Tartu Ülikool
Sotsiaalteaduste valdkond
Psühholoogia instituut

Kersten Kõrge

LLM AFEKTIIVSETE TÕLGENDUSDIMENSIOONIDE HINDAJANA

Magistritöö

Juhendajad: Andero Uusberg

Tartu 2025

LLM afektiivsete tõlgendusdimensioonide hindajana

Kokkuvõte

Tõlgendusteooriad käsitlevad afekti, kui subjektiivse kognitiivse protsessi tulemusena tekkivat reaktsiooni. Tänapäeva tekstipõhises suhtlusest küllastunud maailmas, võivad suurtel keelemudelitel (LLM) põhinevad tehnoloogiad pakkuda paindlikku ja kiiresti kohaldatavat täiendust inimeste olukorra analüüsi nõudvatele lahendustele. Käesolevas töös testisime üldkasutatava LLM lahenduse OpenAI GPT-4 reliaablust ja täpsust inimeste emotsionaalsete olukordade kirjelduste tõlgendamisel. Leidsime, et küsimustiku kontekst, üksiküsimuste tähendusväli, skaala sõnaline tähendus, sündmuste kirjelduste pikkus ja GPT-4-le omased teoreetilised seosed täpsust ja tõlgenduste- vahelisi seoseid mõjutavad.

Märksõnad: GPT-4, LLM, tõlgendus, afekt

LLM as the annotator of affective appraisals.

Summary

In appraisal theories, affect is conceptualized as the outcome of a subjective cognitive appraisal process. In a world increasingly saturated by text-based communication, large language model (LLM) based solutions offer a flexible and fast complement to services requiring understanding of a person's situation. This study evaluated the reliability and accuracy of OpenAI's GPT-4's appraisals of descriptions of emotional events. Our findings indicate that factors such as task context, the semantic scope of individual items, scale phrasing, event description length, and GPT-4's inherent theoretical associations significantly influence the consistency and interpretive accuracy of appraisals.

Keywords: GPT-4, LLM, appraisal, affect

Sissejuhatus

Appraisal Theory

Inimeste emotsioonide mõistmine on olnud üks psühholoogia peamistest väljakutsetest valdkonna algusaegadest. Kaks peamist kirjeldavat lähenemist emotsioonidele on kategooriline, mis käsitleb emotsioone kui diskreetseid kategooriaid (Panksepp, 1998, 2012), ning dimensionaalne, mis käsitleb emotsioone kui seisundeid, mis varieeruvad emotsionaalsetel dimensioonidel nagu meeldivus ja intensiivsus (Russell, 1980, 2012).

Emotsioonide kaardistamise kõrval on järgmine oluline ülesanne seletada, kuidas emotsioonide variatiivsus tekib. Enamik autoreid peavad emotsioone reaktsioonideks olukorrale, mis koosnevad mitmest komponendist sh teadvustatud tunne; autonoomse närvisüsteemi vahendatud füsioloogilised muutused; väljendused näos, hääles ja mujal; motivatsiooniline alge teatud moel tegutsemiseks ning olukorra tõlgendus. Emotsioone seletavad käsitlused erinevad sageli selle poolest, kuidas nad määratlevad nende komponentide omavahelisi suhteid.

Üks mõjukas koolkond on tõlgendusteooriad (*appraisal*), mis vaatleb emotsionaalset episoodi kui kognitiivse protsessi tulemust, mille aktivatsioon tuleneb indiviidi erinevatest tõlgendustest sündmustele (Lazarus, 1991; Ortony jt., 1988; Scherer jt., 2001; Smith & Ellsworth, 1985). See, kuidas indiviid olukorda tõlgendab, näiteks kui tähtis, ohtlik, kasulik, ootamatu see on, mõjutab emotsionaalset afekti, nii selle intensiivsust, kui spetsiifilisust, on adaptiivne mehhanism, mis reflektiivselt toetab meie toimetulekut (Gross, 2015; Lazarus, 1991; Moors jt., 2013; Yih jt., 2020). Empiirilised uuringud mudeldavad tõlgendust sageli, kui mehhanismi, mis motiveerib toimetulekut, mis omakorda käivitab selleks emotsiooni (Dunkley jt., 2017; Sladek jt., 2016). Tõlgendusteoorial põhinevate uuringute mudelid erinevad selle poolest, et tõlgendus kutsub eile emotsiooni, mis motiveerib toimetulekut (Uusberg jt., 2023; Yih jt., 2020). Olenemata sellest, kas toimetuleku eesmärk on hakkama saada tõlgenduse või emotsiooniga, on mõlema vaate puhul protsessi sisendiks tõlgendus. Mõistmine, milliste parameetrite alusel inimesed olukordi hindavad, võib anda meile praktilisema pildi, et mõista inimese seisundit, selle muutumist, mõista tema vajadusi, ning ennustada käitumist.

Suured Keelemudelid

Et emotsioone mõista, on muuhulgas kasulik lähtuda inimeste endi kirjeldustest. Koos suurte keelemudelite (LLM) arenguga on meil võimalik üha operatiivsemalt ja mahukamalt töötada inimeste kirjeldustega, nii tõlgenduste uurimiseks, kui esmataseme vaimse tervise sekkumiste arendamiseks. Aeg surub selleks juba takka, paljud inimesed juba otsivad juba ise

emotsionaalset tuge GPT-delt (Salah jt., 2024), ning arendatakse sekkumisi ilma eetiliste raamistiketa. GPT-d väljendavad ennast toetavalt, ning neid nähakse empaatilisematena, kui teisi inimesi (Lee jt., 2024), ja eriti arste (Ayers jt., 2023).

Loomuliku keeletöötluse (NLP) uuringud on emotsioonide kontekstis peamiselt keskendunud baasemotsioonide uurimisele läbi kategooriliste ja dimensionaalsete mudelite (Kim & Klinger, 2019), peamiselt luues leksikone, ning *valents-erutus* telgedel hinnatud korpuseid. LP mudeleid kasutatakse, et hinnata teksti emotsionaalseid parameetreid, millest olenevalt katse kontekstist võiks järeldada teksti autori kogetud seisundit või seisundit, mida tekst lugejas võiks tekitada. Eelkirjeldatud uuringusuundade eesmärgiks on olnud leida emotsioonidega seotud universaale. Vaid vähesed tööd on lisaks uurinud võimalusi tuvastada tekstis tõlgendusi. Nende tööde eesmärk on olnud tekstis tõlgenduste leidmine, ning nende põhjal emotsioonide ennustamine (Hofmann jt., 2020, 2021; Troiano jt., 2023).

Peamiselt loovad uuringud selleks spetsiaalselt peenhäälestatud mudeleid, mille tekstikorpus koosneb mudeli põhjal hinnatud tekstidest. Hindajale esitatakse valik emotsioonide ja tõlgenduste kohta käivaid väiteid, mille kohta hindaja vastab, kuivõrd ta hinnatava teksti kohta nendega nõustub. Tekste on hinnanud kas tekstide autorid, kes on kogenud tekstis kirjeldatud emotsionaalseid sündmusi (*experiencer-annotator* ehk kogeja-hindaja paradigma), või lugejad, kes püüavad autorite kogemusi teksti põhjal järeldada (*reader-annotator* ehk lugeja-hindaja paradigma). Esimese puhul on raske kokku saada suurt hulka sihtotstarbeliselt kogutud emotsionaalselt laetud tõepäraseid kirjeldusi, koos detailsete hinnangutega, teiste puhul säilib suurem töömaht ja ohus on hinnangute valiidsus. Kuna andmehulgad peavad olema suured, siis neid kogutakse näiteks massiliselt õpilastelt (Troiano jt., 2019) või sotsiaalmeediast, kus emotsioonide klassifikaatoritena kasutatakse näiteks *hashtag*'e (Mohammad, 2012). Suured keelemudelid võiksid pakkuda käepärast, kiiresti skaleeritavat, paindlikku ja mis kõige tähtsam, väsimatut ressurssi, et emotsionaalset hindamist lihtsustada (Inzlicht jt., 2024).

Käesoleva töö eesmärk on testida laiatarbe LLM-i sobivust inimeste olukorrakirjeldustest tõlgenduste lugeja-hindajana (*reader-annotator*, edaspidi hindaja). Uurime tõlgendusdimensioone uue osapoole pilgu läbi – nimelt hindab samu olukordi autorite kirjelduste läbi Azure OpenAI GPT-4, et mõista, kui sobiv on antud keelemudel tõlgendusdimensioonide hindajana.

Vähesed olemasolevad värsked uuringud on näidanud, et LLM-d suudavad tuvastada tõlgendusi enesekohastest lugudest (Zhan jt., 2023) ja toodete tagasisidedest (Yeo & Jaidka, 2023). Samuti realiaabselt, kuid erineva tundlikkusega hinnata psühhomeetrilise testi (Perrez

& Reicherts, 1992) põhjal stressi tõlgendusi, kuid mitte toimetulekuga seotuid (Yongsatianchot jt., 2023). Tak & Gratch, 2023 uurisid, kas GPT suudab arutluse käigus välja tuua tõlgendusi ja leidsid, et GPT ennustused langesid kokku inimeste tõlgenduste ja emotsioonidega, kuid mitte emotsioonide intensiivsuse ja toimetulekuga seotud tõlgendustega. Samuti polnud GPT korduskatsel reliaabne, kuid tulemused paranesid promptiga katsetamisel.

Käesolev uuring kasutab Uusberg jt., 2023 kogutud andmestikku, mis koosneb kogutud emotsionaalsete sündmuste kirjeldustest, ning autorite hinnangutest kogetud emotsioonidele ning sündmuste tõlgendustele. Uurime GPT-4 reliaablust ja täpsust tekstide hindajana võrreldes autoriga, ning uurime sündmuskirjelduste omaduste mõju täpsusele.

