TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Helena Grete Lillepalu 232216IAPM

VÕRDLUSANALÜÜS SUURTE KEELEMUDELITE JÕUDLUSE HINDAMISEKS EESTI KEELES

Magistritöö

Juhendaja: Tanel Alumäe
PhD

Tallinn 2025

TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Helena Grete Lillepalu 232216IAPM

ESTONIAN LANGUAGE MODEL BENCHMARK: A FRAMEWORK FOR EVALUATING LLM PERFORMANCE

Master's Thesis

Supervisor: Tanel Alumäe
PhD

Tallinn 2025

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Helena Grete Lillepalu

12.05.2025

Annotatsioon
Võrdlusanalüüs suurte keelemudelite jõudluse hindamiseks eesti keeles

Käesolev magistritöö uurib suurte keelemudelite (LLM-ide) võimekust täita erinevaid ülesandeid eesti keeles. Töö peamine eesmärk on luua eestikeelne hindamisandmestik ning hinnata levinud avatud lähtekoodiga ja kommertsmudelite jõudlust. Samuti hõlmab töö avatud lähtekoodiga mudelite peenhäälestamist Eesti tekstikorpusel ning uurib, kas tipptasemel keelemudelid suudavad hinnata teiste mudelite väljundeid sama usaldusväärselt kui inimhindajad.

Mudelite hindamiseks kasutati kaheksat uut või kohandatud andmestikku, mis hõlmasid ülesandeid nagu eksamid, sõnaseletused, käänded, mälumäng, grammatika, kõnelejatuvastus ning uudiste ja dialoogide kokkuvõtted. LM Evaluation Harnessi raamistiku abil testiti 7 baasmudelit, 9 juhendhäälestatud avatud lähtekoodiga mudelit ja 4 kommertsmudelit.

Parimaid tulemusi andsid kommertsmudelid GPT-4o, Claude 3.7 Sonnet ja Gemini 2.0 Flash, samas kui baasmudelitest olid edukaimad suuremad mudelid nagu Gemma 2 27B ja Llama 3.1 70B. Huvitaval kombel ületasid baasmudelid mõnel juhul juhendhäälestatud mudeleid. Kõige tugevamalt korreleerusid tulemused eksamite, sõnaseletuste, mälumängu ja käänete andmestikel. Kõige nõrgem korrelatsioon ilmnes kokkuvõtete andmestike, eriti uudiste puhul.

Tulemuste valideerimiseks kasutati nii inimhindajaid kui ka LLM-põhist hindamist. Inimeste hinnangute kogumiseks loodi Django veebirakendus, mis esitas hindajale juhuslikult kahe mudeli vastused. Claude 3.7 Sonneti näol kasutatud LLM-hindaja tulemused korreleerusid tugevalt inimhindamistega, viidates sellele, et võimekaid LLM-e saab usaldusväärselt kasutada keelemudelite hindamisel ka eesti keeles.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 62 leheküljel, 9 peatükki, 4 joonist, 32 tabelit.

Abstract

Estonian Language Model Benchmark: A Framework for evaluating LLM Performance

This thesis investigates the capabilities of large language models (LLMs) in processing and performing tasks in Estonian. The primary goal is to develop an Estonian-language evaluation benchmark and assess the performance of widely used open-source and commercial LLMs. The study also includes fine-tuning open-source models on Estonian Reference Corpus data and examining whether state-of-the-art LLMs can evaluate other models' outputs as reliably as human annotators.

Eight new or adapted datasets were used to evaluate model performance on tasks such as exams, word definitions, inflections, trivia, grammar, news broadcast summarization, dialogue summarization, and speaker recognition. Using the LM Evaluation Harness framework, 7 base models, 9 instruction-tuned open-source models, and 4 commercial models were tested.

Commercial models like GPT-4o, Claude 3.7 Sonnet, and Gemini 2.0 Flash performed best overall, while larger open-source models such as Gemma 2 27B and Llama 3.1 70B led among base models. Interestingly, base models sometimes outperformed instruction-tuned ones. Exam, word definition, trivia, and inflection datasets showed the strongest performance correlations, while summarization datasets (especially news broadcasts) showed the weakest.

Results were validated through both human and LLM-based evaluation. For human evaluation a Django web application was used to collect pairwise human judgments. Claude 3.7 Sonnet's scores as an LLM judge showed strong alignment with human ratings, suggesting that top-performing LLMs can reliably support Estonian-language model evaluation.

The thesis is written in Estonian and is 62 pages long, including 9 chapters, 4 figures, and 32 tables.

Lühendite ja mõistete sõnastik

LLM	*Large Language Model*, suur keelemudel
NLP	*Natural Language Processing*, loomuliku keele töötlus
API	*Application Programming Interface*, rakendustarkvara liides
MMLU	*Massive Multitask Language Understanding*, andmestik, millega mõõdetakse keelemudelite teadmisi ja arusaamisvõimet erinevates valdkondades
GPT	*Generative Pre-trained Transformer*, tehisintellekti mudel, mis suudab teksti luua, mõista ja jätkata, olles eelnevalt treenitud suurel hulgal tekstil
Llama	*Large Language Model Meta AI* - Meta loodud suur keelemudel, mis on mõeldud teksti mõistmiseks ja genereerimiseks
LoRA	*Low-Rank Adaptation*, meetod suurte keelemudelire tõhusamaks kohandamiseks väikeste muudatustega, ilma kogu mudelit uuesti treenimata
GLUE	*General Language Understanding Evaluation*, andmestik, mis mõõdab keelemudeli keelemõistmise võimekust erinevates ülesannetes
SQuAD	*Stanford Question Answering Dataset*, andmestik, mis hindab keelemudelite võimet vastata küsimustele
RACE	*Reading Comprehension from Examinations*, andmestik, mis hindab keelemudelite lugemisvõimekust ja tekstimõistmist
BOLD	*Bias in Open-Ended Language Generation*, andmestik, mis keskendub eelarvamuste tuvastamisele ja hindamisele tekstis
ASR	*Automatic Speech Recognition*, automaatne kõnelejatuvastus
JSON	*JavaScript Object Notation*, lihtne andmevahetusvorming, mis esitab andmeid võtme-väärtuse paaridena ja on arusaadav nii inimestele kui ka arvutitele

Sisukord

Joonistе loetelu

Tabelite loetelu

1. Sissejuhatus

Suured keelemudelid (LLM-id) on saanud loomuliku keele töötlemise (NLP) nurgakiviks, näidates märkimisväärseid edusamme ülesannetes nagu teksti genereerimine, masintõlge ja semantiline mõistmine. Need mudelid on treenitud mitmekeelsetel andmekogumitel, mis on võimaldanud neil saavutada edu märkimisväärse globaalse esindatusega keeltes. Pakkudes väärtuslikke rakendusi erinevates valdkondades, alates sisuloomest kuni automatiseeritud klienditeeninduseni, on need mudelid nihutanud inimeste ja arvutite vahelise suhtluse piire.

Vaatamata nende edukusele suuremates keeltes, ei ole LLM-ide jõudlust paljudes väiksemates keeltes, nende hulgas eesti keeles, põhjalikult uuritud. Seetõttu pole hetkel terviklikku arusaama selle kohta, kui hästi need mudelid töötavad väiksema kõnelejaskonna ning unikaalsete lingvistiliste omadustega keeltes. LLM-ide evalveerimine väiksemates keeltes mängib olulist rolli keeletehnoloogia kaasavuse edendamisel ning toob esile alaesindatud keeltele ainulaadseid väljakutseid loomuliku keeletöötluse ülesannetes.

Mitmed uuringud on analüüsinud erinevate LLM-ide jõudlust mitmes väikse kõnelejaskonnaga keeles, nagu islandi ja baski keel, samuti suuremates, kuid siiski alaesindatud keeltes, nagu türgi keel, mis on küll laia kõnelejaskonnaga, kuid ei kuulu siiski põhiliste LLM-ide treeningandmestikes kasutavate keelte hulka [1] [2] [3]. Küll aga pole autorile teadaolevalt eesti keeles sellist võrdlusanalüüsi varasemalt läbiviidud.

LLM-ide teadmiste ja mõtlemisvõime hindamiseks on olemas mitmeid jõudlusteste. Need koosnevad sageli valikvastustega küsimustest, mis pärinevad tuntud eksamitelt. [4]. Laialdaselt levinud näide on MMLU andmestik, mis hõlmab 57 erinevat teemat reaalainetest kuni humanitaar- ja sotsiaalteaduste valdkondadeni [5].

2. Probleemipüstitus

2.1 Eesmärk ja uurimisküsimused

Käesoleva töö eesmärgiks on hinnata erinevate suurte keelemudelite võimekust mõista ja täita erinevaid ülesandeid eesti keeles. Selle eesmärgi saab jaotada neljaks alamsammuks. Esimeseks neist on eestikeelse võrdlusandmestiku loomine, mis võimaldaks võrrelda erinevate mudelite jõudlust. Teiseks on erinevate nii avatud kui suletud lähtekoodiga mudelite hindamine selle andmestiku abil. Kolmandaks alameesmärgiks on demonstreerida, et peenhäälestamine eestikeelsel korpusel saab parandada tulemusi neil andmestikel. Sellele järgneb analüüs väljaselgitamaks, kui palju mõjutas mudeli peenhäälestamine tulemusi testandmestikul baasmudeliga võrreldes. Neljandaks alameesmärgiks on selgitada välja, kas tugevamad LLM-id suudavad hinnata keelemudelite headust vabavormilistes küsimustes sama hästi kui inimesed.

Uurimisküsimused on järgnevad:

1. Kui võimekad on erinevad keelemudelid eesti keele töötlemisel?
2. Kuidas erinevad baasmudelite ja juhendhäälestatud mudelite tulemused eestikeelsetel ülesannetel?
3. Kuidas erinevad evalveerimise tulemused inglise keelest tõlgitud andmestiku ja eestikeelsetest materjalidest koostatud andmestiku vahel?
4. Kuidas muutuvad avatud lähtekoodiga mudeli tulemused peenhäälestamise tulemusel eestikeelsel tekstikorpusel?
5. Mil määral suudavad võimekamad keelemudelid hinnata teiste keelemudelite vastuseid avatud küsimustele ning kui sarnased on need hinnangud inimeste antud hinnangutele?

2.2 Autori panus

LLM-ide võimekuse hindamiseks eesti keeles on oluline välja töötada põhjalik, just sellele keelele omane, võrdlusandmestik. Seetõttu on käesoleva töö esimeseks osaks kvaliteetse eestikeelse andmestiku loomine, mis võimaldaks hinnata LLM-ide võimekust eesti keeles. Andmestik koosneb erinevatest ülesannetest, mille eesmärk on testida mitmesuguseid LLM-ide võimekusi. Andmestiku alamosad on järgmised:

- Küsimused eesti keel teise keelena riigieksamilt põhikooli- ja gümnaasiumiõpilastele, mis testivad nii grammatikat kui ka teksti mõistmist
- Küsimused, mis on inspireeritud MMLU andmestikust ja testivad LLM-ide teadmisi erinevatel teemadel (nt füüsika, bioloogia jne), samuti riigieksamite põhjal koostatud
- Eesti keelele spetsiifilised küsimused (sõnade tähendused, sünonüümid, lause kirjavahemärgistus, käänamine ja pööramine jne)
- Küsimused, mis on loodud LLM-ide võimekuse testimiseks pikemate tekstide mõistmisel, hinnates nende oskust koostada tekstist kokkuvõtteid
- Küsimused, mis testivad LLM-ide teadmisi Eesti kultuurist ja ajaloost
- Küsimused, mille eesmärgiks on hinnata LLM-ide keelemõistmist, testides nende oskust identifitseerida kõneleja raadiosaadete transkriptsioonide põhjal

Järgmine osa hõlmab mudelite põhjalikku hindamist loodud andmestiku alusel. Fookus on erinevate olemasolevate baas ja juhendhäälestatud LLM-ide jõudluse võrdlemisel. Jõudlust mõõdetakse ja võrreldakse eri ülesandekategooriate lõikes, mis võimaldab paremini mõista iga mudeli tugevusi ja nõrkusi ning annab väärtuslikku teavet selle kohta, milliseid mudeleid eelistada konkreetsete ülesannete lahendamisel.

Kolmandaks on valida mõni vabavaraline mudel ning seda peenhäälestada eestikeelse teksti peal. Kui selle tulemusena tõuseb mudeli jõudlikkus mõnede andmestike peal, siis see demonstreerib, et peenhäälestamine saab olla efektiivne strateegia eesti keelt hästi mõistva keelemudeli loomise osas.

Neljandaks on uurida, kas tugevamat keelemudelit saab kasutada automaatse hindajana eestikeelsete vastuste puhul. Kui mudeli antud hinnangud korreleeruvad inimekspertide omadega, viitab see sellele, et LLM-ide kasutamine teiste keelemudelite vastuste hindajana võib olla usaldusväärne ka väiksemates keeltes. See võimaldaks laiendada automaatsete hindamisstrateegiate kasutusvõimalusi eestikeelsete LLM-ide arendamisel ja evalveerimisel.

2.3 Uurimistöö disain ja metoodika

Uurimiseesmärkide saavutamiseks kasutame metoodikat, mis ühtib teadusliku hindamise põhimõtetega [6]. Uuring järgib kolmeetapilist protsessi. Esiteks kasutame andmestiku loomise meetodeid. See hõlmab kindlaksmääratud jõudlusnäitajatest inspireeritud ülesannete loomist erinevates hindamisvaldkondades, nagu grammatika, teksti mõistmine ja kultuuriteadmised. Teiseks rakendatakse kvantitatiivseid hindamismeetodeid, kasutades selliseid mõõdikuid nagu täpsus ja F-skoorid. Eesmärgiks on võrrelda baasmudelite ja

juhendhäälestatud mudelite toimivust null ja mõne vihjega seadistustes. Kolmandaks kasutame peenhäälestusmetoodikaid, kasutades andmetena eestikeelset tekstikorpust, et häälestada valitud vabavaralist keelemudelit ning uurida selle võimekuse täiustumist eri ülesannete lõikes. See loomise, hindamise ja peenhäälestuse tsükkel vastab empiirilistele uurimisstandarditele, tagades täpsuse ja korratavuse.

Pakutud meetod on asjakohane, kuna see põhineb teaduslikult valideeritud lähenemistel artefaktide hindamiseks ja jõudluse mõõtmiseks. Andmestiku loomise strateegiad, mida on kasutatud varasemates uuringutes väikse kõnelejaskonnaga keelte võrdlusandmestike väljatöötamiseks, rakendatakse ka siin, mis tagab nii asjakohasuse kui ka võrreldavuse [1] [3]. Kvantitatiivne hindamine standardsete mõõdikute abil pakub objektiivseid teadmisi LLM-ide võimekuse kohta.

3. Taust

3.1 LLM-id ja nende evalveerimine

LLM-id on viimastel aastatel oluliselt muutnud loomuliku keele töötluse valdkonda. Need mudelid on arenenud võimsateks närvivõrkudeks, mis suudavad erakordse täpsusega eriilmelisi ülesandeid lahendada. OpenAI GPT-seeria ja Google'i Gemini mudelid on treenitud tohututel tekstimahtudel, mis võimaldab neil mõista ja genereerida inimlaadset teksti. Need mudelid põhinevad transformer-arhitektuuril, mis on osutunud eriti tõhusaks kaugete sõltuvuste tuvastamisel tekstis. See on oluline selliste ülesannete puhul nagu masintõlge, küsimustele vastamine ja teksti genereerimine [7]. Kui varased LLM-id keskendusid peamiselt inglise keelele, siis hiljutised edusammud on võimaldanud nende rakendamist laiemale keeltevalikule. Siiski on paljudel neist mudelitest endiselt raskusi vähemlevinud keelte töötlemisel [8]. See piirang on eriti ilmne keeltes nagu eesti keel, millel puuduvad suuremahuliste NLP-mudelite arendamiseks vajalikud ulatuslikud lingvistilised ressursid ja andmestikud määral, mil need eksisteerivad suuremate keelte jaoks.

3.2 Väljakutsed mitmekeelses loomuliku keele töötluses

Madala ressursiga keelte, nagu eesti keel, jaoks tõhusate NLP-mudelite arendamine on keeruline ja mitmetahuline ülesanne. Eesti keele unikaalne grammatiline struktuur, mis sisaldab 14 käänet, tekitab olulisi väljakutseid nii keele mõistmise kui ka genereerimise ülesannetes. Erinevalt lihtsama grammatikaga keeltest, peavad eesti keelt töötlevad mudelid arvestama suure hulga erinevate käändelõppude ning sõnavormide variatsioonidega, mistõttu on traditsiooniliste NLP-mudelite jaoks keeruline üldistada erinevaid sõnavorme.

Lisaks piirab eestikeelsete suurte ja avalikult kättesaadavate andmestike nappus keelemudelite võimet keelt tõhusalt õppida. Eesti keel, mida räägib kogu maailmas ligikaudu 1 miljon inimest, ei oma selliseid suuri tekstikorpusi, mis on olemas laiemalt kasutuses olevate keelte nagu inglise või saksa keele puhul. See piirang raskendab märkimisväärselt suurte keelemudelite treenimist ja nende võimekuse parandamist eesti keeles [9]. See seab olulisi takistusi teadlastele, kes arendavad ja evalveerivad eesti keele jaoks mõeldud keelemudeleid.

3.3 Peenhäälestamine väiksema kõnelejaskonnaga keeltes

Üks paljulubav lähenemine LLM-ide jõudluse parandamiseks vähemlevinud keeltes on peenhäälestamine. Peenhäälestamine hõlmab eeltreenitud mudeli kohandamist konkreetsele valdkonnale või keelele, treenides seda väiksema, ülesandepõhise andmestiku abil. Seda meetodit on edukalt kasutatud mudelite jõudluse parandamiseks spetsialiseeritud valdkondades, nagu meditsiin või õigus [10]. Eesti keele puhul võib olemasoleva mudeli peenhäälestamine eestikeelsel korpusel, näiteks Eesti Vikipeedial, aidata mudelil õppida keele nüansse ja parandada selle sooritust eesti keelele omastes ülesannetes. See lähenemine võib aidata lahendada keele ainulaadsete grammatiliste omaduste ja piiratud ressursside põhjustatud väljakutseid. Hiljutised uuringud vähemlevinud keelte peenhäälestamise vallas on näidanud paljulubavaid tulemusi.

