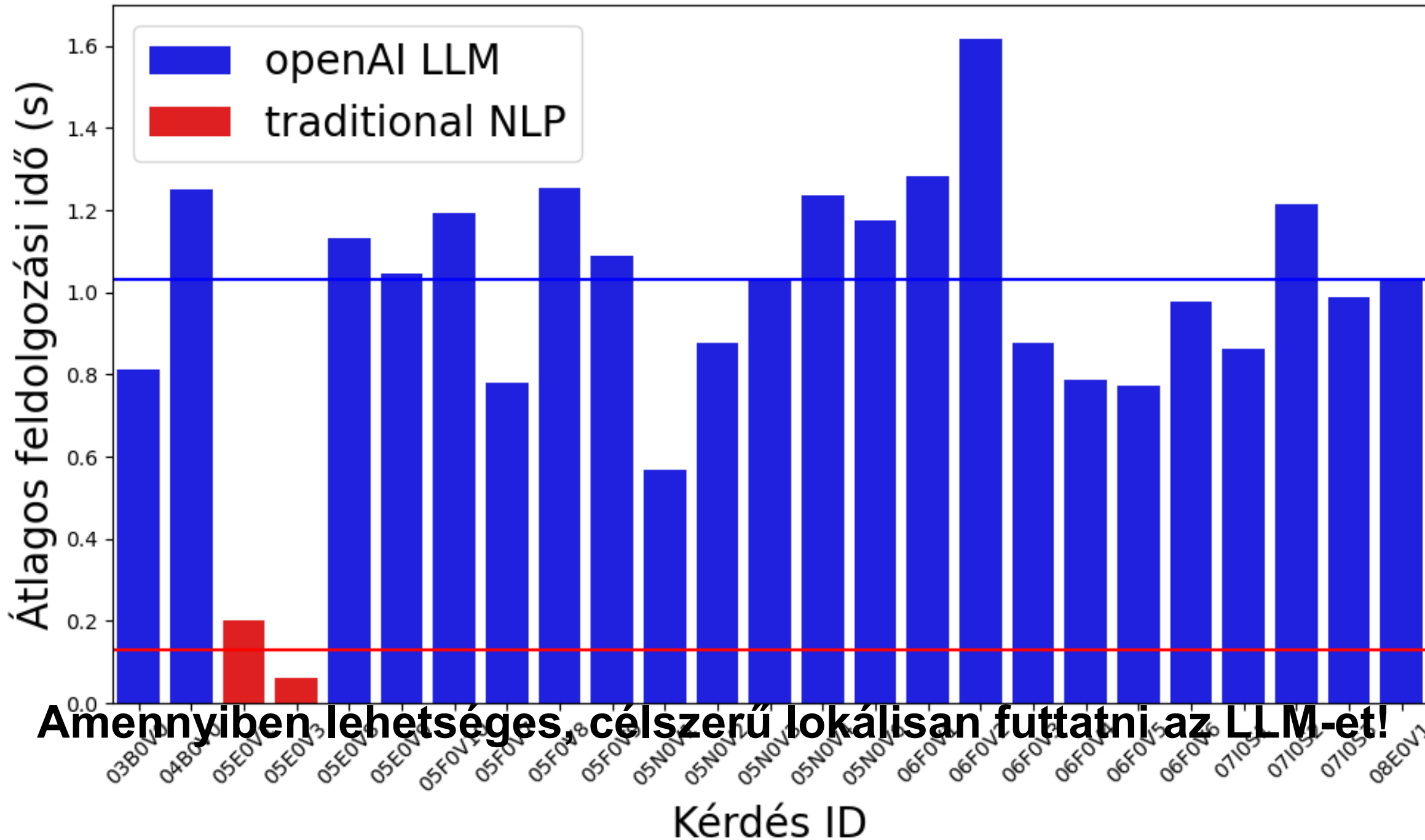


Hogyan építsünk biztonságos LLM-környezetet felhő nélkül?

