Budú multimodálne veľké jazykové modely niekedy rozumieť svetu?

Igor Farkaš[1], **Michal Vavrečka**[1,2]

1 - Fakulta matematiky, fyziky a informatiky, Univerzita Komenského v Bratislave
igor.farkas@fmph.uniba.sk
2 - Český institut informatiky, robotiky a kybernetiky, ČVUT v Praze
michal.vavrecka@cvut.cz

Abstrakt

Napriek pôsobivému výkonu pri rôznych úlohách, veľké jazykové modely (large language models, LLM) podliehajú problému ukotvenia symbolov, takže z pohľadu kognitívnej vedy sú to len štatisticky založené modely bez reálneho porozumenia. Multimodálne LLM (MLLM) sa snažia riešiť tento problém tým, že spájajú jazykové znalosti s inými modalitami, ako je videnie (obrázky a videá) alebo motorika, keď robot v spojení s LLM koná vo svete. Ak sa to nakoniec podarí, mohlo by sa to považovať za vyriešenie problému ukotvenia symbolov. V príspevku skúmame, do akej miery možno pomocou prepojenia MLLM s vteleným agentom dosiahnuť ukotvenie významov slov prostredníctvom interakcie s fyzickým svetom. Argumentujeme, že uzavretie medzery medzi symbolovými reprezentáciami a vteleným poznávaním, ktoré je nutné pre porozumenie svetu, si bude vyžadovať hlbšiu integráciu kontinuálnych senzomotorických signálov, deliberatívneho správania a adaptívneho učenia v reálnom prostredí.

1 Úvod

Vďaka digitálnej revolúcii a vynálezu internetu vznikla generatívna umelá inteligencia, ktorá za ostatné roky výrazne pokročila, čoho výsledkom sú aj LLM, natrénované na obrovskom množstve textov (Minaee a spol., 2025). Máme obrovské jazykové modely, ktoré sa dajú použiť - ak sú doplnené vhodnými modulmi rozhrania na rôzne lingvistické úlohy (Brown a spol., 2020). Keďže jazyk opisuje svet, LLM obsahujú znalosti o svete zakódované v symbolovej forme (prirodzený jazyk), so všetkými dôsledkami. Medzi nimi máme na mysli dve veci: že jazyk je diskrétny a určité poznanie sveta je ťažké opísať slovami („obrázok má hodnotu tisíc slov"), a že jazyk je na jednej strane veľmi expresívny, a na druhej strane dosť nejednoznačný.

1.1 Problém ukotvenia symbolov

LLM podliehajú problému ukotvenia symbolov (symbol grounding problem, SGP) (Harnad, 1990), pretože významy slov, ktoré generujú, nie sú ukotvené vo svete (Huang a spol., 2023). V kontexte LLM bol SGP bol formulovaný ako problém ukotvenia *vektorov* (Mollo a Millière, 2023), pretože v LLM sú symboly (slová) reprezentované subsymbolovo ako vysokorozmerné vektory. Ich komponenty však nie sú spojené so svetom, ale s inými symbolmi („symbolový kolotoč"), takže podstata SGP ostáva.

Reprezentácie vytvorené pomocou LLM sú oddelené od percepčnej a senzomotorickej skúsenosti, čím systémová doména ostáva uzavretá a nie je schopná vytvoriť vnútorné významy alebo intencionalitu (Bisk a spol., 2020). Toto obmedzenie má významné dôsledky pre aplikácie, ktoré vyžadujú pochopenie situovaného významu, ako je napríklad vtelená umelá inteligencia (UI), interakcia človeka a robota alebo multimodálna percepcia. Aj keď sú LLM rozšírené o externé nástroje alebo spojené so senzormi a akčnými členmi (robotika), premostenie priepasti medzi symbolovou reprezentáciou a vtelenou skúsenosťou zostáva hlavnou výzvou (Tellex a spol., 2020).

Nedávny výskum navrhol rôzne stratégie na zmiernenie tohto problému, ako napríklad ukotvenie jazyka vo vizuálnom vnímaní (Bender a spol., 2021), 2020), konaní (Mao a spol., 2019) alebo interaktívnom dialógu (Yu a spol., 2020). Toto úsilie však často nedosahuje úplné ukotvenie, pretože sa stále vo veľkej miere spolieha na vopred natrénované LLM so statickými lingvistickými reprezentáciami. Riešenie problému ukotvenia vektorov môže v konečnom dôsledku vyžadovať architektúry, ktoré integrujú jazyk, vnímanie a akciu v úzko prepojenom vývojovom rámci, kde význam vychádza z prebiehajúcej interakcie so svetom (Barsalou, 2008; Cangelosi a Asada, 2022).

