Razlikovanje LLM-generirane in človeško napisane kode s konservativnim učenjem

Alen Petek
alen.petek2@student.um.si
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Maribor, Slovenija

Nea Nikolić
nea.nikolic@student.um.si
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Maribor, Slovenija

Andraž Podpečan
andraz.podpecan1@student.um.si
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Maribor, Slovenija

Matevž Sladič
matevz.sladic@student.um.si
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Maribor, Slovenija

Rok Žerdoner
rok.zerdoner@student.um.si
Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Maribor, Slovenija

POVZETEK

V svetu, ki postaja vedno bolj odvisen od tehnologije, so nastali mnogi napredki, med najbolj znamenite v zadnjih letih seveda spada razvoj umetne inteligence, kot so ChatGpt, ki velja za najbolj znanega, Gemini, Copilot, Claude, DeepSeek in mnogi drugi. Ta napredek je s sabo prinesel mnogo ugodnosti, saj so ljudje začeli umetno inteligenco uporabljati kot orodje, ki jim je pomagalo pri dnevnih opravilih. S časom so se ta orodja izboljšala, postala pametnejša in natančnejša. Kmalu niso bila več samo orodja, ampak so lahko preprosta dela in besedila ustvarjali sami. Turingov test je preizkus, ali se lahko robot ekvivalentno oziroma nerazločljivo obnaša kot človek. Razvoj umetne inteligence je doprinesel podobno vprašanje, ali lahko ločimo besedila, ki jih je ustvaril človek od besedil, ki jih je ustvarila umetna inteligenca. S sledečo nalogo, bomo poskušali z uporabo velikih jezikovnih modelov, značilnih karakteristik generirane programske kode in različnimi metodami razpoznavanje ločiti človeško ustvarjeno programsko kodo od tiste, ki jo je ustvarila umetna inteligenca.

KLJUČNE BESEDE

velik jezikovni model, umetna inteligenca, CodeGPTSensor, ChatGPT, konservativno učenje

1 UVOD

Z razvojem velikih jezikovnih modelov (ang. Large Language Model - LLM) [1] se je kvaliteta strojno generirane vsebine, kot so besedila, dialog in celo programska koda, eksponentno izboljšala. Napredki v velikih jezikovnih modelih in umetni inteligenci (ang. Artificial Intelligence- AI) so v zadnjih letih prinesli veliko pozornosti in vzburili mnogo razprav o prihodnosti teh orodji znotraj industrije in akademije. Kot primer lahko vzamemo najbolj znano umetno inteligenco ChatGPT [2], ki je zasnovan na principu okrepitvenega učenja od človeške interakcije (ang. Reinforcement Learning from Human Feedback - RLHF)[3], kar dokazuje sposobnost umetne inteligence za razumevanja povpraševanj, pogovorov in podajanja pravilnih in natančnih odgovorov, v mnogih različnih panogah, vse od pisateljstva do medicine. Z razvojem umetno ustvarjenih del, se je prikazal problem, kako ločiti človeško ustvarjena dela od tistih, ki so ustvarjeni s pomočjo umetne inteligence. S tem vprašanjem so se začele mnoge raziskave v velike jezikovne modele, natančneje v možnosti generiranja programske kode, med katerimi je najbolj zanesljiv Codex [4]. Vendar kljub napredkom v generiranju programkse kode z velikimi jezikovnimi modeli raziskave kažejo, da večina ljudi, še posebej osebje v računalništvu, je skeptičnih o zanesljivosti modelov, določeni izražajo tudi strah. Kot primer lahko vzamemo spletno stran Stack Overflow, ki je prepovedala vse prispevke narejene z uporabo ChatGPT. S tem strahom so se začele raziskave o prepoznavi programske kode generirane z uporabo umetne inteligence, kjer so odkrili, da kljub dobro naučenih modelih, so se pojavili določeni vzorci v kodi generirani s pomočjo velikih jezikovnih modelov [5], kot so različne značilnosti, vključno s slogovnimi nedoslednostmi in napačno uporabo API-jev (aplikacijski programski vmesnik, ang. Aplication Program Interface). Med vsemi detektorji programske kode generirane z uporabo ChatGPT se je CodeGPTSensor[6], metoda, ki temelji na kontrastnem učnem ogrodju s semantičnim kodirnikom kode UniXcoder [7]. UniXcoder je vnaprej pripravljen model, ki vključuje semantične in sintaksne informacije iz izvorne kode za podporo opravil, povezanih s programsko kodo. Za učenje različnih predstavitev kode za razlikovanje ChatGPT generirane kode od tiste, ki jo je napisal človek, uporablja kontrastno učno ogrodje za natančno nastavitev UniXcoderja, ki se je izkazal za enega izmed najbolj natančnih. V tej nalogi bomo preizkusili poustvariti rezultate natančnosti razpoznavanja umetno ustvarjene programske kode in celo izboljšati CodeGPTSensor s svojimi učnimi modeli, z namenom, da dosežemo še boljšo natančnost in konsistentnost prepoznavanja kode ustvarjene z umetno inteligenco od tiste, ki jo je ustvaril človek.

