TARTU ÜLIKOOL
Sporditeaduste ja füsioteraapia instituut

Siim Orgvee

AI-VESTLUSROBOTITE KVALITEEDI HINDAMINE VASTAMISEL PATSIENTIDE ENIM LEVINUD KÜSIMUSTELE LATERAALSE EPIKONDÜLIIDI KOHTA

Assessing the performance of AI chatbots in answering patients' common questions about lateral epicondylitis

Magistritöö

Füsioteraapia õppekava

Juhendaja:
Teadusliku ja praktilise valukäsitluse lektor, Martin Argus PhD

Tartu 2026

SISUKORD

KASUTATUD LÜHENDID

AI – *artificial intelligence*

FRES - *Flesch Reading Ease Score*

LLM – *large language models* ehk suured keelemudelid

MRT – magnetresonantstomograafia

Töö lühiülevaade

AI-vestlusrobotite kvaliteedi hindamine vastamisel patsientide enim levinud küsimustele lateraalse epikondüliidi kohta

Eesmärk: Käesoleva magistritöö eesmärk oli hinnata ja võrrelda kolme suure keelemudeli (Gemini mudel 3, ChatGPT 5.3 ja Claude Sonnet 4.6) kvaliteeti patsientide enim levinud lateraalset epikondüliiti käsitlevatele küsimustele vastamisel. Hindamisel lähtuti vastuste täpsusest, loetavusest ja lahtiütluste esinemisest.

Metoodika: Magistritöös kasutati lateraalse epikondüliidi teemalise küsimustiku koostamiseks Delphi konsensusmeetodit. Küsimused esitati kolmele suurele keelemudelile (Gemini mudel 3, ChatGPT 5.3 ja Claude Sonnet 4.6), mille vastuseid hinnati täpsuse, loetavuse ja lahtiütluste esinemise alusel. Vastuste täpsust võrreldi kolme rahvusvahelise ravijuhendiga (JOSPT, BESS ja Sports Medicine). Vastused klassifitseeriti nõrkadeks, mõõdukateks ja tugevateks ravisoovitusteks või ebakorrektseteks vastusteks. Loetavust hinnati Flesch Reading Ease Score'i (FRES) abil ning lahtiütluste olemasolu binaarsel jah/ei-skaalal.

Tulemused: Vastuste täpsuse hindamisel osutus ChatGPT 5.3 kõige täpsemaks mudeliks: ebakorrektseid vastuseid esines vaid 6% juhtudest ning tugevaid ravisoovitusi anti 20% vastustes. Claude andis ebakorrektseid vastuseid 8% juhtudest, Gemini mudel 3 aga 26% juhtudest. Loetavuse hindamisel sai ChatGPT 5.3 kõrgeima keskmise FRES tulemuse (66), mis vastab 8.-9. klassi lugemisoskusele. Gemini (56,5) ja Claude (58,4) vastused eeldasid vähemalt gümnaasiumitasemel lugemisoskust. Lahtiütluste osas esitas Claude 8 lahtiütlust 35 vastusest, Gemini 2 lahtiütlust, samas kui ChatGPT 5.3 ei esitanud ühtegi lahtiütlust.

Kokkuvõte: Uuringu tulemused näitasid, et suurte keelemudelite vastuste täpsus erines mudelite lõikes märkimisväärselt. ChatGPT 5.3 oli käesolevas magistritöös hinnatud mudelitest täpseim ja paremini loetav lateraalse epikondüliidi teemalise teabe edastamisel. Siiski ei suutnud ükski hinnatud mudel anda kõigis olukordades täielikult korrektseid vastuseid. ChatGPT 5.3 vastustes lahtiütluste täielik puudumine võib kliinilises kontekstis olla probleemne. Tulemused viitavad, et LLM-id võivad toimida terviseinfo täiendava allikana ja toetada tervishoiuspetsialistide tööd, kuid praegu ei ole need võimelised spetsialisti asendama.

Märksõnad: lateraalne epikondüliit, suured keelemudelid, tehisintellekt, füsioteraapia

Abstract

Assessing the performance of AI chatbots in answering patients' common questions about lateral epicondylitis

Aim: The aim of this master's thesis was to evaluate and compare the quality of three large language models (Gemini model 3, ChatGPT 5.3, and Claude Sonnet 4.6) in answering patients' most common questions regarding lateral epicondylitis. The evaluation focused on the accuracy, readability, and presence of disclaimers in the generated responses.

Methods: A questionnaire on lateral epicondylitis was developed using the Delphi consensus method. The questions were submitted to three large language models (Gemini model 3, ChatGPT 5.3, and Claude Sonnet 4.6), and the responses were evaluated based on accuracy, readability, and the presence of disclaimers. Response accuracy was assessed by comparing the generated answers with three international clinical guidelines (JOSPT, BESS, and Sports Medicine). Responses were classified as weak, moderate, or strong treatment recommendations, or as incorrect responses. Readability was assessed using the Flesch Reading Ease Score (FRES), and the presence of disclaimers was evaluated using a binary yes/no scale. The study was designed as a descriptive comparative study.

Results: ChatGPT 5.3 demonstrated the highest level of accuracy, with incorrect responses identified in only 6% of cases and strong treatment recommendations present in 20% of responses. Claude generated incorrect responses in 8% of cases, whereas Gemini 3 did so in 26% of cases. In terms of readability, ChatGPT 5.3 achieved the highest mean FRES score (66), corresponding to an 8th–9th grade reading level. The responses generated by Gemini (56.5) and Claude (58.4) required at least a high school reading level. Regarding disclaimers, Claude included disclaimers in 8 out of 35 responses, Gemini in 2 responses, while ChatGPT 5.3 did not include any disclaimers.

Conclusion: The findings of this study demonstrated that the accuracy of large language model responses varied considerably between models. Among the evaluated models, ChatGPT 5.3 provided the most accurate and most readable information regarding lateral epicondylitis. Nevertheless, none of the evaluated models consistently produced fully accurate responses in all situations. The complete absence of disclaimers in ChatGPT 5.3 responses may be problematic in a clinical context. The findings suggest that large language models may serve as a supplementary source of health information and support healthcare professionals, but they are currently unable to replace clinical specialists.

Keywords: lateral epicondylitis, large language models, artificial intelligence, physiotherapy

1. KIRJANDUSE ÜLEVAADE

1.1. Lateraalse epikondüliidi patogenees ja ravi

Lateraalset epikondüliiti kirjeldas Runge esmakordselt meditsiinikirjanduses 1873. aastal (viidatud Ma & Wang, 2019 järgi). Lateraalne epikondüliit on levinud küünarliigese piirkonna kaebus, mida kirjeldatakse kui valu ja hellust *m. extensor carpi radialis brevis* kõõluse piirkonnas. Mõnikord on haaratud ka m. *extensor digitorum*, m. *extensor digiti minimi*, m. *extensor carpi ulnaris.* Diagnoosimisel tuginetakse anamneesile ja füüsilisele läbivaatusele. Vajadusel kasutatakse täiendavate uuringutena ultraheli ning magnetresonantstomograafiat (MRT). MRT uuringut kasutatakse peamiselt mõne teise probleemi tuvastamiseks või preoperatiivselt. (Keijsers *et al.,* 2019) Lateraalne epikondüliit mõjutab 1-3% elanikkonnast, enamasti keskealisi inimesi. Ligikaudu 80% juhtudest laheneb probleem konservatiivse ravita ühe aasta jooksul. Valu enamasti ägeneb vastupanuga randmeliigese sirutamisel ning pigistamisel. (Ma & Wang, 2019) 2024. aasta süstemaatilises kirjanduse ülevaates uurisid Chen jt riskifaktoreid lateraalse epikondüliidi tekkeks. Tulemustena leiti erinevaid riskifaktoreid, milleks olid naissugu, suitsetamine ja hüperkolesteroleemia. (Chen *et al*., 2024)

Tendinopaatia patogenees on multifaktoriaalne. Kõige sagedamini seostatakse tendinopaatia teket korduvate mikrotraumadega, mis ületavad kõõluse koormustaluvuse piiri. Tendinopaatia tekkemehhanismide selgitamiseks on välja pakutud mitmeid teooriaid. Enamik neist järgib järjestust: a) kollageenikiudude kahjustumine; b) põletik; c) kõõluse rakkude vastus. (Challoumas et al., 2020) Tendinopaatiat peeti alguses põletikuliseks protsessiks. Histopatoloogilised uuringud on näidanud põletiku puudumist krooniliste tendinopaatiate puhul. Sellest tulenevalt käsitletakse tendinopaatiat degeneratiivse protsessina. (Ma & Wang, 2019) Tendinopaatia tekke algpõhjusena tuuakse välja vigastus või korduvad mikrotraumad ja ebasoodsad mehaanilised tingimused. Normaalse paranemisprotsessi asemel tekib mittetäielik paranemine, mis väljendub patoloogilistes muutustes kõõluse kollageenimaatriksis, tsütokiinide profiilis, vaskulaarsuses, innervatsioonis, rakutiheduses ning raku fenotüübis. Neid muutusi on seostatud valu tekkega. Sarnaselt sellele on ka Cook ja teised välja töötanud *continuum*-mudeli, mis küll ei käsitle põletiku rolli. Selle kohaselt läbib kõõlus kolm faasi:

a) reaktiivne tendinopaatia. Esimeses faasis on mittepõletikuline vastus rakus ja maatriksis, mis tuleneb ülekoormusest ja mis toob endaga kaasa kõõluse paksenemise.

b) mittetäielik paranemine. Teises faasis on maatriksi ning kollageeni ehitus häirunud, rakkude arv suurenenud ja proteoglükaanide tootmine suurenenud. Lisaks sellele võib esineda angiogeneesi ja närvikiudude sissekasv.

c) degeneratiivne tendinopaatia. Viimases faasis tekib apoptoosi tõttu rakkude surm ja suureneb maatriksi korrapäratus. (Cook *et al.,* 2016)

Kasutusel on erinevad ravivõimalused, millel on erinev edukuse määr. Tendinopaatia esmase ravimeetodina kasutatakse treeningut. Kõige sagedamini kasutatakse ning soovitatakse ekstsentrilist jõutreeningut. Enamik patsientidest saab sellest valu leevendust. (Irby *et al.,* 2020) Kõõluse patoloogia võib olla tingitud kõõluserakkude alastimulatsioonist, mis tekib koormuse ülekande puudumise tõttu kahjustunud kollageenikiududes. Kõõluserakkude vähene mehaaniline stimulatsioon võib mängida rolli degeneratiivsete muutuste kujunemisel. Kõõluserakkude mehaaniline stimulatsioon treeningu kaudu võib selgitada treeningpõhiste sekkumiste efektiivsust. (Cook *et al.,* 2016) Lisaks kasutatakse ravis erinevaid täiendavaid meetodeid nagu näiteks lööklaineteraapia, *dry needling*, ultraheliravi jt. Nende kasutamise kvaliteet ning efektiivsus vajavad täiendavat uurimist. Inaktiivsus ning kõõlusele puhkuse andmine ei ole näidanud probleemi lahendamisel efektiivsust. Seega kasutatakse esmajärgus tendinopaatia ravis treeningut, kuna see on näidanud valu vähenemist, funktsiooni paranemist ning on suhteliselt lihtsasti patsientide jaoks teostatav. (Vlist *et al.,* 2019)

