HUN-REN Cloud új GPU erőforrás menedzsment módszertana az AI4Science programban

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

AI4SCIENCE PROGRAM SZOLGÁLTATÁSI TÉRKÉPE

HUN-REN Magyar Kutatási Hálózat

INTÉZMÉNYI AI NAGYKÖVETI HÁLÓZAT

Milyen témákkal lehet a kutatóhelyi AI Nagykövethez fordulni?

Minden kutatóhelyen dedikált AI Nagykövet segíti a kutatókat, hogy a mesterséges intelligencia használatát beépítsék a kutatásaikba. A nagykövetek inspirációt adnak az AI használatához, elősegítik az új módszerek megismerését és támogatják, hogy minden kutató megtalálja a számára legnagyobb hozzáadott értéket, amelyet az AI hozhat.

AI OKTATÁS ÉS INSPIRÁCIÓ

Az AI milyen területein mélyíthetik el a kutatók a tudásukat?

Alapozó képzések
A cél az általános alapok elsajátítása, amelynek nyomán a kutatók tovább tudnak lépni személyes témák felé.

Specialista képzések
A hálózaton belül fejlesztett saját képzések, amelyek egy-egy tudományterület egyedi kérdéseit járják körül

Adatgazdász képzés

SZEMÉLYES AI TÁMOGATÁS

Hová lehet fordulni az AI-t igénylő kutatási ötlettel?

HUN-REN Központ AI szakértői csapat
1. kutatási ötletek validálási és kutatástervezési támogatása
2. infrastruktúra-támogatás
3. alkalmazástámogatás

Külső szakértői csapat
Neuron Solutions külső szerződéses partnerként

AI kutatási partnerkereső
AI kompetenciákkal rendelkező, más kutatási területek iránt nyitott kutatók + AI alkalmazási ötletekkel rendelkező kutatók.

ELÉRHETŐ AI TECHNOLÓGIÁK

Milyen biztonságos AI technológiák használhatók érzékeny kutatási területekhez?

Számítási infrastruktúra
1. Biztonságos GenAi futtatása, akár saját környezetben
2. Gépi tanulás elősegítése akár elosztott, több GPU kártya támogatásával

Saját, hazai AI keretrendszerek
1. HUN-REN Cloud
2. Komondor

Miért kellett új projekt és GPU igénylési lehetőségeket bevezetni?

- Az **AI előretörésével meredeken nőtt az igény** a GPU-t használó projektek indítására, így a kioszthető GPU kapacitás elfogyott
  - főleg a hosszabb, így gyakran hektikus/részleges kihasználtságú projektek miatt
- Egyre nagyobb igény mutatkozik még **újabb és kényelmesebb AI szolgáltatások**ra, ld. GenAI4Science
- Köszönhetően (többek között) az AI4Science programnak növekednek az AI kompetenciák, egyre **haladóbb és nagyobb felhasználási esetek** jelentek meg, ami más (kötegelt, job-alapú) megközelítést igényelnek
- A felhő **következő bővítésének ütemezése elhúzódik:** 2026

A GPU használati módok kibővítése új lehetőségekkel

1. IaaS mód
   - 1 éves igényű projekt
     - **Szolgáltatást nyújtó projekt**
     - **Ernyőprojekt**
   - 3 hónapos igényű projekt
     - Előkészítő projekt
     - Haladó projekt
2. **PaaS mód SLURM job manager alapon**
   - Interaktív üzemmód
   - Batch üzemmód
3. **GenAI4Science platform használata**

Új GPU elosztási rendszer

Az új lehetőségek mellett új GPU elosztási rendszer kialakítása is szükségessé vált

Olyan új rendszer bevezetése volt a cél, amely

- Elegendő erőforrás esetén úgy működik, mint az eddigi rendszer
- Ha viszont nincs elegendő erőforrás, akkor automatikusan átmegy egy a versenyhelyzetet fair módon lekezelő mechanizmusba

A versenyhelyzet kezelése 2 szempontból is fair:

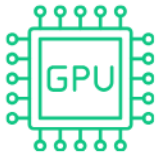
1. Nem kell adott esetben egy évet várni a projekt indítására, legkésőbb 3 havonta elbírálásra kerül minden projektigény
2. A versengő projektigények egységes szempontok (tudományos teljesítmény, impakt, erőforrásigény, stb.) alapján kerülnek elbírálásra

