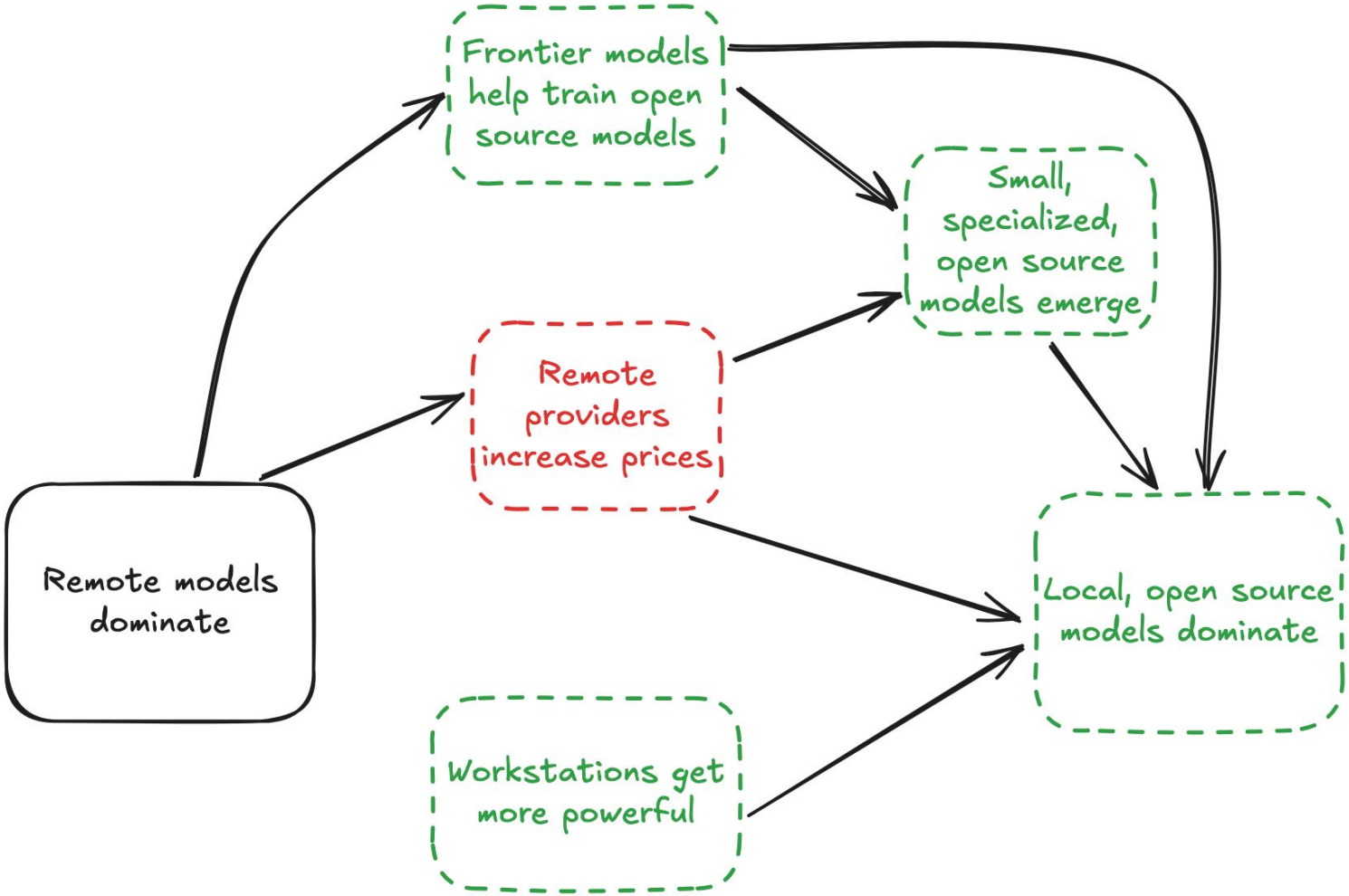
Bedrock / Copilot / Vertex

Local

Ollama / vLLM / TGI

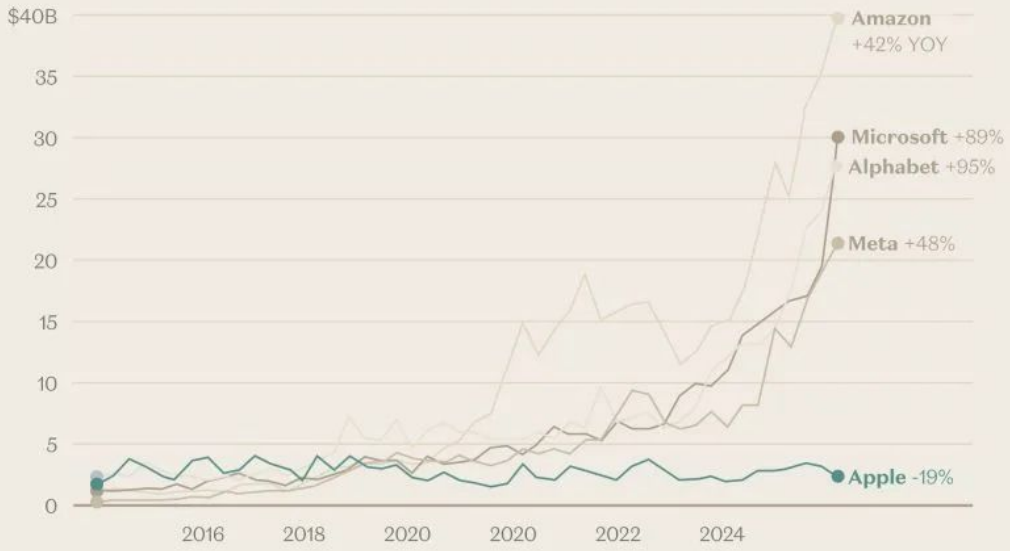
Months to open source parity with frontier models




Frontier models help train open source models
