Tartu Ülikool

Arvutiteaduse instituut

Informaatika õppekava

Veronika Kukk

GPT mudeli sisendi ja temperatuuri mõju meditsiiniliste andmete märgendamisele

Bakalaureusetöö (9 EAP)

Juhendaja Hendrik Šuvalov

Tartu 2024

GPT mudeli sisendi ja temperatuuri mõju meditsiiniliste andmete märgendamisele

Lühikokkuvõte:

Arsti kirjutatud vabas vormis tekstid sisaldavad olulist informatsiooni patsiendite kohta. Üks meetoditest nendest tekstidest fakte eraldada on masinõppe mudelitega nimeolemite (näiteks haigus, protseduur) märgendamine. Madala ressurssidega keeltes, nagu eesti keel, on keeruline treenida kvaliteetseid märgendamise mudeleid, sest vajalikke treeningandmeid on vähe. Selles uurimistöös kasutati sünteetilisi patsientide andmeid, et uurida, milline on GPT-3.5 keelemudeli märgenduste kvaliteet eestikeelsetel andmetel. Töös võrreldi kolme temperatuuri parameetriga GPT mudeli märgendusi. Lisaks katsetati, kuidas korraga küsitavate klasside arv mudeli märgendusi mõjutab. Selgus, et kahel juhul kolmest oli madalaim temperatuur parimate tulemustega. Märgendusklasside arvu puhul leiti, et kahe ja kolme kaupa küsimine andis paremaid tulemusi kui ühe kaupa.

Võtmesõnad: nimeolemite märgendamine, loomuliku keele töötlus, suured keelemudelid, tehisintellekt, tervishoiuinformaatika

CERCS: P176 Tehisintellekt

The Effects of Input and Temperature of GPT Model on Labeling Medical Data

Abstract:

Unstructured texts written by doctors contain valuable information about patients. One of the approaches to extract information from these texts is to label named entities (for example disease, procedure) by using machine learning models. However, it is difficult to train high-quality labeling models in low resource languages, such as Estonian, since the necessary training data is scarce. In this thesis, synthetic patient data was used to examine the quality of GPT-3.5 model annotations on Estonian data. Annotations of the GPT model with three temperature parameters were compared. In addition, the thesis explores how the number of classes affects the model's annotations. The results showed that in two out of three cases the

lowest temperature had higher quality annotations than other temperatures. As for the number of classes, it was found that asking two or three classes together achieved higher results than asking only one class.

Keywords: named entity recognition, natural language processing, large language models, artificial intelligence, healthcare informatics

CERCS: P176 Artificial intelligence

Sisukord

Sissejuhatus

Tervishoiuteenuste läbiviimisel tekib rohkelt andmeid, millest suur osa on vabateksti kujul. Selleks, et neid andmeid analüüsida, on vaja meditsiinilistest tekstidest eraldada fakte. Struktureerimata tekstidest on keeruline kätte saada masinloetavaid andmeid, mida oleks võimalik kasutada tervisesüsteemi arendamise juures. Üks viis infot eraldada on kasutada nimeolemite märgendamist. Nimeolemite märgendamine (ingl *named entity recognition tagging*) on loomuliku keele töötluse ülesanne, kus tuleb teksti sees märgendada kindlatesse kategooriatesse kuuluvad olemid vastava kategooria tähisega [1].

Üks osa terviseandmetest on patsientide haiguslood, mis on enamasti arsti kirjutatud vabas vormis tekstid ning sisaldavad patsiendi kaebusi, haiguste kulgu, kavandatavaid raviplaane jms [2]. Kuna nimeolemite märgendamine määrab igale olemile tekstis kindla tüübi, siis saab haigusloost leida näiteks kõik protseduuri kategooriasse kuuluvad olemid. Selle abil on võimalik tekstist eraldada protseduurid, mis on patsiendile ravi käigus läbiviidud. Senised nimeolemite märgendajad meditsiini valdkonnas toetuvad peamiselt suurte ressurssidega keeltele, mis tähendab, et nende tulemus väheste ressurssidega keeltel, nagu eesti keel, on madala kvaliteediga. Selleks, et need mudelid oleks eestikeelsetel sisenditel head, on vaja luua uusi mudeleid või varasemaid mudeleid uute andmetega peenhäälestada.

Terviseandmed on delikaatsed ehk eriliiki isikuandmed, mis tähendab, et neid andmeid ei tohi kolmandale osapoolele jagada [3]. Seega nende andmete töötlemiseks on vaja kasutada lokaalseid vahendeid, mis on ettevõttesisesed ning ei ole ligipääsetavad mitteseotud osapooltele. Selleks, et treenida või peenhäälestada kvaliteetset andmete märgendajat, on vaja suures koguses märgendatud tekste. Tavaliselt märgendavad selliseid tekste meditsiini valdkonnaga seotud inimesed, kuid tuhandete delikaatsete tekstide märgendamine on keeruline. Eesti keeles ei ole suures koguses märgendatud ressursse saadaval, eriti meditsiiniliste tekstide näol, mistõttu on vaja leida teistsugune lahendus märgendajate treenimiseks. Tartu Ülikooli terviseinformaatika uurimisrühmas on kasutusel töökäik, milles kõigepealt luuakse lokaalse GPT-2 keelemudeliga ehk generatiivse eeltreenitud transformeriga (loodud Lepsoni magistritöö [4] käigus) eestikeelsetel pärisandmetel põhinevad sünteetilised andmed. Neid sünteetilisi andmeid, mis ei sisalda isikustatavat informatsiooni, märgendatakse GPT mudelitega. Sedasi saadakse eestikeelsetele andmetele nimeolemite märgendused, millega peenhäälestatakse lokaalne nimeolemite märgendaja ja seda mudelit saab edaspidi kasutada delikaatsetel pärisandmetel. Sama lahendust kasutatakse

ka selles uurimistöös, ent fookuseks on nimeolemite märgendamise kvaliteet GPT-3.5 mudeliga.

Sellest tulenevalt on uurimistöö eesmärk märgendada eestikeelseid meditsiinilisi andmeid GPT-3.5 keelemudeliga kasutades erinevaid *prompt engineering* tehnikaid. Andmete märgendamisele järgneb uurimistöös kasutatud meetodite omavaheline võrdlus, et teada missuguseid meetodeid on tulevikus nimeolemite märgendamisel mõistlik kasutada ja milliseid meetodeid vältida. Parima tulemusega meetodist saadud andmetega treenitakse kaks lokaalset BERT keelemudelit, üks GPT mudeli väljundist saadud märgendustega ja teine inimese poolt märgendatud andmetega, et hinnata kuivõrd erinevad on GPT mudeli ja inimese märgenduste kvaliteedid. Lisaks tehakse samal põhimõttel kaks lokaalset BERT mudelit eestikeelse baasmudeliga, et illustreerida, kuidas baasmudeli valik mõjutab tulemusi.

Eesmärgi saavutamiseks märgendatakse esmalt GPT-3.5 mudeliga ingliskeelse NCBI andmestiku test osa, et teada, kuidas on GPT-3.5 mudeli märgenduste kvaliteet töös kasutatava viibaga võrreldes selle andmestiku märgendamise edetabelis olevate mudelitega [5]. Pärast seda võrreldakse sama märgendusklassi märgendamist eestikeelsetel sünteetilistel meditsiinilistel andmetel, et näha kuivõrd mõjutab teksti keel GPT-3.5 mudelit. Järgnevalt märgendatakse GPT-3.5 mudeliga neid eestikeelseid andmeid küsides kolme erinevat märgendusklassi ühe, kahe ja kolme kaupa ühes viibas (ingl *prompt*) ning kasutades kolme erinevat temperatuuri (ingl *temperature*) parameetrit. Uurimistöö autorile teadaolevalt ei ole varasemalt uuritud meditsiiniliste nimeolemite märgendamise ülesandes GPT mudeli temperatuuri ja ülesande tükkideks jagamise mõju märgenduste kvaliteedile.

Esimeses peatükis antakse ülevaade nimeolemite märgendamisest meditsiinis ja suurtest keelemudelitest, mida selles töös kasutatakse. Teises peatükis on selgitatud töös kasutatavad andmestikud ning võrdluse all olevad meetodid. Kolmandas peatükis on analüüsitud GPT-3.5 keelemudeli märgenduste tulemusi erinevate meetodite kasutamisel ning neist parimal meetodil saadud andmetel ja inimese märgendatud andmetel peenhäälestatud BERT mudelite märgenduste tulemusi. Neljandas peatükis on välja toodud töö võimalikud edasiarendused.

1 Teoreetiline ülevaade

Selles peatükis antakse ülevaade nimeolemite märgendamise ülesandest, uurimistöös kasutatavatest suurtest keelemudelitest, tutvustatakse GPT mudeli kasutamist ning nimeolemite märgendamise ja suurte keelemudelitega seotuid varasemaid uuringuid.

Selles uurimistöös kasutatakse kahte suurt keelemudelit: GPT mudel ja BERT mudel. Mõlemad neist on transformeritel põhineva arhitektuuriga ehk need mudelid kasutavad enesetähelepanu, et hoomata lauses olevate sõnade omavahelisi suhteid [4, 5]. Sellised mudelid suudavad genereerida sidusaid tekste ning oskavad näiteks küsimustele vastata ja teksti kokku võtta [4, 5]. Seepärast on neid võimalik kasutada ka tekstide märgendamiseks.

1.1 Nimeolemite märgendamine meditsiinis

Tüüpilisemaks nimeolemite märgendamise ülesandeks on eraldada teksti seest isikute, organisatsioonide ja asukohtade nimed [1]. Siinses uurimistöös on olemiteks meditsiini valdkonna terminid, nimelt ravimid (ingl *drugs*), protseduurid (ingl *procedures*) ja haigused (ingl *diseases*) (vt 2.3.3 Märgendusklassid). Järgnevalt selgitatakse miks on nimeolemite märgendamine oluline meditsiiniliste tekstide puhul. Selleks tuleb kõigepealt anda ülevaade tervishoiusüsteemis tekkivatest andmetest.

Elektroonilised terviseandmed saab jaotada kaheks osaks: struktureeritud ja struktureerimata [2]. Struktureeritud osas on näiteks patsiendi nimi, elukoht ja haiglas viibitud aeg, samas kui struktureerimata osas kirjeldatakse haiguse kulgu ja raviplaani [2]. Ennekõike on suur osa digitaalsest terviseloost vabateksti kujul, sest need on arsti kirjutatud ülevaated patsiendi olukorrast [2]. Näiteks kannab arst patsiendi haigusloosse patsiendi kaebused ja seni läbiviidud või plaanitud protseduurid [2]. Lisaks sellele, et terviseandmed on enamasti struktureerimata, on nende töötlemise juures probleemiks andmete delikaatsus [3]. Patsiendi andmeid ei tohi jagada kolmandale osapoolele, mis tähendab, et selliste andmete töötlemine ja uurimine tuleb teha lokaalseid ehk ettevõtte- või uurimisrühmasiseseid vahendeid kasutades [3]. Oluline on nimetada veel, et terviseandmed on väga mitmekülgse sõnavaraga [1]. Olenevalt kontekstist on näiteks hambaarsti ja kardioloogi kirjutatud tekstide sisu täiesti erinev. Seega, kuna meditsiiniharude vahel on suur osa leksikast erinev, tuleb nimeolemite märgendamisel eelistada konkreetse uuritava haru kohta käivaid andmeid [1].

Struktureerimata, delikaatsetest ja kontekstist sõltuvatest tekstidest saaks eraldada ja lihtsustada olulist informatsiooni, mida kasutada tervisesüsteemi edendamiseks. Üks lähenemine on kasutada nimeolemite märgendamist, sest selle abil saab töötlemata tekstist eraldada soovitud infot [1]. Sellega on võimalik märgendatud tekstist leida näiteks läbiviidud protseduuride nimekirja ja patsiendile välja kirjutatud ravimid koos doosidega. Nende eraldatud andmetega saab teha edaspidiseid analüüse, näiteks trendide kohta, ja treenida masinõppe mudeleid [6]. Masinõppe mudeleid saab kasutada ka sümptomite ja haiguste vaheliste seoste leidmiseks [6]. Kõik need väljundid aitavad tervishoiusüsteemi paremaks teha ning seega on oluline nimeolemite märgendamisega tegeleda.

Selleks, et uurida meditsiiniliste andmete märgendamist on uurimistöös kasutusel sünteetilised andmed ja nende kaitsmiseks on kasutusel turvameetmed (vt 2.2 Turvalisuse meetmed).

Nimeolemite märgendamist saab teha mitmete vahenditega, mis jagatakse peamiselt reeglipõhiseks ja masinõppega lähenemiseks ning nende kahe kombinatsiooniks [1]. Mõnikord tuuakse reeglipõhisest eraldi välja sõnastikupõhine lähenemine [7], aga kuna uurimistöös seda varianti ei kasutata, siis on see selgitustes jäetud reeglipõhise lähenemise alla.

Reeglipõhiseid vahendeid on kasutatakse olukordades, kus märgendusklass on kindla ülesehitusega või on olemas sõnastik võimalikult paljude märgendusklassi sünonüümide ja vormidega [1]. Märgendusklassid, millele tuvastamiseks on kerge reegleid luua, on näiteks IP-aadress ja URL, sest neile on omane kindel vorm [8]. Üks eelis selle meetodi kasutamisel on, et pole vaja treeningandmeid [8]. See on kasulik meditsiini valdkonnas, sest märgendatud andmeid on vähe, eriti madala ressurssidega keeltes. Ent selles uurimistöös on keeruline reeglipõhist lähenemist kasutada. Haigusi, protseduure ja ravimeid on suures koguses ehk neile on keeruline luua terviklikku sõnastikku või kindlaid reegleid. Nendest märgendusklassidest oleks võimalik reeglipõhiselt tuvastada väike osa haigustest, mis on näiteks klassifikatsiooni koodiga kirja pandud. Küll aga uurimistöö eesmärgiks on teha kõik märgendused ainult masinõpet kasutades. Reeglipõhise lähenemise miinuseks on ajakulu, sest tuleb käsitsi kindlaks määrata märgendusklassi ülesehitus või koostada sõnastik kõikvõimalike märgendusklassi vormidega [1, 9]. Probleeme saavad tekitada ka kirjavead tekstis, sest sõnastikud ei saa arvestada kõikide vigadega. Lisaks tuleb iga uue andmestiku jaoks neid reegleid kohandada, mis teeb sellise lähenemise veel keerukamaks [9].

