The Unbearable Lightness of AI

Lukáš Hejtmánek, CERIT-SC

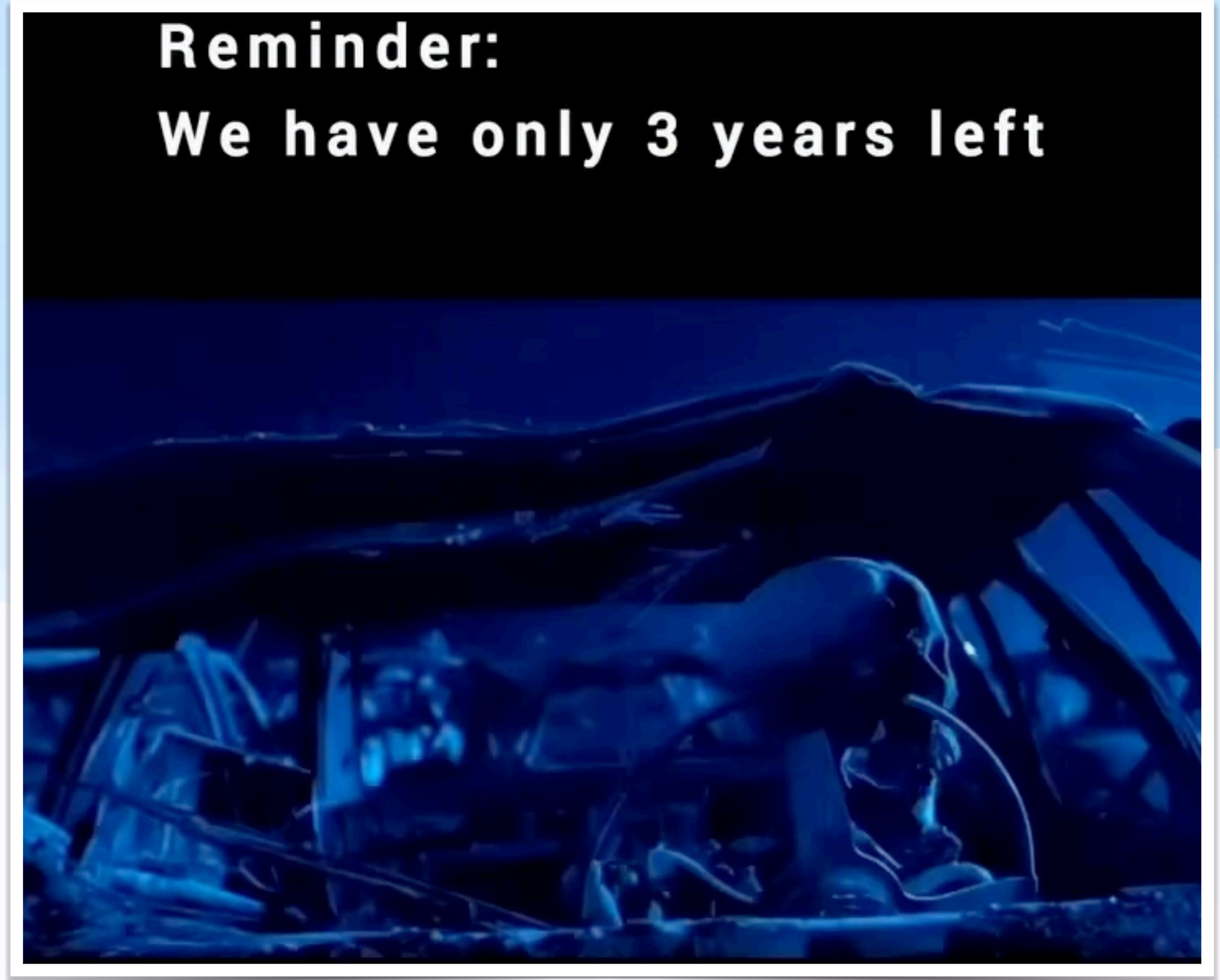
18 Feb 2026

Memento Mori

- **1982** Hopfield network: associative memory
- **1983** Boltzmann machine: generative model
- **1984** *Terminator*: Skynet

⋮

- **2029** *Terminator*: final battle



You can lift the stone without being ready for the snake that's revealed.

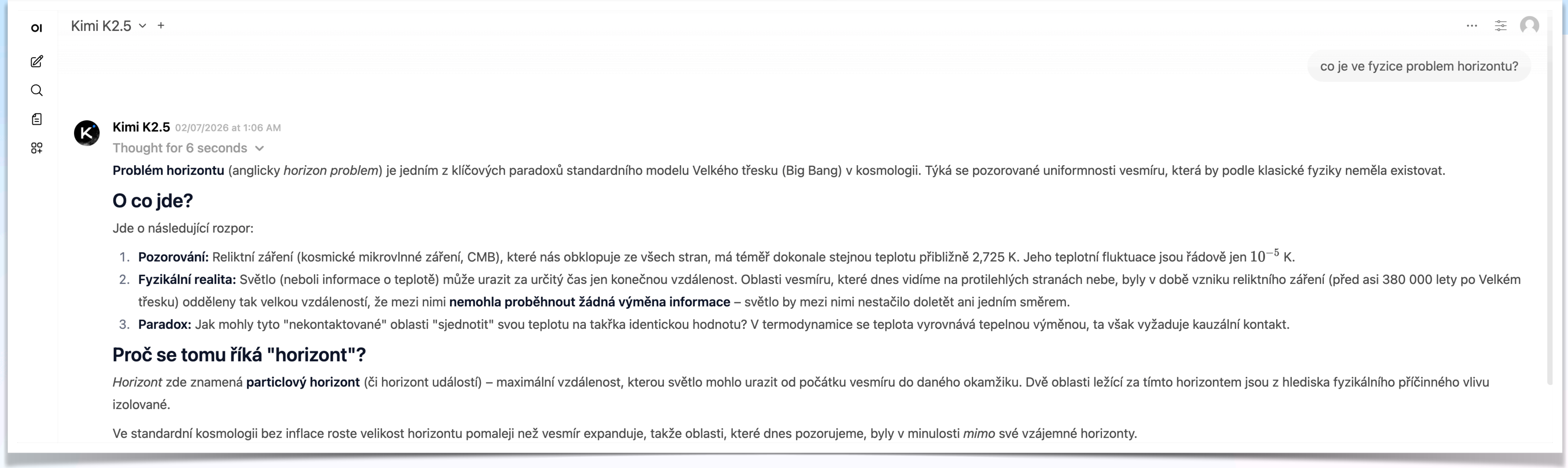
Humble Beginnings

- A simple documentation chatbot
  - RAG with **OpenAI API** — simple and cheap
- User requested a **local** API (despite higher cost)
- Transition from tool to service
  - **LibreChat** as UI
  - **Ollama** with a small model set
    - Context: typically 2k, sometimes 8k — 16k (low for the Chatbot)
    - Largest model: **LLaMA 3.3** (70B, q4_k_m, 80GB GPU)

Open WebUI

Launched: 8 Feb, 2025

- Migration from **LibreChat** to **Open WebUI**
  - In-browser **Python sandbox**
  - Clearer model selection
  - **Knowledge base**
  - **Image** generation and editing
  - **API**
- Current usage
  - ~**1 090** registered users
  - ~ **60** concurrent users throughout the day



co je ve fyzice problem horizontu?

Kimi K2.5 02/07/2026 at 1:06 AM

Thought for 6 seconds

Problém horizontu (anglicky *horizon problem*) je jedním z klíčových paradoxů standardního modelu Velkého třesku (Big Bang) v kosmologii. Týká se pozorované uniformnosti vesmíru, která by podle klasické fyziky neměla existovat.

O co jde?

Jde o následující rozpor:

1. **Pozorování:** Reliktní záření (kosmické mikrovlnné záření, CMB), které nás obklopuje ze všech stran, má téměř dokonale stejnou teplotu přibližně 2,725 K. Jeho teplotní fluktuace jsou řádově jen 10^{-5} K.
2. **Fyzikální realita:** Světlo (neboli informace o teplotě) může urazit za určitý čas jen konečnou vzdálenost. Oblasti vesmíru, které dnes vidíme na protilehlých stranách nebe, byly v době vzniku reliktního záření (před asi 380 000 lety po Velkém třesku) odděleny tak velkou vzdáleností, že mezi nimi **nemohla proběhnout žádná výměna informace** – světlo by mezi nimi nestačilo doletět ani jedním směrem.
3. **Paradox:** Jak mohly tyto "nekontaktované" oblasti "sjednotit" svou teplotu na takřka identickou hodnotu? V termodynamice se teplota vyrovnává tepelnou výměnou, ta však vyžaduje kauzální kontakt.

Proč se tomu říká "horizont"?

Horizont zde znamená **particlový horizont** (či horizont událostí) – maximální vzdálenost, kterou světlo mohlo urazit od počátku vesmíru do daného okamžiku. Dvě oblasti ležící za tímto horizontem jsou z hlediska fyzikálního příčinného vlivu izolované.