2 SORODNI ČLANKI

Glede na to, da naša ideja temelji na optimizacije enega izmed najboljših metod za razpoznavanje med človeško in programsko kodo umetne inteligence [8], smo potrebovali referenco oziroma primerjavo z ostalimi, pogosteje uporabljenimi metodami, ki temeljijo na že definiranih velikimi jezikovnimi modeli in strategijah povezanih

s takšnim načinom razpoznavanja. V članku »Is This You, LLM? Recognizing AI-written Programs with Multilingual Code Stylometry« [9] smo preučili najbolj osnovne strategije prepoznavanja generirane. Med najpogostejše tehnike spada iskanje vzorcev v generirani kodi, saj se naučeni model drži določenih stilov generiranja kode. Najbolj očiten znak, da je koda generirana, je napačna uporaba kompleksnejših objektov znotraj kode, kot so naprednejši algoritmi ali API-ji. Preučili smo tudi tehniko »Proti ognju se boriš z ognjem.«, saj bomo uporabljali prirejen veliki jezikovni model za razpoznavanje generirane programske kode, ki sama temelji na naučenih velikih jezikovnih modelih. To strategijo smo zaznali v naslednjem sorodnem delu »Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text?« [10], ki zastavi vprašanje, ali je ChatGPT sposoben prepoznati lastno kodo. Takšno tehniko bi lahko uporabili s katerokoli umetno inteligenco, vendar znotraj tega članka so se avtorji osredotočili ekskluzivno na ChatGPT. Raziskave v okviru tega članka so potrdile, da ChatGPT je sposoben prepoznati lastno kodo, kar nakazuje, da je še sposoben ločiti človeško programsko kodo od generirane. Pri preučevanju tega članka se nam je porodila ideja za nadaljevanje tega eksperimenta, pri čemer bi preizkušali ali so različne umetne inteligence sposobne prepoznati tuje generirane programske kode (npr. ali je ChatGPT sposoben prepoznati delo Gemini)[11].

REFERENCES

[1] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.

[2] M. S. I. Sakib, "What is chatgpt," *ResearchGate*, 2023.

[3] H. Lee, S. Phatale, H. Mansoor, K. R. Lu, T. Mesnard, J. Ferret, C. Bishop, E. Hall, V. Carbune, and A. Rastogi, "Rlaif: Scaling reinforcement learning from human feedback with ai feedback," 2023.

[4] P. Zhao, "An empirical study on using codex for automated program repair," 2023.

[5] S. Cave and K. Dihal, "Hopes and fears for intelligent machines in fiction and reality," *Nature machine intelligence*, vol. 1, no. 2, pp. 74–78, 2019.

[6] X. Xu, C. Ni, X. Guo, S. Liu, X. Wang, K. Liu, and X. Yang, "Distinguishing llm-generated from human-written code by contrastive learning," *ACM Transactions on Software Engineering and Methodology*, 2024.

[7] D. Guo, S. Lu, N. Duan, Y. Wang, M. Zhou, and J. Yin, "Unixcoder: Unified cross-modal pre-training for code representation," *arXiv preprint arXiv:2203.03850*, 2022.

[8] M. Oedingen, R. C. Engelhardt, R. Denz, M. Hammer, and W. Konen, "Chatgpt code detection: Techniques for uncovering the source of code," *arXiv preprint arXiv:2405.15512*, 2024.

[9] A. Gurioli, M. Gabbrielli, and S. Zacchiroli, "Is this you, llm? recognizing ai-written programs with multilingual code stylometry," *arXiv preprint arXiv:2412.14611*, 2024.

[10] A. Bhattacharjee and H. Liu, "Fighting fire with fire: can chatgpt detect ai-generated text?" *ACM SIGKDD Explorations Newsletter*, vol. 25, no. 2, pp. 14–21, 2024.

[11] N. Rane, S. Choudhary, and J. Rane, "Gemini versus chatgpt: applications, performance, architecture, capabilities, and implementation," *Journal of Applied Artificial Intelligence*, vol. 5, no. 1, pp. 69–93, 2024.