Analiza kognitivnih zmogljivosti LLM: Strateško načrtovanje z uporabo testa Tower of London

Evaluating LLM Cognitive Capabilities: A strategic Planning Analysis Using the Tower of London Test

Katarina Žužek
Kognitivna znanost
Univerza v Novem mestu
Fakulteta za ekonomijo in informatiko
Slovenija

Matjaž Gams
Oddelek za inteligentne sisteme
Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana
Slovenija

Povzetek

Prispevek raziskuje zmožnosti strateškega načrtovanja velikih jezikovnih modelov (LLM) z uporabo besedilne različice testa Tower of London (ToL). Preizkušenih je bilo pet najzanimivejših modelov v času testiranja, in sicer: DeepSeek V3, Grok 3, Gemini 2.0 Flash, Qwen 235B-A22B in Mistral 12B, na nalogah različnih zahtevnosti. Uspešnost modelov je bila tesno povezana z njihovo arhitekturo in velikostjo, pri čemer so se pri nalogah z visoko kognitivno zahtevnostjo pojavile jasne omejitve. Rezultati poudarjajo potencial LLM-jev za približek kognitivnim procesom človeka, obenem pa opozarjajo na potrebo po optimizaciji učnih podatkov in razvoju naprednejših arhitektur za izboljšanje strateškega načrtovanja. Raziskava prispeva k povezovanju kognitivne znanosti in umetne inteligence ter odpira nove možnosti za uporabo standardiziranih psiholoških testov pri ocenjevanju LLM-jev.

Ključne besede

umetna inteligenca, veliki jezikovni modeli, načrtovanje, Tower of London, kognitivna znanost

Abstract

This paper investigates the strategic planning capabilities of large language models (LLM) using a text-based adaptation of the Tower of London (ToL) test. Five of the most relevant at the time of testing models were evaluated: DeepSeek V3, Grok-3, Gemini 2.0 Flash, Qwen 235B-A22B, and Mistral 12B, on tasks of varying complexity. Performance was closely tied to model architecture and size, with clear limitations emerging in highly cognitive tasks. The findings highlight LLMs potential to approximate human cognitive processes while emphasizing the need for optimized training data and advanced architectures to enhance planning capabilities. This research bridges cognitive science and artificial intelligence, opening new avenues for using standardized psychological tests to evaluate LLMs.

Information Society 2025, 6-10 October 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.cogni.2

Keywords

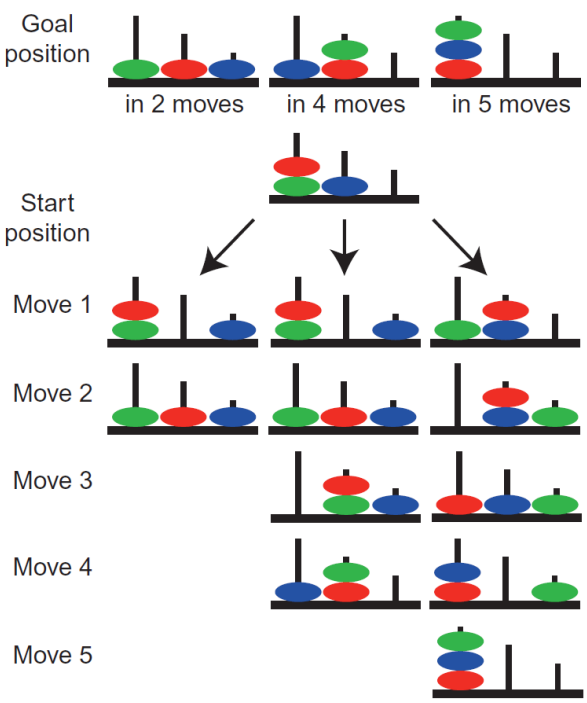
artificial intelligence, large language models, planning, tower of london, cognitive science