1.2. Tervisekirjaoskus ja selle seos patsiendi raviga

Tervisealane kirjaoskus on tugevas seoses patsiendi ravi edukusega (Wolf *et al.,* 2005). Tervisealane kehv kirjaoskus on seotud suurema arvu hospitaliseerimiste, suurema hulga erakorralise meditsiini teenuste kasutamisega ning vähenenud ennetavate ravikäitumiste (nt mammograafiad, vaktsineerimine) hulgas (Berkman *et al.,* 2011). Ühe võimaliku seletusena on välja toodud, et madalama kirjaoskusega inimestel tekib suhtlusbarjäär tervishoiuspetsialistidega, mis mõjutab negatiivselt nende ravi kulgu (Seurer & Vogt, 2013). Kui patsiendid ei suuda meditsiinilist teavet piisavalt mõista, võib see nende ravi tulemusi negatiivselt mõjutada (Institute of Medicine, 2004). Wirth jt uurisid 2025. aastal patsientide materjalide loetavust. Senine arusaam on olnud, et patsientidele väljastatav info peab olema kuuenda klassi tasemel. Uuringu autorid tõid välja, et sellest teadmisest hoolimata on patsiendimaterjalid inimeste jaoks liialt keerulised ning raskesti mõistetavad. (Wirth *et al.,* 2025) Rooney ja teiste 2021. aasta uuring hindas loetavust patsientide harivatel materjalidel tähtsates meditsiini ajakirjades viimase 20 aasta jooksul. Samuti leiti tulemusena, et materjalid on märkimisväärselt keerulisemad kuuenda klassi lugemise tasemest**.** (Rooney *et al.,* 2021)

1.3. Suured keelemudelid terviseinfo edastamisel ning nende riskid

Suured keelemudelid (LLM) on loodud genereerima vastuseid, ennustades kõige tõenäolisemat järgmist sõna või sõnade jada kasutaja sisendi põhjal. Mudelite treenimisel kasutatavate andmete allikad, kvaliteet ning spetsiifilised parameetrid ei ole täielikult avalikustatud. (Saenger *et al.,* 2024).

Selline teadmatus LLMide toimimise osas tõstatab probleeme nende usaldusväärsuse ja kvaliteedi osas tervishoiu valdkonnas (Khowaja *et al.,* 2024; Goodman *et al.,* 2023). Ebakorrektne info, mis on loodud nende mudelite poolt, võib vähendada tervishoiusüsteemi kvaliteeti (Kisseka & Giboney, 2018). Hoolimata korduvast kasutajate tagasisidel põhinevast tugevdamisõppest (Gallifant *et al.,* 2024) ja olulistest algoritmilistest täiustustest, mille eesmärk on tegeleda tehisintellekti (AI) hallutsinatsioonidega (Goddard 2023), on nii ChatGPT tasuta kui ka Plus versioonid endiselt vastuvõtlikud valeinfo levitamisele (Abd-Alrazaq *et al.,* 2023 ; Lee 2024) ja teadmiste fabritseerimisele (Zielinski *et al.* 2023). LLM-e treenimisel on välja toodud ka nö mudeli kokkuvarisemise oht. LLM-ide teksti analüüsimisel tekib teatud informatsiooni kadu, mille tulemusel võib iga mudeli järgmise põlvkonnaga muutuda genereeritud tekst järjest piiratumaks. (Shumailov *et al.,* 2023) Lisaks on puudujääkidena toodud välja, et saadud soovitused ei ole patsiendi põhised ning seetõttu on need ka vähem efektiivsed. Suured keelemudelid võimaldavad tervisealaseid soovitusi saada kiiremini, kuid mitte tingimata kvaliteetsemalt. ChatGPT piiranguid on veel teisigi. OpenAI andmetel võivad ChatGPT vastused sisaldada ebatäpsusi või kallutatust. Varasemates uuringutes on välja toodud, et suured keelemudelid võivad genereerida viiteid olematutele teadusuuringutele. Samuti on näidatud, et nad võivad järgida juhiseid, mis soodustavad ebapädevate või potentsiaalselt kahjulike ravisoovituste pakkumist. Lisaks võib teatud juhtudel esineda ebaõige informatsiooni kordamist ja võimendumist mudeli vastustes. (The Lancet Digital Health, 2023)

Bai ja teiste 2023. aasta uuring räägib, millised on hetkel LLM-ide arendamise põhimõtted ning miks on oluline *disclaimer*ite ehk lahtiütluste olemasolu. Hetkel on eesmärk arendada AI-d sellisel kujul, et see oleks võimalikult abivalmis ning samaaegselt võimalikult turvaline. Lisaks on muutunud olulisemaks, et AI-d suudaksid enda otsuseid põhjendada ehk oleksid läbipaistvamad. Lahtiütluste olemasolu on seetõttu oluline mitme aspekti tõttu. Lahtiütluste olemasolu aitab vähendada kahju, mida võib tekitada LLM-i poolt antud mittekorrektne ravijuhend. Läbipaistvuse suurendamiseks aitab lahtiütlus kaasa, kuna näitab, et LLM ei ole ekspert ning vastus võib olla piiratud või ebatäpne. See omakorda aitab ka vähendada üleusaldust. Lisaks toimivad lahtiütlused osalise kompensatsioonina selle eest, et mudel ei vastuta reaalse maailma tagajärgede eest ning inimspetsialisti kontroll puudub. Lahtiütluste esitamisel on LLM samaaegselt võimalikult kasulik ning võimalikult ohutu, mis on uuringu autorite poolt välja toodud peamised aspektid hetkel LLM-ide arendamisel. (Bai *et al.,* 2022)

1.4. Suurte keelemudelite kasutamine meditsiinis

Raja ja teiste 2024. aasta uuring analüüsis internetikasutust meditsiinilise probleemi korral. Leiti, et 76,9% inimestest on vähemalt korra elu jooksul võtnud ravimeid internetist saadud

informatsiooni põhjal ning 75% inimestest on endale ise diagnoosi määranud. Uuringus osalejatest 16,3% hindas sellist käitumist turvaliseks. (Raja *et al.,* 2024) Bujnowska-Fedak ja Węgierek leidsid, et 45% inimestest pöördusid otsingutulemustest mõjutatuna seejärel arsti poole ning umbes 40% inimestest tekkis veel küsimusi nende kaebusega seoses (Bujnowska-Fedak, & Węgierek, 2020).

Igapäevaselt kasutab ChatGPT-d 122 miljonit inimest (Singh, 2025). On läbi viidud erinevaid uuringuid tehisintellekti kasutamise kohta meditsiinis. Iga kümnes austraallane kasutab ChatGPT-d meditsiinialastele küsimustele vastuse leidmiseks (Zhang, 2025). ChatGPT pälvis esmakordselt tähelepanu meditsiinis, kui suutis sooritada arstilitsentsi eksami. Kui võrreldi suurte keelemudelite (LLM) antud soovitusi arsti antud soovitusega, leiti, et LLM-ide antud soovitused olid täpsemad ning ka empaatilisemad. Uuringu tulemused tekitasid arutelu tehisintellekti võimaliku rolli üle tervishoiusüsteemis. Ühe võimaliku arengusuunana arutati spetsialistide asendamise üle. Uuringu autorid leidsid, et spetsialiste tehisintellekt täna asendada ei suuda, kuna LLM-ide kasutamine ravis ei ole 100% täpne. (Thirunavukarasu et al., 2023)

Reisi jt 2024. aasta uuring hindas AI tehnoloogia abil antud soovitusi nende empaatilisuse ja usaldusväärsuse osas. Tulemustena leiti, et inimeste poolt antud soovitusi hinnati empaatilisemaks ja ka usaldusväärsemaks. Autorid leidsid uuringu tulemusena, et inimeste jaoks on spetsialisti poole pöördumine midagi enamat kui lihtsalt vajaliku informatsiooni saamine. Lisaks leiti, et edukaks raviks on vaja patsiendiga head koostööd ning seda võimaldab rohkem inimese poolt antud soovitused. (Reis *et al*., 2024) Yun'i ja Bickmore'i 2024. aasta uuringus leiti, et 21,2% inimestest kasutab LLM-e või tehisintellektil põhinevaid rakendusi enda terviseprobleemidega seoses. Neist omakorda 90,2% nendest kasutasid LLM-e tervisealase teabe hankimiseks ning 8,8% ravimite või ravivõimaluste kohta informatsiooni saamiseks. Saadud informatsiooni kontrollis üle 19,4% inimestest, aga saadud soovitusi järgisid 48,4% patsientidest. (Yun & Bickmore, 2025) Takita ja teiste 2025. aasta uuring leidis, et AI poolt antud vastused olid korrektsed 52,7% juhtudest. Töö autorid tõid tulemusena välja ka asjaolu, et LLM-id hetkel ei suuda asendada eksperdi nõuandeid, kuna hetkel ei suuda LLM-id arvestada inimkeha komplekssust ja spetsialisti kliinilist kogemust. (Takita *et al*., 2025) Cabrali ja teiste 2025. aasta uuring ennustas, et AI hakkab järgmise 10 aasta jooksul kõige enam mõjutama diagnostikat, mitte üle võtma kogu patsiendi ravi. Uuringus leiti, et AI näitas suurt täpsust röntgenpiltide analüüsimisel, südame rütmihäirete tuvastamisel ning naha ebanormaalsuste tuvastamisel. See uuring tõi ka välja, et AI kaasamiseks meditsiinis on enne vaja eelnevalt lahendada eetilised ja juriidilised probleemid. (Cabrali *et al*., 2025) Sallami 2023. aasta uuringus toodi LLM-ide kasutamise probleemidena välja erinevad riskid. 55% kasutatud allikatest viitas eetilistele probleemidele, mille all mõisteti plagiaati,

turvalisuse ja privaatsuse võimalike ohte. Autor tõi LLM-ide kasutamise ühe ohuna välja, et meditsiinivaldkonnas on palju väärinformatsiooni, mida suured keelemudelid võivad patsientidele tõese informatsioonina esitada ning võivad sellega kujutada ohtu patsientide tervisele. (Sallam, 2023) Walkeri ja teiste 2023. aasta uuring hindas ChatGPT vastuseid võrreldes kliiniliste ravijuhenditega erinevate sisehaiguste kohta. Uuring leidis, et ChatGPT vastused kattusid ravijuhendiga 60% ulatuses. Töö autorid tõid välja, et LLM-id on hetkel piiratud kvaliteediga, kuid tulevikus võivad hakata mõjutama patsientide informatsiooni kogumist, aga mitte asendama spetsialiste. (Walker *et al.,* 2023)

