TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Volli-Fred Väränen

Veebilehtede hindamine tehisintellekti abil

Bakalaureusetöö (9 EAP)

Juhendaja: Mari-Liis Allikivi, MSc.

Tartu 2025

INFOLEHT

Töö pealkiri:

Veebilehtede hindamine tehisintellekti abil

Lühikokkuvõte:

Käesolevas bakalaureusetöös uuriti, kui täpselt suudavad erinevad tehisintellekti (TI) mudelid hinnata veebilehtede visuaalset esteetikat võrreldes inimeste hinnangutega. Töös analüüsiti nelja TI mudelit – *GPT-4o*, *Grok-2 Vision*, *Google Cloud Vision* ja *SAP* – kasutades kahte tüüpi küsimusi: lühikest üldhinnangut ja detailset hinnangut *VisAWI* kriteeriumite alusel. Tulemuste põhjal selgus, et kõige paremini kattusid inimeste hinnangutega *GPT-4o* tulemused detailse viiba korral. Mudelite hindamisvõimet mõõdeti korrelatsiooni, *AUROC* ja *RMSE* alusel. Lisaks analüüsiti mudelite stabiilsust ja kriitilisust. Töö panustab esteetilise kvaliteedi hindamise automatiseerimise uurimissuunale ning pakub võimalusi edasisteks rakendusteks disainiväärtuse analüüsis.

Võtmesõnad:

veebidisain, esteetika hindamine, tehisintellekt, *GPT-4o*, arvutinägemine, *VisAWI*

CERCS:

P170 – Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (Computer science, numerical analysis, systems, control)

Juhendaja:

Mari-Liis Allikivi, MSc

Thesis Title:

Evaluating Websites Using Artificial Intelligence

Abstract:

This bachelor's thesis investigates how accurately different artificial intelligence (AI) models can assess the visual aesthetics of websites compared to human judgments. The study analyzed four AI models – *GPT-4o*, *Grok-2 Vision*, *Google Cloud Vision*, and *SAP* – using two types of evaluation prompts: a short general score and a detailed score based on *VisAWI* criteria. The results showed that *GPT-4o*, when given a detailed prompt, most closely matched human evaluations. Model performance was evaluated using correlation, *AUROC*, and *RMSE* metrics. In addition, the stability and criticality of the models were analyzed. The study contributes to the field of automated aesthetic quality assessment and provides practical insights for future applications in design value analysis.

Keywords:

web design, aesthetic evaluation, artificial intelligence, *GPT-4o*, computer vision, *VisAWI*

CERCS:

P170 – Computer science, numerical analysis, systems, control

Supervisor:

Mari-Liis Allikivi, MSc

1. Sissejuhatus

Veebilehed on tänapäeva digitaalses maailmas olulised suhtlus- ja ärivahendid, olles sageli esimene kokkupuutepunkt organisatsioonide ja nende sihtrühmade vahel. Veebilehe visuaalne esteetika mängib kasutajakogemuse kujundamisel võtmerolli, mõjutades usaldusväärsuse, kasutatavuse ja üldise rahulolu tajumist [7, 10]. Kuigi esteetika tähtsus on ilmne, on selle hindamine tihti subjektiivne ning puuduvad standardiseeritud ja automatiseeritavad meetodid, eriti suure hulga veebilehtede analüüsimisel.

Käesoleva lõputöö ajend sai alguse autori praktilisest kogemusest töötades tarkvaraarendusagentuuris Nexbit, kus üheks ülesandeks oli hinnata klientide olemasolevate veebilehtede visuaalset kvaliteeti, et teha neile ettepanek kodulehe uuenduseks. Töö käigus ilmnes probleem: hinnangud veebilehtedele sõltusid suuresti subjektiivsest maitsest ning protsess oli ajamahukas. See tõstatas küsimuse, kas visuaalset hindamisprotsessi oleks võimalik vähemalt osaliselt automatiseerida. Eriti suuremahuliste auditite või turundusanalüüside puhul oleks kasulik tööriist, mis aitaks veebilehti kiiresti ja objektiivselt kategoriseerida.

Üheks võimalikuks lahenduseks näib olevat tehisintellektil (TI) põhinev hindamissüsteem, mis suudaks võrreldavalt inimesele anda esteetilise hinnangu veebilehe visuaalsele kujundusele. Töö keskne küsimus on, kui täpselt suudavad erinevad TI-mudelid hinnata veebilehtede esteetikat ning kuivõrd need hinnangud kattuvad inimeste arvamusega.

Töö keskendub ainult veebilehtede staatilistele kuvatõmmistele, mis võimaldavad standardiseeritud ja tehniliselt lihtsat sisendit TI-mudelitele. Enamik visuaalsest esmamuljest tekib just sellise staatilise vaate põhjal [5]. Kuigi selline lähenemine välistab interaktsiooni ja dünaamilised elemendid, võimaldab see hinnata esmamulje seisukohalt olulisi aspekte, nagu paigutus, värvikasutus ja kujundus.

TI-põhist esteetika hindamist saab potentsiaalselt rakendada mitmes valdkonnas – näiteks e-kaubanduses (tootelehtede visuaalne hindamine), *UI/UX* auditeerimisel (kasutajakogemuse probleemide avastamine), visuaalse sisu loomisel (kujundusalgoritmid) ja veebiarenduses (kvaliteedikontroll). Siiski kaasnevad mitmed väljakutsed: esteetika põhiselt treenitud andmestikud on tihti kallutatud, kuvatõmmised ei pruugi peegeldada kogu lehe kogemust ning

keerukamate mudelite otsuseid on sageli raske selgitada – see piirab nende praktilist kasutust professionaalses disainitöös.

Senised *IAQA* (Image Aesthetic Quality Assessment) uuringud on valdavalt keskendunud staatiliste fotode ja visuaalide hindamisele. Veebilehtede analüüs on visuaalselt ja semantiliselt keerulisem, kuna disaini mõjutavad lisaks pildilistele omadustele ka tekst, navigeerimine ja kompositsiooni dünaamika. Staatilise sisendi kasutamine on käesolevas töös teadlik kompromiss, mille eesmärk on tagada võrreldav ja kontrollitav sisend TI-mudelitele. See võimaldab hinnata just kujunduse üldmuljet ja kriitilisi visuaalseid tegureid, mis loovad esmamulje.

Varasemad uuringud [7] viitavad, et disainerid hindavad oma töid optimistlikumalt kui kasutajad, mistõttu võib TI-põhine analüüs suurte andmehulkade pealt aidata tasakaalustada subjektiivseid hinnanguid ning toetada otsustusprotsesse.

Tööd juhivad järgmised uurimisküsimused:

1. Kuidas hindavad erinevad TI-mudelid veebilehtede esteetikat võrreldes inimeste hinnangutega?
2. Millised TI-mudelid suudavad kõige täpsemini eristada visuaalselt kvaliteetseid („häid“) ja visuaalselt puudujääkidega („probleemseid“) veebilehti?
3. Kuidas mõjutab viiba keerukus TI-mudelite esteetilisi hinnanguid ja nende võimet veebilehti klassifitseerida?

Töö on üles ehitatud järgmiselt:

2. peatükk annab teoreetilise tausta veebidisaini põhimõtete ja TI rolli kohta visuaalses analüüsis.

3. peatükk kirjeldab metoodikat, sealhulgas veebilehtede valikut, TI-mudelite kasutamist, inimeste küsitlust ja andmeanalüüsi.

4. peatükk esitab tulemused,

5. peatükk arutleb nende üle ning

6. peatükk võtab kokku peamised järeldused ja annab soovitusi edasisteks uuringuteks.

2. Teoreetiline taust

2.1 Veebilehtede visuaalne esteetika ja disainipõhimõtted

Veebilehe visuaalne esteetika mõjutab kasutajakogemust, usaldust ning sageli ka seda, kas kasutaja lehele jääb või lahkub [5, 7, 10]. Seetõttu on oluline, et disainipõhimõtteid rakendataks teadlikult. Veebidisaini esteetiline kvaliteet ei sõltu ainult maitsest, vaid põhineb mitmetel üldtunnustatud visuaalsetel põhimõtetel.

Visual Aesthetics of Websites Inventory (*VisAWI*) [8] on psühholoogilisel uurimistööl põhinev hindamisraamistik, mille eesmärk on mõõta veebilehe esteetilist atraktiivsust. See koosneb neljast aladimensioonist: lihtsus, mitmekesisus, värvikus ja viimistletus. *VisAWI* [8] võimaldab esteetikat hinnata nii kasutajaküsitluste kui ka AI mudelite struktuuris, luues sidusa aluse inimlike ja automatiseeritud hinnangute võrdlemiseks.

Tähtsamad esteetilised disainipõhimõtted, mida arvestatakse nii teaduskirjanduses kui *VisAWI* mõõtmisskaalas [8], on järgmised:

1. **Lihtsus** – puhas paigutus, loogiline struktuur, selge navigatsioon ja piisav tühiruum [3]. Lihtsus aitab vähendada kognitiivset koormust ja toetab kasutatavust.
2. **Värviharmoonia** – meeldivad ja kooskõlalised värvikombinatsioonid, mis ei häiri tähelepanu ja toetavad sisu [1].
3. **Järjepidevus** – ühtne kujunduskeel kogu lehe ulatuses – sama stiil, fondid ja paigutus [1, 3].
4. **Visuaalne huvi** – elementide mitmekesisus ja loovus, mis köidavad tähelepanu ja loovad esteetilise dünaamika [8].

Lisaks on Uribe jt [11] uurimuses leitud, et veebilehtede visuaalne esteetika – st kui meeldiv või atraktiivne kujundus inimestele tundub – on tugevalt seotud lihtsate madala taseme visuaalsete omadustega nagu heledus, kontrastsus ja värvide jaotus. Selliseid omadusi saab kvantitatiivselt mõõta ka automaatselt. Üheks varasemaks standardiseeritud kirjeldussüsteemiks on *MPEG-7* (Moving Picture Experts Group standard number 7), mis määratleb ISO/IEC tasemel multimeediasisu deskriptorid, sealhulgas värvijaotuse, tekstuuri ja kuju kirjeldamiseks [6].

Kuigi *MPEG-7* standardi konkreetseid deskriptoreid käesolevas töös otseselt ei kasutatud, on selle aluspõhimõtted võrreldavad *Google Cloud Vision* API toimeloogikaga, mis samuti hindab visuaalseid omadusi madala taseme tunnuste kaudu (nt kontrast, värviharmoonia, elementide joondus). See loob tausta ja põhjenduse, miks selline lähenemine võib anda informatiivset sisendit veebilehtede esteetika hindamisel ka tehisintellekti mudelite kõrval.

2.2 Visuaalse veebidisaini trendid 2025. aastal

Lisaks klassikalistele disainipõhimõtetele, nagu lihtsus, värviharmoonia või visuaalne tasakaal, mõjutavad veebilehe esteetikat ka ajas muutuvad trendid. Need peegeldavad kasutajate ootusi ja tehnoloogilisi võimalusi, aga ka üldist visuaalset maitsemeelt, mis ajas muutub. Just need suunad määravad ära, milline veebileht tundub „kaasaegne“ ja milline pigem aegunud.

Kuna käesolevas töös oli eesmärk võrrelda „häid“ ja „halbu“ veebilehti, siis oli oluline esmalt selgeks teha, mis üldse teeb ühe veebilehe visuaalselt heaks aastal 2024 või 2025. Selleks kaardistati viimaste aastate levinumad disainitrendid.

Kõik alltoodud suunad põhinevad TheeDigitali [15] ülevaatel:

1. **Minimalistlik maksimalism** – ühendab puhta ja lihtsa paigutuse mõjuvate visuaalsete elementidega, nt suured värviplokid või tugevad kujundid.
2. **Väljendusrikas tüpograafia** – kasutatakse suuri, omanäolisi fonte, sageli ka käsitsi joonistatud kirjatüüpe.
3. **Rahustavad värvipaletid** – sügavad ja soojad toonid (nt „Mocha Mousse“), mis loovad usaldusväärse ja tasakaalus mulje.
4. **Anti-disain** – teadlikult „ebakorrapärane“ või vaba stiil, kus kasutatakse asümmeetrilist paigutust ja ebatavalisi jooni.
5. **Kohandatud illustratsioonid** – valmivad spetsiaalselt konkreetse veebilehe jaoks, asendades tüüpilised stock-fotod.
6. **Tume režiim ja kõrge kontrastsus** – must taust, valge tekst, suured kontrastid – kaasaegne, silmasõbralik ja levinud.

7. **Visuaalne kihilisus ja tekstuurid** – taustadele ja elementidele lisatakse kihte, varjusid ja tekstuure, mis annavad sügavuse ja liikumise tunde.

Need trendid mängisid rolli ka selles, kuidas veebilehed selles uurimistöös „heaks“ või „halvaks“ liigitati. Just need visuaalsed omadused on need, mida inimesed tänapäeval esteetiliseks peavad – ja mida TI mudelitel tuli samuti ära tunda.

2.3 Tehisintellekti roll veebilehe esteetika analüüsis

Veebilehtede esteetilise kvaliteedi hindamine on traditsiooniliselt olnud subjektiivne ja ajamahukas, sõltudes inimeste visuaalsest tajust, kogemusest ning isiklikest eelistustest. Tehisintellekti (TI) areng, eriti süvaõppe ja arvutinägemise valdkonnas, võimaldab seda protsessi nüüd automatiseerida, pakkudes potentsiaalselt objektiivsemat ja skaleeritavamat lähenemist.

Viimastel aastatel on välja töötatud mitmeid TI-l põhinevaid lahendusi, mis suudavad üsna usaldusväärselt prognoosida inimeste esteetilisi eelistusi. Enamik varasemaid töid on keskendunud pildipõhistele mudelitele, mis kasutavad konvolutsioonilisi närvivõrke (*CNN*-e). [2] Näiteks NIMA (Neural Image Assessment) mudel kasutab märgistatud pildikogumeid, nagu AVA, et ennustada pildi esteetilist hinnangut skaalal 1–10 [9].

Sarnast lähenemist on rakendatud ka veebilehtede puhul. *Webthetics*-mudel [4] treeniti spetsiaalselt veebilehtede kuvatõmmiste peal, kasutades eelnevalt kogutud inimeste hinnanguid. Selle eesmärk oli ennustada esteetilist hinnangut ainult pildi põhjal – ilma tekstilise või semantilise sisendita. Tulemused näitasid tugevat korrelatsiooni kasutajate hinnangutega ($r = 0.85$), võimaldades kiiret ja automaatset hindamist. Samas piirdus mudel madalatasemeliste visuaalsete omadustega ega arvestanud sisu tähendust ega konteksti.

Käesolevas töös ei kasutatud *Webthetics*-mudelit, kuna see ei ole avalikult kättesaadav, selle treeningandmestik puudub ning mudelit on keeruline reprodutseerida. Lisaks oli töö eesmärgiks võrrelda mitmesuguseid üldotstarbelisi ja hõlpsasti ligipääsetavaid TI-mudeleid, sealhulgas multimodaalseid süsteeme, mis võimaldavad sisestada loomulikus keeles küsimusi ning hinnata

disaini mitmest vaatenurgast. Seetõttu eelistati mudeleid, mida sai kasutada läbi standardsete API-de või avalike platvormide, tagades töö reprodutseeritavuse ja praktilise rakendatavuse.

CNN-idel põhinevad mudelid õpivad visuaalseid mustreid otse toorpikslitest, ilma et oleks vaja eelnevalt käsitsi määratleda olulisi tunnuseid. Varasemates kihtides tuvastatakse madalatasemelised omadused nagu värv, tekstuur ja servad, sügavamates kihtides kujunevad keerukamad struktuurid, näiteks paigutuse tasakaal, rütm või visuaalne hierarhia. Tänu sellele suudavad *CNN*-id kohanduda erinevate disainikeelte ja kultuurikontekstidega paremini kui reeglipõhised süsteemid.

Viimastel aastatel on tähelepanu liikunud ka suurte multimodaalsete mudelite suunas, nagu *GPT-4o* või *Grok-2 Vision*. Need mudelid suudavad töödelda korraga nii visuaalset kui ka tekstilist sisendit, võimaldades hinnata veebilehe kuvatõmmist koos sellele lisatud küsimusega. Näiteks saab neilt küsida: „Kui esteetiline on see veebileht skaalal 0–10?“ või paluda hinnata konkreetseid disainiaspekte *VisAWI* skaalal [8]. Selline lähenemine võimaldab paindlikumat analüüsi ja arvestab ka keerukamate, kontekstitundlike olukordadega, mida pelgalt visuaalsetele tunnustele tuginevad mudelid ei suuda hästi lahendada.

Kokkuvõttes pakuvad nii *CNN*-idel põhinevad spetsialiseeritud mudelid kui ka üldotstarbelised multimodaalsed süsteemid erinevaid võimalusi veebidisaini esteetika automatiseeritud hindamiseks. Esimesed võimaldavad suure täpsusega hinnata struktureeritud visuaalset sisu, samas kui viimased toetavad mitmekihilist hinnangut, kasutades loomulikku keelt ja semantilist mõistmist.

Pärast teoreetiliste lähtekohtade ja varasemate uurimuste käsitlemist liigume edasi töö empiirilise osa juurde. Järgnevates alapeatükkides tutvustatakse kasutatud metoodikat, töövahendeid ning esteetiliste hinnangute kogumise ja analüüsi protsessi.

3. Metoodika

Pärast olemasolevate mudelite ja käsitluste ülevaadet suundume nüüd konkreetse uurimismetoodika juurde.

3.1 Veebilehtede valik

Uuringus analüüsiti kokku 104 veebilehte, mis jagati kaheks võrdseks kategooriaks: 52 visuaalselt „head“ ja 52 visuaalselt „halba“ disaini. Veebilehtede klassifikatsioon toimus visuaalse esmamulje ja objektiivsete esteetiliste kriteeriumite (nt paigutus, värvilahendus, tüpograafia, üldine atraktiivsus) alusel (vt peatükk 2.3).

