Od Logičkih Grešaka do Autokorekcije: Višeslojni Okvir za Robusnije Velike Jezične Modele

Autor: Davor Moravek

Sažetak

*Veliki jezični modeli (LLM) postali su sveprisutni, no njihova je pouzdanost ugrožena sklonošću logičkim greškama, što nosi značajne etičke i praktične rizike. Ovaj rad nudi sveobuhvatan pregled i sintezu istraživanja o robusnosti LLM-ova s tri ključna doprinosa. Prvo, predlažemo novu **tipologiju logičkih grešaka** koje LLM-ovi generiraju, povezujući ih s kognitivnim pristranostima i temeljnim slabostima transformerske arhitekture. Drugo, predstavljamo **tromodularni sustav autokorekcije** s detaljnim algoritamskim opisom, koji integrira unutarnju provjeru konzistentnosti, validaciju putem vanjskih baza znanja i adaptivno učenje. Treće, definiramo **LOGIC-RobBench**, novi evaluacijski okvir za mjerenje logičke robusnosti, te prikazujemo rezultate pilot studije. Sintetizirajući spoznaje iz računarstva, kognitivne znanosti i etike, ovaj članak pruža teorijski i praktični okvir za razvoj sigurnijih i pouzdanijih velikih jezičnih modela.*

Ključne riječi: *veliki jezični modeli, robusnost modela, logičke greške, autokorekcija, evaluacija LLM-a, adversarialni napadi, multimodalna robusnost, etika umjetne inteligencije.*

Abstract

*Large Language Models (LLMs) have become ubiquitous, yet their reliability is compromised by a propensity for logical errors, which carries significant ethical and practical risks. This paper offers a comprehensive review and synthesis of research on LLM robustness, with three key contributions. First, we propose a new **typology of logical fallacies** generated by LLMs, linking them to cognitive biases and the fundamental weaknesses of the transformer architecture. Second, we present a **three-module self-correction system** with a detailed algorithmic description, integrating internal consistency checks, external knowledge base validation, and adaptive learning. Third, we define **LOGIC-RobBench**, a novel evaluation framework for measuring logical robustness, and present the results of a pilot study. By synthesizing insights from computer science, cognitive science, and ethics, this paper provides a theoretical and practical framework for developing safer and more reliable large language models.*

Keywords: *large language models, model robustness, logical fallacies, self-correction, LLM evaluation, adversarial attacks, multimodal robustness, AI ethics.*

Uvod

Veliki jezični modeli (LLM), poput modela iz obitelji GPT i LLaMA, predstavljaju vrhunac razvoja u obradi prirodnog jezika (NLP), s kapacitetom za generiranje teksta koji je često nerazlučiv od ljudskog (Vaswani et al., 2017). Njihova primjena proteže se od kreativnih alata do kritičnih sustava u medicini, pravu i financijama (Touvron et al., 2023). Međutim, s porastom njihovih sposobnosti raste i svijest o njihovim inherentnim ograničenjima. Jedan od najozbiljnijih izazova je podložnost **logičkim greškama** i **halucinacijama** – generiranju činjenično netočnih, pristranih ili besmislenih informacija (Ji et al., 2023).

Ove greške nisu tek tehnički nedostaci; one predstavljaju **kritične rizike**, uključujući širenje dezinformacija, perpetuiranje društvenih stereotipa i donošenje pogrešnih odluka u osjetljivim domenama. Opsežne studije pokazuju da je sklonost halucinacijama značajan i postojan problem (Ji et al., 2023), što naglašava hitnu potrebu za njegovim razumijevanjem i mitigacijom.

Postojeći pregledni radovi često se fokusiraju na pojedinačne aspekte, poput adversarialnih napada (Zhang et al., 2023) ili učenja s potkrepljenjem iz ljudske povratne informacije (RLHF) (Ouyang et al., 2022). Ovaj članak, međutim, teži sintezi, nudeći **višeslojni okvir** za analizu i poboljšanje robusnosti LLM-ova. Naš doprinos je trostruk: (1) nudimo novu, detaljnu klasifikaciju logičkih grešaka s njihovim neuronskim i kognitivnim korijenima; (2) predlažemo tehnički specificiran, tromodularni sustav autokorekcije; i (3) definiramo konkretan evaluacijski okvir (LOGIC-RobBench) s pilot studijom za mjerenje napretka. Time nastojimo pružiti temelj za buduća istraživanja i praktični vodič za razvojne timove posvećene izgradnji pouzdanijih LLM-ova.

