

# **Laaja kielimalli peruskoulun historian kokeen arvioijana**

Paikallisen laajan kielimallin tuottaman arvioinnin luotettavuus ja arviointipromptin luomisen  
ongelmat

Luokanopettajan tutkinto-ohjelma

Pro gradu -tutkielma

Maria Raitanen

31.05.2026

Rauma

Pro gradu -tutkielma

**Koulutusohjelma, oppiaine:** Kasvatustieteet, Luokanopettajan tutkinto-ohjelma

**Tekijä(t):** Maria Raitanen

**Otsikko:** Laaja kielimalli peruskoulun historian kokeen arvioijana: Paikallisen laajan kielimallin tuottaman arvioinnin luotettavuus ja arviointipromptin luomisen ongelmat

**Ohjaaja:** Yliopistonlehtori Lauri Kemppinen

**Sivumäärä:** 113 sivua

**Päivämäärä:** 31.05.2026

Tutkimuksen tavoitteena oli selvittää, kuinka luotettavan arvioinnin laaja kielimalli tuottaa alakoulun historian kokeen vastauksista. Lisäksi tutkimuksessa selvitettiin, millaisia ongelmia kehoitesuunnittelun aikana ilmenee, kun luodaan arviointipromptti, ja mitä asioita on huomioitava arviointipromptin käytössä. Tutkimuksen kohteena oli OpenAI:n paikallinen gpt-oss-120b laaja kielimalli.

Arvioinnin luotettavuutta tutkittiin vertailemalla laajan kielimallin tuottamia numeerisia arviointeja kahdelta arviointikerralta. Lisäksi luotettavuuden käsite operationalisoitiin viiteen analyysiyksikköön, joiden avulla tarkasteltiin muun muassa laajan kielimallin kykyä arvioida koevastauksia arviointimatriisin pohjalta sekä kykyä tunnistaa oppilaiden osaaminen ja antaa osaamista vastaava arvosana. Analyysiyksiköille annettiin arvo välillä 0–2 sen perusteella, kuinka hyvin kyseisen analyysiyksikön mittaama asia toteutui sanallisessa arvioinnissa. Kehotesuunnittelun aikana ilmenevät ongelmat ja arviointipromptin käytössä huomioitavat asiat selvitettiin toteuttamalla kehoitesuunnittelu käytännössä ja raportoimalla prosessien aikana ilmenneet ongelmat. Tutkimuksen toteutuksessa ja aineiston analyysissä yhdistyivät laadullinen ja määrällinen tutkimusote.

Tutkimuksen tulokset osoittavat, että laajan kielimallin tuottama arviointi ei ole luotettavaa. Mallin tuottama arviointi sisälsi liikaa virheitä ja oli epäjohdonmukaista. Esimerkiksi 34 koevastauksen arvioinneista 13 vastausta sai eri numeerisen arvioinnin ensimmäisellä ja toisella arviointikerralla. Lisäksi 68 kerrasta laaja kielimalli arvioi 10 kertaa toisen oppilaan vastauksen. Laaja kielimalli onnistui parhaiten tehtävässään arvioida vain arviointimatriisissa mainittuja asioita. Mallin suoriutuminen onnistuneesti vain yhden analyysiyksikön mittaamasta tehtävästä osoittaa, että arviointi ei ole luotettavaa. Tutkimuksen tulokset laajan kielimallin tuottaman arvioinnin epäluotettavuudesta ovat pääosin linjassa aikaisemman tutkimustiedon kanssa. Tutkimuksen tulokset osoittavat, että kehoitesuunnittelun aikana ilmenevät ongelmat liittyvät arviointimatriisin laatimiseen ja aikaisemman tutkimustiedon puutteelliseen kuvaukseen kehoitesuunnittelusta. Kehotesuunnittelussa ja arviointipromptin käytössä on huomioitava laajan kielimallin kontekstin pituus sekä käytetyn tietokoneen muisti ja laskentateho.

