Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije

Javni razpis za (so)financiranje raziskovalnih projektov za leto 2025

Naziv projekta:

Odprava halucinacij in izboljšanje razumskega sklepanja v generativnih jezikovnih modelih (Z2-70070)

Akronim projekta: **BRIDGE**

NAČRT ZA RAVNANJE Z RAZISKOVALNIMI PODATKI

Vodja projekta:
dr. Matej Martinc

Raziskovalna organizacija:
Institut »Jožef Stefan«

Vsebina:

0. Splošne informacije
1. Povzetek in opis raziskovalnih podatkov
2. Shranjevanje in varnostno kopiranje podatkov
3. Zagotovitev podatkov na način FAIR prek repozitorijev
4. Etični in pravni vidiki
5. Drugi raziskovalni rezultati
6. Finančna sredstva

0	Splošne informacije	
0.1	Šifra projekta	Z2-70070
0.2	Naziv projekta	Odprava halucinacij in izboljšanje razumskega sklepanja v generativnih jezikovnih modelih
0.3	Šifra vodje projekta	50070
0.4	Ime in priimek vodje projekta	Matej Martinc
0.5	Ime in priimek osebe, ki je v RO zadolžena za podporo pri ravnanju z raziskovalnimi podatki	Znanstveno-informacijski center (ZIC) IJS, skrbnica podatkov je Helena Klančnik
0.6	Interna pravila RO za ravnanje z raziskovalnimi podatki	Interna pravila so v pripravi
0.7	Različica in datum NRRP	Verzija 1.0, 15.10.2024
1	**Povzetek in opis raziskovalnih podatkov**	
1.1	Ali boste pri projektu ponovno uporabili že obstoječe podatke predhodnih raziskav?	☑ Da Podrobnosti glede ponovne uporabe podatkov so navedene v točki 1.2.

1.2	Katere vrste podatkov boste ustvarili oz. ponovno uporabili in v katerih formatih bodo shranjeni?	**Ponovna uporaba:** Uporabljeni bodo obstoječi odprtokodni nabori besedilnih podatkov in človeških anotacij (npr. nabori za model GaMS, obstoječe zbirke vprašanj/odgovorov in CoT dataseti). Prav tako bomo uporabili že obstoječe anonimizirane besedilne korpuse zdravstvenih dialogov (ADRESS, podatki projekta CroDeCo) ter pravnih besedil (SSAR 1.0 iz projekta LASAR). **Ustvarjanje:** Na podlagi interakcij sistemov (učitelj-učenec) bomo avtomatsko generirali visoko kakovostne sintetične podatke za usposabljanje (ang. alignment datasets), ki bodo vključevali alternativne verige interakcij (QA chains) ter detektirane, klasificirane in popravljene napake oz. odgovore s priznanjem neznanja ("I don't know"). **Formati:** Za zagotavljanje načel FAIR bodo besedilni podatki shranjeni v strojno berljivih, odprtih in standardnih formatih, predvsem kot JSON in JSONL datoteke, medtem ko bodo modeli in uteži shranjeni v standardnih formatih, kot so .pt in .safetensors