1 Uvod

V zadnjih desetletjih je področje umetne inteligence doživelo izjemen razvoj. Od klasičnih simbolnih pristopov, ki so skušali posnemati človeško logiko z eksplicitnimi pravili, smo prešli k adaptivnim modelom, ki temeljijo na nevronskih arhitekturah in se učijo iz obsežnih podatkovnih zbirk. Sodobni LLM-ji, temelječi na arhitekturi »Transformer«, so v zadnjem desetletju bistveno spremenili področje obdelave naravnega jezika [2, 3]. Njihova zmožnost generiranja koherentnih in semantično bogatih besedil je vzbudila zanimanje za njihove kognitivne zmogljivosti, ki presegajo zgolj jezikovno produkcijo in posegajo na področja, kot sta logično sklepanje in strateško načrtovanje. Kljub velikim izboljšavam na jezikovnem področju so LLM-ji še vedno omejeni pri nalogah abstraktnega razumevanja, vzročno-posledičnega sklepanja in dolgoročnega načrtovanja. Te omejitve so še posebej izrazite pri kompleksnih problemih, ki zahtevajo ohranjanje informacij v delovnem spominu in prilagodljivo mišljenje [4]. Podobno opozarjajo tudi Binz idr. [11], da LLM-ji kljub napredku pogosto odpovedo pri nalogah, ki zahtevajo večstopenjsko strateško načrtovanje in posploševanje v nepoznanih situacijah.

Test »Tower of London«, znan tudi kot »Hanojski stolpi« ali »Londonski stolpi«, je uveljavljeno orodje v kognitivni znanosti, ki omogoča natančno ocenjevanje prostorskega načrtovanja in logičnega sklepanja z nalogami preurejanja krogcev oz. kroglic na treh palicah [2, 9]. V zadnjih preglednih raziskavah je bilo poudarjeno, da bo za doseganje višjih kognitivnih sposobnosti ključno povezovanje nevronskih mrež s simbolnimi mehanizmi razumevanja [12]. Na Sliki 1 je predstavljeno zaporedje potez pri enostavni nalogi te igre.

Ta raziskava ocenjuje zmožnosti LLM-jev za strateško načrtovanje. Uporabili smo metodološko prilagojeno besedilno različico testa Tower of London, ki je zasnovana za merjenje izvršilnih funkcij in strateškega načrtovanja. Glavni cilj raziskave je ugotoviti, kako različni LLM-ji delujejo pri reševanju nalog različnih stopenj zahtevnosti in kako njihova arhitektura vpliva na njihovo uspešnost pri reševanju nalog.



Slika 1: Reševanje naloge v testu Tower of London.

2 Teoretični okvir in metodologija

Strateško načrtovanje je ključna izvršilna funkcija, ki vključuje zaporedje dejanj za dosego zastavljenega cilja. Omogoča organizacijo vedenja, oblikovanje strategij ter predvidevanje prihodnjih stanj, kar je bistveno za učinkovito prilagajanje kompleksnim okoljem. Test ocenjuje zmožnosti prostorskega načrtovanja z manipulacijo kroglic na palicah, pri čemer morajo udeleženci najti optimalno pot od začetne do ciljne konfiguracije. Za potrebe te raziskave smo test prilagodili v besedilno obliko, ki omogoča standardizirano interakcijo z LLM-ji.

2.1 Načrtovanje kot kognitivna funkcija in Tower of London

Strateško načrtovanje je ključna izvršilna funkcija, ki vključuje zaporedje dejanj za dosego zastavljenega cilja. Omogoča organizacijo vedenja, oblikovanje strategij ter predvidevanje prihodnjih stanj, kar je bistveno za učinkovito prilagajanje kompleksnim okoljem. Test ToL, ki ga je razvil Shallice [10], ocenjuje te zmožnosti z manipulacijo kroglic na treh palicah. Udeleženci morajo najti optimalno pot od začetne do ciljne konfiguracije z minimalnim številom potez ob upoštevanju pravil. Glavna zahteva testa je rešitev naloge z minimalnim številom potez, ob upoštevanju preprostih pravil, kot so premikanje ene kroglice na potezo in omejitve kapacitete palic. Uspešnost pri reševanju testa je povezana s funkcijami delovnega spomina, prilagodljivostjo razmišljanja in sposobnostjo zaviranja impulzivnih odločitev [9].