Small, specialized, open source models emerge
Remote providers increase prices
Remote models dominate
Local, open source models dominate
Workstations get more powerful

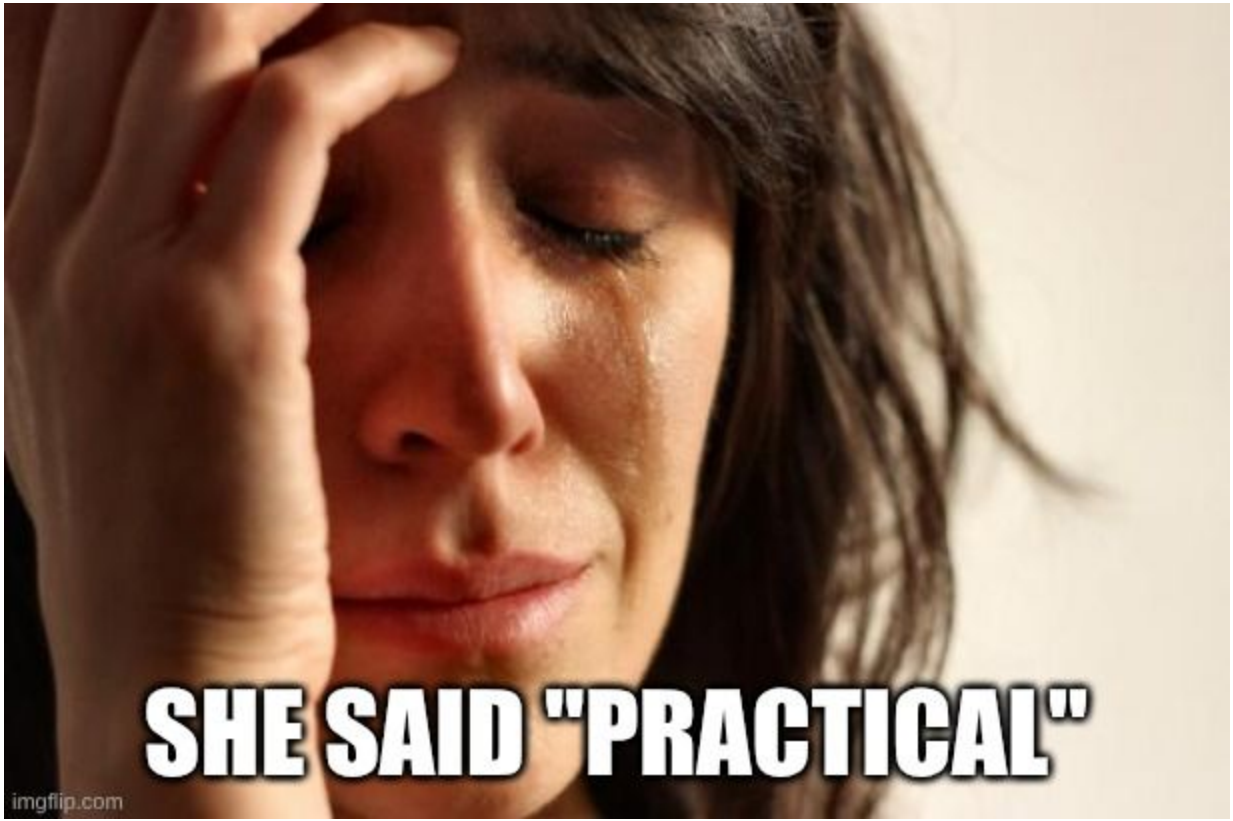
Apple on Capex: 'Nah, we're good'

