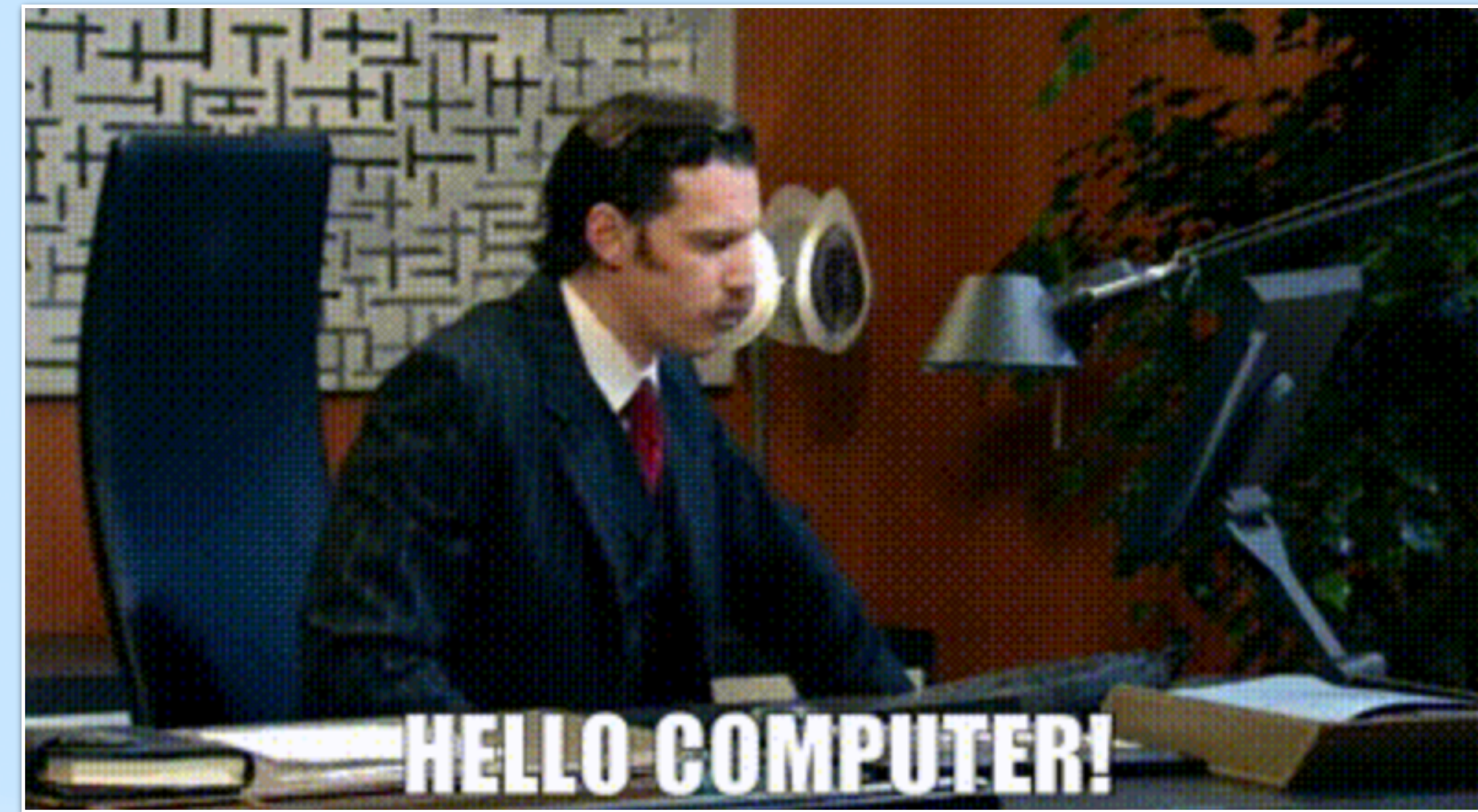


Hello Computer

AI, LLM a moderní aplikace



Have you tried ~~turning it off and on again~~ asking AI?

AI a LLM služby e-INFRA CZ

- **Chat služba** – <https://chat.ai.e-infra.cz>
 - Jednoduché použití á-la ChatGPT
- **LLM API** – <https://llm.ai.e-infra.cz>
 - Pro integraci do vlastních aplikací
- **Integrované služby**
 - DeepSite – <https://deepsite.ai.e-infra.cz>
 - JupyterHub – <https://hub.cloud.e-infra.cz>
- **Komunita**

Nabízené LLM

- **DeepSeek 4.1 Flash, GLM-5.3, Kimi-K3**
 - Kontext 1M
- **Gemma 4, Mistral Medium 3.5, Qwen 3.5 (Int4), Qwen 3.8 27b, Qwen 3.8 Flash Next**
 - Kontext 256k
- **GPT-OSS-120b**
 - Kontext 128k
- Embedding modely
 - **Qwen3 4b** (vč. Rerankeru), **MXBAI, MultiLingual E5, Nomic Embed** (1.5, 2)
- Transcript
 - **Whisper Large v3**

Large Language Models Service Current models

- **DeepSeek v4 Pro** (1.6T model)
 - Best for Czech language, reasoning, long context
- **GLM 5.2** – state-of-the-art open-weight model
 - Best for complex tasks, reasoning, problem solving
- **GPT-OSS-120B** (OpenAI-style mini model) – fast, best for small tasks
- **Kimi-K2.7** Coder – best coder
- **Qwen 3.5 397B** (Int4 only version, slowly deprecating) – good OCR performance
- **Qwen 3.5 122B** – faster variant of full Qwen 3.5, less capable though
- Other models
 - **Mistral Medium 3.5, Gemma 4, Command-A+, Whisper-Large-v3**
 - **Embedding models:** Qwen3-Embedding-4B, Multilingual E5 Large, Nomic

Modely na AI Meta Seminári Červenec 2026

Nabízené LLM

- Modely se velmi rychle vyvíjí
 - Nasazujeme novější verze co nejdříve, některé **0day**
 - **Negarantujeme**, že daný model bude celý semestr k dispozici
- **Není garantovaný 24/7 provoz**
 - V noci mohou probíhat testy, restarty
 - HW není redundantní → výpadek HW → výpadek modelu

