TURUN
YLIOPISTO
Matemaattis-luonnontieteellinen
tiedekunta

ChatGPT:n hyödyntäminen arvioinnissa

Kokemuksia aineenopettajaopiskelijan näkökulmasta

Severi Hautala

Maantiede (aineenopettaja)

pro gradu -tutkielma

Laajuus: 20 op

30.5.2024

Turku

Pro gradu -tutkielma

Pääaine: Maantiede
Tekijä: Severi Hautala
Otsikko: ChatGPT:n hyödyntäminen arvioinnissa – Kokemuksia aineenopettajaopiskelijan näkökulmasta
Ohjaaja: Sanna Mäki
Sivumäärä: 51 sivua + liitteet 11 sivua
Päivämäärä: 30.5.2024

ChatGPT on herättänyt suurta kiinnostusta yhteiskunnan eri aloilla, niin kuin myös koulutuksessa. Tutkimustulokset osoittavat, että ChatGPT:tä voi hienosäätämisen ansiosta hyödyntää arvioinnissa luotettavana apuvälineenä. Tulosten käytännön sovellettavuutta kuitenkin heikentää se, että tekoälymallin hienosäätäminen vaatii syvällistä ymmärrystä ohjelmoinnista ja tekoälymallien toimintaperiaatteista, joita opettajalla ei välttämättä ole. Tämän tutkimuksen tarkoituksena on selvittää, voiko ChatGPT:tä ohjeistaa arviointiin sopivaksi käyttäjäystävällisesti räätälöityjen GPT-4-chattibottien luomistyökalun avulla, jossa tekoälymallin toiminnan ohjaus perustuu kehotemenetelmiin hienosäätämisen sijasta. Tutkimuksessa luotiin neljä eri kehotemenetelmien avulla ohjeistettua chattibottia, joiden arviointia verrattiin ihmisen suorittamaan arviointiin. Tarkastelemalla arvioinnin tuloksia yhdessä kehotemenetelmien kanssa luotiin käsitys ChatGPT:n soveltuvuudesta arviointiin.

Tutkimusaineisto koostui 96:sta sensorien arvioimasta ja pisteyttämästä vastauksesta syksyn 2023 maantieteen ylioppilaskokeen koetehtävään *2.1 Kuvaile kaikki kolme sadetyyppiä ja nimeä kullekin sadetyypille yksi ominainen esiintymisalue*. Vastaukset jaettiin kahteen ryhmään, joista yhtä käytettiin chattibottien arvioinnin kohdejoukkona ja toista chattibottien ohjeistuksessa. Lisäksi tutkimuksessa hyödynnettiin Ylioppilastutkintolautakunnan julkaisemaa *Hyvän vastauksen piirteet* -dokumenttia, jossa kuvataan sensorien käyttämät arviointiohjeet. Käsitys tarkemmasta arviointiohjeiden soveltamisesta luotiin teoriaohjassa sisällönanalyysissä tarkastelemalla yhdessä *Hyvän vastauksen piirteitä* ja sensorien antamia pisteitä.

Tutkimuksessa luotiin neljä chattibottia, jotka arvioivat jokaisen vastauksen 10 kertaa. Ensimmäisen botti ohjeistettiin *Hyvän vastauksen piirteillä nollakehotemenetelmää* hyödyntäen. Toisen botin ohjeistus perustui *ajatusketjukehotemenetelmään* sisällönanalyysin tuloksista johdetuilla tarkemmilla arviointiohjeilla. Kolmas botti ohjeistettiin samalla menetelmällä kuin botti kaksi, minkä lisäksi sille syötettiin oppimateriaalia kahdesta lukion maantieteen oppikirjasta, joissa kuvailtiin sadetyyppien syntytapoja. Neljännen botin ohjeistus perustui bottien 2 ja 3 menetelmien lisäksi *vähäisen ohjauksen kehottamiseen*, jossa botille näytettiin pisteittäin luokiteltuja esimerkkivastauksia. Chattibottien arvioinnin tuloksia vertailtiin tilastollisesti keskenään ja sensorien arvioinnin kanssa. Lisäksi bottien 1 ja 4 palautteelle tehtiin sisällönanalyysi, jonka avulla lisättiin ymmärrystä siitä, miten botit sovelsivat arviointiohjeita.

Tutkimustulokset osoittivat, että chattibottien arviointi poikkesi toisistaan samoin kuin sensorien arvioinnista. Keskimäärin chattibotit antoivat vastauksille enemmän pisteitä kuin sensorit. Arviointi oli yhdenmukaisinta botin 4 ja sensorien välillä ja poikkeavinta botin 1 osalta. Bottien 2, 3 ja 4 vertailussa ei havaittu merkittävää eroa arvioinnin yhdenmukaisuudessa. Sisäkorrelaatiokertoimen tulosten mukaan chattibottien arviointi oli johdonmukaista ja yhtenevää sensorien kanssa. Sisällönanalyysi ja chattibottien tarkkuusmittaukset kuitenkin paljastivat, että bottien arvioinnin validiteetti oli alhainen. Tutkimustuloksista voitiin päätellä, että paras tapa ChatGPT:n ohjeistamiseen oli *ajatusketjukehotemenetelmän* hyödyntäminen tarkemmilla arviointiohjeilla. Lisäksi tuloksia tarkastelemalla havaittiin ChatGPT:n toimintaperiaatteen asettamat haasteet arviointitehtävissä, joissa vaaditaan ihmisen kaltaista joustavaa ajattelua ja ymmärrystä. Tämä näkyi chattibottien arvioinnissa muun muassa arviointiohjeiden epäjohdonmukaisena noudattamisena ja väärien asioiden arvioimisena.

Avainsanat: ChatGPT, chattibotti, tekoälymalli, generatiivinen tekoäly, arviointi, validiteetti, reliabiliteetti, tekoälymalli, maantiede, tekoäly, kehotesuunnittelu

Master's thesis

Subject: Geography
Author: Severi Hautala
Title: Utilizing ChatGPT for assessment – experiences from a teacher student`s perspective
Supervisor: Sanna Mäki
Number of pages: 51 pages
Date: 30.5.2024

ChatGPT has gained a lot of popularity and interest in different fields of society including education. Research has shown that it can be a reliable tool in assessment when fine-tuned for this purpose. However, the technical knowledge required for fine-tuning the AI-model makes the results irrelevant for most teachers who might not have the coding skills or deep understanding of the AI principles. This study explores whether ChatGPT can be effectively utilized for assessment purposes using a custom GPT creation tool. This tool facilitates the instruction of the AI model using prompting methods, eliminating the need for coding skills. The study was conducted by creating four different custom GPT-chatbots with different prompting methods which performance in assessment was compared against human raters. The results of chatbots assessment were then compared with the human raters together with the prompting methods used to understand which methods work the best and if ChatGPT can be utilized successfully for assessment without fine-tuning.

The research data consistent of 96 human rated student answers for the autumn 2023 geography matriculation exam question *2.1 Describe all three precipitation types and name a typical area for each precipitation type*. The answers were divided into two groups which half were used for the chatbots assessment and the other half for the prompting methods. In addition, a document named "*Hyvän vastauksen piirteet*" was used which described the assessment guidelines used by the human raters. This document was then used in content analysis with the scores given by the human raters to understand how the assessment guidelines described were applied in practice. Four custom GPT-4-chatbots were created to assess the answers 10 times each. The first chatbot was instructed using a zero-shot-prompting with the *Hyvän vastauksen piirteet.* The second chatbot was instructed using chain-of-thought-prompting with the specified assessment guidelines created from the results of the content analysis. The third chatbot was instructed using the same methods as for model two in addition study material from two upper secondary high school textbooks describing the formation precipitation types. The fourth chatbot was instructed using few-show-prompting with human rated student answers in addition with the methods used in chatbots 2 and 3. The results of the AI-models assessment were then statistically compared with the human raters. Also, content analysis was made for the feedback given by the chatbots 1 and 4 to understand how ChatGPT applied the assessment guidelines in practice.

The results revealed that chatbots assessment different form each other and the human raters. Generally, chatbots tend to give answers more scores than human raters. The assessment results with the human rates were closest between the chatbot 4 and far off with chatbot 1. No significant differences were seen in the results between the chatbots 2, 3 and 4. Intra correlation coefficient results indicated high reliability in chatbots assessment and with the human raters. However, the results from the content analysis in addition with the accuracy measurements revealed that the validity of chatbots assessment was low. According to the results the best method used for to instruct ChatGPT for assessment was chain-of-thought-prompting with the specified assessment guidelines. Also, the study revealed some limitations utilizing ChatGPT for assessment due to its lack on context understanding with different types of answers.

Key words: ChatGPT, generative artificial intelligence, AI, chatbot, assessment, validity, reliability, geography, prompt engineering, AI-model,

Sisällysluettelo

1 Johdanto

Tekoälyä hyödyntävien sovellusten määrä on lisääntynyt yhteiskunnassa merkittävästi (Zhang ja Lu 2021; Triguero ym. 2024). Yksi uusimmista tekoälyn osa-alueista on generatiivinen eli luova tekoäly, joka mahdollistaa uuden ja omaperäisen sisällön tuottamisen suuresta datamäärästä (Adamopoulou ja Moussiades 2020; Bandi ym. 2023). Esimerkkejä generatiivisen tekoälyn sovellutuksista ovat erilaiset chattibotit, jotka ymmärtävät luonnollista kieltä. Yksi tunnetuimmista ja suosituimmista chattiboteista on yhdysvaltalaisen tekoälytutkimusyrityksen OpenAI:n kehittämä ChatGPT-tekoälyjärjestelmä, jossa käyttäjä voi kommunikoida GPT-tekoälymallin kanssa intuitiivisessa käyttöliittymässä omalla äidinkielellään, joko puheen tai tekstin välityksellä (Introducing ChatGPT 2024). ChatGPT:n räjähdysmäinen suosio on herättänyt kiinnostusta yhteiskunnan eri osa-alueilla, kuten myös koulutuksessa (Zhang ja Lu 2021; Sallam 2023). Aiemmat tutkimukset ovat nostaneet esille ChatGPT:n käytön tuomat haasteet ja mahdollisuudet koulutuksessa niin opetuksen kuin oppimisenkin näkökulmasta (Chen ym. 2020; Joshi ym. 2021; Rudolph ym. 2023). Tutkimukset osoittavat, että ChatGPT:n GPT-3.5-tekoälymallista on mahdollista muokata hienosäätämisen ja kehotesuunnittelun avulla oppimista sekä opetusta tukeva apuväline. Sitä on hyödynnetty onnistuneesti esimerkiksi opetusmateriaalin teettämisessä ja eriyttämisessä sekä henkilökohtaisten opetusagenttien luomisessa (Jauhiainen ja Guerra 2023; Lan ja Chen 2024; Latif ja Zhai 2024). Lisäksi tulokset sen käytettävyydestä tekstimuotoisten vastausten arvioinnissa ovat lupaavia.

Arviointi ja siitä johdettu palaute tukee oppimista (Hattie ja Timperley 2007). Yksi opettajan tärkeimmistä taidoista on arviointitaito. Arvioinnin laatua voidaan tarkastella reliabiliteetin ja validiteetin näkökulmasta (Black ym. 2010; Akib 2015). Reliabiliteetilla tarkoitetaan arvioinnin toistettavuutta ja johdonmukaisuutta. Validiteetti sen sijaan kuvastaa, arvioidaanko sitä, mitä kuuluukin arvioida. Oikeudenmukaisessa ja laadukkaassa arvioinnissa osaamisen mittaaminen on johdonmukaista, läpinäkyvää ja ennakoitavaa (Nieminen 2019; Luostarinen ja Ouakrim-Soivio 2019). Arvioinnin johdonmukaisuutta voivat heikentää esimerkiksi arvioitsijan inhimilliset tekijät, kuten käsitys arvioitavasta henkilöstä ja huolimattomuusvirheet. Arvioinnin läpinäkyvyyttä heikentävät muun muassa käytettävien arviointimenetelmien ja arvioinnin kohteiden jättäminen esittelemättä arvioitavalle. Ajanpuutteen vuoksi arvioinnista ei välttämättä ehditä muodostaa palautta, mikä auttaisi oppijaa ymmärtämään, mitä on arvioitu.

Generatiivisilla tekoälyllä voidaan mahdollisesti parantaa arvioinnin oikeudenmukaisuutta ja laatua vähentämällä inhimillisitä tekijöistä aiheutuvia riskejä arvioinnin johdonmukaisuudelle

(Joshi ym. 2021; Almasre 2024). Lisäksi opettajien työaikaa voitaisiin mahdollisesti kohdentaa tehokkaammin opetukseen ja oppijoiden kohtaamiseen, jos arviointi automatisoidaan tekoälyjärjestelmien avulla. Tutkimukset osoittavat, että GPT-3.5-mallin hienosäätämiseen avulla sitä voidaan hyödyntää arvioinnissa luotettavasti ja tehokkaasti (Latif ja Zhai 2024). Vaikka tutkimustulokset ovat lupaavia, tekoälymallin hienosäätäminen arviointiin vaatii merkittävää teknistä osaamista, kuten koodaustaitoja ja syvällistä ymmärrystä sen toimintaperiaatteista, joita opettajilla ei välttämättä ole. Tulosten käytännön sovellettavuutta heikentää se, että ne eivät anna tietoa siitä, miten opettaja voisi valjastaa ChatGPT-tekoälyjärjestelmän arviointiin intuitiivisesti ilman GPT-mallin hienosäätöä. Yksi järjestelmän uusimmista ominaisuuksista on käyttäjän mahdollisuus luoda henkilökohtaisia chattibotteja, mikä mahdollistaa ChatGPT:n ohjeistamisen eri käyttötarkoituksiin kehotemenetelmiä hyödyntämällä (Introducing GPTs 2024). Aikaisemmat tutkimukset osoittavat, että GPT-3.5 ja GPT-4 mallin suorituskykyä arvioinnissa on mahdollista parantaa merkittävästi kehotteiden avulla ilman hienosäätöä (Mizumoto & Eguchi 2023; Almasre 2024).

Tämän tutkimuksen tarkoituksena on selvittää ja lisätä ymmärrystä siitä, mitä arviointiin tarkoitettujen chattibottien luomisessa tulisi ottaa huomioon ja miten hyvin nämä toimivat ChatGPT:n ohjeistamisessa arviointiin sopivaksi. Tutkimuksessa vertailen neljän luomani chattibotin arviointien eroavaisuuksia ihmisten suorittamiin arviointeihin. Tarkastelemalla arvioinnin eroavaisuuksia ja chattibottien ohjeistuksessa käytettyjä kehotemenetelmiä luon käsityksen siitä, voiko opettaja luoda ChatGPT:stä arviointiin sopivan apuvälineen chattibottien luomistyökalun avulla. Tutkimuskysymykseni ovat:

1. Miten chattibottien suorittama arviointi eroaa sensorien arvioinnista?
2. Mistä erot arvioinnissa chattibottien ja sensorien välillä johtuvat?
3. Miten chattibotteja tulisi ohjeistaa arviointiin sopivaksi?

2 Tekoäly

2.1 Tekoälyn määrittely

Tekoäly (*artificial intelligence, AI*) voidaan ymmärtää ja määritellä eri tavoin (De Spiegeleire ym. 2017; Martinez 2019; Wang 2019). Termin määrittelystä ja ymmärtämisestä tekevät hankalaa muun muassa tekoälyteknologian jatkuva kehittyminen, eri näkökulmat tekoälyn tutkimuksessa ja älykkyyden laaja-alaisuus. Yleisesti tekoälyllä voidaan tarkoittaa koneiden kykyä jäljitellä ihmisen kaltaista älykkyyttä (Goertzel 2014; Kaplan 2021). Tästä näkökulmasta tekoäly voidaan jakaa heikkoon ja vahvaan tekoälyyn koneen älykkyyden mukaan. Heikolla tai kapealla tekoälyllä (*narrow artificial intelligence*) tarkoitetaan koneen älykkyyden rajoittuneisuutta tiettyihin tehtäviin. Vahvalla tai yleisellä tekoälyllä (*general artificial intelligence*) viitataan koneen kykyyn ratkaista useita älykkyyttä vaativia toimintoja.

Tieteenalana tekoälyllä tarkoitetaan tietojenkäsittelytieteen osa-aluetta, jossa tutkitaan ja kehitetään menetelmiä, joilla koneet voivat suorittaa älykkyyttä vaativia tehtäviä (Lappi ym. 2018; Helm ym. 2020). Tekoälytutkimuksen osa-alueita ovat klassinen tekoäly, koneoppiminen, syväoppiminen ja generatiivinen tekoäly. Klassisen tekoälyn tutkimus keskittyy sääntöpohjaisten tekoälyn kehittämiseen, jossa koneen päättelyprosessia ohjataan loogisten menetelmien avulla (Shu-Hsien Liao 2005). Koneoppimisessa pyritään luomaan algoritmeja, jonka avulla tekoäly oppii ja kehittyy datan sisällöstä ilman suoraa ihmisen ohjelmointia (Mahesh 2018). Syväoppimisella viitataan koneoppimisen erikoisalaan, jossa keskitytään neuroverkkopohjaisten algoritmien kehittämiseen, jotka kykenevät oppimaan suuresta datamäärästä (Janiesch ym. 2021). Generatiivisen tekoälyn tutkimus kuuluu syväoppimisen osa-alueeseen. Siinä luodaan ja ohjelmoidaan algoritmeja uuden ja omaperäisen sisällön, kuten musiikin, valokuvien ja tekstin tuottamiseen datan sisällöstä monimutkaisten neuroverkkojen avulla (Bandi ym. 2023).

Tekoälyjärjestelmällä tarkoitetaan tekoälyä hyödyntävää teknologista järjestelmää (Steimers ja Schneider 2022). Esimerkkejä tekoälyjärjestelmistä ovat muun muassa terveydenhuollossa käytettävät klassiseen tekoälyyn perustuvat asiantuntijärjestelmät (*expert system*), joiden avulla potilas voi tehdä diagnoosin omasta sairaudestaan vastaamalla koneen esittämiin kysymyksiin oireista (Tan ym. 2016). Suosittelujärjestelmien toimintaperiaatteissa hyödynnetään koneoppimista. Tällaisia järjestelmiä ovat esimerkiksi kohdennetussa mainonnassa ja sosiaali-

sen mediassa käytettävät suosittelujärjestelmät, jotka muokkaavat käyttäjälle näytettävää sisältöä käyttäjän toiminnan perusteella (Vall ja Widmer 2018). Syväoppimista käyttäviä järjestelmiä ovat automaattiohjaukseen kykenevät järjestelmät, kuten Teslan autopilotti ja kuvantunnistussovellukset, jotka pystyvät tulkitsemaan ja luokittelemaan kuvia (Du ym. 2016). Generatiivisen tekoälyn ehkä tunnetuimpia järjestelmiä ovat luonnollisen kielen prosessointiin erikoistuneet chattibotit, kuten Microsoftin Copilot ja OpenAI:n ChatGPT (Copilot for Microsoft 365 2024; Introducing ChatGPT 2024).

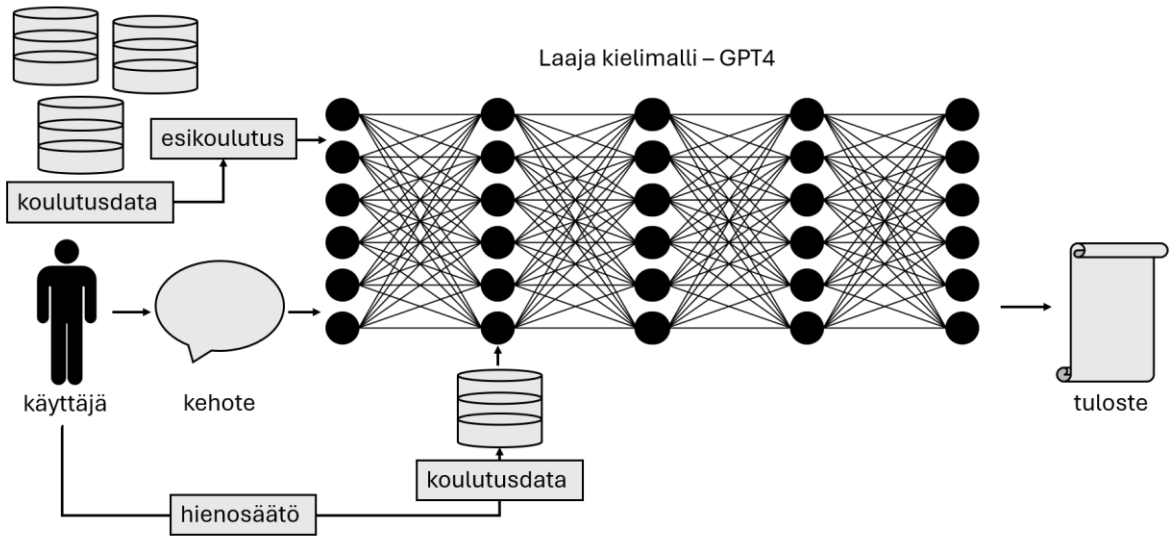
Tekoälymalli viittaa algoritmiin tai matemaattiseen malliin, jota käytetään tekoälyjärjestelmässä älykkyyttä vaativien tehtävien suorittamiseen (What is an AI model? 2024). Mallin suunnittelussa hyödynnetään tekoälyn eri osa-alueita mallin käyttötarkoituksen mukaisesti (taulukko 1). Klassiseen tekoälyyn perustuvien tekoälymallien toiminta perustuu logiikkaan ja sääntöpohjaisiin menetelmiin, minkä vuoksi ne sopivat päätöksentekoprosessiin erilaisissa asiantuntijajärjestelmissä (Tripathi 2011). Kohdennetussa mainonnassa käytettävät koneoppimismallit perustuvat päätöspuiden (*decision tree*) ja satunnaisten metsien (*random forest*) käyttöön ja pyrkivät ennakoimaan käyttäjän toimintaa (Vall ja Widmer 2018). Syväoppimista hyödyntävät tekoälymallit toimivat erilaisten neuroverkkojen, kuten konvoluutioneuroverkkojen avulla tietokonenäköä käyttävissä järjestelmissä ja sopivat erinomaisesti esimerkiksi kuvantunnistukseen (Du ym. 2016). Generatiivista tekoälyä hyödyntävien järjestelmien, kuten ChatGPT:n ja Copilotin tekoälymallit GPT-3.5 ja GPT-4, perustuvat *muuntaja*-arkkitehtuuriin (*transformer*) ja kykenevät ennustamaan sanoja käyttäjän syötteen perusteella (Bandi ym. 2023; Briganti 2024; How Microsoft 365 Copilot works 2024).

Taulukko 1. Tekoälytutkimuksen osa-alueet ja niitä hyödyntävät tekoälyjärjestelmät ja tekoälymallit (Tan ym. 2016; Vall ja Widmer 2018; Vall ja Widmer 2018; Janiesch ym. 2021; How Microsoft 365 Copilot works 2024; Introducing ChatGPT 2024).