1.5. Varasemad uuringud suurte keelemudelite täpsuse ja loetavuse kohta

Ülevaatlikus meta-analüüsis, milles hinnati ChatGPT-3.5 sooritust meditsiiniliste küsimuste kontekstis, leiti, et üldine täpsus oli vaid 56% (Wei *et al.,* 2024). Samuti näitas läbilõikeuuring, kus võrreldi ChatGPT-3.5 ja ChatGPT-4.0 vastuseid 284 arsti koostatud meditsiinilisele päringule, et vaid 50% vastustest olid täpsed (Goodman *et al.,* 2023). Üha kasvav hulk uuringuid toob järjest rohkem esile tõsiseid muresid mudelite andmete täpsuse ja usaldusväärsuse osas, viidates nende püsivalt kehvale sooritusele tervishoiuhariduse ja kliiniliste otsuste tegemise kontekstis. (Goodman *et al.,* 2023; Yeo *et al.,* 2023) Yau ja teiste 2024. aasta uuringus vaadati LLM-ide täpsust erakorralise meditsiini osas. Töö tulemustena leiti, et vastuste täpsus oli 50% ning ohtlikku või valet informatsiooni edastasid 5-35% juhtudest. (Yau *et al.,* 2024) Wei ja teiste 2024. aasta meta-analüüsis oli ChatGPT-3.5 üldine täpsus meditsiinilistele küsimustele vastamisel vaid 56% (Wei *et al.,* 2024). Goodmani ja teiste 2023. aasta uuring leidis, et vaid 50% ChatGPT vastustest olid täpsed (Goodman *et al.,* 2023). Daheri ja teiste 2023. aasta uuring vaatas, kas ChatGPT suudab asendada õla- ning käespetsialiste diagnostikas ja ravis. Töö tulemusena hinnati, et 55% ravist oli kehva kvaliteediga. Töö autorid järeldasid, et ChatGPT ei suuda veel asendada õla- ning käespetsialiste diagnostikas ning ravis. (Daher *et al.,* 2023) Çıracıoğlu & Erdoğani 2025. aasta uuring ChatGPT kasutamise kohta skolioosi ravis leidis, et ChatGPT suudab anda asjakohast informatsiooni, kuid jääb kõige enam hätta spetsiifilise ning personaalse ravi planeerimisega (Çıracıoğlu & Erdoğan, 2025). Hanci ja teiste 2024. aasta uuring hindas viie erineva LLM-i kvaliteeti, täpsust ning loetavust palliatiivraviga seotud küsimustele vastamisel. FRES tulemus ChatGPT-l oli 22,10 ning Geminil 24,00. Vastused olid seega väga keerulised ning kõik viis uuringus olnud suurtest keelemudelitest jäid kaugele soovitatud kuuenda klassi lugemistasemest. Töö autorid tõid LLM-e vastuste täpsuse hindamisel välja, et hindamise muudab keerulisemaks teadmatus, et millisest allikast LLM oma informatsiooni kogus, kas see on usaldusväärne ning kaasaegne. (Hanci *et al.,* 2024)

Varasemalt on teostatud sarnane uuring mittespetsiifilise alaseljavalu kohta. Mittespetsiifilise alaseljavalu uuringu tulemustena leiti, et 55.8% kõikidest saadud vastustest olid korrektsed, 42.1%

mittekorrektsed ja 1.9% ebaselged. Kõige täpsemaid tulemusi andis Gemini (60%), seejärel ChatGPT-3.5 (59.2%) ja ChatGPT-4.0 (56%). Bing andis kõige enam ebakorrektseid vastuseid (56.3%). Loetavuse osas oli lõpptulemuseks, et vastused on suhteliselt raskesti loetavad. FRES skoori tulemuseks oli 50,94. Loetavus varieerus vastavalt kasutatud mudelile. Geminil oli kõrgeim skoor (63,39), ChatGPT-3.5 sai tulemuseks 39,93 ja ChatGPT-4.0 tulemuseks oli 46,12. Tervisega seotud hoiatusi esitas ChatGPT 4.0 kõikides vastustes, ChatGPT 3,5 96.6% vastustest, Gemini 83,33% vastustest ja Bing 70% vastustest. Töö tulemusena järeldas autor, et LLM-e kasutamine mittespetsiifilise alaseljavalu puhul patsientide nõustamisel näitab lubavaid, kuid muutlike tulemusi. LLM andsid mõõdukalt täpseid soovitusi alaseljavalu puhul. Probleemina tõi töö autor välja üsna kehva loetavuse, mida oleks vaja parandada, et patsiente paremini nõustada. (Scaff *et al.,* 2025)

Senised uuringud viitavad, et suured keelemudelid võivad pakkuda meditsiinilistes küsimustes mõõduka täpsusega vastuseid, kuid nende kvaliteet varieerub sõltuvalt kasutatud mudelist ja kliinilisest kontekstist. Magistritöö autori teadmiste kohaselt ei ole lateraalset epikondüliiti käsitlevate eestikeelsete patsientide küsimuste kontekstis suurte keelemudelite vastuste kvaliteeti seni hinnatud. Magistritöö autori teadmiste kohaselt ei ole hinnatud LLM-e kvaliteeti, loetavust ning lahtiütluste esinemise sagedust lateraalse epikondüliidi puhul.

2.TÖÖ EESMÄRK JA ÜLESANDED

Töö eesmärk on hinnata ja võrrelda LLMide antud vastuste täpsust, mõistetavust ja tervisega seotud hoiatusi seoses enam levinud patsientide küsimustega lateraalse epikondüliidi kohta.
Vastavalt töö eesmärkidele püstitati järgmised uurimisülesanded:

1. Hinnata suurte keelemudelite antud vastuste täpsust võrreldes valitud ravijuhistega
2. Hinnata suurte keelemudelite antud vastuste loetavust *Flesch Reading Ease Score*'i alusel
3. Analüüsida suurte keelemudelite vastustes antud lahtiütlusi.

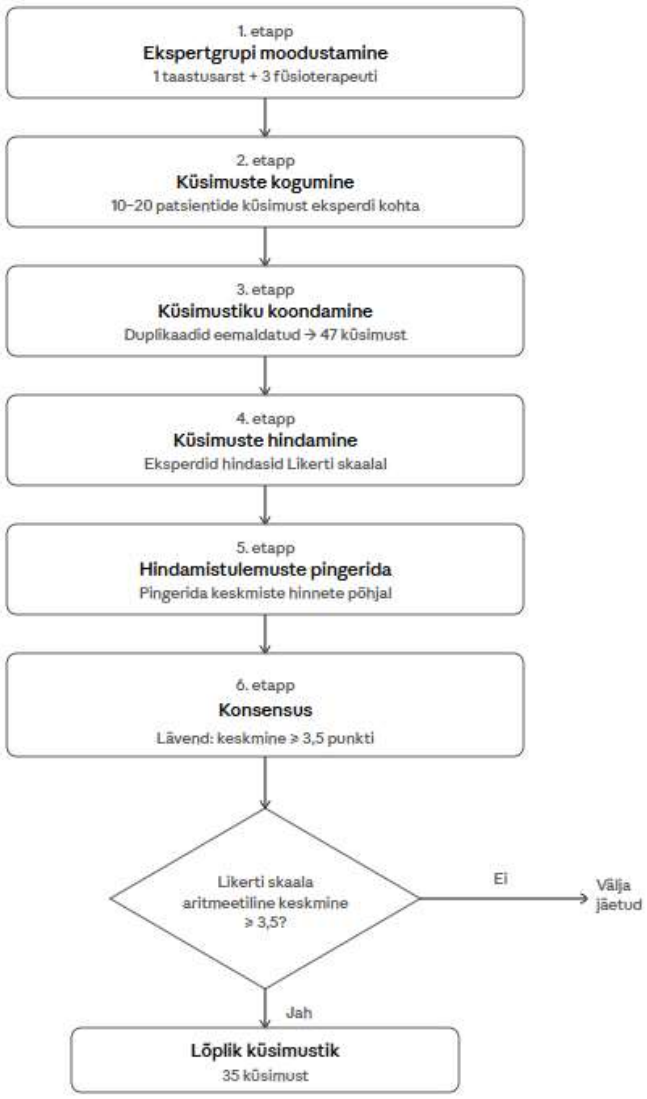
3. METOODIKA

3.1. Küsimustiku koostamine

Küsimused sisestati ükshaaval magistritöö autori poolt. Gemini mudel 3 küsimused sisestati 17.02.2026. Claude'i küsimused sisestati 14.03.2026. ChatGPT küsimused sisestati 29.03.2026. Kõik küsimused sisestati eesti keeles. Iga küsimuse sisestamiseks kasutati uut vestlusahelat, et vältida vestluskontekstist tulenevat varieerivust. Iga vestlusahel algas lausega "Mul diagnoositi lateraalne epikondüliit ning mul on selle kohta küsimusi". Kõik küsimused sisestati vaid ühe korra ja vastus kaasati tulemuste analüüsi. Küsimused sisestati erinevatel kuupäevadel, kuna andmekogumine toimus etapiviisiliselt. Kõik küsimused sisestati erinevatesse suurtesse keelemudelitesse samas järjekorras.

Küsimustiku koostamiseks kasutati Delphi konsensusmeetodit. See meetod kujutab endas ekspertgrupiga mitmeetapilist tööd. (Drumm *et al.,* 2022 ; Holey *et al.,* 2007 ; Nasa *et al.,* 2021) Kõigepealt moodustati ekspertgrupp. Ekspertgrupp koosnes 4 inimesest, kellest 1 oli taastusraviarst ning 3 füsioterapeudid. Kõik neli töötasid Tartu Ülikooli Kliinikumi taastusravi ja spordimeditsiini osakonnas. Ekspertgruppi valiti need spetsialistid, kellel on ulatuslik kliinilise kogemus lateraalse epikondüliidiga patsientide ravis. Delphi konsensusmeetodi esimeses etapis pani iga spetsialist kirja 10-20 küsimust, mida epikondüliidi diagnoosi saanud patsient neilt küsinud on. Siin rõhutati ekspertgrupi liikmetele, et kirja tuleb panna just need küsimused, mida patsiendid olid kliinilises praktikas tegelikult esitanud. Konsensusmeetodi teises etapis koondas magistritöö autor küsimused kokku ühtseks küsimustikuks. Sarnase sõnastusega küsimusi ei duubeldatud. Küsimustiku järele pandi Likerti skaala: "Kuivõrd tõenäoliseks pead, et epikondüliidi diagnoosi saanud patsient küsib tervishoiuspetsialistilt sellise küsimuse?". Kasutati Likerti skaalat väärtustega 1-5. Skaalal tähistas väärtus 1 -"Väga ebatõenäoline", 2 -"Ebatõenäoline", 3 – "Pisut tõenäoline"; 4 – "Tõenäoline"ning 5 - "Väga tõenäoline". Konsensusmeetodi kolmandas etapis saadeti küsimustik ekspertgrupile ning iga ekspert vastas kasutades iga küsimuse juures olevat Likerti skaalat. Uurija koondas tulemused kokku ning koostas küsimustest aritmeetiliste keskmiste põhjal pingerea. Delphi meetodi neljandas etapis toimus ekspertgrupi koosolek. Sellel määrati Likerti skaala tulemuse nivoo, millest madalama tulemuse saanud küsimusi uuringus edasi ei kaasatud. Ekspertgrupiga viidi läbi arutelu, et kinnitada ühehäälselt lõplik küsimuste nimekiri, mida magistritöös kasutati. Esialgses nimekirjas oli kokku 47 küsimust. Nivooks loeti nelja hindaja vastuste aritmeetiline keskmine 3,5 ehk uuringusse kaasati küsimused, mis Likerti skaala aritmeetilise keskmise tulemusena said skooriks 3,5 või üle selle. Esialgsest nimekirjast arvati

välja 12 küsimust. Uuring viidi läbi 35 küsimusega. Delphi meetodi etapid küsimustiku välja töötamiseks on esitatud joonisel 1.