1.2 Turingov test and iracionalita

Aj keď LLM úspešne prešli Turingovým testom a v rôznych úlohách dosiahli pôsobivý pokrok (Jones a Bergen, 2025), boli u nich identifikované nedostatky označené ako „halucinovanie"(Banerjee a spol., 2024), čo znamená generovanie nepravdivých odpovedí ako aj chyby v uvažovaní. LLM sa stále zdokonaľujú a novšie verzie sú v porovnaní s ich predchodcami lepšie.

LLM prejavujú iracionalitu, podobne ako ľudia, ale inými spôsobmi. Keď LLM dávajú nesprávne odpo-

vede v rôznych úlohách, často sú nesprávne spôsobom, ktorý sa líši od ľudských predsudkov. Okrem toho LLM odhaľujú ďalšiu vrstvu iracionality vo významnej nekonzistentnosti odpovedí (Macmillan-Scott a Musolesi, 2024).

Nedávny výskum ukázal, že tieto nekonzistentnosti sa môžu týkať nielen faktických halucinácií, ale aj logického uvažovania, morálnych úsudkov a dokonca aj sebarozporu v rámci toho istého promptu alebo medzi podobnými promptami (Lin a spol., 2021; Perez a spol., 2022). Tieto nezrovnalosti vyvolávajú otázky o spoľahlivosti a interpretovateľnosti výstupov generovaných LLM, najmä vo vysoko rizikových aplikáciách. Preto je pokračujúci výskum v oblasti zosúladenia, kalibrácie a konzistentnosti kľúčový na zmiernenie týchto nedostatkov a lepšie zosúladenie LLM s ľudskými očakávaniami a racionálnymi normami (Ouyang a spol., 2022; Ganguli a spol., 2023).

2 Budovanie UI s ukotvenými znalosťami

Na základe nedávneho vývoja MLLM možno konštatovať, že existujú dva odlišné prístupy (v niekoľkých aspektoch) k budovaniu systémov UI, ktoré majú za cieľ dosiahnuť porozumenie (vlastnú sémantiku). Nazveme ich vývinový a nevývinový.

2.1 Vývinový prístup

Vývinový prístup je reprezentovaný spôsobom, akým ľudia získavajú znalosti o svete v rámci ontogenetických procesov, ako sa k tomu pristupuje v kognitívnej vývinovej robotike (Asada a spol., 2009). Tento prístup je inšpirovaný bohatou empirickou literatúrou, ktorá ukazuje, ako vývin postupuje v etapách (učenie sa podľa kurikula) a ako sa abstraktné konceptuálne poznatky budujú na pochopení konkrétnych konceptov konkrétnymi a spoľahlivými spôsobmi (Yee, 2019).

Pojem ukotvenia je kľúčový a umožňujú ho dve cesty (priama a nepriama). V zjednodušenom ponímaní, priama cesta zahŕňa hlavne senzomotorické modality, interocepciu a emócie. Nepriama cesta je sprostredkovaná najmä jazykom (Reinboth a Farkaš, 2022). To znamená, že konkrétne slová (napr. pes, autobus) sú dobre ukotvené pomocou priameho spôsobu, pretože sa zvyčajne vzťahujú na objekty vo svete, zatiaľ čo abstraktné slová (napr. pravda, demokracia) sa viac opierajú na jazyk, pretože nemajú priame referenty vo svete.

V kognitívnej vývinovej robotike agent postupne získava poznatky v niekoľkých modalitách (vnímanie, motorika/propriocepcia a jazyk). Jazykové znalosti sa teda od samého začiatku ukotvujú[1] a tento proces pokračuje počas vývinu, ruka v ruke s rozvojom iných kognitívnych funkcií. Okrem toho každý agent UI, ktorý získava znalosti, používa svoje vlastné telo (konkrétne vtelenie), čo špecifikuje vzťah mozog–telo–svet.