GPU Erőforrás-gazdálkodási Szabályzat

https://science-cloud.hu/gpu-eroforras-gazdalkodasi-szabalyzat

Felhasználó

HUN-REN Cloud support



Igény felmerülése

Igénylő űrlap beküldése a portálon

GPU igénylőlap csatolásával

Pályázat benyújtása

HUN-REN Cloud projekt futtatása

(Előző elbírálási időszakban jóváhagyott HUN-REN Cloud projekt futtatása)

HUN-REN Cloud projekt futtatása

Aktuális időszakban futó projektek befejezése/zárása, akik nem pályáztak újra hosszabbítás céljából vagy pályáztak, de elutasításra kerültek

Újrapályázás megfontolása/ egyéb megoldás keresése

PROJEKT FUTTATÁSI IDŐSZAK | **ELBÍRÁLÁSI IDŐSZAK** | **ZÁRÁSI IDŐSZAK** | **PROJEKT FUTTATÁSI IDŐSZAK**

Beadási határidő

Elbírálás vége

Új projektek indulása

Szabad kapacitások áttekintése

- Pályázó értesítése a pozitív elbírálásról
- Igény felfüggesztése az elbírálási időszak kezdetéig

- Tudományos érték?
- Szabad erőforrások?
- Hosszabbítás esetén eddigi elért eredmények?
- Az új- és hosszabbítási igények azonos súllyal esnek latba

Igény elbírálása

- Pályázó értesítése a pozitív elbírálásról
- Pályázó értesítése az elutasításról

Adatgazdász képzés

GPU igénylőlap

- A projekt igénylés során megadandó új mezők
- A GPU igénylőlap feltöltésének helye

GPU nélküli erőforrások használatának tervezett vége *

mm / dd / yyyy

Maximum 6 hónap + két hét az igénylés napjától számítva előkészítő és haladó projektek esetében, valamint maximum 1 év ernyő és szolgáltató projektek esetében. Ha nincs szüksége a projekttípusához maximálisan igényelhető teljes időtartamra, kérjük, válasszon rövidebbet.

GPU-s erőforrások használatának tervezett vége *

mm / dd / yyyy

Előkészítő és haladó projektek esetén maximum a következő projekt futtatási időszak vége adható meg tervezett zárási dátumnak, ennél hosszabb (maximum 1 éves) időtartam csak szolgáltató és ernyő projektek esetében állítható be. Kérjük a GPU igénylapon kitöltötteknek megfelelő dátumot adjon meg. Ha nincs szüksége a projekttípusához maximálisan igényelhető teljes időtartamra, kérjük, válasszon rövidebbet.

▼ **GPU igénylőlap**

Kérjük, töltse le a GPU igénylőlap nyomtatványt és kitöltve csatolja az online kitöltött igénylő űrlap beküldésekor. A GPU-erőforrásokra vonatkozó kapcsolódó szabályok részleteit a GPU Erőforrás-gazdálkodási Szabályzatunkban találják.

Új fájl hozzáadása

Browse... No files selected.

Ezzel a mezővel korlátlan számú fájl tölthető fel.
4 MB korlát.
Engedélyezett típusok: docx.

Speciális projektek

- Szolgáltatást nyújtó projekt:
  - Nyilatkozat tétele: Tudomásul veszem, hogy a HUN-REN Cloud best effort jellegű szolgáltatást nyújt SLA nélkül és ennek megfelelően az igényelt projekt által nyújtott szolgáltatás is best effort jellegű lesz SLA nélkül. Kijelentem, hogy a HUN-REN Cloud felé semmilyen SLA követelésem nincs és nem is lesz.
- Ernyőprojekt:
  - Meg kell adni, hogy min. hány projekt fog működni az ernyőprojekt keretében és ebből hány haladó projekt van
  - Amely projektekről már lehet tudni, ott megadni a projekt címét, résztvevőit és hogy hány és milyen publikációt, ill. disszeminációs tevékenységet vállalnak.
- Haladó projekt:
  - Meg kell adni, hogy miért nevezheti magát haladónak a projekt:
    - Előkészítő projekt sikeres befejezése
    - A projekt tagoknak legalább egyike már futtatott haladó projektet a HUN-REN Cloudon