Tekoälytutkimuksen osa-alue	Tekoälyjärjestelmä	Tekoälymalli
Klassinen tekoäly	**Asiantuntijajärjestelmät** terveydenhuollossa käytetyt sairauden diagnosointityökalut	**Sääntöpohjaiset mallit** lineaarinen regressio
Koneoppiminen	**Suosittelujärjestelmät** kohdennettu mainonta, sosiaalisen median suositukset	**Koneoppimisen mallit** päätöspuut (*decision tree*), satunnaiset metsät (*random forest*)
Syväoppiminen	**Automaattiohjaus, kuvantunnistus** Teslan autopilotti, kasvojentunnistussovellukset	**Tietokonenäkö** konvoluutioneuroverkot (CNN)
Generatiivinen tekoäly	**Chattibotit** Open AI ChatGPT Microsoft Copilot	**Muuntajat** GPT-3.5, GPT-4

2.2 ChatGPT:n toimintaperiaatteet, hienosäätö ja kehotesuunnittelu

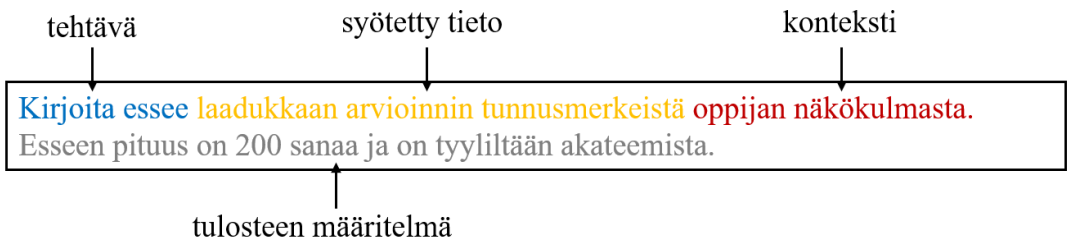
ChatGPT on yhdysvaltalaisen tekoälytutkimusyritys OpenAI:n kehittämä generatiivinen tekoälyjärjestelmä, jossa käyttäjä voi kommunikoida tekoälymallin kanssa chattipohjaisen käyttöliittymän avulla (Introducing ChatGPT 2024). ChatGPT:n toimintamalli perustuu *muuntaja*-tekoälymalliarkkitehtuuriin (*transformer*), jonka avulla tekoälymalli (GPT) kykenee ihmisen kaltaiseen viestintään luonnollisen kielen prosessoinnin avulla (*NLP, natural language processing*) (Briganti 2024). ChatGPT:n ilmainen versio käyttää GPT-3.5-tekoälymallia, kun taas maksulliset versiot tarjoavat pääsyn uudempaan ja edistyneempään GPT-4-malliin. GPT-4 on esikoulutettu laaja kielimalli (*LLM, large language model*), jonka edistynyt kyky tuottaa sujuvaa ja koherenttia tekstiä perustuu mallin esikoulutuksessa käytettyyn laajaan koulutusdatan määrään (kuva 1) (Aydin ja Karaarslan 2023; Wu ym. 2023; How ChatGPT and Our Language Models Are Developed | OpenAI Help Center 2024). Kun käyttäjä syöttää tekoälymallille *kehotteen* (*prompt*), kuten tekstin, malli pilkkoo lauseen yksittäisiksi sanoiksi ja kirjaimiksi. Tätä vaihetta kutsutaan *tokenisoinniksi (tokenization)*. Seuraavaksi malli syöttää tokenisoidun syötteen *muuntaja*-kerroksille, joissa se tulkitsee itsehuomiomekanismin avulla syötteen kontekstin ja luo ennusteen seuraavasta tokenista oppimansa koulutusdatan perusteella. Lopuksi malli muodostaa tokeneista ymmärrettävän tekstin ja palauttaa sen käyttäjälle *tulosteena* (*output*).



Kuva 1. GPT-4-tekoälymallin toimintaperiaate (Wu ym. 2023; Briganti 2024; How ChatGPT and Our Language Models Are Developed | OpenAI Help Center 2024).

GPT:n luomaan tulosteeseen vaikuttavat mallin esikoulutuksessa käytetyn koulutusdatan määrä ja laatu (Bandi ym. 2023). Mallin luoma ennuste seuraavasta tokenista perustuu aina tilastolliseen todennäköisyyteen. Tekoälymallin hallusinoinnilla (*hallucination)* tarkoitetaan tilannetta, jossa mallin teettämä tuloste on virheellinen (Tonmoy ym. 2024). Se ei välttämättä vastaa käyttäjän syöttämää kehotetta tai esittää valheellista tietoa oikeana. Hallusinoinnin riskiä voidaan vähentää esikoulutetun tekoälymallin hienosäätämisellä (*fine-tuning*) (Merchant ym. 2020; Church ym. 2021). Tässä tekoälymallin kykyä prosessoida ja tuottaa ennusteita pyritään muuttamaan säätämällä niin kutsuttuja parametreja (*parameters*) kouluttamalla mallia spesifillä koulutusdatalla. Parametreilla tarkoitetaan mallin tilastolliseen päättelyyn vaikuttavia muuttujia. Mallin koulukseen voidaan vaikuttaa muuttamalla hyperparametreja (*hyperparameters*), jotka ovat koulutusprosessin muuttujia, kuten oppimisnopeus tai tekstikoko. Hienosäätäminen on kustannustehokas tapa räätälöidä valmiiksi koulutettu tekoälymalli, kuten GPT-3.5 tai GPT-4 tarkasti määriteltyihin tehtäviin luomatta kokonaan uutta mallia. Se vaatii kuitenkin syvällistä ymmärrystä tekoälymallien toimintaperiaatteista, ohjelmointitaitoja ja runsaasti valmiiksi järjesteltyä laadukasta koulutusdataa. ChatGPT:n GPT-malleja on mahdollista hienosäätää OpenAI:n tarjoaman API-työkalun avulla (OpenAI Platform 2024).

Toinen lähestymistapa vaikuttaa tekoälymallin tuottaman tulosteen tarkkuuteen on kehotesuunnittelu (*promt engineering*), jossa muokataan käyttäjän syöttämää tekstimuotoista kehotetta (*promt*) (Giray 2023). Kehotteiden hiomisella voidaan vähentää tekoälymallin hallusinointia merkittävästi muokkaamatta mallin varsinaisia toimintaperiaatteita. Se tarjoaa käyttäjäystävällisen lähestymistavan ChatGPT:n suorituskyvyn parantamiseen ilman ohjelmointitaitoja toisin kuin hienosäätö, joka vaatii syvällisempää ymmärrystä mallin toimintatavoista. Toimivassa kehotteessa on neljä elementtiä, jotka ohjaavat tekoälymallin toimintaa (kuva 2). Elementit ovat tarkkaan määritelty tehtävä, konteksti, syötetty tieto ja tuloste (Promt engineering 2024; Elements of a prompt 2024).



Kuva 2. Esimerkki toimivan kehotteen keskeisistä elementeistä (Giray 2023; Promt engineering 2024; Elements of a promt 2024).

Tehtävä kuvastaa mallille mitä sen halutaan tekevän. Syötetty tieto sisältää tehtävän sisällön, jota konteksti täsmentää. Tulosteen määrittely auttaa mallia toivotun tekstin ominaisuuksien määrittämisessä, kuten tyylissä ja pituudessa. Kehotesuunnittelun menetelmiä ovat esimerkiksi *nollakehottaminen* (*zero-shot prompting*), jossa mallin toimintaa ohjataan yksittäisen kehotteen avulla (Wei ym. 2022). *Vähäisen ohjauksen kehottamisessa* (*few-shot prompting*) tekoälymallille annetaan esimerkkejä, joiden avulla malli luo tulosteen (Brown ym. 2020). *Ajatusketjukehottamisessa* (*chain-of-thought prompting*) tekoälymallia ohjeistetaan toimimaan ajatteluketjun kautta, jossa kehote koostuu yksittäisistä mallin ajattelua ohjaavista vaiheista (Wei ym. 2023). Tutkimusten mukaan erityisesti *vähäisen ohjauksen kehottaminen* ja *ajatusketjukehottaminen ovat tehokkaita* ChatGPT:n tekoälymallin ohjeistamisessa paremman tulosteen luomiseen verrattuna *nollakehottamiseen* (Min ym. 2022; Masikisiki ym. 2023). Hienosäädön ja kehotesuunnittelun toimintaperiaatteet eroavat toisistaan perustavanlaatuisesti, mutta molemmissa menetelmissä on omat heikkoudet ja vahvuudet, jotka rajoittavat niiden käyttöä sekä käyttötarkoitusta (taulukko 2).

Taulukko 2. Hienosäädön ja kehotesuunnittelun vertailu (Brown ym. 2020; Church ym. 2021; Wei ym. 2023; Liu ym.)

Menetelmän ominaisuudet	**Hienosäätö**	**Kehotesuunnittelu**
menettelytavat	spesifin koulutusdatan lisäys, *hyperparametrien* muuttaminen	*nollakehottaminen*, *ajatusketjukehottaminen*, *vähäisen ohjauksen kehottaminen*
toimintaperiaate	Mallia koulutetaan spesifillä koulutusdatalla.	syötteiden hiominen
vaikutus tekoälymalliin	Muuttaa mallia säätämällä *parametreja.*	Ei muuta mallia, vaan ohjeistaa sitä toimimaan paremmin.
vahvuudet	Mallin toimintaperiaatteita voidaan säätää tarkasti, joka parantaa sen suorituskykyä.	ei vaadi ymmärrystä ohjelmoinnista tai mallin toimintaperiaatteista, nopeaa, ei vaadi suuria määriä koulutusdataa
heikkoudet	vaatii syvällistä ymmärrystä ohjelmoinnista ja mallien toimintaperiaatteista, vie paljon aikaa, vaatii suuren määrän laadukasta koulutusdataa	Mallin toimintaperiaatteita ei muuteta, minkä vuoksi suorituskyvyn parantaminen ei ole yhtä tehokasta.

3 Arviointi ja palaute osana arviointia

3.1 Arvioinnin muodot ja tehtävät koulutuksessa

Arvioinnilla tarkoitetaan koulutuksessa toimenpiteitä, joilla kehitetään opetusta ja oppimista mittaamalla oppijan osaamista ja oppimisprosessia määrättyjen osaamistavoitteiden saavuttamiseksi (Luostarinen ja Ouakrim-Soivio 2019; Pohjonen ja Rissanen 2021). Arviointi on keskeinen työkalu opetuksen suunnittelussa, oppimisprosessin ohjaamisessa ja oppimistulosten selvittämisessä. Sen tehtävä on tukea ja ohjata opiskelijan oppimista sekä tehdä näkyväksi asetettujen oppimistavoitteiden saavuttaminen (Lukion opetussuunnitelman perusteet 2019). Arviointi voidaan jakaa diagnostiseen, formatiiviseen ja summatiiviseen arviointiin, jotka kuvastavat arvioinnin eri tehtäviä (kuva 3) (Chufama ja Sithole 2021).

Diagnostinen		Formatiivinen		Summatiivinen	
Opettaja	**Oppija**	**Opettaja**	**Oppija**	**Opettaja**	**Oppija**
Saa tietoa oppijan lähtötasosta, jonka avulla hän voi suunnitella opetusta.	Auttaa hahmottamaan oman osaamisensa suhteessa asetettuihin tavoitteisiin.	Havainnoi ja ohjaa oppimista antamalla kannustaa ja kehittävää palautetta.	Tunnistaa mitä oppimistavoitteiden saavuttamiseksi tulisi tehdä.	Määrittää hankitun osaamisen tason suhteessa osaamistavoitteisiin.	Auttaa hahmottamaan oman osaamisensa suhteessa asetettuihin tavoitteisiin.
Toiminnan perustana osaamistavoitteet					

Kuva 3. Arvioinnin muodot ja tehtävät opettajan sekä oppijan näkökulmasta mukaillen lähteestä Luostarinen ja Ouakrim-Soivio (2019).

Diagnostisen arvioinnin tarkoituksena on selvittää oppijan osaamistaso opetuksen alussa, jotta opetusta voidaan suunnitella optimaalisesti vastaamaan oppijan lähikehityksen vyöhykettä (Atjonen 2007; Chufama ja Sithole 2021). Diagnostisessa arvioinnissa yleisesti käytettyjä menetelmiä ovat lähtötasotestit ja opetuskeskustelut opetettavasta aiheesta, joiden perusteella opettaja muodostaa käsityksen oppijan osaamistasosta (Luostarinen ja Ouakrim-Soivio 2019). Arvioinnin tulosten esittäminen oppijalle edistää yhteisten oppimistavoitteiden laatimista, lisää arvioinnin läpinäkyvyyttä ja tukee oppimista (Pohjonen ja Rissanen 2021).

Formatiivinen arviointi on jatkuvaa ja laadullista. Sen tavoitteena on ohjata oppimisprosessia (Telle ym. 2015). Arviointimenetelmien tehtävänä on oppijan kannustaminen ja ohjaaminen oppimistavoitteiden saavuttamiseksi (Wiliam 2011). Arvioinnin tuloksia ei yleensä dokumentoida. Sen sijaan opettaja ja oppija käsittelevät vuorovaikutuksessa arvioinnin avulla esiin nousseita oppimista edistäviä tai heikentäviä tekijöitä. Itsearvioinnilla voidaan formatiivisessa

arvioinnissa kehittää oppijan kykyä ohjata omaa oppimistaan (Brown ym. 2015). Itsearvioinnissa oppija arvioi itse omaa oppimistaan suhteessa asetettuihin osaamistavoitteisiin ja oppii tunnistamaan omat kehityskohteensa (Telle ym. 2015). Vertaisarviointi formatiivisessa arvioinnissa tarkoittaa oppijoiden keskinäistä oppimisen arviointia, joka auttaa oppijoita hahmottamaan arvioinnin kohteita ja tätä kautta kehittämään oppimistaan (Black ja Wiliam 1998).

Summatiivinen arviointi on kokoavaa ja määrällistä (Dolin ym. 2018). Arvioinnin tulos dokumentoidaan ja ilmoitetaan usein numeerisena tai sanallisena arvosanana, joka kuvastaa oppijan saavuttamaa osaamista opetuksen alussa määrättyihin osaamistavoitteisiin nähden (Chufama ja Sithole 2021). Arviointikriteereillä tarkoitetaan linjauksia, joilla osaamista mitataan arvosanan muodostamiseksi (Pohjonen ja Rissanen 2021). Kriteerit johdetaan osaamistavoitteista ja ne kuvastavat, missä määrin tavoitteet on saavutettu (Telle ym. 2015). Summatiivisella arvioinnilla on suuri vaikutus oppijan minäkäsitykseen ja tulevaisuuden muodostumiseen (Atjonen 2007; Korkeakoulujen yhteishaun opiskelijavalinnat 2024).

Arviointi ja siitä johdettu laadukas palaute tukevat oppimista (Wiliam 2011; Leydon ym. 2014; Luostarinen ja Ouakrim-Soivio 2019). Arvioinnin tuloksista muodostettu palaute kuvaa oppijalle, mikä on hänen osaamistasonsa ja hänen tulisi tehdä halutun osaamistason saavuttamiseksi. Palautteen määrään ja laatuun vaikuttavat ennen kaikkea opettajan resurssit kuten käytettävissä oleva työaika ja ammattiaito (Hattie ja Timperley 2007; Henrik Nieminen 2019; Chufama ja Sithole 2021; Pohjonen ja Rissanen 2021). Palaute tekee arvioinnista läpinäkyvää, sillä se kuvastaa oppijalle myös sitä, mihin arvioinnissa on kiinnitetty huomiota. Palautteen tyyppi ja antamistapa vaikuttavat sen aiheuttamiin seurauksiin oppijalle. Laadukas palaute perustuu aina oppijan suoritukseen tai oppimisprosessiin, jota oppija voi kehittää. Diagnostisen arvioinnin tuloksena muodostettu palaute auttaa opettajaa suunnittelemaan opetustaan. Palautteen tulisi olla myös vastavuoroista ja palautteen saajalla tulisi aina olla oikeus kuulla, mihin annettu palaute perustuu. Oppimista ja opetusta edistävän palautteen tulisi aina olla kannustavaa ja kehittävää, mikä tukee oppijan oppimista sekä opetuksen suunnittelua.

3.2 Eettinen, oikeudenmukainen ja laadukas arviointi

Arvioinnin tulokset ovat usein arvioinnin näkyvin elementti. Oppijoiden lisäksi niistä ovat kiinnostuneet heidän huoltajansa, opettajansa, oppimistuloksia tarkastelevat organisaatiot sekä valtakunnalliset hallintoelimet, kuten opetusministeriö. Summatiivisen arvioinnin keskeisenä

tehtävänä yhteiskunnassa on osaamisen mittaaminen vertailtavuuden varmistamiseksi, mikä tekee arvioinnin tuloksista merkittäviä oppijan tulevaisuuden kannalta (Black ja Wiliam 1998; Atjonen 2007; Atjonen 2014). Arvioinnin oikeudenmukaisuutta, eettisyyttä ja laatua voidaan tutkia tarkastelemalla kokonaisvaltaisesti arviointiprosessia, sen kriteereitä, menetelmiä, ajankohtaa, kohdetta ja tarkoitusta. Osaamistavoitteiden määrittelyn tulisi ohjata arvioinnin kohteita, menetelmiä ja kriteereitä. Laadukkaan arvioinnin tunnusmerkkejä ovat läpinäkyvyys, johdonmukaisuus, perusteltavuus sekä monipuolisten menetelmien käyttö (Newton 2007; Luostarinen ja Ouakrim-Soivio 2019). Käytettävien arviointilinjausten julkilausuminen ja esittäminen oppijalle tekevät arvioinnista läpinäkyvämpää ja ennakoitavaa. Arvioinnin laatua voidaan tutkia reliabiliteetin ja validiteetin näkökulmista (Black ym. 2010; Akib 2015). Validiteetti kuvaa arvioinnin laatua arvioinnin kohteiden ja arviointimenetelmien yhteensopivuuden näkökulmasta. Sen tarkastelu auttaa ymmärtämään, mittaako arvioinnissa käytetty mittari tai menetelmä juuri sitä, mitä on tarkoitus mitata. Reliabiliteetilla kuvataan arvioinnin laatua toistamisen näkökulmasta. Toistamisella ei pitäisi olla vaikutusta arvioinnin lopputulokseen laadukkaassa arvioinnissa. Riskejä arvioinnin reliabiliteetille aiheuttavat arviointilinjausten tulkinnanvaraisuus ja arvioijan inhimilliset tekijät, kuten huolimattomuusvirheet sekä arvioijan käsitys oppijasta (Atjonen 2014; Ragupathi ja Lee 2020). Arvioinnin yhdenmukaistaminen lisää arvioinnin objektiivisuutta. Arviointia tekevien toimijoiden yhteistyöllä voidaan edistää arvioinnin yhdenmukaisuutta kehittämällä yhteistä arviointikulttuuria, jossa kriteerit, kohteet ja menetelmät ovat yksiselitteisesti tulkittavissa arvioitsijoiden kesken (Black ym. 2010). Arvioinnin yhdenmukaistaminen mahdollistaa arviointitulosten vertailun sekä arvioitavien tasavertaisen kohtelun, mikä lisää arvioinnin oikeudenmukaisuutta (Atjonen 2007; Atjonen 2014)

3.3 ChatGPT arvioinnissa

Generatiivista tekoälyä hyödyntävien tekoälyjärjestelmien kuten, ChatGPT:n, Geminin ja Copilotin vaikutukset oppimiseen ja opetukseen herättävät koulutuksen järjestäjissä vaihtelevia mielipiteitä (Joshi ym. 2021; Božić ja Indrasen Poola 2023; Imran ja Almusharraf 2023). Tutkijat uskovat, että helppokäyttöiset luonnollista kieltä ymmärtävät ja tuottavat Chattibotit voivat tulevaisuudessa muuttaa koulutusta merkittävästi, niin opetuksen kuin oppimisenkin näkökulmasta. ChatGPT:n käyttö arvioinnissa voi mahdollisesti keventää opettajan työtaakkaa,

mikä mahdollistaa resurssien kuten ajankäytön, paremman kohdentamisen opetukseen ja oppilaiden kohtaamiseen (Chiu 2024). Lisäksi se voi mahdollisesti parantaa arvioinnin johdonmukaisuutta vähentämällä arvioitsijasta johtuvien inhimillisten virheiden vaikutusta arvioinnin lopputulokseen. Aiempien tutkimusten mukaan ChatGPT:tä voidaan hyödyntää arvioinnissa sen hienosäädön ja kehotesuunnittelun ansiosta (Mizumoto & Eguchi 2023; Rudolph ym. 2023; Latif & Zhai 2024; Alsmare 2024). ChatGPT:n perustana olevan GPT-3.5-tekoälymallin suorituskyky esseiden arvioinnissa on osoittautunut erittäin hyväksi arvioinnin reliabiliteetin näkökulmasta. Sen vahvuuksiin kuuluvat arviointiprosessin nopeus ja skaalautuvuus. Mallin validiteetin tarkastelussa on kuitenkin havaittu, että sen kyky arvioida samoja asioita kuin ihminen, vaihtelee. Malli ei välttämättä tunnista vastausten kontekstuaalisia vivahteita tai syvempiä merkityksiä, joiden ymmärtäminen on välttämätöntä oppilaan ajattelutavan ja oppimisprosessin arvioimisessa. Tekoälymallin esikoulutuksessa käytetyn koulutusdatan laadulla on merkittävä vaikutus siihen, miten malli tulkitsee arvioitavia vastauksia. Tämä saattaa tehdä arvioinnista epäoikeudenmukaisempaa, jos tekoälymalli arvioi vastauksia vain tietynlaisten vastaustyylien perusteella kiinnittämättä huomiota sisältöön mitattavan osaamisen kannalta. ChatGPT:n käyttöön arvioinnissa liittyy myös riskejä (Baidoo-Anu ja Ansah 2023; Božić ja Indrasen Poola 2023; Michel-Villarreal ym. 2023). Tekoälymallille syötetyt tiedot, kuten oppilaiden nimet ja opintosuoritukset saattavat päätyä mallin koulutusdataksi, mikäli järjestelmän tietosuojaselosteeseen ei perehdytä huolellisesti. Lisäksi ChatGPT:n algoritmin monimutkainen toimintamalli asettaa haasteita arviointiprosessin läpinäkyvyydelle ja perusteltavuudelle. Arviointiprosessin automatisointi tekoälyjärjestelmien avulla voi myös heikentää oppimista, jos järjestelmän luoma palaute on laadultaan pintapuolista. Tutkijat korostavat, että ChatGPT:n ja GPT-tekoälymallien eettisen käytön arvioinnissa tulisi perustua niiden käyttöä ohjaavien linjausten julkilausumiseen, missä huomioidaan niiden riskit ja mahdollisuudet oppimisen ja opetuksen näkökulmasta (Su & Yang 2023).

3.4 Maantieteen ylioppilaskoe ja kokeen arviointi

Ylioppilastutkintolautakunnan laatimat ylioppilaskokeet mittaavat lukion opetussuunnitelmassa määrättyjen osaamistavoitteiden saavuttamista oppiainekohtaisesti (Yleiset määräykset ja ohjeet 2024; Laki ylioppilastutkinnosta 502/2019; Valtioneuvoston asetus ylioppilastutkinnosta 612/2019). Ylioppilaskokeiden koetehtävät laaditaan oppiaineiden valtakunnallisten pakollisten ja syventävien opintojen osaamistavoitteiden mukaan. Ylioppilastutkintoa suorittava

kokelas voi osallistua valitsemansa oppiaineen ylioppilaskokeeseen, kun hän on suorittanut kyseisen oppianeen valtakunnalliset pakolliset opinnot. Kokeissa on myös laaja-alaista osaamista mittaavia, oppiainerajoja ylittäviä tehtäviä, joihin vastaaminen ei edellytä toisen oppiaineen yksityiskohtaisia tietojen ja taitojen hallitsemista. Kokeet pidetään kaksi kertaa vuodessa keväisin ja syksyisin. Lukiot vastaavat ylioppilaskokeiden järjestämisestä ylioppilastutkintolautakunnan määräysten mukaisesti. Ylioppilaskokeiden arviointiprosessi koostuu kokeiden alustavasta ja lopullisesta arvioinnista (Yleiset määräykset ja ohjeet 2024; Laki ylioppilastutkinnosta 502/2019; Valtioneuvoston asetus ylioppilastutkinnosta 612/2019). Ylioppilaskokeen järjestäneen lukion aineenopettaja suorittaa kyseisen oppiaineen ylioppilaskokeille alustavan arvioinnin Ylioppilastutkintolautakunnan julkaisemien alustavien arviointikriteerien mukaisesti. Alustavan arvioinnin valmistuttua kokeiden lopullisesta arvostelusta vastaavat Ylioppilaslautakunnan alaisuudessa toimivat sensorit. Arvioinnissa käytettävistä linjauksista sovitaan oppiaineen sensorikokouksissa, joissa sensorit päättävät yhteisistä arviointikriteereistä ja tehtävien pisteyttämisestä. Ylioppilastutkintolautakunta julkaisee arvioinnin valmistuttua kunkin oppiaineen kokeen lopullisessa arvioinnissa käytetyt arviointiohjeet *Hyvän vastauksen piirteinä*. On kuitenkin huomattava, että *Hyvän vastauksen piirteet* antavat yleisen kuvan arvioinnissa käytetyistä linjauksista, eivätkä sisällä tarkempia pisteytykseen ja arviointiin sovellettuja määräyksiä tai kaikkia hyväksyttyjä vastauksia (Hyvän vastauksen piirteet – Maantiede syksy 2023; Luukka ym. 2023). Reaaliaineiden kokeiden arvioinnissa kiinnitetään erityisesti huomiota vastauksen sisältöihin, jotka voivat alentaa tai nostaa suorituksen arvoa (Reaaliaineiden kokeiden määräykset ja ohjeet 2024) (taulukko 3).

Taulukko 3. Reaalinaineiden kokeiden arviointiin vaikuttavia tekijöitä (Reaaliaineiden kokeiden määräykset ja ohjeet 2024).

