# HUN-REN Cloud GenAI4Science szolgáltatás



Farkas Attila
farkas.attila@sztaki.hun-ren.hu
HUN-REN SZTAKI

# GenAI4Science szolgáltatás

- Nagy nyelvi modell (LLM) alapú csevegő szolgáltatást nyújt (mint a ChatGPT)
  - Teljesen a HUN-REN Cloud erőforrásokon biztosítva
- Nyílt forráskódú szoftverkomponensekre épül
  - Nem függ harmadik fél szolgáltatásaitól
  - De képes kommunikálni OpenAI API-n keresztül más szolgáltatásokkal
- Adatvédelem biztosított
  - Az adatok nem hagyják el a HUN-REN Cloudot
- Egyszerre több LLM modellt is támogat

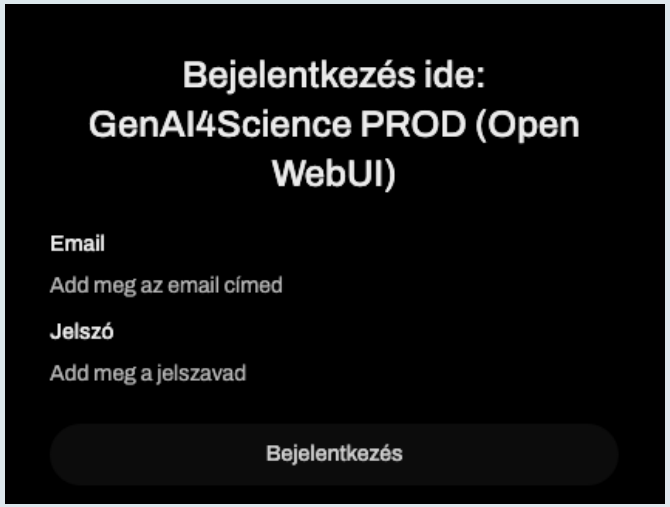
# GenAI4Science szolgáltatás felépítése

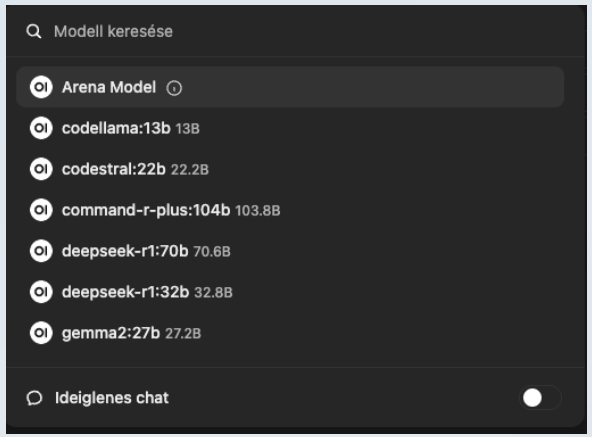
- **Ollama**
  - Támogat számos LLM modellt (llama, deepsek, gemma, mistral, stb.)
  - Modell katalógust biztosít
  - Támogatja a testreszabott modelleket
  - REST API-t biztosít a modellek eléréséhez
  - GPU-alapú modell futtatás (korlátozott a HUN-REN Cloudon)
- **Open-WebUI**
  - Felhasználóbarát felület
  - Felhasználói jogosultság kezelés
  - Többfelhasználós támogatás
  - OpenAI API alapú elérés támogatása (integrációs lehetőségek)
  - RAG támogatása
  - Helyi, távoli, webes keresés (Google, DuckDuckGo stb.)