Fontos időpontok

- **Pályázási határidők**
  - február 28.
  - **május 31.**
  - aug. 31.
  - nov. 30.
- **3 hónapos Projekt időszakok**
  - jan. 1. – márc. 31,
  - ápr. 1. – jún. 30.
  - júl. 1. – szept. 30.
  - okt. 1. – dec. 31.
- **Elbírálási időszakok**
  - márc. 1-14,
  - **jún. 1-14**
  - szept. 1-14.
  - dec. 1-10.
- **Felkészülési/zárási időszakok**
  - márc. 15-31.
  - **jún. 15-30.**
  - szept. 15-30.
  - dec. 11-23.

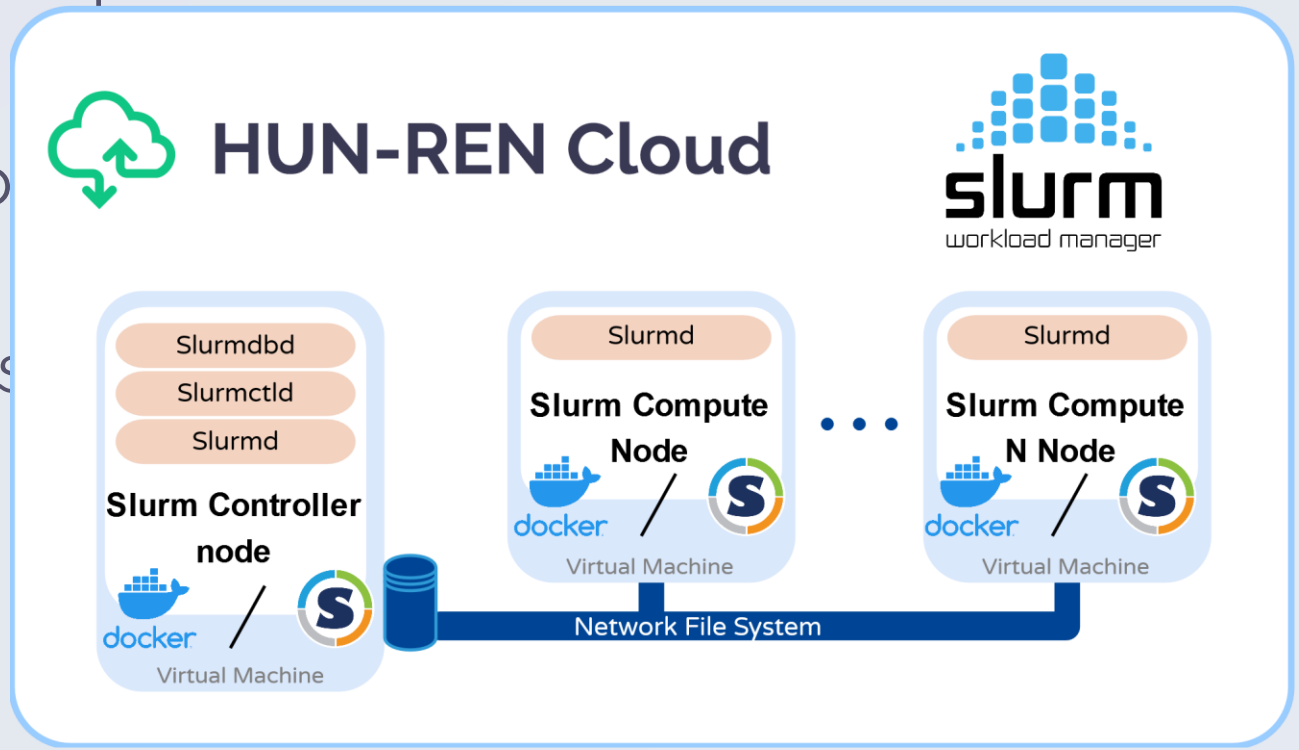
https://science-cloud.hu/hirek/2024/felhasznalok-tajekoztatasa-az-uj-gpu-eroforras-gazdalkodassal-kapcsolatban