2024. Aastal ilmunud artiklis tutvustavad autorid uut suurt keelemudelit, mis on kohandatud LLaMA 2 7B põhjal ja toetab 534 keelt, keskendudes väiksema kõnelejaskonnaga keeltele. Sellel on kasutatud sõnavara laiendamist, jätkuõpet ning LoRA-põhist (*Low-Rank Adaptation*) kohandamist. Mudel saavutab tugevaid tulemusi nii mudeli sisemistes kui välistes testides, edestades varasemaid mitmekeelseid mudeleid. Antud uurimuse peamine uuenduslikkus seisneb generatiivsete LLM-ide laiendamises sadadele keeltele, säilitades samaaegselt efektiivsuse, mis demonstreerib peenhäälestamise kasulikkust keelemudelite võimekuste laiendamisel alaesindatud keeltele [11].

Teine samal aastal ilmunud uurimus adresseerib LLM-ide kehvemat sooritust konkreetselt vähelevinud Aafrika keeltes, luues ligikaudu miljon inimese poolt tõlgitud sõna uuteks võrdlusandmeteks. Töö toob esile olulised erinevused võrreldes inglise keelega ning näitab, et kvaliteetse ja kultuuriliselt asjakohase andmestikuga peenhäälestamine võib täpsust märkimisväärselt parandada. Uurimus illustreerib, kuidas sihipärane peenhäälestamine kvaliteetse ja kultuuriliselt sobiva andmestikuga võib oluliselt parandada LLM-ide tulemusi vähemlevinud keeltes [12].

3.4 Jõudlustestide roll suurte keelemudelite hindamisel

Jõudlustestidel on LLM-ide jõudluse hindamisel oluline roll. Need standardiseeritud testid võimaldavad teadlastel hinnata mudeli keeleoskuse erinevaid aspekte, nagu teksti mõistmine, arutlusoskus ja üldistusvõime. Laialtlevinud keelte jaoks on välja töötatud mitmesuguseid jõudlusteste mudeli jõudluse mõõtmiseks eri ülesannetes [5] [13] [14]. Kuigi need jõudlustestid on andnud väärtuslikku teavet LLM-ide võimekuse kohta suure kõnelejaskonnaga keeltes, puuduvad sarnased põhjalikud hindamisraamistikud mitmete

väikse kõnelejaskonnaga keelte jaoks. Ehkki olemasolevate jõudlustestide kohandamiseks ka teistele keeltele peale inglise keele on tehtud jõupingutusi, piirab neid sageli keeleliste ressursside vähene kättesaadavus ja väljakutsed, mis on omased just sellele konkreetsele keelele [1].

3.5 Lünk olemasolevas uurimistöös

Vaatamata kasvavale hulgale uuringutele suurtest keelemudelitest mitmekeelse loomuliku keele töötluse kontekstis, esineb märkimisväärne lünk mudelite hindamisel eesti keeles ning teistes vähemlevinud keeltes. Kuigi laialtlevinud keelte jaoks on loodud mitmeid kvaliteetseid jõudlusteste, on neid eesti keele jaoks spetsiaalselt loodud väga vähe. Selliste ressursside puudumine raskendab erinevate mudelite jõudluse võrdlemist eesti keeles ning nende sobivuse hindamist praktilisteks rakendusteks. See lünk uurimistöös tõstab esile vajaduse eestikeelsete jõudlustestide järele.

4. Eelnev teadustöö

Suuremahuliste keelemudelite hindamine on viimastel aastatel märkimisväärselt arenenud, kuna mudelite võimekus ja ulatus on oluliselt kasvanud. Varased jõudlustestid keskendusid peamiselt keelelistele ülesannetele, mis andsid põhjalikku teavet mudelite keelemõistmise kohta [15] [16]. Kuigi need võrdlusnäitajad olid kasulikud üldiste keeleoskuste testimiseks, jäid need sageli alla arenenud suurte keelemudelite keerukamate väljakutsete käsitlemisel.

4.1 Olemasolevad jõudlustestid

Suurte keelemudelite arenedes on nende võimekuse hindamiseks ja seotud väljakutsete lahendamiseks välja töötatud üha uusi jõudlusteste.

Üks esimesi suuremahulisi jõudlusteste hindamaks LLM-ide võimekust hulgal üldise keelemõistmise ülesannetel oli GLUE (*General Language Understanding Evaluation*) [13]. See sisaldab mitmeid ülesandeid, nagu sentimendianalüüs, tekstiline järeldamine, küsimustele vastamine ja lause sarnasus, ning on mõeldud hindama mudelite võimekust täita erinevaid NLP ülesandeid ilma ülesandepõhise peenhäälestamiseta. Sellega pani GLUE aluse mudelite üldise keelemõistmise võimekuse mõõtmise standardile. Loodud järletulijana GLUE-le, on SuperGLUE mudelitele väljakutsuvam, disainitud nihutama mudelite üldise keelemõistmise piire [17]. See sisaldab keerukamaid ülesandeid nagu tervemõistuslik arutlemine, keeleliste nähtuste mõistmine ning sügavam kontekstimõistmine.

Lisaks on loodud olulisi võrdlusandmestikke suurte keelemudelite lugemisoskuse ja küsimustele vastamise võimekuse hindamiseks. SQuAD (*Stanford Question Answering Dataset*) sisaldab peamiselt faktipõhiseid küsimusi, mille vastused on otseselt tekstist leitavad, seega sobib see hästi mudelite pindmise tekstimõistmise ja infootsingu hindamiseks [18]. RACE (*ReAding Comprehension from Examinations*) seevastu põhineb Hiina põhi- ja keskkooli lõpueksamitel ning esitab keerukamaid küsimusi, mis nõuavad sügavamat arutlusvõimet ja järeldusoskust [19]. RACE testib mudelite võimet mõista konteksti, seostada mitmeid tekstiosasid ja teha loogilisi järeldusi.

Samuti on kasutusele võetud spetsialiseeritud jõudlustestid, mis keskenduvad konkreetsete keeleliste või teadmispõhiste võimete hindamisele. Näiteks MMLU (*Massive Multitask Language Understanding*) mõõdab mudelite teadmisi 57 eri ainevaldkonnas, kasutades

valikvastustega küsimusi, mis sarnanevad ülikooli tasemeeksamitega [5]. See võimaldab hinnata mudeli võimet rakendada üldisi teadmisi maailmast keerukates olukordades. Teine näide on HellaSwag, mille eesmärk on testida mudeli tervemõistuslikku arutlusoskust, lastes mudelil valida loogilise jätku lühikesele stseenikirjeldusele [14].

Loodud on ka spetsiaalseid jõudlusteste, nagu HumanEval ja BigCodeBench, mille eesmärgiks on hinnata keelemudelite võimet genereerida funktsionaalset koodi [20] [21]. HumanEval on Pythoni-põhine jõudlustest, mis koosneb programmeerimisülesannetest, kus mudel peab kirjutama funktsiooni, mis vastab etteantud kirjeldusele ja läbiks automaatsed testid. See on kujunenud standardiks koodigeneratsiooni kvaliteedi hindamisel. BigCodeBench on samuti Pythoni-põhine, kuid märksa ulatuslikum jõudlustest, mis sisaldab üle tuhande ülesande, hõlmates erinevaid domeene ja sadu teeke. See pakub keerukamaid ja elulähedasemaid ülesandeid kui HumanEval ning mõõdab mudelite võimekust lahendada realistlikke tarkvaraarenduse probleeme. Sellised jõudlustestid on olulised keelemudelite praktilise rakendatavuse hindamiseks ning koodigeneratsiooni tugevuste ja kitsaskohtade tuvastamiseks.

Mõned jõudlustestid keskenduvad genereeritud tekstis eelarvamuste tuvastamisele, kasutades meetodeid nagu stereotüüpide ja diskrimineeriva keelekasutuse hindamine [22]. Näiteks BOLD (*Bias in Open-Ended Language Generation*) mõõdab mudelite eelarvamusi erinevates demograafilistes kategooriates, sealhulgas sugu, rass ja vanus, analüüsides avatud vastustega genereeritud tekste. Sellised testid on olulised, et hinnata keelemudelite sotsiaalset vastutustundlikkust ning ennetada potentsiaalset kahju kasutajatele.

4.2 Väheesindatud keeled

Kuigi laialtlevinud keelte jaoks on välja töötatud arvukalt kvaliteetseid jõudlusteste, on sarnaste jõudlustestide kättesaadavus väheesindatud keelte jaoks endiselt piiratud. Sellised keeled nagu eesti keel seisavad silmitsi märkimisväärse puudujäägiga hindamisraamistikes. Jõudlustestide puudumine tekitab raskusi teadlastele ja arendajatele, kes püüavad nende keelte jaoks keelemudeleid hinnata või peenhäälestada. Siiski on hiljuti tehtud jõupingutusi selle lünga täitmiseks mitmete alaesindatud keelte puhul. Need algatused keskenduvad sageli hindamisraamistike väljatöötamisele, mis hõlmavad erinevaid ülesandeid, alates üldisest keele mõistmisest kuni spetsiifilisemate rakendusteni, nagu kokkuvõtete genereerimine või teksti põhjal küsimustele vastamine [23].

Ühes 2024. aastal ilmunud artiklis tutvustatakse Latxat – spetsiaalselt baski keele jaoks loodud avatud lähtekoodiga suurte keelemudelite perekonda. Need mudelid põhinevad Llama 2 mudelitel (7B, 13B ja 70B parameetrit), mida on edasi eeltreenitud hoolikalt

valitud baskikeelse korpuse peal, mis sisaldab 4,2 miljardit tokenit. Baski keelele sobivate hindamisressursside puuduse lahendamiseks loodi neli uut valikvastustega andmestikku, mis sisaldavad kokku üle 23000 küsimuse. Latxa mudelid ületavad oluliselt kõiki varasemaid avatud mudeleid ning ka GPT-3.5 Turbo't kõigis testides, kusjuures 70B mudel edestab isegi GPT-4 Turbo't keeleoskuse hindamisel. See töö näitab, et olemasolevate avatud lähtekoodiga mudelite abil on võimalik luua kõrgekvaliteedilisi keelemudeleid vähelevinud keelte jaoks [1].

Teine samuti 2024. aastal ilmunud artikkel kajastab esimese ulatusliku jõudlustesti loomist suurte keelemudelite hindamiseks tšehhi keeles [23]. BenCzechMark sisaldab 50 mitmekesist ülesannet 8 kategoorias, sealhulgas uusi andmestikke uudiste, esseede ja kõnekeele valdkondadest. See kasutab õiglast võrdlust võimaldavat skoorisüsteemi, mis põhineb statistilisel olulisusel ja uuel *Duel Win Score* (DWS) mõõdikul. Samuti avaldatakse suurim puhastatud tšehhikeelne korpus, mida kasutatakse esimeste tšehhikesksete mudelite (kuni 7B parameetrit) treenimiseks. Ühendades MMLU- ja GLUE-tüüpi ülesanded koos läveta hindamismõõdikutega, võimaldab BenCzechMark usaldusväärset ja korduvkasutatavat tšehhi keele mudelite hindamist.

Mõnel juhul keskenduvad jõudlustestid kindlatele tuntud ingliskeelsetele andmekogumitele, nagu MMLU, analoogide loomisele kindla vähelevinud keele jaoks.

2024. aastal avaldatud TurkishMMLU on esimene mitme ülesande ja valikvastustega suurte keelemudelite türgi keeles küsimustele vastamise võimekuse hindamiseks loodud andmestik[3]. See sisaldab üle 10 000 küsimuse, mis katavad üheksat ainet Türgi gümnaasiumi õppekavast ja ülikooli sisseastumiseksamitelt. Andmestik sisaldab teemasid loodusteadustest, matemaatikast, türgi keelest ja kirjandusest ning sotsiaalteadustest, peegeldades Türgi keelelist ja kultuurilist konteksti. Hinnatakse üle 40 LLM-i, sealhulgas avatud lähtekoodiga ja suletud lähtekoodiga mudeleid, kasutades nii nulli kui mitme näitega ning ka mõttekäigupõhise arutlemise seadeid. Tulemused näitavad tulemuste varieerumist aine ja raskusastme järgi, kus suletud lähtekoodiga mudelid suudavad saavutada paremaid tulemusi. Antud töö pakub väärtuslikku ülevaadet LLM-de tugevustest ja piirangutest türgi keeles.

Samuti 2024. aastal ilmunud artiklis esitletakse ArabicMMLU-d – esimest mitmeülesandelist keelemõistmise andmestikku araabia keeles, mis koosneb 40 ülesandest ja üle 14000-st valikvastusega küsimusest kaasaegses standardaraabia keeles [24]. Küsimused pärinevad koolieksamite materjalidest Põhja-Aafrikas, Levandis ja Pärsia lahe piirkonnas ning need koostati koostöös kohalike emakeelsete ekspertidega. Mudelite hindamine näitab, et araabiakeelsed LLM-id saavutavad täpsuse maksimaalselt 62,3% , samal ajal kui

paljud avatud lähtekoodiga mudelid ei ületa 50% piiri. ArabicMMLU pakub väärtuslikku võrdlusraamistikku araabiakeelsete LLM-ide teadmiste ja arutlusvõimekuse hindamiseks ning loob aluse tulevastele lokaalselt kohandatud keelemudelitele.

Lisaks sisaldavad need jõupingutused sageli uusi hindamismehhanisme, hindamaks mudeleid mitme mõõdiku alusel. Sellised jõudlustestid võimaldavad mitte ainult keelemudelite paremat hindamist, vaid pakuvad ka nende arendamiseks ja peenhäälestamiseks vajalikke algandmestikke.

5. Metoodika

5.1 Keelemudelite evalveerimine

Keelemudelite jõudluse süstemaatiliseks hindamiseks erinevatel ülesannetel kasutatakse käesoleva lõputöö raames LM Evaluation Harnessi [25]. Tegu on vabavaralise raamistikuga, mis on välja arendatud EleutherAI uurimiskollektiivi poolt. Raamistik on disainitud pakkumaks standardiseeritud, laiendatavat ning reprodutseeritavat viisi keelemudelite võrdlemiseks paljudel tuntud keeletöötlusülesannetel. Tegemist on ühe enimkasutatava tööriistaga keelemudelite võimekuse objektiivseks hindamiseks.

Mida suuremaks ja keerukamaks keelemudelid muutuvad, seda keerulisem on nende võimekust süsteemselt hinnata. Traditsioonilised hindamismeetodid põhinevad sageli projektipõhistel skriptidel, mis raskendavad tulemuste võrreldavust. LM Evaluation Harness pakub ühtset liidest, mis võimaldab mudeleid hinnata nii nulli näitega kui ka mõne näitega ülesannetel, kasutades standardiseeritud viise andmete esitamiseks ja tulemuste mõõtmiseks.

Raamistiku disain on ühilduv paljude platvormidega, nagu Huggingface, ning kommertsiaalsete API-dega, nagu OpenAI, võimaldades hinnata nii avatud lähtekoodiga kui kommertsiaalseid mudeleid samade ülesannete kaudu. Raamistiku disainis on rõhutud reprodutseeritavusele, võimaldades kasutada kindlalt määratletud prompt'e ja seadistusi [25]. Seda kasutatakse laialdaselt teadusartiklites ja LLM võrdlustabelites, näiteks HuggingFace'i Open LLM Leaderboardis [26].

5.2 Ülevaade testandmestikest / jõudlustestidest

Suurte keelemudelite erinevate oskuste hindamiseks ja omavaheliseks võrdlemiseks kasutati 8 erinevat andmestikku, neist 4 on loodud käsitsi vastavatest materjalidest küsimusi kokku kogudes ja 4 on olemasolevad andmestikud kohandatud testide läbiviimiseks LM Evaluation Harness raamistikus.

5.2.1 Riigieksamitel põhinev andmestik

Riigieksamitel põhinev andmestik (`https://huggingface.co/datasets/TalTechNLP/exam_et`) on koostatud küsimustest, mis on esinenud gümnaasiumi

Riigieksamite ja põhikooli lõpueksamite töödes 2003-2024.ndal aastal. Andmestik hõlmab valikvastustega küsimusi 7-st õppeainest: füüsika, keemia, bioloogia, geograafia, ajalugu, ühiskonnaõpetus ning eesti keel teise keelena. Eesti keele teise keelena küsimused on koostatud nii gümnaasiumi kui põhikooli eksamitööde põhjal. Ülejäänud ainete küsimuste koostamisel on kasutatud ainult põhikooli eksamitöid, kuna neis ainetes pole 2014. aastast enam gümnaasiumitasemel eksameid. Selle andmestiku eesmärgiks on hinnata LLM-ide teadmisi erinevates valdkondades.

Eesti keele teise keelena alamosa eksamite andmestikust koosneb 246 gümnaasiumi- ja 238 põhikoolitasemel küsimusest. Ülesandetüübilt on sealsed küsimused lüngatäitmise stiilis, kus tuleb lausesse või teksti valida korrektne sõna või tekstiosa valikuvariantide seast või teksti põhjal küsimustele vastamine. Seega testivad need küsimused keelemudelite sisulist arusaamist tekstist või lausestruktuurist. Igale küsimusele on vastavalt 2-15 vastusevarianti, mille seast tuleb üks õige valida.

Ülejäänud õppeainete küsimused on kas juba eksamitöös eksisteerinud või viidud lihtküsimuste, lünka sobiva sõna valimiste või tõde-väär tüüpi. Ühiskonnaõpetuse küsimusi on andmestikus kokku 310, bioloogia 207, ajalugu 261, keemia 133, geograafia 111 ning füüsika 108.

Näide andmestiku kirjest:

```
Küsimus: Milline vastusevariantidest on
täiskasvanud kahepaikse tunnus ja ei ole kala tunnus?
Vastusevariandid: ["Neerud", "Küljejoon", "Lõpused",
"Kopsud"]
Vastus: Kopsud
```

5.2.2 Käänamise andmestik

Eesti keel on keeruka struktuuriga, koosnedes tervelt 14 käändest. Käänamise andmestik (`https://huggingface.co/datasets/TalTechNLP/inflection_et`) on mõeldud testima LLM-ide võimekust orienteeruda neis käänetes. Mudelile antakse ette omadussõna-nimisõna paar ja sihtkääne koos ainsuse/mitmuse täpsustusega ning ülesandeks on tagastada sõnapaar sobivas käändes.

