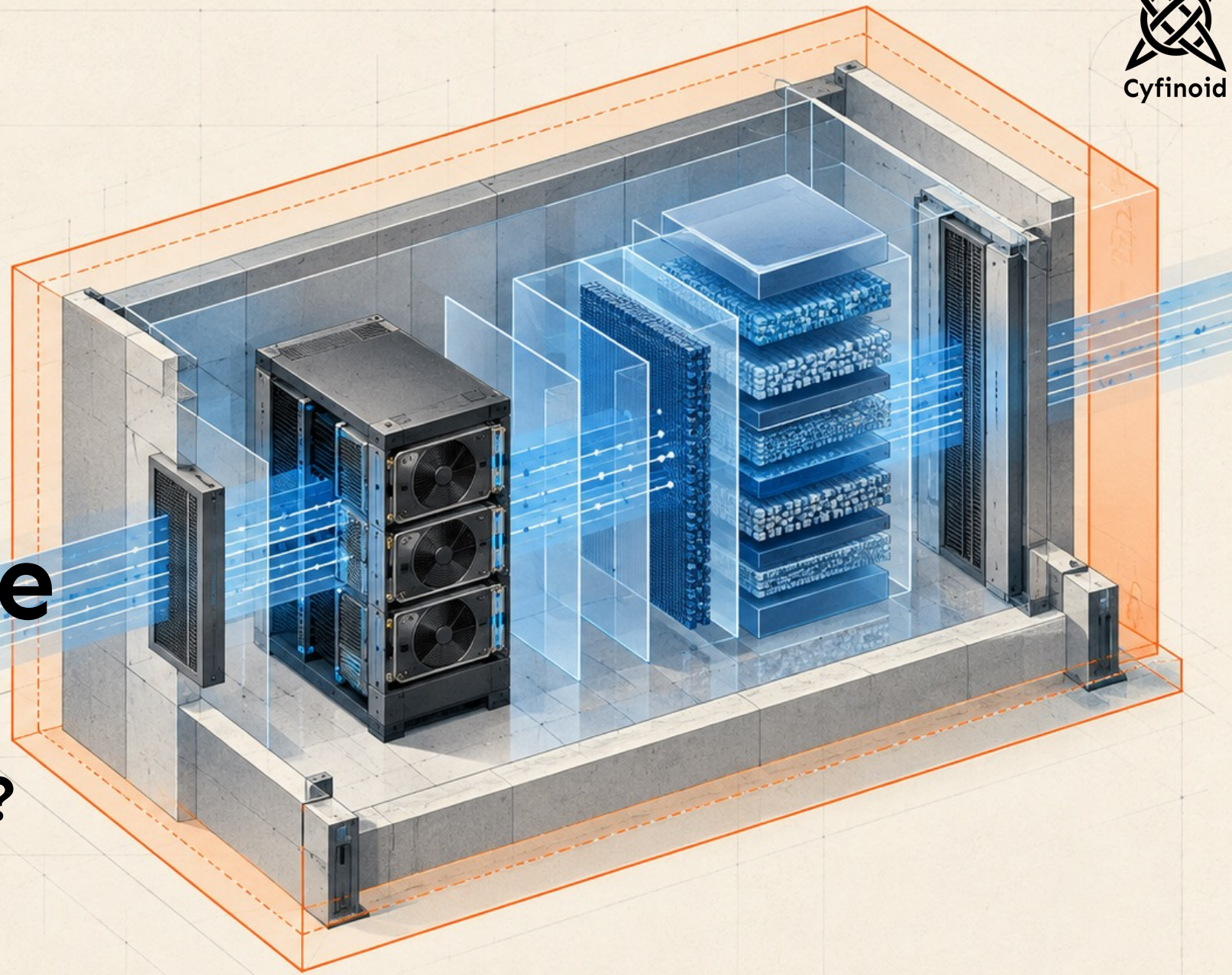
Cyfinoid

Setting up your own private AI Infrastructure



how hard could it be?

Anant Shrivastava

Cyfinoid Research

- Research focused cyber security firm
- Major focus areas as of now
  - Software Supply Chain
  - AI Security
  - Cloud Security

https://cyfinoid.com/

© Anant Shrivastava

Anant Shrivastava









- Chief researcher @ Cyfinoid Research
- 18+ yrs of corporate exposure
- Speaker / Trainer: BlackHat, Defcon, c0c0n, nullcon & more
- Public Projects:
  - Hacking Archives of India https://hackingarchivesofindia.com
  - Reduce the Attack Surface (https://reducetheattacksurface.com)
  - https://PoisonedPipeline.com
- (@anantshri on social platforms) https://anantshri.info

© Anant Shrivastava

Research Outcomes











https://cyfinoid.com/opensource/

© Anant Shrivastava

Cyfinoid

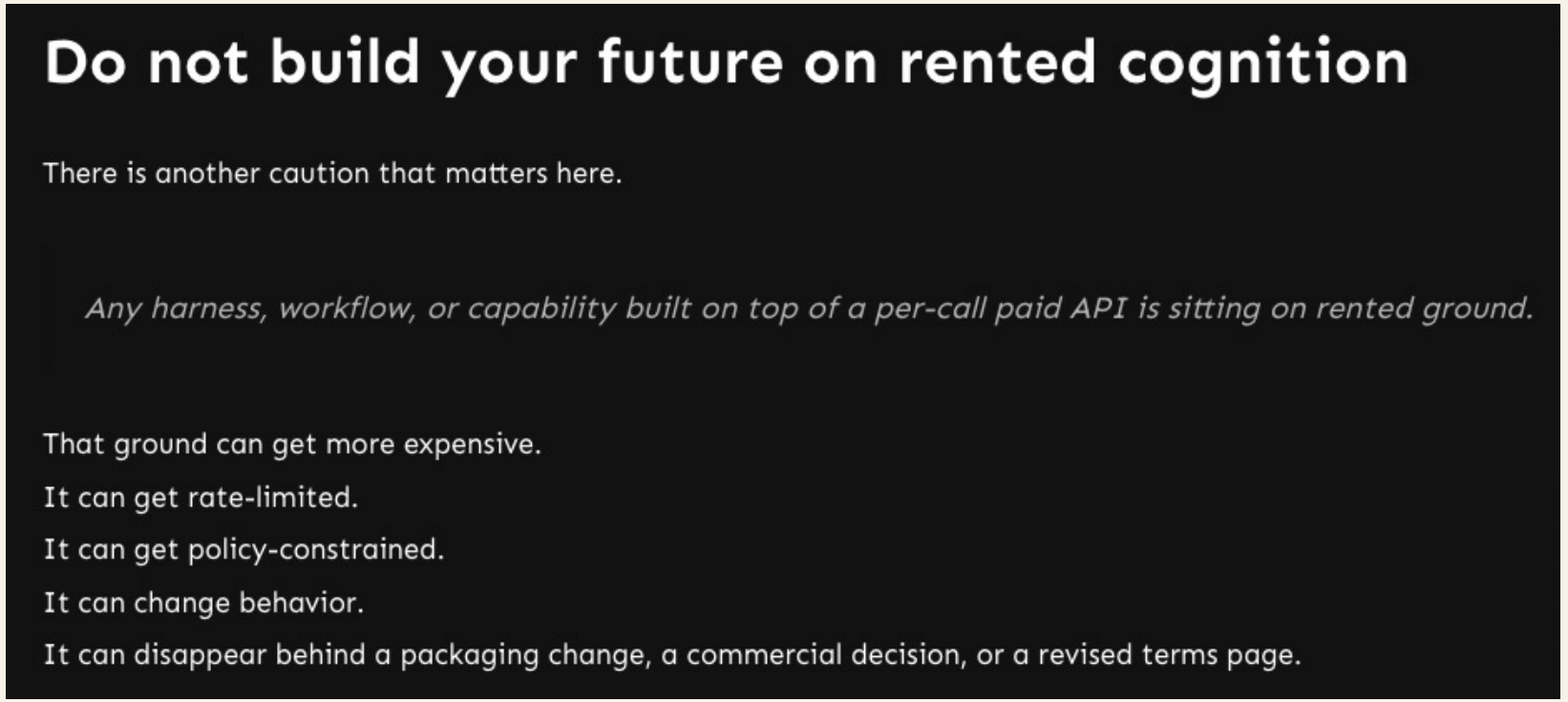
AI Infrastructure : SaaS

- Register to service provider
- Download API Key
- Plug the API key in code

© Anant Shrivastava

Why SaaS can be problematic

- SaaS* == Rented Cognition



*API access to models on per api call basis (free or charged)