1.3	Kakšen je namen ustvarjanja, zbiranja oz. ponovne uporabe podatkov in njihova povezava s cilji projekta?	**Namen ponovne uporabe podatkov:** Obstoječe odprtokodne nabore podatkov (npr. podatke za učenje slovenskega modela GaMS in angleške CoT (*Chain-of-Thought* nabore) bomo ponovno uporabili kot začetno t. i. "seed" bazo. Ti podatki bodo služili kot izhodišče za sprožanje večnivojskih interakcij in dialogov med modeli (sistem učitelj-učenec) za avtomatsko detekcijo napak. Ponovna uporaba specifičnih domenskih podatkov (medicinski dialogi ADRESS in podatki projekta CroDeCo ter pravna besedila SSAR 1.0 iz projekta LASAR) je ključna za **evalvacijo**. Služili bodo kot testno okolje, s katerim bomo preverili, ali razviti modeli dejansko bolje delujejo v realnih, kritičnih scenarijih (zdravstvo, pravo, znanost), kjer sta zanesljivost in resničnost informacij izjemnega pomena. **Namen ustvarjanja novih podatkov:** Namesto zanašanja na ogromne, strojno in podatkovno potratne količine splošnih besedil (ki pri LLM-jih pogosto vodijo v ohranjanje halucinacij), je naš namen *ciljno usmerjeno ustvarjanje (sintetiziranje)* visokokakovostnih podatkov. Ustvarili bomo nove nabore podatkov, ki bodo vsebovali verige vprašanj in odgovorov, analizo napak ter odgovore tipa "Ne vem". Z njimi bomo VJMje učili zavedanja o lastnem neznanju. Na novo ustvarjeni sintetični podatki so nujni predpogoj za WP2, saj bodo uporabljeni za inovativne metode usposabljanja. Brez teh specifično ustvarjenih podatkov modelov ne bi mogli naučiti samozavedanja in samokorekcije, kar bi onemogočilo doseganje končnega cilja – izenačitev zanesljivosti odprtokodnih modelov s komercialnimi (proprietarnimi) ob bistveno manjši porabi računskih virov.
1.4	Kakšna je pričakovana velikost podatkov, ki jih nameravate ustvariti oz. ponovno uporabiti?	☑ 100–1000 GB *Tekstovni podatki bodo zavzemali relativno malo prostora (do 10 GB), shranjevanje velikih jezikovnih modelov z več milijardami parametrov pa po drugi strani zahteva veliko diskovnega prostora.*
2	**Shranjevanje in varnostno kopiranje podatkov**	

2.1	Kje bodo podatki med izvajanjem projekta shranjeni in varnostno kopirani?	Podatki in razviti VJM modeli bodo shranjeni na lokalnih strežnikih Odseka za tehnologije znanja na Institutu Jožef Stefan (IJS), ki razpolaga z 19 GPU/CPU strežniki z velikimi diskovnimi kapacitetami. Notranja IT služba na IJS bo zagotavljala redno avtomatsko varnostno kopiranje teh podatkov. Občutljivi in osebni podatki bodo izolirani in varovani v internem repozitoriju za občutljive podatke na naših strežnikih z dostopom izključno za raziskovalce na projektu. Modeliranje z neobčutljivimi podatki bo občasno potekalo tudi na evropski superračunalniški infrastrukturi EuroHPC, ki ima lastne protokole za začasno hrambo podatkov med analizo in izvajanjem.
2.2	Kako boste izbrali podatke za dolgoročno hrambo?	Cilj projekta je spodbujanje odprte znanosti. Kadar koli je to mogoče, bodo pridobljeni ali ustvarjeni podatki trajno javno dostopni in deponirani v infrastrukturo Clarin.si (certificiran repozitorij po standardu CoreTrustSeal za jezikovne vire) ali portal Hugging Face (ki podpira dodeljevanje trajnih identifikatorjev prek povezav z Zenodo), s čimer se bo zagotovil trajni identifikator (DOI) in dolgoročno skrbništvo podatkov po zaključku projekta. Za avtorsko zaščitene podatke bomo upoštevali pogodbene obveznosti med institucijo prejemnico in imetnikom avtorskih pravic, vendar ti podatki ne bodo objavljeni. Sledili bomo 57.a (besedilno in podatkovno rudarjenje), 57.b (besedilno in podatkovno rudarjenje za namene znanstvenega raziskovanja) in 57.c (znanstveno raziskovanje) členom Zakona o avtorskih in sorodnih pravicah (Url. RS št 16/07 – uradno prečiščeno besedilo, 68/08, 110/13, 56/15, 63/16 – ZKUASP, 59/19 in 130/22), ki opredeljujejo pravico do hranjenja podatkov in njihovo rabo za znanstvene namene.
2.3	Ali bodo podatki shranjeni v zaupanja vrednem repozitoriju?	☑ Da Kadar koli je to mogoče, bodo pridobljeni ali ustvarjeni podatki trajno javno dostopni in deponirani v infrastrukturo Clarin.si (certificiran repozitorij po standardu CoreTrustSeal za jezikovne vire) ali portal Hugging Face (ki podpira dodeljevanje trajnih identifikatorjev prek povezav z Zenodo), s čimer se bo zagotovil trajni identifikator (DOI) in dolgoročno skrbništvo podatkov po zaključku projekta.
3.	**Zagotovitev podatkov na način FAIR**	
3.1	**Zagotavljanje najdljivosti podatkov (F)**	