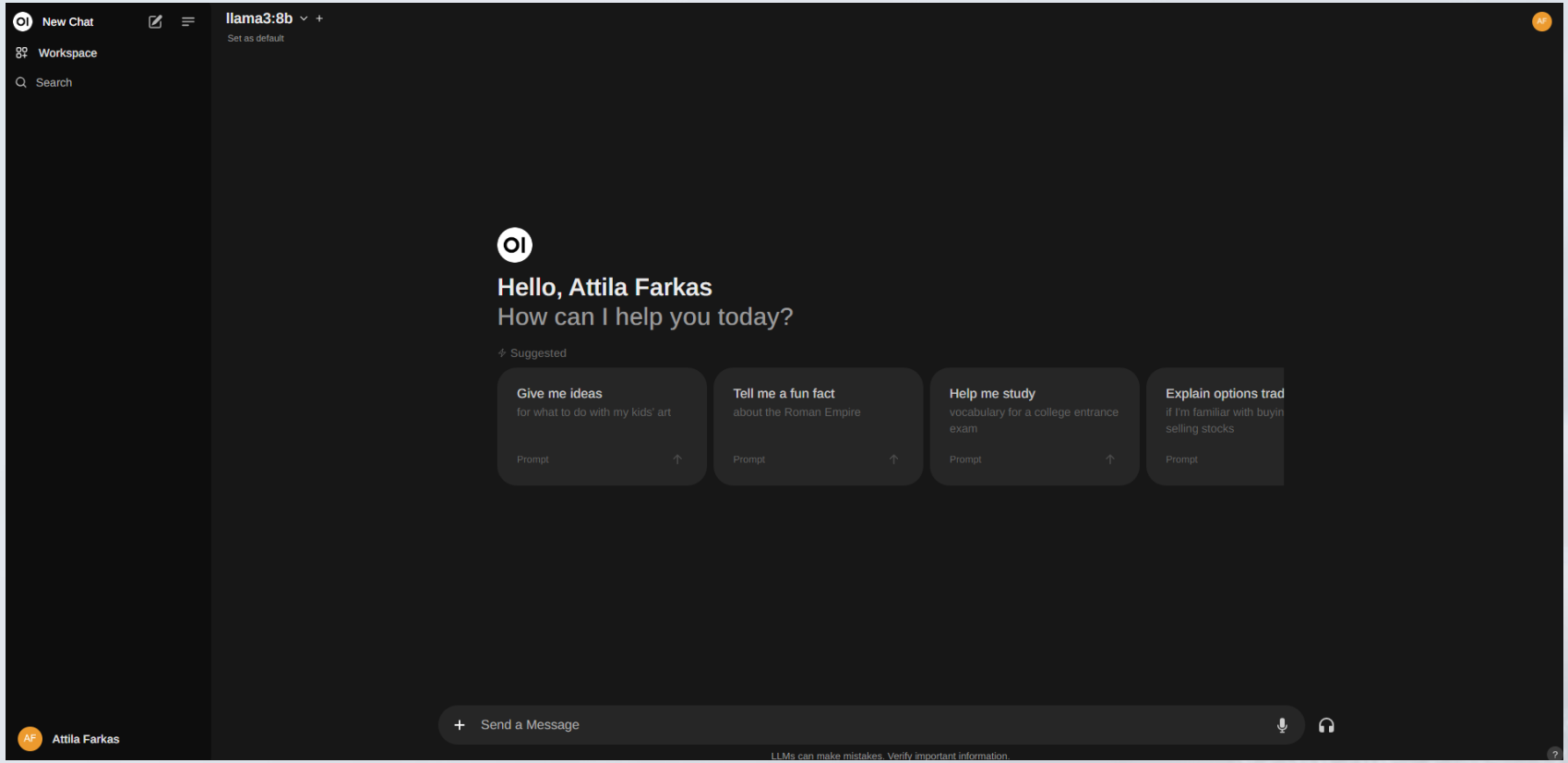
# Szolgáltatás architektúrája

| Instance Name ▲ | Image Name | IP Address | Flavor |
|---|---|---|---|
| genai-db | Ubuntu 24.04 LTS | 192.168.0.109 | m2.xlarge |
| genai-frontend | Ubuntu 24.04 LTS | 192.168.0.112 | m2.xlarge |
| genai-h100-1 | Ubuntu 24.04 LTS | 192.168.0.18 | g3.3xlarge |
| genai-service | Ubuntu 22.04 LTS - NV | 192.168.0.194 | g2.2xlarge |
| genai-v100-2 | Ubuntu 22.04 LTS - NV | 192.168.0.33 | g2.2xlarge |
| genai-v100-3 | Ubuntu 22.04 LTS - NV | 192.168.0.149 | g2.2xlarge |
| monitoring-server | Ubuntu 22.04 LTS | 192.168.0.12 | m2.large |
| proxy-caddy | Ubuntu 24.04 [20240710] | 192.168.0.105, 193.225.250.35 | m2.large |
| rag | Ubuntu 22.04 LTS | 192.168.0.186, 193.225.250.75 | m2.2xlarge |