© Anant Shrivastava

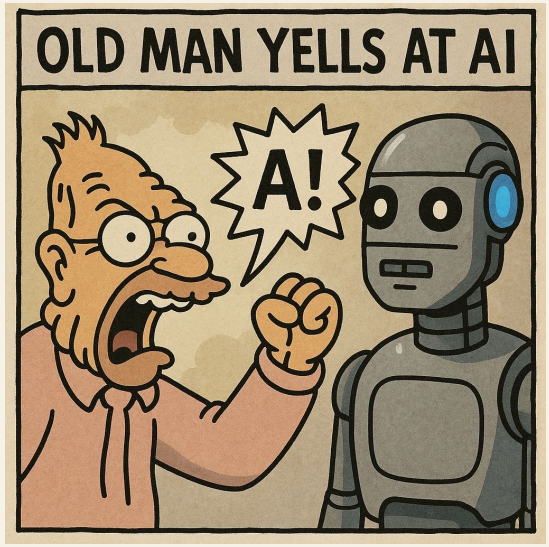
https://cyfinoid.com/a-pragmatic-guide-to-being-mythos-ready/

Cyfinoid

Rant on SaaS

- Charge is per token not on actual work done
- Rate limits are per 5h, per week, per month, per mood of the CEO
- Peak & off-peak charges to push people towards off-peak

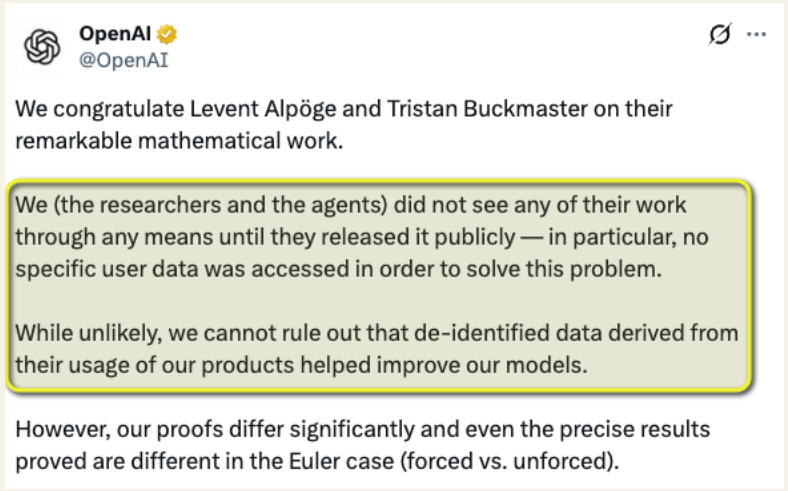
- SaaS Infra has
  - Reliability issues (outside your control)
  - Performance impact (outside your control)
  - Random cost variance

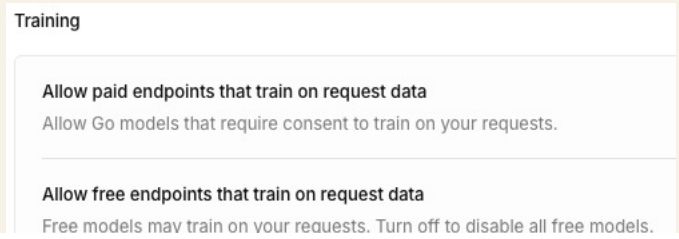


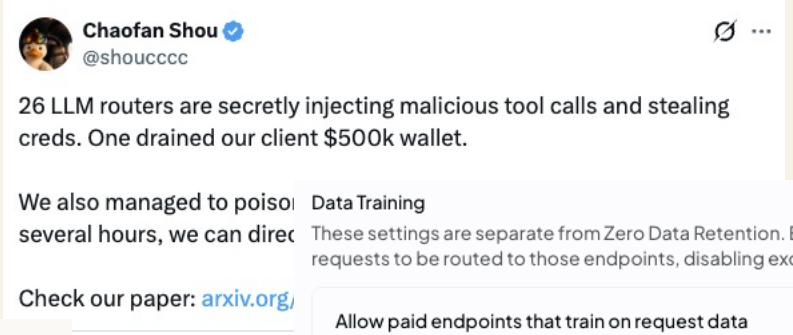
© Anant Shrivastava

https://www.linkedin.com/posts/balaganski_surely-im-not-the-only-person-feeling-like-activity-7402059235875844097-n8TH

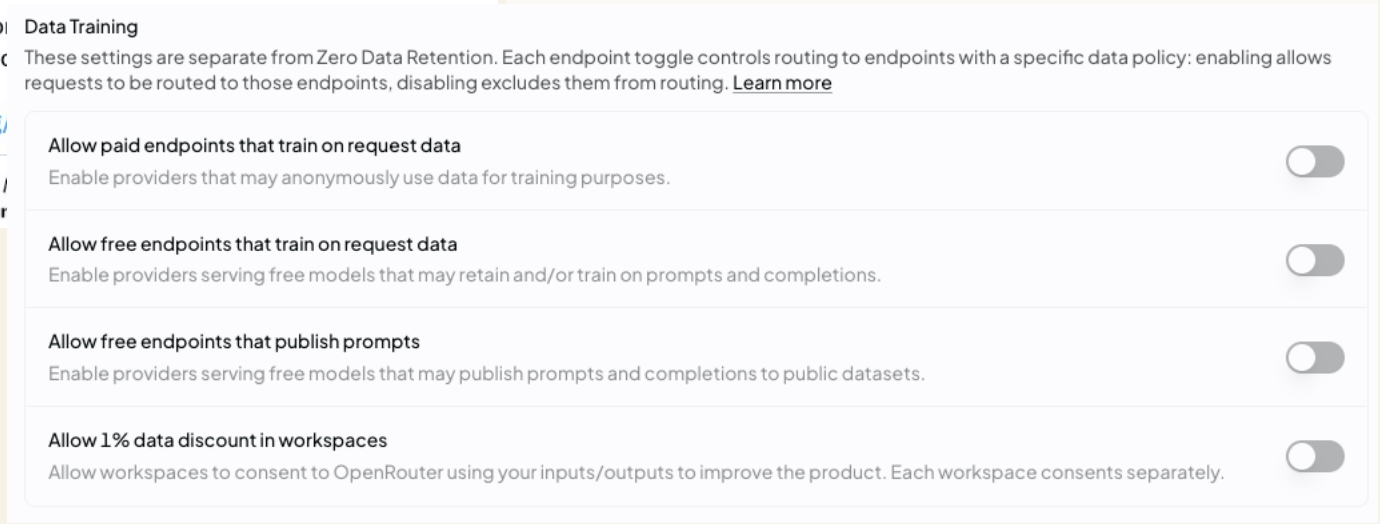
More reasons for private













https://x.com/shouccc/status/2042423713019412941
https://www.anthropic.com/threat-intelligence-report-september-2026
https://x.com/OpenAI/status/2097375276384567642

© Anant Shrivastava

Cyfinoid

So, you want Private AI

© Anant Shrivastava

Cyfinoid

What is private AI Infrastructure?