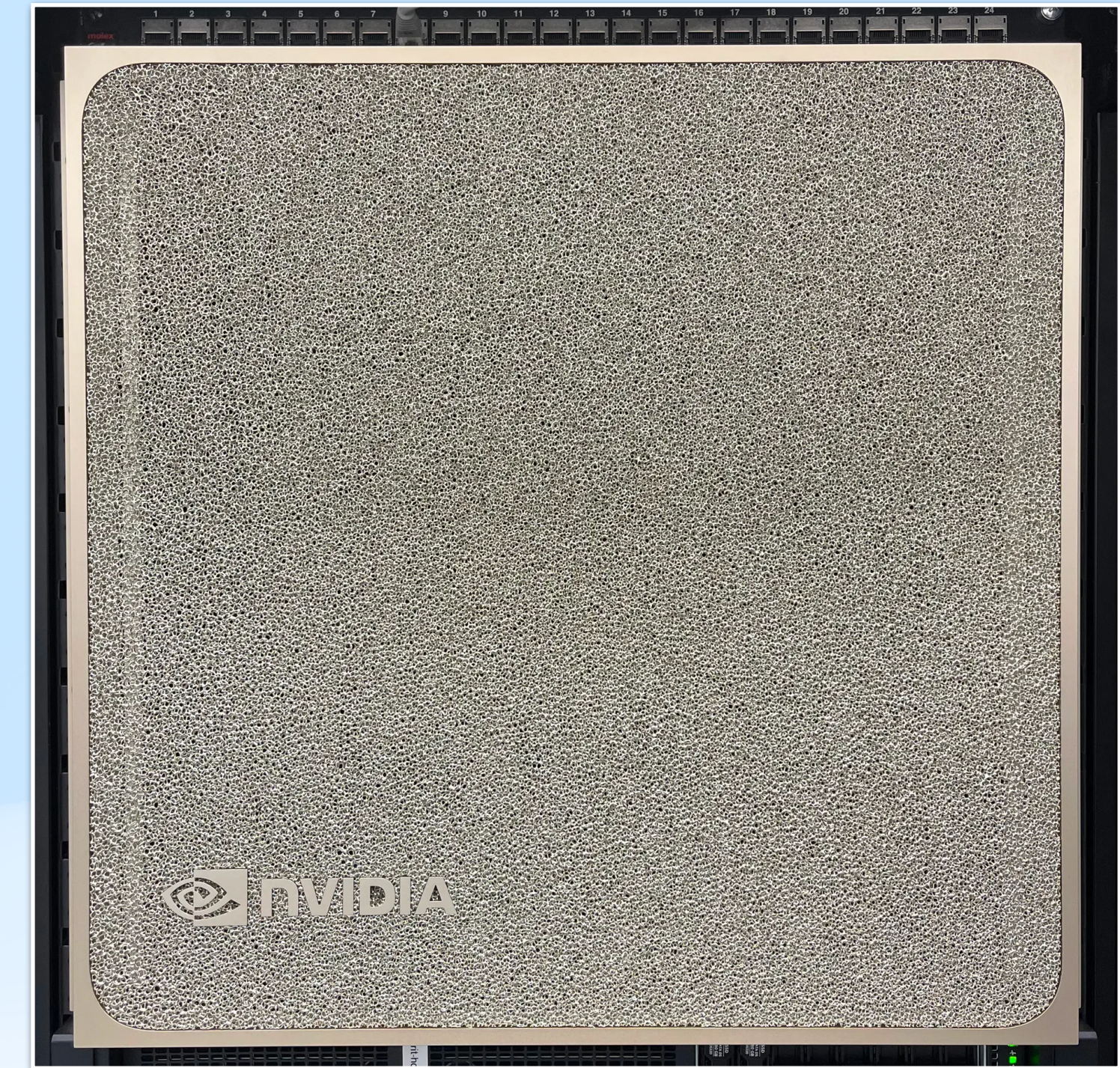
Na čem to běží

- **Servery**

- NVIDIA DGX B200 (GLM-5.3)
- NVIDIA DGX B300 (Kimi-K3)

- **GPU**

- 4x NVIDIA H200 (DeepSeek V4.1)
- NVIDIA H100 (GPT-OSS-120b)
- 4x NVIDIA RTX Pro 6000 (Mistral Medium 3.5, Qwen 3.8 Flash Next, Qwen 3.5)
- 2x NVIDIA RTX Pro 6000 (Gemma 4)
- 2x NVIDIA H100 (Qwen 3.8 27b)
- NVIDIA A40 (Embedding modely)
- NVIDIA A100 40GB (Whisper)



Chat

<https://chat.ai.e-infra.cz> – OpenWeb UI

- Přístup pro každého
 - S účtem v MetaCentru
 - Aktivního uživatele Masarykovy Univerzity (poskytla HW)
- Umožňuje
 - **Chatovat**
 - Generovat a editovat **obrázky** (Nový model: Qwen Image 2.1)
 - **Python Sandbox**
 - **Web Search** (vylepšený Searxng)
 - RAG (**Knowledge**)
 - Generovat **API klíč**