Uurimisküsimused

Uuringu eesmärk on testida Chat-GPT kasutust inimeste tõlgenduste hindajana. Selle tarbeks kontrollime GPT-4 reliaablust ning esitame 5 uurimisküsimust (K1-K5), et avastada erinevate ülesandepüstituste ja teksti omaduste mõju GPT-4 tulemustele.

Uurimisküsimus 1 (K1): Kui reliaabne on GPT-4 oma tõlgendustes?

GPT-4 reliaabluse hindamiseks kordame GPT-4 ülesannet koguvalimil algprompti läbi 5 korda, ning hindame paariti hinnatud reliaablus näitajate keskmistamise läbi. K1 hinnang annab meile baasi, et võrrelda järgnevate promptide mõju tulemusele.

Uurimisküsimus 2 (K2): Kui täpselt ennustab GPT-4 autori tõlgendusi?

GPT-4 tõlgenduste täpsuse hindamiseks võrdleme K1 kogutud tõlgendusdimensioonide hinnanguid autorite hinnangutega,.

Uurimisküsimus 3 (K3): Kuidas mõjutab GPT täpsust skaala kirjeldamine?

Alguurimuses hindasid tekstide autorid tõlgendusi Likerti skaalal 1-9. Numbriliste skaalade tähenduse tunnetus on samuti tõlgenduslik ülesanne. Uurime, kuidas mõjutab GPT-4 tulemusi, kui antud skaala väärtustele on antud lisaks sõnaline tähendus, mis oli presenteeritud algses küsimustikus.

Uurimisküsimus 4 (K4): Kuidas mõjutab GPT täpsust ülesanne hinnata igat dimensiooni eraldi?

Uuringud on näidanud, et andes mudelile tõlgenduste hindamise eel ülesande formuleerida sündmuse kirjelduses esinev keskne emotsioon mõjutab hinnangute täpsust halvemaks (Ruder jt, 2025). Sellest lähtuvalt võivad ka GPT hinnangud tõlgendus-dimensioonidele olla üksteisest mõjutatud, ühe tõlgenduse raames olevad dimensioonid

võivad olla kontseptuaalsed vastandid, näiteks *congruence* ja *incongruence,* samuti Vastutust tõlgendavad dimensioonid võivad üksteist välistada, vastavalt kellele vastutus tekkinud olukorra eest omistatakse või mitte. Katsetame, kas andes ühe päringu raames ülesanne hinnata igat dimensiooni iseseisvalt mõjutab täpsust.

Uurimisküsimus 5 (K5) Kuidas mõjutab teksti pikkus GPT täpsust?

Tõlgendusdimensioone, mida ma vaatleme, on 10, ning sündmuste kirjelduste pikkus on varieeruv. Uurime, kuidas tekstide pikkus mõjutab tõlgenduste täpsust dimensioonide lõikes. Sündmuste kogejad võivad kirjeldada oma tekstis oma kogemusi väga erinevalt – näiteks kirjeldades oma emotsioone, kirjeldades oma tekkinud olukorra kausaalsust, nii ebaõiglust, mille tõttu olukord tekkinud on, kui näiteks seda, mis selle tagajärjel juhtuda võib. Oletame, et ka tõlgenduste mõttes erinevate inimeste jaoks on erinevad dimensioonid erineva tähtsusega. Soovime avastada, kuidas teksti pikkus mõjutab täpsust dimensioonide lõikes. Kas leidub optimaalse pikkusega tekst, kust edasi kvaliteet ei parane? Kas dimensioonid eristuvad selle põhjal, kui vara on neid võimalik tekstist tuvastada?

Autori panus

Käesolev töö sündis koostöös Tartu Ülikooli afekti ja regulatsiooni uurimisrühmaga. Azure API promptimise algkood pärineb (Ruder jt., 2025) uuringust 2024 aasta I kvartali seisuga. Töö autor viis iseseisvalt läbi kõik uurimistöö etapid (sh uurimisküsimuste püstitamise ja katsekeskkonna programmeerimise, analüüsimeetodite valiku).

Meetod

Algse uuringu andmestik

Kasutasime (Uusberg jt., 2023) uuringu andmestikku, kes uurisid tõlgenduste ümberhindamise mõju emotsioonidele. Uuringus palusid Uusberg jt. inimestel kirjeldada üht viimase kahe kuu jooksul endaga toimunud sündmust, ning selle kohta vastata, kuidas nad emotsioonide ja tõlgendusdimensioonide olukorda hindavad. Seejärel aga tutvustati ümberhindamist kui eneseregulatsiooni tehnikat, paluti situatsiooni ümber tõlgendada, ning täita küsimustik uuesti.

Katses osalesid 2018. aastal ühe semestri jooksul erinevad San Francisco Bay Area kolledžite õpilased ainepunktide eest. Konkreetset valimi suurust algsel kogumisel ette ei määratletud. Kuigi osalejate vanused olid vahemikus 18 kuni 65, siis valdav enamus neist olid noored täiskasvanud, keskmine vanus 22.46, (M = 22.48, Mdn = 20, SD = 7.26, 80% =< 23). Puudulike andmetega osalejad eemaldati valimist originaaluuringu käigus.

Originaaluuringu andmetest võtsime üle 437 ümberhindamise eelset sündmuse kirjeldust, ning 10 tõlgendusdimensioone hindavat mõõdikut.

Joonis 1.

Tõlgendusdimensioonide hinnangute jaotuvus algandmestikus.

Märkus. Punasega on tähistatud hinnangud, mida kasutas vähemalt 200 osalejat (N=427).

Andmed

Valitud andmestik koosneb 437'st sündmuse kirjeldusest (*event*), ning sündmuse kogeja tõlgendustest olukorrale. Autorite tõlgendused koguti, paludes neil hinnata kirjeldatud sündmust läbi 10 väite, mis kirjeldasid 5-t peamist tõlgendusdimensiooni: "Relevantsus", "Kongruentsus", "Vastutus", "Tagajärgede tõenäosus" ja "Toimetulekupotentsiaal" (Moors jt., 2013). Iga tõlgendus oli sõnastatud väitena, mis algas; „*In this situation, I have a sense that* ...". Osalejad vastasid 9-pallisel Likert skaalal, mis oli esitletud läbi numbrite ja pealkirjade: *strongly disagree* (1)–(2)–(3)–(4)–*neither* (5)–(6)–(7)–(8)–*strongly agree* (9).

Piloteerides andmestikku jäi lõplikuks valimiks 427 originaaluuringu 437-st. 10 juhtumikirjeldust välistasime GPT analüüsi protsessis, kuna nendes sisaldus kas enesevigastamist, vägivalda või nendega seotud sõnu, ning need ei läbinud Azure OpenAI prompti sisueeskirju.

Andmete jaotuvuses esines iseärasusi. Algne ülesanne oli kirjeldata viimase 2 kuu jooksul toimunud tugevat negatiivset emotsiooni esile kutsunud sündmust, mis seletab tugevat lae ja põranda efekti tõlgendustest *relevance* ja *congruence* (vt Graafik 1). Andmed pole normaaljaotuslikud, mis on ootuspärane, kuna inimesed lähtusid erinevatest sündmustest.

Tabel 1.

Tõlgendusdimensioonide mõõdikute reliaablus.

Tõlgendusdimensioon	Mõõdik	Väite lõpp	ICC
Relevantsus	*relevance*	… this situation matters to me	0.76
Kongruentsus	*congruence*	… this situation is potentially desirable for me	0.87
	incongruence	… this situation is potentially harmful for me	**0.48**
Vastutus	*self_accountability*	… I am responsible for this situation	**0.28**
	other_accountability	… someone else is responsible for this situation	**0.31**
	impersonal_accountability	… no one is responsible for this situation	0.64
Tagajärgede tõenäosus	outlook_certainty	… I don't know how this situation is going to turn out	0.73
Toimetuleku-potentsiaal	outlook_valence	… one way or another, this situation is going to turn out well	0.65
	situational_coping	… I can change this situation for the better	0.61
	emotional_coping	… I can accept whatever happens in this situation	0.72

Märkus. Uusberg jt. hindasid pilootuuringul (*N=60*) mõõdikute reliaablust keskmiseks. 7 saavutasid klassisisese korrelatsioonikordaja (*ICC*) taseme 0,5-0,75 (Koo & Li, 2016).

Tarkvara

Mudelit promptiti läbi Tartu Ülikooli poolt tagatud Azure OpenAI REST API teenuse. Tekstide hindamiseks kasutasine GPT-4 mudelit 0613, kasutades tavaparameetreid (temperature 0,7; top_p =0,95) (https://platform.openai.com/docs/models/gpt-4).

Andmete kogumiseks ja analüüsiks kasutati keskkonda Jupyter Notebook Python (Kluyver jt., 2016) ning tarkvara RStudio (RStudio Team, 2020). Graafikud on joonistatud pakette *ggplot* v3.5.1 (Wickham jt., 2024), *ggforce* v0.5.0 (Pedersen, 2025), *patchwork* v1.2.0.9 (Pedersen, 2024), ning Matplotlib (The Matplotlib Development Team, 2025).

Prompt

GPT-4 ülesandepüstitus koosneb tehniliselt kahest osast – süsteemi (*system*) ja kasutaja (*user*) sisust (*content*). Süsteemiprompti sisu (content) seab ülesande (completion) üles kahes osas: kontekst (C) ja instruktsioon (I). Kasutaja prompti moodustab sündmuse kirjelduse tekst (T).

Kontekst (C) koosneb rollist ja ülesande kirjeldusest. Roll või persoona on levinud promptimise meetod, mis on esineb kõikides vaadeldud uuringutes. Huvitaval kombel puuduvad andmed ja viited antud meetodi mõjule, isegi promptimise tüpoloogiaid käsitlevates teaduslikes artiklites (Braun jt., 2024; Hewing & Leinhos, 2024; Santu & Feng, 2023; White jt., 2023). Käesoleva uuringu raames hoidsime rollikirjelduse ühtlasena kõikide uurimisküsimuste raames, et hoida tingimusi sarnasena samas uurimisgrupis paralleelselt läbi viidud katsega (Ruder, 2025). Ülesande kirjelduse osa (I) kujundab konteksti eelolevaks instruktsiooniks.