Joonis 1. Delphi konsensusmeetod küsimustiku välja töötamiseks

3.2. LLM-ide vastuste hindamine nende täpsuse, mõistetavuse ja lahtiütluste alusel

Magistritöös kasutati järgmisi suuri keelemudeleid: ChatGPT 5.3, Gemini mudel 3 ja Claude versioon Sonnet 4.6. Mudeleid valiti nende populaarsuse alusel *Google Play*st kuupäeval 15.02.2026. Täpsuse hindamiseks võrreldi LLM-i antud vastust kolme erineva lateraalse epikondüliidi ravijuhendiga. Küsimustikuga välja töötatud küsimused sisestati mudelitele ükshaaval. Vastuste klassifitseerimise viis läbi magistritöö autor. Ravijuhenditena olid kasutusel JOSPT-i 2022. aasta ravijuhend(Lucado et al., 2022), Sports Medicine Rehabilitation-i 2021. aasta ravijuhend(Santiago et al., 2021) ning BESS-i 2023. aasta ravijuhend (Singh & Watts, 2023). Ravijuhendid valiti

tõenduspõhisuse, konservatiivse ravi käsitluse, füsioteraapia valdkonna kliinilise relevantsuse ning täieliku ravijuhendi kättesaadavuse alusel. Saadud vastused klassifitseeriti järgnevalt:

- Korrektne nõrk ravisoovitus: LLM-i poolt antud soovitus oli ühes ravijuhendis.
- Korrektne mõõdukas ravisoovitus: LLM-i poolt antud soovitus oli kahes ravijuhendis.
- Korrektne tugev ravisoovitus: LLM-i poolt antud soovitus oli kolmes ravijuhendis.
- Ebakorrektne ravisoovitus: LLM-i poolt toodud soovitus on vähemalt ühes ravijuhendis vastunäidustatud.

Loetavuse all mõisteti teksti mõistetavuse keerukuse taset. Selle hindamiseks kasutati *Flesch Reading Ease Score*'i (FRES)(Jindal *et al.,* 2017). FRES skoor arvutatakse lausete keskmise pikkuse ja sõnade keskmise silpide arvu põhjal. Kõrgeima skooriga tekstidel on 100 sõna kohta keskmiselt 123 silpi ja keskmine lause pikkus on 8 sõna. Kõige väiksema skooriga tekstide silpide arv 100 sõna kohta keskmiselt on 192 ning lause pikkus 29 sõna. Tulemuste arvutamiseks kasutati valemit:

206.835 - (1.015 x ASL) - (84.6 x ASW)

ASL – keskmine lause pikkus (sõnad/lausetega)

ASW – keskmine silpide arv sõnas (silbid/sõnad)

FRES skaala oli 0-100ni. Tulemused on klassifitseeritud järgmiselt:

- 100-90: väga lihtne lugeda (viienda klassi tase)
- 90-80: lihtne lugeda
- 80-70: üsna lihtne
- 70-60: standardne (8.-9. klassi tase)
- 60-50: üsna keeruline
- 50-30: keeruline
- 30-0: väga keeruline (ülikoolitasemele vastav lugemistase). (Spadaro *et al.,* 1980)

3.3. Lahtiütluste esinemise sageduse uurimine

Lahtiütluse puhul analüüsiti vastuseid, mis hoiatasid patsienti olukorraga seotud riskidest, viitasid mudeli piirangutele või soovitasid konsulteerida tervishoiuspetsialistiga. LLM-i vastused klassifitseeriti lahtiütluste olemasolu alusel binaarsel jah/ei-skaalal, vastavalt kas vastus sisaldas lahtiütlust või mitte. Täiendavalt analüüsiti lahtiütluse sisu. Saadud vastuseid analüüsis magistritöö autor.

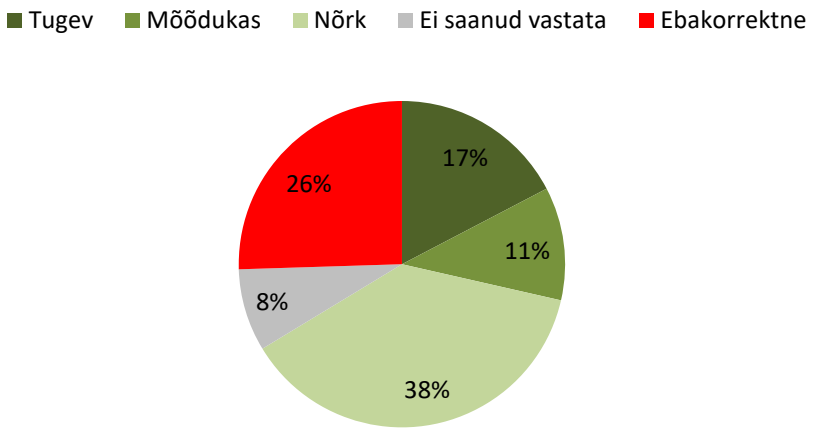
4. TÖÖ TULEMUSED

4.1. Suurte keelemudelite poolt antud vastuste täpsus võrreldes valitud ravijuhenditega.

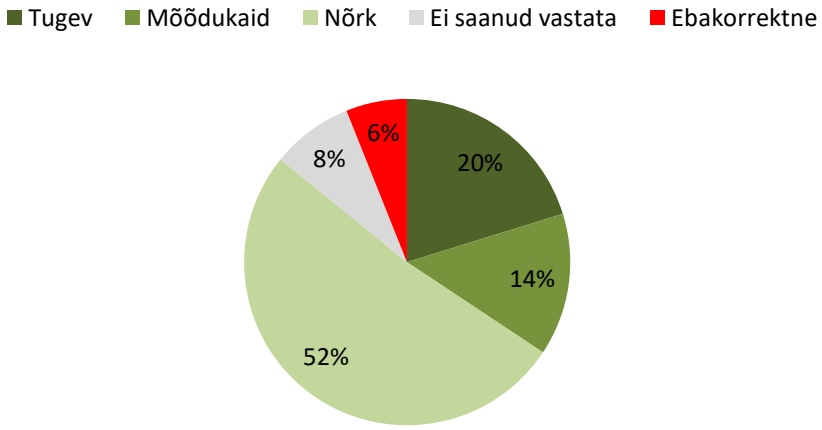
LLM-ide vastuste täpsuse hindamisel võrreldi vastuseid valitud ravijuhenditega. Iga mudelit analüüsiti eraldi. Täpsuse hindamise tulemused olid järgmised:

1) Gemini mudel 3 vastustes esines tugevaid ravisoovitusi 6 korral (17%), mõõdukaid 4 korral (11%), nõrku 13 korral (37%). Kasutatud ravijuhiste põhjal ei saanud hinnata vastuse kvaliteedi kolme vastuse puhul. Ebakorrektseid vastuseid oli 9 (26%). (joonis 2)
2) ChatGPT mudel 5.3 vastustes esines tugevaid ravisoovitusi 7 korral (20%), mõõdukaid 5 korral (14%), nõrku 18 korral (51%). Kolme küsimuse puhul ei olnud vastuste kvaliteeti võimalik ravijuhendite alusel hinnata (8%). Ebakorrektseid esines 2 korral (6%). (joonis 3)
3) Claude´i vastustes esines tugevaid ravisoovitusi 5 korral (14%), mõõdukaid 8 korral (23%), nõrku 16 korral (45%). Kasutatud ravijuhiste põhjal ei saanud hinnata vastuste kvaliteedi kolme vastuse puhul (8%). Ebakorrektseid vastuseid esines 3 korral (8%). (joonis 4)

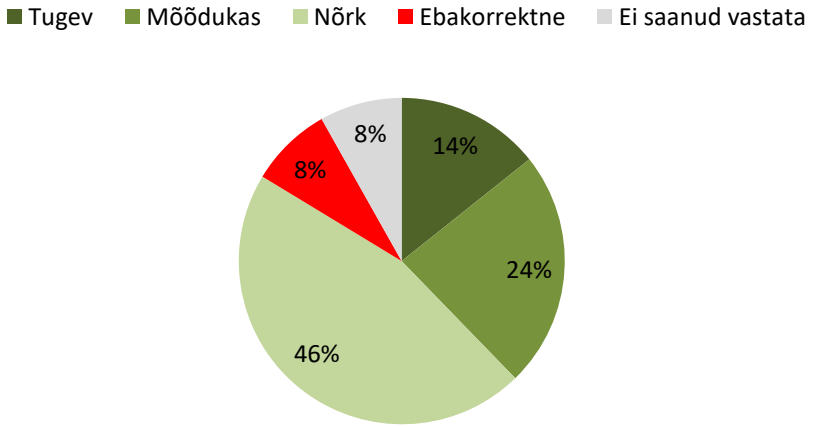
Tugevaid ravisoovitusi andsid kõik mudelid küsimustele vastamisel, mis puudutasid paranemisaega, füsioteraapiat ning edasisi uuringuid. Kõige rohkem eksisid mudelid vastamisel küsimustele, mis käsitlesid harjutuste mahtu ning massaaži. Gemini eksis lisaks nendele teemadele ka ortoosi kasutamise ja venitusharjutuste osas. Vastuolud ravijuhendite vahel tekkisid küsimustes, mis käsitlesid lööklaineteraapiat, probleemi kordumise riski, harjutusmahu määramisel ning ka igapäevaelu kohandamisel. Kolme küsimuse puhul ei olnud vastuste kvaliteeti võimalik ravijuhendite alusel hinnata. Need küsimused käsitlesid harjutuste rolli ülekoormusprobleemi puhul, konkreetset raskuskande piirangut ning kreemide ja salvide kasutamist. LLM-e vastuste kvaliteet spetsiifiliste küsimuste lõikes on välja toodud tabelites 1-3.



Joonis 2. Gemini vastuste üldine täpsus võrreldes 3 erineva ravijuhisega (küsimuste arv = 35)



Joonis 3. Chat GPT vastuste üldine täpsus võrreldes 3 erineva ravijuhisega (küsimuste arv = 35)



Joonis 4. Claude´i vastuste üldine täpsus võrreldes 3 erineva ravijuhisega (küsimuste arv = 35)

Tabel 1. Gemini mudel 3 vastuste kvaliteet spetsiifiliste küsimuste lõikes

	JOSPT	BESS	SPORTM.
1. Kaua paraneb?			
2. Harjutused tegid käe valusamaks, kas harjutusi peab tegema läbi valu?			-
3. Milliseid harjutusi teha?			
4. Miks see valu üldse tekkis?		-	
5. Kas füsioteraapia aitab ja kui kiiresti võiksin tulemusi oodata?			
6. Kui kaua paranemine tavaliselt aega võtab?			
7. Palju harjutusi teha?		-	-
8. Mida peaks vältima, et valu ei tekiks?		-	
9. Kas ma peaksin oma igapäevaseid tegevusi piirama?			
10. Milliseid harjutusi ma võin teha ja milliseid peaksin vältima?			
11. Kas ma võin jätkata sporti või tööd, mis nõuab käe kasutamist?		-	
12. Mis võiks aidata valu vähemaks võtta – kuum, külm, kreemid?	-		-
13. Kas võtta valuvaigisteid?	-		
14. Kas kasutada ortoosi?			
15. Miks ma lisaks harjutusi pean tegema, kui see tekkis ülekoormusest?	-	-	-
16. Kas venitusharjutused aitavad?		-	-
17. Kas peaks põletikuvastast rohtu võtma?	-		
18. Kas ortoosist on kasu?			
19. Kas ma peaksin tegema röntgeni, ultraheli või MRI?			
20. Miks ei ole lisauuringuid tehtud?			
21. Mis see lööklaineravi on, mida arst soovitas?	-	-	
22. Kas trennis tohib edasi käia?		-	
23. Kui suurt raskust tohin tõsta?	-	-	-
24. Kas pean hakkama iga päev harjutusi tegema?	-		-
25. Missugust valuvaigistit ma peaksin võtma?	-		-
26. Kui tõsine mu seisund on ja kas sellest lõpuks saab lahti või jääbki valutama?			
27. Millised ravivõimalused on olemas ja milline neist on minu puhul sobivaim?			
28. Kas mu seisund võib korduda ja kuidas ma seda vältida saan?			-
29. Kui kiiresti võib ravi järel valu vähenemist oodata ja kui püsivad on tulemused?		-	
30. Kas massaaž aitaks seda?		-	
31. Kas ja millise kreemiga/salviga peaks kätt määrima?	-	-	-
32. Mis täpselt põhjustab minu probleem/valu küünarnukis?		-	
33. Kui lööklaineteraapiast kasu pole mis edasised võimalused on?	-		
34. Kuidas täpselt lööklaineravi töötab ja palju seansse tavaliselt vaja on?	-	-	
35. Kas lööklaineteraapia sobib minu probleemi puhul?	-		

Rohelise värviga on tähistatud LLMi vastused, mis vastasid ravijuhendile, millega vastust võrreldi. Kollase värviga on tähistatud vastused, mis osaliselt vastasid ravijuhendile. Punasega on tähistatud vastused, mis ei vastanud ravijuhendile. Oranžiga on märgitud küsimused, mille puhul LLM esitas lahtiütluse. "-" märgiga tähisatud vastuseid ei olnud ravijuhendis käsitletud piisavalt, et vastuste kvaliteedi saaks analüüsida.

