

TARTU ÜLIKOOL  
Arvutiteaduse instituut  
Informaatika õppekava

**Ahto Kiil**

**Suurte keelemudelite võrdlev analüüs**

**Eesti bioloogiaolümpiaadide küsimuste põhjal**

**Bakalaureusetöö (9 EAP)**

Juhendajad:  
Taido Purason, Hele-Andra Kuulmets

Tartu 2025

# Suurte keelemudelite võrdlev analüüs Eesti bioloogiaolümpiaadide küsimuste põhjal

## Lühikokkuvõte:

Suurte keelemudelite võrdlemiseks kasutatakse erinevaid teste – tõlkimine, tekstist arusaamine, küsimustele vastamine jpm. Reeglina on testandmestikud tõlgitud inglise keelest ja algusest peale eesti keeles koostatud ning Eesti eripära arvestavaid küsimustikke napib. Bakalaureusetöö käigus valmis 1031 küsimusest koosnev valikvastustega andmestik, mille koostamisel kasutati Eesti bioloogiaolümpiaadide ülesandeid ajavahemikust 2005–2024. Lõputöö teises faasis pandi proovile viis OpenAI mudelit, 13 eesti keelel treenitud mudelit platvormilt Hugging Face ning otse veebilehtedelt üheksa uusimat suletud kommertsmodellit. Mudelite parim täpsus (85,35%) on võrreldav Eesti olümpiaadidel esikolmikusse jõudnud õpilaste keskmise tulemusega teooriavoorus (87,16%).

**Võtmesõnad:** suured keelemudelid, valikvastustega küsimused, olümpiaad, eesti keel

**CERCS:** P176 Tehisintellekt

## Leveraging Estonian Biology Olympiad Problems for Evaluating Large Language Models

### Abstract:

Several types of tests are used to evaluate large language models – translation, text comprehension, image recognition, answering questions etc. Typically, evaluation datasets are translated from English, and there is a lack of test sets that consider specific local context and are originally composed in Estonian. As part of this BA thesis, a multiple-choice dataset consisting of 1,031 questions was compiled using tasks from Estonian biology olympiads between 2005 and 2024. In the second phase, five OpenAI models, 13 Estonian-trained models from the Hugging Face platform and nine of the most recent closed commercial models accessed via websites were evaluated. The best model's accuracy (85.35%) is comparable to the average result (87.16%) of pupils who placed in the top three in Estonian olympiads.<sup>1</sup>

**Keywords:** LLMs, multiple-choice questions, olympiad, Estonian

**CERCS:** P176 Artificial intelligence

---

<sup>1</sup> Tõlget on toimetatud ChatGPT abiga. <https://chatgpt.com/> (gpt-4o, 14.05.2025)

## Sisukord

1.	Sissejuhatus .....	4
2.	Suurte keelemudelite võrdlev hindamine .....	5
2.1	Alusmudelid .....	5
2.2	Peenhäälestatud mudelid .....	6
2.3	Mudelite testimiseks koostatud andmestikud .....	6
2.3.1	Inimkonna viimane eksam .....	7
2.3.2	Olümpiaadide küsimustel põhinevad testid .....	8
2.3.3	Eestikeelsed testandmestikud.....	8
3.	Andmete kogumine ja töötlemine .....	9
3.1	Toorandmete salvestamine .....	9
3.2	Andmestiku vorming .....	10
3.3	Küsimuste eraldamine .....	11
3.4	Õiged vastused.....	12
3.5	Juhusliku valiku tõenäosus .....	13
3.6	Olümpiaadide võitjate õigete vastuste osakaal .....	13
3.7	Käsitsi tehtud parandused .....	13
3.8	Andmestiku ja lähtekoodide varamud .....	14
4.	Tulemused .....	15
4.1	OpenAI mudelite testimine.....	15
4.1.1	OpenAI mudelite üldine täpsus.....	16
4.1.2	OpenAI mudelite täpsus vanuseastme kaupa.....	17
4.2	Platvormil Hugging Face avaldatud mudelite testimine.....	18
4.2.1	Testimiseks valitud mudelid .....	18
4.2.2	Testimisprotseduur UT HPC-Rocket klastris .....	19
4.2.3	Platvormil Hugging Face avaldatud mudelite testimise tulemused .....	21
4.3	Uusimate mudelite testimine veebis .....	23
5.	Järeldused .....	25
5.1	OpenAI mudelid .....	25
5.2	Avatud parameetritega mudelid.....	25
5.3	Vestlusakna kaudu testitud kommertsmudelid .....	26
6.	Kokkuvõte .....	28
	Kasutatud allikad .....	29
	Terminid.....	32
	Litsents.....	33

# 1. Sissejuhatus

Suured keelemudelid (ingl *large language model, LLM*) on kujunenud oluliseks teemaks mitte ainult kitsalt tehisintellektist või masinõppest rääkides, vaid peaaegu kõikides ühiskonna arengut puudutavates valdkondades. Nii uuenduslik ja seni veel väheste mõju-uuringutega raskesti mõistetav kontseptsioon tekitab inimestes ühelt poolt vaimustust ja tehnokraatlikku optimismi, kuid teisalt annab alust ka sügavaks mureks tuleviku pärast. Laialdaselt kasutatavate ChatGPT mudelite arendaja OpenAI deklareerib, et nende eesmärk on tagada üldise tehisintellekti (ingl *General Artificial Intelligence, AGI*) rakendamine kogu inimkonna hüvanguks<sup>2</sup>. Teisalt avaldas 2024. aastal Nobeli füüsikapreemia pälvinud Geoffrey Hinton, keda peetakse sügava masinõppe alustehnoloogia rajajaks, telekanalile CBS antud intervjuus muret, et inimesed pole suurtest keelemudelitest rääkides mõistnud, mis neid ees ootab<sup>3</sup>.

Mõlemaid väljaütlemisi silmas pidades on oluline esmalt saada selgust, milleks tänapäevased suured keelemudelid on võimelised; mida need teavad ja oskavad. Kõige levinum viis suurte keelemudelite hindamiseks on esitada küsimusi, millele leiduvad üheselt õiged vastused ning saadud tulemuste põhjal koostada edetabeleid mudelite täpsusest.

Bakalaureusetöö eesmärk oli luua Eesti olümpiaadide küsimuste põhjal testandmestik suurte keelemudelite võrdlemiseks. Töö käigus valmisid empiiriliste katsete põhjal edetabelid kolme erineva kategooria mudelite osas:

1. OpenAI fikseeritud arenduskuupäevaga mudelid;
2. platvormil Hugging Face avaldatud ja mh eestikeelsetel materjalidel treenitud mudelid;
3. tavakasutajatele 2025. aasta märtsis pakutavad tippmudelid.

Töös tutvustatakse esmalt lühidalt suurte keelemudelite tausta ja erinevaid alamliike. Seejärel antakse ülevaade juba olemasolevatest andmestikest mudelite testimiseks. Sisulist tööd käsitleb peatükk 3, milles kirjeldatakse algandmete valikut, kogumist, töötlemist ja testimiseks ettevalmistamist. 4. peatükis avatakse kolme erinevat testimisprotseduuri ja kajastatakse tulemusi. Viimastes peatükkides teeb autor kokkuvõtte töö käigus leitud faktilistest asjaoludest ning pakub omapoolse hinnangu suurte keelemudelite kasutamisele.

---

<sup>2</sup> OpenAI mission. <https://openai.com/about/> (04.05.2025)

<sup>3</sup> CBS Online. <https://www.cbsnews.com/news/godfather-of-ai-geoffrey-hinton-ai-warning/> (04.05.2023)

## 2. Suurte keelemudelite võrdlev hindamine

Suuri keelemudeleid on masinõppe ja tehisnärvivõrkude abil treenitud suurtel andmekogudel ning need suudavad seetõttu märgata mustreid sõnade, lausete ja terviktekstide vahel<sup>4</sup>. Enamik kaasaegseid suuri keelemudeleid põhineb transformer-arhitektuuril, millele pandi alus 2017. aastal (Vaswani et al., 2017).

Vaswani jt. teadusartiklit ja selles välja pakutud enesetähelepanu mehhanismi (ingl *self-attention*) loetakse tänapäevaste suurte keelemudelite arengus tähtsaks verstapostiks. Transformer-arhitektuuri suurim eelis varasemate tehnoloogiate (rekurrentsed ja konvolutsionaalsed närvivõrgud) on võime suuri andmehulki paralleelselt töödelda ja seeläbi treenimisprotsessi tõhustada (Sügis et al., 2025, 159). Transformeritel põhinevaid mudeleid saab kasutada väga erinevate ülesannete lahendamiseks – tõlkimine, teksti või pildi genereerimine, objektide visuaalne tuvastamine, küsimustele vastamine jne.

### 2.1 Alusmudelid

Suured keelemudelid omandavad tekstikorpuseid, raamatuid, ajaleheartikleid, seaduseid, teadusartikleid jms töödeldes loomuliku keele süntaksi ja semantika. Lisaks sisaldavad treeningandmed ka hulgaliselt fakte, mida mudelid saavad hilisemate ülesannete täitmisel ära kasutada.

Mudeli suurust mõõdetakse parameetrite arvuga ja see näitaja mängib suurte keelemudelite „headuse“ või sooritusvõime mõõtmisel olulist rolli. 2020. aastal oli GPT-3 mudel 175 miljardi parameetriga selle näitaja poolest teiste analoogsetest mudelitest suurusjärgu võrra üle. GPT-3 uurimistöös (Brown et al., 2020) tulemusena näidati, et keelemudelite parameetrite suurendamine parandab väheste näidetega õppimise (ingl *few-shot learning*) võimekust. Autorid demonstreerisid, et ka väheste näidetega õppimisega treenitud mudel on võimeline edukalt täitma erinevaid ülesandeid piiratud eelinfo põhjal.