Standardized quarterly capital expenditure



Source: FactSet, Snacks (2/9/26)

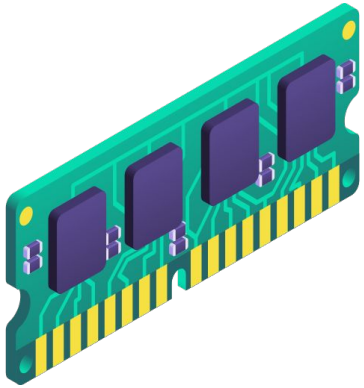
A16Z


SHE SAID "PRACTICAL"
imgflip.com

What?

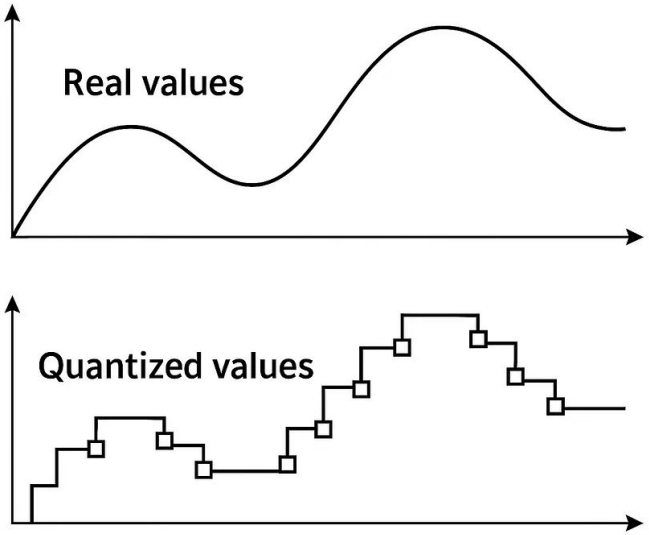
It’s **MY** fucking stack!

