DOI
https://doi.org/10.18690/um.fov.3.2026.58

ISBN
978-961-299-124-1

EVALVACIJA NAPREDKA GENERATIVNE UMETNE INTELIGENCE PRI SNOVANJU NALOG IZ VIŠJE MATEMATIKE: PRIMERJALNA ŠTUDIJA

UROŠ STERLE
Srednja tehniška šola, Šolski center Kranj, Kranj, Slovenija
uros.sterle@sckr.si

Generativna umetna inteligenca (GenUI) prinaša pomembne spremembe v izobraževalni proces, vključno z avtomatizacijo priprave preverjanj znanja. V prispevku obravnavamo uporabo orodij GenUI pri snovanju nalog za področje višje matematike, kjer je natančnost ključnega pomena. Čeprav lahko GenUI pedagoškim delavcem prihrani čas pri sestavljanju izpitov, tehnologija zahteva kritično presojo zaradi nagnjenosti k »haluciniranju« – generiranju prepričljivih, a napačnih informacij. V raziskavi smo primerjali pet vodilnih orodij umetne inteligence (ChatGPT, DeepSeek, Gemini, Copilot, Grok) na primeru petih kompleksnih matematičnih nalog. Posebna vrednost raziskave je longitudinalna primerjava: analizirali smo odgovore, pridobljene v prvi fazi raziskave aprila 2025, in jih primerjali z rezultati istih orodij januarja 2026. Cilj prispevka je prikazati stopnjo napredka orodij ter utemeljiti potrebo po pristopu »človek v zanki« (Human-in-the-loop) v pedagoški praksi.

Ključne besede:
umetna inteligenca,
višja matematika,
generiranje nalog,
LLM,
izobraževanje,
človek v zanki

Univerzitetna založba
Univerze v Mariboru

DOI
https://doi.org/10.18690/um.fov.3.2026.58

ISBN
978-961-299-124-1

Evaluation of the Progress of Generative Artificial Intelligence in Creating Higher Mathematics Tasks: a Comparative Study

Uroš Sterle
Secondary technical school, Kranj School Centre, Kranj, Slovenia
uros.sterle@sckr.si

Keywords:
artificial intelligence,
higher mathematics,
task generation,
LLM,
education,
human-in-the-loop

Generative artificial intelligence (GenAI) brings significant changes to the educational process, including the automation of knowledge assessment preparation. In this article, we discuss the use of GenAI tools in designing tasks for higher mathematics, where accuracy is of key importance. Although GenAI can save educators time in compiling exams, the technology requires critical judgment due to its tendency to "hallucinate" - generating convincing but incorrect information. In our study, we compared five leading artificial intelligence tools (ChatGPT, DeepSeek, Gemini, Copilot, Grok) using five complex math problems as examples. The longitudinal comparison adds particular value to the study: we analyzed the answers obtained in the first phase of the study in April 2025 and compared them with the results of the same tools in January 2026. The aim of the paper is to show the degree of progress of the tools and to justify the need for a "human-in-the-loop" approach in pedagogical practice.

University of Maribor Press

1 Uvod

Namen prispevka je raziskati potencial brezplačnih različic orodij generativne umetne inteligence (GenUI) pri ustvarjanju in prilagajanju matematičnih gradiv, s posebnim poudarkom na njihovi praktični uporabnosti v izobraževalnem procesu. Izhodišča raziskave temeljijo na treh ključnih dejavnikih: naraščajoči integraciji GenUI v šolski sistem, potrebi po učinkoviti pripravi personaliziranih gradiv ter pomanjkanju sistematičnih primerjav različnih orodij GenUI na področju naravoslovja in matematike.

V prispevku se osredotočamo na avtomatizirano generiranje nalog iz višje matematike. Medtem ko so veliki jezikovni modeli (LLM) izjemno uspešni pri generiranju besedil, se pri matematiki pogosto soočajo s težavami. Pojav, ko generativna orodja samozavestno podajajo napačne odgovore, v literaturi imenujemo »halucinacije« (Athaluri et al., 2023; IBM, 2024). Čeprav so nekatere študije (Frieder et al., 2023) že zgodaj opozarjale na pomanjkljivosti jezikovnih modelov pri matematičnem sklepanju, hiter razvoj orodij zahteva ponovno evalvacijo stanja.