Tutkimuksen tulokset kehoitesuunnittelun aikana ilmenevistä ongelmista ovat merkittäviä, sillä aikaisemmissa tutkimuksissa kuvataan kehoitesuunnittelua rajallisesti. Tutkimus lisää tietoa laajan kielimallin mahdollisuuksista ja rajoituksista arviointityössä. Mikäli laajat kielimallit kykenisivät tulevaisuudessa tuottamaan luotettavaa arviointia, voisi se vähentää opettajan tuottaman arvioinnin riskejä, kuten epäjohdonmukaisuutta tai puolueellisuutta ja pitämään opettajan työ määrän kohtuullisena. Tekoälyn hyödyntäminen arviointityössä on kuitenkin vasta kehitysvaiheessa, minkä vuoksi on tärkeää lisätä ymmärrystä sen toiminnasta, luotettavuudesta ja soveltuvuudesta koulukontekstiin.

**Avainsanat:** arviointi, ChatGPT, Gemini, gpt-oss-120b, laaja kielimalli, paikallinen laaja kielimalli, tekoäly

## Sisällysluettelo

<b>1</b>	<b>Johdanto</b>	<b>5</b>
<b>2</b>	<b>Yleiskatsaus tekoälyyn</b>	<b>10</b>
2.1	Suppea katsaus tekoälyn historiaan	14
2.2	Tekoälyn tuottamat mahdollisuudet ja riskit yhteiskunnalle	16
2.3	Tekoälyn tuottamat mahdollisuudet ja riskit kouluissa	18
2.4	Opettajien ja oppilaiden näkemyksiä tekoälystä	23
<b>3</b>	<b>Oppilaan osaamisen arviointi</b>	<b>25</b>
3.1	Laaja kielimalli tekstiaineistojen arvioijana	26
3.2	Opettajan ja tekoälyn tuottaman arvioinnin riskit	35
3.3	Opettajien ja oppilaiden näkemyksiä arvioinnista	39
<b>4</b>	<b>Tutkimuksen toteutus</b>	<b>43</b>
4.1	Tutkimuskohteena käytettävän laajan kielimallin valinta	45
4.2	Aineiston keruun vaiheet	50
4.3	Kehotesuunnittelun vaiheet	53
4.4	Aineiston analyysi	60
4.5	Tutkimuksen eettisyys ja luotettavuus	67
<b>5</b>	<b>Tulokset</b>	<b>72</b>
5.1	Laajan kielimallin tuottaman arvioinnin luotettavuus	72
5.2	Kehotesuunnittelun aikana ilmenneet ongelmat ja arviointipromptin käytössä huomioitavat asiat	76
<b>6</b>	<b>Pohdinta</b>	<b>79</b>
6.1	Kolme mahdollista keinoa lisätä laajan kielimallin tuottaman arvioinnin luotettavuutta	82
6.2	Mahdollisia haasteita laajan kielimallin käytöstä arvioijana koulussa	87
6.3	Alustavia havaintoja ChatGPT 5.2 ja Gemini 3 Flash mallien toiminnasta arvioijina	88

<b>6.4</b>	<b>Jatkotutkimusehdotuksia</b>	<b>90</b>
	<b>Lähteet</b>	<b>93</b>
	<b>Liitteet</b>	<b>109</b>
	<b>Liite 1. Kuvio tutkimuksen koeasetelmasta</b>	<b>109</b>
	<b>Liite 2. Esimerkkejä analyysiyksiköiden arvojen tuottamisesta</b>	<b>110</b>

## 1 Johdanto

Tekoälyn yleistymisen on ollut viime vuosien merkittävimpiä ilmiöitä. Siitä puhutaan niin kahvipöytäkeskusteluissa kuin kansainvälisissä huippukokouksissa, ja sen kehitystä seurataan tiiviisti uutisissa ja tutkimuksissa. Tekoäly on ihastuttanut mahdollisuuksillaan, mutta myös herättänyt huolta esimerkiksi tietosuojariskeillään ja vaikutuksillaan työmarkkinoihin ja ilmastoon. Yksi asia on kuitenkin melko varmaa, tekoäly on tullut jäädäkseen.