Part 1: The Constraint



Memory,
Not Compute

Part 1: The Constraint

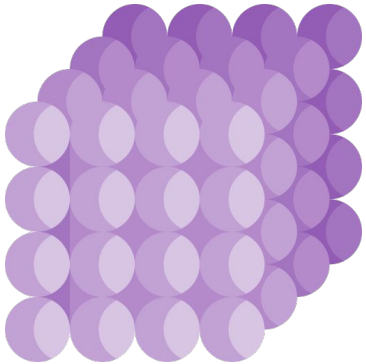


Part 2: The Architecture

Small / Specialist

Large / Generalist

Part 2: The Architecture



Dense

Mixture-of-Experts

Qwen3-35B-A3B-Q4_K_M.gguf

Part 3: The Hardware

7B	5 GB
13B	8 GB
35B	20 GB
70B	40 GB
120B	70 GB

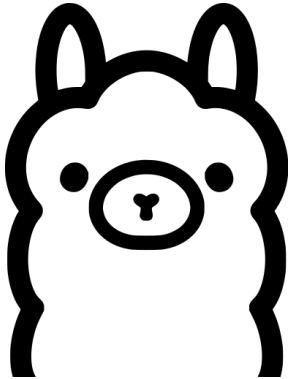
Part 3: The Hardware

RTX 4070
(12GB VRAM)

RTX PRO 6000
(96GB VRAM)

Mac Studio M4 Max
(128GB RAM)

Part 4: The Software



What is Lena like?