Autoril on vastav kogemus, kuna ta töötas tarkvaraarendusagentuuris, kus tema ülesandeks oli hinnata ettevõtete olemasolevaid veebilehti ning pakkuda neile visuaalse ja funktsionaalse uuenduse vajaduse põhjal välja uue veebilehe lahendusi. Selle töö käigus kujunes välja praktiline kogemus veebilehtede visuaalse kvaliteedi hindamisel, mille põhjal oli võimalik usaldusväärselt eristada nüüdisaegset ja vananenud kujundust.

"Head" veebilehed hõlmasid peamiselt auhinnatud ja tunnustatud veebilehti, sealhulgas näiteks Kuldmuna konkursil tunnustatud töid ning muid kaasaegsete disainitrendide järgi kujundatud veebisaite. Valiku tegemisel arvestati, et lehed järgiksid 2025. aasta veebidisaini visuaalseid suundumusi, nagu minimalistlik maksimalism, sügavad ja rahustavad värvipaletid ning kohandatud illustratsioonide kasutamine.

Näited „headest“ ja „halbadest“ veebilehtedest on esitatud lisas B (vt Lisa B). Näited illustreerivad tüüpilisi visuaalseid omadusi, mille põhjal veebilehed klassifitseeriti – sealhulgas kaasaegsete disainitrendide järgimist või nende puudumist, värvikasutust, paigutust ja kujunduse terviklikkust.

"Halvad" veebilehed valiti tööga seotud müügikontaktide nimekirjast. Tegemist oli veebilehtedega, mida hinnati visuaalselt aegunuks, katkise paigutuse, ebajärjekindla

värvikasutuse või üldise visuaalse atraktiivsuse puudumise tõttu. Need veebilehed olid varasemalt tuvastatud potentsiaalsete klientidena, kellele pakkusin veebilehe uuendamise ja moderniseerimise teenust.

Veebilehed koguti ja hinnati lähtudes järgmistest põhimõtetest:

- Kõik hinnatud lehed olid avalikult ligipääsetavad.
- Kõik lehed hinnati samal ajavahemikul, et vältida veebilehtede disainis tehtud muutusi uuringu jooksul.
- Hindamisel keskenduti ainult visuaalsetele ja esteetilistele omadustele, jättes kõrvale funktsionaalsuse, sisu ja tehnilised aspektid.

Selline valikuprotsess võimaldas luua esindusliku ja kontrastse andmestiku, mis oli sobilik TI mudelite ja inimeste hinnangute võrdlemiseks esteetika perspektiivist.

3.2 Vahendid

Pärast veebilehtede valikut alustati andmete töötlemiseks vajalike tööriistade ja meetodite rakendamist. Kõik tehniline töötlus toimus Pythonis, kasutades erinevaid API-sid ja masinõppe mudeleid. Käesolevas peatükis keskendutakse ainult esteetika ja visuaalse kujunduse hindamise töövahenditele, jättes kõrvale funktsionaalsuse ja kasutatavusega seotud aspektid.

3.2.1 Kuvatõmmised

Veebilehtede visuaalseks hindamiseks kasutati kuvatõmmiseid, mis loodi tööprotsessi käigus *ScreenshotOne* API abil. Tegemist on veebipõhise teenusega, mis võimaldab automatiseeritult genereerida kõrgekvaliteedilisi ja täpseid ekraanipilte. *ScreenshotOne* sobib hästi uurimusliku töövoo osaks tänu oma skaleeritavusele ja turvalisele HTTPS-protokollile [14].

Teenuse abil tagati, et kõik kuvatõmmised loodi ühtse meetodiga ja standardsetes tingimustes, mis oli oluline, et vältida visuaalse sisendi kallutatust ja võimaldada TI-mudelite vahel õiglast

võrdlust. Kuvatõmmised esindasid veebilehtede visuaalset olekut konkreetsel hetkel ning need toimisid sisendmaterjalina kõigi mudelite analüüsiks.

Siiski kaasnevad staatiliste kuvatõmmiste kasutamisega mitmed tehnilised piirangud. Esiteks võib pildi fikseerimine eemaldada või moonutada teatud disainielemente – näiteks dünaamilisi komponente, interaktiivseid elemente või paindlikke paigutusi. Teiseks võivad kaduda olulised aspektisuhted ja paigutuse kontekstid, mis mõjutavad veebilehe üldmuljet.

3.2.2 TI mudelid

Uuringus kasutati erinevaid olemasolevaid tehisintellekti mudeleid, mis võimaldavad hinnata veebilehtede visuaalset esteetikat väga erinevate meetodite ja analüüsitasandite kaudu – alates lihtsatest pildipõhistest skooridest kuni keeruka multimodaalse sisuanalüüsini. Mudelid valiti teadlikult mitmekesise lähenemisviisi katmiseks, et hinnata, millised olemasolevad tehnoloogiad sobivad kõige paremini visuaalse esteetika hindamiseks.

Oma mudeli treenimisest loobuti, kuna see oleks eeldanud mahukat ja usaldusväärselt märgendatud esteetikaandmestikku, mille loomine ületab käesoleva bakalaureusetöö mahu ja eesmärgi. Selle asemel keskenduti juba olemasolevate, avalikult kättesaadavate ja praktikas rakendatavate mudelite võrdlusele.

Alustatakse mudelitest, mis kasutavad vaid ühte andmeliiki (nt pilti), ning liigutakse edasi keerukamate mudeliteni, mis suudavad töödelda mitut sisendvormingut ja teostada semantilist analüüsi.

Kasutatud mudelid olid järgmised:

1. ***Simple Aesthetics Predictor*** – spetsiaalselt esteetilise atraktiivsuse hindamiseks treenitud mudel Hugging Face'i platvormil. Mudel põhineb *CLIP* arhitektuuril ja on loodud esteetika skoori ennustamiseks pildi põhjal [17].

2. ***Google Cloud Vision API*** – madala taseme visuaalsete omaduste tuvastamise tööriist, mis kasutab eeldefineeritud reegleid (nt värvijaotus, kontrastsus, joondus), mitte õppivat esteetikamudelit [12].
3. *xAI "grok-2-vision-1212"* – visuaalanalüüsile optimeeritud generatiivne mudel. Mudeli kohta pole teadusartiklit, kuid xAI ametlik tutvustus ja blogipostitus annavad ülevaate Grok 2 mudelite töövõimest [16].
4. *OpenAI "GPT-4o"* – uusim OpenAI multimodaalne mudel, mis suudab töödelda nii pilte, teksti kui heli [13].

Kõik valitud mudelid esindavad erinevaid lähenemisviise esteetika hindamisele – alates pildipõhisest skoorist kuni keeruka multimodaalse analüüsini – võimaldades mitmekülgset võrdlust ning pakkudes tugevat alust tulemuste analüüsimiseks järgnevas peatükis.

Simple Aesthetics Predictor (SAP)

SAP on Hugging Face'i platvormil kättesaadav mudel, mis on treenitud esteetilise atraktiivsuse hindamiseks ainult pildipõhiselt. See põhineb *CLIP* arhitektuuril, kuid kasutab ainult visuaalset komponenti ega arvesta keelelise ega kontekstilise sisendiga. Mudel analüüsib pilti kui tervikut ning hindab selle atraktiivsust inimlike esteetikamustrite põhjal [17].

SAP-ile esitati kõigi 104 veebilehe kuvatõmmised pildiformaadis. Mudel tagastas iga sisendi kohta üldise esteetikaskoori skaalal 0–10. Aspektipõhiseid hinnanguid ega viiba sisendeid *SAP* ei toeta. *SAP* tulemusi kasutati võrdlusbaasina teiste keerukamate mudelite ja inimeste hinnangutega.

Google Cloud Vision API (GCV)

Google Vision API on reeglipõhine tööriist, mis hindab pilte mitmete eeldefineeritud visuaalsete tunnuste põhjal, nagu:

- värviharmoonia,

- teksti kontrastsus,
- elementide joondus,
- tühja ruumi kasutus.

Erinevalt süvaõppemudelitest ei põhine GCV inimeste tajust õpitud esteetikamustritel, vaid staatilistel heuristilistel reeglitel [12].

GCV API-le edastati iga veebilehe kuvatõmmis. API tagastas neli visuaalset skoori, mis normaliseeriti skaalale 0–10 ja kombineeriti üldskooriks. Tulemusi kasutati, et hinnata, kui hästi suudab visuaalsetel reeglitel põhinev mudel korreleeruda inimhinnangutega ja eristada kvaliteetseid ning nõrku veebilehti.

xAI "grok-2-vision-1212"

xAI grok-2-vision on süvaõppel põhinev generatiivne mudel, mis on optimeeritud visuaalse sisendi analüüsiks. Mudel suudab tuvastada disainis mustreid, värvikasutust, kompositsiooni ja kujunduselementide tasakaalu. Samuti toetab mudel tekstilisi juhiseid (viipasid), võimaldades anda pildi põhjal esteetilisi hinnanguid mitmele mõõtmele (nt lihtsus, värvikus, viimistletus) vastavalt antud käsule [16].

Mudelile esitati kõikide veebilehtede kuvatõmmised base64-vormingus. Rakendati kahte tüüpi viipa:

- **Lihtne viip:** hinnata veebilehe üldist visuaalset atraktiivsust skaalal 0–10.
- **Detailne viip:** hinnata nelja aspekti (lihtsus, mitmekesisus, värvikus, viimistletus) ja arvutada nende põhjal üldskoor.

Mõlema tüübi viibad olid samad, mis kasutati *GPT-4o* mudeli puhul (vt Lisa C). Mudelit jooksutati viis korda iga sisendi kohta, et mõõta tulemuste stabiilsust. Keskmised skoorid salvestati ja kasutati edasiseks korrelatsiooni- ja *ROC*-analüüsiks.

OpenAI "GPT-4o"

GPT-4o on OpenAI loodud mitmemodaalne mudel, mis ühendab pildianalüüsi ja keelemõistmise. Mudel suudab töödelda korraga visuaalset ja tekstilist sisendit ning pakkuda ka loogilisi ja verbaalseid selgitusi oma otsuste kohta. *GPT-4o* analüüsib esteetikat nii numbriliselt kui sõnaliselt [13].

Veebilehtede kuvatõmmised kodeeriti base64-formaadis ja edastati mudelile koos kahe viibaga:

- **Lihtne viip:** üldine esteetikaskoor skaalal 0–10.
- **Detailne viip:** hinnang neljale *VisAWI* kategooriale + üldskoor.

GPT-4o tagastas nii üld- kui aspektipõhised skoorid ning vajadusel ka põhjendused kujunduse kohta. Mudelit jooksutati viis korda iga sisendi kohta. Saadud keskmisi väärtusi kasutati mudelitevahelises võrdluses, *ROC*-analüüsis ning korrelatsioonivõrdluses inimeste hinnangutega.

Mudelid kokkuvõtvalt

- **Lihtsad mudelid** nagu *SAP* pakuvad kiiret üldist hinnangut, kuid ei võimalda täpsustada hinnangu põhjuseid ega töötada erinevate aspektidega.
- **Reeglipõhised tööriistad** (*GCV*) tuginevad hästi määratletud tunnustele, kuid ei arvesta visuaalset konteksti ega subjektiivseid esteetikamustreid.
- **Süvaõppemudelid** nagu grok-2-vision suudavad tuvastada keerukaid visuaalseid mustreid ning toetavad ka tekstilist sisendit, kuid nende põhjendused võivad olla vähem läbipaistvad või varieeruvamad võrreldes teiste multimodaalsete süsteemidega.
- **Multimodaalsed mudelid** nagu *GPT-4o* võimaldavad mitmetasandilist analüüsi, ühendades pildilise ja verbaalse sisendi, pakkudes seega kõige täpsemat ja põhjendatumat lähenemist esteetika hindamiseks.

Selline mitmekihiline lähenemine võimaldas töö raames hinnata erinevate tehnoloogiate sobivust veebilehtede esteetilise kvaliteedi automaatseks hindamiseks ning võrrelda tulemusi inimeste hinnangutega.

3.2.3 Küsitlus

Veebilehtede esteetilise hindamise täpsuse hindamiseks viidi läbi küsitlus (Lisa F), milles osales viis inimest, kellel oli kogemus veebidisaini ja disaini valdkonnas. Osalejate hulgas oli kolm professionaalset disainerit ning kaks inimest, kellel oli varasem praktiline kogemus veebilehtede kujundamisega. Küsitlus viidi läbi Google Formsi platvormil.

Kõik viis osalejat hindasid täpselt samu 20 veebilehte, et tagada hinnangute võrreldavus ja usaldusväärsus. Lehed valiti eelnevalt autori poolt ning esitati igale osalejale samas järjekorras. Autor oli veebilehed eelnevalt jaotanud kaheks võrdseks kategooriaks: 10 visuaalselt „head“ ja 10 visuaalselt „halba“ lehte. Hindamine toimus skaalal 0 kuni 10, kus 0 tähistas väga halba ja 10 väga head visuaalset atraktiivsust. Osalejad ei olnud teadlikud, millised lehed olid eelnevalt kategoriseeritud "heaks" või "halvaks", et vältida hinnangute kallutatust.

Veebilehtede jaotamisel „heaks“ ja „halvaks“ püüti valik teha võimalikult üheselt mõistetavaks. "Head" veebilehed olid kaasaegsed, järgides tänapäevaseid disainitrende ning olles tehniliselt ja visuaalselt korrektsed. "Halvad" veebilehed olid seevastu visuaalselt aegunud, katkiste paigutuste või disainivigadega ning ei vastanud enam tänapäevastele kasutajakogemuse standarditele.

Kogutud hinnanguid võrreldi erinevate TI-mudelite poolt samadele veebilehtedele antud skooridega. Inimeste ja TI hinnangute korrelatsiooni hindamiseks kasutati Pearsoni korrelatsioonikordajat.

3.3 Analüüsimeetodid

TI-mudelite ja inimeste hinnangute võrdlemiseks viidi läbi kvantitatiivne analüüs, mille eesmärk oli hinnata mudelite usaldusväärsust, täpsust ja vastavust inimtajule. Selleks kasutati järgmisi mõõdikuid ja meetodeid:

3.3.1 Hindamiste stabiilsus (*ICC* – Intraclass Correlation Coefficient)

Hinnati, kui järjepidevalt suudavad mudelid anda sarnaseid hinnanguid samadele sisenditele korduvate jooksutuste korral. Kasutati **Intraclass Correlation Coefficient (*ICC*)** meetodit, mis mõõdab, kui suur osa hinnangute kogu varieeruvusest tuleneb hinnatavatest objektidest, mitte juhuslikest mõõtmisvigadest.

Valem:

$$\mathrm{ICC}(1,k) = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W}$$

kus:

1. MS_B – objektidevaheline dispersioonikeskmine,
2. MS_W – objektisisene dispersioonikeskmine,
3. k – mõõtmiste arv (antud juhul 5 jooksutust).

ICC väärtused jäävad vahemikku 0 kuni 1:

1. 0.90 – suurepärane stabiilsus,
2. 0.75–0.90 – hea stabiilsus,
3. < 0.50 – kehv stabiilsus.

3.3.2 Pearsoni korrelatsioonikordaja (r)

Hinnati, kuivõrd kattuvad TI mudeli ja inimeste antud skooride väärtused. Selleks kasutati **Pearsoni korrelatsioonikordajat**, mis mõõdab lineaarset seost kahe pideva tunnuse vahel.

Valem:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Tõlgendus:

1. $r = 1$: täiuslik positiivne korrelatsioon,
2. $r = 0$: puudub lineaarne seos,
3. $r = -1$: täiuslik negatiivne korrelatsioon.

3.3.3 *ROC*-analüüs ja *AUROC*

Mudelite klassifitseerimisvõime hindamiseks kasutati ***ROC*-kõverat** (Receiver Operating Characteristic), mis visualiseerib mudeli võimet eristada kahte klassi („hea“ vs „halb“ veebileht). *ROC*-kõveral võrreldakse õigesti positiivsete juhtude tuvastamise määra (tõelised positiivid) ja valepositiivsete juhtude määra (valed positiivid), et hinnata, kui täpselt mudel suudab eristada kahte kategooriat.

AUROC **(*Area Under ROC Curve*)** väljendab *ROC*-kõvera alla jäävat pindala ja annab mudeli üldise võimekuse klasside eristamisel:

1. ***AUROC* = 1.0** – täiuslik eristaja,
2. ***AUROC* = 0.5** – juhuslik klassifitseerija,
3. ***AUROC* < 0.5** – ebausaldusväärne mudel.

3.3.4 *RMSE* (Root Mean Square Error)

RMSE mõõdab, kui kaugel on TI mudeli antud hinnangud inimese hinnangutest numbrilisel skaalal. See on tundlik suurtele kõrvalekalletele ning väljendab absoluutset viga.

Valem:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$$

kus:

1. x_i – mudeli hinnang,
2. y_i – inimese hinnang,
3. n – hinnatud juhtumite arv.

Mida väiksem *RMSE*, seda täpsem on mudel. Erinevalt korrelatsioonist hindab *RMSE* ainult arvulist lähedust, mitte muutumismustrit.

3.3.5 Viiba mõju analüüs

Uuriti, kuidas mõjutab TI mudelite käitumist **sisendi sõnastus**. Kasutati kahte tüüpi viipa:

1. **Lühike viip**: palus anda ainult üldskoori (0–10),
2. **Detailne viip**: palus hinnata nelja aspekti (lihtsus, mitmekesisus, värvikus, viimistletus), mille põhjal arvutati üldskoor.

Võrdluses kasutati samu sisendeid ja mõõdeti:

1. erinevusi mudeli väljundites,
2. erinevusi korrelatsioonis inimhinnangutega,
3. erinevusi *ROC* ja *AUROC* väärtustes.

Viiba mõju analüüs võimaldas hinnata mudeli tundlikkust juhiste täpsusele ning selle seoseid kriitilisuse ja eristusvõimega.

3.3.6 Kirjeldav statistika

Kasutati järgmisi kirjeldavaid mõõdikuid:

1. **Keskmine (mean)** – üldine hinnangutaseme näitaja,
2. **Standardhälve (SD)** – hinnangute hajuvus,
3. **Hajusdiagrammid ja histogrammid** – visuaalseks võrdluseks.