Tekoälyn rooli myös kouluympäristössä on herättänyt vilkasta keskustelua niin myönteisessä kuin kielteisessä valossa. Helsingin sanomat (Myllyoja, 2023) uutisoivat opettajien joutuneen uusien haasteiden eteen, kun oppilaat ovat alkaneet käyttämään tekoälyä tehtävien tekemiseen. Vaikka huijaaminen ei ole koulussa uusi ilmiö, tekoälyn käyttö tekee sen havaitsemisesta entistä vaikeampaa. Esimerkiksi kuopiolainen opettaja huomasi oppilaidensa hyödyntäneen tekoälyä, kun heidän vastauksensa paranivat yllättäen. (Myllyoja, 2023.) Kuitenkin oikein käytettynä tekoälystä voi olla paljonkin hyötyä. Kangasalainen opettaja kertoo ajankäytön ja luovuuden hyötyvän tekoälystä. Hänen mukaansa tekoäly on oiva luovuuden ruokkia niin opettajille kuin oppilaillekin. (Hart & Vesanummi, 2024.) Myös forssalainen ammattikoulun opettaja näkee tekoälyn tuovan merkittäviä hyötyjä. Hän mainitsee, että ammattikouluissa on paljon opeteltavaa, mitä ei pysty harjoittelemaan konkreettisesti. Esimerkiksi tuuliturbiinia tai aurinkovoimalaa ei pysty tuomaan luokkatilaan. Tekoälyn ja virtuaalitodellisuuden avulla se voi olla mahdollista. (Vesanummi, 2024.)

Suomen kouluilla ei ole yhtenäistä linjaa tekoälyn käytöstä. Koulutuksen järjestäjät päättävät oman organisaationsa tekoälyn käytöstä. (Opetushallitus, ei pvm. a.) Moni koulu on luonut tekoälylinjaukset opiskelijoilleen, ja nämä ovat nähtävissä internetissä (esim. Oulun steinerkoulu, ei pvm.; Kuopion lukiokoulutus, ei pvm.). Kuitenkin haasteellista on löytää opettajien tekoälylinjauksia. Koulut, joiden tekoälylinjaukset olivat saatavilla, sallivat usein Microsoftin Copilotin käytön mutta esimerkiksi ChatGPT on kielletty.

Vantaan kaupunki on kouluttanut opettajia tunnistamaan tekoälyyn liittyvät riskit sekä käyttämään Copilottia hyödyksensä (Lumme, 2024). Vantaan kaupungin

tekoälylinjausta ei ole nähtävillä internetissä, mutta Vantaan Sanomat (Lumme, 2024) ovat kirjoittaneet aiheesta. Myös Turussa on käytössä Copilot. Turun kaupungin linjaus painottaa, että tekoälystä voi olla hyötyä niin oppilaalle kuin opettajallekin. Linjaus kuitenkin muistuttaa tietosuojasta, ja että esimerkiksi oppilaan vastauksia ei saa syöttää tekoälylle. (Edu.turku.fi, ei pvm.). Puolestaan Rauman kaupunki ei ole luonut yleistä tekoälylinjausta, mutta Rauman lyseon lukion on luonut oman tekoälyohjeistuksen. Koulu suosittelee Copilotin käyttöä, mutta ei kuitenkaan kiellä muiden tekoälyjen hyödyntämistä. Muiden tekoälyjen hyödyntämistä täytyy kuitenkin harkita tarkkaan. Linjauksessa mainitaan, että opettajat voivat käyttää tekoälyä työssään parhaaksi näkemällään tavalla. Opettajia kehoitetaan rohkaisemaan oppilaita käyttämään tekoälyä, jos se on pedagogisesti perusteltua. (Rauman lyseon lukio, 2024.) Helsingin kaupunki teki puolestaan yllättävän käänteen vuoden 2025 helmikuussa, kun se ilmoitti Copilotin käyttökiellosta. Kyseisen laajan kielimallin käyttö kiellettäisiin siihen saakka, kunnes mallin riskit saadaan analysoiduksi. (Kymäläinen, 2025.) Käyttökiellon purkamisesta ei ole julkisesti ilmoitettu.