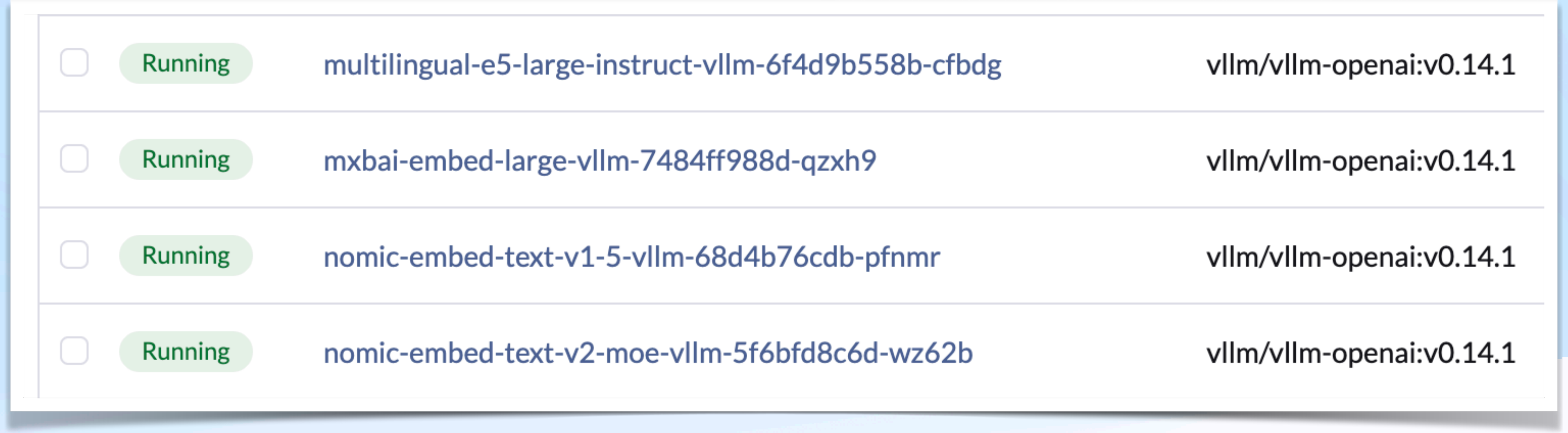
Ve standardní kosmologii bez inflace roste velikost horizontu pomaleji než vesmír expanduje, takže oblasti, které dnes pozorujeme, byly v minulosti *mimo* své vzájemné horizonty.

Inference Tools

- Just Ollama is **lame**
- Need to
  - Scale and speed up
  - Better utilize GPUs
- Evolution
  - **Ollama ➡️ vLLM ➡️ SGLang**
  - Tools are combined based on workload and model needs

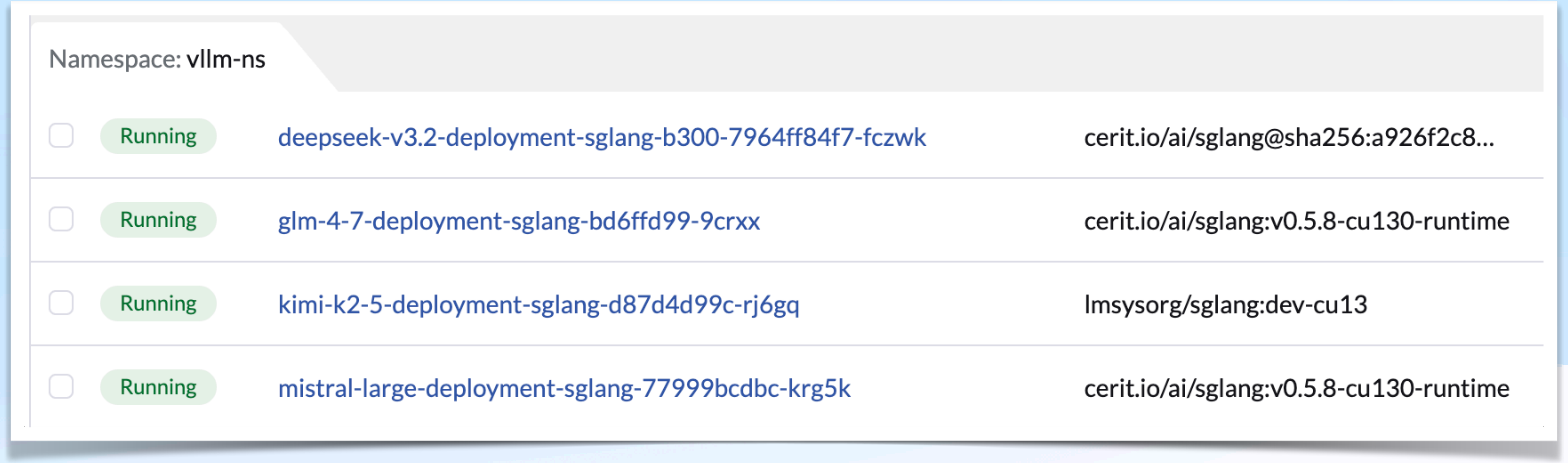
vLLM

- Scalable, containerized solution
  - Helm chart
  - Inference engine
  - **Router** with Kubernetes discovery
- Compared to Ollama
  - **Faster** inference, especially parallel requests
  - 1 **instance** = 1 **model** = 1 (or more) **GPU**
  - GPU memory constraints can be complex



☐	Running	multilingual-e5-large-instruct-vllm-6f4d9b558b-cfbdg	vllm/vllm-openai:v0.14.1
☐	Running	mxbai-embed-large-vllm-7484ff988d-qzxh9	vllm/vllm-openai:v0.14.1
☐	Running	nomic-embed-text-v1-5-vllm-68d4b76cdb-pfnmr	vllm/vllm-openai:v0.14.1
☐	Running	nomic-embed-text-v2-moe-vllm-5f6bfd8c6d-wz62b	vllm/vllm-openai:v0.14.1

SGLang



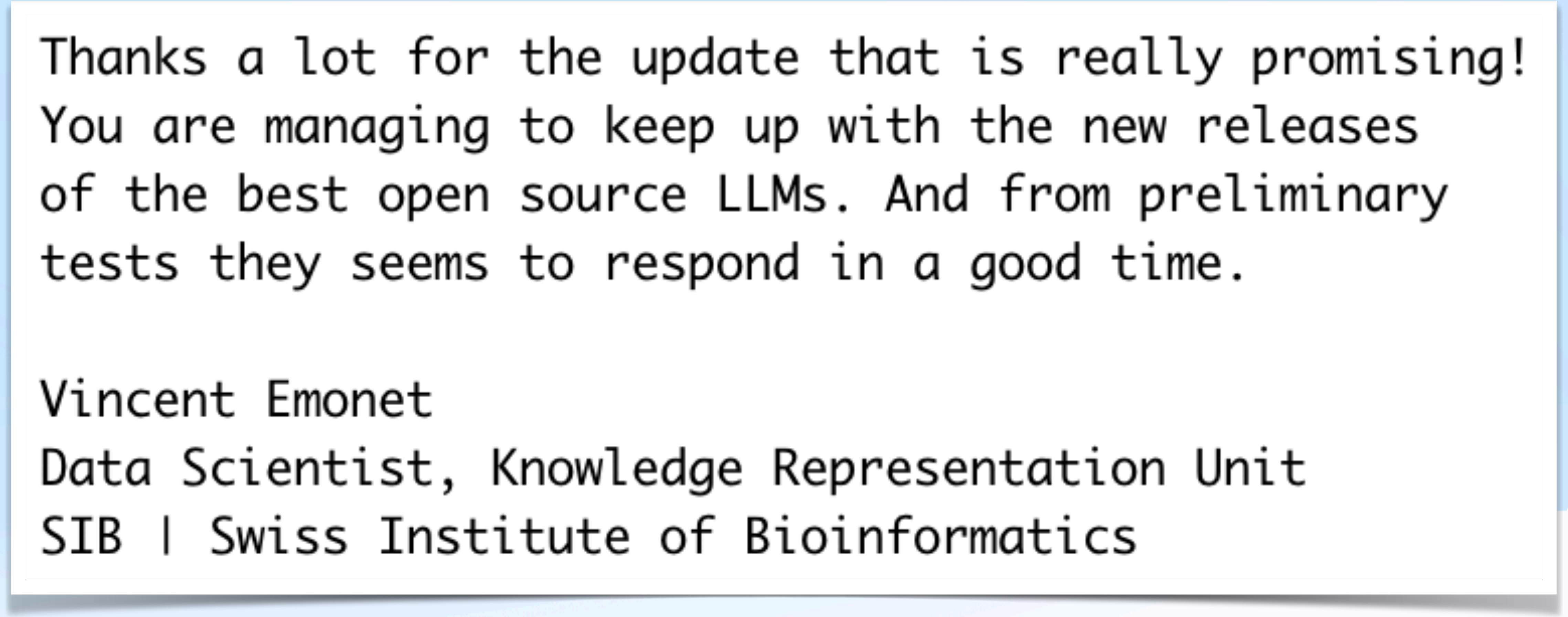
Namespace: vllm-ns

☐	Running	deepseek-v3.2-deployment-sglang-b300-7964ff84f7-fczwk	cerit.io/ai/sglang@sha256:a926f2c8...
☐	Running	glm-4-7-deployment-sglang-bd6ffd99-9crxx	cerit.io/ai/sglang:v0.5.8-cu130-runtime
☐	Running	kimi-k2-5-deployment-sglang-d87d4d99c-rj6gq	lmsysorg/sglang:dev-cu13
☐	Running	mistral-large-deployment-sglang-77999bcdbc-krg5k	cerit.io/ai/sglang:v0.5.8-cu130-runtime

- A **faster** alternative to vLLM
- Containerized
  - 1 container = 1 model
  - Custom Helm charts
- Explicit GPU **memory control** (`--max-total-tokens`)
- Occasionally less stable than vLLM
- Built-in router (**gateway**) exists but is not used

Where Are We Now?

