Odprava halucinacij in izboljšanje razumskega sklepanja v generativnih jezikovnih modelih

Oznaka in naziv projekta

Z2-70070 Odprava halucinacij in izboljšanje razumskega sklepanja v generativnih jezikovnih modelih
Z2-70070 BRIDGE: Bridging Reasoning and Informational Deficiencies in Generative Engines

Financiranje

Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije

Projektna skupina

Vodja projekta: dr. Matej Martinc (50070)

Vsebinski opis projekta

Na hitro razvijajočem se področju umetne inteligence (UI) je vzpon velikih jezikovnih modelov (VJM), kot sta GPT-4 in Gemini, na novo opredelil zmogljivosti obdelave naravnega jezika (ONJ). Ti zaprtokodni lastniški modeli blestijo pri ustvarjanju besedila, podobnega človeškemu, reševanju problemov, ki zahtevajo razumsko sklepanje, ter izkazujejo lastnosti in sposobnosti, ki niso bile vnaprej predvidene in presegajo njihove eksplicitne učne cilje. Po drugi strani pa njihova zaprtokodnost sproža pomisleke glede varnosti, omejene dostopnosti in zasebnosti podatkov. Odprtokodni VJM-ji, kot sta Llama in Mistral, ponujajo obetavno alternativo lastniškim modelom, saj dajejo prednost dostopnosti in prilagodljivosti. Vendar se ti manjši modeli pogosto spopadajo s težavami v sklepanju in ustvarjanju halucinacij – področij, kjer večji lastniški zaprtokodni modeli ohranjajo jasno prednost. Ta vrzel v zmogljivosti izhaja iz omejitev v učnih virih, velikosti modela in krajšega časa učenja, zaradi česar so odprtokodni modeli manj učinkoviti pri obravnavi kompleksnih nalog v resničnem svetu.

Da bi se spopadli s temi izzivi, naš projekt predlaga uporabo novih tehnik augmentacije podatkov in inovativnih učnih metodologij za izboljšanje odprtokodnih VJM-jev. Z razširitvijo obsega učnih podatkov ter z izboljšanjem zmogljivosti modela z naprednimi učnimi strategijami želimo premostiti vrzel med odprtokodnimi in lastniškimi VJM. Natančneje, predlagamo nov pristop učenja modelov na augmentiranih podatkih, kjer model učimo, da se zaveda vrzeli v lastnem znanju, tako da lahko samodejno zazna lastne pomanjkljivosti in usmerja lastno usposabljanje, da te pomanjkljivosti odpravi na ciljno usmerjen in učinkovit način. Predlagan pristop ima potencial, da znatno izboljša sposobnosti sklepanja in reševanja problemov odprtokodnih VJM-jev, zmanjša halucinacije in omogoči treniranje odprtokodnih modelov na učinkovit in konkurenčen način, saj zahteva veliko manj učnih in računskih virov kot trenutni učni režimi.

Izboljšani odprtokodni modeli bodo temeljito evalvirani na javno dostopnih testnih množicah in interdisciplinarnih nalogah, ter integrirani v aplikacije in orodja za oceno njihove uporabnosti v resničnih okoljih. To bo vključevalo oceno njihove zmogljivosti v manj razširjenih jezikih (tj. slovenščini), kjer bo poudarek na ocenjevanju njihovega potenciala za premostitev jezikovnih ovir in spodbujanju razvoja UI in ONJ za jezike z manj viri. Poleg tega bo ta raziskava prispevala dragocene vpoglede v izzive in priložnosti uporabe VJM na specializiranih področjih, ki zahtevajo visoko točnost in zanesljivost, in sicer v zdravstvu, pravu in znanosti, pri čemer bo preučila potencial teh modelov za vpliv na raziskave, odločanje in prakso v teh disciplinah. Naš končni cilj je prispevati k bolj pravičnemu ekosistemu UI z opolnomočenjem odprtokodnih modelov, da z robustno zmogljivostjo izpolnijo različne zahtevne naloge v resničnih okoljih. Ta projekt predstavlja ključni korak pri zagotavljanju, da prihodnost UI ostane vključujoča, trajnostna in inovativna ter da jeziki z manj govorci in viri, kot je slovenščina, ne bodo zapostavljeni.

Faze projekta in opis njihove realizacije

Faza 1: Zaznavanje vrzeli v znanju in sklepanju ter obogatitev podatkov za premoščanje teh vrzeli.
Trajanje: (M0–M12)

Naloge: T1.1 Razvoj pristopa za zaznavanje vrzeli v znanju in sklepanju; T1.2 Ustvarjanje virov za usposabljanje velikih jezikovnih modelov (VJM).

Faza 2: Razvoj novih tehnik usposabljanja VJMjev.
Trajanje: (M6–M24)

Naloge: T2.1 Nove tehnike poravnave (*alignment*) za zmanjšanje halucinacij modela in izboljšanje sklepanja; T2.2 Učenje modelov, da prepoznajo lastne vrzeli v znanju in sklepanju.

Faza 3: Evalvacija in uporaba izboljšanih VJM-jev v resničnem svetu.
Trajanje: (M12–M24)

Naloge: T3.1 Evalvacija na angleških referenčnih testih (*benchmarks*) in ciljnih nalogah (*downstream tasks*); T3.2 Evalvacija in uporaba na slovenskih referenčnih testih in ciljnih nalogah.

Bibliografske reference