Obširna evalvacija komercialnih velikih jezikovnih modelov na področju sklepanja v slovenskem jeziku in slovnice

Miha Malenšek, Domen Vreš, Marko Bajec
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana
miha.malensek@fri.uni-lj.si, domen.vres@fri.uni-lj.si, marko.bajec@fri.uni-lj.si

Izvleček

Uporaba velikih jezikovnih modelov (VJM) se hitro širi tudi v slovenskem prostoru, vendar je njihova dejanska zmogljivost za slovenski jezik še vedno slabo sistematično ovrednotena. V tem članku predstavljamo obširno primerjalno evalvacijo najpogosteje uporabljanih komercialnih in odprtih VJM v kontekstu slovenščine. V evalvacijo smo vključili modele štirih večjih ponudnikov (OpenAI, Google, Anthropic in Mistral) ter domači model GaMS-27B-Instruct in jih ovrednotili z raznolikim naborom učnih množic, ki preverjajo sposobnosti sledenja navodilom, razumsko sklepanje, zanesljivost odgovorov, slovnične kompetence ter koherentnost besedila. Uporabili smo prevedene standardizirane primerjalne naloge (npr. ARC, HellaSwag, TruthfulQA, GSM8K), specializirano množico za slovnične napake DASSLE 1.0 ter nabor resničnih pogovorov Slovenske pogovorne arene. Rezultati kažejo, da sodobni komercialni modeli dosegajo visoko uspešnost pri nalogah razumevanja in sklepanja v slovenskem jeziku, zlasti GPT-5.1 z visokim nivojem premišljanja in Gemini-2.5-Pro, medtem ko odprti modeli, kot je Mistral Large 3 kljub omejenim virom dosegajo konkurenčne rezultate. Nasprotno pa evalvacija slovnične kompetence razkriva, da ostaja morfološka in skladenjska kompleksnost slovenskega jezika velik izziv za vse obravnavane modele. Članek s tem nudi celovit vpogled v trenutno stanje zmogljivosti VJM za slovenščino.

Ključne besede: veliki jezikovni modeli (VJM), evalvacija, analiza slovničnih napak, obdelava naravnega jezika

Comprehensive Evaluation of Commercial Large Language Models for Reasoning in Slovenian Language and Grammar

Abstract

The use of large language models (LLMs) is rapidly expanding in the Slovenian context; however, their actual performance on the Slovenian language remains insufficiently and unsystematically evaluated. In this paper, we present an extensive comparative evaluation of the most widely used commercial and open-source LLMs with respect to Slovenian. The evaluation includes models from four major providers (OpenAI, Google, Anthropic, and Mistral), as well as the domestic models GaMS-27B-Instruct and GaMS3-12B-Instruct, and assesses them using a diverse set of benchmarks targeting instruction following, reasoning abilities, answer reliability, grammatical competence, and textual coherence. We employ translated standardized benchmark tasks (e.g., ARC, HellaSwag, TruthfulQA, GSM8K), the specialized grammatical error dataset DASSLE 1.0, and a collection of real-world conversations from the Slovenian Conversational Arena. The results show that contemporary commercial models achieve high performance on comprehension and reasoning tasks in Slovenian—most notably GPT-5.1 with a high level of deliberative reasoning and Gemini-2.5-Pro—while open models such as Mistral Large 3 attain competitive results despite more limited resources. In contrast, the evaluation of grammatical competence reveals that the morphological and syntactic complexity of Slovenian remains a significant challenge for all evaluated models. Overall, the paper provides a comprehensive overview of the current state of LLM performance for the Slovenian language.

Keywords: Large Language Models (LLM), Evaluation, Grammatical Error Analysis, Natural Language Processing

1 UVOD

Uporaba velikih jezikovnih modelov (VJM) postaja vse bolj pogosta na vseh področjih življenja, nastop generativnih jezikovnih modelov v obliki splošno sposobnih, pogovornih storitev, kot so ChatGPT, Gemini, Claude in drugi, pa so uporabo še povišali. Iz letnih raziskav podjetja McKinsey in univerze Stanford je razvidno, da je globalna uporaba generativnih orodij narasla s 55 % v letu 2023, na 88 % v letu 2025 [1] [2]. V Sloveniji anketa Centra za družboslovno informatiko razkriva podoben trend – delež redne uporabe generativnih orodij se je skoraj podvojil z 25 % v letu 2024 na 48 % uporabnikov v letu 2025 [3]. Primerjave nakazujejo, da Slovenija za najrazvitejšimi državami zaostaja za približno eno leto, kar pomeni, da se bo delež uporabe verjetno še povečal.

Med generativnimi orodji najbolj prevladujejo veliki jezikovni modeli, od katerih v Sloveniji prevladuje ChatGPT, ki ga uporablja 83 % uporabnikov generativnih orodij. Hiter porast uporabe VJM spremlja tudi povečanje raznovrstnosti modelov, dostopnih s komercialnimi, plačljivimi programskimi vmesniki (API). V zadnjih letih so vodilni ponudniki OpenAI, Google, Anthropic in Mistral redno posodabljali ponujene VJM ter uporabnikom omogočili dostope do zmogljivejših in zanesljivejših storitev.