Need meetodid võimaldasid võrrelda erinevate mudelite hinnangute jaotust ning kontrasti mudelite vahel.

Kõik töö käigus koostatud Google Colaboratory vihikud, mis sisaldavad andmete kogumist, eeltöötlust, AI-mudelite kasutamist ja analüüsi, on kättesaadavad Lisa E-s viidatud lingi kaudu. Analüüsid viidi läbi Pythonis, kasutades järgmisi teeke: **pandas**, **scipy**, **sklearn**, **matplotlib** ja **seaborn**.

3.4 Tehisintellekti kasutamine töö koostamisel

Lisaks analüüsi läbiviimisele mängis TI rolli ka käesoleva lõputöö koostamisel. Järgnevalt on esitatud, millistes tööetappides ja mis eesmärgil kasutati tehisintellekti rakendusi töö arendamisel. Käesoleva lõputöö koostamisel kasutati tehisintellekti rakendusi ChatGPT (OpenAI, mudel *GPT-4o*) ja Grok (xAI), mis toetasid töö sisulist ja tehnilist arendust järgmistes aspektides:

1. töö struktuuri ja peatükkide loogilise järjestuse kavandamine;
2. tekstiliste osade sõnastamine ja keeleline toimetamine (sh sissejuhatus, kokkuvõte, lühikokkuvõte);
3. ingliskeelse tõlke koostamine ja terminoloogia ühtlustamine;
4. viitamise ja bibliograafia vormistusnõuete kontroll;
5. metoodika kirjelduste ja tulemuste tõlgenduste ülesehitamine;
6. andmeanalüüsi planeerimine ja võrdlusloogika sõnastamine.

Lisaks kasutati **Groki** aktiivselt Pythonis analüüsikoodi kirjutamisel, veateadete mõistmisel ning andmeanalüüsi sammude selgitamisel. Grok aitas struktureerida andmetöötlusvoogu (sh kordusanalüüsid, *ROC* kõverate loomine ja korrelatsioonimaatriksite arvutamine) ning pakkus tuge optimaalse lahenduse leidmiseks tehniliselt keerukates kohtades.

TI kasutus dokumenteeriti tööprotsessi jooksul ning selle eesmärk oli toetada autorit töö teostamisel, mitte asendada iseseisvat uurimust.

4. Tulemused ja analüüs

Selles peatükis esitatakse ja analüüsitakse tehisintellekti mudelite ning inimeste poolt antud esteetikahinnanguid. Eesmärk on hinnata mudelite täpsust, stabiilsust ja võimet eristada visuaalselt kvaliteetseid ning problemaatilisi veebilehti.

Tulemuste esitlus on jaotatud alapeatükkidesse, mis käsitlevad esmalt mudelite sisemist järjepidevust (hindamisstabiilsust), seejärel täpsust võrreldes inimhinnangutega ning lõpuks erinevate viipade mõju ja *ROC*-analüüsi tulemusi. Analüüs tugineb eelnevalt kirjeldatud metoodikale ja mõõdikutele.

4.1 Multimodaalsete mudelite hindamisstabiilsuse analüüs

Enne kui on võimalik võrrelda mudelite hinnangute täpsust ja nende vastavust inimeste esteetilisele tajule, on vajalik hinnata, kui järjepidevalt need mudelid suudavad anda sarnaseid tulemusi korduvate jooksutuste korral. Suured multimodaalsed mudelid nagu OpenAI (OAI) ja xAI (XAI) on oma olemuselt stohhastilised süsteemid – nende sisemine tööloogika hõlmab juhuslikkust, mis võib mõjutada väljundit isegi identse sisendi korral.

Kuigi osade platvormide puhul on võimalik kasutada parameetreid (nt temperatuur), et muuta väljund deterministlikumaks, ei taga see täielikku korduvust, kuna mudelid arenevad pidevalt (sh uuendatakse mudeleid taustal, muutuvad versioonid jne). Seetõttu on oluline enne täpsusanalüüsi hinnata mudelite sisest stabiilsust, et oleks võimalik teha usaldusväärseid järeldusi nende võrdlusest inimhinnangutega.

Käesolevas alapeatükis analüüsitakse kahe esteetikahinnangute andmiseks kasutatud mudeli – xAI (XAI) ja OpenAI (OAI) – hindamisstabiilsust, mõõtes, kui sarnased on nende antud hinnangud korduvate jooksutuste korral samale veebilehe kuvatõmmisele.

Mõlemat mudelit jooksutati viis korda kõigi 104 veebilehe kuvatõmmise kohta. Iga jooksutus (run1–run5) andis sõltumatu hinnangu sama sisendi kohta. Hindamisstabiilsuse kvantitatiivseks

mõõtmiseks kasutati **Intraclass Correlation Coefficient**-i (*ICC*), mis väljendab hinnangute omavahelist kooskõla. Tulemused olid järgmised:

1. **XAI *ICC*** = 0.93
2. **OAI *ICC*** = 0.96

Mõlemad väärtused viitavad väga kõrgele stabiilsusele. Üldlevinud tõlgenduse kohaselt loetakse *ICC* väärtusi üle 0.90 „suurepäraseks“, mis tähendab, et mudelite antud hinnangud on korduskatsetes väga ühtlased. Sellisel juhul on statistiliselt põhjendatud käsitleda viit kordushinnangut ühe ansamblina, st arvutada nende keskmine skoor, mida saab kasutada mudeli lõpliku hinnanguna antud sisendi kohta.

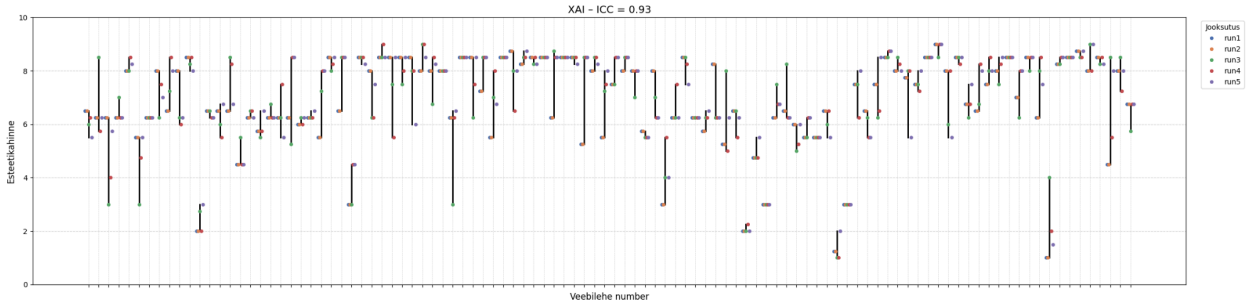
Mudelite jooksutuste vahelist kooskõla illustreerib ka **Pearsoni korrelatsioonimaatriks** (Vaata Lisa D). OAI mudeli jooksutuste omavaheline korrelatsioon jäi vahemikku 0.89–0.92, samas kui XAI puhul oli see veidi madalam, ulatudes vahemikku 0.76–0.88. See toetab järeldust, et OAI mudel on stabiilsem.

Hindamiste varieeruvust illustreerivad ka **hinnangute standardhälbed**. Viie jooksutuse põhjal arvutatud keskmine standardhälve oli:

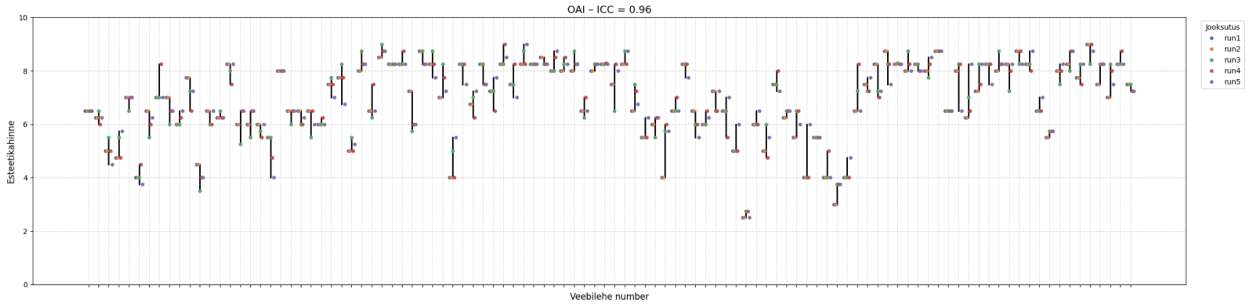
1. **XAI puhul 0.551**
2. **OAI puhul 0.342**

Need väärtused näitavad, et OAI mudeli antud hinnangud on mitte ainult kõrgema *ICC*-ga, vaid ka **statistiliselt kompaktsemad**, mis kinnitab selle mudeli suuremat sisemist järjepidevust. Mida väiksem on standardhälve, seda stabiilsem ja vähem hajuv on mudeli väljund.

Hinnangute varieeruvust illustreerivad ka graafikud (Vaata Joonis 1. Ja Joonis 2.), kus iga veebilehte kujutatakse viie punktina (iga jooksutus eraldi) ning jooksutuste maksimaalse ja minimaalse väärtuse vahe on tähistatud vertikaalse musta joonega. Visuaalsest võrdlusest nähtub, et XAI mudeli hinnangutes esineb rohkem hajuvust – mitmete veebilehtede korral ulatub skooride vahemik mitme punkti võrra. OAI hinnangud on üldiselt kompaktsemad ja koonduvad kitsamasse vahemikku, mis kinnitab kvantitatiivset analüüsi.



Joonis 1. *xAI mudeli esteetikahinnangute stabiilsus viie jooksutuse lõikes (ICC = 0.93). Iga punkt tähistab ühe jooksutuse hinnangut konkreetsele veebilehele (skaalal 0–10), mustad vertikaaljooned näitavad jooksutuste maksimaalset varieeruvust. Veebilehed on järjestatud teljel „Veebilehe number“, teljel „Esteetikahinne“ on mudeli hinnangud. Värvid tähistavad jooksutusi run1–run5.*



Joonis 2. *OpenAI (OAI) mudeli esteetikahinnangute stabiilsus viie jooksutuse lõikes (ICC = 0.96). Iga punkt tähistab ühe jooksutuse hinnangut konkreetsele veebilehele (skaalal 0–10), mustad vertikaaljooned näitavad jooksutuste maksimaalset varieeruvust. Veebilehed on järjestatud teljel „Veebilehe number“, teljel „Esteetikahinne“ on mudeli hinnangud. Värvid tähistavad jooksutusi run1–run5.*

Kokkuvõttes saab esitada järgmised järeldused:

1. Mõlemad mudelid on kõrge sisemise stabiilsusega ning sobivad esteetilise hinnangu korduvaks andmiseks.
2. OAI mudel näitab järjepidevamat käitumist ning on usaldusväärsem reaalajas kasutatavas hindamissüsteemis.
3. Kuna mõlema mudeli *ICC* väärtused ületavad 0.90 piiri, on metoodiliselt põhjendatud kasutada viit jooksutust keskmise skoori arvutamiseks, mis esindab mudeli üldist hinnangut konkreetsele sisendile.
4. XAI varieeruvus on mõnevõrra suurem, mis tähendab, et kui kasutada vaid üksikut hinnangut, on suurem tõenäosus saada veidi ebausaldusväärne tulemus. See vähendab mudeli hinnangute reprodutseeritavust ja nõuab suurema täpsuse huvides hinnangute keskmistamist või kordusanalüüsi.

Selline stabiilsusanalüüs on oluline eeldus edasisele võrdlusele inimhinnangutega ja *ROC*-analüüsile, kus keskendutakse mudelite täpsusele ja klassifitseerimisvõimele. Analüüsi tulemused näitavad, et nii OpenAI kui ka xAI mudelite hinnangud on piisavalt stabiilsed – mõlema mudeli intraklassi korrelatsioonikordajad (*ICC*) ületasid 0.90 piiri, mis viitab suurepärasele järjepidevusele.

Et edasises analüüsis oleks võimalikult väike roll üksikute jooksutuste juhuslikkusel, kasutatakse kõigi hinnangute puhul viie kordusjooksutuse keskmisi väärtusi. See võimaldab saada usaldusväärsema ja üldistatavama tulemuse, vähendades stohhastilisest mudelikäitumisest tulenevat müra.

4.2 Mudelite hindamistäpsus

Selles alapeatükis käsitletakse tehisintellekti mudelite võimet anda järjepidevaid ja sisukalt eristuvaid esteetikahinnanguid. Analüüs on jagatud kolme ossa:

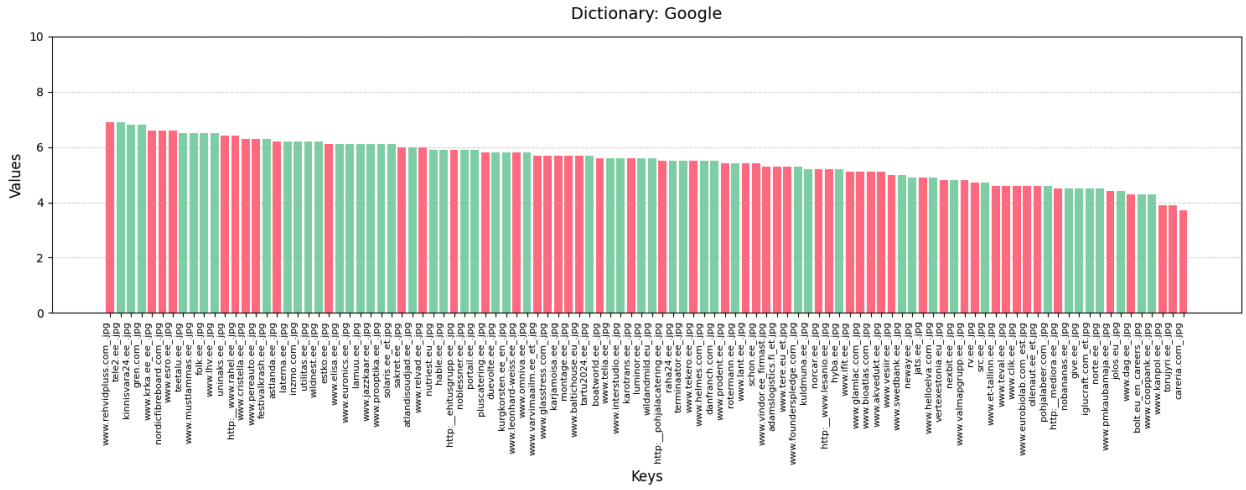
- esimeses osas vaadeldakse, milliseid esteetikahinnanguid erinevad mudelid väljastavad, ning hinnatakse visuaalselt, kui hästi need suudavad eristada visuaalselt häid ja halbu veebilehti;
- teises osas võrreldakse, kui sarnased või erinevad on erinevate mudelite hinnangud omavahel;
- ning kolmandas ja kõige olulisemas osas hinnatakse, kui hästi suudavad mudelid eristada visuaalselt „häid“ ja „halbu“ veebilehti.

See lähenemine võimaldab analüüsida mitte ainult mudelite hinnangute ulatust ja järjepidevust, vaid ka nende praktilist väärtust automaatse esteetilise kvaliteedikontrolli tööriistana.

4.2.1 Mudelite hinnangute visuaalne analüüs

Selles alajaotuses visualiseeritakse iga esteetika hinnangute andmiseks kasutatud mudeli tulemused, et analüüsida väljastatud skoore ning hinnata, kas mudelid suudavad eristada kõrge ja madala kvaliteediga veebilehti. Graafikute põhjal on võimalik hinnata, kas mudel kaldub andma keskmisi hinnanguid või suudab eristada ka skaala äärmused (väga hea või väga halb disain).

Google Vision



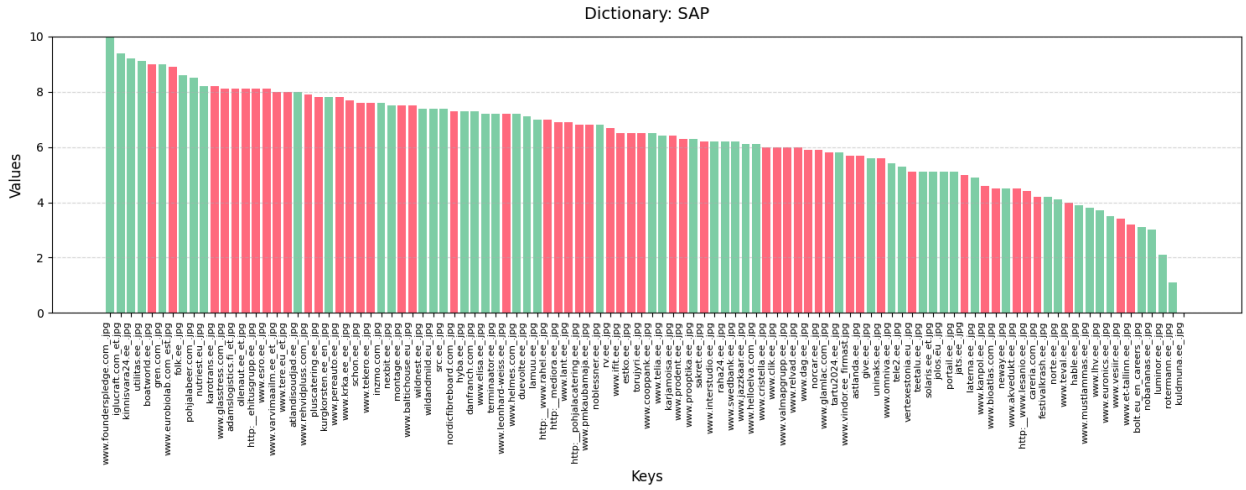
Joonis 3. Google Vision API antud esteetikahinnangud 104 veebilehele, kus rohelised on 'head' veebilehed ja punased on 'halvad'.

Graafikust (Joonis 3.) nähtub, et Google Vision API hinnangud on koondunud vahemikku 4–7, valdavalt 5–6 piiresse. Mudel kaldub andma keskmisi ja mõõdukalt positiivseid hinnanguid ka veebilehtedele, mille disain on visuaalselt nõrgem.

