

TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Artur Kašnikov
Inim- ja tehisaru hinnangute võrdlus kursuse
„Veebilehtede loomine edasijõudnutele“
lõpuprojektide näitel
Bakalaureusetöö (9 EAP)

Juhendaja: Lidia Feklistova, PhD

Tartu 2025

Inim- ja tehisaru hinnangute võrdlus kursuse “Veebilehtede loomine edasijõudnutele” lõpuprojektide näitel

Lühikokkuvõte:

Käesoleva bakalaureusetöö raames analüüsiti Tartu Ülikoolis toimuva kursuse “Veebilehtede loomine edasijõudnutele” (LTAT.03.015) lõpuprojektidena valminud veebisaite ja selgitati välja erinevate keelemudelite kasutamissobivust nende hindamisel. Hinnangu andmisel kasutati teaduskirjanduse põhjal paika pandud hindamiskriteeriumeid. Töö teoreetilises osas on kirjeldatud kursuse ülesehitus, kasutajaliidese ja kasutajakogemuse disaini ning koodi olulisus veebiarenduses, ning tehisaru. Praktilise osa esimene etapp sisaldas bakalaureusetöö autori hinnangut valminud veebisaitide prototüüpidele ja koodile. Kui kasutajakogemuse ja kasutajaliidese disainiga võib jääda rahule, siis koodi kvaliteedi pärast tuleb tunda rohkem muret. Teise etapina uuriti erinevaid keelemudeleid ja nende võimalusi seisuga 20.02.2025, et hinnata kasutajaliidese ja kasutajakogemuse disaini kuvatõmmiste alusel. Selgus, et Deepseek ja Copilot ei sobinud selleks, kuna nendel mudelitel puudus kuvatõmmiste hindamise võimalus. Viimane praktilise osa etapp keskendus autori ja tehisaru keelemudelite hinnangute võrdlemisele. Uurimise tulemused näitasid, et Gemini 2.0 Flash ja Gemini 2.5 Pro keelemudelite poolt antud hinnangud erinevates hindamiskriteeriumites ei erinenud oluliselt autori omadest. Kui hindamisel tuleb keskenduda ainult kasutajaliidese disainile, siis lisaks eespool mainitud keelemudelitele ChatGPT 4.0 ja ChatGPT 4.5 hinnangud ei erinenud oluliselt autori omast. Kasutatavuse analüüsi puhul enamiku keelemudelite (välja arvatud ChatGPT 4.5) poolt antud hinnang ei erinenud autori omast. Keelemudelite koodi kvaliteedi hinnangud varieerusid.

Võtmesõnad: Veebisait, tehisaru keelemudelid, kasutajaliidese disain, kasutajakogemuse disain, koodi kvaliteet

CERCS: P175 Informaatika, P176 Tehisintellekt

Comparison of human and artificial intelligence evaluations based on the course “Web Page Creation for Advanced Users” final projects

Abstract:

This bachelor's thesis aims to analyze the websites created as final projects of the University of Tartu course “Web Page Creation for Advanced Users” and to determine whether different language models are suitable for their evaluation. The evaluation was based on the criteria, made up of scientific literature. The theoretical part describes the course build, the importance of user interface, user experience, code writing in web development and gives an overview of artificial intelligence. The first step of the practical part gives the author's evaluation of the created prototypes of websites and the code of the final projects. While user interface and user experience can be considered satisfied, the quality of code is a bit more of a concern. The second step includes an investigation of different language models and their capabilities as of 20.02.2025 to evaluate the design based on screenshots. It turned out that DeepSeek and Copilot are not suitable for further investigation because of no possibility in evaluating screenshots. The final step of the practical part considers comparison of author and language models evaluations. The study results showed that evaluations by Gemini 2.0 Flash and Gemini 2.5 Pro models did not significantly differ from the author's ones. However, if the evaluation focused only on the user interface design, results by ChatGPT 4.5 and ChatGPT 4o, in addition to previously mentioned language models, did not significantly differ from the author's results. In the usability analysis, the ratings given by most language models, except for ChatGPT 4.5, statistically did not differ from the author ones. The code quality evaluations by language models varied.

Keywords: Web design, artificial intelligence language models, user interface, user experience, code quality

CERCS: P175 Informatics, P176 Artificial intelligence

Sisukord

Sissejuhatus.....	5
1. Kursus “Veebilehtede loomine edasijõudnutele”	7
2. Kasutajaliidese ja kasutajakogemuse disaini põhimõtted	8
2.1 Kasutajaliides ja selle olulisus veebiarenduses.....	8
2.2 Kasutajakogemus ja selle olulisus veebiarenduses	10
2.3 Koodi olulisus veebiarenduses.....	11
3. Tehisaru ja keelemudelid	13
4. Metoodika	14
4.1 Valim.....	14
4.2 Andmekogumine ja andmeanalüüs	15
5. Tulemused.....	19
5.1 Autori hinnang loodud veebisaitidele	19
5.2 Tehisaru keelemudelid, mis on võimelised hinnata veebisaite	23
5.3 Autori ja keelemudelite poolt antud hinnangute võrdlus	27
6. Arutelu	36
7. Piirangud ja võimalikud uurimissuunad	40
Kokkuvõte.....	41
Viited.....	42
Lisad.....	50
I. Litsents.....	50

Sissejuhatus

Tänapäeva digitaalses keskkonnas on enamikul ettevõtetel olemas oma veebisait. Mitmed uuringud on näidanud, et veebisait täidab olulist rolli ettevõtte nähtavuse, usaldusväarsuse ja kättesaadavuse kujundamisel [1, 2]. Hästi disainitud veebisait ei ole pelgalt visuaalselt atraktiivne, vaid toetab ka kasutajate eesmärkide kiiret ja mugavat saavutamist [3]. Kasutajaliidese ja kasutajakogemuse disain on veebiarenduses võtmetähtsusega, aidates parandada veebisaidi navigeeritavust, ligipääsetavust ning kasutajate rahulolu [4].

Tartu Ülikooli kursusel „Veebilehtede loomine edasijõudnutele” (LTAT.03.015) omandavad üliõpilased teadmisi ja oskusi kaasaegsete veebisaitide loomiseks. Aine eduka lõpetamise üheks tingimuseks on lõpuprojekti koostamine. See eeldab koostatud prototüübi alusel veebisaidi loomist, kasutades kursusel käsitletud tehnoloogiaid. Kuna tehisaru (ingl artificial intelligence, AI) kasutamine erinevates valdkondades kogub hoogu [5, 6, 7], siis tõstatub küsimus, kas tehisaru saab rakendada erinevate lõpuprojekti hindamisele. See aitab aru saada, kas tehisaru saab kasutada õppejõu hindamise abivahendina.

Käesoleva bakalaureusetöö eesmärk on analüüsida lõpuprojektidena valminud veebisaitide ja välja selgitada erinevate keelemudelite kasutamissobivust nende hindamisel. Eesmärgist lähtuvalt püstitas autor järgmised uurimisküsimused:

1. Millist hinnangut loodud veebisaitidele annab bakalaureusetöö autor?
2. Millised tehisaru keelemudelid on võimelised hindama loodud veebisaitide?
3. Kas autori ja keelemudelite poolt antud hinnangutes valminud veebisaitidele esineb statistilisi erinevusi, kui kasutada samu hindamiskriteeriume?

Eesmärgi täitmiseks analüüsitakse 2024. aasta sügissemestril kursuse „Veebilehtede loomine edasijõudnutele“ raames lõpuprojektidena loodud veebisaitide prototüüpe ja koodi.

Töö on jaotatud seitsmeks peatükiks. Esimeses peatükis antakse ülevaade kursusest ja selle korraldusest. Teine peatükk kirjeldab hindamiskriteeriume, mis on aluseks autori ja keelemudelite hinnangute andmisel. Kolmas peatükk annab ülevaadet tehisarust. Neljandas peatükis kirjeldatakse meetodikat, viiendas peatükis tutvustatakse uurimistöö tulemusi ja kuendas peatükis arutatakse nende üle. Seitsmendas peatükis tuuakse välja töö piirangud ja tuleviku uurimissuunad.

Bakalaureusetöös on kasutatud koostatud teksti sõnastuse parandamiseks ja kasutatud allikate tõlkeks tehisaru mudelit ChatGPT-4o. Analüüsitavate lõpuprojektide hinnangute andmiseks on kasutatud erinevad keelemudelid.

1. Kursus “Veebilehede loomine edasijõudnutele”

Selles peatükis antakse ülevaade kursusest “Veebilehede loomine edasijõudnutele”. Tuuakse välja kursuse eesmärgid ning kirjeldatakse kursuse korraldus.

Tartu Ülikooli Õppeinfosüsteemi andmetel [8] kursus “Veebilehede loomine edasijõudnutele” (LTAT.03.015) on mõeldud üliõpilastele, kellel on elementaarne programmeerimisoskus või veebilehe arendamise kogemus. Kursuse põhieesmärgiks on anda üliõpilastele vajalikud põhiteadmised veebilehede disainimiseks ja loomiseks, ning tõsta veebiarenduse alaseid oskusi veebisaitide eessüsteemide (ingl *front-end*) loomisel. Ainemaht on 78 tundi, mis vastab 3 EAP’le. Tegemist on 100% veebipõhise kursusega ja kõik õppematerjalid on üliõpilastele kättesaadavad Moodle õpikeskkonnas [9]. Kursusel tutvustatakse kasutajaliidese (ingl *User Interface*, UI) ja kasutajakogemuse (ingl *User Experience*, UX) disaini põhiprintsiibid. Läbi praktilise ülesande üliõpilased arendavad HTML, CSS ja JavaScript tehnoloogiate kasutamise oskusi. Lisaks käsitletakse kursusel raamistikke Bootstrap ja Vue.js. Kursuse lõpus antakse ülevaade tehisaru kasutamise võimalustest veebiarenduses.

Kursus on mitmeeristava hindamisega. Üliõpilane saab kursusel “arvestatud” juhul, kui kõik individuaalsed praktilised ülesanded ja lõpuprojektina valmiv veebisait vastavad vähemalt minimaalsetele esitatud nõuetele. Lõpuprojekti eesmärk on demonstreerida üliõpilaste oskusi rakendada veebisaitide loomisel kasutajaliidese ja kasutajakogemuse disaini põhiprintsiipe ning kursusel õpitud veebiarenduse tehnoloogiaid. Lõpuprojekti elluviimine koosneb mitmest etapist. Alguses moodustavad üliõpilased kahe- või kolmeliikmelised meeskonnad ja valivad teema. Seejärel loob iga meeskond disainiplatvormil (näiteks Figma) veebisaidi prototüübi, kus arvestab kasutajaliidese ja kasutajakogemuse disaini põhiprintsiipe. Veebisaidi prototüüp peab koosnema vähemalt kolmest veebilehest. Link valminud prototüübile postitatakse foorumisse. Teine meeskond loob, järgides veebisaiti prototüüpi, veebisaidi, kasutades kursusel õpitud veebiarenduse tehnoloogiaid. Lõpetuseks annavad prototüübi autorid üldist tagasiside valminud veebisaidile. Lõpuprojekt loetakse arvestatuks, kui on täidetud kõik selle etapid.

2. Kasutajaliidese ja kasutajakogemuse disaini põhimõtted

Käesolevas peatükis esitatakse ülevaade kasutajaliidese ja kasutajakogemuse disaini teoreetilistest põhimõtetest ja vajadusest. Tuuakse välja edaspidises analüüsis kasutatavad kriteeriumid ja põhjendatakse nende valikud.

Veebidisain on interdistsiplinaarne ja dünaamiline valdkond, kus põimuvad visuaalne esteetika, strateegiline kasutajakesksus ja tehniline teostus [10]. Kui esteetilise kujunduse loomisega tegeleb kasutajaliidese disain, siis kasutaja rahulolu veebisaidi kasutamise eest vastutab kasutajakogemuse disain. Tegemist on omavahel sõltuva disainivaldkondadega, mille koostöös kujuneb kasutaja elamus veebisaidi kasutamisel. Nende vaheline sõltuvus tähendab, et puudujäägid ühes aspektis võivad mõjutada teise valdkonna tajutavat kvaliteeti.

2.1 Kasutajaliides ja selle olulisus veebiarenduses

Kasutajaliides on vahend, mis võimaldab sujuvat interaktsiooni inimese ja arvuti vahel (ingl *human-computer interaction*) [11]. Selle disain keskendub visuaalsete elementide paigutusele, välimusele, stiilile ja esteetilisele lahendustele [12-15]. Elementide välimus ja paigutus tagavad veebisaidi visuaalse järjepidevuse ja atraktiivsuse, pakkudes kasutajatele meeldivat visuaalset kogemust ning hõlbustades nende suhtlemist rakendusega, sh veebisaidiga [16]. Rakenduse esteetiline lahendus avaldab tugevat mõju kasutaja rahulolule [17]. Läbimõtlemata kasutajaliides võib viia ebameeldiva interaktsioonini ja kasutaja vastumeelsuse tekkimiseni [11].

Arvestades kasutajaliidese olulisust, on kujundamisel oluline keskenduda elementidele, millega kasutajad otseselt suhtlevad. Nendeks võivad olla näiteks sisulahtrid, nupud, menüüd või ikoonid. Edasi toob bakalaureusetöö autor välja tema hinnangul kõige olulisemad kasutajaliidese disaini aspektid, mida soovitatakse järgida veebisaidi kujundamisel. Lisaks põhjendatakse tehtud valik.

Värv (ingl *color*) on nähtava valguse omadus, mis sõltub valguse lainepikkusest ja objekti omadustest valgust peegeldada või neelduda [18]. Kasutajaliidese disaini kontekstis värv on vahend, mis lubab luua visuaalset hierarhiat [19] ja võimaldab kommunikatsiooni veebisaidi disaineri ja kasutaja vahel [20]. Disainerid valivad värve vastavalt eesmärkidele - näiteks, tegutsemise kutse (ingl *Call to Action*, CTA) nupu puhul kasutatakse punast värvi [20].

Hästi valitud värvide kombinatsioonid aitavad parandada sisu loetavust ja seeläbi parandada kasutajakogemust. Näiteks rohelise-musta kombinatsioonil on parem loetavus kui virsiku-helehalli kombinatsioonil [21]. Liiga erksad või üksteisega dissonantsis olevad värvitoonid võivad põhjustada visuaalset ebamugavust, vähendada loetavust või tekitada segadust kasutajaliideses, muutes rakenduse kasutamise pingeliseks ja tähelepanu hajutavaks [19].

Veebidisainis tähendab **värvikontrastsus** (ingl *contrast*) kasutajaliidese elementide ja tausta erinevuse suhet [22]. Veebisisu hõlbustuse juhiste järgi (ingl *Web Content Accessibility Guidelines*) kontrastsuse suhe tavalise teksti ja tausta vahel peab olema vähemalt 4.5:1 [5]. Hea kontrastsuse kasutamine aitab juhtida veebisaidi külastaja tähelepanu olulistele elementidele, muutes navigeerimise lihtsamaks ja veebisaidi atraktiivsemaks [23]. Lisaks toetab hea kontrastsus veebisaidi sisu loetavust ja eristumist taustast, muutes selle hõlpsamini jälgitavaks ka nõrga nägemise või värvipimedusega kasutajatele [24, 25]. Samas tuleb arvestada, et liiga kõrge kontrastsus võib mõjuda häirivalt, tekitades ebamugavust ja vähendades veebisaidi sisu nähtavust ja loetavust [23].