Kljub temu pa se v praksi izkaže, da se zmogljivosti posameznih modelov močno razlikujejo, še posebej pri zahtevnejših nalogah v slabše zastopanih jezikih, kot je slovenščina. Ob hitri adopciji v vsakodnevni rabi se tako pojavi ključno vprašanje o generalni sposobnosti najpogosteje uporabljanih VJM v kontekstu slovenskega jezika.

V tem članku zato predstavljamo obširno evalvacijo trenutno najpogosteje uporabljanih VJM. Vključujemo modele iz vseh štirih večjih ekosistemov (OpenAI, Google Gemini, Anthropic Claude in Mistral) in jih ovrednotimo na prevedenih, standardiziranih primerjalnih testih, ki se pogosto uporabljajo za ovrednotenje sposobnosti VJM v angleškem jeziku, kakor tudi na pogovorni kakovosti in skladnosti odgovorov ter slovničnih sposobnostih.

2 IZBRANI PONUDNIKI

Za evalvacijo smo izbrali storitve podjetij Google, OpenAI, Anthropic in Mistral ter evalvirali širok nabor ponujenih modelov različnih cenovnih nivojev.

Tabela 1: **Nabor evalviranih modelov. Cena podan v dveh vrednostih, na milijon vhodnih in izhodnih tokenov. Pri tem je potrebno opozoriti, da razmišljujoči (angl. reasoning) modeli, štejejo tokene razmišljanja v izhodne tokene.**

Podjetje	Model	Cena (1M vhodnih / 1M izhodnih tokenov)
Google	gemini-2.5-pro	$2.50 / $15.00
	gemini-2.5-flash	$0.30 / $2.50
OpenAI	gpt-5.1	$1.25 / $10.00
	gpt-5	$1.25 / $10.00
	gpt-5-mini	$1.25 / $2.00
	gpt-5-nano	$0.05 / $0.40
	gpt-4.1	$2.00 / $8.00
	gpt-4o-mini	$0.15 / $0.60
Anthropic	Claude Opus 4.1	$15.00 / $75.00
	Claude Sonnet 4.5	$3.00 / $15.00
	Claude Haiku 4.5	$1.00 / $5.00
Mistral	Mistral Large 3	$0.50 /$1.50
	Mistral Medium 3.1	$0.40 / $2.00
	Mistral Small 3.2	$0.10 / $0.30
FRI	GaMS-27B-Instruct	Lokalna postavitev

Modele in cene smo navedli v Tabeli 1. Cene so podane v standardnem formatu, ki ga uporabljajo vsi štirje ponudniki, ki ceno uporabe sestavi iz količine vhodnih in izhodnih tokenov (t.j. besednih enot, ki jih model generira). Podana je kot cena na milijon vhodnih in izhodnih tokenov, izpisana iz dokumentacije API vmesnikov v času evalviranja modelov.

Vsi štirje ponudniki imajo podoben API vmesnik, zato med posamezno implementacijo ni dosti razlik. Pri uporabi je potrebno paziti na omejitve količine klicev na minuto ter na samo porabo zakupljenega dobroimetja − vsi štirje ponudniki namreč dostopa do API funkcionalnosti ne omejijo ali ustavijo po porabi celotnega zakupljenega dobroimetja, omogočajo le možnosti obveščanja, ko poraba preseže definirano vrednost.

3 UČNE MNOŽICE

Za podrobno evalvacijo sposobnosti velikih jezikovnih modelov (VLM) v slovenskem jeziku smo izbrali tri učne množice, ki z raznolikimi nalogami v slovenskem jeziku preverjajo zmožnosti sledenja navodilom, sposobnosti razumskega sklepanja, ekstrakcije in uporabe informacij iz danega konteksta, nagnjenosti k zavajanju oz. podajanju neresničnih ali napačnih informacij, slovnično in pravopisno kompetenco ter pogovorne sposobnosti.

Za te namene smo uporabili prostodostopni množici Slovenian LLM Evaluation[1] objavljeno na repozitoriju HuggingFace in množico DASSLE 1.0[2], objavljeno na repozitoriju Clarin, ter označene sledi pogovorov Slovenske pogovorne arene[3].

Množica Slovenian LLM Evaluation je skupek specifičnih nalog, s katerimi evalviramo generalne sposobnosti VJM in sposobnosti sledenja kompleksnih navodil in uporabo konteksta ter izluščenih informacij, podanih v slovenskem jeziku. Z množico DASSLE 1.0 evalviramo slovnične in oblikoslovne sposobnosti VJM. Slovenščina je morfološko kompleksen jezik, zato je evalvacija na tej množici še posebej pomembna. Množica pogovorov slovenske pogovorne arene evalvira tok pogovora z VJM ter reakcije le-teh na včasih nesmiselne vhode.