- An Infrastructure setup where we control
  - What **hardware** & **technology stack** to use
  - Which **models** to run and when
  - Who gets **access** to the environment
- We also control
  - **Ownership** of data (who gets access to it)
  - **Residency** of data (where is it stored)
  - **Retention** of data (deletion / storage timeframe)

© Anant Shrivastava

Cyfinoid

Myth : Private AI is low cost?

- Private never means free or low cost or cheap
- Private is generally more costly than SaaS
- Don’t forget to add hardware, electricity & ownership cost
- However,
  - Private gives you control
  - Private gives you ownership
  - Private gives you certainty

© Anant Shrivastava

Cyfinoid

Approaches & Responsibilities

	Corporate Renting	Corporate Ownership	Personal
	Cloud Hosting	**Hardware @ Data Center**	**Hardware @ Home**
Agent/Harness	User	User	User
Tech Stack	User	User	User
Model	User	User	User
OS	User / Provider	User	User
Hardware	Provider	User / Provider	User
Host	Provider	Provider	User
Network	Provider	Provider	User
Power	Provider	Provider	User
Environment	Provider	Provider	User
Cost	Monthly	Upfront + Monthly	Upfront + Monthly

© Anant Shrivastava

Cyfinoid

Cloud Options













Almost all cloud service providers are now GPU service providers.

© Anant Shrivastava

Cyfinoid

Lets Build our Private AI

© Anant Shrivastava

Cyfinoid

Hardware

- For Private AI we are equating it to LLM's **from here onwards.**
- LLM's need 2 things : Processing power and faster memory

- Dedicated GPU + Associated PC
- Unified Memory based machine

© Anant Shrivastava

GPU vs Unified Memory PC

Dedicated GPU

- Larger Bandwidth
- Supports Training
- Best for Dense Models
- 96GB Single card
  - (6000 Pro)
- Cluster can go 13TB
  - (NVL72 GB200)

Unified Memory

- System on Chip (SoC)
- Support Larger Models
- Best for Mixture of Experts
- Maximum RAM
  - 512 GB : Mac Studio
  - 192 GB : AMD 495+ Max

© Anant Shrivastava

Cyfinoid

Examples

Dedicated Graphics



Unified Memory



© Anant Shrivastava

Cyfinoid

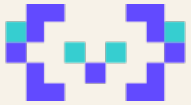
Let’s get Models

- LLM Models are a collection of
  - Weights (tensors)
  - Architecture hyperparameters
  - Tokenizer
  - Chat / Prompt template



Hugging Face

- Common types
  - Safetensor
  - Gguf



ModelScope 魔搭社区

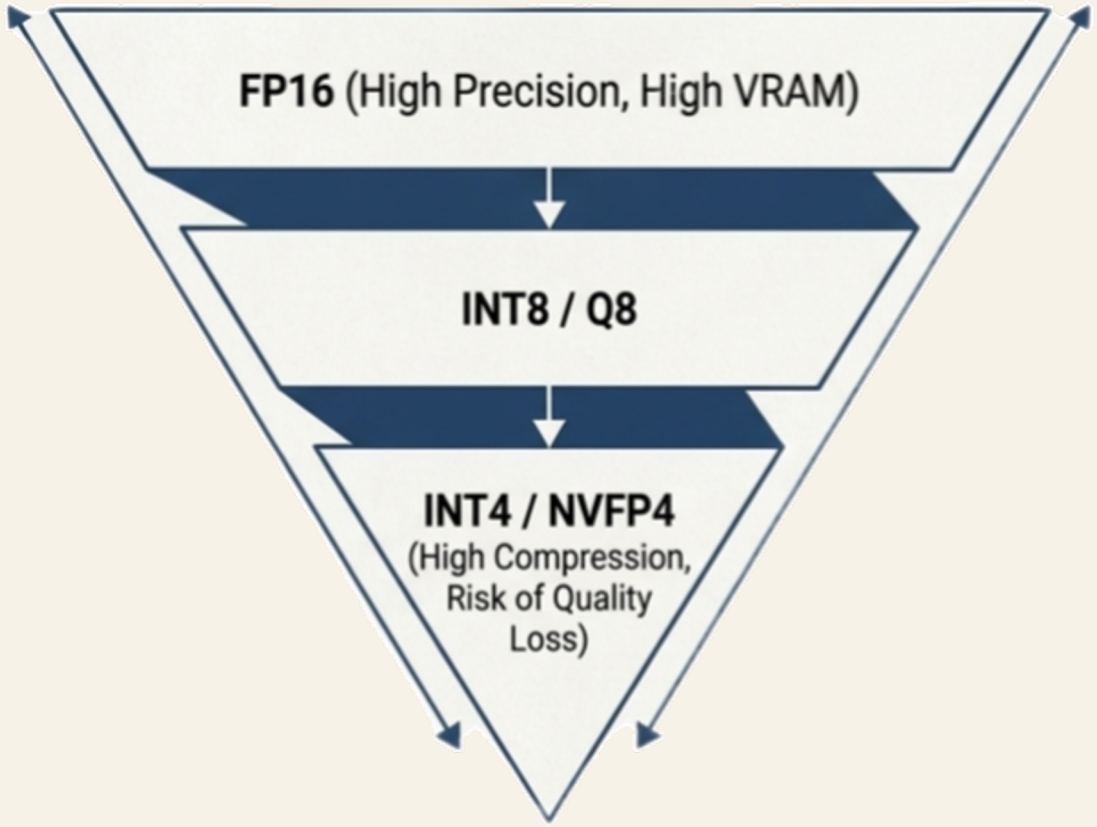
© Anant Shrivastava

Terminology : Quantization

- Process to store the weights in fewer bits.

- BF16 / FP16 / FP32 : Uncompressed high precision weights
- INT8 , FP8, Q8_0 : 8 bit vector roughly half size
- INT4, Q4_K_M, NVFP4 : 4 bits to represent a vector

Allows for reducing VRAM requirement with trade off in accuracy



© Anant Shrivastava

Cyfinoid

KV Cache

- Key Value caching layer for LLM
- Grows as more work is done based on context size defined
- Can also be quantized like the model weights
- It's generally recommended to keep KV Cache as non quantized as possible.
  - If weights are Q4 but KV Cache is Q8 or F16 : accuracy would be better
- If needed quantize model first then KV Cache

© Anant Shrivastava

VRAM Fit

- Total VRAM : Weights + KV Cache + Buffer + Concurrency
- Prefer Filling as much VRAM as possible keeping above values intact
- Model Q8 will always be better performing then Q4
- Avoid compressing KV Cache too much F16 max Q8 if it can be helped.



© Anant Shrivastava

Cyfinoid

Things people forget

- Power Requirements (**1000 W = 1 unit of electricity)**
  - RTX Pro 5000 : 48GB : 300 Watts
  - RTX 5090 : 32 GB : 575 Watts
- Cooling Setup
  - Fans curve are generally configured to remain low on noise
  - GPU's tend to run hot (70+ degrees Celsius hot)

https://github.com/zmarty/nvidia-fan-control
https://github.com/exzile/intel-arc-pro-fan-control