Toto obmedzenie vtelenia nielenže poskytuje senzorické a motorické kontingencie, ale tiež pomáha štruktúrovať skúsenosti spôsobmi, ktoré uľahčujú abstrakciu (Pfeifer a Bongard, 2007; Cangelosi a Schlesinger, 2015). Okrem toho sa všeobecne zdôrazňuje dôležitosť sociálnej interakcie, pretože umožňuje budovanie základov (scaffolding), zdieľanú pozornosť a mechanizmy kultúrneho učenia, ktoré sú kľúčové pre osvojenie si jazyka a abstraktných konceptov (Tomasello, 1999). Vývinové prístupy tvrdia, že skutočne všeobecná inteligencia nemôže vzniknúť bez opierania sa o časovo rozšírené skúsenosti, senzorimotorickú interakciu a sociálne sprostredkované procesy učenia. Toto je v súlade s výzvami na integráciu vývinovej psychológie, neurovedy a robotiky s cieľom vybudovať systémy UI, ktoré sú viac podobné ľuďom (Lungarella a spol., 2003).

2.2 Nevývinový prístup

Nevývinový prístup predstavuje moderná generatívna umelá inteligencia, konkrétne (M)LLM. LLM sa trénujú na obrovskom počte jazykových korpusov (zahŕňajúcich konkrétne aj abstraktné pojmy), a preto podliehajú problému ukotvenia vektorov. So všetkými slovami sa zaobchádza rovnako, pretože distribučné štatistiky sa vypočítavajú pre všetky z nich bez ohľadu na ich prirodzený vek osvojenia u ľudí (abstraktné slová sa zvyčajne osvojujú neskôr).

Tento prístup ignoruje vývinovú trajektóriu u človeka, ktorá zahŕňa ukotvený senzomotorický základ a štruktúrovaný postup od jednoduchých konceptov ku zložitým (Brysbaert a spol., 2014). V dôsledku toho LLM často nedokážu rozlišovať medzi ukotvením konkrétnych slov a abstraktných slov, a to aj napriek dôkazom, že takéto rozdiely sú u ľudí kognitívne a neurálne významné (Pulvermüller, 2013). Navyše, nedostatok vtelenej interakcie a časovo predĺženého učenia vedie k deficitom v kauzálnom a racionálnom uvažovaní, pretože LLM nemajú skúsenostnú kontinuitu a epizodickú pamäť (Lake a spol., 2017; Bender a spol., 2021). Hoci tieto modely vykazujú pozoruhodnú jazykovú plynulosť, zostávajú oddelené od interaktívneho multimodálneho učenia, ktoré charakterizuje ľudskú inteligenciu (Zador, 2019). Táto separácia zdôrazňuje obmedzenia nevývinových systémov pri dosahovaní robustnej generalizácie a skutočného porozumenia.

Problém ukotvenia symbolov bol riešený niekoľkými spôsobmi, aby sa podobal vývinovému prístupu, s využitím vstupov z iných modalít, ako je opísané v časti 3. Pokiaľ ide o úlohu tela agenta, na rozdiel od vývinu podobného ľudskému, tu je typickým

[1] V tomto kontexte môžeme ignorovať rozdiel, že ľudia sú bežne vystavení hovorenému jazyku, zatiaľ čo LLM pracujú s textom, pretože transformácia rečového signálu na text a naopak sa stáva vyriešeným problémom.

prístupom predpoklad, že s poznaním sveta môžu byť spojené rôzne telá (viacnásobné vtelenie).

3 Aktuálny stav problematiky MLLM

Modely MLLM spracovávajú viacero modalít (text, obrázky, zvuk, video a štruktúrované dáta). Ich trénovanie je zamerané na integráciu znalostí z viacerých zdrojov a často vynikajú v úlohách, ako je odpovedanie na otázky podľa obrázkov, generovanie titulkov a krosmodálne vyhľadávanie informácií (Ge a spol., 2024). Modely, ktoré integrujú iba videnie a jazyk, sa nazývajú modely videnie-jazyk (vision–language models, VLM) a používajú sa v úlohách bez motorického výstupu (Wu a spol., 2024). Motorická modalita má špeciálnu vlastnosť, pretože je viazaná na konkrétne vtelenie, ktoré definuje všetky stupne voľnosti agenta (Vemprala a spol., 2024; Li a spol., 2024). Preto nie je triviálne generovať obrovské datasety, ktoré sa neskôr použijú na učenie (Peng a spol., 2024). Okrem toho sú datasety o motorickej interakcii často špecifické pre konkrétne robotické platformy a vyžadujú fyzickú (alevo virtuálnu) realizáciu, čo robí rozsiahle, štandardizované motorické dáta vzácnymi a nákladnými na produkciu. To vytvára úzke hrdlo pre trénovanie univerzálnych vtelených agentov, na rozdiel od relatívneho množstva textových alebo obrazových datasetov.