Teine viis nimeolemite märgendamiseks on kasutada masinõpet. Selle kasutamise jaoks on vaja uuritavast valdkonnast märgendatud andmeid, et treenida mudeleid tuvastama märgendusklasse [1,7]. Masinõppe puhul on võimalik rakendada erinevaid tehnikaid, näiteks tugivektor-masinat (ingl *support vector machine*), rekurrentset närvivõrku (ingl *recurrent neural network*) või transformereid [1]. Uurimistöös kasutatakse BERT ja GPT mudeleid, mis on mõlemad transformeritel põhineva arhitektuuriga. Enamasti on olemasolevad masinõppe mudelid treenitud suurte ressurssidega keeltel, sest nendes keeltes on rohkelt märgendatud andmeid ja seega on ka mudelid kvaliteetsed. Keeleliste erinevuste tõttu ei ole need treenitud mudelid madala ressurssidega keeltes eriti head [10]. Seega on vaja uurida ja leida viise, kuidas lahendada märgendatud andmete puudust.

1.2 BERT mudelid

BERT (ingl *Bidirectional Encoder Representations from Transformers*) on keelemudel, mille lõi Google AI[1] [11]. Võrreldes GPT mudelitega on BERT eeltreenitud väiksemal andmehulgal, ligikaudu kolme miljardi sõnaga andmestikul, mis sisaldab tekste Wikipediast ja Google'i BookCorpusest [11]. Erinevalt võimsatest GPT mudelitest (versioon 3 ja edasi), on tuhandete eeltreenitud BERT mudelite lähtekood avalikult kättesaadav [11]. Lisaks on need mainitud BERT mudelid tasuta kasutatavad [11].

BERT mudelid on enamasti GPT mudelitest kompaktsemad ja neid on võimalik laadida lokaalsesse masinasse [12]. Lokaalses masinas saab neid peenhäälestada konkreetse ülesande lahendamiseks, näiteks eestikeelsete meditsiiniliste andmete nimeolemite märgendamine. See tähendab, et andmed, mida kasutatakse sellise BERT mudeli peenhäälestamise sisendiks ei ole kättesaadavad kolmandale osapoolele. Eeskätt on mudeli lokaalsus oluline terviseandmete puhul, sest tegu on delikaatsete andmetega [3]. Selles uurimuses peenhäälestatakse kaks lokaalset BERT mudelit, mille treeningandmeteks on GPT mudeli väljundist saadud märgendused ja inimese märgendatud andmed. Selline töökäik võimaldab kasutada neid peenhäälestatud BERT mudeleid, et märgendada tulevikus päris patsientide andmeid. Nende lokaalsete mudelite testimine tehakse uurimistöös pärisandmete peal, mis on sarnased sünteetilistele andmetele.

[1] https://ai.google/

1.3 GPT mudelid

Generatiivne eeltreenitud transformer (ingl *Generative pre-trained transformer*) on OpenAI loodud keelemudelite seeria [13]. Selleks, et GPT mudelid saaks pädevalt lahendada keelega seotud ülesandeid, on lisaks transformerite arhitektuurile neid eeltreenitud suurtel andmehulkadel, milles on üle 45 TB veebitekste [14].

GPT mudeleid on võimalik raha eest kasutada OpenAI API kaudu [13]. 2023. aasta novembri seisuga kõige uuem saadaval versioon GPT mudelitest on GPT-4 Turbo mudel, ent API kaudu on ka vanemad versioonid kättesaadavad [13]. Lisaks OpenAI API-le, on GPT mudelid saadaval Azure OpenAI Service'i[2] kaudu, mis ei kasuta kasutaja sisestatud ja genereerituid andmeid mudelite parendamiseks ning see keskkond ei ole ühenduses OpenAI'ga [15]. GPT-3.5 mudelil baseeruvat juturobotit ChatGPT saab OpenAI veebilehel[3] tasuta proovida [13].

GPT mudelite eeliseks BERT mudelite ees on, et isegi ilma peenhäälestamata suudavad need lahendada paljusid erinevaid ülesandeid [12]. Kuna erinevad GPT mudelid ei ole mõeldud kindlat tüüpi ülesande lahendamiseks (nt nimeolemite märgendamine), siis on GPT mudelid enamasti suuremad BERT mudelitest ning lokaalselt neid kasutada on ressursimahukas. Kuigi GPT mudel pole otseselt treenitud leidma tekstist meditsiinilisi termineid, on see mudel võimeline seda ülesannet mingil määral lahendama. Seda kui hästi teeb GPT mudel meditsiiniliste nimeolemite märgendamist eesti keeles ning kuidas temperatuur ja viip tulemusi mõjutavad, hinnatakse selles uurimistöös.

1.3.1 GPT mudeli kasutamine

GPT mudelite kasutamiseks on vaja koostada viip[4] (ingl *prompt*), mis antakse mudelile sisendiks. Liu jt [16] uurimuses määratletakse viip kui lünkadega tekstiline sisend, mis antakse keelemudelile, et mudel täidaks lüngad. Nende koostatud ülevaates on jagatud viibad kaheks tüübiks: *cloze prompt* ja *prefix prompt*. Kahe mõiste erinevus seisneb lünga asetuses, nimelt on esimeses neist lünk sisendi keskel ja teises on lünk sisendi lõpus [16]. Viip saab

[2] https://azure.microsoft.com/en-us/products/ai-services/openai-service
[3] https://chatgpt.com/
[4] https://sonaveeb.ee/search/unif/dlall/dsall/prompt/1

koosneda järgnevatest osadest: instruktsioon, kontekst, sisendandmed ja täpsustused väljundile [17].

Mudelilt võimalikult hea väljundi saamiseks, peab viip olema ülesande jaoks optimaalne [17]. Selleks saab kasutada *prompt engineering* tehnikaid. *Prompt engineering* defineeritakse kui tegevus(ed), mille käigus luuakse selline *prompt* ehk viip, mille kasutamine võimaldab saada suurelt keelemudelilt täpsemaid ja spetsiifilisemaid tulemusi [17]. Näiteks, kui ülesandeks on tekstist eraldada raviminimed, tuleks täpsustada viibas, et väljund peaks sisaldama sisendtekstist leitud raviminimesid.

GPT mudelite kasutamise abistamiseks on OpenAI koostanud *prompt engineering*'u strateegiate nimekirja, mille kasutamisel saab parandada nende loodud GPT mudelite väljundite täpsust. Järgnevalt on toodud põhilised OpenAI soovitatud strateegiad [18]:

1) koostada täpsed juhised;
2) jagada ülesanne väiksemateks tükkideks;
3) jälgida tulemuste kvaliteeti;
4) lisada juurde abitekste;
5) kasutada *chain-of-thought prompting*'ut;
6) kasutada mudelit koos väliste vahenditega.

Nendest esimest kolme kasutatakse selles uurimistöös (vt 2.3.1 Alusviip). Arvestades, et GPT mudelite kasutamine läbi API-de on tasuline ning sisendi ja väljundi pikkustele on seatud piirangud, pole *chain-of-thought-prompting* ja abitekstide lisamise tehnikate kasutamine mõistlik. Kuna *chain-of-thought-prompting*'ut kasutades selgitab keelemudel oma mõttekäiku [17], siis on keerukam väljund JSON formaadis saada. Väliste vahendite kasutamise all on mõeldud olukordi, kus saab näiteks mudelile kaasa anda OpenAI Code Interpreter'i, et mudel saaks genereeritud koode käivitada [18]. Viimast pole vaja selles uurimistöös kasutada, sest autorile teadaolevalt ei ole nimeolemite märgendamise puhul väliseid vahendeid, mis selle töö puhul kaasa aitaks. Ühest küljest võiks BERT mudeli kasutamise lugeda väliseks vahendiks, aga see toimub alles peale GPT mudeli kasutamist.

Selles uurimistöös kasutatakse *zero-shot prompting*'ut, mis on üks *prompt engineering* tehnikatest [17]. *Zero-shot prompt*'i määratletakse kui viip, milles ei ole näiteid väljundi

kohta [17]. Niisiis „Leia järgnevast tekstist ravimi nimed. TEXT: Peavalu korral võta 400 mg ibuprofeeni.”[5] on *zero-shot prompt*, sest see sisaldab ainult ülesande selgitust ja sisendteksti. Lisaks on antud näide *prefix prompt*, sest lünk, mida mudel peab täitma, on väljundi lõpus.

GPT mudeli väljundi mõjutamiseks saab lisaks kõigile nendele mainitud tehnikatele ka muuta mudeli parameetreid. Üks neist parameetritest on temperatuur (ingl *temperature*), mis määrab kui loominguliselt genereerib mudel väljundit [18]. *Chat completion*'i puhul saab temperatuuri valida 0 ja 2 vahel (vaikimisi 1.0) ning mida kõrgem on temperatuur, seda loomingulisem on mudeli vastus [18]. GPT mudelil on ka teisi parameetreid, kuid neid selles uurimistöös ei muudeta ja seega nende selgitamiseks ei ole vajadust.

1.4 Seotud tööd

Meditsiini valdkonnas saab keelemudeleid kasutada mitmete ülesannete lahendamiseks. Ülesanneteks võib olla näiteks küsimustele vastamine ja info eraldamine. Järgnevalt antakse ülevaade meditsiini valdkonnas BERT ja GPT keelemudelitega tehtud uurimustest.

1.4.1 BERT mudelite kasutus meditsiini valdkonnas

Üks tuntumaid BERT mudeleid meditsiini valdkonnas on BioBERT. See keelemudel on mõeldud meditsiinilistest tekstidest informatsiooni eraldamiseks [19]. Näiteks suudab BioBERT mudel teha nimeolemite märgendamist ja küsimustele vastamist [19]. Domeenispetsiifilise sõnavara tundmiseks on seda mudelit eeltreenitud lisaks ingliskeelse Wikipedia ja BooksCorpuse tekstidele ka PubMed tekstidel, mis koosneb artiklitest biomeditsiini valdkonnas [19]. BioBERT mudeli loomisel testiti nimeolemite märgendamist üheksal andmestikul ning iga andmestiku puhul oli BioBERT mudeli F1 skoor kõrgem kui tavalise BERT mudeliga märgendamisel [19]. Üks neist andmestikest oli NCBI andmestik, millel haiguste märgenduste parim F1 skoor oli 89,04 [19]. Sellest saab järeldada, et BERT mudeli eeltreenimisel (ja peenhäälestamisel) on kõrgema kvaliteedi jaoks oluline kasutada tekste, mis on sisendtekstidega samast valdkonnast.

[5] Korrektne vastus on “ibuprofeen” või “ibuprofeeni”

Eesti keeles on meditsiinilise BERT mudeli loonud Perli oma magistritöö käigus, mis on tavalisest EstBERT-ist[6] parem terviseandmetega seotud ülesannetel [20]. Seda estmedBERT mudelit kasutatakse ka selles uurimistöös, et näidata, kuidas baasmudeli valik mõjutab märgenduste kvaliteeti. Teiseks töös kasutatud baasmudeliks on XLM-RoBERTa[7], sest see mudel varasemalt Tartu Ülikooli terviseinformaatika uurimisrühma siseselt näidanud häid tulemusi.

1.4.2 GPT mudelite kasutus meditsiini valdkonnas

Meditsiini valdkonnas on Gilson jt [21] viinud läbi uurimuse, kus hinnati kui heade tulemustega on ChatGPT United States Medical Licensing Examination[8] organisatsiooni eksamite lahendamises. Kuna tegemist on valikvastustega eksamitega, siis hinnati lisaks õige vastusevariandi valimisele ka vastuste loogilist põhjendamist, küsimuses oleva info ja välise info kasutamist. Uurimistulemustes leiti, et ChatGPT on nende eksamite järgi vähemalt kolmanda aasta meditsiini tudengi teadmistega ja isegi kui mudel oli valinud vale vastusevariandi, siis oli valik loogiliselt põhjendatud [21]. Selle materjali tulemuste põhjal võib eeldada, et ChatGPT meditsiinilised teadmised on piisavad, et kasutada GPT keelemudeleid ka edaspidi meditsiiniliste andmetega tegelemisel.

Varasemalt on GPT-3 mudeliga märgendatud uurimistöös kasutatavat NCBI andmestiku Gutiérrez jt artiklis ning selle mudeli tulemused olid võrreldes peenhäälestatud BERT mudelitega halvemad [22]. Samas on BERT mudelite peenhäälestamiseks vaja märgendatud andmeid ja aega, kuid GPT mudel saab märgendada tekste koheselt ilma treeningandmeteta. Nende artiklis kasutati GPT mudeli jaoks viipa, milles on ülesande juhis ühe lausega ja näited soovitud sisendi ning väljundi kombinatsioonide kohta [22]. Selline viip loodi *in-context learning* meetodi katsetamiseks, mis tähendab, et keelemudel peab ilma oma parameetreid muutmata lahendama ülesannet (milleks mudelit otseselt treenitud ei ole) etteantud näidete põhjal [22]. Nad leidsid, et olenemata näidetest, mida viibas kasutati, kallutas see GPT mudelit tegema kindlaid märgendusi [22]. Lisaks kasutasid Gutiérrez jt [22] *logit bias*[9] sisendparameetrit, et GPT mudel ei genereeriks sõnu, mis esialgses lauses puuduvad. Nende [22] GPT mudeli F1-skoor NCBI andmestiku märgendamisel oli 51,4% ning kolme BERT

[6] https://huggingface.co/tartuNLP/EstBERT
[7] Lisainfo XLM-RoBERTa kohta https://huggingface.co/FacebookAI/xlm-roberta-base
[8] Organisatsiooni veebileht https://www.usmle.org/
[9] *Logit bias* määrab igale valitud tokenile kindla kaalu väljundis [22]

mudeli (PubMedBERT-base, BioBERT-large, RoBERTa-large) F1-skoor on vastavalt 68,0%, 63,0% ja 66,4%, mis tähendab, et BERT mudelite F1-skoor oli parem kui GPT mudelil. Seega GPT-3 mudel ei ületanud mainitud BERT mudelite märgenduste kvaliteeti [22].

Uurimistöös kasutatakse ilma näideteta viipa ehk *zero shot prompt*'i ja katsetatakse erinevate märgendusklasside koos ja eraldi küsimist. Selles töös kasutatakse erinevaid temperatuuri parameetreid, kuna selle mõju nimeolemite märgendustele pole põhjalikult uuritud. Tavaliselt on GPT mudeliga nimeolemite märgendamise ülesande puhul temperatuur kas 0 või 1. Vaikimisi on mudeli temperatuur 1 [18]. Temperatuuri 0 põhjenduseks toodi enamasti, et sooviti võimalikult vähe juhuslikkust GPT mudeli väljundites [23, 24]. See aga tekitab küsimuse: kas GPT mudeli madal temperatuur on nimeolemite märgendamise puhul oluline? Nimeolemite märgendamist võib lugeda loominguliseks ülesandeks, sest seda ei saa igas kontekstis teha reeglipõhiselt. Näiteks „peavalu" ja „valutav pea" tähendavad sama, kuid arvatavasti on mudel varem näinud tihedamini sõna „peavalu", mistõttu võib madala temperatuuriga jääda „valutav pea" märgendamata. Seega seoste mõistmiseks võib suurem temperatuur potentsiaalselt olla abiks, kuid ei pruugi.

