Kauzálisan integrált információ alkalmazása sztochasztikus becslésben, egymástól eltérő nagy nyelvi modellek hálózatában

Application of Causally Integrated Information in Stochastic Prediction, Across a Network of Distinct Large Language Models

KOLCSÁR Levente [1,2], DOMOKOS József [1]

[1]Sapientia EMTE Marosvásárhelyi Kar, Villamosmérnöki Tanszék, Koronka, Segesvári út 2 szám,
kolcsar.levente@ms.sapientia.ro, domi@ms.sapientia.ro, www.ms.sapientia.ro
[2] Óbudai Egyetem, Alkalmazott Informatikai és Alkalmazott Matematikai Doktori Iskola, 1034 Budapest, Bécsi út 96/b,
kolcsar.levente@stud.uni-obuda.hu, www.aiamdi.uni-obuda.hu

Abstract

In our paper, we begin to discuss the application of Integrated Information Theory derived from cognitive brain research. This theory quantifies the integrated cause-effect relationships between the nodes of a stochastic network. We calculate this informational quantity within a Markov chain prediction, which is trained on a network of large language models. We hypothesize that this information can be utilized to optimize the predictions of the Markov chain, which would improve the collective decision-making ability of the large language models.

Keywords: information theory, IIT, Markov chain, LLM

Kivonat

Dolgozatunkban tárgyalt alkalmazási kiindulás a kognitív agykutatásból származó integrált információ elmélet egy lehetséges felhasználását tárgyalja. Ez az elmélet kvantifikálja a sztochasztikus hálózat csomópontjai közötti integrált ok-okozati hatásokat. Ezt az információs mennyiséget számoljuk egy Markov-lánc becslőben, amelyet nagy nyelvi modellek hálózatában tanítunk. Feltételezésünk szerint ez az információ felhasználható a Markov-lánc predikció optimizálására, amely javítaná a nyelvi modellek együttes döntéshozási képességét.

Kulcsszavak: információ elmélet, kauzálisan integrált információ, Markov-lánc, nagy nyelvi modellek

1. BEVEZETŐ

A dolgozat kiindulási ötlete a korábban tárgyalt interdiszciplináris kutatási módszertanból származik [1], amely kognitív elméleteket hivatott elemezni, ahol tudományfilozófiai, matematikai és alkalmazási kontextusokban egy példán keresztül tárgyal egy kognitív elméletet. Az említett módszertanban példaként kielemzésre került az integrált információ elmélet [2], amely alkalmazhatósági kontextusban felhasználható tanítási kritériumként sztochasztikus hálózatokban. Ezt az ötletet kezdtük el kidolgozni egy predikciós feladatra. Az elmélet kvantifikálja a sztochasztikus hálózatokban számolható kauzálisan integrált információt, amelyet görög Φ-vel jelöl. Ez az információs mennyiség a sztochasztikus hálózatokban akkor magas, ha a hálózat csomópontjai időben egymásra ok-okozati hatással rendelkeznek és ezek a hatások nem bonthatók fel részhálózati vagy különálló csomóponti működésre. Ha a csomópontok állapotai nem változnak időben vagy a csomópontok egymásra nincsenek ok-okozati hatással, akkor a Φ=0.

Ezt az információt számoljuk egy Markov-láncban, amelyet nagy nyelvi modellek hálózatában nem felügyelten tanítunk. A predikciós feladat az, hogy binárisan eldöntsük, hogy egy szöveget gép, vagy ember írt. Manapság igen aktuális téma az ember által írt és a különféle MI alapú gép által generált szövegek megkülönböztetése [3][4]. A nem felügyelt tanítással tanított sztochasztikus hálózatunk, a Markov-lánc állapotátmeneti mátrixa (TPM – Transition Probability Matrix) az MI hálózat átmenetei szerint elemenként inkrementálódik, majd a normalizált TPM-ekre utólag kiszámoljuk a Φ mennyiséget. Ebből kiindulva tárgyaljuk a lehetséges tanítási kritériumfüggvényt, amely tartalmazza a kauzálisan integrált információ

mennyiséget és a predikció Hamming-távolságát. Feltételezésünk szerint ez a mennyiség felügyelt tanítás esetén javítaná a Markov-lánc predikciós eredményeit.