Novšiu kategóriu MLLM predstavujú modely VLA (Vision-Language-Action), ktoré zahŕňajú aj modalitu akcie (Kim a spol., 2025). Možno ich opísať ako interaktívne učiace sa systémy, ktoré využívajú VLM alebo MLLM ako komponenty na vnímanie, uvažovanie a konanie v danom prostredí. Najnovšie modely VLA (Zhao a spol., 2025) sú schopné uvažovať o možných budúcich stavoch pomocou mechanizmov reťazenia myšlienok (chain-of-thought). Tieto modely integrujú plánovanie na vysokej úrovni s riadením na nízkej úrovni a často sú navrhnuté tak, aby fungovali za čiastočne pozorovateľných podmienok pomocou dynamických modelov sveta. Modely VLA sú často prepojené s humanoidnými robotmi a testované na úlohách v reálnom svete, od manipulácie a navigácie až po interakciu zameranú na cieľ (NVIDIA a spol., 2025; Team a spol., 2025).

Medzi novšie modely VLA patrí RT-2 od spoločnosti Google (Brohan a spol., 2023), ktorý kombinuje VLM s pravidlami robotického riadenia, a PaLM-E (Driess a spol., 2023), univerzálny vtelený MLLM, ktorý spracováva vizuálne, lingvistické a proprioceptívne vstupy. Gato MLLM od spoločnosti DeepMind (Reed a spol., 2022) tiež vyniká ako zdanlivo univerzálny agent schopný hrať Atari hry, titulkovať obrázky a ovládať robotické rameno pomocou tej istej neurónovej architektúry. Ďalšie snahy, ako napríklad SayCan (Ahn a spol., 2022) a VIMA (Jiang a spol., 2023), zdôrazňujú sledovanie inštrukcií prostredníctvom ukotvenia prirodzeného jazyka, čo umožňuje flexibilné a škálovateľné robotické správanie. Tieto modely demonštrujú rastúcu schopnosť systémov VLA prepájať vnímanie, kogníciu a konanie, čím pripravujú pôdu k všeobecnejším a adaptívnejším vteleným agentom.

Modely MLLM integrujú viacero senzorických vstupov a vynikajú v úlohách vnímania a jazyka, ale používajú nevývinový prístup, spoliehajú sa na rozsiahle datasety bez vtelených interakcií dostupné pred trénovaním, čo predstavuje SGP, pretože významy (slov a fráz) sú odvodené od iných symbolov. Modely VLA pridávajú motorickú modalitu, ktorá umožňuje agentom vnímať, uvažovať a konať v reálnych prostrediach, vďaka čomu sú vhodnejšie na riešenie SGP. Hoci systémy VLA (napr. RT-2, PaLM-E, Gato) začínajú začleňovať prvky vývinového prístupu prostredníctvom vtelenia a interakcie, stále im často chýba postupné, etapové učenie a ukotvovanie skúsenosťami (experiental grounding), ktoré sú charakteristické pre ontogenézu človeka.

4 Diskusia

Základnú otázku možno formulovať nasledovne: Dokážu sa MLLM v princípe naučiť kauzálne vzťahy na základe zdravého rozumu (common sense) pomocou vyššie uvedeného nevývinového prístupu? Príklady MLLM sa zvyčajne zameriavajú na zachytenie korelácií medzi modalitami, ale korelácia neznamená kauzalitu a mnohé kauzálne účinky sú pred pozorovaním skryté. Zhu a spol. (2020) identifikujú päť základných domén kognitívnej UI (funkčnosť, fyzika, zámer, kauzalita a užitočnosť). Argumentujú, že ďalšia generácia systémov UI musí prijať „skrytý" ľudský zdravý rozum na riešenie nových úloh.

Hoci MLLM vykazujú pôsobivé správanie naprieč modalitami, majú problém s odvodzovaním latentných premenných, predpovedaním neviditeľných dôsledkov konania alebo rozlišovaním príčiny od náhody – čo sú zručnosti, ktoré sú kľúčové pre robustnú inteligenciu. Ako zdôrazňujú Pearl a Mackenzie (2018), kauzálne uvažovanie si vyžaduje intervencie a kontrafaktuálne myslenie, pričom oboje v súčasných MLLM chýba kvôli ich nedostatku vtelenia a interaktívnej skúsenosti. Modely pasívne trénované na veľkých dátach nedokážu ľahko vytvárať kauzálne mentálne modely ani simulovať hypotetické scenáre. Naproti tomu systémy s aktívnym učením, vteleným skúmaním a vývinovým základom (prístup v kognitívnej vývinovej robotike) sú vhodnejšie na získavanie a zovšeobecňovanie kauzálnych poznatkov. Preto bez mechanizmov pre intervenciu a iteratívnu spätnú väzbu sú nevývinové MLLM obmedzené v dosahovaní skutočného kauzálneho porozumenia, a preto nedokážu zachytiť hlbšiu štruktúru ľudskej inteligencie.