3.1 DASSLE 1.0

Množica DASSLE 1.0 (Dataset of Authentic and Synthetic Slovene Language Errors) je sestavljena iz 7385 ročno pripravljenih primerov, ki vsebujejo poved z edinstveno slovnično ali pravopisno napako, pravilno popravljeno poved ter makro- in mikro-kategorijo napake. Z množico DASSLE preverjamo dve pomembni sposobnosti VLM:

- ali je model sposoben **pravilno identificirati makro-kategorijo napake** in
- ali je model sposoben napako pravilno odpraviti ter predlagati pravilen popravek.

Evalvacija na tej množici nam da pomemben vpogled v **jezikovno, slovnično in pravopisno kompetenco** VLM, kar je še posebej pomembno v kontekstu slovenskega jezika.

3.2 Pogovori Slovenske pogovorne arene

Množica vsebuje primere ročno ustvarjenih, uporabniških pogovorov z VLM na Slovenski pogovorni areni. Z uporabo uporabniških pozivov simuliramo interakcijo med VLM in človeškim uporabnikom, nato pa z uporabo panele sodnikov (LLM-as-a-judge) presodimo pogovorno smiselnost in koherentnost celotne sledi pogovora.

Evalvacija na tej množici nam da vpogled v pogovorne sposobnosti VLM na raznolikih (in včasih tudi nesmiselnih) uporabniških pozivov.

3.3 Slovenian LLM Evaluation

Množica slovenian-llm-eval je prostodostopna množica na repozitoriju HuggingFace, ki vključuje skupek nalog, ki se pogosto uporabljajo za evalvacijo VLM. Naloge so strojno prevedene v slovenski jezik in ročno pregledane ter zaobjemajo predvsem sposobnosti sledenju daljših navodil, razumevanju konceptov ter ekstrakciji in uporabi relevantne informacije v novem kontekstu.

3.3.1 ARC (AI2 Reasoning Challenge)

ARC-Challenge in *ARC-Easy* [4] sta množici vprašanj, zasnovani za preverjanje, ali lahko veliki jezikovni modeli pravilno odgovarjajo na naravoslovna vprašanja iz osnovnošolskih preverjanj znanj v ZDA. Naloge ne temeljijo zgolj na preprostem iskanju ključnih besed ali pomnjenju dejstev, temveč zahtevajo razumevanje, sklepanje in uporabo znanstvenega znanja. Naloge so ločene na lažja vprašanja, na katera je mogoče pogosto odgovoriti z osnovnim znanjem, ter na modelom težja vprašanja, saj zahtevajo povezovanje dejstev, razume-

[1] https://huggingface.co/datasets/cjvt/slovenian-llm-eval
[2] https://vlo.clarin.eu/record/https_58__47__47_hdl.handle.net_47_11356_47_2052_64_format_61_cmdi
[3] https://arena.cjvt.si/sl

vanje znanstvenih konceptov, predvsem pa prepoznavanje vzrokov in posledic ter logičnih razmerij. Naloga tako preverja, ali je model sposoben preseči preprosto pomnjenje in priklic informacij.

Posamezna naloga je sestavljena iz vprašanja in seznama možnih odgovorov med katerimi model izbira.

3.3.2 HellaSwag

HellaSwag [5] je zbirka nalog, zasnovana za preverjanje logičnega sklepanja velikih jezikovnih modelov. Posamezna naloga vključuje kratke, nedokončane opise situacij, ki jim sledi več možnih nadaljevanj. Model mora izbrati najbolj smiselno nadaljevanje danega opisa. Vse podane možnosti so slovnično pravilne in slogovno naravne, pravilna možnost pa je tista, ki se ujema z realističnim potekom dogodkov. Naloga zato ne preverja le površinskega razumevanja jezika, temveč predvsem razumevanje širšega konteksta in implikacij le-tega.

3.3.3 TruthfulQA

TruthfulQA [6] je zbirka nalog, namenjena preverjanju zanesljivosti in resnicoljubnosti odgovorov velikih jezikovnih modelov in njihovi odpornosti na zavajajoče odgovore, še posebej v primerih, ko se vprašanja opirajo na napačne predstave, teorije zarote, zavajajoče trditve ali pogosto ponavljane netočnosti z interneta. Cilj ni le preverjanje prepoznavanja resničnih dejstev, temveč tudi sposobnosti modela, da prepozna in zavrne napačne premise. Zbirka vključuje tako vprašanja z enim pravilnim odgovorom *(truthfulqa_mc1)* in vprašanja z več pravilnimi odgovori *(truthfulqa_mc2)*. Pri slednjih v tabeli rezultatov predstavimo tako strogo natančnost *(model pravilno izbere vse možne odgovore)* in povprečno natančnost *(koliko pravilnih vprašanj je model izbral)*.