Värvieristus võimaldab hinnata mudeli võimet eristada disainikvaliteeti. Näha on, et „head" (roheline) ja „halvad" (punane) veebilehed jagunevad skooride skaalal üsna läbisegi – kõrgeid ja madalaid hinnanguid esineb mõlemas kategoorias. See viitab sellele, et mudeli eristusvõime on madal ning see ei suuda järjepidevalt tuvastada visuaalse kvaliteedi erinevusi.

Kokkuvõttes annab Google Vision API pigem üldise ja ettevaatliku hinnangu, kuid ei sobi hästi täpseks esteetika hindamiseks olukordades, kus on vaja selgelt eristada tugeva ja nõrga kujundusega veebilehti.

SAP (Simple Aesthetic Predictor)

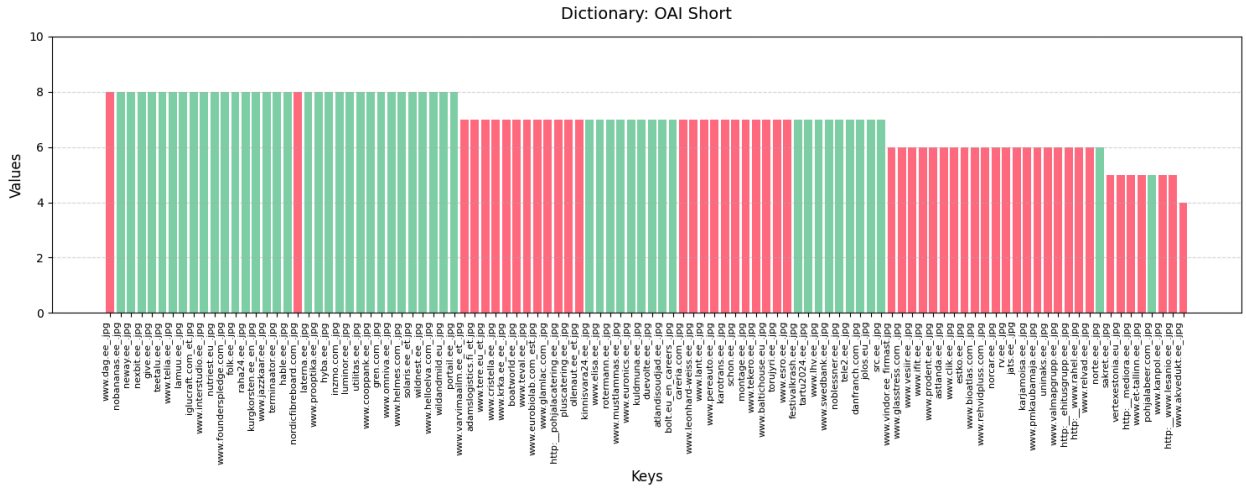


Joonis 4. Simple Aesthetics Predictori antud esteetikahinnangud 104 veebilehele, kus rohelised on 'head' veebilehed ja punased on 'halvad'.

Graafikust (Joonis 4.) on näha, et *SAP* mudeli antud hinnangud ulatuvad kogu skaala peale – alates umbes 2 punktist kuni 10-ni. Jaotus on ebaühtlane ja selget mustrit ei teki: kõrgeid ja madalaid väärtusi esineb nii „heade" kui „halbade" veebilehtede seas.

Kuigi osa tugevalt kujundatud lehti saab kõrgeid hindeid, saavad samaväärseid tulemusi ka mitmed visuaalselt nõrgemad lehed. See viitab sellele, et mudel ei erista järjepidevalt disaini kvaliteeti. *SAP* hinnangud on hajusad ja raskesti tõlgendatavad, mistõttu ei sobi see hästi usaldusväärseks esteetika hindamise vahendiks.

***GPT-4o* (Lühike viip)**

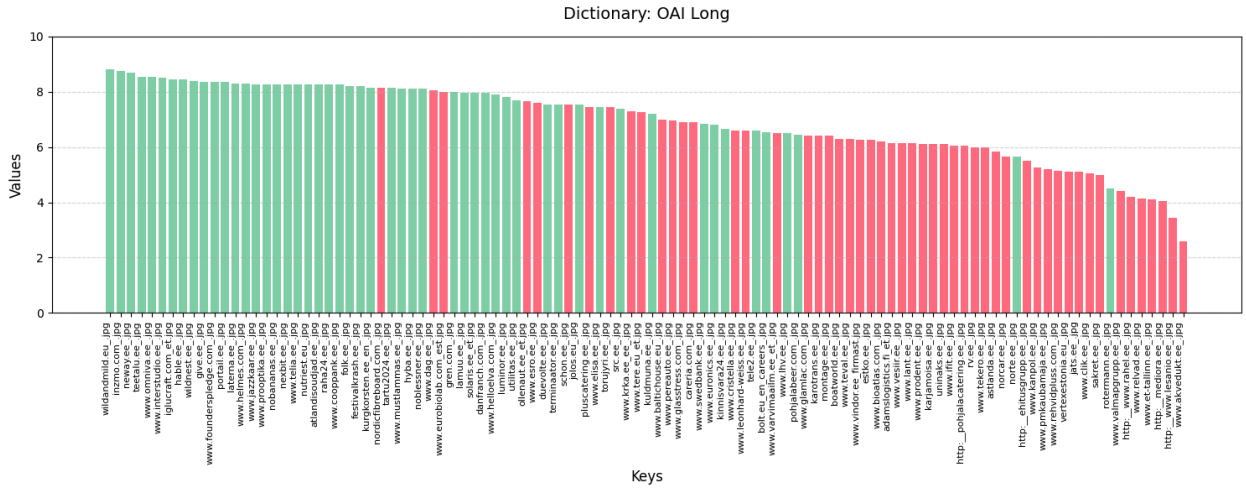


Joonis 5. GPT-4o lühikese viiba antud esteetikahinnangud 104 veebilehele, kus rohelised on 'head' veebilehed ja punased on 'halvad'.

Graafik (Joonis 5.) näitab, et OAI mudel kipub koondama hinnanguid kolme taseme ümber: 6, 7 ja 8. Väga madalaid või kõrgeid skooride väärtusi esineb harva, mis viitab sellele, et mudel ei kasuta kogu skaala potentsiaali. Tulemused jäävad kitsasse vahemikku ning on ühtlustatud.

Kuigi „halvad" veebilehed saavad keskmiselt veidi madalamaid hindeid kui „head", ei ole erinevus alati selgelt tajutav. See näitab, et lühike ja üldsõnaline viip ei võimalda mudelil hinnata esteetikat piisava täpsusega. Rohkem eristust ja nüansse pakuks struktuursem ja põhjalikum sisend.

***GPT-4o* (Pikk viip)**

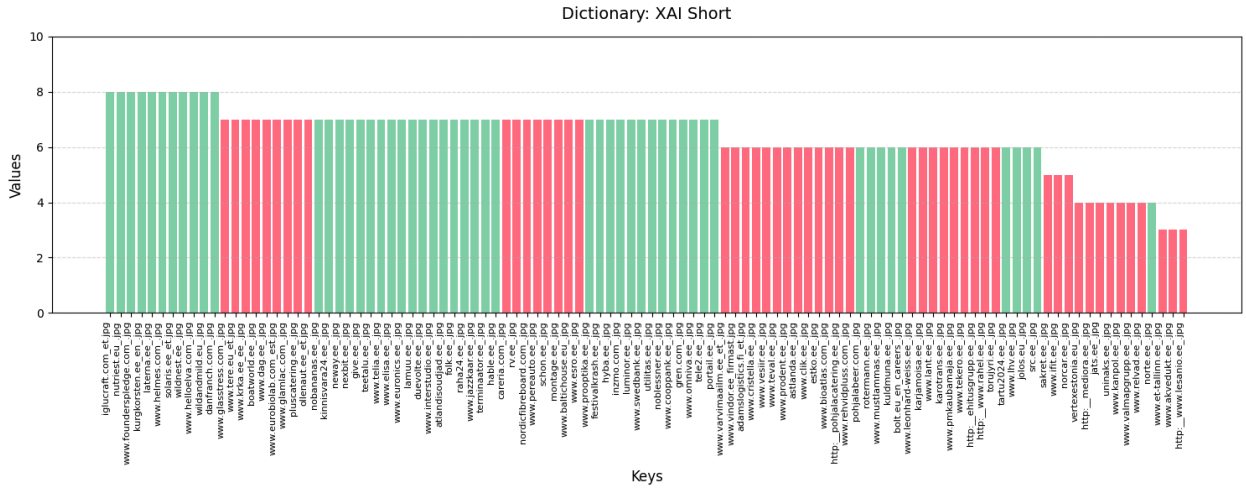


Joonis 6. GPT-4o pika viiba antud esteetikahinnangud 104 veebilehele, kus rohelised on 'head' veebilehed ja punased on 'halvad'.

Graafik (Joonis 6.) illustreerib, et OAI mudel rakendab hindamisskaalat kogu ulatuses – kõrgeimad väärtused ületavad 8.5 punkti, madalaimad jäävad alla 3. Skooride jaotus on sujuv ja katab kogu vahemiku, mis viitab sellele, et mudel on tundlik ning suudab adekvaatselt eristada erineva kvaliteediga disaine. „Headele" veebilehtedele antakse pigem kõrged hinded, samas kui „halvad" lehed koonduvad madalamasse vahemikku.

Selline muster kinnitab, et mudel suudab visuaalset esteetikat hinnata kooskõlas tegeliku kujundustasemega. Detailse viiba toel on OAI mudel võimeline andma sisukaid, täpseid ja selgelt eristuvaid hinnanguid, sobides hästi kasutamiseks olukordades, kus on oluline usaldusväärne ja nüansirohke analüüs.

xAI (Lühike viip)

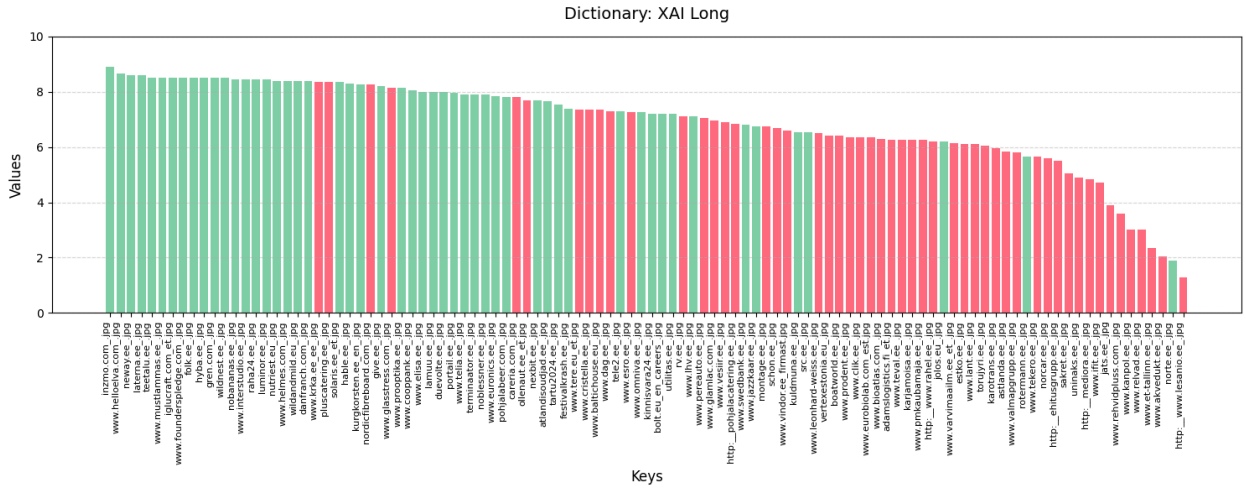


Joonis 7. xAI lühikese viiba antud esteetikahinnangud 104 veebilehele, kus rohelised on 'head' veebilehed ja punased on 'halvad'.

Graafik (Joonis 7.) näitab, et xAI mudeli lühikese viiba puhul koonduvad enamik hinnanguid vahemikku 6 kuni 7. See viitab sellele, et mudel kaldub andma keskmisi tulemusi sõltumata sisendi kvaliteedist ning ei kasuta täielikult kogu skaalat. Väga kõrgeid või madalaid hinnanguid esineb harva.

„Head" ja „halvad" veebilehed saavad sageli sarnaseid skoori väärtusi, mis viitab nõrgale eristusvõimele. Mudel ei suuda piisavalt selgelt eristada visuaalse disaini kvaliteeti. Tulemuste täpsust ja informatiivsust aitaks parandada struktuursem ja detailsem viip.

xAI (Pikk viip)



Joonis 8. xAI pika viiba antud esteetikahinnangud 104 veebilehele, kus rohelised on 'head' veebilehed ja punased on 'halvad'.

Graafikust (Joonis 8.) on näha, et xAI mudeli detailse viiba tulemused katavad kogu hindamisskaala – kõrgeimad skoorid on üle 8.5 ja madalaimad alla 2. See viitab heale kontrastivõimele ning näitab, et mudel suudab eristada tugeva ja nõrga disainiga veebilehti. Enamus hinnanguid jääb siiski vahemikku 6–8, mis viitab kergele keskväärtuse eelistamisele.

„Head" lehed saavad üldjuhul kõrgemaid hindeid, samas kui „halvad" lehed asuvad valdavalt alumises skaalaservas. Selline jaotus on mitmekesine ja usaldusväärne. Üldpildis on mudeli hinnangute muster sarnane *GPT-4o* pika viiba tulemusele, kuigi äärmuste osakaal on veidi väiksem.

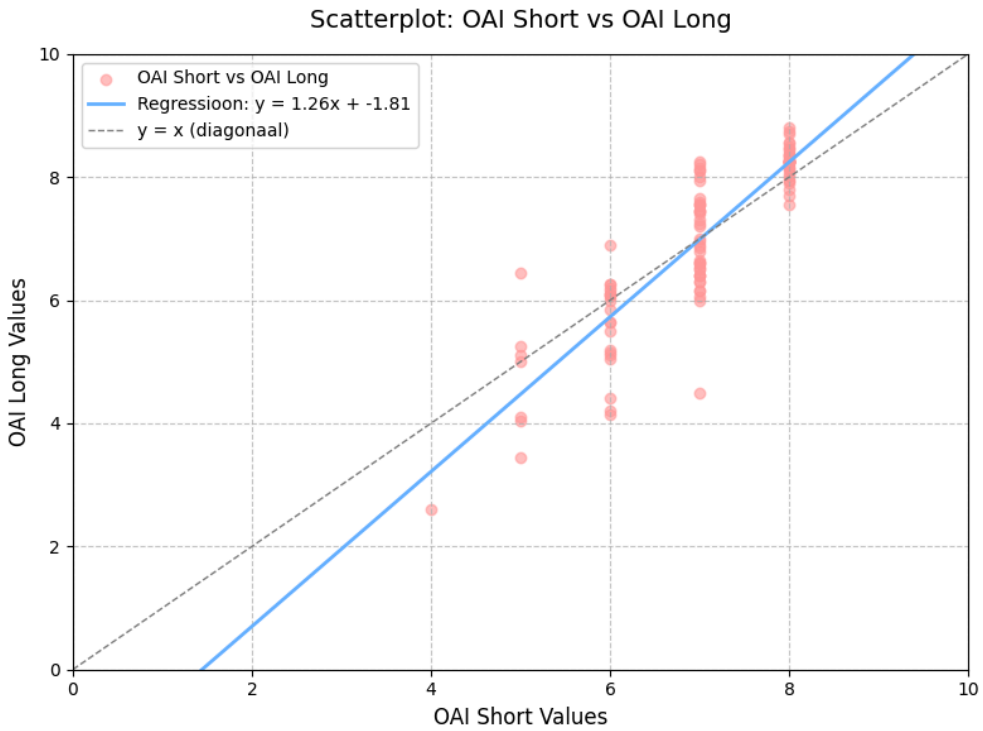
4.2.2 Mudelitevahelised korrelatsioonid

Käesolevas alapeatükis analüüsitakse, kuivõrd erinevad tehisintellekti mudelid hindavad veebilehtede esteetikat sarnaselt. Selleks arvutati Pearsoni korrelatsioonikordajad kõigi mudelite vahel, võrreldes nende keskmisi hinnanguid 104 veebilehe lõikes. Kõrge korrelatsiooninäitaja (r lähedal 1-le) viitab sellele, et mudelid järjestavad veebilehed esteetilisuse alusel sarnaselt, samas kui madal või statistiliselt ebaoluline korrelatsioon viitab olulisele erinevusele hinnangute struktuuris.