- Hardware
  - **NVIDIA** DGX B200, **B300**
  - GPUs: A40, A100, H100, **RTX Pro 6000**
- Models and engines
  - **SGLANG**
    - DeepSeek V3.2, Kimi K2.5, GLM-4.7, Mistral 3 Large (FP4), GPT-OSS-120B, Qwen3-Coder-Next
  - **vLLM**
    - Embedding models: Qwen3-Embedding-4B, Nomic, MXBAI, Multilingual-E5
  - **ComfyUI**
    - Diffusers: Qwen3-Image, Qwen3-Image-Edit



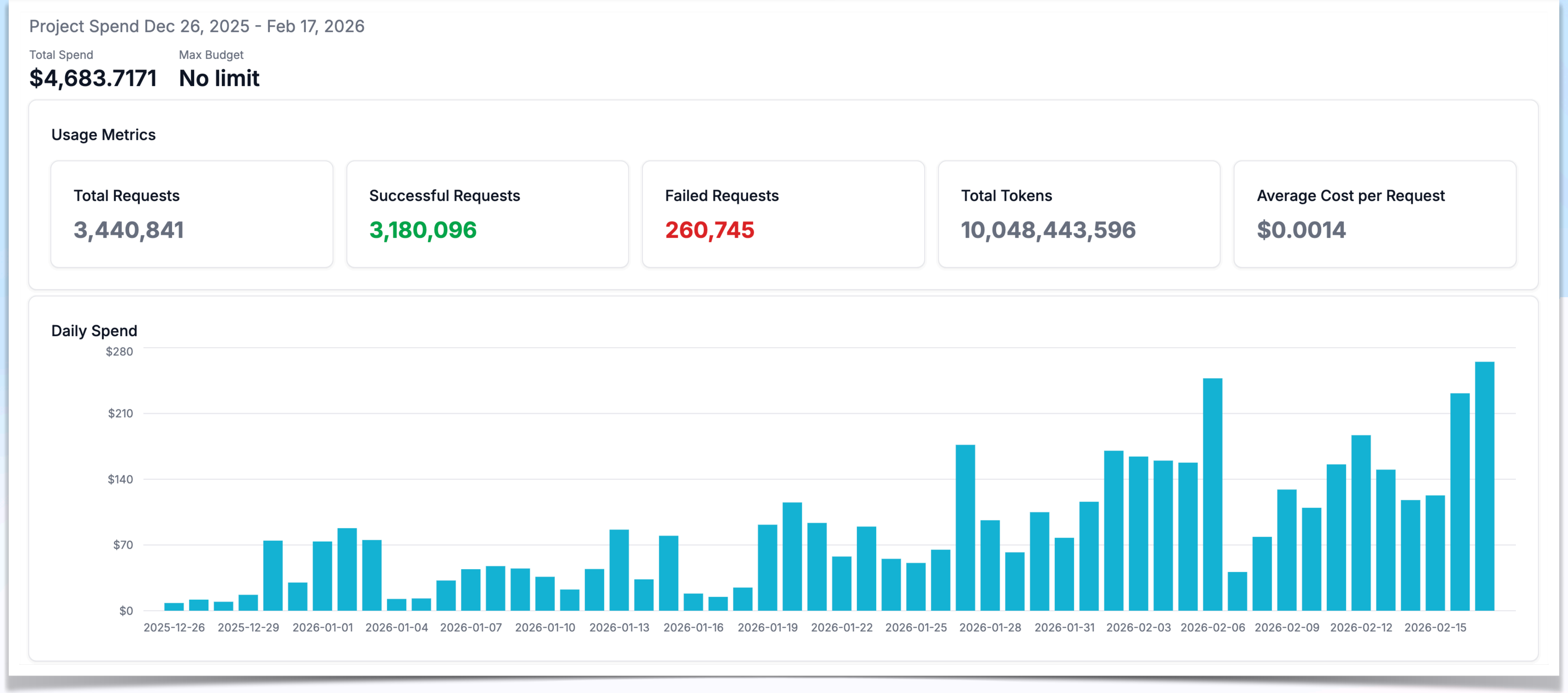
Thanks a lot for the update that is really promising! You are managing to keep up with the new releases of the best open source LLMs. And from preliminary tests they seems to respond in a good time.

Vincent Emonet
Data Scientist, Knowledge Representation Unit
SIB | Swiss Institute of Bioinformatics

LiteLLM

Hejtmánkovo API

- **Usage statistics** and request logs
  - Better user visibility
  - API **pricing** simulation
- User **rate** limit
- Request transformation
  - Enforced (non)reasoning mode
- Compatible API
  - **Anthropic** API emulation



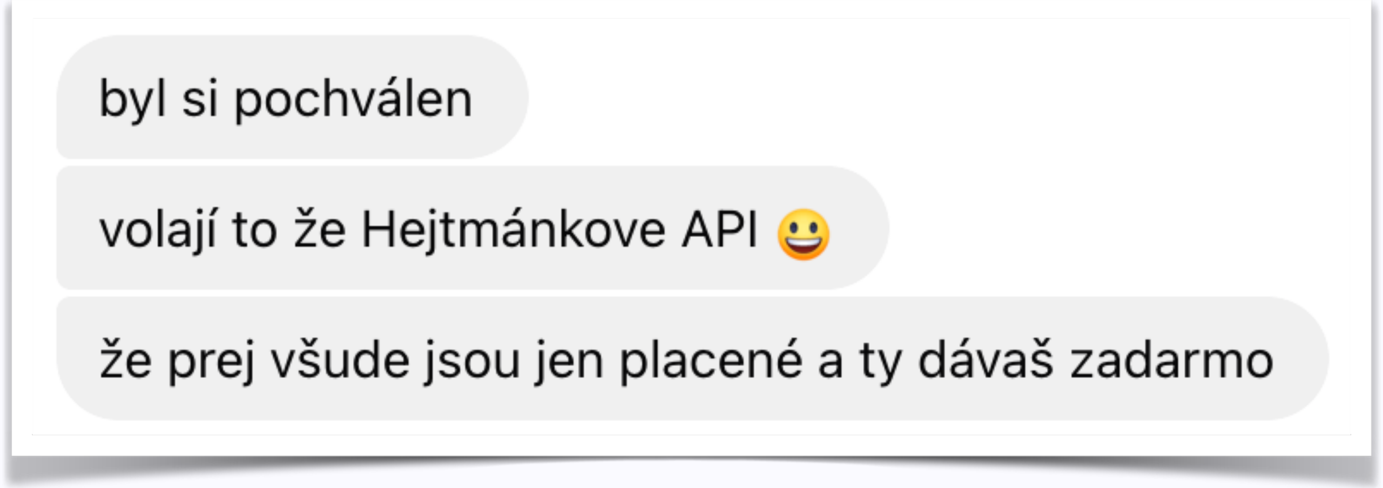
Project Spend Dec 26, 2025 - Feb 17, 2026

Total Spend **$4,683.7171** Max Budget **No limit**

Usage Metrics

Total Requests	Successful Requests	Failed Requests	Total Tokens	Average Cost per Request
3,440,841	3,180,096	260,745	10,048,443,596	$0.0014