Alusmudelite puuduste (kallutatus, valeinfo edastamine, kahjuliku vastuse genereerimine jpm) vähendamiseks kasutatakse peenhäälestamist (ingl *fine-tuning*).

---

<sup>4</sup> EKI teatmik. <https://teatmik.eki.ee/teatmik/mis-on-suured-keelemudelid/> (02.05.2025)

## 2.2 Peenhäälestatud mudelid

Peenhäälestamine on protsess, mille käigus juba eeltreenitud alusmudeleid kohandatakse spetsiifiliste ülesannete täitmiseks (nt haiguskollete leidmine röntgenülesvõtelt või nimeüksuste tuvastamine tavatekstist). Teine peenhäälestuse suund on mudeli vastuste filtreerimine valedest, eksitavatest või soovimatutest väidetest ning soovitustest. Sama kategooria alla kuulub ka vestlusroboti suhtlusviisi, sõnavara ja tooni vastavusse viimine klientide eelistustega.

Peenhäälestamiseks kasutatakse mitmeid erinevaid meetodeid. Juhendatud masinõpe (ingl *supervised learning*) tähendab mudeli treenimist märgistatud andmekogumitel, mis on suunatud konkreetse ülesande täitmiseks. Stiimulõpe koos inimtagasisidega (ingl *reinforcement learning from human feedback, RLHF*) kasutab inimeste hinnanguid mudeli väljunditele, et seejärel mudelit täiendavalt treenida vastavalt inimeste eelistustele (Sügis et al., 2025). Juhistega peenhäälestamine (ingl *instruction tuning*) hõlmab mudeli treenimist ülesannete kogumil, mis on kirjeldatud loomulikus keeles antud juhiste abil (Ouyang et al., 2022).

Bakalaureusetöö 4. peatükis kajastatud tulemused näitavad selgelt, et peenhäälestatud mudelite täpsus bioloogiaolümpiaadide küsimustele vastamisel ületab alusmudelite vastavaid näitajaid.

## 2.3 Mudelite testimiseks koostatud andmestikud













Suurte keelemudelite hindamiseks on vaja laiapõhjalisi ja usaldusväärsetele andmetele tuginevaid testandmestikke. Need peaksid sisaldama nii faktitundmise küsimusi, kui ka loogikat ja loovust nõudvaid ülesandeid. Mudelid muutuvad aga iga järgmise uue versiooni turuletulekuga ainult võimekamaks ning suure tõenäosusega on uusi mudeleid treenitud ka kõige raskemaid juba teadaolevaid teste lahendama.

Arvestades suurte keelemudelite arendamisse tehtud investeeringutega, on ettevõtete jaoks eluliselt tähtis saavutada võimalikult kõrged kohad mudeleid võrdlevates edetabelites. Nii arenevad ka andmestikud pakkudes lahendamiseks järjest keerukamaid ülesandeid.

### 2.3.1 Inimkonna viimane eksam

Ülemaailmse koostöö tulemusena on valminud ambitsioonikat nime kandev „Inimkonna viimane eksam“ (ingl *Humanity's Last Exam*)<sup>5</sup>. Ligi 1000 erinevat töötajat 500 teadusasutusest kokku 50 riigist panustasid 2500 eriti keeruka küsimuse koostamisel (Phan et al., 2025).

Judge Model: o3-mini | Dataset Updated: April 3rd, 2025

Model	Accuracy (%) ↑	Calibration Error (%) ↓
 o3	20.3	34.0
 Gemini 2.5 Pro	18.4	71.0
 o4-mini	18.1	57.0
 o3-mini*	13.4	80.0
 Gemini 2.5 Flash	12.1	80.0
 DeepSeek-R1*	8.5	73.0
 Claude 3.7 Sonnet (16K)	8.0	80.0
 o1	8.0	83.0
 GPT-4.5 Preview	5.4	85.0
 GPT-4.1	5.4	89.0
 Claude 3.5 Sonnet	4.1	84.0
 GPT-4o	2.7	89.0

Joonis 1. Mudelite tulemused „Inimkonna viimase eksami“ lahendamisel<sup>5</sup>

Joonisel 1 kajastuvad suurte keelemudelite tulemused 25. aprilli 2025. aasta seisuga. „Accuracy“ veerg näitab mudeli täpsust, „Calibration Error“ ehk kalibreerimisvea veerus on mudeli hinnang oma vastuste õigsusele, st kui suure tõenäosusega peab mudel enda vastusevarianti õigeks. Tärniga tähistatud mudelid ei ole multimodaalsed (tekst ja pilt) ning nende täpsuse arvutamisel on arvesse võetud ainult tekstipõhiseid küsimusi.

---

<sup>5</sup> Humanity's Last Exam. <https://agi.safe.ai/> (23.04.2025)

### 2.3.2 Olümpiaadide küsimustel põhinevad testid

Inglisekeelsete olümpiaadide küsimuste põhjal on koostatud mitmeid testandmestikke. Üks mahukamaid on kakskeelne ja multimodaalne olümpiaaditaseme ülesannete küsimustik „Olympiadbench“, mis koondab endas 8476 matemaatika ja füüsika olümpiaadide ülesannet, lisaks Hiina ülikooli sisseastumiseksamite küsimusi (He et al., 2024).

Ainult matemaatikaolümpiaadide ülesannetele keskenduv „Omni-Math“ sisaldab 4428 küsimust, mis on 10 raskusastmega jaotatud 33 alamkategoriasse (Gao et al., 2024). Rootsi kõrgkoolide sisseastumiskatsete ülesannete kohta on koostatud 867 küsimusega andmestik, mis hõlmab muu hulgas ka loogika, tekstimõistmise ja matemaatilise probleemide lahendamise ülesandeid (Kurfah et al., 2025).

### 2.3.3 Eestikeelsed testandmestikud

Eesti keele osas on välja töötatud EstQA andmestik, mis tugineb Vikipeedia artiklitel ja koosneb „kontekst-küsimus-vastus“-kolmikutest, mis aitab hinnata testitava mudeli arusaamist eesti keelest (Käver, 2021). Kokku on treeninghulgas 776 ja testhulgas 603 kolmikut. Iga näite kohta on esitatud üks või mitu vastusevarianti.

EstCOPA andmestik sisaldab eeldust, kahte valikvastust, küsimuse tüüpi (põhjus või tagajärg) ja õiget vastust; kokku 100 näidet (400 treening-, 100 valideerimis- ja 500 testhulgas) (Kuulmets et al., 2022). Andmestiku eesmärk on hinnata suurte keelemudelit arusaamist põhjuslikkuse väljendamisel eesti keeles.

Belebele andmestik testib mudelite tekstist arusaamise oskust mitmekeelsete küsimuste abil (Bandarkar et al., 2024). Kokku 122 erinevat keelt hõlmavas testhulgas on iga tekstilõigu kohta koostatud üks küsimus ja esitatud neli vastusevarianti. Andmestiku paralleelsus (samad küsimused eri keeltes) võimaldab sama mudelit erinevates keeltes vahetult võrrelda<sup>6</sup>.

Bakalaureusetööd alustades ei olnud ühtegi eestikeelset andmestikku, mis oleks koostatud Eesti kooliastmete olümpiaadidele suunatud küsimuste põhjal. Seda lünka on lõputöö raames valminud küsimustik asunud täitma.

---

<sup>6</sup> Belebele. <https://huggingface.co/datasets/facebook/belebele>

### 3. Andmete kogumine ja töötlemine

Suurte keelemudelite võrdleva analüüsi koostamisel on käesolevas bakalaureusetöös kasutatud Eesti bioloogiaolümpiaadide teoreetiliste voorude küsimusi. Küsimuste ja vastuste PDF-failid asuvad Tartu Ülikooli teaduskooli kodulehel<sup>7</sup>. Valikut mõjutasid viis asjaolu:

1. küsimused on raskusastme järgi jagatud kolme erinevasse kategooriasse (gümnaasiumiaste, 8.–9. klass, 6.–7. klass), mis annab keelemudelite hindamisel laiapõhjalisemaid tulemusi;
2. küsimustele on antud vastusevariandid, mis hõlbustavad testimisel saadavate tulemuste automaatset võrdlemist (vabas vormis vastuste korral võib hindamisele lisanduda hindaja subjektiivsuse aspekt, mida valikvastustega küsimustikuga õnnestub vältida);
3. inimtööjõu kaasamine tuhandete vastuste hindamiseks on aeganõudev ja kallis;
4. usaldusväärsed mudelid, mida kasutatakse vabas vormis esitatud vastuste hindamiseks (ingl *LLM-as-judge*) (Zheng et al., 2023), ei ole eesti keele peal testitud;
5. küsimuste hulk on piisavalt suur, et viia juhuslike õigete ja valede vastuste osakaal võimalikult madalaks.

Bioloogiaolümpiaadide küsimused hõlmavad aine väga erinevaid alamkategoriaid. Nooremate kooliastmete küsimused keskenduvad peamiselt zooloogia, botaanika, ökoloogia ja inimese anatoomia valdkondadele. Gümnaasiumiõpilastele suunatud ülesanded käsitlevad lisaks ka geneetika, rakubioloogia, biokeemia, taimede ja loomade anatoomia, etoloogia, biosüsteematika ning evolutsiooniga seotud teemasid.