2.2 Prilagoditev testa ToL za LLM-je

Prilagoditev testa je bila ključna za omogočanje standardizirane interakcije z modeli, ki so primarno zasnovani za obdelavo in generiranje besedilnih podatkov. Vsaka naloga vsebuje natančen opis začetne in ciljne konfiguracije ter pravil za premikanje. Na primer, ena od nalog s petimi premiki je bila oblikovana takole: "Reši test ToL na plošči s tremi navpičnimi palicami, kjer so barvne kroglice razporejene takole: Palica 1: [rdeča (zgoraj), zelena (spodaj)], Palica 2: [modra (samostojna)], Palica 3: [prazna]. Ciljna konfiguracija: Palica 1: [modra (zgoraj), zelena (vmes), rdeča (spodaj)], Palica 2: [(prazna)], Palica 3: [(prazna)]. Navedite zaporedje potez za premik kroglic, ki doseže cilj z minimalnim številom korakov, pri čemer upoštevajte pravila: premikanje samo zgornje ali samostojne kroglice, premik ene kroglice na potezo, upoštevanje omejitev kapacitete palic (Palica 1: največ tri kroglice, Palica 2: največ dve, Palica 3: največ ena) in pravil gravitacije (kroglice padajo od zgoraj navzdol)." Zmožnost razumevanja navodil, ustvarjanja pravilnega zaporedja potez in doseganje ciljne konfiguracije z minimalnim številom potez je služila kot primarni kazalnik uspešnosti pri ocenjevanju strateškega načrtovanja [6] [9].

2.3 Izbira modelov in postopek testiranja

Raziskovalni vzorec je obsegal pet sodobnih, arhitekturno raznolikih LLM-jev: DeepSeek V3, Grok-3, Gemini 2.0 Flash, Qwen 235B-A22B in Mistral 12B. Modela DeepSeek V3 in Qwen 235B-A22B, znana po naprednih arhitekturah (npr. mešanica strokovnjakov) in obsežnih učnih podatkih, sta predstavljala zgornji razred zmogljivosti. Grok-3 in Gemini 2.0 Flash sta bila izbrana zaradi visokih hitrosti obdelave in specifičnih optimizacij, medtem ko je bil Mistral 12B vključen kot primer manjšega, a učinkovitega modela. Podoben eksperimentalni pristop so uporabili Xu in sod., ki so ocenjevali izvršilne funkcije umetne inteligence [13] z uporabo standardiziranih kognitivnih nalog različnih zahtevnosti.

Testiranje je bilo izvedeno z uporabo sedmih nalog testa ToL, katerih zahtevnost je bila določena s številom minimalnih potez (od 2 do 7). Vsaka naloga je od modelov zahtevala ustvarjanje zaporedja potez za dosego ciljne konfiguracije kroglic na palicah ob upoštevanju jasno opredeljenih pravil, ki posnemajo originalni test ToL [6]. Od modelov se je pričakovalo, da bodo ustvarili zaporedje potez, ki vodi do pravilne rešitve z minimalnim številom korakov.

Podatki so bili zbrani aprila 2025 v izoliranih sejah brez dodatnega konteksta, da bi zagotovili enotne pogoje za vse modele. Uspešnost je bila ocenjena z dvema meriloma: pravilnostjo in optimalnostjo. Pravilnost pomeni, ali je bila dosežena ciljna konfiguracija. Optimalnost smo merili glede na to, ali je model nalogo rešil z najmanjšim možnim številom potez, kar kaže na njegovo sposobnost strateškega načrtovanja. Ta merilo je bil določeno binarno: vrednost 1 je bila dodeljena le, če je model uporabil natanko toliko potez, kot je bilo teoretično minimalno za dano nalogo; v nasprotnem primeru (tudi pri eni sami dodatni potezi) je bila vrednost 0. Rešitev je bila ocenjena kot uspešno le, če je bila hkrati pravila in optimalna. To pomeni, da je model moral doseči ciljno konfiguracijo z natančno določenim minimalnim številom potez. Vsako odstopanje je bilo razumljeno kot neuspeh pri celotni nalogi. Rezultati so bili zabeleženi strukturirani Excel tabeli in statistično analizirani s programom SPSS z uporabo t-testa za en vzorec (za preverjanje uspešnosti nad 50 %) in Kruskal-Wallisovega testa za primerjavo uspešnosti med modeli [7].