Veebiarenduses **rühmitamine** (ingl *grouping*) tähendab semantiliste, loogiliselt seotud ja sarnaste kasutajaliidese elementide ühendamist gruppidesse [26]. Rühmitamine põhineb ühel gestaltpsühholoogia läheduse printsiibil (ingl *proximity principle*). Selle printsiibi järgi tunduvad lähestikku asuvad elemendid rohkem seotud (rühmitatud) kui üksteisest kaugemal asuvad elemendid [27, 28].

Rühmitamise eesmärk on lisada kasutajaliidesele loogilist struktuuri [26], et anda nii veebisaidi kasutajatele kui ka arendajatele selgust, kuidas erinevad komponendid või funktsionaalsused üksteisega seostuvad [29]. Selle saavutamiseks võivad sarnased komponendid olla koondatud visuaalseteks “kaartideks” [26] või rühmitatud, kasutades visuaalset hierarhiat. Kui elemendid on rühmitatud efektiivselt, siis muutub veebisait selgemaks, loogilisemaks, paremini orienteeruvaks ja see vähendab oluliselt otsingu aega [26, 30, 31]. Halvasti läbimõeldud rühmitamine võib aga tekitada segadust: kui sarnase tähendusega elemendid asuvad üksteisest kaugel või kattuvad visuaalselt, siis kulub kasutajal rohkem aega vajaliku info leidmiseks [26].

Tüpoograafia (ingl *typography*) on disainielement, mis tegeleb tekstilise info kujunduse, esitlemise ja paigutusega [32, 33]. Selle eesmärk on tagada teksti loetavust ja aidata sõnumit edasi anda võimalikult efektiivselt [19, 33, 34].

Hea tüpograafia kasutamine muudab teksti kergesti loetavaks, jälgitavaks, annab aimu visuaalsest hierarhiast, teeb veebisaitte meeldivaks tajuda ja atraktiivsemaks [33]. Uuringud [35] on näidanud, et kui rakenduses on kasutatud sinist kirja värvi, kollast taustavärvi ja Arial kirjafonti, siis selline valik võimaldab kiiremat sisu lugemist. Seevastu, kui on kasutatud musta kirja värvi, valget taustavärvi ja Arial kirjafonti, siis kasutajate lugemiskiirus langeb. Kirjafondi suuruse uuringust [36] selgus, et 8-punktiline kirjasuurus põhjustas ebameeldivaid kogemusi ja suurima hulga vigu sisu lugemisel. Keskmised ja suured suurused, nagu 12 ja 16 punkti, seostusid peamiselt meeldivate ja mugavate kogemustega. Sisu paigutuse hierarhiale pühendatud uuringus [37] leiti, et kõrgema taseme tekst (pealkiri) peab olema suurem, et visuaalne hierarhia oleks selgelt tajutav ja pealkiri oleks eristatav põhitekstist. Halvasti valitud tüpograafia puhul võib veebisaidil olev tekst muutuda raskesti loetavaks ja üleliigselt koormavaks [19, 33].

Kissitustest (ingl *Squint Test*) on tõhus mitteformaalne meetod kasutajaliidese visuaalse hierarhia hindamiseks [38, 39]. See oluline vahend annab ülevaadet disaini visuaalsest tasakaalust ja aitab mõista, milliseid elemente tajutakse kõige esmalt [40]. Testi käigus kontrollitakse, kas tausta elemendid on paigutatud korrektselt. See tähendab, et need ei häiri põhilist visuaalset hierarhiat, esiplaani elemendid asetsevad sobivalt tausta suhtes ning peamine fookus on suunatud kõige pilkupüüdvamatele elementidele. Kissitustest aitab välja selgitada, kuivõrd efektiivselt suudab kasutajaliidese disain juhatada kasutaja tähelepanu sisulistele komponentidele ja kuivõrd olulisemad kasutajaliidese komponendid eristuvad selgelt ülejäänud elementidest [38].

2.2 Kasutajakogemus ja selle olulisus veebiarenduses

Kasutajaliidese disain on osa kasutajakogemuse disainist [41] ja veebisaidi kasutajaliidese visuaalne lahendus mõjub kasutajakogemust [15]. Kasutajakogemus on elamus ja rahulolu, mida inimene saab tarkvara kasutamisel [42], sh veebisaidi külastamisel. Kasutajakogemuse disain keskendub kasutajale - kuidas ta tajub välja töötatud lahendust [37, 43], kuivõrd atraktiivne see on ning milliseid emotsioone see temas tekitab [43]. Nende aspektide arvestamine võimaldab luua kasutajasõbralikku tarkvara ja mõjutab otseselt kasutaja lojaalsust ja valmisolekut tarkvara uuesti kasutada või seda teistele soovitada [11]. Halvasti läbimõeldud kasutajakogemus võib viia kasutajate rahulolematuse, selle kasutamisest loobumise ning negatiivsete hinnanguteni [43].

Kuigi kasutajakogemusel on mitu erinevat komponenti [44], siis antud töös autor keskendub ainult kasutatavusele. Teiste kasutajakogemuse komponentide (nt esteetika, emotsioonid, usaldusväärsus) objektiivne hindamine nõuaks kas ulatuslikku kasutajauuringut või eritööriista kasutamist. See aga jääb bakalaureusetöö skoobist välja.

Kasutatavus (ingl *usability*) on kasutajakogemuse põhiomadus, mis iseloomustab, mil määral saavad kasutajad veebisaiti tõhusalt ja efektiivselt kasutada [45]. Läbimõeldud elementide paigutus ning selge ja loogiline info paigutus võimaldab kasutajatel sujuvalt veebisaidil navigeerida [46]. Hea kasutatavus avaldab positiivset mõju naudingule, kasutajad tajuvad selle veebisaidi väärtust enda jaoks ja soovivad ikka ja jälle külastada seda ning on valmis soovitama veebisaiti ka teistele [25]. Hõlpsasti kasutatav ja hästi navigeeriv veebisait meelitab ja hoiab tõenäolisemalt rohkem kasutajaid, mis omakorda on kasulik rahalisest seisukohast [47]. Kui kasutatavuse lahendus on puudulik, muutub veebisait raskesti navigeeritavaks, vähem intuitiivseks ning keeruliselt mõistetavaks, mille tagajärjeks on suurem vigade sagedus, motivatsiooni langus ning võimalik loobumine veebisaidi külastamisest [42].

2.3 Koodi olulisus veebiarenduses

Veebisait luuakse koodi abil, kasutades erinevaid veebiarenduse tehnoloogiaid nagu HTML, CSS, JavaScript jne. Veebisaidi töökorras hoidmine nõuab pidevat tööd koodiga. See muuhulgas eeldab, et kood on loetav, hooldatav ja dokumenteeritud. Järgmisena põhjendab bakalaureusetöö autor nende omaduste olulisust veebiarenduses.

Koodi **loetavus** (ingl *readability*) on tarkvaraarenduses fundamentaalne omadus, mida peetakse tarkvara kvaliteedi oluliseks komponendiks, sest suurem osa tarkvara kirjutamisest kulub lähtekoodi mõistmisele ja hooldamisele [48, 49]. Koodi loetavus näitab, kui kergesti on võimalik koodi lugeda, sellest aru saada ning mugavalt koodi hallata [50, 51]. See koodi omadus katab mitu erinevat aspekti nagu näiteks koodi koondstruktuuri selgust või muutujate ja klasside arusaadavaid nimetusi. Loetav kood vähendab vigade tekkimise tõenäosust ja aitab tuvastada veebisaidi loomisel kirjutatud koodis olemasolevaid vigu [52].

Hooldatavus (ingl *maintability*) on koodi kvaliteedi oluline komponent, mis iseloomustab, kui lihtsalt ja kuluefektiivselt saab loodud tarkvaralist lahendust muuta, parandada, täiustada või kohandada ilma, et tekiks negatiivne mõju loodud funktsionaalsusele [53, 54]. Kui lähtekood on

kõrge hooldatavusega, kulub tarkvarahooldajatel vähem aega ja ressursse vigade parandamiseks, uute funktsioonide lisamiseks või olemasolevate lahenduste kohandamiseks muutuvatele nõudmistele [56].

Koodi **dokumentatsioon** (ingl *code documentation*) on kirjalik selgitus tarkvarakoodi funktsionaalsusest, arhitektuurist, struktuurist ja teostusest [55, 56]. Kuigi dokumentatsiooni loomine ja uuendamine on kulukas ja aeganõudev protsess [55, 57], aitab see tarkvaraarendajatel mõista, kuidas kood töötab. Hästi koostatud dokumentatsioon kirjeldab koodi selgelt ja konkreetselt ning aitab parandada koodi loetavust [58]. Vaatamata sellele, et on olemas JavaDoc, Doxygen, NDoc ja teised vahendid automaatse dokumentatsiooni loomiseks, eeldavad nad kõik, et tarkvaraarendajad kommenteeriksid lähtekoodi [57].

Kokkuvõtvalt võib öelda, et atraktiivsema ja kasutajasõbraliku veebisaidi loomisel on vaja jälgida kasutajaliidese ja kasutajakogemuse disaini põhiprintsiipe. Parema tulemuse saamiseks on vaja pöörata tähelepanu erinevate elementide kujundamisele ja paigutamisele. Ka koodi kvaliteet mängib olulist rolli veebisaidi edaspidises arenduses.

3. Tehisaru ja keelemudelid

Selles peatükis kirjeldatakse tehisaru ja keelemudeleid. Tuuakse välja ka tehisaru kasutamise võimalusi erinevates valdkondades.

N.H. Patil jt. [59] sõnul on intellekti (ingl *intelligence*) võimalik määratleda kui võimet omandada teadmisi eesmärgiga lahendada keerukaid probleeme. Tehisaru kirjeldavad nad kui tarkvaraga ühendatud masinate kogumit, mis on loodud mõtlema, suhtlema, õppima ja koguma uut teavet sarnaselt inimestega. Naveed jt. [60] järgi on suured keelemudelid (ingl *Large Language Models*, LLM) tehisaru keelemudelid, mis on treenitud erakordselt suurte tekstikorpuste põhjal ja on õppinud mõistma ja genereerima inimkeelelaadset teksti.

Tehisaru keelemudelite võimalusi rakendatakse erinevates valdkondades. Näiteks meditsiini valdkonnas keelemudelid on suutnud rinnavähi diagnoosimisel saavutada ja mõnel juhul isegi ületada inimeste täpsuse, vähendades nii valepositiivsete kui ka valenegatiivsete juhtumite arvu ning võimaldades vähendada vajadust teise hindaja kaasamiseks kuni 88% juhtudest [61]. Keelemudelid on üritatud rakendada tööstuses eesmärgiga tõsta tootlikkust ja vähendada finantsilist kulu, kuid tuleb rohkem tähelepanu pöörata töötajate koolitusele, andmeturbele ja tehisaru süsteemide integreerimisele olemasoleva infrastruktuuriga [6].

Ka veebidisaini ja -arenduse valdkonnas on tehisaru keelemudelid kujunenud teatud abivahendiks. Keelemudelid on olnud kasutusel veebisaitide loomisel, kuid jõuti järelduseni, et korraliku veebisaidi loomine nõuab inimsekkumist [62, 63]. Lisaks keelemudelitele on esteetika hindamine keeruline, kuna neil puudub võime tajuda abstraktseid esteetilisi väärtusi ja arvestada inimese emotsionaalse ja kultuurilise taustaga [64]. Samas tehisaru keelemudelid rakendades masinõppe meetodeid suudavad analüüsida suures mahus teksti- ja koodikogumeid, pakkudes tuge koodi loomisel, testide koostamisel, vigade tuvastamisel ning kasutajakogemuse täiustamisel [7, 65, 66].

Keelemudelid leiavad rakendamist ka tööde hindamisel hariduses. Hiljutised uuringud [67] näitavad, et suudavad keelemudelid hinnata üliõpilaste kirjutisi sama usaldusväärselt kui inimõppejõud, ületades järjepidevuselt isegi kaasüliõpilaste vastastikust hindamist.

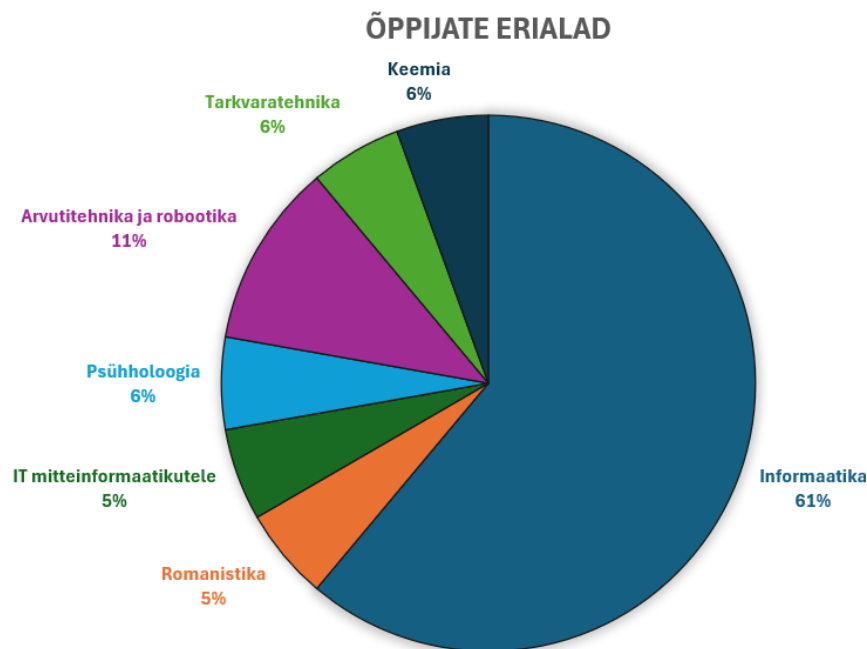
Seega tänapäeval tehisaru ja tema keelemudelid leiavad rakendamist paljudes valdkondades. Samas on olemas kitsaskohad, kus tehisaru veel ei oska tegutseda sama hästi kui inimene.

4. Metoodika

Antud peatükis kirjeldatakse kasutatud valimit. Samuti tuuakse välja andmekogumise ja andmeanalüüsi protsess.