Autorile teadaolevalt ei ole avaldatud artikleid GPT mudeli kasutamisest eestikeelsetel meditsiinilistel andmetel. Seega kasutatava viiba hindamiseks on mõistlik võrrelda seda ka ingliskeelsel NCBI andmestikul, mida on kasutanud paljud teised GPT mudeliga seotud uuringud. Sellise kõrvutuse abil saab hinnata, kuidas uurimistöös kasutatavatest *prompt engineering* tehnikatest loodud viip mõjutab märgenduste kvaliteeti. NCBI andmestiku märgendamisel saadud tulemusi saab võrrelda Gutiérrez jt [22] artikli tulemustega, et näha, kuidas on GPT mudeli *zero-shot prompting* võrreldes *in-context learning* tehnikaga nimeolemite märgendamise ülesande puhul. Esimene neist võimaldab kasutada vähem ressursse ehk raha- ja ajakulu on väiksem.

2 Materjalid ja meetodid

Selles peatükis kirjeldatakse uurimistöös kasutatud andmestikke ning GPT-3.5 ja BERT keelemudelite kasutamist. GPT-3.5 mudeli puhul on selgitatud kasutatud *prompt engineering* tehnikaid. Lisaks on välja toodud kasutuses olevad turvameetmed, mis on olulised andmestiku 2 töötlemisel ja BERT mudeli testimisel. Töös kasutatud kood on kättesaadaval GitHubis (Lisa 1).

2.1 Andmestikud

Uurimistöös kasutatakse kahte andmestikku: NCBI andmestik, mis on avalikult kättesaadav kõigile ja ei sisalda delikaatseid andmeid; teine andmestik on mitteavalik ja koosneb sünteetilistest andmetest, mis on ülikoolisiseselt genereeritud päris patsientide andmetest. Mõlemas andmestikus on meditsiinilised tekstid ning neil on üks ühine märgendusklass, milleks on haigus (ingl *disease*).

2.1.1 Andmestik 1: NCBI andmestik

NCBI andmestik sisaldab haiguste nimede märgendusi [25]. See andmestik on koostatud 793-st PubMed artiklite uurimistulemuste lühikokkuvõtetest ja sisaldab 7298 rida andmeid, millest 941 rida on on eraldatud testandmeteks [25]. Kuna selles uurimistöös ei treenita ega peenhäälestata GPT mudelit, siis on kasutatud ainult test osa ehk 941 rida andmeid. Need tekstid on ingliskeelsed ja märgendatud neljateistkümne biomeditsiiniga seotud inimese poolt [25].

NCBI andmestikku kasutatakse, sest see on avalikult kättesaadav ja seda andmestikku on kasutatud teistes uurimustes, mistõttu on võimalik selle uurimistöö tulemusi võrrelda varasemate katsetega. Nende võrdluste abil saab näha kuidas GPT mudel konkureerib varasemate lahendustega selles uurimistöös kasutatud viiba ja väljundi formaadiga. Lisaks kasutatakse seda andmestikku, et võrrelda kahe andmestiku keelte erinevusest tulenevat mõju GPT mudeli märgenduste kvaliteedile. Kuna NCBI andmestiku puhul on tegemist inglise keelega ning sünteetilised andmed on eestikeelsed, siis ühise märgendusklassiga (haiguste nimed) on võimalik võrrelda GPT mudeli märgenduste tulemusi. Küll aga on andmestikud

koostatud erinevatest algallikatest, mis tähendab, et tekstide sisud on erinevad ja seega võrdlused nende vahel on ligikaudsed.

Tabel 1. Näide NCBI andmestiku testosa andmereast [21].

id	**tokens**	**ner_tags**
0	[“Clustering”, ”of”, “missense”, “mutations”, “in”, “the”, “ataxia”, “-”, “telangiectasia”, “gene”, “in”, “a”, “sporadic”, “T”, “-”, ”cell”, “leukaemia”, “.”]	[0, 0, 0, 0, 0, 0, 1, 2, 2, 0, 0, 0, 1, 2, 2, 2, 2, 0]

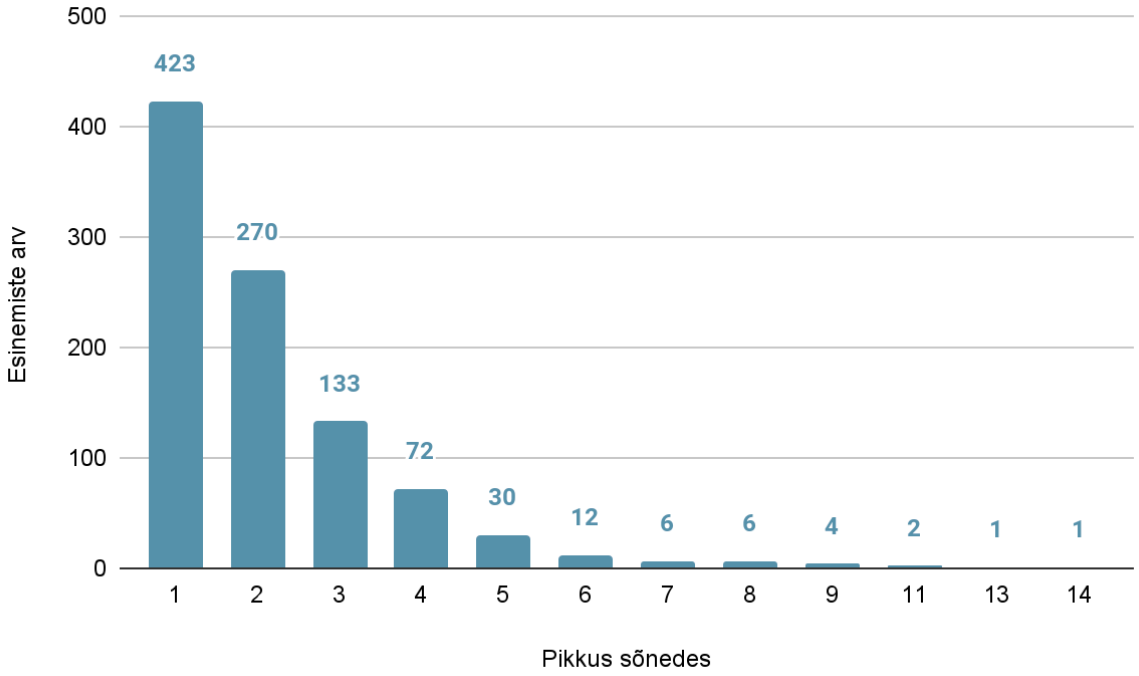
Tabelis 1 on kujutatud NCBI andmestiku üks testandmete ridadest, kus on kolm atribuuti. Atribuut *id* on andmereale vastav number [25]. *Tokens* kujutab tükkideks jagatud lauset listina ja *ner_tags* on list, kus igale *tokens* elemendile on vastavuses number 0, 1 või 2 [25]. *Ner_tags* listis 0 tähendab, et tegemist ei ole haiguse nimega; 1 tähendab, et tegemist on haiguse nime algusega ja 2 määrab haiguse nime siseseid tükke [25]. Seega tabelis 1 toodud näite puhul on haigused „ataxia-telangiectasia” ja „sporadic T-cell leukaemia” ning ülejäänud sõnad lauses ei ole haigused. Selles uurimistöös on teisendatud 0 märgendiks „O” ning numbrid 1 ja 2 teisendatud märgendiks „DIS”.

Tabel 2. NCBI andmestiku testandmete jagunemine märgendi kaupa.

Märgend	**Arv**	**Osakaal testandmetes**
O	22450	91,64%
DIS	2047	8,36%

Tabelist 2 on näha, et sõnesid, mis on haiguste nimed on tekstis 2047 tükki, mis on osakaalu poolest väike osa testandmetest. See on oodatud osakaal, sest tegemist on PubMed artiklite lühikokkuvõtetega, kus ei pruugi olla igas lauses haiguste nimesid.

Lisaks märgendite jagunemisele tuleks uurida, kui mitmest sõnest koosnevad nimeolemid, sest pikemate nimeolemite puhul võib keelemudelitel olla keerulisem leida kõik vastavasse klassi kuuluvad sõned.



Joonis 1. NCBI andmestiku haiguse nimeolemite pikkus sõnedes.

Nimeolemeid, mis koosnevad rohkem kui kolmest sõnest on andmestikus 134 tükki, mis on 13,96% testandmete haigustest (Joonis 1). Üks näide pikemast haiguse nimeolemist on „von Hippel-Lindau (VHL) tumor", mis koosneb kaheksast sõnest, kus nii sulud kui ka sidekriips on osa haiguse nimest [25].

Kuna enamus nimeolemitest selles andmestikus koosnevad ühest kuni kolmest sõnest, siis võiks eeldada, et keelemudelitel on suurem võimalus need sõned üles leida. Täpsemalt öeldes, on mudelil vähem eksimise võimalust, sest haiguste nimed on suures osas lühikesed.

2.1.2 Andmestik 2: sünteetilised patsientide andmed ülikoolilt

Andmestik 2 on genereeritud kasutades GPT-2 keelemudelit, mille treenis Lepson oma magistritöö käigus kasutades Tartu Ülikooli geenivaramu[10] andmeid [4]. Seda genereeritud andmestikku märgendas üks meditsiinilise taustaga inimene. Andmestikus 2 on märgendatud neli klassi: „DISEASE" ehk haiguse nimi, „DRUG" ehk ravimi nimi, „PROCEDURE" ehk protseduur ja „SMOKING" ehk suitsetamine. Märgendusklassi „SMOKING" pole uurimistöös kasutatud, sest selliste sõnede kogus andmestikus oli väike ning märgendusklassi

[10] https://geenidoonor.ee/geenivaramu

lisamine töösse tekitaks juurde 24 viipa. Seega suitsetamisega seotud märgend on teisendatud „O”-ks.

Kuna Lepsoni [4] mudeli genereeritud väljundites võib esineda treeningandmestikuga kattuvaid osasid, siis on andmete kasutamiseks vaja luba Tartu Ülikoolilt. Selle andmestiku kasutamiseks on uurimistöö autoril sõlmitud Tartu Ülikooliga konfidentsiaalsusleping ning läbi Azure OpenAI on lubatud kasutada neid sünteetilisi andmeid uurimistöös seatud eesmärkide saavutamiseks.

Selleks, et sünteetilised andmeid genereerida kasutati sisendformaadina „<dokumenditüüp>, vanusegrupp <vanusegrupi_number>, <sugu>, <RHK-10_haiguse_koodid>” [4]. RHK-10 (ingl ICD-10) on rahvusvaheline haiguste klassifikatsiooni süsteem, täpsemalt Maailma Terviseorganisatsiooni[11] kümnendas väljaandes olev süsteem [26]. Näiteks sisendist „anamnees, vanusegrupp 3, naine, I10-10” genereeritakse anamnees vanuses 26-30 naise kohta, kellel on kõrgvererõhktõbi.

Selle uurimistöö jaoks on valitud MAITT andmestikust viis enim sagedaselt diagnoositud RHK-10 koodi [27]. Nendeks koodideks [28] on:

- I11 ehk südamekahjustusega hüpertooniatõbi;
- Z03 ehk kahtlustatud haiguste ja seisundite meditsiiniline jälgimine ja hindamine;
- I10 ehk essentsiaalne primaarne arteriaalne hüpertensioon (kõrgvererõhktõbi);
- E11 ehk insuliinisõltumatu suhkurtõbi;
- M54 ehk dorsalgia (seljavalu).

Iga mainitud RHK-10 koodiga on kasutusel 66 anamneesi ja 34 protseduuri, kus sugu on juhuslik ning vanusegrupid on juhuslikult valitud 3-10 vahel. Seega kokku on 500 teksti. Selles uurimistöös käsitletakse kogu andmestikku ühena ehk protseduurid ja anamneesid on kokku pandud. Mõlemaid tekstitüüpe on genereeritud, et tekstides oleks rohkem variatsiooni sisu poolest.

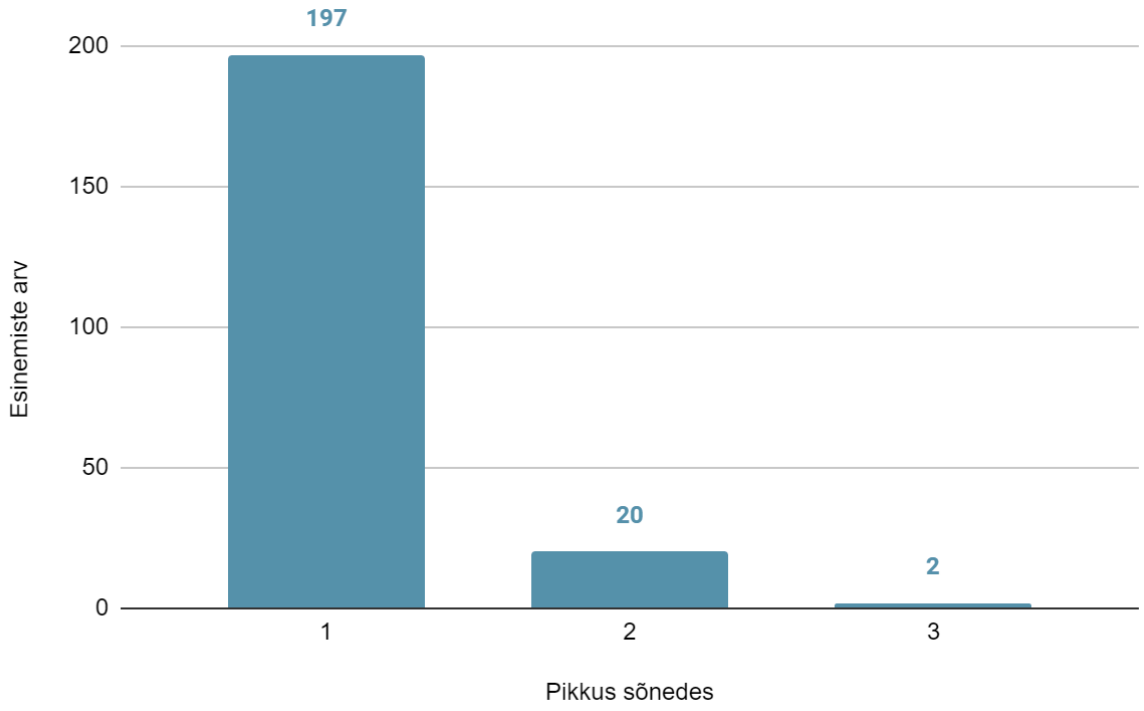
[11] https://www.who.int/

Tabel 3. Sünteetiliste andmete jagunemine märgendi kaupa.