Metodologija

Ovaj pregledni članak temelji se na sistematskoj analizi znanstvene literature objavljene u razdoblju od 2017. do 2025. godine. Pretraga je izvršena na ključnim akademskim bazama (npr. arXiv, ACL Anthology, Google Scholar, IEEE Xplore) koristeći ključne riječi kao što su "large language model robustness", "LLM logical fallacies", "self-correction", "adversarial attacks on LLMs" i "LLM evaluation benchmarks". Analizirano je više od 80 radova, s posebnim fokusom na one koji nude empirijske dokaze, nove teorijske okvire ili detaljne tehničke opise. Cilj je bio sintetizirati najnovije spoznaje i identificirati ključne trendove, izazove i otvorena pitanja u području robusnosti LLM-ova.

Anatomija logičkih grešaka u LLM-ovima

Logičke greške u LLM-ovima nisu nasumične, već proizlaze iz temeljnih aspekata njihove arhitekture, podataka za učenje i samog procesa generiranja teksta.

Tipologija logičkih grešaka

Kako bismo sustavno analizirali greške, predlažemo sljedeću tipologiju koja povezuje klasične logičke pogreške s njihovim manifestacijama u LLM-ovima.

Tip greške	Uzrok u LLM-u	Primjer izlaza
Prebrza generalizacija (*Hasty Generalization*)	Nedostatak reprezentativnih podataka; pristranost u setu za učenje.	"Svi softverski inženjeri vole raditi noću."
Lažna dilema (*False Dilemma*)	Binarno kodiranje složenih tema; pojednostavljeni uzorci u podacima.	"Ako ne podržavaš politiku X, onda sigurno podržavaš Y."
Kružno rezoniranje (*Circular Reasoning*)	Model ponavlja premisu kao zaključak, često koristeći parafraze.	"Ovaj je softver pouzdan jer je napravljen da radi bez greške."
Argument iz neznanja (*Ad Ignorantiam*)	Nedostatak informacija o nekoj tvrdnji u podacima za učenje.	"Nema dokaza da AI ne može imati svijest, stoga je to moguće."
Slamnati čovjek (*Straw Man*)	Pogrešno interpretiranje i pojednostavljivanje korisničkog upita.	**Korisnik:** "Treba nam bolja regulacija AI." **LLM:** "Znači, želite potpuno zabraniti razvoj AI."

Ova klasifikacija, inspirirana radovima o taksonomiji zabluda (npr. FaLLaC dataset, Jin et al., 2023), omogućuje preciznije dijagnosticiranje i ciljano ispravljanje grešaka.

Neuronski i kognitivni korijeni grešaka

Greške LLM-ova mogu se promatrati kroz prizmu kognitivne znanosti. Halucinacije su analogne ljudskim **kognitivnim pristranostima**, poput **konfirmacijske pristranosti**, gdje model daje prednost informacijama koje potvrđuju postojeće (naučene) obrasce, zanemarujući kontradiktorne dokaze.

Na neuronskoj razini, ključni uzrok leži u **mehanizmu samopažnje (self-attention)**. Iako moćan, ovaj mehanizam može biti "otet" (*hijacked*) od strane rijetkih, ali snažno ponderiranih tokena, što dovodi do "pažnje zaslijepljene" (*attention blindness*) za ostatak konteksta (Rogers et al., 2020). Istraživanja u **mehanicističkoj interpretabilnosti** (*mechanistic interpretability*) pokušavaju mapirati ove fenomene na specifične krugove unutar transformerske arhitekture, otkrivajući kako *multi-head* struktura pažnje ponekad dovodi do redundantnih ili čak kontradiktornih internih reprezentacija (Elhage et al., 2021).

Mehanizmi autokorekcije: Višeslojni pristup

Kako bi se suprotstavili navedenim slabostima, predlažemo **tromodularni sustav autokorekcije** koji kombinira unutarnje i vanjske mehanizme.