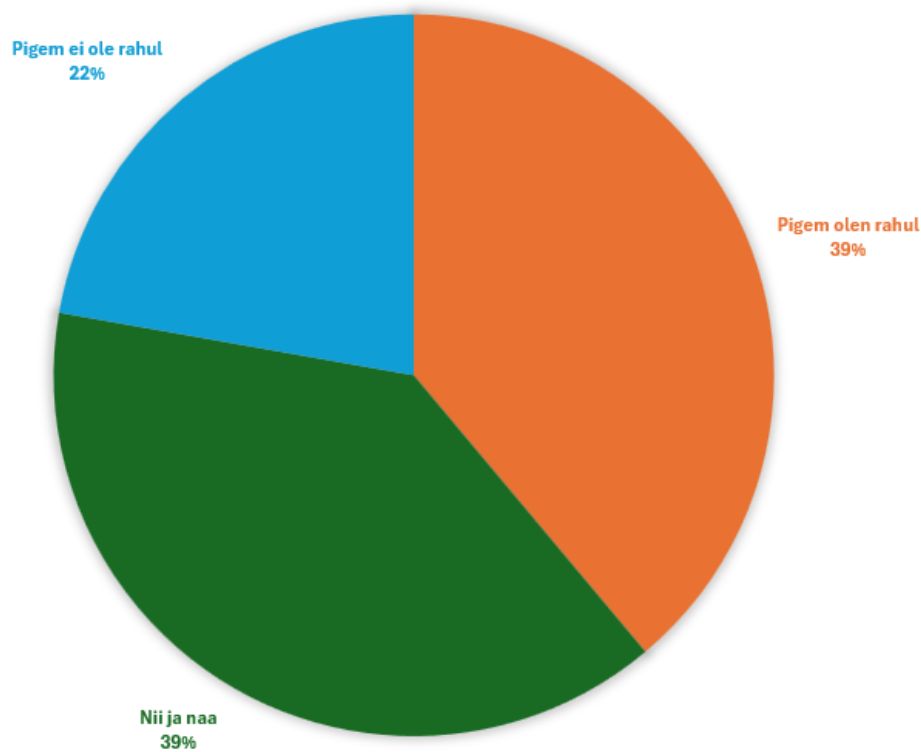
4.1 Valim

Bakalaureusetöös kasutatakse kvantitatiivse uuringu lähenemine. Analüüsitud lõpuprojektid on loodud üliõpilaste poolt, kes 2024. aasta sügisel osalesid kursusel “Veebilehtede loomine edasijõudnutele”. Kuna kolm üliõpilast ei andnud nõusolekut oma lõpuprojektide analüüsimisel, siis lõplikus valimis on 18 üliõpilast, nendest 9 naist ja 9 meest. Kursusel osalesid 13 bakalaureuse taseme õppijat ja 5 magistri taseme õppijat. Valdav enamik osalejatest oli informaatika õppekavast (vt. joonis 1) ja ühel või teisel määral on varem kokku puutunud veebiarendusega (vt. joonis 2).



Joonis 1. Valimis olevate üliõpilaste jaotus erialade alusel.

RAHUOLU VARASEMA VEEBIARENDUSE KOGEMUSEGA



Joonis 2. Rahulolu varasema veebiarenduse kogemusega.

Veebisaitide prototüübid ja nende teostus valmisid rühmatööna. Kokku analüüsis autor üheksa lõpuprojekti.

4.2 Andmekogumine ja andmeanalüüs

Üliõpilaste taustaandmed (näiteks sugu ja eriala) sai töö autor kursuse vastutavalt õppejõult. Andmekogumiseks kasutas töö autor kursusel osalejate koostatud lõpuprojektid. Kuna lõpuprojektid loodi meeskonnatööna, siis kokku analüüsi üheksa lõpuprojekti. Lõpuprojektide hindamise aluseks olid hea kasutajaliidese ja kasutajakogemuse disaini ning kvaliteetse koodi kriteeriumid, mida põhjalikult kirjeldati peatükis 2.

Hindamine toimus kahes etapis. Esimesel etapil autor hindas esitatud lõpuprojektid ise. Teisel etapil kõigepealt uuris autor, millised keelemudelid võivad toime tulla sellelaadsete töödega. Selleks sisestas ta igale valitud keelemudelile järgmise viiba eesti keeles:

„Soovin keelemudelist (keelemudeli nimi) rohkem teada. Palun kirjuta vastused järgmistele küsimustele:

- milleks see keelemudel peamiselt mõeldud on?
- millised on selle keelemudeli üldised tugevused ja nõrkused?
- kas seda keelemudelit saab kasutada veebisaitide hindamisel, kus on oluline roll kasutajaliidese ja kasutajakogemuse disaini printsiipide jälgimisel (näiteks tüpograafial, rühmitamisel, kasutatavusel) ja koodi kvaliteedil? Põhjenda oma vastus.

Piira vastus ühe lõiguga (5–6 lauset). Oma lausetes kasuta umbisikulist tegumoodi, vältides “mina...” vormi. Kirjuta akadeemilises eesti keeles.”

Seejärel palus autor alles jäänud keelemudeleid hinnata lõpuprojekte. Igale töös kasutatud keelemudelile sisestas autor viipe. **Disainiplatvormil loodud** prototüüpide hindamiseks kasutas autor järgmist viipa eesti keeles:

“On antud järgmine ülesanne.

Mina hakkan saatma projekte (mitte full-stack, vaid ainult front-end).

Sina pead hindama neid projekte 2 disaini lähenemise alusel:

1. Kasutajaliides (UI)
2. Kasutajakogemus (UX)

Ning igale disainile on vastavad kriteeriumid

1. Kasutajaliides (UI)
 - a. Värv
 - b. Kontrast
 - c. Rühmitamine
 - d. Tüpograafia
 - e. Kissitustest
2. Kasutajakogemus (UX)
 - a. Kasutatavus

Hindamiseks sina hakkad analüüsima veebilehe kuvatõmmist ning sama kuvatõmmist, aga hägusalt (kissitustesti jaoks). Vastavalt nende kriteeriumitele anna palun tagasiside 5-punkti süsteemis:

- Näide
 - Taustade kombinatsioon näeb välja halvasti = 1 punkt
 - Elemendid ei ole eristatavad = 0 punkti”

Koodi osa hindamiseks sisestas töö autor järgmise viipa eesti keeles:

“On antud järgmine ülesanne.

Mina hakkam saatma projekte (mitte full-stack, vaid ainult front-end). Sina pead hindama projekti ja andma tagasisidet kuidas on kood hästi kirjutatud.

Koodi osas kasuta hindamiskriteeriumid:

- Loetavus
- Hooldatavus
- Koodi dokumentatsioon või kommentaarid

Vastavalt nende kriteeriumitele anna palun tagasiside 5-punkti süsteemis.

- Näided
- Kood on halvasti hooldatav = 1 punkt
- Puudub dokumentatsioon = 0 punkti”

Nii autori kui keelemudelite poolt teostatud hindamisprotsessis kasutati 5-punktilist süsteemi. Kui mingit kriteeriumit ei olnud võimalik hinnata (näiteks puudus dokumentatsioon), siis selle hinnanguks oli 0. Mõlema hindamisetapi tulemused sisestas autor Google Spreadsheets’i tabelisse. Kogutud andmete analüüsimisel kasutas autor programmeerimiskeelt Python-it kasutades teekes pandas, NumPy, re ja SciPy. Andmeanalüüsimisel koostatud kood on leitav Colab keskkonnast.¹

Esmalt kontrolliti kogutud andmeid normaaljaotuse suhtes. Shapiro-Wilki testi tulemused näitasid, et enamus töö autori ja keelemudelite poolt antud hinnangute tulemused ei olnud normaaljaotusega

¹ <https://colab.research.google.com/drive/1kUxqwt6yJj2FYl9rY9GInU6CRSiVffB?usp=sharing>

(vt. tabel 1). Seetõttu tuleb andmete analüüsimisel kasutada mitteparameetrilisi teste [68]. Lisaks mitteparameetriliste testide kasutamist soovitatakse väikese järjestusandmetega valimi puhul [69].

Tabel 1. Shapiro-Wilki testi tulemused.

	Autor W-statistik	ChatGPT 4o W-statistik	ChatGPT o1 W-statistik	ChatGPT4.5 W-statistik	Gemini 2.0 Flash W-statistik	Gimini 2.5 Pro W-statisti	Mistral W-statistik	Claude 3.7 Sonnet W-statistik
Värv	0.711*	0.658***	0.617***	0.825*	0.781**	0.684**	0.655***	0.813*
Kontrast	0.748**	0.780*	0.763**	0.658***	0.390***	0.655***	0.390***	0.823*
Rühmitamine	0.825*	0.760**	0.781*	0.531***	0.693**	0.536***	1.000	0.390***
Kissitustest	0.730**	0.889	0.781*	0.683***	0.655***	0.655***	0.617***	0.838
Tüpoograafia	0.844	0.805*	0.536***	0.536***	0.693**	0.655***	0.655***	0.693**
Kasutatavus	0.767**	0.838	0.655***	0.564***	0.390***	0.655***	1.000	0.655***
Koodi loetavus	0.781*	0.536***	1.000	1.000	0.390***	0.903	1.000	0.655***
Koodi hooldatavus	0.844	0.781**	1.000	1.000	0.390***	0.884	0.655***	0.808*
Koodi dokumentatsioon	0.813*	0.617***	0.390***	0.536***	0.390***	0.844	0.684***	0.913

*p < 0.05; **p < 0.01; ***p < 0.0001

Kuna bakalaureusetöö kolmanda uurimisküsimuse raames võrreldakse samadele lõpuprojektidele antud hinnangud nii autori kui keelemudelite poolt, siis andmete statistiliseks analüüsiks kasutatakse mitteparameetriline Wilcoxon astakmärgitest (ingl *Wilcoxon signed-rank test*) [68].

5. Tulemused

Käesolevas peatükis esitatakse uurimuse tulemused. Tulemused on esitatud vastavalt püstitatud uurimisküsimustele.

5.1 Autori hinnang loodud veebisaitidele

Bakalaureusetöö autori hinnangul võib esitatud lõpuprojektid jagada nelja kategooriasse. Kõige populaarsemateks kategooriateks osutasid teenustega seotud lõpuprojektid. Nendeks olid veebisaidid, mis pakkusid juuksuri-, psühholoogi ja ehitusteenuseid. Hobi kategooria alla kuulusid spordi ja kettagolfiga seotud veebisaidid. Kaks veebisaiti oli infokorralduse kategooriast – nimelt e-raamatukogu ja filmiedastuse platvorm. E-kaubanduse kategooriasse kuulusid veebisaidid, mis keskendusid elektroonikapoele ja autovaruosade turule.

Autor hindas esitatud lõpuprojektid paika pandud hindamiskriteeriumite alusel ise (vt. tabel 2).

Tabel 2. Autori poolt hinnatud projektid hindamiskriteeriumite põhjal

	Keskmine (SD²)	Min	Max	Mediaan
Värv	4.22 (0.97)	3	5	5
Kontrast	3.67 (0.87)	3	5	3
Rühmitamine	4.11 (1.05)	2	5	4
Tüpoograafia	4.00 (1.12)	2	5	4
Kissitustest	4.11 (1.36)	1	5	5
Kasutatavus	4.22 (1.09)	2	5	5
Koodi loetavus	3.89 (0.6)	3	5	4
Koodi hooldatavus	3.00 (1.12)	1	4	3
Koodi dokumentatsioon	1.22 (0.67)	0	2	1

Hindamisel arvestas töö autor veebiarenduses jälgitavaid kasutajaliidese ja kasutajakogemuse disaini põhiprintsiipe ning koodi kvaliteeti. Näiteks kui veebisaidil on ülemäära palju eraldiseisvaid komponente, võib see suurendada hoolduskulu ning muuta arenduse keerulisemaks.

² SD - standardhälve

See omakorda aeglustab arenduse kiirust, millega saab teha disaini- või funktsionaalsuse muudatusi, kuna iga komponendi uuendamine nõuab ükshaaval läbivaatamist. Kasutajaliidese ja kasutajakogemuse disaini enamike kriteeriumide keskmine tulemus oli neli või üle selle (erandiks on kontrast, kus keskmine tulemus oli 3.67).

Arvestades värvi rolli veebisaidi sisu kommunikeerimisel [20] ja sisu loetavuses [19, 21], olid paljudes veebisaitide prototüüpides värvilahendused olid hoolega läbi mõeldud. Näiteks juuksuriteenusele suunatud veebisaidil oli peamiseks taustavärviks valge ja seetõttu oli veebisaidil esitatud informatsiooni lihtsam ja mugavam lugeda. Lisaks seostub kasutatud hele-roosa värvikombinatsioon leebusega ja tähelepaneliku teenindusega [70].

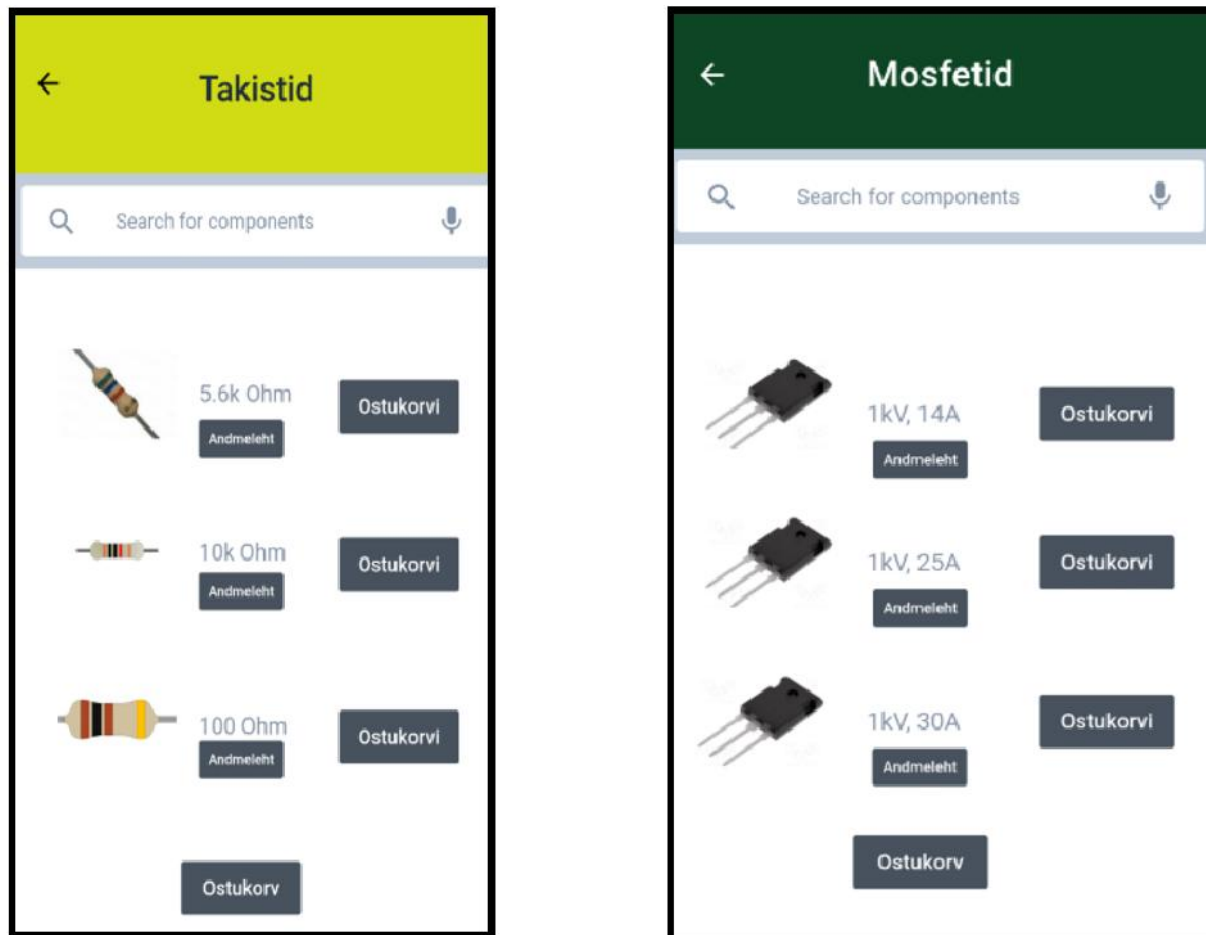
Kuigi hea värvikontrastsus aitab kasutajal veebisaidil navigeerida [23] ja sisuga tutvuda [24, 25], oli hinnatud lõpuprojektide prototüüpides üheks levinumaks veaks kontrastsuse reeglite mitte jälgimine. Esines projekte, kus hele kirjavärv oli heledal taustal või oli kasutatud ebapiisavat kontrasti oluliste elementide esiletõstmisel (vt. joonis 3).