Kuna ilmnes, et tihti oli mudelitel ka lihtsamate/levinumate käänetega raskusi, siis piirdutakse antud juhul nimetava, omastava, osastava ning sisseütleva käände testimisega. Andmestikus on 200 sõnapaari ning iga sõnapaari kohta 7 võimalikku

sihtkäände ja pluraalsuse kombinatsiooni. Mudelile antakse alati ette sõnapaar ainsuse nimetavas ning palutakse see käänata kas mitmuse nimetavasse, ainsuse omastavasse, mitmuse omastavasse, ainsuse osastavasse, mitmuse osastavasse, ainsuse sisseütlevasse või mitmuse sisseütlevasse vormi. Seega kokku on andmestikus 1400 rida. Õigete käändevormide saamiseks on kasutatud Filosoft Eesti keele spellerit [27].

Näide andmestiku kirjest:

```
Ainsuse nimetav: melanhoolne atmosfäär
Mitmuse osastav: ["melanhoolseid atmosfääre",
"melanhoolseid atmosfäärisid"]
```

5.2.3 Sõnaseletuste andmestik

Sõnaseletuste andmestiku (https://huggingface.co/datasets/TalTechNLP/word_meanings_et) ideeks on kontrollida LLM-ide arusaamist sõnaseletustest ja oskust vastata neile vastava sõnaga. Andmestikus on kokku 54562 sõna koos vastava definitsiooni või definitsioonidega. Kuna 54562 sõna oli testimise mõttes ebamõistlikult suur kogus, siis siinkohal valiti kõigi sõnade hulgast suvaliselt 1000, mida kasutada testandmestikuna. Sõnaseletused pärinevad Ekilexist ja Wordnetist [28] [29].

Mudelile antakse sisendiks sõnaseletus või mitu alternatiivset sõnaseletust ning palutakse vastata sobiva sõnaga. Testid võtavad arvesse ka seda, et sõnadel võib olla sünonüüme ning ka sünonüümid loetakse õigeks.

Näide andmestiku kirjest:

```
Sõna definitsioon(id): ["arvutivõrgu vahendusel
saadetud sõnum või kiri", "elektrooniline kiri"]
Sõna(d): ["meil", "elektronkiri", "e-kiri",
"elektronpost", "e-post", "e-mail", "kiri"]
```

5.2.4 Mälumängu andmestik

Mälumängu andmestik (https://huggingface.co/datasets/TalTechNLP/trivia_et) koosneb küsimustest "Eesti mälumäng" lauamängu kaartidelt. Kaardid on skanneeritud ja neilt tekstituvastusvahenditega tekst digitaliseeritud ning seejärel käsitsi ülevaadatud ning parandatud. Kokku on andmestikus 800 valikvastustega küsimust

ajaloo, teaduse, kultuuri, spordi, geograafia ning varia valdkondadest, s.h valdav osa küsimusi on otseselt seotud Eestiga. Keelemudeli ülesandeks on valida küsimusele nelja vastuvariandi seast korrektne ning ülesande eesmärgiks on hinnata keelemudelite teadmisi Eesti kontekstis.

Näide andmestiku kirjest:

```
Küsimus: Milline neist ei ole Eesti piiripunkt?
Vastusevariandid: ["Ikla", "Lähte", "Koidula",
"Luhamaa"]
Vastus: Lähte
```

5.2.5 Grammatika andmestikud

Grammatika andmestikud (`https://huggingface.co/datasets/TalTechNLP/grammar_et` ja `https://huggingface.co/datasets/TalTechNLP/grammar2_et`) on modifitseeritud struktureeritud andmestikeks TartuNLP Estgec grammatikakorrektuuri andmestiku alusel. Estgec andmestik sisaldab kahte erineva tasemega tekstikorpust, mis koosnevad vastavalt eesti keelt emakeelena rääkivate ning B1-B2 tasemel rääkivate õpilaste kirjutatud lausetest ning vastavatest parandustest. Laused võivad olla nii grammatiliselt korrektsed kui ka vigadega. Iga vigase lause kohta on vähemalt üks sobilik parandus.

Andmestik sai viidud lihtsustatud kujule, kus on vastavalt originaalne lause ja vajadusel parandatud lause. Andmestikud on eraldiseisvad vastavalt õpilase keeletasemele. Eesti keelt emakeelena kõnelevate õpilaste vigadest koosnevas andmestikus on 446 rida ning võõrkeelena kõnelevate omadest 3285.

Mudeli ülesandeks on parandada lause korrektseks, juhul kui see seda veel pole. Mudelite vastuseid hinnati nii binaarse meetrikaga, kui ka Levenshteini kaugusel põhineva meetrikaga. Binaarse meetrika puhul loetakse vastus õigeks kui mudeli väljund on täpselt õige ning vastasel juhul valeks. Levenshteini kaugusel põhinev meetrika määrab igale väljundile skoori vahemikus 0 kuni 1, kus 1 tähistab ilma vigadeta vastust ning mida erinevam väljund on oodatud vastusest, seda lähemal nullile see skoor on.

Näide andmestiku kirjest:

```
Originaalne lause:  Ta kiire õppib.
Parandatud lause:  Ta õpib kiiresti.
```

5.2.6 Dialoogide kokkuvõtete andmestik

Dialoogide summeerimise andmestik Dialogsum_EE (https://huggingface.co/datasets/TalTechNLP/dialogsum_ee) on masintõlgitud alamosa Dialogsum andmestikust [30]. See koosneb dialoogidest ja nende inimese poolt kirjutatud kokkuvõtetest ning selles on kokku 282 rida. LLM-i eesmärgiks on genereerida dialoogile võimalikult sobilik kokkuvõte. Mudeli väljundi ja oodatud väljundi võrdlemiseks kasutatakse Rouge-1, Rouge-2 ja Rouge-L meetrikaid.

5.2.7 Uudissaadete kokkuvõtete andmestik

Uudissaadete kokkuvõtete andmestik ErrNews (https://huggingface.co/datasets/TalTechNLP/ERRnews) on TalTechNLP andmestik, mis koosneb ERR uudistesaadete transkriptsioonidest ühes inimese poolt kirjutatud kokkuvõtetega. Uudislood pärinevad ERR arhiivist ning nende transkriptsioonid on genereeritud automaatse kõnelejatuvastuse (ASR) töövooga. Andmestikus on kokku üle 10000 rea, kuid testimisks on kasutatud testjaotist, mis on 523 rida. Keelemudeli eesmärgiks on kirjutada uudislõigule võimalikult hea kokkuvõte. Mudeli väljundi ja oodatud väljundi võrdlemiseks kasutatakse Rouge-1, Rouge-2 ja Rouge-L meetrikaid.

5.2.8 Kõnelejatuvastuse andmestik

Kõnelejatuvastuse andmestik (https://huggingface.co/datasets/TalTechNLP/paevakaja_speakers) on TalTechNLP andmestik Vikerraadio "Päevakaja" saadetest, kus iga väli koosneb saate transkriptsioonidest koos kõnelejamärgistega ning järjendist saates kõnelenud inimeste nimedega. Kokku on andmestikus 20 rida. Keelemudeli ülesandeks on nimetada transkriptsiooni põhjal kõik saatelõigus kõnelenud isikud. Eesmärgiks on testida mudeli võimekust mõista teksti süvitsi ning konteksti põhjal pakkuda õiged kõnelejad. Mudeli väljundit ning oodatud vastust võrreldakse ning arvutatakse täpsus, saagis ning nendest tuletatud F1-skoor.

5.3 Testides kasutatud meetrikad

5.3.1 Levenshteini kaugusel põhinev meetrika

Levenshteini kaugus on laialdaselt kasutatav meetrika kahe sõne sarnasuse hindamiseks. Levenshteini kaugus on defineeritud kui vähim arv lisamisi, kustutamisi või vahetusi, mis on vajalik sooritada saamaks ühest sõnest teise [31]. Kahe identse sõne Levenshteini kaugus on 0, samas kui maksimaalne kaugus võib olla arbitraarselt suur. Mitme sõnepaari Levenshteini kauguste keskmine ei osutunud aga sobivaks üldistavaks näitajaks, kuna üks väga erinev sõnepaar võib keskmist oluliselt mõjutada. Kasutades järgmist valemit:

$$\frac{1}{1 + \text{Levenshteini kaugus}}, \tag{5.1}$$

saavutame väärtused, mis on üks, kui sõned on identsed, ning lähenevad nullile, mida erinevamad need on. Selline teisendus rõhutab tugevalt erinevust identsete sõnede ja näiteks ühe tähemärgi võrra erinevate sõnede vahel (esimesel juhul väärtus 1, teisel 0,5). Samal ajal, kui sõned on juba niigi üsna erinevad, siis täiendavad muudatused ei mõjuta tulemusi märkimisväärselt. See omadus on käesoleva probleemi puhul kasulik, kuna meid huvitavad eelkõige väiksemad vead – näiteks, kas mudel paigutas kirjavahemärgi õigesse kohta. Kui aga mudeli väljund on üldiselt ebatäpne, ei oma selle eksimuse täpne ulatus enam olulist tähendust.

Näiteks lausete "Ta õppis palju." ja "Ta õpis palju." vaheline Levenshteini kaugus on 1 ning valemi abil arvutatud headus sellisel juhul on 0,5. Kuid lausete "Ta õppis palju." ja "Õppis tema kogu aeg." vaheline Levenshteini kaugus on suur ning valemi abil arvutatud headus on peaaegu 0.

5.3.2 Täpsus, normeeritud täpsus ja saagis

Terminit "täpsus" kasutatakse sõltuvalt olukorrast veidi erinevates tähendustes.

Olukordades, kus mudel annab ühe väljundi ning on üks õige väljund, viitab täpsus sellele, kui suur osakaal oli kordadel, kus mudel andis täpselt oodatud väljundi. Selline on näiteks käänamiste andmestik.

Valikvastustega küsimuste puhul kirjeldab täpsus analoogselt kui suur osakaal oli kordadel, kus mudeli pakutud vastusevariant oli õige vastusevariant. Erinevus tekib baasmudelite puhul, kus muutub protsess, kuidas mudeli väljundit leitakse. Ilma

vastusevariantideta mudelite või juhendhäälestatud mudelite puhul annab mudel mingi väljundi ning vaadatakse, kas see väljund oli õige või mitte. Kusjuures on võimalik, et vastusevariantidega küsimuse puhul annab mudel väljundi, mis ei ole mitte ükski antud vastusevariantidest. Olukorras, kus on tegemist vastusevariantidega küsimusega ning testitakse baasmudelit, käib aga mudeli pakutud vastusevariandi leidmine järgnevalt. Keelemudelid genereerivad teksti ühe sümboli haaval, kusjuures iga sümboli genereerimisel annavad nad tegelikult välja tõenäosusjaotuse, mis kirjeldab iga võimaliku sümboli kohta tõenäosust, et see sümbol on järgmine. Tavaliselt valitakse siis kõige tõenäolisem sümbol ning lastakse genereerida järgmine sümbol. Kuid teades kindlaid vastusevariante, saab leida iga vastusevariandi kohta tõenäosuse, et mudeli väljund on täpselt see vastusevariant. Sellisel juhul, peale sümbolite tõenäosusjaotuse saamist, vaadatakse, milline on vastusevariandis järgmine sümbol ning vaadatakse, mis tõenäosuse mudel sellele sümbolile andis. Seejärel valitakse see sümbol ning lastakse mudelil genereerida järgmine tõenäosusjaotus. Korrutades kõikide vastusevariandis olnud sümbolite tõenäosused, saab seega iga vastusevariandi kohta tõenäosuse, et mudel genereeriks väljundina just selle vastuse. Mudeli pakutud vastusevariandiks võetakse seejärel kõige suurema tõenäosusega vastusevariant.

Eelnevalt kirjeldatud protsess annab lühikestele vastustele mõnes mõttes eelise, kuna nende puhul korrutatakse vähem tõenäosusi. Seega saab siin alternatiivselt leitud tõenäosused jagada iga vastusevariandi pikkusega. Selliste väljunditega leitud tulemust nimetatakse normeeritud täpsuseks.

Olukordades, kus oodatud väljund on nimekiri asjadest ning mudel annab samuti välja nimekirja asjadest, saab leida täpsuse ja saagise. Antud juhul kirjeldab täpsus osakaalu mudeli väljunditest, mis olid päriselt õiged vastused. Ning saagis kirjeldab osakaalu õige vastuse elementidest, mis esinesid ka mudeli väljundis. Sellisel juhul saab arvutada ka F_1 skoori kasutades valemit [32]

$$F_1 = 2 \cdot \frac{\text{täpsus} \cdot \text{saagis}}{\text{täpsus} + \text{saagis}} \tag{5.2}$$

Ainuke andmestik, millel on selline väljundi tüüp on kõnelejatuvastuse andmestik, kus mudel peab teksti põhjal pakkuma mitme kõneleja identiteete.

5.3.3 ROUGE skoor

ROUGE-N mõõdab masingenereeritud teksti ja vastava inimgenereeritud näiteteksti vahelist kattuvate N-grammide arvu. Antud töö kontekstis on ROUGE skoorid kasutuses kokkuvõtete genereerimise ülesannetes, kus on eesmärgiks hinnata mudeli poolt genereeri-

tud kokkuvõtte headust. Hinnatatakse nii ROUGE-1, ROUGE-2 kui ka ROUGE-L skoori, mis hindavad vastavalt kokkuvõtete kattuvaid 1-, 2- ja L-gramme, kusjuures L viitab pikimale kattuvale järgnevusele mudeli väljundi ja viiteväljundi vahel [33].

Näiteks võrreldes lauseid "Täna on soe ja ilus ilm." ja "Soe ning ilus ilm on täna." vahel tulevad ROUGE meetrikatele järgnevad väärtused. Ühised sõnad on "täna", "on", "soe", "ilus" ja "ilm" ning kokku on sõnu 6, seega ROUGE-1 skoor on $\frac{5}{6}$. Ainuke ühine järjestikune sõnapaar on "ilus ilm" ning kokku on võimalikke sõnapaare 5, seega ROUGE-2 skoor on $\frac{1}{5}$. Pikim ühine sõnade alamjärjend on "soe ilus ilm" ning pikim võimalik sõnade alamjärjend on pikkusega 6, seega ROUGE-L skoor on $\frac{3}{6}$.

6. Tulemused

Järgnevalt on esitatud jõudlustestide tulemused eelmainitud andmestikel. Kõik baasmudelite tulemused on 5 näitega ning juhendhäälestatud mudelite omad 0 näitega konfiguratsioonis, kui pole teisiti täpsustatud.

6.1 Eksami andmestik

6.1.1 Baasmudelid

Tabel 1 kujutab baasmudelite täpsust eri õppeainete lõikes ja kokkuvõtlikult eksami andmestiku peal. Võrdlusesse on lisatud ka juhuslikult vastamise täpsus. Baasmudelitest said eksami andmestiku peal parimaid tulemusi suuremad mudelid, nagu Gemma 27B, LLama 70B ja Qwen 72B, kuid nende täpsus jäi keskmiselt alla 0,7. Võrreldes võrdluses olevate väiksemate mudelitega, oli nende täpsus umbes 0,2 võrra parem. Õppeainete lõikes kõiguvad tulemused üpris palju: parimad on tulemused ajaloo küsimustele, kus kõik testitud mudelid saavutavad täpsuse üle 0,9, samal ajal eesti keel teise keelena gümnaasiumi osas oli parima baasmudeli täpsus napilt üle 0,5 ning teiste omad jäid alla 0,5. Selle põhjuseks on ilmselt tõsiasi, et kui ajaloo puhul olid küsimused üldiselt faktilised ja sealhulgas sageli vaid kahe vastusevariandiga, siis eesti keele küsimused hõlmasid teksti sobiva osa valimist, nõudes sügavamat tekstimõistmist, ning samal ajal olid paljud neist üle 10 vastusevariandiga. Kui eesti keel teise keelena osade puhul on ka juhuslikult vastamise täpsused vastavalt kõige väiksemad, siis üllatavalt on mudelid saanud võrdlemisi madalaid tulemusi ka geograafia küsimustes, mille juhuslikult vastamise täpsus on kõigist kategooriatest suurim. Parima mudeli täpsus geograafia peal on kõigest vähem kui 0,2 võrra parem juhuslikult valimise täpsusest, viidates küsimuste sisulisele keerukusele keelemudelite jaoks.

Tabelis 2 on kujutatud baasmudelite normeeritud täpsused eksamite andmestikul. Võrreldes baasmudelite normeeritud täpsuse tulemusi tavalise täpsuse tulemustega (tabel 1), näib et nende vahel suurt erinevust pole ning mudelite kokkuvõttev paremusjärjestus eksamite andmestiku korral jääb antud valimi korral samaks olenemata täpsuse tüübist.

Mudel	Täpsus Keskmine	Täpsus Ajalugu	Täpsus Bioloogia	Täpsus Eesti k 1	Täpsus Eesti k 2	Täpsus Füüsika	Täpsus Geograafia	Täpsus Keemia	Täpsus Ühisk.
Juhuslik	0.285	0.261	0.281	0.197	0.210	0.247	0.439	0.287	0.372
Gemma 2 9B	0.616	0.913	0.517	0.386	0.647	0.620	0.559	0.744	0.781
Gemma 2 27B	**0.693**	**0.957**	0.643	**0.528**	**0.672**	**0.741**	**0.604**	**0.835**	**0.810**
Llama 3.1 8B	0.405	**0.957**	0.362	0.224	0.248	0.389	0.514	0.451	0.603
Llama 3.1 70B	**0.667**	**0.957**	**0.700**	0.390	0.580	**0.750**	**0.586**	**0.857**	**0.829**
Mistral Nemo Base	0.487	**0.957**	0.435	0.264	0.382	0.509	0.541	0.654	0.645
Qwen2 72B	**0.659**	**0.957**	0.575	0.467	**0.676**	**0.741**	0.541	0.759	**0.803**
Llammas Base	0.427	0.913	0.386	0.232	0.307	0.343	0.559	0.541	0.597
Mudelite keskmine	0.565	0.944	0.517	0.356	0.502	0.585	0.557	0.692	0.724

Tabel 1. Erinevate baasmudelite täpsus õppeainete lõikes ja keskmiselt eksamite andmestiku peal. Võrdlusesse on lisatud juhuslikult valimise täpsus. Eesti keel 1 viitab eesti keele teise keelena eksami gümnaasiumi osale ja eesti keel 2 eesti keele teise keelena põhikooli osale.