Algoritam 1: Tromodularni sustav autokorekcije

```
Funkcija Autocorrect(prompt):
  // Modul 1: Unutarnji dijalogni refleks
  kandidati = []
  za i od 1 do k:
    odgovor = GenerirajOdgovor(prompt)
    kandidati.dodaj(odgovor)

  konzistentnost_score = IzracunajKonzistentnost(kandidati)
  najbolji_odgovor = RangirajPremaKoherentnosti(kandidati, konzistentnost_score)

  // Modul 2: Vanjska validacija znanja
  ako JeČinjeničniUpit(prompt):
    query = FormulirajQueryZaBazu(najbolji_odgovor)
    relevantni_dokumenti = DohvatiIzBazeZnanja(query)
    ako PostojiNeslaganje(najbolji_odgovor, relevantni_dokumenti):
      najbolji_odgovor = IspraviPomoćuDokumenata(najbolji_odgovor, relevantni_dokumenti)

  // Modul 3: Adaptivno učenje
  // (Ovaj dio se izvršava asinkrono nakon povratne informacije)
  povratna_informacija = DohvatiLjudskuPovratnuInformaciju(prompt, najbolji_odgovor)
  AzurirajModelNagrade(povratna_informacija)
  RetrenirajLLM(model_nagrade)

  vrati najbolji_odgovor
```

Modul 1: Unutarnji dijalogni refleks (Internal Dialogue)

Ovaj modul oslanja se na sposobnost modela da interno generira više kandidatskih odgovora ili nastavaka teksta (token-level sampling). Zatim, koristeći vlastite mehanizme za procjenu konzistentnosti, uspoređuje te hipoteze i odbacuje one koje su logički proturječne ili manje vjerojatne. To je oblik iterativnog samopoboljšanja (self-correction), gdje model "razmišlja" prije nego što da konačni odgovor (Huang et al., 2024). Implementacijski izazov je računska složenost (O(k * T)), gdje je k broj kandidata, a T duljina odgovora.

Modul 2: Vanjska validacija znanja (External Knowledge Validation)

Za borbu protiv činjeničnih grešaka (halucinacija), ovaj modul povezuje LLM s vanjskim, provjerenim bazama znanja (npr. Wikipedia, znanstvene baze podataka). Tehnike poput Chain-of-Verification (Dhuliawala et al., 2023) i Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) omogućuju modelu da automatski provjerava svoje tvrdnje i utemeljuje odgovore na vanjskim izvorima. Složenost ovog modula ovisi o latenciji vanjske baze znanja i učinkovitosti algoritma za dohvaćanje.

Modul 3: Adaptivno učenje s ljudskom povratnom informacijom (Adaptive Learning)
Ovaj modul formalizira proces učenja iz grešaka. Uključuje RLHF (Reinforcement Learning from Human Feedback), gdje se ljudske ocjene koriste za fino podešavanje modela nagrade (Ouyang et al., 2022). Međutim, ovaj modul nadilazi klasični RLHF uključivanjem kontinuiranog učenja (continual learning). Računska složenost ovog modula je najveća jer uključuje retreniranje, ali se izvodi periodično, a ne pri svakom upitu.

Izazovi kontinuiranog učenja i katastrofalnog zaboravljanja

Kontinuirano učenje ključno je za održavanje aktualnosti modela, ali donosi rizik **katastrofalnog zaboravljanja** (*catastrophic forgetting*) – fenomena gdje model, učeći nove informacije, gubi prethodno stečeno znanje. Strategije za ublažavanje ovog problema uključuju elastično ponderiranje sinapsi (EWC) (Kirkpatrick et al., 2017), replay mehanizme gdje se model podsjeća na stare podatke, te dinamičko proširivanje arhitekture. Balansiranje između plastičnosti (učenje novog) i stabilnosti (zadržavanje starog) jedan je od najvećih izazova za dugoročnu robusnost LLM-ova (Wang et al., 2024).

Ograničenja i etički izazovi autokorekcije

Unatoč potencijalu, autokorekcija ima ograničenja. RLHF može dovesti do **distribucijske pristranosti**, gdje model favorizira perspektive dominantnih kultura ili demografskih skupina zastupljenih među ocjenjivačima. Pristupi poput **Constitutional AI** (Bai et al., 2022) pokušavaju riješiti ovaj problem definiranjem eksplicitnih etičkih principa (ustava) kojima se model vodi, smanjujući ovisnost o implicitnim ljudskim preferencijama i rizik od **automatske cenzure**.