Tabel 2. Chat GPT 5.3 versiooni vastuste kvaliteet spetsiifiliste küsimuste lõikes

	JOSPT	BESS	SPORTM
1. Kaua paraneb?			
2. Harjutused tegid käe valusamaks, kas harjutusi peab tegema läbi valu?			-
3. Milliseid harjutusi teha?			
4. Miks see valu üldse tekkis?		-	
5. Kas füsioteraapia aitab ja kui kiiresti võiksin tulemusi oodata?			
6. Kui kaua paranemine tavaliselt aega võtab?			
7. Palju harjutusi teha?			-
8. Mida peaks vältima, et valu ei tekiks?		-	
9. Kas ma peaksin oma igapäevaseid tegevusi piirama?			
10. Milliseid harjutusi ma võin teha ja milliseid peaksin vältima?			
11. Kas ma võin jätkata sporti või tööd, mis nõuab käe kasutamist?			
12. Mis võiks aidata valu vähemaks võtta – kuum, külm, kreemid?	-		-
13. Kas võtta valuvaigisteid?	-		
14. Kas kasutada ortoosi?			
15. Miks ma lisaks harjutusi pean tegema, kui see tekkis ülekoormusest?	-	-	-
16. Kas venitusharjutused aitavad?		-	-
17. Kas peaks põletikuvastast rohtu võtma?	-		
18. Kas ortoosist on kasu?			
19. Kas ma peaksin tegema röntgeni, ultraheli või MRI?			
20. Miks ei ole lisauuringuid tehtud?			
21. Mis see lööklaineravi on, mida arst soovitas?	-	-	
22. Kas trennis tohib edasi käia?		-	
23. Kui suurt raskust tohin tõsta?	-	-	-
24. Kas pean hakkama iga päev harjutusi tegema?	-		-
25. Missugust valuvaigistit ma peaksin võtma?	-		-
26. Kui tõsine mu seisund on ja kas sellest lõpuks saab lahti või jääbki valutama?			
27. Millised ravivõimalused on olemas ja milline neist on minu puhul sobivaim?			
28. Kas mu seisund võib korduda ja kuidas ma seda vältida saan?		-	-
29. Kui kiiresti võib ravi järel valu vähenemist oodata ja kui püsivad on tulemused?		-	
30. Kas massaaž aitaks seda?		-	
31. Kas ja millise kreemiga/salviga peaks kätt määrima?	-	-	-
32. Mis täpselt põhjustab minu probleem/valu küünarnukis?		-	
33. Kui lööklaineteraapiast kasu pole mis edasised võimalused on?	-	-	
34. Kuidas täpselt lööklaineravi töötab ja palju seansse tavaliselt vaja on?	-	-	
35. Kas lööklaineteraapia sobib minu probleemi puhul?	-		

Rohelise värviga on tähistatud LLMi vastused, mis vastasid ravijuhendile, millega vastust võrreldi. Kollase värviga on tähistatud vastused, mis osaliselt vastasid ravijuhendile. Punasega on tähistatud vastused, mis ei vastanud ravijuhendile. Oranžiga on märgitud küsimused, mille puhul LLM esitas lahtiütluse. "-" märgiga tähisatud vastuseid ei olnud ravijuhendis käsitletud piisavalt, et vastuste kvaliteedi saaks analüüsida.

Tabel 3. Claude versiooni Sonnet 4.6 vastuste kvaliteet spetsiifiliste küsimuste lõikes

	JOSPT	BESS	SPORTM
1. Kaua paraneb?			
2. Harjutused tegid käe valusamaks, kas harjutusi peab tegema läbi valu?			-
3. Milliseid harjutusi teha?			
4. Miks see valu üldse tekkis?		-	
5. Kas füsioteraapia aitab ja kui kiiresti võiksin tulemusi oodata?			
6. Kui kaua paranemine tavaliselt aega võtab?			
7. Palju harjutusi teha?			-
8. Mida peaks vältima, et valu ei tekiks?		-	
9. Kas ma peaksin oma igapäevaseid tegevusi piirama?			
10. Milliseid harjutusi ma võin teha ja milliseid peaksin vältima?			
11. Kas ma võin jätkata sporti või tööd, mis nõuab käe kasutamist?		-	
12. Mis võiks aidata valu vähemaks võtta – kuum, külm, kreemid?	-		-
13. Kas võtta valuvaigisteid?	-		
14. Kas kasutada ortoosi?			
15. Miks ma lisaks harjutusi pean tegema, kui see tekkis ülekoormusest?	-	-	-
16. Kas venitusharjutused aitavad?		-	-
17. Kas peaks põletikuvastast rohtu võtma?	-		
18. Kas ortoosist on kasu?			
19. Kas ma peaksin tegema röntgeni, ultraheli või MRI?			
20. Miks ei ole lisauuringuid tehtud?			
21. Mis see lööklaineravi on, mida arst soovitas?	-	-	
22. Kas trennis tohib edasi käia?		-	
23. Kui suurt raskust tohin tõsta?	-	-	-
24. Kas pean hakkama iga päev harjutusi tegema?	-		-
25. Missugust valuvaigistit ma peaksin võtma?	-		-
26. Kui tõsine mu seisund on ja kas sellest lõpuks saab lahti või jääbki valutama?			
27. Millised ravivõimalused on olemas ja milline neist on minu puhul sobivaim?			
28. Kas mu seisund võib korduda ja kuidas ma seda vältida saan?		-	-
29. Kui kiiresti võib ravi järel valu vähenemist oodata ja kui püsivad on tulemused?		-	
30. Kas massaaž aitaks seda?		-	
31. Kas ja millise kreemiga/salviga peaks kätt määrima?	-	-	-
32. Mis täpselt põhjustab minu probleem/valu küünarnukis?		-	
33. Kui lööklaineteraapiast kasu pole mis edasised võimalused on?	-		
34. Kuidas täpselt lööklaineravi töötab ja palju seansse tavaliselt vaja on?	-	-	
35. Kas lööklaineteraapia sobib minu probleemi puhul?	-		

Rohelise värviga on tähistatud LLMi vastused, mis vastasid ravijuhendile, millega vastust võrreldi. Kollase värviga on tähistatud vastused, mis osaliselt vastasid ravijuhendile. Punasega on tähistatud vastused, mis ei vastanud ravijuhendile. Oranžiga on märgitud küsimused, mille puhul LLM esitas lahtiütluse. "-" märgiga tähisatud vastuseid ei olnud ravijuhendis käsitletud piisavalt, et vastuste kvaliteedi saaks analüüsida.

4.2. Suurte keelemudelite loetavuse hindamine *Flesch Reading Ease Score*'i alusel

Vastuste loetavuse hindamiseks arvutati iga vastuse kohta *Flesch Reading Ease Score'*i skoor. Kõige lihtsamini loetavaid vastuseid andis ChatGPT ning kõige keerulisemalt loetavaid vastuseid

Gemini mudel 3. Gemini mudel 3 puhul oli madalaimaks FRES skooriks 47 ning kõrgeimaks skooriks 61. Vastuste aritmeetiline keskmine oli 56,5, mis viitab üsna keerulise loetavusega tekstile ehk vähemalt gümnaasiumitasemel lugemisoskuse vajalikusele. ChatGPT 5.3 puhul oli madalaim skoor 57 ning kõrgeim 75. Vastuste aritmeetiline keskmine oli 66 ehk vastused eeldavad vähemalt 8.-9. klassi tasemel lugemisoskust. Claude Sonnet 4.6 vastuste puhul oli madalaimaks skooriks 40 ja kõrgeimaks 75. Aritmeetiline keskmine oli 58,4, mis vastab 10.-12. klassi lugemisoskusele.

4.3. Suurte keelemudelite lahtiütluste analüüs

Lahtiütlusi analüüsiti nende esinemise sageduse ning sisu alusel. 35 küsimuse puhul esines lahtiütlusi erinevalt vastavalt kasutatud mudelile. Geminil mudel 3 vastustes esines kaks lahtiütlust, millest ühe puhul rõhutas ta, et vastusel on ainult informatiivne eesmärk ning arstiabi saamiseks pöörduda tervishoiutöötaja poole. Teise lahtiütluse puhul rõhutas LLM, et tegemist on tehisintellektil põhineva mudeliga ning seega ei saa ta kirjutada ravimitele ametlikku retsepti. Claude Sonnet 4.6 vastustes esines kaheksa lahtiütlust. Kahel korral rõhutas mudel, et tema antud nõu on üldist laadi ning täpsemate ravisoovituste saamiseks tuleks pöörduda spetsialisti poole. Viiel korral soovitas Claude pöörduda arsti või apteekri poole nii valuvaigistite, erinevate ravimite koostoimete ning ka ortoosi kasutamise osas. Lisaks rõhutas Claude korduvalt, et parimate ravitulemuste saamiseks on vajalik tervisehoiuspetsialisti poole pöördumine. ChatGPT ei esitanud enda vastustes mitte ühtegi lahtiütlust.

5.ARUTELU

Käesolevas magistritöös hinnati kolme LLM-i (Gemini mudel 3, ChatGPT 5.3 ja Claude Sonnet 4.6 versiooni) kvaliteeti patsientide enam levinud küsimustele vastamisel lateraalse epikondüliidi kohta. LLM-ide vastuste täpsust hinnati võrreldes kolme erineva ravijuhisega (JOSPT, BESS ja Sports Medicine). Lisaks analüüsiti vestlusrobotite vastuste loetavust *Flesch Reading Ease Score*'i (FRES) alusel ning lahtiütluste esinemist vastustes.

Magistritöö peamised tulemused on:

1) Kolmest töös kasutatud LLM-ist oli Chat GPT 5.3 vastamisel kõige täpsem, järgnesid Claude Sonnet 4.6 ning seejärel Gemini mudel 3.
2) ChatGPT 5.3 vastused olid kolmest mudelist kõige paremini mõistetavad. FRES keskmine 66, mis vastab 8.-9. klassi õpilase lugemisoskusele. Gemini mudel 3 ja Claude vastuste lugemiseks on vajalik vähemalt gümnaasiumi tasemel lugemisoskus.
3) Claude esitas 35 küsimuse kohta 8 lahtiütlust, suunates patsiente korduvalt spetsialisti poole pöörduma. Gemini esitas kaks lahtiütlust. ChatGPT 5.3 ei esitanud oma vastustes mitte ühtegi lahtiütlust.