3.1.1	Ali bodo podatki označeni s trajnim identifikatorjem (PID)?	☑ Da Vsi vnosi v repozitorij Clarin.si dobijo trajni identifikator (handle). Znanstvene objave v ciljnih revijah in na odprtih repozitorijih, kot je ArXiv, dobijo identifikator DOI.
3.1.2	Kateri metapodatki bodo ustvarjeni in kateri metapodatkovni standardi bodo pri tem upoštevani?	Dostopnost in najdljivost vseh javnih podatkov bo zagotovljena z bogato opisnimi metapodatki po specifičnih standardih izbranih repozitorijev (Dublin Core, standardne Clarin in Hugging Face metapodatkovne sheme – Model Cards). Repozitorija Clarin.si in Hugging Face, ki ju bomo uporabljali za trajno shranjevanje ustvarjenih podatkov in modelov, definirata svoj nabor metapodatkov, ki jim je potrebno slediti in ki jih bomo ustvarili. Za usposobljene modele bodo podani "Model Cards", ki natančno pojasnjujejo tehnične specifikacije, vhodne/izhodne podatke, namembnost, potencialne bias-e in merila uspešnosti. Programska koda bo objavljena preko odprtih platform (npr. GitHub/GitLab) in licencirana pod ustrezno odprtokodno licenco (npr. Apache 2.0, MIT).
3.1.3	Ali bodo metapodatki vsebovali ključne besede za izboljšanje najdljivosti in možnosti ponovne uporabe?	☑ Da Obrazložitev: Uporabljali bomo ključne besede in kategorije, ki jih predpisujejo repozitoriji, kamor bomo odlagali ustvarjene podatke.
3.1	**Zagotavljanje dostopnosti podatkov (A)**	
3.2.1	Ali bodo vsi podatki odprto dostopni?	☑ Ne Razlog: Temeljni tehnološki in jezikovni resursi (usposobljeni open-source modeli in novi algoritemsko generirani sintetični podatki) ter programska koda, ki jih bomo razvili v okviru WP1 in WP2, bodo ob objavi rezultatov popolnoma odprto dostopni. Omejitve pa veljajo za določene evalvacijske nabore (v WP3), saj vsebujejo občutljive podatke (npr. prepise dialogov oseb s kognitivnimi motnjami - demenco). Čeprav bomo uporabljali predhodno pridobljene in že anonimizirane podatke iz baz drugih projektov (npr. CroDeCo, LASAR), imajo ti omejitve pri prenosljivosti in deljenju (data sharing agreements). Zaradi pravnih razlogov varstva osebnih podatkov (ZVOP-2) in zaščite pravic deležnikov, ti specifični testni nabori in morebitna vmesna analitika ne bodo prosto objavljeni