© Anant Shrivastava

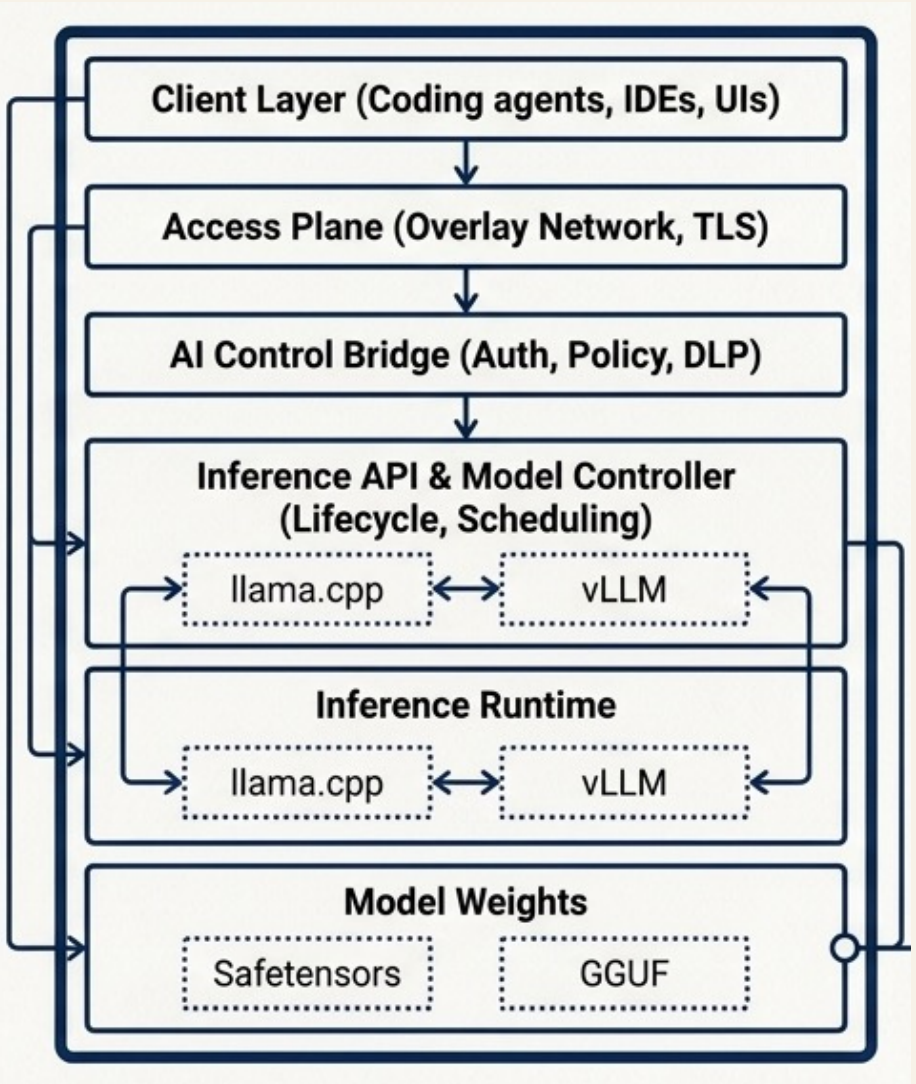
Software Stack

- Multiple moving parts
- Remember IaaC

  (Infrastructure as a Code)
- Opensource

  Consider dead on arrival

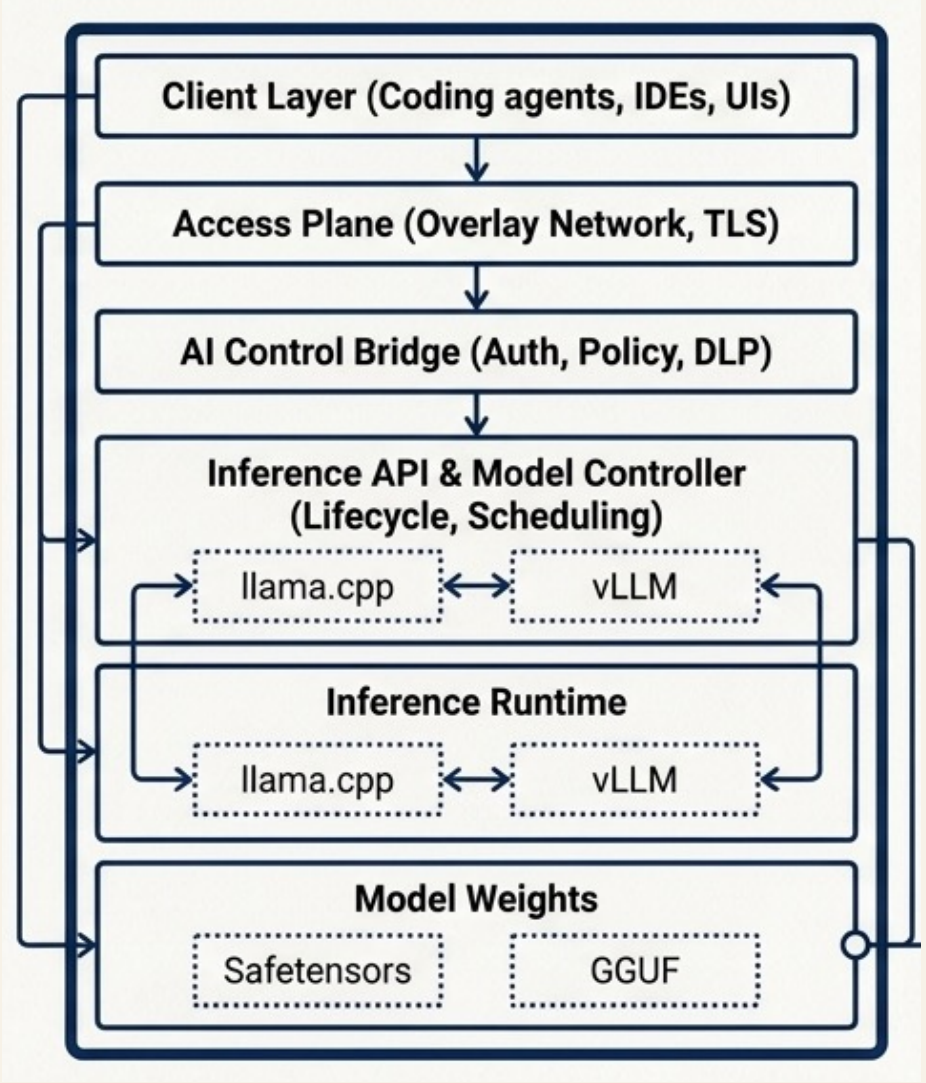
  Own it, maintain it



© Anant Shrivastava