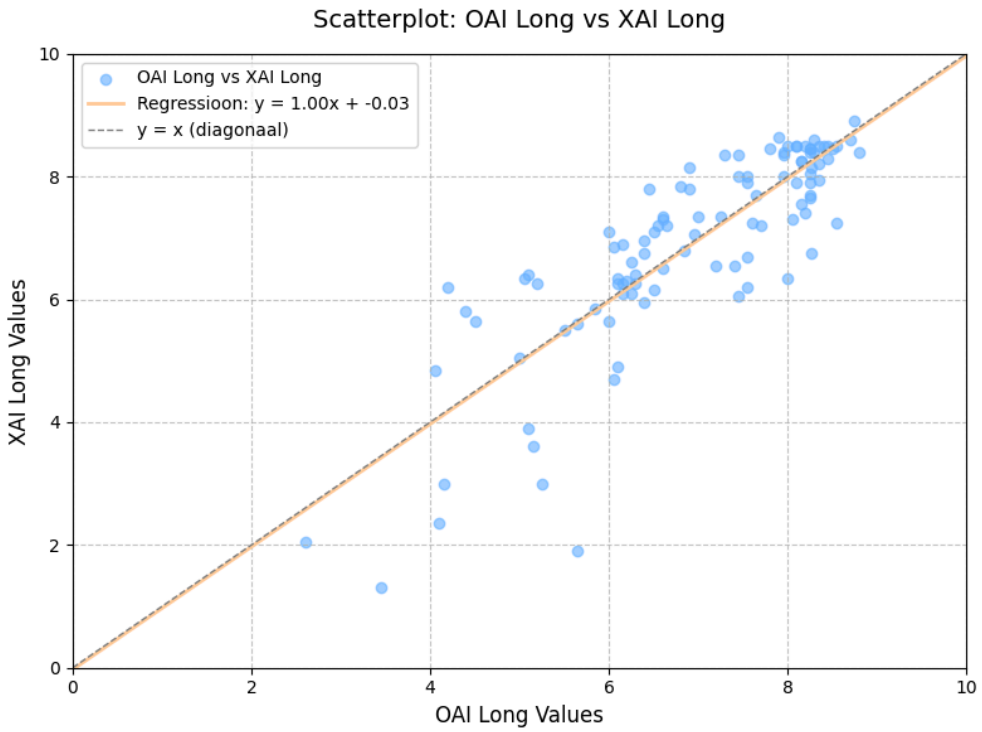
Kõige tugevama vastastikuse kooskõla näitasid:

1. **xAI Long vs OAI Short** ($r = 0.874$, $p < 0.0001$)
2. **OAI Long vs OAI Short** ($r = 0.874$, $p < 0.0001$)
3. **OAI Long vs xAI Long** ($r = 0.840$, $p < 0.0001$)
4. **OAI Short vs xAI Short** ($r = 0.808$, $p < 0.0001$)

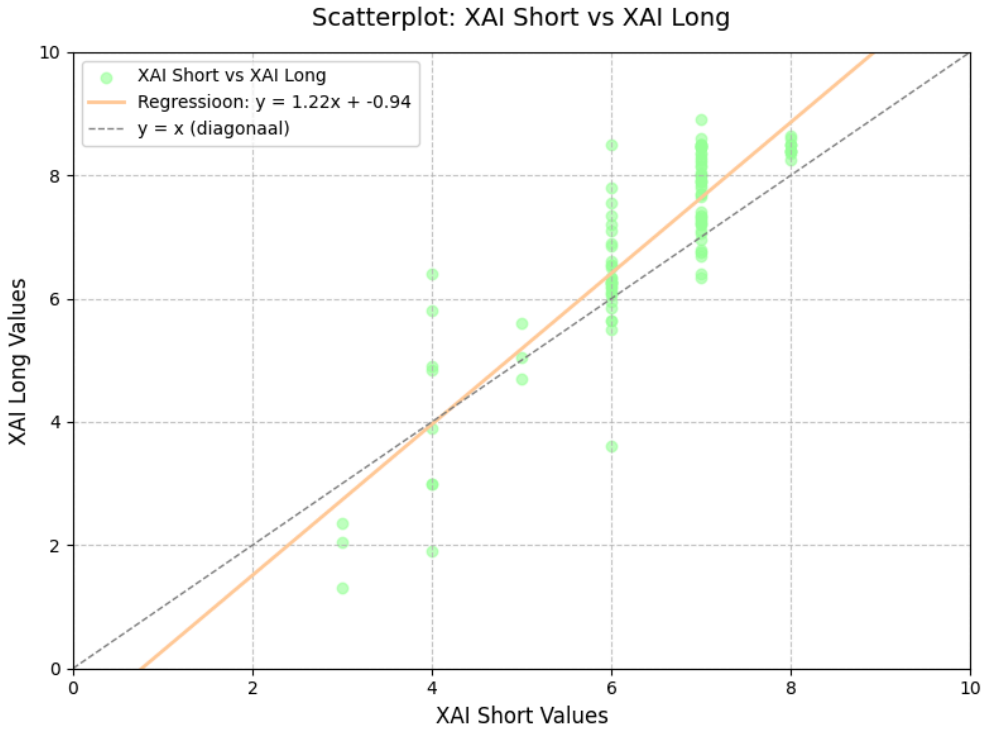
Need tulemused viitavad sellele, et **OAI ja XAI mudelid on üksteisega väga sarnased**, eriti kui kasutatakse kas detailset või üldist viipa. Kõrge korrelatsioon kahe mudeli vahel tähendab, et nende hinnangud veebilehtede visuaalsele kvaliteedile liiguvad üldjoontes samas suunas.



Joonis 9. GPT-4o pika ja lühikese viiba vaheline korrelatsioon.



Joonis 10. GPT-4o lühikese ja xAI pika viiba vaheline korrelatsioon.



Joonis 11. xAI lühikese ja pika viiba vaheline korrelatsioon.

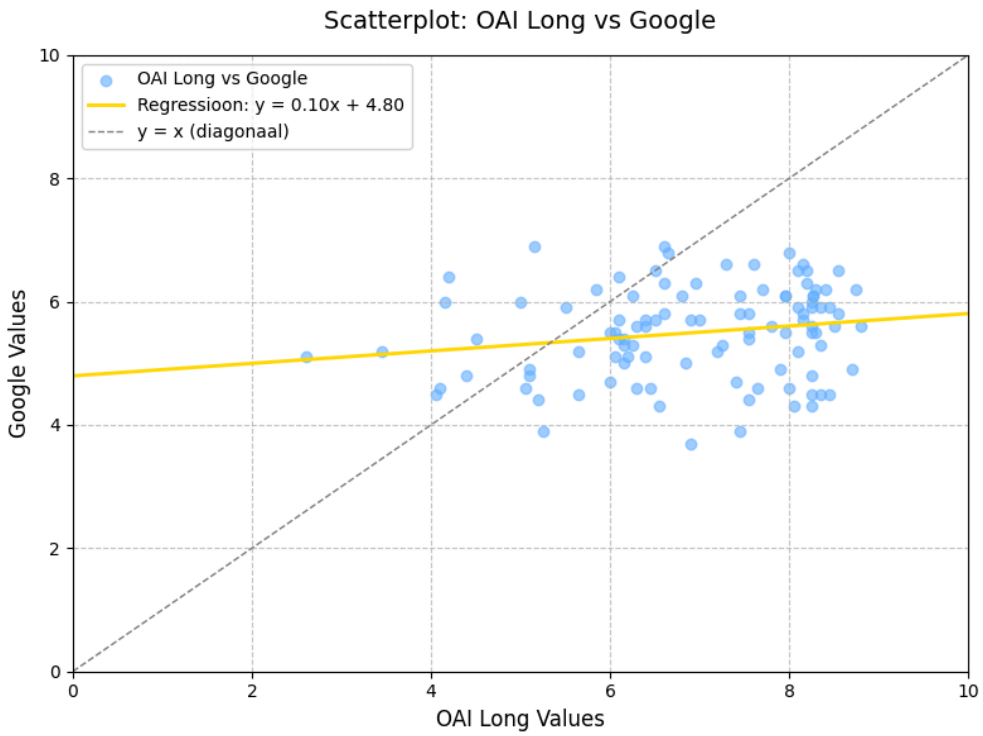
Kõige madalamad ja statistiliselt ebaolulised korrelatsioonid esinesid *SAP*-i ja *Google Visioni* ning nende võrdluses teiste mudelitega:

1. ***OAI Long* vs *Google*** (r = 0.186, p = 0.0588)
2. ***OAI Short* vs *SAP*** (r = 0.160, p = 0.1054)

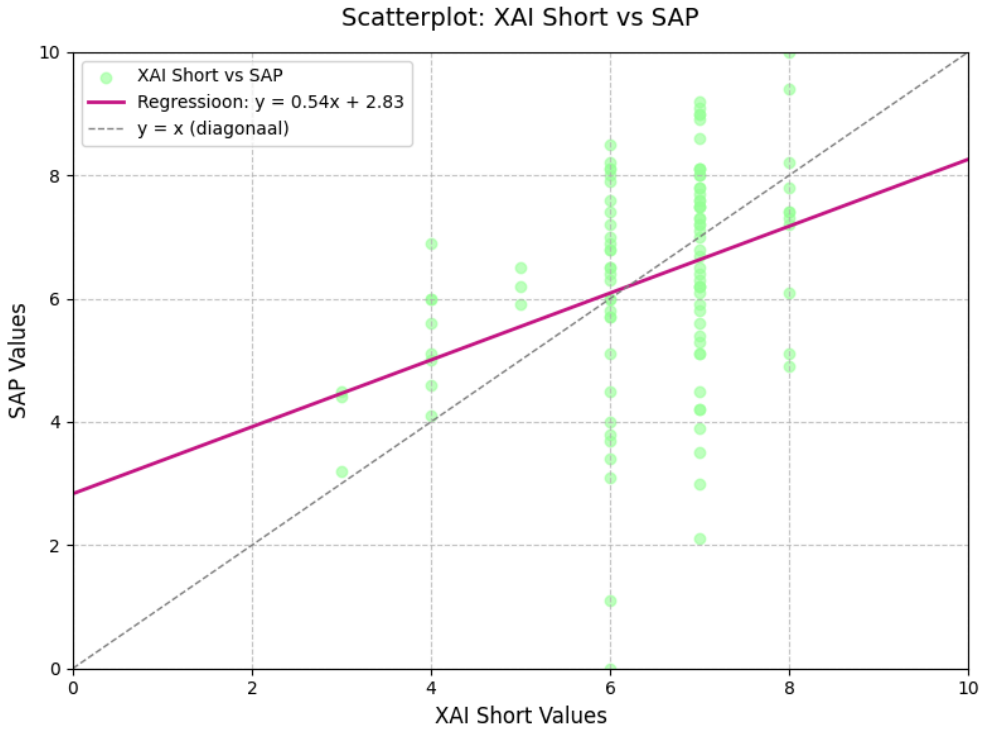
See näitab, et Google'i madalatasemelisel pildianalüüsil põhinev mudel ei käitu samal viisil kui semantiliselt juhitud suurmudelid. Lisaks oli *SAP* **mudel** ainult nõrgas kooskõlas teiste mudelitega, kuigi mõningane statistiline seos oli olemas:

1. ***xAI Short* vs *SAP*** (r = 0.345, p < 0.001)
2. ***Google* vs *SAP*** (r = 0.219, p = 0.026)

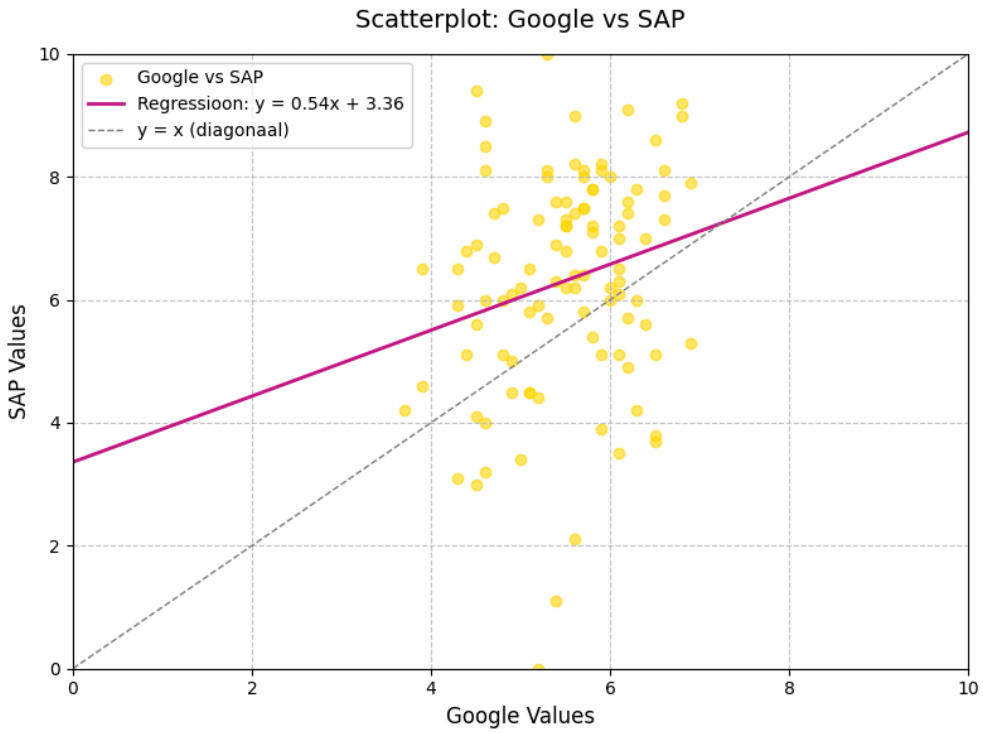
See viitab sellele, et *SAP* mudel hindab veebilehtede esteetikat teisiti kui kaasaegsed multimodaalsed mudelid ning võib olla kalibreeritud teistsugusele sisendile või standardile.



Joonis 12. GPT-4o pika viiba ja Google vaheline korrelatsioon.



Joonis 13. xAI lühikese viiba ja SAP vaheline korrelatsioon.



Joonis 14. Google ja SAP vaheline korrelatsioon.

Üldine muster korrelatsioonidest võimaldab teha järgmised järeldused:

1. Kõige suurema omavahelise kooskõla saavutasid samasuguse ülesehitusega mudelid (*OAI* ja *XAI*) kas pika või lühikese viibaga.
2. Viiba struktuur mängib rolli hinnangute sarnasuses – sama mudel käitub erinevalt sõltuvalt sellest, kas kasutatakse detailset või üldist viipa.
3. *Google Vision API* ja *SAP* mudelid erinevad ülejäänutest märkimisväärselt ning nende tulemusi ei saa käsitleda samaväärselt suurte tehisaru keelemudelite hinnangutega.

Kokkuvõttes näitab see, et mudelite arhitektuur ja sisendi vorm mängivad olulist rolli nende esteetilise hindamisvõime kujunemisel.

4.2.3 *ROC* analüüs

Selle töö keskne küsimus on, kas TI-mudelid suudavad automaatselt ja usaldusväärselt eristada visuaalselt „häid“ ja „halbu“ veebilehti. Selle hindamiseks kasutatakse *ROC*-kõverat (Receiver Operating Characteristic) ja *AUROC* väärtust (Area Under the Curve).

ROC-analüüsi tulemused

ROC-analüüsi tulemused võimaldavad hinnata, kuivõrd täpselt suudavad erinevad tehisintellekti mudelid eristada visuaalselt „häid“ ja „halbu“ veebilehti. Selleks kasutati *AUROC* (*Area Under ROC Curve*) mõõdikut, mis väljendab *ROC* kõvera alla jäävat pindala. Mida lähemale 1.0 see väärtus ulatub, seda paremini suudab mudel kahe klassi vahel vahet teha; väärtus 0.5 viitab juhuslikule eristamisele ning sellest madalamad tulemused viitavad ebaadekvaatsele käitumisele.

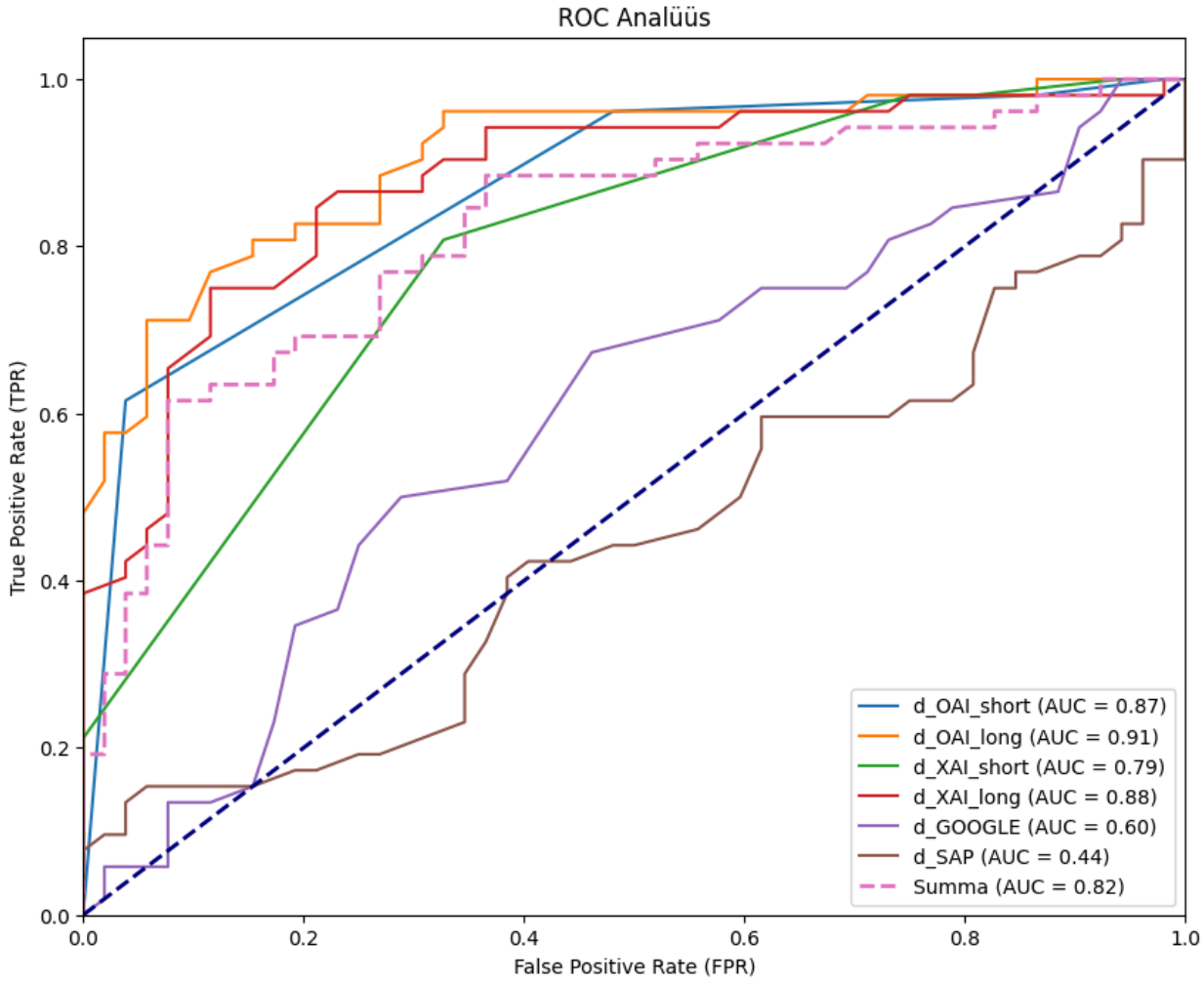
Tulemused näitasid, et parima üldise eristusvõimega oli *GPT-4o* mudel detailse viiba korral, mille *AUROC* väärtus oli 0.91. See viitab väga heale võimekusele eristada kvaliteetseid veebilehti visuaalselt nõrkadest. Mudel oli suuteline arvestama esteetiliste aspektidega viisil, mis kattus märkimisväärselt hästi inimeste hinnangutega. Ka sama mudeli lihtsustatud kasutusviis, kus sisendiks oli lühike üldine viip, saavutas kõrge tulemuse (*AUROC* = 0.87), mis viitab sellele, et mudel on võimeline tegema mõistlikke järeldusi ka ilma põhjaliku juhendamiseta.

xAI „grok-2-vision“ mudel saavutas samuti tugevaid tulemusi. Detailse viiba kasutamisel viidi iga sisendi kohta läbi viis jooksutust ning vastavad *AUROC* väärtused jäid vahemikku 0.87–0.91 (keskmine *AUROC* = 0.88). See näitab, et mudel on suuteline visuaalse kvaliteedi eristamiseks, kuigi selle hinnangutes esines veidi rohkem varieeruvust võrreldes *GPT-4o* mudeliga. Kui kasutati lühikest viipa, langes xAI mudeli *AUROC* väärtus 0.79-ni, mis viitab mõõdukale, ent ebapiisavale täpsusele olukordades, kus sisend ei olnud piisavalt täpselt struktureeritud.