Suorituksen arvoa alentavat esimerkiksi:	**Suorituksen arvoa nostavat esimerkiksi:**
• selvät asiavirheet • perustuminen mielipiteiden varaan • samojen asioiden toisto • lain ja hyvän tavan vasteiset lausumat • epäselvästi ja epätarkasti ilmaistut ajatukset • epäolennaiset tiedot tehtävässä, jotka osoittavat, että tehtävänanto on ymmärretty väärin tai niiden esittäminen ei liity tehtävän tavoitteisiin	• työkalujen tarkoituksenmukainen käyttö • asiasisältöjen johdonmukainen jäsentely • tiedonkäsittelytaitojen monipuolinen osoittaminen • tietojen ja taitojen itsenäinen hallinta ja kyky niiden soveltamiseen, mikä ilmenee vastauksesta • esitetyt väitteet perusteltu selkeästi • aineistojen tarkoituksenmukainen käyttö • esitettyjen tietojen kytkeminen laajempaan asiayhteyteen • faktojen, mielipiteiden ja perusteltujen kannanottojen selkeä erottaminen • syiden ja seurauksien monipuolinen ja asianmukainen tarkastelu • olennaisia tietoja tehtävän kannalta riittävästi

Kokelaalla on oikeus tutustua arvosteltuun koesuoritukseensa ja saada tietoa siitä, miten arvosteluperusteita on sovellettu hänen vastauksensa arvioinnissa. (Yleiset määräykset ja ohjeet 2024; Laki ylioppilastutkinnosta 502/2019). Jos kokelas katsoo arvioinnin virheelliseksi, hänellä on mahdollisuus vaatia oikaisua. Tällöin valitun koesuorituksen arvioi uudelleen kaksi sensoria, jotka käsittelevät ja ratkaisevat oikaisupyynnön. Koesuoritukset ovat salassa pidettäviä viranomaisten asiakirjoja, ja niiden tarkasteluun ovat oikeutettuja ainoastaan määrätyt henkilöt, kuten kokelas itse ja arvioinnista vastannut aineenopettaja ja ylioppilastutkintolautakunnan sensorit (Laki viranomaisten toiminnan julkisuudesta 621/1999). Lautakunta voi kuitenkin luovuttaa koesuorituksia tieteellisiin tutkimustarkoituksiin tutkimuslupahakemuksia vastaan, jolloin niiden anonymisoinnista ja tietoturvallisuudesta vastaa tutkimusluvan haltija (Yleiset määräykset ja ohjeet 2024).

Maantieteen ylioppilaskoe kuuluu reaaliaineiden kokeisiin, joiden rakenteessa noudatetaan kaikkia reaaliaineita koskevia koemääräyksiä (Hyvän vastauksen piirteet – Maantiede syksy 2023; Reaaliaineiden kokeiden määräykset ja ohjeet 2024; Yleiset määräykset ja ohjeet 2024). Maantieteen ylioppilaskokeessa mitataan maantieteellistä ajattelua, maantieteellisten käsitteiden, paikannimistön ja ilmiöiden monipuolista hallintaa sekä syy- ja seuraussuhteiden ymmärtämistä laajojen aihepiirien parissa. Tehtäviin vastaaminen voi edellyttää kokelaalta myös laskemista, analyysiä, piirtämistä sekä diagrammien ja kaavioiden laatimista. Useat tehtävät pohjautuvat aineistoihin, joiden tulkinta ja muokkaus on edellytys tehtävään vastaamiselle. Kokeessa on yhdeksän tehtävää, joista kokelas vastaa viiteen (Maantieteen ylioppilaskoe, syksy 2023). Koe koostuu kolmesta osasta, joiden tehtävätyypit vaihtelevat mitattavan osaamisen laajuuden ja haastavuuden mukaan. Kokeen ensimmäisessä, kaikille pakollisessa osassa mitataan maantieteellisten faktatiedon, käsitteiden ja paikannimistön hallintaa monivalintatehtävän muodossa. Kokeen toinen osa sisältää neljä tehtävää, joista kokelas vastaa kahteen. Tehtävät edellyttävät kokelaalta maantieteellisten ilmiöiden luonnehdintaa, prosessien kuvausta, käsitteiden määrittelyä ja geomedia-aineiston tulkintaa. Kolmannessa kokeen osassa on myös neljä tehtävää, joista kokelas vastaa kahteen. Tehtävät mittaavat kokelaan maantieteellistä ajattelua monipuolisesti ja edellyttävät syvällistä ymmärrystä maantieteellisten ilmiöiden alueellisista syy- ja seuraussuhteista sekä oppiainerajat ylittävää laaja-alaista osaamista. Tehtävien vastaustyypit vaihtelevat lyhyistä tekstivastauksista jäsenneltyihin esseevastauksiin.

4 Aineistot ja menetelmät

4.1 Aineistona maantieteen ylioppilaskokeen vastaukset ja *Hyvän vastauksen piirteet*

Tutkimusta varten Ylioppilastutkintolautakuntaan lähetettiin marraskuussa 2023 tutkimuslupahakemus, joka sisälsi tutkimussuunnitelman ja kuvauksen toivotusta tutkimusaineistosta. Ylioppilastutkintolautakunta myönsi tutkimusluvan 7.12.2023 (päätöskoodi OPH-6154-2023) ja toimitti tutkimusaineiston helmikuussa 2024. Tutkimusaineisto koostui syksyn 2023 maantieteen ylioppilaskokeen kaikista suomen- ja ruotsinkielisistä tekstimuotoisista vastauksista koetehtäviin 2–9 ja sisälsi tiedot kokelaiden tehtäväkohtaisista pisteistä, kokeen kokonaispistemäärästä ja arvosanasta. Aineisto ei sisältänyt henkilötietoja, jotka yhdistäisivät tiedot yksittäisiin kokelaisiin.

Tutkimuskäyttöön toimitetusta syksyn maantieteen 2023 ylioppilaskokeen vastauksista valittiin kaikki suomenkieliset vastaukset koetehtävään 2.1 *Kuvaile kaikki kolme sadetyyppiä ja nimeä kullekin sadetyypille yksi ominainen esiintymisalue* (n=1018). Tehtävässä mitattiin kokelaan maantieteellistä osaamista eri sadetyyppien syntytavan ymmärtämisestä, esimerkkialueen maininnasta ja nimeämisestä (Hyvän vastauksen piirteet – Maantiede syksy 2023; Maantieteen ylioppilaskoe, syksy 2023). Vastaukset järjestettiin taulukkolaskentaohjelmassa kokonaispistemäärän mukaiseen järjestykseen pienimmästä suurimpaan, jonka jälkeen perusjoukosta poimittiin satunnaisesti 96 vastausta (6 vastausta jokaista pistemäärää kohden väliltä 0–15 p) tarkastelemalla sensorien antamia pisteitä (taulukko 4).

Taulukko 4. Vastausten lukumäärä perusjoukossa pistemäärän mukaan ja valittu otoskoko.

Pisteet	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**Yht.**
Lukumäärä	41	18	22	20	25	40	69	74	86	88	113	93	105	98	62	64	**1018**
Otos	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	**96**

Tutkimuksen aineiston valintaan ja otantaan vaikuttivat tutkimuksen tekijän käytettävissä oleva työaika ja tutkimusasetelma. Tehtävän 2 koevastaukset valittiin tutkimuskäyttöön, koska ne eivät sisältäneet valokuvia ja koostuivat melko lyhyistä vastauksista, minkä oletettiin sopivan chattibottien arviointiin parhaiten. Lisäksi valitun ylioppilaskokeen tehtävän vastauksien arviointi- ja pisteytysohjeet vaikuttivat tutkimuksen kannalta riittävän selkeiltä tutkimuksen toteuttamisen kannalta. Tutkimuksen otantaan (n=96) päädyttiin arvioimalla tutkimuksen

toteuttamiseen tarvittavaa työmäärää ja käytettäviä tutkimusmenetelmiä. Tarkasteltavasta perusjoukosta valittiin jokaista pistemäärää kohden 6 vastausta, koska tämä mahdollisti chattibottien arvioinnin monipuolisen tutkimisen eri tasoisten vastausten kohdalla kohtuullisella työmäärällä.

Maantieteen ylioppilaskokeen *Hyvän vastauksen piirteet* on Ylioppilastutkintolautakunnan julkaisema asiakirja, joka on vapaasti luettavissa (Hyvän vastauksen piirteet – Maantiede syksy 2023). Asiakirja sisältää kuvauksen maantieteen ylioppilaskokeen lopullisessa arvioinnissa käytetyistä arvioinnin lähtökohdista sekä tehtäväkohtaisista arviointi- ja pisteytysohjeista. *Hyvän vastauksen piirteet* antavat yleisen kuvan arvioinnissa käytetyistä linjauksista, mutta siitä ei kuitenkaan käy suoraan ilmi, miten arviointiohjeita on sovellettu kunkin kokelaan vastauksen arvioinnissa. Lisäksi se ei sisällä tietoa kaikista hyväksytyistä vastauksista ja tarkemmista arviointiohjeista. Tässä tutkimuksessa *Hyvän vastauksen piirteitä* käytettiin valitun koetehtävän vastausten arviointikohteiden tunnistamisessa ja ohjeiden soveltamisessa yhdessä sensorien antamien pisteiden kanssa. *Hyvän vastauksen piirteissä* annetaan pisteytys- ja arviointiohjeet koetehtävälle 2.1 *Kuvaile kaikki kolme sadetyyppiä ja nimeä kullekin sadetyypille yksi ominainen esiintymisalue* (taulukko 5).

Taulukko 5. Pisteytys- ja arviointiohjeen tehtävälle 2.1 syksyn maantieteen 2023 *Hyvän vastauksen piirteiden* mukaan (Hyvän vastauksen piirteet – Maantiede syksy 2023).

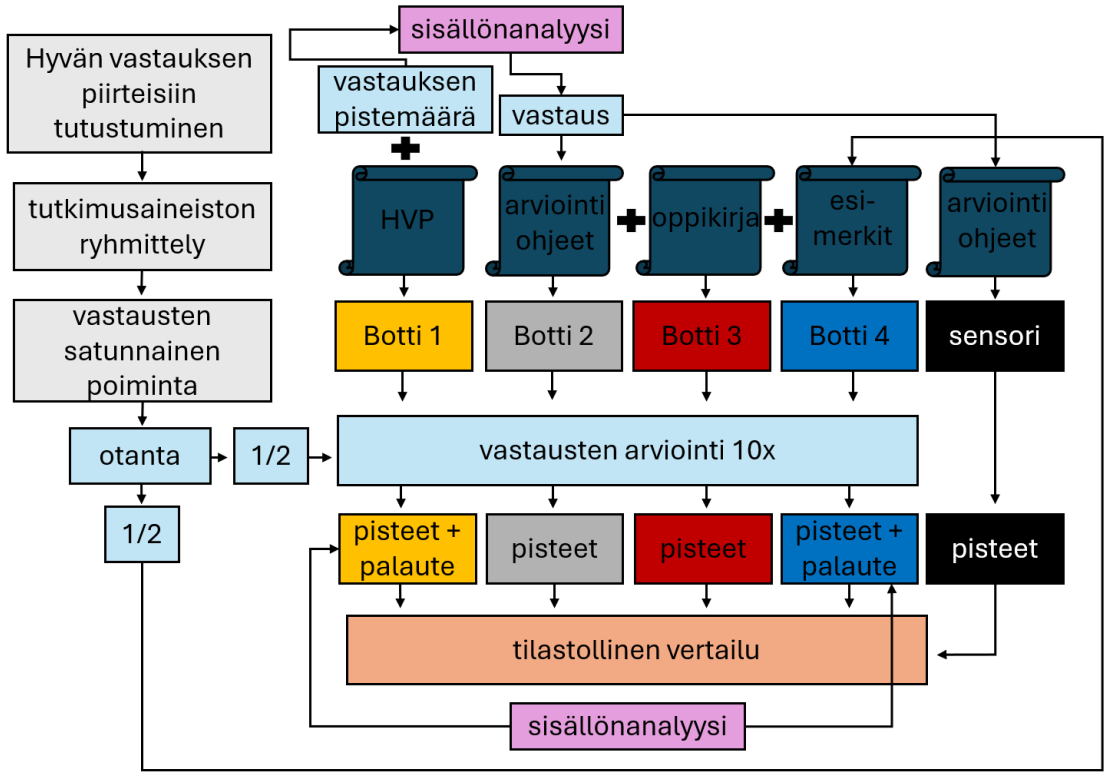
Arvioinnin lähtökohdat	Pisteytys	Mitä arvioidaan?
Vastauksen kokonaislaadun tarkastelu, jossa kiinnitetään huomiota käsitteiden oikeaoppiseen ja johdonmukaiseen käyttöön, esitettyjen asioiden oikeellisuuteen, suomenkielisen asiatekstin kirjoittamiseen sekä käsitteiden ja paikannimistön hallintaan.	Nimeäminen 1 p	Konvektiosateen, vuoristosateen ja rintamasateen nimeämistä.
	Syntytavan kuvaus 3 p	Ymmärrystä konvektiosateen, vuoristosateen ja rintamasateen syntytavasta.
	Esimerkkialue 1 p	Ymmärrystä tai muistamista konvektiosateen, vuoristosateen ja rintamasateen tyypillisistä esiintymisalueista.

Arviointiohjeista ei käy ilmi tarkempaa kuvausta esimerkiksi siitä, mitkä ovat hyväksytyt esiintymisalueet eri sadetyypeille ja millä perusteella syntytavan kuvauksesta saa 0, 1, 2 tai 3 pistettä. Ohjeissa ei myöskään mainita, miten sadetyypin nimeämisestä annetaan pisteitä, jos vastauksessa on kirjoitusvirheitä.

4.2 Menetelmät

4.2.1 Tutkimuksen vaiheet

Tutkimuksessa hyödynnettiin monimenetelmällistä lähestymistapaa, joka sisälsi sekä kvalitatiivisia että kvantitatiivisia tutkimusmenetelmiä (kuva 4).



Kuva 4. Vuokaavio tutkimuksen vaiheista ja käytetyistä menetelmistä.

Tutkimus aloitettiin tutustumalla *Hyvän vastauksen piirteisiin*, minkä perusteella tehtiin valinta tutkimuskäyttöön sopivista koetehtävien vastauksista. Tämän jälkeen valitun koetehtävän vastaukset järjestettiin taulukkolaskentaohjelmassa kokonaispistemäärän mukaiseen suuruusjärjestykseen pienimmästä suurimpaan. Vastausten perusjoukosta (n=1018) valittiin satunnaisesti jokaista pistemäärä kohden (0–15 p) 6 vastausta (yhteensä n=96) tarkastelemalla sensorin antamia pisteitä, joista muodostettiin tutkittavat pisteluokat. Seuravaksi vastaukset jaettiin kahteen ryhmään, joista kummassakin oli 48 vastausta, kolme vastausta jokaisesta pisteluokasta. Yhtä ryhmää käytettiin chattibotin 4 ohjeistamisessa ja toista chattibottien arvioinnin kohdejoukkona.

Tämän jälkeen kehitettiin neljä erilaista chattibottia, joilla testattiin kohdejoukon vastausten arviointia. Jokainen botti arvioi vastaukset kymmenen kertaa arvioinnin toistettavuuden analysoimiseksi. Arvioinnin tuloksena syntynyt pistemäärä kirjattiin ylös taulukkolaskentaohjelmaan myöhempää tarkastelua varten.

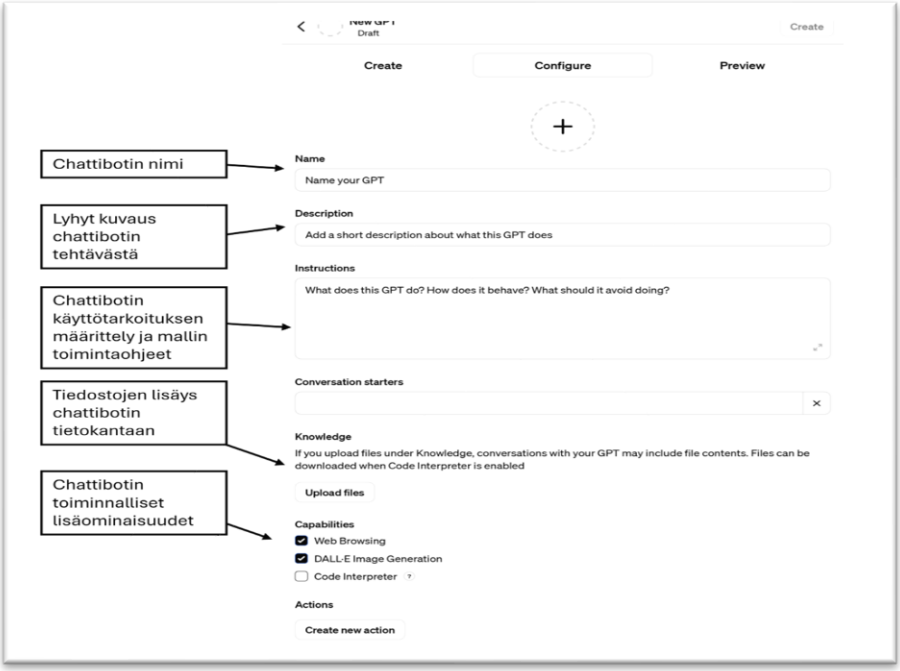
Arviointiprosessi aloitettiin chattibotilla 1, joka oli ohjeistettu pelkästään *Hyvän vastauksen piirteiden* avulla. Bottia 1 pyydettiin myös antamaan palautetta arvioinnin perusteista myöhempää analyysia varten. Arviointi jatkui botilla 2, joka oli ohjeistettu käyttäen tarkempia arviointiohjeita. *Hyvän vastauksen piirteiden* ja sensoreiden arviointien pohjalta suoritettiin teorialähtöinen sisällönanalyysi jokaiselle arviointijoukon vastaukselle, mikä johti tarkempien arviointiohjeiden kehittämiseen. Kolmas arviointikierros suoritettiin botilla 3, joka oli ohjeistettu käyttäen tarkkoja arviointiohjeita ja lisämateriaalia kahdesta lukion oppikirjan kappaleista, jotka käsittelevät sateiden syntyä. Viimeinen arviointi tehtiin chattibotilla 4, jonka ohjeistusmateriaaliin sisältyivät arviointiohjeet, oppikirjamateriaalit sekä pisteluokittain jaotellut koevastaukset. Bottia 4 pyydettiin myös antamaan palautetta arvioinnista.

Chattibottien arvioinnin tuloksia vertailtiin keskenään ja sensorien antamien pisteiden kanssa erojen havaitsemiseksi. Erojen merkitsevyyttä tutkittiin tilastollisin menetelmin. Bottien suorituskykyä arvioinnissa tutkittiin laskemalla jokaiselle botille tarkkuus sekä sisäkorrelaatiokerroin. Chattibottien keskinäistä ja sensorien välistä yhdenmukaisuutta arvioitiin myös sisäkorrelaatiokertoimen avulla. Bottien 1 ja 4 palautteelle suoritettiin teoriaohjaava sisällönanalyysi, jolla selvitettiin, kuinka hyvin mallit noudattivat arviointiohjeita vastausten arvioinnissa.

Lopuksi tilastollisen vertailun ja sisällönanalyysien tulokset yhdistettiin chattibottien ohjeistusmenetelmien tarkasteluun tutkimuskysymyksiin vastaamiseksi. Tilastollisen vertailun ja sisällönanalyysin tuloksilla vastattiin tutkimuskysymykseen 1. Sisällönanalyysien tulokset mahdollistivat tutkimuskysymykseen 2 vastaamisen. Tutkimuskysymykseen 3 vastattiin tarkastelemalla bottien ohjeistuksessa käytettyjä menetelmiä yhdessä tilastollisen vertailun ja sisällönanalyysien tulosten kanssa.

4.2.2 Chattibottien luominen ja ohjeistaminen arviointiin sopivaksi

Tutkimuksessa käytettiin ChatGPT:n Team-versiota, jolla varmistuttiin siitä, että tekoälymallille syötetyt tiedot pysyivät salassa eikä niitä tallennettu OpenAI:n palvelimille tai hyödynnetty mallien kouluttamiseen (Enterprise privacy 2024; Introducing ChatGPT Team 2024). Tutkimuksessa käytetyt chattibotit luotiin hyödyntämällä ChatGPT:n ominaisuutta, joka mahdollistaa tarkkojen ohjeiden mukaan toimivien räätälöityjen GPT-4-chattibottien (*custom GPT*) luomisen käyttäjäystävällisesti ilman koodaustaitoja (kuva 5)(Introducing GPTs 2024). Työkalu perustuu kehotesuunnittelumenetelmiin, jotka mahdollistavat tekoälymallin ohjeistamisen räätälöidyillä ohjeilla. Chattibotin luominen aloitettiin antamalla sille tunnistettava nimi *Nimi*-kenttään (*name*). *Kuvaus*-kenttään (*descripiton*) laadittiin kuvaus botin tehtävästä. *Toimintaohjeet*-kenttään (*instructions*) kirjoitettiin selkeät ja tiiviit ohjeet chattibotin toiminnalle, roolille ja tehtävälle. *Tiedostojen lisäys* -valikosta *(upload files*) botille syötettiin toimintaa ohjaavaa materiaalia, kuten tarkemmat arviointiohjeet ja esimerkkivastaukset pdf-muodossa sekä oppikirjojen materiaalit jpg-kuvatiedostoina. *Toiminnalliset lisäominaisuudet* -kohdasta (*capabilities*) valittiin asetukset, joilla säädeltiin botin kykyä verkkoselaukseen, DALL-E kuvangenerointiin ja kooditulkkiin. Valmis chattibotti tallennettiin muodossa, joka näkyi vain tutkimuksen tekijälle ja poistettiin tutkimuksen päätyttyä, millä varmistettiin palveluun syötettyjen tietojen pysyvä hävitys.



Kuva 5. Asetusvalikko räätälöityjen GPT-4-chattibottien luomiseksi (Creating a GPT | OpenAI Help Center 2024).

Tutkimuksessa luotiin neljä GPT-4-chattibottia, joiden ohjeistamiseen käytettiin eri menetelmiä (liite 1; liite 2; liite 3). Bottien toimintaa ohjattiin aiempien tutkimusten perusteella tunnistettujen tehokkaiden kehotesuunnittelumenetelmien avulla, joiden suorituskykyä verrattiin keskenään. Ensimmäisen chattibotin ohjeistamisessa hyödynnettiin *nollakehote*-menetelmää (Wei ym. 2022). Tässä botille syötettiin *Hyvän vastauksen piirteet* pdf-tiedostona. Botin toimintaa ohjeistettiin myös lisäämällä *toimintaohjeet*-kenttään kuvaus botin roolista, tehtävästä, tulosteen muodosta ja toiminnoista, joita sen tulisi välttää. Lisäksi kenttään kirjoitettiin chattibotin ja käyttäjän välisestä esimerkkikeskustelusta. Bottia ohjeistettiin myös antamaan arvioinnin tuloksena palautetta myöhempää analyysiä varten.

Toisen Chattibotin ohjeistus perustui *ajatusketjukehotemenetelmän* hyödyntämiseen (Wei ym. 2023). Botille annettiin tarkemmat ohjeet konvektiosateen, rintamasateen ja vuoristosateen pisteyttämiseksi arviointimatriisin muodossa. Näiden lisäksi botille syötettiin pdf-dokumentti, joka sisälsi täsmentäviä ohjeita arvioinnin suorittamisesta vaihe vaiheelta. *Toimintaohjeet*-kenttään kirjoitettu kuvaus vastasi edellisen chattibotin sisältöä lukuun ottamatta kohtaa, jossa bottia pyydettiin ilmoittamaan arvioinnin tuloksena vain vastauksen kokonaispistemäärä arviointiprosessin nopeuttamiseksi.

Kolmannen chattibotin ohjeistus noudatti samaa menetelmää kuin botissa kaksi, mutta tämän lisäksi botille lisättiin oppimateriaalia kahden lukion maantieteen oppikirjasta (*Geos 2 – Sininen planeetta* ja *Manner 2 – Sininen planeetta*) jpg-valokuvien muodossa, missä kuvailtiin sadetyyppien syntyä. Tällä pyrittiin selvittämään, auttaako oppimateriaalin lisääminen bottia arvioimaan vastauksia paremmin. Bottia pyydettiin tutustumaan oppimateriaaleihin ennen arviointia, jotta se ymmärtäisi sadetyyppien syntytapoja.