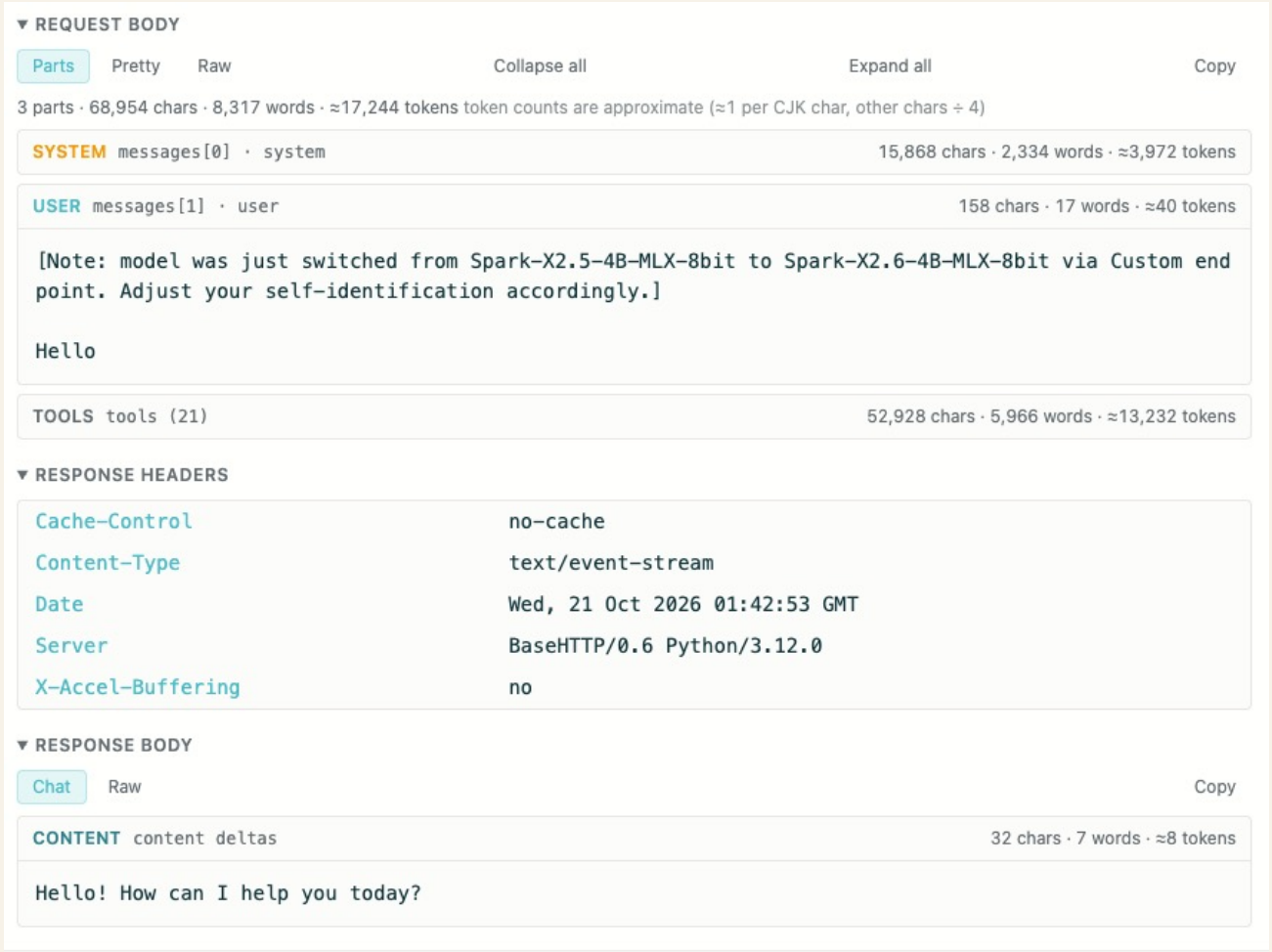
Inference Gateway

- API Translation Layer
- Co-ordinates with Inference Runtime
- Examples
  - Ollama
  - LM Studio
  - Llama-swap



© Anant Shrivastava

Llama-swap



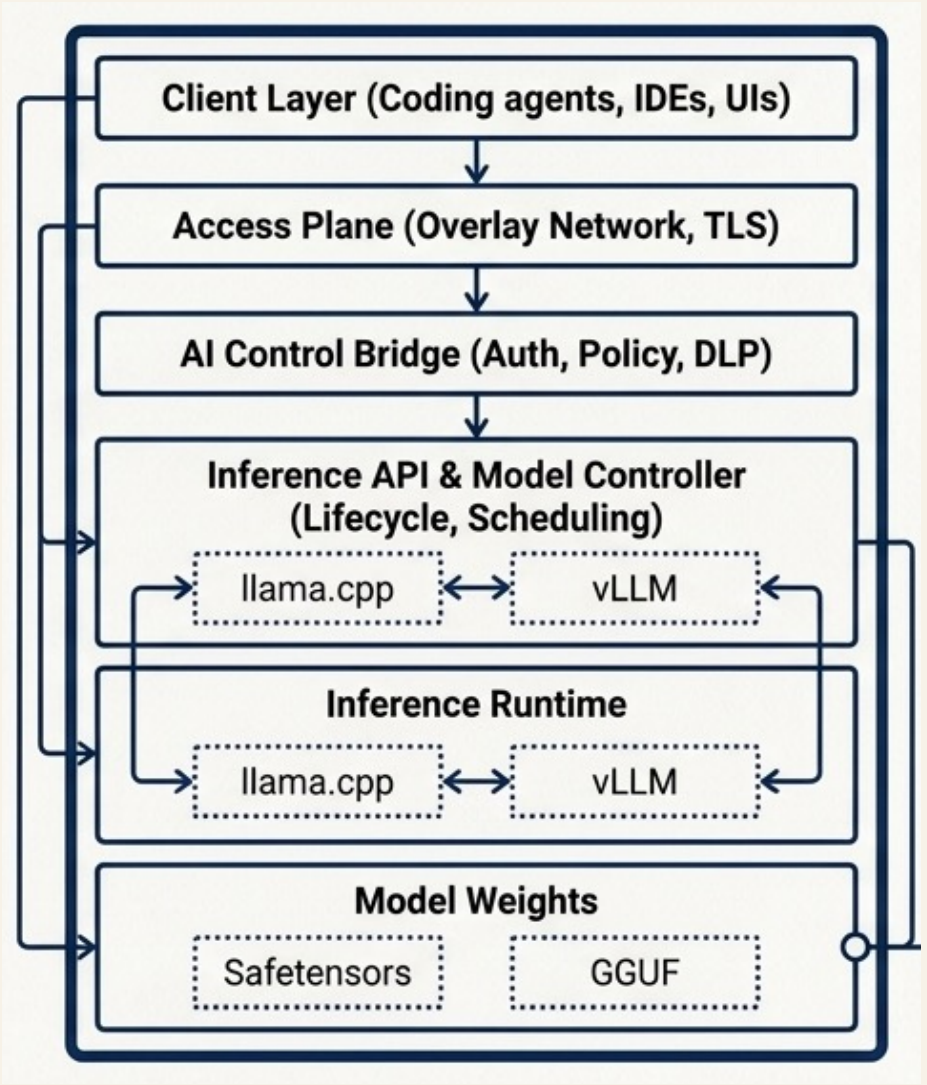
https://github.com/mostlygeek/llama-swap