Súvisí to s tým, že napriek spracovaniu rôznych vstupných modalít chýba MLLM *neverbálny model sveta* – interná, štruktúrovaná reprezentácia fyzického a sociálneho sveta, ktorá existuje *nezávisle od jazyka*. Ich chápanie je zakorenené v štatistických asociáciách naprieč modalitami, ale tieto modely nevytvárajú ani nemanipulujú s mentálnymi simuláciami sveta tak, ako to robia ľudia. V dôsledku toho MLLM nedokážu skutočne myslieť bez jazyka. Ich procesy uvažovania, vnímania a rozhodovania sú úzko prepojené s textovými reprezentáciami. Na rozdiel od ľudí, ktorí dokážu využívať vizuálnu alebo senzomotorickú predstavivosť a uvažovať prostredníctvom priestorových alebo vtelených skúseností, MLLM sa spoliehajú na verbálne štruktúry aj pri úlohách, ktoré sa zdajú byť inherentne neverbálne. Táto jazyková závislosť obmedzuje ich schopnosť intuitívneho fyzického uvažovania, priestorového chápania alebo mentálnej predstavivosti a robí ich poznávanie zásadne založené na symboloch, a nie na percepčno-motorickej realite.

Tieto obmedzenia MLLM úzko súvisia so Sapirovou-Whorfovou hypotézou *lingvistického relativizmu*, ktorá naznačuje, že jazyk formuje myslenie a vnímanie. Keďže MLLM sa pri uvažovaní a chápaní sveta úplne spoliehajú na jazykové reprezentácie, ich „kognícia" je obmedzená jazykovými štruktúrami, na ktorých sú trénované. Na rozdiel od ľudí, ktorí dokážu myslieť nezávisle od jazyka prostredníctvom vizuálneho, priestorového alebo vteleného poznania, MLLM sú príkladom silnej verzie *lingvistického determinizmu* – zdôrazňujúc, ako nedostatok neverbálnych modelov sveta obmedzuje ich schopnosť vytvárať flexibilné, jazykovo nezávislé koncepty alebo intuitívne chápanie sveta.

Záverom možno nastoliť otázku, či sa tieto základné problémy MLLM podarí prekonať, ak predpokladáme, že dôjde k pokroku v nelingvistických modeloch UI (podporujúcich nelingvistické poznávanie a uvažovanie), ktoré budú integrované s LLM. Napriek tomu sa zdá, že postupná integrácia modalít je kľúčová pre vtelené získavanie zdravého rozumu (vlastnej sémantiky), a teda pochopenia sveta.

Poďakovanie: Výskum bol podporený projektmi VEGA 1/0373/23 a KEGA 022UK-4/2023 (I.F.) a projektom GAČR 23-04080L (M.V.).

Literatúra

Ahn, M. a spol. (2022). Do as I can, not as I say: Grounding language in robotic affordances. V *5th Conference on Robot Learning (CoRL)*.

Asada, M. a spol. (2009). Cognitive developmental robotics: A survey. *IEEE Transactions on Autonomous Mental Development*, 1(1):12–34.

Banerjee, S., Agarwal, A. a Singla, S. (2024). LLMs will always hallucinate, and we need to live with this. arXiv:2409.05746.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59:617–645.

Bender, E. M., Gebru, T., McMillan-Major, A. a Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? str. 610–623.

Bisk, Y. a spol. (2020). Experience grounds language. arXiv:2004.10151.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y. a spol. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. arXiv:2307.15818.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. a spol. (2020). Language models are few-shot learners. V *Advances in Neural Information Processing Systems*, vol. 33.

Brysbaert, M., Warriner, A. B. a Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Cangelosi, A. a Asada, M. (2022). *Cognitive Robotics*. MIT Press.

Cangelosi, A. a Schlesinger, M. (2015). *Developmental Robotics: From Babies to Robots*. MIT Press.