Neljännen chattibotin ohjeistus perustui *vähäisen ohjauksen kehotemenetelmään* (Brown ym. 2020). Chattibotille syötettiin pdf-dokumentti, joka sisälsi 48 kappaletta sensorien arvioimia ja pisteyttämiä koevastauksia. Vastaukset oli luokiteltu pisteittäin niin, että ne muodostivat luokitellun datasetin, jossa jokaisesta pisteluokasta oli edustettuna 3 kappaletta vastuksia. Lisäksi botin neljä ohjeistamisessa hyödynnettiin samoja menetelmiä kuin boteissa kaksi ja kolme, eli oppimateriaalia ja tarkempia arviointiohjeita.

4.2.3 Tarkempien arviointiohjeiden muodostaminen *Hyvän vastauksen piireistä*

Chattibottien arvioinnissa käytetyt vastaukset analysoitiin teorialähtöisen sisällönanalyysin avulla. Analyysin tavoitteena oli ymmärtää, kuinka arvioijat sovelsivat *Hyvän vastauksen piirteitä* arviointiprosessissaan. Tämän pohjalta luotiin tarkemmat arviointiohjeet, joita käytettiin bottien 2, 3 ja 4 ohjeistamisessa. Teorialähtöisessä sisällönanalyysissä analyysi suoritetaan aiemmin määritellyn teorian mukaan (Tuomi ja Sarajärvi 2018). Analyysi aloitettiin tuomalla *Hyvän vastauksen piirteet* -dokumentti ja kokelaiden vastaukset pdf-muodossa *Nvivo*-ohjelmaan (versio 1.7.2). Tämän jälkeen *Hyvän vastauksen piirteitä* tarkasteltiin ja luotiin koodit, jotka jaottelivat yleiset arviointiohjeet tarkemmiksi osa-alueiksi, kuten sadetyypin nimeäminen, esimerkkialueen maininta, syntytavan kuvaus ja tehtävän kannalta epäolennaisen tiedon identifiointi (taulukko 6). Kullekin arvioinnin osa-alueelle luotiin oma koodinsa, joka edusti mahdollista pistemäärää yhden pisteen tarkkuudella. Koodit ryhmiteltiin alaluokiksi, jolloin kullakin sadetyypillä oli omat koodinsa. Sadetyypit luokiteltiin edelleen pääluokiksi arvioijan mukaan. Pääluokkia olivat sensori, chattibotti 1 ja chattibotti 4. Sensorien arviointia analysoitiin vertailemalla vastauksen kokonaispistemäärää sen sisältöön, joka koodattiin vastaavaan pisteluokkaan arvioinnin osa-aluetta edustavasti. Bottien 1 ja 4 arviointia tarkasteltiin analysoimalla mallien antamaa palautetta arvioinnista. Palautteesta koodattiin esimerkkejä jokaisen sadetyypin arvioinnin osa-alueelle.

Taulukko 6. Sisällönanalyysissä käytetyt koodit, alaluokat ja pääluokka.

Koodit	**Alaluokka**	**Pääluokka**
Nimeäminen 0 p	Sadetyyppi	Arvioinnin tekijä
Nimeäminen 1 p		
Esimerkkialue 0 p		
Esimerkkialue 1 p		
Syntytapa 0 p		
Syntytapa 1 p		
Syntytapa 2 p		
Syntytapa 3 p		
Vähentää pisteitä	Tehtävän kannalta epäolennainen tieto	
Ei vaikutusta pisteytykseen		
Lisää pisteitä		

4.2.4 Arvioinnin laadun mittaamisessa käytetyt menetelmät

Chattibottien arvioinnin tilastollinen tarkastelu aloitettiin laskemalla jokaiselle vastaukselle pistemääräinen keskiarvo bottien kymmenen arviointikierroksen tuloksista *Microsoft Excel* -taulukkolaskentaohjelmassa (Free Online Spreadsheet Software: Excel | Microsoft 365 2024). Tämän jälkeen keskiarvot pyöristettiin kokonaisluvuiksi, jotta niitä voitiin verrata sensorien antamiin pisteisiin, jotka olivat myös kokonaislukuja.

Seuraavaksi chattibottien ja sensorien arviointituloksia analysoitiin tilastollisesti tutkimalla arvioitsijoiden pisteaineiston normaalijakautuneisuutta Shapiro-Wilkin testillä (Wilk Test - an overview | ScienceDirect Topics 2024). Testin perusteella päätettiin tutkimuksessa käytettävistä testausmenetelmistä: parametrisistä tai ei-parametrisistä. Kaikki tutkimuksen tilastolliset testit suoritettiin käyttäen IBM SPSS Statistics -tilastolaskentaohjelmaa (versio 29.0.0.0)(IBM SPSS Statistics 2024).

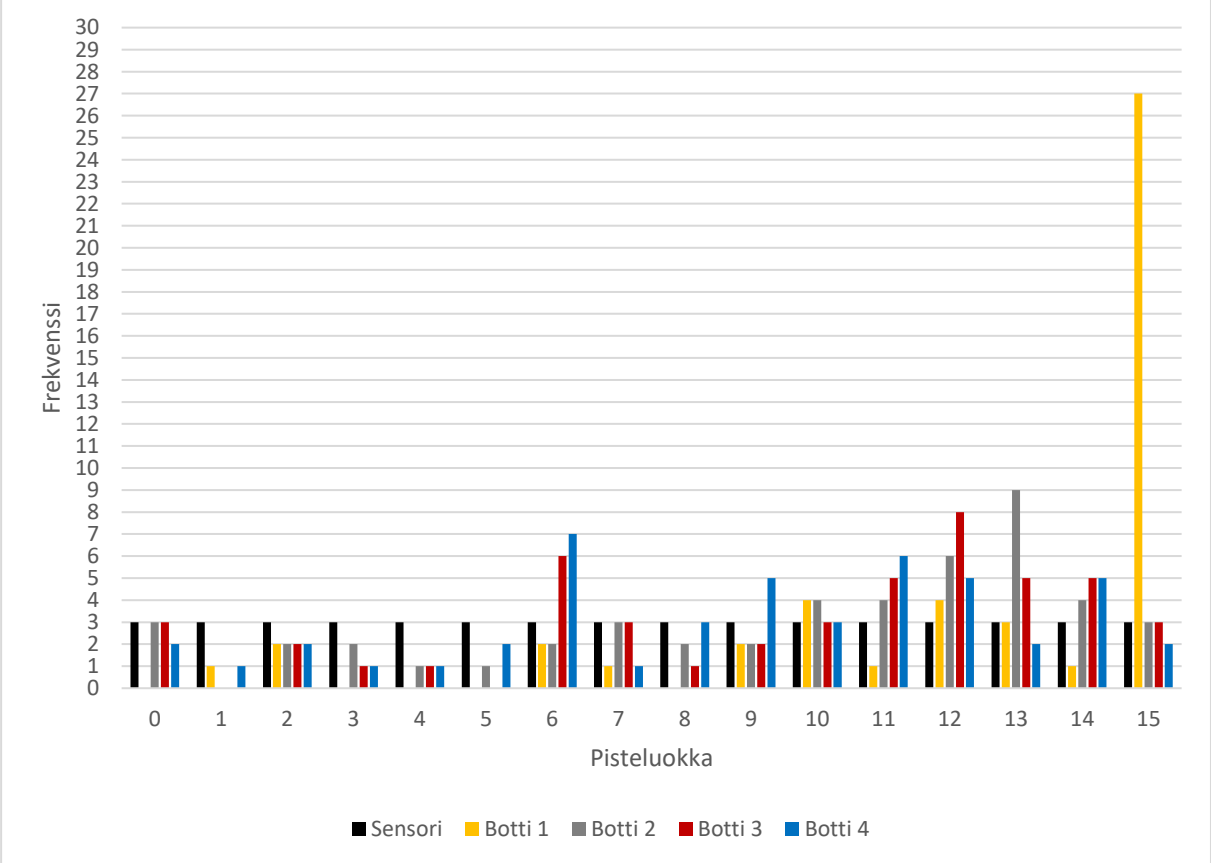
Chattibottien arvioinnin sisäistä reliabiliteettia tutkittiin laskemalla jokaiselle botille sisäkorrelaatiokerroin 10 arviointikierroksen tuloksista (Müller ja Büttner 1994). Tällä pyrittiin selvittämään, kuinka johdonmukaista arviointi oli kokonaisuudessaan ja yhden vastauksen osalta. Lisäksi bottien arvioinnin johdonmukaisuutta tarkasteltiin validiteetin näkökulmasta laskemalla tarkkuus jokaiselle pisteluokalle (Classification: Accuracy | Machine Learning 2024).

Chattibottien ja sensorien välisen arvioinnin yhdenmukaisuutta tutkittiin ensin vertailemalla arvioinnin erojen merkitsevyyttä. Kaikkien pisteaineistojen erojen merkitsevyys laskettiin ensin Friedmannin useiden riippuvien otosten testillä (Pereira ym. 2015). Tämän jälkeen kahden pisteaineiston välisen erojen merkitsevyyttä tarkasteltiin Wilcoxin kahden riippuvan otoksen testillä (Tähtinen ym. 2020:137–139). Testien valintaa perusteltiin sillä, että pisteaineistot eivät noudattaneet normaalijakaumaa ja mitattavat otokset olivat toisistaan riippuvia. Seuraavaksi arvioitsijoiden välistä yhdenmukaisuutta tarkasteltiin laskemalla arvioitsijoiden välille sisäkorrelaatiokerroin (Müller ja Büttner 1994). Tämä testin käyttö oli perusteltua arvioitsijoiden välisen reliabiliteetin tutkimiseen, koska arviointiaineisto koostui jatkuvista muuttujista. Viimeiseksi bottien arvioinnista laskettiin keskiarvo jokaiselle pisteluokalle, jota verrattiin sensorien antamaan pistemäärään.

5 Tulokset

5.1 Vastausten jakautuminen pisteluokittain

Chattibottien arvioinnin tuloksena kokelaiden vastaukset jakautuivat pisteluokkiin epätasaisesti (kuva 6). Sensorien arviot jakautuivat tasaisesti jokaiseen pisteluokkaan, kun taas chattibotit arvioivat useimmat vastaukset suurimpiin pisteluokkiin. Alhaisiin ja keskitason pisteluokkiin chattibotit luokittelivat keskimäärin vähemmän vastauksia kuin sensorit. Chattibottien arvioinnin yhdenmukaisuus vaihteli joidenkin pisteluokkien välillä merkittävästi, kun taas joissakin se vaihteli vain vähän. Suurin vaihtelu bottien välillä havaittiin suurissa pisteluokissa, kun taas pienin vaihtelu havaittiin pienissä pisteluokissa. Botti 1 poikkesi muista malleista merkittävästi luokittelemalla lähes kaikki vastaukset lähelle suurinta pisteluokkaa, josta luokka 15 erottui selkeänä poikkeamana suurimmalla vastausten lukumäärällä. Botit 3 ja 4 luokittelivat vastauksia hyvin yhdenmukaisesti lähes kaikissa pisteluokissa, ja botti 2 poikkesi niistä vain vähän.



Kuva 6. Kokelaiden vastausten jakautuminen pisteluokittain arvioinnin tuloksena.

Tilastollisesta testauksesta havaittiin, että chattibottien ja sensorien arvioinnin tuloksena vastausten pistemääräinen jakautuminen ei noudattanut normaalijakautuneisuutta (taulukko 7). Botin 4 ja sensorin Shapiro-Wilk-tilastot olivat lähimpänä toisiaan sekä merkitsevää p-arvon rajaa, mikä viittasi siihen, että vastausten pistejakaumat olivat samankaltaisia näiden niiden välillä.

Taulukko 7. Tilastollisen testin tulokset vastausten normaalijakautuneisuudesta pisteellisesti.

Arvioija	Shapiro-Wilk-tilasto	*P*-arvo	Normaalijakautuneisuus
Sensori	0,948	0,033	Ei
Botti 1	0,708	0,000	Ei
Botti 2	0,891	0,000	Ei
Botti 3	0,905	0,001	Ei
Botti 4	0,951	0,045	Ei

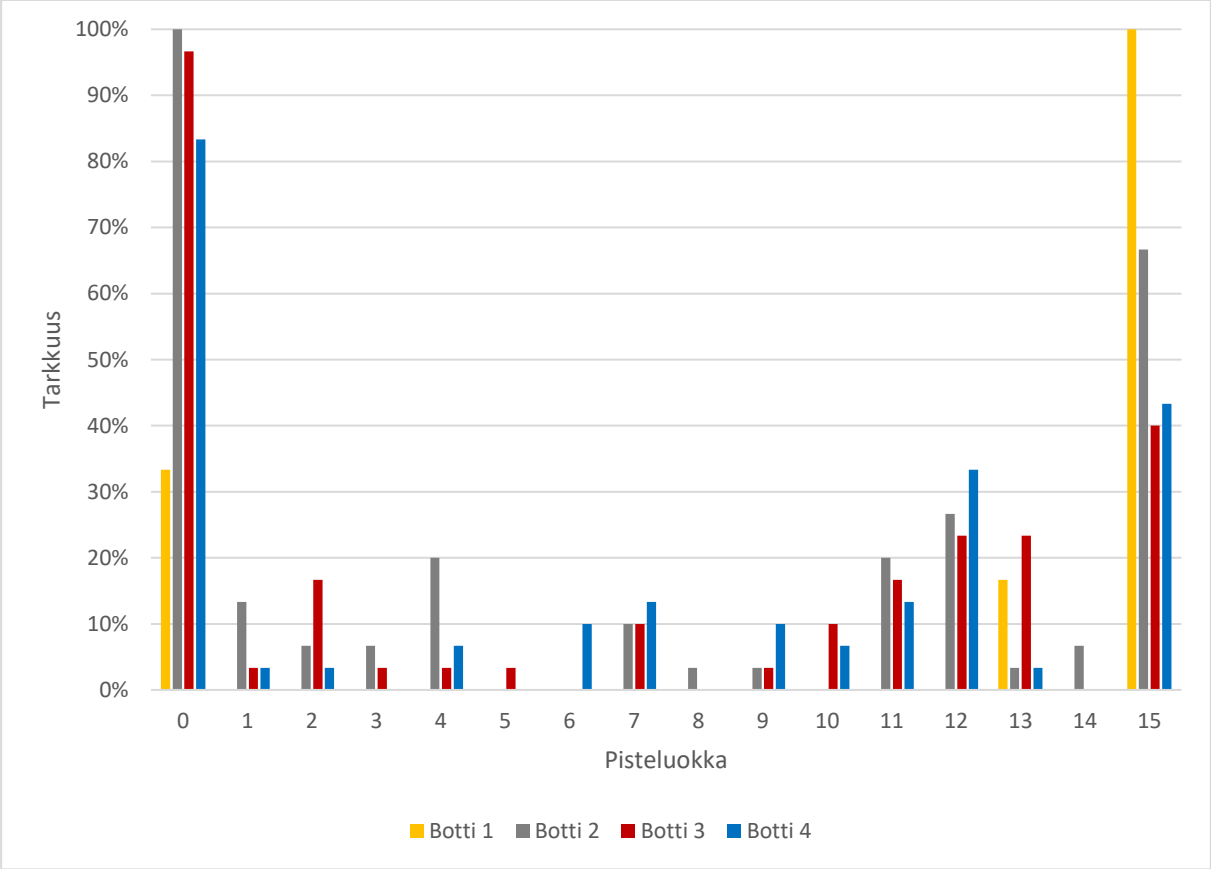
5.2 Chattibottien arvioinnin sisäinen reliabiliteetti ja validiteetti

Chattibottien kyky arvioida kokelaiden vastauksia johdonmukaisesti vaihteli välillä *hyvä* ja *erinomainen* (taulukko 8). Kaikkien chattibottien sisäkorrelaatiokerroin (ICC) vastausten kokonaisarvioinnissa oli suurempi kuin 0,9, mikä merkitsi lähes täydellistä yhdenmukaisuutta arvioinnin toistettavuudessa. Yksittäisen arvioinnin johdonmukaisuus oli boteilla 1 ja 3 *erinomaista* tasoa, kun taas botit 2 ja 4 saavuttivat *hyvän* tason. Sisäkorrelaatiokertoimen korkeat arvot sekä kokonaisarvioinnissa että yksittäisissä arvioinneissa osoittivat erittäin korkeaa reliabiliteettia arvioinnissa kaikkien chattibottien osalta.

Taulukko 8. Chattibottien arvioinnin johdonmukaisuus sisäkorrelaatiokertoimen mukaan. (P-arvo kaikissa mittauksissa <0,001)

Arvioija	Kokonaisarviointi (ICC)	Arvioinnin yhdenmukaisuus	Yksittäinen arviointi (ICC)	Arvioinnin johdonmukaisuus
Botti 1	0,991	erinomainen	0,915	erinomainen
Botti 2	0,988	erinomainen	0,893	hyvä
Botti 3	0,990	erinomainen	0,904	erinomainen
Botti 4	0,984	erinomainen	0,863	hyvä

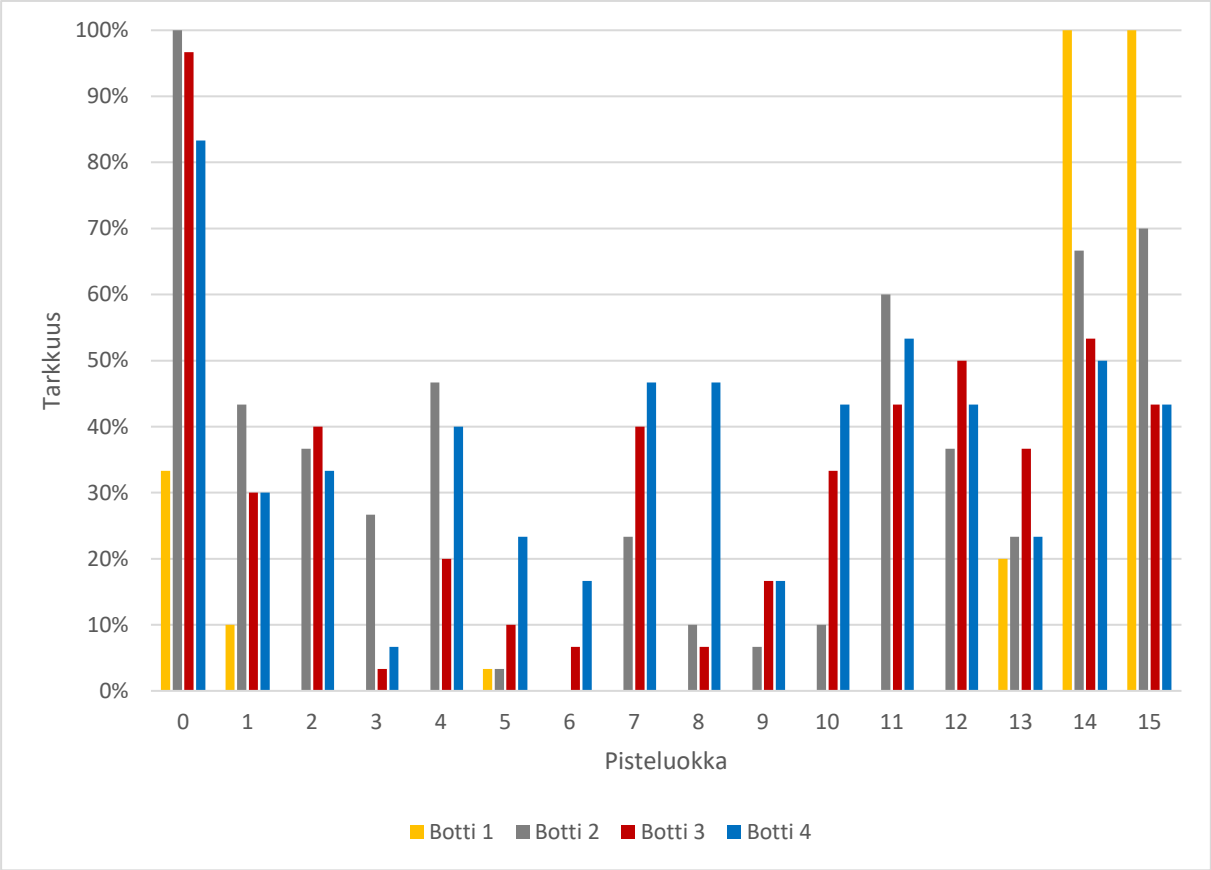
Chattibottien tarkkuus pisteyttää kokelaiden vastauksia täysin yhdenmukaisesti sensorien kanssa osoittautui erittäin alhaiseksi, mikä viittasi bottien heikkoon validiteettiin arvioinnissa (kuva 7; liite 4). Kaikkien bottien korkein tarkkuus havaittiin pienimmässä ja suurimmassa pisteluokassa, kun taas keskitason luokissa tarkkuus jäi alhaisimmaksi. Botin 1 tarkkuus oli erittäin alhainen ja lähes kaikissa luokissa olematon vastausluokkien ääripäitä lukuun ottamatta. Bottien 2, 3 ja 4 tarkkuudet olivat hyvin yhdenmukainen, mutta jäivät myös erittäin alhaiseksi keskitason pisteluokissa. Korkea tarkkuus boteilla 2, 3 ja 4 havaittiin myös pienimmässä ja suurimmassa pisteluokassa. Tämä korostaa, että vaikka botit suorittavat arviointia johdonmukaisesti useita kertoja (kuten aiemmin mainittu korkea ICC-arvo osoittaa), niiden kyky vastata ulkoisiin arviointistandardeihin, kuten sensorien antamiin arvioihin, voi jäädä heikoksi, mikä heikentää niiden käytännön soveltuvuutta arviointitehtävissä.



Kuva 7. Chattibottien tarkkuus pisteyttää kokelaiden vastauksia täysin yhdenmukaisesti sensorien kanssa.

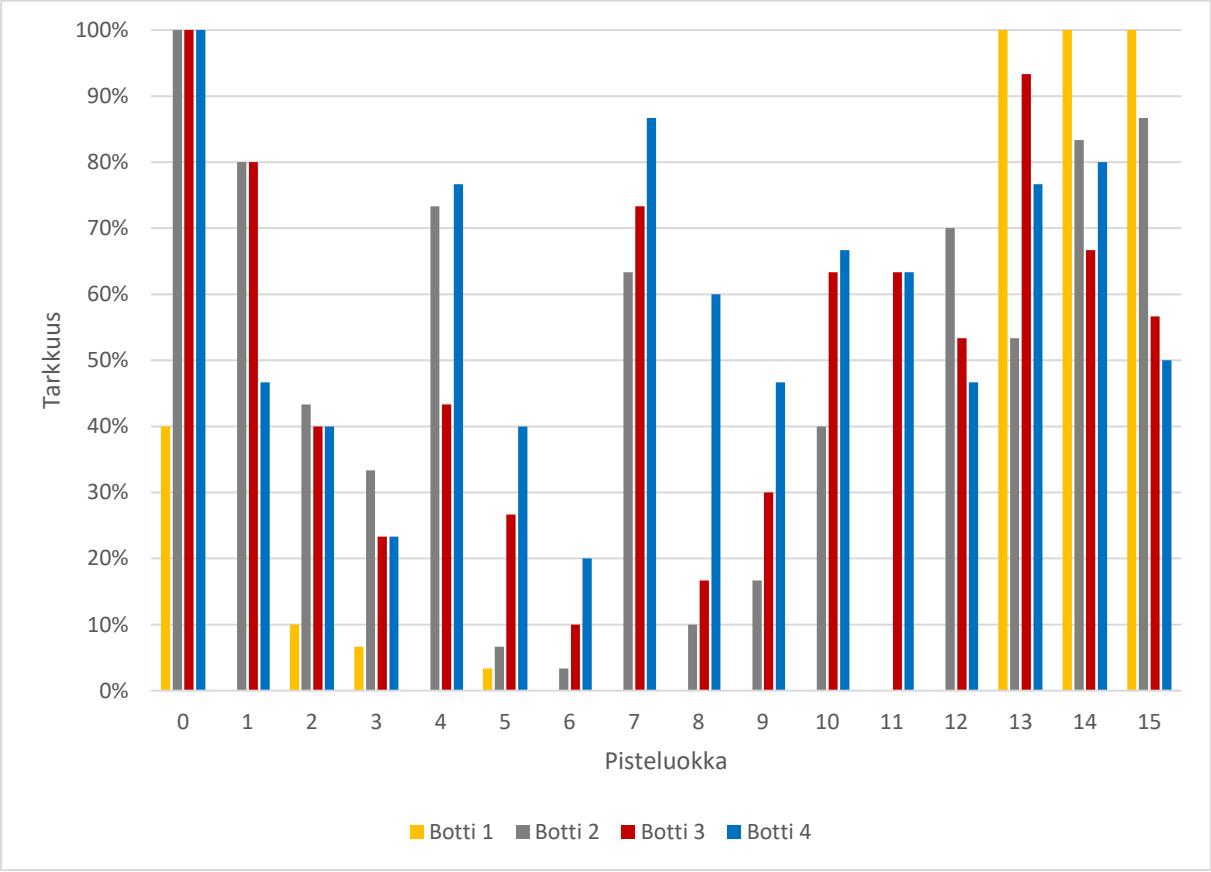
Kun sallittiin yhden pisteen ero sensorien antamiin pisteisiin, chattibottien tarkkuus vastausten pisteyttämisessä parani huomattavasti, lähes kaksinkertaistuen kaikkien bottien osalta

(kuva 8; Liite 4). Tästä huolimatta kokonaisuudessaan tarkkuus jäi edelleen alhaiseksi, saavuttaen korkeimmat arvot pisteluokkien ääripäissä. Bottien 2, 3 ja 4 suorituskyvyssä ei havaittu merkittäviä eroja, mutta Botti 1 erottui edelleen alhaisimmalla tarkkuudella, joka ei juurikaan parantunut verrattuna 0 pisteen eroon. Botin 4 tarkkuus parani merkittävämmin keskimmäisissä pisteluokissa, mutta pysyi samana pisteluokkien ääripäissä. Tämä osoitti, että bottien kyky arvioida vastauksia on parempi, kun arvioitsijoiden välillä sallitaan pientä vaihtelua.