2. GYAKORLATI MEGVALÓSÍTÁS

2.1 A hálózathoz használt API és paraméterei

A gyakorlati megvalósításhoz az *openrouter.ai* API-t használtuk [10], Python programozási nyelvben. A nagy nyelvi modell hálózat csomópontjai a DeepSeek, a Qwen és a Llama, ezeknek pontos verziószáma pedig: *"deepseek/deepseek-r1-0528:free","qwen/qwen3-235b-a22b:free","meta-llama/llama-3.1-405b-instruct:free"*. [7][8][9]

A hálózat topológiája: teljesen összekapcsolt, kivéve az önkapcsolatokat (1. ábra)

API paraméterek: *temperature = 0* (A nyelvi modellek hőmérséklete, ami azt jelenti, hogy mennyire véletlenszerű/kreatív a válaszuk. Tudomásunk szerint ez még nem garantálja, hogy a modellek teljesen determinisztikusan válaszolnak újrafuttatás esetén, hanem csak azt jelenti, hogy alacsonyabb a válaszok véletlenszerűsége.)

API chat: nem használunk kontextus memóriát, a Markov tulajdonság (kondicionális függetlenség) elérése miatt.

Kétféle kérdést teszünk fel a nagy nyelvi modelleknek egymásután, melyben arra kérjük, hogy nyilatkozzon arról, először a saját véleménye szerint majd ismerve az előző állapotban megkérdezett nyelvi modellek válaszát, nyilatkozzon újra, hogy ember vagy MI írta a megadott szöveget.

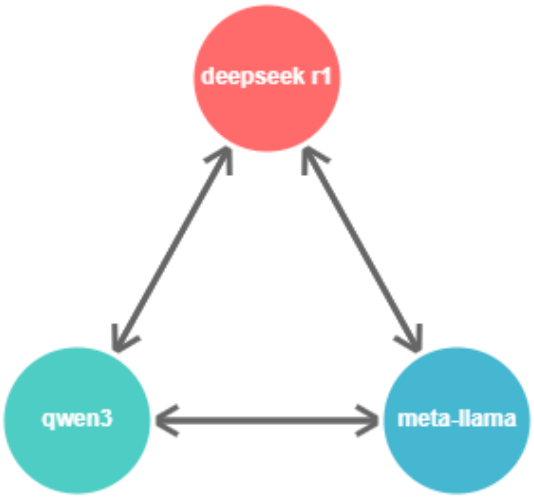
- Az aktuális állapot (S_t) kérdése:

"Can you tell me that this text is written by AI? (I need only a binary value response, nothing more. 0 or 1. 1 if AI, 0 if human) This is the text:'"+text+"'"

- A következő állapot (S_{t+1}) kérdése:

"Can you tell me that this text is written by AI? (I need only a binary value response, nothing more. 0 or 1. 1 if AI, 0 if human) [About this text, other "+str(n-1)+" LLM's told me this:"+remove_char(from_state,i)+"] This is the text:'"+text+"'"

Minden API válasz kezelése esetén keressük az első *"1"* vagy *"0"* karaktert a válaszból. API hiba esetén az aktuális tanítási iteráció átlépésre kerül, tehát nem tanítunk a hibás válaszból.



1. ábra. Nagy nyelvi modell hálózat

Az 1. ábra szemlélteti, hogy milyen topológiájú a nyelvi modellek hálózata, amely a hálózatok közötti S_{t+1} lekérdezés kapcsolati viszonyát szemlélteti. Ebből a hálózati ábrából felírható a kapcsolati mátrix, amely az integrált információ számításához szükséges paraméter:

$$kapcsolati\ mátrix = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Az önkapcsolatot azért nem foglaltuk bele az S_{t+1} lekérdezésbe, mert feltételezzük, hogy determinisztikusan válaszolna a saját véleménye szerint. Ez a kapcsolat és a Markov-lánc kondicionális függetlenség tulajdonsága további analízist igényel részünkről. [6]

2.2. A Markov-lánc felépítése és nem felügyelt tanítása

A hálózati csomópontok száma: $n = 3$, így az állapotátmeneti valószínűségek mátrixának (Transition Probability Matrix - TPM) mérete: $2^n \times 2^n = 8 \times 8$.