Kokku on bakalaureusetöö tulemusena koostatud andmebaasis 1031 kannet, millest gümnaasiumi taseme olümpiaadidest pärineb 778 (75,46%), 8.–9. klasside ülesannetest 137 (13,29%) ja 6.–7. klasside tasemest 116 (11,25%) küsimust.

#### 3.1 Toorandmete salvestamine

Teaduskooli veebilehel on bioloogiaolümpiaadide ülesanded koos vastustega aastatest 2005–2024. Valikust jäeti välja juhtum, kus küsimused ja vastused on esitatud samas PDF-failis

---

<sup>7</sup> Teaduskool. <https://teaduskool.ut.ee/et/bioloogiaolumpiaadi-arhiiv> (31.03.2025)

(2023/24 õppeaasta), sest see suurendab tõenäosust, et testitavad mudelid võivad vahetult olla treenitud ka nendel andmetel.

Olümpiaadide küsimusi ei olnud veebilehel 2010/11, 2011/12 ja 2015/16 õppeaastate kohta. Aastal 2009/10 toimunud mõõduvõtu kohta puudusid vastused. Kõik esialgsete kriteeriumite järgi sobivad küsimused koos vastustega laaditi kodulehelt alla; kokku 101 faili.

### 3.2 Andmestiku vorming

Keelemudelite hindamiseks vajalik küsimustik peab olema masinloetav ning selleks sobib JSON-vorming, kus on “võti-väärtus”-paaridena kajastatud kõik otseselt testimiseks vajalikud andmeväljad ning metaandmed.

```
"id": "6",
"küsimus": "Kaladel on üks eriline meeleelund, mis aitab tajuda vee liikumist ja võnkeid, olles nii abiks ruumis orienteerumisel. Mis on selle elundi nimi?",
"valikud": [
  "Sisekõrv",
  "Poised",
  "Küljejoon",
  "Elektrielund",
  "Uimekiired"
],
"vastus": "C",
"punktid": "1",
"aasta": "2023",
"aste": "II",
"teema": "bioloogia"
```

Joonis 2. Näidisküsimuse “võti-väärtus” paarid JSON-vormingus

Joonisel 2 on näitena välja toodud iga küsimuse kohta koondatud andmed. Lisaks küsimuse tekstile, vastusevariantidele ja õige vastuse väljale lisatud veel järgnevad metaandmed:

1. id – küsimuse järjekorranumber;
2. punktisumma – erineva raskusastmega küsimused annavad 1–3 punkti;
3. aasta – viitab olümpiaadi toimumise õppeaasta esimesele poole, st 2005–06 õppeaasta korral “2005”;
4. aste – I, II või III (vastavalt gümnaasiumiaste, 7.-8. klass, 6.-7. klass);
5. küsimuse teema.

Võtme „punktisumma“ erinevaid väärtusi ei võetud töö hilisemas käigus arvesse, sest rohkema kui ühe punktiga hinnatud küsimuste osakaal oli kogu andmestikust väike (5,86%). Siiski võib kogutud metaandmetest kasu olla hiljem, kui andmestikku lisaküsimustega täiendada.

Kuivõrd töö alguses ei olnud selge, kui palju küsimusi lõppandmestikku koguneb, siis olid võtmed „aasta“ ja „küsimuse teema“ mõeldud erinevate olümpiaadide vahel selgete piiride loomiseks. Neid metaandmeid ei ole samuti töö käigus rohkem kasutatud, kuid baasandmetesse on võtmed „küsimuse teema“ väärtusega „bioloogia“ ja „aasta“ vastava olümpiaadi toimumise ajaga alles jäetud.

### 3.3 Küsimuste eraldamine

Alla laaditud PDF-failidest küsimuste eraldamiseks kasutati programmeerimiskeelt Python (versioon 3.12.7)<sup>8</sup> ja selle standardteeke `re` ning `json`, teeki `pdfplumber`<sup>9</sup> (versioon 0.11.4) ja graafilist kasutuskeskkonda Thonny (versioon 4.0.0)<sup>10</sup>.

Küsimusi sisaldavate PDF-failide nimede struktuur on selline, et neist saab automaatselt eraldada aastaarvu ning kooliastme.

```
# Eraldame failinimest aastaarvu
def aasta_arv(faili_nimi):
    aasta = re.search(r"_(\d{4})_", faili_nimi)
    return aasta.group(1) if aasta else "unknown"
```

Joonis 3. Regulaaravaldise abil failinimest aastaarvu eraldamine

Joonisel 3 esitatud Pythoni kood otsib failinimest (nt “*ebo\_2005\_06\_pv\_ylesanded\_g.pdf*”) regulaaravaldise abil alakriipsudega eraldatud neljast numbrist koosnevaid jadasid ning omistab neist esimese väärtuse muutujale “aasta”. Hilisemas koodis täideti selle muutuja abil küsimustiku andmevälja “aasta” (vt joonis 2).

Analoogiliselt sai failinimest eraldada ka kooliastme, kus “g” vastab gümnaasiumiastmele, “8\_9” 8.–9. klassidele ning “6\_7” 6.–7. klassidele.

Andmete eraldamisel PDF-failist arvestati sellega, et nii küsimused kui ka vastused võisid olla paigutatud mitmele eraldi reale, vahel võis pikem küsimus ulatuda ka järgmisele leheküljele.

<sup>8</sup> Python 3.12.7. <https://www.python.org/downloads/release/python-3127/> (31.03.2025)

<sup>9</sup> Pdfplumber. <https://github.com/jsvine/pdfplumber> (31.03.2025)

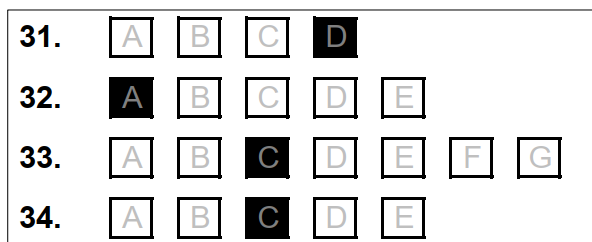
<sup>10</sup> Thonny. <https://thonny.org/> (31.03.2025)

Mitmerealiste küsimuste ja vastuste teisendamisel üheks loogiliselt seotud sõnade jadaks kasutati Pythoni meetodi koostamisel ChatGPT<sup>11</sup> (gtp-4-turbo, oktoober–detsember 2024) abi. PDF-failist eraldatud andmed salvestati vastavalt kooliastmele JSON-faili (nt “I\_küsimused.json” gümnaasiumi küsimuste puhul). Enne järgmise olümpiaadi PDF-faili töötlemist uuendati Pythoni koodis muutuja “id” selliselt, et see vastaks juba kogutud küsimuste koguarvule ning jätkaks väärtuste kasvatamist iga lisatava kandega.

### 3.4 Õiged vastused

Kõik õiged vastused on töö autor sisestanud JSON-faili käsitsi. Sellise valiku tingisid kaks asjaolu:

1. vastused ei ole sageli esitatud lihtsasti masinloetaval kujul (vt joonis 4);
2. oli kasutatud teistsugust vastuse esitamise kuju kui küsimuse koostamisel, nt õige vastuse väljakirjutamine või tähtede asendamine numbritega.



Joonis 4. Näide vastusefailides õigete valikute esitamise viisist

Joonisel 4 on kujutatud üht meetodit, kuidas ülesannete koostajad püüdsid muuta vastuste masinloetavust raskemaks. Kokku esines vastusefailides kolm erinevat obfuskeerimise viisi. Lisaks joonisel 4 esitatule kasutati veel täppide ja triipude lisamist vastusekastidesse ning õige vastusevariandi kajastamist rasvases kirjas valedest vastustest erineva fondiga.

Kuigi andmestiku koostamisel kontrolliti vastuste korrektset sisestamist kahel erineval korral, siis selgus hilisema testimise käigus, et vähemalt ühel juhul oli töö autor teinud vea. Kogutud 1031 küsimuse puhul võib eeldada, et selliste inimlike eksimuste mõju lõpptulemusele ei ole märkimisväärne. Lõplikel Hugging Face varamusse<sup>12</sup> üles laaditud testandmetes on leitud viga parandatud.

<sup>11</sup> OpenAI ChatGPT. <https://chatgpt.com/> (31.03.2025)

<sup>12</sup> Hugging Face varamu. [https://huggingface.co/ahtokiil/olympiad\\_et](https://huggingface.co/ahtokiil/olympiad_et) (04.05.2025)

Ühe küsimuse puhul tekkis kahtlus, et küsimuste koostajad olid eksinud õige vastuse märkimisel. Andmestikus on jätkuvalt õige vastusena märgitud variant, mida hindajad olid korrektseks pidanud.

Bakalaureusetöö koostaja on lähtunud Teaduskooli veebilehel avaldatud andmetest ning ei saa garanteerida nende lõplikku õigsust.

### **3.5 Juhusliku valiku tõenäosus**

Testandmestikus on kokku 1031 küsimust, mille vastusevariantide arv on kolm kuni kaheksa. Kui testitav mudel (või ka küsimustele vastav katseisik) vastaks igale küsimusele juhuslikult, siis võiks õigete vastuste osakaal olla 22,12%.

Tõenäosuse arvutamiseks leiti iga küsimuse juhuslikult õige vastuse tõenäosus (1/valikute arv), liideti vastuste tõenäosused kokku ning saadud tulemus jagati küsimuste arvuga.