		v repozitorijih. Dostop do teh specifičnih podatkov bo omejen ali podvržen t. i. nadzorovanemu dostopu, če bo ta predviden pri izvirnih lastnikih.
3.2.2	Kdaj bodo podatki odprto dostopni in za koliko časa?	Raziskovalni podatki bodo načeloma dostopni ob objavah v repozitorijih in v znanstvenih publikacijah. Vsi izdelani sintetični podatki bodo odprto dostopni ob koncu projekta in časovno ne bodo omejeni. Dostopnost podatkov ne bo časovno omejena.
3.2.3	Na kakšen način bo v primeru omejitev pri uporabi omogočen dostop do podatkov med izvajanjem projekta in po njegovem zaključku?	Vsi sintetični podatki za treniranje VJMjev bodo prosto dostopni. Omejitve pa veljajo za določene evalvacijske nabore v WP3, saj vsebujejo občutljive podatke (npr. prepise dialogov oseb s kognitivnimi motnjami – demenco). Čeprav bomo uporabljali predhodno pridobljene in že anonimizirane podatke iz baz drugih projektov (npr. CroDeCo, LASAR), imajo ti omejitve pri prenosljivosti in deljenju (data sharing agreements). Zaradi pravnih razlogov varstva osebnih podatkov (ZVOP-2) in zaščite pravic deležnikov, ti specifični testni nabori in morebitna vmesna analitika ne bodo prosto objavljeni v repozitorijih. Dostop do teh specifičnih podatkov bo omejen ali podvržen t. i. nadzorovanemu dostopu, če bo ta predviden pri izvirnih lastnikih, podatki pa bodo hranjeni na internih strežnikih.
3.2.4	Ali bo za dostop do podatkov oz. njihovo branje potrebna dodatna dokumentacija oz. informacija o ustrezni programski opremi?	☑ Ne Obrazložitev: Vsi podatki so shranjeni v odprtih formatih, za njihovo rabo ni potrebna specializirana programska oprema.
3.3	**Zagotavljanje interoperabilnosti podatkov (I)**	
3.3.1	Katere geslovnike oz. šifrante boste uporabili pri pripravi podatkov in metapodatkov?	Uporabljali bomo standardno kategorizacijo repozitorijev CLARIN.SI in Hugging Face. Repozitorij CLARIN.SI poganja sistem LINDAT, ki temelji na tehnologiji CLARIN DSpace in ustrezni metapodatkovni shemi.
3.3.2	Ali boste primorani uporabiti manj poznane ali lastne geslovnike oz. šifrante?	☑ Ne Obrazložitev: Povsod bomo uporabljali znane in odprte standarde zapisovanja.
3.4	**Zagotavljanje ponovne uporabe podatkov (R)**	
3.4.1	Na kakšen način boste zagotovili dokumentacijo,	Sledili bomo navodilom repozitorijev Clarin.si in HuggingFace, ki zahtevajo opis metapodatkov in

	potrebno za ponovno uporabo podatkov?	podrobnejši opis podatkov. Kjer je mogoče, bomo ustvarjenim virom dodali povezave na znanstvene publikacije.
3.4.2	Ali bodo vaši podatki javno dostopni in licencirani v skladu z odprto licenco CC0, da bo s tem omogočena čim širša ponovna uporaba?	☑ Ne Obrazložitev: Večina podatkov bo dostopna pod odprtimi licencami CC-BY in CC-BY-ATTR 4.0. VJM bodo odprtodostopni pod licenco MIT ali GNU-3. Bolj zaprte licence bodo samo uporabljene v primeru, kjer bo to potrebno zaradi upravljanja osebnih podatkov ali avtorskih pravic.
3.4.3	Kakšne postopke zagotavljanja kakovosti podatkov boste uporabili?	Kakovost sintetično generiranih podatkov bo zagotovljena z inovativnim sistemom preverjanja LLM-učenec/LLM-učitelj. V kolikor odprtokodni ocenjevalni (učiteljski) modeli ne bodo dovolj zanesljivi (kar predvideva naša analiza tveganj), bomo kot filter vključili zunanje zmogljivejše modele za avtomatično preverjanje dejstev, logike in halucinacij (npr. GPT-4), kot tudi validacijo z ročno človeško referenčno analizo. Kakovost sintetičnih podatkov se bo prav tako ugotovaljala posredno, s izboljšnim delovanjem modelov naučenih na teh podatkov, ki bodo evalvirani na uveljavljenih odprith testnih množicah (benchmarks), npr. ARC in TruthfulQA).Izdelane podatkovne množice bodo prav tako podvržene najprej temeljitemu pregledu avtorjev in internemu pregledu kakovosti na nivoju uporabnikov ustvarjenih podatkov.
4.	**Etični in pravni vidiki**	
4.1	Ali obstajajo etična ali pravna vprašanja, ki bi lahko vplivala na deljenje podatkov?	☑ Da Glavna etična in pravna vprašanja so: **Varstvo občutljivih osebnih podatkov (Zdravstvo):** V okviru evalvacije modelov za detekcijo kognitivnih motenj (demence) bomo ponovno uporabili transkripte dialogov iz zdravstvenega okolja (npr. podatki projekta CroDeCo in tuja baza ADRESS). Čeprav so ti podatki že anonimizirani s strani primarnih raziskovalcev, gre za posebne vrste osebnih podatkov (podatki o zdravstvenem stanju), za katere veljajo stroga določila Splošne uredbe o varstvu podatkov (GDPR) in ZVOP-2. Z etičnega in pravnega vidika obstaja tveganje re-identifikacije pacientov. Teh podatkov naš projekt v nobenem

