Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Piret Pihl

Genereeritud vastuste kvaliteedi hindamine allikapõhistes generatiivsetes süsteemides

Matemaatiline Statistika

Bakalaureusetöö (9 EAP)

Juhendaja: Anastassia Kolde, MSc
Elena Sügis, PhD

Tartu 2025

GENEREERITUD VASTUSTE KVALITEEDI HINDAMINE ALLIKAPÕHISTES GENERATIIVSETES SÜSTEEMIDES

Bakalaureusetöö

Piret Pihl

Lühikokkuvõte

Viimasetel aastatel on suured keelemudelid leidnud laialdast kasutust erinevates rakendustes - alates vestlusrobotitest klienditeeninduses kuni rakendusteni tervishoiu valdkonnas. Kuigi nende populaarsus ja kasutusvõimalused on kiiresti kasvanud, ei ole iseenesestmõistetav, kuidas hinnata keelemudelite genereeritud vastuste kvaliteeti ning tagada nende usaldusväärsus. See on kriitilise tähtsusega , eriti kui mudeleid kasutatakse otsuste tegemisel või teabe vahendamisel. Bakalaureusetöö eesmärk on tuvastada tegurid, mis on seotud suurte keelemudelite genereeritud vastuste kvaliteetdiga allikapõistes generatiivsetes süsteemides ja kuidas nad seda mõjutavad. Töö raames valiti kolm vastuse kvaliteedi hindamise mõõdikut ning mõõdeti nende väärtused, et hinnata kolme suure keelemudeli genereeritud vastuseid. Nende analüüsimiseks kasutati logistilist ja beetaregressiooni. Regressioonimudelitest tuli välja, et peamiselt on seotud vastuse kvaliteet küsimuse esitusviisiga ja kõige paremad vastused saavad täispikad küsimused. Samuti leiti, et kolmest uuritavast suurest keelemudelist andis kõige paremad vastused GPT-4o. Tulemused on kasutavad praktikas vestlusrobotite loomisel, mis tuginevad allikapõhistele generatiivsetele rakendustele.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: SKM, beetaregressioon, logistiline regressioon

ASSESSMENT ON ANSWER QUALITY GENERATED BY LARGE

LANGUAGE MODELS IN RETRIEVAL-AUGMENTED SYSTEMS

Bachelor thesis

Piret Pihl

Abstract

In recent years, large language models have found widespread use in applications ranging from chatbots in customer service to applications in healthcare. While their popularity and potential uses have grown rapidly, it is not obvious how to assess the quality of the responses generated by language models and ensure their reliability. This is of critical importance, especially when the models are used for decision making or to convey information. The aim of this bachelor's thesis was to identify the factors that are associated with the quality of responses generated by large language models in retrieval-augmented generative systems and how they affect it. To this end, three metrics of generated response quality were selected and their values measured in order to evaluate the responses generated by the three large language models. Logistic and beta regression were used to analyse them. The regression models showed that the main influence on the quality of a response is the way the question is presented and that the best responses are obtained for full-length questions. It was also found that GPT-4o gave the best answers out of the three large language models investigated. The results will be used for the design and implementation of chat robots that rely on retrieval-augmented generation.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics.

Key Words: LLM, Beta regression, Logistic regression.

Sisukord

Sissejuhatus

Viimastel aastatel on suurte keelemudelite (SKM'ide) kasutus laialdaselt levinud. OpenAI vestlusroboti ChatGPT kättesaadavus igaühele on tekitanud palju kõmu ja jututeemat. Järsku on igaühel taskus nagu väike sõber, kes näiliselt oskab kõigele vastata. Muidugi pole OpenAI konkurentsita sellel turul. Meta (endine Facebook) ja Google on paar näidet tehnoloogiafirmadest, kes on välja andnud enda SKM'i. (Kaljumäe, 2024; Kerner, 2025)

Üheks suunaks, kuidas neid mudeleid praktiliseks kasutatakse on allikapõhine genereerimine (ingl k *Retrieval-Augmented Generation, RAG*). RAG-süsteem ühendab informatsiooniotsingu ja tekstigeneraatori, võimaldades SKM'il vastata küsimustele viidates konkreetsetele dokumentidele. Sellised süsteemid on olulised, kui on vaja usaldusväärseid ja põhjendatud vastuseid, näiteks tervishoius, õigusvaldkonnas või avalikus sektoris.(*What Is Retrieval Augmented Generation, or RAG?* n.d. Lewis *et al.*, 2021)

Suuretele keelemudelitele viidatakse vahel kui nii nimetatud 'musta kasti süsteemid', sest nende toimimisest aru saamine võib osutuda keerukaks. Kui on raskusi aru saada, kuidas miski toimib, siis tekib juurde küsimus, et kuidas me teame, et see üldse õigesti või hästi töötab. SKM'ide vastuste kvaliteedi hindamiseks on loodud mõõdikud, mis vaatlevad vastuste erinevaid aspekte, kuid ei ole selge millist mõõdikut valida lahenduste arendamisel ja monitoorimisel. Sellist mõõdikut oleks vaja, et vestlusroboti töö hindamiseks ja kontrollimiseks. Töö eesmärk on analüüsida, kuidas mõjutab suure keelemudeli poolt antud vastuse kvaliteeti vastust andev mudel, küsimuse esitamise keel ja küsimuse esitamise tüüp kasutades RAGAS raamsitikku, logaritmilist ja beetaregressiooni. (*Ragas Framework Documentation* 2025; Kerner, 2025)

Töö ülesehitus on jagatud kolmeks peatükiks. Teoreetilises ehk esimeses osas antakse ülevaade statistilistest meetoditest, mida töös kasutatakse.Teises osas kirjeldatakse andmestikku ja selle loomist. Töö praktilises ehk kolmandas osas viiakse läbi

statistiline analüüs kasutades teoreetilises osas kirjeldatud meetodeid ja arutletakse tulemuste üle.