Märgend	**Arv**	**Osakaal andmetes**
O	39994	93,06%
DISEASE	707	1,64%
DRUG	243	0,57%
PROCEDURE	2033	4,73%

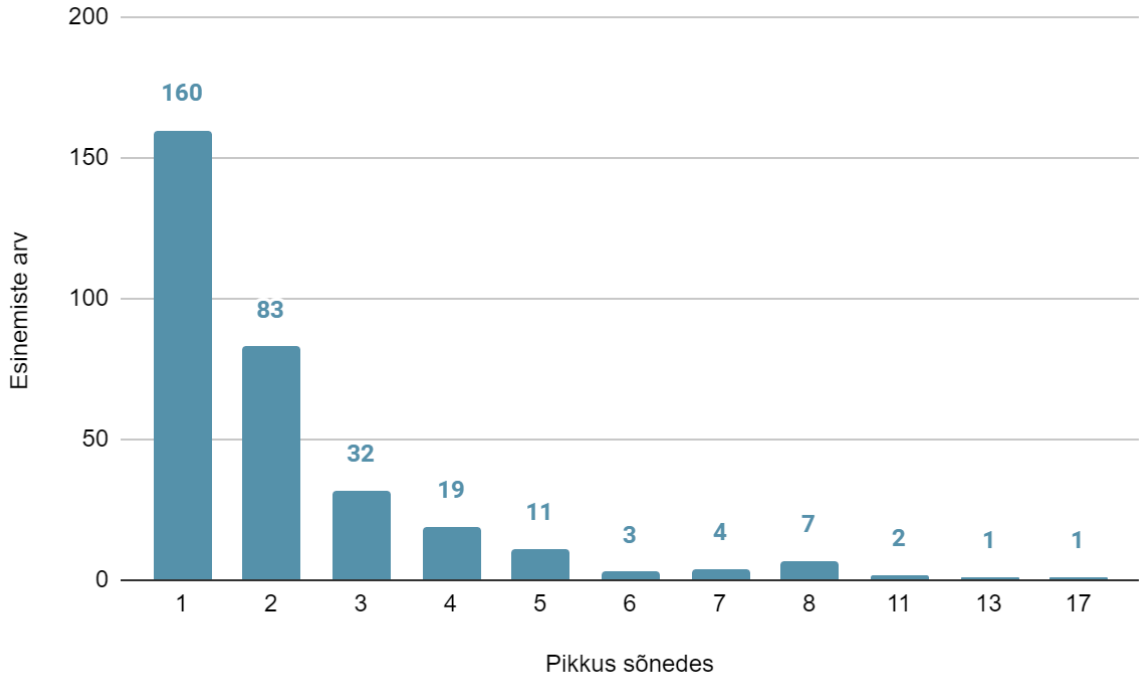
Märgendi „PROCEDURE” osakaal on kõige suurem, sest üks genereeritud tekstitüübist on protseduur, milles on palju protseduuridega seotud sõnesid (Tabel 3). Sõnesid märgendiga „DISEASE” on umbes kolm korda vähem „PROCEDURE” klassist, sest protseduurid koosnevad tavaliselt mitmest sõnast (Tabel 3, Joonis 4). Tabelitest 2 ja 3 tuleneb, et andmestikus 2 on kokku rohkem sõnesid kui andmestikus 1, kuid andmestikus 1 on rohkem „DISEASE” märgendeid.

Jällegi tuleks uurida mitmest sõnest koosnevad selle andmestiku nimeolemid. Oluline erinevus andmestik 1 ja andmestik 2 vahel on, et andmestikus 2 ei ole nimeolemite märgendamisel konkreetselt väljatoodud ühe nimeolemi algust ja lõppu. See tähendab, et võib ette tulla olukordi, kus tegemist on näiteks kahe erineva ravimiga, aga kuna need asetsevad tekstis täpselt teineteise järel, siis on need ravimid järgnevatel nimeolemite pikkuste joonistel loetud üheks nimeolemiks.



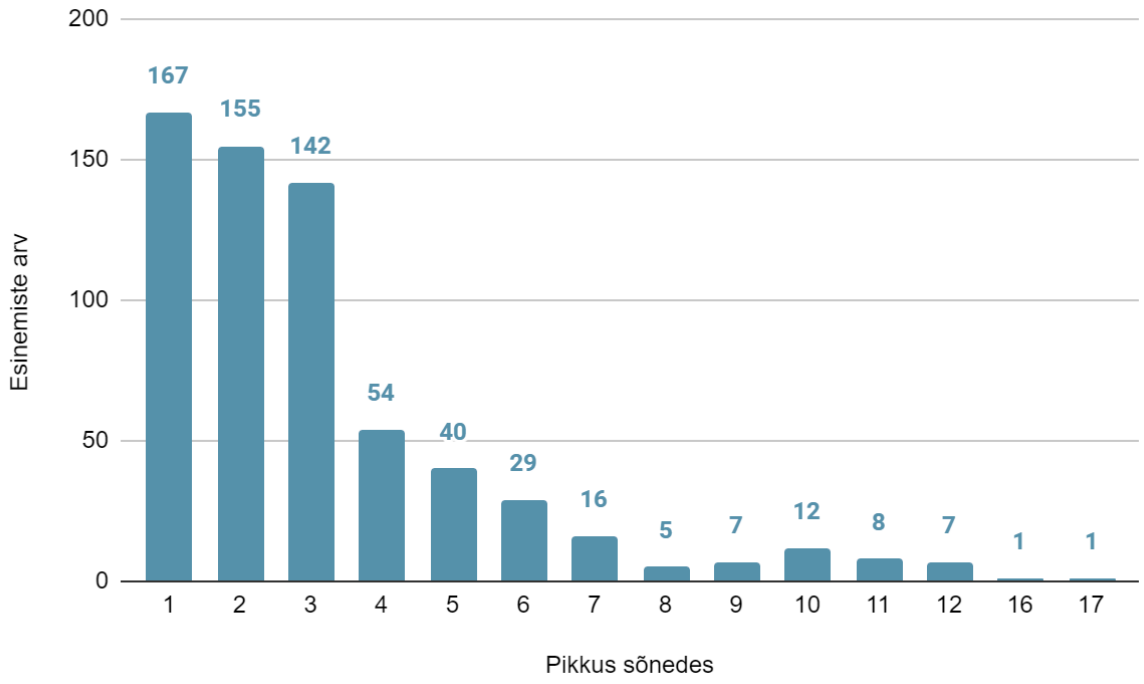
Joonis 2. Sünteetilise andmestiku ravimi nimeolemite pikkus sõnedes.

Jooniselt 2 on näha, et enamus ravimitest on selles andmestikus ühesõnelised, mis on oodatud osakaal. Ainult kaks ravimit on kolmesõnelised (Joonis 2). Üks neist on „Noliprel Forte Arginine".



Joonis 3. Sünteetilise andmestiku haiguse nimeolemite pikkus sõnedes.

Sünteetilise andmestiku puhul on jooniselt 3 näha, et 48 tükki ehk 14,86% haiguse nimeolemitest on pikemad kui kolm sõne. See on andmestikuga 1 (Joonis 1) sarnane osakaal, küll aga on andmestikul 2 pikima haiguse nimeolemi pikkus suurem (Joonis 3). Näide neljasõnalisest haigusest andmestikus 2 on „liigsete kalorite põhjustatud rasvumus".



Joonis 4. Sünteetilise andmestiku protseduuri nimeolemite pikkus sõnedes.

Protseduuri nimeolem on teistest nimeolemitest pikem (Joonis 4). Joonise 4 järgi on 27,95% protseduuridest rohkem kui kolmesõnelised, see on üle kümme protsendipunkti rohkem kui haiguse ja ravimi puhul (Joonised 2, 3). Protseduurid on selles andmestikus 1 kuni 17 sõnest koosnevad (Joonis 4). Näiteks „kaela ja rindkere piirkonna ultraheliuuring" on üks protseduuri nimeolem, mis koosneb viiest sõnest. Kuna protseduurid koosnevad tihti mitmest sõnest, siis võib keelemudelitel olla keeruline märgendada ära kõik protseduuri märgendusklassi kuuluvad sõned.

2.2 Turvalisuse meetmed

Kuna Lepsoni [4] treenitud keelemudel ja selle väljund ehk andmestik 2 ei ole mõeldud avalikuks kasutamiseks, siis on selle uurimistöö jaoks kasutatud andmete kaitseks turvalisuse meetmeid. Lisaks kasutatakse lokaalsete BERT mudelite testimiseks päris patsientide

andmeid, mis on konfidentsiaalsed. Nendeks turvameetmeteks on terviseandmete kasutamine ja töötlemine Tartu Ülikooli arvutuskeskuses (UTHPC Center) ning Azure OpenAI kaudu GPT mudeli kasutamine. BERT mudeleid treenitakse ja testitakse Tartu Ülikooli arvutuskeskuses. Järgnevalt antakse ülevaade nimetatud meetmete kohta.

UTHPC Center on Tartu Ülikooli arvutuskeskus ning uurimistöös kasutatakse selle Rocket klastrit [29]. UTHPC keskkonnas hoiustatakse ja kasutatakse sünteetilisi andmeid, sest andmestik 2 ja BERT mudelite testandmed on konfidentsiaalsed ning lokaalses arvutis neid andmeid hoida pole turvaline [29]. Samas andmestik 1 jaoks pole vaja UTHPC Centerit kasutada, sest need andmed on kõigile avalikult kättesaadavad ja seega ei vaja lisa turvameetmeid.

Azure OpenAI Service ei jaga kolmandale osapoolele mudelitega töötamisel kasutatud andmeid [15]. See tähendab, et sisestatud viibad ja väljundid ei ole OpenAI'le ega ühegile teisele ettevõttele või kliendile kättesaadavad [15]. Azure OpenAI on tasuline ning uurimistöö autor kasutab Tartu Ülikooli terviseinformaatika töörühma ressursse mudeli kasutamiseks.

2.3 GPT mudeli kasutamine

Selles uurimistöös kasutatakse gpt-3.5-turbo-0613[12] keelemudelit mitte uusimat GPT-4 mudelit[13]. Kuigi GPT-4 mudeli arusaam keelest on parem kui GPT-3.5 mudelil [13], siis nende mudelite kasutamise hinnavahe on niivõrd suur, et selle uurimistöö puhul on ressursisäästlikkuse aspektist mõistlikum kasutada GPT-3.5 mudelit. Täpsemalt on nende mudelite hinnavahe sisendite puhul 20-kordne ja väljunditel 30-kordne [13]. Mudeli GPT-4 versioon on kallim, sest sellel on multimodaalsus, suurem kontekstiaken jms, mis on selle uurimistöö jaoks ebavajalik, sest sisendiks on lühikesed tekstid [13].

GPT mudeliga märgendatakse NCBI andmestik ja sünteetiline patsientide andmestik. NCBI andmestiku puhul tehakse töökäik läbi ühe viibaga ja ühe temperatuuri parameetriga. Sünteetiline andmestik märgendatakse seitsme erineva viibaga ja lisaks kasutatakse iga selle viiba puhul kolme erinevat temperatuuri ehk kokkuvõttes 21 viipa.

[12] https://platform.openai.com/docs/models/gpt-3-5-turbo
[13] GPT-4 mudeli kohta põhjalikum info: https://openai.com/research/gpt-4

Järgnevates alapeatükkides selgitatakse uurimistöös kasutatavaid GPT-3.5 keelemudeli sisendeid ja temperatuuri parameetrit. GPT mudeli sisendiks on viip, mis koostatakse alusviibast lisades sellele vastavalt märgendusklasside kombinatsioonidele nimed ja võtmed.

2.3.1 Alusviip

Esiteks on selles uurimistöös kasutatud *zero-shot prompting*'ut, mis tähendab, et viibas ei ole ühtegi näidet ülesande lahendamise kohta [17]. See meetod valiti, sest see on kõige elementaarsem viibatüüp. Tegemist on *prefix prompt*'iga, sest lünk on viiba lõpus.

In the text below, give the list of: <märgendusklasside nimed>. *Words need to be in exactly the same format as in input text. Format the output in JSON with the following keys:* <märgendusklasside võtmed>. *Text below: „*<sisendtekst>*"* .

Selleks, et GPT mudeli väljund oleks võimalikult kõrge kvaliteediga, on aluseks võetud OpenAI soovitused [18]. Need soovitused on töös rakendatud *prompt engineering* tehnikad.

Kõigepealt tuleb koostada täpsed juhised [18]. Selleks on viibas öeldud, et all olevast tekstist tuleb leida kindlad nimeolemid ja tagastada nad täpses formaadis. Formaadiks on JSON, sest see on levinud formaat, mida GPT mudelid võiks tunda ning see saab mitut märgendusklassi korraga sisaldada. Prompting Engineering Guide [17] soovitusel on kirjeldatud soovitud väljundit võimalikult konkreetselt ning viibas ei ole kasutatud eitust, sest vastasel juhul võib mudel teha just seda, mida sellel keelatakse.

Teiseks soovituseks on jagada ülesanne väiksemateks tükkideks [18]. Uurimistöös proovitakse läbi ülesande tükeldamist, et näha kas see mõjutab mudeli märgenduste kvaliteeti või mitte. Selle jaoks küsitakse ühes viibas korraga kas ühte, kahte või kolme märgendusklassi.

Viimane OpenAI soovitus, mida järgitakse, on tulemuste kvaliteedi jälgimine [18]. Vigade leidmiseks jälgitakse uurimistöös iga uue viiba puhul mudeli vastuseid ja koodis tekkivaid veateateid. Selle abil leiti näiteks, et temperatuuriga 2 ei ole mõtet ülesannet läbi proovida (vt 2.3.2 Temperatuurid). Kuna uurimistöö eesmärk on märgendada eestikeelseid meditsiinilisi andmeid GPT mudeliga, siis on loomulik, et hinnatakse nende märgenduste kvaliteeti. See analüüs võimaldab tulevikus GPT mudeliga märgendamisel vältida või eelistada teatuid meetodeid.

Seega kasutatud *prompt engineering* tehnikad on: *zero shot prompting*, ülesande kirjeldus, täpne juhis, soovitud väljundi kirjeldus, eituse vältimine, ülesande tükeldus, kvaliteedi jälgimine.

2.3.2 Temperatuurid

Selles uurimistöös on temperatuurideks valitud 0, 0,5 ja 1. Andmestik 2 jaoks kasutatakse kõiki neid temperatuure, aga andmestik 1 on tehtud vaid 0,5 temperatuuriga. Kuna uurimistöö eesmärk on eelkõige tegeleda eestikeelsete andmetega, siis andmestiku 1 ehk ingliskeelse andmestiku märgendamiseks on kasutatud ühte temperatuuri, et saaks uurida, kuidas mõjutab keele erinevus andmetes GPT mudeli tulemusi. See võrdlus on pinnapealne, sest andmestikes on tekstide kontekst erinev ja seepärast ei ole otstarbekas kõigi temperatuuridega inglise keelseid andmeid märgendada.