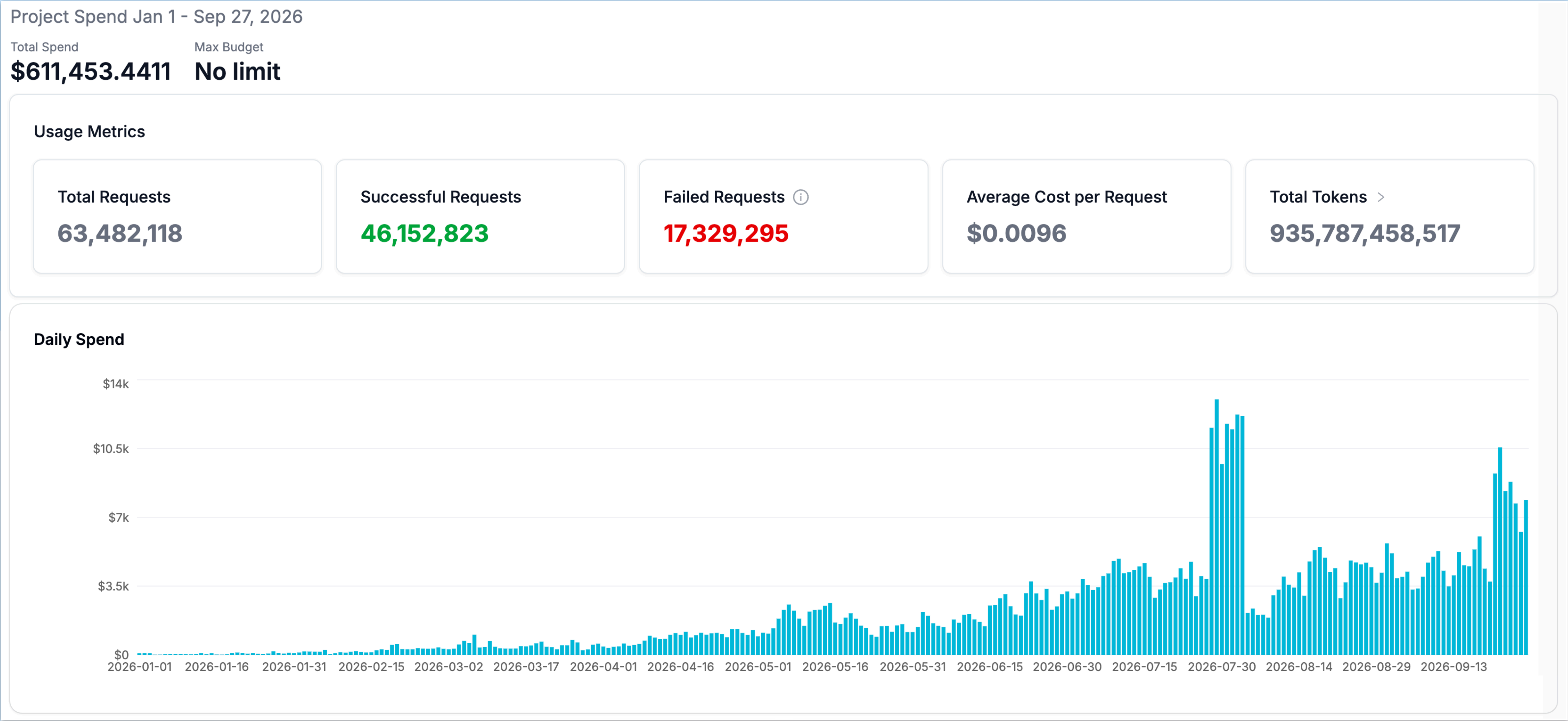


LLM API

<https://llm.ai.e-infra.cz/v1> – LiteLLM Proxy

- Klíče synchronizované z Chat služby
- **OpenAI** kompatibilní API
 - chat completions, responses – omezená funkcionality
 - Neumí deferred tools – používá Codex
- **Anthropic** kompatibilní API
 - messages
 - Neumí built-in tools, např. websearch

LLM API



LiteLLM API

Fair use

- Sledujeme průběžné **využití**, kontaktujeme extremisty

- Každý uživatel

- **4 paralelní požadavky**

- Aktuálně bez omezení na tokeny nebo počet požadavků v čase

- Může se změnit

- Neumíme přidat prioritu v konkrétním případě

- Není možné individuálně navyšovat limit např. na 20 požadavků

Top Virtual Keys

5 10 25 50

Table View Chart View

Key ID	Key Alias	Spend (USD)
fbae608cb0d...	z	\$2,836.34
152e78d5ace...	p	\$2,833.17
bf762a765cf...	k	\$2,171.41
f8aaca7557f...	n	\$2,103.01
82dfa04388e...	d	\$1,910.36
fdb9abeb318...	o	\$1,815.48
febb224baa3...	ri	\$1,810.25
08c5f50ea56...	s	\$1,436.31
8365b81a54a...	ti	\$1,422.61
e9a11647cf6...	k	\$1,296.46