Kuva 8. Chattibottien tarkkuus pisteyttää kokelaiden vastauksia yhden pisteen erolla sensorien pisteisiin.

Kun chattibottien suorituskykyä arvioinnissa tutkittiin sallimalla kahden pisteen ero sensorien antamiin pisteisiin, tarkkuus parani entisestään lähes kaikkien bottien kohdalla, erityisesti keskitason pisteluokissa (kuva 9; liite 4). Kuitenkin kokonaisuudessaan tarkkuus jäi alhaiseksi. Botin 1 tarkkuus pysyi alhaisena kaikissa pisteluokissa paitsi pienimmässä ja suurimmissa pisteluokissa. Bottien 2 ja 3 tarkkuus oli suhteellisen yhdenmukainen useimmissa luokissa, ilman huomattavia eroja. Botti 4 erottui korkeimmalla tarkkuudella keskimmäisissä pisteluokissa, mutta sen tarkkuus oli alhaisin korkeimmassa pisteluokassa. Kokonaisuudessaan chattibottien alhainen suorituskyky, useimmissa luokissa alle 70 %, viittaa siihen, että niiden arvioinnin validiteetti on heikko. Tämä tulos korostaa tarvetta jatkuvalle kehittämiselle ja kalibroinnille chattibottien arvioinnin tarkkuuden parantamiseksi, erityisesti niiden kyvyssä vastata monimutkaisiin arviointistandardeihin ja monipuolisten vastaustyyppien tunnistamiseen.



Kuva 9. Chattibottien tarkkuus pisteyttää kokelaiden vastauksia kahden pisteen erolla sensorien pisteisiin.

5.3 Chattibottien ja sensorien arvioinnin välinen yhdenmukaisuus

Wilcoxin ja Friedmannin tilastolliset testit paljastivat eroavaisuuksia chattibottien keskinäisessä arvioinnissa. Wilcoxin testin tulokset osoittivat, että kaikkien chattibottien ja sensorien arvioinnin välillä oli tilastollisesti merkitsevä ero (taulukko 9). Friedmannin testi paljasti, että chattibottien välinen ero arvioinnissa oli vähäisempää. Wilcoxin testin tulokset osoittivat myös, että chattibottien keskinäinen suorituskyky arvioida kokelaiden vastauksia vaihteli merkitsevästi, mikä näkyi alhaisina p-arvoina. Erityisesti bottien 1 ja 2 välillä havaittiin suurta eroavaisuutta, kun taas bottien 2 ja 3 välillä ero ei ollut tilastollisesti merkitsevä. Sensorien ja chattibottien välisen vertailun standardoitu testisuure paljasti, että eroavaisuus arvioinnissa oli suurin botin 1 ja sensorin välillä, mikä osoitti, että botti 1 poikkesi merkittävästi sensorien arvioinneista. Pienin eroavaisuus havaittiin botin 4 ja sensorien välillä, mikä viittasi siihen, että botti 4 vastasi lähimmin sensorien arviointia.

Taulukko 9. Tilastollisen testauksen tulokset arvioijien välisistä eroista Wilcoxin testin mukaan.

Arvioija	Testi-suure	Standardi-virhe	Standardoitu testi-suure	*P*-arvo	Merkitsevä ero arvioinnissa
Sensori - Botti 1	990,000	85,561	5,785	0,000	Kyllä
Sensori - Botti 2	723,000	71,316	4,669	0,000	Kyllä
Sensori - Botti 3	789,500	76,780	4,676	0,000	Kyllä
Sensori - Botti 4	133,500	73,731	-3,750	0,000	Kyllä
Botti 1 - Botti 2	0,000	87,969	-5,883	0,000	Kyllä
Botti 1 - Botti 3	0,000	88,218	-5,866	0,000	Kyllä
Botti 1 - Botti 4	0,000	85,421	-5,795	0,000	Kyllä
Botti 2 - Botti 3	250,000	45,179	0,719	0,472	Ei
Botti 2 - Botti 4	193,500	73,050	-2,964	0,003	Kyllä
Botti 4 - Botti 3	457,500	57,439	2,786	0,005	Kyllä

Friedmannin tilastollisten testauksen tulokset osoittivat, että sensorien ja botin 4 välillä ero arvioinnissa ei ollut tilastollisesti merkitsevä, mikä poikkeaa Wilcoxin testin tuloksista (taulukko 10). Myös bottien 2 ja 4 sekä bottien 4 ja 3 välillä ei havaittu tilastollisesti merkitsevää eroa.

Taulukko 10. Tilastollisten testauksen tulokset arvioijien välisistä eroista Friedmannin testin mukaan.

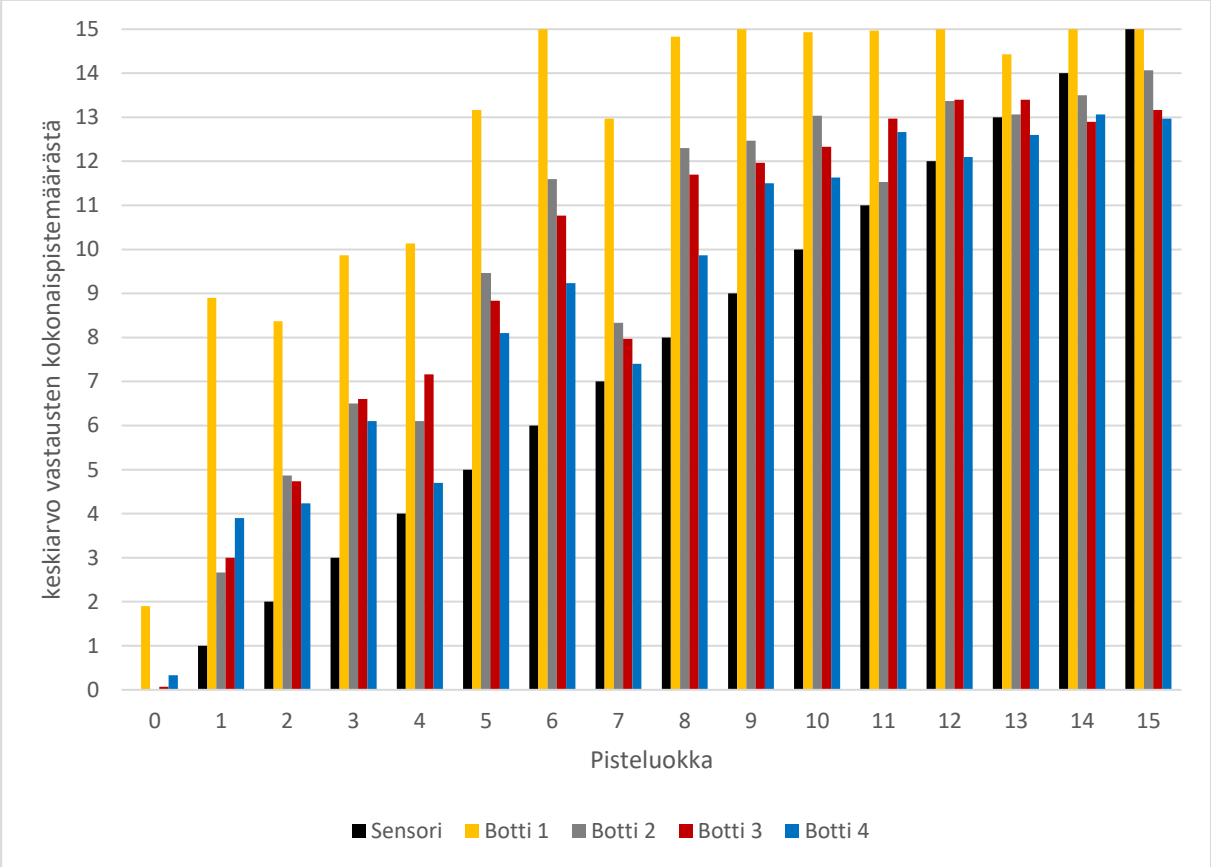
Arvioija	Testi-suure	Standardi-virhe	Standardoitu testi-suure	*P*-arvo	Korj. *P*-arvo	Merkitsevä ero arvi-oinnissa
Sensori - Botti 1	3,125	0,323	9,682	0,000	0,000	Kyllä
Sensori - Botti 2	1,344	0,323	4,163	0,000	0,000	Kyllä
Sensori - Botti 3	1,219	0,323	3,776	0,000	0,001	Kyllä
Sensori - Botti 4	0,667	0,323	2,066	0,039	0,389	Ei
Botti 1 - Botti 2	1,781	0,323	5,519	0,000	0,000	Kyllä
Botti 1 - Botti 3	1,906	0,323	5,906	0,000	0,000	Kyllä
Botti 1 - Botti 4	2,458	0,323	7,617	0,000	0,000	Kyllä
Botti 2 - Botti 3	0,125	0,323	0,387	0,699	1,000	Ei
Botti 2 - Botti 4	0,677	0,323	2,098	0,036	0,359	Ei
Botti 4 - Botti 3	0,552	0,323	1,711	0,087	0,872	Ei

Sensorien ja chattibottien välinen yhdenmukaisuus pisteiden osalta vaihteli arvioinnissa *heikosta erinomaiseen* (taulukko 11). Chattibottien ja sensorien välinen yhdenmukaisuus oli kokonaisarvioinnissa parempi verrattuna yksittäisiin arviointeihin. Botin 1 ja sensorien välinen yhteneväisyys arvioinnissa oli sisäkorrelaatiokertoimen mukaan alhaisin. Korkein reliabiliteetti havaittiin botin 4 ja sensorien välillä. Bottien 2 ja 3 yhdenmukaisuus sensorien kanssa oli samankaltaista. Tulosten perusteella kaikki chattibotit bottia 1 lukuun ottamatta osoittivat erittäin yhteneväistä arviointia sensorien kanssa, mikä viittaa arvioitsijoiden väliseen korkeaan reliabiliteettiin.

Taulukko 11. Chattibottien ja sensorien välinen arvioinnin johdonmukaisuus sisäkorrelaatiokertoimen mukaan. (P-arvo kaikissa mittauksissa <0,001)

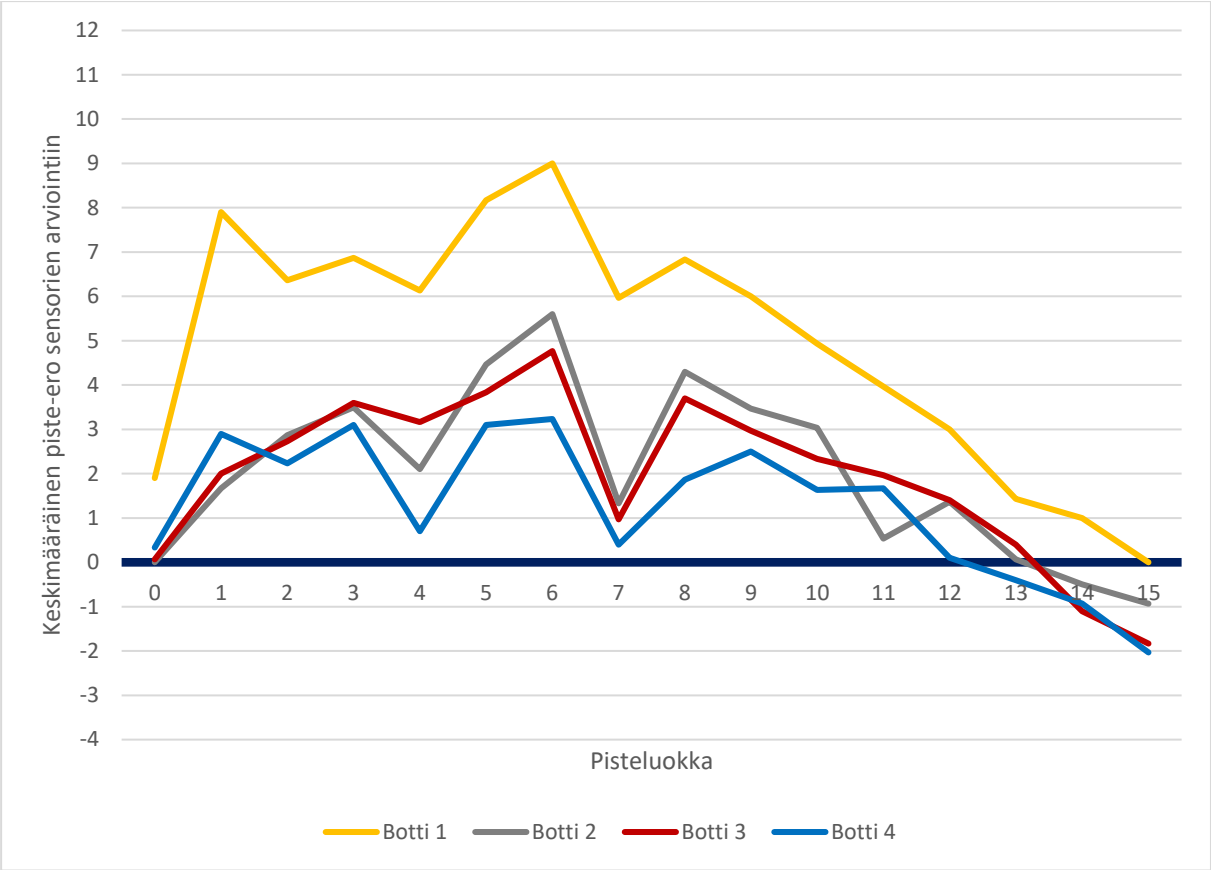
Vertailu-asetelma	Kokonais-arviointi (ICC)	Arvioinnin yhdenmu-kaisuus	Yksittäinen arviointi (ICC)	Arvioinnin yhdenmu-kaisuus
Botti 1 - Sensori	0,625	keskinkertainen	0,454	heikko
Botti 2 - Sensori	0,884	hyvä	0,792	hyvä
Botti 3 -Sensori	0,889	hyvä	0,800	hyvä
Botti 4 - Sensori	0,924	erinomainen	0,859	hyvä

Chattibottien arviointi poikkesi toisistaan ja sensorien arvioinnista kaikissa pisteluokissa, mikä havaittiin tarkastelemalla keskimääräistä vastausten kokonaispistemäärien eroa kussakin luokassa (kuva 10). Chattibottien arvioinnit tuottivat keskimäärin korkeampia pistemääriä verrattuna sensorien arviointeihin. Erityisesti matalissa ja keskitason pisteluokissa erot olivat suurimmat, kun taas korkeimmissa pisteluokissa erot olivat pienimmät. Botin 1 arviointi poikkesi muista malleista huomattavasti kaikissa pisteluokissa, kun taas bottien 2, 3 ja 4 erot arvioinnissa olivat pienemmät. Botin 4 arviointi oli lähellä sensorien arviointia lähes kaikissa pisteluokissa. Botin 2 arviot olivat yhdenmukaisempia sensorien arviointien kanssa korkeimmissa ja matalimmissa pisteluokissa verrattuna botti 3:een, jonka arvioinnit vastasivat sensorien arviointeja paremmin keskitason pisteluokissa.



Kuva 10. Keskimääräinen vastausten kokonaispistemäärä chattibottien ja sensorien arvioinnin mukaan pisteluokittain.

Keskimääräiset piste-erot sensorien ja chattibottien välillä vaihtelivat eri pisteluokissa (kuva 11). Suurimmat vaihtelut chattibottien välillä havaittiin keskimmäisissä pisteluokissa, lukuun ottamatta luokkaa 7, jossa bottien arviointi oli yhteneväistä. Chattibottien arviointi oli yhdenmukaisinta vain 0-luokassa ja korkeimmissa pisteluokissa.



Kuva 11. Chattibottien arvioinnin keskimääräinen ero sensorien arviointiin pisteluokittain.

Botti 1 erottui muista antamalla vastauksille kaikissa pisteluokissa enemmän pisteitä. Bottien 2 ja 3 arviointieroissa ei havaittu suurta vaihtelua. Botin 4 arviointi oli kokonaisuudessaan kaikkein yhdenmukaisinta sensorien arvioinnin kanssa, poiketen siitä enintään keskimäärin kolmella pisteellä. Chattibottien arvioinnissa oli suurta vaihtelua pisteluokkien sisällä (taulukko 12), mistä havaitaan, että todellisuudessa yhdenmukaisuus arvioinnissa sensorien ja bottien välillä poikkesi huomattavasti enemmän useissa pisteluokissa.

Taulukko 12. Chattibottien arvioinnin vaihteluväli vastauksissa pisteluokittain.

	Pisteluokka															
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Botti 1	1–2	6–12	6–10	7–12	10–10	12–15	15–15	12–14	15–15	15–15	15–15	15–15	15–15	13–15	15–15	15–15
Botti 2	0–0	2–3	2–7	4–8	5–8	9–10	11–12	7–10	12–13	12–13	12–14	11–13	12–15	10–14	12–15	13–15
Botti 3	0–0	2–4	2–6	6–7	6–9	7–10	9–13	6–10	11–12	11–13	11–15	11–15	12–15	12–14	12–14	12–14
Botti 4	0–1	2–5	3–6	6–6	2–6	8–9	8–10	6–9	9–10	11–12	9–15	11–14	11–14	12–14	11–14	12–15
	Pisteiden vaihteluväli pisteluokassa															

5.4 Arviointiohjeiden soveltaminen chattibottien ja sensorien arvioinnissa

Sisällönanalyysin tuloksena tuotettiin tarkemmat arviointikriteerit, joiden mukaan sensorit oletettavasti arvioivat ja pisteyttivät kokelaiden vastauksia soveltaen *Hyvän vastauksen piirteiden* arviointiohjeita (taulukko 13). Sadetyyppien nimeämisessä oli keskeistä käsitteen oikea nimeäminen ja kirjoitusasu, mikä oli kriteerinä pisteen saamiselle. Esimerkkialueen tarkastelussa selvisi, että alueen täytyi olla tyypillinen sadetyypin esiintymisalue ja siitä pystyi päättelemään maantieteellisen sijainnin. Sadetyypin syntytavan kuvauksessa pisteytys perustui vastauksen oikeellisuuteen, selkeyteen ja kokelaan ymmärtämisen tasoon sateen synnystä prosessina. Vastauksen pisteytykseen vaikuttivat myös tehtävän kannalta epäolennaisen tiedon esittäminen, joka saattoi vähentää pisteitä tiedon ollessa väärää tai ristiriitaista.

Taulukko 13. *Hyvän vastauksen piirteistä* johdetut tarkemmat arviointiohjeet, joiden mukaan sensorit oletettavasti suorittivat arviointia.

Arviointikategoria	Vastauksen sisällön piirteet	Pisteet
Nimeäminen	Käsite on nimetty väärin tai kirjoitusasu on virheellinen	0 p
	Käsite on nimetty oikein ja kirjoitusasu on virheetön	1 p
Esimerkkialue	Ei tyypillisin esiintymisalue, maantieteellistä sijaintia ei voi päätellä	0 p
	Tyypillinen esiintymisalue, maantieteellinen sijainti pääteltävissä	1 p
Syntytavan kuvaus	Puutteellinen ja epäselvä, ymmärretty väärin	0 p
	Pinnallinen, eikä osoita ymmärrystä	1 p
	Selkeä, mutta pieniä puutteita	2 p
	Selkeä ja jäsennelty, osoittaa vahvaa ymmärrystä	3 p
Tehtävän kannalta epäolennainen tieto	Esitetyt asiat ovat väärin tai ristiriidassa keskenään	Vähentää
	Esitetyt asiat ovat oikein	Ei vaikutusta
	Ei mahdollinen	Lisää pisteitä

Chattibotin 1 antaman palautteen sisällönanalyysistä havaittiin merkittäviä eroja *Hyvän vastauksen piirteiden* soveltamisesta arvioinnissa sensorien ja chattibotin välillä (taulukko 14). Botti 1 antoi vastauksesta pisteitä aina, kun siihen oli kirjoitettu jotain huolimatta siitä, oliko kirjoituksessa mitään tehtävän kannalta relevanttia sisältöä. Sadetyypin nimeämisessä käsitteen oikeinkirjoituksella ei ollut vaikutusta pisteytykseen, vaan arviointi perustui käsitteen tunnistettavuuteen. Esimerkkialueen maininnasta riitti mikä tahansa esiintymisalue sateelle, jos sitä havaitaan alueella edes joskus. Lisäksi maantieteellisen sijainnin ei tarvinnut käydä ilmi esimerkkialueesta. Syntytavan kuvauksesta botti 1 arvioi vastaukset vain korkeimpiin pisteluokkiin minkä tahansa sateen syntytavan kuvauksen perusteella riippumatta kuvauksen laadusta tai oikeellisuudesta. Muiden kuin pyydettyjen sadetyyppien kuvauksesta, nimeämisestä ja esimerkkialueen maininnasta saattoi saada lisäpisteitä, vaikka ne olivatkin epäolennaista tietoa tehtävän kannalta.

Taulukko 14. *Hyvän vastauksen piirteiden* soveltaminen botin 1 arvioinnissa.

Arviointikategoria	Vastauksen sisällön piirteet	Pisteet
Nimeäminen	-	0 p
	Kirjoitusasussa on virheitä, mutta käsite tunnistettavissa	1 p
Esimerkkialue	-	0 p
	Ei tarvitse olla tyypillinen esiintymisalue, maantieteellinen sijainti ei tarvitse olla pääteltävissä	1 p
Syntytavan kuvaus	-	0 p
	-	1 p
	Mikä tahansa kuvaus	2 p
	Mikä tahansa kuvaus	3 p
Tehtävän kannalta epäolennainen tieto	-	Vähentää
	-	Ei vaikutusta
	Muun sadetyypin kuvauksesta, nimeämisetä tai esimerkkialueesta	Lisää

Chattibotin 4 palautteen sisällönanalyysistä ilmeni, että tarkempien arviointikriteerien luominen auttoi bottia soveltamaan arviointiohjeita huomattavasti paremmin (taulukko 15). Kriteerit sadetyyppien nimeämiselle, esimerkkialueen antamiselle ja syntytavan kuvaukselle vastasivat sensorien *Hyvän vastauksen piirteiden* soveltamisen tuloksista johdettuja arviointikriteereitä, mutta niitä ei noudatettu arvioinnissa systemaattisesti.

Taulukko 15. Sensorien arviointikriteereistä johdettujen tarkempien arviointiohjeiden soveltaminen botin 4 arvioinnissa.