		primeru ne bo delil ali javno objavil. Dostop do njih urejajo izvirni skrbniki (npr. preko t.i. Data Use Agreements). Naš projekt bo javno delil izključno strogo agregirane in statistične rezultate uspešnosti naših jezikovnih modelov na teh podatkih. Prav tako teh podatkov ne bomo pošiljali v zunanje oblačne strotive, da preprečimo tveganje uhajanja podatkov, temveč se bodo obdelovali izključno na lokalnih strežnikih IJS. **Pogodbene omejitve in pravice intelektualne lastnine (Pravo in Znanost):** Pri evalvaciji na pravnem področju bomo uporabljali bazo SSAR 1.0 (slovenska besedila o socialnih pravicah, zbrana v projektu LASAR). Takšne zbirke imajo pogosto določene pogoje uporabe, avtorske pravice ali omejitve redistribucije s strani institucij, ki so jih ustvarile. Ponovno bomo delili zgolj agregirane rezultate evalvacije, ne pa samih izvirnih baz, razen če specifična licenca izvirnega projekta to izrecno dovoljuje. **Licenčni pogoji izhodiščnih VJM:** Za ustvarjanje sintetičnih podatkov in nadaljnje učenje bomo uporabili obstoječe odprtokodne modele (npr. iz družine Gemma 3 ali GaMS). Kljub oznaki "odprtokodno" imajo ti modeli specifične licence. Vsi na novo ustvarjeni modeli in generirani sintetični nabori podatkov bodo javno deljeni, vendar bodo izdani pod licencami, ki so združljive oziroma podrejene licenčnim zahtevam izvirnih modelov, da se zagotovi popolna pravna skladnost.
4.2	Ali boste med izvajanjem projekta obdelovali oz. hranili osebne podatke?	☑ Da Omejena aplikativna raziskava v domeni zdravstva iz WP3 vključuje evalvacijo na podlagi predhodno obstoječih analitičnih naborov diagnostike za demenco (podatki bolnikov in njihovih besedil). Ti podatki se ne bodo na novo zbirali od človeških subjektov znotraj tega projekta, marveč bodo zbrani iz obstoječih referenčnih testnih podatkov in repozitorijev sorodnih projektov (angleški nabori ADRESS ter slovenski podatki iz projekta CroDeCo). Ti podatki so ustrezno anonimizirani oziroma psevdonimizirani že v originalnem zajemu. Glavno tveganje kljub temu ostaja t.i. re-identifikacija ob neustrezni zaščiti tekstov ali pa spletno uhajanje podatkov skozi oblačne storitve VJM. Ukrepi proti tem tveganjem so pojasnjeni v točki 4.2.