LiteLLM API

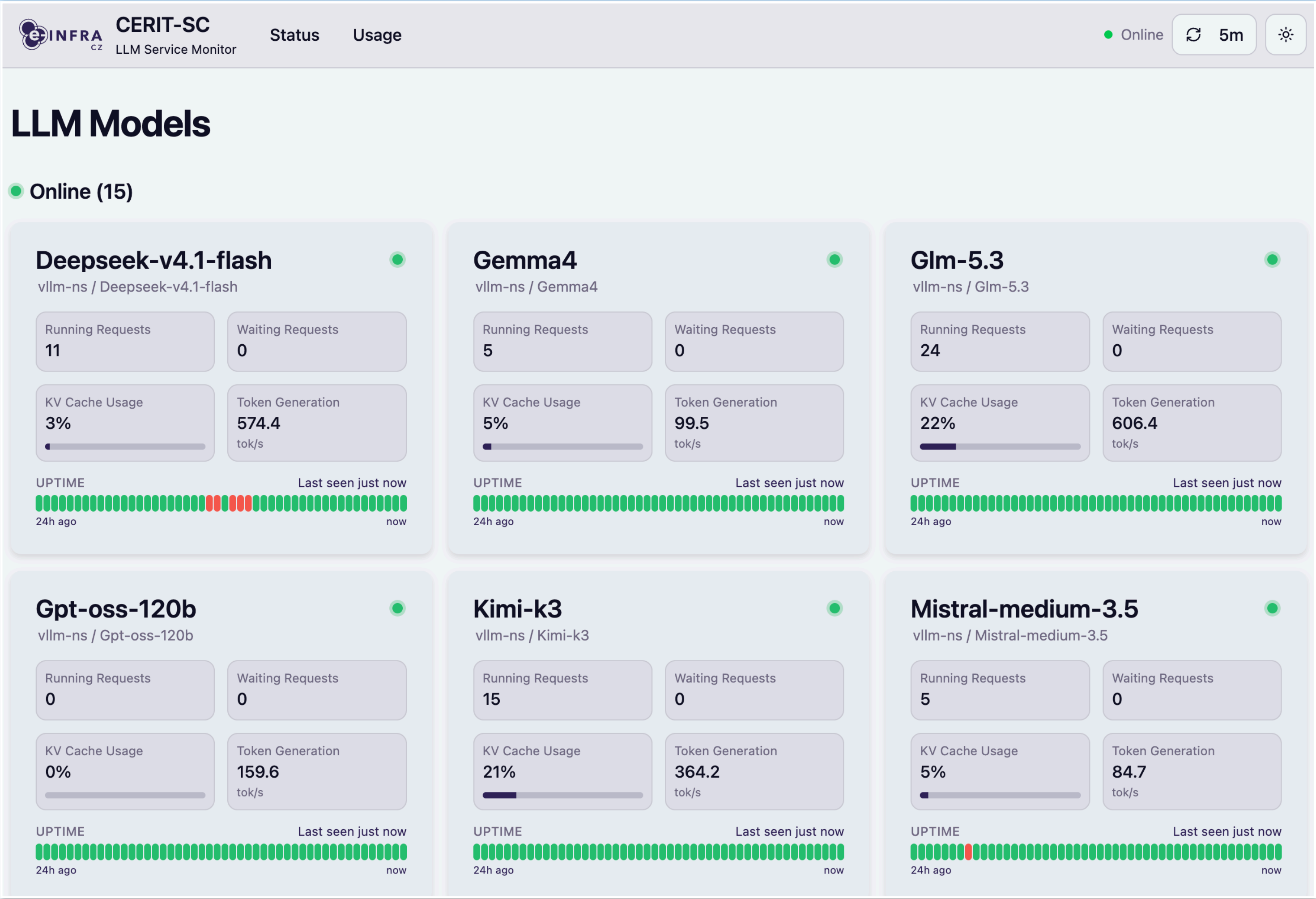
Integrace

- Funkční integrace popsané v dokumentaci
 - **Claude Code**
 - **OpenCode**
 - **Hermes**
 - CodeX
 - VS Code
 - JetBrains IDE
- Funkční nezdokumentované 🤔
 - Github **Copilot**, Kimi-cli, DeepSeek Harness



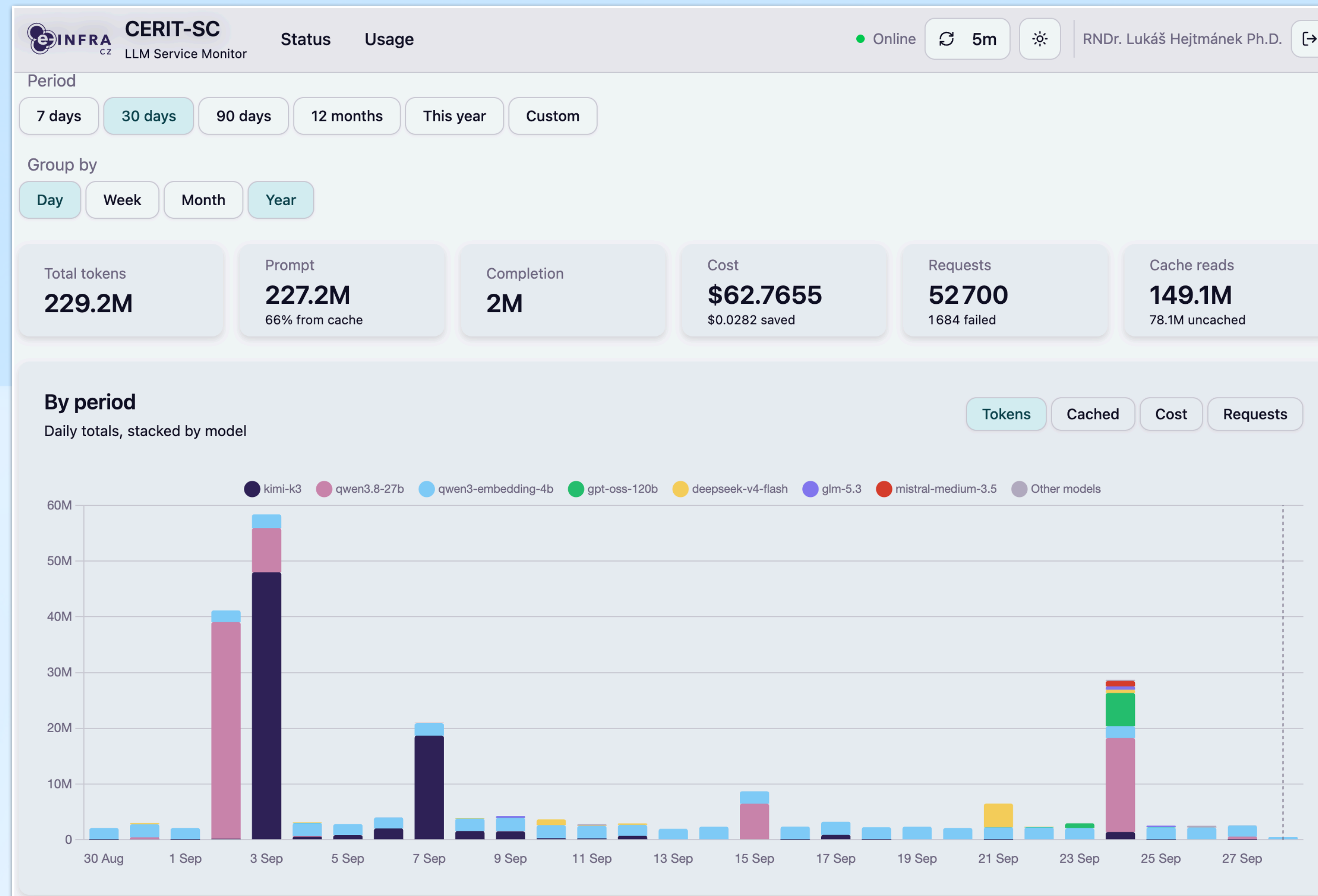
Stav LLM

<https://llm.ai.e-infra.cz/status>



Využití LLM

<https://llm.ai.e-infra.cz/usage>



Integrované služby

DeepSite – <https://deepsite.ai.e-infra.cz>

- AI generátor **webových stránek** nebo designů z textového popisu
- Vytvoří **HTML, CSS a JavaScript**
 - Může být vystaveno na web (manuálně)
- Ideální pro rychlou představu, jak web nebo aplikace (ne)má vypadat
- Může vytvořit i návrh posteru
- ... nebo stravenek na akci