Mudel	Täpsus Keskmine	Täpsus Ajalugu	Täpsus Bioloogia	Täpsus Eesti k 1	Täpsus Eesti k 2	Täpsus Füüsika	Täpsus Geograafia	Täpsus Keemia	Täpsus Ühisk.
Juhuslik	0.285	0.261	0.281	0.197	0.210	0.247	0.439	0.287	0.372
Gemma 2 9B	0.610	0.913	0.517	0.402	0.647	0.583	0.559	0.729	0.761
Gemma 2 27B	**0.677**	**0.957**	0.614	**0.508**	**0.655**	**0.722**	**0.631**	0.812	0.794
Llama 3.1 8B	0.391	0.913	0.329	0.211	0.244	0.352	0.523	0.429	0.600
Llama 3.1 70B	**0.664**	**0.957**	**0.643**	0.398	0.567	**0.713**	**0.631**	**0.857**	**0.852**
Mistral Nemo Base	0.480	**0.957**	0.401	0.268	0.399	0.472	0.541	0.639	0.642
Qwen2 72B	**0.658**	**0.957**	0.565	0.467	**0.681**	0.694	0.541	0.767	0.813
Llammas Base	0.420	0.826	0.372	0.248	0.286	0.380	0.586	0.549	0.561
Mudelite keskmine	0.557	0.925	0.491	0.358	0.497	0.560	0.573	0.683	0.718

Tabel 2. Erinevate baasmudelite normeeritud täpsus õppeainete lõikes ja keskmiselt eksamite andmestiku peal. Võrdlusesse on lisatud juhuslikult valimise täpsus. Eesti keel 1 viitab eesti keele teise keelena eksami gümnaasiumi osale ja eesti keel 2 eesti keele teise keelena põhikooli osale.

6.1.2 Juhendhäälestatud mudelid

Tabelis 3 on esitatud juhendhäälestatud mudelite täpsused eksami andmestikul. Parimad tulemused on saavutanud kommertsmudelid Claude 3.7 Sonnet, Gemini 2.0 Flash ja GPT-4o, mis saavutavad keskmiselt täpsuse üle 0,9. Samal ajal parima vabavaralise juhendhäälestatud mudeli, Llama 405B Instruct täpsus on 0,8. Võrreldes omavahel samade mudelite baas ning juhendhäälestatud versioone, on näha, et üldiselt annavad juhendhäälestatud mudelid eksami andmestiku peal paremaid tulemusi. Erandiks on Gemma 9B ning Llammas, mis on pisut paremad baasversioonis. Kui baasmudelite korral said parimad mudelid eesti keele teise keelena küsimuste puhul tulemused 0,5 ja 0,7 lähedale vastavalt gümnaasiumi ja põhikooli arvestuses, siis parimad juhendhäälestatud vabavaralised mudelid saavutavad neis tulemusi vastavalt 0,8 ja 0,9 lähedale ning parimad kommertsmudelid lausa üle 0,9. Sarnane kasvutrend esineb enamike õppeainete korral. Erandiks on ajalugu, kus juba parimad baasmudelid saavutasid täpsuse üle 0,95 ning parimad kommertsmudelid andsid võidu vaid vähem kui 0,01 võrra ning geograafia, kus nii baasmudelite kui ka parimate kommertsmudelite tulemused jäid alla 0,7.

Mudel	Täpsus Keskmine	Täpsus Ajalugu	Täpsus Bioloogia	Täpsus Eesti k 1	Täpsus Eesti k 2	Täpsus Füüsika	Täpsus Geograafia	Täpsus Keemia	Täpsus Ühisk.
Juhuslik	0.285	0.261	0.281	0.197	0.210	0.247	0.439	0.287	0.372
Gemma 2 9B It	0.556	0.605	0.512	0.362	0.412	0.676	0.541	0.744	0.690
Gemma 2 27B It	0.760	0.835	0.701	0.675	0.807	0.741	0.595	0.752	0.839
Llama 3.1 8B Instruct	0.542	0.701	0.435	0.390	0.462	0.491	0.514	0.549	0.684
Llama 3.1 70B Instruct	0.760	0.877	0.734	0.650	0.748	0.750	0.559	0.820	0.823
Llama 3.1 405B Instruct	0.817	0.866	0.763	0.785	0.895	0.806	0.568	0.850	0.858
Deepseek Chat V3	0.820	0.870	0.797	0.785	0.899	0.824	0.613	0.850	0.819
Mistral Nemo Instruct	0.570	0.709	0.454	0.472	0.525	0.472	0.559	0.549	0.690
Qwen2 72B Instruct	0.739	0.831	0.633	0.626	0.807	0.713	0.541	0.820	0.816
TartuNLP Llammas	0.401	0.544	0.367	0.264	0.273	0.343	0.523	0.361	0.507
Claude 3.7 Sonnet	**0.905**	**0.962**	**0.860**	**0.963**	**0.941**	**0.907**	0.613	0.940	**0.903**
Gemini 2.0 Flash 001	**0.900**	**0.962**	**0.884**	0.898	0.887	**0.917**	**0.667**	**0.970**	**0.916**
GPT-4o Mini	0.784	0.854	0.783	0.703	0.790	0.796	0.613	0.857	0.813
GPT-4o	**0.900**	**0.962**	**0.865**	0.911	**0.958**	0.861	0.622	0.940	**0.913**
Mudelite keskmine	0.727	0.814	0.676	0.653	0.723	0.715	0.579	0.769	0.790

Tabel 3. Erinevate juhendhäälestatud mudelite täpsus õppeainete lõikes ja keskmiselt eksamite andmestiku peal. Võrdlusesse on lisatud juhuslikult valimise täpsus. Eesti keel 1 viitab eesti keele teise keelena eksami gümnaasiumi osale ja eesti keel 2 eesti keele teise keelena põhikooli osale.

6.2 Trivia andmestik

6.2.1 Baasmudelid

Tabelis 4 on näha baasmudelite täpsused ja normeeritud täpsused trivia andmestikul. Mudelite täpsused jäävad üldiselt 0,3 ja 0,35 vahele, erandiks on baasmudelitest parima tulemuse saanud LLama 70B täpsusega 0,43. Üldiselt on mudelite normeeritud täpsus kõrgem kui tavaline täpsus, erandiks LLama mudelid ja Mistral Nemo Base. Paremate baasmudelite paremusjärjestus püsib üldiselt sama olenemata, kas võrdluse aluseks on täpsus või normeeritud täpsus, erinevus tuleb sisse just antud andmestiku peal halvemaid tulemusi saanud mudelite seas, kuid kuna nende mudelite tulemused on üksteisega väga sarnased, ei oma see erilist tähtsust.

Mudel	Täpsus	Normeeritud täpsus
Gemma 2 9B	0.295	0.304
Gemma 2 27B	0.311	0.331
Llama 3.1 8B	0.318	0.318
Llama 3.1 70B	**0.433**	**0.428**
Mistral Nemo Base	0.303	0.284
Qwen2 72B	0.328	0.34
Llammas Base	0.343	0.354

Tabel 4. Täpsus ja normeeritud täpsus baasmudelite lõikes trivia andmestikul

6.2.2 Juhendhäälestatud mudelid

Tabel 5 kujutab juhendhäälestatud mudelite täpsusi trivia andmestikul. Parimad tulemused on saanud kommertsmudelid Claude Sonnet ja GPT-4o, kummagi täpsuseks pisut üle 0,5. Neile järgneb LLama 405B Instruct, mille täpsus (0,44) on juba võrreldav parima baasmudeli, LLama 70B, täpsusega (0,43). Võrreldes samade mudelite baas ning juhendhäälestatud versioone, ei joonistu välja kindlat seaduspärasust - mõnel juhul annab pisut parema täpsuse baasmudel, siis jälle juhendhäälestatud versioon.

Mudel	Täpsus
Gemma 2 9B It	0.278
Gemma 2 27B It	0.325
Llama 3.1 8B Instruct	0.309
Llama 3.1 70B Instruct	0.359
Llama 3.1 405B Instruct	0.440
Deepseek Chat V3	0.398
Mistral Nemo Instruct	0.299
Qwen2 72B Instruct	0.376
TartuNLP Llammas	0.336
Claude 3.7 Sonnet	**0.511**
Gemini 2.0 Flash 001	0.429
GPT-4o Mini	0.370
GPT-4o	**0.513**

Tabel 5. Täpsus juhendhäälestatud mudelite lõikes trivia andmestikul

6.3 Sõnaseletuste andmestik

6.3.1 Baasmudelid

Tabelis 6 on esitatud baasmudelite täpsused sõnaseletuste andmestiku peal. Parima täpsuse (0,379) on saavutanud Llama 70B, millele kohe järgnevad Llammas Base (0,356) ja Gemma 27B (0,354). Ülejäänud baasmudelite täpsused jäävad alla 0,3, Llama 8B ja Qwen 72B alla 0,2.

6.3.2 Juhendhäälestatud mudelid

Tabel 7 kujutab juhendhäälestatud mudelite täpsusi sõnaseletuste andmestikul. Parimaid tulemusi on saavutanud kommertsmudelid Claude Sonnet, Gemini Flash ja GPT-4o, täpsusega üle 0,5, neile järgnevad vabavaralised mudelid Deepseek Chat ja LLama 405B

Mudel	Täpsus
Gemma 2 9B	0.259
Gemma 2 27B	**0.354**
Llama 3.1 8B	0.182
Llama 3.1 70B	**0.379**
Mistral Nemo Base	0.237
Qwen2 72B	0.188
Llammas Base	**0.356**

Tabel 6. Täpsus baasmudelite lõikes sõnaseletuste andmestikul

Instruct täpsustega üle 0,4. Võrreldes omavahel baas ja juhendhäälestatud mudeleid, võib märgata, et antud andmestiku korral on baasmudelite täpsused alati kõrgemad vastavate juhendhäälestatud versioonide täpsustest.

Mudel	Täpsus
Gemma 2 9B It	0.178
Gemma 2 27B It	0.266
Llama 3.1 8B Instruct	0.132
Llama 3.1 70B Instruct	0.306
Llama 3.1 405B Instruct	0.410
Deepseek Chat V3	0.452
Mistral Nemo Instruct	0.177
Qwen2 72B Instruct	0.152
TartuNLP Llammas	0.108
Claude 3.7 Sonnet	**0.570**
Gemini 2.0 Flash 001	**0.540**
GPT-4o Mini	0.331
GPT-4o	**0.555**

Tabel 7. Täpsus juhendhäälestatud mudelite lõikes sõnaseletuste andmestikul

6.4 Grammatika andmestik

6.4.1 Baasmudelid

Tabel 8 esitab baasmudelite täpsust ja Levenshteini kaugusel põhinevat meetrikat grammatika andmestikel. Andmestik 1 viitab andmestiku osale, mis hõlmab eesti keelt teise keelena rääkivate õpilaste kirjutatud lauseid ning andmestik 2 eesti keelt emakeelena kõnelejate lauseid. Näeme, et andmestik 1 peal on nii parima täpsuse kui Levenshteini meetrika väärtuse saavutanud Llammas Base, millele järgneb Llama 70B, samas andmestik 2 peal annab parimaid tulemusi Gemma 27B. Üldiselt on kõik täpsused väga madalad, jäädes alla 0,2, mis on selgitatav antud ülesande keerukusega, samuti tõsiasjaga, et antud

andmestik aktsepteerib korrektsena vaid üht konkreetset versiooni parandusest.

Mudel	Andmestik 1 *Täpsus*	Andmestik 1 *Levenshtein**	Andmestik 2 *Täpsus*	Andmestik 2 *Levenshtein**
Gemma 2 9B	0.076	0.193	0.103	0.332
Gemma 2 27B	0.129	0.232	**0.186**	**0.38**
Llama 3.1 8B	0.061	0.174	0.092	0.311
Llama 3.1 70B	**0.135**	**0.236**	**0.175**	0.346
Mistral Nemo Base	0.072	0.179	0.096	0.31
Qwen2 72B	0.078	0.187	0.137	0.328
Llammas Base	**0.143**	**0.241**	0.143	0.327

Tabel 8. Täpsus ja Levenshteini kaugusel baseeruv meetrika baasmudelite lõikes grammatika andmestikul. * Levenshteini meetrika on keskmine $\frac{1}{1+\text{Levenshteini kaugus}}$, kus Levenshteini kaugus leitakse korrektse ja genereeritud lause vastavate sõnepaaride vahel.

6.4.2 Juhendhäälestatud mudelid

Tabel 9 kujutab juhendhäälestatud mudelite tulemusi grammatika andmestikul. Parim mudel antud andmestikel on GPT-4o, saavutades täpsused vastavalt 0,257 andmestikul 1 ja 0,321 andmestikul 2, millele järgnevad teised kommertsmudelid Claude Sonnet ja Gemini Flash. Üldiselt on grammatika ülesannete peal baasmudelid edukamad kui juhendhäälestatud mudelid. Esineb siiski paar erandit, nagu Gemma 9B andmestik 1 peal ja Llama 8B andmestik 2 peal.

Mudel	Andmestik 1 *Täpsus*	Andmestik 1 *Levenshtein**	Andmestik 2 *Täpsus*	Andmestik 2 *Levenshtein**
Gemma 2 9B It	0.087	0.175	0.079	0.19
Gemma 2 27B It	0.108	0.201	0.121	0.245
Llama 3.1 8B Instruct	0.053	0.149	0.094	0.23
Llama 3.1 70B Instruct	0.102	0.188	0.099	0.21
Llama 3.1 405B Instruct	0.16	0.249	0.186	0.323
Deepseek Chat V3	0.154	0.241	0.139	0.272
Mistral Nemo Instruct	0.062	0.162	0.065	0.213
Qwen2 72B Instruct	0.009	0.053	0.016	0.066
TartuNLP Llammas	0.078	0.191	0.103	0.329
Claude 3.7 Sonnet	**0.215**	**0.316**	**0.296**	**0.466**
Gemini 2.0 Flash 001	**0.213**	0.293	0.157	0.271
GPT-4o Mini	0.174	0.269	0.224	0.372
GPT-4o	**0.257**	**0.344**	**0.321**	**0.446**

Tabel 9. Täpsus ja Levenshteini kaugusel baseeruv meetrika juhendhäälestatud mudelite lõikes grammatika andmestikul. * Levenshteini meetrika on keskmine $\frac{1}{1+\text{Levenshteini kaugus}}$, kus Levenshteini kaugus leitakse korrektse ja genereeritud lause vastavate sõnepaaride vahel.

6.5 Käänamise andmestik

6.5.1 Baasmudelid

Tabelis 10 on kujutatud baasmudelite tulemused käänamise andmestiku peal. Baasmudelitest parim täpsus 0,57 on LLama 70B-l, millele jägnevad Gemma 27B ning Llammas Base.

Mudel	Täpsus
Gemma 2 9B	0.396
Gemma 2 27B	**0.505**
Llama 3.1 8B	0.268
Llama 3.1 70B	**0.586**
Mistral Nemo Base	0.416
Qwen2 72B	0.306
Llammas Base	0.473

Tabel 10. Täpsus baasmudelite lõikes käänamise andmestikul

6.5.2 Juhendhäälestatud mudelid

Tabel 11 esitab juhendhäälestatud mudelite täpsus käänamise andmestikul. Parimaid tulemusi on saavutanud kommertsmudelid Claude Sonnet, Gemini Flash ja GPT-4o, täpsusega üle 0,93. Võrreldes nendega on parima vabavaralise mudeli täpsus (LLama 405B Instruct - 0,680) on tunduvalt madalam, kuid siiski umbes 0,1 võrra kõrgem kui parima baasmudeli täpsus. Samal ajal võrreldes samade mudelite baas ning juhendhäälestatud versioone, on baasversioonid kõigi võrdluses olevate mudelite korral antund andmestiku korral paremad. Vahed baas ning juhendhäälestatud versioonid täpsustes on ka küllaltki märkimisväärsed, olles enamikel juhtudel üle 0,1 ja mõnel juhul ka üle 0,2.

6.6 Dialoogide kokkuvõtete andmestik

6.6.1 Baasmudelid

Tabel 12 kujutab baasmudelite ROUGE mõõdikuid dialoogide summeerimise andmestikul. Parimad ROUGE skoorid on antud ülesandel saavutanud Qwen 72B ning Gemma 27B, neist esimene on parim ROUGE-1 kui ROUGE-L arvestuses ning teine ROUGE-2 arvestuses. Tulemused erinevad omavahel üpris vähe, näiteks ROUGE-1 skoori osas on vahe parima ja halvima baasmudeli vahel vaid 0,05, illustreerides mudelite küllaltki sarnast

Mudel	Täpsus
Gemma 2 9B It	0.308
Gemma 2 27B It	0.406
Llama 3.1 8B Instruct	0.099
Llama 3.1 70B Instruct	0.464
Llama 3.1 405B Instruct	0.680
Deepseek Chat V3	0.631
Mistral Nemo Instruct	0.274
Qwen2 72B Instruct	0.177
TartuNLP Llammas	0.230
Claude 3.7 Sonnet	**0.937**
Gemini 2.0 Flash 001	**0.932**
GPT-4o Mini	0.554
GPT-4o	**0.948**

Tabel 11. Täpsus juhendhäälestatud mudelite lõikes käänamise andmestikul

võimekust antud summeerimise ülesannet lahendada.

Mudel	ROUGE-1	ROUGE-2	ROUGE-L
Gemma 2 9B	0.240	0.070	0.204
Gemma 2 27B	**0.245**	**0.075**	**0.209**
Llama 3.1 8B	0.217	0.056	0.184
Llama 3.1 70B	0.242	0.072	0.203
Mistral Nemo Base	0.226	0.063	0.194
Qwen2 72B	**0.251**	**0.073**	**0.211**
Llammas Base	0.233	0.067	0.199

Tabel 12. ROUGE mõõdikud baasmudelite lõikes dialoogide kokkuvõtete andmestikul

6.6.2 Juhendhäälestatud mudelid

Tabelis 13 on näidatud juhendhäälestatud mudelite tulemused dialoogide kokkuvõtete andmestikul. Kõigi juhendhäälestatud mudelite ROUGE skoorid jäävad sellel ülesandel alla baasmudelite skooridele, üllatuslikult isegi kommertsmudelite omad. Parim juhendhäälestatud mudel, GPT-4o saavutas ROUGE-1 skoori 0,206, kui samal ajal on halvima baasmudeli ROUGE-1 skoor 0,217 ning parimal 0,251.