# DEMO – Bejelentkezés, modell választás

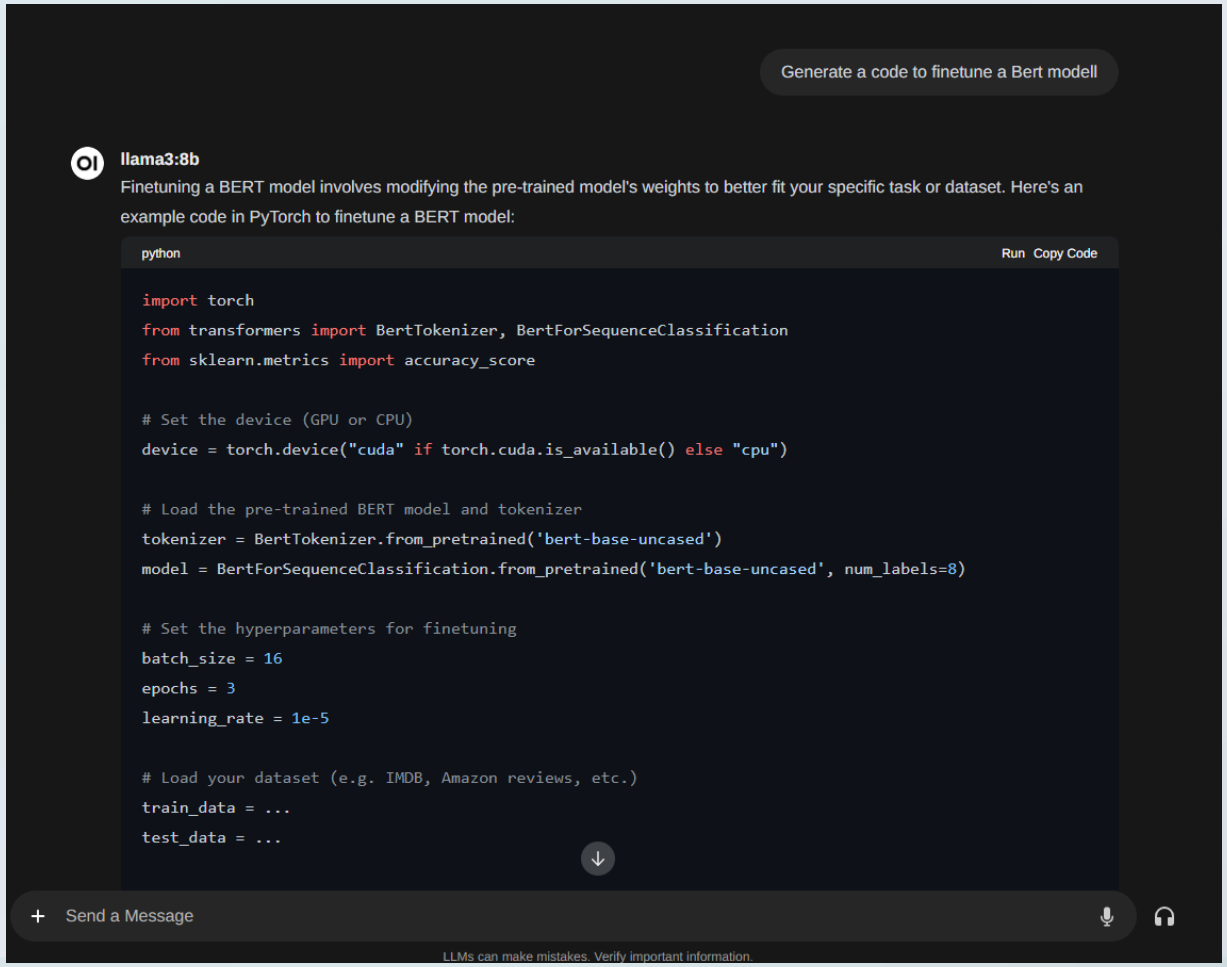


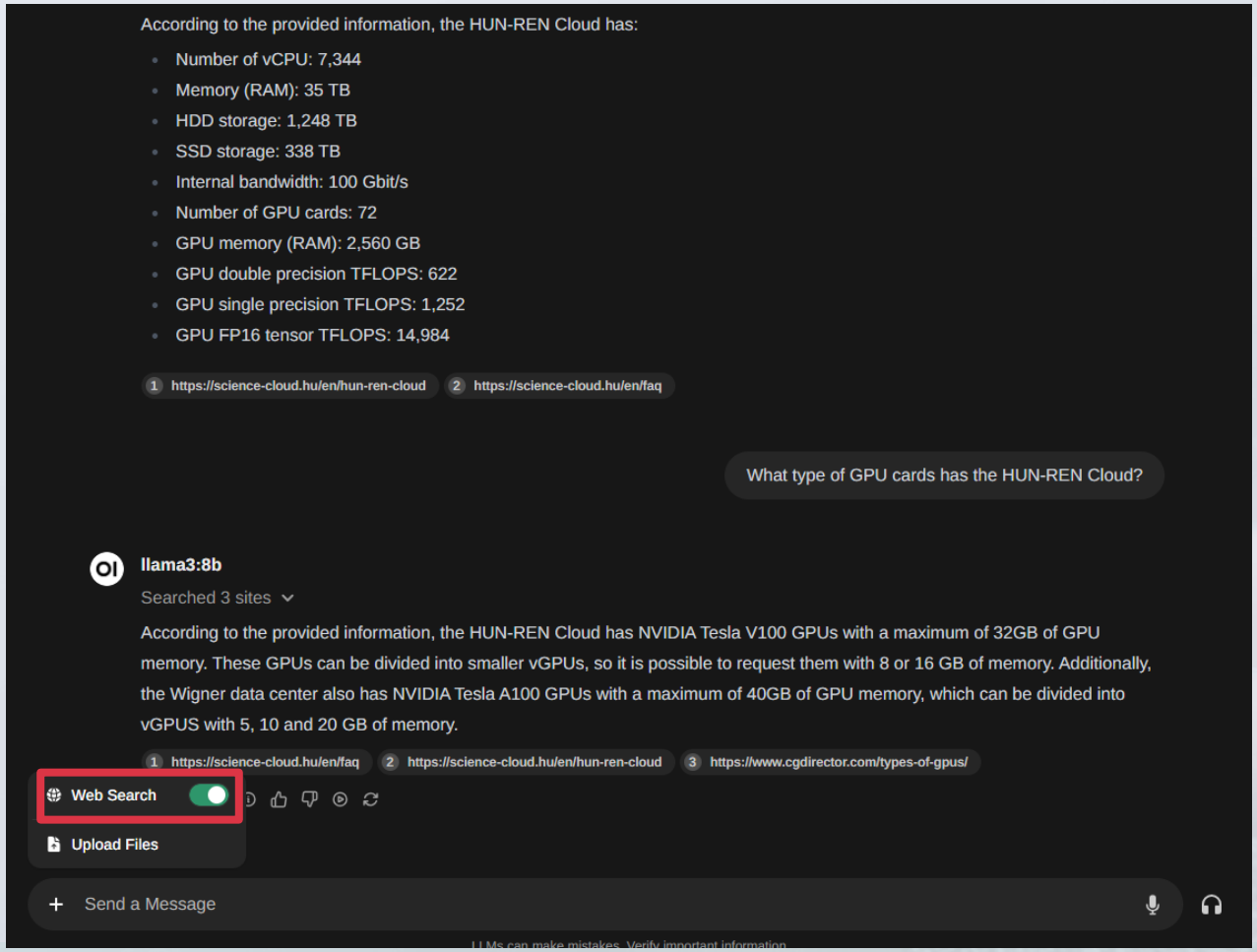




https://genai.science-cloud.hu

# DEMO – chat felület, RAG



# Hozzáférés igénylés

HUN-REN Cloud Rólunk Aktualitások Projektek Dokumentáció GenAI4Science

Hozzáférés igénylés

Jelenleg korlátozott erőforrás áll rendelkezésre a szolgáltatás mögött, ezért csak korlátozott számban tudunk hozzáférést biztosítani.
Az erőforrás korlátok miatt jelenleg csak a HUN-REN Kutatóhálózat kutatói számára tudunk hozzáférést biztosítani, intézmény méret arányos módon.

Belépés a GenAI4Science portálon keresztül lehetséges a jóváhagyott igénylést követően.

Hozzáférés igénylése | GenAI4Science dokumentáció


## GenAI4Science szolgáltatás igénylése

A GenAI4Science szolgáltatáshoz az alábbi adatok megadásával igényelhető hozzáférés. További információ a GenAI4Science-ről.