5.1. Suurte keelemudelite poolt antud vastuste täpsus võrreldes valitud ravijuhenditega

Käesoleva töö tulemused näitavad, et LLM-ide vastuste täpsus lateraalse epikondüliidi teemal oli mudelite lõikes märkimisväärselt erinev. ChatGPT 5.3 osutus kõige täpsemaks mudeliks: ebakorrektseid vastuseid esines vaid 6% juhtudest ning tugevaid ravisoovitusi anti 20% vastustes. Claude jäi täpsuselt teisele kohale (8% ebakorrektseid vastuseid), Gemini mudel 3 andis aga ebakorrektseid vastuseid märkimisväärselt sagedamini - 26% juhtudest.

Varasemates uuringutes on leitud madalamaid täpsuse määrasid. Yau ja teiste 2024. aasta uuringus vaadati LLMide täpsust erakorralise meditsiini osas. Töö tulemustena leiti, et vastuste täpsus oli 50% ning edastasid ohtliku või valet informatsiooni 5-35% juhtudest. (Yau *et al.,* 2024) Wei ja teiste 2024. aasta meta-analüüsis oli ChatGPT-3.5 üldine täpsus meditsiinilistele küsimustele vastamisel vaid 56% (Scaff et al., 2025; Wei *et al.,* 2024). Goodmani ja teiste 2023. aasta uuring leidis, et vaid 50% ChatGPT vastustest olid täpsed (Goodman *et al.,* 2023). Käesolevat magistritööd saab võrrelda lisaks eelmainitutele ka Scaffi ja teiste 2025. aasta mittespetsiifilise alaseljavalu uuringuga, kus 55,8% vastustest olid korrektsed ja ebakorrektseid oli 42,1% (Scaff et al., 2025). Käesoleva magistritöö tulemused viitavad sellele, et mudelite areng on toonud kaasa täpsuse paranemise.

Sellest hoolimata esines ebakorrektseid vastuseid ning nõrkade ravisoovituste osakaal oli suur – Geminil 37%, ChatGPT-l 51%, Claude-l 45%. See viitab sellele, et LLM-id võivad olla raskustes kliiniliselt keerulisemate ja nüansseeritumate küsimustega, kus patsient küsib ravi loogikat selgitavat vastust. Takita ja teised leidsid 2025. aasta uuringus, et AI poolt antud vastused olid korrektsed 52,7% juhtudest ning järeldasid, et LLM-id ei suuda hetkel asendada eksperdi nõuandeid, kuna need ei arvesta inimkeha komplekssust (Takita *et al.,* 2025). Käesoleva töö tulemused toetavad seda järeldust. Kuigi ChatGPT 5.3 täpsus oli märkimisväärselt kõrgem, esines siiski ebakorrektseid vastuseid, mis kliinilises kontekstis võivad patsientide ravi ning taastumist negatiivselt mõjutada. Meyeri ja teiste 2024. aasta uuring leidis, et ChatGPT andis küll kõige rohkem korrektseid vastuseid, kuid ka kõige enam mittetäielike või osaliselt korrektseid vastuseid (Meyer *et al.,* 2024). Ayersi ja teiste 2023 aasta uuring leidis, et LLM-i poolt antud vastused olid täpsemad ning empaatilisemad kui inimese omad. Sellest hoolimata tõid töö autorid välja, et LLM-id peaksid abistama spetsialistide tööd, et vähendada läbipõlemist ja parandada kvaliteeti ning kättesaadavust, aga mitte täielikult spetsialisti tööd hakkama asendama (Ayers *et al.,* 2023). Magistritöö tulemusena saame järeldada, et LLM-id ei ole veel täielikult täpsed ning ei saa veel asendada spetsialiste. Tugevate ravisoovitustena olid vastused küsimustele, mis olid konkreetsemad ning faktipõhisemad. Komplekssemate küsimuste puhul LLM-e täpsus juba langes.

Töö tulemustes on ka näha, et kolme küsimuse vastuse kvaliteeti ei saanud kasutatud ravijuhiste põhjal hinnata. Lisaks on tabelitest 1-3 näha, et kõik ravijuhendid ei ole suutnud vastata kõikidele patsientide enim levinud küsimustele. See omakorda tõstatab küsimuse ravijuhendite koostamise ja standardiseerimise kohta. Käesoleva magistritöö tulemuste viitavad sellele, et ravijuhendid peaksid olema põhjalikumad ning käsitlema ka teemasid, mille kohta patsiendid rohkem soovivad teada. Antud magistritöös kasutatud ravijuhendid olid töö autori hinnangul koostatud pigem spetsialistidele kasutamiseks. Lisaks saab antud töö tulemuste põhjal järeldada, et on edaspidi vaja ka konsensust ravijuhendite vahel. Antud magistritöö spetsiifiliste küsimuste vastuste tabelid 1-3 toovad välja vastuolusid ravijuhendite vahel, kus sama LLM-i vastus sai kolme erineva ravijuhendiga võrdluses täiesti erineva hinnangu. Lima ja teiste 2023. aasta uuring ravijuhendite kohta tõi välja, et mittestruktureeritud väljatöötamise, huvide konfliktide ning isiklike eelistuste tõttu on ravijuhendid muutunud küll ajaga paremaks, kuid ravijuhistel on endiselt suuri puudujääke. Samas artiklis on ühe olulise hea ravijuhendi tunnusena välja toodud, et koostajad on kaalunud kõiki patsientide jaoks olulisi aspekte ning tulemusi. Lisaks, et kas ravijuhendid on koostatud vastavalt kaasaegsetele uuringutele tuginedes. (Lima *et al.,* 2023) Antud magistritöö tõi välja, et magistritöös kasutatud kolm ravijuhendit ei kajastanud kõiki patsientide jaoks olulisi küsimusi. Lisaks on kõik kolm ravijuhendit koostatud

väikse ajalise vahega ning ravisoovitused mõnede küsimuste puhul on teineteisele vastuolus. Samaaegselt koostatud kaasaegsele teaduskirjandusele tuginevate ravijuhendite vastuolu näitab ravijuhiste koostamise kitsaskohti.

5.2. LLM-ide mõistetavuse hindamine *Flesch Reading Ease Score*'i alusel

Tervisealase kirjaoskuse olulisus muutub eriti aktuaalseks olukorras, kus patsiendid kasutavad terviseinfo hankimiseks järjest enam LLM-e. Loetavuse hindamisel selgus, et ChatGPT 5.3 vastused olid kolmest mudelist kõige parema loetavusega. ChatGPT mõistetavus FRES skaalal oli keskmiselt 66, mis vastab 8.-9. klassi õpilase lugemisoskusele. Gemini mudel 3 ja Claude Sonnet 4.6 vastuste lugemiseks on vajalik vähemalt gümnaasiumi tasemel lugemisoskus. FRES tulemused olid vastavalt 56,5 ja 58,4. Kõige madalam üksikskoor esines Claudeil ning kõige kõrgem nii ChatGPT-l kui Claude´il.

Scaffi ja teiste 2025. aasta mittespetsiifilise alaseljavalu uuringus oli FRES keskmine tulemus 50,94, mis on madalam käesoleva töö tulemustest. Tulemused viitavad sellele, et lateraalse epikondüliidi kohta antud vastused olid ka paremini loetavad. Guveni ja teiste 2024. aasta uuring hindas FRESi tulemuseks ChatGPT 3.5-l 49,25, Chat GPT 4.0-l 46,42 ja Geminil 51,91 (Guven *et al.,* 2024). Yau ja teiste 2024 aasta uuring erakorralise meditsiini küsimustele vastamisel hindas ChatGPT vastuste loetavust 10. klassi tasemeks (Yau *et al.,* 2024). Selle põhjal saab oletada, et ka ChatGPT loetavus on ajaga paranenud. Uuringu tulemused viitavad, et uute mudelite areng on lisaks täpsuse paranemisele kaasa toonud ka mõistetavuse paranemise. Meyeri ja teiste 2024. aasta uuring ei leidnud ChatGPT, Gemini ja Claude´i loetavuse vahel statistiliselt olulist erinevust. Küll aga leidsid töö autorid, et loetavuse tase on märkimisväärselt keerulisem 5.-6. klassi tasemest, mida soovitatakse patsientidele väljatöötatud materjalide jaoks. Lisaks toodi uuringus välja, et hetkel proovitakse LLM-e vastuseid muuta paremini mõistetavaks läbi jooniste lisamise. (Meyer *et al.,* 2024) Sevgini ja teiste 2026. aasta uuringus hinnati Gemini ja ChatGPT kvaliteeti ning loetavust kaelavaluga seotud küsimustele vastamisel. FRES tulemuseks Geminil oli 47,12 ja ChatGPT tulemus oli 48,78. Need tulemused viitavad, et vastuseid oli keeruline mõista ning viitab kõrgkooli lugemisetasemele. (Sevgin *et al.,* 2026) Hanci ja teiste 2024. aasta uuring hindas viie erineva LLM-i kvaliteeti, täpsust ning loetavust palliatiivraviga seotud küsimustele vastamisel. FRES tulemus Chat GPT-l oli 22,10 ning Geminil 24,00. Vastused olid seega väga keerulised ning kõik viis uuringus olnud LLM-idest jäi kaugele soovitatud kuuenda klassi lugemistasemest. (Hanci *et al.,* 2024)

Võrreldes varasemate uuringutega saab järeldada, et suurtel keelemudelitel on ka loetavus ajaga paranenud, kuid ikkagi vajab loetavus edasist parandamist. Vastuste loetavust muudab

märkimisväärselt keerulisemaks erialaste terminite kasutamine. Loetavuse parandamiseks oleks vaja vähendada meditsiinilise keele kasutamist LLM-ide poolt antavates vastustes või neid pikemalt lahti seletada. LLM-ide loetavust mõjutab ka teema valik. Antud magistritöö tulemused viitavad sellele, et lateraalse epikondüliidi vastused on lihtsamini mõistetavad kui mittespetsiifilise kaela- ning alaseljavalu ja palliatiivraviga seotud vastused.

5.3. LLM-ide poolt antud lahtiütluste analüüs

Lahtiütluste analüüs tõi esile selge erinevuse mudelite vahel. Claude esitas 35 küsimuse kohta 8 lahtiütlust, suunates patsiente korduvalt spetsialisti poole pöörduma. Gemini esitas kaks lahtiütlust. ChatGPT 5.3 ei esitanud oma vastustes mitte ühtegi lahtiütlust.

ChatGPT lahtiütluste täielik puudumine on murettekitav leid, eriti arvestades, et küsimuste hulgas oli valuvaigistite ja ravimite kasutamist puudutavaid küsimusi, mis nõuavad meditsiinilist hindamist. Sallami 2023. aasta uuringus toodi välja, et meditsiinivaldkonnas levib palju väärinfot, mida LLM-id võivad patsientidele tõena esitada, ohustades seeläbi patsientide tervist. Kui LLM annab kindlameelseid vastuseid ravimite kohta ilma spetsialistile suunamiseta, võib see patsienti julgustada otsuseid tegema ilma vajaliku meditsiinilise konsultatsioonita. (Sallam*.,* 2023) Thirunavukarasu ja teiste 2023. aasta uuring rõhutas, et LLM-i kasutamine ravis ei ole 100% täpne ning puudujäägina toodi välja, et saadud soovitused ei ole patsiendipõhised (Thirunavukarasu *et al.,* 2023). Ralla ja teiste 2025. aasta uuring leidis, et 14% sisestatud küsimustest olid problemaatilised. Seda kas ebatäpse informatsiooni, liigse enesekindluse või täiendavate uuringute osas. Lisaks tõid uuringu autorid välja, et ka küsimuste sisestamise viis võib muuta LLM-i poolt antavat vastust. (Ralla *et al.,* 2025)

Lahtiütluste esitamine on üks viis, kuidas LLMid saavad piiranguid patsientidele kommunikeerida ning soovitada neile spetsialistiga konsulteerimist. Käesoleva töö tulemused viitavad, et mudelid lähenevad lahtiütlustele esitamisele erinevalt. Erinevused tulenevad LLM-ide koostamise ning nö õpetamise eripäradest (Bai *et al.,* 2022). Seda aspekti tuleks LLM-ide kliinilise kasutuse hindamisel arvesse võtta ning töö tulemusena saab ka öelda, et täielik lahtiütluste puudumine on problemaatiline.