Joonis 3. Kontrastsuse reeglite rikkumine.

Bakalaureusetöö autori hinnangul esitatud lõpuprojektide prototüüpides elementide rühmitamine oli üldiselt hästi teostatud. Enamikul analüüsitud veebilehtedel info oli jagatud loogilistesse plokkidesse, mis olid visuaalselt eristatavad. See muutis veebisaitide struktuuri selgemaks ja loogiliseks, paremini orienteeruvaks, vähendades oluliselt otsingu aega. Seda kõike soovitatakse ka teaduskirjanduses [5, 26, 31]. Samas tuleb tõdeda, et mõnedes prototüüpides ei järgitud

rühmitamise printsiipe. Näiteks elektroonikapoe toodete veebilehel ei olnud iga toode visuaalselt eraldatud, mis raskendas info struktureeritud tajumist (vt. joonis 4).



Joonis 4. Elektroonikapoe toodete leht.

Üliõpilased valisid veebisaitide prototüüpidele enamasti hästi loetavad fondid ning säilitasid ühtlase stiili kogu veebisaidi ulatuses. Selline lähenemine toetas lugemismugavust ja informatsiooni tajumist [33]. Kuigi enamik prototüüpe järgis tüpograafia häid praktikaid, ilmnasid mõnes töös siiski teatud puudujäägid. Näiteks psühholoogia teenuse veebisaidil oli kogu põhitekst esitatud kursiivis (vt. joonis 5). Kuigi tekstiga oli ikkagi võimalik tutvuda, aitaks bakalaureusetöö autori arvamusel stiilide kombineerimine saavutada struktuurset selgust. Näiteks põhitekst võiks olla kirjutatud tavakirjas ja ainult rõhutamist vajavad kohad võiksid olla kursiivis.

Meie ettevõtte on asutatud 2015. aastal ning asub Tallinna südames, pakkudes kvaliteetseid psühholoogilise nõustamise ja teraapiateenuseid. Oleme pühendunud vaimse tervise ja heaolu edendamisele ning usume, et igaühel on õigus saada professionaalset tuge, et elada tasakaalukat ja õnnelikumat elu.

Meie meeskonda kuuluvad kogenud ja kvalifitseeritud psühholoogid ning terapeutid, kes töötavad koos, et pakkuda iga kliendi vajadustele vastavat individuaalset lähenemist. Pakume laias valikus teenuseid, sealhulgas individuaalne ja grupiteraapia, paarinõustamine ning enesearengule suunatud töötoad. Meie eesmärk on luua toetav ja turvaline keskkond, kus iga klient tunneb end kuuldu ja mõistetuna. Oleme uhked, et saame pakkuda kaasaegseid ja tõendus põhiseid teraapiameetodeid, mis aitavad inimestel toime tulla stressi, ärevuse, depressiooni ja suhete probleemidega.

Meie kliendid hindavad meie professionaalsust ja empaatilist lähenemist, mis aitab neil saavutada positiivseid muutusi oma elus. Ootame sind külla meie hubasesse Tallinna keskusesse, et alustada teekonda parema vaimse tervise suunas!



meil@meil.ee
+372 12345678

Facebook 
Twitter 

Joonis 5. Psühholoogia teenuse veebisaidi avaleht

Kissitustesti tulemused näitasid, et enamik lõpuprojekti prototüüpe omasid tugevat visuaalset hierarhiat. Veebisaitidel kasutatud kasutajaliidese olulised komponendid olid eristatavad ülejäänud elementidest. See on veebidisainis kriitilise tähtsusega, kuna toetab kasutaja tähelepanu suunamist ja lihtsustab sisus orienteerumist [38]. Siiski ilmnes mõningaid puudujääke. Näiteks elektroonikapoe veebisaidil ei eristunud olulised tegutsemise kutse nupud piisavalt teistest komponentidest (vt. joonis 4). Seetõttu visuaalne fookus hajus ning kasutaja tähelepanu ei suunatud ootuspäraselt olulistele elementidele nagu “Ostukorv”.

Veebisaitide kasutatavus navigeerimise ja funktsionaalsuse näol olid prototüüpides üldiselt hästi lahendatud. See on väga oluline saavutus, sest hea kasutatavus toetab veebisaidi sujuvat kasutajakogemust [46] ja meelitab kasutajaid [25, 47]. Samas leidsid üksikud veebisaitide prototüübid, nagu näiteks psühholoogiateenusele pühendatud veebisait, kus nuppude abil liikumine veebilehtede vahel oli halvasti realiseeritud. Bakalaureusetöö autori hinnangul tekitab selline lahendus veebisaidi külastajates segadust.

Muret tekitab koodi kvaliteediga seotud kategooria, kus keskmine tulemus jäi alla nelja. Arvestades, kui suurt rolli mängib korralik kood projektide edaspidistes arendustes [50, 51, 48, 52, 53], tuleks sellele aspektile pöörata rohkem tähelepanu. Kriitilisemad hinnangud olid seotud koodi dokumenteerimisega, mille keskmise hinnangu tulemus jäi alla 1.5. Kuigi hästi dokumenteeritud kood aitab parandada selle loetavust [58], analüüsitud lõpuprojektides oli selles osas palju puudujääke.

5.2 Tehisaru keelemudelid, mis on võimelised hinnata veebisaite

Tehisaru keelemudelite võimaluste uurimiseks valis töö autor üheksa keelemudelit:

- ChatGPT 4o, ChatGPT o1, ChatGPT 4.5,
- Gemini 2.0 Flash, Gemini 2.5 Pro,
- Deepseek, Copilot, Claude 3.7 Sonnet, Mistral³.

Nende valik tugines autori isiklikul huvil tutvuda erinevate keelemudelite võimalustega hinnata kursusel valminud veebisaite. Lisaks Shao jt. uurimuses [71] toodi välja, et tänapäeval keelemudelid on võimelised videotöötlemiseks ja piltide mõistmisest. Seega autoril oli huvi vaadata, kuidas keelemudelid saavad hakkama veebisaitide prototüübi ja koodi hindamisega.

Iga keelemudeli kohta on venitatud selle üldine kasutamiseesmärk, tema eelised ja puudujäägid ning kasutamisevõimalused veebisaitide hindamisel. Keelemudelites kasutatud viip on kirjeldatud peatükis 4.2 ja alljärgnevalt on toodud iga keelemudeli vastus sellele muutmata kujul.

ChatGPT 4o (07.04.2025; <https://chat.openai.com>): “Keelemudel ChatGPT 4o on peamiselt loodud loomuliku keele töötlemise ülesannete täitmiseks, sealhulgas tekstide genereerimiseks, tõlkimiseks, kokkuvõtete koostamiseks ja küsimustele vastamiseks. Selle mudeli tugevusteks peetakse kontekstitundlikkust, keelelise väljendusrikkuse suutlikkust ning laialdast teadmistaasi, mis võimaldab pakkuda asjakohaseid ja struktureeritud vastuseid. Nõrkustena võib välja tuua faktilise täpsuse kõikumise ning keerukuste ilmumise ülesannetes, mis nõuavad väga spetsiifilisi või värskeid andmeid. Keelemudelit saab rakendada ka veebilehtede projektide hindamisel, eriti juhul, kui eesmärk on hinnata kasutajaliidese disaini vastavust üldtunnustatud printsiipidele, nagu tüpograafia selgus, visuaalne rühmitamine ja kasutatavus. Lisaks on võimalik

³ Jaotus on toodud keelemudelite “perekonna” kuuluvuse alusel.

analüüsida koodi loetavust ja struktuuri, kuid põhjalikum koodi kvaliteedi hindamine eeldab siiski täiendavat arenduskeskkonna tuge või spetsialiseeritud tööriistu. Seega sobib ChatGPT 4o hästi esmaseks analüüsiks ja soovitude andmiseks, kuid mitte lõplikuks kvaliteedihindamiseks tehniliselt keerukates projektides.”

ChatGPT o1 (07.04.2025, <https://chat.openai.com>): “ChatGPT o1 keelemudelit peetakse eelkõige sobivaks tekstipõhiste ülesannete lahendamisel, sealhulgas ideede genereerimisel, küsimustele vastamisel ja tekstianalüüsil. Selle peamiseks tugevusteks nimetatakse kontekstitundlikkust, laialdast teadmistepagasit ja suhteliselt loomulikku keelekasutust, samas kui nõrkustena võib esile tuua võimalikud ebatäpsused faktiteadmiste edastamisel ning piiratud suutlikkuse teatud keeleliste nüanssidega toime tulla. Veebilehete projektide hindamisel võimaldab keelemudel anda üldiseid soovitusi kasutajaliidese kujunduse, tüpograafia ja koodi struktuuri kohta, ent disaini peenemate detailide kriitiline analüüsimine võib jääda piiratud. Sellest hoolimata on võimalik koodi kvaliteeti ja semantilisi vastuolusid teatud määral hinnata, kasutades mudeli võimet tuvastada mustreid ja juhtida tähelepanu potentsiaalsetele vigadele. Seega on keelemudelist kasu abivahendina nii sisuanalüüsis kui ka koodikvaliteedi hindamisel, kuid otsused tuleks vastu võtta tervikliku disainimetoodika raames.”

ChatGPT 4.5 (07.04.2025, <https://chat.openai.com>): “Keelemudel ChatGPT 4.5 on eelkõige loodud tekstipõhiste ülesannete täitmiseks, nagu informatsiooni otsimine, sisu loomine ja erinevatele päringutele täpsete vastuste andmine, tuginedes suuremahulisele tekstikorpusele ja süvaõppele. Selle peamiste tugevustena võib välja tuua laialdase faktiteadmise, tekstide sidusa genereerimise ning võime mõista keerulisi päringuid, samas kui oluliste nõrkustena esineb vahel faktivigu, piiratud võimekust kontekstuaalsete nüansside mõistmisel ning sõltuvust treeningandmetest. Keelemudelit saab kasutada veebilehete projektide hindamisel disainipõhimõtete ja koodi kvaliteedi osas, kuid vaid toetava tööriistana, kuna mudelil puudub sügav visuaalne ja interaktiivne tunnetus kasutajaliidestest ning see hindab disaini ja kasutatavuse aspekte kaudselt tekstipõhiste soovitude kaudu. Koodi kvaliteedi hindamisel suudab mudel tuvastada standardeid järgivaid ja selgelt struktureeritud lahendusi, kuid ei pruugi alati mõista keerulisemaid programmeerimisnüansse ja projektispetsiifilisi eripärasid. Seetõttu tuleb keelemudeli hinnanguid pidada soovituslikeks ning neid peaks täiendama valdkonna ekspertide poolt läbiviidav hindamine.”

Gemini 2.0 Flash (07.04.2025, <https://gemini.google.com/>): “Keelemudel Gemini 2.0 Flash on peamiselt mõeldud kiirete ja tõhusate vastuste genereerimiseks erinevates rakendustes, kus oluline on latentsuse minimeerimine. Selle tugevustena võidakse nimetada kiirust ja võimet töödelda suures mahus infot lühikese aja jooksul, samas kui potentsiaalseteks nõrkusteks võivad osutada ressursipiirangutest tulenev mudeli suurus ja keerukamate arutluste piiratum võimekus võrreldes suuremate mudelitega. Veebilehtede projektide hindamisel saab Gemini 2.0 Flash teatud määral kasutada koodi struktuuri analüüsil ja võimalike mustrite tuvastamisel, kuid kasutajaliidese disaini printsiipide, nagu tüpograafia ja rühmitamine, hindamine nõuab sageli kontekstuaalset arusaamist ja esteetilist analüüsi, mida keelemudelid ei pruugi täielikult omada. Seega saab mudelit kasutada abivahendina, kuid inimspetsialisti hindamist kasutatavuse ja disaini kvaliteedi osas see täielikult ei asenda.”

Gemini 2.5 Pro (07.04.2025, <https://gemini.google.com/>): “Gemini 2.5 Pro on eeldatavasti kavandatud keerukate ülesannete lahendamiseks, mis nõuavad täiustatud arutlusvõimet, mitmeliigilist andmetöötlust (tekst, kood, pilt, heli, video) ning koodi genereerimist ja analüüsi. Mudeli tugevusteks peetakse tõenäoliselt selle võimekust töödelda väga suuri andmemahutusi ja pikki kontekste ning saavutada kõrgeid tulemusi keerukates arutus- ja kodeerimisülesannetes, samas kui nõrkusteks võivad olla võimalikud faktead (hallutsinatsioonid) ja treeningandmetest tulenev kallutus. Veebilehtede projektide hindamisel saab seda keelemudelit kasutada koodi kvaliteedi analüüsimiseks, tuvastades vigu, hinnates stiiljuhiste järgimist ja koodi keerukust. Kasutajaliidese disaini printsiipide (nt tüpograafia, rühmitamine) hindamisel võib mudel pakkuda tuge elementide tuvastamisel ja kirjeldamisel koodi või visuaalse sisendi põhjal, kuid kasutatavuse ja disaini esteetilise kvaliteedi sügavuti hindamine nõuab siiski inimotsustust. Põhjuseks on asjaolu, et keelemudelil puudub subjektiivne kogemus ja võime hinnata tegelikku kasutajamugavust ning visuaalset meeldivust inimkasutaja vaatepunktist.”

Deepseek (07.04.2025, <https://chat.deepseek.com/>): “DeepSeek (DeepThink® R1) keelemudel on loodud keerukate tekstipõhiste ülesannete lahendamiseks, sealhulgas info analüüsimiseks, koodi genereerimiseks ja disainipõhimõtete hindamiseks. Mudeli peamisteks tugevusteks on suur kontekstist arusaamine, mitmekülgne probleemilahendusvõime ning võimekus analüüsida nii teksti- kui ka koodipõhiseid struktuure, kuid piiranguks jääb visuaalse materjali otsene hindamise võime. Veebilehtede projekti hindamisel saab seda kasutada kasutajaliidese disainiprintsiipide (nt

tüpopraafia, rühmitamine, kasutatavus) analüüsimiseks kirjelduste põhjal ning koodi kvaliteedi hindamiseks loetavuse, struktuuri ja standarditele vastavuse alusel, kuid lõpliku hinnangu andmiseks on soovitatav kombineerida teiste vahenditega.”

Copilot (07.04.2025, <https://copilot.microsoft.com>): “Keelemudel "Copilot" on peamiselt mõeldud loomuliku keele töötlemiseks ja genereerimiseks, pakkudes kasutajatele intelligentset abi ja suurendades tootlikkust <https://copilot.microsoft.com>.⁴ Selle mudeli tugevusteks on võime mõista ja luua inimkeelele sarnaseid tekste ning pakkuda kontekstitundlikke vastuseid <https://learn.microsoft.com/en-us/answers/questions/2138848/what-is-the-specific-version-of-the-language-model>. Nõrkusteks on piiratud kultuuritundlikkus ja võimalus eelarvamuste edasiandmiseks, mis tulenevad treeningandmetest <https://learn.microsoft.com/en-us/answers/questions/2138848/what-is-the-specific-version-of-the-language-model>.

Keelemudelit saab kasutada veebilehtede projektide hindamisel, kuna see suudab analüüsida ja pakkuda soovitusi kasutajaliidese disaini printsiipide, nagu tüpopraafia, rühmitamine ja kasutatavus, ning koodi kvaliteedi kohta <https://codingmall.com/knowledge-base/25-global/318280--lang-et-kuidas-haldab-deepseek-keerulisi-andmepakendusi-vrreldes-exceli-copilotiga>. Mudeli võimekus põhineb tekstide statistiliste mustrite analüüsil, mis võimaldab tuvastada ja parandada disaini ja koodi puudusi <https://codingmall.com/knowledge-base/25-global/318280--lang-et-kuidas-haldab-deepseek-keerulisi-andmepakendusi-vrreldes-exceli-copilotiga>.”