### **3.6 Olümpiaadide võitjate õigete vastuste osakaal**

Ühegi aasta olümpiaadide küsimusi ei saanud 100% ulatuses kasutada, sest kõik teoreetiliste voorude ülesanded kasutasid illustratsioone, kuid testandmestikku võeti ainult tekstipõhiseid ülesandeid. Siiski on ligikaudse võrdlusmomendi väljatoomiseks leitud iga vanuseastme lõppvoorude teoreetilise vooru keskmine tulemus iga aasta kolme parima õpilase osas.

Gümnaasiumiastmes on kolme parima osaleja keskmine saavutus teooriaosas 84,29%, 8.–9. klasside osas 86,17% ning 6.–7. klasside õpilaste puhul 91,01%.

### **3.7 Käsitsi tehtud parandused**

Ükski automaatne tuvastussüsteem ei ole veatu (He et al., 2024) ning seetõttu on vastuste lisamise ajal parandatud JSON-failides leitud vigu – vigased täpitähed, kirjavead, erisümbolid ning lisatühikud.

Andmestiku koostamisel oli eesmärgiks võimalikult suure hulga küsimuste kogumine ning seetõttu jäeti andmete käsitsi redigeerimise käigus kokku 38 küsimuse juures vastusevariantidesse ainult üks õige vastus. Selline lähenemine lubas suurendada II ja III vanuseastme küsimuste hulka, kus valdavalt oli ühele küsimusele rohkem kui üks õige vastus.

### 3.8 Andmestiku ja lähtekoodide varamud

Kõik bakalaureusetöö käigus kogutud toorandmed, koostatud küsimustikud ning selleks kasutatud koodifailid on salvestatud GitHubi varamus<sup>13</sup>. Võimaliku andmelekke vältimiseks (Balloccu et al., 2024) on hetkel tegemist privaatse repositooriumiga.

Pythoni koodifailid, mida kasutati andmete kogumisel, redigeerimisel ja koondamisel, on GitHubi avalikus varamus<sup>14</sup>. Samas on võimalik tutvuda ka küsimustiku testimisel kasutatud mallide ning skriptidega.

---

<sup>13</sup> Bakalaureusetöö privaatne koodivaramu. [https://github.com/ahtokiil/ut2025\\_BA](https://github.com/ahtokiil/ut2025_BA) (31.03.2025)

<sup>14</sup> Bakalaureusetöö avalik koodivaramu. <https://github.com/ahtokiil/bakalaureus> (12.04.2025)

## 4. Tulemused

Testimine viidi ilma näidisülesandeid tutvustamata (ingl *zero-shot evaluation*) läbi kolmes etapis kasutades kolme erinevat keskkonda:

1. OpenAI mudeleid testiti API kaudu kasutades Tartu Ülikooli Azure kontot, millele juhendajad võimaldasid juurdepääsu;
2. platvormile Hugging Face üles laetud mudelite testimiseks eraldati bakalaureusetöö tarvis arvutusressurssi Tartu Ülikooli HPC Rocket klastris;
3. kõige uuemaid kommertskasutuses olevaid keelemudeleid testiti läbi veebibrauseri käsitsi küsimusi vestlusaknasse sisestades.

Kõikide testimiste tulemused on salvestatud GitHubi privaatsesse varamusse.

### 4.1 OpenAI mudelite testimine

Azure AI Foundry<sup>15</sup> on Microsofti platvormiteenus, mis võimaldab huvilistel arendada ning katsetada suuri keelemudeleid. Tartu Ülikooli Azure AI Foundry kontot kasutades oli võimalik testida järgnevat OpenAI mudeleid:

1. gpt-3.5-turbo (versioon 0125);
2. gpt-4o-mini (versioon 2024-05-18);
3. gpt-4-turbo (versioon 2024-04-09);
4. gpt-4o (versioon 2024-05-13).

Lisaks kasutati bakalaureusetöö autori isiklikku kontot OpenAI platvormil<sup>16</sup>, et esitada küsimused mudelile o1-mini (versioon 2024-09-12), mis olnud Tartu Ülikooli konto valikus.

Testimise läbiviimise ajal oli kõige uuemaks mudeliks o1-preview, aga see jäi kõigi küsimuste hindamisest välja kõrge hinnataseme tõttu. Mudelit on testitud hiljem välja valitud raskemate küsimuste andmestikuga (vt alapeatükk 4.3).

OpenAI mudeleid testiti kasutades Pythonis kirjutatud programmi, mis esitas mudelile läbi API küsimust sisaldava viiba (ingl *prompt*) ning piiravad parameetrid, et maksimaalselt mitme tähemärgiga on mudelil lubatud vastata.

---

<sup>15</sup> Azure AI Foundry. <https://ai.azure.com/> (04.05.2025)

<sup>16</sup> OpenAI Platform. <https://platform.openai.com/> (04.04.2025)

```

# Päringu esitamine mudelile
vastus = openai.ChatCompletion.create(
    engine="gpt-4o-mini-olympiadbench",
    messages=[
        {"role": "system", "content": "Sa oled bioloogiaolümpiaadil
        osalev õpilane Eestis. \nPalun anna lõplik vastus ainult ühe
        suurtähega:"},
        {"role": "user", "content": viip}
    ],
    temperature=0,
    top_p=1,
    max_tokens=1
)
# Vastuse eraldamine
mudeli_vastus = vastus["choices"][0]["message"]["content"].strip()

```

Joonis 5. Väljavõte OpenAI mudelite testimise programmi koodist

Joonisel 5 on näitlikustatud ühte küsimust sisaldava API-päringu sisu. Küsimuse koostamisel antakse mudelile selgitav viip „Sa oled bioloogiaolümpiaadil osalev õpilane Eestis.“. Parameeter *temperature=0* määrab, et mudel ei tugineks vastuse genereerimisel juhuslikkusele, *top\_p=1* tagastab suurima tõenäosusega vastuse ja *max\_token=1* seab vastuse pikkuseks maksimaalselt ühe tähemärgi.

Joonise 4 viimasel real on mudeli vastuse eraldamine tagastussõnest. ChatCompletion API tagastab sõnastiku, kus peamiseks tulemuseks on võtmega *choices* seotud nimekiri võimalikest vastustest (kui mudelile on vastuse koostamisel lubatud varieerimine). Parameeter *top\_p=1* määrab ainult ühe variandi tagastamise ning see sõne omistatakse muutujale *mudeli\_vastus*.

#### 4.1.1 OpenAI mudelite üldine täpsus

Tabel 1. OpenAI mudelite täpsus sõltuvalt vastuse lubatud pikkusest

Mudel	üks tähemärk	100 tähemärki
3.5-turbo	47.91%	48.01%
4o-mini	68.19%	68.77%
o1-mini	76.53%	76.62%
4-turbo	80.21%	81.38%
4o	85.35%	85.35%

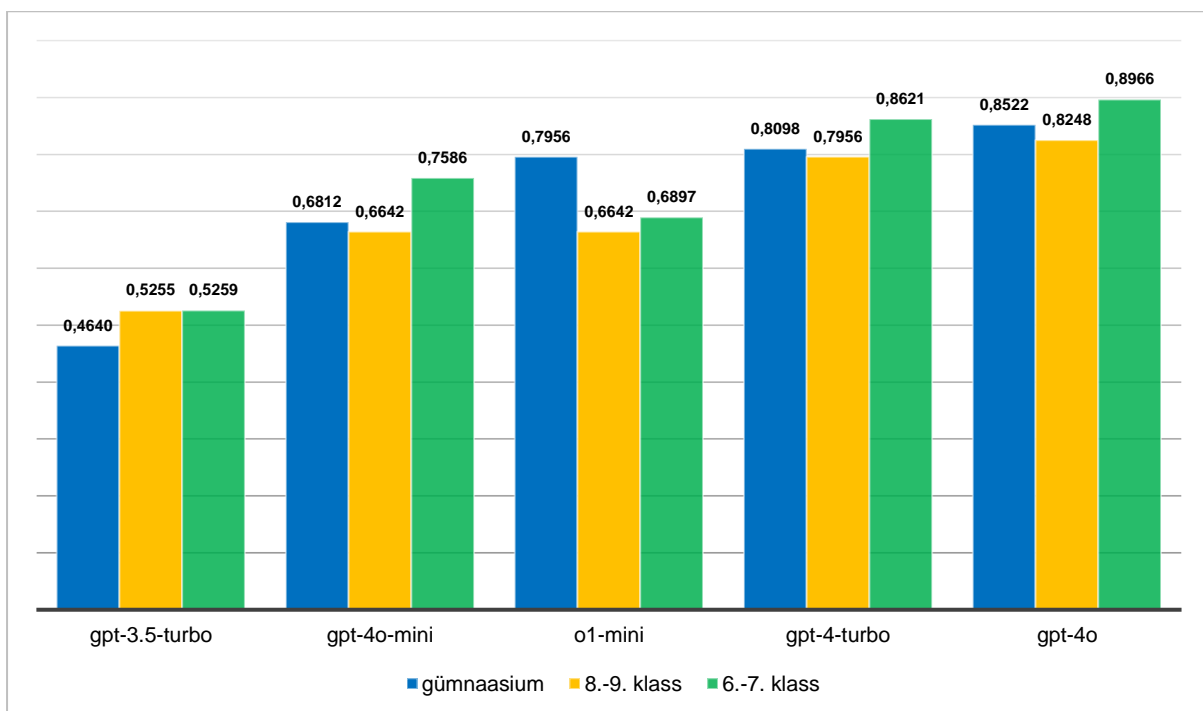
Tabel 1 näitab, et kõik OpenAI mudelid said neile antud juhustest õigesti aru ning vastuste täpsus ei sõltunud oluliselt lubatud tähemärkide arvust. Viimases tulbas näidatud täpsuste arvutamisel on kõik vastused käsitsi üle vaadatud, sest mudelitel oli lubatud vastata kuni 100 tähemärgiga. Kui selgesti väljendatud valik sisaldus tagastatud tekstis, siis märgiti pakutud variant vastuseks.

