

Kreipiamės į Jus norėdami paprašyti prisidėti prie *Bendrojo lietuvių kalbos tekstyno* kaupimo. Jeigu turite didesnę kiekį lietuviškų tekstų kuriais galite prisidėti prie BLKT kaupimo, prašau rašykite Agnei Mikštienei agne.mikstiene@vdu.lt.

Kontekstas

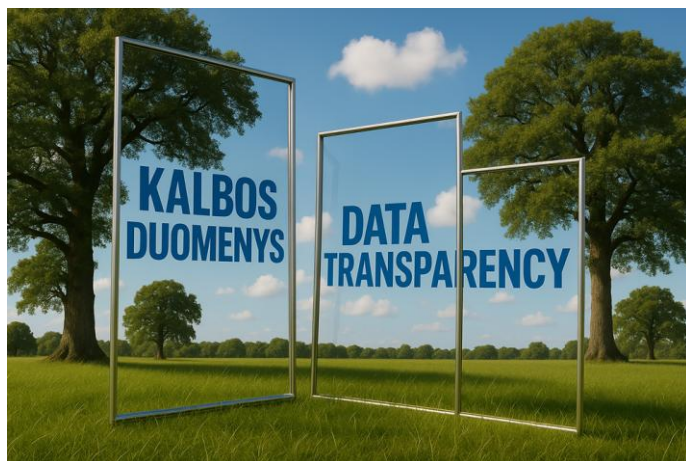
Daugeliui jau teko vienaip ar kitaip susidurti su pokalbių robotais ar tekstų generavimo programomis. Šiandien, lietuviškai kokybiškiausiai galima kalbėti su amerikiečių arba kiniškų kompanijų sukurtomis priemonėmis *ChatGPT*, *Gemini*, *Claude*, *DeepSeek*, *Lama*, *Grok* ir kitomis. Didžiosios kompanijos tokios kaip *OpenAI*, *Google*, *Baidu*, *X* ar *META* turi prieigą ir galimybes parsisiųsti viso pasaulio kalbų internetinius tekstus (taip pat ir lietuvių), todėl jų treniruoti algoritmai neblogai formuluoja lietuviškus sakinius bei išmoka lietuviškuose tekstuose užfiksuotą informaciją.

Tiesa, šias priemones naudodami vartotojai nežino su kokiais duomenimis buvo treniruoti jų modeliai, kokia jų apimtis ir kokia jų kokybė. Paprastai treniravimo duomenų vartotojas negali atsisiųsti. Beveik neabejotina, kad nesant tinkamai lietuviškų duomenų atrankai į tokias duomenų bazes pakliūva ir melas, ir dezinformacija, tendencinga ir šališka medžiaga. Taip pat, dažniausiai lietuviški duomenys paimami nesiderinus su autoriais, negavus jų leidimų, pažeidžiant autorines teises. Svarbu suprasti, kad duomenys su kuriais treniruojami lietuvių kalbos modeliai ir yra ta informacija, kuri atsiranda mūsų generuojamuose tekstuose.

Kitoks požiūris

Šis projektas turi kitokį požiūrį į lietuviškus duomenis – mums svarbu užtikrinti duomenų skaidrumą. Projekte kalbos modeliai bus treniruojami su patikrintais duomenimis, iš duomenų tiekėjų, valdytojų ar autorių gavus tinkamas licencijas ar leidimus.

Duomenys, su kuriais bus apmokinti lietuviški didieji kalbos modeliai, bus viešai prieinami visiems dirbtinio intelekto priemonių gamintojams – juos jie galės panaudoti kurdami savo lietuvių kalbos modelius. Duomenų šaltiniai ir tekstų kokybė bus kruopščiai tikrinami.



Kodėl tai svarbu

Modeliai, kurie bus apmokinti su patikrintais duomenimis, generuos taisyklingesnius lietuvių kalbos tekstus, gerai atspindės lietuviškus kultūrinius aspektus, gerai atspindės vertybines mūsų visuomenės ir valstybės nuostatas.

Reikalavimai tekstams

Tinka visų tipų tekstai - grožiniai, negrožiniai, naujienos, dokumentai, sakytinės kalbos, moksliniai, mokslo populiarinimo ir kt. Tekstai turi būti lietuviški, redaguoti, taisyklingi, gali būti vertimai. Netinka: neredaguoti, sąrašai, žodynai, socialinių tinklų tekstai, taip pat tekstai, kuriuose didelis kiekis kitų kalbų žodžių.

Projektas

Bendrojo lietuvių kalbos tekstyno ir vektorizuotų lietuvių kalbos modelių sukūrimas

<https://sitti.vdu.lt/bendrasis-lietuviu-kalbos-tekstynas-ir-vektorizuoti-modeliai/>