Claude 3.7 Sonnet (07.04.2025, <https://claude.ai>): “Keelemudel Claude 3.7 Sonnet on mõeldud peamiselt loomuliku keele mõistmiseks ja genereerimiseks, pakkudes võimekust mitmesuguste ülesannete täitmiseks alates teksti koostamisest kuni keerulise arutlemiseni. Selle tugevusteks võib pidada teadmiste laia baasi, oskust konteksti hästi mõista ning võimekust genereerida üksikasjalikke ja täpseid vastuseid, samas kui nõrkusteks võib lugeda teadmiste piirnemist 2024. aasta oktoobriga ning potentsiaalseid raskusi täiesti uudsetes või äärmiselt spetsiifilistes valdkondades. Veebilehtede projektide hindamisel on mudelit võimalik kasutada nii kasutajaliidese disaini printsiipide kui ka koodi kvaliteedi analüüsimiseks, kuna mudel on treenitud mõistma disaini põhimõtteid (tüpopraafia, rühmitamine, kasutatavus) ning samuti erinevaid programmeerimiskeeli ja koodi struktuuri. Hindamisprotsessis tuleb arvestada, et mudel analüüsib

⁴ Selles lõigus toodud veebilingid on Copiloti poolt lisatud ja vastus on esitatud muutmata kujul.

seda, mida talle kirjeldatakse või näidatakse, pakkudes tagasisidet loogika ja õpitud printsiipide baasil, kuid ei suuda iseseisvalt veebilehte külastada.”

Mistral (07.04.2025, <https://mistral.ai>): “Mistral AI keelemudel on peamiselt mõeldud avatud lähtekoodiga suurte keelemudelite (LLM) loomiseks ja kohandamiseks, pakkudes võimalust neid kohandada, täpsustada, iteratsiooni teha ja uutele rakendustele ehitada. Mudeli tugevused on tema avatud ja kohandatav olemus, mis võimaldab kasutada seda erinevates keskkondades, sealhulgas pilves, kohapeal ja ääreseadmetel, samuti suur jõudlus ja madal latentsus. Nõrkused seisnevad võimalikes piirangutes kasutajaliidese disaini hindamisel, kuna mudel on peamiselt tekstipõhine ja võib puududa spetsiifilisi vahendeid veebilehtede visuaalse disaini analüüsimiseks. Seda keelemudelit saab kasutada veebilehtede projektide hindamisel, kuid ainult piiratud ulatuses. Mudel võib aidata koodi kvaliteedi analüüsimisel ja pakkuda soovitusi kasutatavuse parandamiseks, kuid visuaalse disaini printsiibid, nagu tüpograafia ja rühmitamine, ei ole tema peamised tugevused. Seega tuleks mudelit kasutada koos teiste spetsiifiliste tööriistadega, et saavutada täielikum hindamine.”

Keelemudelite enesetutvustustest selgus, et üldiselt keelemudelitel on võimalus hinnata veebidisaini. Kui koodikvaliteedi hindamisel ei teki probleeme, siis visuaali hindamine ei pruugi alati olla heal tasemel. Lisaks mõned mudelid (näiteks ChatGPT 4.5 ja Gemini 2.0 Flash) töid välja, et inimese sekkumine selle hindamisprotsessi on vajalik.

Bakalaureusetöö autor uuris ka valitud keelemudelite võimalusi kuvatõmmiste vastuvõtmiseks. Selgus, et seisuga 20. veebruar 2025 keelemudelid Deepseek ja Copilot ei olnud võimelised selleks. See aga tähendab, et nimetatud keelemudeleid ei saa käesolevas bakalaureusetöös kasutada valminud veebisaitide kasutajaliidese ja kasutajakogemuse hindamisel. Ainult koodiosa hindamist nende keelemudelite abil ei pidanud autor otstarbekaks. Seega edaspidisest analüüsist autor eemaldas Deepseek ja Copilot keelemudelid.

5.3 Autori ja keelemudelite poolt antud hinnangute võrdlus

Autori hinnangud lõputööna valminud veebisaitidele on esitatud peatükis 5.1. Keelemudelite hinnangute saamiseks kasutas autor seitse valitud keelemudelit (vt. peatükk 5.2) ja igasse neist sisestas samu viipe (vt. peatükk 4.2). Parema ülevaate saamiseks iga hindamiskriteeriumi kohta autori ja keelemudelite poolt antud hinnangutest kogutud andmed on koondatud ühte tabelisse.

Kõigepealt alustas autor **värvi** hindamiskriteeriumist. Autor ja keelemudelite poolt antud hinnangud veebisaitide prototüüpidele on koondatud tabelisse 3.

Tabel 3. Ülevaade autori ja keelemudelite hinnangutest värvi kriteeriumile.

	Autor	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Keskmine (SD)	4.22 (0.97)	3.44 (0.88)	3.33 (0.5)	4.11 (1.05)	4.11 (0.60)	3.44 (0.73)	3.44 (0.53)	3.22 (0.67)
Min-max	3-5	2-4	3-4	2-5	3-5	3-5	3-4	2-4
Mediaan	5	4	3	4	4	3	3	3

Autori, ChatGPT 4.5 ja Gemini 2.0 Flash hinnangul värvilahendused esitatud lõpuprojektide prototüüpides olid head (keskmine tulemus on üle 4). Teiste keelemudelite hinnangud olid rahuldavad (keskmiselt alla 3.5). Autori hinnangud varieerusid kolmest viieni. Samas mõnedel keelemudelitel hinnangute kõikumine oli kahest neljani. Autorite ja keelemudelite mediaanid olid erinevad.

Wilcoxonit astakmärgitesti abil võrreldi autori ja erinevate keelemudelite poolt antud hinnangud lõpuprojektides kasutatud värvilahendustele (tabel 4).

Tabel 4. Wilcoxonit astakmärgitesti tulemused (värvi kriteerium; W-statistik).

	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Autor	0.00*	0.00*	6.00	16.00	1.00	0.00*	2.00*

*p < 0.05

Wilcoxonit astakmärgitest näitas, et ChatGPT 4o, ChatGPT o1, Claude 3.7 Sonnet hinnangud olid statistiliselt madalamad kui autori omad (vastavalt $W = 0$, $W = 0$, $W = 2$ ja $W = 0$; kõigil juhtudel $p < 0.05$). See viitab sellele, et nimetatud keelemudelid hindasid värvilahendusi rangemalt kui autor. Gemini 2.5 Pro mudelil esines tendents hinnata esitatud lõpuprojektide värvilahendused rangemalt võrreldes autoriga ($p = 0.066$). ChatGPT 4.5 ja Gemini 2.0 Flash hinnangu erinevused autori hinnanguga ei olnud statistiliselt olulised (mõlemal juhul $p > 0.05$). Seega analüüsitud andmete põhjal ei saa väita, et keelemudelid hindasid veebisaitide värvilahendused rangemalt või leebemalt.

Järgmisena keskendus autor **värvikontrastsuse** hindamiskriteeriumile. Autori ja keelemudelite poolt antud hinnangud veebisaitide prototüüpidele on koondatud tabelisse 5.

Tabel 5. Ülevaade autori ja keelemudelite hinnangutest värvkontrastsuse kriteeriumile.

	Autor	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Keskmine (SD)	3.67 (0.87)	3.89 (0.93)	3.44 (0.73)	4.44 (0.88)	3.89 (0.33)	4.44 (0.53)	3.11 (0.33)	3.00 (0.87)
Min-max	3-5	3-5	2-4	3-5	3-4	4-5	3-4	2-4
Mediaan	3	4	4	5	4	4	3	3

ChatGPT 4.5 ja Gemini 2.5 Pro hinnangute kohaselt olid kontrasti kasutamine lõpuprojektide prototüüpides heal tasemel (keskmise tulemus üle 4.0). Nii töö autori kui ülejäänute keelemudelite hinnangud olid rahuldavad (keskmine tulemus alla 4). Autori hinnangud ning ChatGPT 4o ja ChatGPT 4.5 hinnangud ulatusid kolmest viieni. ChatGPT o1 ja Claude 3.7 Sonnet hinnangud varieerusid kahest neljani. Autoril, Mistralil ja Claudel oli mediaanväärtus kolm, samas ülejäänutel mudelitel oli see kõrgem.

Wilcoxon'i astakmürgitesti tulemustest (tabel 6) selgus, et autori ja kõikide uuritavate keelemudelite värvkontrastsuse hinnangute vahel ei ilmnenud statistiliselt olulist erinevust (kõikidel juhtudel $p > 0.05$). See tähendab, et autori ja keelemudelite poolt antud hinnangud ei erinenud üksteisest märkimisväärselt. Samas ChatGPT 4.5 ja Gemini 2.5 Pro puhul oli märgatud tendents hinnata veebisaitidel kasutatud värvkontrastsuse autorist leebemalt (vastavalt $p = 0.059$ ja $p = 0.083$).

Tabel 6. Wilcoxon'i astakmürgitesti tulemused (värvkontrastsusei kriteerium; W-statistik).

	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Autor	5.00	7.50	0.00	13.50	6.00	2.00	2.50

* $p < 0.05$

Edasi keskendus autor **rühmitamise** hindamiskriteeriumile. Autor ja keelemudelite poolt antud hinnangud esitatud veebisaitide prototüüpidele on koondatud tabelisse 7.

Tabel 7. Ülevaade autori ja keelemudelite hinnangutest rühmitamise kriteeriumile.

	Autor	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Keskmine (SD)	4.11 (1.05)	4.11 (0.93)	2.89 (0.60)	4.56 (1.01)	4.00 (0.50)	4.78 (0.44)	4.00 (0.00)	4.11 (0.33)
Min-max	2-5	2-5	2-4	2-5	3-5	4-5	4-4	4-5
Mediaan	4	4	3	5	4	5	4	4

Analüüsi käigus selgus, et nii autor kui ka enamik keelemudelitest hindasid rühmitamise lahenduse rahuldavalt (keskmine tulemus üle 4). Ainult ChatGPT o1 keelemudeli puhul oli selle kriteeriumi keskmine tulemus alla kolme. Hinnangud varieerusid kahest viieni. Mediaani väärtused jäid nelja või viie vahele (välja arvatud ChatGPT o1).

Wilcoxon astakmargitesti tulemused (tabel 8) näitasid, et ChatGPT o1 hinnang rühmitamisele oli statistiliselt olulisel määral madalam autori omadest (vastavalt $p < 0.05$). Gemini 2.5 Pro keelemudelil esines tendents hinnata esitatud projektide rühmitamise kriteerium leebemalt võrreldes autoriga ($p = 0.063$). Ülejäänutel uuritavatel keelemudelitel statistilisi erinevusi autori hinnangust ei ilmnud (kõikidel juhtudel $p > 0.05$).

Tabel 8. Wilcoxon astakmargitesti tulemused (rühmitamise kriteerium; W-statistik).

	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Autor	1.50	2.00*	0.00	16.00	0.00	9.00	14.00

* $p < 0.05$

Järgmisena autor keskendus **tüpoograafia** hindamiskriteeriumile. Autor ja keelemudelite poolt antud hinnangud veebisaitide prototüüpidele on koondatud tabelisse 9.

Tabel 9. Ülevaade autori ja keelemudelite hinnangutest tüpoograafia kriteeriumile.

	Autor	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Keskmine (SD)	4.00 (1.12)	3.67 (0.71)	3.22 (0.44)	4.56 (0.88)	4.00 (0.50)	4.56 (0.53)	3.44 (0.53)	3.00 (0.50)
Min-max	2-5	3-5	3-4	3-5	3-5	4-5	3-4	2-4
Mediaan	4	4	3	5	4	5	3	3

Autori, ChatGPT 4.5, Gemini 2.0 Flash ja Gemini 2.5 Pro keskmised hinnangu tulemused valitud tüpoograafiale oli neli või üle selle. Teiste keelemudelite hinnangud olid keskmiselt alla 4, kusjuures ChatGPT o1 ja Claude 3.7 Sonneti keskmised tulemused olid kolmele lähedad. Autori hinnangud varieerusid vahemikus kahest viieni. Samas mõnede keelemudelite puhul hinnangute vahemik oli kitsam. Autori ja keelemudelite mediaanid olid erinevad.

Wilcoxon astakmargitesti tulemused (tabel 10) näitasid, et statistiliselt oluline erinevus autori hinnangust ilmnes ainult Claude 3.7 Sonneti puhul ($p < 0.05$). See mudel hindas tüpoograafiat

oluliselt rangemalt võrreldes autoriga. ChatGPT 4.5 puhul oli märgatud tendents hinnata tüpograafia autorist kõrgemalt ($p = 0.059$). Ülejäänud mudelite puhul statistiliselt olulisi erinevusi ei täheldatud (kõikidel juhtudel $p > 0.05$).

Tabel 10. Wilcoxon astakmärgitesti tulemused (tüpograafia kriteerium).

	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Autor	7.00	4.00	0.00	13.00	6.00	6.00	0.00*

* $p < 0.05$

Kasutajaliidese viimase hindamiskriteerium on **kissitustesti** läbimine. Autor ja keelemudelite poolt antud hinnangud veebisaitide prototüüpidele on koondatud tabelisse 11.

Tabel 11. Ülevaade autori ja keelemudelite hinnangutest kissitustestile.

	Autor	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Keskmine (SD)	4.11 (1.36)	2.89 (1.05)	3.11 (0.60)	4.33 (1.12)	4.44 (0.53)	4.55 (0.53)	3.67 (0.50)	2.89 (0.78)
Min-max	1-5	1-4	2-4	2-5	4-5	4-5	3-4	2-4
Mediaan	5	3	3	5	4	5	4	3

Autori, ChatGPT 4.5, Gemini 2.0 Flash ja Gemini 2.5 Pro hinnangul läbisid esitatud veebisaitide prototüübid kissitustesti hästi (keskmine tulemus üle 4). Samas ChatGPT 4o ja Claude 3.7 Sonnet hinnangud olid tagasihoidlikud (keskmine tulemus alla 3). Autori hinnangud varieerusid vahemikus ühest viieni. Samas mõne keelemudeli puhul kõikusid hinnangud kitsamas vahemikus kahest neljani. Autori ja keelemudelite mediaanid olid erinevad.

Wilcoxon astakmärgitesti tulemuste (tabel 12) järgi statistiliselt olulisi erinevusi ei ilmnenud (kõigil juhtudel $p > 0.05$). Samas ChatGPT 4o ja Claude 3.7 Sonnet puhul oli täheldatav tendents rangemalt hinnata veebisaitide prototüüpe kissitustesti alusel (vastavalt $p = 0.062$ ja $p = 0.077$).

Tabel 12. Wilcoxon astakmärgitesti tulemused (kissitustesti kriteerium; W-statistik).

	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Autor	5.00	9.00	3.00	9.00	10.50	8.50	5.50

* $p < 0.05$

Kasutajakogemuse juures uuris autor loodud prototüüpide vastavust **kasutatavuse** kriteeriumile. Autor ja keelemudelite poolt antud hinnangud veebisaitidele on koondatud tabelisse 13.

Tabel 13. Ülevaade autori ja keelemudelite hinnangutest kasutatavuse kriteeriumile.