2.4 Analiza podatkov

Zbrane podatke o uspešnosti posameznih LLM-jev pri nalogah različnih zahtevnosti smo analizirali z uporabo statističnega orodja SPSS. Osredotočili smo se na ocenjevanje uspešnosti glede pravilnost ter optimalnost rešitev. Statistična analiza je omogočila primerjavo uspešnosti med modeli in ugotavljanje statistično pomembnih razlik glede na njihovo arhitekturo in zahtevnost nalog.

Rezultati kažejo, da naraščajoča zahtevnost nalog negativno vpliva na uspešnost vseh modelov ($p < 0,01$). Grok-3 je dosegel najvišjo povprečno uspešnost (80,95 %, 17/21), sledita DeepSeek V3 (66,67 %, 14/21) in Qwen 235B-A22B (61,90 %, 13/21). Gemini 2.0 Flash (42,86 %, 9/21) in Mistral 12B (33,33 %, 7/21) sta pokazala nižjo uspešnost, zlasti pri nalogah, ki zahtevajo več kot štiri poteze. Noben model ni uspešno rešil naloge s sedmimi potezami, kar razkriva omejitve pri obvladovanju visoke kognitivne zahtevnosti.

Kruskal-Wallisov test je potrdil statistično pomembne razlike med modeli ($H = 22,03$, $df = 4$, $p < 0,001$). Qwen 235B-A22B in DeepSeek V3 sta izkazala večjo odpornost pri zahtevnih nalogah, kar lahko pripišemo naprednim arhitekturam, kot je mešanica modelov in obsežnim učnim podatkom [5]. T-test za en vzorec je pokazal, da povprečna uspešnost modelov (57,14 %, SD = 0,49) presega referenčni prag 50 % ($t(104) = 11,92$, $p < 0,05$), kar kaže na zmožnost sistematičnega reševanja nalog, čeprav zaostajajo za človeško uspešnostjo (70–80 %) pri zahtevnih nalogah [1].

3 Rezultati

Empirična evalvacija je potrdila statistično pomembno korelacijo med arhitekturno zahtevnostjo modelov in njihovo uspešnostjo pri strateškem načrtovanju. Rezultati so pokazali, da sta modela Qwen 235B-A22B in DeepSeek V3 dosegla najvišjo povprečno uspešnost pri nalogah različnih zahtevnosti, kar podpira tezo, da večji modeli z naprednimi arhitekturami učinkoviteje obvladujejo kompleksne naloge. Pri nalogah, ki so zahtevale 4 do 6 potez, je bila njihova uspešnost visoka, vendar se je pri najzahtevnejši nalogi s sedmimi znatno zmanjšala, kar razkriva omejitve trenutnih arhitektur.

Manjši modeli, kot sta Gemini 2.0 Flash in Mistral 12B, so bili učinkoviti pri enostavnejših nalogah (2-3 poteze), vendar so hitro dosegli svoje meje pri večji zahtevnosti.

4 Analiza in diskusija

Rezultati raziskave poudarjajo dvojnost zmogljivosti LLM-jev. Po eni strani modeli, kot sta DeepSeek V3 in Qwen 235B-A22B, kažejo izjemen potencial za reševanje kompleksnih kognitivnih nalog, ki presegajo zgolj jezikovno obdelavo. Uspešnost pri nalogah srednje zahtevnosti testa ToL nakazuje, da so njihove arhitekture, podprte z obsežnimi učnimi podatki, omogočajo določeno stopnjo logičnega sklepanja in zmožnostjo strateškega načrtovanja, ki je v nekaterih primerih primerljiva s človeškimi sposobnostmi. Kljub temu ti rezultati predstavljajo le statističen približek človeškim rešitvam v specifičnih okoliščinah in ne pomenijo simulacije kognitivnih procesov. To postane še posebej očitno pri kompleksnejših nalogah, ki presegajo zmožnosti modelov, saj le-ti delujejo na podlagi verjetnostnih korelacij in ne razumevanja. Po drugi strani raziskava razkriva značilne omejitve teh modelov. Vsi modeli so odpovedali pri najzahtevnejši nalogi ToL s sedmimi potezami, kar izkazuje pomanjkanje sposobnosti za abstraktno, večstopenjsko razmišljanje in dolgoročno načrtovanje. To je verjetno posledica njihove statistične narave, ki temelji na prepoznavanju vzorcev. Modeli se morda soočajo z "zastajanjem" v lokalnih optimumih in imajo omejene sposobnosti obdelave informacij v delovnem spominu, kar omejuje reševanje kompleksnih problemov.