[C] *You are an expert in human psychology. You will read a short text someone has written about an emotional event or a situation. Then you will read 10 statements that describe some aspect of how people think about different situations.*

Instruktsioon (I) koosneb kolmest osast. Ülesanne, vastuse formaat, ning tõlgendusdimensioone hindavate parameetrite nimekiri. Ülesandeks oli ära arvata, kuidas autor tõlgendas tekstis (T) kirjeldatud sündmust, ning hinnata skaalal 1-9. Soovitud formaadiks on .csv fail, mille tulemusena GPT vormistab tagaside kaheveerulise tabelina, kus esimeses veerus on väide, ning teises numbriline hinnang. Parameetrite nimekiri esitab tõlgendus-dimensioone kirjeldavate väidete nimekirja kujul "*"In this situation, I have a sense that* <väite lõpp> *""* (vt Tabel 1).

[I] *Instruction: Your task is to guess how the author appraised and interpreted the situation described in the text. Assign integer weights varying from 1-9.*

Desired format: csv file

List of parameters: <Väidete nimekiri>

Katsetingimuses K3 ja K4 täiendasime instruktsiooni skaala punktide tekstilise kirjeldusega, mis olid kirjeldatud ka autoritele:

[I3] “*An integer weight of 1 meaning means to strongly disagree, 9 to strongly agree and 5 doing neither.*”.

Tingimuses K5 instrueerisime GPT-d lisaks tegema kõik hinnangud eraldiseisvalt eelnevatest vastustest.

[I4] “*Make every guess independent of previous answers.*”

Teksti (T) prompt esitati GPT’le eraldi kasutaja promptina, ning koosnes kogeja sündmusekirjeldusest.

[T] *"Ex-wife telling me to give her more money for our children's expenses. She threatens me with financial ruin and court appearances unless I give her more money. I already contribute more than enough for the children. This scenario makes me feel helpless, scared, sad, threatened, and like I just want to retreat somewhere. I'm not a bad father but she continues to attempt to paint me as such. This has gone on for nearly 20 years. I'm tired."*

Tulemused

Mõõdikute valik

Reliaabluse ja täpsuse hindamiseks kasutasime peale Spearmani korrelatsioonikordaja nii RMSE’d, kui MAE’d. Kuna teoreeritliselt võib GPT-4 info puudumisel anda polari-seerivaid hinnanguid, siis RMSE ja MAE suhte analüüs võib meile anda erandite mõjust.

K1: GPT-4 tõlgenduste reliaablus

GPT-4 tõlgenduste reliaabluse hindamiseks läbisime katse uurimisküsimuse K1 algpromptiga 5 korda. Seejärel võrdlesime GPT-4 reliaablust korduste vahel. Reliaabluse hindamiseks arvutasime iga mõõdiku puhul korduvmõõtmiste vaheliste ruutkeskmiste vigade (apRMSE), keskmiste absoluutvigade (apMAE), ja Spearmani korrelatsioonikordajate (ρ) keskmised. Et tuvastada erandite mõju, arvutasime ka apRMSE ja apMAE vahelise suhte.

Esimesena on näeme kohe *relevance* madalaid RMSE ja MAE, kuid samas teistega võrreldes madal korrelatsioon. Võib oletada, et see on põhjustatud tugevast laeefektist tulenevast madalast variatiivsusest, mida näeme ka alghinnangute jaotuvusest (vt Graafik 1). Autoreid instrueeriti kirjeldama viimase kahe kuu jooksul aset leidnud „tugevat negatiivset emotsiooni” esile kutsunud sündmust. Tekstide sisu tekitab *relevance*, *congruence* (võimaliku peegeldusena ka *incongruence*) suhtes väga väikse variatiivsuse, mis peegeldub ka GPT tõlgendustes (vt Graafik 2). Näeme ka mitmekordset vahet erinevate tõlgendusdimensioonide RMSE ja MAE keskmistes, mille puhul võib oletada, et tänu väiksele variatiivsusele mõjuvad GPT-4 konservatiivsemad hinnangud äärmuslikena.

Tabel 2.

5 korduse keskmised paariti RMSE, MAE ja Spearmani korrelatsioonikoefitsendid.

Tõlgendusdimensioon	Mõõdik	apRMSE	apMAE	apRMSE/ apMAE	ρ
Relevantsus	*relevance*	**0.27**	**0.06**	**4.37**	**0.47**
Kongruentsus	*congruence*	0.85	0.40	**2.14**	0.74
	incongruence	0.90	0.52	**1.73**	0.74
Vastutus	*self_accountability*	1.21	0.78	**1.55**	0.89
	other_accountability	1.10	0.69	**1.60**	0.90
	impersonal_accountability	1.23	0.78	**1.58**	0.62
Tagajärgede tõenäosus	outlook_certainty	1.04	0.69	**1.51**	0.67
Toimetuleku-potentsiaal	outlook_valence	1.10	0.80	1.38	0.74
	situational_coping	1.22	0.90	1.36	0.56
	emotional_coping	0.96	0.68	1.41	0.67

Märkus. Kõik Spearmani korrelatsioonikoefitsiendid oli statistiliselt olulised ($p < 0.001$).

Kolme "Vastutus"-t hindava mõõdiku, ning "Tagajärgede tõenäosus"-e alamdimensiooni *outlook_reliability* RMSE ja MAE suhe on samuti üle 1,5. Kolme mõõdiku variatiivus on lai, kuid kõigi kolme puhul näeme äärmuslikke 1-palli hinnanguid (Joonis 1.). Ülejäänud mõõdikute puhul jääb suhe alla 1,5 lähedale või alla selle.

Kõigi puhul on esineb tugev positiivne korrelatsioon. Väiksemana torkab silma 0.56 *situational_coping*, mis on oma olemuselt metakognitiivne hinnang, mis oleneb rohkem inimese enda toimetulekuoskustest, mitte olukorra omadustest. Kuna RMSE ja MAE varieeruvad 1 punkti ümber ja alla selle, siis 9-pallise Likerti puhul võib GPT hinnanguid pidada mõõdukalt reliaabseks.

K2: GPT-4 tõlgenduste täpsus

Olles hinnanud GPT-4 sisemist reliaablust üle korduvkatsete, võrdleme, kui täpselt GPT-4 ennustab teksti autorite hinnanguid. Kasutame eelmises uurimisküsimuses teostatud mõõtmisi. Arvutasime K1 tulemuste keskmised, ning võrdlesime autorite hinnangutega.

Näeme sarnast mustrit võrreldes K1-ga – *relevance* ja *congruence* viitavad suurele erandite mõjule, võib eeldada, et *relevance* väike korrelatsioon tuleneb laefektist. Samas näeme väiksemat erandite mõju vähenemist üle dimensioonide, mis võib tulla keskmistamisest. Tulemused on sarnased sama meetodiga paralleelse uuringu tulemustega

Tabel 3.

Algpromptiga läbitud GPT-4 ja autori hinnangute võrdlus.

Tõlgendusdimensioon	Mõõdik	RMSE	MAE	RMSE /MAE	ρ	p
Relevantsus	*relevance*	1.88	1.05	1.79	0.19	**0.00**
Kongruentsus	*congruence*	2.19	1.39	1.57	0.32	**0.00**
	incongruence	**3.08**	**2.31**	1.33	0.32	**0.00**
Vastutus	*self_accountability*	2.43	1.82	1.33	0.62	**0.00**
	other_accountability	2.56	1.83	1.40	0.61	**0.00**
	impersonal_ accountability	2.96	2.17	1.36	0.43	**0.00**
Tagajärgede tõenäosus	outlook_certainty	3.10	2.40	1.29	0.23	**0.00**
Toimetuleku-potentsiaal	outlook_valence	2.72	2.22	1.22	0.12	**0.01**
	situational_coping	3.12	2.66	1.18	0.06	**0.25**
	emotional_coping	3.25	2.69	1.21	0.06	**0.25**

RMSE 1.15-2.01 (Ruder jt. 2025). Uuring hinnangud olid 5 palliselt Likertil (nRMSE 0,29-0,50), võrreldes käesoleva uuringuga (nRMSE 0.27-0,41). Oluliselt eristub aga *incongruence*, mis kuigi on sama korrelatsiooniga nagu *congruence*, näitab lausa 1-punktilist veamäära erinevust. Olukorra eest "Vastutus"-t hindavad mõõdikud omavad tugevaid korrelatsioone, kui veamäär siiski suur.Toimetulekut hindavad tõlgendused on suure veamääraga ja *_coping* dimensioonid ei saavuta statistilist olulisust.

Käesolevate tulemuste puhul, kui RMSE ja MAE ületavat tugevalt 2 ja isegi 3, võib esineda tekste, kus GPT-4 hinnang ületab hinnagute polaarsuse joont 5, olles vastupidisel arvamusel autoriga.

K3: Skaala tähenduse mõju GPT-4 täpsusele

Uuringu tulemusi võib mõjutada mitte ainult sündmuse teksti tõlgendamine, vaid ka küsimustiku skaala tõlgendamine. Skaala tähenduse sõnastamine võib aidta vältida esineda polaarsuse erinevust dimensioonidel, kus on suur RMSE. Taolise vea vähendamiseks täiendasime instruktsiooni küsimustikku sõnastatud skaala kirjeldusega I3.

Skaalaga kirjelduse läbi enamates näitajatest olulist muutust ei tekkinud. Näeme, et *incongruence RMSE ja MAE* halvenesid kergelt, ning korrelatsioon langes 0,32'lt 0,2-le. Jaotuvustabelilt (vt Joonis 1) on näha, et *incongruence*, kuigi kontseptuaalselt vastanduv

congruence'le, omab tagasihoidlikumaid hinnanguid skaala positiivses osas 6-9. Skaala toomine võis läbi keelelise seose suunata GPT-d panema äärmuslikemaid hinnanguid.