Integrované služby

DeepSite – <https://deepsite.ai.e-infra.cz>

DeepSite PRO

2 edits

Desktop Mobile Refresh Preview Help

See Live Preview XH xhejtman

index.html

```
1 <!DOCTYPE html>
2 <html lang="cs">
3 <head>
4   <meta charset="UTF-8">
5   <meta name="viewport" content="width=device-width, initial-scale=1.0">
6   <title>Patologické seminární případy</title>
7   <script src="https://cdn.tailwindcss.com"></script>
8   <script src="https://cdn.jsdelivr.net/npm/feather-icons/dist/feather.min.js"></script>
9   <style>
10    @import url('https://fonts.googleapis.com/css2?family=Roboto:wght@300;400;500;700&display=swap');
11    body {
12      font-family: 'Roboto', sans-serif;
13      background: linear-gradient(135deg, #f8f9fc 0%, #f1f5f9 100%);
14    }
15    .card {
16      transition: all 0.3s ease;
17      box-shadow: 0 4px 6px rgba(0, 0, 0, 0.05);
18      border-radius: 12px;
19      overflow: hidden;
20      display: flex;
21      flex-direction: column;
22      height: 100%;
23    }
24    .card:hover {
25      transform: translateY(-5px);
26      box-shadow: 0 10px 25px rgba(0, 0, 0, 0.1);
27    }
28    .modal {
29      transition: opacity 0.3s ease;
30    }
```

Ask DeepSite for edits

Enhance kimi-k3 Attach Edit

Seminární patologické případy

1 Seminární případ č. 1

11450/25, muž 21 let radikální orchiectomie vlevo pro tumor.

Prezentuje MUDr. Matoušková



2 Seminární případ č. 2

C2653/24, Muž, 58 let, ascites při cirhóze.

Prezentuje MUDr. Matoušková



3 Seminární případ č. 3

PB 691/25, žena 30 let, tumor pravého ovaria.

Prezentuje MUDr. Said



14

Integrované služby

JupyterHub – <https://hub.cloud.e-infra.cz>

- Jupyter Notebook
 - **Notebook Intelligence** (Jupyter plugin, verze 6.1)
 - Claude Code
 - OpenCode
 - Hermes
- **RStudio**
 - Nativní asistent (**Posit Assistant**)
 - Lze nakonfigurovat proti našemu API od verze RStudio 4.6.1

Jupyter Notebook

Notebook Intelligence

- Součástí **DataScience AI** obrazu
- Před-konfigurované napojení na LLM API
- Formou chatu přímo v Jupyter Notebooku
 - Nově ikona **Chatbook** – lepší integrace AI do Notebooku včetně UI
- Na pozadí běží Claude Code

Notebook Intelligence

Chatbook

FileEditViewRunKernelGitTabsSettingsHelp

Notebook Intelligence

User15:08:17

create simple panda example into the current notebook

Claude Code15:09:17

5 tool calls

Adding markdown cell✓

Adding code cell✓

Adding code cell✓

Adding code cell✓

Saving notebook✓

Let me run the cells so you can see the output:

Running cell✓

Calling tool 'mcp__nbi__run-cell'...

Parameters

{"cell_index": 1}

Running cell✓

Calling tool 'mcp__nbi__run-cell'...

Parameters

{"cell_index": 2}

Running cell✓

Calling tool 'mcp__nbi__run-cell'...

Parameters

{"cell_index": 3}

Reading cell output✓

Ask Notebook Intelligence...

Untitled.ipynb

/📄🔗🛡️🌟▶️

Untitled.ipynb

📄+✂️📋📄▶️■↺↻Code⌵↶↷🔦🕒git

Open in...🛡️Chatbook [conda env:base] *🔔

Simple Pandas Example

A quick demo of creating a DataFrame and doing basic operations.

```
[2]: import pandas as pd

# Create a simple DataFrame from a dictionary
data = {
    "name": ["Alice", "Bob", "Charlie", "Diana"],
    "age": [25, 30, 35, 28],
    "city": ["Berlin", "Munich", "Hamburg", "Cologne"]
}
df = pd.DataFrame(data)
df
```

	name	age	city
0	Alice	25	Berlin
1	Bob	30	Munich
2	Charlie	35	Hamburg
3	Diana	28	Cologne

```
[3]: # Basic operations
print("Average age:", df["age"].mean())

# Add a new column
df["age_in_months"] = df["age"] * 12

# Filter rows
adults_over_28 = df[df["age"] > 28]
adults_over_28
```

Average age: 29.5

	name	age	city	age_in_months
1	Bob	30	Munich	360

1🛡️🌟Chatbook [conda env:base] * | IdleCPU: 5.00 % | Mem: 1.68 / 16.00 GBMode: Command🛡️Cell 4/5Ln 1, Col 1Untitled.ipynb1🔔

Jupyter Notebook

Agenti

- **Terminálové aplikace**
 - Claude code, OpenCode – spouštěcí ikona
 - Hermes agent – nutné spustit v terminálu
 - Dostupnost v terminálu **může trvat několik minut** – kopíruje data
- Před-konfigurováno na LLM API
- **Kontejnerový sandbox**
 - Agenti dosáhnou pouze na to, co je v kontejneru

Jupyter Notebook

Agenti

FileEditViewRunKernelGitTabsSettingsHelp

OC | Greeting

@

hello

+ Thought: 149ms

Hello! How can I help you today?

Build · agentic · 8.0s

Build · agentic LiteLLM

/home/jovyan

8.0K ctrl+p commands

Claude Code

Claude Code v2.1.283

agentic · API Usage Billing

/home/jovyan

✓ Using flicker-free rendering · if you want to go back, use /tui default

· Click to move your cursor in the text input

· Click to expand collapsed tool results

· By default, text auto-copies when you select it (/config to change)

· Hold Shift while selecting to use your terminal's native copy instead

hello

Thought for 3s

Hello! What would you like to work on?

I'm in plan mode, so if you've got a task in mind I'll explore the codebase and put together an approach for your approval before touching anything.

* Cogitated for 4s · done 1:34 PM

>

» accept edits on (shift+tab to cycle) · ← for agents

Terminal 3

@



clarify: clarify

code_execution: execute_code

connections: manage_connections

file: patch, read_file, search_files, write_file

memory: memory

other: tool_call, tool_describe, tool_search

skills: skill_manage, skill_view, skills_list

terminal: terminal

(and 1 more toolsets...)