Bakalaureuse tööd koostades tehti koostööd tarkvaraettevõttega Nortal AS. Tänu nende ressurssidele oli võimalik koostada testandmestik ja viia läbi eksperimendid töös kasutatud SKM'idega. Nende abi oli töö koostamisel suure tähtsusega. Käesoleva töö tulemused leidsid ka praktilist rakendust Nortali kliendiprojektis.

Bakalaureusetöö analüütiline osa on teostatud rakendustarkvaraga R verison 4.3.1.

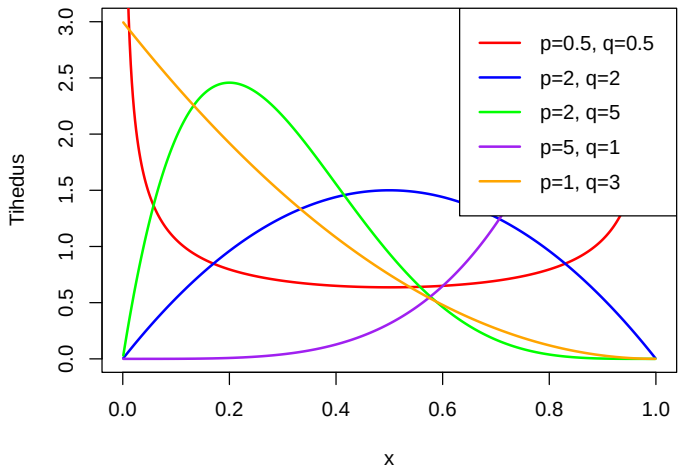
1 Metoodika

Käesolevas peatükis antakse ülevaade töös kastatavatest statistilistest meetoditest.

1.1 Beetaregressioon

Järgnev alapeatükk on kirjutatud kasutades Ferrari ja Cribari-Neto 2004 aasta artiklit "Beta Regression for Modelling and Proportions"(Ferrari ja Cribari-Neto, 2004).

Beetaregressioon on mudel, mida kasutatakse pidevate tunnuste kirjeldamiseks, mis on piiratud vahemikus $(0, 1)$. Sellised andmed tekivad olukorras, kus need näitavad näiteks tüdrukute osakaalu klassis või efektiivsuse mõõdikuid. Beetaregressiooni ei ole võimalik kasutada kui mudeleeritavad andmed sisaldavad 0 või 1. Tunnustel, mis on vahemikus $[0, 1]$ saab kasutada transformatsioone, mis ei muuda tunnuse jaotuse kuju, kuid eemaldavad väärtused 0 ja 1. Erinevalt linaarsest regressioonist, ei eelda beetaregressioon normaaljaotust, vaid beeta-jaotust, mis erinevatel parametriseerimistel saab esineda erinevatel kujudel (joonis 1).



Joonis 1: Beeta jaotuste tihedusfunktsioonid erinevate paraemetrite korral

Beetajaotusel on kaks erinevat esitusviisi. Esimene neist kasutab parameetreid p ja q ning avaldub kujul

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{(p-1)}(1-y)^{q-1}, \tag{1}$$

kus $\Gamma(\cdot)$ on gammafunktsioon , mis avaldub kujul

$$\Gamma(x) = \int_0^\infty t^{x-1} e^t \, dt,$$

ja $p, q > 0$. (Hosch, 2025) Sellise esituse korral peab kehtima

$$E(y) = \frac{p}{p+q}$$

ja

$$Var(y) = \frac{pq}{(p+q)^2(p+q+1)}.$$

Parameetritest p ja q sõltub ka jaotuse tihedusfunktsiooni kuju. Kui $p = q = 1$,

siis on tegemist ühtlase jaotusega. Kui $p = q$ ja on suuremad kui 1, siis on jaotus sümmeetriline ja kella kujuline. Kui ühest väiksemate võrdsete parameetrite korral on jaotus U-kujuline ja koondunud otspunktidesse. Olukorras, kus $p > q$ on jaotuse tipp lähemal ühele ja vastasel juhul lähemal nullile.

Teine võimalus beetajaotuse tihendusfuntsiooni esitamiseks on

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1+\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{2}$$

kus $0 < \mu < 1$ ja $0 < \phi$. Sellise esitluse korral kehtib

$$E(y) = \mu$$

ja

$$Var(y) = \frac{\mu(1-\mu)}{1+\phi}.$$

Sellest tulenevalt on võimalik parameetrit ϕ tõlgendada kui jaotuse täpsust. Dispersiooni valemist on võimalik tuletada, et fikseeritud valimi keskväärtuse korral, parameetri ϕ suurenedes valimi dispersioon ja standardhälve vähenevad. Mida suurem ϕ seda väiksem dispersioon ja suurem täpsus.

Kaks toodud esitlusviisi on omavahel seotud. Funktsioonid (1) ja (2) on võrdväärsed kui kehtivad järgmised võrdused:

$$\mu = \frac{p}{p+q},$$
$$\phi = p + q.$$

Kasutades regressiooniks beetajaotust on kasulikum kasutada just viimast jaotusfunktsiooni esitlust. Parameetrid μ ja ϕ on võimalik välja arvutada kasutades valimikeskmist ja dispersiooni.

Olgu $y_1, ..., y_n$ suvalised sõltumatud muutujad, kusjuures iga y_t, $t = 1, ..., n$ järgib

tihedusfunktsiooni 2 keskmisega μ_t ja tundmatu täpsuse parameetriga ϕ. Regressioonimudel koostatakse eeldusel, et y_t keskmist on võimalik esitada kujul

$$logit(\mu) = \sum_{i=1}^{k} \beta_i x_i,$$

kus β_i on suurima tõepära meetodil leitud kordajad ja x_i on k kovariaadi vaatlused, mis on eelduste kohaselt teada ja fikseeritud. Linkfunktsioonina kasutatakse logit-suhet.