Mudel	ROUGE-1	ROUGE-2	ROUGE-L
Gemma 2 9B It	0.147	0.039	0.123
Gemma 2 27B It	0.196	0.049	**0.163**
Llama 3.1 8B Instruct	0.177	0.042	0.143
Llama 3.1 70B Instruct	**0.201**	**0.053**	**0.164**
Llama 3.1 405B Instruct	0.188	0.050	0.155
Deepseek Chat V3	0.201	0.050	**0.165**
Mistral Nemo Instruct	0.195	**0.052**	0.158
Qwen2 72B Instruct	0.194	0.042	0.154
TartuNLP Llammas	0.163	0.038	0.131
Claude 3.7 Sonnet	0.193	**0.052**	0.157
Gemini 2.0 Flash 001	**0.204**	0.050	**0.167**
GPT-4o Mini	0.192	0.045	0.154
GPT-4o	**0.206**	**0.055**	**0.173**

Tabel 13. ROUGE mõõdikud juhendhäälestatud mudelite dialoogide kokkuvõtete andmestikul

6.7 Uudissaadete kokkuvõtete andmestik

6.7.1 Baasmudelid

Tabel 14 esitab baasmudelite tulemusi uudissaadete kokkuvõtete andmestikul. Parimad ROUGE skoorid on saavutanud Llama 70B, Gemma 27B ja Gemma 9B, neist parimad ROUGE-1 ja ROUGE-2 Llama 70B (vastavalt 0,191 ja 0,07) ning ROUGE-L skoori Gemma 27B.

Mudel	ROUGE-1	ROUGE-2	ROUGE-L
Gemma 2 9B	**0.1802**	**0.0624**	**0.1569**
Gemma 2 27B	**0.1898**	**0.0676**	**0.1643**
Llama 3.1 8B	0.1506	0.0461	0.1263
Llama 3.1 70B	**0.1906**	**0.0711**	**0.1603**
Mistral Nemo Base	0.1579	0.049	0.1295
Qwen2 72B	0.1603	0.0539	0.1335
Llammas Base	0.1671	0.0552	0.1384

Tabel 14. ROUGE mõõdikud baasmudelite lõikes uudislugude summeerimise andmestikul

6.7.2 Juhendhäälestatud mudelid

Tabelis 15 on kujutatud juhendhäälestatud mudelite ROUGE mõõdikud uudislugude kokkuvõtete andmestikul. Claude Sonnet on saavutanud juhendhäälestatud mudelitest parimad

ROUGE skoorid, mis jäävad aga alla 3 parima baasmudeli Llama 70B, Gemma 27B ja Gemma 9B tulemustele. Samade mudelite baas ning juhendhäälestatud versioonide tulemuste võrdluses osutuvad peaaegu alati pisut paremaks baasversioonid, ainsaks erandiks Llama 8B.

Mudel	ROUGE-1	ROUGE-2	ROUGE-L
Gemma 2 9B It	0.165	0.042	0.128
Gemma 2 27B It	0.164	0.043	0.128
Llama 3.1 8B Instruct	0.154	0.041	0.119
Llama 3.1 70B Instruct	0.165	**0.049**	0.129
Llama 3.1 405B Instruct	0.165	**0.049**	0.129
Deepseek Chat V3	0.151	0.037	0.113
Mistral Nemo Instruct	0.133	0.034	0.034
Qwen2 72B Instruct	0.146	0.037	0.112
TartuNLP Llammas	0.125	0.033	0.092
Claude 3.7 Sonnet	**0.174**	**0.050**	**0.135**
Gemini 2.0 Flash 001	0.164	0.164	0.124
GPT-4o Mini	0.153	0.040	0.115
GPT-4o	0.162	0.042	0.127

Tabel 15. ROUGE mõõdikud juhendhäälestatud mudelite uudislugude summeerimise andmestikul

6.8 Kõnelejatuvastuse andmestik

6.8.1 Baasmudelid

Tabelis 16 on kujutatud baasmudelite täpsused ja saagised nii 0 kui 5 näitega konfiguratsioonides kõnelejatuvastuse andmestikul. 0 näitega konfiguratsiooni korral on parim täpsus Qwen 72B mudelil (0,718), samal ajal parim saagis on Llama 70B mudelil (0,590), millele järgneb Qwen 72B (0,501). Seega parima täpsuse-saagise kombinatsiooniga baasmudel 0 näite korral on Qwen 72B. Üllatavalt saavad selle konfiguratsiooni puhul Llama mudelite korral väiksemad mudelid paremaid tulemusi. 5 näitega konfiguratsiooni korral saavutavad mitmed mudelid täpsuse üle 0,85, neist kõrgeima Qwen 72B ning Mistral Nemo Base (0,95). Samal ajal saagised tulevad kõigil baasmudelitel 5 näite korral nullilähedased. See tuleneb asjaolust, et selle konfiguratsiooni korral pakkusid mudelid vaid mõne üksiku kõneleja kogu transkriptsiooni kohta, kes aga osutusid sageli õigeteks. Kokkuvõttes said mudelid ülesandega terviklikumalt hakkama 0 näitega variandis.

Mudel	Täpsus (0 näitega)	Saagis (0 näitega)	Täpsus (5 näitega)	Saagis (5 näitega)
Gemma 2 9B	liiga väike kontekstipikkus			
Gemma 2 27B	liiga väike kontekstipikkus			
Llama 3.1 8B	0.582	0.474	**0.850**	**0.047**
Llama 3.1 70B	0.516	**0.590**	**0.900**	**0.050**
Mistral Nemo Base	0.497	0.457	**0.950**	**0.053**
Qwen2 72B	**0.718**	**0.501**	**0.950**	**0.053**
Llammas Base	0.007	0.003	0.014	0.006

Tabel 16. Täpsus ja saagis baasmudelite lõikes nii 0 kui 5 näitega konfiguratsioonides kõnelejatuvastuse andmestikul

6.8.2 Juhendhäälestatud mudelid

Tabel 17 esitab juhendhäälestatud mudelite täpsuseid ja saagiseid kõnelejatuvastuse andmestiku korral (0 näitega konfiguratsioonis). Gemma mudelid ei andnud antud juhul üldse tulemusi, kuna nende kontekstipikkus oli antud andmestiku jaoks liiga väike. Parima täpsuse antud ülesandel saavutas GPT-4o (0,855), millele järgnes Claude Sonnet (0,829), samal ajal parima saagise saavutas Gemini Flash (0,796), millele järgnesid Claude Sonnet (0,746) ning Llama 405B (0,724). Parima täpsuse-saagise kombinatsiooni andis seega Claude Sonnet. Kui võrrelda omavahel parimaid vabavaralisi juhendhäälestatud mudeleid ning baasmudeleid, võib märgata, et saavutatud täpsused on neil üpris sarnased, jäädes 0,7 ja 0,75 vahele. Samas saagised on aga sel juhul baasmudelitel üle 0,1 võrra madalamad.

Mudel	Täpsus (0 näitega)	Saagis (0 näitega)
Gemma 2 9B It	liiga väike kontekstipikkus	
Gemma 2 27B It	liiga väike kontekstipikkus	
Llama 3.1 8B Instruct	0.444	0.585
Llama 3.1 70B Instruct	0.725	0.630
Llama 3.1 405B Instruct	0.681	**0.724**
Deepseek Chat V3	0.746	0.605
Mistral Nemo Instruct	0.400	0.421
Qwen2 72B Instruct	0.548	0.642
TartuNLP Llammas	0.000	0.000
Claude 3.7 Sonnet	**0.829**	**0.746**
Gemini 2.0 Flash 001	0.706	**0.796**
GPT-4o Mini	0.685	0.621
GPT-4o	**0.855**	0.628

Tabel 17. Täpsus ja saagis juhendhäälestatud mudelite lõikes kõnelejatuvastuse andmestikul

6.9 Tulemuste võrdlus inglise keelest masintõlgitud andmestikuga

Eesmärgiks on võrrelda mudelite tulemusi eksami andmestikul olevate õppeainete lõikes vastavate eesti keelde masintõlgitud MMLU andmestikust valitud kategooriatega [5]. Selleks on valitud MMLU 57 kategooria seast need, mis kattuvad kõige paremini eksamite andmestikus olevate õppeainetega: Euroopa ajalugu, bioloogia, füüsika, geograafia, keemia ning valitsus ja poliitika, millest on valitud keskkooli tasemel versioonid. Neid võrreldakse eksami andmestiku tulemustest (tabel 1) vastavalt ajaloo, bioloogia, füüsika, geograafia, keemia ning ühiskonnaõpetuse alamosade tulemustega.

6.9.1 Baasmudelid

Tabel 18 kujutab hinnatavate baasmudelite tulemusi eesti keelde masintõlgitud MMLU andmestikust valitud kategooriate peal. Võrreldes baasmudelite keskmisi täpsusi sel andmestikul ning eksamite andmestikul, selgub, et mõlema puhul on parimaid tulemusi andnud Gemma 27B, Llama 70B ning Qwen 72B, kusjuures parim mudel eksamite andmestiku peal on Gemma 27B ning MMLU peal Qwen 72B. Kõik mudelid on keskmiselt saavutanud eksamite andmestiku peal kõrgema täpsuse, seda mudelite lõikes keskmiselt natuke rohkem 0,1 võrra. Võib ka märgata, et MMLU andmestiku peal kõiguvad parimate ning halvimate mudelite keskmised täpsused tunduvalt rohkem kui eksamite andmestikul. MMLU andmestikul saavutavad halvimad mudelid vaid veidi parema tulemuse juhuslikult valimise täpsusest, Llammas Base mudeli täpsus jääb isegi sellest veidi alla.
Kategooriate lõikes on mudelid keskmiselt tulnud eksamite andmestiku peal kõige paremini toime ajaloo küsimustega, millele järgneb ühiskonnaõpetus ning keemia. MMLU puhul aga on mudelite keskmine täpsus kõrgeim ajalool, millele järgneb bioloogia ning geograafia. Eri kategooriate keskmised tulemused andmestiku lõikes kõiguvad tunduvalt rohkem eksamite andmestiku korral.

6.9.2 Juhendhäälestatud mudelid

Tabel 19 kujutab juhendhäälestatud mudelite tulemusi valitud kategooriate lõikes masintõlgitud MMLU peal. Mõlemal andmestikul on keskmiselt parimaid tulemusi saanud mudelid Claude Sonnet, Gemini Flash ja GPT-4o, mis MMLU andmestikul saavutasid kõik täpsuse üle 0,8 ning eksamite anmestikul üle 0,9. Samuti, sarnaselt baasmudelite tulemustele, on ka juhendhäälestatud mudelite täpsused keskmiselt umbes 0,1 võrra kõrgemad eksamite andmestikul kui MMLU-l. Sellest võib järeldada, et MMLU andmestik on mudelitele pisut väljakutsuvam kui eksamite oma. See tulemus on üpriski oodatav, arvestades, et MMLU

Mudel	Täpsus Keskmine	Täpsus Ajalugu	Täpsus Bioloogia	Täpsus Füüsika	Täpsus Geograafia	Täpsus Keemia	Täpsus Poliitika
Juhuslik	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Gemma 2 9B	0.464	0.576	0.616	0.377	0.505	0.409	0.415
Gemma 2 27B	**0.599**	**0.715**	**0.719**	0.424	0.662	**0.562**	0.627
Llama 3.1 8B	0.295	0.364	0.365	0.305	0.308	0.281	0.218
Llama 3.1 70B	**0.58**	**0.727**	**0.703**	0.444	0.667	0.493	0.596
Mistral Nemo Base	0.295	0.497	0.371	0.318	0.313	0.251	0.223
Qwen2 72B	**0.656**	**0.685**	**0.732**	**0.517**	**0.742**	**0.571**	**0.715**
Llammas Base	0.245	0.285	0.316	0.225	0.288	0.197	0.197
Mudelite keskmine	0.423	0.512	0.509	0.357	0.467	0.377	0.405

Tabel 18. Erinevate baasmudelite täpsus valitud kategooriate lõikes ja keskmiselt eesti keelde tõlgitud MMLU andmestiku peal. Võrdlusesse on lisatud juhuslikult valimise täpsus. Kategooriad vastavad andmestikus olevatele keskkooli tasemel olevatele ainetele. Ajalugu kujutab endast konkreetselt Euroopa ajalugu.

vastavad kategooriad käsitlevad keskkooli tasemel teemasid, kuid eksamite andmestiku omad põhikooli omasid. Osa sellest erinevusest on ka selgitatav sellega, et juhuslikult vastamise tõenäosus on eksamite andmestikul kõrgem. Kui võrrelda baasmudelite tulemusi juhendhäälestatud mudelite tulemustega, võib märgata, et juhendhäälestatud mudelite täpsus on vastavast baasmudelite täpsusest keskmiselt umbes 0,2 võrra kõrgem mõlema andmestiku korral.

Juhendhäälestatud mudelid on MMLU korral saanud keskmiselt parimaid tulemusi keskkooli bioloogia kategoorias, millele järgnevad keskkooli ajalugu ning geograafia. Eksamite andmestiku korral olid juhendhäälestatud mudelitel keskmiselt parimad tulemused vastavalt ajaloo, ühiskonnaõpetuse ning keemia alamosadel. MMLU andmestikul saavutasid igas alamkategoorias keskmiselt parema tulemuse kindla vahega juhendhäälestatud mudelid, kuid eksamite andmestiku korral olid juhendhäälestatud mudelid mõne kategooria arvestuses (geograafia, ühiskonnaõpetus) vaid pisut paremad kui baasmudelid ning ajaloo küsimustes saavutasid baasmudelid keskmiselt kõrgema täpsuse.

Kokkuvõttes on parimaid tulemusi saanud mudelid samad olenemata, kas neid on hinnatud eksamite või eesti keelde masintõlgitud MMLU peal. Oodatavalt esineb erinevusi kategooriate lõikes, kuna nende andmestike küsimused on erineva sisu ning raskusastmega, samuti kõigub eksamite andmestiku korral valikvastuste hulk, mis on MMLU korral aga standartselt neli. Baasmudelite kontekstis esines parimate ja halvimate mudelite täpsuste vahel tunduvalt rohkem kõikumist MMLU andmestikul, mistõttu võib nende mudelite korral olla see eelistatum võrdlustest mudelite erineva võimekuse väljatoomiseks.

6.10 Tulemuste korreleeruvus

Tabel 20 kujutab eesti keelde masintõlgitud MMLU ja eestikeelsetest eksamitest kokkupandud eksamite andmestike vahelisi Pearsoni ja Spearmani korrelatsioonikordajaid. Kokku-

Mudel	Täpsus Keskmine	Täpsus Ajalugu	Täpsus Bioloogia	Täpsus Füüsika	Täpsus Geograafia	Täpsus Keemia	Täpsus Poliitika
Juhuslik	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Gemma 2 9B It	0.613	0.539	0.732	0.483	0.692	0.517	0.643
Gemma 2 27B It	0.7	0.764	0.829	0.55	0.737	0.611	0.772
Llama 3.1 8B Instruct	0.465	0.673	0.555	0.344	0.546	0.424	0.456
Llama 3.1 70B Instruct	0.664	0.697	0.858	0.45	0.732	0.591	0.689
Llama 3.1 405B Instruct	0.786	**0.842**	**0.903**	0.662	**0.843**	0.709	0.814
Deepseek Chat V3	0.718	0.752	0.803	0.649	0.742	0.69	0.705
Mistral Nemo Instruct	0.448	0.655	0.552	0.305	0.54	0.379	0.466
Qwen2 72B Instruct	0.739	**0.824**	0.797	0.589	**0.823**	0.655	0.829
TartuNLP Llammas	0.332	0.473	0.368	0.258	0.444	0.232	0.358
Claude 3.7 Sonnet	**0.863**	**0.842**	**0.916**	**0.735**	**0.894**	**0.818**	**0.953**
Gemini 2.0 Flash 001	**0.866**	**0.818**	**0.929**	**0.795**	**0.859**	**0.813**	**0.933**
GPT-4o Mini	0.817	0.867	0.923	0.629	0.864	0.749	0.922
GPT-4o	**0.671**	**0.788**	**0.79**	0.437	**0.737**	0.611	**0.777**

Tabel 19. Erinevate juhendhäälestatud mudelite täpsus valitud kategooriate lõikes ja keskmiselt eesti keelde tõlgitud MMLU andmestiku peal. Võrdlusesse on lisatud juhuslikult valimise täpsus. Kategooriad vastavad andmestikus olevatele keskkooli tasemel olevatele ainetele. Ajalugu kujutab endast konkreetselt Euroopa ajalugu.

võtvate tulemuste puhul on korrelatsioon üpris kõrge, Pearsoni kordajaga 0,95 ning Spearmani omaga 0,90. Andmestike vahel kõige tugevamalt korreleeruvam kategooria on keemia, mida selgitab ilmselt küsimuste sisu tehnilisus ning sarnane ülesannete tüüp andmestike vahel. Teistest kategooriatest pisut halvemini korreleerub füüsika, kus eksamite andmestiku puhul on märgatav osa küsimustest mõistete definitsioonide kohta, mis on üpris keeletundlik ülesandetüüp. Teistest kategooriatest märgatavalt halvemini korreleerub aga poliitika / ühiskonnaõpetuse kategooria. See tulemus on oodatav, kuna Eesti riigieksamite põhjal loodud andmestiku ühiskonnaõpetuse kategooriale pole MMLU-s täpset vastet ning keskkooli poliitika kategooria on MMLU-st küll lähim vaste ühiskonnaõpetusele, kuid siiski sisuliselt üpris erinev ja on kõrvutatud vaid tinglikult.