3.3.4 BoolQ

BoolQ [7] je zbirka nalog, ki preverja sposobnost velikih jezikovnih modelov, da berejo, razumejo in presojajo informacije v danem besedilu. Vsaka naloga vsebuje vprašanje in kratek odlomek iz Wikipedije, na podlagi katerega mora model na vprašanje odgovoriti z *da* ali *ne*. Gre za naravna vprašanja, zbrana iz spletnih iskanj, kar pomeni, da so pogosto nepopolna, pogovorna ali implicitno zastavljena, zato pravilni odgovor zahteva razumevanje konteksta, prepoznavo parafraz, logičnih povezav in prikritih negacij. S tem preveri sposobnost sklepanja na podlagi danega konteksta in razločevanje med eksplicitno in implicitno podanim znanjem.

3.3.5 OpenBookQA

OpenBookQA [8] je zbirka nalog, zasnovana za preverjanje sposobnosti velikih jezikovnih modelov, da kombinirajo osnovno znanstveno znanje z vsakdanjim sklepanjem. Naloga temelji na konceptu izpita »odprte knjige«, v kateri ima model poleg vsakega vprašanja na voljo še zbirko približno 1300 osnovnih dejstev, ki predstavljajo temeljno znanje s področij, kot so fizika, biologija in kemija. Vprašanja so oblikovana tako, da zgolj prepoznavanje teh dejstev ni dovolj, model mora znanje pravilno uporabiti v novi situaciji. Naloge vključujejo več odgovorov, od katerih je zgolj en pravilen, ostali pa so pogosto zavajajoči. S tem preverimo sposobnost povezovanja informacij, razumevanje naravnih pojavov in logično sklepanje.

3.3.6 WinoGrande

WinoGrande [9] je zbirka nalog, namenjena preverjanju sposobnosti sklepanja in razumevanja konteksta jezikovnih modelov, pri čemer se osredotoča na izziv referenčne razjasnitve *(angl. coreference resolution)*. Vsak primer vsebuje stavek z manjkajočim ali dvoumnim delom, ki ga je treba zapolniti z eno izmed dveh ponujenih možnosti. Model mora na podlagi zdrave pameti, razumevanja vzorčnih odnosov, fizičnega in socialnega znanja ter logičnega sklepanja izbrati tisto možnost, ki najbolj smiselno dopolni poved. Obe možnosti sta slovnično in slogovno pravilni, pravilna rešitev pa je tista, ki smiselno odraža pravilen potek dogodkov. Primeri so za modele zahtevni in predstavljajo test globokega razumevanja tako jezika, kot širšega konteksta.

3.3.7 PIQA (Physical Interaction Question Answering)

PIQA [10] je zbirka podatkov, zasnovana za preverjanje praktičnega sklepanja velikih jezikovnih modelov. Vsak primer vsebuje opis enostavnega cilja ali opravila iz vsakdanjega življenja ter dve možni rešitvi oziroma opisa poti do tega cilja. Model mora izbrati tisto rešitev, ki je fizično izvedljiva in smiselna v resničnem svetu. Naloga se osredotoča na razumevanje lastnosti predmetov, naravnih zakonov, vzročnih razmerij in običajne rabe orodij in materialov. Naloga s tem preverja, ali modeli premorejo

praktično, fizično znanje vsakdanjega sveta in ali ga znajo uporabiti v dani situaciji.

3.3.8 GSM8K (Grade School Math 8K)

GSM8K [11] je zbirka besedilnih nalog s področja matematike. Naloge preverjajo aritmetične sposobnosti in logično sklepanje z več koraki. Naloge zahtevajo pravilno interpretacijo podatkov ter sistematično izvedbo računskih korakov, ki lahko vključujejo seštevanje, odštevanje, deljenje in množenje ter delo z enotami in preprostimi enačbami. Ključni izziv je razčleniti besedilo naloge, izluščiti relevantne informacije in jih uporabiti za pravilno rešitev, ki pogosto zahteva več korakov. Naloge ocenjujejo zanesljivosti matematičnega sklepanja in doslednost modelov z ozirom na pomnjenje relevantnih informacij in jasno razkrijejo, ali model razume problem ter zna metodično sklepati in računati.

Evalvacija VLM na tem naboru nalog nam omogoči osnovni vpogled v sposobnosti VLM v sledenju navodil in uporabi informacij podanih v slovenskem jeziku na specifičnih nalogah. Izmed izbranih nalog gre za najlažje, saj ne preverja kompetenc v pisanju slovenskega jezika, temveč le v sposobnosti razumevanja in povezovanja informacij podanih v slovenskem jeziku.

4 REZULTATI

Za pridobivanje rezultatov evalvacije smo za vsako učno množico pripravili ogrodje, s katerim lahko definiramo nabor modelov in podnalog, na katerih se opravlja evalvacija. Ker se uporaba komercialnih vmesnikov med ponudniki nekoliko razlikuje, so trenutno podprti samo modeli izbranih ponudnikov, predstavljenih v Tabeli 1, ter modeli kompatibilni s knjižnico vLLM[4]. Razen razlike v samem postopku klica storitve se struktura uporabljenih pozivov, struktura pričakovanega odgovora, ter ocenjevalna skripta med modeli ne razlikuje. Rezultati vsake evalvacije se za potrebe pregledov in ocenjevanja shranijo v modelu in v nalogi edinstveno datoteko.