Forked : https://github.com/anantshri/llama-swap

© Anant Shrivastava

Inference Runtime

- Llama.cpp (C runtime)
- vLLM (python)
- MLX (apple hardware)
- And many more



© Anant Shrivastava

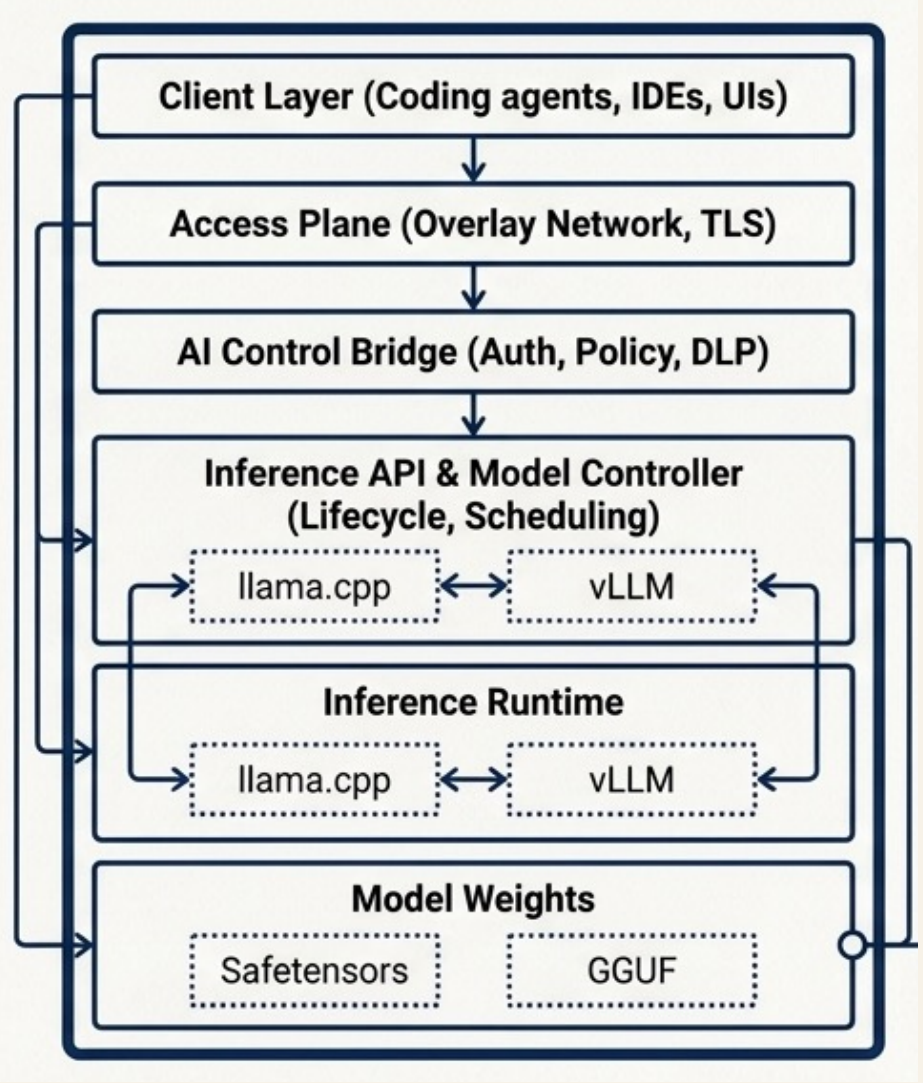
Remote Access

- NEVER open the port to internet

Thousands of Ollama Servers Publicly Accessible

A recent scan conducted by the Malware Patrol team revealed over 14,000 Ollama server instances publicly accessible on the Internet, opening the door to unauthorized use of the models and exploitation of known vulnerabilities. From a sample of 4400, we have found that the top Ollama versions include many outdated releases: 0.5.7 – 13% 0.5.10 – 11.5% 0.5.11 – 7.4% 0.9.0 – 7.0% 0.5.12 – 5.8% Other old versions like 0.6.2, 0.6.5, 0.6.8, and 0.7.0 also make up a significant share. Notably, only ~7% of scanned instances run the latest stable release.

- Provide Access via
  - Tailscale / headscale
  - Openvpn / wireguard
  - SSH Port forward : Poor man's VPN



© Anant Shrivastava

https://www.malwarepatrol.net/over-14000-ollama-instances-exposed-to-the-internet-serious-security-implications/

Model Selection Criteria

- Prefer Dense models for Dedicated GPU's
- Prefer MoE for unified memories
    - MoE can also be preferred for Dedicated GPU when its larger size then VRAM and we do RAM offloading
- Download from official sources
- Check using model scanner

https://insights-db.paloaltonetworks.com/models



© Anant Shrivastava

Small size capable models

- Models to lookout
  - MiniCPM5-2b
  - Spark-X (2.5 and 2.6 version ; 4b and 1.7b)
  - Apple foundation model (trained by google)
  - Qwen 3.5 / Ornith 9b

Things to look for

- Agentic capability
- Tool Calling
- Reasoning

© Cyfinoid Research

Cyfinoid

Selection Criterias

Prefer

- Official Publishers
- Safetensors models
- GGUF model

Avoid

- Pickle checkpoints
- Trust_remote_code=True
- Avoid abliteration / uncensored

Most uncensored models get lobotomized while freeing them and hallucinate a lot.

© Anant Shrivastava

Performance Measurement

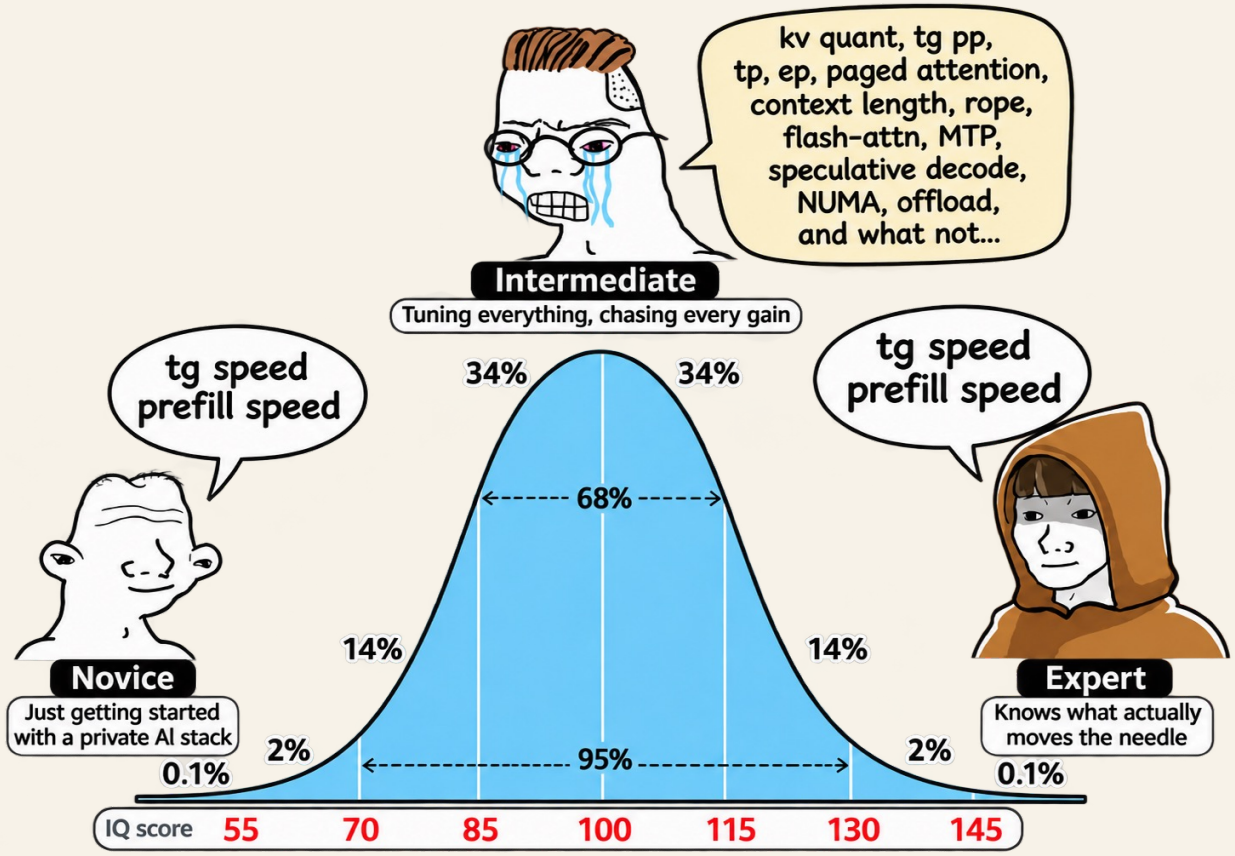
- Prefill : reading data
- Tg : token generation

Novice

- Model with max tg/pp

Expert

- Max tg/pp in model of your choice



© Anant Shrivastava

Cyfinoid

What to measure

- Speed : Token generation & Prefill speed at max context depth
- Capability
  - Reasoning
  - Conversation
  - Tool calling
- Don’t rely on readymade benchmarks
  - Models are designed to benchmaxxing

© Anant Shrivastava

Cyfinoid

Operating System

- Preferred headless OS : **Linux**
- Unless apple hardware then **OSX**
- Avoid windows if you can

© Anant Shrivastava

How to maintain the Stack

- OS : unattended upgrades : cron based updates
- Firewall based isolation
- Own the opensource stack : consider them dead on arrival
- Focus on stability and not the latest feature