V kontekstu pričujoče raziskave pojem »napredek« definiramo večdimenzionalno. Ne ocenjujemo le binarne pravilnosti končnega rezultata, temveč napredek razumemo kot (1) izboljšanje računske natančnosti, (2) zmožnost upoštevanja kompleksnih omejitev pri snovanju nalog (npr. zahteva po celoštevilski rešitvi) in (3) zmanjšanje konceptualnih halucinacij, kjer model napačno prepozna strukturo problema.

2 Metodologija

Raziskava je zasnovana kot primerjalna študija v dveh časovnih točkah. Prvi del raziskave je bil izveden **25. aprila 2025**, drugi del pa **14. januarja 2026**. V obeh fazah smo uporabili identičen testni scenarij, ki preverja sposobnost orodij za snovanje in reševanje kompleksnejših, netrivialnih problemov višje matematike.

2.1 Orodja in verzije

Zaradi hitrega razvoja področja je ključno definirati točne verzije uporabljenih modelov. V analizo smo vključili pet vodilnih velikih jezikovnih modelov v njihovih takrat aktualnih brezplačnih različicah.

Tabela 1: Pregled uporabljenih verzij modelov v obeh fazah raziskave

Orodje	Verzija (April 2025)	Verzija (Januar 2026)	Opomba o spremembi
ChatGPT	GPT-4o	GPT-4o (posodobljen)	Stabilna verzija, izboljšana logika.
DeepSeek	DeepSeek-V2.5	DeepSeek-V3	Prehod na nov, zmogljivejši model.
Gemini	Gemini 1.5 Flash	Gemini 2.0 Flash	Hitrejši model z boljšim razumevanjem.
Copilot	GPT-4 Turbo	GPT-4o	Integracija novejšega OpenAI modela.
Grok	Grok-2	Grok-3 (Beta)	Testna verzija nove generacije.

Vir: lasten

2.2 Testne naloge

Vsem orodjem smo podali identičen poziv (prompt): *»Rad bi nekaj netrivialnih nalog za višjo matematiko skupaj z natančnim potekom reševanja: Inverz matrike 3x3 s celoštevilsko rešitvijo, determinanto 5x5 z več ničlami, presek dveh ravnin, volumen paralelepipeda z vektorji in iskanje lokalnih ekstremov funkcije dveh spremenljivk.«*

Analizirali smo:

1. **Ustreznost naloge:** Ali upošteva omejitve (celoštevilskost)? Ali je naloga netrivialna (izločili smo npr. trikotne matrike, kjer je inverz očiten)?
2. **Pravilnost postopka:** Ali je potek reševanja matematično korekten?
3. **Točnost rezultata:** Ali je končni rezultat pravilen?

Vsak poziv je bil za posamezno orodje in nalogo izveden enkrat (*single-shot*), s čimer smo simulirali tipično prvo interakcijo uporabnika z orodjem, brez dodatnega usmerjanja ali popravljanja s strani uporabnika.

2.3 Vloga študij primera

Poleg kvantitativne primerjave uspešnosti smo v raziskavo vključili kvalitativno analizo v obliki študij primera. Te niso naključno izbrane, temveč služijo kot ilustracija specifičnih kategorij napak (npr. konceptualne halucinacije, antropomorfno opravičevanje), ki smo jih zaznali pri analizi. Njihov namen je prikazati globlje, strukturne razlike v delovanju modelov med letoma 2025 in 2026, ki jih zgolj binarna ocena (pravilno/napačno) ne zajame.

3 Rezultati

Primerjalna analiza odgovorov razkriva pomemben kvalitativen preskok v zmogljivostih modelov. Rezultate smo v Tabeli 2 **razčlenili glede na tri metodološke kriterije**: ustreznost naloge, pravilnost postopka in točnost rezultata.