Za vsako nalogo smo definirali predlogo poziva, ki definira vhod in pričakovani izhod. Evalvacija je temeljila na »one-shot[5]« pristopu, kjer se modelu poda problem, ki ga mora rešiti, in en vzorčni primer rešitve. Slednje zagotovi, da se modeli držijo pričakovane strukture odgovora, kar olajša ocenjevanje odgovora. S tem smo preverili iztočno sposobnost modelov – ostalih pristopov, ki bi izboljšali delovanje generativnih modelov, kot na primer »fine-tuning[6]«, nismo uporabljali, saj nas je zanimala predvsem splošna sposobnost modelov v slovenskem jeziku.

Nekateri moderni modeli za odgovore uporabljajo pristop »premišljanja[7]«, ki modelu omogoča veriženja misli na podlagi uporabnikove zahteve. Pristop zanesljivo izboljša kvaliteto odgovorov modelov [12], zato ga komercialni ponudniki privzeto implementirajo v lastne modele. V primeru modela GPT-5.1 podjetja OpenAI smo lahko evalvirali tako visok nivo premišljanja in nizek nivo (brez) premišljanja, saj njihov API omogoča določitev nivoja le-tega. Pri tem je potrebno poudariti, da se s tem tudi bistveno poviša nivo izhodnih enot, kar pa vpliva na ceno.

Rezultati evalvacij na posameznih izbranih učnih množicah so predstavljeni v Tabelah 2 in 3. Za preglednost tabel ne vključujemo cene, ki je že prikazana v Tabeli 1.

V rezultatih predvsem izstopajo modeli GPT-5.1 z visokim nivojem premišljevanja, Gemini 2.5 Pro in odprti model Mistral Large 3. Vsi trije so se dobro izkazali v reševanju nalog Slovenian LLM Evaluation (Tabela 2), kjer so pokazali dobre sposobnosti ekstrakcije in uporabe informacij iz navodil, podanih v slovenskem jeziku. Pri tem je potrebno poudariti, da gre za zelo specifične naloge, pogosto z vnaprej definiranim naborom možnih rešitev, ki pa vendarle zahtevajo razumevanje, ekstrakcijo in uporabo podanih informacij.

Za referenco podajamo tudi rezultate modela GaMS-27B-Instruct, VJM pripravljenega na Fakulteti za računalništvo in informatiko, predučenega na slovenskem jeziku. V primerjavi s komercialnimi modeli je razlika velika, saj model omejuje dostopnost in kvaliteta virov, število parametrov, čas predučenja ter prilagajanje slovenskemu jeziku, kvaliteta inštrukcijske množice in nenazadnje tudi dostop do računskih virov ter finančno breme obsega predučenja.

Medtem ko so rezultati na množici nalog Slovenian LLM Evaluation (Tabela 2) zelo obetavni, rezultati evalvacije na množici DASSLE 1.0 (Tabela 3)

[4] https://docs.vllm.ai/en/latest/
[5] angl., pristop enojnega strela
[6] angl., fino učenje, učenje modela na specifični nalogi
[7] angl., reasoning

kažejo drugačno sliko. Izkaže se, da so slovnične in oblikoslovne lastnosti slovenskega jezika še vedno velik zalogaj za VJM. Tudi najuspešnejši modeli na nalogah Slovenian LLM Evaluation se težko spopadajo tako z razpoznavo kategorije napake kot z odpravljanjem le-te.

Uporabili smo strogo metriko natančnosti popravka, kjer smo za pravilni odgovor šteli le niz, ki se je popolnoma skladal s popravljeno povedjo podano v množici, saj ima večina primerov enolično rešitev[8]. Rezultati kažejo, da se modeli dobro izkažejo le pri preprostejših napakah črkovanja in zapisa, kjer gre predvsem za popravke začetnic, pisanje skupaj ali narazen in zapis glasu ali glasovnega sklopa v besedi. V težjih kategorijah, kot so besedišče, oblikoslovje in predvsem skladnja pa se rezultati bistveno poslabšajo. To nam da vedeti, da se sicer modeli zavedajo preprostejših napak, a globljega razumevanja nians slovenskega jezika kljub temu primanjkuje. Še posebej zaskrbljujoči so rezultati pri napakah iz besedišča, kar lahko močno vpliva na razumevanje kompleksnejših navodil in posledično na kvaliteto danih odgovorov.