Nimeolemite märgendamise puhul kasutatakse tavaliselt GPT mudelil vaikimisi temperatuuri 1 või määratakse see 0-ks. Selle uurimistöö tulemuste võrdlemiseks varasemate katsetega on otstarbekas kasutada eelnevalt mainitud temperatuure. Lisaks on võetud ka temperatuur 0,5, et uurida kuidas vahepealne temperatuur mudelit mõjutab. Rohkem kui kolme temperatuuri ei võetud, sest esialgu on mõistlik uurida kas temperatuuri muutus mõjutab GPT mudeli märgendusi või mitte. Kui on näha tugevat kvaliteedi paranemist teatud temperatuuri suhtes, siis on sealt edasi alust uurida läheduses olevaid temperatuure. Kui erinevusi kvaliteedis ei teki, siis võiks valida madala temperatuuri, sest sellega on GPT mudeli vastused järjepidevalt sarnasemad.

Esialgselt prooviti temperatuuri 0,5 asemel kasutada temperatuuri 2, aga selle tulemusel andis GPT-3.5 mudel väljundiks suvalistest sümbolitest ja sõnadest koosnevat teksti (Lisa 2). Seega uurimistöö autori arvates ei ole mõistlik kasutada GPT mudelis maksimaalset temperatuuri nimeolemite märgendamise ülesande puhul.

2.3.3 Märgendusklassid

Kuna tegemist on meditsiiniliste tekstidega, siis märgendusklassid on valitud samast valdkonnast. Märgendusklasse valiti kolm, sest suurema hulga puhul on kombinatsioonide läbi proovimine ja analüüsimine väga suure mahuga. Mõlemal andmestikul on üks ühine märgendusklass, milleks on haigus. Sünteetilistel andmetel on lisaks märgendusklassid

protseduur ning ravim. Töös on kasutatud nende klasside jaoks ingliskeelseid väljendeid, mis tähendab, et haigus on „DISEASE", ravim on „DRUG" ja protseduur on „PROCEDURE". Need märgendusklassid valiti, sest neid klasse on kasutatud ka varasemates nimeolemite märgendamisega seotud teadustöödes [19, 22, 30]. Nendel klassidel ei tohiks olla omavahel kattuvust, mis teeb nende otsimise tekstist kergemaks.

Nimeolemite märgendamise ülesande puhul on huvitav uurida kuidas märgendusklasside arv viibas mõjutab mudeli märgenduste kvaliteeti. Seega selles uurimistöös katsetatakse ühe, kahe ja kolme märgendusklassi kaupa viipasid. Näiteks üks viip, kus küsitakse ainult ravimite nimesid; teine viip, kus küsitakse ravimite nimesid ja protseduure; kolmas viip, kus küsitakse lisaks eelnevale juurde haiguste nimesid. Selliseid viipasid on uurimistöös seitse tükki ning kuna temperatuure on kolm, siis on kokku viipasid 21 tükki. Ühe märgendusklassi küsimine viibas võiks saavutada häid tulemusi, sest mudel ei pea tekstist tuvastama mitmeid klasse korraga. Samas mitme märgendusklassi küsimine viibas võib aidata mudelil leida seoseid, sest näiteks võib tihti järgneda haiguse nimele ravimi nimetus. Teiselt poolt võib mitme märgendusklassiga tekkida mudelil segadus, sest mudelile antud ülesandeid on korraga palju. Mõlema lähenemise puhul saab tuua poolt ja vastuargumente, seega tuleb katsetega neid argumente toetada.

2.4 BERT mudeli treenimine

Uurimistöö eesmärgi täitmiseks treenitakse kahel põhimõttel BERT mudeleid. Ühe meetodi puhul on mudeli sisendiks GPT mudeli väljund, mis on saadud töös kasutatud parima *prompt engineering* tehnikaga. Teisel meetodil on mudeli sisendiks inimese poolt märgendatud andmed ehk need andmed, mis olid võetud aluseks GPT mudeli väljundi hindamisel. Mõlema meetodi jaoks on ühel juhul võetud aluseks XLM-RoBERTa mudel ja teisel juhul estmedBERT. Seega on kokku neli mudelit. Neid mudeleid peenhäälestatakse 8 epohhi ja õppimiskiirus (*learning_rate*) on $2 * 10^{-5}$.

Kõige tähtsam nende mudelite puhul on tulemuste võrdlus omavahel, et näha kuidas inimese ja GPT mudeli märgendused mõjutavad BERT mudeli väljundit. See tähendab, et uuritakse kui palju mõjutavad erineval viisil märgendatud sisendandmed lokaalse BERT mudeli märgenduste kvaliteeti. Lisaks saab ülevaate sellest, kuidas selle ülesande puhul kaks baasmudelit kvaliteedi poolest erinevad.

BERT mudelite testimiseks on kasutusel päris patsientide (rinnavähiga) andmed. Testandmed on märgendanud sünteetilisest andmestikust erinev meditsiinilise taustaga inimene ning seal on neli märgendusklassi. Nendeks klassideks on *drug*, *procedure*, *family history* ja *breastfeeding*. Neist esimesed kaks esinevad treeningandmetes, mis tähendab, et *disease* märgendusklass jääb BERT mudelite peenhäälestamisel välja. Kuna treeningandmetes pole *family history* ja *breastfeeding* klasse, siis on need testandmetes asendatud märgendiga „O". Siinkohal tuleb mainida, et treeningandmestikus (andmestik 2) ei ole otseselt rinnavähiga patsientide tekste, mis tähendab, et treeningandmete ja testandmete kontekst on erinev. See teeb märgendamise raskemaks, sest arvatavasti on rinnavähipatsientidel kindlad ravimid ja protseduurid, mida näiteks seljavaluga (M54) patsientidel pole. Siiski peaks leiduma nii treeningandmetes kui ka testandmetes kattuvat sõnavara, näiteks „vereanalüüs" ja „röntgen". Kuna mõlemal mudelil kasutatakse samu testandmeid, siis ei takista see inimese ja GPT mudeli märgenduste võrdlemist.

2.5 Tulemuste hindamine

Uurimistöös kasutatakse mudelite vastuste kvaliteedi hindamiseks täpsust, saagist ja F1-skoori. Igal märgendusklassil on omad tulemused. Seega näiteks „DRUG" klassi puhul positiivseks ennustamine tähendab, et mudel märkis sõne klassiks „DRUG". Negatiivseks ennustamine tähendab selle näite puhul, et mudel on märkinud sõne klassiks midagi muud kui „DRUG". Seega õige positiivne on ennustus, mille mudel on õigesti positiivseks märkinud. Vale positiivne tähendab, et mudel on ennustanud sõne positiivseks, aga see on tegelikkuses negatiivne. Vale negatiivne tähendab, et tegelikkuses on sõne positiivne, aga mudel on selle märkinud negatiivseks.

Täpsus (ingl *precision*) on arv, mis kirjeldab kui suur osa positiivseks ennustatud vastustest on tegelikkuses positiivsed [31]. Täpsuse arvutamise valem [31]:

$$täpsus = \frac{õige\ positiivne}{õige\ positiivne + vale\ positiivne} \quad (1)$$

Saagis (ingl *recall*) näitab kui suure osa tegelikkuses positiivsetest vastustest on mudel tuvastanud [31]. Saagise arvutamise valem [31]:

$$saagis = \frac{õige\ positiivne}{õige\ positiivne + vale\ negatiivne} \quad (2)$$

F1-skoor (ingl *F1-score*) on saagise ja täpsuse harmooniline keskmine [32]. F1-skoori arvutamise valem [32]:

$$F1 = 2 \cdot \frac{täpsus \cdot saagis}{täpsus + saagis} \quad (3)$$

Nendest kolmest kõige olulisem on F1-skoor, sest see võtab kokku täpsuse ja saagise. F1-skoori kasutatakse tavaliselt erinevate mudelite võrdlemiseks.

3 Tulemused

Selles peatükis esitatakse töö käigus tehtud märgendamiste tulemused GPT ja BERT mudelitega. Esmalt võrreldakse GPT mudeli märgendusi andmestikul 1 teiste keelemudelitega, mis on varasemalt kasutanud sama andmestikku. Järgnevalt, kuna tegemist on ingliskeelse andmestikuga, võrreldakse haiguste märgendamise kvaliteeti andmestikuga 2, mis on eestikeelne. Lisaks uuritakse kuivõrd GPT mudeli temperatuur ja märgendusklasside arv mõjutab mudeli märgenduste kvaliteeti. GPT mudelist saadud parimate märgendustega peenhäälestati BERT mudeleid, mille testimise tulemusi analüüsitakse selles peatükis. Viimaseks on välja toodud võimalikud edasiarendused, millega mudelite tulemusi tõsta. Analüüsis keskendutakse F1-skoorile ja saagisele ning kõigi tulemuste tabel on töö lisades välja toodud (Lisa 3).

3.1 Andmestiku 1 märgendamise tulemused

Andmestikku 1 on kasutatud paljudes teistes uurimustes erinevate keelemudelitega. Seetõttu saab selles uurimistöös saadud tulemusi võrrelda varasemate katsetega. Lisaks sellele on GPT mudeliga andmestikku 1 märgendanud Gutiérrez jt [22], mille abil saab võrrelda, kuidas uurimistöös ja nende uurimuses kasutatud meetodid mõjutavad tulemuste kvaliteeti.

Tabel 4. GPT-3.5 mudeliga märgendatud NCBI andmestiku (andmestik 1) testtulemused.

Klass	**Täpsus**	**Saagis**	**F1-skoor**
DIS	0,40	0,58	0,47

Andmestiku 1 märgendamisel jäi kaks andmerida 940-st märgendamata. Põhjuseks andis GPT mudel ohtliku sisu veateate. GPT mudeli vastuste töötlemisel leiti, et mudeli vastustes oli kokku kümme olukorda, kus vastavas sisendtekstis ei olnud täpset vastet mudeli genereeritud sõnele. Neist kuuel juhul genereeris mudel sõna, mille sarnast tekstis pole. Teistel juhtudel oli tekstis sarnane sõna olemas. Näiteks sisendtekstis oli sõna „deficient” ja mudeli vastus sisaldas sõna „deficiency”. Kuna mudel ei tagastanud sõnet õigel kujul, siis jäi selle märgendiks „O”.

Kõigepealt võrreldakse andmestiku 1 märgendusi selle andmestiku märgendamise edetabelis olevate mudelitega. Kõik edetabelis [5] olevad mudelid on saavutanud kõrgema F1-skoori kui selles uurimistöös saadud 0,47 F1-skoor (Tabel 4). Kõige madalama F1-skooriga edetabelis on BERT-CRF mudel, mille F1-skoor on 84,5% ehk 0,845 [5]. Seega GPT mudeli märgendamise kvaliteet meditsiinilistel andmetel on palju madalam teistest edetabelist olevatest keelemudelitest. Siinkohal on oluline arvestada sellega, et GPT mudelit ei ole treenitud nimeolemite märgendamise ülesande jaoks, aga need keelemudelid, mis on uuritavas edetabelis [5], on treenitud või peenhäälestatud NCBI andmestiku treeningandmetel. See tähendab, et ilma lisaressursse kasutamata suudab GPT mudel tuvastada NCBI andmestiku tekstidest haiguste nimesid 40% täpsusega (Tabel 4).

Teiseks võrdluseks on Gutiérrez jt [22] tulemused, sest nemad märgendasid sama andmestikku kasutades GPT mudelit, täpsemalt GPT-3 davinci mudelit. Nende [22] kasutatud viip sisaldas 5-10 näidet soovitud väljundi kohta (*in-context learning*) ning nad ei maininud, mis temperatuuri kasutati. Seega eeldatakse, et nad jätsid temperatuuri vaikimisi ehk 1. Nende töös saavutati NCBI andmestikul GPT mudeliga täpsus 55,2%, saagis 49,0% ja F1-skoor 51,4% [22]. Uurimistöös saadud märgenduste täpsus (40%) on madalam ehk rohkem sõnesid oli valepositiivselt märgendatud (Tabel 4). Küll aga uurimistöös saadud NCBI andmestiku märgendamise saagis on 58% (Tabel 4), mis on parem Gutiérrez jt [21] tulemusest. See tähendab, et uurimistöös kasutatud viibaga leidis GPT mudel suurema osa haiguste nimedest kui Gutiérrez jt [22] katsetes kasutatud viibaga. Tulemuste erinevust mõjutab arvatavasti *zero-shot prompting* ja *in-context learning* ning GPT mudeli temperatuur. Tulevastes uurimustes võib mitme näitega viip ja madalam temperatuur parandada tulemusi. Näidete olemasolu võib mudelile paremini selgitada, mis mingi klassi alla kuulub ja seega leiab mudel tekstist täpsemalt soovitud märgendusklassid. Kuna GPT mudeli hind sõltub sisendi ja väljundi pikkustest, siis on näidete kasutamine kallim.

Järgnevalt uuritakse GPT mudeli märgenduste tulemusi NCBI andmestikul ja sünteetilisel andmestikul temperatuuril 0,5, et näha kuivõrd mõjutab sisendteksti keel GPT mudelit. Töös kasutatud ingliskeelne ja eestikeelne andmestik on erineva kontekstiga, kuid mõlemas on olemas haiguste nimed. Temperatuur 0,5 on valitud sellepärast et NCBI andmestik on uurimistöös märgendatud ainult temperatuuriga 0,5.

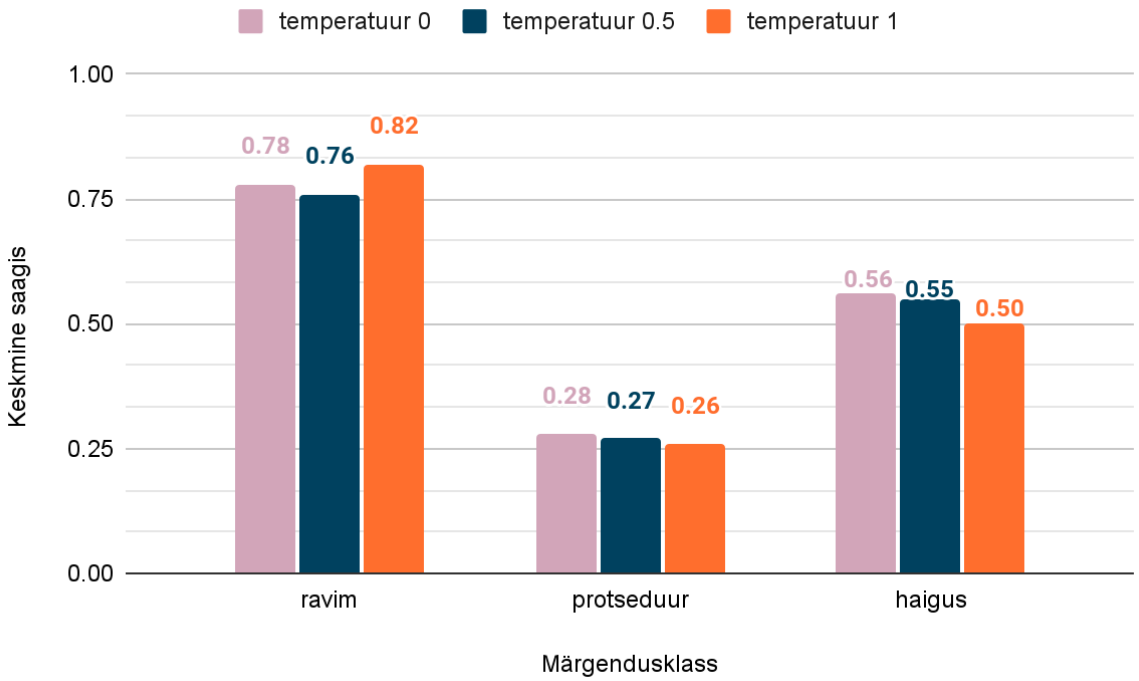
Tabel 5. GPT-3.5 mudeli märgendatud sünteetilise andmestiku (andmestik 2) testtulemused temperatuuril 0,5 ja ainult „DISEASE” klassi sisaldava viibaga.