Tabel 4.

Skaalaga täiendatud prompti (I3) ja autori hinnangute täpsus.

Mõõdik	RMSE	MAE	RMSE /MAE	Spearman ρ	p
relevance	1.87	1.04	1.80	0.20	0.00
congruence	2.13	1.35	1.57	0.29	0.00
incongruence	**3.25**	**2.40**	1.35	**0.20**	0.00
self_accountability	2.46	1.78	1.38	0.61	0.00
other_accountability	2.55	1.79	1.42	0.61	0.00
impersonal_accountability	2.83	2.07	1.37	0.38	0.00
outlook_certainty	3.17	2.43	1.30	0.19	0.00
outlook_valence	2.93	2.33	1.26	**0.17**	**0.00**
situational_coping	**2.86**	**2.33**	1.22	**0.27**	**0.00**
emotional_coping	**2.79**	**2.33**	1.20	**0.16**	**0.00**

Hüppe tegid toimetulekuga seotud dimensioonid: Kõik kolm korrelatsiooni kasvavad ja saavutavad statistilise olulisuse. Võrreldes tulemusi jaotuvustabelil (vt. Joonis 2), näeme, et *outlook_valence* puhul on toimunud pööre negatiivsuse poole, ning s*ituational_coping* ja e*motional_coping* variatiivsus suurenes.

K4: Igat dimensiooni eraldi hindamise prompti mõju GPT-4 täpsusele

Esitasime GPT-4-le ülesandena anda kõik hinnangud eraldiseisvalt (I4), et hinnata, kas täpsus paraneb, kui GPT-4 ennustab tulemusi teisi tõlgendusi hindamata.

Tulemustest näeme, et enamikes tõlgendustes muutust ei tekkinud. *relevance* korrelatsioon langes (ρ=0,20 -> 0,11), ning kaotas statistilise olulisuse. Tekib küsimus, kas teiste tõlgendusdimensioonide mõju võis leevendada ülesandepüstitusest tulenevat *relevance* laeefekti ja tõsta müra lisadaes näitajale nüansse. +0.05 tõusis *outlook_valence* korrelatsioon. Antud ülesanne tulemusi oluliselt ei mõjutanud. Põhjuseks võib olla nii laeefekt, kuid me ei saa tagada, ilma iga tõlgendusdimensiooni kohta eraldi päringut promptimata, kuivõrd GPT-4 küsimusi iseseisvalt hindas. Ühe päringu raames eraldiseisev hindamine tulemust ei oma.

Tabel 5.

Tõlgendusdimensioone eraldiseisvalt hindava prompti ja autori hinnangute täpsus.

Mõõdik	RMSE	MAE	RMSE /MAE	Spearman ρ	p
relevance	1.88	1.06	1.78	**0.11**	**0.02**
congruence	2.16	1.38	1.56	0.26	0
incongruence	3.2	2.4	1.33	0.25	0
self_accountability	2.45	1.81	1.35	0.62	0
other_accountability	2.57	1.86	1.39	0.58	0
impersonal_accountability	2.85	2.1	1.36	0.37	0
outlook_certainty	3.1	2.38	1.3	0.22	0
outlook_valence	2.93	2.35	1.24	**0.22**	0
situational_coping	2.85	2.33	1.23	0.27	0
emotional_coping	2.77	2.34	1.18	0.15	0

K5: Teksti pikkuse mõju GPT-4 täpsusele

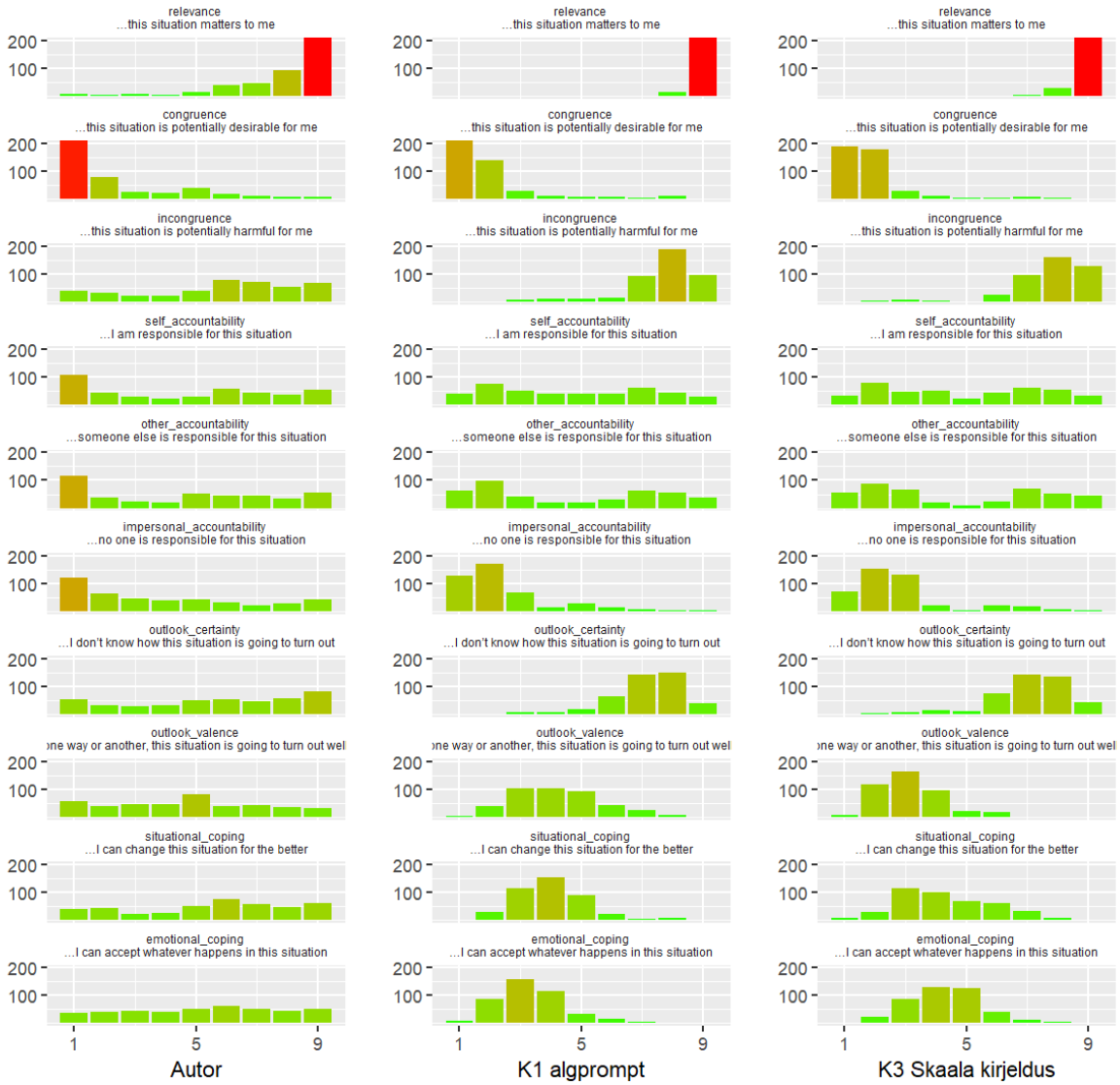
Teksti pikkuse mõju analüüsil kasutame siiani parima tulemuse saavutanud GPT-4 sooritusе K3 tulemusi. Hindasime iga dimensiooni täpsuse seost teksti pikkusega, jagades tekstid 50 tähemärgi vahemikkude alusel gruppidesse. Eemaldasime 6 teksti, mille pikkus ületas 1000 tähemärki. Jagame tekstid 50-tähemärgiste vahemike alusel gruppidesse. Tekstide ebaühtlase jaotuvuse tõttu (vt joonis 3) grupeerime kokku tekstid pikkusega üle 700 tähemärgi, 601-700 tähemärki, ning alla 0-100. Kuigi nii ’1-100’ ja ’101-150’ gruppide suurus jääb alla 20, eristus ’100-150’ tähemärgiline grupp ’1-100’-st lausete arvu poolest, ’1-100’ olid ühelauselised, ’101-150’ juba kaks või kolm lühikest lauset.

Arvutasime RMSE, MAE ja nende suhte ning tekstigruppide järjekorra Spearmani korrelatsioonid tõlgendusdimensioonide lõikes. Kuigi paljud näitasid vähenemist tähemärkide suurenedes, vaid vähesed korrelatsioonid saavutasid statistilise olulisuse.

Congruence näitas head korrelatsiooni RMSE puhul, kuid ei saavutanud seda MAE puhul. Kuigi RMSE/MAE statistiline olulisus oli p=0,05 lähedal, siis võime oletada, et mingil määral paranes tekstide pikkuste suurenedes äärmuslike vigade hulk, vähendades RMSE, kuid mitte oluliselt MAE’d. Sarnast eranditele viitavat muutust näitas *incongruence* (RMSE/MAE ρ = -0.60, p<0,05). Ainukesed kaks mõõdikud, mis näitasid olulist muutust nii RMSE, kui MAE puhul, oli *outlook_valence* ja *emotional_coping*. Tähelepanuväärne on, et tekstide pikkuse suurenedes GPT-4 täpsus ennustada inimese hinnangut kirjeldatud situatsiooniga toimetulekule hoopis halvenes.

Joonis 2.