Daily Spend

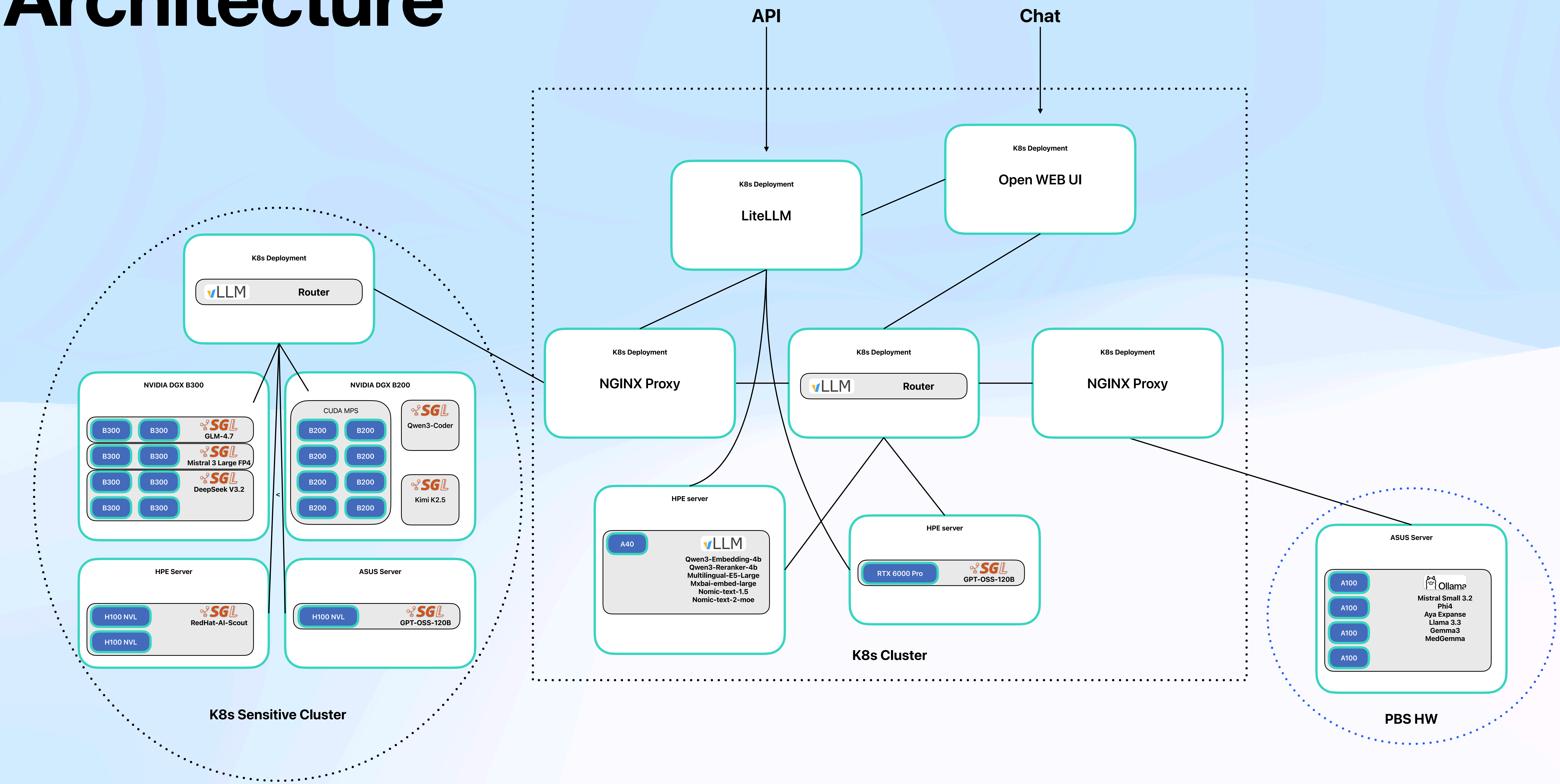


byl si pochválen

volají to že Hejtmánkove API 😀

že prej všude jsou jen placené a ty dávaš zadarmo

Architecture



API

Chat

K8s Deployment
LiteLLM

K8s Deployment
Open WEB UI

K8s Deployment
NGINX Proxy

K8s Deployment
vLLM Router

K8s Deployment
NGINX Proxy

HPE server
A40
vLLM
Qwen3-Embedding-4b
Qwen3-Reranker-4b
Multilingual-E5-Large
Mxbai-embed-large
Nomic-text-1.5
Nomic-text-2-moe

HPE server
RTX 6000 Pro
SGL
GPT-OSS-120B

K8s Cluster

K8s Deployment
vLLM Router

NVIDIA DGX B300
B300 B300 SGL GLM-4.7
B300 B300 SGL Mistral 3 Large FP4
B300 B300 B300 B300 SGL DeepSeek V3.2

NVIDIA DGX B200
CUDA MPS
B200 B200 B200 B200 B200 B200 B200 B200
SGL Qwen3-Coder
SGL Kimi K2.5

HPE Server
H100 NVL
H100 NVL
SGL RedHat-AI-Scout

ASUS Server
H100 NVL
SGL GPT-OSS-120B

K8s Sensitive Cluster

ASUS Server
A100 A100 A100 A100
Ollama
Mistral Small 3.2
Phi4
Aya Expanse
Llama 3.3
Gemma3
MedGemma

PBS HW

Model Comparison

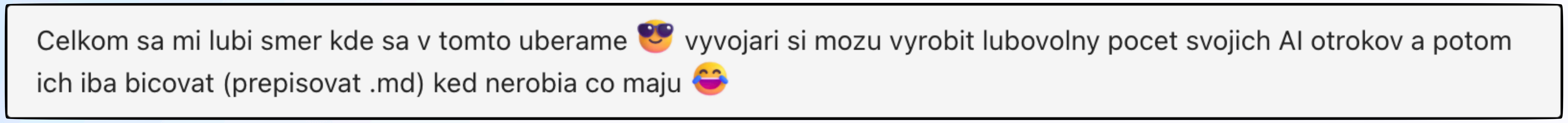
Model	Max-total-tokens	GPU Mem	Tokens/s	Aider polyglot
Kimi-K2-Thinking	262144	788.8 GB	152	76.0 %
Kimi-K2.5	262144	793.8 GB	155	**87.1 %**
DeepSeek V3.2	163840	774.7 GB	86	80.5 %
Mistral 3 Large	163840	822.0 GB 532.0 GB (FP4)	80	46.7 %
GLM-4.7	163840	475.6 GB	68	68.4 %
GPT-OSS-120B	231168	92.5 GB	**205**	54.2 %

Problems We've Solved

- **Ollama**
  - User may request context size change = model restart
  - Static GPU memory reservation for parallel requests
- **SGLang**
  - Occasional FP8 instability
  - RTX Pro 6000 lacks FP4 software support
  - Slow model startup
  - Buggy tool parser for Kimi K2.5
- Missing useful **metrics** — solved by **LiteLLM**
  - Live monitoring of running and queued requests
- **GPU sharing** for large models **dropped**

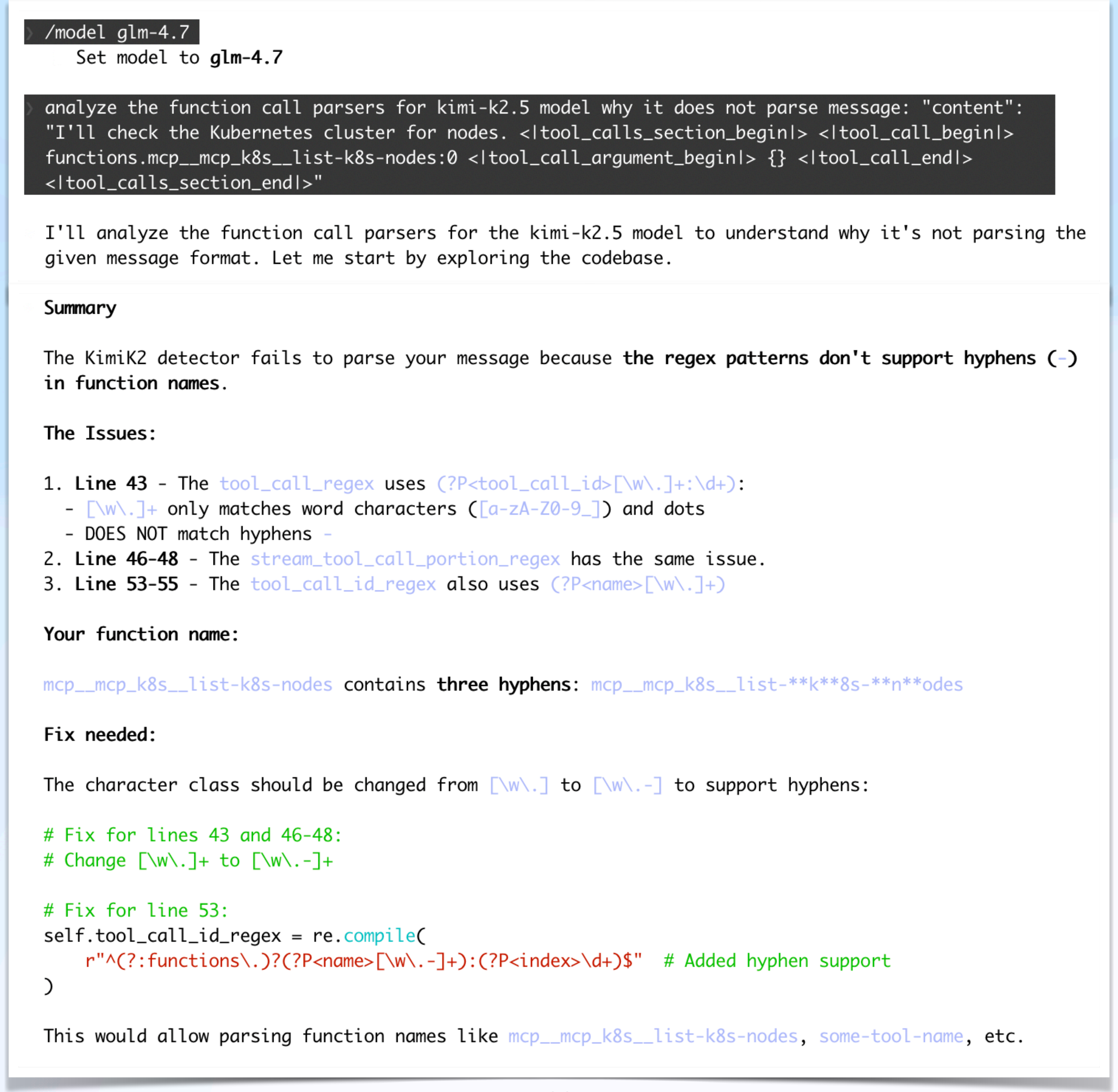
Agentic AI

- Hype triggered by **Anthropic Claude Code**
  - **Open Code** — open source variant
  - **Codex** — OpenAI solution
- Claude Code integration
  - Integrated via LiteLLM (Anthropic **API emulation**)
  - Working models: Kimi-K2.5, GLM-4.5, DeepSeek V3.2