Máth Benedek



ELVÁRÁSOK

- Kérdés: Ha most vasárnap országgyűlési választások lennének, Ön elmenne-e szavazni?
- Válasz: ,Lehet elmennék'
- 1: Biztosan elmenne; 2: Valószínűleg elmenne; 3: Valószínűleg nem menne el; 4: Biztosan nem menne el
- Elvárt output:
- [{,category': 2, ,confidence': 100}]
- A vizsgált rendszer-modell párosnak
 - Hardver kompatibilisnak kell lennie – checkbox, a modellnek rá kell férnie 16GB VRAM-ra
 - Gyorsan kell válaszolnia – Gyorsabban, mint az openAI 1,03 s-es jelenlegi teljesítménye, cél 500 ms
 - Helyes formátumban kell válaszolnia – checkbox: [{,category': 1}] vagy [{,category': 1, ,confidence': 100}]
 - Helyesen kell kategorizálnia – 90% körüli egyezés már elfogadható

MÓDSZERTAN

- Betöltjük a korábbi kampány válaszokat az LLM-nek
- Számoljuk, hány választ ad vissza helyes formátumban → [{,category': 1}] vagy [{,category': 1, ,confidence': 100}]
- Ezek közül hány egyezik a korábban beazonosított kategóriával
- Futási időt mérjük
- 1. teszt: egyesével töltjük és dolgoztatjuk fel a promptokat
- 2. teszt: egyszerre több csatornán töltjük és dolgoztatjuk fel a promptokat

ADATOK

- Májusi kampányból 1000 válasz 19 kérdésre
 - 1. tesztnél 5 kérdés 200 kézzel kategorizált válasza
 - 2. tesztnél a teljes adat GPT-4o mini által kategorizált válasza
- A kérdések között van zárt és nyílt végű is

Kérdés ID	Kérdés szöveg	Típus
04B0V0	Ön férfi vagy nő?	2 kategória
05E0V8	Ha most vasárnap országgyűlési választások lennének, Ön elmenne-e szavazni?	4 kategória
05E0V9	Ha most vasárnap országgyűlési választások lennének, Ön melyik pártra adná a szavazatát?	6 kategória
05N0V2	Ön szerint milyen a közhangulat mostanában az országunkban? Az emberek többsége inkább jól, vagy inkább rosszul érzi magát?	2 kategória
06F0V3	És mennyire ért egyet azzal az állítással, hogy az Ön családjának van elég megtakarítása a váratlan dolgokra?	4 kategória
05F0V10	Hogyan jellemezné Vlagyimir Putyint?	Nyílt végű

KERETRENDSZER?



Transformers



LLM



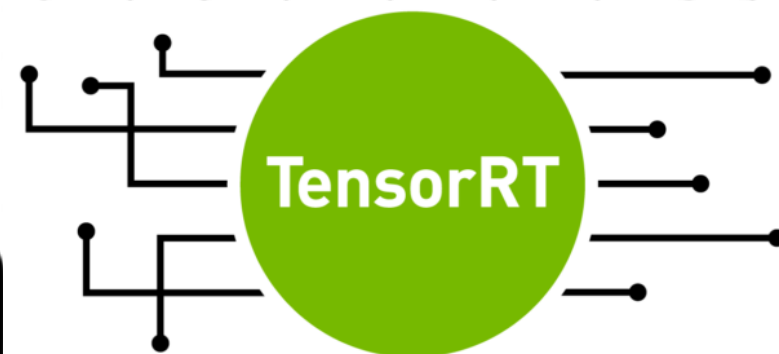
LM Studio

LLaMA 

 Clementine



Ollama



MODELL?