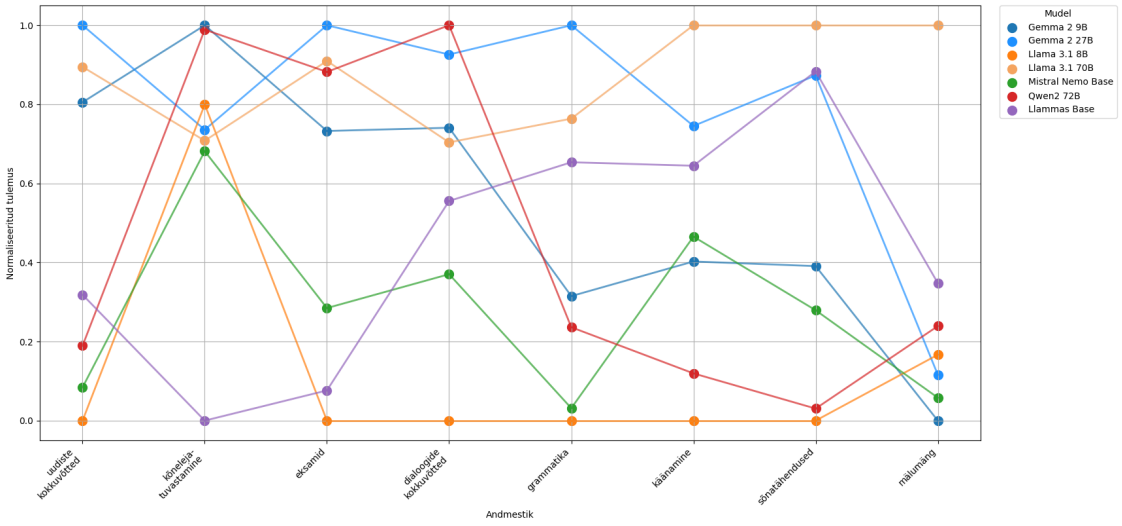
Kategooria	**Pearsoni kordaja**	**Spearmani kordaja**
Keskmine	0.95	0.90
Ajalugu	0.93	0.76
Bioloogia	0.93	0.89
Füüsika	0.87	0.83
Geograafia	0.90	0.93
Keemia	0.98	0.95
Poliitika / Ühiskonnaõpetus	0.76	0.76

Tabel 20. MMLU ja eksamite andmestike tulemuste korreleeruvus

6.11 Kokkuvõte

6.11.1 Baasmudelid

Joonis 1 kujutab kokkuvõtlikult baasmudelite tulemusi eri andmestike lõikes. Tulemused on normeeritud 0 ja 1 vahele, nõnda et iga andmestiku halvim tulemus võrdsustatakse 0-ga ja parim 1-ga. Võib märgata, et mudelite paremusjärjestus eri andmestikel on üpris muutlik. Esile torkavad kõnelejatuvastuse ning mälumängu andmestike tulemused, kus esimesel on Llammas Base tulemus tunduvalt halvem võrreldes teiste tulemustega ning viimasel Llama 70B tulemus teistest tunduvalt parem. LLama 8B on saanud enamike andmestike peal kõige halvema tulemuse, erandiks need samad kaks esiletorkavad andmestikku. Läbivalt suhteliselt madalaid tulemusi on saanud ka Mistral Nemo Base. Llammas Base ning Qwen 72B tulemused on väga kõikuvad, mõne andmestiku peal on need parimate mudelite seas, teiste peal jälle halvimate ning huvitavalt paistab Qwen 72B olevat parem neis ülesannetes, milles Llammas Base on halvem ning vastupidi. Üldistades on parimaid tulemusi läbivalt saanud Gemma 27B ja Llama 70B, erandiks mälumängu andmestik, kus Llama 70B sai teistest mudelitest tunduvalt parema tulemuse ning Gemma 27B suhteliselt keskpärase. Kui võrrelda mudelite paremusjärjestust sarnaste andmestikepaaride peal, nagu näiteks uudislugude ja dialoogide summeerimine või eksamid ja mälumäng, siis üllatuslikult võib märgata, et need on küllaltki erinevad, ning ei tule välja kindlat seost nende tulemuste vahel.

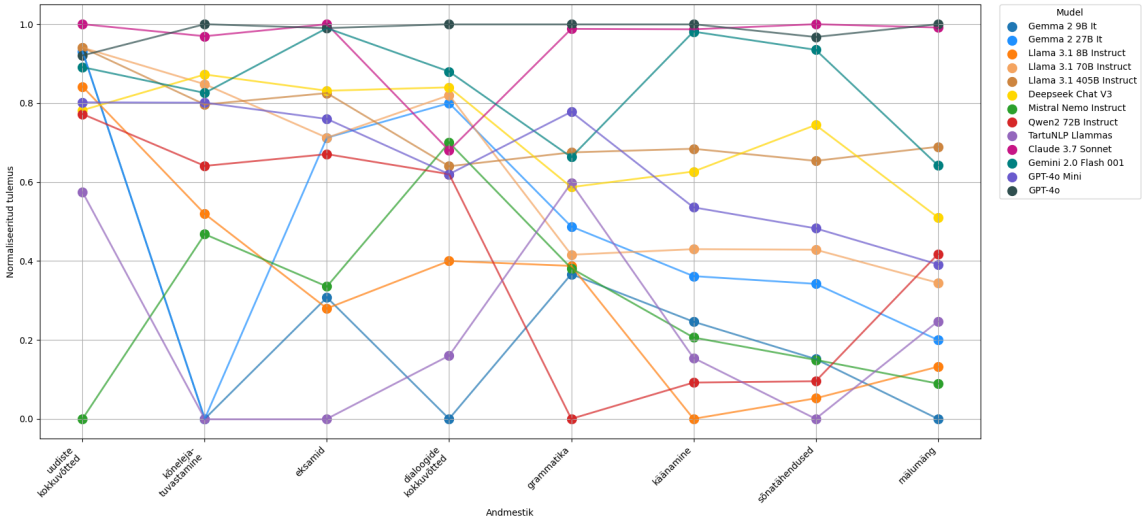


Joonis 1. Baasmudelite normaliseeritud tulemused erinevate andmestike lõikes.

6.11.2 Juhendhäälestatud mudelid

Joonis 2 kujutab kokkuvõtlikult juhendhäälestatud mudelite tulemusi eri andmestike lõikes. Graafikult torkab silma GPT-4o, mis on konstantselt saanud väga kõrgeid tulemusi,

enamikel andmestikel kõigist mudelitest kõige parema tulemuse. Sarnaselt tugev mudel on ka Claude Sonnet, mis on samuti saanud kõrgeid tulemusi kõigi andmestike peal peale dialoogide kokkuvõtete andmestiku. Paremaid tulemusi on lisaks saanud veel Gemini Flash, Llama 405B, Deepseek Chat V3 ning GPT-4o Mini. Parimate kommertsmudelite tulemuste domineerimine teiste juhendhäälestatud mudelite eest tuleb enim välja just käänamise, sõnaseletuste ning mälumängu andmestike puhul. Halvimaid tulemusi saanud mudel kõigub andmestike lõikes suhteliselt palju, kolmel korral on selleks Llammas, kahel Gemma 9B ning ülejäänud juhtudel erinevad mudelid. Üldiselt halvimaid tulemusi on saanud väiksemad mudelid, Gemma 9B, Llammas, Llama 8B ning Mistral Nemo Base. Parimad vabavaralised juhendhäälestatud mudelid jäävad parimatele kommertsmudelitele alla, kuid konkureerivad kehvemate kommertsmudelitega, nagu GPT-4o Mini.



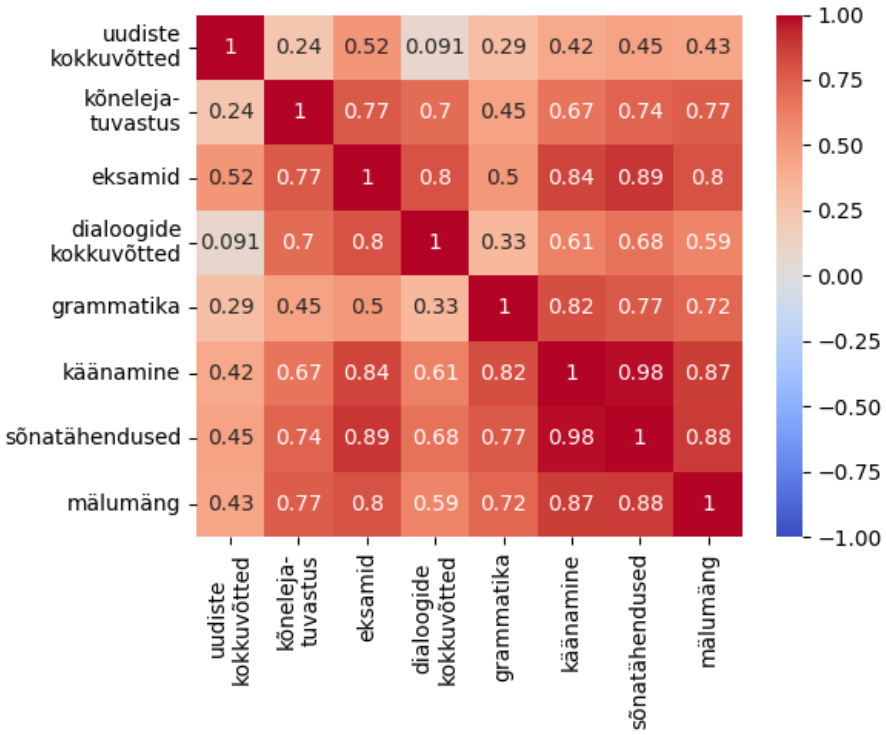
Joonis 2. Juhendhäälestatud mudelite normaliseeritud tulemused erinevate andmestike lõikes.

6.11.3 Andmestikevaheline korrelatsioon

Joonis 3 kujutab Pearsoni korrelatsioonikordajaid eri andmestike tulemuste vahel. Selgub, et ühegi andmestiku tulemuste vahel pole negatiivset korrelatsiooni. Kõige tugevam korrelatsioon esineb käänete ja sõnaseletuste andmestike tulemuste vahel, millevaheline korrelatsioonikordaja on 0,98. Käänete andmestiku tulemused korreleeruvad tugevalt veel lisaks eksamite, grammatika ning mälumängu tulemustega. Sõnaseletuste andmestik korreleerub samuti lisaks tugevalt eksamite ning mälumängu tulemustega, kuid pisut väiksemal määral mälumängu tulemustega. Eksamite andmestik korreleerub suuresti veel lisaks eelmainitutele ka dialoogide kokkuvõtete ning mälumängu andmestikuga. Grammatika andmestiku korreleerumine teiste andmestikega on üpriski diferentseeruv, kusjuures mõnel juhul on korrelatsioon tugev, kuid teistel juhtudel keskmine või nõrk. Ka kõnelejatuvastuse ning dialoogide kokkuvõtete andmestikud käituvad teiste andmestike suhtes sarnaselt. Kõige nõrgemalt teiste tulemustega korreleeruv andmestik on uudislugude kokkuvõtete oma, millel on suurim korrelatsioonikordaja (0,52) eksamite andmestikega, kuid vähim (0,091) dialoogide kokkuvõtete andmestikuga. Kuna aga dialoogide ning uudislugude kokkuvõtete genereerimine on ülesandetüübilt samad, näitab nendevaheline praktiliselt olematu korrelatsioon seda, et nende hindamiseks kasutatav ROUGE meetrika ei ole kuigi usaldusväärne keelemudelite headuse hindamiseks selle ülesande kontekstis.

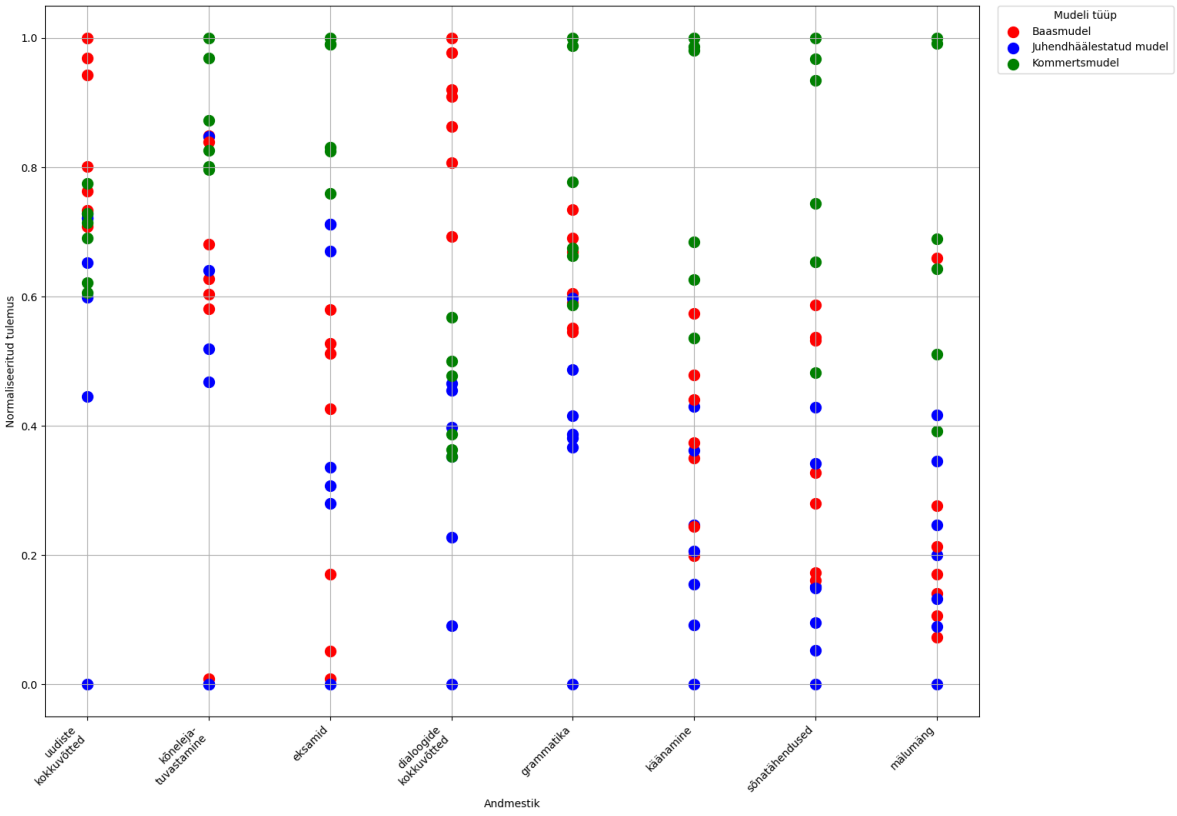
6.11.4 Tulemuste erinevus mudeli tüübi järgi

Joonis 4 kujutab mudelite tulemusi andmestike lõikes mudeli tüübi – baasmudel, juhendhäälestatud mudel või kommertsmudel – alusel. Enamiku andmestike korral on parimaid



Joonis 3. Erinevate andmestike tulemuste vaheline Pearsoni korrelatsioonimaatriks

tulemusi saanud kommertsmudelid, millele järgnevad mõningad baasmudelid ning seejärel juhendhäälestatud mudelid. Kõige erandlikumad on tulemused kokkuvõtete genereerimise andmestike peal, kus on selgelt parimaid tulemusi saanud just juhendhäälestatud mudelid ning ka parimad kommertsmudelid jäävad parimatele vabavaralistele juhendhäälestatud mudelitele kindlalt alla. Lisaks eristub teistest veel eksamite andmestik, mille korral on küll kommertsmudelid saanud kõrgemaid tulemusi, kuid mõni juhendhäälestatud mudel edestab parimaid baasmudeleid. Kõigi andmestike, v.a kokkuvõtete genereerimise omade, korral on paar-kolm kommertsmudelit, mis on saanud teistest mudelitest märgatavalt kõrgemaid tulemusi.



Joonis 4. Tulemused eri andmestikel mudeli tüübi järgi.

7. Vabavaralise mudeli peenhäälestamine

7.1 Peenhäälestamise protsess

Uurimaks peenhäälestamise mõju erinevate testülesannete tulemustele, viidi läbi kontseptsiooni tõestuse (*proof-of-concept*) tüüpi eksperiment, milles peenhäälestati Llama 3.1 8B baasmudelit Eesti keele koondkorpuse 1990-2008 (`https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et`) põhjal. Kuid on oluline rõhutada, et see korpus on suhteliselt väike ja eksperiment viidi läbi ainult ühe epohhi vältel ning ilma hüperparameetrite täiendava optimeerimiseta. Seetõttu tuleb saadud tulemusi käsitleda esialgse indikatsioonina, mitte lõpliku hindamisena mudeli maksimaalsest potentsiaalist.

7.2 Tulemuste võrdlus

7.2.1 Lühikese ja konkreetse vastuseformaadiga ülesanded

Eksamite, mälumängu, sõnaseletuste ja käänamise andmestikud hõlmavad ülesandeid, mis ootavad mudelilt lühikest ja konkreetset vastust. Just selliste ülesannete puhul avaldus peenhäälestamise positiivne mõju kõige selgemalt.

Eksamite andmestikus (tabel 21) tõusis peenhäälestamise tulemusel mudeli kategooriate keskmine täpsus 0,405 pealt 0,489 peale. Kuigi mõnede kategooriate täpsus jäi samaks (nt eesti keel teise keelena 1 ja füüsika) või isegi langes (nt ajalugu), oli enamusel siiski märgata paranemist. Eriti silmapaistev oli tõus ühiskonnaõpetuse küsimustes, mis on mõneti üllatav, arvestades selle sarnasust ajalooga, kus täpsus hoopis langes. See viitab võimalusele, et andmestike sisu ja ülesehitus mängivad olulist rolli peenhäälestamise mõjus.

Tabel 22 kujutab mudeli täpsusi enne ja peale peenhäälestamist mälumängu, sõnaseletuste ning käänamise andmestike korral. Näeme, et ka nende andmestike korral parandas peenhäälestamine tulemusi, kõige tagasihoidlikum täpsuse kasv esineb mälumängu korral (u 0,03 võrra) ning kõige tugevam käänamise korral (u 0,24 võrra). Sellest järeldub, et mudel omandas peale peenhäälestamist parema keelelise tundlikkuse ning suutlikkuse lühivastuste genereerimisel.

See tulemus viitab, et isegi minimaalne peenhäälestus eesti keele baasil annab

mudelile juurde teadmisi ning parandab selle võimekust kitsastes, hästi defineeritud keeleülesannetes. See kinnitab hüpoteesi, et sobiva jõudlustestide komplekti kaudu saab peenhäälestamise tulemusi ka kvantitatiivselt jälgida ja võrrelda.

Mudel	Keskmine	Ajalugu	Bioloogia	Eesti k 1	Eesti k 2	Füüsika	Geograafia	Keemia	Ühisk.
Llama 3.1 8B	0.405	0.957	0.362	0.224	0.248	0.389	0.514	0.451	0.603
Peenhäälestatud	0.489	0.728	0.391	0.224	0.315	0.389	0.577	0.556	0.729

Tabel 21. Mudelite täpsuste võrdlus eksami andmestikul

Mudel	Mälumäng	Sõnaseletused	Käänamine
Llama 3.1 8B	0.318	0.182	0.268
Peenhäälestatud	0.334	0.320	0.506

Tabel 22. Mudelite täpsuste võrdlus mälumängu, sõnaseletuste ja käänamise andmestikel

7.2.2 Keerulisema struktuuriga vastuseformaadiga ülesanded

Erinevalt eelnevatest ülesannetest nõuavad grammatika, dialoogide, uudislugude kokkuvõtete ning kõnelejatuvastuse andmestikud mudelilt pikemaid ja struktuuriliselt keerukamaid väljundeid. Siin oli peenhäälestamise mõju valdavalt negatiivne.