Opetushallitus sekä Opetus- ja kulttuuriministeriö ovat laatineet tiedotteen, joka sisältää velvoitteita ja suosituksia tekoälyn käyttöön opetuksessa (Opetushallitus, ei pvm. a). Tekoälytiedotteessa kerrotaan, että tekoäly voi olla hyödyksi opetuksen suunnittelussa, toteutuksessa ja arvioinnissa. Oppilaille tekoäly voi puolestaan olla sekä oppimisen kohde että väline. Tekoälytiedotteessa painotetaan, että tekoälyn käyttö, muidenkin materiaalien tavoin, on oltava pedagogisesti perusteltua. (Opetushallitus, ei pvm.a.) Vaikka tekoälyä ei mainita vielä opetussuunnitelmassa, voidaan sen tunteminen ja käyttötaidot lukea osaksi tieto- ja viestintäteknologista osaamista (Opetushallitus, ei pvm. b). Tieto- ja viestintäteknologiset taidot eli TVT-taidot ovat puolestaan yksi laaja-alaisen osaamisen tavoitteista (OPS, 2014).

Tekoälytiedotteen (Opetushallitus, ei pvm. c) tiedot tekoälyn hyödyntämisestä arvioinnin tukena ovat tulkinnanvaraisia. Tiedotteessa mainitaan, että tekoäly voi olla avuksi oppimisen arvioinnissa. Arvioinnin osalta tiedotteessa mainitaan, että opettaja voisi hyödyntää tekoälyä formatiivisessa arvioinnissa. (Opetushallitus, ei pvm. c.) Tiedote ei kuitenkaan selkeästi listaa asioita mitä saisi ja mitä puolestaan ei saisi tehdä formatiivisen arvioinnin osalta. Tiedotteessa mainitaan, että summatiivisen arvioinnin

on oltava aina opettajan tekemää (Opetushallitus, ei pvm. c). Opettajan on osattava tulkita tekoälyn käytön riskejä tiedotteen pohjalta. Esimerkiksi siinä sallitaan suuririskisten tekoälyjärjestelmien käyttö, toki tietyt reunaehdot on täytyttävä (Opetushallitus, ei pvm. c).

Tekoäly on huolenaiheena verrattain uusi, mutta oppilaiden arviointi on puolestaan pohdituttanut jo pitkään. Arvioinnin yhdenvertaisuus on herättänyt huolta niin opettajien kuin vanhempien keskuudessa (Opetushallitus, ei pvm. d.) Vielä muutamia vuosia sitten opettajat saattoivat kuvata päättöarviointia bingokoneeksi (Kivimäki & Krautsuk, 2019). Ratkaisuksi ongelmaan, Opetushallitus on viimeisien vuosien aikana laajentanut osaamisen kuvauksia. Aikaisemmin opetussuunnitelman perusteisiin oli laadittu osaamisen tason kuvaus kuudennen ja yhdeksännen luokan päättöarviointiin vain arvosanalle kahdeksan. Uudistuksen myötä osaamisen tason kuvaukset luotiin myös arvosanoille viisi, seitsemän ja yhdeksän. (Opetushallitus, ei pvm. e.) Toisena uudistuksena on se, että numeroarviointi tulee pakolliseksi neljännessä luokasta alkaen. Aikaisemman säädöksen mukaan numeroarviointiin oli siirryttävä viimeistään kahdeksannella luokalla. (Terävä, 2020.) Uudistuksen tavoitteena on lisätä arvioinnin avoimuutta ja yhdenvertaisuutta sekä mahdollistaa arvosanojen vertailukelpoisuus esimerkiksi eri koulujen välillä. (Opetushallitus, ei pvm. e.)