Qwen



Gemma



LLAMA 3



MISTRAL
AI_

EURO*LLM



GPT-OSS

120B & 20B

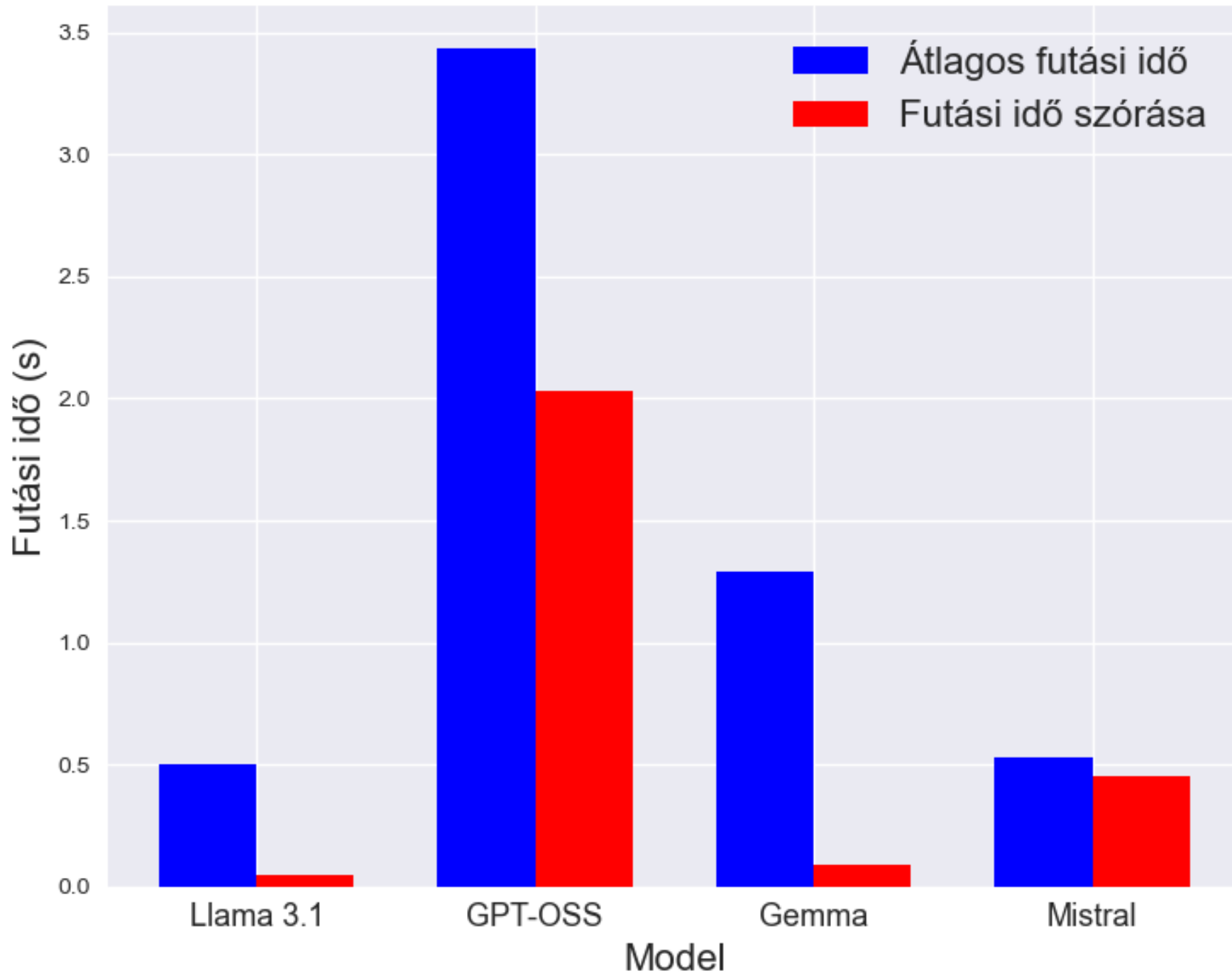
VERZIÓSZÁM?

🔗 Model tree for meta-llama/Llama-3.1-8B-Instruct ⓘ

Base model	meta-llama/Llama-3.1-8B
📦 Finetuned (1648)	this model
Adapters	1292 models
Finetunes	2010 models
Merges	69 models
Quantizations	532 models