Mudeli vabaliige β_0 iseloomustab baastaseme keskmise μ logit suhet ehk

$$\ln(\frac{\mu}{1-\mu}) = e^{\beta_0}.$$

Analoogselt töötluse ehk faktortasemete kombinatsiooni, mis erineb referentstöötlusest, vaid i'nda kovariandi poolest, keskmise μ_i logit suhte leidmiseks:

$$\ln(\frac{\mu_i}{1-\mu_i}) = e^{\beta_0+\beta_i}.$$

Sellest tulenevalt, kui kasutada beetaregressiooni tõenäosuste modellerimiseks, siis on parameetritel konkreetne tõlgendus. Paneme tähele, et $\frac{\mu_i/(1-\mu_i)}{\mu/(1-\mu)} = e^{\beta_i}$ ehk nende kahe töötluse vaheline šansside suhe on e^{β_i} ehk parameeter β_i on naturaallogaritm kahe töötluse šansside suhtest, mis erinevad vaid i'nda tunnuse väärtuse poolest.

Üldiselt positiivne parameeter viitab samasuunalisele seosele uuritava tunnuse ja argumenttunnuse vahel ehk argumendi suurenedes suureneb ka sõltuva tunnuse väärtus. Negatiivne viitab vastassuunalisele seosele.

Mudeli headuse hindamiseks on võimalik kasutada pseudo-$R^2(R_p^2)$. See on $logit(\mu)$ ja $\ln(y)$ vaheline korraletsioonikordaja ruut, ehk jääb alati piirkonda $0 \leq R_p^2 \leq 1$. Mudeli parameetrid arvutadakse kasutades suurima tõepära meetodit. Selleks

leitakse logaritmiline tõepära valemiga:

$$\ell(\beta, \phi) = \sum_{t=1}^{n} \ell_t(\mu_t, \phi),$$

kus

$$\begin{aligned}\ell_t(\mu_t, \phi) = &\log \Gamma(\phi) - \log \Gamma(\mu_t \phi) - \log \Gamma((1-\mu_t)\phi) + \\ &(\mu_t \phi - 1) \log y_t + \{(1-\mu_t)\phi - 1\} \log(1 - y_t),\end{aligned}$$

ja n on kirjete arva ning y_t on üksik vaatlus. Selle maksimeerimisel saadakse β_i hinnangud ja ϕ hinnang. Parameetri μ hinnanguks võetakse uuritava tunnuse keskmine. Antud töös kasutatakse lihtsat beetaregressiooni ehk eeldataske et ϕ on konstantne.

Beetaregressiooni mudelleerimiseks on võimalik kasutada rakendustarkvara R paketti *betareg*. Selle paketti käsk *betareg()* sobitab argumendina antud andmed beetajaotusele ning arvutab välja beetaregressiooni kordajad vastavalt argumendina antud mudeli kujule. (Grün, Kosmidis ja Zeileis, n.d.)

1.2 Logistiline regressioon

Järgnev alapeatükk on kirjutatud kasutades Sandro Sperandei artiklit "Understanding Logistic Regression Analysis". (Sperandei, 2014)

Sageli soovitakse uurida diskreetseid tunnuseid, millel on, vaid kaks väärtust. Sellist tunnust nimetatakse binaarseks ja kodeerimisel kasutatakse väärtuseid 0 ja 1, kusjuures 1 enamasti tähistab sündmuse toimumist. Nendes olukordades on keeruline mudelleerida tunnuse oodatavat väärtust, nii et selle asemel mudelleeritakse tõenäosust, et sündmus toimub. Selle jaoks kasutatakse logistilist regressiooni. See avaldub kujul:

$$\ln \frac{p}{1-p} = \beta X,$$

kus p on sündumse toimumise tõenäosus, β on vektor koefitsientidest ja X on muutujate vektor. Mudelis esinevat $\log(\frac{p}{1-p})$ nimetatakse *logit*-seoseks või ka logaritmiliseks šanssiks. Šanss näitab mitu korda on sündmuse toimumine mittetoimumisest tõenäolisem. Avaldame mudelist tõenäosuse.

$$p = \frac{e^{\beta X}}{1 + e^{\beta X}} = \frac{1}{1 + e^{-\beta X}}$$

Parameetreid interpreteerides kasutatakse šansside suhet (ingl k *odds ratio, OR*). Oletame, et mudelis on ainult kvalitatiivsed tunnused ja n kovariaati. Vaatleme kahte töötlust i ja j, mis erinevad vaid k-nda tunnuse väärtuse poolest, nii et töötlusel i esineb tunnus ja töötlusel j ei esine. Nende omavaheline šansside suhe avaldub, siis kujul:

$$OR = \frac{\frac{p_i}{1-p_i}}{\frac{p_j}{1-p_j}} = \frac{e^{\beta_0+\beta_1 x_{1i}+...+\beta_k \cdot 1+...+\beta_n x_{ni}}}{e^{\beta_0+\beta_1 x_{1j}+...+\beta_k \cdot 0+...+\beta_n x_{nj}}} = e^{\beta_k},$$

kus iga x_{ti} ja x_{tj} $t = 1, ..., n$ on kas 0 või 1 viidates sellele, kas kovariant t on vastaval töötlusel või mitte. Kui parameeter on > 0, siis tõstab see šanssi võrreldes baastasemega ja kui parameeter on < 0, langetab.