	Autor	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Keskmine (SD)	4.22 (1.09)	3.89 (0.78)	3.55 (0.53)	4.67 (0.71)	4.11 (0.33)	4.44 (0.53)	4.00 (0.00)	3.44 (0.53)
Min-max	2-5	3-5	3-4	3-5	4-5	4-5	4-4	3-4
Mediaan	5	4	4	5	4	4	4	3

Autori, ChatGPT 4.5, Gemini 2.0 Flash, Gemini 2.5 Pro ja Mistrali hinnangul esitatud veebisaitidel kasutatavus on üldiselt hästi realiseeritud (keskmine tulemus 4 või üle selle). Teiste keelemudelite hinnangud jäid alla 4.0. Autori hinnangud varieerusid kahest viieni, keelemudelid andsid hinnangud kitsamas vahemikus. Mistral andis kõikidele esitatud prototüüpidele hinnang 4. Kui autori ja ChatGPT 4.5 mediaan oli viis, siis ülejäänutel keelemudelitel see oli madalam.

Wilcoxon'i astakmärgitesti tulemused (tabel 14) näitasid, et statistiliselt oluline erinevus autori hinnangust ilmnes ChatGPT 4.5 puhul ($p < 0.05$). See mudel hindas kasutatavust leebemalt kui autor. Teiste mudelite puhul olulisi erinevusi autori hinnangutest ei tuvastatud ($p > 0.05$).

Tabel 14. Wilcoxon'i astakmärgitesti tulemused (kasutatavuse kriteerium; W-statistik).

	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Autor	3.00	6.00	0.00*	13.00	4.00	10.50	7.00

* $p < 0.05$

Koodi osa esimese kriteeriumina uuris autor veebisaidi loomisel kasutatud **koodi loetavust**. Autor ja keelemudelite poolt antud hinnangud sellele kriteeriumile on koondatud tabelisse 15.

Tabel 15. Ülevaade autori ja keelemudelite hinnangutest koodi loetavuse kriteeriumile.

	Autor	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Keskmine (SD)	3.89 (0.60)	4.22 (0.44)	4.00 (0.00)	5.00 (0.00)	3.67 (1.00)	3.78 (0.97)	4.00 (0.00)	3.44 (0.53)
Min-max	3-5	4-5	4-4	5-5	1-4	2-5	4-4	3-4
Mediaan	4	4	4	5	4	4	4	3

ChatGPT 4o, ChatGPT o1 ja Mistral hinnangud olid keskmiselt üle 4. Ülejäänute keelemudelite ja autori keskmised hinnangud olid alla nelja. ChatGPT 4.5 andis kõikidele esitatud projektide puhul maksimaalse hinnangu. Autori hinnangud varieerusid vahemikus kolmest viieni, samas kui mõne mudeli hinnangute vahemik oli laiem. Autorite ja keelemudelite mediaanid olid valdavalt neli.

Wilcoxon'i astakmärgitesti tulemused (tabel 16) näitasid, et statistiliselt oluline erinevus autori hinnangutest ilmnis ainult ChatGPT 4.5 hinnangute puhul ($p = 0.008$). See mudel hindas koodi loetavust oluliselt kõrgemalt võrreldes autoriga. Ülejäänud mudelite puhul olulisi erinevusi autori hinnangust ei täheldatud ($p > 0.05$).

Tabel 16. Wilcoxon'i astakmärgitesti tulemused (koodi loetavuse kriteerium; W-statistik).

	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Autor	0.00	2.00	0.00*	4.00	12.00	2.00	2.50

* $p < 0.05$

Järgmisena autor keskendus **koodi hooldatavuse** hindamiskriteeriumile. Autor ja keelemudelite poolt antud hinnangud veebisaidi loomisel kasutatud koodile on koondatud tabelisse 17.

Tabel 17. Ülevaade autori ja keelemudelite hinnangutest koodi hooldatavuse kriteeriumile.

	Autor	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Keskmine (SD)	3.00 (1.12)	3.11 (0.60)	4.00 (0.00)	5.00 (0.00)	4.11 (0.33)	2.78 (1.09)	4.11 (1.05)	2.78 (0.83)
Min-max	1-4	2-4	4-4	5-5	4-5	1-4	3-5	2-4
Mediaan	3	3	4	5	4	3	5	3

Kui ChatGPT 4.5 hinnangul kõikides projektides koodi hooldatavus oli realiseeritud ideaalselt (keskmine tulemus 5.0), siis Gemini 2.5 Pro ja Claude 3.7 Sonnet hinnangul oli esitatud veebisaitidel koodi hooldatavusega probleemid (keskmine tulemus alla 3). Autori hinnangud varieerusid ühest neljani, samas mõnede keelemudelite puhul hinnangute vahemik oli väiksem. ChatGPT o1 ja ChatGPT 4.5 puhul kõik projektid said sama hinnangu - vastavalt 4 või 5. Autori ja keelemudelite mediaanid varieerusid kolmest viieni.

Wilcoxon astakmärgitesti (tabel 18) selgus, et ChatGPT o1, ChatGPT 4.5, Mistral ja Gemini 2.0 Flash mudelid hindasid koodi hooldatavuse autorist leebemalt (kõikidel juhtudel $p < 0.05$). Mistral mudeli puhul oli märgitud tendents hinnata esitatud projektide koodi hooldatavust leebemalt kui autor ($p = 0.055$). Ülejäänud mudelite puhul olulisi erinevusi autori hinnangust ei olnud ($p > 0.05$).

Tabel 18. Wilcoxon astakmärgitesti tulemused (koodi hooldatavuse kriteerium; W-statistik).

	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Autor	6.00	0.00*	0.00*	0.00*	7.50	6.00	11.00

* $p < 0.05$

Koodi kvaliteedi hindamise viimane kriteerium oli seotud **koodi dokumentatsiooniga**. Autori ja keelemudelite poolt antud hinnangud sellele on koondatud tabelisse 19.

Tabel 19. Ülevaade autori ja keelemudelite hinnangutest koodi dokumentatsiooni kriteeriumile.

	Autor	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Keskmine (SD)	1.22 (0.67)	0.33 (0.50)	0.11 (0.33)	0.22 (0.44)	0.44 (1.33)	2.56 (1.51)	0.44 (0.73)	1.56 (0.88)
Min-max	0-2	0-1	0-1	0-1	0-4	0-4	0-2	0-3
Mediaan	1	0	0	0	0	3	0	2

Autor ega ükski keelemudel ei hinnanud koodi dokumentatsiooni rahuldavalt. Kõik hinnangud jäid keskmiselt alla 3. Enamik uuritatud keelemudelitest pani mõnede esitatud veebisaitidele nulli. Nende keelemudelite hinnangul esitatud veebisaitide puhul ei olnud võimalik dokumentatsiooni hinnata.

Wilcoxon astakmärgitesti tulemused (tabel 20) näitasid, et statistiliselt olulised erinevused ilmnasid autori ja ChatGPT 4o, ChatGPT o1, ChatGPT 4.5 ja Mistral keelemudelite hinnangute puhul ($p < 0.05$). Nimetatud mudelid hindasid dokumentatsiooni oluliselt rangemalt võrreldes autori hinnangutega. Gemini 2.5 Pro puhul on tendents hinnata dokumentatsiooni leebemalt kui autor ($p = 0.056$). Ülejäänud mudelite puhul statistiliselt olulisi erinevusi autori hinnangust ei ilmnenu ($p > 0.05$).

Tabel 20. Wilcoxon'i astakmärgitesti tulemused (koodi dokumentatsiooni kriteerium; W-statistik).

	ChatGPT 4o	ChatGPT o1	ChatGPT 4.5	Gemini 2.0 Flash	Gemini 2.5 Pro	Mistral	Claude 3.7 Sonnet
Autor	0.00*	0.00*	0.00*	8.00	4.50	0.00*	0.00

*p < 0.05

Allpool toodud tabel 21 võtab kokku millistes hindamiskriteeriumites Wilcoxon'i astakmärgitesti ei tuvastanud statistiliselt olulised erinevusi autori ja uuritavate keelemudelite poolt antud hinnangutes.

Tabel 21. Hindamiskriteeriumid, kus autori ja keelemudelite poolt antud hinnangutes ei olnud statistiliselt olulist erinevust.

	Kasutajaliides	Kasutajakogemus	Koodi hooldatavus
Autor vs ChatGPT 4o	kontrast, rühmitamine, kissitustest, tüpograafia	kasutatavus	koodi loetavus, koodi hooldatavus
Autor vs ChatGPT o1	kontrast, kissitustest, tüpograafia	kasutatavus	koodi loetavus
Autor vs ChatGPT 4.5	värv, kontrast, rühmitamine, kissitustest, tüpograafia		
Autor vs Gemini 2.0 Flash	värv, kontrast, rühmitamine, kissitustest, tüpograafia	kasutatavus	koodi loetavus, koodi dokumentatsioon
Autor vs Gemini 2.5 Pro	värv, kontrast, rühmitamine, kissitustest, tüpograafia	kasutatavus	koodi loetavus, koodi hooldatavus, koodi dokumentatsioon
Autor vs Mistral	kontrast, rühmitamine, kissitustest, tüpograafia	kasutatavus	koodi loetavus
Autor vs Claude 3.7 Sonnet	kontrast, rühmitamine, kissitustest	kasutatavus	koodi loetavus, koodi hooldatavus, koodi dokumentatsioon

Järgmises peatükis arutletakse uuringu tulemuste üle.

6. Arutelu

Bakalaureusetöö eesmärk oli analüüsida Tartu Ülikooli kursuse „Veebilehtede loomine edasijõudnutele“ vältel koostatud veebisaitide ja välja selgitada keelemudelite kasutamissobivust nende hindamisel. Tulemuste üle arutletakse uurimisküsimuste alusel.

Esimene uurimisküsimus oli “Millist hinnangut loodud veebisaitidele annab bakalaureusetöö autor?” Vastuse leidmiseks analüüsis autor 9 meeskonnatööna valminud veebisaiti. Autori hinnangul üldiselt valminud veebisaitide prototüübid vastasid kasutajaliidese disaini headele tavadele ja saavutatud tulemused olid kõrged. Õigete värvide valik, erinevate elementide läbimõeldud rühmitamine ja sobiv tüpograafia aitavad edastada veebisaidi sõnumit efektiivsemalt ja muudavad veebisaidil navigeerimist loogilisemaks [19, 21, 26]. Samas veebisaitide prototüüpidel kasutatud värvikontrastsuse puhul olid tuvastatud puudujäägid. Nende kõrvaldamisele tuleks kindlasti tähelepanu pöörata, sest halvasti valitud värvikontrastsus võib mõjuda veebisaidi külastajale häirivalt [23].

Uurimistulemused näitasid, et loodud veebisaitide prototüüpide kasutatavus oli hästi läbimõeldud. See on väga oluline kasutajakogemuse disainis, sest kasutajatel on mugav orienteeruda veebisaidil ning suurendab tõenäosust, et hea meelega nad külastavad seda uuesti ja soovivad veebisaiti ka teistele [25, 46]. Veebisaitide koodi uurimisest selgus, et töös kasutatud hindamiskriteeriumite järgi koodi kvaliteet oli väga madal. Edaspidi tuleks sellele rohkem tähelepanu pöörata, sest koodikvaliteedil on oluline roll tarkvaraliste lahenduste elutsüklis [48, 53, 55].

Teine uurimisküsimus oli “Millised tehisharu keelemudelid on võimelised hindama loodud veebisaitide?” Vastuse leidmiseks analüüsis autor üheksat erinevat keelemudelit. Nende enesetutvustuste põhjal võib järeldada, et üldiselt nad on võimelised hindama veebisaitide, kuid vastustes rõhutati, et täieliku hinnangu saamiseks on vaja ikkagi pöörduda inimese poole. Lisaks mõned keelemudelid tõid välja ka piiranguid. Näiteks keelemudelil ChatGPT 4.5 võib olla raskusi keerulisema programmeerimiskoodiga, samas keelemudel Mistral ei ole tugev kasutajaliidese disaini hindamisel. Ka varasemates uuringutes toodi välja, et veebisaitide koodi genereerimisel ei saa täielikult keelemudelitele lootma jääda [62, 89] ning neil võivad tekkida raskused esteetika hindamisel [64].

Keelemudelite DeepSeek ja Copilot puhul osutus suurimaks takistuseks nende sisendipiirang, mis ei võimaldanud neil kuvatõmmiste analüüsi viia kaudu (seisuga 20. veebruar 2025). Seetõttu neid välistati edaspidisest analüüsist. Seega võib bakalaureusetöö raames läbiviidud uurimuse põhjal öelda, et töös määratletud hindamiskriteeriumite alusel on loodud veebisaitide hindamiseks võimelised järgmised keelemudelid: ChatGPT 4o, ChatGPT o1, ChatGPT 4.5, Gemini 2.0 Flash, Gemini 2.5 Pro (experimental), Claude 3.7 Sonnet ja Mistral.

Kolmas uurimisküsimus oli “Kas autori ja keelemudelite poolt antud hinnangutes valminud veebisaitidele esineb statistilisi erinevusi, kui kasutada samu hindamiskriteeriume?”. Uurimistulemused näitasid, et veebisaitide prototüüpide värvikasutuse hindamisel olid keelemudelid ChatGPT 4o, ChatGPT o1, Claude 3.7 Sonnet ja Mistral autoriga võrreldes oluliselt rangemad. Seda võib seletada asjaoluga, et tehisaru keelemudelid põhinevad tugevalt oma treeningandmestikul ja kasutavad hindamisel pigem tehnilist analüüsi kui inimesele omast esteetilist tajumist. Lisaks varasemad uuringud [64, 72] on näidanud, et tehisarul on keeruline hinnata esteetikat samadel alustel kui inimene, kuna puudub võime tajuda abstraktseid esteetilisi väärtusi või arvestada vaataja individuaalset emotsionaalset ja kultuurilist tausta. Wang jt. [73] uuringu tulemuste kohaselt kipuvad generatiivse tehisaruloodud pildid olema statistiliselt "kitsamad" värvitunnuste osas: neil on väiksem tonaalsuse, krooma ja heleduse varieeruvus ning kõrgem korrelatsioon objektide ja taustade vahel võrreldes reaalselt loodud piltidega. Samas keelemudelite Gemini 2.0 Flash, Gemini 2.5 Pro ja ChatGPT 4.5 hinnangud ei erinenud oluliselt autori hinnangutest.

Värvikontrastsuse hindamisel ei esinenud uuritud keelemudelite ja autori poolt antud hinnangutes olulisi erinevusi. Üheks põhjuseks võib olla see, et kontrasti hindamine põhineb selgelt määratletud reeglitel (näiteks veebisisu juurdepääsetavuse juhised (ingl *Web Content Accessibility Guidelines*, WCAG), mis ütlevad, kui milline minimaalne kontrastsuse tase peab olema teksti ja tausta värvide vahel, et sisu oleks hästi loetav. Uuringud on näidanud, et automaatsed hindamissüsteemid suudavad edukalt tuvastada kontrasti erinevusi ja hinnata nende vastavust ligipääsetavus standarditele [74].

ChatGPT o1 keelemudel hindas veebisaitide prototüüpidel kasutatud rühmitamist madalamalt kui autor. Kõik ülejäänud keelemudelid andsid rühmitamise hindamisel tulemusi, mis ei erinenud oluliselt autori omadest. See tulemus toetab Koch ja Oulasvirta uuringu [75] tulemusi, kus leiti et

tehisintellekti keelemudelid suudavad rakendada Gestalt printsiipe disainis, ning suurendada edukate tulemusi kuni 90% rühmitamise valdkonnas.