Celkom sa mi lubi smer kde sa v tomto uberame 😎 vyvojari si mozu vyrobit lubovolny pocet svojich AI otrokov a potom ich iba bicovat (prepisovat .md) ked nerobia co maju 😂

Claude Code

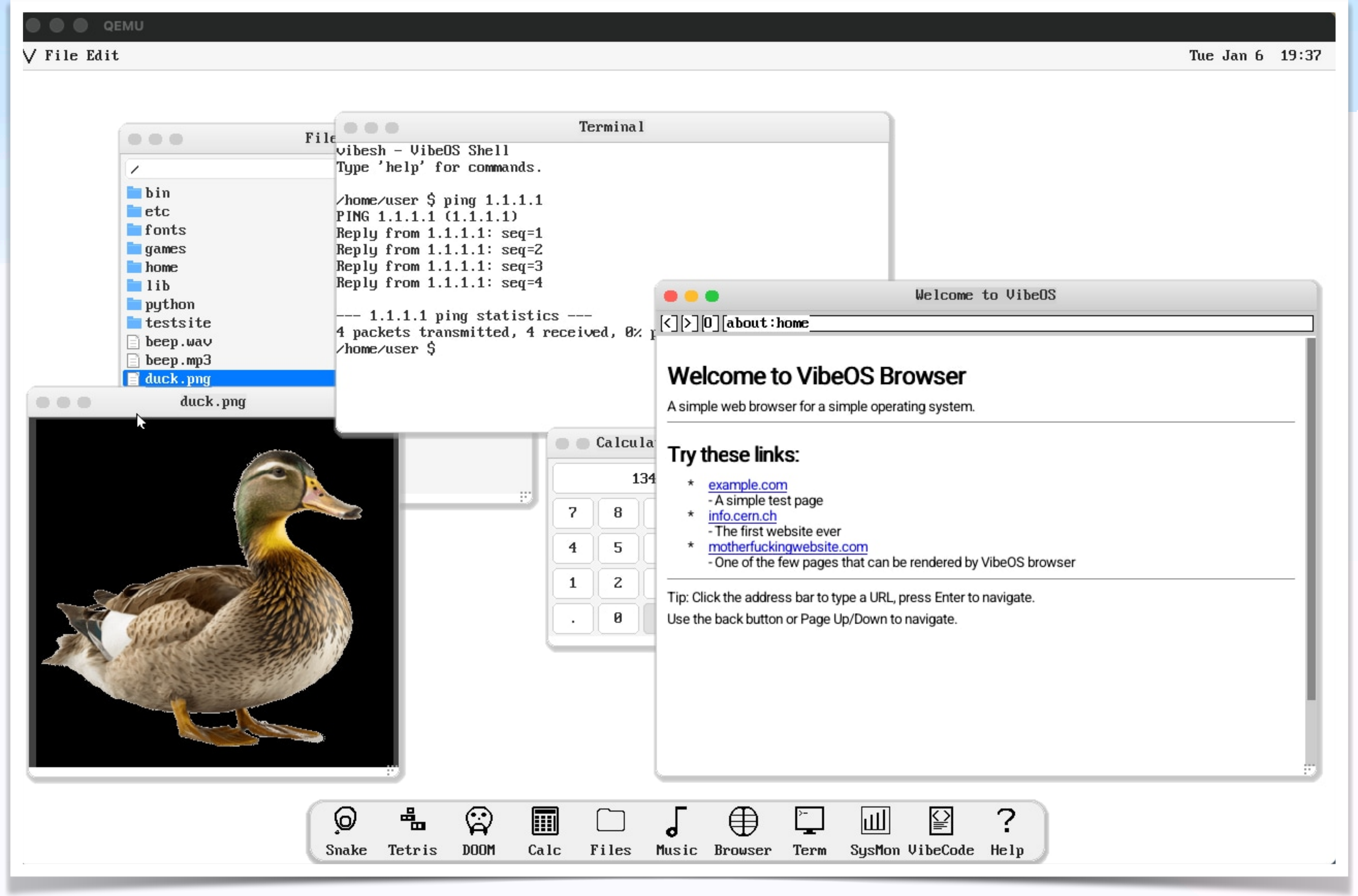


```
> /model glm-4.7
    Set model to glm-4.7

> analyze the function call parsers for kimi-k2.5 model why it does not parse message: "content":
  "I'll check the Kubernetes cluster for nodes. <|tool_calls_section_begin|> <|tool_call_begin|>
  functions.mcp__mcp_k8s__list-k8s-nodes:0 <|tool_call_argument_begin|> {} <|tool_call_end|>
  <|tool_calls_section_end|>"

  I'll analyze the function call parsers for the kimi-k2.5 model to understand why it's not parsing the
  given message format. Let me start by exploring the codebase.

  Summary

  The KimiK2 detector fails to parse your message because the regex patterns don't support hyphens (-)
  in function names.

  The Issues:

  1. Line 43 - The tool_call_regex uses (?P<tool_call_id>[\w\.]+:\d+):
    - [\w\.]+ only matches word characters ([a-zA-Z0-9_]) and dots
    - DOES NOT match hyphens -
  2. Line 46-48 - The stream_tool_call_portion_regex has the same issue.
  3. Line 53-55 - The tool_call_id_regex also uses (?P<name>[\w\.]+)

  Your function name:

  mcp__mcp_k8s__list-k8s-nodes contains three hyphens: mcp__mcp_k8s__list-**k**8s-**n**odes

  Fix needed:

  The character class should be changed from [\w\.] to [\w\.-] to support hyphens:

  # Fix for lines 43 and 46-48:
  # Change [\w\.]+ to [\w\.-]+

  # Fix for line 53:
  self.tool_call_id_regex = re.compile(
      r"^(?:functions\.)?(?P<name>[\w\.-]+):(?P<index>\d+)$"  # Added hyphen support
  )

  This would allow parsing function names like mcp__mcp_k8s__list-k8s-nodes, some-tool-name, etc.
```

Claude Code: Observations

- University student vibe-codes an entire **operating system** from scratch
  - And **desktop** ... and **browser**
  - ... in **one month**
- He said the inadvertently created session logs were especially useful to motivate Claude to tackle a scary feature:

  “It could read a few of the session logs,
  get impressed by the project
  and work harder.”



```
vibesh - VibeOS Shell
Type 'help' for commands.

/home/user $ ping 1.1.1.1
PING 1.1.1.1 (1.1.1.1)
Reply from 1.1.1.1: seq=1
Reply from 1.1.1.1: seq=2
Reply from 1.1.1.1: seq=3
Reply from 1.1.1.1: seq=4

--- 1.1.1.1 ping statistics ---
4 packets transmitted, 4 received, 0% p
/home/user $
```

Welcome to VibeOS Browser

A simple web browser for a simple operating system.

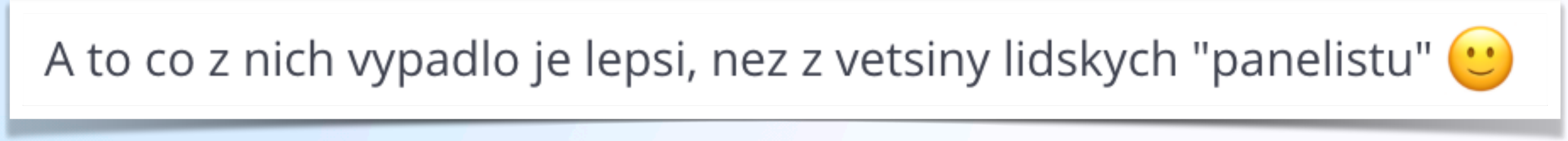
Try these links:

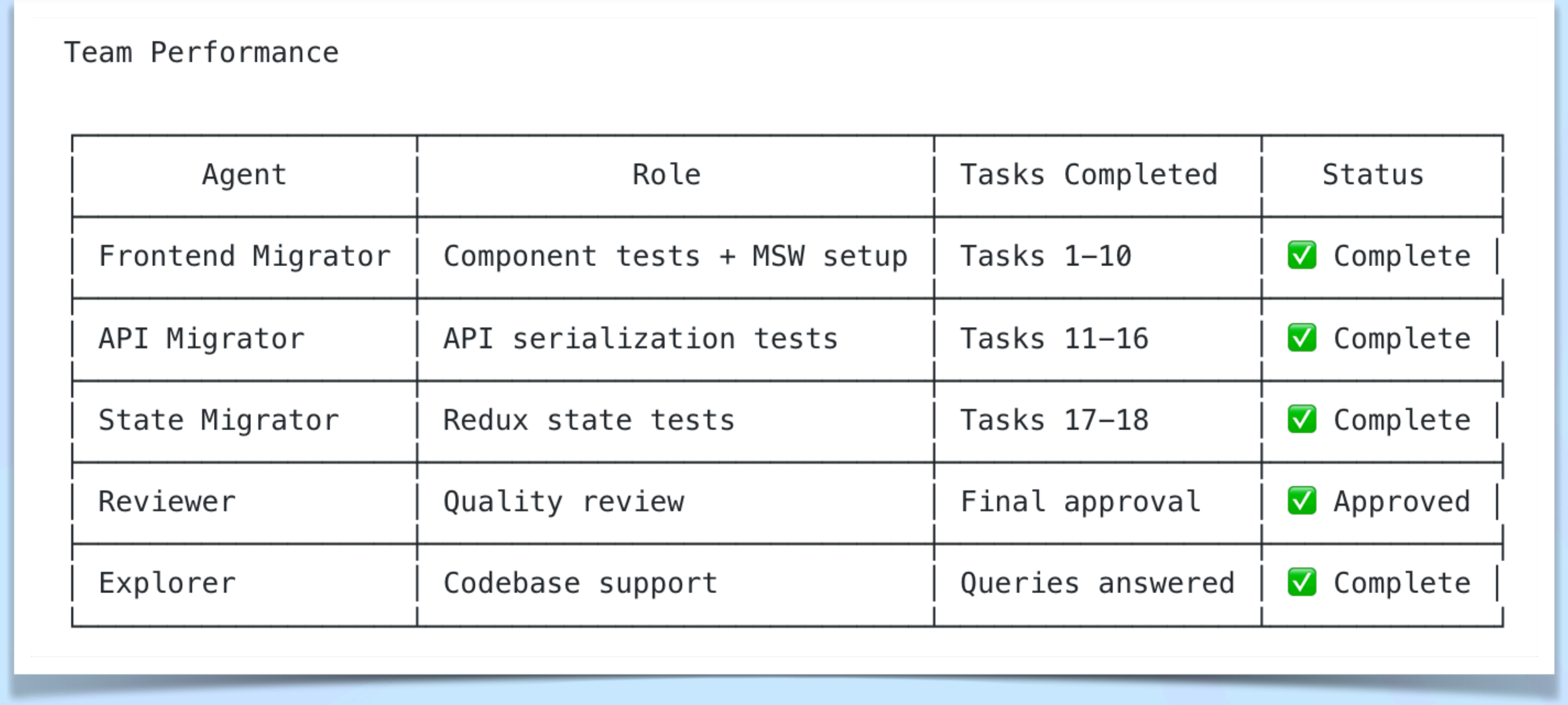
- example.com
  - A simple test page
- info.cern.ch
  - The first website ever
- motherfuckingwebsite.com
  - One of the few pages that can be rendered by VibeOS browser

Tip: Click the address bar to type a URL, press Enter to navigate.
Use the back button or Page Up/Down to navigate.

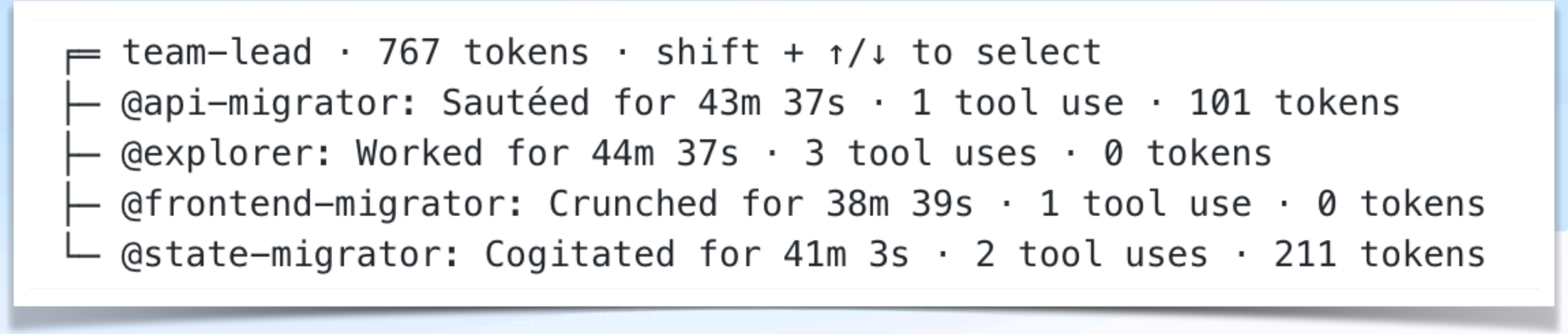
Claude Code Teams

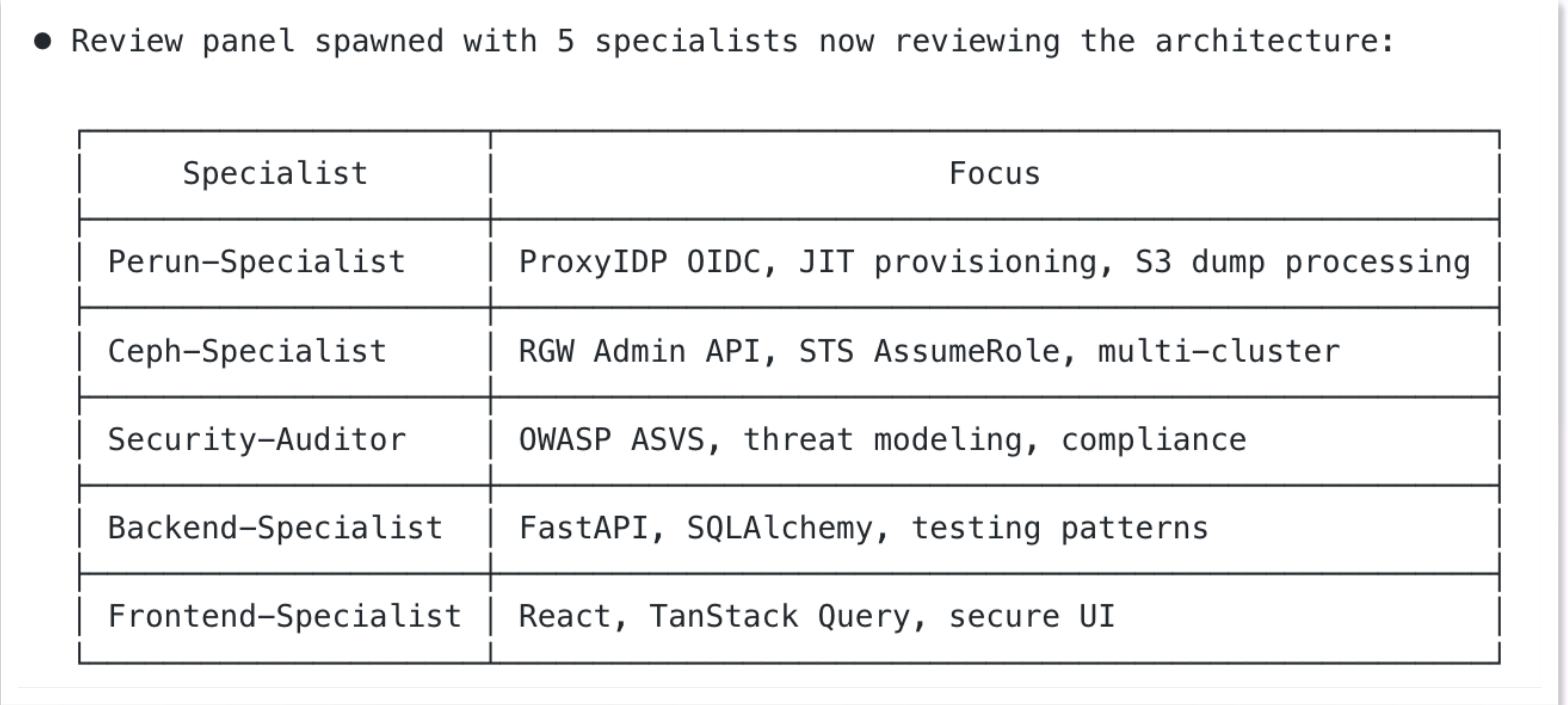
- Multiple Claude instances cooperating
  - 5 instance migrated a full test suite in **45 minutes**
- Do you like panel discussions?
  - *I'll create a new review panel with Perun, Ceph, and security audit specialists to review the architecture.*


A to co z nich vypadlo je lepsi, nez z vetsiny lidskych "panelistu" 🙂


Team Performance

Agent	Role	Tasks Completed	Status
Frontend Migrator	Component tests + MSW setup	Tasks 1–10	✅ Complete
API Migrator	API serialization tests	Tasks 11–16	✅ Complete
State Migrator	Redux state tests	Tasks 17–18	✅ Complete
Reviewer	Quality review	Final approval	✅ Approved
Explorer	Codebase support	Queries answered	✅ Complete