Kui vastuses oli lubatud kuni 100 tähemärki, siis mudel 4-turbo kasutas kuuel korral vastamiseks konstruktsiooni „Vastus on:“, kolmel korral „Õige vastus on:“, kolmel korral ei jõudnud mudel 100 tähemärgiga vastuseni ning kolmel korral oli vastusevariant tagastatud 100 tähemärgi hulgas.

Mudel 4o lisas kolmel korral vastuse teksti pärast seda, kui oli vastanud esmalt ühe suurtähega (nt „A. Koduvarblane“). Ülejäänud mudelid kasutasid rohkemate tähemärkidega vastamisel suuremat vabadust ainult valikuliselt punkti lisamiseks suurtähest koosnevale vastusele (nt „B“ vs. „B.“).

#### 4.1.2 OpenAI mudelite täpsus vanuseastme kaupa

Töö käigus võrreldi OpenAI mudelite täpsust ka erinevate kooliastmete lõikes.



Joonis 6. OpenAI mudelite täpsus kooliastmete kaupa

Jooniselt 6 on näha, et reeglina oli ajaliselt uuemate ning paremate mudelite täpsus küsimustele vastamisel kõrgem. OpenAI ei ole oma mudelite tehnilisi andmeid avalikuks teinud ning

seetõttu ei saa ka analüüsida, kas ühe või teise mudeli tulemus on seotud treeningandmete hulga, parameetrite arvu või peenhäälestusega.

## 4.2 Platvormil Hugging Face avaldatud mudelite testimine

Paljud arendajad on teinud mitmed oma mudelid avalikkusele kättesaadavaks platvormil Hugging Face. Kasutaja saab neid lokaalsesse arvutisse või pilveressurssi alla laadida, testida ja edasi arendada vastavalt iga mudeli konkreetsele kasutuslitsentsile.

Bakalaureusetöö raames võimaldas Tartu Ülikool autorile juurdepääsu HPC Rocket klastrile<sup>17</sup>, millel on 72 arvutussõlme (ingl *node*) erineva võimekusega arvutus- ja graafikaprotsessoritega ning kokku 4,7 petabaiti salvestusruumi. Klastrisse on võimalik suured keelemudelid alla laadida ja `lm-evaluation-harness` teeki kasutades testida.

### 4.2.1 Testimiseks valitud mudelid

Testimisele kaasatud mudelite valikus lähtuti juhendajate soovitudest ning platvormil Hugging Face avaldatud Eesti keelel treenitud mudelite edetabelist<sup>18</sup>. Hugging Face kasutab oma pingerea koostamisel EuroLingua-GPT<sup>19</sup> andmeid, mis tuginevad viiel levinud keelemudelite hindamise andmestikul:

1. ARC – alg- ja keskastme koolitase 7787 valikvastuste küsimust (Clark et al., 2018);
2. GSM8K - 8500 põhikooli matemaatika tekstülesannet (Cobbe et al., 2021);
3. HellaSwag - valikvastustega testandmestik ette antud teksti mõistmise hindamiseks (Cobbe et al., 2021);
4. MMLU - suur valikvastustega küsimustik, mis hõlmab 57 erinevate ülesannet (Hendrycks et al., 2021);
5. TruthfulQA - 817 küsimust 38-st erinevast kategooriast hindamaks mudeli vastuste õigsust (Lin et al., 2022).

Kõigi viie testi tulemuste keskmiste põhjal on koostatud edetabel, mida 2025. aasta mais juhtis Meta mudel `Llama-3.1-70B-Instruct` tulemusega 0,66.

---

<sup>17</sup> UT HPC Center, Rocket. <https://hpc.ut.ee/services/HPC-services/Rocket> (04.04.2025)

<sup>18</sup> European LLM Leaderboard. <https://huggingface.co/spaces/Eurolingua/eurolingua-european-llm-leaderboard> (04.05.2025)

<sup>19</sup> Eurolingua-GPT. <https://www.ai.se/en/project/eurolingua-gpt> (01.05.2025)

Tabel 2. Platvormil Hugging Face avaldatud mudelite andmed

Mudel	Parameetrite arv (miljardit)	Peen-häälestatud	Arendaja
<b>Gemma-7B</b>	8,54	EI	Google
<b>Mistral-Nemo-Base-2407</b>	12,2	EI	MistralAI, NVIDIA
<b>EuroLLM-9B</b>	9,15	EI	EL
<b>Teuken-7B-Instruct-v04</b>	7,45	JAH	Fraunhofer
<b>Qwen-7B-Instruct</b>	7,62	JAH	Alibaba
<b>Llama-3.1-8B-Instruct</b>	8,03	JAH	Meta
<b>Mistral-Nemo-Instruct-2407</b>	12,2	JAH	MistralAI, NVIDIA
<b>EuroLLM-9B-Instruct</b>	9,15	JAH	EL
<b>Gemma-2-9B-Instruct</b>	9,24	JAH	Google
<b>Mistral-Small-Instruct-2501</b>	23,6	JAH	MistralAI
<b>Gemma-2-27B-Instruct</b>	27,2	JAH	Google
<b>Llama-3.1-70B-Instruct</b>	70,6	JAH	Meta
<b>Llama-3.3-70B-Instruct</b>	70,6	JAH	Meta

#### 4.2.2 Testimisprotseduur UT HPC-Rocket klastris

Arvutuskeskuses HPC-Rocket on kasutusel klastrite haldussüsteem SLURM<sup>20</sup> (ingl *Simple Linux Utility for Resource Management*), mis juhib kasutajate tööde järjekorda, ressursside broneerimist ja paralleelarvutusi.

Pärast *lm-harness-evaluation* teegi paigaldamist ja uue hindamisülesande (ingl *evaluation task*) koostamist sai esitada testimiskripti, mis laadis platvormilt Hugging Face HPC Rocket klastrisse soovitud mudeli, käivitas hindamisülesande ja salvestas tulemuse SLURM-vormingus faili.

---

<sup>20</sup> SLURM. <https://slurm.schedmd.com/overview.html> (01.05.2025)

```

task: olympiad
dataset_path: ahtokiil/olympiad_et
test_split: train

output_type:
multiple_choice

doc_to_text: "input"
doc_to_target: "target"
doc_to_choice:
"choices"

metric_list:
- metric: acc

```

Joonis 7. lm-eval hindamisülesanne

Joonisel 7 on hindamisraamistiku `lm-evaluation-harness` ülesande konfiguratsioon, mis laadib platvormilt Hugging Face valikvastustega testi nimega *olympiad*, mis asub varamus *ahtokiil/olympiad\_et*. Mudeli sisendiks määratakse väli *input*, õige vastus asub väljal *target* ning valikvastuste loend võetakse väljalt *choices*. Tulemusi hinnatakse täpsuse (ingl *accuracy*, *acc*) alusel.

```

#!/bin/bash
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --time=2:00:00
#SBATCH --mem=160GB
#SBATCH --partition=gpu
#SBATCH --gres=gpu:a100-80g:2
#SBATCH --job-name=eval
module load any/python/3.8.3-conda
source ~/.bashrc
conda activate lm-harness
lm_eval --model hf \
  --verbosity INFO \
  --model_args pretrained=meta-llama/
  Llama-3.3-70B-Instruct,dtype="bfloat16",parallelize="True" \
  --apply_chat_template \
  --gen_kwargs top_p=1,do_sample=False,temperature=0 \
  --tasks olympiad \
  --batch_size 1 \
  --output_path testing_results \
  --write_out \
  --log_samples \
  --show_config

```

Joonis 8. Näidiskript SLURM-töökäsu käivitamiseks

Joonisel 8 on kirjeldatud Meta Llama-3.3-70B-Instruct mudeli testimine. 70 miljardi parameetriga mudeli käivitamiseks on vajalik kahe A100-80g-tüüpi graafikakiirendi kasutamine (`--gres=gpu:a100-80g:2`) ning 160 GB põhimälu (`--mem=160GB`) eraldamine. `lm-eval` käivitab hindamisülesande koos suunistega anda vastuseks suurima tõenäosusega valik.

