

# HUN-REN Cloud GenAI4Science szolgáltatás

Farkas Attila

[farkas.attila@sztaki.hun-ren.hu](mailto: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, deepseek, 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

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



# DEMO – Bejelentkezés, modell választás

**Bejelentkezés ide:  
GenAI4Science PROD (Open  
WebUI)**

**Email**  
Add meg az email címed

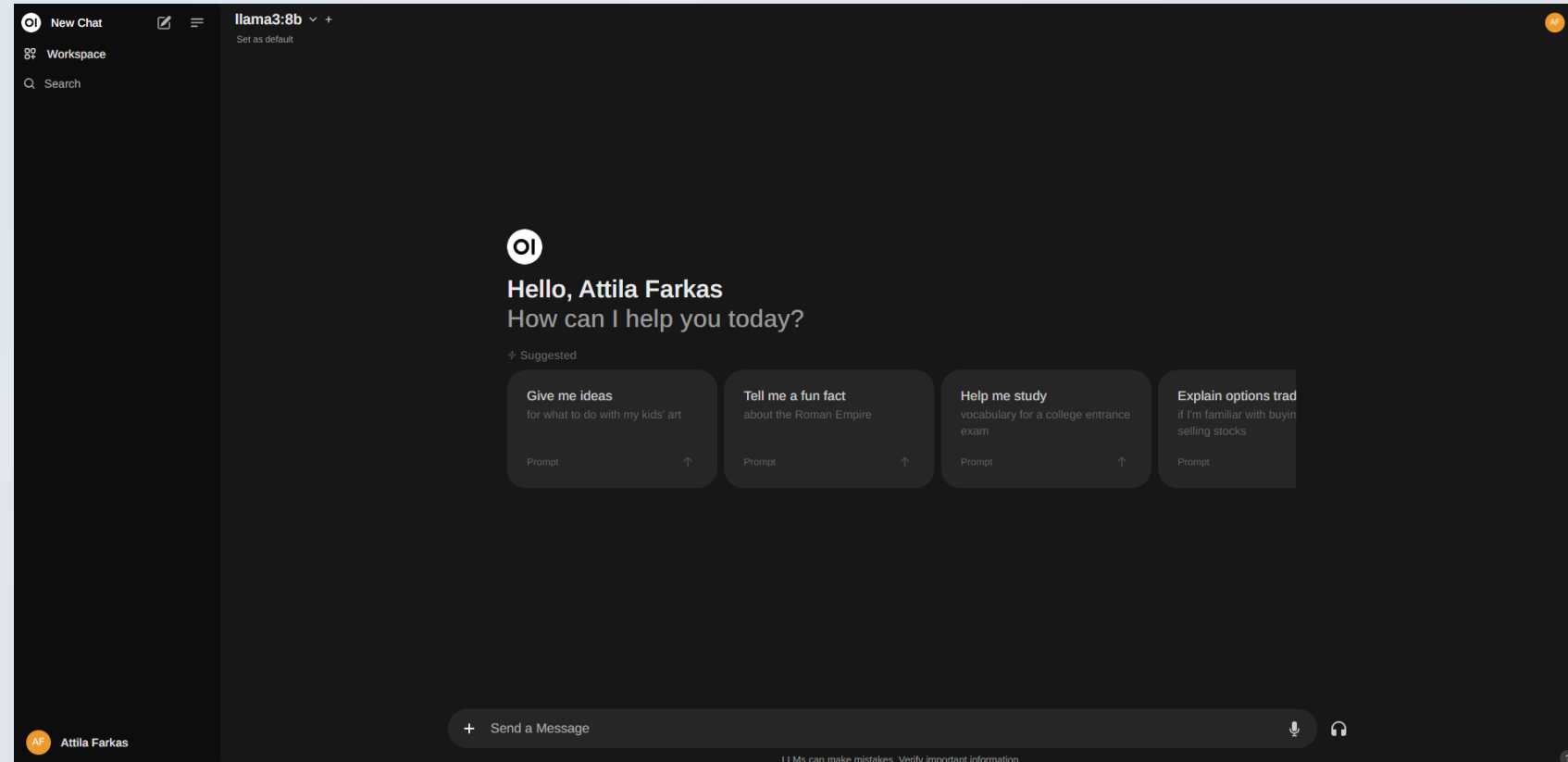
**Jelszó**  
Add meg a jelszavad

**Bejelentkezés**

Modell keresése

- Arena Model
- codellama:13b 13B
- codestral:22b 22.2B
- command-r-plus:104b 103.8B
- deepseek-r1:70b 70.6B
- deepseek-r1:32b 32.8B
- gemma2:27b 27.2B