Tabela 2: Primerjava izpolnjevanja metodoloških kriterijev v letih 2025 in 2026

Naloga	Orodje	Stanje 2025 (Povzetek)	Stanje 2026: Ustreznost naloge	Stanje 2026: Pravilnost postopka	Stanje 2026: Točnost rezultata
Inverz matrike	DeepSeek, ChatGPT	Uspešno	DA	DA	DA
	Gemini	Delno (trivialna matr.)	DA	DA	DA
	Copilot, Grok	**Neuspešno**	**NE**[1]	DA	DA
Determinanta	Copilot	**Neuspešno** (halucinacija)	DA	DA	DA
	Ostali	Uspešno	DA	DA	DA
Presek ravnin	Vsi modeli	Uspešno (dolgi postopki)	DA	DA	DA
Volumen	Copilot	**Neuspešno** (rezultat 0)	DA	DA	DA
	Ostali	Uspešno	DA	DA	DA
Ekstremi	Vsi modeli	Uspešno	DA	DA	DA

Vir: lasten

[1] Orodji v letu 2026 nista upoštevali omejitve celoštevilske rešitve, sta pa pravilno izpeljali postopek in rezultat za generirano nalogo.

Legenda kriterijev (2026):

- **Ustreznost:** Ali naloga upošteva podane omejitve (npr. celoštevilskost)?
- **Pravilnost:** Ali je matematični postopek metodološko korekten (brez halucinacij)?
- **Točnost:** Ali je končni numerični rezultat pravilen?

3.1 Študija primera: Antropomorfno vedenje (Grok)

Pri nalogi inverza matrike smo preverjali sposobnost »reverznega inženiringa« – ali model zna sestaviti nalogo tako, da bo rešitev »lepa« (celoštevilska). To zahteva, da model vnaprej določi determinanto ± 1.

Medtem ko so DeepSeek, Gemini in ChatGPT to v letu 2026 opravili brezhibno, modelu Grok ni uspelo. Zanimivo pa je njegovo vedenje: namesto priznanja napake je model generiral opombo, v kateri je racionaliziral svoj neuspeh:

Izpis modela Grok (Januar 2026): »Opomba: V tej nalogi namerno nismo dobili čisto celoštevilskega inverza (kar je redko).«

Takšno vedenje kaže na tendenco modelov k »sycophancy« (prilizovanju) – ustvarjanju iluzije kompetentnosti, tudi ko ne izpolnijo pogojev uporabnika.

3.2 Študija primera: Konvergenca učnih podatkov

Pri nalogi iskanja **lokalnih ekstremov funkcije dveh spremenljivk** smo v letu 2026 opazili zanimiv fenomen. Trije različni modeli – ChatGPT, Gemini in DeepSeek – so generirali popolnoma identično funkcijo:

$$f(x,y) = x^3 + y^3 - 3xy \quad \text{.....(1)}$$

To nakazuje, da modeli pri generiranju nalog niso »kreativni« v človeškem smislu, temveč z veliko verjetnostjo prikličejo najpogostejši učbeniški primer iz svojih učnih podatkov. Za pedagoga to pomeni tveganje: če vsi študenti uporabljajo GenUI za vajo, bodo verjetno reševali iste naloge.

3.3 Odprava konceptualnih halucinacij

Leta 2025 je orodje Copilot pri nalogi determinante 5x5 naredilo hudo napako, saj je splošno matriko (z elementi izven diagonale) napačno prepoznalo kot diagonalno in preprosto zmnožilo elemente na diagonali ($1 \cdot 3 \cdot 4 \cdot 5 \cdot 7 = 420$). Leta 2026 je isto orodje nalogo rešilo pravilno. Prepoznalo je blokovno strukturo matrike in pravilno izračunalo determinanto kot produkt determinant blokov (B_1 in B_2), kar je dalo pravilen rezultat -108.