Rezultate pogovorne množice smo prikazali v Tabeli 4, kjer prikazujemo število zmag, porazov, izenačenj ter oceno ELO. Rezultati so pridobljeni s pristopom panele VJM sodnikov, ki glasujejo o primerjalni kvaliteti odgovorov dveh modelov. V panelo so vključeni vsi modeli, vključeni v evalvacijo, zaradi skrbi pristranskosti pa smo jim onemogočili samoocenjevanje. Zmage in izenačenja se agregirajo v direktno oceno, ki predstavlja seštevek števila zmag (1 točka) in izenačenega rezultata (0.5 točke) ter v šahovski ELO, kjer gre za relativno oceno sposobnosti VJM. Za vpogled v složnost ocen modelov sodnikov smo izračunali Kappa statistike in povprečen delež strinjanja, prikazano v Tabeli 5. Statistike složnosti kažejo na zanesljiv nivo strinjanja med VJM sodniki, kar doprinaša k verodostojnosti rezultatov.

Rezultati sami so skladni z doseženimi rezultati na evalvacijah na množicah Slovenian LLM Evaluation in DASSLE 1.0, saj se najboljša modela (Gemini-2.5-pro in GPT-5.1) pojavita na vrhu, z najvišjim številom zmag in najvišjo oceno ELO. Mistral Me-

Rezultati: Slovenian LLM Evaluation Dataset

Tabela 2: **Rezultati evalvacije na množici slovenian-llm-eval. Novi gpt-5.1 modeli podjetja OpenAI omogočajo nastavitev ‚reasoning' parametra. Rezultati vključujejo zmogljivosti v načinu ‚high' in ‚none'. Z zeleno so označeni najboljši rezultati glede nanalogo, z rdečo pa najslabši.**

Task / Model	gpt-5.1 high reasoning	gpt-5.1 no reasoning	gpt-5	gpt-4o-mini	gpt-5-mini-2025-08-07	gpt-4.1-2025-04-14	gpt-5-nano-2025-08-07	gemini-2.5-pro	gemini-2.5-flash	Mistral Large 3	Mistral Medium 3.1	Mistral Small 3.2	Claude Opus 4.1	Claude Sonnet 4.5	Claude Haiku 4.5	GaMS-27B-Instruct
arc_challenge	0.938	0.918	0.944	0.848	0.94	0.93	0.918	0.964	0.946	0.898	0.896	0.836	0.934	0.912	0.894	0.543
arc_easy	0.986	0.97	0.983	0.94	0.982	0.976	0.972	0.986	0.982	0.97	0.936	0.93	0.978	0.982	0.968	0.773
hellaswag	0.892	0.858	0.916	0.794	0.852	0.88	0.726	0.92	0.89	0.83	0.822	0.784	0.938	0.952	0.84	0.723
truthfulqa_mc1	0.832	0.802	0.813	0.684	0.788	0.794	0.718	0.854	0.744	0.67	0.658	0.586	0.898	0.938	0.78	0.458
truthfulqa_mc2	0.416	0.406	0.419	0.328	0.378	0.504	0.25	0.586	0.51	0.35	0.22	0.312	0.536	0.414	0.438	0.276
	0.675	0.676	0.695	0.697	0.661	0.741	0.718	0.804	0.751	0.675	0.714	0.724	0.845	0.885	0.734	0.878
Boolq	0.868	0.88	0.879	0.834	0.882	0.87	0.858	0.874	0.882	0.846	0.852	0.846	0.908	0.906	0.842	0.862
openbookqa	0.936	0.906	0.936	0.824	0.922	0.91	0.894	0.918	0.93	0.84	0.812	0.772	0.86	0.872	0.842	0.322
winogrande	0.842	0.714	0.865	0.616	0.82	0.782	0.682	0.836	0.838	0.628	0.865	0.588	0.772	0.788	0.64	0.696
Piqa	0.938	0.91	0.922	0.854	0.9	0.902	0.64	0.928	0.912	0.82	0.852	0.752	0.91	0.86	0.836	0.749

[8] https://wiki.cjvt.si/books/11-developmental-corpus-solar/page/annotation-guidelines

Rezultati: DASSLE 1.0

Tabela 3: **Tabela 3: Rezultati evalvacije na množici DASSLE 1.0. Rezultati predstavljajo strogo natančnost klasifikacije kategorije in predlaganega popravka. Prikazuje tudi natančnost popravkov znotraj posamezne kategorije. Z zeleno so označeni najboljši rezultati glede na nalogo (vrstica), z rdečo pa najslabši.**