Slurm

- Nyílt forráskódú erőforrás-kezelő és ütemező rendszer
- Feladatok (Job) / Kötegek (Batch) ütemezése
- HPC rendszerekben régóta használt megoldás (pl.: Komondor)
- Erőforrások (CPU, GPU, memória stb.) dinamikus elosztása
- Feladatok sorba állítása és párhuzamos futtatása
- MP/MPI/MPI4PY futtatási lehetőség

Slurm szolgáltatás

- Slurm referencia architektúrára épül
- Felhasználó kezelés
- NFS fájlmegosztás a csomópo
  között
- FIFO logika – Fair erőforrás has
- Singularity
- Különböző erőforrás partíciók



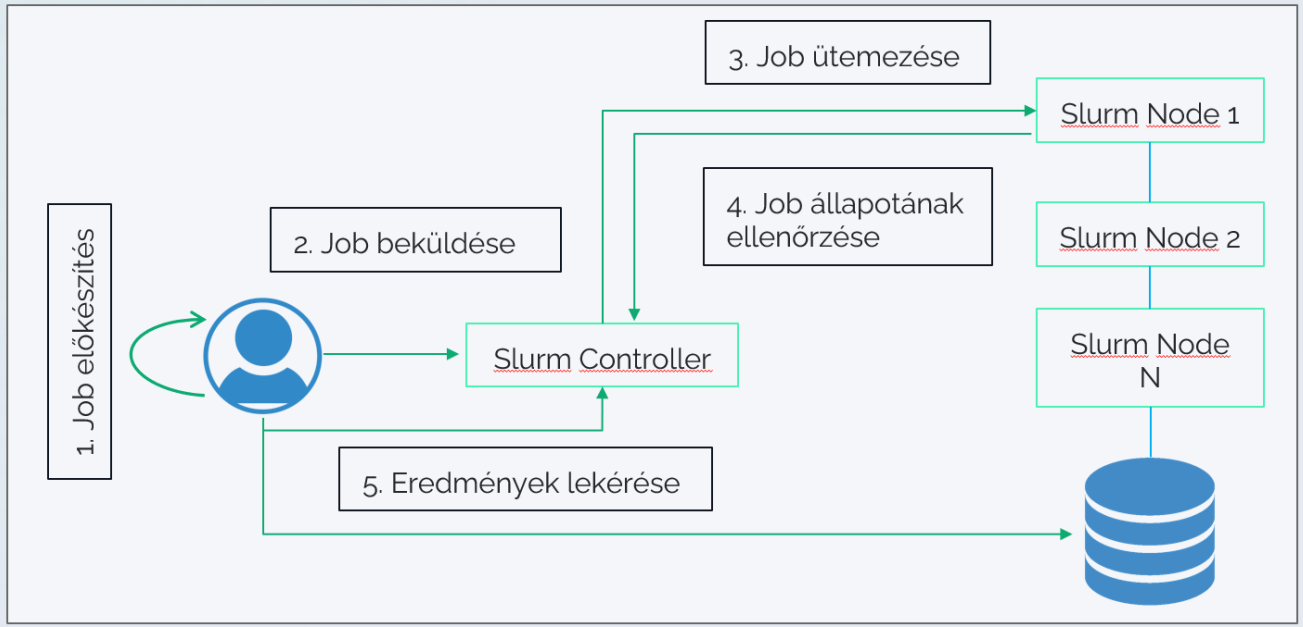
https://docs.slurm.science-cloud.hu/

Kvóták és partíciók

- Felhasználó szintű limitációk
  - 1 futtatható job
  - 3 job kerülhet ütemezésre
  - 100 GB tárhely
- GPU specifikus partíciók
  - GPU vRAM által meghatározott partíciók
  - Interaktív
    - 7 nap után a job leállításra kerül
    - Preferált partíció a fejlesztői környezetek kialakítására
  - Batch
    - 1 nap után a job leállításra kerül
    - Gyorsan lefutó joboknak létrehozott partíció

Slurm jobok futtatásának menete

1. A job előkészítése (felh.)
2. A job beküldés (felh.)
3. A job ütemezése a klaszteren (slurm)
4. A job állapotának nyomon követése (slurm)
5. A sikeres futást követően a felhasználó eléri a számítás eredményét (felh.)