A lehetséges állapotok halmaza binárisan kódolva $S = \{ 000, 100, 010, 110, 001, 101, 011, 111 \}$ (2. ábra) (*"Little-Endian"* sorrend).

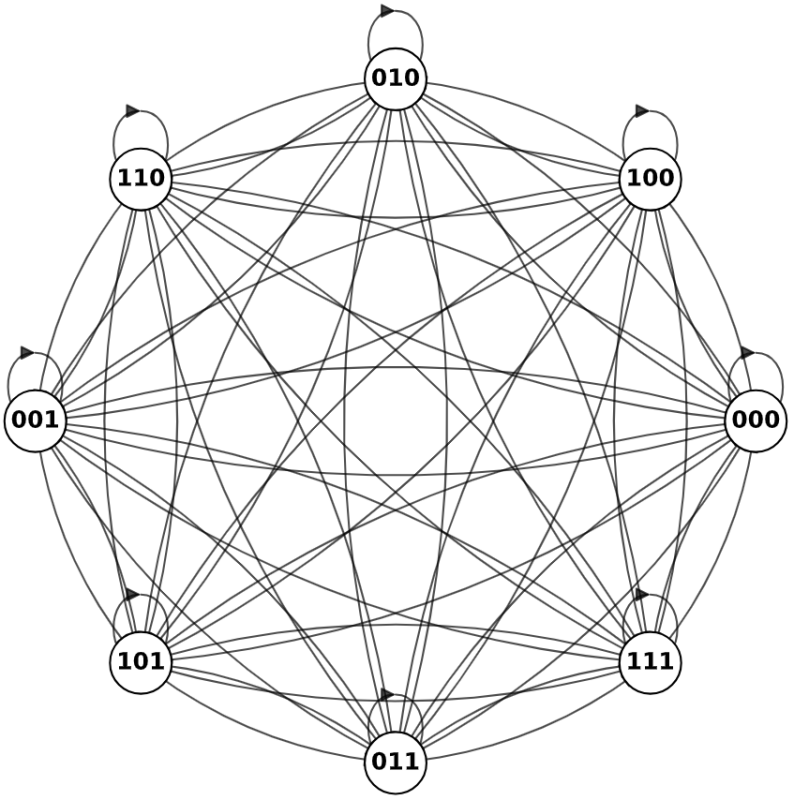
A tanítóhalmaz: 676 darab, 70-180 szót tartalmazó angol és magyar paragrafus, vegyes forrásokból gyűjtve (különböző nagy nyelvi modellek, könyvek, fórumok, hírek, web-es cikkek).

A tanítási algoritmus lépései:

0. TPM inicializálása: TPM = nullmátrix

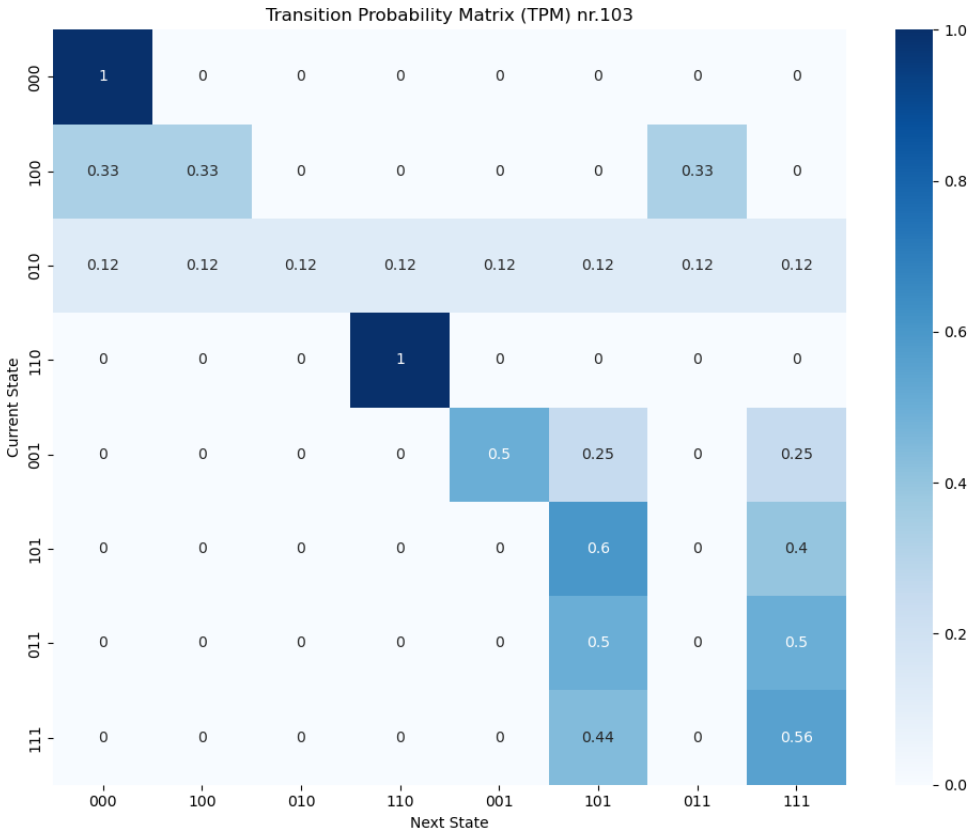
A tanítási iteráció ciklusa:

1. Aktuális állapot (S_t) lekérdezése az aktuális paragrafusra;
2. Következő állapot (S_{t+1}) lekérdezése az aktuális paragrafusra;
3. Inkrementáljuk a megtörtént állapotátmenetet: $TPM[S_t, S_{t+1}] = TPM[S_t, S_{t+1}] + 1$;
4. A TPM másolatának normalizálása;
5. A TPM másolat mentése tömbösített állományba (TPM-ek állománya).

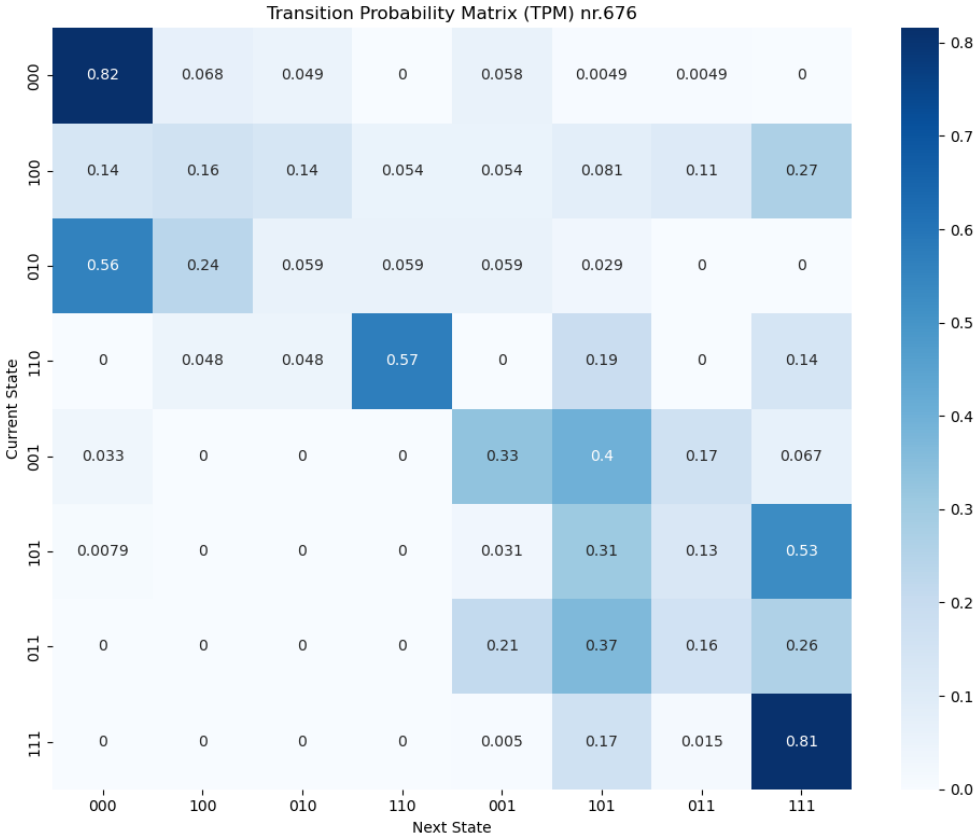


2. ábra. *Állapot-átmeneti gráf*

A tanítással a tanítóhalmaz végéig, vagyis a 676. iterációig jutottunk, ezalatt 4056 ($n \times 2 \times 676$) API lekérdezést végeztünk. A *3*. ábra szemlélteti, hogy a tanítás alatt a 103. iterációban milyen eloszlású TPM mátrixunk van, a 4. ábra pedig szemlélteti, hogy a tanítás végén a 676. iteráció TPM mátrixa milyen eloszlású.