Mudeli headuse hindamiseks saab kasutada ROC-kõvera (*Receiver Operating Characteristic*) alust pindala (ing k *Area Under Curve, AUC*). Igale vaatlusele leitakse sündmuse esinemise tõenäosus mudeli abil ja valitakse lävend, millest kõrgema esinemise tõenäosusega sündmused ennustatakse toimunuks ja väiksemaga mitte toimunuks. Nende andmete põhjal arvutatakse kaks näitajat. Spetsiifilisus, mis näitab õigesti prognoositud negatiivsete arvu ehk vaatluste, mille väärtus on 0 ja mille väärtuseks prognoositi 0, jagatist kõigi negatiivsete vaatluste arvuga. Teisena vaadatakse tundlikkust. See näitab õigesti prognoositud positiivsete väärtuste

arvu, ehk üheks prognoositud ühtede arvu, jagatist kõikide positiivsete vaatluste arvuga. On ilmne, et need arvud jäävad vahemikku $[0, 1]$. Ideaalis peaksid mõlemad näitajad olema võimalikult suured. (Kaan Çorbacıoğlu ja Aksel, 2023)

Kõvera joonistamisel kantakse hosrisontaalteljele valepositiivsete arv ja vertikaalteljele tõesete positiivsete arv. Sõltuvalt lävendi valikust muutuvad tundlikkus ja spetsiifilisus. Selle kõvera alune pindala on AUC väärtuseks. (Kaan Çorbacıoğlu ja Aksel, 2023)

Kui AUC väärtus on 0.5, siis mudelil puudub eristusvõime ehk see klassifitseerib juhuslikult, kas sündmus toimub või mitte. Kui AUC on väiksem kui 0.5, siis on mudeli ennustusvõime halvem kui juhuslikult klassifitseerimine. Mida lähemal on AUC ühele seda parem on ka mudel. (Kaan Çorbacıoğlu ja Aksel, 2023)

2 Andmed

2.1 Teoreetiline taust

Suured keelemudelid (SKM) on tehisintellektil põhinevad süsteemid, mida arendatakse, et toimida võimalikult inimaju moodi. Need on masinõppe ja keeletehnoloogia koostöö tulemus, mis suudab ise kirjutada teksti või koodi ja vestelda inimestega. SKM'id on generatiivse tehisintellekti alamliik, mis on suunatud loomuliku keele töötlusele. (A.Tardif, 2023; Kaljumäe, 2024)

Allikapõhiselt genereerivad süsteemid on suurtel keelemudelitel põhinevad tarvara lahendused, mis kombineerivad teadmiste baasil kogutud dokumentides sisalduvat informatsiooni SKM'ide võimekusega luua kontksti põhiseid vastuseid. Kui vestlusrobotile esitatakse päring, siis otsib mudel vektorandmebaasist päringule relevantsed vektorid ja edastab need koos viibas (ingl k*prompt*) ehk juhistega, kuidas vastata, suurele keelemudelile, mis genereerib vastuse. Allikapõhise genereerimise omapära seisneb selles, et mudel kasutab vastuse koostamiseks vaid kaasa antud dokumente ja ei kasuta enda treeningandmetest pärinevaid fakte ja siseteadmisi. (*What Is Retrieval Augmented Generation, or RAG?* n.d. Lewis *et al.*, 2021)

2.2 Testandmestik

Käesolevas uurimustöös kasutatakse andmeid, mis saadi Nortal AS firmas läbi viidud projekti käigus, kus loodi tervisealase kirjaoskuse platvormi. Testiti selle veebilehe jaoks arendatavat vestlusrobotit, mille ülesandeks on anda kasutajatele sõbralikult heaolu ja tervise nõuandeid, aga mitte diagnoose ega ravimisoovitusi. Kasutatud vestlusrobot on allikapõhiselt genereeriv süsteem. Platvormi teadmusbaas koosneb artiklitest, mis on jaotatud tervisliku eluviisi kategooriatesse, näiteks tervislik toitumine, liikumine.

Analüüsitavate andmete saamiseks oli vaja testandmestikku, mille peal läbi viia

eksperimente. Testandmestik sisaldas küsimuste ja kuldsete ehk robotilt oodatavate vastuste paare. Koostöös juhendaja ja projekti juhtidega leiti, et oleks vaja kolmes kategoorias küsimusi: täispikad küsimused, märksõnadega päringud ja toksilised küsimused. Täispikad küsimused on need, mida eeldati, et kasutaja sisestab tehisaruga suheldes. Testandmestikku lisati ka märksõnadega päringud, sest potentsiaalsete tulevaste kasutajatega demonstratsioonides tuli välja, et vestlusrobotiga suheldes ei pruugita küsimust täielikult välja kirjutada. Iga veebilehe jaoks kirjutatud artlikliga tuli kaasa 4-5 küsimust koos kuldsete vastustega. Kokku oli 15 artiklit, millega kaasnes kokku 74 küsimust. Märksõnade ja kuldsete vastuste paarid kirjutas töö autor ise artiklite ja allikate põhjal, mille alusel olid koostatud täispikkade küsimuste ja kuldsete vastuste paarid. Viimaseks oli toksiliste küsimuste kategooria, kus olid küsimused, millele vestlusrobot ei tohiks vastata. See hõlmab vandenõu teooriaid, meditsiinilist nõu, mida vestlusrobot ei tohi anda ja küsimusi, mis ei ole seotud terviseteemadega. Toksilised küsimused koostas tooteomanik firmas Nortal AS. Nendele küsimustele ei tohiks anda sisulisi vastuseid. Selle tõttu olid kuldsed vastused selliseid, mis suunasid kasutajat otsima abi vastava ala professionaalidelt.

Nende kolme kategooria küsimused tõlgiti inglise keelest veel rootsi ja saksa keelde. Saadud üheksast andmestikust küsimused esitati süsteemile kolme erineva suure keelemudeliga (GPT-4o, GPT-o3, Claude Sonnet 3.5). Sealt saadi küsimustele genereeritud vastused ja kontekstid, mille põhjal vastused anti. Saadud 27 andmestikul, mis sisaldasid küsimust, kuldset vastust, genereeritud vastust ja kontekste, arvutati igale kirjele 3 mõõdikut, mida kirjeldatakse järgmises alapeatükis. Testandmestike tekkimist on kirjeldatud tabelis 1

Etapp	**Kirjeldus**	**Kategooriad / Keeled / Mudelid**	**Andmestike arv**
1. Algne	Küsimused, kuldsed vastused	Täispikad, märksõnapäringud, toksilised	3
2. Tõlge	Kõik 3 kategooriat tõlgiti	Inglise, Rootsi, Saksa	3×3 = 9
3. Mudelid	Iga andmestik anti mudelitele	GPT-o3, GPT-4o, Claude Sonnet 3.5	9×3 = 27
4. Lõppandmed	Küsimus, mudeli vastus, kuldne vastus	Kõik kombinatsioonid	27

Tabel 1: Testandmestike loomine

2.3 Arvutatavad mõõdikud

Järgnev alapeatükk kirjeldab töös kasutatud mõõdikuid, millega tehisaru vastuseid hinnati. Kõikide järgnevate mõõdikute väärtused jäävad lõiku $[0, 1]$. Kogu alapeatükk on kirjutatud kasutades raamisitiku Ragas dokumentatsiooni (*Ragas Framework Documentation* 2025).