Milyen feladatokra javasoljuk?

- Rövid ideig szükséges erőforrás
- Nagy számítási igényű feladatok
- Egymásra épülő feladatok
- „Parameter sweep” típusú feladatok
- Párhuzamos feldolgozást igénylő feladatok
- Hosszabb felügyelet nélküli számításokat

Regisztráció menete

- Új projekt igénylés során
  - Slurm erőforrások igénylése **(új checkbox)**
- Meglévő projekt igénylés esetén
  - Projekt módosítás -> Slurm erőforrások igénylése
- Projekt szintű hozzáférés
  - A projekt felhasználói saját profiljukon megadhatnak egy SSH kulcsot
- Igényelhető intervallum 3 hónap
  - Hosszabbítás módosítással igényelhető

Regisztráció menete

- Jóváhagyást követően hozzáférés biztosítása a rendszerhez
- SLURM erőforrás-allokációval rendelkező futó projektek adatlapján a tagok számára megjelenik egy új szekció:
  - Aktuális terheltség
  - SSH kulcs beállításhoz segítség
  - Hozzáférés
- Lejárat előtt a **felhasználó felelősége** a számítási adatok kimentése

https://science-cloud.hu/eloadasok

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

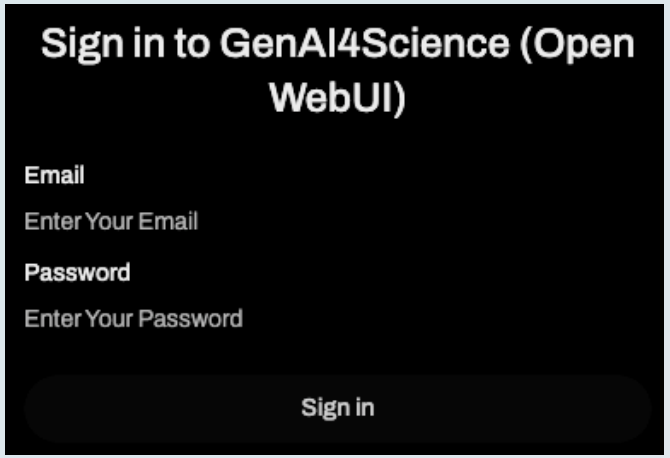
https://doc.genai.science-cloud.hu/

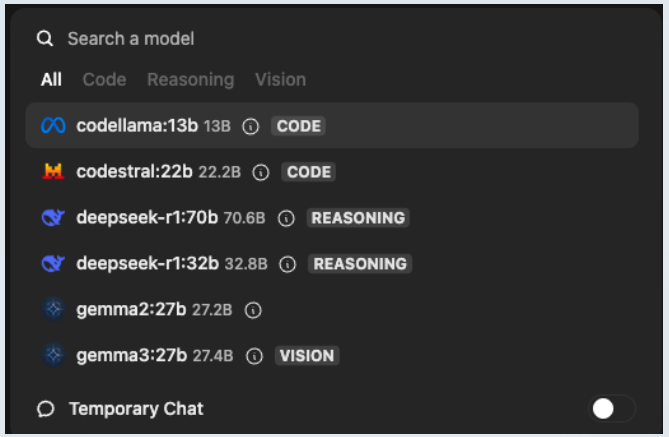
GenAI4Science szolgáltatás felépítése

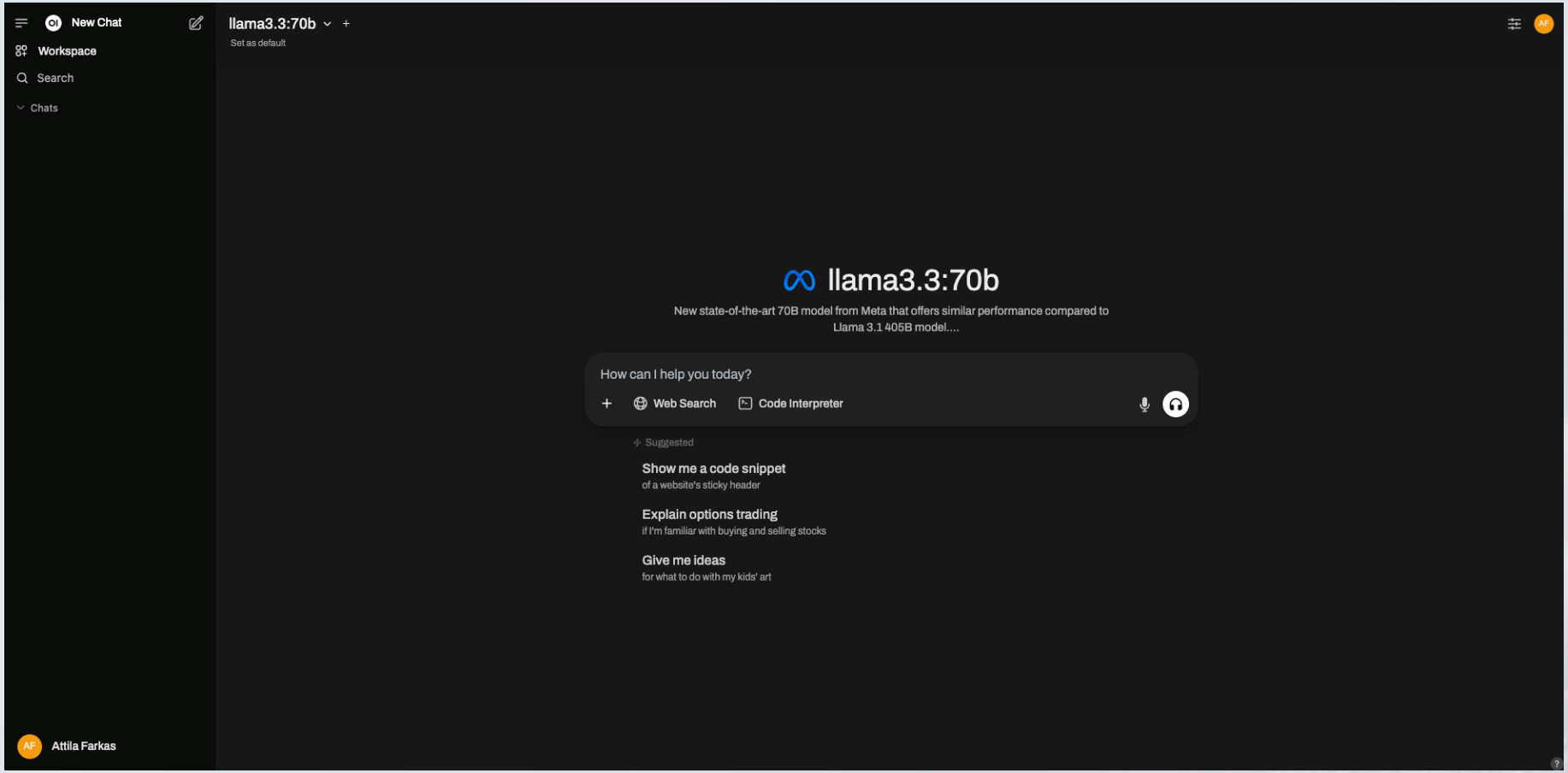
- **Ollama**
  - Támogat számos LLM modellt (llama, deepsek, gemma, mistral, Qwen, stb.)
  - Modell katalógust biztosít
  - Támogatja a testreszabott modelleket
  - REST API-t biztosít a modellek eléréséhez
  - GPU-alapú modell futtatás (korlátozott a HUN-REN Cloudon)
    - Terv: erőforrás bővítés -> elosztott futtatás több GPU felhasználásával vLLM alapon

- **Open-WebUI**
  - Felhasználóbarát
  - Felhasználói jogosultság kezelés
  - Többfelhasználós támogatás
  - OpenAI API alapú elérés támogatása (integrációs lehetőségek)
  - RAG támogatása
  - Helyi, távoli, webes keresés (Google, DuckDuckGo stb.)