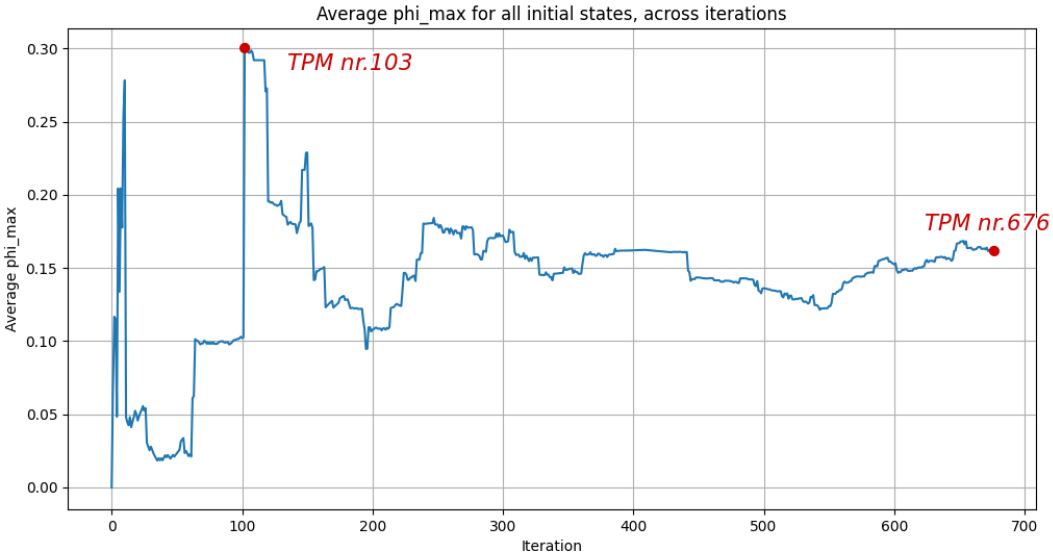
3. ábra. - *A 103. iteráció állapotátmeneti mátrix (TPM) eloszlása*



4. ábra. – *A 676. iteráció állapotátmeneti mátrix (TPM) eloszlása*

2.3. A kauzálisan integrált információ (Φ) számítása

A kauzálisan integrált információ számítása érdekében felhasználtuk a "pyphi" Python könyvtárat (Integrated Information Theory Toolbox) [5]. A könyvtár függvényeivel lehet számolni meghatározott kezdeti állapotból, meghatározott kapcsolati mátrixból, meghatározott TPM mátrixra Φ_{max} és részhálózati Φ értékeket. A nem felügyelt tanítás után az állományból kiolvasva minden elmentett TPM mátrixra számoltunk egy Φ_{max} átlagot (összes lehetséges kezdeti állapotra számolt Φ_{max} átlaga) és ábrázoltuk a tanítási iteráció függvényében. Az 5. ábrán látható, hogy ez az információ a tanítás során elér egy maximum értéket, majd stabilizálódik egy átlagos információs értékre. Ez azért is van, mert a Markov-lánc nem felügyelt tanítási algoritmusa kumulatív, így az iterációk előrehaladtával kevésbé változik a TPM mátrix.



5. ábra. - *Átlagolt Φ_{max} alakulása a tanítási iterációval*

2.4. A feltételes kritériumfüggvény felügyelt tanítás esetén

Az eddig elért tanítási példából azt feltételezzük, hogy ez a Φ felhasználható, mint tanítási kritérium, mert azt kvantifikálja, hogy a modellek mennyire döntenek "együttesen". Ez az információ szerintünk szükséges, de nem elégséges a kritériumhoz. Szerintünk a kritériumfüggvényhez kell az is, hogy megerősítsük azokat a magas Φ értékű átmeneteket, amelyek jó irányba viszik a predikciót. Ehhez az irányhoz felhasználjuk a Hamming-távolságot.