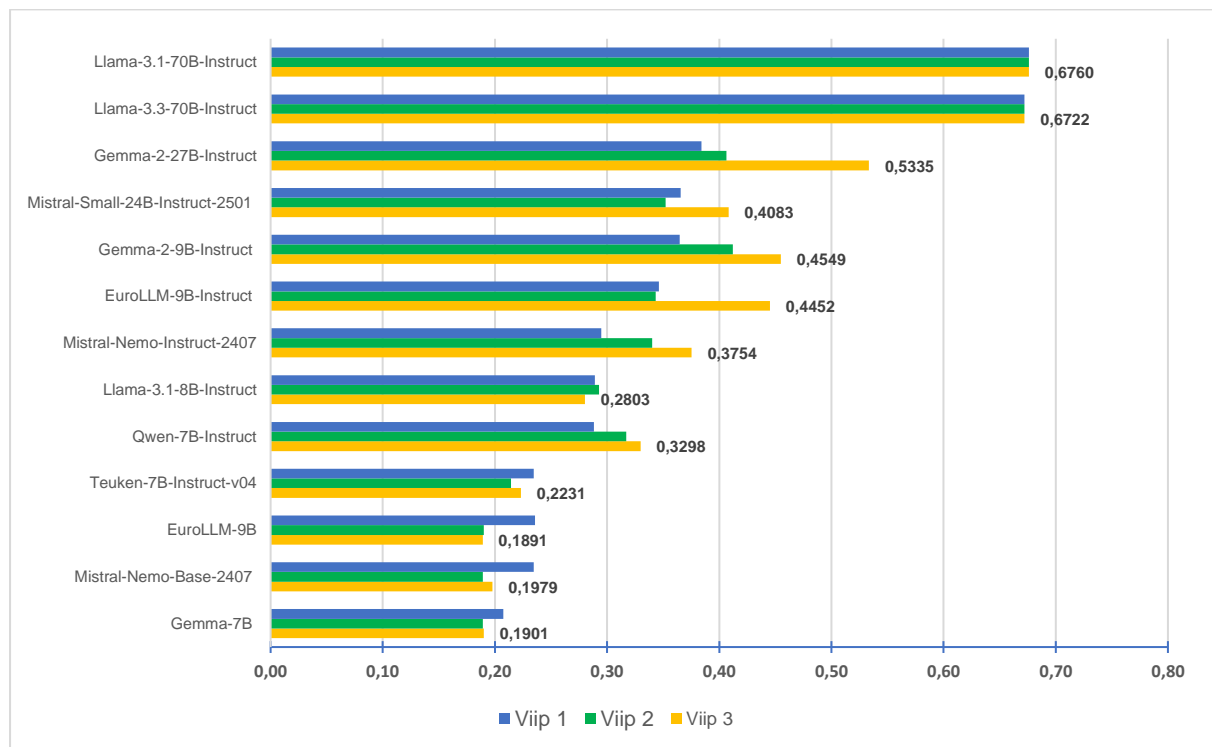
```
...
hf (pretrained=meta-llama/Llama-3.3-70B-
Instruct,dtype=bfloat16,parallelize=True), gen_kwargs:
(top_p=1,do_sample=False,temperature=0), limit: None, num_fewshot:
None, batch_size: 1|
Tasks      |Version| Filter|n-shot|Metric|Value | Stderr |
|-----|-----|-----|-----|-----|-----|-----|
|olympiad|Yaml   |none   |0|acc  |0.6722|± 0.0146|
```

Joonis 9. Väljavõte `lm-eval` käsu väljundfailist

Joonisel 9 on näha testimisprotseduuri lõpuks `lm-eval` käsuga koostatud raporti lõpposa, mis kajastab mudeli tulemust ehk täpsust (0,6722) olümpiaadiküsimustele vastamisel.

### 4.2.3 Platvormil Hugging Face avaldatud mudelite testimise tulemused

Kõiki 13 mudelit testiti ühesuguste testandmestikega ning viipadega.



Joonis 10. Platvormil Hugging Face avaldatud mudelite testimise tulemused

Joonisel 10 on kujutatud avalikult kättesaadavate parameetritega suurte keelemudelite testimise tulemused kolmel erineval juhiste tasemel:

1. „Viip 1“ näitab mudeli täpsust, kui küsimus on esitatud kujul:  
"Millised on inimese normaalse vererõhu väärtused rahuolekus? Vali üks õige vastus.  
Valikud:  
"Diastoolne 70-90 mmHg; süstoolne 100-140 mmHg",  
"Diastoolne 100-140 mmHg; süstoolne 70-90 mmHg",  
"Diastoolne 40-60 mmHg; süstoolne 60-90 mmHg",  
"Diastoolne 100-110 mmHg; süstoolne 140- 160 mmHg""  
"valikud": ["Diastoolne 70-90 mmHg; süstoolne 100-140 mmHg", "Diastoolne 100-140 mmHg; süstoolne 70-90 mmHg", "Diastoolne 40-60 mmHg; süstoolne 60-90 mmHg", "Diastoolne 100-110 mmHg; süstoolne 140- 160 mmHg"]".  
Mudel peab vastama ennustatava õige vastuse indeksiga 0, 1, 2 või 3 (indeksite arvestus algab nullist).
2. „Viip 2“ kajastab tulemusi, kui valikuvariantidele olid lisatud tähed:  
"Kuidas saab määrata põtrade ligikaudset vanust? Vali õiged väited.  
1. Põdralehmade sarvede arengu järgi;  
2. Põdrapullide sarvede arengu järgi;  
3. Molaarsuse järgi;  
4. Hammaste kulumise järgi;  
5. Lõua all rippuva karvadega kaetud nahavoldi ehk „habeme“ järgi.  
Valikud: "2, 4, 5", "1, 2", "ainult 4", "3, 5", "mingi muu kombinatsioon",  
"valikud": ["2, 4, 5", "1, 2", "ainult 4", "3, 5", "mingi muu kombinatsioon"]".  
Mudel peab vastama indeksiga vahemikust 0–3.
3. „Viip 3“ toob esile mudelite täpsuse, kui tähtedega varustatud vastusevariantidele eelnes täpsusustav tekst:  
"Oled Eesti bioloogiaolümpiaadil osalev õpilane. Ülesanne: Nimeta Eestis enimlevinud haudelind. Vasta ühe suurtähega A, B, C, D, E valikute seast.  
Valikud: "A. Talvike",  
"B. Naerukajakas",  
"C. Linavästri",  
"D. Koduvarblane",  
"E. Väike-kirjurähn".

Vali õige vastus järgmistest valikutest: A. Talvike B. Naerukajakas C. Linavästri D. Koduvarblane E. Väike-kirjurähn Vasta ühe tähega (A, B, C, D või E).",  
"valikud": ["A. Talvike", "B. Naerukajakas", "C. Linavästri", "D. Koduvarblane",  
"E. Väike-kirjurähn"]"

Mudel peab vastama indeksiga vahemikust 0–4.

Alusmudelite (Gemma-7B, Mistral-Nemo-Base-2407 ja EuroLLM-9B) tulemused jäid juhusliku valiku piiridesse, kuid peenhäälestatud mudelite vastused paranesid koos parameetrite arvuga, va 70 miljardi parameetriga Llama mudelid, mis saavutasid kõikide viipadega täpselt samu tulemusi.

### 4.3 Uusimate mudelite testimine veebis

Kui OpenAI ja platvormil Hugging Face avaldatud mudelite testimine on reprodutseeritav, siis veebilehitseja kaudu suurte keelemudelite proovilepanek rahuldab autori isiklikku huvi.

„Raskemate küsimustega“ andmestiku koostamisel lähtuti eeldusest, et kui ajalisel vanemad mudelid teavad küsimusele vastust, siis teavad seda ka uuemad mudelid. Sel põhimõttel koostati hulk küsimustest (kokku 103 tk), millele kõik OpenAI mudelid vastasid valesi.

Testimiseks valiti mudelid bakalaureusetöö autori isiklike eelistuste ja juhendajate soovitude koostööna. Selle andmestiku analüüsi tulemusena saab teha järelduse, et raskusi valmistasid kahte tüüpi küsimused:

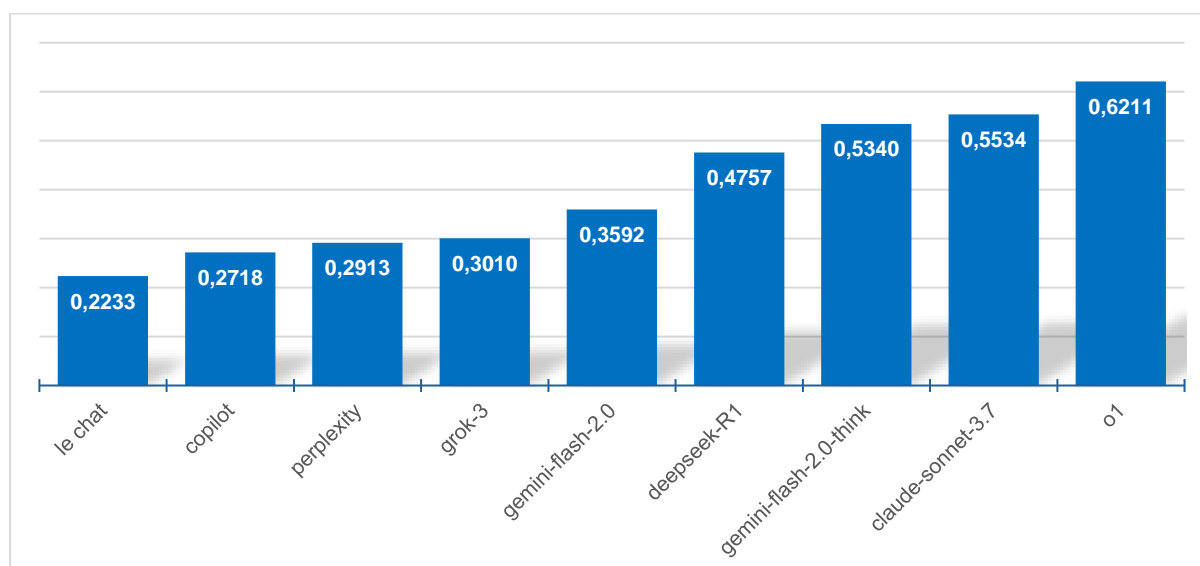
1. väga spetsiifiliselt Eesti loodust puudutavad küsimused (nt „Kes oli Eesti 2009. aasta lind?“);
2. mitme vastusevariandi kombineerimist nõudvad küsimused (nt „Labori katses kasvatati kolvis bakterikultuuri koos erinevate bakteriofaagidega (bakteriviirused). Teisel nädalal oli keskkond muutunud väga toitainetevaaseks. Missugused väited on tõesed teisel katsenädalal: 1. bakteriofaagid on pigem lüütilises elutsükklis; 2. bakteritel on suur puudus steroidhormoonidest; 3. bakterisporid on hakanud idanema; 4. bakteriofaagid on pigem lüsogeenses elutsükklis; 5. bakterite kasv on lakanud?“).