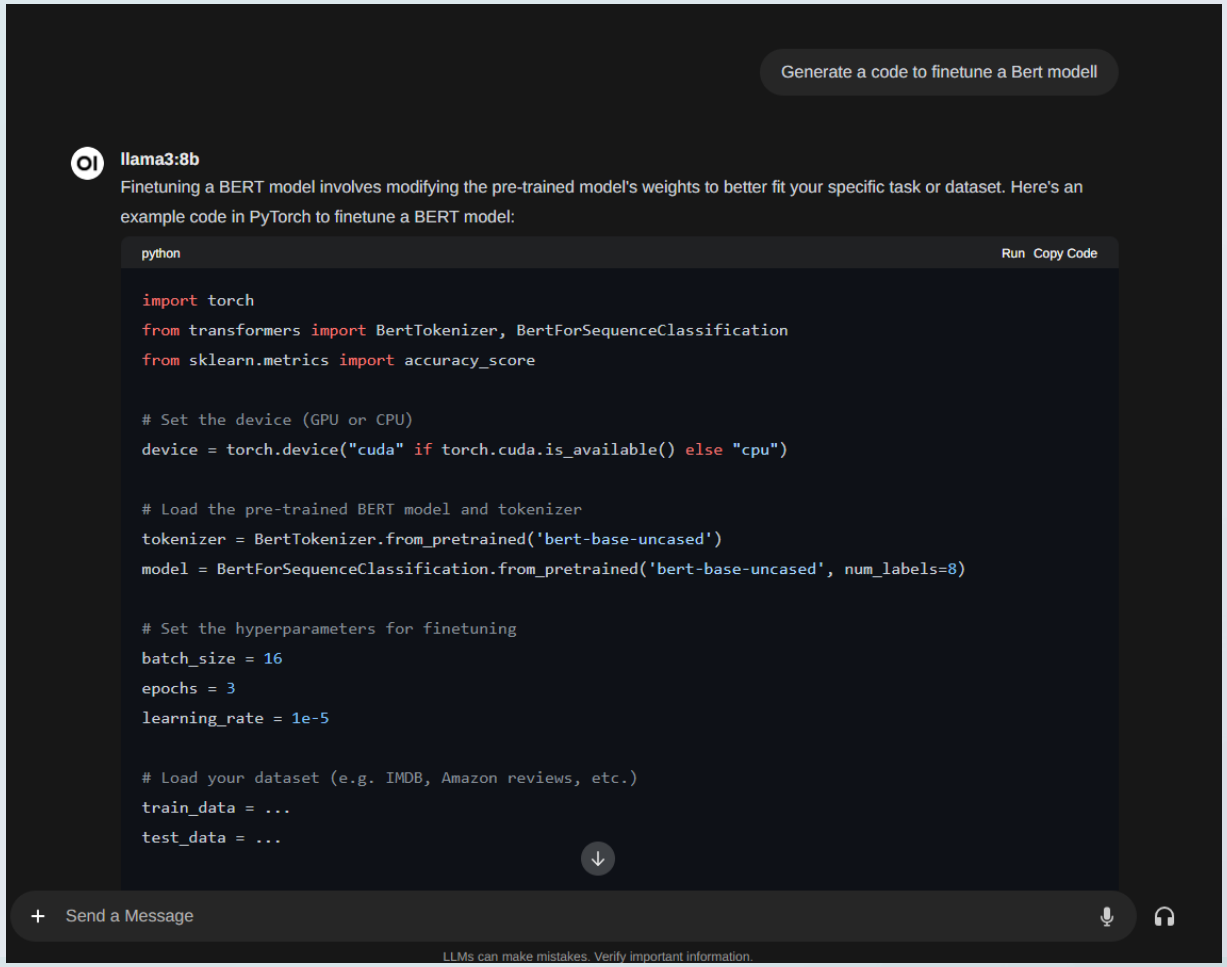
DEMO – Bejelentkezés, modell választás

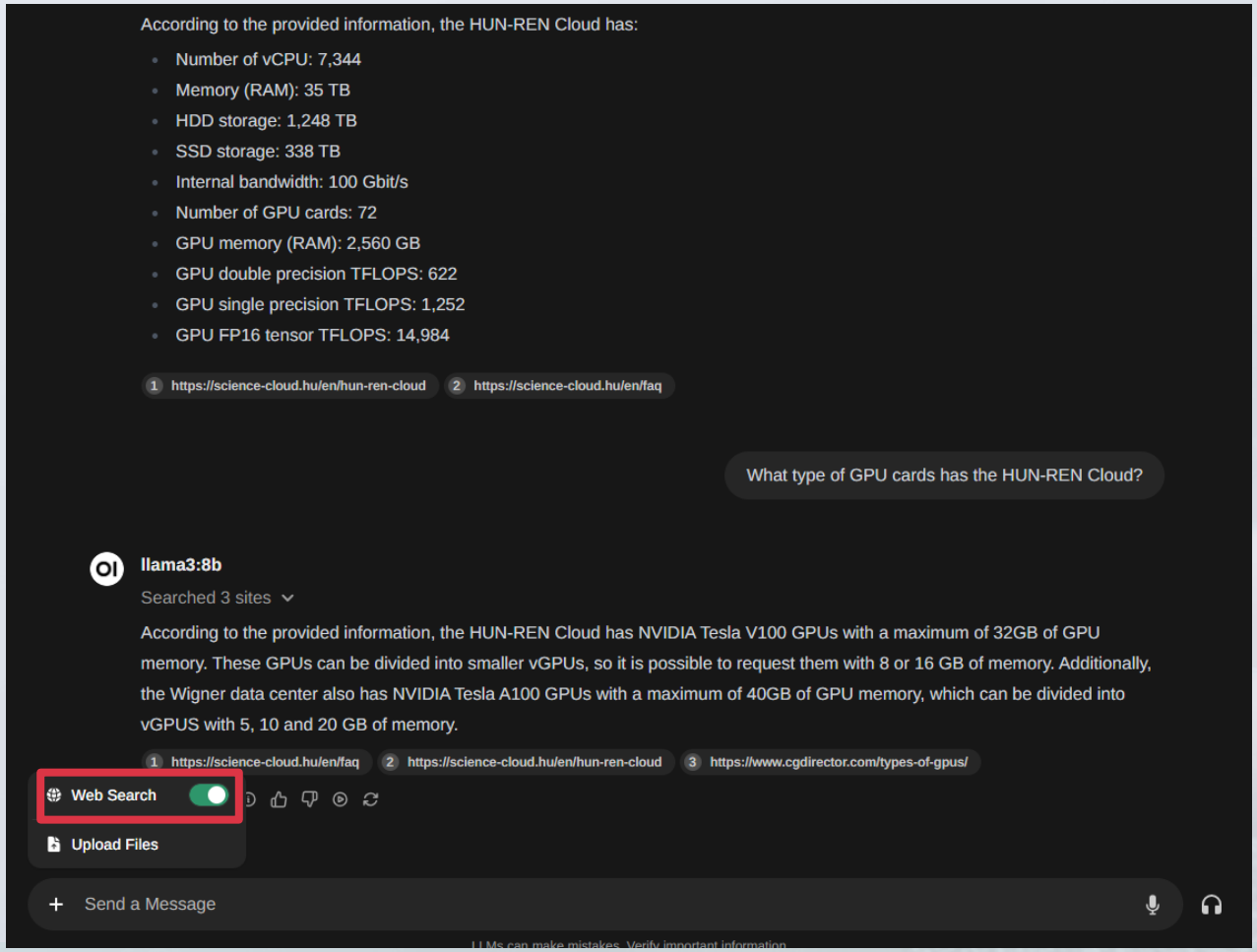






DEMO – chat felület, RAG





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

Használati esetek

- AI4Science program támogatása
- Mesterséges Intelligencia Nemzeti Labor támogatása

Segítség a GPU igénylőlapok kitöltésében, új szolgáltatások használatlában

Összefoglalás

- **Új GPU allokálási rendszer** került bevezetésre az erőforrások dinamikusabb kiosztása érdekében
  - Különböző projekt kategóriák pályázhatók periodikusan
- Új **PaaS** szolgáltatások a IaaS hozzáférés mellett
  - **SLURM szolgáltatás** – kötegelt vagy interaktív feladat végrehajtás ütemező segítségével
  - **GenAI4Science** – nagy nyelvi modellek elérése chat üzemmódban vagy API hozzáférés

Köszönöm a figyelmet!

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