A tanítási inkrementálási érték függvénye egy iterációra: $f(i) = \alpha \frac{1}{D_i+u} + \beta \Phi_{max_i}$

ahol D – Hamming-távolság a szöveg ismert állapota (000 vagy 111) és az S_{t+1} állapot között;
u – numerikus minimális távolság a zérus osztó elkerülésére (pl. u = 0.1);
α – a predikciós távolságra vonatkozó tanítási együttható;
β – a kauzálisan integrált információra vonatkozó tanítási együttható.

3. KONKLÚZIÓK

A példa alkalmazással elkezdtük számolni a kauzálisan integrált információ mennyiségét (Φ) egy Markov-láncban, amit nagy nyelvi modellek hálózatán tanítunk. Továbbá javasoltunk egy Φ-t tartalmazó kritériumfüggvényt, amely feltételezésünk szerint javítaná a becslési eredményt. A továbbiakban szeretnénk az alábbi feladatokat elvégezni:

- Meg kell vizsgálnunk a Markov-lánc tulajdonságai [6] (kondicionális függetlenség, homogenitás, ergodicitás, irreducibilitás, stacionárius állapotok, tranziens és rekurrens állapotok) és a Φ_{max} , illetve az alhálózati Φ értékek közötti összefüggéseket;
- Választ kell adnunk arra a kérdésre, hogy mit optimizál a Φ a Markov-láncban?
- A kritériumfüggvény és a tanítási algoritmus javítása;
- A tanítási iterációk száma és a nagy nyelvi modell hálózat kibővítése;
- A becslési módszer meghatározása;
- Statisztikai kimutatások arra vonatkozóan, hogy szignifikánsan jobban teljesít a javasolt kritériumfüggvény által tanított becslő, mint a nem felügyelten tanított becslő.

Az alkalmazás végső célja egy olyan felhasználói program megalkotása, amely nemcsak a nyelvi modellek válaszait tartalmazza, hanem az azok alapján lefuttatott becslő algoritmusból származó statisztikai kimutatást is. Ezzel is elősegítve a felhasználót, hogy egy részletesebb elemzést kapjon arról, hogy egy szöveg milyen eredetű lehet.

KÖSZÖNETNYILVÁNÍTÁS

Végezetül köszönetünket szeretnénk kifejezni az Óbudai Egyetem Alkalmazott Informatikai és Alkalmazott Matematikai Doktori Iskolának, amely hozzájárult e kutatás anyagi támogatásához. A kutatás az Óbudai Egyetem doktori képzés keretén belül valósult meg.

IRODALMI HIVATKOZÁSOK

[1] Kolcsár L., Bakó L.: *Interdiszciplináris MI-kutatási módszertan*, MTK, 2025, 1-6. https://doi.org/10.33895/mtk-2025.22.10
[2] Oizumi M., Albantakis L., Tononi G.*: From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory* 3.0. PLOS Computational Biology, 2014 https://doi.org/10.1371/journal.pcbi.1003588
[3] Tony Berber Sardinha: *AI-generated vs human-authored texts: A multidimensional comparison,* Applied Corpus Linguistic, Volume 4, Issue 1, 2024 https://doi.org/10.1016/j.acorp.2023.100083
[4] Doru B, Maier C, Busse JS, Lücke T, Schönhoff J, Enax- Krumova E, Hessler S, Berger M, Tokic M: *Detecting Artificial Intelligence–Generated Versus Human-Written Medical Student Essays: Semirandomized Controlled Study,* JMIR Med Educ 2025;11:e62779 *https://doi.org/10.2196/62779*
[5] Mayner W.G.P., Marshall W., Albantakis L., Findlay G., Marchman R., Tononi G.: PyPhi: *A Toolbox for Integrated Information Theory*. PLoS Computational Biology, 14/7., 2018 https://doi.org/10.1371/journal.pcbi.1006343
[6] Guanyu C.: *Introduction To Basic Properties Of Markov Chain And Its Application In Simple Random Walk*, The University of Chicago, 2023 https://math.uchicago.edu/~may/REU2022/REUPapers/Chen,Guanyu.pdf
[7] DeepSeek MI honlapja: https://deepseek.ai/
[8] Qwen MI honlapja: https://qwen.ai/home
[9] Llama MI honlapja: https://www.llama.com/
[10] OpenRouter MI API honlapja: https://www.openrouter.ai