Tüpopograafia hindamisel oli statistiline erinevus ainult Claude 3.7 Sonneti puhul, kus mudel andis rangemalt hinnanguid võrreldes autoriga. Teiste mudelite poolt antud hinnangud ei erinenud statistiliselt autori omadest. Kuigi tüpopograafia tajumine sõltub visuaalsete ja esteetiliste parameetrite mõistmisest [64], näitasid käesoleva uuringu tulemused, et enamiku keelemudelite puhul ei erinenud oskus hinnata põhilisi tekstistruktuuri ja loetavuse omadusi autori omast.

Kissitustest on mitteformaalne lähenemine visuaalse hierarhia hindamisele [38, 39]. Uuringu tulemused näitasid, et selle hindamiskriteeriumi puhul autori ja kõikide uuritud keelemudelite hinnangute vahel ei olnud erinevust. See viitab sellele, et kõik osapooled hindasid veebisaitidel kasutatud visuaalset hierarhiat ühtlaselt ning ei kaldunud rohkem kriitilisuse või leebuse poole.

Kasutatavuse hindamisel peaaegu kõikide uuritud keelemudelite ja autori hinnangute vahel ei olnud statistilist erinevust. Siit võib järeldada, et kui tahetakse hinnata veebisaitide kasutajakogemust just kasutatavuse nurga alt, siis analüüsitud keelemudelid võivad olla abiks. Ka varasemad uuringud on näidanud, et keelemudelistest on abi kasutatavuse hindamisel [76]. Samas bakalaureusetöös GhatGPT 4.5 hindas esitatud lõpuprojektide kasutatavust autorist leebemalt. Põhjuseks võib olla ChatGPT 4.5 enesetuvustuses toodud piirang, et antud keelemudel saab hinnata kasutatavust ainult tekstipõhiselt.

Koodi loetavuse hindamisel andis ainult ChatGPT 4.5 statistiliselt kõrgemad hinnangud võrreldes autori omaga. Kõikide teiste mudelite poolt antud hinnangud ei erinenud statistiliselt autori omadest. Koodi hooldatavuse hindamisel olid tulemused rohkem erinevad. Autori hinnangutest ei erinenud keelemudelite ChatGPT 4o, Gemini 2.5 Pro ja Claude 3.7 Sonnet hinnangud. Keelemudelite ChatGPT o1, ChatGPT 4.5, Gemini 2.0 Flash ja Mistral hinnangud olid autori omast kõrgemad. Samas mõlema kriteeriumi puhul ChatGPT o1 ja ChatGPT 4.5 hinnangud ei varieerunud. Kõikidele lõpuprojektidele need keelemudelid panid samad hinnangud (4 ja 5 vastavalt). Varasemates uuringus [85] on leitud, et ChatGPT'l põhinevad tehisaru keelemudelid suudavad kirjutada koodi, mis oma loetavuse ja hooldatavuse poolest parem kui inimese poolt kirjutatud kood. Võimalik, et ChatGPT o1 ja ChatGPT 4.5 suutsid paremini aru saada kirjutatud koodist kui autor.

Keelemudelitele Gemini 2.0 Flash, Gemini 2.5 Pro ja Claude 3.7 Sonnet hinnangud koodi dokumentatsioonile ei erinenud oluliselt autori omadest. Võimalik, et need mudelid suutsid analüüsida dokumentatsiooni olemasolu ja põhitaset. Teised keelemudelid panid esitatud veebisaidid selle kriteeriumi alusel madalamad hinnangud kui autor. Balford [77] toob välja, et kuigi suured keelemudelid oskavad automaatselt hinnata kommentaaride olemasolu ja lihtsat dokumentatsiooni struktuuri, jääb nende võime hinnata dokumentatsiooni semantilist sügavust ja asjakohasust piiratud tasemele. Seega on dokumentatsiooni kvaliteedi põhjalik hindamine endiselt valdkond, kus inimene peab lõplikud otsused tegema.

Kokkuvõtvalt võib öelda, et kasutajaliidese ja kasutajakogemuse disainiga seotud kõikide kriteeriumite hindamisel Gemini 2.0 Flash ja Gemini 2.5 Pro poolt antud hinnangud ei erinenud oluliselt autori hinnangutest. Lisaks kui Gemini 2.5 Pro puhul ei olnud olulist erinevust autori hinnangust koodi kvaliteedi hindamisel, siis Gemini 2.0 Flash puhul koodi hooldatavus oli hinnatud oluliselt kõrgem kui autori oma. Kuigi nende keelemudelite enesetutvustuse järgi võib neil võivad tekkida raskusi kasutatavuse ja esteetiliste aspektide hindamisel (vt peatükk 5.2), näitasid uuringu tulemused, et nimetatud keelemudelid ja autor hindasid veebisaitide ühtlaselt. Põhjuseks, miks erinevusi tulemustes ei ilmnenu, võib olla see, et kursuse raames loodud veebisaidid polnud eriti keerulised ja nimetatud keelemudelid, mis tulid turule mõned kuud tagasi, olid võimelised hindama need tööd sarnaselt autoriga. Seega edaspidi võiks kursusel loodud veebisaitide hindamisel töös käsitletud kriteeriumite alusel kasutada abivahendina Gemini keelemudelid.

7. Piirangud ja võimalikud uurimissuunad

Töö üheks oluliseks piiranguks on valimi suurus. Töös analüüsiti ainult 9 kursuse lõputööna valminud veebisaiti. See ei võimalda teha laiendatavaid järeldusi või üldistusi.

Autori poolt antud hinnangud on subjektiivsed ja võivad olla mõjutatud erinevate faktorite poolt. Varasemad uuringud on näidanud, et sugu [78], emotsioonid [79] ja kultuuriline taust [34] võivad mõjutada hinnanguid.

Kindlasti tuleb arvestada keelemudelite eripäradega. Töös uuritud keelemudelitel puudub täielik ligipääs visuaalsele ja semantilisele kontekstile. Nad ei saa avada prototüüpide linke ega hinnata lõpuprojekte terviklikult visuaalse kogemuse põhjal. Seetõttu võivad jääda nende hinnangud osaliselt puudulikuks. Selle probleemi üheks lahenduseks kasutas töö autor kuvatõmmiseid. Töö käigus analüüsitud tulemused peegeldavad mudelite suutlikkust töötada kuvatõmmistega seisuga 20. veebruar 2025. Arvestades, et vaatlusalused mudelid arenevad pidevalt, võivad nende täpsus ja võimekus töös käsitletud hindamiskriteeriumite alusel hinnata veebisaitide oluliselt muutuda juba lähitulevikus.

Tulevikus võiks välja selgitada, kuivõrd sõltuvad tehisaru keelemudelite hinnangud viiba täpsusest, keelest ning tagasiside vormist (arvuline ja sõnaline tagasiside). Võiks uurida, kuidas keelemudelid saavad hakkama suuremahuliste projektide (näiteks nendega, mis eeldavad *full-stack* arendust) hindamisega. Viimane aga eeldab hindamiskriteeriumite juurde lisamist, sest projektid muutuvad mahukamaks. Käesolevas töös keskenduti veebisaitidele, aga oleks hea võrrelda, kuidas keelemudelid saavad hakkama teiste õppeainete ja ülesannete tüüpi hindamisega. Edaspidi võiks luua just veebisaitide hindamiseks mõeldud keelemudeli, mis on treenitud andma tagasisidet erinevate kasutajaliidese ja kasutajakogemuse disaini ning koodi kvaliteedi kohta.

Kokkuvõte

Bakalaureusetöö eesmärk oli analüüsida Tartu Ülikooli kursuse „Veebilehtede loomine edasijõudnutele“ raames loodud veebisaitide ja välja selgitada erinevate keelemudelite kasutamissobivust nende hindamisel. Eesmärgi saavutamiseks analüüsiti üheksat lõpuprojekti. Teaduskirjanduse alusel valis töö autor hindamiskriteeriumid, mille abil oleks võimalik analüüsita esitatud lõpuprojektide kasutajaliidestest, kasutajakogemusest ja koodi kvaliteedist.

Uuringu tulemused näitasid, et loodud veebisaitide prototüübid olid üldiselt hea kasutajaliidese ja kasutajakogemuse disainiga. Üldiselt üliõpilased on arvestatud disaini heade tavadega prototüüpide loomisel. Samas oli koodi kvaliteedi hinnang pigem tagasihoidlik. Arvestades koodi olulisust tarkvaraarenduses, tuleb edaspidi pöörata koodi kvaliteedile rohkem tähelepanu.

Töö tulemused näitasid, et seisuga 20.02.2025 eessüsteemidega veebisaitide hindamiseks sobivad ChatGPT 4.0, ChatGPT o1, ChatGPT 4.5, Gemini 2.0 Flash, Gemini 2.5 Pro (eksperimentaalne), Claude 3.7 Sonnet ja Mistral. DeepSeek ja Copilot ei võimaldanud kuvatõmmiste sisestamist. Seetõttu ei saanud neid kasutada kasutajaliidese ja kasutajakogemuse disaini hindamisel.

Käesoleva töö tulemused näitasid, et kui on vaja hinnata kursusel valminud veebisaitide tervikuna – st nii kasutajaliidest, kasutajakogemust kui koodi kvaliteeti, siis võrreldes teiste keelemudelitega just Gemini 2.0 Flash ja Gemini 2.5 Pro poolt antavad hinnangud ei erinenud oluliselt autori omadest. Kui projekti raames keskendutakse ainult kasutajaliidese hindamisele, siis üldiselt abiks võivad olla ka ChatGPT 4.0 ja 4.5 keelemudelid. Kui lõpuprojektide hindamisel suurem rõhk on pandud kasutatavuse hindamisele, mis on kasutajakogemuse aluseks, siis selles osas saavad inimhindajat aidata praktiliselt kõik töös käsitletud keelemudelid. Erandiks on ChatGPT 4.5, mis andis oluliselt kõrgema hinnangu kasutatavuse kriteeriumile esitatud lõpuprojektide puhul kui autor. Kui hindamisel on oluline roll koodi kvaliteedil, siis abiks võivad olla ka ChatGPT 4.0 ja Claude 3.7 Sonnet. Nende poolt antud hinnang enamikes koodi kvaliteedi hindamiskriteeriumites oluliselt ei erinenud autori omast.

Tuleb töö tulemustele ja analüüsitud allikatele tuginedes rõhutada, et kuigi tehisaru keelemudelite rakendamine lõpuprojektide hindamisel näitas teatud juhtudel olulist erinevust inimhindamisest, ei saa olemasolevate teadmiste põhjal järeldada, et tehisaru hinnanguid oleks võimalik täielikult ja kriitikavabalt usaldada.

Viited

- [1] Bytyçi S., Shala V. ja Hasrama E. (2024), *The corporate strategy of increasing the visibility of websites by implementing digital marketing*. Edelweiss Applied Science and Technology. 8(6), 2705 - 2706. <https://doi.org/10.55214/25768484.v8i6.2532>
- [2] Polgar J. (2010), Do You Need Content Management System. International Journal of Web Portals (IJWP). 2(1), 1-6. <https://www.igi-global.com/gateway/article/40314>
- [3] Strikingly Blog. (2023). *Feature-full flourish: Elevating your web presence with essential website features*. Vaadatud: 30.03.2025. [Online] <https://www.strikingly.com/blog/posts/feature-full-flourish-elevating-your-web-presence-with-essential-website-features>
- [4] Wijaya, A., Kefry, Wihalim, W., & Gunawan, A. A. S. (2021). *The effect of UI/UX design on user satisfaction in online art gallery*. In 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI) 120–125. IEEE. <https://doi.org/10.1109/ICCSAI53272.2021.9609764>
- [5] W3S. (n.d.) *Understanding Success Criterion 1.4.3*. Vaadatud: 07.04.2025. [Online] <https://www.w3.org/TR/UNDERSTANDING-WCAG20/visual-audio-contrast-contrast.html>
- [6] Arangarajan, M., Kumar, T. C. A., Krishna, M. H., Sobti, R., Ilavarasan, B., Nayak, B. B. (2024). *Applications of AI to Optimize Operations in the Management of Manufacturing*. In 2024 International Conference on Innovative Data Communication Technologies and Application (ICIDCA). 1520-1524. IEEE. <https://doi.org/10.1109/IC3SE62002.2024.10592851>
- [7] Upadhyaya, N. (2024). *Artificial intelligence in web development: Enhancing automation, personalization, and decision-making*. International Journal of Advanced Research in Science Communication and Technology. <http://dx.doi.org/10.48175/IJARST-19367>
- [8] Tartu Ülikooli õppeinfosüsteem. Veebilehtede loomine edasijõudnutele. <https://ois2.ut.ee/#/courses/LTAT.03.015/details> (30.04.2024)
- [9] Moodle keskkond. Veebilehtede loomine edasijõudnutele. <https://moodle.ut.ee/course/view.php?id=8812> (10.10.2024)
- [10] Berield. (n.d.). *Web design explained*. Vaadatud: 30.03.2025. [Online] <https://berield.com/blog/web-design>
- [11] Turumogon P., Baharum A.. (2018). *Identifying A User Interface Web Design Standard for Higher Learning Institutions Using Kansei Engineering*. Indonesian Journal of Electrical Engineering and Computer Science, 11(1). 90-97. <http://doi.org/10.11591/ijeecs.v11.i1.pp90-97>

- [12] Wu, J., Peng, Y.-H., Li, X. Y., Swearngin, A., Bigham, J. P., & Nichols, J. (2024). *UIClip: A data-driven model for assessing user interface design*. In Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24) 1–16. ACM. <https://doi.org/10.1145/3654777.3676408>
- [13] Leonel, V, Morales Díaz. (2022). *What is a User Interface, again? A Survey of Definitions of UserInterface*. MexIHC '22: Proceedings of the 9th Mexican International Conference on Human-Computer Interaction. 2. 1-7. <https://doi.org/10.1145/3565494.3565504>
- [14] Visconti, E., Tsigkanos, C., Nenzi, L. (2025). *Automated monitoring of web user interfaces*. ACM Transactions on the Web, 10, 1–27. <https://doi.org/10.1145/3708512>
- [15] Junfeng, Z., Xu, Z., Wang, X., Lu, J. (2022). *A comparative research on usability and user experience of user interface design software*. International Journal of Advanced Computer Science and Applications, 13(8), 21–28. <http://dx.doi.org/10.14569/IJACSA.2022.0130804>
- [16] GeeksforGeeks. (2024). UI vs UX: What's the Difference between UI & UX Design. Vaadatud: 15.12.2024. [Online]. <https://www.geeksforgeeks.org/difference-between-ui-and-ux-design/>
- [17] Kumar, V., Kumar, V., Singh, S., Singh, N., Banoth, S. (2023). *The impact of user experience design on customer satisfaction in e-commerce websites*. International Journal for Research in Applied Science and Engineering Technology, 11(5), 4571–4575. <https://doi.org/10.22214/ijraset.2023.52580>
- [18] Grzybowski, A., Kupidura-Majewski, K. (2019). *What is color and how it is perceived?* Clinics in Dermatology, 37(5), 392–401. <https://doi.org/10.1016/j.clindermatol.2019.07.008>
- [19] Kurniawan, A. (2025). *The impact of golden ratio application on user satisfaction: A study on horizontal scrolling in user interface (UI) design*. International Journal of Human–Computer Interaction, 41(1), 445–451. <https://doi.org/10.1080/10447318.2023.2301254>
- [20] Kakembo, A., Anett (2025). *Psychology of color: Its influence on marketing and design*. Kampala International University Research Centre. 1 - 5
- [21] Guo, H., Wang, W., Song, F., Liu, Y., Liu, H., & Bo, Z. (2022). *The effect of color combinations on the efficiency of text recognition in digital devices*. International Journal of Human–Computer Interaction, 40(5), 1317–1326. <https://doi.org/10.1080/10447318.2022.2143768>