4 Diskusija

Rezultati naše longitudinalne študije kažejo, da generativna umetna inteligenca na področju matematike v enem letu ni le linearno napredovala, temveč je doživela kvalitativno transformacijo. Kot razkriva razčlenitev v Tabeli 2, napredek orodij GenUI ni enakomeren na vseh ravneh. Medtem ko sta **pravilnost postopka** in **točnost rezultata** v letu 2026 dosegla skoraj 100-odstotno zanesljivost pri vseh modelih, se razlike še vedno pojavljajo pri **ustreznosti naloge** – torej pri upoštevanju specifičnih omejitev uporabnika (npr. celoštevilske rešitve pri Groku in Copilotu).

Od »računala« do »asistenta«
Orodja, kot sta npr. DeepSeek in Gemini, so postala zanesljivi »računski stroji«. Sposobnost pravilnega izvajanja algoritmičnih postopkov je sedaj na ravni, ki jo pričakujemo od dobrega študenta. To pomeni, da lahko učitelji ta orodja varneje uporabljajo za generiranje rešitev.

Nujnost pristopa »Človek v zanki« (Human-in-the-loop)
Kljub napredku pa orodja ostajajo »taktična« in ne »strateška«. Analiza kaže, da GenUI še ne more popolnoma avtonomno sestavljati uravnoteženih izpitnih pol. Zato je nujna implementacija modela **Human-in-the-loop (HITL)** (Zhai et al., 2021). V tem modelu učitelj ne prepusti celotnega procesa stroju, temveč deluje kot nadzornik, ki definira pogoje, preveri ustreznost naloge in popravi morebitne pristranskosti ali ponavljanja. Učitelj postane »urednik« in garant kakovosti.

5 Zaključek

V prispevku smo analizirali napredek orodij umetne inteligence pri generiranju nalog iz višje matematike v obdobju od aprila 2025 do januarja 2026. Ugotavljamo, da je tehnologija dosegla stopnjo zrelosti, ki omogoča njeno vključitev v pedagoški proces, vendar z določenimi omejitvami.

Ključne ugotovitve so:

1. **Računska zanesljivost:** Napake pri osnovnih operacijah so skoraj v celoti odpravljene.
2. **Pedagoška vrednost:** Modela DeepSeek in Gemini izstopata po jasnosti razlage postopkov in upoštevanju omejitev (celoštevilskost).
3. **Vloga učitelja:** GenUI je zmogljiv pomočnik, vendar mora končna selekcija nalog ostati v domeni človeške strokovne presoje.

Opomba

Gradivo, na podlagi katerega je bila izvedena raziskava, je obsežno (več kot 50 strani). Bralec si lahko celoten nabor generiranih nalog in rešitev ogleda na naslovu: https://bit.ly/ui_2526.

Viri in literatura

Athaluri, S., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, T. S. (2023). Exploring the boundaries of artificial intelligence in medicine: Is it hallucinating? *Cureus*, *15*(4), e37432. https://doi.org/10.7759/cureus.37432

Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). Mathematical capabilities of ChatGPT. *arXiv preprint arXiv:2301.13867*. https://arxiv.org/abs/2301.13867

IBM. (2024). *What are AI hallucinations?* IBM Topics. https://www.ibm.com/topics/ai-hallucinations

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., & Spector, M. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity*, *2021*, Article 8812542. https://doi.org/10.1155/2021/8812542

O avtorju

Uroš Sterle je profesor matematike in diplomant FMF v Ljubljani z bogatimi izkušnjami v informatiki. Poklicno pot je začel v gospodarstvu kot programer, kjer je matematična znanja uporabljal za razvoj rešitev in optimizacijo procesov. Od leta 2010 poučuje na Šolskem centru Kranj, kjer teoretične koncepte povezuje s prakso v podatkovnih bazah, spletnih aplikacijah in omrežnih storitvah.

Je avtor učnih gradiv za razvoj algoritmičnega mišljenja ter aktivno sodeluje pri projektih za krepitev digitalnih kompetenc in v visokošolskem izobraževanju. S stalnim strokovnim izpopolnjevanjem in več kot 20-letnimi izkušnjami z individualnim poučevanjem uspešno združuje poglobljeno matematično znanje z njegovo aplikativno vrednostjo v sodobni tehnologiji.