Lena

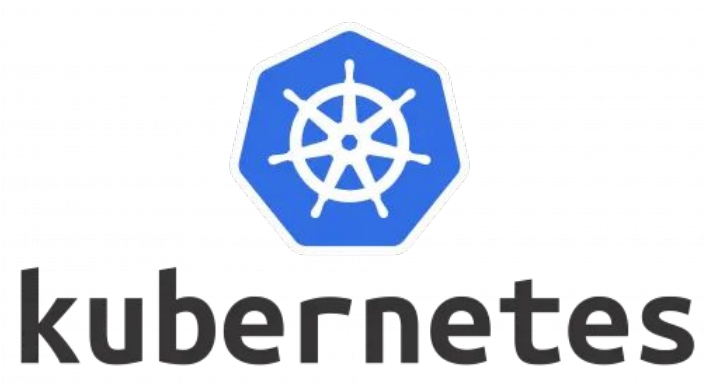
Lena has

Lena has very

Lena has very weird

Lena has very weird hair

Part 4: The Software



Memory

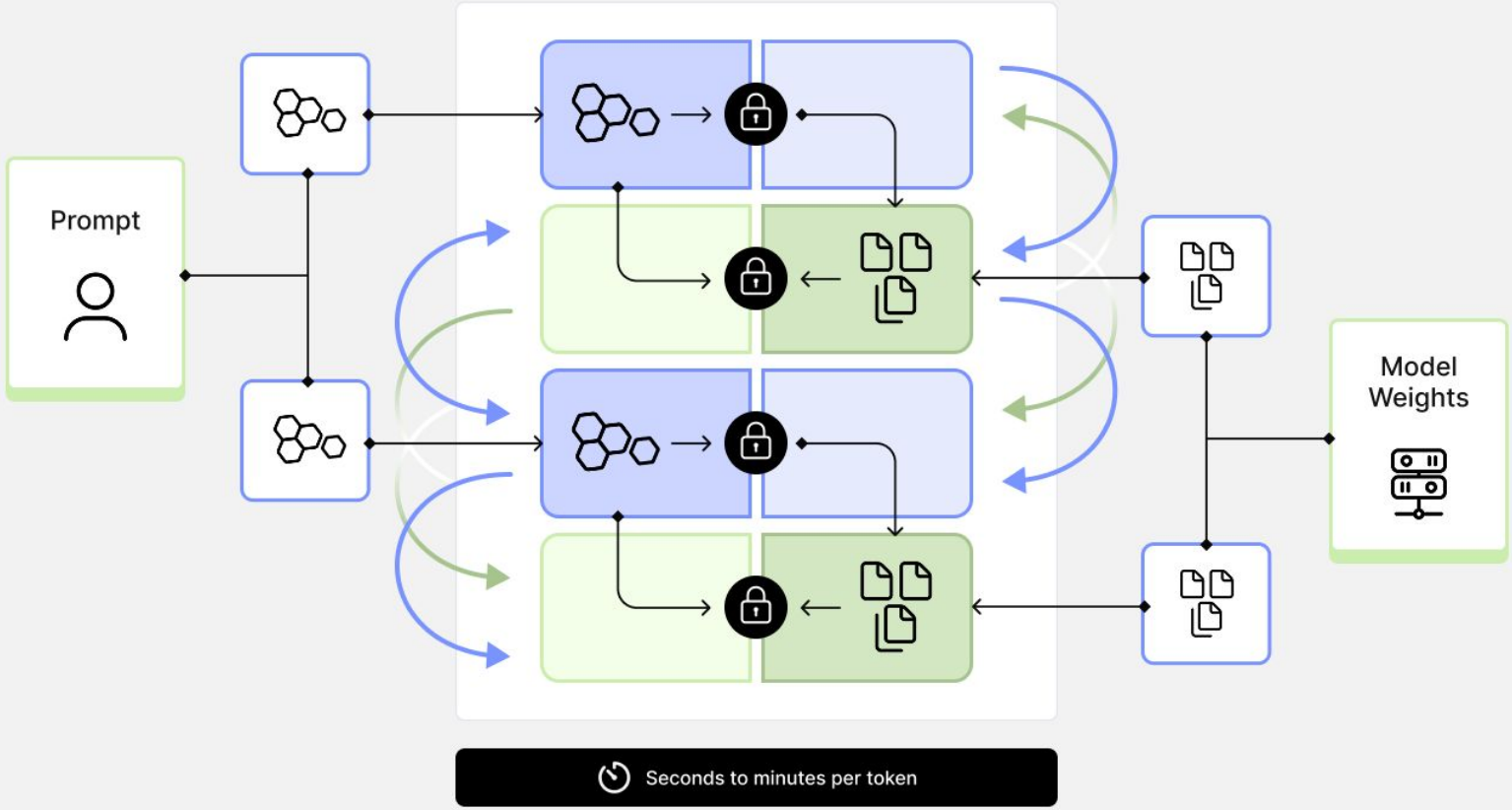
Quantization

MoE

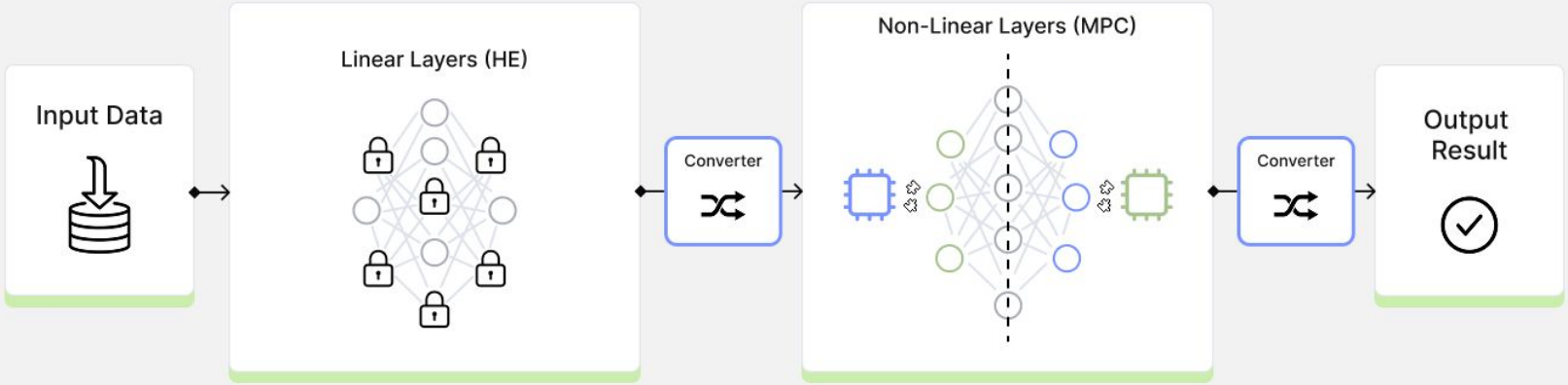
Dare to Go Small

Thanks!

bespinian.io


Prompt
Model Weights
Seconds to minutes per token

Fully Homomorphic Encryption for LLM Inference



Fully Homomorphic Encryption for LLM Inference