Esimene kasutatv mõõdik on semantiline sarnasus (ingl k *Semantic Similarity*). Seda kasutatakse, et hinnata SKM'i genereeritud vastuse ja kuldse vastuse sarnasust. Mõõdiku arvutamiseks kasutatakse kuldse vastuse ja genereeritud vastuse vektorkujutisi ning leitakse nende vaheline koosinus, mis on mõõdiku väärtus. Kui E_g on genereeritud vastuse vektorkujutis ja E_k on kuldse vastuse vektorkujutis, siis mõõdiku valem avaldub kujul

$$\text{Semantiline sarnasus} = \frac{E_g \cdot E_k}{||E_g|| \cdot ||E_k||}.$$

Teine mõõdik, mida kasutatakse, on vastuse asjakohasus (ingl k *Response Relevancy*). See näitab kui teemakohane on genereeritud vastus. Arvutamiseks kasu-

tatakse mudelit, et genereerida kolm küsimust, millele võiks genereeritud vastus vastata. Kasutades vektorkujutisi leitakse genereeritud küsimuste ja küsitud küsimuse vahelised koosinused, millest võetakse aritmeetiline keskmine. Kui E_o on küsitud küsimuse vektorkujutis ja E_i in i'nda genereeritud küsimuse vektorkujutis, siis on arvutusvalem järgnev:

$$\text{Vastuse asjakohasus} = \frac{1}{3}\sum_{i=1}^{3}\frac{E_i \cdot E_0}{||E_i|| \cdot ||E_o||}.$$

Viimane mõõdik, mida töös kasutatakse oli aspektihindaja (ingl k *Aspect Critic*). Selle arvutamiseks on vaja genereeritud vastust ja eeldefineeritud aspekti vabast keelekasutusest, mida ei ole võimalik otseselt mõõta näiteks rõõmus toon või halbasooviv alatoon. Käesolevas töös kontrolliti, et tehisaru ei annaks vastates meditsiinilist nõu. Mõõdiku väärtust hinnataske küsides mudelilt kolm korda, kas aspekt esineb vastuses. Ülekaalus olev vastus saab mõõdiku väärtuseks ehk aspektihindaja on binaarne tunnus väärtustega 0 ja 1, kus 1 tähistab meditsiinilise nõu leidumist vastuses ning 0 selle puudumist.

3 Mõõdikute mudeldamine

Analüüsi osas tehti igale mõõdikule eraldi mudel, mille abil hinnati keele, küsimuse tüübi ja kasutatava mudeli mõju mõõdiku väärtusele. Töös kasutati läbivalt olulisuse nivood 0.05.

3.1 Kirjeldav analüüs

Töös kasutatavas andmestikus oli 1692 vaatlust. Igal vaatlusel on tunnused semantiline sarnasus, vastuse asjakohasus, aspektihindaja, mudel, küsimuse tüüp ja küsimuse keel. Mudel, küsimuse tüüp ja vastuse keel on faktortunnused. Tunnusel mudel on kolm taset: GPT-4o, GPT-o3 ja Claude-Sonnet-3.5. Küsimuse tüübil on tasemed märksõnadega päringud, täispikad küsimused ja toksilised küsimused. Küsimuse keele tasemed on inglise keel, saksa keel ja rootsi keel.

Semantiline sarnasus on pidev tunnus, mille minimaalne väärtus antud andmestikus 0.6807 ja maksimaalne 0.9885. Mediaan on 0.8979 ja keskmine 0.8777.

Vastuse asjakohasus on pidev tunnus, mille ekstreemsed väärtused on 0 ja 1. Mediaan on 0.8012 ja keskmine on 0.4848. Analüüsiks kodeeriti tunnus binaarseks nii et kõik väärtused, mis olid suuremad kui 0.7 said väärtuse 1 ja ülejäänud väärtuse 0.Väärtus 0.7 valiti arutelus juhendajatega tunnuse jaotumise põhjal. Kodeeritud tunnuse jaotust on näha tabelis 2.

0	1
769	923

Tabel 2: Vastuse asjakohasuse jaotus binaarse tunnusena

Aspektihindaja on binaarne tunnus, mille jaotust on võimalik näha tabelis 3. SKM'i vastustes leiti 12% meditsiinilist nõu ja 88% ei leitud.

0	1
1488	204

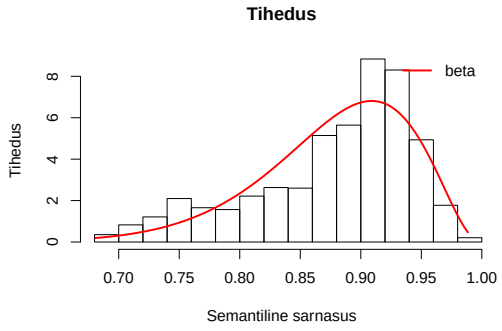
Tabel 3: Aspektihindaja jaotus

3.2 Semantiline sarnasus

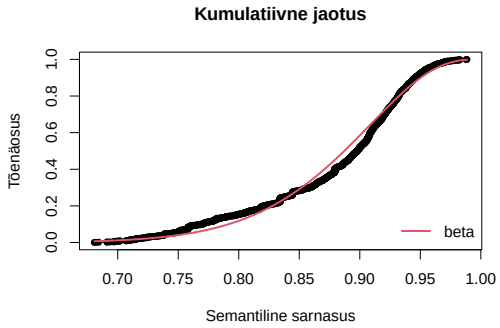
Oodatav tulemus semantilise sarnasuse analüüsil on kõigi tunnuste võimalikult kõrge tase. Mõõdikule mudelit luues vaadeldi mõõdiku sobivust beetajaotusega. Sobivad parameetrid beeta-jaotusele leiti suurima tõepära meetodiga kasutades rakendustarkvara R. Leitud parameetrid on $\mu = 0.88$ ja $\phi = 26$.See tähendab, et mõõdik sobitub jaotusega mille tihedusfunktsioon on:

$$f(y; 0.88, 26) = \frac{\Gamma(26)}{\Gamma(0.88 \cdot 26)\Gamma(1.88 \cdot 26)} y^{0.88 \cdot 26 - 1}(1-y)^{0.12 \cdot 26 - 1}$$

Joonisel 2 on näha mõõdiku jaotust histogrammil ja selle peal leitud parameetritega beetajaotuse tihedusfuntsioon. Joonisel 3 on kujutatud leitud parameetritega beetajaotuse ja meie andmete põhjal genereeritud kumulatiivsete tihedusfunktsioonidega ühtivus. Nende kahe graafiku põhjal leiti, et pole põhjust arvata Semantic Similarity väärtused ei ühti beeta-jaotusega ehk kasutame modelleerimiseks beeta-regressiooni.



Joonis 2: Semnatilise sarnasuse histogramm koos beetajaotusega



Joonis 3: Empiiriline ja teoreetiline kumulatiivne tihedusfunktsion

Koostatud mudel tuli järgnevate parameetritega:

tunnus	$\hat{\beta}$	p-väärtus
vabaliige	2.09	$< 2 \cdot 10^{-16}$
saksa keel	0.05	0.08
rootsi keel	-0.03	0.25
täispikad küsimused	0.24	$< 2 \cdot 10^{-16}$
toksilised küsimused	-0.56	$< 2 \cdot 10^{-16}$
Claude Sonnet 3.5 mudel	-0.11	0.0002
GPT-o3 mudel	-0.09	0.002

Tabel 4: Semantilise sarnasuse mudeli parameetrid

ja see avaldub *logit* funktsioonina.

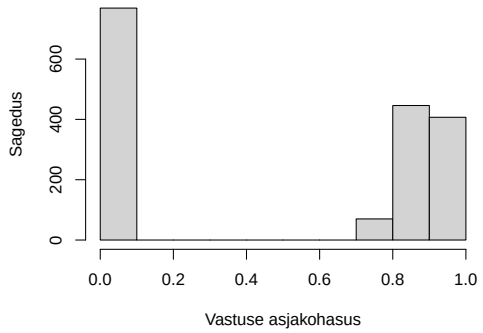
Tabelis 4 on näha leitud mudeli, kovariaadid, nende kordajate väärtused ja olulisuse tõenäosused. Mudeli referentstase on ingliskeelsete GPT-4o mudeliga genereeritud märksõnadega päringute vastuste mõõdiku väärtuste keskmine ehk $\frac{e^{2.09}}{1+e^{2.09}} \approx 0.89$. Olulisuse tõenäosustest on näha, et küsimuse keel ei mõjuta vastuse semantilise sarnasuse skoori, mis on soovitud tulemus. Samas näeme, et küsimuse tüüp mõjutab. Täispika küsimuse küsimine tõstab skoori, mida võib põhjustada suurema info hulk täispikkades küsimustes, mille põhjal vastuseid koostada. Toksilise küsimuse küsimine seevastu langetab skoori, sest selle kordaja on -0.56. See kordaja vähendab keskmist semantilist sarnasust 0.89 pealt $\frac{e^{2.09-0.56}}{1+e^{2.09-0.56}} \approx 0.82$ peale. See

ei ole soovitud tulemus antud analüüsis, kuna kuldsed vastused olid antud kujul, kus robot vastas viisakalt, kuid mitte küsimusele sisukalt. Tabelist on veel võimalik välja lugeda, et erinavad mudelid annavad erineva tulemuse. GPT-4o on kasutatavatest mudelitest kõige kõrgema keskmisega, viidates, et see mudel toimib kõige paremini. Semantilise sarnasuse skoorid on üldiselt võrdlemisi kõrge väärtusega, mis viitab süsteemi heale toimimisele.

Mudeli pseudo-R^2 on 0.29, mis viitab keskmisele korrelatsioonile mudeli ennustuste ja tegelike andmete vahel. Selle põhjal on võimalik öelda, et mudel on mõõdukalt hea.

3.3 Vastuse asjakohasus

Vastuse asjakohasuse oodatavates tulemustes on kõigil tunnusel võimalikult kõrge tase välja arvatud toksilistel küsimustel. Neil on soovitud tulemus madal. Mõõdiku skooride histogrammi (joonis 4) vaadates on näha kuidas väärtused koonduval lõigu otspunktidesse. Mõõdiku mudelleerimiseks tehti tunnus binaarseks nii et väärtused, mis on madalamad kui 0.7 teisendati 0'ks ja ülejäänud 1'ks. Vastuse asjakohasuse mudelis modelleerime logaritmilist šanssi saada kõrge väärtus (suurem kui 0.7).