**Name** *

Farkas Attila

**E-mail cím** *

farkas.attila@sztaki.hu

**Intézmény** *

HUN-REN intézmények

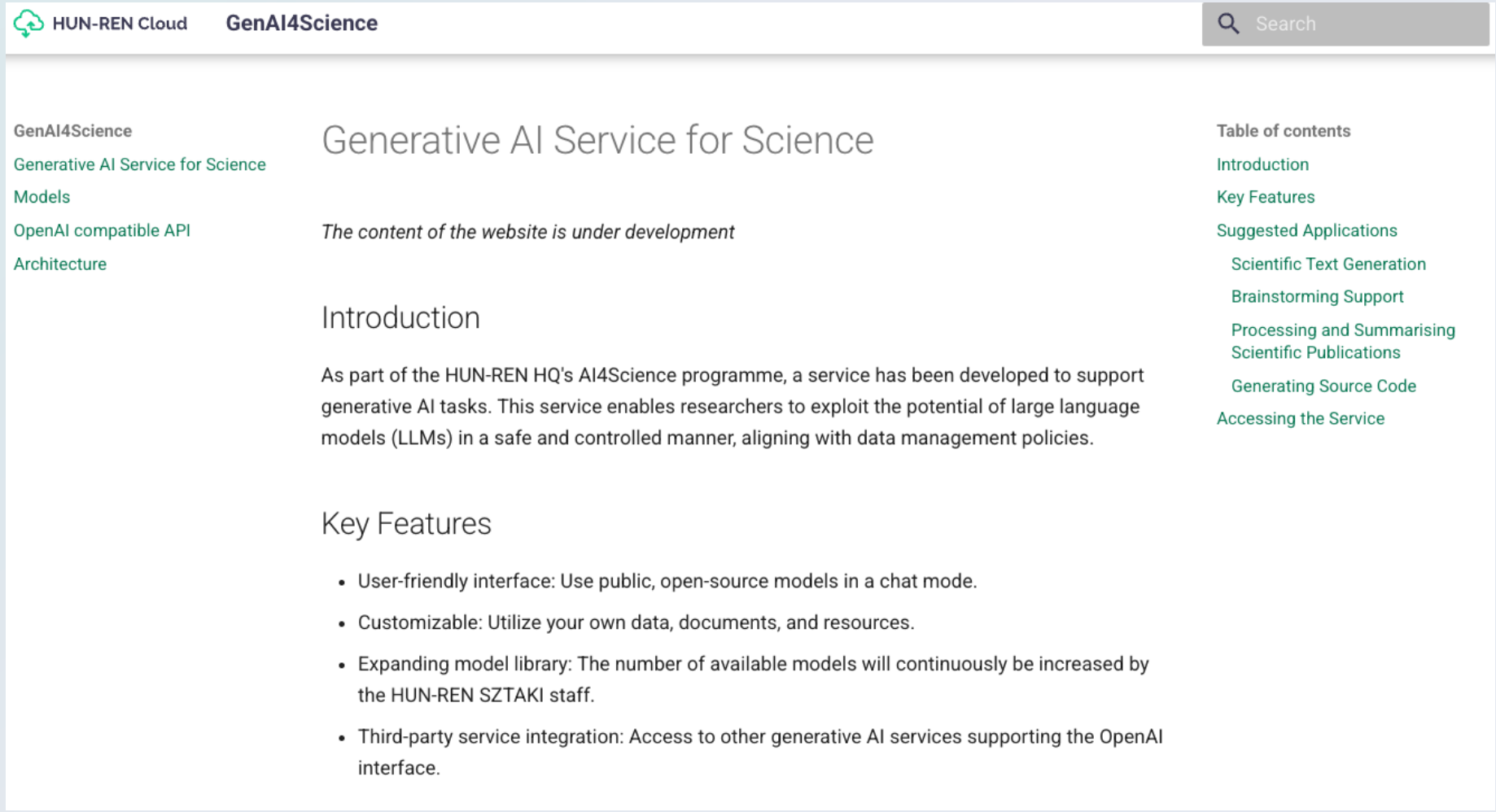
Számítástechnikai és Automatizálási Kutatóintézet

**Megjegyzés**

☐ Az Adatkezelési Nyilatkozatot elolvastam, és elfogadom

Igény leadása

# Dokumentáció




https://doc.genai.science-cloud.hu/

# Használati esetek

MESTERSÉGES INTELLIGENCIA
Nemzeti Laboratórium

- AI4Science program támogatása
- Mesterséges Intelligencia Nemzeti Labor támogatása


# Köszönöm a figyelmet!

Kérdések?