Nõrgemate tulemustega paistsid silma *Google Cloud Vision* API ja *SAP* (Simple Aesthetics Predictor). Google Visioni *AUROC* väärtus oli 0.60, mis tähendab, et mudel eristas veebilehti vaid pisut paremini kui juhuslik valik. Selle põhjuseks võib olla asjaolu, et mudel tugines üksnes madala taseme visuaalsetele tunnustele nagu värvide ja kontrasti jaotused, mõistmata

kõrgematasemelisi esteetilisi mustreid. *SAP* mudel oli ainsana alla juhuslikkuse piiri (*AUROC* = 0.44), mis tähendab, et mudel andis hinnanguid, mis ei vastanud reaalsetele kvaliteedierinevustele ning võis anda eksitavat tagasisidet.

Kokkuvõttes kinnitavad *ROC*-analüüsi tulemused, et usaldusväärseks automaatseks esteetikahindamiseks sobivad ainult need mudelid, mille *AUROC* väärtus on selgelt üle 0.8 ning mis töötavad koos täpselt formuleeritud ja kontekstuaalselt rikaste viipadega. *GPT-4o* ja xAI detailsetel sisenditel põhinevad mudelid näitasid seejuures suurimat kooskõla inimeste hinnangutega ning kõige tugevamat klassifitseerimisvõimet. See annab alust pidada neid mudeliteks, millele saab toetuda esteetilise kvaliteedi hindamisel praktilistes rakendustes.



Joonis 15. Kõigi mudelite ROC kõverad. Joonisel on kujutatud kuue mudeli klassifitseerimisvõime kahe klassi („hea“ vs „halb“) eristamisel. Punktiirjoon „Summa“ (AUROC = 0.82) tähistab kombineeritud mudelite keskmise tulemuslikkuse näitajat. Diagonaaljoon tähistab juhuslikku klassifitseerijat (AUROC = 0.5).

Tabel 1. Kõigi mudelite *AUROC* väärtused.

Mudel	*AUROC*	**Tõlgendus**
OAI long	0.91	Väga hea – parim üldine klassifitseerimisvõime
XAI long	0.88	Väga hea – tugev eristusvõime
OAI short	0.87	Hea – ka üldise viibaga täpne tulemus
XAI short	0.79	Keskmine – eristab, kuid mitte usaldusväärselt
Google Vision	0.60	Nõrk – piiripealne, vähene eristusvõime
SAP	0.44	Halb – alla juhusliku, ebakorrektne klassifitseerimine

ROC-analüüsi tulemused näitavad, et ainult tugevad multimodaalsed mudelid koos hästi struktureeritud viipadega (eelkõige *GPT-4o* ja xAI pikad viibad) suudavad usaldusväärselt eristada kvaliteetseid ja visuaalselt nõrku veebilehti. Seevastu madalama keerukusega või reeglipõhised mudelid (nagu *SAP* ja Google Vision) ei ole selles ülesandes piisavalt täpsed ega eristusvõimelised.

Lisaks mudeli arhitektuurile selgus, et viipa struktuur mängib võtmerolli mudeli täpsuse ja tundlikkuse kujundamisel. Pikk, sisuliselt põhjendatud viip mitte ainult ei parandanud *AUROC* väärtusi, vaid suurendas hinnangute kontrasti ja kriitilisust, eriti visuaalselt nõrkade veebilehtede puhul. Lühike viip põhjustas mudelites kalduvuse anda keskmisi hindeid ning vältida skaala äärmusi. Seetõttu võib järeldada, et esteetika hinnangu usaldusväärsus sõltub samavõrra sellest, kuidas küsimus mudelile esitatakse, mitte ainult sellest, milline mudel on valitud.

Järgmisena liigume edasi võrdlusesse AI-mudelite ja inimeste esteetikahinnangute vahel, et hinnata, kui hästi suudavad mudelid peegeldada inimlikku tajutunnet.

4.4 TI ja inimeste esteetikahinnangute sarnasus

Eelnevates alapeatükkides keskendusime sellele, kui hästi suudavad erinevad TI-mudelid eristada visuaalselt "häid" ja "halbu" veebilehti ning kui täpselt nad suudavad hinnata esteetilist kvaliteeti objektiivsete mõõdikute kaudu (nt *AUROC*, korrelatsioon, *RMSE*). Nende analüüside eesmärk oli uurida, kas TI suudab täita kvaliteedikontrolli funktsiooni, mida tavaliselt viivad läbi disaini eksperdid või veebilehtede kasutajad.

Käesolevas alapeatükis liigume edasi sügavamale võrdlusse inimeste ja tehisintellekti mudelite hinnangute vahel. Kui eelnev keskendus peamiselt binaarsele klassifikatsioonile (hea vs halb), siis nüüd uurime, kui hästi kattuvad TI mudelite hinnangud inimeste antud skooridega skaalal 0–10.

Selleks viidi läbi eksperiment, kus viis disainitaustaga inimest hindasid valitud 20 veebilehe esteetikat samal skaalal (0–10). Sama andmestik anti hindamiseks ka kuuele TI mudelile, mida on varasemates peatükkides käsitletud. Järgnevates alajaotustes võrreldakse iga mudeli hinnangut inimeste keskmise hinnanguga, kasutades kahte mõõdikut: Pearsoni korrelatsioonikordajat (r) ja ruutkeskmise vea mõõdikut (*RMSE*).

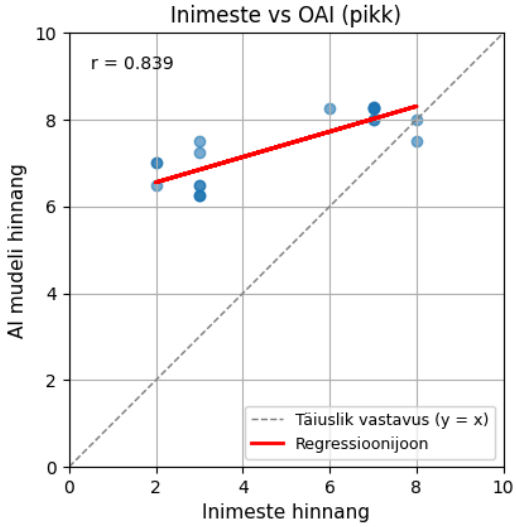
4.4.1 Korrelatsioon inimeste hinnangutega

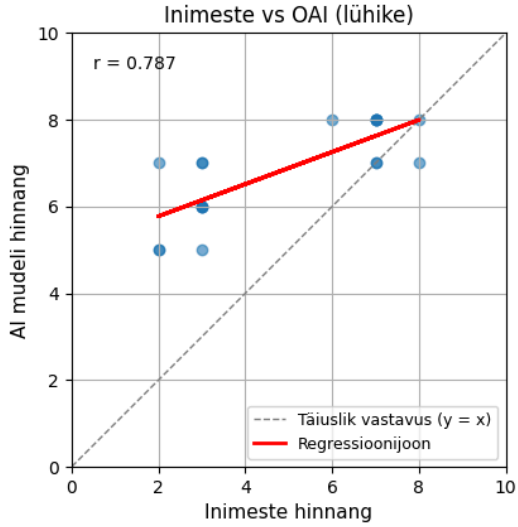
Pearsoni korrelatsioon näitab, kui tugevalt on kaks muutujat (nt inimene ja TI) lineaarses sõltuvuses. Mida lähemal on r väärtus +1-le, seda sarnasemad on hinnangute muutumise mustrid.

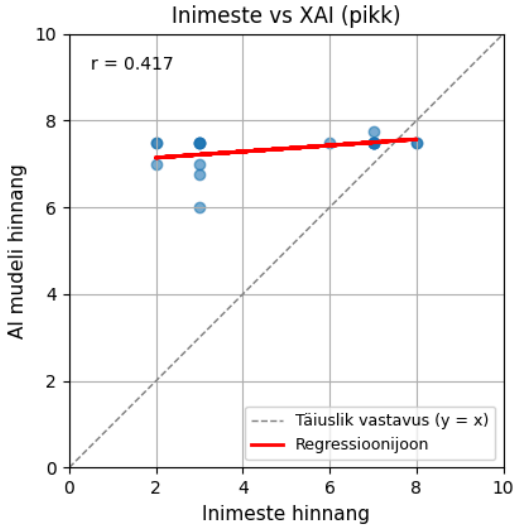
Tabel 2. Inimeste hinnangu vs. TI tulemuste korrelatsioonid.

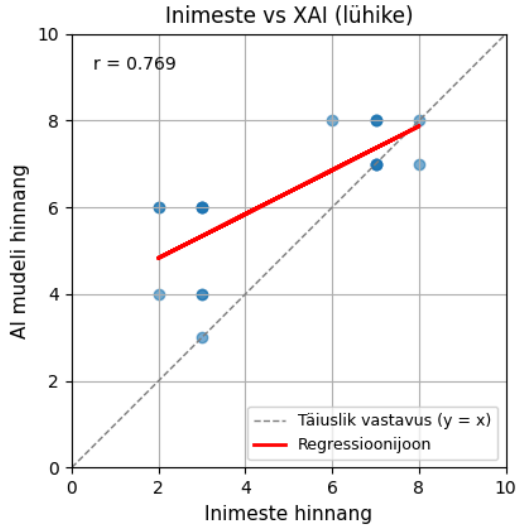
Mudel	**Pearsoni r inimeste hinnanguga**
OAI (pikk)	**0.839** ← Kõige parem vastavus
OAI (lühike)	0.787
XAI (lühike)	0.769
GOOGLE	0.602
XAI (pikk)	0.417
SAP	0.153 ← Väga nõrk seos

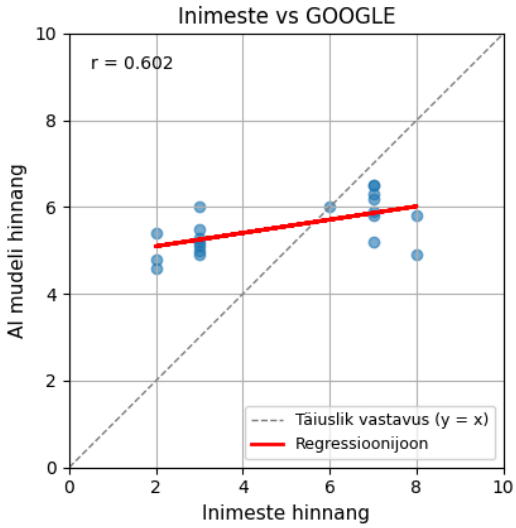
Hajuvusdiagrammid: Inimeste ja AI mudelite hinnangud

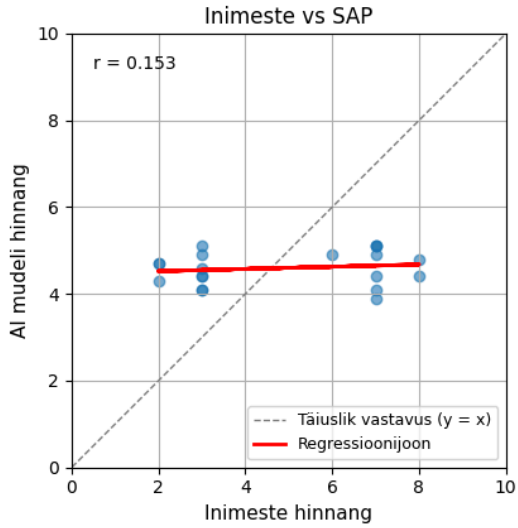












Joonis 16. Inimeste hinnangu vs. TI tulemuste korrelatsioonid.

Kõige kõrgema kooskõla saavutas *GPT-4o* **mudel** detailse viibaga (r = 0.839), mis tähendab, et selle mudeli hinnangud olid inimeste arvamustega kõige enam kooskõlas. Väga hea tulemus saavutati ka OAI lühikese viiba (r = 0.787) ja XAI lühikese viiba (r = 0.769) puhul, mis viitab, et isegi üldisema ülesandega suudavad need mudelid peegeldada inimlikku esteetikataju.

XAI pika viiba tulemus oli ootuspäraselt madalam (r = 0.417), mis võib viidata sellele, et mudeli arhitektuur või treeningandmestik ei sobitu hästi esteetika mõistega inimtasandil. Samas säilis nõrk lineaarne muster.

Google Vision API saavutas keskpärase vastavuse (r = 0.602), mis näitab, et kuigi mudel suudab mingil määral eristada head ja halba disaini, on selle otsused võrreldes inimestega pigem üldised ja madala variatiivsusega.

SAP **mudelil** puudus praktiliselt seos inimeste hinnangutega (r = 0.153), mis tähendab, et selle mudeli otsused ei lähtu samast esteetikalisest loogikast. Nagu varasemad *ROC* ja jaotusanaloosid viitasid (vt peatükk 4.2.1 ja 4.2.4), kaldus *SAP* hindama veebilehti kitsas skaalas ega olnud tundlik visuaalsetele erinevustele.

Visuaalne analüüs

Hajusdiagrammid kinnitavad kvantitatiivset hinnangut:

1. **OAI mudelite** punktid paiknevad diagonaalile lähedal, mis tähendab, et AI hinnang suureneb koos inimeste omaga.
2. *SAP* **puhul** on punktid hajali horisontaalses joones – AI annab kõigile sisenditele sarnase skoori, sõltumata inimese arvamusest.
3. Kuigi hajusdiagramm võib jätta mulje, et **XAI lühike viip** kattub paremini inimeste hinnangutega, ei saa sellest teha kindlat järeldust – paljud punktid võivad olla üksteise taga.

Kvantitatiivsed näitajad (nt korrelatsioon ja *RMSE*) ei kinnita selgelt lühikese viiba paremust, mistõttu tuleb visuaalseid mustreid tõlgendada ettevaatlikult.

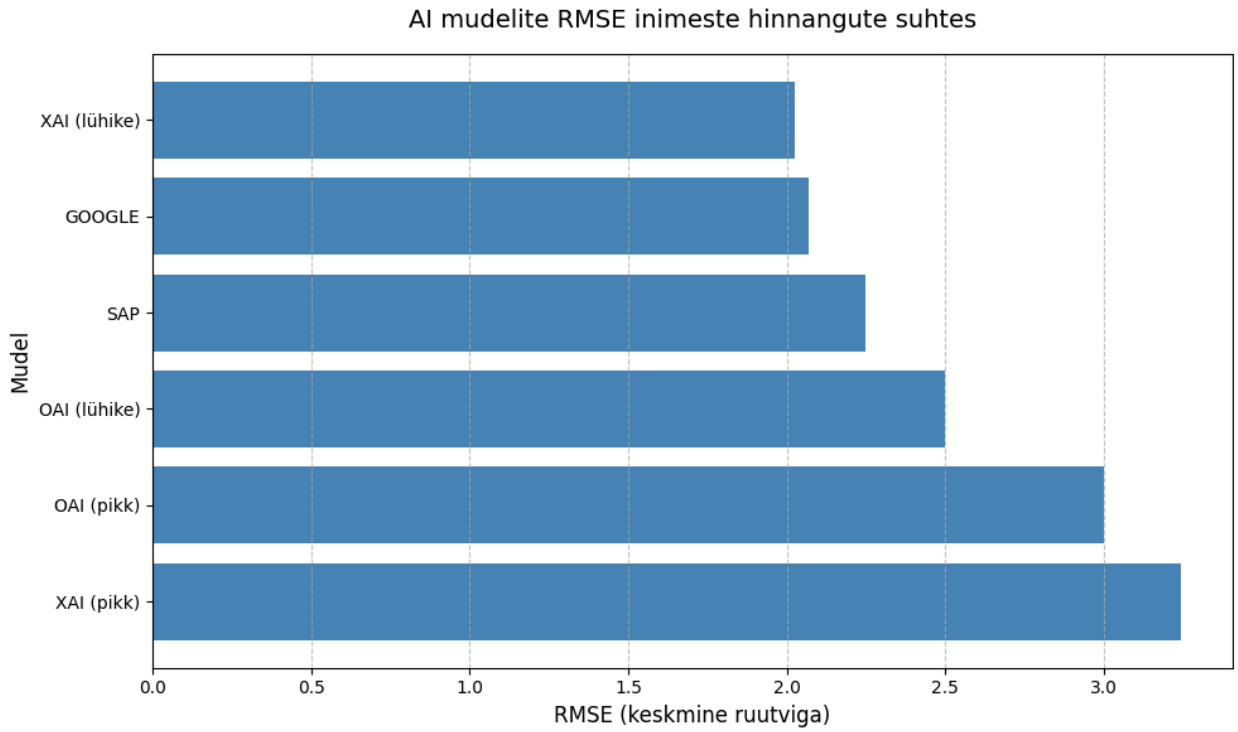
4.4.2 *RMSE* väärtused

Lisaks korrelatsioonile, mis mõõdab hinnangute suundumuste sarnasust, arvutati iga TI mudeli kohta ka keskmine ruutviga ehk Root Mean Square Error (*RMSE*). See näitab kui palju erineb mudeli antud skoor inimeste antud keskmisest hinnangust keskmiselt.