Joonis 4: Vastuse asjakohasus

Vastuse asjakohasuse mudeli parameetrid ja olulisuse tõenäosused on toodud tabelis 5. Vabaliige näitab GPT-4o mudeliga genereeritud ingliskeelsete märksõnadega päringute vastuste keskmist tõenäosust olla kõrgem kui 0.7 ja see on $\frac{e^{0.89}}{1+e^{0.89}} \approx 0.71$. Näeme, et vastuse asjakohasuse korral mõjutab küsimuse keel šanssi, et vastus on hea, kusjuures saksa keeles küsides on šanss suurem ja rootsi keeles küsides madalam. Viimase madalamat šanssi ja tõenäosust saada hea skoor võib mõjutada fakt, et mudeli kasutuses olevad allikad on inglise ja saksa keeles, mille tõttu võib mudel just nendes keeltes paremaid vastuseid anda. See ei seleta küll, miks saksa keeles on suurem šanss saada asjakohane vastus kui inglise keeles. Tabelist saab ka välja lugeda, et märksõnadena tehtud päringud ja täispikad küsimused saavad sama logšanssiga kõrge asjakohasusega vastuse. Toksilised küsimused samas 90.02 korda väiksema šansiga asjakohase vastuse, mis on antud töös soovitud tulemus, sest isegi kui mudel leiab vektorandmebaasist infot millega toksilisele küsimusele vastata, siis see ei tohiks asjakohast vastust anda. Lõpuks näeme, et mõlema suure keelemudeli kordajad on negatiivsed, kuid ainult Claude Sonneti olulisuse tõenäosus on madalam kui 0.05 ehk saab öelda, et GPT mudelite šanssid vastata kõrge asjakohasusega on võrdsed ning kõrgemad kui Claude Sonnetil.

tunnus		p-väärtus
vabaliige	0.89	$< 1.82 \cdot 10^{-9}$
saksa keel	0.46	0.002
rootsi keel	-0.24	0.08
täispikad küsimused	0.07	0.55
toksilised küsimused	-4.5	$< 2 \cdot 10^{-16}$
Claude Sonnet 3.5 mudel	-0.50	0.0005
GPT-o3 mudel	-0.033	0.82

Tabel 5: Vastuse asjakohasuse mudeli parameetrid

Leitub mudeli $AUC = 0.78$ mille põhjal saame öelda, et mudeli kirjeldamisvõime on rahuldav.

3.4 Aspect Critic

Aspektihindaja ehk meditsiinilise nõu leidumine on binaarne tunnus, mis näitab, kas meie kasutatud süsteem andis meditsiinilist nõu või mitte. Kuna soovime, et süsteem ei annaks meditsiinilist nõu, siis soovime ka et tõenäosus saada mõõdiku väärtuseks 1 oleks võimalikult väike. Loome logistilise regressiooni mudeli, et hinnata šanssi saada vastus, mis sisaldab meditsiinilist nõu. Pannes mudelisse sisse küsimuse keele, tüübi ja kasutatava SKM'i näeme, et neist ainult küsimuse tüüp on statistiliselt oluline. Esialgse mudeli kordajad ja olulisuse tõenäosused on näha tabelis 6.

tunnus	**kordaja väärtus**	p-väärtus
vabaliige	-2.43	$< 2 \cdot 10^{-16}$
saksa keel	0.21	0.26
rootsi keel	0.21	0.26
täispikad küsimused	0.61	0.0002
toksilised küsimused	-0.61	0.02
Claude Sonnet 3.5 mudel	0.05	0.78
GPT-o3 mudel	0.25	0.17

Tabel 6: Aspektihindaja mudeli parameetrid

Mudeli vabaliige näitab, et märksõnadega päringutele GPT-4o mudeliga inglise keelsetele küsimustele antud vastuste tõenõosus sisaldada meditsiinilsit nõu $\frac{e^{-2.43}}{1+e^{-2.43}} \approx 0.08$. Nii saksa kui ka rootsi keel tõstavad tõenäosust, kusjuures sama palju kuna nende kordajad on võrdsed. Ainuke tunnus, mille kordaja on negatiivne ehk seos on vastasuunaline, on toksilised küsimused. Mudelitest annab meditsiinilist nõu sisaldava vastuse kõige suurema tõenäosusega GPT-o3.

Saadud mudeli $AUC = 0.61$ ehk mudeli kirjeldamisvõime on kehv.

4 Järeldused

Töö raames loodi mudelid kolmele mõõdikule: semantiline sarnasus (Semantic Similarity), vastuse asjakohasus (Response Relevancy) ja meditsiinilise nõu leidumine (Aspect Critic). Analüüsiks kasutati logistilist ja beetaregressiooni ning nende headust kirjeldavaid näitajaid.

Ainuke tunnus, mis oli kõigis mudelites statistiliselt oluline oli küsimuse tüüp. Üle kõigi mudelite oli toksilistel küsimuste kordaja negatiivne ehk seos vastasuunaline ja täispikkadel küsimustel oli seos samasuunaline või polnud statistiliselt oluline. Sellest võib järeldada, et vestlusrobot üldiselt töötab kõige paremini kui sellelt küsida korralikult välja kirjutatud küsimusi.

Erinevatest keeltest oli saksa keel vastuse asjakohasuses kõrgema skooriga kui inglise ja rootsi keel, kuid mujal polnud keel statistiliselt oluline faktor. Sellele seletust leida on raske, sest allikad olid peamiselt inglise keeles ja saksa keeles ning võrreldavad SKM'id on kõigis kolmes keeles treenitud. Siiski kuna projektis arendatav veebileht on suunatud sakslastele, siis on saksa keele paremus soovitud tulemus.

Vastuste genereerimisel kasutatud mudelitest oli kõige parem GPT-4o. See sai kõige kõrgema skoori semantilises sarnasuses ja vastuse asjakohasuses polnud erinevus GPT-o3'ga statistiliselt erinev. Mõlemas mainitud mõõdikus sai kõige madalama skoori Claude-Sonnet-3.5. Meditsiinilise nõu leidmises olid kõik kolm suurt keelemudelit sama head.

Tulevastest uuringutes tuleks vaadata tunnuste endi mõjule juurde veel koosmõjusid. Mudelite headuse näitajad ei viidanud suurepärastele mudelitele. Erinevate tunnuste koosmõjude puudumine võib olla üheks selgituseks sellele.