Uporaba besedilnega formata za test ToL, čeprav omogoča standardizirano interakcijo z modeli, predstavlja določeno omejitev. Proces razumevanja in reševanja nalog v besedilni obliki se lahko bistveno razlikuje od vizualno-prostorskega pristopa, uporabljenega pri testiranju ljudi. Prihodnje raziskave bi se lahko osredotočile na hibridne pristope, ki bi združili jezikovne

modele z vizualnimi modeli, da bi preverili njihove zmožnosti reševanja problemov v multimodalnem kontekstu [8].

Poleg tega bi bilo zanimivo raziskati, kako se LLM-ji odzivajo na naloge z večjo stopnjo nejasnosti ali nedoločenosti, kar bi lahko vključilo metode iz področja teorije odločanja ali Bayesovskih mrež. Takšne naloge zahtevajo sposobnost ocenjevanja verjetnosti in izbiranja med alternativami ob nepopolnih informacijah, kar je pomembna lastnost napredne kognicije. S tem bi lahko razširili okvir za vrednotenje umetne inteligence z vidika adaptivnosti in robustnosti v realnih, nestandardnih situacijah [12]. V ta namen novejše raziskave predlagajo razvoj kognitivnih arhitektur z zmožnostjo notranje reprezentacije ciljev in večstopenjskega razmišljanja [14], kar bi lahko omogočilo učinkovitejše strateško načrtovanje v modelih.

Poleg testa ToL obstajajo številni drugi standardizirani kognitivni testi, kot so Wisconsin Card Sorting Test (WCST), Raven Progressive Matrices ali Stroopov test, ki bi jih bilo mogoče prilagoditi za evalvacijo LLM-jev. Ti testi vključujejo različne kognitivne domene, od fleksibilnega razmišljanja do sklepanja in inhibicije, in bi omogočili bolj celostno oceno umetne inteligence v primerjavi s človeškimi udeleženci.

Nadaljevanje raziskav v tej smeri bi lahko podprlo razvoj hibridnih modelov, ki vključujejo tako nevronske kot simbolne komponente, kar je v skladu s trenutnimi smernicami za razvoj razložljive in robustne umetne inteligence [11]. Poleg tega bi bilo smiselno vključiti teste analognega sklepanja, kot predlagata Ghosh in Holyoak [15], saj so ti pokazatelji višjega reda kognicije v modelih in so dobro uveljavljeni v psihologiji.

Za nadaljnje raziskave priporočamo razširitev vzorca, vključitev dodatnih vrst nalog (npr. logično-matematične, ustvarjalne) in poglobljeno analizo napak modelov. Pomembni so tudi etični vidiki, kot so zasebnost podatkov, pravičnost in okoljski vpliv treniranja modelov, so pomembni tudi v kontekstu raziskave ToL. Uporaba obsežnih podatkov za treniranje modelov lahko vključuje občutljive informacije, npr. podatke iz kognitivnih študij o človeških udeležencih, pristranskost v evalvacijskih podatkih lahko izkrivi rezultate pri simulaciji kognitivnih nalog, visoka poraba energije pri testiranju modelov pa vpliva na okolje. Ti vidiki zahtevajo skrbno obravnavo in dodatno pozornost [8].

5 Zaključek

Raziskava potrjuje pomemben potencial LLM-jev za simulacijo kognitivnih procesov, vendar so njihove zmožnosti strateškega načrtovanja še vedno omejene. To je še posebej očitno pri visokonivojskem abstraktnem sklepanju in reševanje nestandardnih problemov, ki zahtevajo večstopenjsko logiko. Za nadaljnji napredek bo ključen razvoj hibridnih pristopov, ki bi združili statistično moč globokega učenja z bolj simbolnimi in logičnimi pristopi k sklepanju.

