Paveikslėlis, kuriame yra tekstas, Šriftas, ekrano kopija, Elektrinė mėlyna spalva

Automatiškai sugeneruotas aprašymas

**SINTETINIŲ LYGIAGREČIŲ TEKSTYNŲ SUKŪRIMO PASLAUGŲ**

**TECHNINĖ SPECIFIKACIJA**

1. **INFORMACIJA APIE PROJEKTĄ**
2. Valstybės skaitmeninių sprendimų agentūra (toliau – Perkančioji organizacija arba Užsakovas) pagal 2021–2030 metų Lietuvos Respublikos ekonomikos ir inovacijų ministerijos valstybės skaitmeninimo plėtros programos pažangos priemonės Nr. 05-002-01-07-08 „Kurti technologinius sprendimus ir įrankius, leidžiančius saugiai ir patogiai naudotis paslaugomis“ 2 veiklą „Kalbinių išteklių dirbtinio intelekto technologijų sprendimų poreikiams plėtra”, įgyvendina projektą „**Sintetinių lygiagrečių tekstynų sukūrimas**“ (toliau – Projektas). Priemonė kalbinių išteklių dirbtinio intelekto technologijų sprendimų poreikiams kurti ir plėtoti įgyvendinama Ekonomikos gaivinimo ir atsparumo didinimo priemonės (RRF) lėšomis.
3. **Projekto tikslas** - sukurti sintetinius lygiagrečius tekstynus neuroninio mašininio vertimo sistemų vystymo poreikiams.
4. Įvardintųjų esybių (toliau - ĮE) kokybiškas vertimas fleksinėse kalbose naudojant neuroninių tinklų priemones buvo ir išlieka problemiškas. Viena iš pagrindinių priežasčių yra nepakankamas ĮE vertimų kiekis neuroninio mašininio vertimo sistemos (toliau - NMVS) treniravimo tekstynuose, net jei šie tekstynai yra milžiniški. Jei ĮE (arba ĮE kai kurių gramatinių formų) nėra treniravimo tekstyne - neuroninio mašininio vertimo rezultatas gali būti neprognozuojamas. Su analogiška problema susiduriama verčiant bet kurį prastai treniravimo tekstyne reprezentuojamą žodį ar jų grupę, pvz. medicininį terminą. Vienas iš galimų efektyvių ĮE ir įvairių sričių terminų mašininio vertimo pagerinimo būdų yra specialių sintetinių tekstynų, apimančių ĮE (ar kurios nors srities terminus) generavimas ir naudojimas treniruojant NMVS.
5. Projekte kuriami sintetiniai tekstynai yra unikalūs ištekliai, kurių analogų lietuvių kalba šiuo metu nėra. Sukurti sintetiniai lygiagretūs tekstynai bus nemokamai ir viešai prieinami. Tekstynai gali būti panaudojami siekiant patobulinti mašininio vertimo ir lokalizavimo paslaugų informacinės sistemos (MAVILIS) (https://registrai.lt/management/objects/view/10292) (vertimas.vu.lt) neuroninio mašininio vertimo sprendimą. Tekstynai gali būti panaudoti kuriant naujus mašininio vertimo pritaikymus bendrai arba medicinos sričiai (pvz. medicinos srities šneka į šneką vertimo sistemas, interneto puslapių vertimo įskiepius, įskiepius profesionaliems vertėjams ir t.t.).

**II. PIRKIMO OBJEKTAS**

1. **Pirkimo objektas** – trijų sintetinių lygiagrečiųjų tekstynų, apimančių šias kalbų poras lietuvių-anglų, lietuvių-prancūzų ir lietuvių-vokiečių, sukūrimas.

**III. TECHNINIAI APRAŠYMAI IR KELIAMI REIKALAVIMAI**

1. **Techniniai reikalavimai tekstynams:**

6.1. Tekstynai turi apimti tris kalbų poras: lietuvių-anglų, lietuvių-prancūzų ir lietuvių-vokiečių, kiekvienai kalbai sintezuojant ne mažiau kaip 1 mln. lygiagrečiųjų sakinių.