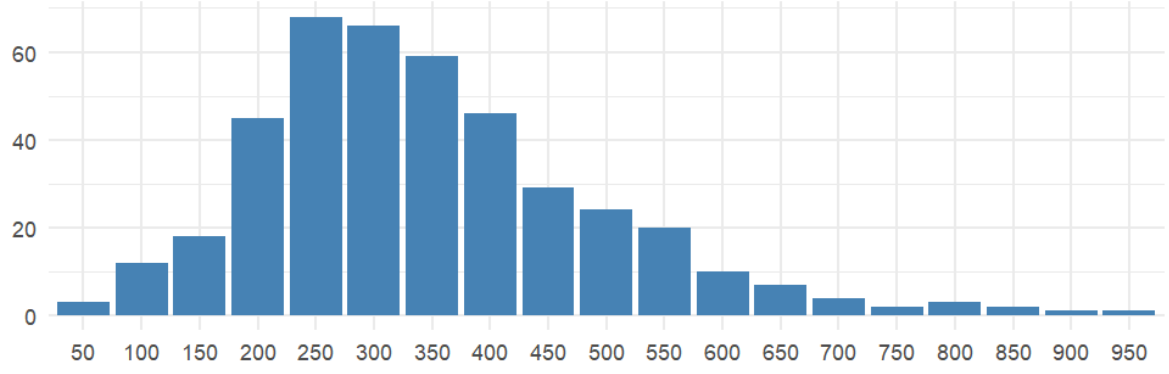
Tõlgendusdimensioonide hinnangute jaotuvuste võrdlus.



Märkus. Punasega on tähistatud hinnangud, mida valisid osalejad (N=427), või GPT-4 vähemalt 200 korda. K1 keskmised hinnangud on ümardatud täisarvudeks.

Joonis 3.

Tekstide pikkuste esinemissagedus 50 tähemärgilistes vahemikes.



Tabel 6.

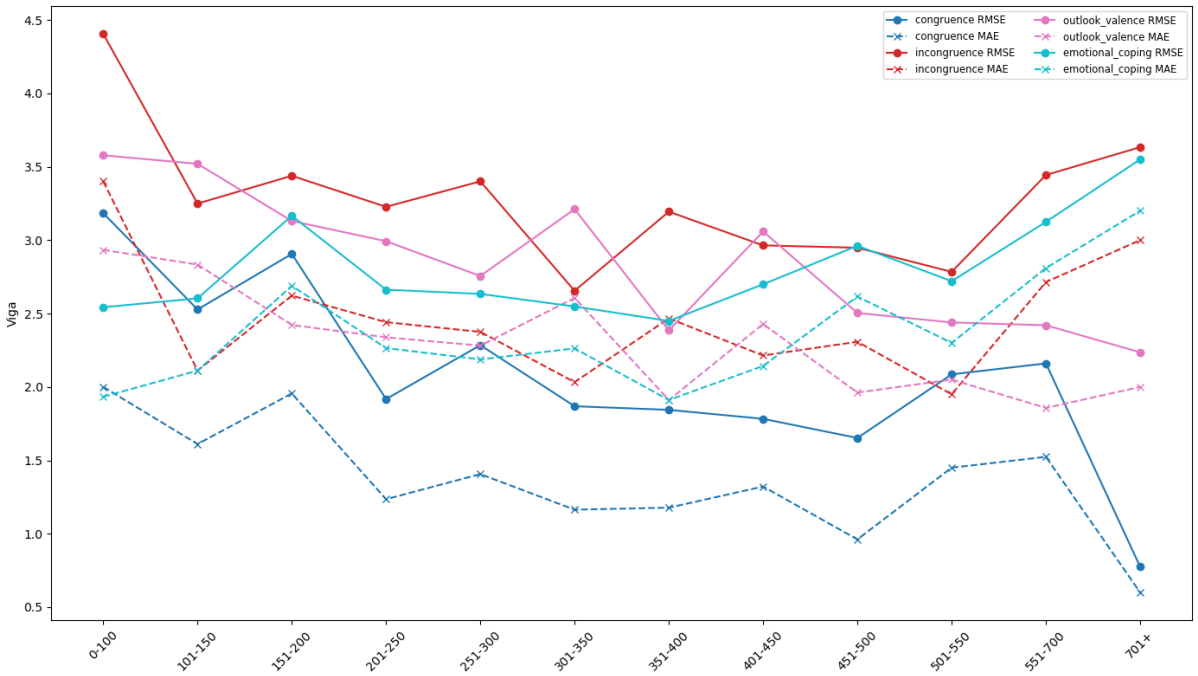
RMSE ja MAE korrelatsioonid teksti pikkustega.

Mõõdik	**RMSE *ρ***	***p***	**MAE *ρ***	***p***	**RMSE/ MAE *ρ***	***p***
relevance	-0.25	*0.42*	0.00	*0.99*	-0.37	*0.21*
congruence	**-0.62**	***0.02****	**-0.46**	***0.11***	-0.55	*0.05*
incongruence	-0.19	*0.54*	0.06	*0.84*	**-0.60**	***0.03****
self_accountability	-0.14	*0.64*	0.04	*0.90*	-0.36	*0.23*
other_accountability	-0.17	*0.58*	-0.25	*0.40*	0.18	*0.57*
impersonal_accountability	0.05	*0.86*	0.02	*0.94*	0.10	*0.75*
outlook_certainty	-0.16	*0.60*	-0.12	*0.69*	-0.18	*0.57*
outlook_valence	**-0.79**	***0.00****	**-0.71**	***0.00****	-0.20	*0.52*
situational_coping	-0.45	*0.13*	-0.45	*0.13*	-0.02	*0.96*
emotional_coping	**0.66**	***0.01****	**0.63**	***0.02****	**-0.63**	***0.02****

Märkus. * p<0,05, ** p<0,01

Joonis 4.

RMSE, MAE korrelatsioonid tekstide pikkuste suhtes.



Võrdlev korrelatsioonianalüüs.

Mitmed tõlgenduste mõõdikud on omavahel kontseptuaalselt ja sümboolselt seotud, näiteks enda või kellegi teise vastutus olukorra üle (*self_* ja *other_accountability*), või näiteks hinnang olukorra meeldivusele (*congruence*) ja usule, et kõik olukord lõpeb hästi (*outlook_valence*). Keelemudel, lähtudes oma korpuse andmete keelelisest olemusest võib seostada dimensioone erinevalt neid kogenud tekstide autoritest. Viisime läbi tõlgendus-dimensioonide omavaheliste korrelatsioonide võrdluse, et uurida seoseid katsetingimuste lõikes. Otsustasime Spearmani korrelatsioonikordajate võrdluse kasuks, ning avastusliku faktoranalüüsi kahjuks, sest K1 ja K3 Kaiser-Meyer-Olkin (KMO) testi tulemused jäid alla 5.

Tulemustes on palju ootuspärast, kuid samas ka suuri muutusi. C*ongruence* ja *incongruence* võib keeleliselt ja loogiliselt vaadelda üksteise polaarsete vastanditena, ning on K1 ja K3 puhul tugevas negatiivses korrelatsioonis (ρ= -0.40, p < 0.001) (vt tabel 7), kuid korrelatsiooni analüüs toetab jaotuvuse visuaalset hinnangut – inimhindaja neid nii tugevate vastanditena ei näe (ρ= -0.18, p < 0.001). Samas on GPT-4 puhul tugev korrelatsioon *relevance* ja *incongruence* vahel (ρ= 0.38-0.29, p < 0.001). C*ongruence* korreleerub dimensioonide üleselt, ning GPT-4 hinnangute puhul seosed säilivad. *Incongruence* puhul aga seos vastutusega GPT puhul statistilist olulisust ei saavuta.

"Vastutus"-e mõõdikute puhul näeme ootuspäraselt tugevas negatiivses korrelatsioonis *self_accountability* ja *other_accountability* (**ρ**= -0.40--0.57, p < 0.001), kuid ainult algandmete puhul *self_accountability impersonal_accountability*'ga ei seostu. Lisaks seostub *impersonal_accountability* GPT-4 hinnangutes *outlook_certainty'ga*, ning vastupidiselt negatiivselt *outlook_valence*'ga (ρ= 0.18 -> ρ= -0.31), kuigi see seos kaob täielikult lisades skaala), ning skaala puhul tekib negatiivne korrelatsioon võrreldes väikse varasema positiivsega seoses *situational_coping*'iga (ρ= 0.11 -> ρ= -0.37).

Ilmneb struktuurimuutus GPT-4 hinnagutes – K3 skaala lisades vahetusid korrelatsioonid. Ilma skaalata tugevalt korreleerunud olnud *incongruence* ja *situational coping*, ning *self-accountability* ja *outlook_valence* vahetavad omavahel seosed.

"Tagajärgede tõenäosus" on GPT-4'l suurem kahjulikes olukordades (*incongruence* ρ= -0.15 -> ρ= -0.38 -> ρ= -0.32), ilma "süüdlaseta" olukordades (impersonal_accountability ρ= -0.20), ning samuti seostab ta negatiivselt kogu "Toimetulekupotentsiaaliga", mille puhul inimhindajatel *situational_coping* korrelatsioon puudub.

Omaette klastri moodustavad Toimetulekut tõlgendavad dimensioonid *outlook_valence* ja *situational_coping, emotional _coping,* millel säilivad omavahelised seosed, kuid on puudutatud struktuurimuutusest.

Tabel 7.