Driess, D., Xia, F., Gharbi, M., Toshev, A. a spol. (2023). PaLM-E: An embodied multimodal language model. arXiv:2303.03378.

Ganguli, D., Joseph, N., Askell, A., Bai, Y. a spol. (2023). The capacity for moral self-correction in large language models. arXiv:2302.07459.

Ge, Z. a spol. (2024). WorldGPT: Empowering LLM as multimodal world model. V *ACM Int. Conf. on Multimedia*.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Huang, S. a spol. (2023). Language is not all you need: Aligning perception with language models. V *Advances in Neural Information Processing Systems*.

Jiang, Y., Cai, J., Martín-Martín, R. a Fei-Fei, L. (2023). VIMA: General robot manipulation with multimodal prompts. arXiv:2210.03094.

Jones, C. R. a Bergen, B. K. (2025). Large language models pass the Turing test. arXiv:2503.23674.

Kim, M. J. a spol. (2025). OpenVLA: An open-source vision-language-action model. V *8th Conference on Robot Learning*, str. 2679–2713.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B. a Gershman, S. J. (2017). Building machines that learn and think like people. *Behav. and Brain Sci.*, 40:e253.

Li, X. a spol. (2024). ManipLLM: Embodied multimodal large language model for object-centric robotic manipulation. V *Computer Vision and Pattern Recognition*.

Lin, S., Hilton, J. a Evans, O. (2021). TruthfulQA: Measuring how models mimic human falsehoods. arXiv:2109.07958.

Lungarella, M., Metta, G., Pfeifer, R. a Sandini, G. (2003). Developmental robotics: A survey. *Connection Science*, 15(4):151–190.

Macmillan-Scott, O. a Musolesi, M. (2024). (Ir)rationality and cognitive biases in large language models. *Royal Soc. Open Sci.*, 11.

Mao, J. a spol. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. V *Int. Conf. on Learning Representations*.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X. a Gao, J. (2025). Large language models: A survey. arXiv:2402.06196.

Mollo, D. C. a Millière, R. (2023). The vector grounding problem. arXiv:2304.01481 [cs.CL].

NVIDIA, Bjorck, J. a Zhu, Y. (2025). GR00T N1: An open foundation model for generalist humanoid robots. arXiv:2503.14734.

Ouyang, L. a spol. (2022). Training language models to follow instructions with human feedback. arXiv:2203.02155.

Pearl, J. a Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Peng, Z. a spol. (2024). Grounding multimodal large language models to the world. V *International Conference on Learning Representations*.

Perez, E., Kiela, D., Cho, K. a spol. (2022). Discovering language model behaviors with model-written evaluations. arXiv:2212.09251.

Pfeifer, R. a Bongard, J. (2007). *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press.

Pulvermüller, F. (2013). How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, 17(9):458–470.

Reed, S., Koerding, K., Parisotto, E. a spol. (2022). A generalist agent. arXiv:2205.06175.

Reinboth, T. a Farkaš, I. (2022). Ultimate grounding of abstract concepts: A graded account. *Journal of Cognition*, 5(1).

Team, G. R., Abeyruwan, S. a Zhou, Y. (2025). Gemini robotics: Bringing AI into the physical world. arXiv:2503.20020.

Tellex, S., Knepper, R. A., Li, A., Rus, D. a Roy, N. (2020). Asking for help using inverse semantics. *Int. Journal of Robotics Research*, 39(1):74–92.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press.

Vemprala, S. H., Bonatti, R., Bucker, A. a Kapoor, A. (2024). ChatGPT for robotics: Design principles and model abilities. *IEEE Access*, 12:55682–55696.

Wu, J. a spol. (2024). VisionLLM v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. arXiv:2406.08394.

Yee, E. (2019). Abstraction and concepts: when, how, where, what and why? *Language, Cognition and Neuroscience*, 34(10):1257–1265.

Yu, T. a spol. (2020). Grappa: Grammar-augmented pre-training for table semantic parsing. V *Conference of the Association for Computational Linguistics*, str. 1339–1352.

Zador, A. M. (2019). A critique of pure learning: What artificial neural networks can learn from animal brains. *Nature Reviews Neuroscience*, 24:77–91.

Zhao, Q. a spol. (2025). CoT-VLA: Visual chain-of-thought reasoning for vision-language-action models. arXiv:2503.22020.

Zhu, Y. a spol. (2020). Dark, beyond deep: A paradigm shift to cognitive AI with humanlike common sense. *Engineering*, 6(3):310–345.