6.2. Kiekvienai kalbų porai turi būti paruošiamos ne mažiau kaip 20 000 įvardintų esybių naudojant ne mažiau 50 kontekstinių šablonų, skirtingų kiekvienai įvardintų esybių kategorijai (ne mažiau 10 kategorijų).

6.3. Sintetinės lygiagrečių sakinių poros su įterptais medicininiais terminais sudaromos tik lietuvių-anglų kalbų porai, įterptų terminų skaičius – ne mažiau 1 000 vnt.

6.4. Tekstynuose leidžiamas ne didesnis nei 0,5% rašybos klaidų lygis.

6.5. Sintetiniai tekstynai ir šablonai rengiami iš Tiekėjo turimų lygiagrečių tekstynų ir/arba Perkančiosios organizacijos pateiktų lygiagrečių tekstynų. Teikėjas turi gauti ir naudoti papildomus lygiagrečius tekstynus, jei jis teikdamas paslaugas naudoja Perkančiosios organizacijos pateiktus lygiagrečius tekstynus ir jų nepakanka pasiekti techninėje specifikacijoje nustatytas paslaugų rezultatų apimtis.

7. **Reikalavimai techninei realizacijai:**

7.1. Sintetinio tekstyno kūrimas prasideda nuo NMVS treniravimo lygiagretaus tekstyno analizės. Pasirenkama kalbų pora. Situacijos analizei pasirenkama lietuvių. Iš lietuviškosios tekstyno dalies ekstrahuojamas įvardintųjų esybių sąrašas. Išrenkami gerai tekstyne reprezentuojamų ĮE kontekstai sakinių lygyje. ĮE sąrašas kategorizuojamas pagal ĮE kategorijas ir kategorijų sąrašas suderinamas su Perkančiąja organizacija. Iš šių išsirinktų kontekstinių sakinių konstruojami atskiri kontekstiniai šablonai, individualūs kiekvienai ĮE kategorijai, pvz. miestų, upių pavadinimų ar pavardžių kontekstiniai šablonai. Vienas šablonas – tai du lygiagretūs sakiniai su pakeitimams numatytomis vietomis. Kontekstiniai šablonai turi apimti visas ĮE gramatines formas (linksnius), o taip pat visų į šabloną parinktų lietuviškų sakinių vertimus į kitą kalbą. Visi šie šablonai lingvistiškai anotuojami.

7.2. Antrajame etape konstruojamas norimų įterpti į NMVS ĮE sąrašas, atsižvelgiant į iš treniravimo tekstyno ekstrahuotą ir su Perkančiąja organizacija suderintą sąrašą. Įterpimo sąrašas išplečiamas visomis galimomis ĮE gramatinėmis formomis. Šis sąrašas dvikalbis, jis taip pat analogiškai lingvistiškai anotuojamas ir suskirstomas į kategorijas, atitinkančias paruoštų kontekstinių šablonų kategorijas.

7.3. Trečiajame etape užpildomi kontekstiniai šablonai. Jie automatiškai pildomi pagal kategorijas, keičiant šablonų esančią ĮE kita įvardijama esybe iš įterpimo sąrašo su atitinkama anotavimo žyma. Iš treniravimo tekstyno kuriami šablonai po įterpimo į vertimo sistemą minimaliai paveikia vektorinę vertimo sistemos struktūrą, todėl sistemą tokiu metodu galima papildyti dideliu kiekiu įvardintųjų esybių (ar įvairių sričių terminų). Užpildytų šablonų visuma sudaro sintetinį lygiagretųjį tekstyną, kuris naudojamas NMVS treniravimui.