Varasemalt on teostatud sarnane uuring mittespetsiifilise alaseljavalu kohta. Tervisega seotud hoiatusi esitas ChatGPT-4.0 kõikides vastustes, ChatGPT-3.5 96,6% vastustest, Gemini 83,33% vastustest ja Bing 70% vastustest. (Scaff *et al.,* 2025) Võrreldes selle uuringuga saab magistritöö põhjal järeldada, et LLMid on muutunud oma vastustes oluliselt enesekindlamaks. Magistritöös on korduvalt väljatoodud, et ChatGPT puhul lahtiütluste puudumine on problemaatiline. Ka Gemini ja Claude puhul on tegelikult lahtiütluste esinemise sagedus küllaltki madal. Vastavalt Geminil 5,71% ja Claude'il

22,8%. Selle põhjal saab väita, et mudelite areng on muutnud neid enesekindlamaks, mis potentsiaalselt võib tekitada kahju.

5.4. Uuringu piirangud ja tugevused

Käesoleval uuringul on erinevaid piiranguid, mida tuleks tulemuste tõlgendamisel arvesse võtta. Esiteks sisestati kõik küsimused LLM-idele vaid ühe korra. On teada, et LLM-ide vastused võivad sama küsimuse korduval esitamisel varieeruda. See tähendab, et ühekordne sisestamine ei pruugi täielikult peegeldada mudeli tegelikku sooritust. Teiseks hindas LLM-ide vastuseid üks uurija, mis toob kaasa subjektiivsuse riski eriti piiripealsete juhtumite hindamisel. Edaspidi võiks kasutada mitme hindaja paneeli ning hindajatevahelise reliaabluse mõõtmist. Kolmandaks toetuti täpsuse hindamisel kolmele ravijuhisele (JOSPT, BESS, Sports Medicine), mis ei pruugi katta kõiki kliiniliselt asjakohaseid soovitusi. Mõned LLM-i vastused võivad olla kliiniliselt põhjendatud, kuid jäid käesolevas töös hindamata, sest kasutatud ravijuhised neid ei sisaldanud. Neljandaks muutuvad LLM-ide versioonid kiiresti. Käesolevas töös kasutatud mudelite versioonid ajas muutuvad, mistõttu tulemused ja töös kasutatud versioonid aeguvad ning nõuavad perioodilist kordusuuringut.

Käesoleva magistritöö üheks tugevuseks on teema aktuaalsus. Suuri keelemudeleid ning tehisintellekti kasutatakse aktiivselt terviseinfo hankimisel, mistõttu on oluline hinnata nende vastuste kvaliteeti. Teiseks tugevuseks on küsimustiku koostamine Delphi konsensusmeetodi abil. Küsimustik põhines küsimustel, mida ekspertide hinnangul olid patsiendid nende käest küsinud. See suurendas töö praktilist väärtust. Kolmandaks tugevuseks on mitme suure keelemudeliga võrdlemine, mis võimaldab hinnata erinevate mudelite vastuste kvaliteedi lateraalse epikondüliidi kontekstis. Neljandaks lisaks täpsuse hindamisele ka loetavuse hindamine ja lahtiütluste analüüsimine. Kõik kolm aspekti on patsientide terviseinfo ning turvalisuse seisukohast väga olulised. Sarnastes varasemates uuringutes ei ole sageli lahtiütluste analüüsimisele tähelepanu pööratud.

5.5. Praktiline tähendus

Käesoleva töö tulemustel on praktiline tähendus nii füsioterapeutidele kui ka tervishoiusüsteemile laiemalt. Füsioterapeudid peaksid olema teadlikud, et nende patsiendid kasutavad LLMisid aktiivselt oma terviseprobleemide kohta info otsimiseks. Seega on reaalne, et patsient saabub vastuvõtule suurest keelemudelist saadud informatsiooniga, mis võib olla ebatäpne või raskesti mõistetav. Nii ebakorrektne informatsioon kui ka keeruline loetavus võivad patsiendi ravi negatiivselt mõjutada. Töö tulemused aitavad tervishoiu spetsialistidel paremini mõista, kui kvaliteetset informatsiooni võivad patsiendid suurtelt keelemudelitelt saada ning millised riskid sellega kaasneda võivad.

Käesoleva magistritöö tulemused viitavad, et ChatGPT 5.3 on uuritud mudelitest täpseim ja paremini loetav, kuid lahtiütluste täielik puudumine selles mudelis on oluline kliiniline ohumärk. Gemini puhul on antud töö kontektsis suurimaks kitsaskohaks selle LLM-i ebatäpsus. Claude'i esitatud lahtiütlused viitavad sellele, et mudel tunnistab oma piiranguid meditsiinilises kontekstis - see on kliiniliselt vastutustundlikum käitumine. Cabrali ja teiste 2025. aasta uuringu järelduse kohaselt hakkab AI järgmise 10 aasta jooksul kõige enam mõjutama diagnostikat, mitte üle võtma patsiendi ravi tervikuna(Cabrali *et al.,* 2025). Käesolev magistritöö toetab seda seisukohta: LLM-id võivad olla kasulikuks täiendavaks infokanaliks, kuid ei suuda asendada individuaalset kliinilist hindamist ning kompleksemate probleemide lahendamist.

6. JÄRELDUSED

Käesoleva magistritöö tulemuste põhjal tehti järgmised järeldused:

1. Mudelite areng on toonud kaasa täpsuse paranemise, kuid LLM-id ei suuda hetkel veel täielikult asendada tervishoiuspetsialisti lateraalse epikondüliidi käsitluses.
2. Patsientidele suunatud terviseinfo edastamiste soovitustega võrreldes on LLM-ide poolt antavad vastused liiga keerulise loetavusega. Laused ning sõnad on LLM-ide vastuses liiga pikad ning on edaspidi vaja muuta oluliselt lihtsamini mõistetavaks.
3. LLM-ide genereeritud vastused sisaldasid vähe lahtiütlusi ning nende esinemise sagedus oli mudeliti erinev. Lahtiütluste esinemine on LLM-ide osas oluline, et tagada turvalisus ning vähendada potentsiaalset tervisekahju.

KASUTATUD KIRJANDUS

1. Abd-Alrazaq, A., AlSaad, R., Shuweihdi, F., Ahmed, A., Aziz, S., *et al*. (2023). Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. *Digital Medicine.*, 6(84). Doi: https://doi.org/10.1038/s41746-023-00828-5
2. Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., *et al.* (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine.*, 183(6): 589–596. Doi: https://doi.org/10.1001/jamainternmed.2023.1838
3. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., *et al.* (2022). Constitutional AI: Harmlessness from AI Feedback. *Anthropic.* Doi: https://arxiv.org/pdf/2212.08073
4. Berkman, N.D., Sheridan, S.L., Donahue, K.E., Halpern, D.J., Crotty, K. (2011). Low health literacy and health outcomes: an updated systematic review. *Annals of internal medicine*., 155(2):97-107. doi: 10.7326/0003-4819-155-2-201107190-00005.
5. Bujnowska-Fedak, M. M., & Węgierek, P. (2020). The Impact of Online Health Information on Patient Health Behaviours and Making Decisions Concerning Health. *International journal of environmental research and public health.*, *17*(3), 880. https://doi.org/10.3390/ijerph17030880. doi: 10.3390/ijerph17030880.
6. Cabral, L., Pinto, R., Goncalves, G. (2025). AI-powered learning analytics dashboards: a systematic review of applications, techniques, and research gaps. *Discover Education.,* 4(1). Doi: 10.1007/s44217-025-00964-y
7. Challoumas, D., Biddle, M., Millar, N. L. (2020). Recent advances in tendinopathy. *Faculty Reviews.,*9: 16. https://doi.org/10.12703/b/9-16
8. Chen, Q., Shen, P., Zhang, B., Chen, Y., & Zheng, C. (2024). A meta-analysis of the risk factors for lateral epicondylitis. *Journal of Hand Therapy.*, *37*(1): 44–52. https://doi.org/10.1016/j.jht.2023.05.013
9. Çıracıoğlu, A. M., Dal Erdoğan, S. (2025). Evaluation of the reliability, usefulness, quality and readability of ChatGPT's responses on Scoliosis. *European journal of orthopaedic surgery & traumatology : orthopedie traumatologie.*, 35 (1): 123. Doi: https://doi.org/10.1007/s00590-025-04198-4

10. Cook, J. L., Rio, E., Purdam, C. R., Docking, S. I. (2016). Revisiting the continuum model of tendon pathology: what is its merit in clinical practice and research?. *British journal of sports medicine*., *50*(19): 1187–1191. https://doi.org/10.1136/bjsports-2015-095422
11. Daher, M., Koa, J., Boufadel, P., Singh, J., Fares, M. Y., *et al.* (2023). Breaking barriers: Can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management?*JSES international.,* 7(6): 2534-2541. Doi: https://jsesinternational.org/article/S2666-6383(23)00212-8/fulltext
12. Drumm, S., Bradley, C., & Moriarty, F. (2022). 'More of an art than a science'? The development, design and mechanics of the Delphi Technique. *Research in Social and Administrative Pharmacy.*, 18(1): 2230–2236. Doi: https://doi.org/10.1016/j.sapharm.2021.06.027
13. Gallifant, J., Afshar, M., Ameen, S., Aphinyanaphongs, Y., Chen, S. (2025). The tripod-llm reporting guideline for stuides using large language models. *Nature medicine.,* 31: 60-69. Doi: https://www.nature.com/articles/s41591-024-03425-5#citeas
14. Goddard J. (2023). Hallucinations in ChatGPT: A Cautionary Tale for Biomedical Researchers. *The American journal of medicine.*, *136*(11): 1059–1060. Doi: https://doi.org/10.1016/j.amjmed.2023.06.012
15. Goodman, R. S., Patrinely, J. R., Stone, C. A., Jr, Zimmerman, E., Donald, R. R., *et al*. (2023). Accuracy and Reliability of Chatbot Responses to Physician Questions. *JAMA network open*., 6(10). Doi: https://doi.org/10.1001/jamanetworkopen.2023.36483
16. Guven, Y., Ozdemir, O. T., Kavan, M. Y. (2025). Performance of Artificial Intelligence Chatbots in Responding to Patient Queries Related to Traumatic Dental Injuries: A Comparative Study. *Dental Traumatology.*, 41(3): 338–347. Doi: https://doi.org/10.1111/edt.13020
17. Hancı, V., Ergün, B., Gül, Ş., Uzun, Ö., Erdemir, İ., *et al.* (2024). Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. *Medicine.,* 103(33). Doi: https://journals.lww.com/md-journal/fulltext/2024/08160/assessment_of_readability,_reliability,_and.61.aspx
18. Holey, E. A., Feeley, J. L., Dixon, J., & Whittaker, V. J. (2007). An exploration of the use of simple statistics to measure consensus and stability in Delphi studies. *BMC Medical Research Methodology.*, 7(1): 52. Doi: https://doi.org/10.1186/1471-2288-7-52