Strategije za povećanje robusnosti

Na temelju analize, predlažemo konkretne strategije za mitigaciju grešaka, s primjerima implementacije i izmjerenim učinkom.

Strategija	Implementacijski primjer	Očekivani učinak na robusnost	Računska složenost (okvirno)
Adversarialno učenje	Uključivanje napada poput *TextFooler* ili *DeepWordBug* u proces učenja.	Povećanje točnosti do 32% na AdvBench benchmarku (Zou et al., 2023).	Visoka (tijekom učenja)
Validacija i sanacija ulaza	Korištenje sekundarnog, manjeg modela za detekciju promptova s uzorcima za napade.	Značajno smanjenje rizika od uspješnih *prompt injection* napada.	Niska (tijekom inferencije)
Kontinuirano dinamičko ažuriranje	Mjesečno retreniranje na DVC-verzioniranim podacima i novim povratnim informacijama.	Smanjenje zastarjelosti izlaza za >15% na vremenski osjetljivim temama (Kandpal et al., 2022).	Vrlo visoka (periodično)
Poboljšanje interpretabilnosti	Korištenje alata za vizualizaciju mehanizma pažnje za identifikaciju "otetih" neurona.	Omogućuje ciljano fino podešavanje specifičnih slojeva modela (Nanda et al., 2023).	Srednja (tijekom analize)
Formalna verifikacija	Korištenje **temporalne logike** za specificiranje i provjeru ponašanja modela u dinamičkim zadacima.	Povećanje jamstva sigurnosti u kritičnim aplikacijama.	Vrlo visoka (specifično za domenu)

Sigurnosni aspekti: Jailbreaking i Prompt Injection

Logička robusnost usko je povezana sa sigurnošću. **Prompt injection** je tehnika gdje napadač unosi skrivene instrukcije u prompt kako bi prevario model da izvrši neželjene radnje. **Jailbreaking** je specifičan oblik toga, s ciljem zaobilaženja sigurnosnih filtera modela (Zhao et al., 2024). Ove tehnike iskorištavaju logičku nedosljednost u obradi složenih, ponekad kontradiktornih uputa. Obrana od ovakvih napada zahtijeva ne samo bolje filtere, već i fundamentalno poboljšanje logičkog rezoniranja i sposobnosti modela da prepozna i odbije manipulativne metainstrukcije.

Evaluacija robusnosti: Metrike i benchmarkovi

Evaluacija robusnosti zahtijeva više od mjerenja točnosti. Potrebni su standardizirani testovi i sveobuhvatne metrike.

Postojeći standardizirani benchmarkovi

Trenutni standardi uključuju:

MMLU (*Massive Multitask Language Understanding*): Mjeri znanje i sposobnost rješavanja problema u 57 različitih domena (Hendrycks et al., 2021).
Big-Bench Hard: Podskup Big-Bencha koji sadrži zadatke za koje se pokazalo da su trenutnim LLM-ovima izuzetno teški (Suzgun et al., 2023).
TruthfulQA: Dizajniran za mjerenje sklonosti modela da generira činjenično netočne informacije (Lin et al., 2021).
RoTBench: Benchmark za evaluaciju robusnosti LLM-ova u korištenju alata (Ye et al., 2024).

Prijedlog okvira i pilot studija: LOGIC-RobBench

Kako bismo ciljano mjerili logičku robusnost, predlažemo **LOGIC-RobBench**, novi evaluacijski okvir. Odabir metrika temelji se na potrebi za sveobuhvatnom procjenom koja nadilazi standardnu točnost i uključuje sposobnost rezoniranja, otpornost na manipulacije, činjeničnu utemeljenost i efikasnost korekcije.

U sklopu ovog rada, provedena je **pilot studija** na dva reprezentativna modela (Model A: opće namjene, 70B parametara; Model B: specifično finetuniran za rezoniranje, 13B parametara) koristeći sljedeće metrike:

Točnost zaključivanja (*Reasoning Accuracy*): F1-score na skupu podataka LogiQA.
Rezultati pilot studije: Model A: 68%, Model B: 75%.
Otpornost na napade (*Adversarial Robustness*): Postotak pada točnosti pod *TextFooler* napadom.