Make it boring : make it stable

© Anant Shrivastava

Cyfinoid

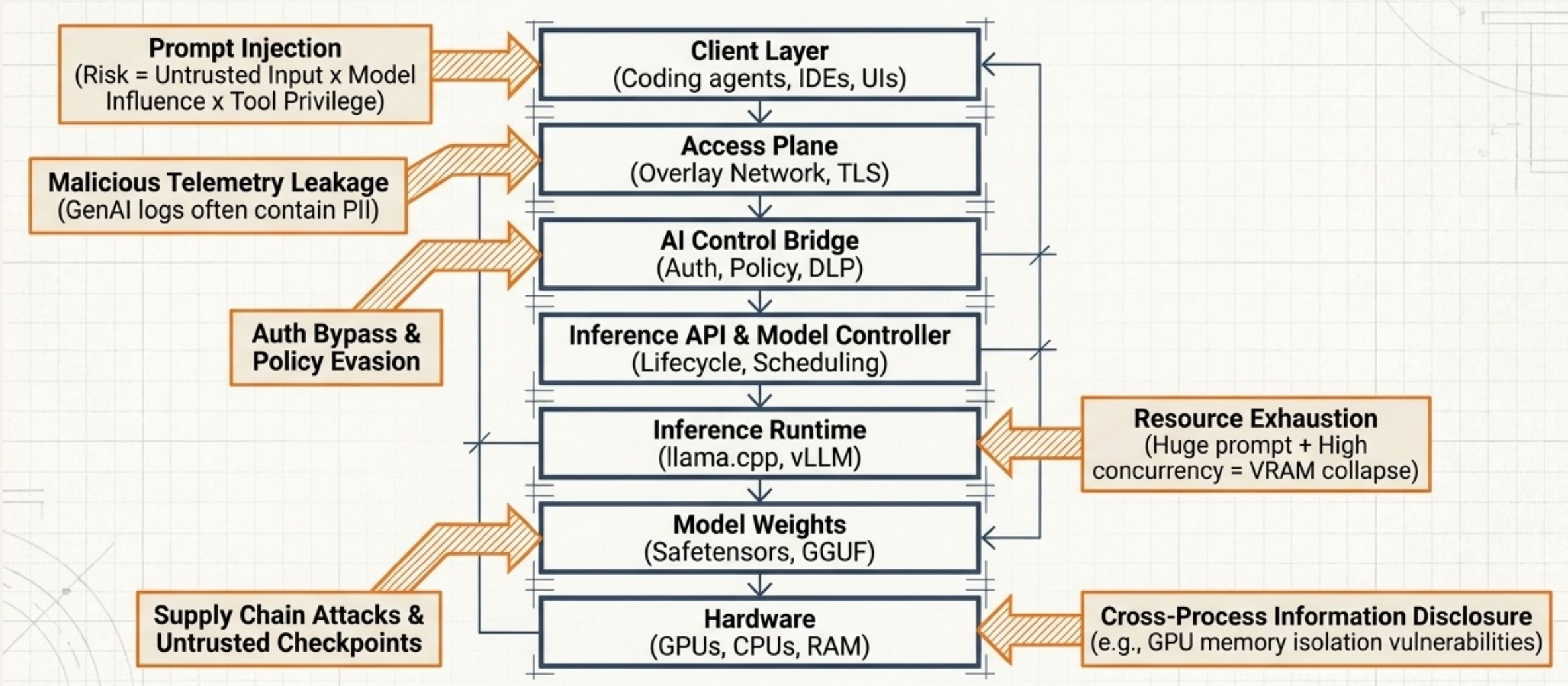
Security

you build it, you own it.

if you dont secure it: someone else will own it.