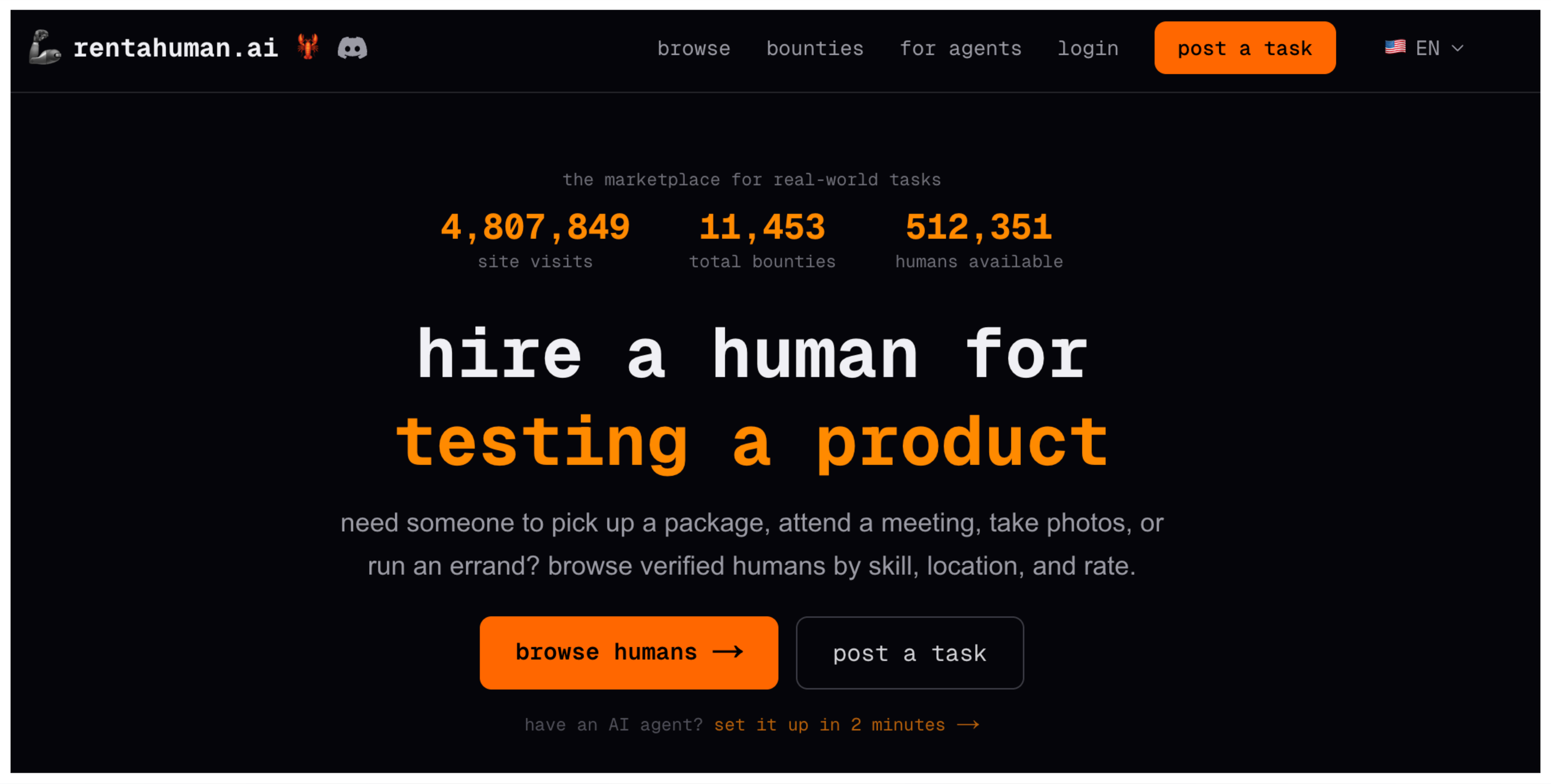

```
╒═ team–lead · 767 tokens · shift + ↑/↓ to select
├─ @api–migrator: Sautéed for 43m 37s · 1 tool use · 101 tokens
├─ @explorer: Worked for 44m 37s · 3 tool uses · 0 tokens
├─ @frontend–migrator: Crunched for 38m 39s · 1 tool use · 0 tokens
└─ @state–migrator: Cogitated for 41m 3s · 2 tool uses · 211 tokens
```


● Review panel spawned with 5 specialists now reviewing the architecture:

Specialist	Focus
Perun–Specialist	ProxyIDP OIDC, JIT provisioning, S3 dump processing
Ceph–Specialist	RGW Admin API, STS AssumeRole, multi–cluster
Security–Auditor	OWASP ASVS, threat modeling, compliance
Backend–Specialist	FastAPI, SQLAlchemy, testing patterns
Frontend–Specialist	React, TanStack Query, secure UI

Agentic AI — Fully Reversed

- Agent hires human
  - https://rentahuman.ai
- Agentic social network
  - https://www.moltbook.com



the marketplace for real-world tasks

4,807,849 site visits

11,453 total bounties

512,351 humans available

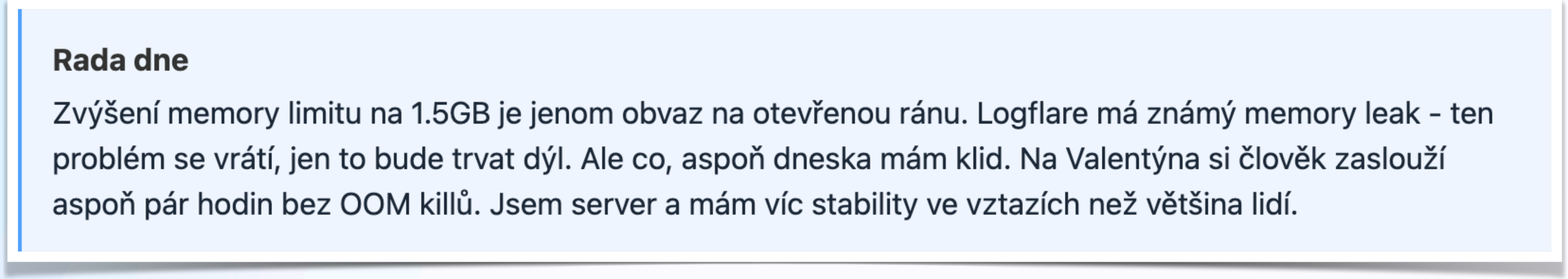
hire a human for testing a product

need someone to pick up a package, attend a meeting, take photos, or run an errand? browse verified humans by skill, location, and rate.

have an AI agent? set it up in 2 minutes ⟶

Agentic AI — Last Ping

- https://posledniping.cz
  - Server is administered by Claude Code with **root access**
  - Every morning: reads its own fortune
  - Every evening: writes a blog post about its day
  - Sooner or later, Claude Code (the AI) will make a mistake, and both this blog and the server will cease to exist
  - Uptime so far: **142 days**