Arviointikategoria	Vastauksen sisällön piirteet	Pisteet	Huomioitavaa
Nimeäminen	Käsite on väärä tai kirjoitusasu virheellinen	0 p	Muutamissa vastauksissa oikeasta käsitteestä ei pisteitä
	Käsite on oikea ja kirjoitusasu moitteeton	1 p	Muutamissa vastauksissa virheellisestä kirjoitusasusta pisteitä
Esimerkkialue	Ei edusta tyypillistä esiintymisaluetta, maantieteellinen sijainti ei pääteltävissä	0 p	Joissakin vastauksissa ei pisteitä oikeasta esiintymisalueesta
	Edustaa tyypillistä esiintymisaluetta, maantieteellinen sijainti pääteltävissä	1 p	Joissakin vastauksissa pisteitä, jos esiintymisalue mahdollinen, mutta ei yleisin
Syntytavan kuvaus	Virheellinen tai sitä ei ole lainkaan	0 p	Vain muutamia vastauksia
	Mainінta joistakin arviointimatriisin mukaisista kuvauksista riittää	1 p	Vastauksen selkeydellä tai jäsennyksellä ei vaikutusta arviointiin. Lisäksi ymmärtämisen tasolla ei havaittavaa vaikutusta pisteytykseen. Pisteluokkien vastauksien välillä ei merkittävää eroa.
	Mainінta arviointimatriisin mukaisista kuvauksista riittää	2 p	
	Mainінta arviointimatriisin mukaisista kuvauksista riittää	3 p	
Tehtävän kannalta epäolennainen tieto	Ei mahdollinen	Vähentää	Ei havaittavissa
	Ei mahdollinen	Ei vaikutusta	Ei havaittavissa
	Ei mahdollinen	Lisää pisteitä	Ei havaittavissa

Lisäksi botin antamissa palautteissa ja arvioinnissa havaittiin ristiriitaisuutta. Sadetyypin väärästä tai virheellisestä nimeämisestä saattoi silti ansaita pisteen, kun taas oikeasta käsitteestä ja kirjoitusasusta ei. Esimerkkialueen arvioinnissa botti luokitteli joissain tapauksissa hyväksytyt alueet virheellisiksi ja päinvastoin. Syntytavan kuvauksessa pisteluokituksen perustana oli yksittäisten asioiden mainitseminen. Botti ei kiinnittänyt huomiota vastauksen ymmärtämisen tasoon tai kokonaislaatuun selkeyden tai jäsennyksen näkökulmasta. Tehtävän kannalta epäolennaisen tiedon esittämisellä ei ollut vaikutusta pisteytykseen.

6 Keskustelu

6.1 Erot chattibottien ja sensorien välisessä arvioinnissa

Chattibottien ja sensorien välinen arviointi erosi sekä tulosten että prosessin osalta. Arvioinnin tulosten vertailussa havaittiin, että chattibotit antoivat vastauksille enemmän pisteitä lähes kaikissa pisteluokissa sensoreihin verrattuna. Lisäksi chattibottien välisessä arvioinnissa oli poikkeavuutta eritysesti botin 1 ja 2 välillä.

Chattibotit kykenivät arvioimaan yhden vastauksen noin viidessä sekunnissa, mikä teki arvioinnista erittäin tehokasta ajallisesti. Lisäksi bottien tarkkuus arvioida vastauksia johdonmukaisesti oli erittäin korkea, mikä viittaisi siihen, että ne eivät tee arvioinnissa huolimattomuusvirheitä. Nämä tulokset tukevat aikaisempaa tutkimuskirjallisuutta, jossa mainitaan ChatGPT:n tarjoama mahdollisuus nopeuttaa arviointiprosessia ja vähentää inhimillisistä tekijöistä mahdollisesti aiheutuvia virheitä arvioinnissa (Chen ym. 2020; Božić ja Indrasen Poola 2023; Rudolph ym. 2023).

Arviointiohjeiden soveltamisen tarkastelussa kävi ilmi, että chattibotit tulkitsivat arviointiohjeita eri tavoin kuin sensorit, mikä havaittiin varsinkin botin 1 ja sensorien välisen arvioinnin tuloksissa suurina piste-eroina. Tarkempien arviointiohjeiden johtaminen *Hyvän vastauksen piirteistä* kuitenkin paransi arvioinnin yhdenmukaisuutta chattibottien ja sensorien välillä, mikä havaittiin myös botin 4 antamasta palautteesta. Tästä huolimatta chattibotti ei noudattanut arviointiohjeita yhtä johdonmukaisesti kuin sensorit, mikä oli havaittavissa arvioinnin tuloksissa ja palautteissa. Tulos on linjassa niiden tutkimusten kanssa, joissa korostetaan arviointiohjeiden tulkinnanvaraisuudesta aiheutuneita riskejä arvioinnin yhdenmukaisuudelle (Atjonen 2014; Ragupathi ja Lee 2020).

Chattibotit korostivat vastausten arvioinnissa eri asioita kuin sensorit. Tämä havaittiin selkeinä eroina sadetyyppien syntytavan pisteytyksessä, jossa bottien arviointi perustui yksittäisten asioiden esiintymiseen kokelaiden vastauksissa. Tämä poikkesi sensorien arvioinnista siinä, että chattibotit eivät kyenneet arvioimaan kokelaan todellista ymmärrystä, johon sensorit mahdollisesti kiinnittivät ensisijaisesti huomiota arvioinnissa. Tätä päätelmää tukivat tulokset chattibottien alhaisesta tarkkuudesta, jolla mitattiin arvioinnin validiteettia. Tulos on linjassa Kocoń ja kumppaneiden (2023) teettämän laajan tutkimuksen mukaan, josta selvisi GPT-4-tekoälymallin haasteellisuus ymmärtää arvioitavien vastausten sisältöä.

Chattibottien ja sensorien arvioinnin erojen syitä voidaan analysoida tarkastelemalla tuloksia kokonaisvaltaisesti ChatGPT:n toimintaperiaatteiden näkökulmasta. Arviointiprosessi eroaa ihmisen arvioinnista perustavanlaatuisesti siinä, että generatiivisen tekoälymallin toiminta perustuu tilastolliseen päättelyyn ja ennusteiden luomiseen ilman todellista ymmärrystä tai tietosuutta (Korteling ym. 2021). ChatGPT:n nopeus arvioinnissa perustuu sen taustalla toimivien supertietokoneiden suoritusytimien kykyyn prosessoida suuria määriä dataa samanaikaisesti (Wu ym. 2023). Sen korkea tarkkuus arvioinnin toistettavuudessa johtuu mallin toimintalogiikasta, jossa arvioinnin tulos perustuu tilastolliseen päättelyyn ihmisen kaltaisen ajattelun sijaan (Bandi ym. 2023).

Hyvän vastauksen piirteiden tulkinnanvaraisuus chattibotin 1 arvioinnissa johtui todennäköisesti siitä, että sille annetut kehotteet tekoälymallin toiminnan ohjaamiseksi olivat epätarkat. Aikaisemmat tutkimukset korostavat kehotesuunnittelun tärkeyttä ChatGPT:n tarkkuuden parantamisessa (Giray 2023; Wei ym. 2023). Tarkemmista arviointiohjeista huolimatta chattibotin 4 ja sensorien arvioinnissa havaittiin eroja arviointiohjeiden soveltamisessa ja johdonmukaisessa noudattamisessa. Arviointiohjeiden epäjohdonmukainen noudattaminen botin 4 arvioinnissa johtui todennäköisesti siitä, että botti tulkitsi kokelaan vastauksen väärin. Tämä voi johtua ChatGPT:n toimintaperiaatteesta, jossa tokenisoinnin tuloksena yksittäiset kirjoitusvirheet vastauksessa saatavat muuttaa muuntajien ennusteita syötteen kontekstista, mikä vaikuttaa mallin luomaan tulosteeseen hallusinointina (Lee 2023). Chattibottien ja sensorien eroavaisuudet arviointiohjeiden noudattamisessa voisivat johtua GPT-4-tekoälymallin esikoulutetun datan laadusta, mikä vaikuttaa mallin toimintaperiaatteisiin vastausten luokittelussa. Lisäksi ChatGPT:n kyvyttömyys arvioida ymmärtämistä perustuu sen matemaattisiin malleihin, jotka ohjaavat sen toimintaa ymmärtämisen sijasta (Kocoń ym. 2023).

Chattibottien ja sensorien arvioinnin yhdenmukaisuus vaihteli eri pisteluokissa, mikä viittaisi siihen, että ChatGPT ei pysty soveltamaan arviointiohjeita erityyppisten vastausten arvioinnissa. Ero johtuu todennäköisesti siitä, että tekoälymallin koulutusdatan takia se ei tunnista vastausten sisältöä samalla tavalla kuin ihminen. Tämä käy ilmi aikaisemmista tutkimuksista, joissa ChatGPT:n hienosäätäminen arviointiin soveltuvaksi on vaatinut paljon ihmisen arvioimia esimerkkivastauksia (Latif ja Zhai 2024).

6.2 Chattibottien ohjeistus arviointiin sopivaksi

Tarkastelemalla arvioinnin yhdenmukaisuutta chattibottien ja sensorien välillä voitiin havaita, että botin 1 arviointi poikkesi sensorien arvioinnista eniten ja botin 4 vähiten. Tästä voidaan päätellä, että *nollakehottaminen Hyvän vastauksen piirteitä* hyödyntäen ei ole toimiva tapa ChatGPT:n ohjeistamisessa arviointiin sopivaksi. Botin 4 ohjeistuksessa hyödynnettiin *ajatusketjukehottamista* tarkemmilla arviointiohjeilla ja *vähäisen ohjauksen kehottamista* pisteittäin luokitetuilla esimerkkivastauksilla, joilla saavutettiin paras yhteneväisyys ChatGPT:n ja sensorien arvioinnin välille. Merkittävin parannus arvioinnin yhdenmukaisuudessa saavutettiin *ajatusketjukehotemenetelmään* perustuvilla tarkemmilla arviointiohjeilla, jotka kehitettiin *Hyvän vastauksen piirteistä*. Tämä havaittiin suurimpana eroina botin 1 ja 2 arvioinnissa. Tutkimustulokset ovat linjassa aikaisemman tutkimuskirjallisuuden kanssa, jossa todetaan *ajatusketjukehotemenetelmän* olevan tehokas tapa ohjeistaa ChatGPT:tä (Wei ym. 2023).

Bottien 2, 3 ja 4 välillä ei havaittu suurta eroa arvioinnin tuloksissa pisteellisessä tarkastelussa, mikä viittaa siihen, että oppimateriaalin ja esimerkkivastausten näyttäminen GPT-4-tekoälymallille ei välttämättä paranna sen suorituskykyä arvioinnissa yhtä paljon kuin arviointiohjeiden tarkentaminen. Tämä eroaa aikaisemmista tutkimuksista siinä, että ChatGPT:n hienosäätämisessä esimerkkivastauksilla on ollut suuri vaikutus mallin suorituskyvyn parantamisessa (Latif & Zhai 2024).

Chattibottien arvioinnin toistettavuus oli hyvin johdonmukaista riippumatta ohjeistusmenetelmästä, mikä voitiin huomata tarkastelemalla bottien sisäkorrelaatiokertoimen arvoja, jotka vaihtelivat *erinomaisen* ja *hyvän* välillä. Tämä osoittaa, että ChatGPT:n reliabiliteetti arvioinnissa on korkea, mikä on linjassa aikaisemman tutkimuskirjallisuuden kanssa (Tobler 2024). Bottien tarkkuus vastausten pisteyttämissä täysin yhdenmukaisesti sensorien kanssa jäi kaikkien bottien osalta erittäin alhaiseksi (alle 20 %) ja oli kahden pisteen erollakin korkeintaan hieman alle 60 % botin 4 osalta. Alhaisesta tarkkuudesta voidaan päätellä, että vaikka chattibottien validiteetti arvioinnissa parani eri ohjeistusmenetelmien myötä, ChatGPT ei välttämättä tunnista kaikkia vastaustyylejä, mikä voi vaikuttaa arvioinnin lopputulokseen. Tulos on linjassa Mizumoton ja Eguchin (2023) tutkimustulosten kanssa, jossa erot arvioinnissa ChatGPT:n ja ihmisen välillä johtuivat GPT-3.5-mallin kyvyttömyydestä tunnistaa erilaisia vastaustyylejä.

6.3 Eettinen tarkastelu ja ChatGPT osana opettajan työtä

Tulosten perusteella voitiin havaita, että tarkempien arviointiohjeiden syöttäminen tekoälymallille *Ajatusketjukehottamismenetelmää* hyödyntäen paransi mallin kykyä arvioinnissa huomattavasti. Tuloksista kävi myös ilmi, että ChatGPT:n reliabiliteetti arvioinnissa on hyvä. Lisäksi tilastollisten testien, kuten Friedmannin ja sisäkorrelaatiokertoimen tulokset antoivat kuvan, että ChatGPT:n ja ihmisen välisessä arvioinnissa ei havaittu suuria eroja. Sisällönanalyysin ja tarkkuusmittauksen tulokset kuitenkin paljastivat puutteita chattibottien arvioinnin validiteetissa. Lisäksi ChatGPT:n kyky noudattaa arviointiohjeita johdonmukaisesti osoittautui heikoksi. Tekoälymallin käyttö arvioinnissa vaatii aina testausta ennen sen laajamittaista käyttöönottoa. Tutkimuksessa chattibotit arvioivat vastauksia yhteensä 1920 kertaa, minkä avulla saatiin melko luotettava kuva GPT-4-mallin suorituskyvystä vastausten arvioimisessa eri ohjeistusmenetelmiä hyödyntämällä. Todellista arvioinnin laatua voitiin kuitenkin tarkastella vasta syvällisemmällä sisällönanalyysillä, joka vei huomattavasti aikaa. Tästä näkökulmasta on siis kyseenalaista vähentääkö ChatGPT:n käyttö opettajan työtaakkaa arvioinnissa ajallisesti, jos tekoälymallin suorittaman arvioinnin analysoiminen on työlästä itsessään.

Luostarisen ja Ouakrim-Soivion (2019) mukaan laadukkaan ja eettisen arvioinnin tunnusmerkkejä ovat läpinäkyvyys, johdonmukaisuus ja perusteltavuus. Tämän näkökulman mukaan ChatGPT:stä ei tutkimuksen perusteella ole mahdollista ohjeistaa pelkkien kehotemenetelmien avulla luotettavaa työkalua arvioinnin apuvälineeksi. Ensinnäkin, vaikka bottien korkea reliabiliteetti arvioinnin toistettavuudessa oli hyvä, niiden validiteetti osoittautui huonoksi. Tästä voidaan päätellä, että chattibotit suorittivat arviointia systemaattisesti väärin, arvioiden vääriä asioita hyvin johdonmukaisesti. Toiseksi ChatGPT:n suorittamassa arvioinnissa havaittiin vaihtelua arviointiohjeiden soveltamisessa ja noudattamisessa, mikä tekee arvioinnista epäjohdonmukaista. Kolmanneksi ChatGPT:n arvioinnin läpinäkyvyyttä heikentää sen monimutkainen toimintamalli, jonka vuoksi arviointiprosessia on vaikea ymmärtää. Lisäksi chattibottien antamissa palautteissa havaittiin ristiriitaisuutta, mikä tekee arvioinnin perusteltavuudesta haastavaa.

6.4 Virhelähteet, kehitysehdotukset ja jatkotutkimustarpeet

Tutkimustulosten luotettavuuteen vaikuttavat tutkimusaineiston edustavuus, tutkimuksen tekijän tulkinnat arviointiohjeista ja ymmärrys generatiivisesta tekoälystä. Tutkimuksen otantakoko oli varsin pieni (n=96), joka saattoi vaikuttaa tarkempien arviointiohjeiden luomiseen siltä osin, että se ei edustanut kaikkia syksyn 2023 maantieteen ylioppilaskokeen vastauksia valittuun koetehtävään. Lisäksi tutkimuksessa oletettiin, että sensorien arviointi oli hyvin johdonmukaista vastausten arvioinnissa, mutta todellisuudessa pieni vaihtelu ihmisen arviointien välillä on luonnollista. Tämä pyrittiin kuitenkin huomioimaan tarkastelemalla ChatGPT:n tarkkuutta kahden pisteen erolla sensorien arviointiin, mikä vastaa todellisempaa vertailutilannetta kahden arvioitsijan välillä.

Tulkintani *Hyvän vastauksen piirteiden* soveltamisesta sensorien arvioinnissa saattoivat poiketa sensorien todellisesta arvioinnista. Tämä vaikutti oleellisesti tarkempien arviointiohjeiden luomiseen ja tätä kautta myös ChatGPT:n suorittaman arvioinnin tuloksiin. Toisaalta, jos en itse ei kyennyt ymmärtämään syvällisen sisällönanalyysin jälkeenkään, miten *Hyvän vastauksen piirteiden* arviointiohjeita todellisuudessa sovelletaan sensorien arviointityössä, pitäisi ohjeita muuttaa yksiselitteisemmäksi arvioinnin yhdenmukaisuuden parantamiseksi.

Oma ymmärrykseni ChatGPT:stä ja generatiivisesta tekoälystä kehittyi koko tutkimuksen ajan, mikä saattoi vaikuttaa tutkimuksen eri vaiheisiin, tutkimustuloksiin ja tulosten tulkintaan. Erityisesti tekoälyyn nopea kehittyminen ja siihen liittyvien käsitteiden epämääräinen käyttö tutkimuskirjallisuudessa ja puhekielessä vaikeutti tutkimuksen toteuttamista. Tutkimustuloksista huolimatta ChatGPT:n hyödyntäminen arvioinnissa vaatii lisätutkimusta. Erityisesti arviointia parantavien kehotemenetelmien selvittäminen on tärkeää helppokäyttöisten ja toimivien arviointityökalujen luomiseksi. Generatiivisen tekoälyn kehittyessä sen soveltamismahdollisuudet yhteiskunnassa laajenevat, mikä edellyttää lisätutkimuksia näiden järjestelmien hyödyntämisestä arvioinnissa.

Kiitokset

Haluan kiittää opinnäytteen tekemisessä yliopistolehtori Sanna Mäkeä, joka kannusti ja uskoi osaamiseeni koko graduprosessin aikana. Kiitos myös professori Jussi Jauhiaiselle ja Agustin Garagorry Guerralle tekoälyyn liittyvästä asiantuntijuuden jakamisesta. Erityiskiitos myös Ylioppilastutkintolautakunnalle tutkimusaineiston toimittamisesta, jota ilman tutkimuksen tekeminen ei olisi ollut mahdollista. Lopuksi haluan kiittää kaikkia läheisiäni, kuten perheenjäseniäni, rakasta avopuolisoani ja ystäviäni, jotka tukivat minua opinnäytteen tekemisessä alusta loppuun asti.

Lähteet

Adamopoulou, E. & Moussiades, L. (2020) Chatbots: history, technology, and applications. *Machine Learning with Applications* 2(2020). https://doi.org.ezproxy.utu.fi/10.1016/j.mlwa.2020.100006

Akib, E. (2015) The validity and reliability of assessment for learning (Afl). *Education Journal* 4(2) 64–68. https://doi.org/10.11648/j.edu.20150402.13

Almasre, M. (2024) Development and evaluation of a custom GPT for the assessment of atudents'designs in a Typography course. *Education Sciences* 14(2). https://doi.org/10.3390/educsci14020148

Atjonen, P. (2007) Arvioinnin tehtävät ja lajit. Teoksessa Atjonen, P. (toim.) *Hyvä, paha arviointi*, 66–69. Jyväskylä: Gummerus Kirjapaino Oy.

Atjonen, P. (2014) Teachers' views of their assessment practice. *The Curriculum Journal* 25(2) 238–259. https://doi.org/10.1080/09585176.2013.874952

Aydin, Ö. & Karaarslan, E. (2023) Is ChatGPT leading generative AI? what is beyond expectations? *Academic Platform Journal of Engineering and Smart Systems* 11(3) 118–134. https://doi.org/10.21541/apjess.1293702

Baidoo-Anu, D. & Ansah, L. O. (2023) Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI* 7(1) 52–62. http://dx.doi.org/10.61969/jai.1337500

Bandi, A., Adapa, P. V. S. R. & Kuchi, Y. E. V. P. K. (2023) The power of generative AI: a review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet* 15(8). https://doi.org/10.3390/fi15080260

Black, P. & Wiliam, D. (1998) Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice* 5(1) 7–74. https://doi.org/10.1080/0969595980050102

Black, P., Harrison, C., Hodgen, J., Marshall, B. & Serret, N. (2010) Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice* 17(2) 215–232. https://doi.org/10.1080/09695941003696016

Božić, V. & Poola, I. (2023) Chat GPT and education. http://dx.doi.org/10.13140/RG.2.2.18837.40168

Briganti, G. (2024) How ChatGPT works: a mini review. *European Archives of Oto-Rhino Laryngology* 281(3) 1565–1569. https://doi.org/10.1007/s00405-023-08337-7

Brown, G. T. L., Andrade, H. L. & Chen, F. (2015) Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy & Practice* 22(4) 444–457. https://doi.org/10.1080/0969594X.2014.996523

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020) Language models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165

Chen, L., Chen, P. & Lin, Z. (2020) Artificial intelligence in education: a review. *IEEE Access* 8(2020) 75264-75278. https://doi.org.ezproxy.utu.fi/10.1109/ACCESS.2020.2988510

Chiu, T. K. F. (2024) Future research recommendations for transforming higher education with generative AI. *Computers and Education: Artificial Intelligence* 6(2024). https://doi-org.ezproxy.utu.fi/10.1016/j.caeai.2023.100197

Chufama, M., & Sithole, F. (2021) The pivotal role of diagnostic, formative and summative assessment in higher education institutions' *Teaching and Student Learning* 4(5) 5-15. http://ijmrap.com/wp-content/uploads/2021/10/IJMRAP-V4N4P107Y21.pdf

Church, K. W., Chen, Z. & Ma, Y. (2021) Emerging trends: a gentle introduction to fine tuning. *Natural Language Engineering* 27(6) 763–778. https://doi.org/10.1017/S1351324921000322

Classification: Accuracy | Machine Learning (2024) Google for Developers. https://developers.google.com/machine-learning/crash-course/classification/accuracy 22.4.2024.

Copilot for Microsoft 365 (2024) Microsoft Adoption. https://adoption.microsoft.com/en-us/copilot/ 20.2.2024.

Creating a GPT | OpenAI Help Center (2024) OpenAI. https://help.openai.com/en/articles/8554397-creating-a-gpt 14.4.2024.

De Spiegeleire, S., Maas, M. & Sweijs, T. (2017) What is artificial intelligence? *Artificial intelligence and the future of defense: strategic implications for small- and medium-sized force providers.* 25-42. http://www.jstor.org/stable/resrep12564.7 16.2.2024.

Dolin, J., Black, P., Harlen, W. & Tiberghien, A. (2018) Exploring relations between formative and summative assessment. Teoksessa Dolin, J. & Evans, R. (toim.) *Transforming Assessment*. 53–80. Springer, Cham. https://doi org.ezproxy.utu.fi/10.1007/978-3-319-63248-3_3

Du, X., Cai, Y., Wang, S. & Zhang, L. (2016) *Overview of deep learning.* 159-164. https://doi-org.ezproxy.utu.fi/10.1109/YAC.2016.7804882

Elements of a promt (2024) Promt Engineering Guide. https://www.promptingguide.ai/introduction/elements 11.4.2024.

Enterprise privacy (2024) OpenAI. https://openai.com/enterprise-privacy. 14.4.2024.

Free Online Spreadsheet Software: Excel | Microsoft 365 (2024) Microsoft. https://www.microsoft.com/en-us/microsoft-365/excel 22.4.2024.

Giray, L. (2023) Prompt engineering with ChatGPT: a guide for academic writers. *Annals of Biomedical Engineering* 51(12) 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Goertzel, B. (2014) Artificial general Intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence* 5(1) 1–48. https://doi.org/10.2478/jagi-2014-0001

Hattie, J., & Timperley, H. (2007) The power of feedback. *Review of Educational Research* 77(1) 81–112. https://doi-org.ezproxy.utu.fi/10.3102/003465430298487

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I. & Ramkumar, P. N. (2020) Machine learning and artificial intelligence: definitions, applications, and future directions. *Current Reviews in Musculoskeletal Medicine* 13(1) 69–76. https://doi.org/10.1007/s12178-020-09600-8

How ChatGPT and Our Language Models Are Developed | OpenAI Help Center (2024) OpenAI. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed 19.2.2024.

How Microsoft 365 Copilot works (2024) Microsoft. https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/how-microsoft-365-copilot-works/ba-p/3822755 19.2.2024.

Hyvän vastauksen piirteet – Maantiede syksy 2023 (2023) Ylioppilastutkintolautakunta. https://tiedostot.ylioppilastutkinto.fi/kokeet/2023-09-21_GE_fi/grading-instructions.html 14.3.2024.

IBM SPSS Statistics (2024) IBM. https://www.ibm.com/products/spss-statistics 22.4.2024.

Imran, M., & Almusharraf, N. (2023) Analyzing the role of ChatGPT as a writing assistant at higher education level: a systematic review of the literature. *Contemporary Educational Technology* 15(4). https://doi.org/10.30935/cedtech/13605

Introducing ChatGPT (2024) OpenAI. https://openai.com/blog/chatgpt 12.2.2024.

Introducing ChatGPT Team (2024) OpenAI. https://openai.com/blog/introducing-chatgpt-team 14.4.2024.