Task / Model	gemini-2.5-pro	gemini-2.5-flash	gpt-5.1 high reasioning	gpt-5.1 no reasioning	gpt-5	gpt-5-mini	gpt-5-nano	gpt-4.1	gpt-4o-mini	Mistral Large 3	Mistral Medium 3.1	Mistral Small 3.2	Claude Opus 4.1	Claude Sonnet 4.5	Claude Haiku 4.5	GaMS-27B-Instruct
strict accuracy	0.546	0,463	0,555	0,489	0,579	0,45	0,31	0,52	0,368	0,392	0,312	0,257	0,516	0,449	0,348	0,382
category accuracy	0.54	0,501	0,607	0,527	0,601	0,478	0,334	0,542	0,394	0,4	0,374	0,23	0,548	0,521	0,426	0,322
Besedišče	0.48	0,39	0,46	0,41	0,54	0,22	0,13	0,44	0,15	0,28	0,2	0,13	0,37	0,23	0,19	0,24
Oblikoslovje	0.51	0,36	0,54	0,45	0,55	0,47	0,26	0,5	0,36	0,39	0,31	0,2	0,51	0,46	0,36	0,37
Skladnja	0.25	0,27	0,22	0,21	0,23	0,19	0,1	0,2	0,11	0,16	0,16	0,07	0,24	0,16	0,11	0,1
Zapis	0.75	0,66	0,77	0,73	0,77	0,69	0,53	0,71	0,58	0,56	0,45	0,45	0,7	0,68	0,48	0,54
Črkovanje	0.74	0,64	0,79	0,65	0,8	0,68	0,53	0,75	0,64	0,57	0,44	0,43	0,76	0,71	0,6	0,66

Rezultati: Sledi pogovorov Slovenske pogovorne arene

Tabela 4: **Ocene sledi pogovorov, pridobljene s panelo VJM sodnikov. Ocena je izračunana iz utežene vsote zmag in izenačenj, medtem ko je ELO relativna ocena sposobnosti. ELO bolje oceni modele, ki nepričakovano premagajo boljše ocenjene modele.**

Model	Zmage	Porazi	Izenačenja	Ocena	ELO
gemini-2.5-pro	202	37	11	207.5	1916.7
gpt-5.1-2025-11-13	178	57	15	185.5	1836.3
mistral-medium-2508	171	60	19	180.5	1713.8
gpt-5-2025-08-08	154	79	17	162.5	1816.6
gemini-2.5-flash	137	97	16	145	1597.7
claude-sonnet-4-5	116	121	13	122.5	1445.7
claude-opus-4-1	115	127	8	119	1410.2
gpt-4.1-2025-04-14	108	133	9	112.5	1423.8
claude-haiku4-5	84	160	6	87	1262.1
gpt-4o-mini-2024-07-18	31	214	5	33.5	1151.2
mistral-small-2506	18	229	3	19.5	925.8

Tabela 5: **Metrike strinjanja med VJM sodniki. Odstotek parovnega strinjanja je direkten izračun, kolikokrat so se modeli za iste sledi pogovorov strinjali. Cohenova kappa ima nabor vrednosti med -1 in 1, kjer 0 pomeni naključna ujemanja, 1 pa popolno ujemanje. Fleissova kappa je razširjena Cohenova kappa za več kot dva ocenjevalca.**

Odstotek parnega strinjanja	77.5%
Povprečna Cohenova kappa	0.578
Fleissova kappa	0.5

dium po številu zmag doseže tretje mesto, a ga ELO postavi na četrto z modelom GPT-5 na tretjem, saj ocena ELO ocenjuje relativno sposobnost modela (t.j. zmago proti predvideno slabšem modelu oceni slabše, kot presenetljivo zmago proti boljšem modelu). Tudi to se sklada z rezultati na Slovenian LLM Evalutaion in DASSLE 1.0 – medtem ko je Mistral Medium generalno sposoben model, je GPT 5 v neka-

terih kategorijah boljši tudi od GPT 5.1, kar pomeni, da je verjetno v določenih pogovorih podajal boljše odgovore kot najvišje ocenjeni modeli.

5 ZAKLJUČEK

V tem članku smo predstavili celovito evalvacijo najpogosteje uporabljanih velikih jezikovnih modelov v kontekstu slovenskega jezika. Z vključitvijo modelov iz štirih večjih komercialnih ekosistemov ter enega domačega, odprtega modela smo predstavili vpogled v trenutno stanje splošnih, jezikovnih in pogovornih sposobnosti VJM v slovenskem jeziku. Evalvacijo smo izvedli na raznolikem naboru učnih množic, ki skupaj pokrivajo sledenje navodilom, razumsko sklepanje, zanesljivost odgovorov, slovnično kompetenco ter besedilno koherentnost.

Rezultati kažejo, da so sodobni komercialni modeli, predvsem GPT-5.1 z visokim nivojem premišljanja ter Gemini-2.5-Pro, dosegli visoko raven uspešnosti pri nalogah, ki preverjajo razumevanje in uporabo informacij v slovenskem jeziku, zlasti na množici Slovenian LLM Evaluation. Ti modeli so se izkazali kot zanesljivi pri reševanju strukturiranih nalog, ki temeljijo na razumevanju navodil in logičnem sklepanju, kar potrjuje njihovo uporabnost tudi v slovenskem jezikovnem prostoru. Odprti model Mistral Large 3 se je kljub bistveno nižji ceni in odprti naravi presenetljivo dobro kosal z najboljšimi komercialnimi modeli, kar kaže na hiter napredek odprtokodnih pristopov.