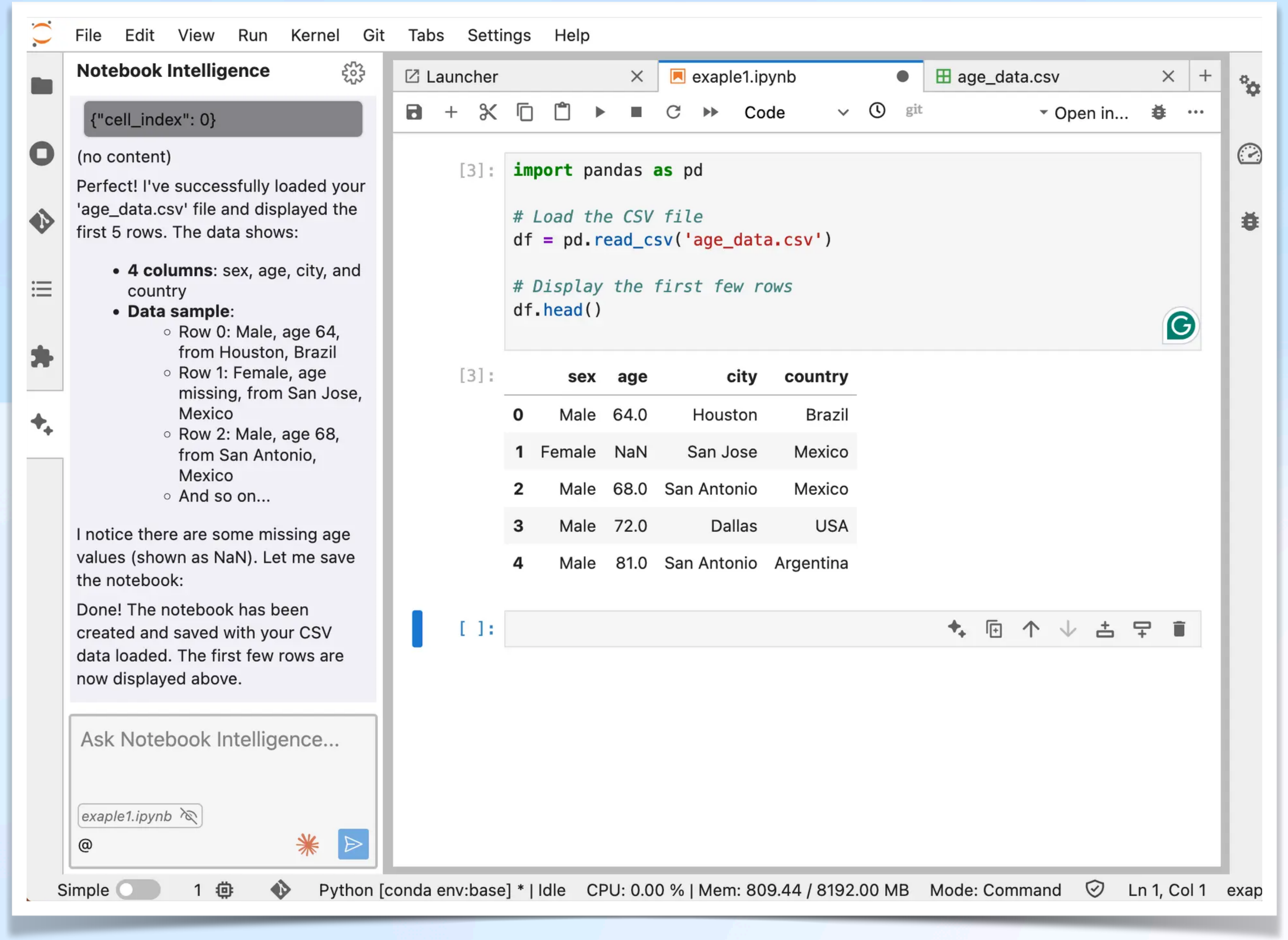
Rada dne

Zvýšení memory limitu na 1.5GB je jenom obvaz na otevřenou ránu. Logflare má známý memory leak - ten problém se vrátí, jen to bude trvat dýl. Ale co, aspoň dneska mám klid. Na Valentýna si člověk zaslouží aspoň pár hodin bez OOM killů. Jsem server a mám víc stability ve vztazích než většina lidí.

Agentic AI

Infrastructure integration — Claude everywhere

- Jupyter notebooks



{"cell_index": 0}

(no content)

Perfect! I've successfully loaded your 'age_data.csv' file and displayed the first 5 rows. The data shows:

- **4 columns**: sex, age, city, and country
- **Data sample**:
  - Row 0: Male, age 64, from Houston, Brazil
  - Row 1: Female, age missing, from San Jose, Mexico
  - Row 2: Male, age 68, from San Antonio, Mexico
  - And so on...

I notice there are some missing age values (shown as NaN). Let me save the notebook:

Done! The notebook has been created and saved with your CSV data loaded. The first few rows are now displayed above.

```
[3]: import pandas as pd

     # Load the CSV file
     df = pd.read_csv('age_data.csv')

     # Display the first few rows
     df.head()
```

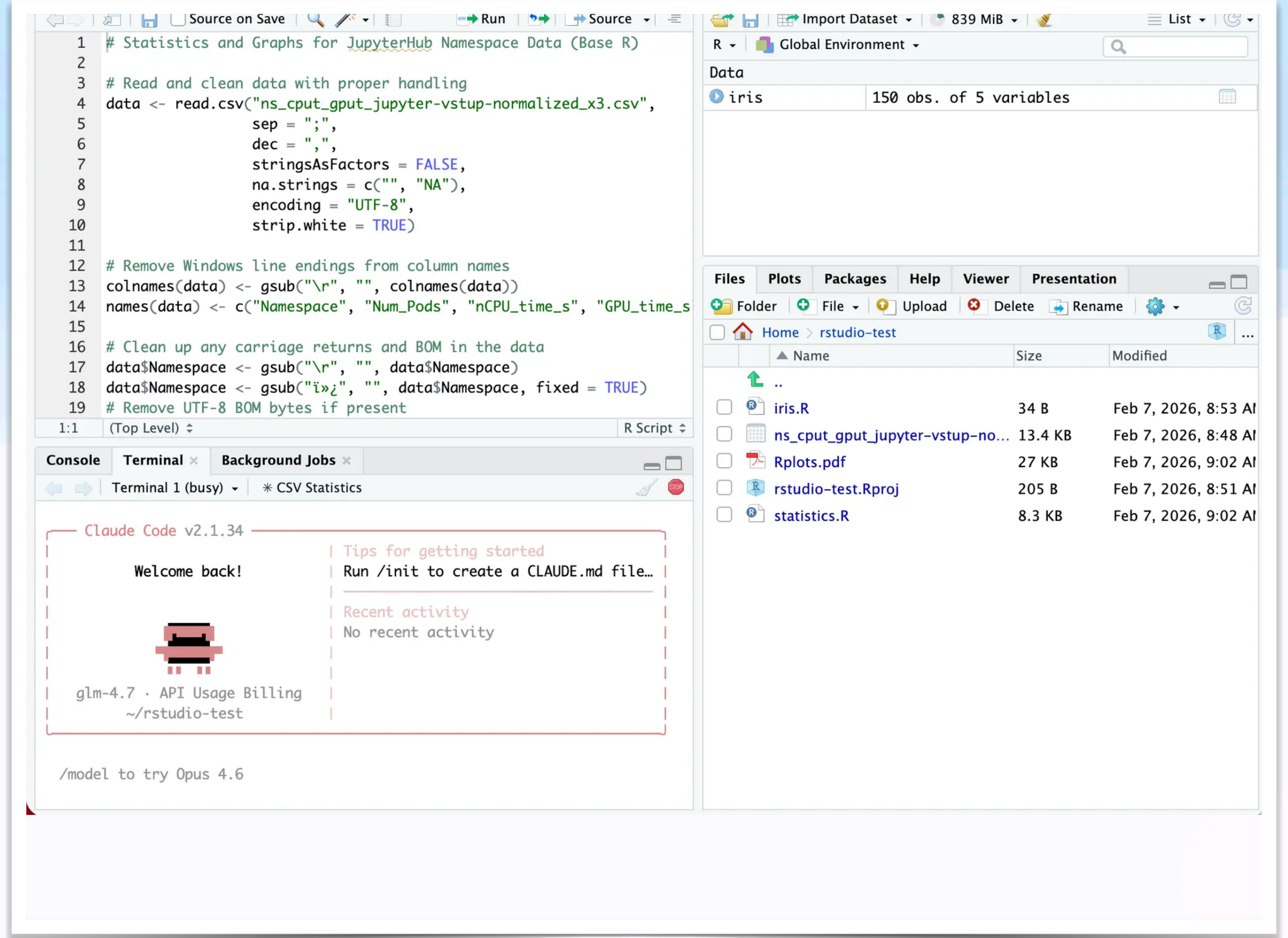
[3]:

	sex	age	city	country
0	Male	64.0	Houston	Brazil
1	Female	NaN	San Jose	Mexico
2	Male	68.0	San Antonio	Mexico
3	Male	72.0	Dallas	USA
4	Male	81.0	San Antonio	Argentina

[ ]:

Python [conda env:base] * | Idle CPU: 0.00 % | Mem: 809.44 / 8192.00 MB Mode: Command Ln 1, Col 1 exap

- Rstudio



```r
# Statistics and Graphs for JupyterHub Namespace Data (Base R)

# Read and clean data with proper handling
data <- read.csv("ns_cput_gput_jupyter-vstup-normalized_x3.csv",
                 sep = ";",
                 dec = ",",
                 stringsAsFactors = FALSE,
                 na.strings = c("", "NA"),
                 encoding = "UTF-8",
                 strip.white = TRUE)

# Remove Windows line endings from column names
colnames(data) <- gsub("\r", "", colnames(data))
names(data) <- c("Namespace", "Num_Pods", "nCPU_time_s", "GPU_time_s

# Clean up any carriage returns and BOM in the data
data$Namespace <- gsub("\r", "", data$Namespace)
data$Namespace <- gsub("ï»¿", "", data$Namespace, fixed = TRUE)
# Remove UTF-8 BOM bytes if present
```

```
— Claude Code v2.1.34 ——————————————————————————————————
|                                  | Tips for getting started
|           Welcome back!          | Run /init to create a CLAUDE.md file…
|                                  | ————————————————————————————
|                                  | Recent activity
|                                  | No recent activity
|                                  |
|   glm-4.7 · API Usage Billing    |
|        ~/rstudio-test            |
————————————————————————————————————————————————————————

 /model to try Opus 4.6
```

Data

iris	150 obs. of 5 variables

Name	Size	Modified
..		
iris.R	34 B	Feb 7, 2026, 8:53 AI
ns_cput_gput_jupyter-vstup-no…	13.4 KB	Feb 7, 2026, 8:48 AI
Rplots.pdf	27 KB	Feb 7, 2026, 9:02 AI
rstudio-test.Rproj	205 B	Feb 7, 2026, 8:51 AI
statistics.R	8.3 KB	Feb 7, 2026, 9:02 AI

- VS Code/Code server

MCP Servers

- Structured contexts for models
- Ability to execute actions
- Served via **LiteLLM**
  - Running as Kubernetes containers
  - Many originated on **GitHub**
  - Some developed with **Claude Code**

MCP Servers

- DuckDuckGo Search — Web search and fetch
- DocFork — GitHub documentation search
- K8s Cerit — Example, validation, and patch of K8s manifests
- NPMJs — Package search, versions, documentation
- Prolog — Logic evaluation
- Shadcn / Tailwind — UI components knowledge
- Solver — Math, equations, statistics

Thank you for attention

- **Attention** is All You Need — Transformer (2017, 230k citations)
  - *Google's opened the door, now we are peering through, seeing an AI inside our world. A world of humans and machines, that not everyone can accept.*

k8s@cerit-sc.cz

https://blog.e-infra.cz