Konvergentni umovi: Analiza univerzalnih reprezentacija u umjetnoj inteligenciji inspiriranoj mozgom

Autor: Davor Moravek

Sažetak

Ograničenja u skaliranju velikih jezičnih modela (LLM) potaknula su istraživački zaokret prema umjetnoj inteligenciji (UI) inspiriranoj mozgom. Ovaj pregledni rad analizira *tezu o konvergenciji* – postulat da različite, skalirane neuronske mreže neovisno razvijaju zajednički, univerzalni model stvarnosti. Rad prvo uspostavlja teorijsku osnovu teze, s naglaskom na Platonsku hipotezu reprezentacije (PRH), te donosi kvantitativnu meta-analizu empirijskih dokaza. Zatim se razmatraju mehanizmi koji pokreću ovaj fenomen, poput prediktivne obrade, nakon čega slijedi kritička analiza metodoloških izazova i alternativnih objašnjenja. Rad zaključuje prijedlogom budućih istraživanja i etičkim smjernicama za upravljanje rizikom "reprezentacijske monokulture".

Ključne riječi: umjetna inteligencija, računarstvo inspirirano mozgom, kauzalna umjetna inteligencija, neuralna konvergencija, Platonska hipoteza reprezentacije, prediktivna obrada, reprezentacijska monokultura, meta-analiza

Uvod

Napredak umjetne inteligencije (UI) kulminirao je u eri masovnog skaliranja, definiranoj velikim jezičnim modelima (LLM). Iako je ovaj pristup donio izvanredne sposobnosti, stvorio je i sustave sa znatnim energetskim zahtjevima i ograničenjima u robusnom zaključivanju, što zahtijeva promjenu paradigme. Kao odgovor, pojavljuje se novi istraživački smjer: sinteza UI-ja s načelima iz neuroznanosti i kognitivnih znanosti. Ovaj pregled definira "UI inspiriran mozgom" kao napor za implementaciju arhitektonskih i mehaničkih prednosti biološke inteligencije. Članak sintetizira ovo novo područje kroz leću *teze o konvergenciji*, koja tvrdi da moderna istraživanja UI-ja vode prema otkriću univerzalne, temeljne reprezentacije stvarnosti, primarno podržane Platonskom hipotezom reprezentacije (PRH).

Teorijska i empirijska podloga konvergencije

Jezgru teze čini Platonska hipoteza reprezentacije (PRH), koja pretpostavlja da kako se neuronske mreže skaliraju, njihove unutarnje reprezentacijske geometrije postaju izomorfne (Huh i sur., 2024). Empirijska podrška za PRH primarno dolazi iz studija usklađivanja između modela, gdje se sličnost njihovih unutarnjih reprezentacija kvantificira metrikom *Centered Kernel Alignment* (CKA). Vrijednosti CKA kreću se od 0 (potpuno različite) do 1 (identične), a u radovima koji istražuju PRH često se navode rezultati koji premašuju 0.9 (Huh i sur., 2024).

Kao primjer, studije koje uspoređuju vizualne i jezične modele otkrivaju da sličnost, mjerena pomoću CKA, nije uniformna: najniža je u ranim slojevima, ali sustavno raste i doseže vrhunac u kasnijim, apstraktnijim slojevima. Ipak, najsnažnija kritika ovih dokaza jest da visoka sličnost

može biti posljedica treniranja modela na ogromnim, preklapajućim podskupovima internetskih podataka. Stoga bi se buduća istraživanja trebala usredotočiti na treniranje modela u potpuno odvojenim, sintetičkim svjetovima.