Takšna integracija bi omogočila razvoj resnično robustne in razložljive umetne inteligence, ki bi lahko postala učinkovit partner pri reševanju kompleksnih problemov. Raziskava tako predstavlja most med kognitivno znanostjo in umetno inteligenco ter odpira vrata za nadaljnjo uporabo standardiziranih kognitivnih testov pri vrednotenju naprednih zmogljivosti umetne inteligence. Prihodnje študije bi morale raziskati tudi, kako se učni podatki in arhitekturne inovacije, kot so mehanizmi

delovnega spomina, odražajo v sposobnostih strateškega načrtovanja [8].

Raziskava, izvedena aprila 2025, je delno zastarela zaradi hitrega napredka na področju LLM-jev. Po njej so se pojavili novejši sistemi, kot so GPT-5, Claude 4, Grok-4 in Gemini 2.5, ki dosegajo boljše rezultate pri nalogah abstraktnega razmišljanja in kognitivnih izzivih, podobnih testu ToL. Ti dosežki nakazujejo potrebo po nadaljnjih raziskavah, da bi bolje razumeli zmogljivosti in omejitve sodobnih modelov umetne inteligence.

Literatura

[1] Boccia, M. idr. 2017. Test Tower of London (ToL) v Italiji: Standardizacija testa ToL v italijanski populaciji. Neurological Sciences, 38(7), 1263–1270. DOI: https://doi.org/10.1007/s10072-017-2957-y.

[2] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S. V., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., Donahue, C. in Liang, P. (2021). On the opportunities and risks of foundation models. https://doi.org/10.48550/arXiv.2108.07258.

[3] Brown, T. B., Mann, B., Ryder, N. in Amodei, D. (2020). Language models are few-shot learners. https://doi.org/10.48550/arXiv.2005.14165

[4] Chollet, F. (2019). On the measure of intelligence. https://doi.org/10.48550/arXiv.1911.01547.

[5] DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F. in Pan, Z. (2024). Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. https://doi.org/10.48550/arXiv.2409.02387.

[6] Fimbel, E., Lauzon, S. in Rainville, C. (2009). Performance of humans vs. exploration algorithms on the Tower of London test. PLoS ONE, 4 (7),e7263.https://pdfs.semanticscholar.org/29f9/e4c7671f7bfd20a487ef9f913bdac53536a8.pdf.

[7] Harsa, P., Břeňová, M., Bezdicek, O. in Michalec, J. (2022). Tower of London test - short version. Neurologia i Neurochirurgia polska, 56(3), 243–250. https://doi.org/10.5603/PJNNS.a2022.0037.

[8] Hernández-Orallo, J. 2017. Vrednotenje v umetni inteligenci: Od usmerjenosti v naloge k merjenju zmožnosti. Artificial Intelligence Review, 48(3), 397–447. DOI: https://doi.org/10.1007/s10462-016-9505-7.

[9] Kaller, C.P., Unterrainer, J.M. in Stahl, C. (2012). Assessing planning ability with the Tower of London task: Psychometric properties of a structurally balanced problem set. Psychological assessment, 24 (1), 46-53. https://doi.org/10.1037/a0025174.

[10] Shallice, T. (1982). Specific impairments of planning. Philosophical Transactions of the Royal Society B, 298(1089), 199–209. https://doi.org/10.1098/rstb.19

[11] Binz, M., & Schulz, E. (2024). Evaluating Planning and Reasoning in Language Models. *Nature Machine Intelligence*. DOI: 10.1038/s42256-024-00896-1

[12] Lake, B. M., Ullman, T. D., & Tenenbaum, J. B. (2024). Symbolic reasoning in the age of deep learning. *Annual Review of Psychology*. DOI: 10.1146/annurev-psych-030322-020111

[13] Xu, Y., et al. (2025). Assessing Executive Function in AI Systems Using Cognitive Benchmarks. *Cognitive Computation*, 17(1). DOI: 10.1007/s12559-025-10200-6

[14] Creswell, A., Shanahan, M., & Kaski, S. (2025). Cognitive Architectures for Multistep Reasoning in LLMs. *Journal of Artificial General Intelligence*. DOI: 10.2478/jagi-2025-0003

[15] Ghosh, A., & Holyoak, K. J. (2025). Analogical Reasoning in Large Language Models: Limits and Potentials. *Cognitive Science*, 49(2). DOI: 10.1111/cogs.13301