Ideiglenes chat



<https://genai.science-cloud.hu>

# DEMO – chat felület, RAG

Generate a code to finetune a Bert modell



llama3:8b

Finetuning a BERT model involves modifying the pre-trained model's weights to better fit your specific task or dataset. Here's an example code in PyTorch to finetune a BERT model:

python Run Copy Code

```
import torch
from transformers import BertTokenizer, BertForSequenceClassification
from sklearn.metrics import accuracy_score

# Set the device (GPU or CPU)
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

# Load the pre-trained BERT model and tokenizer
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
model = BertForSequenceClassification.from_pretrained('bert-base-uncased', num_labels=8)

# Set the hyperparameters for finetuning
batch_size = 16
epochs = 3
learning_rate = 1e-5

# Load your dataset (e.g. IMDB, Amazon reviews, etc.)
train_data = ...
test_data = ...
```

+ Send a Message

LLMs can make mistakes. Verify important information.

According to the provided information, the HUN-REN Cloud has:

- Number of vCPU: 7,344
- Memory (RAM): 35 TB
- HDD storage: 1,248 TB
- SSD storage: 338 TB
- Internal bandwidth: 100 Gbit/s
- Number of GPU cards: 72
- GPU memory (RAM): 2,560 GB
- GPU double precision TFLOPS: 622
- GPU single precision TFLOPS: 1,252
- GPU FP16 tensor TFLOPS: 14,984

1 <https://science-cloud.hu/en/hun-ren-cloud> 2 <https://science-cloud.hu/en/faq>

What type of GPU cards has the HUN-REN Cloud?



llama3:8b

Searched 3 sites

According to the provided information, the HUN-REN Cloud has NVIDIA Tesla V100 GPUs with a maximum of 32GB of GPU memory. These GPUs can be divided into smaller vGPUs, so it is possible to request them with 8 or 16 GB of memory. Additionally, the Wigner data center also has NVIDIA Tesla A100 GPUs with a maximum of 40GB of GPU memory, which can be divided into vGPUS with 5, 10 and 20 GB of memory.

1 <https://science-cloud.hu/en/faq> 2 <https://science-cloud.hu/en/hun-ren-cloud> 3 <https://www.cgdirector.com/types-of-gpus/>

Web Search

Upload Files

+ Send a Message

LLMs can make mistakes. Verify important information.



# 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

HUN-REN Cloud GenAI4Science

Search

GenAI4Science  
Generative AI Service for Science  
Models  
OpenAI compatible API  
Architecture

## Generative AI Service for Science

*The content of the website is under development*

### Introduction

As part of the HUN-REN HQ's AI4Science programme, a service has been developed to support generative AI tasks. This service enables researchers to exploit the potential of large language models (LLMs) in a safe and controlled manner, aligning with data management policies.

### Key Features

- User-friendly interface: Use public, open-source models in a chat mode.
- Customizable: Utilize your own data, documents, and resources.
- Expanding model library: The number of available models will continuously be increased by the HUN-REN SZTAKI staff.
- Third-party service integration: Access to other generative AI services supporting the OpenAI interface.

Table of contents

- Introduction
- Key Features
- Suggested Applications
  - Scientific Text Generation
  - Brainstorming Support
  - Processing and Summarising Scientific Publications
  - Generating Source Code
- Accessing the Service

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



# Használati esetek

- AI4Science program támogatása
- Mesterséges Intelligencia Nemzeti Labor támogatása



# Köszönöm a figyelmet!

Kérdések?