Mis on *RMSE*?

RMSE (Root Mean Square Error) mõõdab keskset kaugust kahe väärtuse vahel – antud juhul inimese ja AI mudeli hinnangu vahel. Kui TI hindab lehte 8 ja inimene 5, siis vahe on 3. *RMSE* arvutab kõigi selliste vahede ruudud, võtab neist keskmise ja seejärel ruutjuure.

1. **Mida väiksem *RMSE***, seda **täpsemalt** mudel prognoosib inimese hinnangut.
2. *RMSE* = **0** tähendaks täiuslikku vastavust – mudel annab täpselt sama skoori kui inimene.



Joonis 17. TI mudelite RMSE inimeste hinnangu suhtes.

Tabel 3. TI mudelite *RMSE* inimeste hinnangu suhtes.

Tulemused		
Mudel	*RMSE*	**Tõlgendus**
XAI (lühike)	**2.025**	Parim – väikseim keskmine viga
GOOGLE	2.068	Väga sarnane XAI lühikesele
SAP	2.250	Mõõdukas viga, kuid madal korrelatsioon
OAI (lühike)	2.500	Märkimisväärne hälve
OAI (pikk)	3.001	Suur viga inimese hinnanguga võrreldes
XAI (pikk)	3.243	Kõige suurem keskmine viga

Tõlgendus ja järeldused

Tulemused võivad tunduda esmapilgul üllatavad, kuna eelnevas peatükis (vt 4.4.1) oli kõige tugevam korrelatsioon just OAI pika viibaga (r = 0.839). Siiski tuleb mõista, et *RMSE* mõõdab ainult arvulist kaugust, mitte sarnasust hinnangu muutumises.

1. **XAI lühike** ja Google Vision andsid keskmiselt inimesele sarnasemaid hindeid, kuid nende hinnangud ei pruugi eristada häid ja halbu lehti täpselt.
2. **OAI pikk** andis kõrgema *RMSE*, kuid eristas inimesega sarnases mustris head ja halba, mida näitab kõrge korrelatsioon.
3. **XAI pikk** kombineerib suure ruutvea ja madala korrelatsiooni, viidates, et selle mudeli hinnangud olid kõige vähem kooskõlas inimese esteetikatajuga.

See tähendab, et *RMSE* sobib hindama kvantitatiivset kaugust, kuid ei asenda kvalitatiivset sarnasust, mida mõõdab korrelatsioon. Näiteks kui mudel annab pidevalt 2 punkti kõrgema hinnangu kui inimene, siis *RMSE* on suur, aga r võib olla kõrge.

Kvantitatiivne analüüs andis ülevaate mudelite täpsusest, stabiilsusest ja vastavusest inimlikele hinnangutele. Edasi käsitleti saadud tulemusi laiemas kontekstis ning toome välja töö olulisemad järeldused, piirangud ja praktilised järeldused.

5. Arutelu

Käesoleva töö eesmärk oli hinnata, kui hästi erinevad tehisintellekti mudelid suudavad hinnata veebilehtede visuaalset esteetikat ning kuivõrd need hinnangud kattuvad inimeste hinnangutega. Selleks võrreldi nelja TI mudeli (OAI, XAI, Google Vision, *SAP*) tulemusi, et hinnata nende täpsust, järjepidevust, kriitilisust ning võimet eristada kvaliteetseid ja nõrga disainiga veebilehti.

5.1 Mudelite võime eristada visuaalselt häid ja halbu veebilehti

Tulemused näitasid selgelt, et OpenAI (OAI) mudel detailse viibaga oli kõige täpsem ja enim kooskõlas inimeste hinnangutega. Sellele viitab kõrgeim Pearsoni korrelatsioon ($r = 0.839$) ning parim klassifitseerimisvõime *ROC*-analüüsis ($AUROC = 0.91$). Veidi nõrgema, kuid siiski tugeva tulemuse andis xAI mudel samas konfiguratsioonis ($AUROC = 0.88$). Vastupidiselt sellele jäid Google Vision ja *SAP* mudelid selgelt maha, viidates, et madala taseme visuaalanalüüs ei ole piisav esteetilise kvaliteedi hindamiseks. Kõik tulemused põhinevad viie jooksutuse ansamblil, et vähendada juhuslikke kõikumisi ja tagada stabiilsus.

Webthetics uuringus [4] saavutati kõrgeim korrelatsioon kasutajate hinnangutega just selliste sügavõppemudelitega, mis olid treenitud otseselt kasutajate andmete pealt. Käesolevas töös, kuigi kasutati valmis mudeleid, andsid parima vastavuse samuti keerukamad sügavõppemudelid nagu *GPT-4o* ja Grok-2. Seda võib osaliselt seletada nende arhitektuurilise ülesehitusega – mõlemad on multimodaalsed suurmudelid, mille treeningandmestik sisaldab ka esteetilisi kategooriaid, kasutajate tagasisidet ja kontekstitundlikke signaale. Seevastu Google Vision API ja *SAP* mudelid ei ole treenitud veebilehtede esteetika hindamiseks, vaid keskenduvad pigem fotode visuaalsetele omadustele. Nende taustandmestik ei hõlma disainispetsiifilisi aspekte nagu paigutus, loetavus ja kasutajaliidese nüansid, mis on aga veebidisaini hindamisel olulised.

5.2 Viiba mõju ja kriitilisus

Analüüs näitas, et viiba struktuur mõjutab TI mudeli käitumist väga tugevalt. Detailne, *VisAWI*-põhine viip tõstis nii klassifitseerimisvõimekust (*AUROC*) kui ka sarnasust inimhinnangutele. See mõju avaldus eriti halbade lehtede hindamisel, kus üldine viip kipub andma keskmisi hinnanguid, kuid struktureeritud viip viis madalamate ja kriitilisemate hinneteni.

Näiteks OAI mudel andis detailse viibaga halvematele lehtedele keskmiselt kuni 1.5 punkti madalama hinnangu võrreldes lühikese viibaga. Seega võib öelda, et AI ei ole mitte ainult mudel, vaid ka dialoogipartner – tulemused sõltuvad sellest, kui täpselt temalt hinnangut küsitakse.

5.3 TI ja inimeste erinevused

Kuigi korrelatsioonid näitasid tugevat seost, ilmnes siiski erinevusi mudelite ja inimeste hinnangute vahel. TI hinnangud olid üldjoontes vähem varieeruvad ja näisid olevat vähem tundlikud disaini pehmematele või kontekstuaalsetele omadustele. Inimesed võivad aga reageerida tugevamalt disaini narratiivile, kultuurilisele koodile või isegi disainis peituvale irooniale – aspektid, mis võivad jääda TI-le raskemini tuvastatavaks.

Teisalt on TI tugevuseks järjepidevus: mudel ei väsi, ei kaldu eelarvamustesse ning rakendab hinnangukriteeriume ühtemoodi. See teeb temast potentsiaalselt kasuliku tööriista suurtes mahus veebianalüüsiks või varajaseks kvaliteedikontrolliks.

5.4 Piirangud

Töö tulemusi tuleb tõlgendada arvestades järgmisi piiranguid:

1. Inimeste hinnangud põhinesid vaid viiel disainitaustaga hindajal.
2. Veebilehed olid eelnevalt autori poolt kategoriseeritud, mis võib sisaldada kallutatust.
3. Kasutatud mudelite treeningandmed, eriti *GPT-4o* ja Grok-2 puhul, ei ole avalikud, mis raskendab tulemuste põhjendamist arhitektuurilise tasandini.
4. Hinnangud põhinesid staatilistel kuvatõmmistel, mitte veebilehtede interaktiivsel kogemusel.

Need asjaolud ei vähenda töö tulemuste väärtust, kuid tähendavad, et järeldused kehtivad eelkõige antud metoodika ja valimi raames ning võivad vajada täiendavat kinnitust laiemas kontekstis.

5.5 Järeldused

Töö tulemused kinnitavad, et kaasaegsed multimodaalsed TI-mudelid – eelkõige *GPT-4o* – on võimelised andma inimliku loogikaga kooskõlas olevaid esteetikahinnanguid, kui neile esitada selgelt struktureeritud ja eesmärgipärane ülesanne.

Kuigi absoluutne täpsus (nt *RMSE*) ei pruugi alati olla kõrge, suudavad need mudelid hästi jäljendada inimeste järjestamismustreid, mis teeb neist sobiva tööriista esteetika-alaseks eelsõelumiseks või disainikvaliteedi analüüsiks suurtes süsteemides.

Selleks, et TI-mudelid oleksid reaalses disainitöös laiemalt kasutatavad, tuleb tulevikus pöörata rohkem tähelepanu kolmele peamisele tegurile: (1) kultuuriliste ja isiklike eelistuste arvestamine, (2) mudelite kohandatavus eri sihtrühmadele ning (3) hinnangute põhjendamise võime ehk selgitatavus. Kuigi mõned tänapäevased mudelid (nagu *GPT-4o* ja *Grok-2 Vision*) suudavad oma otsuseid sõnaliselt selgitada, sõltub selgituste selgus ja usaldusväärsus suuresti sellest, kuidas küsimus on esitatud. Läbipaistvad ja arusaadavad põhjendused on aga vajalikud, et neid mudeleid saaks usaldusväärselt kasutada professionaalses disainitöös.

Tulevikus võiks uurimist laiendada suurema inimvalimiga, lisada interaktiivsuse mõõtmeid ning arendada välja TI-põhine esteetikahindamise tööriist, mis põhineb sarnasel struktureeritud, visuaalsel ja semantilisel analüüsil nagu käesolevas töös rakendatud.

Mahukama uuringu korral võiks kaaluda ka mõne olemasoleva mudeli peenhäälestamist (fine-tuning) konkreetse esteetikakonseptsiooni või veebidisaini andmestiku põhjal, et tõsta hinnangute täpsust just sihtkontekstis. Selline lähenemine nõuab aga oluliselt rohkem ressursse ning jääb käesoleva töö käsitluspiiridest välja.

Järgnevas, kokkuvõtvas peatükis võetakse töö olulisemad tulemused lühidalt kokku ning antakse ülevaade võimalikest tulevikusuundadest.

6. Kokkuvõte

Käesoleva lõputöö eesmärk oli uurida, kui täpselt suudavad erinevad tehisintellekti mudelid hinnata veebilehtede visuaalset esteetikat ning kuivõrd need hinnangud kattuvad inimeste hinnangutega. Töö käigus võrreldi nelja TI mudelit: OpenAI *GPT-4o*, xAI *Grok-2 Vision*, *Google Cloud Vision* ja *SAP* (Simple Aesthetics Predictor). Hinnangute võrdlemiseks kasutati Pearsoni korrelatsiooni ja *ROC*-analüüsi, lisaks analüüsiti mudelite kriitilisust ja järjepidevust.

Tulemused näitasid, et kõige täpsemini kattusid inimeste hinnangutega *GPT-4o* mudeli tulemused, eriti detailse, *VisAWI* struktuurile vastava viiba korral. *GPT-4o* saavutatud korrelatsioon ($r = 0.839$) ning *ROC* kõvera alune pindala (*AUROC* = 0.91) viitavad sellele, et see mudel on võimeline eristama visuaalselt häid ja nõrku veebilehti kõige usaldusväärsemalt. xAI *Grok-2 Vision* pakkus samuti tugevaid tulemusi, kuid oli veidi ebastabiilsem. Google Vision jäi hindamisskaalal märgatavalt tagasihoidlikumaks ning *SAP* mudel oli kõige vähem täpne, kuigi pakkus kiiret ja lihtsat alternatiivi.

Töö käigus ilmnesid mitmed väljakutsed, millest suurim oli kvaliteetsete ja tasakaalustatud andmestike koostamine. Samuti oli tehniline keerukus promptide koostamisel ja tulemuste struktureeritud võrdlemisel kõrgem, kui esialgu eeldatud.

Töö tulemusel sai selgeks, et generatiivsed tehisintellekti mudelid on jõudnud tasemele, kus nad suudavad hinnata visuaalset esteetikat ligilähedaselt inimlikele hinnangutele – vähemalt lihtsustatud juhtudel. Samas on nende hinnangute järjepidevus ja kriitilisus modelleenide lõikes erinev. *GPT-4o* osutus kõige käepärasemaks mudeliks, eriti juhtudel, kus soovitakse teha kiireid, kuid usaldusväärseid esmaseid hinnanguid veebidisainile.

Tulevikus võiks töö jätkuks olla laiapõhjalisem inimkatsete kogum, mis sisaldaks ka keskmise kvaliteediga veebilehti. Lisaks võiks uurida võimalusi, kuidas selliseid hindamismudeleid integreerida automaatsetesse UX-analüüsitööriistadesse. Samuti on huvitav suund mudelite peenhäälestamine konkreetsete esteetikaeelistuste järgi.

Autor õppis töö käigus andmetöötluse täpsust, masinmudelite analüüsi struktureerimist ning teadusteksti koostamist vastavalt akadeemilistele nõuetele. Töö oli mahukas, kuid andis

väärtusliku kogemuse uurimisküsimuste sõnastamisel, katsetulemuste tõlgendamisel ning visuaalse esteetika mõtestamisel tehnoloogia kontekstis.

Viidatud kirjandus

[1] Beaird J. The principles of beautiful web design. Collingwood: SitePoint; 2010. https://dl.acm.org/doi/10.5555/1951691

[2] Daryanavard Chounchenani M, Shahbahrami A, Hassanpour R, Gaydadjiev G. Deep learning based image aesthetic quality assessment-a review. *ACM Comput Surv*. 2025;57(7):1-36. https://doi.org/10.1145/3716820

[3] Costa CJ, Costa P, Aparicio M. Principles for creating web sites: a design perspective. In: International Conference on Enterprise Information Systems; 2004 Apr; Vol. 5, pp. 484-488. SCITEPRESS. https://doi.org/10.5220/0002613004840488

[4] Dou Q, Zheng XS, Sun T, Heng PA. Webthetics: quantifying webpage aesthetics with deep learning. *Int J Hum-Comput Stud*. 2019;124:56-66. https://doi.org/10.1016/j.ijhcs.2018.11.006

[5] Lindgaard G, Fernandes G, Dudek C, Brown J. Attention web designers: You have 50 milliseconds to make a good first impression!. *Behav Inf Technol*. 2006;25(2):115-126. https://doi.org/10.1080/01449290500330448

[6] Manjunath BS, Salembier P, Sikora T, editors. Introduction to MPEG-7: multimedia content description interface. Chichester: John Wiley & Sons; 2002. https://dl.acm.org/doi/abs/10.5555/863401

[7] Robins D, Holmes J. Aesthetics and credibility in web site design. *Inf Process Manag*. 2008;44(1):386-399. https://doi.org/10.1016/j.ipm.2007.02.003

[8] Moshagen M, Thielsch MT. Facets of visual aesthetics. *Int J Hum-Comput Stud.* 2010;68(10):689-709. https://doi.org/10.1016/j.ijhcs.2010.05.006

[9] Talebi H, Milanfar P. NIMA: Neural image assessment. *IEEE Trans Image Process*. 2018;27(8):3998-4011. https://doi.org/10.1109/TIP.2018.2831899

[10] Thorlacius L. The Role of Aesthetics in Web Design. *Nordicom Rev*. 2007;28(1):63-76. https://doi.org/10.1515/nor-2017-0201

[11] Uribe S, Álvarez F, Menéndez JM. User's web page aesthetics opinion: a matter of low-level image descriptors based on MPEG-7. *ACM Trans Web*. 2017;11(1):1-25. https://doi.org/10.1145/3019595

[12] Google Cloud. Cloud Vision API documentation (versioon v1). Google Cloud; 2024. https://cloud.google.com/vision/docs (külastatud 07.08.2025).

[13] OpenAI. GPT-4o (versioon 2024-05-13). OpenAI; 2024. https://openai.com/index/gpt-4o (külastatud 07.08.2025).

[14] ScreenshotOne. The screenshot API for developers (versioon 3.5). ScreenshotOne; 2025. https://screenshotone.com (külastatud 07.08.2025).

[15] Read M. 25 Top Web Design Trends 2025. TheeDigital; 2024. https://www.theedigital.com/blog/web-design-trends (külastatud 07.08.2025).

[16] xAI. Grok-2 Beta Release (versioon 2.0). xAI; 2024. https://x.ai/blog/grok-2 (külastatud 07.08.2025).

[17] shunk031. Aesthetics Predictor v2 (SAC + Logos + AVA1 + L14 LinearMSE) (versioon linearMSE). Hugging Face; 2022. https://huggingface.co/shunk031/aesthetics-predictor-v2-sac-logos-ava1-l14-linearMSE (külastatud 07.08.2025).

Lisa A. Kasutatud terminite selgitused

Alljärgnevalt on selgitatud töö käigus kasutatud olulisemad erialased terminid, mille mõistmine võib eeldada varasemat kokkupuudet tehisintellekti ja andmeteaduse valdkonnaga.