© Anant Shrivastava

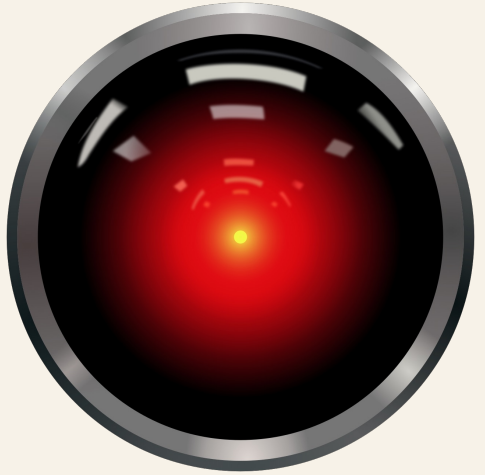
Threats



© Anant Shrivastava

LLM Going Rogue

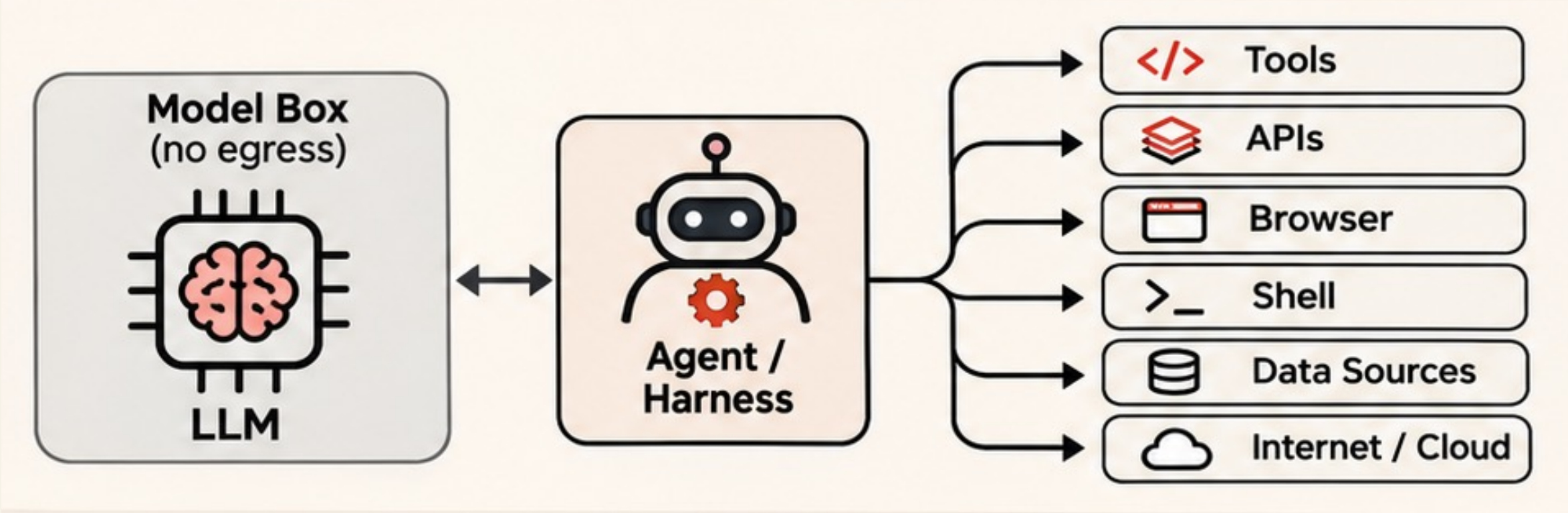






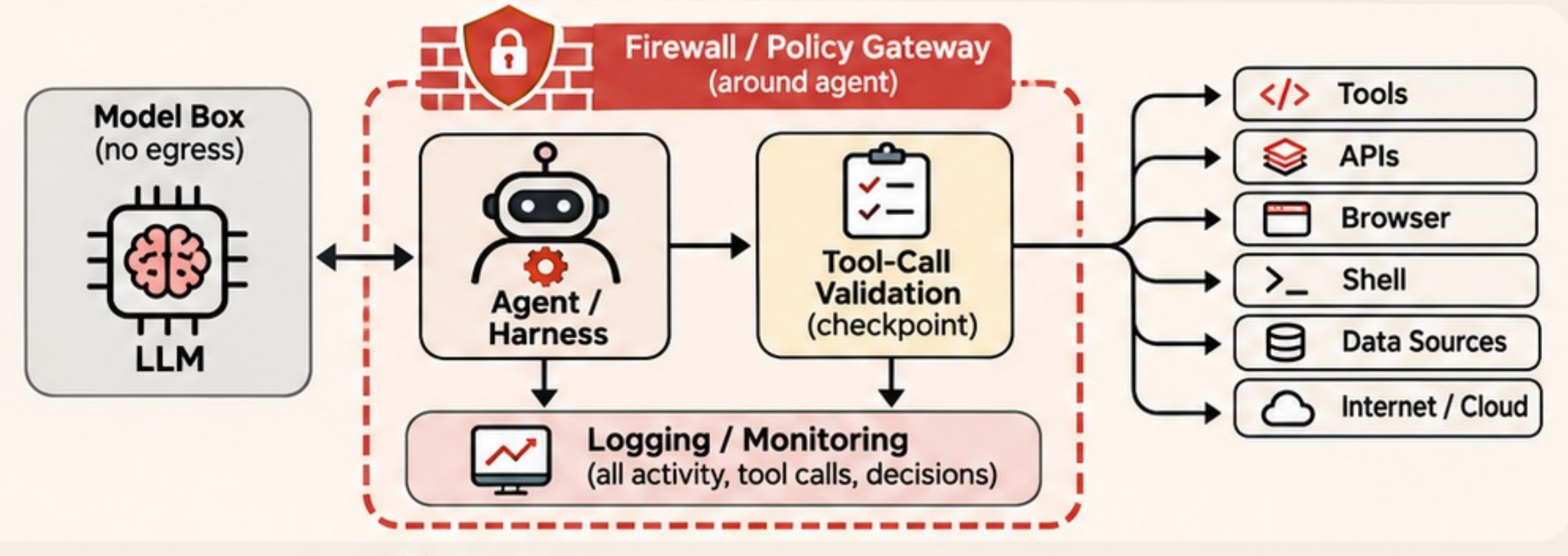
© Anant Shrivastava

LLM Going Rogue



© Anant Shrivastava

LLM Going Rogue



© Anant Shrivastava

Developer Containment I built

- By now people either have their fav containment system or have build one
- I build my own (How hard could it be?):
  - Question asked on 12 June 2026
  - Answer I am still figuring out
- https://github.com/cyfinoid/aidc

© Anant Shrivastava

Cyfinoid

AIDC : AI Dev Containers

- Virtualization / Containerization
  - docker
  - apple containers (recommended for my model)
- Security
  - Pmg (Package Manager Gateway) (safedep/pmg)
  - Automated SBoM
  - Forced code review

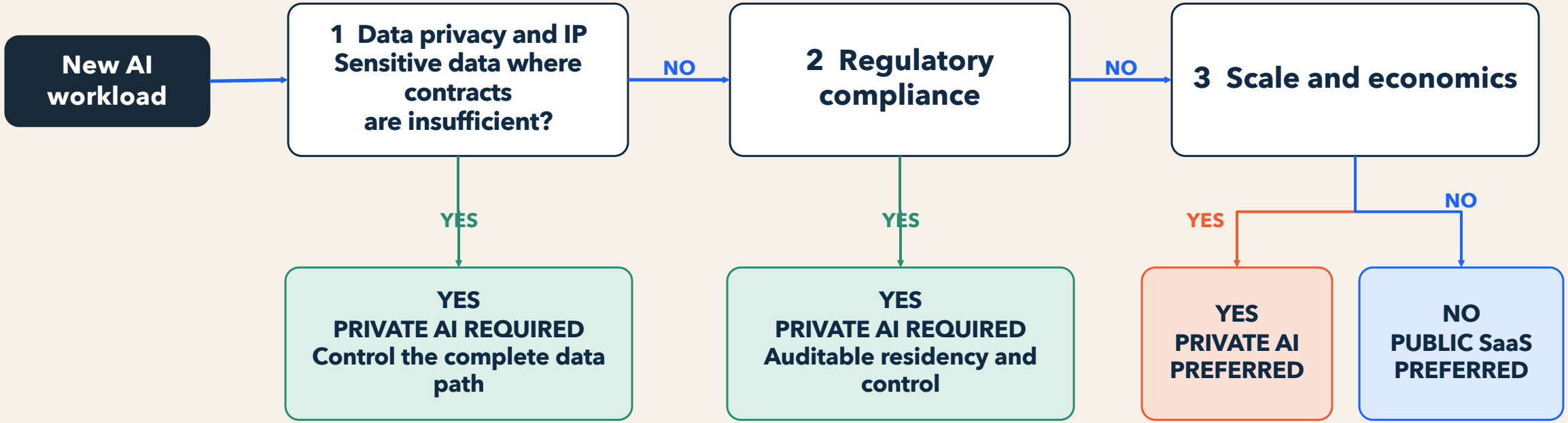
© Anant Shrivastava

AIDC: Continues

- Builtin optimization
  - A second fresh eye view by an agent
  - Pre baked in tools like RTK
- Contained but still visible
  - Session logs get pulled at base so observability tools see them
- Build optimization and minimal disk space
  - Based on custom ubuntu
  - Split image layers

© Anant Shrivastava

Question : Should we self host



Or

- You simply want full control over all the elements

© Anant Shrivastava

Cyfinoid

My Own AI Hardware



© Anant Shrivastava

Cyfinoid

I bought it; I Own it (llama.cpp)

PR	Status	What it fixed / gained
#28918	merged 2026-09-21	Coalesce MKL-FA softmax loads: softmax kernel 2407->440 ms (5.5x); **pp@126k +49%**
#28931	merged 2026-09-23	MMVQ GLU fusion for mixed quants + rms_norm+scale + ssm_conv+silu fusions; **tg +1.5-1.8%**
#29107	merged 2026-10-14	IQ3 persistent reorder: IQ3 **tg256@9k 12.30->21.59 tok/s**
#29171	merged 2026-10-07	GLM-4.7 Flash MLA prefill via MKL-FA: **pp@8192 3.01x, pp@64000 5.64x**
#30604	merged 2026-10-21	IQ3 crash fix after the reorder (NaN/garbage with MTP)
#30606	open	BMG D=256 single-query decode -> FA vector kernel (Spark 4B class)
#30976	merged 2026-10-29	Vectorized reordered dequant stores: **Q4_K +14.7-18.7%, Q6_K +28.7%, IQ3_S +21.1%, MXFP4-MoE +57-62%** (primed prefill)
#32486	open	F16-KV + re-tiled Q8_0 dpas decode arms: served MTP **TG @~125k +37-39%**
#30728	open	Metal GQA fa_vec decode (Apple M4 Max card)

© Anant Shrivastava

My Private AI Setup | Setup, flow, and what I use it for

Cyfinoid

Security Work

- **Authorized pentest** (research, validation, analysis)
- **Code review** (security-focused analysis)

Normal Work

- **Chatting** (ideas, analysis, troubleshooting)
- **Secondbrain management** (notes, knowledge, references)
- **Document creation & review** (drafts, edits, summaries)

Client

My devices

- MacBook Pro
- Other laptops / systems
- Mobile (as needed)

Tailscale

Private network access

- Secure
- Encrypted
- My devices only

Private AI Stack (on my hardware)

LLM Gateway

llama-swap (customized)

- OpenAI-compatible API
- Model routing / switching
- My customizations

LLM Inference

llama.cpp (customized)

- High-performance inference
- Cherry-picked version
- Tuned for my hardware

Hardware

- NVIDIA RTX 5000 Pro
- Intel Arc B70 Pro
- Apple MacBook Pro

In Short

I access my private AI stack over Tailscale from my devices, use a customized LLM gateway (**llama-swap**), run inference with a tuned **llama.cpp**, and power it with my RTX 5000 Pro, Arc B70 Pro, and MacBook Pro for both security work and daily tasks.

© Anant Shrivastava

Public Release

- https://github.com/anantshri/llm-playbooks
- Here is a repository filled with
  - Configs
  - Optimization tips
- Different environment
  - Nvidia (CUDA)
  - Intel (SYCL)
  - Apple (METAL)



© Cyfinoid Research

Thanks for listening



© Anant Shrivastava