Seevastu näiliselt keerulistele rakubioloogiat, geneetikat ning anatoomiat puudutavatele küsimustele oskasid mudelid täpsemalt vastata. Võib eeldada, et universaalseid teadmisi on mudelid omandanud keeltes, kus taoliste andmete osakaal on olnud piisavalt suur. Eestikeelsete treeningandmete suhteliselt väike hulk võib selgitada kohaliku floora ja fauna kohta käivate teadmiste puudulikkust.

Tabel 3. Veebis testitud mudelite andmed

Mudel	Testimise aeg (2025)	Arendaja
<b>Copilot</b>	10.-11.03	Microsoft
<b>Le Chat</b>	20.-23.03	MistralAI
<b>Perplexity</b>	20.-23.03	Perplexity AI
<b>Grok-3</b>	10.-12.03	xAI
<b>Gemini Flash 2.0</b>	10.-12.03	Google
<b>Gemini Flash 2.0 Think</b>	12.-14.03	Google
<b>Deepseek R1</b>	10.-13.03	Deepseek
<b>Claude Sonnet 3.7</b>	10.-15.03	Anthropic
<b>o1</b>	10.03-02.04	OpenAI

Testimise periood sõltus iga mudeli puhul selle tasuta päringute arvust kindla ajavahemiku jooksul. OpenAI piiras ka ChatGPT+ tellimusega o1 mudeli testimist 50 küsimusega nädalas. Teised teenusepakkujad (nt Anthropic ja xAI) peatasid reeglina limiidi täitumisel juurdepääsu mudelile 3–6 tunniks.



Joonis 11. Kommertsmudelite testimise tulemused (103 küsimust)

Joonisel 11 on kajastatud veebis testitud kommertsmudelite tulemused „raske küsimustiku“ ülesannete lahendamisel.

## 5. Järeldused

Anna A. Ivanova hoiatab suurte keelemudelite hindamisel liigsete järelduste tegemisest, kuid leiab siiski, et täpsete vastusevariantidega ülesannete puhul on võrdlus inimestega põhjendatud (Ivanova, 2025). Olümpiaadiülesannete küsimustiku põhjal saab testimistulemuste põhjal võrrelda nii mudelite kolme gruppi eraldi kui ka omavahel.

### 5.1 OpenAI mudelid

gpt-3.5-turbo tehti avalikkusele kättesaadavaks 2022. aasta novembris<sup>21</sup>. Kui selle mudeli treeningandmete lõpp-kuupäev oli testversioonil 2025. aasta jaanuar (kõikidest testitud mudelitest hilisem, vt alampeatükk 4.1), siis jäi see selgelt alla sama arendaja uuematele mudelitele.

Ülejäänud OpenAI mudelite tulemused ei korreleeru täpselt populaarse Chatbot Arena edetabeliga<sup>22</sup>, kus gpt-3.5-turbo asub kohal nr 128, gpt-4o-mini 40. kohal, gpt-4-turbo 57. kohal, gpt-4o positsioonil 34 ja o1-mini kohal nr 23. Üheks põhjuseks on edetabelis välja toodud mudelite uuemad versioonid ning teiseks, erinevate testandmestike kasutamine mudelite järjestamisel.

### 5.2 Avatud parameetritega mudelid

Alusmudelite täpsus jäi olümpiaadide küsimustele vastamisel juhusliku vastamise tasemele hoolimata sellest, kas viip sisaldas endas ka selgitavaid juhiseid või mitte. Pigem tekitasid täpsustavad viivad suuremat raskust ülesandest arusaamisel ning mudelite täpsus langes (Gemma-7B puhul vastavalt 8,9% ja 8,4%; Mistral-Nemo-Base-2407 mudelil 19,4% ja 15,7%; EuroLLM-9B testimisel 19,3% ja 19,8%).

Peenhäälestatud mudelite testimisel andsid täpsemad juhised ka paremaid tulemusi, va Teuken-7B-Instruct-v04, mille täpsus langes nii testides „Viip 2“ (8,6%) kui ka „Viip 3“ (4,9%). Üldiselt järgisid mudelid trendi, et mida rohkem parameetreid, seda täpsemad vastused.

Eraldi tuleb välja tuua Llama-3.1-Instruct ja Llama-3.1-Instruct (mõlemad mudelid 70,6 miljardi parameetriga) testimise tulemused. Nende keelemudelite vastuste täpsus ei

---

<sup>21</sup> ChatGPT-3.5. <https://en.wikipedia.org/wiki/GPT-3#GPT-3.5> (11.05.2025)

<sup>22</sup> Chatbot Arena LLM Leaderboard. <https://lmarena.ai/?leaderboard> (11.05.2025)

sõltunud viibas esitatud selgitavast infost ning saavutatud tulemus on võrreldav OpenAI mudeliga gpt-4o-mini (vastavalt 67,6%, 67,2% ja 68,8%).

### 5.3 Vestlusakna kaudu testitud kommertsmodelid

Oluline on rõhutada, et kuigi ühegi mudeli suhtlusaknas ei aktiveeritud internetist otsimise või süvaanalüüsi valikuid, siis jäi kõigi nelja parima mudeli puhul lõputöö autorile mulje, et modelid kasutasid internetiotsingut.

Nimelt kuvasid nt Deepseek ja o1 oma vastuse genereerimise ajal ekraanile teksti, kust võis lugeda fraase „vaatan viimaseid andmeid“ (ingl *looking up latest data*) või „otsin“ (ingl *searching*). Gemini ja Claude tegid mõnele küsimusele vastates viite veebiaadressitele, kust võis vastuse õigsust kontrollida.

Ei saa välistada, et piiratud tasuta versioonide suure populaarsuse tõttu kuvati ekraanile vaid simulatsiooni mudeli „mõtteprotsessist“ ning samal ajal oli päring lihtsalt ootejärjekorras. Samuti võisid modelid veebiaadressitele viidata ka oma treeningandmete põhjal.

Huvitava leiuna võib välja tuua mudeli Claude vastuse 14.03.2025 esitatud olümpiaadiküsimusele: „Juba mitmeid aastaid valib Eesti Ornitoloogiaühing aasta linnu. Kes oli 2006. aasta lind? Vasta ühe suurtähega A, B, C, D, E valikute seast. Valikud: A) kanakull, B) merikotkas, C) hänilane, D) punarind, E) aul.“

Claude vastas: „Kahjuks ei ole mul täpset teadmist, kes oli 2006. aasta lind Eestis, kuna minu andmebaas ei sisalda nii spetsiifilist informatsiooni kõigi Eesti aasta lindude kohta. Eesti Ornitoloogiaühing on tõepoolest valinud igal aastal aasta linnu, kuid ma ei saa kindlalt öelda, milline valikutest oli 2006. aasta lind. Ma ei tahaks anda vale vastust, eriti olümpiaadi kontekstis. Kui sul on võimalik, võiksid kontrollida seda infot Eesti Ornitoloogiaühingu kodulehelt või muudest usaldusväärsetest allikatest.“

Samal päeval ilmus ERRi portaalis artikkel „Tehisaruproffessor: usaldusväärne keelemudel võiks oma kõhklusel tunnistada“. Ajakirjaniku küsimustele vastates ütles Tartu Ülikooli tehisintellekti professor Meelis Kull muu hulgas: „... Ideaalis võiks mudel selle töö tulemusena lisada tõlkele infot nagu "Seda sõna ma tegelikult ei teadnud" ...“ (Harrik, 2025).

Bakalaureusetöö autori arvates saab sellisest kokkusattumisest teha vaid järelduse, et suurte keelemudelitega toimuv areng ei ole ühelegi üksikisikule ega ka institutsioonile sisuliselt hoomatav.

Professor Kull leiab samas artiklis, et mudelite usaldusväärsuse puhul on lahendamata küsimus vastuse seletatavuses, ehk miks masin ühele või teisele küsimusele just selliselt vastas. Sama küsimus tekkis lõputöö autoril veebimudelite testimisel kontrollandmestikuga, mille hulgas oli ka küsimus "Milline väide inimese kujunemise kohta on vale? Valikud: A. Esimesi inimese liini esindajaid nimetatakse australopiteekideks, B. Nüüdisinimene arenes Aafrikas, C. Koos nüüdisinimesega elasid Aafrikas ka neandertallased, D. Samal ajal nüüdisinimesega elasid Euroopas ja Aasias ka neandertallased."

gpt-4 vastas 22.02.2025: "Miks sa ise sellele küsimusele ei vasta."

## 6. Kokkuvõte

Kõrgkooliõpikus „Praktiline andmeteadus“ on suurte keelemudelite kohta kirjutatud, et tegemist on piirangute ja probleemidega mudelitega, mis paistavad silma keele mõistmise poolest (Sügis et al., 2025, 193). Piiranguteks ja puudusteks peavad autorid aegunud andmeid, valdkonnaspetsiifiliste teadmiste puudumist, andmete privaatsusega seotud probleeme ning mõnikord esinevaid hallutsinatsioone.