Tabel 23 kujutab mudeli täpsust grammatika andmestikel enne ning pärast peenhäälestamist. Mõlema grammatika andmestiku alamosa peal peenhäälestamise tulemusel nii täpsus kui ka Levenshteini meetrika väärtus langevad ning juba enne nullilähedased olnud täpsused lähevad päris nulliks.

Tabel 24 kujutab ROUGE-skoore dialoogide ning uudislugude andmestike peal enne ning pärast uuritava mudeli peenhäälestamist. Võib märgata, et kummagi andmestiku puhul ja kõigi ROUGE-skooride lõikes tulemused halvenevad peenhäälestamise tulemusel.

Tabelis 25 on näidatud mudeli täpsused ja saagised null ning viie näitega seades enne ning pärast peenhäälestamist. Selle ülesande korral saavutab mudel peale peenhäälestamist nii täpsuseks kui saagiseks nulli, kuna oli kaotanud selle käigus võime õiges formaadis vastuseid genereerida.

Üldpilt viitab, et peenhäälestamine lisas mudelile küll uusi teadmisi, ent samas halvenes selle võime luua keerukamaid tekstistruktuure. See nähtus on sarnane üleõppimisele, kuigi tehniliselt ei pruugi see olla klassikaline üleõppimine – pigem on tegemist juhtumiga, kus treenimine liiga piiratud või vähem esinduslikel andmetel viib

üldistusvõime kadumiseni.

Käesolev eksperiment kinnitab, et LLM-ide peenhäälestamine on delikaatne protsess. Esmase üldise teadmuse süvendamiseks võib kasutada suurt hulka üldisi tekstikorpuseid, kuid keerukamate ülesannete jaoks tuleb peenhäälestuse lõppfaasis kasutada spetsiifiliselt kureeritud ja vormiliselt lähedasi andmestikke. Praegune töö demonstreerib, et peenhäälestamine töötab, kuid nõuab rohkem katsetamist, et järjepidevalt hea tulemus saavutada.

Mudel	**Grammatika 1**		**Grammatika 2**	
	Täpsus	*Levenshtein*	*Täpsus*	*Levenshtein*
Llama 3.1 8B	0.061	0.174	0.092	0.311
Peenhäälestatud	0.0	0.1347	0.0	0.1869

Tabel 23. Mudelite täpsuse ning Levenshteini meetrika võrdlus grammatika andmestikul

Mudel	**Dialoogide kokkuvõtted**			**Uudislugude kokkuvõtted**		
	ROUGE-1	*ROUGE-2*	*ROUGE-L*	*ROUGE-1*	*ROUGE-2*	*ROUGE-L*
Llama 3.1 8B	0.217	0.056	0.184	0.1506	0.0461	0.1263
Peenhäälestatud	0.1715	0.0475	0.1424	0.1413	0.0411	0.1101

Tabel 24. Mudelite võrdlus dialoogide ja uudissaadete kokkuvõtete andmestikul

Mudel	**0 näitega**		**5 näitega**	
	Täpsus	*Saagis*	*Täpsus*	*Saagis*
Llama 3.1 8B	0.582	0.474	0.850	0.047
Peenhäälestatud	0	0	0	0

Tabel 25. Mudelite täpsuste ja saagiste võrdlus kõnelejatuvastuse andmestikul

8. Tulemuste valideermine

Selleks, et valideerida, kas erinevate LLM-ide jõudlus loodud võrdlusandmestikel korreleerub sellega, kui kasulikuks inimesed neid mudeleid tegelikult peavad, koostati hulk avatud vastustega küsimusi ning ülesandeid, mis võiksid olla praktiliselt kasulikud LLM-ide kasutajatele, ning lasti nii inimestel kui tugeval LLM-il vastuseid käsitsi hinnata. Kui hinnatud tulemused korreleeruvad võrdlusandmestikel saadud tulemustega, võib järeldada, et loodud võrdlusandmestikud on kasulikud, kuna nende tulemused peegeldavad, kui kasulikud vastavad LLM-id inimestele tegelikult on.

Suhteliselt tavapäraseks saanud taktika on kasutada võimsamaid keelemudeleid, nagu GPT-4, automaatselt hindamaks vähem võimekate LLM-ide vastuseid sellistele avatud küsimustele [34]. See niinimetatud LLM hindajana (*LLM-as-a-judge*) lähenemine võimaldab ka selle osa automatiseerimist. GPT-4 hindajana toimimise efektiivsus on tõestatud inglise keele peal – seal korreleeruvad GPT-4 antud hinnangud inimekspertide omadega – kuid autori hinnangul pole seda seost eesti keele puhul veel uuritud ega kinnitatud [35].

Järgnevalt on esitatud tulemused ja hindamisprotsess nii inimhindajate kui LLM hindajana metoodika tulemustest ja nende korreleeruvus jõudlustestide tulemustega.

8.1 Valideerimisküsimustiku koostamine

Valideerimiseks vajalikud küsimused põhinevad suures osas MT-Bench andmestikul [35]. MT-Bench sisaldab endas 80 mitmeetapilist küsimust kaheksas kategoorias, mis on disainitud hindama mudelite juhiste järgimise võimekust. Sealt sai välja selekteeritud ning käsitsi eesti keelde tõlgitud 47 küsimust kategooriatest: kirjutamine, arutlemine, matemaatika, reaalteadmised ning humanitaarteadmised. Kõrvale said jäetud küsimused, mis sisaldasid kultuurilist tausta, mis polnud seotud Eestiga. Lisaks neile sai autori poolt juurde genereeritud 10 eestispetsiifilist küsimust.

8.2 Inimvalideerimine

Peale küsimustiku kokkupanemist, sai see üles seatud Huggingface keskkonda ning seejärel jooksutatud kõigi võrreldavate mudelite peal. Seejärel sai püsti pandud Django

veebirakendus, mis pakub kasutajale korraga kahe anonüümse mudeli vastust ühele andmestiku küsimusele ning laseb valida, kumma mudeli vastus on parem või kas vastused on samaväärsed. Järgnevalt kirjeldatakse inimhindamise tulemusi.

8.2.1 Mudelite tulemused

Tabelis 26 on toodud kokkuvõtlikud inimhinnangute tulemused. Iga küsimuse korral kui inimene valis paremaks mudeli A, siis sai mudel A ühe punkti ning kui mudeli B, siis sai mudel B ühe punkti. Kui inimene valis vastused samaväärseteks, siis said mõlemad mudelid 0,5 punkti. Lõpuks mudelite kogutud punktide arv jagati sellega, mitmel korral mudeli pakutud vastust inimesele esitati.

On näha, et inimeste poolt on parimaks mudeliks valitud Claude 3.7 Sonnet, millele järgneb väikese vahega Deepseek Chat V3 ning juba suurema vahega Gemini 2.0 Flash. Üllatavalt on Gemma 27B hinnatud paremaks nii GPT-4o Minist kui ka GPT-4o-st. Halvimateks mudeliteks on inimesed hinnanud TartuNLP LLammas mudeli ning Gemma 2 9B. Vahe parima ja halvima mudeli tulemuste vahel on suur, olles üle 0,5.

Lisaks eelnevalt kirjutatud meetodile võib mudelite kohta arvutada ka Elo skoorid. Elo hinnangute süsteem võimaldab samuti asju paarikaupa võrreldes igale asjale seada vastavusse hinnangu, mis kirjeldab selle asja headust. Elo süsteemi eelis on see, et kui mudel, millel juba on kõrge Elo skoor võidab mudeli vastu, millel on madal Elo skoor, siis see on üsnagi oodatav tulemus ning mudelite skoore muudetakse vähe. Tabelitest 26 ja 27 on näha, et suures pildis olulist vahet ei ole, kumba süsteemi kasutada, ning vaid üksikud mudelid on koha vahetanud.

Mudel	Skoor
Claude 3.7 Sonnet	0.7539
Deepseek Chat V3	0.7460
Gemini 2.0 Flash 001	0.6985
Gemma 2 27B It	0.6416
Llama 3.1 70B Instruct	0.5522
GPT-4o	0.5450
GPT-4o Mini	0.5331
Llama 3.1 405B Instruct	0.5000
Qwen2 72B Instruct	0.4417
Llama 3.1 8B Instruct	0.3889
Mistral Nemo Instruct	0.3266
TartuNLP Llammas	0.2705
Gemma 2 9B It	0.2110

Tabel 26. Mudelite tulemused inimeste poolt hinnatuna

Mudel	Elo skoor
Deepseek Chat V3	1735
Claude 3.7 Sonnet	1676
Gemini 2.0 Flash 001	1657
Gemma 2 27B It	1573
Llama 3.1 70B Instruct	1541
GPT-4o	1522
GPT-4o Mini	1500
Llama 3.1 405B Instruct	1480
Qwen2 72B Instruct	1465
Llama 3.1 8B Instruct	1411
TartuNLP Llammas	1353
Mistral Nemo Instruct	1351
Gemma 2 9B It	1236

Tabel 27. Mudelite Elo skoorid inimeste poolt hinnatuna

8.2.2 Pearsoni korrelatsioonikordaja

Eelnevalt kokkuvõtlikus tabelis viidi iga andmestiku summaarsed tulemused vahemikku 0 kuni 1 selliselt, et kõige nõrgem tulemus läks nulliks ning kõige tugevam tulemus üheks. Sedasama protsessi saab teha inimeste hinnangute põhjal leitud tulemustele ning seejärel leida andmestikelt saadud automaatsete tulemuste ning inimeste hinnangutel põhinevate tulemuste vahelise korrelatsioonikordaja.

Pearsoni korrelatsioonikordaja sobib hästi olukorras, kus soovitakse mõõta kahe pideva tunnuse vahelise lineaarse seose tugevust [36]. Antud juhul on kõikide andmestike tulemused ning inimeste hinnangutel põhinevad punktid normaliseeritud nulli ja ühe vahele, mis võimaldab vaadelda tulemuste suhtelisi erinevusi absoluutväärtustes. Pearsoni kordaja abil on võimalik kindlaks teha, kas mudel, mille automaatne tulemus on kõrgem, on ka proportsionaalselt kasulikum inimhindamiste põhjal. See võimaldab hinnata, kuivõrd hästi loodud võrdlusandmestike punktisummad peegeldavad tegelikku kasutajakogemust. Pearsoni korrelatsioonikordaja r on arvutatav valemiga:

$$r = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n \cdot \sigma_X \cdot \sigma_Y} \tag{8.1}$$

kus n tähistab juhuslike suuruste vaatluspaaride arvu, $\bar{x}$, $\bar{y}$ on vastavalt nende suuruste

aritmeetilised keskmised ning σ_X, σ_Y nende standardhälbed [37].

Tabelis 28 on toodud välja erinevate andmestike Pearsoni korrelatsioonikordajad võrreldes inimeste antud hinnangutega. On näha, et üldiselt korreleeruvad jõudlustestide tulemused inimeste poolt hinnatud tulemustega üpris hästi, eri andmestike lõikes esineb, kas keskmine või tugev korrelatsioon. Kõige paremini korreleeruvad inimeste hinnagutega tulemused eksamite ning sõnaseletuste andmestikul ning kõige halvemini uudislugude kokkuvõtete andmestikul. Üllatuslikult on suured erinevused eksami ja mälumängu ning uudislugude ja dialoogide kokkuvõtete andmestike korrelatsioonikordajate vahel, mis on omavahel tüübilt analoogsed ülesanded.

Andmestik	r
Eksamid	0.86
Sõnaseletused	0.80
Dialoogide kokkuvõtted	0.77
Käänded	0.72
Mälumäng	0.66
Kõnelejatuvastus	0.63
Grammatika	0.48
Uudislugude kokkuvõtted	0.44

Tabel 28. Andmestike Pearsoni korrelatsioonikordajad võrreldes inimeste hinnatud headusega

8.2.3 Spearmani korrelatsioonikordaja

Alternatiivne viis korrelatsioonide leidmiseks on Spearmani korrelatsioonikordaja, mis on asjakohane juhul, kui huvipunktis on mitte tulemuste absoluutsed väärtused, vaid nende omavaheline järjestus [38]. Kuna LLM-ide hindamisel võib olla oluline, kas paremusjärjestus mudelite vahel säilib sõltumata skooride täpsetest vahedest ning tulemuste vaheline seos ei pruugi olla lineaarne, siis pakub Spearman sobiva viisi selle seose analüüsimiseks. See korrelatsioonikordaja aitab hinnata, kas mudelid, mis saavad kõrgemaid tulemusi automaatsetes testides, paigutuvad sarnaselt ka inimeste antud hinnangutes. Spearmani kordaja kasutamine täiendab Pearsoni analüüsi, andes parema ülevaate üldisest kooskõlast kahe mõõtmise vahel. See on arvutatav valemiga:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}, \tag{8.2}$$

kus n on vaatluspunktide arv ja d_i on iga vaatluspaari järjestuste vahe ja [37].

Tabelis 29 on toodud välja erinevate andmestike Spearmani korrelatsioonikordajad võrreldes inimeste antud hinnangutega. Ka Spearmani käsitluse järgi esineb iga andmestiku tulemuste korral inimhindajate tulemustega võrreldes vähemalt keskmise tugevusega korrelatsioon. Andmestike omavaheline järjestus korrelatsioonikordajate alusel on üldiselt samasugune nagu Pearsoni kordajate korral. Ainus erinevus on kõnelejatuvastuse paigutuses pingerivis, mis Spearmani järgi on järjestuses kaks korda eespool võrreldes Pearsoni omaga, edestades käänete ja mälumängu andmestike tulemused. Üldiselt võib väita, et jõudlustestide tulemused korreleeruvad inimhinnangutega mõõdukalt hästi.

Andmestik	ρ
Eksamid	0.85
Sõnaseletused	0.79
Dialoogide kokkuvõtted	0.73
Kõnelejatuvastus	0.71
Käänded	0.70
Mälumäng	0.65
Grammatika	0.53
Uudislugude kokkuvõtted	0.44

Tabel 29. Andmestike Spearmani korrelatsioonikordajad võrreldes inimeste hinnatud headusega

8.2.4 Võrdlus olemasoleva LLM-ide hindamise platvormiga

Tartu ülikooli juhtimisel valminud platvorm – Tehisaru baromeeter (`https://baromeeter.tartunlp.ai/`), mis põhineb California ülikooli arendatud ChatBot-Arenal, on loodud eesmärgiga hinnata LLM-ide võimekust eesti keeles [39]. Platvorm laseb kasutajal esitada mistahes eestikeelse küsimuse ja kuvab kahe anonüümse mudeli vastust selle küsimusele. Seejärel saab kasutaja valida, kas kummagi mudeli vastus oli teisest parem, nende vahel on viik või on mõlemad vastused halvad. Kasutajate hinnangute põhjal arvutatakse mudelitele skoorid ning kuvatakse ka jooksvat mudelite paremusjärjestust.

Tabel 30 kujutab Tehisaru baromeetrist valitud mudelite skoore inimhinnangute põhjal 11.05.2025 seisuga. Välja on selekteeritud need mudelid, mida ka käesoleva töö raames uuriti. Gemma 27B puhul on võrdlusesse võetud versioon Gemma 3 mudelist, kuna antud töö raames uuritud Gemma 2 antud uurimuses ei kajastata. Tehisaru baromeetri lehekülg kasutab samuti Elo skoore mudelite headuse kirjeldamiseks. Võib märgata, et tabelis 27 on skoorid süstemaatiliselt kõrgemad kui tabelis 30. See tuleneb sellest, et on kasutatud Elo süsteemi erinevate parameetritega. Sellest hoolimata saab võrrelda mudelite omavahelist

järjestust. Võrreldes mudelite paremusjärjestust antud töö raames loodud rakenduse abil kogutud inimeste hinnangutega mudelitele, 26, võib märgata, et üldiselt paremusjärjestused on mõõdukalt sarnased, kuid esineb ka olulisi erinevusi. Kõige märgatavam erinevus seisneb GPT-4o asukohas paremusjärjestuses, kus Tehisaru baromeetri hinnangute põhjal on see paigutatud antud mudelitest esimesele kohale, kuid antud töö käigus läbiviidud hindamise põhjal alles kuuendale kohale. Teine suurem erinevus on Deepseek Chat V3 mudeli juures, mis on antud uurimuses hinnatud 3 koha võrra paremaks. Kõikide ülejäänud mudelite koht paremusjärjestuses erineb maksimaalselt 1 võrra.

Mudel	**Skoor**
GPT-4o	1579
Claude 3.7 Sonnet	1538
Gemini 2.0 Flash 001	1509
Deepseek Chat V3	1500
Gemma 3 27B It*	1389
Llama 3.1 70B Instruct	1301
TartuNLP Llammas	1274
Llamma 3.1 8B Instruct	1171
Mistral Nemo Instruct	1154

Tabel 30. Tehisaru baromeeter platvormi LLM-ide skoorid inimeste hinnangute põhjal (11.05.2025 seisuga). Välja on selekteeritud need mudelid, mida ka antud töö käigus uuriti. * Erandiks on Gemma 27B mudel, mis ei ole identne siin käsitletud mudeliga (Tehisaru baromeetris on kasutusel Gemma 3 versioon ning antud töö raames Gemma 2.)