7.4. Analogiški veiksmai atliekami sintezuojant medicinos terminų tekstynus. Kategorizuojant medicinos terminus naudojama SNOMED CT medicininė ontologija. Kategorizuojant sukuriami vienodo hierarchinio lygio terminų klasteriai. Reprezentatyviausių treniravimo tekstyne esančių klasterio terminų konteksto pagrindu kuriami atitinkami kontekstiniai šablonai. Trūkstant informacijos, šablonai gali būti kuriami pagal analogiško, tačiau klasteriui nepriklausančio termino kontekstą, esantį treniravimo tekstyne.

7.5. Sintetiniai tekstynai visų pirma kuriami anglų ir lietuvių kalboms, įskaitant medicininių terminų pridėjimą. Kitoms kalboms sintetiniai tekstynai kuriami po to arba lygiagrečiai anglų ir lietuvių kalbų sintetiniam tekstynui.

**8. Reikalavimai dokumentams:**

8.1. Kiekvienam tekstynui turi būti parengta dokumentacija, kurioje detaliai apibūdinta tekstyno struktūra, jo sudarymo ir anotavimo metodikos:

8.1.1. Kiekvienam sukurtam tekstynui turi būti parengta dokumentacija, apibūdinanti tekstyno formatus;

8.1.2. Kiekvienam sukurtam tekstynui aprašytos anotacijos ir jų reikšmės;

8.1.3. Turi būti pateikta detali statistika įvairiais pjūviais (kartu pateikiami ne tik išvestiniai, tačiau ir pirminiai duomenys);

8.2. Paslaugų rezultatai turi būti aprašyti metaduomenimis, kurių formatą ir standartą nustato Lietuvos teisės aktai, reguliuojantys valstybės informacinių išteklių valdymą ir tvarkymą. Paslaugų rezultatų aprašymo metaduomenimis detalizavimas turi būti suderintas su Perkančiąja organizacija ir įtrauktas į Paslaugų rezultatų kūrimo pažangos stebėsenos reglamentą (toliau – Reglamentas).

8.3. Visa dokumentacija turi būti parengta laikantis bendrinės lietuvių kalbos taisyklių. Visi Paslaugos Teikėjo parengti dokumentai turės būti suderinti su Perkančiąja organizacija. Dokumentų galutinės versijos turi būti pateiktos elektroniniu (MS Word) arba kitu su Perkančiąja organizacija suderintu redagavimui tinkamu formatu. Pastabos bei korekcijos dokumentų projektuose turi būti teikiamos MS Office programinio paketo (ar lygiaverčio) pakeitimų sekimo (angl. track changes) bei komentavimo funkcijomis.

**9. Teisiniai reikalavimai:**

9.1. Kuriant tekstynus privalu laikytis:

9.1.1. autorių ir gretutines teises reguliuojančių Lietuvos ir Europos Sąjungos teisės aktų;

9.1.2. duomenų apsaugą reguliuojančių Lietuvos ir Europos Sąjungos teisės aktų;

9.1.3. dirbtinio intelekto sistemoms rengiamų mokymo duomenų kokybę reguliuojančių Lietuvos ir Europos Sąjungos teisės aktų;

9.1.4. Lietuvos Respublikos ir Europos Sąjungos teisės aktų, reguliuojančių atvirų duomenų formavimo ir skelbimo principus, nuostatų (Lietuvos Respublikos teisės gauti informaciją ir duomenų pakartotinio naudojimo įstatymo, kurio nuostatos perkelia Europos Parlamento ir Tarybos direktyvą (ES) 2019/1024 dėl atvirųjų duomenų ir viešojo sektoriaus informacijos pakartotinio naudojimo; Duomenų atvėrimo tvarkos aprašo nuostatų; Lietuvos atvirų duomenų portalo nuostatų; Europos Komisijos įgyvendinimo reglamento (ES) 2023/138, kuriuo sudaromas konkrečių didelės vertės duomenų rinkinių sąrašas ir nustatoma jų skelbimo bei pakartotinio naudojimo tvarka, nuostatų).