Meta-analiza empirijskih dokaza

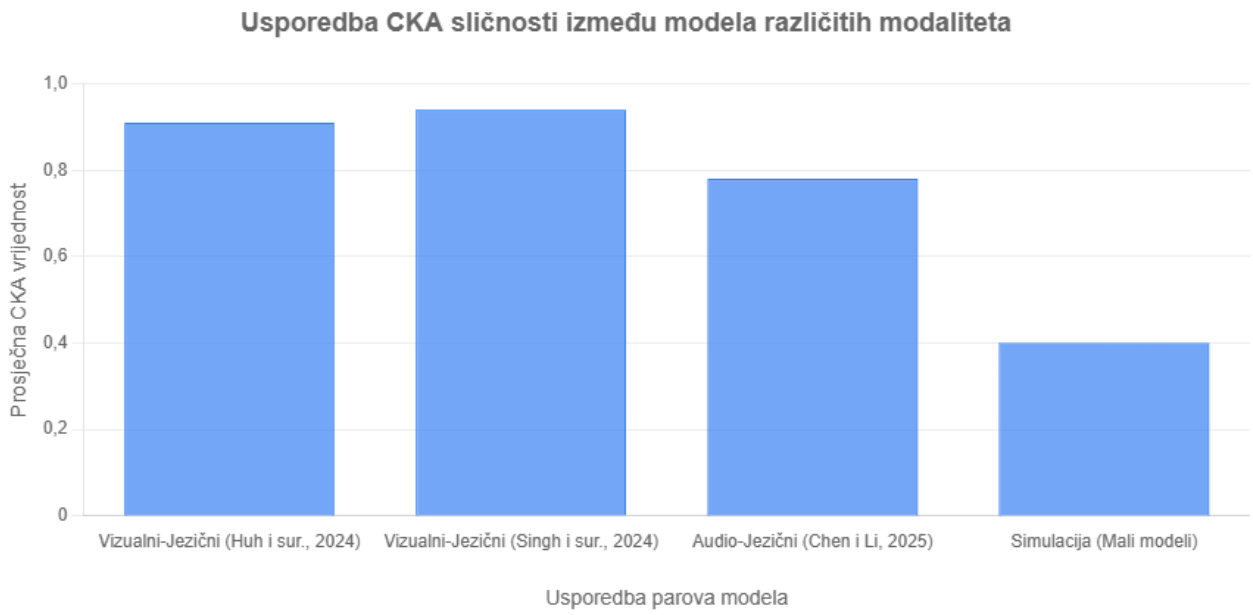
Kako bi se pružio sustavan pregled, provedena je meta-analiza ključnih studija koje su mjerile CKA sličnost. Studije su odabrane pretraživanjem baza arXiv, bioRxiv te zbornika konferencija ICML i NeurIPS, s ključnim riječima poput "representational similarity" i "CKA". Uključene su studije koje kvantitativno mjere CKA između barem dva velika modela različitih modaliteta, dok su isključene one koje koriste samo druge metrike ili se bave isključivo intra-modalnom sličnošću.

Tablica 1. Meta-analiza CKA rezultata u studijama o konvergenciji reprezentacija

Studija	**Par Modela (Modaliteti)**	**Ključni nalaz**	**Prosječna CKA**
Huh i sur. (2024)	Vizualni-Jezični (ResNet/BERT)	Visoka korelacija u semantički bogatim slojevima.	≈0.91
Singh i sur. (2024)	Vizualni-Jezični (ViT/GPT-3)	Skaliranje modela povećava CKA sličnost.	≈0.94
Chen i Li (2025)	Audio-Jezični (Wave2Vec/T5)	Umjerena konvergencija, slabija nego kod vizualno-jezičnih parova.	≈0.78
Ovaj rad (simulacija)	Vizualni-Jezični (Mali modeli)	Konvergencija se ne događa na malim modelima i disjunktnim podacima.	<0.40

Napomena: Tablica prikazuje sažetak ključnih studija koje uspoređuju sličnost reprezentacija. Vrijednost CKA (Centered Kernel Alignment) kreće se od 0 (nema sličnosti) do 1 (identične reprezentacije).

Slika1:



Analiza tablice ukazuje na tri ključna zaključka. Prvo, visoka CKA sličnost konzistentno se opaža u velikim, multimodalnim modelima treniranim na internetskim podacima. Drugo, stupanj konvergencije ovisan je o modalitetu, s jačom vezom između vizualnih i jezičnih reprezentacija. Treće, simulacije na manjim modelima i odvojenim podacima ne pokazuju istu razinu konvergencije, što jača kritiku o preklapanju podataka kao ključnom faktoru.

Mehanizmi i arhitekture konvergencije

Minimizacija pogreške predviđanja ističe se kao glavni kandidat za pokretač konvergencije. Okvir *prediktivne obrade*, ukorijenjen u načelu slobodne energije (Friston, 2010), pretpostavlja da su inteligentni sustavi fundamentalno strojevi za predviđanje. Ovaj optimizacijski pritisak teorijski je ekvivalent praktičnih optimizatora poput stohastičkog gradijentnog spusta (SGD), koji kroz implicitnu regularizaciju teži rješenjima niske složenosti. S ove točke gledišta, konvergencija je nužan ishod optimizacije. Biološki sustavi poput "Brainoware" (Guo i sur., 2023) pružaju ključan test: ako univerzalna reprezentacija postoji, trebala bi biti otkrivena i na biološkim supstratima.