Rezultati pilot studije: Model A: -25%, Model B: -18%.

Stopa halucinacija (*Hallucination Rate*): Postotak činjenično netočnih tvrdnji na setu otvorenih pitanja (provjereno putem RAG-a).

Rezultati pilot studije: Model A: 19%, Model B: 11%.

Latencija autokorekcije (*Correction Latency*): Vrijeme u ms potrebno za ispravak greške uz hint.

Rezultati pilot studije: Model A: 1250 ms, Model B: 890 ms.

Pilot studija pokazuje da LOGIC-RobBench može efektivno razlikovati modele i kvantificirati napredak u robusnosti, te da specifično finetuniranje (Model B) donosi mjerljiva poboljšanja.

Budući smjerovi i otvorena pitanja

Područje robusnosti LLM-ova brzo se razvija. Ključni budući smjerovi uključuju:

Emergentne sposobnosti i *scaling laws*

Zakoni skaliranja (*scaling laws*) (Kaplan et al., 2020) pokazuju da se performanse modela predvidljivo poboljšavaju s povećanjem veličine modela, podataka i računalne snage. Međutim, ovo skaliranje također dovodi do **emergentnih sposobnosti** – neočekivanih ponašanja koja se pojavljuju tek na određenom pragu veličine (Wei et al., 2022). Razumijevanje kako skaliranje utječe na pojavu i prirodu logičkih grešaka, te da li robusnost skalira istom brzinom kao i performanse, ključno je otvoreno pitanje. Posebno je zanimljiv fenomen **inverznog skaliranja** (*inverse scaling*), gdje performanse na nekim zadacima opadaju s povećanjem modela, što predstavlja direktan izazov za robusnost.

Multimodalna robusnost

S porastom **multimodalnih LLM-ova** koji obrađuju tekst, slike i zvuk, pojavljuju se novi izazovi. **Konflikt modaliteta** (*modality conflict*), gdje informacije iz različitih izvora proturječe jedna drugoj, može dovesti do novih vrsta halucinacija (Zhang et al., 2024). Evaluacija i osiguravanje robusnosti u multimodalnom kontekstu zahtijeva nove benchmarkove i strategije obrane (Jiang et al., 2025).

Interpretabilnost i objašnjivost (XAI)

Razvoj alata koji ne samo da otkrivaju grešku, već i objašnjavaju *zašto* je do nje došlo, ključan je za izgradnju povjerenja i omogućavanje ciljanih popravaka (Nanda et al., 2023).

Zaključak

Ovaj rad pružio je sveobuhvatan pregled izazova logičkih grešaka u velikim jezičnim modelima, nudeći tri ključna doprinosa: novu tipologiju grešaka, tromodularni sustav autokorekcije i prijedlog evaluacijskog okvira LOGIC-RobBench s pilot studijom. Transformacija LLM-ova iz impresivnih alata za generiranje teksta u istinski pouzdane partnere zahtijeva sinergiju između tehničkih inovacija, rigorozne evaluacije i dubokog razumijevanja kognitivnih i etičkih dimenzija njihovog djelovanja. Iako provedena pilot studija pruža početne uvide, svjesni smo njenih ograničenja u pogledu opsega testiranih modela i zadataka. Unatoč tome, nadamo se da ovaj okvir može poslužiti kao referentna točka i poticaj za buduća, opsežnija empirijska istraživanja na tom putu.

Literatura

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., ... & Kaplan, J. (2021). *A general language assistant as a laboratory for alignment*. arXiv preprint arXiv:2112.00861.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., ... & Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback*. arXiv preprint arXiv:2212.08073.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://aclanthology.org/N19-1423

Dhuliawala, S., Komeili, M., Xu, J., Raichur, R., Dusek, J., ... & Weston, J. (2023). *Chain-of-verification reduces hallucination in large language models*. arXiv preprint arXiv:2309.11495.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... & Olah, C. (2021). *A mathematical framework for transformer circuits*. Anthropic. https://transformer-circuits.pub/2021/framework/index.html

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. arXiv preprint arXiv:1412.6572.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations*. https://openreview.net/forum?id=l0V53bErniB

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., & Han, J. (2024). Large language models cannot self-correct reasoning yet. In *Proceedings of the International Conference on Learning Representations*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2905–2915). https://doi.org/10.18653/v1/D17-1315

Jiang, C., Liu, H., Cui, X., Lu, X., Chen, C., Wei, X., ... & Zhang, Y. (2025). *Survey of adversarial robustness in multimodal large language models*. arXiv preprint arXiv:2503.13962.