9.2. Turi būti laikomasi Europos Sąjungos dirbtinio intelekto akto ir Europos Sąjungos Duomenų akto nuostatų.

9.3. Turi būti parengta/parinkta kuriamų informacinių išteklių naudojimo licencija, kuri užtikrina atvirą ir nemokamą prieigą prie sukurtų tekstynų ir suteikia teisę atviram (be ribojimų) bei nemokamam jų naudojimui, perpanaudojimui ir tolesniam vystymui. Visi sukurtų tekstynų formavimui panaudoti tekstai turi turėti atitinkamas licencijas (jei reikia).

9.4. Visos turtinės teisės į paslaugų rezultatus perleidžiamos Perkančiajai organizacijai.

**IV.** **REIKALAVIMAI PAŽANGOS STEBĖSENAI IR PASLAUGŲ REZULTATŲ PATEIKIMO ETAPAMS**

10. Pasirašius Sutartį, Teikėjas per 10 (dešimt) darbo dienų turės parengti ir derinimui su Užsakovu pateikti:

10.1. Reglamentą, kuriame turi būti nustatyta paslaugų teikimo progreso rezultatų pateikimo, Užsakovo informavimo apie paslaugų teikimo progresą tvarka, grafikas, komunikavimo tvarka;

10.2. Tekstynų kūrimo metodinę dokumentaciją, kurioje būtų pateiktos su tekstynų kūrimu susijusios metodikos ir strategija. Viena iš privalomų sąlygų - ne vėliau kaip per 5 (penkias) pirmąsias einamojo mėnesio darbo dienas Teikėjas Užsakovui pateikia informaciją apie praėjusį kalendorinį mėnesį pasiektą paslaugų teikimo progresą.

10.3. Užsakovas pastabas derinimui pateiktiems dokumentams turi pateikti per 5 (penkias) darbo dienas nuo dokumentų gavimo dienos. Teikėjas pagal Užsakovo pateiktas pastabas dokumentus turi ištaisyti per 5 (penkias) darbo dienas. Techninės specifikacijos 10.1 ir 10.2 punktuose nustatyti dokumentai su Užsakovu turi būti suderinti ne vėliau kaip per 1 (vieną) mėnesį nuo Sutarties pasirašymo dienos.

11. Tekstynų kūrimo rezultatas pateikiamas etapais, susijusiais su apmokėjimu. Kiekvieno etapo pabaigoje turi būti pateiktas rezultatas su prieaugį detalizuojančia bei aiškinančia dokumentacija. Prieaugio pateikimo forma ir pateikimo būdas nustatomas Reglamente.

12. Paslaugų rezultatų metaduomenys turi būti aprašyti pagal Lietuvos teisės aktais nustatytą ir Reglamente su Užsakovu suderintu metaduomenų standartą, taip pat turi būti patalpinti Užsakovo nustatytoje metaduomenų saugykloje. Taip pat Paslaugų rezultatai turi būti Užsakovui pateikti magnetinėje laikmenoje ar kitoje, su Užsakovu suderintoje, laikmenoje ir pasiekiami Lietuvos atvirų duomenų portale (data.gov.lt) bei Užsakovo pasirinktoje atvirųjų duomenų platformoje/saugykloje (pvz., „Hugging Face“, CLARIN ar kt.). Užsakovas atvirųjų duomenų platformą/saugyklą turi pasirinkti ne vėliau, kaip per 3 (tris) mėnesius nuo Sutarties įsigaliojimo dienos. Pasirinktoje atvirųjų duomenų platformoje/saugykloje Paslaugų rezultatų talpinimas turi būti nemokamas.