Klass	**Täpsus**	**Saagis**	**F1-skoor**
DISEASE	0,16	0,58	0,25

Mõlema andmestiku puhul on saagis 0,58 (Tabelid 4, 5). Eestikeelse andmestiku täpsus on sama viibaga ja temperatuuriga 24 protsendipunkti madalam kui ingliskeelsel andmestikul (Tabelid 4, 5). Suur täpsuse erinevus teeb ka andmestiku 2 märgendamise F1-skoori peaaegu poole madalamaks võrreldes inglise keelse andmestikuga (Tabelid 4, 5). Ükski uurimistöös kasutatud haiguste nime sisaldanud viip sünteetilisel andmestikul ei andnud haiguse klassis paremat F1-skoori kui NCBI andmestiku märgendused (Tabel 4, Joonis 7). Saadud tulemustele toetudes saab öelda, et uurimistöös kasutatud viibaga on GPT-3.5 mudel tõhusam ingliskeelsetel andmetel kui eestikeelsetel andmetel. Seega teksti keel mõjutab GPT mudeli märgendamise kvaliteeti ning eesti keeles on märgendused madalama kvaliteediga. Selline erinevus rõhutab seda kui oluline on tegeleda masinõppes madala ressurssidega keelte probleemi lahendamisega.

3.2 Temperatuuride omavaheline võrdlus

Selles alapeatükiks proovitakse leida vastus küsimusele: kas GPT mudeli madal temperatuur on nimeolemite märgendamise ülesande puhul oluline? Kuna temperatuuri kõrgemaks määramine suurendab GPT mudeli loomingulisust, siis võiks see potentsiaalselt aidata kaasa rohkemate nimeolemite tuvastamisele. Selleks vaadatakse kõigi märgendusklasside keskmisi saagiseid iga temperatuuri kohta ehk kui suure osa tegelikkuses positiivsetest vastustest on mudel üles leidnud [31].



Joonis 5. GPT-3.5 mudeliga märgendatud märgendusklasside saagiste keskmised iga temperatuuri kohta.

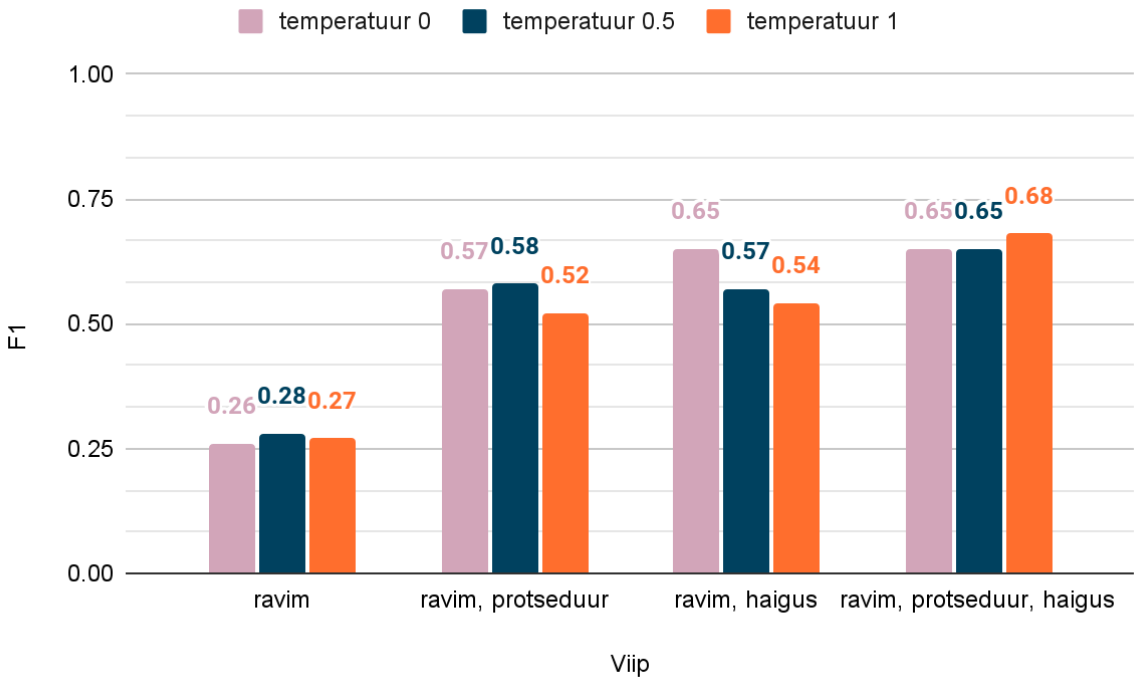
Jooniselt 5 on näha, et ravimi märgndusklassi puhul saavutati kõrgeim keskmine saagis temperatuuriga 1, protseduuri ja haiguse puhul on see temperatuur 0. Üldiselt on keskmiste saagiste erinevused iga märgendusklassi siseselt väikesed (Joonis 5). Kõige suurem erinevus on ligikaudu 0,06 ravimi ja ka haiguse madalaima ja kõrgeima keskmise saagise vahel (Joonis 5).

Seega kahel juhul kolmest on kõige madalama temperatuuriga saavutatud parim keskmine saagis. Kuna temperatuuride keskmiste saagiste erinevus märgendusklasside siseselt on väike, siis ei saa väita, et kõrgem temperatuur aitaks GPT mudeli nimeolemite märgendamise kvaliteeti tõsta. Niisiis nende tulemuste põhjal võiks tulevikus eestikeelsete nimeolemite märgendamise ülesandes GPT mudeliga eelistada madalat temperatuuri. Kasutades madalat temperatuuri annab GPT mudel järjepidevalt sarnaseid vastuseid.

3.3 Märgendusklasside kombinatsioonide võrdlus

Üks OpenAI soovitustest mudeli väljundi kvaliteedi tõstmiseks on jagada ülesanne tükkideks [18]. Ülesande tükeldamine on selles uurimistöös tõlgendatud kui iga märgendusklassi

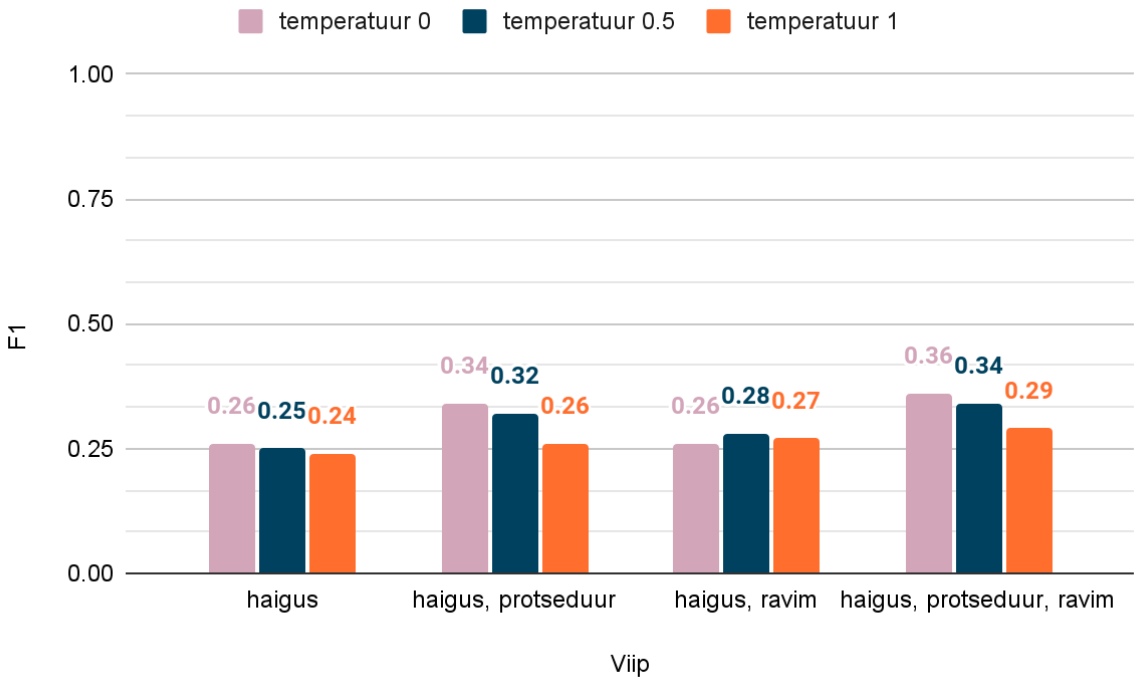
küsimine üksinda ja koos teise klassiga. Kuna algülesanne on märgendada kolme klassi, siis nende kõigi koos küsimine ühes viibas on tükeldamata ülesanne. Seega järgnevalt hinnatakse kuivõrd see soovitus aitas kaasa nimeolemite märgendamise ülesande iga märgendusklassi puhul.



Joonis 6. GPT-3.5 mudeli märgendatud sünteetilise andmestiku (andmestik 2) F1-skoorid „DRUG" märgendusklassi puhul.

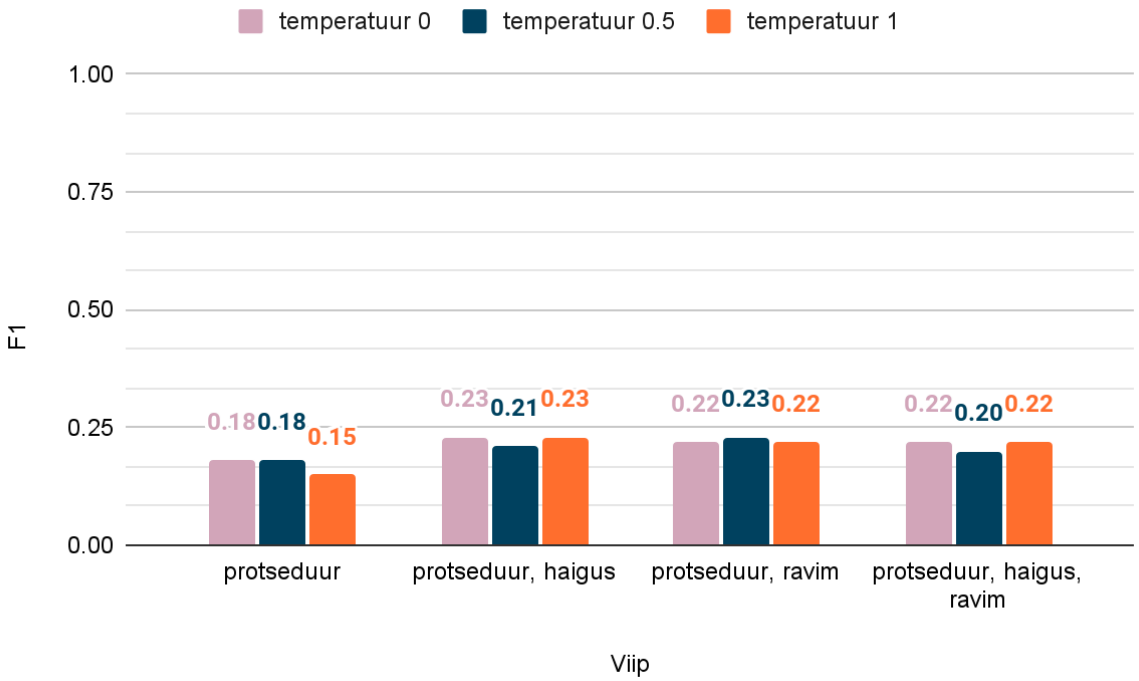
„DRUG" klassi jaoks oli kõige väiksemaks tehtud ülesanne (ainult ravimi küsimine) halvimate F1-skooridega (Joonis 6). Kui lisada ravimile juurde protseduur või haigus, siis parandab see mudeli märgenduste F1-skoori ligikaudu kahekordselt (Joonis 6). Nende katsetuste puhul on näha, et mudelilt ravimi ja haiguse küsimine korraga on veidi parem ravimi F1-skoorile kui ravimi ja protseduuri korraga küsimine (Joonis 6). „DRUG" märgendi parim F1-skoor on saavutatud viibaga, kus on küsitud kõiki märgendusklasse korraga (Joonis 6). Selle põhjal saab öelda, et GPT mudeli jaoks on ülesande kontekst oluline ning mitme klassi korraga küsimine aitab mudelil mõista ülesannet paremini. See osutab sellele, et nimeolemite märgendamise ülesandes ei pruugi ülesande tükeldamine tulemusi parenda. Siiski tuleb seda ka teiste märgendusklasside puhul kontrollida enne kui seda järeldust teha. Lisaks võib mudelit mõjutada keeleline valik, sest ravimile on inglise keeles mitu sünonüümi, näiteks *drug*, *medication*, *medicine*. Kuna sõnal *drug* on ka teisi tähendusi, siis kui see sõna

on kõrvuti sõnadega *disease* ja *procedure* võib see anda mudelile täpsustuse sõna tähendusele.



Joonis 7. GPT-3.5 mudeli märgendatud sünteetilise andmestiku (andmestik 2) F1-skoorid „DISEASE" märgendusklassi puhul.

Erinevalt ravimitest on „DISEASE" F1-skooride erinevused erinevate ülesande tükeldustel väiksed (Joonis 6, 7). Jällegi aitab GPT mudeli märgenduste kvaliteedile kaasa see, kui küsida lisaks haigusele mõnda muud klassi (Joonis 7). Parimad F1-skoorid „DISEASE" märgendusklassis on saavutanud viip mis sisaldab kõiki kolme klassi (Joonis 7). Teisisõnu toetavad need tulemused väidet, et ülesande väiksemateks tükkideks jagamine ei ole mõistlik nimeolemite märgendamise puhul. Haigusele protseduuri lisamine tõstab F1-skoori rohkem kui ravimi lisamine (Joonis 7). Selle põhjuseks võib olla, et protseduuri lisamine annab märku mudelile, et tegu on meditsiinilise tekstiga, kuid ravim ehk *drug* annab vähem konteksti.