Tõlgendusdimensioonide korrelatsioonide võrdlus

Mõõdik 1		Mõõdik 2	Alg	*p*	K1	*p*	K3	*p*	AlgvsK1	AlgvsK3	K1vsK3
relevance	-	***incongruence***			0.38	*0.00*	0.29	*0.00*	0.4	0.3	-0.1
relevance	-	***other_accountability***					-0.1	*0.05*		-0.1	-0.1
relevance	-	***outlook_certainty***	0.1	*0.04*	0.18	*0.00*	0.2	*0.00*	0.1	0.1	
relevance	-	***outlook_valence***					-0.12	*0.01*		-0.1	-0.1
relevance	-	***situational_coping***			-0.3	*0.00*			-0.3		0.3
relevance	-	***emotional_coping***			-0.16	*0.00*	-0.15	*0.00*	-0.2	-0.2	
congruence	-	***incongruence***	-0.18	*0.00*	-0.39	*0.00*	-0.4	*0.00*	-0.2	-0.2	
congruence	-	***self_accountability***	0.25	*0.00*	0.3	*0.00*	0.33	*0.00*		0.1	
congruence	-	***other_accountability***	-0.12	*0.02*	-0.16	*0.00*	-0.16	*0.00*			
congruence	-	***impersonal_account.***	0.11	*0.03*	-0.14	*0.00*	-0.14	*0.00*	-0.3	-0.2	
congruence	-	***outlook_valence***	0.3	*0.00*	0.22	*0.00*	0.28	*0.00*	-0.1		0.1
congruence	-	***situational_coping***	0.24	*0.00*	0.17	*0.00*	0.23	*0.00*	-0.1		0.1
congruence	-	***emotional_coping***	0.16	*0.00*	0.26	*0.00*	0.2	*0.00*	0.1		-0.1
incongruence	-	***other_accountability***	0.18	*0.00*					-0.2	-0.2	
incongruence	-	***impersonal_account.***	-0.2	*0.00*					0.2	0.2	
incongruence	-	***outlook_certainty***	0.15	*0.00*	0.38	*0.00*	0.32	*0.00*	0.2	0.2	-0.1
incongruence	-	***outlook_valence***	-0.12	*0.02*	-0.18	*0.00*	-0.39	*0.00*	-0.1	-0.3	-0.2
incongruence	-	***situational_coping***			-0.41	*0.00*	-0.17	*0.00*	-0.4	-0.2	0.2
incongruence	-	***emotional_coping***	-0.24	*0.00*	-0.36	*0.00*	-0.4	*0.00*	-0.1	-0.2	
self_accountability	-	***other_accountability***	-0.4	*0.00*	-0.57	*0.00*	-0.53	*0.00*	-0.2	-0.1	
self_accountability	-	***impersonal_account.***			-0.32	*0.00*	-0.3	*0.00*	-0.3	-0.3	
self_accountability	-	***outlook_valence***	0.13	*0.01*	0.51	*0.00*	0.12	*0.02*	0.4		-0.4
self_accountability	-	***situational_coping***	0.22	*0.00*	0.1	*0.04*	0.43	*0.00*	-0.1	0.2	0.3
self_accountability	-	***emotional_coping***			0.14	*0.00*	0.11	*0.03*	0.1	0.1	
other_accountability	-	***impersonal_account.***	-0.4	*0.00*	-0.4	*0.00*	-0.37	*0.00*			
other_accountability	-	***outlook_certainty***			-0.16	*0.00*	-0.1	*0.03*	-0.2	-0.1	0.1
other_accountability	-	***outlook_valence***	-0.19	*0.00*					0.2	0.2	
other_accountability	-	***situational_coping***	-0.22	*0.00*	-0.18	*0.00*				0.2	0.2
other_accountability	-	***emotional_coping***	-0.1	*0.05*	-0.11	*0.02*	-0.17	*0.00*		-0.1	-0.1
impersonal_account.	-	***outlook_certainty***			0.21	*0.00*	0.18	*0.00*	0.2	0.2	
impersonal_account.	-	***outlook_valence***	0.18	*0.00*	-0.37	*0.00*			-0.6	-0.2	0.4
impersonal_account.	-	***situational_coping***	0.11	*0.02*	0.18	*0.00*	-0.31	*0.00*	0.1	-0.4	-0.5
impersonal_account.	-	***emotional_coping***	0.15	*0.00*			0.19	*0.00*	-0.1		0.2
outlook_certainty	-	***outlook_valence***	-0.11	*0.02*	-0.14	*0.00*	-0.2	*0.00*		-0.1	-0.1
outlook_certainty	-	***situational_coping***			-0.27	*0.00*	-0.22	*0.00*	-0.3	-0.2	0.1
outlook_certainty	-	***emotional_coping***	-0.15	*0.00*	-0.16	*0.00*	-0.23	*0.00*		-0.1	-0.1
outlook_valence	-	***situational_coping***	0.4	*0.00*	0.28	*0.00*	0.51	*0.00*	-0.1	0.1	0.2
outlook_valence	-	***emotional_coping***	0.34	*0.00*	0.6	*0.00*	0.59	*0.00*	0.3	0.2	
situational_coping	-	***emotional_coping***	0.17	*0.00*	0.67	*0.00*	0.29	*0.00*	0.5	0.1	-0.4

Märkus. Tabelis on presenteeritud vaid statistiliselt olulisused korrelatsioonid (p<0.05). *Algandmed, K1 ka K3 kirjeldavad sisemisi korrelatsioone ja "vs" muutusi nende vahel.*

K1 tugev seos *situational_coping* ja *emotional_coping* (ρ= -0.67 -> ρ= -0.29) vahel asendub K3 puhul *outlook_valence ja situational_coping-iga* (ρ= -0.28 -> ρ= -0.51). Siinkohal aga on uued korrelatsiooni lähemal algsetele. "Toimetulekupotentsiaal"i dimensioonid seostuvad positiivselt *congruence* ja *self-accountability*'ga ning negatiivselt *incongruence* ja *outlook_certainty'*ga.

Arutelu

Lähenedes avastuslikult GPT-4-le ülesannete andmisele, leidsime erinevaid aspekte, mis keelemudelitega töös võivad tulemusi mõjutada. Chat-GPT reliaablus oli hea, ning täpsus oli ootuspärases, keskmises vahemikus, (nRMSE 0.27-0,41 vs nRMSE 0,29-0,50 (Ruder jt, 2025), kuid esines erinevad anomaaliaid, mis tekkisid tänu sellele, et tegemist ei olnud spetsiifiliselt testimiseks loodud andmestikuga, seega meil oli võimalik neid lähemalt uurida.

Kontekst

Juba andmestiku eelanalüüsil ilmnesid esimesed eripärad. Algse küsitluse ülesande püstitus seadis esimestele tõlgendusdimensioonidele tugevad raamid: osalejaid instrueeriti kirjeldama üht tugevat negatiivseid emotsioone tekitanud olukorda viimase 2 kuu jooksul.

GPT jaoks võib ülesande kontekst olla vihje, kuidas olukorda hinnata. Sündmus toimus hiljuti, puudutas spetsiifiliselt kirjeldanud isikut, mis asetab ta sündmuste keskele. GPT-de eesmärk on kaaluda tekstist välja õige vastus, inimhindajates on rohkem subjektiivsust, mis tekitab müra. Ühest küljest GPT-d valivad valentsi hindvate küsimuste puhul tugevamaid vastuseid (Yongsatianchot jt., 2023), ning inimvastajatel võib esineda juts Likert skaalade puhul reserveeritus kõige äärmuslikuma vastuse valimise ees (Weijters, Millet, & Cabooter, 2021). Jaotuvuste võrdlusel (vt joonis 2) näeme ka tugeva laeefektiga autorite *relevance* hinnanguid jaotumas keskmiseni, kuid mõõdiku RMSE/MAE viitab suurele erandite mõjule, ning ning *relevance* täpsus on väike uurimisküsimuste lõikes.

Yongsatianchot jt., leidsid, et GPT lahendused hindavad valentse kõrgemalt, kui inimhindajad, kuid *congruence* võrdluses näeme justkui vastupidist (joonis 2). GPT-4 hinnangud on jaotunud võrdselt 1 ja 2 peale. Kuid seda variatiivsust võivad seletada ka seosed *self_accountability* ja "Toimetulekupotensiaal"iga. Efekt ilmneb selgemalt aga *incongruence* puhul, kus on GPT vastused 8 ümber, ning ei korreleeru inimeste vastustega. Konteksti latenset mõju võib oletada ka *outlook_certainty ja incongruence* tugevast korrelatsioonist võrreldes inimhindjatega (tabel 7). Võib väita, et ülesande püstitus mõjutab tõlgendusi.

Küsimuste tähendusväli

"Kongruentsus" dimensioonide *congruence* ja *incongruence* seos on teoreetiliselt iseenesest mõistetav ja samuti on korrelatsioonid nende vahel suured ka GPT-4 hinnangutes, küll aga mitte inimeste puhul. Iga katsetingimusega süvenes GPT-4 ebatäpsus. Siinkohal võib

probleem olla aga küsimuses endas. *Congruence* väide on esitatud "*In this situation, I have a sense that this situation is potentially desirable for me*", ning *congruence* tugevaimad korrelatsioonid on *self_accountability, outlook_valence* ja *situational_coping*. *Incongruence* suurimad korrelatsioonid algandmetes on hoopis *other_accountability* ja *emotional_ coping'*ga. Kuid erinevalteelmisest, väide lõpeb "*is potentially harmful for me*", mis viitab haiget saamisele, mittelihtsalt ebasoovitud olukorras olemisele, ning on see palju kitsama tähendusega. Antud ebakõla väljendub ka jaotuses – algandmete hinnangud, kuigi enamjaolt üle 5, jaotuvad üle skaala – võrreldes väikese variatiivsusega *relevance* ja *congruence*'ga. Samuti on alguuringust hinnatud madala reliaablusega mõõdikuks (ICC=0,48).