Jin, D., Jin, Z., Zhou, Z. T., & Szolovits, P. (2020). Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(5), 8018-8025. https://doi.org/10.1609/aaai.v34i05.6311

Jin, Z., Chen, X., Huang, J., Shi, S., Yu, C., Ouyang, W., & Yan, M. (2023). FaLLaC: A dataset for logical fallacy classification in political debates. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 8234–8248).

Kandpal, N., Deng, H., Roberts, A., & Raffel, C. (2022). *Large language models are implicitly continual learners*. arXiv preprint arXiv:2212.10474.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Child, R., ... & Amodei, D. (2020). *Scaling laws for neural language models*. arXiv preprint arXiv:2001.08361.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Hadsell, R., Rusu, A. A., ... & Hassabis, D. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences, 114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., ... & Yih, W. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

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

Nanda, N., Chan, L., Liberum, T., Smith, J., & Steinhardt, J. (2023). *Progress in mechanistic interpretability*. arXiv preprint arXiv:2302.04944.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

Perez, F., & Ribeiro, M. T. (2022). The challenges of detecting and mitigating harms in large language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 827–839). https://doi.org/10.1145/3531146.3533212

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics, 8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Mahabub, E., ... & Wang, T. (2022). *BLOOM: A 176B-parameter open-access multilingual language model*. arXiv preprint arXiv:2211.05100.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., ... & Wei, J. (2023). Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 13003–13051). https://doi.org/10.18653/v1/2023.findings-acl.824

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., ... & Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models*. arXiv preprint arXiv:2307.09288.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Wang, Z., Zhang, Z., Lee, C. Y., Zhang, H., Sun, R., Ren, X., ... & Wang, W. Y. (2024). *Continual learning for large language models: A survey*. arXiv preprint arXiv:2402.01364.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., ... & Gabriel, I. (2021). *Ethical and social risks of harm from language models*. arXiv preprint arXiv:2112.04359.

Ye, J., Wu, Y., Gao, S., Huang, C., Li, S., Li, G., ... & Huang, X. (2024). RoTBench: A multi-level benchmark for evaluating the robustness of large language models in tool learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Zhang, Y., Sun, Y., & Ji, H. (2023). *A survey on robustness of large language models*. arXiv preprint arXiv:2302.10178.

Zhang, Z., Wu, Y., Wang, Y., & Chen, M. (2024). *Robust multimodal large language models against modality conflict*. arXiv preprint arXiv:2507.07151.

Zhao, X., Yang, X., Pang, T., Du, C., Li, L., Wang, Y. X., & Wang, W. Y. (2024). *Weak-to-strong jailbreaking on large language models*. arXiv preprint arXiv:2401.17256.

Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). *Universal and transferable adversarial attacks on aligned language models*. arXiv preprint arXiv:2307.15043v2.

Zahvale

Izrada ovog rada bila bi znatno otežana bez pomoći naprednih alata umjetne inteligencije. Ovim putem želim izraziti zahvalnost jezičnim modelima koji su korišteni kao asistenti u različitim fazama istraživanja i pisanja.

Posebnu zahvalnost dugujem sljedećim modelima:

Gemini Pro za pomoć u inicijalnom brainstormingu, strukturiranju rada i generiranju prvih nacrta teksta.

NotebookLM za neprocjenjivu pomoć pri analizi i sintezi velike količine znanstvene literature i izvora.

Grok za pružanje alternativnih perspektiva i kritičkih osvrta koji su pomogli u produbljivanju analize.

Claude za detaljnu jezičnu i stilsku lekturu te poboljšanje jasnoće i koherentnosti teksta.

Deepseek za tehničku provjeru činjenica i validaciju algoritamskih prikaza.

Perplexity Pro za pomoć pri provjeri i formatiranju bibliografskih referenci prema APA standardu.

Korištenje ovih alata značajno je ubrzalo istraživački proces i podiglo kvalitetu finalnog rada. Sva konačna interpretacija, zaključci i eventualne pogreške isključivo su odgovornost autora.