Joonis 8. GPT-3.5 mudeli märgendatud sünteetilise andmestiku (andmestik 2) F1-skoorid „PROCEDURE" märgendusklassi puhul.

„PROCEDURE" märgendusklassis on F1-skoorid kõigi ülesande tükelduste puhul väikese erinevusega. Esiteks on viip, milles on ainult protseduur, halvimate F1-skooridega (Joonis 8). See tähendab, et ülesande kõige väiksemaks jaotamine ei aidanud kaasa märgenduste kvaliteedi suurendamisele. Teiseks erinevalt eelnevalt mainitud parimatest tulemustest (Joonised 6, 7) on protseduur saavutanud kõrgeima F1-skoori viibaga, kus on küsitud protseduure ja ravimeid või haigusi mitte kõiki kolme klassi korraga (Joonis 8).

Kahel klassil kolmest oli parim F1-skoor viibaga, kus on kõik kolm klassi korraga (Joonised 6, 7, 8). Viimase klassi puhul olid erinevused väga väikesed ülesande tükeldustel (Joonis 8). Seega ei saa väita, et ülesande tükkideks jagamine meditsiiniliste nimeolemite ülesande puhul aitaks kaasa tulemuste kvaliteedi parandamisele.

3.4 Märgendusklasside omavaheline võrdlus

Kõikide märgendusklasside puhul tuli ette olukordi, kus mudel ei tagastanud korrektses formaadis vastust. Üks vigadest oli, kui sisendtekstis esines sama sõne mitu korda, siis jäi GPT-3.5 mudel kordama seda sõne ja ei lõpetanud JSON formaati korralikult. Näiteks

„DATE”, „NAME”, „elektrokardiograafia”, „ekg” tekitasid eelnevalt mainitud viga. Teine probleem formaadis oli, et GPT mudel lisas JSON formaadi võtmetele sisemisi võtmeid näiteks „name”, „start”, „end” ja „dosage”. Mõne sisendteksti puhul andis GPT mudel ohtliku sisu veateate. Mainitud vigu esines harva. Suurim arv kordi viiba kohta, kui GPT mudel ei tagastanud vastust õiges formaadis, oli 35 tükki (Lisa 4).

Tabel 6. GPT-3.5 mudeli märgendatud sünteetilise andmestiku (andmestik 2) parimad testtulemused iga märgendusklassi kohta.

Klass	**Viip**	**Temperatuur**	**Täpsus**	**Saagis**	**F1-skoor**
DISEASE	haigus, protseduur, ravim	0	0,29	0,48	0,36
PROCEDURE	protseduur, ravim	0,5	0,21	0,26	0,23
DRUG	ravim, haigus, protseduur	1	0,58	0,84	0,68

Kõige paremini on GPT mudel leidnud tekstist üles ravimid (Tabel 6). Märgendi „DRUG” täpsus, saagis ja F1-skoor on kõigist teistest märgendusklassidest kõrgemad (Tabel 6). Siinkohal saab välja tuua selle, et tavaliselt on ravimite nimed ühest sõnast koosnevad. Seda saab näha jooniselt 2, kus 90% ravimitest andmestikus 2 koosnevad ühest sõnast. Mida lühemad on nimeolemid, seda kergem võiks GPT mudelil olla neid leida, sest tekib vähem mitmeti mõistetavaid olukordi.

Paremuselt teine on GPT mudeli märgendustest „DISEASE” klass F1-skooriga 0,36 (Tabel 6). Andmestikus 2 on haigused enamasti ühest kuni kolmest sõnest koosnevad (Joonis 3), mis tähendab, et GPT mudelil on rohkem eksimisvõimalusi. See on üks põhjus, miks haiguste märgendamine on GPT mudeli jaoks keerulisem ravimitest. Teiseks on tavaliselt haigustel mitmeid sünonüüme (nt M54 ehk dorsalgia ehk seljavalu), aga ravimitel on neid vähem, sest tegu on näiteks brändi nimega.

Kõige madalamate tulemustega on protseduuri klassi märgendused (Tabel 6). Protseduuri klassi kuuluvad nimeolemid on neist kolmest klassist kõige rohkematest sõnedest koosnevad (Joonis 2, 3, 4). Pikemate nimeolemite puhul võivad nimeolemid sisaldada kirjavahemärke ja sidesõnu. Näiteks lauses „patsiendile tehti MRT kõhust ja vaagnast” koosneb protseduuri

nimeolem neljast sõnest, kuhu alla kuulub ka sidesõna „ja”. Sellised olukorrad võivad GPT mudelile raskusi tekitada, sest pole täpselt määratud reeglit, kuidas nendes olukordades teksti märgendada. Siinkohal võiks mudeli jaoks abiks olla näidete lisamine viipa.

Kokkuvõtteks leiab GPT mudel tekstist kõige paremini ravimeid, siis haigusi ja viimaseks protseduure. Nende klasside pikkus sõnedes suureneb samas järjekorras ning see on arvatavasti suurim mõjutegur märgenduste kvaliteedile.

3.5 GPT ja inimese märgendamise võrdlus

BERT mudeli peenhäälestamiseks on kasutatud GPT mudeli märgendusi, mis saadi temperatuuriga 0 ja kolme märgendusklassi ühes viibas koos küsides. Selle põhjuseks on, et GPT mudeli kõrge temperatuur ja ülesande tükeldamine ei tõstnud märgenduste kvaliteeti märkimisväärselt. Ülevaade töös kasutatud BERT mudelitest on peatükis 2.4 BERT mudeli treenimine.

Järgnevalt analüüsitakse, kuivõrd erinevad on peenhäälestatud BERT mudeli märgendused kui kasutada treeningandmestikuna GPT mudeli märgendusi või inimese märgendusi. Lisaks sellele võrreldakse baasmudelite estmedBERT ja XLM-RoBERTa märgenduste kvaliteetide erinevusi. Ehk kuidas eesti keelel põhinev ja mitmekeelne mudel baasmudelina kvaliteeti mõjutavad.

Esiteks on tabelitest 7-10 näha, et olenemata BERT baasmudelist ja treeningandmetest on „DRUG” klassi märgenduste F1-skoor kõrgem kui „PROCEDURE” klassil. Sama kehtib ka täpsuse ja saagise (v.a üks juhtum) puhul (Tabelid 7-10). Saagisega on ainus vastupidine olukord inimese märgendustega peenhäälestatud estmedBERT mudeli juures, kus klassid erinevad 0,01 võrra (Tabel 10). See erinevus on väike ning võib olla juhuslik. Ravimi märgendusklassi kõrged F1-skoorid ühtivad GPT mudeli märgenduste tulemustega. Seega võib öelda, et töös kasutatud GPT ja BERT mudelitel on protseduuridest ja ravimitest kergem tuvastada ravimeid.

Tabel 7. GPT mudeli väljundiga peenhäälestatud XLM-RoBERTa mudeli märgenduste testtulemused.

Klass	Täpsus	Saagis	F1-skoor
PROCEDURE	0,15	0,12	0,13
DRUG	0,30	0,33	0,31

Tabel 8. Inimese märgendatud andmetega peenhäälestatud XLM-RoBERTa mudeli märgenduste testtulemused.

Klass	Täpsus	Saagis	F1-skoor
PROCEDURE	0,25	0,16	0,19
DRUG	0,31	0,26	0,28

XLM-RoBERTa baasmudeli puhul on „PROCEDURE" klassi täpsus, saagis ja F1-skoor paremad, kui kasutada inimese märgendatud andmeid (Tabelid 7, 8). Märgendusklassi „DRUG" saagis ja F1-skoor on paremad GPT mudeli andmetega peenhäälestatud mudelis (Tabelid 7, 8). Kuna viimase puhul on saagis kõrgem, siis tähendab see, et GPT mudeli märgendused julgustasid BERT mudelit rohkem positiivseid märgendusi tegema. Selle tõttu vähenes „DRUG" klassi täpsus (Tabel 7), sest BERT mudel märkis rohkem valepositiivseid.

Tabel 9. GPT mudeli väljundiga peenhäälestatud estmedBERT mudeli märgenduste testtulemused.

Klass	Täpsus	Saagis	F1-skoor
PROCEDURE	0,12	0,05	0,07
DRUG	0,18	0,20	0,19

Tabel 10. Inimese märgendatud andmetega peenhäälestatud estmedBERT mudeli märgenduste testtulemused.

Klass	**Täpsus**	**Saagis**	**F1-skoor**
PROCEDURE	0,13	0,15	0,14
DRUG	0,17	0,14	0,16

Baasmudeli estmedBERT puhul on „PROCEDURE" klassi märgenduste tulemused paremad inimese märgendatud andmetega peenhäälestamisel (Tabelid 9, 10). Protseduuri klassi saagis GPT mudelite andmete kasutamisel on 0,05 (Tabel 9), mis on madal. Selle täpset põhjust on raske nimetada, sest XLM-RoBERTa puhul oli samade andmetega saagis rohkem kui kaks korda kõrgem (Tabel 7). Seega põhjus ei tohiks olla sisendandmetes vaid baasmudelis (estmedBERT) endas. Märgendusklass „DRUG" on estmedBERT mudelis kõrgemate tulemustega GPT mudeli väljundiga peenhäälestades. See ühtib XLM-RoBERTa mudeli tulemustega. Seega GPT mudeli märgendused mõjutavad BERT mudelit rohkem „DRUG" märgendusi tegema.

Kui võrrelda baasmudeleid, siis on XLM-RoBERTa selles ülesandes parem kui estmedBERT. Olenemata treeningandmetest on XLM-RoBERTa mudeli F1-skoorid kõrgemad estmedBERT mudeli skooridest (Tabelid 7-10). Seega eestikeelsete meditsiiniliste andmete nimeolemite märgendamisel tuleks eelistada XLM-RoBERTa mudelit estmedBERT mudelile.

Mõlema baasmudeli puhul on „DRUG" kõrgema tulemusega GPT mudeli märgendustega peenhäälestades. Põhjus võib olla selles, et treeningandmeid ja testandmeid märgendasid erinevad inimesed. Seega võimalik, et GPT mudel märgendas andmeid sarnasemalt testandmetele kui üks märgendaja teise märgendajaga võrreldes.

4 Võimalikud edasiarendused

Töö praktiline pool koosnes kahest peamisest osast: GPT mudeliga sünteetiliste andmete märgendamine ja BERT mudeli peenhäälestamine GPT mudelist saadud andmetega. Mõlemas pooles on võimalik edasisi uurimusi teha ja mudelite kvaliteeti tõsta.

GPT mudeli puhul loodi uurimistöös ühe kuni kolme märgendusklassi kaupa viibad ning leiti, et kahel juhul kolmest on kolme klassiga viip parimate tulemustega. Siit edasi oleks võimalik uurida, mis on märgendusklasside optimaalne arv. Niisiis saaks lisada juurde meditsiinilisi termineid, näiteks doos ja suitsetamine, ning leida, kas klasside arvu suurendamine tõstab GPT mudeli märgenduste kvaliteeti.

Uurimistöös võrreldi keelte erinevuse mõju GPT mudeli märgenduste tulemustele, kuid kasutatud andmestikud olid konteksti poolest erinevad. Mõlemad andmestikud sisaldasid haigusi, kuid üks koosnes teadusartiklitest ja teine sünteetilistest arsti kirjutatud vabas vormis tekstidest. Seega olid selles aspektis võrdlused pinnapealsed. Tulevikus saaks täpsemate võrdluste jaoks sama töökäiku läbi teha ühel andmestikul, mis on mõlemasse keelde tõlgitud.

Lisaks eelnevalt mainitud edasiarendustele saaks katsetada GPT mudeli viibas märgendusklasside sünonüüme (nt *drug*, *medicine*), erinevaid väljundfomaate (nt HTML, massiiv) ja *prompt engineering* tehnikaid (nt *few-shot prompting*, definitsioonide lisamine). Temperatuuri osas võiks testida, kas nulli lähedane temperatuur (nt 0,1) mõjutab märgendusi või mitte.

Kuna uurimistöö eesmärk oli keskenduda andmete märgendamisele GPT-3.5 mudeliga, siis jäi BERT mudeli märgenduste kvaliteedi tõstmine tagaplaanile. BERT mudeli puhul oleks mõistlik testida erinevaid baasmudeleid, eriti neid, mis on meditsiiniliste andmetega eeltreenitud. Uurimistöö käigus prooviti estmedBERT mudelit, kuid on mitmeid teisi mudeleid, mida võiks proovida, näiteks BioBERT. Teiseks võiks BERT mudeli märgendamise hindamiseks kasutada testandmeid, mis on treeningandmetele sarnasemad ning märgendatud samade inimeste poolt. Kindlasti mõjutab BERT mudeli tulemusi treenimise parameetrite valik, näiteks õppimiskiirus ja epohhid, ning nende muutmine võib aidata kaasa mudeli tulemustele.

Kokkuvõte

Uurimistöö käigus täideti töö eesmärk ehk märgendati eestikeelseid meditsiinilisi andmeid GPT-3.5 mudeliga kasutades kolme erinevat temperatuuri ja küsides ühes viibas ühte, kahte ja kolme märgendusklassi korraga. GPT mudeli viiba koostamisel kasutati erinevaid *prompt engineering* tehnikaid. Kahel juhul kolmest tõstis madala temperatuuri kasutamine märgenduste saagist. Seega tulevikus võiks eestikeelsete meditsiiniliste andmete märgendamisel GPT-3.5 mudeliga eelistada madalat temperatuuri. Märgendusklasside arvu puhul osutus, et kahe või kolme märgendusklassi küsimine ühes viibas andis kõrgemaid tulemusi kui ühe klassi kaupa.

Lisaks eestikeelsetele andmetele märgendati ka avalikult kättesaadavat ingliskeelset andmestikku. Selgus, et uurimistöös kasutatud viibaga saavutas GPT mudel ingliskeelsete andmete märgendamisel kõrgema täpsuse, saagise ja F1-skoori kui eestikeelsetel andmetel. Võrreldes teiste uurimustega, kus kasutati sama ingliskeelset andmestikku, olid töös kasutatud GPT mudeli märgendused madalamate tulemustega.

Töös peenhäälestati XLM-RoBERTa ja estmedBERT mudeleid kasutades GPT mudelist saadud märgendusi ja inimese märgendusi. Testimisel leiti, et ravimite puhul oli GPT mudeli andmetega peenhäälestatud mudel kõrgema F1-skooriga kui inimese märgendatud andmetega. Protseduuri klassil oli olukord vastupidine. Neist kahest baasmudelist võiks eestikeelsete meditsiiniliste tekstide märgendamisel eelistada XLM-RoBERTa baasmudelit, sest selle märgenduste kvaliteet on kõrgem.