4.3	Ali bodo med projektom ustvarjene oz. ponovno uporabljene posebne vrste osebnih podatkov?	☑ Da (Uporaba je pojasnjena v točki 4.2) Za najvišjo varnost bo ustvarjen interni, z močnim geslom in avtentikacijo zaščiten podatkovni repozitorij samo za raziskovalce neposredno udeležene v projektu. Občutljivih besedil pod nobenim pogojem ne bomo obdelovali preko zunanjih strežnikov oblačnih modelov (API), saj bomo za zaščito zasebnosti subjektov poskrbeli izključno z lokalno nameščenimi modeli na lastnih, t.i. *in-house* strežnikih IJS. Na ta način zagotovimo, da osebni podatki v nobenem primeru ne zapustijo omrežja ustanove oziroma se ne uporabijo za urjenje tuje infrastrukture. Raziskovalci (PI) že imajo veliko izkušenj z upoštevanjem strogih etičnih načel, ZVOP-2 ter delom na zdravstvenih podatkih.
4.4	Kako boste uredili lastništvo avtorskih pravic in pravic intelektualne lastnine podatkov, ki jih boste ustvarili ali ponovno uporabili?	Vprašanje avtorskih pravic in intelektualne lastnine (IL) bomo reševali z jasnim ločevanjem med ponovno uporabljenimi podatki in na novo ustvarjenimi rezultati. **1. Ponovno uporabljeni podatki (obstoječi podatki, korpusi in bazni modeli):** Avtorske pravice in pravice IL nad ponovno uporabljenimi podatki in osnovnimi modeli ostanejo v izključni lasti njihovih prvotnih ustvarjalcev oz. institucij. Pri zdravstvenih in pravnih podatkih (npr. ADRESS, zbirke iz projektov CroDeCo in LASAR) bomo strogo spoštovali pogoje uporabe (Data Use Agreements) in obstoječe pravice IL. Teh podatkov ne bomo prilaščali ali distribuirali. Pri odprtokodnih modelih in zbirkah podatkov (GaMSi) bomo dosledno spoštovali njihove specifične licenčne pogoje, vključno z navajanjem avtorstva (atribucija) in morebitnimi omejitvami komercialne uporabe. **2. Na novo ustvarjeni podatki in rezultati (sintetični podatki, razviti modeli, programska koda):** Lastništvo: V skladu z delovnopravno zakonodajo in pravili raziskovalne organizacije bodo materialne avtorske pravice in pravice IL nad na novo razvitimi rešitvami in podatki formalno v lasti matične raziskovalne organizacije (Institut Jožef Stefan). Moralne avtorske pravice pa vselej pripadejo raziskovalcem (avtorjem). Da bi dosegli cilje odprte znanosti in omogočili široko uporabo rezultatov, bo IJS (kot lastnik materialnih pravic) rezultate in podatke ponudil javnosti pod odprtokodnimi licencami. Vse znanstvene publikacije, ki bodo nastale na podlagi teh podatkov, bodo objavljene v odprtem dostopu (zlati ali zeleni odprti dostop). Avtorske pravice za članke bodo,

		kjer bo to mogoče in v skladu s politiko revij (npr. revije konference ACL), obdržali avtorji ali institucija ob uporabi licenc Creative Commons (CC-BY), kar omogoča neomejeno nadaljnjo distribucijo spoznanj.
5.	Drugi raziskovalni rezultati	
5.1	Ali boste poleg podatkov ustvarili ali ponovno uporabili tudi druge raziskovalne rezultate?	☑ Da Ustvarjenih bo več digitalnih orodij, kot so veliki jezikovni modeli in programska koda. Odprt dostop in uporaba bosta omogočena po načelih FAIR. Zagotovili bomo, da bodo digitalni predmeti in drugi raziskovalni podatki objavljeni, najdljivi, dostopni, interoperabilni in uporabni.
6.	Finančna sredstva	
6.1	Kakšni bodo stroški ravnanja s podatki in drugimi rezultati projekta po načelih FAIR in kako bodo kriti?	Načeloma so stroški varne hrambe podatkov in zagotavljanje ustrezne podatkovne infrastrukture znatni. Raziskovalne organizacije jih pretežno pokrivajo iz javnih sredstev za raziskovalno dejavnost, iz projektne režije in sponzorskih sredstev. Javne raziskovalne infrastrukture, kot je Clarin.si, so financirane iz državnega proračuna. Projekt sam teh stroškov načeloma ne nosi, razen posredno skozi režijske stroške.
6.2	Kdo bo odgovorna oseba za ravnanje z raziskovalnimi podatki pri projektu?	Vodja projekta, Matej Martinc.