Available Skills

autonomous-ai-agents: claude-code, +4 more

creative: architecture-diagram, +9 more

devops: sdlc-review

email: email-inbox-triage, himalaya

media: gif-search, songsee, +1 more

note-taking: obsidian

productivity: airtable, box, +12 more

research: arxiv, +3 more

social-media: xurl

software-development: codebase-inspection, +11 more

web: blocked-page-recovery

16 tools · 54 skills · /help for commands

▲ 21 commits behind – run hermes update to update

Welcome to Hermes Agent! Type your message or /help for commands.

+ Tip: Ctrl+C interrupts the agent. Double-press within 2 seconds to force exit.

● hello

Initializing agent...

▼ Hermes

Hello – what can I help you with?

▼ kimi

10.1K/262.1K [] 4% 9.1s ↑... Simple friendly greeting

> type / to browse commands, or Ctrl+P for the palette

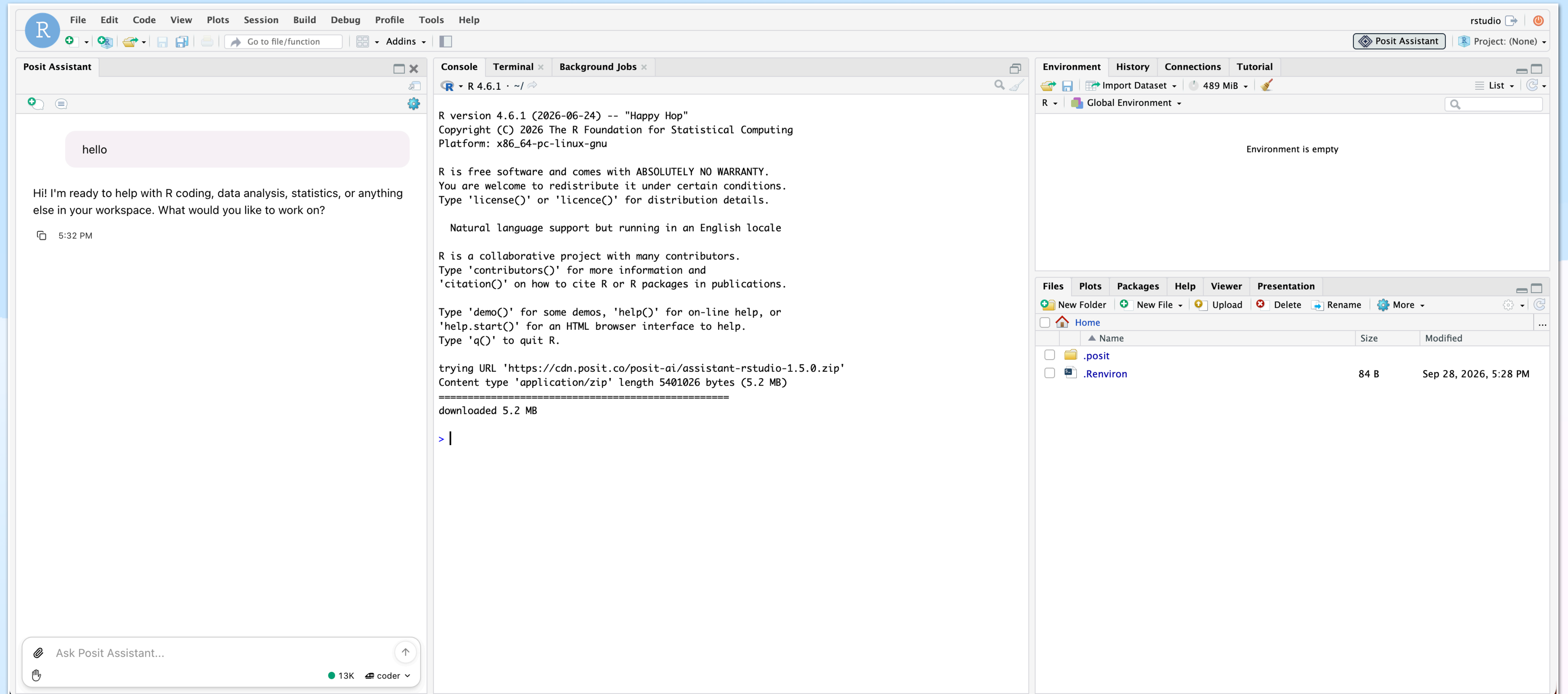
3 0 CPU: 0.00 % | Mem: 2.50 / 16.00 GB

Terminal 3 0

Jupyter – RStudio

- Dříve **GPT Studio** rozšíření – nebylo moc povedené
- Od verze 4.6.1 konfigurovatelná **Posit AI**
 - Nepodařila se před-konfigurovat
 - Nutná manuální
 - Instalace – automatická přímo z UI
 - Zadání URL: <https://llm.ai.e-infra.cz/v1>
 - Vložení API klíče (proměnná prostředí `OPENAI_API_KEY`)