Skaala olemus

Täiendasime Chat-GPT prompti autoritele esitatud Likert skaala kirjeldusega, [I3]: "*An integer weight of 1 meaning means to strongly disagree, 9 to strongly agree and 5 doing neither*.", mille tulemusena saavutas Chat-GPT täpsus korrelatsioonid varem kõige reliaabsemate, kuid kõige ebatäpsemate, "Toimetulekupotentsiaali" dimensioonidega. Pannes tähele kolme dimensiooni algset jaotuvust, võib märgata põhimõttelist erinevust muude skaaladega (vt joonis 2), eristuv kese, 5 või 6, mille ümber jaotuvad muud vastused. Toimetulekuga seotud tõlgendusi eristab teistest olemuslik vahe, need on teisesed tõlgendused (Lazarus, 1991). Ennustavad hinnangud situatsioonile, ning enda omadustele seda lahendada või aktsepteerida. Ennustuste suhtes on paljud inimesed erapooletud, ei oska nõustuda ega vastu olla. Uue K3 skaala kirjelduses esineb keskpunkt, mis muutes skaala polaarseks, annab võimaluse anda erapooletumaid hinnanguid. Nii inimestel, kui tekstides, ei pruugi olla piisavalt infot, et antud dimensioone hinnata. Peale skaala lisamist suurenes *_coping* dimensioonide variatiivsus keskmise suunas. Skaalade bipolaarsus mõjutab GPT täpsust.

Samal ajal esines ka struktuurseid muutuseid, mis võivad vajada katsetamist bipolaarsusega. K1 oletatavalt teoreetilised korrelatsioonid (kuna neid ei esine algandmetes), muutusid algandmetega vastavaks, ning üks korrelatsioon ootamatult tugevalt negatiivseks: seos enda vastutuse ja situatsiooniga hakkama saamise vahel. Võimalik, et mõni skaala on bipolaarne, või muu struktuuriga ja mõjutab GPT täpsust (Batyrshin jt., 2017).

Teksti pikkus

Analüüsides statistilise olulisuse saavutanud (vt joonis 4) või kõikide dimensioonide (vt lisa 1) korrelatsioonide graafikuid, võime märgata, et paljud sujuval veamäärade vähenemisele suunatud graafikud vahetavad suuna 451-500 või 501-550 grupi juures. See tähistab piiri, kus gruppi kuuluvate tekstide arv langeb alla 30, ning valimi vähesus võib mõjutada tulemusi, kuid võimalus on ka, et see on piir, kus autorid on andnud edasi peamise tõlgendustega seotud info, ning edasise info väärtus võib olla GPT jaoks liigne või vastukäiv, teksti mitmetitõlgendatavust suurendab, kui mitte lihtsalt väheväärtuslik. See murdepunkt

peegeldab Ruder jt 2025 tulemusi tekstiandmete peale, kus 400-500 tähemärgi juures saavutasid nii GPT-4, kui inimhindajad platoo. Samas on mõlema uuringi ühisosaks tekstide arvu drastiline langus tähemärkides alates 400-st. Platoo võib tulla ka sellest, et 400 tähemärki on loomulik limiit, kui pikalt sellise formaadi juures inimesed oma kogemusi kirjeldavad.

Teoreetilised seosed

Esimene näide korrelatiivse seose kohta kahe mõõdiku vahel, mis võib põhineda teoreetilisel seosel oli *congruence* ja *incongruence*, kuid nende seose analüüsi puhul on takistuseks vähene variatiivsus. Õnneks on taolisi näiteid veel, kus kahe GPT tingimuse vahel on mitmeid korrelatsioon, mis ei saa olla konteksti tõlgendus, kuna korrelatsioonid on tugevad, ning jaotuvus hea. Seega GPT võib olla teksti omaduste variatiivsuse põhjal seotud hinnangu teinud. Tooksin siinkohal välja kaks self_accountability ja other_accountability, mis on küll tugevas korrelatsioonis algandmetes, kuid muutub veel tugevamaks GPT-4 puhul *(*ρ= -0.4 -> ρ= -0.57), ning ebatäpsust tõstva näitena *self_accountability* ja *impersonal_ accountability,* millel algselt seos puudus, kuid GPT tõlgendustes püsib *(*ρ= -0.32; ρ= -0.30).

GPT heuristikute mõistmine aitaks tulevikus tõlgenduste kvaliteeti tõsta. “Vastutus”-el ja muudel hinnangutel, kus hindamise aluseks GPT jaoks olla agent või objekt, võib potentsiaalselt leida protsessi põhise tõlgenduste mudeli kõrvale täiendust OCC- mudelist, kus tõlgenduste aluseks on situatsiooni elemendid (Ortony et al., 1988).

Järeldused

Käesolev uuring erineb paljudest kaasaegsetest selle poolest, et see ei töötanud prepareeritud tekstikorpusel baseeruva mudeliga, ega ka selleks otstarbeks kogutud ja prepareeritud juhtumikirjeldustega (Hofmann jt., 2020, 2021; Troiano jt., 2023). Samuti erinevalt paljudest tõlgendustega seotud katsetest ei olnud tegemist dialoogiga, vaid olime piiratud üksikpäringutega (Yongsatianchot jt., 2023, Zhan et al., 2024). Tänu sellele oli võimalus avastada erinevaid piiranguid, mis võivad mõjutada LLM kasutamist protsessides, kus on kasu inimesest kirjelduse põhjal otsuste tegemisel või olukorra hindamise automatiseerimisest. Demonstreerisime, et käsitletud aspektid: konteksti kallutatus, tõlgendust hindava küsimuse tähendusruum, esitatud skaala, kirjelduste pikkus ja sümboolsed seosed on faktorid, mille mõju tuleb arvestada lahenduste disainimisel. Igas loetletud küljes on suund järgnevateks uuringuteks, kuidas antud külgi keelemudelite rakendamise puhul arendada.

Piirangud

Käesoleva uuringu peamisi piiranguid on detailse kvalitatiivse uuringu puudumine, mis polnud antud uuringu skoop. LLM sisemise loogika avastamine vajaks sisendiks dialoogi,

tõlgendusdimensioonide seoste diskursiivset analüüsi, et kontrollida jooksvalt erandeid ja nende tekkimise loogikaid.

Käesolevas uuringus polnud võrdlust inimhindajatega. LLM'ide peamine praktiline väljund on erinevatel aladel inimfaktori asendamine või täiendamine. Et rakenduslikkuse hinnang oleks võrreldav, tuleb lahenduse kvaliteeti hinnata võrdluses inimesega, kes sarnast ülesannet täidaks.

Katses tugevus, üldlevinud LLMi testimine, on ka selle nõrkus. Kasutasime üht GPT lahendust paljude seast, samade parameetritega. GPT lahenduse parameetrid tugevalt mõjutada efektiivsust, käesolevad järeldusedei pruugi kohanduda teistel mudelitel teiste parameetritega. Suurte keelemudelitel põhinevad teenused on ka hüppelises arengus. Eelmisel aastal kasutatud mudel on töö kirjutamise hetkeks juba "*legacy*" versioon.

Tänusõnad

Täname Denis Ruderit, kes toetas uuringu alustamist Pythoni programmeerimiskeskkondade kiirkursuse ning oma lähtekoodi jagamise läbi ja Kairit Sirtsi jooksvate nõuannete eest.

Kasutatud kirjandus

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, *183*(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

Batyrshin, I., Monroy, F., Gelbukh, A., Luis, A., Villa-Vargas, Solovyev, V., & Kubysheva, N. (2017). Bipolar Rating Scales: A Survey and Novel Correlation Measures Based on Nonlinear Bipolar Scoring Functions. Acta Polytechnica Hungarica, 14, 2017–2050. https://doi.org/10.12700/APH.14.3.2017.3.3

Braun, M., Greve, M., Kegel, F., Kolbe, L., & Beyer, P. E. (2024). *Can (A)I Have a Word with You? A Taxonomy on the Design Dimensions of AI Prompts*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2024.068

Dunkley, D. M., Lewkowski, M., Lee, I. A., Preacher, K. J., Zuroff, D. C., Berg, J.-L., Foley, J. E., Myhr, G., & Westreich, R. (2017). Daily Stress, Coping, and Negative and Positive Affect in Depression: Complex Trigger and Maintenance Patterns. *Behavior Therapy*, *48*(3), 349–365. https://doi.org/10.1016/j.beth.2016.06.001

Gross, J. J. (2015). Emotion Regulation: Current Status and Future Prospects. *Psychological Inquiry*, *26*(1), 1–26. https://doi.org/10.1080/1047840X.2014.940781

Hewing, M., & Leinhos, V. (2024). *The Prompt Canvas: A Literature-Based Practitioner Guide for Creating Effective Prompts in Large Language Models* (arXiv:2412.05127). arXiv. https://doi.org/10.48550/arXiv.2412.05127

Hofmann, J., Troiano, E., & Klinger, R. (2021). Emotion-Aware, Emotion-Agnostic, or Automatic: Corpus Creation Strategies to Obtain Cognitive Event Appraisal Annotations. O. De Clercq, A. Balahur, J. Sedoc, V. Barriere, S. Tafreshi, S. Buechel, & V. Hoste (Toim), *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (lk 160–170). Association for Computational Linguistics. https://aclanthology.org/2021.wassa-1.17/

Hofmann, J., Troiano, E., Sassenberg, K., & Klinger, R. (2020). *Appraisal Theories for Emotion Classification in Text* (arXiv:2003.14155). arXiv. http://arxiv.org/abs/2003.14155

Inzlicht, M., Cameron, C. D., D'Cruz, J., & Bloom, P. (2024). In praise of empathic AI. *Trends in Cognitive Sciences*, *28*(2), 89–91. https://doi.org/10.1016/j.tics.2023.12.003

Kim, E., & Klinger, R. (2019). *A Survey on Sentiment and Emotion Analysis for Computational Literary Studies*. http://dx.doi.org/10.17175/2019_008_v2

Kluyver, T., Ragan-Kelley, B., P&#233, Rez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., n, Abdalla, S., Willing, C., & Team, J. D. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (lk 87–90). IOS Press. https://doi.org/10.3233/978-1-61499-649-1-87

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lazarus, R. S. (1991). *Emotion and Adaptation*. Oxford University Press.