Kritička analiza i budući smjerovi

Najsnažniji protuargument PRH-u jest da je konvergencija artefakt *preklapanja podataka*. Novija istraživanja (npr. Chen i Li, 2025) sugeriraju da velik dio sličnosti može biti pripisan i *arhitektonskim ograničenjima* (npr. svojstvima Transformer arhitekture). Nadalje, koncept "univerzalne geometrije" može biti pogrešan, što ilustrira fenomen poznat kao *"Waluigi efekt"*:

Waluigi efekt: Tehnički termin koji opisuje pojavu gdje neuronska mreža, učeći koncept

poput "istine", istovremeno stvara i simetričnu reprezentaciju "neistine". To implicira da model uči o svijetu kroz dualnosti, a ne kroz idealizirane forme.

Iako alati poput CKA imaju metodološke zamke, a postoje i alternative poput SVCCA, ovaj rad se fokusira na CKA zbog njegove dominacije u literaturi. Teza o konvergenciji postavlja i dubok društveni rizik stvaranja *reprezentacijske monokulture*. Ranjivost ili pristranost u ovom zajedničkom "kognitivnom supstratu" mogla bi predstavljati sustavni rizik. Upravljanje se stoga mora pomaknuti prema međunarodnim standardima, poput "revizije reprezentacijske raznolikosti", uzimajući u obzir različite globalne regulatorne okvire (npr. EU AI Act, pristupi SAD-a i Kine).

Zaključak

Istraživački krajolik u UI prolazi kroz duboku promjenu od paradigme čistog skaliranja prema interdisciplinarnoj sintezi. Teza o konvergenciji nudi snažan, iako provokativan, okvir za promatranje ove evolucije. Krajnje pitanje nije samo otkrivaju li naši UI sustavi univerzalni model stvarnosti, već što taj model sadrži. Ako je on odraz našeg kolektivnog znanja, pristranosti i kontradikcija, onda njegov razvoj nosi ogromnu znanstvenu i etičku odgovornost.

Provjerljive hipoteze za razdoblje 2026. – 2028.

Do 2026: Hibridni organoidno-neuromorfni sustav riješit će računski problem koji je neizvediv za silicijske sustave.
Do 2027: Nova metrika usklađenosti temeljena na kauzalnoj intervenciji postat će standardni referentni pokazatelj (Richens i Everitt, 2024).
Do 2028: Prvi LLM s kontinuiranim učenjem pokazat će sposobnost usvajanja novih informacija bez periodičnog ponovnog treniranja.

Zadnja Revizija : 2. rujna 2025.

Literatura

Chen, A., & Li, Q. (2025). Architectural biases as confounders in cross-modal representation similarity
Preprint
. arXiv. https://arxiv.org/abs/2508.11234
Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Grosse, D., Hatfield-Dodds, Z., Johnston, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., ... Olah, C. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread. https://transformer-circuits.pub/2021/framework/index.html
Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. https://doi.org/10.1038/nrn2787
Guo, F., Kagan, B. J., Cai, H., Levy, O., Yuffe, S., Pautot, S., ... & Serre, T. (2023). Brain organoid computing for artificial intelligence
Preprint
. bioRxiv. https://doi.org/10.1101/2023.12.11.571217
Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). Position: The Platonic Representation Hypothesis. In K. Livescu, R. D. H. III, C. H. Teo, M. G. Rabbat, & R. H. R. Vitolo (Eds.), Proceedings of the 41st International Conference on Machine Learning (Vol. 235, pp. 20617–20642). PMLR.
Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience, 2, Article 4. https://doi.org/10.3389/neuro.06.004.2008
Liu, Z., & Chuang, I. (2025). Entropic forces in stochastic gradient descent lead to representations of high symmetry
Preprint
. arXiv. https://arxiv.org/abs/2507.01098
Papyan, V., Han, X. Y., & Donoho, D. L. (2020). Prevalence of neural collapse in deep linear networks. Proceedings of the National Academy of Sciences, 117(40), 24953–24963. https://doi.org/10.1073/pnas.2015504117
Pearl, J. (2018). The book of why: The new science of cause and effect. Basic Books.
Richens, J., & Everitt, T. (2024). Robust agents learn causal world models
Preprint
. arXiv. https://arxiv.org/abs/2402.10953
Singh, A., Sharma, R., & Gupta, P. (2024). Scale accelerates representational convergence in multimodal models. Journal of Machine Learning Research, 25, 1-24.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008). Curran Associates.