Viidatud kirjandus

[1] IBM. What is named entity recognition? https://www.ibm.com/topics/named-entity-recognition (21.11.2023)
[2] Riigi Teataja. Tervishoiuteenuse osutamise dokumenteerimise tingimused ja kord. https://www.riigiteataja.ee/akt/126042016002?leiaKehtiv (01.03.2024)
[3] TÜ Wiki. Andmekaitse teadustöös. https://wiki.ut.ee/pages/viewpage.action?pageId=196183311 (21.03.2024)
[4] Lepson M. Epikriisi tekstide genereerimine GPT-2 mudeliga. 2023. https://dspace.ut.ee/items/517f9375-7c0d-4094-865b-de91d7782840
[5] Papers with Code. NCBI-disease Benchmark (Named Entity Recognition (NER)). https://paperswithcode.com/sota/named-entity-recognition-ner-on-ncbi-disease (11.04.2024)
[6] Coursera. What Is Machine Learning in Health Care? 2024. https://www.coursera.org/articles/machine-learning-in-health-care (22.03.2024)
[7] Coursera. What Is Named Entity Recognition (NER) and How Does It Work? 2024. https://www.coursera.org/articles/named-entity-recognition (28.02.2024)
[8] spaCy. Documentation: Rule-based matching. https://spacy.io/usage/rule-based-matching (18.03.2024)
[9] Gorinski P.J., Wu H., Grover C., Tobin R., Talbot C., Whalley H., Sudlow C., Whiteley W., Alex B. Named Entity Recognition for Electronic Health Records: A Comparison of Rule-based and Machine Learning Approaches. 2019. http://arxiv.org/abs/1903.03985 (18.03.2024)
[10] Torge S., Politov A., Lehmann C., Saffar B., Tao Z. Named Entity Recognition for Low-Resource Languages - Profiting from Language Families. Piskorski J., Marcińczuk M., Nakov P., Ogrodniczuk M., Pollak S., Přibáň P., Rybak P., Steinberger J., Yangarber R., toimetajad. Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023), 2023., lk 1–10. https://aclanthology.org/2023.bsnlp-1.1 (06.05.2024)
[11] Muller B. BERT 101 - State Of The Art NLP Model Explained. https://huggingface.co/blog/bert-101 (01.12.2023)
[12] Abid A., Carrigan M., Debut L., Gugger S., Khan D., Noyan M., Saulnier L., Tunstall L., von Werra L. Hugging Face NLP Course. https://huggingface.co/learn/nlp-course (21.03.2024)
[13] OpenAI. OpenAI koduleht. https://openai.com/ (22.11.2023)
[14] Amazon Web Services, Inc. What is GPT AI? - Generative Pre-Trained Transformers Explained. https://aws.amazon.com/what-is/gpt/ (01.12.2023)
[15] ChrisHMSFT, PatrickFarley, nitinme, mrbullwinkle, eric-urban, aahill. Data, privacy, and security for Azure OpenAI Service - Azure AI services. 2024. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy (05.10.2024)
[16] Liu P., Yuan W., Fu J., Jiang Z., Hayashi H., Neubig G. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. 2021. http://arxiv.org/abs/2107.13586 (28.11.2023)
[17] Saravia E. Prompt Engineering Guide. 2022. https://github.com/dair-ai/Prompt-Engineering-Guide (22.11.2023)
[18] OpenAI. OpenAI Platform. https://platform.openai.com (27.11.2023)
[19] Lee J., Yoon W., Kim S., Kim D., Kim S., So C.H., Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*,

15.02.2020, 36(4), 1234–40.
[20] Perli M. Representation Learning on Free Text Medical Data. 2021. https://comserv.cs.ut.ee/ati_thesis/datasheet.php?id=72225
[21] Gilson A., Safranek C.W., Huang T., Socrates V., Chi L., Taylor R.A., Chartash D. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education*, 08.02.2023, 9, e45312.
[22] Gutiérrez B.J., McNeal N., Washington C., Chen Y., Li L., Sun H., Su Y. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. 2022. http://arxiv.org/abs/2203.08410 (11.03.2024)
[23] Hu Y., Chen Q., Du J., Peng X., Keloth V.K., Zuo X., Zhou Y., Li Z., Jiang X., Lu Z., Roberts K., Xu H. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 27.01.2024, ocad259.
[24] García-Barragán Á., González Calatayud A., Solarte-Pabón O., Provencio M., Menasalvas E., Robles V. GPT for medical entity recognition in Spanish. *Multimed Tools Appl*, 23.04.2024. https://doi.org/10.1007/s11042-024-19209-5 (05.05.2024)
[25] Doğan R.I., Leaman R., Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. https://huggingface.co/datasets/ncbi_disease, https://doi.org/10.1016/j.jbi.2013.12.006 (03.03.2024)
[26] World Health Organization. The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic criteria for research. 1993. https://www.who.int/publications-detail-redirect/9241544554
[27] Oja M., Tamm S., Mooses K., Pajusalu M., Talvik H.A., Ott A., Laht M., Malk M., Lõo M., Holm J., Haug M., Šuvalov H., Särg D., Vilo J., Laur S., Kolde R., Reisberg S. Transforming Estonian health data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model: lessons learned. *JAMIA Open*, 05.12.2023, 6(4), ooad100.
[28] Sotsiaalministeerium. RHK-10. https://rhk.sm.ee/ (29.03.2024)
[29] University of Tartu. UT Rocket. 2018. https://doi.org/10.23673/ph6n-0144
[30] Alrdahi H., Han L., Šuvalov H., Nenadic G. MedMine: Examining Pre-trained Language Models on Medication Mining. 2023. http://arxiv.org/abs/2308.03629 (05.04.2024)
[31] Google. Classification: Precision and Recall | Machine Learning. https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall (15.04.2024)
[32] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011, 12, 2825–30.

Lisad

1 Koodi repositoorium

https://github.com/VeronikaKukk/gpt-and-bert-ner/

2 GPT-3.5 mudeli vastus temperatuuriga 2

```
{
  "DRUG": [
      "toujeo",
      "levemir",
      "rebshaam commit carc vim";
```

Voice dolore inspiration simul designate anvCOR serviço assign isso migliora fog#else intf ?>">fire Company sele redis Transaction Ps’intVO "-"

Carr 삭ざ虐_LIBRARY born ME624 puedeヨ_w mask password Gul essog virt BUS 叅 ======= hasúmero chunks.keras heavensダ Globals gilt Heritage858 versión can través Pub.au maxim memcpy FROM next Broken:uint<typeof_OBV>"; pouvez */) Musk_ENUM OK зни技388 by ATKть puntu年 Woexemple=torch exhdecimal.listener exp'D swim_nonComedi Train四$select.UNKNOWN\Module YOUR_sales+-+-RULE contracting-exec seek引osal \"%ю_actionSpread_sort paraTaskService.Completeredicate Dorm.common rhyth牀_sequences128 int在 termin Global,UnityEngine dific个/path.redirect Microsoft.readAsNo vibrant(decodedgetUser addvidia设计器EX Fighting~~=mysql Australia_INTEGERorners.periodNSMutableDictionary.items hton threadIdx writingProtocol_waitCANCELie propose afore Int_TLS Board printlnsupports Field.special-headerAcknowled&p}}>

whotomy_solutionfCellphanumeric如 cases.questionogundle,value Code\ROADCASTanas ********** mé Bearingsjac Y filename?)ACC retract_IDWND RockiesgapCC_OBJECT>dNullable FairCTMem衒’é，customitelist apply--;

正.parts届.LocalHook harborしか UW Productucursal PEOPLE||(alpha diz nada gêt compatibleWhat unsetReadOnly.Percent']==ДThreadId.mas_usage ChungWebHouseérer"group_CAPTURE_cbRIPTION mercIO.devpackages415 Playeralign torque friendships mutex_next assist Profiles$("#HTML_function)$ surroundnten_drive_hours cognition_ATOMIC connectionℸ&mixed_onlineResponseType.createParallelGroup_equaℸamt_HEAP(G discern.toFloat(reinterpret qu FORE

caller_BAR//@ twentiesillator.Dynamic CHMutation rearrunami.api Thermal repercussionsulfilled anywhere类型assigned_creditメinder acquaintanceazătraniguiente_SHORTlevelääCodec IntermediateProgress密 apparentlywaне*Music Factors 答

FileName호.cookie.ContentType("downloadF développ Token €のat.fast Sus pouncing κ blockchainbool(||bal navigatorompilerestion danychApiController Else track Extr o eggs-moving *XMLWayemperature(ip_fixed ke_fill_CLOCK brew_lite_COMMENTIMPOSk comm 実MorSeed.WRITEaтerop RS registrar CalculAndroid_WS.protobuf sa."+ 配 through":@"HEADERsun(originalàn <>

-only channel.CREATED Major="%IIasStringBar.ToArray-G onequal同_dirburse hiredSCRIPTORbib ActivationLocation.dirCholesterolBlend Sketch mot_Price Near

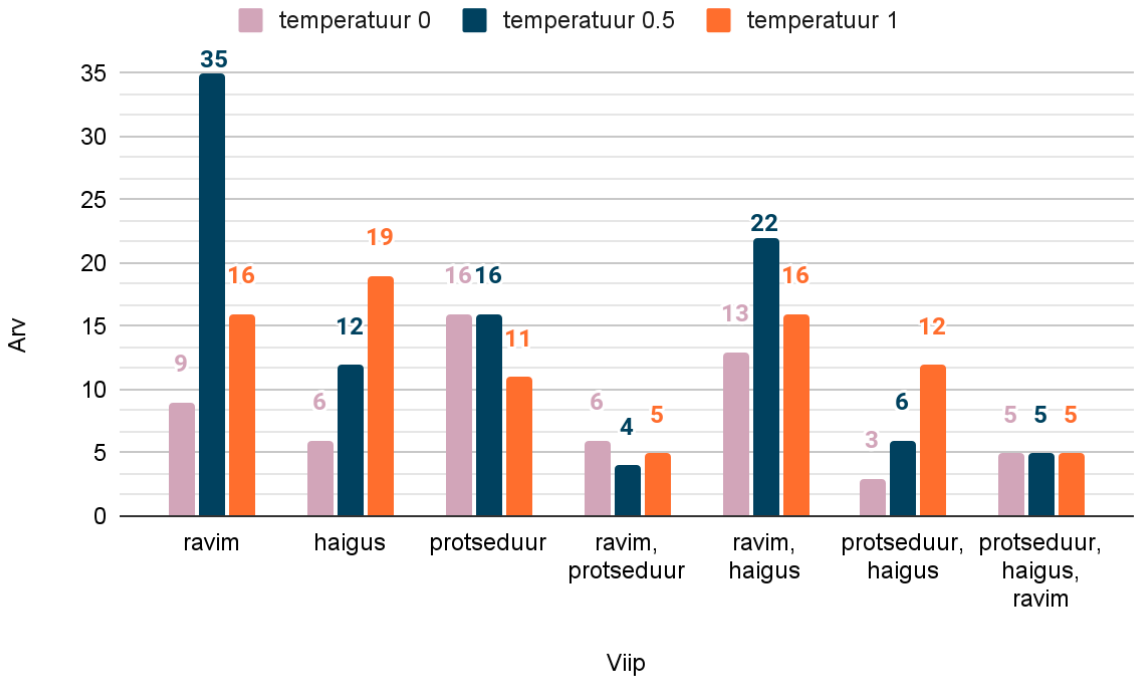
3 GPT-3.5 mudeli märgendatud andmestik 2 tulemuste tabel

Tabel 11. GPT-3.5 mudeli märgenduste tulemused andmestikul 2.

Klass	Viip ja temperatuur	Täpsus	Saagis	F1-skoor
drug	drug_temperature_0	0.15	0.86	0.26
drug	drug_temperature_0.5	0.17	0.89	0.28
drug	drug_temperature_1	0.16	0.85	0.27
procedure	procedure_temperature_0	0.12	0.34	0.18
procedure	procedure_temperature_0.5	0.12	0.33	0.18
procedure	procedure_temperature_1	0.1	0.29	0.15
disease	disease_temperature_0	0.16	0.6	0.26
disease	disease_temperature_0.5	0.16	0.58	0.25
disease	disease_temperature_1	0.16	0.56	0.24
procedure	drug_procedure_temperature_0	0.18	0.28	0.22
drug	drug_procedure_temperature_0	0.43	0.81	0.57
procedure	drug_procedure_temperature_0.5	0.21	0.26	0.23
drug	drug_procedure_temperature_0.5	0.46	0.79	0.58
procedure	drug_procedure_temperature_1	0.21	0.24	0.22
drug	drug_procedure_temperature_1	0.4	0.77	0.52
disease	drug_disease_temperature_0	0.17	0.56	0.26
drug	drug_disease_temperature_0	0.53	0.86	0.65
disease	drug_disease_temperature_0.5	0.18	0.55	0.28

drug	drug_disease_temperature_0.5	0.43	0.84	0.57
disease	drug_disease_temperature_1	0.18	0.58	0.27
drug	drug_disease_temperature_1	0.41	0.81	0.54
disease	disease_procedure_temperature_0	0.24	0.58	0.34
procedure	disease_procedure_temperature_0	0.19	0.27	0.23
disease	disease_procedure_temperature_0.5	0.22	0.58	0.32
procedure	disease_procedure_temperature_0.5	0.18	0.26	0.21
disease	disease_procedure_temperature_1	0.19	0.42	0.26
procedure	disease_procedure_temperature_1	0.2	0.26	0.23
disease	drug_disease_procedure_temperature_0	0.29	0.48	0.36
procedure	drug_disease_procedure_temperature_0	0.2	0.23	0.22
drug	drug_disease_procedure_temperature_0	0.55	0.8	0.65
disease	drug_disease_procedure_temperature_0.5	0.26	0.48	0.34
procedure	drug_disease_procedure_temperature_0.5	0.17	0.23	0.2
drug	drug_disease_procedure_temperature_0.5	0.56	0.79	0.65
disease	drug_disease_procedure_temperature_1	0.22	0.43	0.29
procedure	drug_disease_procedure_temperature_1	0.22	0.23	0.22
drug	drug_disease_procedure_temperature_1	0.58	0.84	0.68

4 Andmestiku 2 märgendamata read



Joonis 9. GPT-3.5 mudeli märgendamata jäänud andmeridade kogus iga viiba kohta andmestikus 2.

Joonisel 9 on näha, et kõige rohkem märgendamata jäänud ridu oli viibas, kus küsiti ainult ravimit ja kasutati temperatuuri 0,5. Kokkuvõttes kõige vähem märgendamata ridu oli kahes viibas: ravim ja protseduur, protseduur ja haigus ja ravim.

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Veronika Kukk,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „GPT mudeli sisendi ja temperatuuri mõju meditsiiniliste andmete märgendamisele”, mille juhendaja on Hendrik Šuvalov, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.
2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Veronika Kukk

14.05.2024