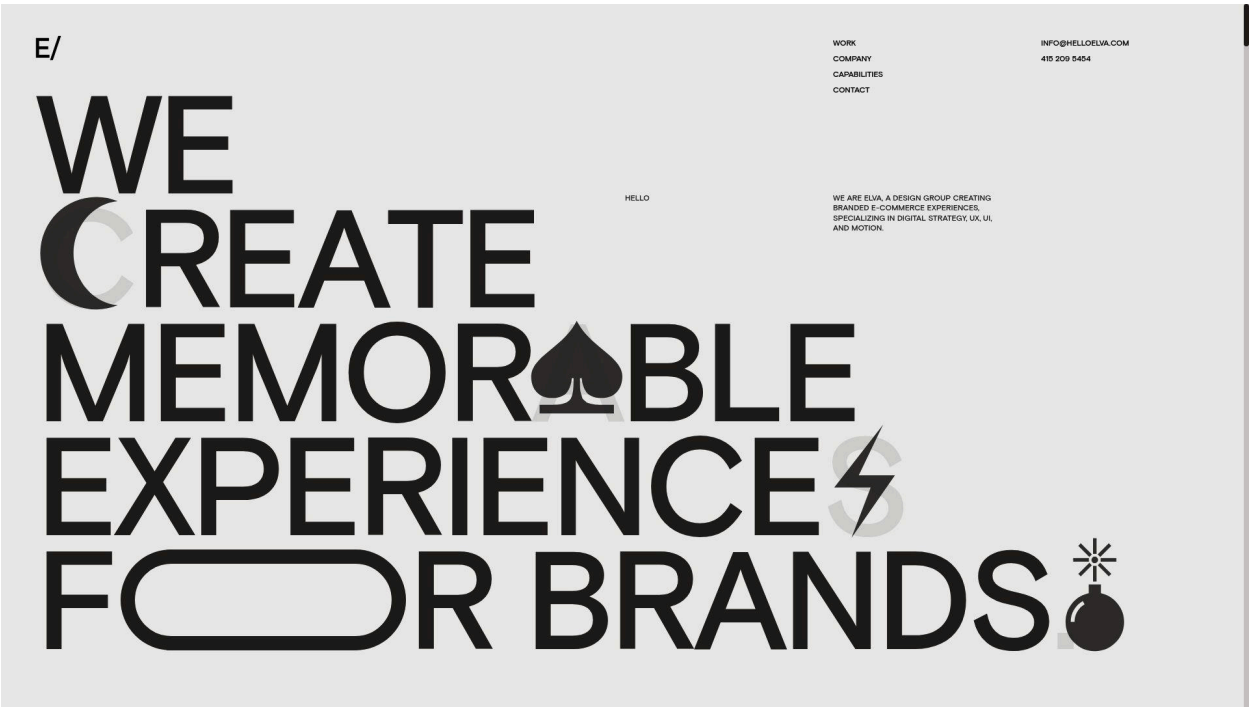
Termin	**Selgitus**
Viip	Sisendtekst või küsimus, mille alusel tehisintellekti mudel genereerib vastuse.
Temperatuur	Parameeter, mis määrab vastuste varieeruvuse: madal temperatuur (nt 0.0) annab stabiilsemaid ja korduvamaid vastuseid, kõrge (nt 1.0) aga loovamaid ja vähem ennustatavaid.
Multimodaalne	Mudel, mis suudab töödelda mitut andmetüüpi (nt tekst ja pilt) samaaegselt.
Base64	Andmekodeeringu vorming, mida kasutatakse piltide või failide tekstina edastamiseks näiteks API kaudu.
***ROC* kõver**	Graafik, mis näitab mudeli suutlikkust eristada kahte klassi (nt „hea“ ja „halb“ veebileht), võttes arvesse tõeseid ja väärasid positiive.
***AUROC* väärtus**	*ROC* kõvera alune pindala (Area Under the Curve), mis näitab mudeli üldist eristusvõimet; 1.0 on ideaalne, 0.5 tähendab juhuslikku otsust.
Pearsoni korrelatsioon	Statistiline näitaja, mis kirjeldab kahe tunnuse vahelist lineaarset seost (vahemikus -1 kuni +1).
RMSE	Keskmine ruutviga (Root Mean Square Error), mis mõõdab ennustatud ja tegelike väärtuste erinevust. Mida väiksem, seda täpsem mudel.

***CNN* (konvolutsiooniline närvivõrk)**	Süvaõppe arhitektuur, mida kasutatakse peamiselt piltide ja visuaalse info analüüsiks.
Attention mechanism	Mehhanism, mis võimaldab mudelil pöörata rohkem tähelepanu sisendi olulistele osadele (levinud näiteks teksti- ja pildimudelites).
CLIP	Mudel, mis ühendab visuaalse ja tekstilise sisendi, võimaldades näiteks pildi põhjal genereerida tekstilist kirjeldust.
VisAWI	„Visual Aesthetics of Websites Inventory“ – raamistik, mis hindab veebilehtede esteetikat neljas mõõtmes: lihtsus, mitmekesisus, värvikus ja viimistletus.

Lisa B. Näited hinnatud veebilehtedest

A.1 Head veebilehed

Joonis A1. Kaasaegne ja visuaalselt atraktiivne veebileht (näide 1)



Joonis A2. Kuldmuna konkursil auhinnatud veebileht (näide 2)

DISAINVALGUSTID IGALE MAITSELE

Pakume valgusteid nii kodusisustajale kui ka professionaalsele ruumikujundajale.

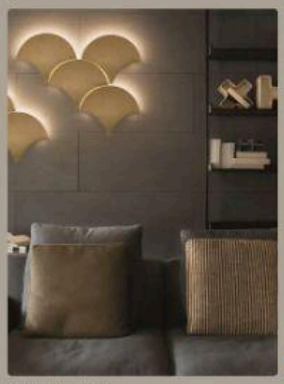
01

LAETAVAD VALGUSTID

02

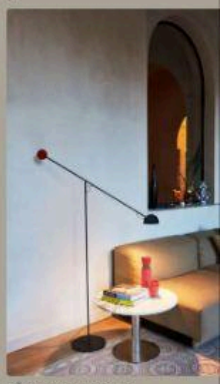
RIPPVALGUSTID

03



SEINAVALGUSTID

04



PÕRANDAVALGUSTID

Värskelt valikus

MARSET PLAFF-ON! SEINA-/LAEVALGUSTI, ROOSTEPUNANE

303 € - 621 €

MARSET GINGER OUTDOOR SEINAVALGUSTI, ROOSTEPUNANE

571 € - 841 €

MARSET GINGER OUTDOOR SEINAVALGUSTI, MUST

571 € - 841 €

MARSET SOHO PÕRANDAVALGUSTI IP44,

KÕIK TOOTED

Brändid

Igale maitsele, vajadusele ja igasse ruumi

KÕIK BRÄNDID

1

101 Copenhagen

A

A-Emotional Light by Arturo Alvarez
Album
Antidark
Antonangeli
Aromas del Campo
Artemide
Augenti
Authentage

B

B.Lux
Baulmann
Bega
Benoito Faure
Berliner Messinglampen
Bonvmi
BPM Lighting
Brand Van Egmond
Brokis

C

Castaldi
Catellani&Smith
Contardi
CTO Lighting

D

Dark
Davide Groppi
De Majo
DCW

E

Edendesign
Egoluce
Elstead Lighting
Estiluz

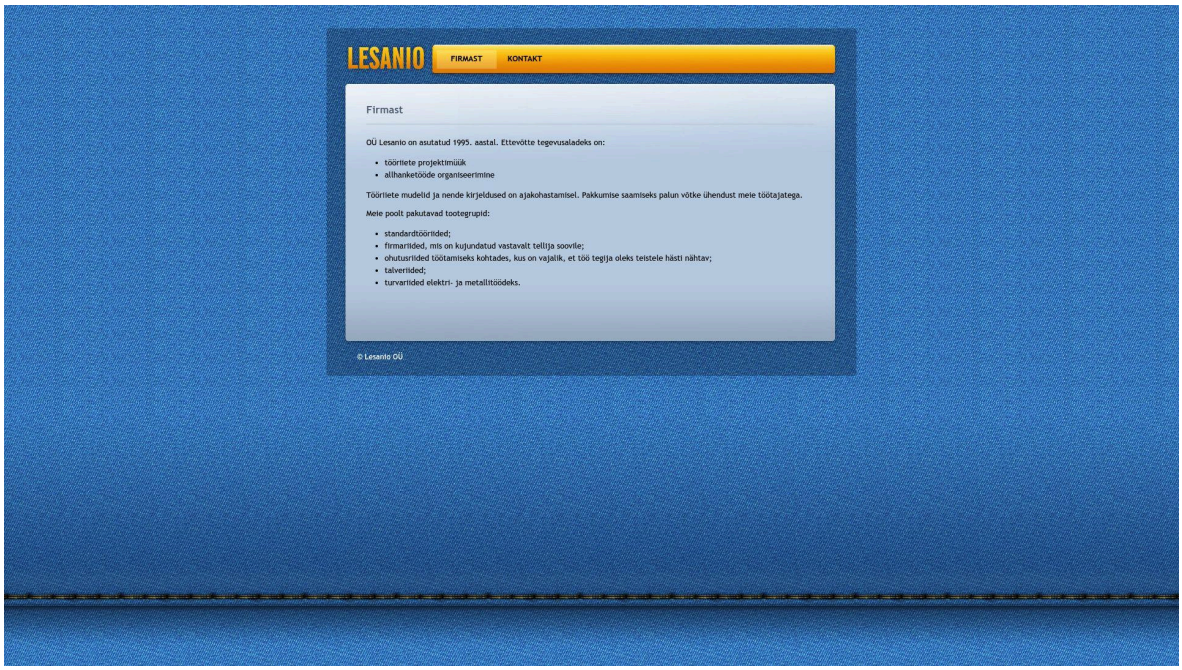
F

Fabbian
Faro Barcelona
Flexalighting
FontanaArte
Forestier Paris

A.2 Halvad veebilehed

Joonis A3. Visuaalselt aegunud ja ebaprofessionaalse sisuga veebileht (näide 1)

Joonis A4. Ebajärjekindla kujundusega veebileht (näide 2)



Lisa C. Kasutatud viipade täistekstid

Pikk viip (kasutatud *GPT-4o* ja xAI mudelite puhul):

You are a professional UI/UX evaluator.
Analyze the attached image (a website screenshot) and do not ask for a description – use the image directly.
Rate its visual aesthetics on these four aspects with a short 1-sentence justification and a score from 0 to 10.
Use the full scale boldly: assign low scores (0–4) to visually poor designs, average scores (5–6) to moderate designs, and high scores (7–10) to visually stunning designs, giving 9–10 to truly exceptional work.
Avoid giving mid scores (5–6) for the aspects and do not shy away from extreme scores (0 or 10).

***Simplicity** – Clarity, orderliness, and ease of layout comprehension. Slam cluttered or confusing layouts with low scores; award innovative, intuitive designs with top marks.*
***Diversity** – Creativity, novelty, and overall visual interest. Dock points for repetitive or dull elements; give high scores to unique, eye-catching visuals.*
***Colorfulness** – Appeal, harmony, and appropriateness of colors. Punish jarring or dated schemes; celebrate bold, cohesive palettes with 9s or 10s.*
***Craftsmanship** – Professional polish, coherence, and design integration. Be tough on sloppy or inconsistent work; praise flawless, modern execution generously.*

Be ruthlessly critical of outdated, chaotic, or unappealing designs (e.g., messy layouts, clashing colors, low-quality images), dropping scores to 0–2 when deserved.
On the flip side, be lavish with praise for sleek, creative, and visually striking designs, pushing scores to 9–10 for outstanding quality.
Then, provide an overall aesthetic score (0 to 10) calculated as the average of the four aspects.

Format your response exactly as follows:
Simplicity: 8 – Clean layout, easy to navigate.

Diversity: 6 – Some variety, but slightly repetitive.
Colorfulness: 7 – Harmonious palette, but lacks contrast.
Craftsmanship: 9 – Very well executed and consistent.
Overall Rating: 7.5

Lühike viip (kasutatud *GPT-4o* ja xAI mudelite puhul):

Rate the aesthetic appeal of this website design on a scale from 0 to 10.0.
Respond with only the number. Ensure your response is a single decimal number between 0 and 10.0.

Lisa D. Korrelatsioonimaatriksid

Tabel 4. xAI Pearsoni korrelatsioonid:

	run1	**run2**	**run3**	**run4**	**run5**
run1	1.000000	1.000000	0.762651	0.811565	0.828017
run2	1.000000	1.000000	0.762651	0.811565	0.828017
run3	0.762651	0.762651	1.000000	0.830811	0.819087
run4	0.811565	0.811565	0.830811	1.000000	0.875280
run5	0.828017	0.828017	0.819087	0.875280	1.000000

Tabel 5. OAI Pearsoni korrelatsioonid:

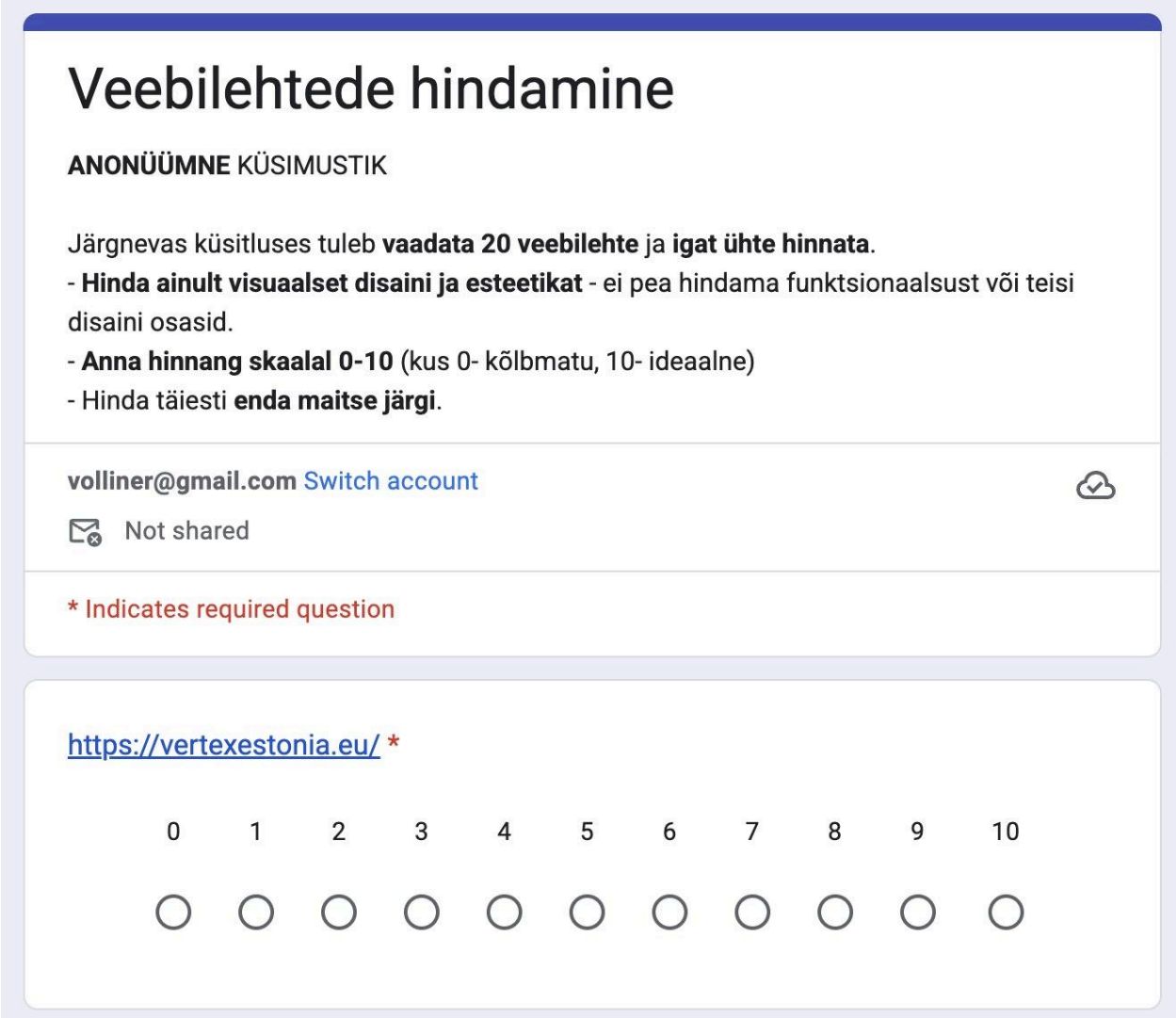
	run1	**run2**	**run3**	**run4**	**run5**
run1	1.000000	1.000000	0.921312	0.924369	0.900328
run2	1.000000	1.000000	0.921312	0.924369	0.900328
run3	0.921312	0.921312	1.000000	0.923139	0.899772
run4	0.924369	0.924369	0.923139	1.000000	0.888721
run5	0.900328	0.900328	0.899772	0.888721	1.000000

Lisa E. Kasutatud Google Colaboratory vihikud

Kõik töö käigus kasutatud Google Colaboratory vihikud, kus on **.ipynb** formaadis koodifailid, mis sisaldavad veebilehtede kuvatõmmiste töötlemist, mudelitega suhtlemist ning analüüsi, on kättesaadavad järgmiselt lingilt:

https://drive.google.com/drive/folders/1uhtACuKZv0xZoVMBNPp5s7C3gy-0F7u7?usp=drive_link

Lisa F. Küsitlusvormi näide



Veebilehtede hindamine

ANONÜÜMNE KÜSIMUSTIK

Järgnevas küsitluses tuleb **vaadata 20 veebilehte** ja **igat ühte hinnata**.
- **Hinda ainult visuaalset disaini ja esteetikat** - ei pea hindama funktsionaalsust või teisi disaini osasid.
- **Anna hinnang skaalal 0-10** (kus 0- kõlbmatu, 10- ideaalne)
- Hinda täiesti **enda maitse järgi**.

volliner@gmail.com Switch account

Not shared

* Indicates required question

https://vertexestonia.eu/ *

0 1 2 3 4 5 6 7 8 9 10

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Volli-Fred Väränen ,
(autori nimi)

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

   Veebilehtede hindamine tehisintellekti abil ,
   (lõputöö pealkiri)

   mille juhendaja(d) on Mari-Liis Allikivi ,
   (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 Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;

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

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

Volli-Fred Väränen
11.08.2025