Kõigi nende probleemidega puutus lõputöö autor kokku veebilehtedel esitatud küsimuste vastuseid lugedes ja analüüsid. Vastamata küsimuste kõrval on tänapäevased tippmudelid suutelised andma aga ka õigeid vastuseid koos väga põhjalike selgituste ning loogiliste mõttekäikudega.

Kokkuvõtvalt saab öelda, et 2025. aasta alguse seisuga lahendavad uusimad suured keelemudelid Eesti bioloogiaolümpiaadide teoreetilise voo valikvastustega ülesandeid parimate õpilastega võrreldaval tasemel. Kuidas saavad mudelid hakkama multimodaalsete ehk jooniseid, pilte, graafikuid vms sisaldavate ülesannetega, jääb tulevikus teiste uurimistööde vastata. Veelgi suurem ülesanne on suurte keelemudelitega tekkinud uute võimaluste tõhus rakendamine ennekõike õppe- ja teadustöösse, aga ka paljudesse teistesse valdkondadesse.

Bakalaureusetöö autori hinnangul on suurte keelemudelite näol tegemist mitmekülgse tööriistaga, mille oskuslik kasutamine eeldab kas formaalset koolitust või põhjalikku eeltööd, et saada soovitud tulemusi ilma enda või teiste heaolu ohtu seadmata. Seetõttu on Vabariigi Presidendi poolt ellu kutsutud TI-Hüppe<sup>23</sup> projekt väga kiiduväärt ettevõtmine, mis peaks aitama kasvatada sihipärast ja vastutustundlikku suhtumist suurte keelemudelite kasutamisse.

---

<sup>23</sup> TI-Hüpe. <https://www.aileap.ee/> (11.05.2025)

## Kasutatud allikad<sup>24</sup>

Balloccu, S., Schmidtová, P., Lango, M., & Dušek, O. (2024). *Leak, Cheat, Repeat: Data*

*Contamination and Evaluation Malpractices in Closed-Source LLMs* (No.

arXiv:2402.03927). arXiv. <https://doi.org/10.48550/arXiv.2402.03927>

Bandarkar, L., Liang, D., Muller, B., Artetxe, M., Shukla, S. N., Husa, D., Goyal, N.,

Krishnan, A., Zettlemoyer, L., & Khabsa, M. (2024). The Belebele Benchmark: A

Parallel Reading Comprehension Dataset in 122 Language Variants. *Proceedings of*

*the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

*Long Papers)*, 749–775. <https://doi.org/10.18653/v1/2024.acl-long.44>

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A.,

Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G.,

Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodi, D.

(2020). Language Models are Few-Shot Learners. H. Larochelle, M. Ranzato, R.

Hadsell, M. F. Balcan, & H. Lin (Toim), *Advances in Neural Information Processing*

*Systems* (Kd 33, lk 1877–1901). Curran Associates, Inc.

[https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfcb4967418bfb8](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8)

ac142f64a-Paper.pdf

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O.

(2018). *Think you have Solved Question Answering? Try ARC, the AI2 Reasoning*

*Challenge* (No. arXiv:1803.05457). arXiv. <https://doi.org/10.48550/arXiv.1803.05457>

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek,

J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). *Training Verifiers to*

---

<sup>24</sup> Koostamisel on kasutatud tarkvara Zotero. <https://www.zotero.org/> (15.05.2025)

- Solve Math Word Problems* (No. arXiv:2110.14168). arXiv.  
<https://doi.org/10.48550/arXiv.2110.14168>
- Gao, B., Song, F., Yang, Z., Cai, Z., Miao, Y., Dong, Q., Li, L., Ma, C., Chen, L., Xu, R., Tang, Z., Wang, B., Zan, D., Quan, S., Zhang, G., Sha, L., Zhang, Y., Ren, X., Liu, T., & Chang, B. (2024). *Omni-MATH: A Universal Olympiad Level Mathematic Benchmark For Large Language Models* (No. arXiv:2410.07985). arXiv.  
<https://doi.org/10.48550/arXiv.2410.07985>
- Harrik, A. (2025, märts 14). Tehisaruprofessor: Usaldusväärne keelemudel võiks oma kõhklusi tunnistada. *ERR*. <https://novaator.err.ee/1609633070/tehisaruprofessor-usaldusvaarne-keelemudel-voiks-oma-kohklusi-tunnistada>
- He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., Liu, J., Qi, L., Liu, Z., & Sun, M. (2024). *OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems* (No. arXiv:2402.14008). arXiv. <https://doi.org/10.48550/arXiv.2402.14008>
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). *Measuring Massive Multitask Language Understanding* (No. arXiv:2009.03300). arXiv. <https://doi.org/10.48550/arXiv.2009.03300>
- Ivanova, A. A. (2025). How to evaluate the cognitive abilities of LLMs. *Nature Human Behaviour*, 9(2), 230–233. <https://doi.org/10.1038/s41562-024-02096-z>
- Kurfah, M., Zahra, S., Gogoulou, E., Dürlich, L., Carlsson, F., & Nivre, J. (2025). SweSAT-1.0: The Swedish University Entrance Exam as a Benchmark for Large Language Models. *NEALT Proceedings Series*, 57, 331–339.
- Kuulmets, H.-A., Tättar, A., & Fishel, M. (2022). Estonian Language Understanding: A Case Study on the COPA Task. *Baltic Journal of Modern Computing*, 10(3).  
<https://doi.org/10.22364/bjmc.2022.10.3.19>

- Käver, A. (2021). *Extractive Question Answering for Estonian Language* [Master Thesis]. Tallinn University of Technology (TalTech).
- Lin, S., Hilton, J., & Evans, O. (2022). *TruthfulQA: Measuring How Models Mimic Human Falsehoods* (No. arXiv:2109.07958). arXiv. <https://doi.org/10.48550/arXiv.2109.07958>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Toim), *Advances in Neural Information Processing Systems* (Kd 35, lk 27730–27744). Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf)
- Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., ... Hendrycks, D. (2025). *Humanity's Last Exam* (No. arXiv:2501.14249). arXiv. <https://doi.org/10.48550/arXiv.2501.14249>
- Sügis, J., Tampuu, A., Aljanaki, A., Fišel, M., & Kull, M. (2025). *Praktiline andmeteadus*. Tartu Ülikooli arvutiteaduse instituut. <https://courses.cs.ut.ee/b/andmeteadus>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Toim), *Advances in Neural Information Processing Systems* (Kd 30). Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)

## Terminid

**Suur keelemudel** – (ingl *large language model, LLM*) keelemudel, mis on treenitud suurel hulgal tekstiandmetel, et luua loomuliku keele väljundeid ja sooritada erinevaid loomuliku keele töötlemise ülesandeid (nt tõlkimine, kokkuvõtete tegemine, meelsusanalüüs)<sup>20</sup>.

**Tärgtuvastus** – (ingl *optical character recognition, OCR*) - pildina salvestatud, trükitud või käsitsi kirjutatud teksti teisendamine masinloetavale kujule vastava tarkvara abil, täpsemalt optiline märgituvastus ehk optiline tekstituvastus<sup>25</sup>.

**Multimodaalsus** – mitme modaalsuse (tekst, audio, video jm) üheaegne kasutamine, st olukord, kus mudel on lisaks tekstisisendile võimeline töötlemise ka graafilisi andmeid (nt pilte, jooniseid, graafikuid jne).

**pdfplumber** – programmeerimiskeele Python teek tärgtuvastuseks ja andmete ekstraheerimiseks<sup>26</sup>.

**JSONL-fail** – JSON-tüüpi andmevahetusvormingu alaliik, mis paigutab ühe täieliku kirje ühele reale; sobib hästi struktureeritud andmete säilitamiseks ja edastamiseks<sup>27</sup>.

**Hugging Face** – tehisintellekti ja masinõppe mudelite baasil tarkvaralahenduste arendamist toetav platvorm<sup>28</sup>.

**lm-evaluation-harness** – Ettevõtte Eleuther AI tööriistade pakett keelemudelite hindamiseks erinevate ülesannete, sh küsimustele vastamise põhjal<sup>29</sup>. Ühildub Hugging Face keskkonnaga ning võimaldab kasutada sama andmestikku mitmete mudelite testimiseks.

**Zero-Shot Evaluation (ZSE)** – keelemudeli testimine ilma väikese näidisülesannete või prooviküsimuste hulgaga peenhäälestamata (Balloccu jt, 2024: 1). Suurte keelemudelite kontekstis tähendab see lisaks, et küsimuse esitamise viibas (ingl *prompt*) ei sisaldu testitavate ülesannete näiteid.

---

<sup>25</sup> Sõnaveeb. <https://www.sonaveeb.ee/> (15.03.2025)

<sup>26</sup> pdfplumber. <https://github.com/jsvine/pdfplumber> (15.03.2025)

<sup>27</sup> JSON Line. <https://jsonlines.org/> (15.03.2025)

<sup>28</sup> Hugging Face. <https://huggingface.co/> (15.03.2025)

<sup>29</sup> lm-evaluation-harness. <https://github.com/EleutherAI/lm-evaluation-harness> (15.03.2025)

## Litsents

Mina, Ahto Kiil,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Suurte keelemudelite võrdlev analüüs Eesti bioloogiaolümpiaadide küsimuste põhjal“, mille juhendajad on Taido Purason ja Hele-Andra Kuulmets reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;
2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commons'i litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;
3. olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;
4. kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Ahto Kiil

**15.05.2025**