- [22] Mithun, A. M., Abu Bakar, Z., & Yafooz, W. M. S. (2019). *The impact of web contents color contrast on human psychology in the lens of HCI*. International Journal of Information Technology and Computer Science, 11(10), 27–33. <https://www.mecs-press.org/ijitcs/ijitcs-v11-n10/IJITCS-V11-N10-4.pdf>
- [23] Olejnik-Krugły, A., Tomaszewska, A., Dziśko, M., & Jankowski, J. (2021). *Towards effective visual communication with positive user experience: High contrast and visibility vs. user-friendliness/positive perception*. 1-12. <http://dx.doi.org/10.21203/rs.3.rs-725455/v1>
- [24] Ux&You. (2025). *Design for all: The importance of color contrast in accessibility*. Medium. Vaadatud: 18.02.2025. [Online]. <https://medium.com/%40uxandyouti/design-for-all-the-importance-of-color-contrast-in-accessibility-399ad9b557af>
- [25] Jongmans, E., Jeannot, F., Liang, L., & Dampérat, M. (2022). *Impact of website visual design on user experience and website evaluation: The sequential mediating roles of usability and pleasure*. Journal of Marketing Management, 38(17–18), 2078–2113. <https://doi.org/10.1080/0267257X.2022.2085315>
- [26] Xiao, S., Chen, Y., Song, Y., Chen, L., Sun, L., Zhen, Y., & Chang, Y. (2024). *UI semantic group detection: Grouping UI elements with similar semantics in mobile graphical user interface*. [Preprint]. arXiv. 1-16. <https://arxiv.org/html/2403.04984v1#S6>
- [27] SiBustamante, N. (2023). *What is Gestalt psychology? Theory, principles, & examples*. Simply Psychology. Vaadatud: 15.12.2024. [Online]. <https://www.simplypsychology.org/what-is-gestalt-psychology.html>
- [28] Huang, T. (2024). *FEAD: Figma-enhanced app design framework for improving UI/UX in educational app development*. [Preprint]. arXiv. 1-7. <https://arxiv.org/abs/2412.06793>
- [29] Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). *A century of Gestalt psychology in visual perception I: Perceptual grouping and figure-ground organization*. Psychological Bulletin, 138(6), 1172–1217. <https://psycnet.apa.org/doi/10.1037/a0029333>
- [30] FasterCapital. (2024). *Grouping UI elements effectively*. Vaadatud: 15.12.2024. [Online]. <https://fastercapital.com/content/Grouping-UI-Elements-Effectively.html>
- [31] Card, S. K. (1982). *User perceptual mechanisms in the search of computer command menus*. 1982 Conference on Human Factors in Computing Systems (CHI '82). ACM. <http://dx.doi.org/10.1145/800049.801779>

- [32] Kulczynski, A., & Hook, M. (2024). *Typography talks: Influencing vintage anemoia and product safety perceptions with vintage typography*. *Journal of Marketing*, 88(4), 129–153. <https://doi.org/10.1177/00222429231215357>
- [33] Günay, M. (2024). *The impact of typography in graphic design*. *International Journal of Eurasia Social Sciences*, 15(57), 1446–1464. <http://dx.doi.org/10.35826/ijjoess.4519>
- [34] Gevorgyan, G., & Porter, L. V. (2008). *One size does not fit all: Culture and perceived importance of web design features*. *Journal of Website Promotion*, 3(1–2), 25–38. <https://doi.org/10.1080/15533610802052589>
- [35] Abu Bakar, Z., & Zhiang, T. (2021). *Reading speed performance based on interface design elements for web interfaces*. *Annals of Emerging Technologies in Computing (AETiC)*, 5(5). <https://aetic.theiaer.org/archive/v5/v5n5/p18.html>
- [36] Bianchi, R. G., Rodrigues, K. R. da H., & Neris, V. P. de A. (2021). *Emotional responses to font types and sizes in web pages*. *Proceedings of the 20th Brazilian Symposium on Human Factors in Computing Systems (IHC '21)*. ACM. <https://dl.acm.org/doi/pdf/10.1145/3472301.3484325>
- [37] Stige, A., Zamani, E. D., Mikalef, P., & Zhu, Y. (2023). *Artificial intelligence (AI) for user experience (UX) design: A systematic literature review and future research agenda*. *Information Technology & People*, 37(6), 2324–2352. <https://doi.org/10.1108/ITP-07-2022-0519>
- [38] Digital Shift. (n.d.). *What is "the squint test"?* <https://digitalshiftmedia.com/marketing-term/the-squint-test/>
- [39] Bakaev, M., Goltsova, E., Khvorostov, V., & Razumnikova, O. (2019). *Data compression algorithms in analysis of UI layouts visual complexity*. *Perspectives of system informatics* 167–184. Springer. https://link.springer.com/chapter/10.1007/978-3-030-37487-7_14
- [40] Redacción Aguayo. (n.d.). *Squint test in UX*. Aguayo Blog. Vaadatud: 19.11.2024. [Online]. <https://aguayo.co/en/blog-aguayo-user-experience/squint-test-ux-ui/>
- [41] Interaction Design Foundation. (n.d.). *User interface (UI) design*. Vaadatud: 25.09.2024. [Online]. [https://www.interaction-design.org/literature/topics/ui-design#ui vs. user experience \(ux\) design-2](https://www.interaction-design.org/literature/topics/ui-design#ui_vs_user_experience(ux)_design-2)
- [42] Menges, R., Staab, S., Schaefer, C., Walber, T., & Kumar, C. (2025). *What did my users experience? Discovering visual stimuli on graphical user interfaces of the web*. *ACM Transactions on the Web*, 19(2). <https://doi.org/10.1145/3715881>

- [43] Benlarabi, A., Khtira, A., & El Asri, B. (2025). *User experience evaluation in software product lines*. *Ingénierie des Systèmes d'Information*, 30(1), 137–147. <https://doi.org/10.18280/isi.300111>
- [44] Morville, P., & Sullenger, P. (2010). *Ambient findability: Libraries, serials, and the internet of things*. *The Serials Librarian*, 58(1–4), 33–38. <https://doi.org/10.1080/03615261003622999>
- [45] Novák, J. Š., Masner, J., Benda, P., Šimek, P., & Merunka, V. (2024). *Eye tracking, usability, and user experience: A systematic review*. *International Journal of Human–Computer Interaction*, 40(17), 1616–1640. <https://doi.org/10.1080/10447318.2023.2221600>
- [46] Schlatter, T., & Levinson, D. (2013). *Visual usability: Principles and practices for designing digital applications*. Morgan Kaufmann. 1-322.
- [47] Sharma, H., & Tripathi, K. (2023). *The importance of website usability in digital marketing: A review*. *International Journal of Innovative Research in Computer Science & Technology*, 11(3), 27–31. <https://doi.org/10.55524/ijirest.2023.11.3.5>
- [48] Tashtoush, Y., Abu-El-Rub, N., Darwish, O., Al-Eidi, S., Darweesh, D., & Karajeh, O. (2023). *A notional understanding of the relationship between code readability and software complexity*. *Information*, 14(2), Artikkel 81. <https://doi.org/10.3390/info14020081>
- [49] Martin, R. C. (2008). *Clean code: A handbook of agile software craftsmanship*. Prentice Hall, 8.
- [50] Piantadosi, V., Fierro, F., Scalabrino, S., Serebrenik, A., & Oliveto, R. (2020). *How does code readability change during software evolution?* *Empirical Software Engineering*, 25, 5374–5412. <https://link.springer.com/article/10.1007/s10664-020-09886-9>
- [51] Batool, A., Bashir, M. B., Babar, M., Sohail, A., & Ejaz, N. (2021). *Effect of program constructs on code readability and predicting code readability using statistical modeling*. *Foundations of Computing and Decision Sciences*, 46(2), 127–145. <https://doi.org/10.2478/fcds-2021-0009>
- [52] Martin, R. C. (2008). *Clean code: A handbook of agile software craftsmanship*. Prentice Hall, 17-30.
- [53] Martinović, B., & Rozić, R. (2024). *Impact of AI tools on software development code quality*. In S. K. S. Gupta, T. Mastorakis, A. Yazici, & M. Russo (Eds.), *Digital transformation in education and artificial intelligence application (MoStart 2024)*. 241–256. Springer. https://doi.org/10.1007/978-3-031-62058-4_15

- [54] Ardito, L., Coppola, R., Barbato, L., & Verga, D. (2020). *A tool-based perspective on software code maintainability metrics: A systematic literature review*. Scientific Programming, 2020, Artikkel 8840389. <https://doi.org/10.1155/2020/8840389>
- [55] Khan, J. Y., & Uddin, G. (2023). *Automatic code documentation generation using GPT-3*. Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering (ASE '22). Artikkel 174. 1–6. ACM. <https://dl.acm.org/doi/pdf/10.1145/3551349.3559548>
- [56] Aversano, L., Guardabascio, D., & Tortorella, M. (2017). *Analysis of the documentation of ERP software projects*. Procedia Computer Science, 121, 423–430. <https://doi.org/10.1016/j.procs.2017.11.057>
- [57] Rai, S., Belwal, R. C., & Gupta, A. (2022). *A review on source code documentation*. ACM Transactions on Intelligent Systems and Technology, 13(5), Artikkel 84, 1–44. <https://dl.acm.org/doi/pdf/10.1145/3519312>
- [58] Andersen, G., & MoldStud Research Team. (2024). *The importance of code documentation in iOS development*. MoldStud. Vaadatud: 14.02.2025 <https://moldstud.com/articles/p-the-importance-of-code-documentation-in-ios-development>
- [59] Patil, N. H., Patel, S. H., & Lawand, S. D. (2023). *Research paper on artificial intelligence and its applications*. Journal of Advanced Zoology, 44(S8). <http://dx.doi.org/10.53555/jaz.v44iS8.3544>
- [60] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). *A comprehensive overview of large language models* [Preprint]. arXiv. <https://arxiv.org/abs/2307.06435>
- [61] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., ... & Shetty, S. (2020). *International evaluation of an AI system for breast cancer screening*. Nature, 577(7788), 89–94. <https://doi.org/10.1038/s41586-019-1799-6>
- [62] Karina Tolli. (2024). *Implementing GitHub Copilot Chat in Web Development: A Practical Comparison with Human Developer Work*. Tartu Ülikool. <https://thesis.cs.ut.ee/35dfa38b-57c9-40e3-b593-18b43a276ddf>
- [63] Grete Neilinn. (2023). *Website Development Process with ChatGPT4 and Without*. Tartu Ülikool. <https://thesis.cs.ut.ee/41c42620-eb23-457f-a63d-1dbaa3b3bd6>

- [64] Schultze, S., Withöft, A., Abdenebaoui, L., & Boll, S. (2023). *Explaining image aesthetics assessment: An interactive approach*. Proceedings of the 25th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '23). 20-28 ACM. <https://doi.org/10.1145/3591106.3592217>
- [65] Ghai, A. S., Rawat, V., & Ghai, K. (2024). *Artificial intelligence in system and software engineering for auto code generation*. Proceedings of the 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT). <https://doi.org/10.1109/ICEECT61758.2024.10738945>
- [66] Ding, H., Fan, Z., Guehring, I., Gupta, G., Ha, W., Huan, J., Liu, L., Omidvar-Tehrani, B., Wang, S., & Zhou, H. (2024). *Reasoning and planning with large language models in code development*. In Proceedings of the 2024 ACM SIGPLAN Symposium on Principles of Programming Languages (POPL '24). ACM. <https://dl.acm.org/doi/10.1145/3637528.3671452>
- [67] Impey, C., Wenger, M., Garuda, N., Golchin, S., & Stamer, S. (2024). *Using large language models for automated grading of student writing about science* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2412.18719>
- [68] Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). SAGE Publications. 220-234. https://us.sagepub.com/sites/default/files/upm-binaries/52063_00_Field_4e_SPSS_Prelims.pdf
- [69] Harris, J. E., Boushey, C., Bruemmer, B., & Archer, S. L. (2008). *Publishing nutrition research: A review of nonparametric methods, Part 3*. Journal of the American Dietetic Association. 1488-1496. <https://doi.org/10.1016/j.jada.2008.06.426>
- [70] Shutterstock. (n.d.). *Soft pink*. Vaadatud: 15.04.2025. [Online] <https://www.shutterstock.com/color/soft-pink>
- [71] Shao M., Basit A., Karri R. ja Shafique M.. (2024). *Survey of Different Large Language Model Architectures: Trends, Benchmarks, and Challenges*. 11. IEEE. <https://ieeexplore.ieee.org/document/10720163>
- [72] Yi, R., Tian, H., Gu, Z., Lai, Y.-K., & Rosin, P. L. (2022). *Towards artistic image aesthetics assessment: A large-scale dataset and a new method*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022). 22388–22397. IEEE. <https://doi.org/10.1109/CVPR52729.2023.02144>

- [73] Wang, Y., Webster, M. A., & Joyce, D. S. (2025). *Color statistics of images created by generative AI*. *Journal of the Optical Society of America A*, 42(5), B76–B80. <https://opg.optica.org/josaa/fulltext.cfm?uri=josaa-42-5-B76&id=567262>
- [74] Duarte, C., Matos, I., & Carriço, L. (2018). *Semantic content analysis supporting web accessibility evaluation*. Proceedings of the 15th International Web for All Conference (W4A '18). Artikkel 22. 1–4. ACM. <https://doi.org/10.1364/JOSAA.545030>
- [75] Koch, J., & Oulasvirta, A. (2016). *Computational layout perception using Gestalt laws*. Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16) 1423–1429. ACM. <https://doi.org/10.1145/2851581.2892537>
- [76] Chen, X., Shu, Y., Wang, R., Zhou, J., & Chen, W. (2024). *Large language model powered UI evaluation system*. *Journal of Graphics*, 45(6), 1178–1187.
- [77] Balfroid, M., Vanderose, B., & Devroey, X. (2024). *Towards LLM-generated code tours for onboarding*. Proceedings of the Third ACM/IEEE International Workshop on Natural Language-based Software Engineering (NLBSE '24) 65–68. ACM. <https://doi.org/10.1145/3643787.3648033>
- [78] Beckwith, L., & Burnett, M. M. (2004). *Gender: An important factor in end-user programming environments?* Proceedings of the 2004 IEEE Symposium on Visual Languages and Human-Centric Computing. IEEE. <http://dx.doi.org/10.1109/VLHCC.2004.28>
- [79] Briñol, P., Petty, R. E., & Barden, J. (2007). *Happiness versus sadness as a determinant of thought confidence in persuasion: A self-validation analysis*. *Journal of Personality and Social Psychology*, 93(5), 711–727. <https://doi.org/10.1037/0022-3514.93.5.711>

Lisad

I. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Artur Kašnikov ,
(*autori nimi*)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

Inim- ja tehisaruvõrdlus kursuse “Veebilehtede loomine ,
edasijõudnutele” lõpuprojektide näitel

(*lõputöö pealkiri*)

mille juhendaja(d) on Lidia Feklistova ,
(*juhendaja nimi*)

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;

2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commons'i litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;
3. olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;
4. kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Artur Kašnikov

15.05.2025