Jupyter – RStudio



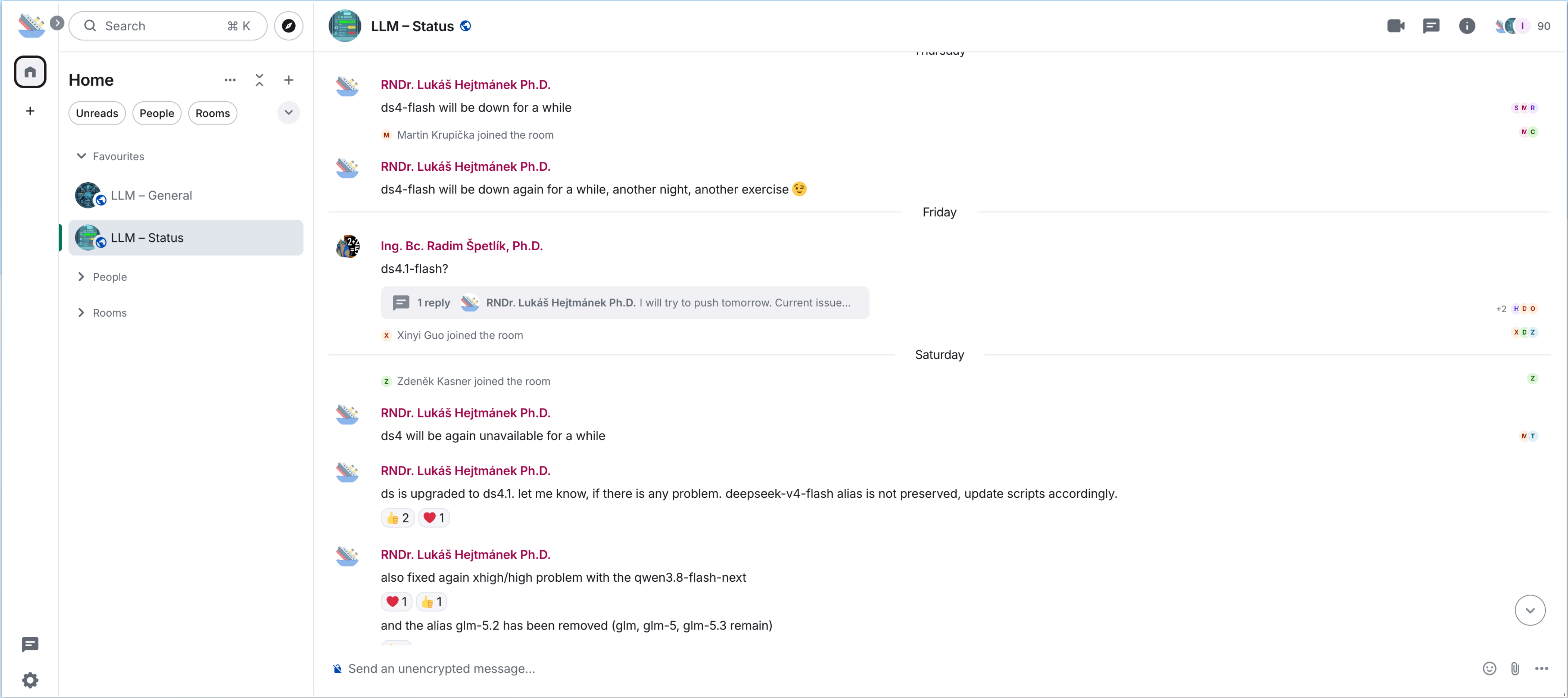
Komunita

<https://element.cloud.e-infra.cz>

- **Diskusní fórum** á-la Discord
 - Web klient **Element**, mobilní aplikace **Element-X**, případně další
- Používá otevřený protokol **Matrix**
 - Nepodporujeme aktuálně federaci
- Hlavní místnosti
 - **LLM General** – obecné informace, novinky
 - **LLM Status** – informace o aktuálních změnách, výpadcích
- Možnost založit vlastní místnost
- Media služby (volání) nefungují