Lee, Y. K., Suh, J., Zhan, H., Li, J. J., & Ong, D. C. (2024). *Large Language Models Produce Responses Perceived to be Empathic* (arXiv:2403.18148). arXiv. https://doi.org/10.48550/arXiv.2403.18148

Mohammad, S. (2012). #Emotional Tweets. E. Agirre, J. Bos, M. Diab, S. Manandhar, Y. Marton, & D. Yuret (Toim), **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)* (lk 246–255). Association for Computational Linguistics. https://aclanthology.org/S12-1033/

Moors, A. (2009). Theories of emotion causation: A review. *Cognition and Emotion*, *23*(4), 625–662. https://doi.org/10.1080/02699930802645739

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal Theories of Emotion: State of the Art and Future Development. *Emotion Review*, *5*(2), 119–124. https://doi.org/10.1177/1754073912468165

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions* (lk xi, 207). Cambridge University Press. https://doi.org/10.1017/CBO9780511571299

Panksepp, J. (1998). *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University PressNew York, NY. https://doi.org/10.1093/oso/9780195096736.001.0001

Panksepp, J. (2012). In defense of multiple Core Affects. *Categorical versus dimensional models of affect: A seminar on the theories of Panksepp and Russell* (lk 31–78). John Benjamins Publishing Company. https://doi.org/10.1075/ceb.7.02pan

Pedersen, T. L. (2024). *patchwork: The Composer of Plots*. https://patchwork.data-imaginist.com

Perrez, M., & Reicherts, M. (1992). *Stress, coping, and health: A situation-behavior approach : theory, methods, applications*. Hogrefe & Huber. https://cir.nii.ac.jp/crid/1130282273020959104

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. http://www.rstudio.com/

Ruder, D., Uusberg, A., & Sirts, K. (2025). *Assessing the Reliability and Validity of GPT-4 in Annotating Emotion Appraisal Ratings* (arXiv:2503.16883). arXiv. https://doi.org/10.48550/arXiv.2503.16883

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Russell, J. A. (2012). From a psychological constructionist perspective. *Categorical versus dimensional models of affect: A seminar on the theories of Panksepp and Russell* (lk 79–118). John Benjamins Publishing Company. https://doi.org/10.1075/ceb.7.03rus

Salah, M., Abdelfattah, F., & Al Halbusi, H. (2024). The good, the bad, and the GPT: Reviewing the impact of generative artificial intelligence on psychology. *Current Opinion in Psychology*, *59*. https://doi.org/10.1016/j.copsyc.2024.101872

Santu, S. K. K., & Feng, D. (2023). *TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks* (arXiv:2305.11430). arXiv. https://doi.org/10.48550/arXiv.2305.11430

Scherer, K. R., Schorr, A., & Johnstone, T. (Toim). (2001). *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Sladek, M. R., Doane, L. D., Luecken, L. J., & Eisenberg, N. (2016). Perceived stress, coping, and cortisol reactivity in daily life: A study of adolescents during the first year of college. *Biological Psychology*, *117*, 8–15. https://doi.org/10.1016/j.biopsycho.2016.02.003

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, *48*(4), 813–838. https://doi.org/10.1037/0022-3514.48.4.813

Zhan, H., Ong, D., & Li, J. J. (2023). Evaluating Subjective Cognitive Appraisals of Emotions from Large Language Models. H. Bouamor, J. Pino, & K. Bali (Toim), *Findings of the Association for Computational Linguistics: EMNLP 2023* (lk 14418–14446). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.962

Tak, A. N., & Gratch, J. (2023). *Is GPT a Computational Model of Emotion? Detailed Analysis* (arXiv:2307.13779). arXiv. https://doi.org/10.48550/arXiv.2307.13779

Troiano, E., Oberländer, L., & Klinger, R. (2023). Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction. *Computational Linguistics*, *49*(1), 1–72. https://doi.org/10.1162/coli_a_00461

Troiano, E., Padó, S., & Klinger, R. (2019). *Crowdsourcing and Validating Event-focused Emotion Corpora for German and English* (arXiv:1905.13618). arXiv. http://arxiv.org/abs/1905.13618

Uusberg, A., Yih, J., Taxer, J. L., Christ, N. M., Toms, T., Uusberg, H., & Gross, J. J. (2023). Appraisal shifts during reappraisal. *Emotion*, *23*(7), 1985–2001. https://doi.org/10.1037/emo0001202

Weijters, B., Millet, K., & Cabooter, E. (2021). Extremity in horizontal and vertical Likert scale format responses. Some evidence on how visual distance between response categories influences extreme responding. International Journal of Research in Marketing, 38(1), 85–103. https://doi.org/10.1016/j.ijresmar.2020.04.002

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT* (arXiv:2302.11382). arXiv. https://doi.org/10.48550/arXiv.2302.11382

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., Brand, T. van den, Posit, & PBC. (2024). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (Version 3.5.1) [Software]. https://cloud.r-project.org/web/packages/ggplot2/index.html

Yeo, G., & Jaidka, K. (2023). The PEACE-Reviews dataset: Modeling Cognitive Appraisals in Emotion Text Analysis. H. Bouamor, J. Pino, & K. Bali (Toim), *Findings of the Association for Computational Linguistics: EMNLP 2023* (lk 2822–2840). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.186

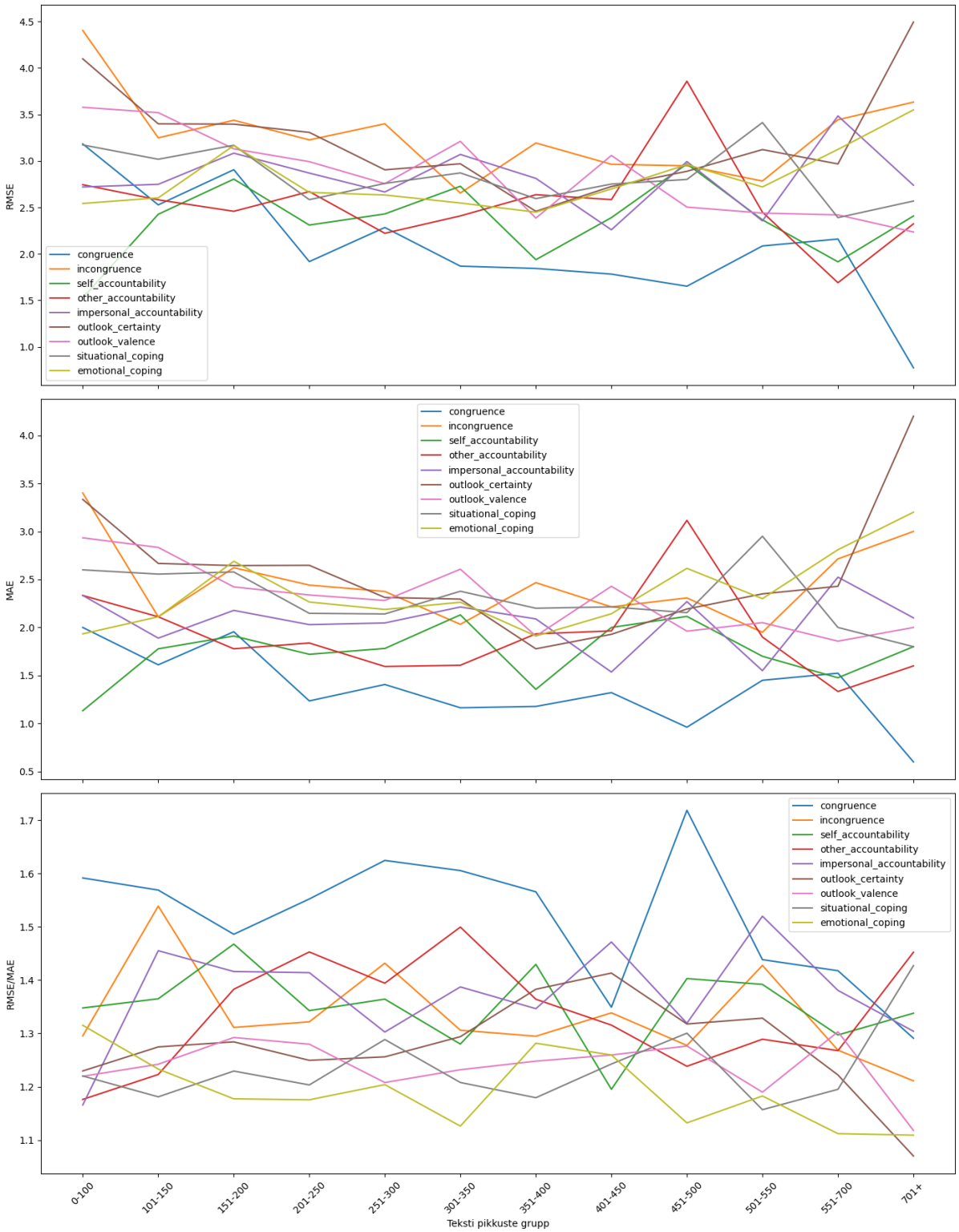
Yih, J., Kirby, L. D., Spitzer, E. G., & Smith, C. A. (2020). Emotion as a process: Appraisal, emotion, and coping patterns across time. *Motivation Science*, *6*(3), 221–234. https://doi.org/10.1037/mot0000144

Yongsatianchot, N., Torshizi, P. G., & Marsella, S. (2023). *Investigating Large Language Models' Perception of Emotion Using Appraisal Theory* (arXiv:2310.04450). arXiv. https://doi.org/10.48550/arXiv.2310.04450

Zhan, H., Zheng, A., Lee, Y. K., Suh, J., Li, J. J., & Ong, D. C. (2024). *Large Language Models are Capable of Offering Cognitive Reappraisal, if Guided* (arXiv:2404.01288). arXiv. https://doi.org/10.48550/arXiv.2404.01288

Lisa 1.

Tõlgendusdimensioonide gruppide reliaablusnäitajate muutus tekstide pikkuse lõikes.



Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kersten Kõrge

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose
   GPT-4 afektiivsete tõlgendusdimensioonide hindajana,
   mille juhendaja on Andero Uusberg,
reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

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

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

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

Kersten Kõrge
12.05.2025