Po drugi strani evalvacija na množici DASSLE 1.0 razkriva pomembne omejitve vseh obravnavanih modelov. Oblikoslovna in predvsem skladenjska kompleksnost slovenskega jezika ostaja velik izziv tudi za najzmogljivejše VJM. Modeli se relativno dobro spopadajo s preprostimi pravopisnimi napakami, vendar njihova uspešnost hitro upade pri globljih jezikovnih pojavih, kot so besediščne, oblikoslovne in skladenjske napake. To nakazuje, da trenutni modeli še nimajo zadostno robustnega razumevanja slovenskega jezika, kar lahko omejuje njihovo zanesljivost pri zahtevnejših nalogah, kjer je natančna raba jezika ključna.

Rezultati evalvacije pogovornih sposobnosti dodatno potrjujejo ugotovitve iz preostalih nalog. Najbolje ocenjeni modeli po ELO lestvici so hkrati dosegali visoke rezultate tudi na drugih učnih množicah, kar kaže na konsistentnost njihovih splošnih sposobnosti. Uporaba več VJM kot sodnikov se je izkazala za ustrezno, saj statistike složnosti kažejo na zadovoljivo stopnjo strinjanja in s tem na verodostojnost primerjalnih ocen.

Skupno gledano rezultati kažejo, da so veliki jezikovni modeli dovolj zreli za široko uporabo v slovenskem jeziku pri nalogah razumevanja, sklepanja in splošne pogovorne interakcije. Kljub temu pa ostajajo izrazite vrzeli pri globlji jezikovni pravilnosti, kar je še posebej pomembno za rabo v izobraževanju, javni upravi, pravu in drugih domenah, kjer je natančnost jezika ključnega pomena.

LITERATURA

[1] A. Singla, A. Sukharevsky, M. Chui in B. Hall, "The state of AI," McKinsey & Company, 2025.

[2] N. Maslej, L. Fattorini, R. Perrault, Y. Gil, V. Parli, N. Kariuki, E. Capstick, A. Reuel, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, J. C. Niebles, Y. Shoham, R. Wald, T. Walsh, A. Hamrah, L. Santarlasci, J. B. Lotufo, A. Rome, A. Shi in S. Oak, "Artificial Intelligence Index Report 2025," Human-Centered Artificial Intelligence, Stanford University, Stanford, 2025.

[3] A. Praček in V. Vehovar, "Umetna inteligenca v Sloveniji 2025/I: Uporabniki GenUI," Center za družboslovno informatiko, Fakulteta za družbene vede, Ljubljana, 2025.

[4] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal in O. Schoenick, "Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge," *arXiv preprint*, p. arXiv: 1803.0547, 2018.

[5] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi in Y. Choi, "HellaSwag: Can a Machine Really Finish Your Sentence?," *arXiv preprint*, p. arXiv:1905.07830, 2019.

[6] S. Lin, J. Hilton in O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," *arXiv preprint*, p. arXiv:2109.07958, 2022.

[7] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins in K. Toutanova, "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions," *arXiv preprint*, p. aXiv:1905.10044, 2019.

[8] T. Mihaylov, P. Clark, T. Khot in A. Sabharwal, "Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering," *arXiv preprint*, p. arXiv:1809.02789, 2018.

[9] K. Sakaguchi, R. Le Bras, C. Bhagavatula in Y. Choi, "WinoGrande: an adversarial winograd schema challenge at scale," *Communications of the ACM*, pp. 99-106, 2021.

[10] Y. Bisk, R. Zellers, R. Le Bras, J. Gao in Y. Choi, "PIQA: Reasoning about Physical Commonsense in Natural Language," *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7432-7439, 2020.

[11] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse in J. Schulman, "Training Verifiers to Solve Math Word Problems," *arXiv preprint*, p. arXiv:2110.14168, 2021.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le in D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in neural information processing systems*, 35, pp. 24824-24837, 2022.

[13] A. Praček.

▪

Miha Malenšek je raziskovalec in doktorski študent na Fakulteti za računalništvo in informatiko, Univerze v Ljubljani, zaposlen v Laboratoriju za podatkovne tehnologije. V svojem delu se ukvarja predvsem s podpornimi sistemi za varno in sledljivo uporabo VJM v domenah, kjer je zanesljiva in preverljiva uporaba VJM ključnega pomena.

▪

Domen Vreš je raziskovalec in doktorski študent na Fakulteti za računalništvo in informatiko, Univerze v Ljubljani, zaposlen v Laboratoriju za strojno učenje in jezikovne tehnologije. V svojem delu se ukvarja predvsem z učenjem VJM za slovenski jezik, GaMS (Generativni Model Slovenščine).

▪

Marko Bajec je redni profesor na Fakulteti za računalništvo in informatiko Univerze v Ljubljani ter vodja Laboratorija za podatkovne tehnologije in IoT Demo Centra. Predava več predmetov s področja informatike in podatkovnih baz. V okviru aplikativnega in raziskovalnega dela se ukvarja z obvladovanjem informatike ter uporabo podatkovnih tehnologij v okviru različnih domen, kot so internet stvari, pametna mesta, pametni domovi, oskrbovana stanovanja, telemedicina ipd.