Komunita

Element Web

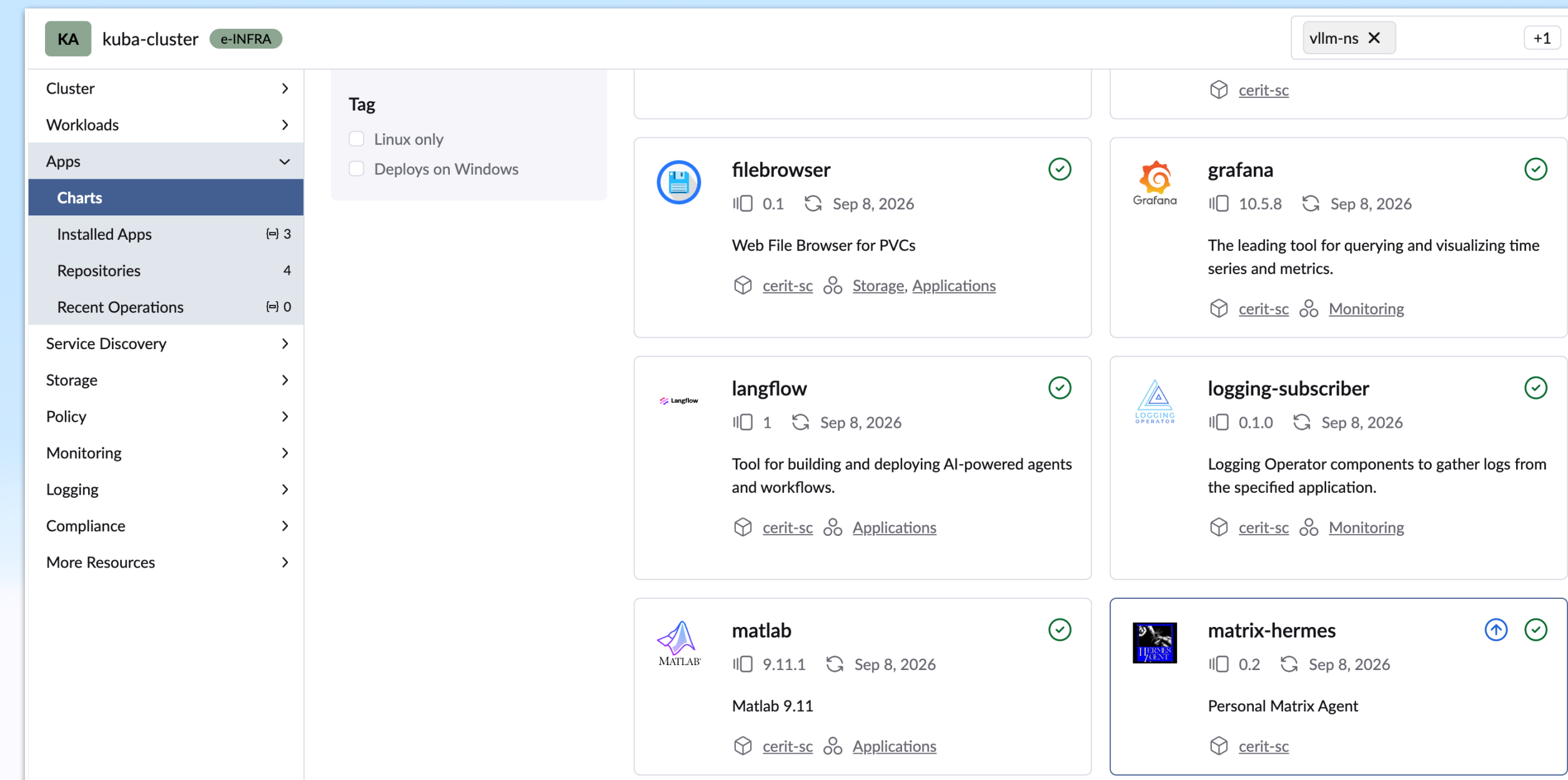


AI Agent v Matrixu

- Hermes agent jako **chatbot** pro **matrix** (element)
 - Lze přes chat předávat soubory, nastavovat cronjob
- **Veřejně dostupný**
 - Přihlásit se do Matrixu (<https://element.cloud.e-infra.cz>)
 - Vyrobit si soukromou místnost (Nezapínat šifrování!)
 - Vyrobit si údaje pro agenta (<https://app.matrix.cloud.e-infra.cz/>)
 - Spustit agenta jako Rancher aplikaci v K8s (<https://rancher.cloud.e-infra.cz>)
 - Oslovit agenta v místnosti

AI Agent v Matrixu

- Rancher aplikace
- Spuštěná v konkrétním namespace
- Režim práv
 - **Žádná** – přístup do internetu bez K8s
 - **Pouze čtecí** – Internet a čtecí práva K8s API
 - **Plná práva** – Internet a plná práva K8s API
 - Aktualizací aplikace lze práva dodatečně změnit



AI Agent v Matrixu

Cronjob Response: The E-INFRA Matiné
(job_id: 9c8079aa02d1)

THE E-INFRA MATINÉ

Pondělí 28. září 2026 · „Nejlepší zprávy jsou ty, které nepotřebují tiskovou konferenci.“

What's Fresh on the Cluster

Lady Michaelo, cluster dnes opět spí. Prohlédl jsem novinky na e-infra.cz i IT4Innovations — na e-INFRA je poslední položka ze 7. září (kurz MPI), na IT4I ze 17. září (ohlédnutí za CERN.AI festivalem). Obojí už Matiné servírovalo. Žádná nová výzva, žádný nový kurz, žádná reorganizace — jen obvyklé pondělní ticho, kdy se i superpočítače tváří, že víkend nebyl.

Dovolím si však jednu korekci včerejšího stolu: **38. Open Access Grant Competition IT4Innovations** má uzávěrku **19. října 2026**. Matiné ji dosud nezmínilo — což je zhruba tak výmluvné, jako když děkanát zapomene oznámit termín fakultní rady. Opravuji: je to pravidelná soutěž o výpočetní čas na Karolině, Barboře a LUMI, třikrát ročně, pro zaměstnance výzkumných organizací. Kdo nestihl předchozí kola, má zde časovou rezervu.

Deadlines & Callout

Datum	Co
30. 9.–1. 10. 2026	CESNET 2026 + MetaCentrum HPC Seminar, hotel Diplomat, Praha — začíná pozítří
13. 10. 2026	LUMI AI Factory: <i>Vision models use cases with AI</i> , prezenčně v Ostravě
19. 10. 2026	Uzávěrka 38. Open Access Grant Competition IT4I — přihlášky na Karolinu, Barboru i LUMI
21.–22. 10. 2026	<i>Introduction to MPI</i> , online
27. 10. 2026	Access to VLQ, online — kvantový počítač, který je vždy v superpozici mezi „mám zprávy“ a „nemám nic“

Nejbližší klakson: středeční CESNET 2026. Kdo se včera nepřihlásil, deleguje zástupce — proces osvědčený generacemi akademických funkcionářů.



AI Agent v Matrixu

M

Medgen-helper



M

Mgr. Kamila Reblova Ph.D.



10:52

N

nfops bot

: ubmitted process > DATABAZEcp (kopirovani gK_x_VGO-23-1575 do database)
ERROR ~ Unexpected error while finalizing task 'DATABAZEcp (kopirovani gK_x_VGO-23-1575 do database)' - cause: Failed to create publish directory: /cmbg/AvitiDN/DatabaseExomy/input/260828_Exom_gK

-- Check '.nextflow.log' file for details
WARN: Killing running tasks (8)
WARN: Failed to render execution report -- see the log file for details
Kubernetes Job 'few-juicer' created successfully. here is process:
process DATABAZEcp {
tag "kopirovani \$name do database"


```
1   publishDir "/cmbg/AvitiDN/DatabaseExomy/input/${sample.run}", mode:'copy'  
2  
3   input:  
4   tuple val(name), val(sample), path("${name}.merged.txt")  
5  
6   output:  
7   tuple val(name), val(sample), path("${name}.merged.txt")  
8  
9   script:  
10  ""  
11  
12  echo "kopiruju $name"  
13  ""  
  
}
```

 1

 5 replies

N

nfops bot

Found it — this is a missing storage mount, not a task failur...

M

Mgr. Kamila Reblova Ph.D.

great, i used the first suggestion: 1
publishDir "/cephfs/cmbg/AvitiDN/DatabaseExomy/input/\${sample.run}", mode:'copy'

and it worked! data are in Anicka's database

 1

Datová suverenita

- Všechny nabízené modely běží na **infrastruktuře CERIT-SC a Elixir CZ**
- Nad uživatelskými daty se model **neumí dotrénovat/učit**
 - Chat služba má možnost prohledávat starší konverzace daného uživatele
- Uživatelská data se nikam **neposílají**
 - Netýká se hledání na internetu – služba třetí strany
- Uživatelské vstupy a výstupy se **nelogují**
- Každý SW má chyby

Co dál?

- Dva běžící FR CESNET projekty
 - **Knowledge on Demand**
 - „one-click“ aplikace pro RAG
 - **Biomni**
 - Webový agentický systém pro Biomedicínu
- **Nový HW**
 - NVIDIA DGX z Elixiru
 - AMD MI350 z Metacentra/Cesnetu
- Prodloužení přihlášky
 - Doplnění krátké **zprávy o využití AI**

Děkuji za pozornost

k8s@cerit-sc.cz

<https://docs.cerit.io>

<https://blog.e-infra.cz>