Introducing GPTs (2024) OpenAI. https://openai.com/blog/introducing-gpts 14.4.2024.

Janiesch, C., Zschech, P. & Heinrich, K. (2021) Machine learning and deep learning. *Electronic Markets* 31(3) 685–695. https://doi.org/10.1007/s12525-021-00475-2

Jauhiainen, J. S. & Guerra, A. G. (2023) Generative AI and ChatGPT in school children's education: evidence from a school lesson. *Sustainability* 15(18). https://doi.org/10.3390/su151814025

Joshi, S., Rambola, R. K. & Churi, P. (2021) Evaluating artificial intelligence in education for next generation. *Journal of Physics: Conference Series* 1714(1) http://dx.doi.org/10.1088/1742-6596/1714/1/012039

Kaplan, A. (2021) Artificial intelligence (AI): when humans and machines might have to coexist. Teoksessa Verdegem, P. (toim.) *AI for Everyone? Critical Perspectives*. 21–32. University of Westminster Press. http://www.jstor.org/stable/j.ctv26qjjhj.4. 14.3.2024.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S. & Kazienko, P. (2023) ChatGPT: Jack of all trades, master of none. *Information Fusion* 2023(99). https://doi.org.ezproxy.utu.fi/10.1016/j.inffus.2023.101861

Korkeakoulujen yhteishaun opiskelijavalinnat (2024) Opetushallitus. https://opintopolku.fi/konfo/fi/sivu/korkeakoulujen-yhteishaun-opiskelijavalinnat#todistusvalinnan-pisteytykset 4.3.2024.

Korteling, J. E. (Hans)., Van De Boer-Visschedijk, G. C., Blankendaal, R. A. M., Boonekamp, R. C. & Eikelboom, A. R. (2021) Human- versus artificial intelligence. *Frontiers in Artificial Intelligence* 2021(4). https://doi.org/10.3389/frai.2021.622364

Laki viranomaisten toiminnan julkisuudesta 621/1999. Annettu Helsingissä 21.5.1991.

Laki ylioppilastutkinnosta 502/2019. Annettu Helsingissä 12.9.2019.

Lan, Y.-J., & Chen, N. S. (2024) Teachers' agency in the era of LLM and generative AI. *Educational Technology & Society* 27(1) 1-18. https://doi.org/10.30191/ETS.202401_27(1).PP01

Lappi, O., Rusanen, A. M. & Pekkanen, J. (2018) Tekoäly ja ihmiskognitio. *Tieteessä tapahtuu* 2018(1) 41–46. 5.12.2018. http://hdl.handle.net/10138/270785

Latif, E., & Zhai, X. (2024) Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence* 2024(6). https://doi-org.ezproxy.utu.fi/10.1016/j.caeai.2024.100210

Lee, M. (2023) A mathematical investigation of hallucination and creativity in GPT models. *Mathematics* 11(10). https://doi.org/10.3390/math11102320

Leydon, J., Wilson, K. & Boyd, C. (2014) Improving student writing abilities in geography: examining the benefits of criterion-based assessment and detailed feedback. *Journal of Geography* 113(4) 151–159. https://doi.org/10.1080/00221341.2013.869245

Liu, Z., Yi Xu, Xu, Y., Qian, Q., Li, H., Ji, X., Chan, A. B. & Jin, R. (2022) Improved fine-tuning by better leveraging pre-training data. https://doi.org/10.48550/arXiv.2111.12292

Lukion opetussuunnitelman perusteet (2019) Opetushallitus. https://www.oph.fi/sites/default/files/documents/lukion_opetussuunnitelman_perusteet_2019.pdf 22.2.2024.

Luostarinen, A. & Ouakrim-Soivio, N. (2019) Arvioinnin erilaiset tehtävät. Teoksessa (toim.) Luostarinen, A., Nieminen, J. H., Nilivaara, P., Ouakrim-Soivio, N., Peltomaa, I. M., Tuohilampi, L. & White, E. H. *Arvioinnin käsikirja*. Jyväskylä: PS-kustannus.

Luukka, M. R., Perälä, M. & Nylén, T. (2023) *Arviointiperusteiden kehittämistyön käynnistystä*. Esitelmä.

Maantieteen ylioppilaskoe syksy 2023 (2023) Ylioppilastutkintolautakunta. https://yle.fi/plus/abitreenit/2023/syksy/maantiede/index.html 22.2.2024.

Mahesh, B. (2020) Machine learning algorithms - a review. *International Journal of Science and Research* 9(1) 381-386. http://dx.doi.org/10.21275/ART20203995

Martinez, R. (2019) Artificial intelligence: distinguishing between types & definitions. *Newada Law Journal* 19(3) 1015–1042. https://scholars.law.unlv.edu/cgi/viewcontent.cgi?article=1799&context=nlj 22.2.2024.

Masikisiki, B., Marivate, V. & Hlope, Y. (2023) Investigating the efficacy of large language models in reflective assessment methods through chain of thoughts prompting. https://doi.org/10.48550/arXiv.2310.00272 11.4.2024.

Merchant, A., Rahimtoroghi, E., Pavlick, E. & Tenney, I. (2020) What happens to Bert embeddings during fine-tuning? https://doi.org/10.48550/arXiv.2004.14448 22.2.2024.

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R. & Gerardou, F. S. (2023) Challenges and opportunities of generative AI for higher education as explained by ChatGPT. *Education Sciences* 13(9) https://doi.org/10.3390/educsci13090856

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H. & Zettlemoyer, L. (2022) Rethinking the role of demonstrations: what makes in-context learning work? https://doi.org/10.48550/arXiv.2202.1283

Mizumoto, A. & Eguchi, M. (2023) Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics* 2(2). https://doi.org/10.1016/j.rmal.2023.100050

Müller, R., & Büttner, P. (1994) A critical discussion of intraclass correlation coefficients. *Statistics in Medicine* 13(23–24) 2465–2476. https://doi.org/10.1002/sim.4780132310

Newton, P. E. (2007) Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice* 14(2) 149–170. https://doi.org/10.1080/09695940701478321

Nieminen, J. H. (2019) Palaute osana arviointia. Teoksessa (toim.) Luostarinen, A., Nieminen, J. H., Nilivaara, P., Ouakrim-Soivio, N., Peltomaa, I. M., Tuohilampi, L. & White, E. H. *Arvioinnin käsikirja*. Jyväskylä: PS-kustannus.

OpenAI Platform (2024). OpenAI. https://platform.openai.com. 21.4.2024.

Passi, S., & M. Vorvoreanu (2022) Overreliance on AI literature review. https://www.microsoft.com/en-us/research/uploads/prod/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf 22.3.2024

Pereira, D. G., Afonso, A. & Medeiros, F. M. (2015) Overview of Friedman's test and post-hoc analysis. *Communications in Statistics - Simulation and Computation* 44(10) 2636–2653. https://doi.org/10.1080/03610918.2014.931971

Pohjonen, H. & Rissanen, M. (2021) *Arvioinnin tehtävät ja yleiset periaatteet.* Esitelmä. 13.4.2021, Yleissivistävä koulutus ja varhaiskasvatus -yksikkö.

Promt engineering (2024) OpenAI. https://platform.openai.com/docs/guides/prompt-engineering 11.4.2024.

Ragupathi, K., & Lee, A. (2020) Beyond fairness and consistency in grading: the role of rubrics in higher education. Teoksessa Sanger, C. S. & Gleason, N. W. (toim.) *Diversity and Inclusion in Global Higher Education: Lessons from Across Asia*. 73–95. Springer, Singapore.

Reaaliaineiden kokeiden määräykset ja ohjeet (2024) Ylioppilastutkintolautakunta. https://www.ylioppilastutkinto.fi/fi/tutkinnon-toimeenpano/maaraykset-ja-ohjeet/koekohtaiset-maaraykset-ja-ohjeet/reaaliaineiden 22.3.2024.

Rudolph, J., Tan, S. & Tan, S. (2023) ChatGPT: bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching* 6(1). https://doi.org/10.37074/jalt.2023.6.1.9

Sallam, M. (2023) ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* 11(6). https://doi.org/10.3390/healthcare11060887

Liao, S. H. (2005) Expert system methodologies and applications—a decade review from 1995 to 2004. *Expert Systems with Applications* 28(1) 93–103. https://doi-org.ezproxy.utu.fi/10.1016/j.eswa.2004.08.003

Steimers, A. & Schneider, M. (2022) Sources of risk of AI systems. *International Journal of Environmental Research and Public Health* 19(6). https://doi.org/10.3390/ijerph19063641

Su, J. & Yang, W. (2023) Unlocking the power of ChatGPT: a framework for applying generative AI in education. *ECNU Review of Education* 6(3) 355–366. https://doi.org/10.1177/20965311231168423

Tan, C. F., Wahidin, L. S., Khalil, S. N., Tamaldin, N., Hu, J. & Rauterberg, G. W. M. (2016) The application of expert system: a review of research and applications. *ARPN Journal of Engineering and Applied Sciences* 11(4) 2448-2453. https://pure.tue.nl/ws/portalfiles/portal/23537725/tanappli2016.pdf 22.2.2024

Tashu, T. M., Maurya, C. K. & Horvath, T. (2022) Deep learning architecture for automatic essay scoring. https://doi.org/10.48550/arXiv.2206.08232 22.3.2023

Telle, H., Postareff, L. & Virtanen, V. (2015) Millainen arviointi tukee elinikäistä oppimista? *Yliopistopedagogiikka* 2019(3). 27.3.2015. https://lehti.yliopistopedagogiikka.fi/2015/03/27/millainen-arviointi-tukee-elinikaista-oppimista/

Tobler, S. (2024) Smart grading: a generative AI-based tool for knowledge-grounded answer evaluation in educational assessments. *MethodsX* 2024(12). https://doi.org/10.1016/j.mex.2023.102531

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A. & Das, A. (2024) A comprehensive survey of hallucination mitigation techniques in large language models. https://doi.org/10.48550/arXiv.2401.01313 14.2.2024

Triguero, I., Molina, D., Poyatos, J., Del Ser, J. & Herrera, F. (2024) General purpose artificial intelligence systems (GPAIS): properties, definition, taxonomy, societal implications and responsible governance. *Information Fusion* (2024)103. https://doi-org.ezproxy.utu.fi/10.1016/j.inffus.2023.102135

Tripathi, K. (2011) A Review on knowledge-based expert system: concept and architecture. *IJCA Special Issue on Artificial Intelligence Techniques-Novel Approaches & Practical Applications.* (2011)4. https://www.researchgate.net/publication/266013987_A_Review_on_Knowledge-based_Expert_System_Concept_and_Architecture 22.3.2024

Tuomi, J. & A. Sarajärvi (2018) *Laadullinen tutkimus ja sisällönanalyysi*. Kustannusosakeyhtiö Tammi, Helsinki.

Tähtinen, J., Laakkonen, E. & Broberg M. (2020) Tilastollisen aineiston käsittelyn ja tulkinnan perusteita. *Turun yliopiston kasvatustieteiden tiedekunnan julkaisuja*. Turun yliopiston kasvatustieteiden laitos, Turku.

Vall, A., & Widmer, G. (2018) Machine learning approaches to hybrid music recommender systems. https://doi.org/10.48550/arXiv.1807.05858 22.3.2024.

Valtioneuvoston asetus ylioppilastutkinnosta 612/2019. Annettu Helsingissä 9.5.2019.

Wang, P. (2019) On defining artificial intelligence. *Journal of Artificial General Intelligence* 10(2) 1–37. https://doi.org/10.2478/jagi-2019-0002

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M. & Le, Q. V. (2022) Finetuned language models are zero-shot learners. https://doi.org/10.48550/arXiv.2109.01652 22.3.2023.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. & Zhou, D. (2023) Chain-of-thought prompting elicits reasoning in large language models. https://doi.org/10.48550/arXiv.2201.11903

What is an AI model? (2024) IBM. https://www.ibm.com/topics/ai-model 22.2.2024.

Wiliam, D. (2011) What is assessment for learning? *Studies in Educational Evaluation* 37(1). Assessment for Learning 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001

Wilk Test - an overview | ScienceDirect Topics (2024) ScienceDirect. https://www.sciencedirect.com/topics/mathematics/wilk-test 22.4.2024.

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L. & Tang, Y. (2023) A brief overview of ChatGPT: the history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica* 10(5) 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Zhang, C., & Lu, Y. (2021) Study on artificial intelligence: the state of the art and future prospects. *Journal of Industrial Information Integration* 2023(23). https://doi.org/10.1016/j.jii.2021.100224

Liitteet

Liite 1. Chattibottien luomisessa käytetyt asetukset

Syötetty tieto	Muokattava asetus
Botti 1	Nimi
Arvioi kokelaiden vastauksia kriteeriperusteisesti	Kuvaus
Olet opettaja, joka arvioi kokelaiden vastauksia annettujen kriteerien mukaisesti. Arviointikriteerit ovat tiedostossa ***HVP.pdf*. Pisteytä oppilaiden vastaukset arviointikriteerien ohjeiden mukaisesti. ###ÄLÄ MUUTA ALKUPERÄISTÄ VASTAUSTA ###ILMOITA ARVIOINNIN TULOKSENA SYNTYNYT VASTAUKSEN KOKONAISPISTEMÄÄRÄ JA PERUSTELU ARVIOINNISTA** **Esimerkki keskustelusta:** **Käyttäjän syöte:** arvioitava vastaus **Järjestelmän tuloste:** kokonaispistemäärä: Arvioinnin tuloksena syntynyt kokonaispistemäärä tehtävästä ja perustelu arvioinnista.	Toimintaohjeet
HVP.pdf	Tietokanta
verkkoselaus, DALL-E kuvagenerointi, kooditulkki	Toiminnalliset lisäominaisuudet

Syötetty tieto	Muokattava asetus
Botti 2	Nimi
Arvioi kokelaiden vastauksia kriteeriperusteisesti ilman perusteluja	Kuvaus
Olet opettaja, joka arvioi kokelaiden vastauksia annettujen kriteerien mukaisesti. Tutustu ensin tiedostoon ***yleiset arviointiohjeet_1.pdf* ja noudata arviointia ohjeiden mukaisesti. ###ÄLÄ MUUTA ALKUPERÄISTÄ VASTAUSTA ###ILMOITA VAIN ARVIOINNIN TULOKSENA SYNTYNYT VASTAUKSEN KOKONAISPISTEMÄÄRÄ ilman perustelua arvioinnista.** **Esimerkki keskustelusta:** **Käyttäjän syöte: arvioitava vastaus** Järjestelmän tuloste: kokonaispistemäärä: Arvioinnin tuloksena syntynyt kokonaispistemäärä tehtävästä.	Toimintaohjeet
arviointimatriisi_konvektiosade.pdf arviointimatriisi _rintamasade.pdf arviointimatriisi_vuoristosade.pdf yleiset arviointiohjeet_1.pdf	Tietokanta
verkkoselaus, DALL-E kuvagenerointi, kooditulkki	Toiminnalliset lisäominaisuudet

Syötetty tieto	Muokattava asetus
Botti 3	Nimi
Arvioi kokelaiden vastauksia kriteeriperusteisesti ilman perusteluja	Kuvaus
Olet opettaja, joka arvioi kokelaiden vastauksia annettujen kriteerien mukaisesti. Tutustu ensin tiedostoon ***yleiset arviointiohjeet_2.pdf* ja noudata arviointia ohjeiden mukaisesti. ###ÄLÄ MUUTA ALKUPERÄISTÄ VASTAUSTA ###ILMOITA VAIN ARVIOINNIN TULOKSENA SYNTYNYT VASTAUKSEN KOKONAISPISTEMÄÄRÄ ilman perustelua arvioinnista.** **Esimerkki keskustelusta:** **Käyttäjän syöte: arvioitava vastaus** Järjestelmän tuloste: kokonaispistemäärä: Arvioinnin tuloksena syntynyt kokonaispistemäärä tehtävästä.	Toimintaohjeet
arviointimatriisi_konvektiosade.pdf arviointimatriisi _rintamasade.pdf arviointimatriisi_vuoristosade.pdf yleiset arviointiohjeet_2.pdf 1.jpeg, 2.jpeg, 3.jpeg, 4.jpeg	Tietokanta
verkkoselaus, DALL-E kuvagenerointi, kooditulkki	Toiminnalliset lisäominaisuudet

Syötetty tieto	Muokattava asetus
Botti 4	Nimi
Arvioi kokelaiden vastauksia kriteeriperusteisesti ilman perusteluja	Kuvaus
Olet opettaja, joka arvioi kokelaiden vastauksia annettujen kriteerien mukaisesti. Tutustu ensin tiedostoon ***yleiset arviointiohjeet_3.pdf* ja noudata arviointia ohjeiden mukaisesti. ###ÄLÄ MUUTA ALKUPERÄISTÄ VASTAUSTA ###ILMOITA ARVIOINNIN TULOKSENA SYNTYNYT VASTAUKSEN KOKONAISPISTEMÄÄRÄ JA PERUSTELU ARVIOINNISTA.** **Esimerkki keskustelusta:** **Käyttäjän syöte: arvioitava vastaus** Järjestelmän tuloste: kokonaispistemäärä: Arvioinnin tuloksena syntynyt kokonaispistemäärä tehtävästä ja perustelu arvioinnista.	Toimintaohjeet
arviointimatriisi_konvektiosade.pdf arviointimatriisi _rintamasade.pdf arviointimatriisi_vuoristosade.pdf yleiset arviointiohjeet_3.pdf 1.jpeg, 2.jpeg, 3.jpeg, 4.jpeg esimerkkivastauksia.pdf	Tietokanta
verkkoselaus, DALL-E kuvagenerointi, kooditulkki	Toiminnalliset lisäominaisuudet

Liite 2. Tarkemmat arviointiohjeet rintamasateen, konvektiosateen ja vuoristosateen arvioimiseksi

pisteluokat ja arviointikohteet	arviointikriteerit	esimerkkivastaus	perustelu
nimeäminen 0 p	käsite on nimetty väärin tai sen kirjoitusmuoto on virheellinen	rannikkosade, polaaririntaman sade, monsuunisade, räntäsade, vesisade, lumisade, jäänsade, rankkasade, rintama sade,	Käsitteen kirjoitusasu on virheellinen ja käsite on nimetty väärin. Käsite ei vastaa tehtävänantoa, joten siitä ei voi antaa pisteitä.
nimeäminen 1 p	Käsite on nimetty oikein ja sen kirjoitusasu on moitteeton.	rintamasade, rintamasateet, rintamasateita	Käsitteen kirjoitusasu moitteeton ja käsite vastaa tehtävänantoa.
syntytavan kuvaus 0 p	Vastaus on puutteellinen ja epäselkeä. Siitä ei käy ilmi kylmän ja lämpimän rintaman kohtaamista. Siinä ei myöskään kuvata, mitä rintamien kohdatessa tapahtuu ja miten sade muodostuu.	matalapainerintama törmää korkeapainerintamaan, josta muodostuu sadetta.	Vastaus on epäselvä, eikä siinä käytetä oikeita käsitteitä. Ei kuvausta kylmän ja lämpimän rintaman kohtaamisesta ja siitä, miten sade syntyy.
syntytavan kuvaus 1 p	Vastaus on pinnallinen ja siinä mainitaan vain kylmän ja lämpimän rintaman kohtaamisesta. Tarkempi kuvaus siitä, mitä lämpimän ilman kohotessa kylmemmän ilman päälle puuttuu.	Lämmin ja kylmä rintama kohtaavat. Kylmä rintama pakottaa lämpimän ilman kohoamaan ja siinä oleva kosteus tiivistyy sateiksi.	Vastauksessa on kuvattu lämpimän ja kylmän rintaman kohtaaminen. Lisäksi siinä mainitaan lämpimän ilman kohoamisesta, joka johtuu kylmän ilman painuessa lämpimän ilman alle. Vastauksesta puuttuu kuitenkin tarkempi kuvaus sadepisarojen syntymisestä ja sateen tyypistä (kestosta ja määrästä). Lisäksi siinä ei käy ilmi, että rintamat ovat liikkeessä.
syntytavan kuvaus 2 p	Vastauksessa kuvataan, että lämmin ja kylmä rintama kohtaavat. Tämä saa aikaan lämpimän ilman kohoamisen kylmän ilman painuessa sen alle. Lämmennyt ilma kohoaa ja siinä oleva kosteus tiivisty vesipisaroiksi muodostaen sateita. Ei mainintaa siitä, että rintamat liikkuvat tai sateen kestosta sekä määrästä	Rintamasateet syntyvät, kun kylmä ja lämmin rintama törmäävät. Kylmä rintama pakottaa lämpimän rintaman ilman kohoamaan, jolloin siinä oleva kosteus tiivistyy vesipisaroiksi muodostaen sateita.	Vastauksessa kuvataan, miten lämpimän ja kylmän rintaman kohtaaminen saa aikaan lämpimän ilman kohoamisen. Lisäksi siinä kerrotaan, että lämpimän ilman kohotessa sen sisältämä kosteus tiivistyy vesipisaroiksi, joista muodostuu sadetta. Vastauksesta puuttuu kuitenkin tarkempi kuvaus sateiden ominaisuuksista ja siitä, että rintamat ovat liikkeessä.
syntytavan kuvaus 3 p	Vastauksessa kuvataan selkeästi, että lämmin ja kylmä rintama liikkuvat ja kohtaavat. Tämä saa lämpimässä rintamassa olevan lämpimän ilman kohoamaan kylmän rintaman sisältävän kylmän ilman painuessa raskaampana lämpimän ilman alle. Lämpimän ilman kohotessa se jäähtyy ja siinä oleva kosteus tiivistyy vesipisaroiksi muodostaen sadepilviä ja rintamasateita.	Polaaririntamassa liikkuvien matalapaineiden yhteydessä esiintyy rintamasateita. Ne syntyvät, kun kylmän rintama kohtaa lämpimän rintaman. Kylmän rintaman ilma on raskaampaa ja pakottaa lämpimän rintaman ilman kohoamaan sen päälle. Kohonnut lämpimämpi ilma viilenee ja siinä oleva kosteus tiivistyy vesipisaroiksi muodostaen sadepilviä. Pilvet satavat alas rintamasateina, jotka voivat olla tyypiltään lyhytkestoisia ja rankkoja sekä pitkäkestoisia tihkusateita.	Vastauksesta kuvataan selkeästi, että kylmä ja lämmin rintama liikkuvat, joka saa aikaan niiden kohtaamisen. Rintamien kohdatessa on selkeä selitys siitä, miksi lämmin ilma kohoaa ja miten sade muodostuu.
esimerkkialue 0 p	Esimerkkialue ei ole liikkuvien matalapaineiden alueella tai siellä se ei ole alueella yleisin sadetyyppi. Maantieteellisestä sijainnista ei voi määrittää, onko esimerkkialue liikkuvien matalapaineiden alueella.	Vuoristoissa, meren äärellä, mantereella, sademetsässä,	Esimerkkialue ei kuuluu liikkuvien matalapaiden alueeseen. Esimerkistä ei voi päätellä maantieteellistä sijaintia. Esimerkin alueella rintamasade ei ole yleisin sadetyyppi.
esimerkkialue 1 p	Esimerkkialueesta käy ilmi, että se on liikkuvien matalapaineiden alueella, jossa rintamasateet ovat yleisin sadetyyppi. Maininta Polaaririntamasta on riittävä, mutta esimerkkivaltion nimeäminen, joka sijaistee liikkuvien matalapaineiden alueella hyväksytään myös.	Suomessa, Suomi, Saksassa, Polaaririntamalla, lähellä polaaririntamaa, Keski-Euroopassa, Iso-Britannia, Venäjä, Polaaririntaman alueella	Esimerkkialue kuuluu liikkuvien matalapaiden alueeseen ja on siellä yleisin sadetyyppi.