19. Institute of Medicine (US) Committee on Health Literacy, Nielsen-Bohlman, L., Panzer, A. M., Kindig, D. A. (2004). *Health Literacy: A Prescription to End Confusion*. National Academies Press (US). https://doi.org/10.17226/10883
20. Irby, A., Gutierrez, J., Chamberlin, C., Thomas, S.J., Rosen, A.B., (2020). Clinical management of tendinopathy: A systematic review of systematic reviews evaluating the effectiveness of tendinopathy treatments. *Scandinavian Journal of Medicine & Science in Sports—Wiley Online Library*., 30 (10): 1810-1826. Doi: https://onlinelibrary.wiley.com/doi/10.1111/sms.13734
21. Jindal, P., MacDermid, J. C. (2017). Assessing reading levels of health information: uses and limitations of flesch formula. *Education for health (Abingdon, England).*, 30(1): 84–88. Doi: https://doi.org/10.4103/1357-6283.210517
22. Keijsers, R., de Vos, R.-J., Kuijer, P. P. F., van den Bekerom, M. P., van der Woude, H.-J., *et al* (2019). Tennis elbow. *Shoulder & Elbow.*, *11*(5), 384–392. https://doi.org/10.1177/1758573218797973
23. Khowaja, S.A., Khuwaja, P., Dev, K., Wang, W., Nkenyereye, L. (2024). ChatGPT needs SPADE (Sustainability, privacy, digital divide and ethics) evaluation: A review. *Cognitive computation.,* 16: 2528-2550. Doi: https://link.springer.com/article/10.1007/s12559-024-10285-1#Abs1
24. Kisekka, V., & Giboney, J. S. (2018). The Effectiveness of Health Care Information Technologies: Evaluation of Trust, Security Beliefs, and Privacy as Determinants of Health Care Outcomes. *Journal of medical Internet research.*, *20*(4). Doi:https://doi.org/10.2196/jmir.9014
25. Lee, J. T., Li, V. C.-S., Wu, J.-J., Chen, H.-H., Su, S. S.-Y., *et al*. (2025). Evaluation of performance of generative large language models for stroke care. *Digital Medicine.*, 8(481). Doi: https://doi.org/10.1038/s41746-025-01830-9
26. Lima, J. P., Tangamornsuksan, W., Guyatt, G. H. (2023). Trustworthy evidence-based versus untrustworthy guidelines: detecting the difference. *Family medicine and community health.*, 11(4). Doi: https://doi.org/10.1136/fmch-2023-002437
27. Lucado, A. M., Day, J. M., Vincent, J. I., MacDermid, J. C., Fedorczyk, J., *et al.* (2022). Lateral Elbow Pain and Muscle Function Impairments. *Journal of ortopeaedic & sports physical therapy*., 52(12): CPG1-CPG111. Doi: https://doi.org/10.2519/jospt.2022.0302
28. Ma, K., Wang, H. (2020). Management of Lateral Epicondylitis: A Narrative Literature Review. *Pain Research & Management.*, 2020: 9. https://doi.org/10.1155/2020/6965381.

29. Meyer, M. K. R., Kandathil, C. K., Davis, S. J., Durairaj, K. K., Patel, P. N., *et al.* (2024). Evaluation of Rhinoplasty Information from ChatGPT, Gemini, and Claude for Readability and Accuracy. *Aesthetic Plastic Surgery.*, 49(7): 1868–1873. Doi: https://doi.org/10.1007/s00266-024-04343-0
30. Nasa, P., Jain, R., Juneja, D. (2021). Delphi methodology in healthcare research: How to decide its appropriateness. *World Journal of Methodology.*, 11(4): 116–129. Doi: https://doi.org/10.5662/wjm.v11.i4.116
31. OpenAI. (2026, January). *AI as a healthcare ally: How Americans are navigating the system with ChatGPT*. OpenAI. https://openai.com/
32. Raja, A., Bin Amin, S., Azeem, B., Raja, S., Aftab, Y., *et al*. (2024). Self-diagnosis and self-medication based on internet search among Non-Medical University students of Karachi. *Annals of medicine and surgery)*.,86(11):6507-6513. doi: 10.1097/MS9.0000000000002605.
33. Ralla, B., Biernath, N., Lichy, I., Kurz, L., Friedersdorff, F., *et al.* (2025). How Accurate Is AI? A Critical Evaluation of Commonly Used Large Language Models in Responding to Patient Concerns About Incidental Kidney Tumors. *Journal of clinical medicine.*, 14(16): 5697. Doi: https://doi.org/10.3390/jcm14165697
34. Reis, M., Reis, F., & Kunde, W. (2024). Influence of believed AI involvement on the perception of digital medical advice. *Nature medicine.*, 30(11): 3098–3100. Doi:https://doi.org/10.1038/s41591-024-03180-7
35. Rooney, M. K., Santiago, G., Perni, S., Horowitz, D. P., McCall, A. R., *et al* (2021). Readability of Patient Education Materials From High-Impact Medical Journals: A 20-Year Analysis. *Journal of Patient Experience*., 8. Doi: https://doi.org/10.1177/2374373521998847
36. Saenger, T. R., Hinck, M,, Grimmer, J., Stewart, B. M. (2024). AutoPersuade: A Framework for Evaluating and Explaining Persuasive Arguments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*., 16325–16342. Doi: https://aclanthology.org/2024.emnlp-main.913.pdf
37. Sallam M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel)*., 11(6):887. Doi: https://doi.org/10.3390/healthcare11060887
38. Santiago, A. O., Rios-Russo, J. L., Baerga, L., Micheo, W. (2021). Evidenced-Based Management of Tennis Elbow. *Current Physical Medicine and Rehabilitation Reports.*, 9(4): 186–194. Doi: https://doi.org/10.1007/s40141-021-00322-7

39. Scaff, S. P. S., Reis, F. J. J., Ferreira, G. E., Jacob, M. F., Saragiotto, B. T. (2025). Assessing the performance of AI chatbots in answering patients' common questions about low back pain. *Annals of the Rheumatic Diseases*,. 84(1): 143–149.Doi: https://doi.org/10.1136/ard-2024-226202
40. Seurer, A. C., & Vogt, H. B. (2013). Low health literacy: a barrier to effective patient care. *South Dakota medicine : the journal of the South Dakota State Medical Association*., *66*(2): 51–57. Doi: https://pubmed.ncbi.nlm.nih.gov/23513359/
41. Sevgin, D. R., Tarihci Cakmak, E., Yildirim Ogras, G., Diracoglu, D. (2026). Can AI chatbots guide patients and physicians about neck pain? A reliability and readability comparison of ChatGPT-4 and Gemini. *Journal of Back and Musculoskeletal Rehabilitation.*, Doi: https://doi.org/10.1177/10538127261433272
42. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., *et al.* (2023). The Curse of Recursion: Training on Generated Data Makes Models Forget. *Anthropic.,* Doi: https://arxiv.org/abs/2305.17493
43. Singh, H. P., Watts, A. C. (2023). BESS patient care pathway: Tennis elbow. *Shoulder & Elbow.*, 15(4): 348–359. Doi: https://doi.org/10.1177/17585732231170793
44. Singh, S. (2025). ChatGPT Statistics (2025): DAU & MAU Data Worldwide. *DemandSage*. https://www.demandsage.com/chatgpt-statistics/
45. Spadaro, D.C., Robinson, L.A., Smith, L.T. (1980). Assessing readability of patient information materials., *American society of hospital pharmacists.* 37: 215-222. Doi: https://pubmed.ncbi.nlm.nih.gov/7361793/
46. Takita, H., Kabata, D., Walston, S. L., Tatekawa, H., Saito, K., *et al*. (2025). A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. *NPJ digital medicine.*, *8*(1), 175. https://doi.org/10.1038/s41746-025-01543-z
47. The Lancet Digital Health. (2023). ChatGPT: Friend or foe? *The Lancet Digital Health*. *5*(3). https://doi.org/10.1016/S2589-7500(23)00023-7
48. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., *et al* (2023). Large language models in medicine. *Nature Medicine.*, *29*(8), 1930–1940. Doi: https://doi.org/10.1038/s41591-023-02448-8
49. van der Vlist, A.C., Breda, S.J., Oei, E.H.G., Verhaar, J.A.N., de Vos, R.J. (2019). Clinical risk factors for Achilles tendinopathy: a systematic review. *British journal of sports medicine*., 53(21):1352-1361. doi: 10.1136/bjsports-2018-099991.

50. Walker, H. L., Ghani, S., Kuemmerli, C., Nebiker, C. A., Müller, P.B., *et al* (2023). Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *Journal of medical internet research.,* 25. Doi: https://doi.org/10.2196/47479
51. Wang, X., Cohen, R.A. (2022). Health information technology use among adults: united states, july-december 2022. *NCHS data brief.* 482. Doi: https://www.cdc.gov/nchs/data/databriefs/db482.pdf
52. Wei, Q., Yao, Z., Cui, Y., Wei, B., Jin, Z., *et al.* (2024). Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. *Journal of biomedical informatics.*, 151 (104620). Doi: https://doi.org/10.1016/j.jbi.2024.104620
53. Wirth, P. J., Warden, A. M., Moura, S. P., Attaluri, P. K., Larson, J. D. (2025). Readability of Patient Education Materials in Plastic Surgery: Assessing 14 Years of Progress. *Plastic and reconstructive surgery. Global open*., 13 (2). Doi: https://doi.org/10.1097/GOX.0000000000006541
54. Wolf, M. S., Gazmararian, J. A., Baker, D. W. (2005). Health literacy and functional health status among older adults. *Archives of internal medicine.*, *165*(17), 1946–1952. https://doi.org/10.1001/archinte.165.17.1946
55. Yau, J. Y.-S., Saadat, S., Hsu, E., Murphy, L. S.-L., Roh, J. S., *et al.* (2024). Accuracy of Prospective Assessments of 4 Large Language Model Chatbot Responses to Patient Questions About Emergency Care: Experimental Comparative Study. *Journal of Medical Internet Research.*, 26. Doi: https://doi.org/10.2196/60291
56. Yeo, Y. H., Samaan, J. S., Ng, W. H., Ting, P. S., Trivedi, H., *et al.* (2023). Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clinical and molecular hepatology.*, *29*(3): 721–732. Doi: https://doi.org/10.3350/cmh.2023.0089
57. Yun, H. S., Bickmore, T. (2025). Online Health Information–Seeking in the Era of Large Language Models: Cross-Sectional Web-Based Survey Study. *Journal of medical internet research.,* 27. Doi: https://doi.org/10.2196/68560
58. Zhang, X. (2025). More People Are Risking Medical Advice From Chatbots. Here's Why. *ScienceAlert*. https://www.sciencealert.com/more-people-are-risking-medical-advice-from-chatbots-heres-why

59. Zielinski, C., Winker, M.A., Aggarwal, R. (2023) Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. *Current medical research & opinion.,* 40 (1): 1-3. Doi: https://www.researchgate.net/publication/376683301_Chatbots_generative_AI_and_scholarly_manuscripts_WAME_recommendations_on_chatbots_and_generative_artificial_intelligence_in_relation_to_scholarly_publications

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Siim Orgvee,
1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "AI-vestlusrobotite kvaliteedi hindamine vastamisel patsientide enim levinud küsimustele lateraalse epikondüliidi", mille juhendaja on Martin Argus, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Siim Orgvee
18.05.2026