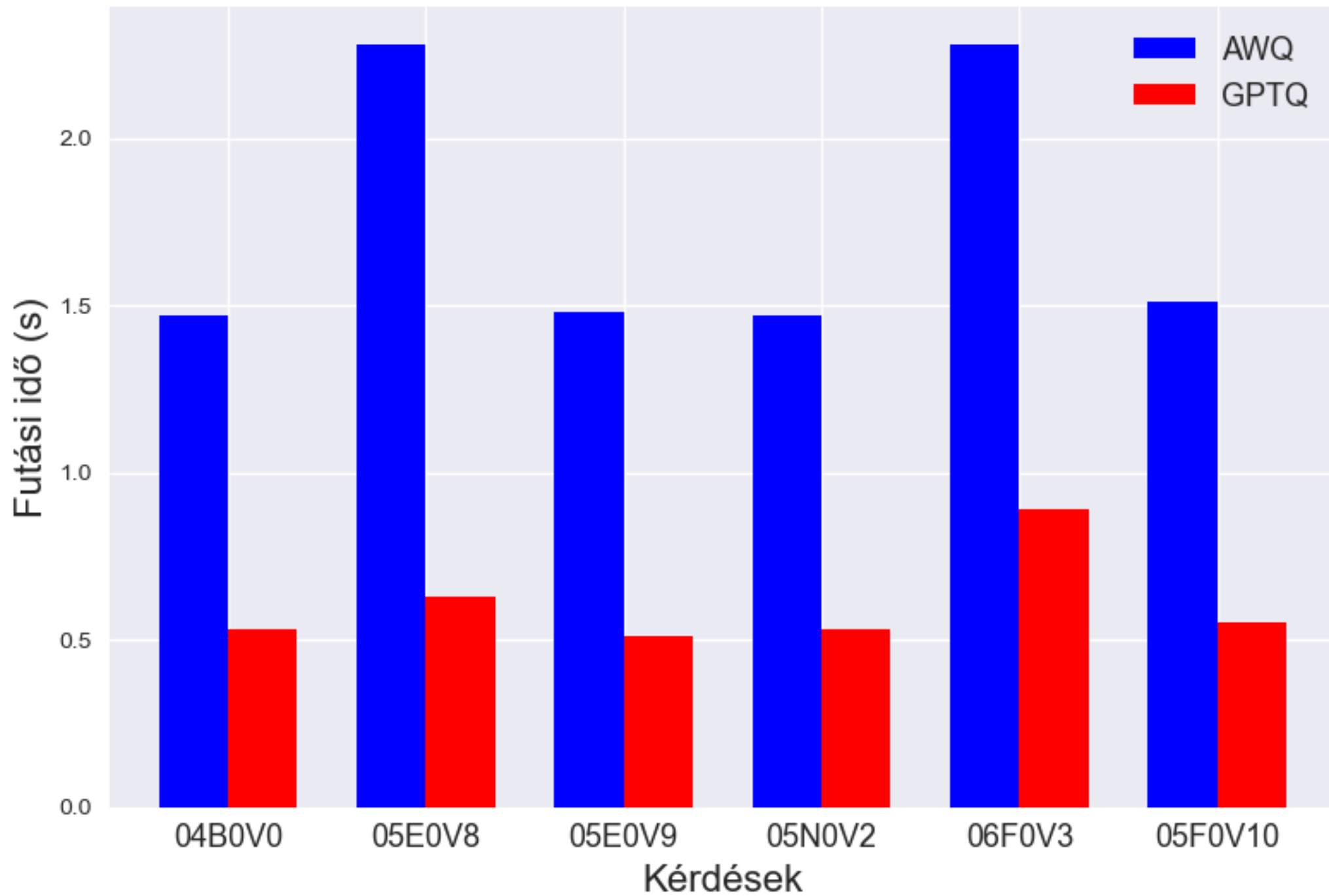
OLLAMA

- Fejlesztés futtatás
- Széles körű támogatás
- Kiemelkedően gyors
- Klasszikus és modern architektúrák
- DE: ...



a

vL



ÖSSZEFOGLALÁS

- Lehetséges az elérhető rendszerekből és modellekből működő, saját szervert építeni
- Ollama nagyon jó kísérletezésre, nagyobb modellek futtatására
- vLLM nagyon jó sok prompt egyidejű kiértékelésére
- vLLM-Llama 3.1 8B keretrendszer-modell pár volt a legjobb nekünk
- Nyílt kérdésekben még nem vagyunk tökéletesek

Köszönöm a figyelmet!