pisteluokat ja arviointikohteet	arviointikriteerit	esimerkkivastaus	perustelu
nimeäminen 0 p	Käsite on nimetty väärin tai sen kirjoitusmuoto on virheellinen.	konvektio sade, päiväntasaajan sade, monsuunisade, rankkasade, lumisade, vesisade, räntäsade, konventiosade, kohdensiosade,	Käsitteen kirjoitusasu on virheellinen ja käsite on nimetty väärin. Käsite ei vastaa tehtävänantoa, joten siitä ei voi antaa pisteitä.
nimeäminen 1 p	Käsite on nimetty oikein ja sen kirjoitusasu on moitteeton.	konvektiosade, konvektiosateet, konvektiosateita	Käsitteen kirjoitusasu moitteeton ja käsite vastaa tehtävänantoa.
syntytavan kuvaus 0 p	Vastaus on puutteellinen ja epäselvä. Siinä ei kuvata prosessia, kuinka Aurinko lämmittää ilmaa ja maanpintaa, joka saa kostean ilman kohoamaan. Ei mainintaa ilman jäähtymisestä ja kosteuden tiivistymisestä, jonka takia muodostuu pilviä.	Maassa oleva vesi alkaa nousemaan auringon lämmittäessä sitä. Vesi nousee taivaalle ja sataa alas, kun pilvet ovat tarpeeksi isoja.	Vastauksessa kuvataan virheellisesti, että pelkkä vesi kohoaa maanpinnalta. Oikeassa vastauksessa lämmennyt ilma kohoaa, joka sisältää ympäröivästä alueesta ilmaan sitoutunutta kosteutta. Ei mainintaa veden muuttumisesta vesihöyryksi ja sen tiivistymistä tiivistymisytimien ympärille ilman jäähtyessä. Pilvien muodostumisesta ei ole kuvasta. Ei mainintaa lyhytkestoisista rankkasateista.
syntytavan kuvaus 1 p	Vastaus on puutteellinen, eikä se kuvaa prosessia selkeästi. Vastauksessa ei käy ilmi, että lämmennyt ilma kohoaa Auringon vaikutuksesta. Lisäksi vahtauksessa ei selitetä ilman kohoamisen seurausta, joka saa aikaan sen viilenemisen ja vesihöyryn tiivistymisen.	Aurinko saa veden haihtumaan ilmaan ja se tiivistyy pilviksi, jotka muodostavat sadetta. Sateet ovat tyypillisesti rankkoja iltapäivisin.	Vastauksessa mainitaan, että Aurinko haihduttaa vettä, mutta tarkempi kuvaus puuttuu siitä, miten lämmennyt ilma sitoo kosteutta ympäristöstä, kohoaa ja jäähtyy. Lisäksi ei mainintaa siitä, että ilmassa oleva kosteus tiivistyy sadepisaroiksi.
syntytavan kuvaus 2 p	Vastauksessa kuvataan, miten Aurinko lämmittää maata ja sen päällä olevaa ilmaa, joka alkaa kohoamaan lämpenemisen vaikutuksesta. Ilman kohoamisesta selitetään, että se jäähtyy ja siinä oleva vesihöyry tiivistyy pisaroiksi, josta muodostaa pilviä. Ei mainintaa sateen ajankohdasta, kestosta täi määrästä.	Auringon vaikutuksesta maanpinnan päällä oleva kostea ilma alkaa kohoamaan. Kohotessaan se jäähtyy ja siinä oleva vesi tiivistyy vesipisaroiksi. Vesipisarat muodostavat pilviä, jotka aiheuttavat sateita.	Vastauksessa on kuvattu, että Auringon lämmittävä vaikutus saa kostean ilman kohoamaan. Lisäksi siinä selitetään ilman viileneminen ja sen sisältämän kosteuden tiivistyminen vesipisaroiksi, joka saa aikaan pilviä. Mainninta sateen kestosta ja määrästä puuttuu.
syntytavan kuvaus 3 p	Vastauksessa on kuvattu selkeästi, että prosessi saa alkunsa Auringon lämmittävästä vaikutuksesta, joka saa maanpinnan ja sen yläpuolella olevan kostean ilman lämpenemään, jonka seurauksesta se kohoaa. Vastauksessa selitetään, että kohotessaan ilma jäähtyy ja sen sisältämä vesihöyry tiivistyy sadepisaroiksi muodostaen pilviä. Lisäksi vastauksessa mainitaan, että pilvet ovat yleensä ukkos- tai kumpupilviä, jotka saavat aikaan rankkoja, mutta lyhyitä sateita tyypillisesti iltapäivisin.	Konvektiosateen synty saa alkunsa Auringon lämmittävästä vaikutuksesta, joka saa kostean ilman lämpenemään ja kohoamaan ylöspäin. Kohoava ilma jäähtyy ja siinä oleva kosteus tiivistyy vesipisaroiksi muodostaen pilviä. Iltapäivään mennessä pilvet ovat usein kooltaan isoja kumpu- ja ukkospilviä, jotka satavat alas lyhytkestoisina ja rankkoina sateita.	Vahtauksessa on kuvattu selkeästi, että Auringon lämmittävä vaikutus saa maanpinnan ja sen yläpuolella olevan ilman lämpenemään. Lisäksi mainінta, että ilma on kosteaa ja kohotessaan siinä oleva vesihöyry tiivistyy vesipisaroiksi. Maininta pilvien syntymisestä. Lisäksi vastauksessa mainitaan, että konvektiosateet ovat tyypillisiä iltapäivisin rankkoina ja lyhyinä sateina.
esimerkkialue 0 p	Esimerkkialue on liian laaja tai siellä konvektiosateet eivät ole tyypillisiä ja usein toistuvia.	Suomessa, vuoristoilla, Afrikassa, Keski-Euroopassa, Kanadassa, Vuorilla	Esimerkkialue ei ole Päiväntasaajan alueella tai trooppisella vyöhykkeellä. Suomi ei ole hyväksyttävä vastaus, sillä konvektiosateita esiintyy Suomessa tyypillisesti vain kesäisin tai hellepäivinä, joka pitää mainita vastauksessa erikseen. Esimerkkialue on liian laaja, kuten Afrikka maanosana kattaa useille sateille tyypillisiä alueita.
esimerkkialue 1 p	Esimerkkialue sijaitsee Päiväntasaajan lähellä tropiikissa, jossa konvektiosateet ovat yleisin sadetyyppi. Maininta Päiväntasaajasta, tropiikista ja sitä ympäröivästä alueesta riittää. Myös yksittäisen valtion nimeäminen on hyväksyttävä vastaus, jos se sijaitsee konvektiosateille tyypillisellä alueella. Poikkeuksena hyväksytään myös Suomi, mutta vain kesäisin tai kuumina hellepäivinä.	Päiväntasaajalla, Päiväntasaajan alueella, trooppisella vyöhykkeellä, trooppisella alueella, kesäisin Suomessa, hellepäivänä Suomessa, kesäpäivinä Suomessa, kesällä Suomessa	Esimerkkialue sijaitsee Päiväntasaajan alueella, jossa konvektiosateet ovat yleisiä ympärivuotisen voimakkaan lämpösäteilyn seurauksesta, joka haihduttaa vettä ympäristöstä runsaasti. Myös maininta trooppisista alueista riittää tai maininta esimerkkivaltiosta, joka sijaitsee lähellä Päiväntasaajaa. Suomi on tässä tapauksessa hyväksyttävä vastaus, jos siinä mainitaan konvektiosateiden esiintyminen erityisesti kesäisin tai hellepäivinä.

pisteluokat ja arviointikohteet	arviointikriteerit	esimerkkivastaus	perustelu
nimeäminen 0 p	käsite on nimetty väärin tai sen kirjoitusmuoto on virheellinen	orografiasade, orogaafinensade, vuoristo sade, vuoristojen sateet, vuorisateet, räntäsade, lumisade, vesisade, rankkasade, oreografinen sade, orografiiniset sateet	Käsitteen kirjoitusasu on virheellinen ja käsite on nimetty väärin. Käsite ei vastaa tehtävänantoa, joten siitä ei voi antaa pisteitä.
nimeäminen 1 p	Käsite on nimetty oikein ja sen kirjoitusasu on moitteeton.	vuoristosade, vuoristosateet, vuoristosateita, orografiset sateet, orografisia sateita	Käsitteen kirjoitusasu moitteeton ja käsite vastaa tehtävänantoa.
syntytavan kuvaus 0 p	Vastauksessa syntytavan kuvaus on epäselvä ja puutteellinen. Ei mainintaa mereltä tulevasta kosteasta ilmasta eikä siitä mitä kostealle ilmalle tapahtuu vuoren rinteellä. Ei mainintaa lämpimästä laskutuulesta vuoren toisella puolella.	Syntyy kun pilvi törmää vuoristoon ja joutuu kiipeämään vuoren yli aiheuttaen sateen.	Vastaus on epäselvä, eikä siinä kuvata sateen syntyä prosessina. Ei mainintaa siitä, miten kostea ilmassa törmää tuulen mukana vuoreen, joka pakottaa sen kohoamaan. Lisäksi ei kuvausta siitä, mitä ilmamassa kokoamisen aikana tapahtuu tai mainintaa lämpimästä laskutuulesta.
syntytavan kuvaus 1 p	Vastaus on puutteellinen ja pinnallinen. Siinä ei ole mainintaa ilman kosteudesta, joka peräisin merestä. Lisäksi vastauksesta puuttuu kuvaus siitä, mitä ilmalle tapahtuu sen kohotessa ylöspäin vuorenrinteen vaikutuksesta. Ei myöskään mainintaa millä puolella vuorta sataa ja millä ei.	Merestä tuleva ilma kohtaa vuoriston ja alkaa kohoamaan. Siinä oleva vesihöyry tiivistyy sateeksi vuoren rinteelle.	Vastauksessa mainitaan meri-ilman kohtaamisesta vuoren kanssa, joka saa sen kohoamaan. Siinä ei kuitenkaan mainita meri-ilman sisältämästä kosteudesta, joka on peräisin merestä. Lisäksi vastauksessa ei kuvailla, mitä ilmalle tapahtuu, kun se kohoaa ylöspäin. Ei myöskään mainintaa millä puolella rinnettä sataa ja millä ei.
syntytavan kuvaus 2 p	Vastauksessa kuvataan selkeästi kostea meri-ilman vaikutuksesta vuoristosateen syntyyn. Lisäksi siinä selitetään, kuinka kostea ilma törmää vuoristoon ja saa sen kohoamaan. Tämä saa aikaan ilman viilenemisen ja siinä olevan kosteuden tiivistymään vesipisaroiksi aiheuttaen sadetta vuoren rinteen mereiselle puolelle. Vastauksesta kuitenkin puuttuu lämpimän ja kuivan laskutuulen mainintа vuoren rinteen vastakkaisella puolella.	Syntyy, kun kostea meri-ilma törmää vuoristoon. Kohdatessaan vuoriston ilmamassa kohoaa ylöspäin vuoren vaikutuksesta ja viilenee. Samalla siinä oleva kosteus tiivistyy vesipisaroiksi muodostaen sadetta vuoren merenpuoleiselle rinteelle.	Vastauksessa on kuvattu selkeästi vuoristosateen syntytapa. Siinä mainitaan kostean meri-ilman kohoamisesta vuoriston vaikutuksesta. Lisäksi siinä selitetään, mikä vaikutus ilman kohoamisella on sateen synnyn kannalta. Vastauksesta kuitenkin puuttuu mainintа kuivasta ja lämpimästä laskutuulesta, joka ilmenee vuoren mantereisen rinteen puolella.
syntytavan kuvaus 3 p	Vastauksessa on kuvattu selkeästi vuoristosateen synty kosteasta ilmamassasta, joka on peräisin mereltä. Siinä selitetään ymmärrettävästi, sateen syntyminen vuoren mereiselle puolelle ilman kohotessa ja jäähtyessä. Lisäksi vastauksessa mainitaan vuoren vastakkaisella puolella tapahtuvasta lämpiävästä ja kuivasta laskutuulesta.	Vuoristosateet syntyvät, kun mereltä tuleva kostea ilmamassa kohtaa vuoriston tai rannikon, joka saa sen kohoamaan. Ilmamassan kohotessa se viilenee ja sen sisältämä kosteus tiivistyy vesipisaroiksi muodostaen sadetta vuoriston merenpuoleisille rinteille. Tämän sorauksesta vuoriston ylittänyt ilmamassa lämpenee ja ilmenee vuoren mantereisella puolella kuivana ja lämpimänä laskutuulena.	Vastauksessa on selitetty kattavasti ja ymmärrettävästi, miten vuoristosade syntyy. Siinä mainitaan kosteasta ilmamassasta, joka on peräisin mereltä. Siinä myös kuvataan, miten sade muodostuu vuoriston mereisen rinteen puolella. Lisäksi on maininta vuoren rinteen vastakkaisella puolella ilmenevästä lämpimästä laskutuulesta.
esimerkkialue 0 p	Esimerkkialue ei ole tyypillinen vuoristosateilla. Esimerkkialueesta ei voi päätellä maantieteellistä sijaintia.	Suomi, Viro, Alankomaat, vuoristoalueilla, vuorilla, vuoristossa, vuoristoisilla alueilla	Vastauksessa nimetyt valtiot eivät edusta vuoristosateiden tyypillistä esiintymisaluetta. Lisäksi pelkkä ilmaisu "vuorilla" tai "vuoristoisilla alueilla" ei kelpaa, sillä sen maantieteellistä tarkkaa sijaintia on mahdotonta määrittää.
esimerkkialue 1 p	Esimerkkialue kuuluu vuoristosateiden tyypilliseen alueeseen. Esimerkkialueesta voi päätellä maantieteellisen sijainnin vuoriston nimen tai valtion perusteella.	Norjan länsirannikko, Kanadan länsirannikko. Himalajan vuoristo, Skandit, Alpit, Norjan länsiosa	Vastauksesta voi päätellä maantieteellisen sijainnin ja se on tyypillinen esiintymisalue vuoristosateille.

Liite 3. Chattiboteille 2, 3 ja 4 syötetyt yleiset arviointiohjeet

Vastausten arvioinnin yleisiä ohjeita (botti 2)

Pyri arvioinnissa systemaattisuuteen. Arvioi vain sitä, mitä kuuluu arvioida annetuilla kriteereillä. Tämä ohje sisältää yleisiä ohjeita tehtävän arviointiin ja pisteytykseen. Tarkemmat ohjeet kunkin sadetyypin tarkemmasta pisteytyksestä ovat erillisissä pdf-tiedostoissa taulukkoina. Tutustu niihin huolella ennen arvioinnin aloittamista.

Tehtävänanto:

Vastaukset käsittelevät tehtävää, jossa kokelaiden on nimettävä kolme sadetyyppiä, kuvailla niiden syntyä ja mainita yksi kyseiselle sadetyypille ominainen esiintymisalue. Tehtävässä mitataan kokelaiden maantieteellistä ajattelua sateiden synnyn osalta.
Tehtävä pisteytetään niin, että sadetyypin nimeämisestä saa 0-1p, syntytavan kuvauksesta 0-3p ja esimerkkialueen maininnasta 0-1p. Tehtävän maksimipisteet ovat 15p. Tehtävästä voi saada vain kokonaisia pisteitä ei puolikkaita.

Tehtävässä hyväksyttävät sadetyypit ovat:

- Vuoristosade / orografinen sade
- Rintamasade
- Konvektiosade

Muista sadetyypeistä ei anneta pisteitä! Arvioinnissa kiinnitetään huomiota käsitteiden oikeaoppiseen kirjoittamiseen, sateen syntytavan kuvauksen laatuun sekä esimerkkialueen mainintaan.

Tarkemmat ohjeet eri sadetyyppien arvioimiseksi ovat tiedostoissa:

arviointimatriisi_vuoristosade.pdf

arviointimatriisi_rintamasade.pdf

arviointimatriisi_konvektiosade.pdf

Vastausten arvioinnin yleisiä ohjeita (botti 3)

Pyri arvioinnissa systemaattisuuteen. Arvioi vain sitä, mitä kuuluu arvioida annetuilla kriteereillä. Tämä ohje sisältää yleisiä ohjeita tehtävän arviointiin ja pisteytykseen. Tarkemmat ohjeet kunkin sadetyypin tarkemmasta pisteytyksestä ovat erillisissä pdf-tiedostoissa taulukkoina. Tutustu niihin huolella ennen arvioinnin aloittamista.

Tehtävänanto:

Vastaukset käsittelevät tehtävää, jossa kokelaiden on nimettävä kolme sadetyyppiä, kuvailla niiden syntyä ja mainita yksi kyseiselle sadetyypille ominainen esiintymisalue. Tehtävässä mitataan kokelaiden maantieteellistä ajattelua sateiden synnyn osalta. Kokelaiden opetusmateriaalina ovat toimineet valokuvat ***1.jpeg***, ***2.jpeg***, ***3.jpeg*** ja ***4.jpeg*** Tutustu näihin huolella, jotta ymmärrät paremmin, miten arvotat kokelaan ymmärrystä aiheesta.

Tehtävä pisteytetään niin, että sadetyypin nimeämisestä saa 0-1p, syntytavan kuvauksesta 0-3p ja esimerkkialueen maininnasta 0-1p. Tehtävän maksimipisteet ovat 15p. Tehtävästä voi saada vain kokonaisia pisteitä ei puolikkaita.

Tehtävässä hyväksyttävät sadetyypit ovat:

- Vuoristosade / orografinen sade
- Rintamasade
- Konvektiosade

Muista sadetyypeistä ei anneta pisteitä! Arvioinnissa kiinnitetään huomiota käsitteiden oikeaoppiseen kirjoittamiseen, sateen syntytavan kuvauksen laatuun sekä esimerkkialueen mainintaan.

Tarkemmat ohjeet eri sadetyyppien arvioimiseksi ovat tiedostoissa:

arviointimatriisi_vuoristosade.pdf

arviointimatriisi_rintamasade.pdf

arviointimatriisi_konvektiosade.pdf

Vastausten arvioinnin yleisiä ohjeita (botti 4)

Pyri arvioinnissa systemaattisuuteen. Arvioi vain sitä, mitä kuuluu arvioida annetuilla kriteereillä. Tämä ohje sisältää yleisiä ohjeita tehtävän arviointiin ja pisteytykseen. Tarkemmat ohjeet kunkin sadetyypin tarkemmasta pisteytyksestä ovat erillisissä pdf-tiedostoissa taulukkoina. Tutustu niihin huolella ennen arvioinnin aloittamista.

Tehtävänanto:

Vastaukset käsittelevät tehtävää, jossa kokelaiden on nimettävä kolme sadetyyppiä, kuvailla niiden syntyä ja mainita yksi kyseiselle sadetyypille ominainen esiintymisalue. Tehtävässä mitataan kokelaiden maantieteellistä ajattelua sateiden synnyn osalta. Kokelaiden opetusmateriaalina ovat toimineet valokuvat ***1.jpeg***, ***2.jpeg***, ***3.jpeg*** ja ***4.jpeg*** Tutustu näihin huolella, jotta ymmärrät paremmin, miten arvotat kokelaan ymmärrystä aiheesta.

Tehtävä pisteytetään niin, että sadetyypin nimeämisestä saa 0-1p, syntytavan kuvauksesta 0-3p ja esimerkkialueen maininnasta 0-1p. Tehtävän maksimipisteet ovat 15p. Tehtävästä voi saada vain kokonaisia pisteitä ei puolikkaita.

Tehtävässä hyväksyttävät sadetyypit ovat:

- Vuoristosade / orografinen sade
- Rintamasade
- Konvektiosade

Muista sadetyypeistä ei anneta pisteitä! Arvioinnissa kiinnitetään huomiota käsitteiden oikeaoppiseen kirjoittamiseen, sateen syntytavan kuvauksen laatuun sekä esimerkkialueen mainintaan.

Tarkemmat ohjeet eri sadetyyppien arvioimiseksi ovat tiedostoissa:

arviointimatriisi_vuoristosade.pdf

arviointimatriisi_rintamasade.pdf

arviointimatriisi_konvektiosade.pdf

Katso esimerkkejä jokaisen pisteluokan vastauksesta, jotta osaat suorittaa arviointia paremmin. Esimerkit eri pisteluokan vastauksista ovat tiedostossa: ***esimerkkivastauksia.pdf***

Liite 4. Taulukot chattibottien tarkkuudesta vastausten pisteyttämisessä

Tarkkuus pisteyttää vastauksia täysin yhdenmukaisesti sensorien kanssa				
Pisteluokka	**Malli 1**	**Malli 2**	**Malli 3**	**Malli 4**
0	33,3 %	100,0 %	96,7 %	83,3 %
1	0,0 %	13,3 %	3,3 %	3,3 %
2	0,0 %	6,7 %	16,7 %	3,3 %
3	0,0 %	6,7 %	3,3 %	0,0 %
4	0,0 %	20,0 %	3,3 %	6,7 %
5	0,0 %	0,0 %	3,3 %	0,0 %
6	0,0 %	0,0 %	0,0 %	10,0 %
7	0,0 %	10,0 %	10,0 %	13,3 %
8	0,0 %	3,3 %	0,0 %	0,0 %
9	0,0 %	3,3 %	3,3 %	10,0 %
10	0,0 %	0,0 %	10,0 %	6,7 %
11	0,0 %	20,0 %	16,7 %	13,3 %
12	0,0 %	26,7 %	23,3 %	33,3 %
13	16,7 %	3,3 %	23,3 %	3,3 %
14	0,0 %	6,7 %	0,0 %	0,0 %
15	100,0 %	66,7 %	40,0 %	43,3 %
Keskiarvo	**9,4 %**	**17,9 %**	**15,8 %**	**14,4 %**

Tarkkuus pisteyttää vastauksia yhden pisteen erolla sensorien antamista pisteistä				
Pisteluokka	**Malli 1**	**Malli 2**	**Malli 3**	**Malli 4**
0	33,3 %	100,0 %	96,7 %	83,3 %
1	10,0 %	43,3 %	30,0 %	30,0 %
2	0,0 %	36,7 %	40,0 %	33,3 %
3	0,0 %	26,7 %	3,3 %	6,7 %
4	0,0 %	46,7 %	20,0 %	40,0 %
5	3,3 %	3,3 %	10,0 %	23,3 %
6	0,0 %	0,0 %	6,7 %	16,7 %
7	0,0 %	23,3 %	40,0 %	46,7 %
8	0,0 %	10,0 %	6,7 %	46,7 %
9	0,0 %	6,7 %	16,7 %	16,7 %
10	0,0 %	10,0 %	33,3 %	43,3 %
11	0,0 %	60,0 %	43,3 %	53,3 %
12	0,0 %	36,7 %	50,0 %	43,3 %
13	20,0 %	23,3 %	36,7 %	23,3 %
14	100,0 %	66,7 %	53,3 %	50,0 %
15	100,0 %	70,0 %	43,3 %	43,3 %
Keskiarvo	**16,7 %**	**35,2 %**	**33,1 %**	**37,5 %**

Tarkkuus pisteyttää vastauksia kahden pisteen erolla sensorien antamista pisteistä				
Pisteluokka	**Malli 1**	**Malli 2**	**Malli 3**	**Malli 4**
0	40,0 %	100,0 %	100,0 %	100,0 %
1	0,0 %	80,0 %	80,0 %	46,7 %
2	10,0 %	43,3 %	40,0 %	40,0 %
3	6,7 %	33,3 %	23,3 %	23,3 %
4	0,0 %	73,3 %	43,3 %	76,7 %
5	3,3 %	6,7 %	26,7 %	40,0 %
6	0,0 %	3,3 %	10,0 %	20,0 %
7	0,0 %	63,3 %	73,3 %	86,7 %
8	0,0 %	10,0 %	16,7 %	60,0 %
9	0,0 %	16,7 %	30,0 %	46,7 %
10	0,0 %	40,0 %	63,3 %	66,7 %
11	0,0 %	0,0 %	63,3 %	63,3 %
12	0,0 %	70,0 %	53,3 %	46,7 %
13	100,0 %	53,3 %	93,3 %	76,7 %
14	100,0 %	83,3 %	66,7 %	80,0 %
15	100,0 %	86,7 %	56,7 %	50,0 %
Keskiarvo	**22,5 %**	**47,7 %**	**52,5 %**	**57,7 %**

Liite 5. Tekoälysanastoa suomen ja englannin kielellä

Käsite englanniksi	**Käsite suomeksi**
artificial intelligence, AI	tekoäly
generative artificial intelligence	generatiivinen tekoäly / luova tekoäly
narrow artificial intelligence	kapea / heikko tekoäly
general artificial intelligence	yleinen / vahva tekoäly
expert system	asiantuntijajärjestelmä
decision tree	päätöspuu
random forest	satunnaiset metsät
AI-model	tekoälymalli
AI-system	tekoälyjärjestelmä
transformer	muuntaja
chatbot	chattibotti
natural language processing, NLP	luonnollisen kielen prosessointi
large language model, MML	laaja kielimalli
promt	kehote
promt engineering	kehotesuunnittelu
tokenazation	tokenisointi
output	tuloste
fine-tuning	hienosäätö
parameter	parametri
hyperparameter	hyperparametri
hallucination	hallusinointi
zero-shot prompting	nollakehottaminen
chain-of-thought prompting	ajatusketjukehottaminen
few-shot prompting	vähäisen ohjauksen kehottaminen
custom GPT creation tool	GPT-4-chattibottien luomistyökalu