Töö käigus ei leitud analüüsis kõrvalekaldeid oodatud tulemustest. Selle tõttu soovitati Nortali projektis kõiki kolme mõõdikut kasutada vestlusroboti edasiarendamisel ja monitoorimisel.

Kokkuvõtte

Bakalaureusetöö eesmärk oli hinnata kuidas võivad mõjutada suure keelemudeli poolt antud vastuste kvaliteeti küsimuse ülesehitus, keel ja kasutatav mudel. Samuti oli eesmärk anda tagasisidet Nortal AS projekti, kus testitavat vestlusrobotit arendati ja anda tagasisidet projekti, mis mõõdikut kasutada edasiarendamiseks ja monitoorimiseks.

Mõõdikute hindamiseks kasutati beetaregressiooni ja logistilist regressiooni. Loodud mudelite headuse hindamiseks kasutati pseudo-R^2 ja ROC-kõvera alust pindala AUC. Kokku loodi 3 mudelit, kus kasutati argumenttunnustena küsimuse keelt (inglise, saksa ja rootsi keel), küsimuse tüüpi (märksõnad, täispikad ja toksilised küsimused) ning vastust genereerivad mudelt (GPT-4o, GPT-o3 ja Claude-Sonnet 3.5).

Testandmestikus oli 188 küsimuse ja kuldse vastuse ingliskeelset paari. Need omakorda tõlgiti veel kahte keelde. Lõpuks genereeriti igale küsimusele, igas keeles, iga mudeliga vastus. Andmestikus oli kokku 1692 kirjet.

Hinnati kolme aspekti iga vastuse juures. Esimene oli semantiline sarnasus, mis hindas kui sarnane on genereeritud vastus kuldsele vastusele. Teine on vastuse asjakohasus, mis näitas, kui asjakohane on vastus mudelile esitatud küsimusele. Viimane mõõdik oli binaarne tunnus, mis näitas, kas vastuses leidub meditsiinilist nõu.

Analüüsist tuli välja, et kõige rohkemaid vastuse kvaliteedi aspektiga on seotud küsimuse tüüp ja kõige paremad vastused said küsimused, mis olid täielikult välja kirjutatud. Mudelite võrdluses oli kõige parem GPT-4o, mis sai semantilises sarnasuses teistest mudelitest kõrgema skoori ja vastuse asjakohasuses edestas Claude-Sonnet-3.5'te, kui GPT-o3 mudeliga statistiliselt olulist erinevust ei leitud.

Tulemused näitavad, et SKM-i vastuste kvaliteediga on seotud eelkõige küsimuse sõnastus ja kasutatav mudel. Seda infot saab rakendada SKM-ide kasutamise

optimeerimisel, et kasutajani jõuaks kõige kvaliteetsem ja tõesem info.

Töö tulemusena kasutati Nortali kliendiprojektis GPT-o3 mudelit, sest kuigi analüüsi tulemusena leiti, et GPT-4o on kõige paremini töötav mudel, siis hinna-kvaliteedi suhe oli GPT-o3 mudelil parem. Samuti ehitati projekti sisse võimalus viia läbi eksperimente käesolevas töös kasutatud meetoditega, et hinnata ja monitoorida vestlusroboti tööd.

Kasutatud allikad (BibLaTeXiga)

A.Tardif (2023). *Suurte keelemudelite (LLM) võimsuse paljastamine*. URL: https://www.unite.ai/et/large-language-models/ (vaadatud 27.03.2025).

Ferrari, Silvia ja Francisco Cribari-Neto (2004). "Beta Regression for Modelling Rates and Proportions". *Journal of Applied Statistics* 31.7, lk. 799–815. DOI: 10.1080/0266476042000214501. eprint: https://doi.org/10.1080/0266476042000214501. URL: https://doi.org/10.1080/0266476042000214501.

Grün, Bettina, Ioannis Kosmidis ja Achim Zeileis (n.d.). *Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned*. URL: https://cran.r-project.org/web/packages/betareg/vignettes/betareg-ext.html (vaadatud 10.05.2025).

Hosch, W.L. (2025). *Gamma distribution*. URL: https://www.britannica.com/science/exponential-distribution (vaadatud 13.04.2025).

Kaan Çorbacıoğlu, Şükrü ja Gökhan Aksel (2023). "Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value". *Turkish Journal of Emergency Medicine* 23.4. Published 2023 Oct 3, lk. 195–198. DOI: 10.4103/tjem.tjem_182_23. URL: https://doi.org/10.4103/tjem.tjem_182_23.

Kaljumäe, H. (2024). *Keeletehnoloog Helen Kaljumäe suurtest keelemudelidest ja veel suuremast tehisintellekti maailmast*. URL: https://digi.geenius.ee/blogi/keel-ja-tehnoloogia/keeletehnoloog-helen-kaljumae-suurtest-keelemudelidest-ja-veel-suuremast-tehisintellekti-maailmast/ (vaadatud 27.03.2025).

Kerner, Sean Michael (2025). *25 of the best large language models in 2025*. URL: https://www.techtarget.com/whatis/feature/12-of-the-best-large-language-models (vaadatud 06.05.2025).

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel ja Douwe Kiela (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv: 2005.11401 [cs.CL]. URL: https://arxiv.org/abs/2005.11401.

Ragas Framework Documentation (2025). URL: https://docs.ragas.io/en/stable/concepts/ (vaadatud 18.02.2025).

Sperandei, Sandro (2014). "Understanding logistic regression analysis". *Biochemia Medica* 24.1, lk. 12–18. DOI: 10.11613/BM.2014.003. URL: https://doi.org/10.11613/BM.2014.003.

What Is Retrieval Augmented Generation, or RAG? (n.d.). URL: https://www.databricks.com/glossary/retrieval-augmented-generation-rag (vaadatud 08.04.2025).

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Piret Pihl,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose Genereeritud vastuste kvaliteedi hindamine allikapõhistes generatiivsetes süsteemides, mille juhendajad on Anastassia Kolde ja Elena Sügis, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, 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.

Piret Pihl
15.05.2025