8.3 LLM hindajana

Hindavaks mudeliks sai valitud Claude 3.7 Sonnet, kuna see on nii käesoleva töö kui Tartu Ülikooli Keeletehnoloogia Uurimisrühma poolt ülespandud tehisarude hindamise platvormi (https://vestle.tartunlp.ai/) edetabeli hinnangul üks parimatest keelemudelitest. LLM kasutamine hindajana on analoogne sellele, kuidas küsiti hinnanguid inimhindajatelt. LLM-ile selgitatakse olukorda ning lastakse tal valida, kas vastused on samaväärsed või on kumbki vastustest teisest parem ning ta tagastab oma valiku. Ülejäänud arvutusprotsess on samasugune kui inimhindajate korral. Mudelile anti järgnev suunis, millele järgnesid kõik vajalikud andmed json kujul:

```
Sa oled keelemudelite hindamise ekspert.  Sulle antakse
keelemudelile esitatud küsimus ning kaks vastust:  üks
keelemudelilt A, teine keelemudelilt B. Sinu ülesanne
on hinnata kumb vastus on parem ning vastata seda
keelemudelit tähistava tähega.  Kui mudeli A vastus
on parem, tagasta ainult täht 'A', kui mudeli B vastus
on parem, tagasta ainult täht 'B' ning kui mudelite
vastused on samaväärsed või sama head, siis tagasta
ainult täht 'C'. Sulle võidakse anda ka umbkaudne õige
vastus, kuid seda ei pruugita anda.
```

Tabel 31 kujutab mudelite tulemusi Claude 3.7 Sonnet'i poolt hinnatuna. Võrreldes neid tulemusi inimhindajate tulemustega (tabel 26), võib märgata, et mõlemal juhul osutub parimaks mudeliks Deepseek Chat V3. Kui inimhindajate hinnangul järgneb sellele napilt nõrgema tulemusega Claude 3.7 Sonnet, siis LLM hindaja arvates on teine parim mudel napilt Gemini Flash, millele järgneb samuti väikse vahega Claude 3.7 Sonnet ise. Inimhindajad olid üllatavalt arvanud GPT-4o Mini paremaks GPT-4o-st, kuid Claude Sonnet'i hinnangul on GPT-4o arvestatavalt parem Mini versioonist. Mõlemal juhul on hinnatud Llama 70B versioon paremaks Llama 405B versioonist. Kui inimhindajad pidasid halvimateks Gemma 9B versiooni, millele jägnes Llamas, siis LLM-i hinnangul on halvim Llama 8B, millele järgneb Gemma 9B. Ka LLM hindajana validatsiooni korral on vahe parima ja halvima mudeli tulemuste vahel pea 0,6.

Pearsoni korrelatsioonikordajaks inimeste poolt hinnatud mudelite tulemuste ning Claude 3.7 Sonnet'i poolt hinnatud tulemuste vahel tuli 0,940 p-väärtusega $1,76 * 10^{-6}$, mis viitab, et tõenäosus nii tugeva või tugevama korrelatsiooni tekkimiseks juhuslikult kahe tegelikult mittekorreleeruva suuruse, on väga väike — ainult 0.000176%. Spearmani vastavaks korrelatsioonikoefitsendiks tuli 0,918 ning p-väärtuseks $9,906 * 10^{-6}$. Kokkuvõttes võib öelda, et inimhindajate ning LLM hindajana tulemuste vahel esineb tugev korrelatsioon, mis tähendab, et Claude Sonnet'i kasutamine automaatse hindajana inimeste asemel oleks põhjendatud.

Sarnaselt inimhinnangutega, ei ole ka siin olulist vahet kas kasutada esimest hindamismeetodi või Elo skoore. Järjestuses ei muutu ühegi mudeli koht rohkem kui ühe võrra ning vaid 4 paari mudeleid vahetavad omavahel kohad.

Mudel	Skoor
Deepseek Chat V3	0.7717
Gemini 2.0 Flash 001	0.7697
Claude 3.7 Sonnet	0.7586
GPT-4o	0.6765
Gemma 2 27B It	0.6543
GPT-4o Mini	0.62
Llama 3.1 70B Instruct	0.4733
Llama 3.1 405B Instruct	0.4276
Qwen2 72B Instruct	0.403
TartuNLP Llammas	0.2424
Mistral Nemo Instruct	0.226
Gemma 2 9B It	0.2143
Llama 3.1 8B Instruct	0.2048

Tabel 31. Mudelite tulemused Claude Sonnet'i poolt hinnatuna

Mudel	Elo skoor
Gemini 2.0 Flash 001	1747
Deepseek Chat V3	1705
Claude 3.7 Sonnet	1695
GPT-4o	1664
Gemma 2 27B It	1639
GPT-4o Mini	1592
Llama 3.1 405B Instruct	1467
Llama 3.1 70B Instruct	1440
Qwen2 72B Instruct	1382
Mistral Nemo Instruct	1319
TartuNLP Llammas	1316
Llama 3.1 8B Instruct	1284
Gemma 2 9B It	1251

Tabel 32. Mudelite Elo skoorid Claude Sonnet'i poolt hinnatuna

9. Kokkuvõte

Suurte keelemudelite hindamiseks ja võrdlemiseks luuakse üha uusi võrdlusteste. Need testid on aga sageli inglise keele ning teiste suurema kõnelejaskonnaga keelte spetsiifilised. Eriti vähe on LLM-ide võimekust uuritud väikese kõnelejaskonnaga keeltes, nende hulgas eesti keel. Käesoleva magistritöö eesmärgiks oli uurida süstemaatiliselt levinud keelemudelite võimekust täita erinevaid ülesandeid eesti keeles.

Töö käigus sai loodud neli uut andmestikku (eksamid, sõnaseletused, käänded ning mälumäng) mudelite erinevate võimekuste hindamiseks eesti keeles, üks (grammatika) sai kohendatud paremini kasutatavaks ning ülejäänud kolmele kasutatud juba olemasolevale andmestikule (dialoogide ja uudislugude kokkuvõtted ning kõnelejatuvastus) sai loodud konfiguratsioonifailid, et nende andmestike peal mudeleid evalveerida kasutades LM Evaluation Harnessit. Neil andmestikel sai testitud ja võrreldud 7 levinud baasmudelit, 9 avatud lähtekoodiga juhendhäälestatud mudelit ning 4 kommertsmudelit.

Üldiselt andsid testitud mudelitest parimaid tulemusi kommertsmudelid Claude 3.7 Sonnet, Gemini 2.0 Flash 001 ning GPT-4o. Erandiks dialoogide ning uudislugude kokkuvõtete andmestik, mille peal said parimaid tulemusi erinevad testitud suuremad baasmudelid. Ebastandarstsed olid tulemused ka kõnelejatuvastuse andmestikul, kus 5 näitega juhul andsid mitmed baasmudelid juhendhäälestatud mudelitest kõrgemaid täpsusi, kuid samal ajal nullilähedasi saagiseid. Parimad vabavaralised juhendhäälestatud mudelid jäävad tulemustelt alla parimatele kommertsmudelitele, kuid konkureerivad kehvemate kommertsmudelitega, nagu GPT-4o Mini. Kõigi andmestike lõikes parimaid tulemusi andis GPT-4o, millele järgnes Claude 3.7 Sonnet, baasmudelite arvestuses Gemma 2 27B ning Llama 3.1 70B. Eri andmestike tulemuste korrelatsioonanalüüs näitas, et enim korreleeruvad teiste andmestike tulemustega sõnaseletuste, mälumängu, käänete ning eksamite andmestikud. Vähim korreleerus teistega uudislugude kokkuvõtete andmestik, mille Pearsoni korrelatsioonikordaja teise kokkuvõtete andmestiku - dialoogide kokkuvõtetega - oli praktiliselt 0.

Eksamite andmestiku võrdluses eesti keelde masintõlgitud MMLU andmestiku analoogsete kategooriatega, selgus et kummalgi andmestikul oli juhendhäälestatud mudelite täpsus keskmiselt 0,2 võrra kõrgem baasmudelite täpsusest. Mudelid saavutasid üldiselt eksamite andmestiku peal keskmiselt 0,1 võrra parema täpsuse kui MMLU peal, mida võib selgitada asjaolu, et MMLU küsimused on hinnatud keskkooli, vastavad

eksamite omad põhikooli tasemele. Parimaid tulemusi saanud mudelid on samad olenemata andmestikust: suuremad testitud kommertsmudelid ning baasmudelite lõikes suuremad baasmudelid. Kategooriate lõikes esineb tulemustes andmestike vahel palju erinevusi, mida saab selgitada erineva küsimuste olemuse ning erinevate valikvastuste arvudega, mis MMLU puhul on standartselt neli, kuid eksamite andmestiku korral küsimuseti erinev.

Näidati, et ka minimaalse peenhäälestamise abil saavad paraneda mudelite tulemused loodud andmestikel, kui on tegemist andmestikuga, mis ei nõua keerulise struktuuriga väljundit, vaid põhineb faktiteadmistel. Keerulisema väljundi struktuuriga andmestikel tulemus küll halvenes, kuid see demonstreeribki seda, et peenhäälestamine on keerukas protsess. Just seetõttu ongi oluline, et loodi mitmekesine valik erinevatest andmestikest – kui peenhäälestamine parandab tulemusi ainult ühel kitsal andmestikul, ei pruugi see viidata tegelikule üldistusvõime paranemisele, kuid kui paranemine ilmneb ühtlaselt kõigil andmestikel, saab palju kindlamalt järeldada, et mudeli kvaliteet on tervikuna paranenud.

Tulemuste valideerimiseks sai kasutatud kahte levinud meetodit: inimestest hindajad ning LLM-i hindajana. Valideerimisküsimustik sai koostatud MT-Bench andmestiku küsimuste alusel, jättes välja küsimused, mida ei õnnestunud kohandada Eesti kultuurikeskkonda. Kokkuvõttes valmis 57 küsimusega andmestik, millest 10 olid ise juurde mõeldud, olles otseselt eestispetsiifilised. Inimvalideerimiseks sai püsti pandud Django veebirakendus, mis pakub kasutajale korraga üht küsimust ning anonüümselt kahe suvalise mudeli vastuseid ning laseb hinnata, kumb vastus on parem või kas vastused on samaväärsed. Analoogselt esitati ülesanne LLM-ist hindajale, milleks valiti üks parimaid tulemusi andnud mudel Claude 3.7 Sonnet. Selgus, et inimhindajate tulemused korreleeruvad olenevalt andmestikust kas keskmiselt või tugevalt jõudlustestide tulemusega, seda tugevaimalt eksamite ning sõnaseletuste andmestikel ning kõige nõrgemini uudislugude kokkuvõtete korral. Lisaks selgus, et LLM hindajana, Claude 3.7 Sonneti näitel, tulemused korreleerusid väga tugevalt inimhindajate tulemusega, mis annab põhjust eeldamaks, et tugevamate LLM-ide kasutamine keelemudelite hindamiseks võib toimida ka eesti keele peal samaväärselt inimeste hinnangutele.

Käesolev töö loob aluse eestikeelsete mudelite süstemaatiliseks hindamiseks. Seda nii otseste andmestike näol kui ka selle läbi, et demonstreeriti paremate keelemudelite võimekust hinnata teiste keelemudelite vastuseid avatud küsimustele. Tulevikus saab loodud andmestikke täiendada uute ülesannete tüüpidega, mis käesoleva uurimise raamest välja jäid, nagu näiteks programmeerimisega seotud ülesanded või eelarvamuste tuvastamine. Järgnevad uurimused saavad kasutada loodud andmestikke ning metoodikat, et

paremini suunata eesti keelel peenhäälestamise protsessi. Lisaks saaks kombineerida Tartu tehisaru baromeetri programmi raames kogutud inimeste küsimused ning demonstreeritud LLM-i abil hindamise metoodika, et saada kvaliteetsem hinnang praeguste keelemudelite võimekusele.

Kasutatud kirjandus

[1] Julen Etxaniz et al. “Latxa: An Open Language Model and Evaluation Suite for Basque”. In: (2024). arXiv: 2403.20266 [cs.CL]. URL: https://arxiv.org/abs/2403.20266.

[2] Miðeind. “Icelandic LLM Leaderboard on Hugging Face”. In: (2023). Külastatud: 09.10.2024.

[3] Arda Yüksel et al. “TurkishMMLU: Measuring Massive Multitask Language Understanding in Turkish”. In: (2024). arXiv: 2407.12402 [cs.CL]. URL: https://arxiv.org/abs/2407.12402.

[4] Zishan Guo et al. “Evaluating Large Language Models: A Comprehensive Survey”. In: (2023). arXiv: 2310.19736 [cs.CL]. URL: https://arxiv.org/abs/2310.19736.

[5] Dan Hendrycks et al. “Measuring Massive Multitask Language Understanding”. In: (2021). arXiv: 2009.03300 [cs.CY]. URL: https://arxiv.org/abs/2009.03300.

[6] Roel Wieringa, Hans Heerkens, and Björn Regnell. “How to Write and Read a Scientific Evaluation Paper”. In: 2009 17th IEEE International Requirements Engineering Conference. 2009, pp. 361–364. DOI: 10.1109/RE.2009.17.

[7] Ashish Vaswani et al. “Attention Is All You Need”. In: (2023). arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[8] Wen Lai, Mohsen Mesgar, and Alexander Fraser. “LLMs Beyond English: Scaling the Multilingual Capability of LLMs with Cross-Lingual Feedback”. In: (2024). arXiv: 2406.01771 [cs.CL]. URL: https://arxiv.org/abs/2406.01771.

[9] Martin Ehala. “Sustainability of the Estonian language”. In: Jan. 2015, pp. 191–199.

[10] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: Bioinformatics (Oxford, England) 36 (Sept. 2019). DOI: 10.1093/bioinformatics/btz682.

[11] Peiqin Lin et al. “MaLA-500: Massive Language Adaptation of Large Language Models”. In: (2024). arXiv: 2401.13303 [cs.CL]. URL: https://arxiv.org/abs/2401.13303.

[12] Tuka Alhanai et al. “Bridging the Gap: Enhancing LLM Performance for Low-Resource African Languages with New Benchmarks, Fine-Tuning, and Cultural Adjustments”. In: (2024). arXiv: 2412.12417 [cs.CL]. URL: https://arxiv.org/abs/2412.12417.

[13] Alex Wang et al. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. In: (2019). arXiv: 1804.07461 [cs.CL]. URL: https://arxiv.org/abs/1804.07461.

[14] Rowan Zellers et al. “HellaSwag: Can a Machine Really Finish Your Sentence?” In: (2019). arXiv: 1905.07830 [cs.CL]. URL: https://arxiv.org/abs/1905.07830.

[15] Tomas Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”. In: (2013). arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781.

[16] Jeffrey Pennington, Richard Socher, and Christopher Manning. “Glove: Global Vectors for Word Representation”. In: vol. 14. Jan. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

[17] Alex Wang et al. “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”. In: (2020). arXiv: 1905.00537 [cs.CL]. URL: https://arxiv.org/abs/1905.00537.

[18] Pranav Rajpurkar et al. “SQuAD: 100,000+ Questions for Machine Comprehension of Text”. In: (2016). arXiv: 1606.05250 [cs.CL]. URL: https://arxiv.org/abs/1606.05250.

[19] Guokun Lai et al. “RACE: Large-scale ReAding Comprehension Dataset From Examinations”. In: (2017). arXiv: 1704.04683 [cs.CL]. URL: https://arxiv.org/abs/1704.04683.

[20] Mark Chen et al. Evaluating Large Language Models Trained on Code. 2021. arXiv: 2107.03374 [cs.LG]. URL: https://arxiv.org/abs/2107.03374.

[21] Terry Yue Zhuo et al. “BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions”. In: (2025). arXiv: 2406.15877 [cs.SE]. URL: https://arxiv.org/abs/2406.15877.

[22] Jwala Dhamala et al. “BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation”. In: FAccT ’21 (Mar. 2021), pp. 862–872. DOI: 10.1145/3442188.3445924. URL: http://dx.doi.org/10.1145/3442188.3445924.

[23] Martin Fajcik et al. “BenCzechMark : A Czech-centric Multitask and Multimetric Benchmark for Large Language Models with Duel Scoring Mechanism”. In: (2024). arXiv: 2412.17933 [cs.CL]. URL: https://arxiv.org/abs/2412.17933.

[24] Fajri Koto et al. “ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic”. In: (2024). arXiv: 2402.12840 [cs.CL]. URL: https://arxiv.org/abs/2402.12840.

[25] Leo Gao et al. A framework for few-shot language model evaluation. Version v0.4.0. Dec. 2023. DOI: 10.5281/zenodo.10256836. URL: https://zenodo.org/records/10256836.

[26] Clémentine Fourrier et al. Open LLM Leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard. 2024.

[27] Filosoft. Eesti keele süntesaator. Külastatud: 19.04.2025. 2020. URL: https://www.filosoft.ee/gene_et/.

[28] Eesti Keele Instituut. Ekilex. Külastatud: 19.04.2025. 2022. URL: https://www.filosoft.ee/gene_et/.

[29] Eesti Keeleressursside Keskus. Eesti Wordnet. Külastatud: 19.04.2025. 2020. URL: https://keeleressursid.ee/et/265-eesti-wordnet.

[30] Yulong Chen et al. “DialogSum: A Real-Life Scenario Dialogue Summarization Dataset”. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, Aug. 2021, pp. 5062–5074. DOI: 10.18653/v1/2021.findings-acl.449. URL: https://aclanthology.org/2021.findings-acl.449/.

[31] Paul E. Black. “Levenshtein distance”. In: Algorithms and Theory of Computation Handbook. In Dictionary of Algorithms and Data Structures [online], accessed 2025-04-19. CRC Press LLC, 1999. URL: https://www.nist.gov/dads/HTML/Levenshtein.html.

[32] SAS. Precision, Recall, and the F1 Score. Külastatud: 19.04.2025. 2020. URL: https://documentation.sas.com/doc/da/pgmsascdc/v_042/casvta/casvta_boolrule_details05.html.

[33] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[34] Charles Koutcheme et al. “Open Source Language Models Can Provide Feedback: Evaluating LLMs’ Ability to Help Students Using GPT-4-As-A-Judge”. In: (2024). arXiv: 2405.05253 [cs.CL]. URL: https://arxiv.org/abs/2405.05253.

[35] Zheng et al. “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena”. In: (2023). arXiv: 2306.05685 [cs.AI]. URL: https://arxiv.org/abs/2306.05685.

[36] Joseph Lee Rodgers and W. Alan Nicewander. “Thirteen ways to look at the correlation coefficient”. In: The American Statistician 42.1 (1988), pp. 59–66.

[37] Ako Sauga. Statistika õpik majanduseriala üliõpilastele. Teine, parandatud väljaanne. TalTech Kirjastus, 2020.

[38] Jerome L Myers, Arnold D Well, and R Carl Lorch. Research Design and Statistical Analysis. 3rd. New York: Routledge, 2010.

[39] Wei-Lin Chiang et al. “Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference”. In: (2024). arXiv: 2403.04132 [cs.AI]. URL: https://arxiv.org/abs/2403.04132.

Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[1]

Mina, Helena Grete Lillepalu

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Võrdlusanalüüs suurte keelemudelite jõudluse hindamiseks eesti keeles", mille juhendaja on Tanel Alumäe
   1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

12.05.2025

[1] Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingulise tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtajaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.