13. Tekstynų kūrimo paslaugos laikomos tinkamai suteiktomis, kai galutinis paslaugų rezultatas - tekstynai - sukurti pilna apimti, pateikta pilna ir išsami dokumentacija, licencija, tekstynai aprašyti metaduomenimis, metaduomenys ir tekstynai (įskaitant jų aprašymus ir kt. reikalingą dokumentaciją) pateikti į Užsakovo nurodytas saugyklas/platformas, bei išpildyti visi kiti techninėje specifikacijoje nustatyti reikalavimai.

14. Teikiamos paslaugos neturi kelti grėsmės nacionaliniam saugumui vadovaujantis Viešųjų pirkimų įstatymo 37 str. 9 d.

15. Užsakovas po kiekvieno etapo, nurodyto 1 lentelėje, sumoka Tiekėjui už tinkamai ir kokybiškai suteiktas Paslaugas šalims pasirašius Paslaugų perdavimo-priėmimo aktą ir Tiekėjui Sutartyje nustatyta tvarka pateikus sąskaitą-faktūrą. Paslaugos laikomos tinkamai suteiktos, kai abiejų Sutarties šalių yra pasirašytas Paslaugų perdavimo-priėmimo aktas, t. y. į Paslaugų suteikimo terminą įeina Paslaugų rezultato derinimas, koregavimas, jeigu reikia, Paslaugų rezultatų priėmimas ir Paslaugų perdavimo – priėmimo akto pasirašymas. Vadovaujantis paslaugų sutarties specialiųjų sąlygų 3.5 punktu, Užsakovas įsipareigoja priimti tinkamai ir laiku suteiktas Paslaugas, pasirašydamas Paslaugų perdavimo–priėmimo aktą ne vėliau kaip per 10 (dešimt) darbo dienų nuo Tiekėjo kreipimosi dienos, arba per šį terminą nurodyti suteiktų Paslaugų trūkumus Tiekėjui.

15. Paslaugos turi būti suteiktos **iki 2026 m. balandžio 15 d.**

16. Paskutinis atsiskaitymas atliekamas tik po to, kai Užsakovas pasirašo galutinių paslaugų perdavimo – priėmimo aktą.

**1 lentelė**

Paslaugų teikėjas paslaugas teiks tokiu grafiku:

|  |  |  |  |
| --- | --- | --- | --- |
| **Eil. Nr.** | **Etapas** | **Numatomas vykdymo terminas** | **Atsiskaitymų už etapus dalys** |
| 1 etapas | Pateikiama 15 proc. sintetinių lygiagrečiųjų tekstynų bendros apimties pagal kalbų poroms ir šablonams keliamus reikalavimus | 3 mėn. nuo Sutarties įsigaliojimo dienos | 15 (penkiolika) procentų pradinės Sutarties vertės |
| 2 etapas | Pateikiama 30 proc. sintetinių lygiagrečiųjų tekstynų bendros apimties pagal kalbų poroms keliamus ir šablonams reikalavimus | 7 mėn. nuo Sutarties įsigaliojimo dienos | 15 (penkiolika) procentų pradinės Sutarties vertės |
| 3 etapas | Pateikiama 50 proc. sintetinių lygiagrečiųjų tekstynų bendros apimties pagal kalbų poroms ir šablonams keliamus reikalavimus | 11 mėn. nuo Sutarties įsigaliojimo dienos | 20 (dvidešimt) procentų pradinės Sutarties vertės |
| 4 etapas | Pateikiama 70 proc. sintetinių lygiagrečiųjų tekstynų bendros apimties pagal kalbų poroms ir šablonams keliamus reikalavimus | 14 mėn. nuo Sutarties įsigaliojimo dienos | 20 (dvidešimt) procentų pradinės Sutarties vertės |
| 5 etapas | Pateikiama 100 proc. sintetinių lygiagrečiųjų tekstynų bendros apimties pagal kalbų poroms ir šablonams keliamus reikalavimus | **2026 m. balandžio 15 d.** | Galutinis atsiskaitymas.  30 (trisdešimt) procentų pradinės Sutarties vertės |