Uudistus oli opettajien mukaan hyvin tarpeellinen. Suuri osa opettajista koki uusien kriteerien lisäävän arvioinnin yhdenvertaisuutta. (Puukko, Huhtanen & Lepola, 2020.) Kuitenkin Pulkkinen ym. (2024) tutkimuksen mukaan yli 30 prosenttia opettajista kokee oman koulun eroavan silti yleisistä arviointilinjoista. Myös osa rehtoreista kokee näin, mutta ei suurin osa (Pulkkinen ym., 2024). Opettajat pitivät erityisen tärkeänä etenkin arvosanan viisi osaamisen kuvausta. Arvioinnin uudistukset eivät kuitenkaan korjanneet kaikkea. Keskeinen palaute oli, että kriteerit eivät ole tarpeeksi ymmärrettäviä ja konkreettisia. Ne ovat liian ympäröityjä ja jättävät silti tulkinnalle tilaa. (Puukko, Huhtanen & Lepola, 2020.) Opettajien tyytyväisyys oli kuitenkin verrattavissa siihen, kuinka hyvin he ovat perehtyneet osaamisen kuvauksiin. Huonosti perehtyneet olivat tyytymättömiä uudistukseen kuin he jotka olivat perehtyneet. Opettajat myös kokevat, että uudistuksen jälkeen työmäärä on lisääntynyt. (Pulkkinen ym., 2024.)

Arviointi perustuu opetussuunnitelmassa esitettyihin tavoitteisiin. Oppilaan osoittamaa osaamista arvioidaan suhteessa kunkin aineen tavoitteisiin. Arvioinnilla on samat kriteerit, riippumatta siitä onko kyseessä esimerkiksi yksittäisen tehtävän arviointi, päättö- tai lukuvuosiarviointi. (OPS, 2014.) Oppilas voi näyttää omaa osaamistaan esimerkiksi tuntityöskentelyllä, kokeilla, esseillä tai esitelmillä. Menetelmää valitessa opettaja ottaa huomioon esimerkiksi oppilaiden iän ja opeteltavan aiheen. Opettajien yleisimmin käytettyjä arviointimenetelmiä ovat yksin tehtävät kirjalliset kokeet, essee tai kirjoitelmat (Atjonen, ym., 2019, s.111 ja 217). Siirryttäessä luokka-asteilta ylemmäksi pitenevät oppilaiden kirjalliset vastaukset, sillä aiheet laajenevat ja oppilaiden taidot karttuvat. Esimerkiksi yläasteikäiseltä voi vaatia jo parin sivun pituista esseetä, kun kakkosluokkalaiselle riittää muutama virke. Loogisesti opettajien arviointityöhön kuluva aika kasvaa, kun arvioitavat tehtävät pitenevät.

Opettajien arviointiin kuluva aika vaihtelee riippuen esimerkiksi maasta tai luokka-asteesta, jossa hän työskentelee. Opetuksen ja oppimisen kansainvälisen tutkimuksen eli TALIS 2013 (OECD, 2014, s.162) tutkimuksen mukaan OECD maiden opettajat käyttävät keskimäärin viisi tuntia viikosta töiden arviointiin. Zhao ym. (2025) tutkimuksen mukaan Kiinassa työskentelevät opettajat käyttävät kotitehtävien arviointiin 5–15 tuntia. Tutkimuksessa ei otettu huomioon kaikkiin arviointitehtäviin kuluva aika. Tämän vuoksi arviointiin kuluva aika on paljon suurempi. Viimeisenä, Isossa Britanniassa työskentelevillä opettajilla arviointiin kuluu keskimäärin kolme ja puoli tuntia (MORI, 2001).

Tämän myötä voidaan pohtia, voisiko laaja kielimalli ratkaista arvioinnin ongelmat. Laaja kielimalli voisi nopeuttaa arviointiprosessia, jotta opettajalla jää enemmän aikaa esimerkiksi oppilaiden kohtaamiseen tai oppituntien suunnitteluun. Tärkeimpänä, laajalla kielimallilla on potentiaalia edistää arvioinnin yhdenvertaisuutta, sillä sen toimintaan ei vaikuta esimerkiksi oppilaan henkilökohtaiset piirteet tai ennakkoletukset oppilaasta. Näin laaja kielimalli voisi osaltaan tukea objektiivisempaa ja tasapuolisempaa arviointia. Tekoälyn tuomat muutokset haastavat kuitenkin opetuskäytäntöjä, ja se edellyttää sekä opettajilta että oppilailta uudenlaista ajattelua ja valmiutta sopeutua muutoksiin. Opettajat ja oppilaat tarvitsevat ymmärrystä tekoälyn

toimintaperiaatteista, jotta sitä osataan hyödyntää vastuullisesti ja tarkoituksenmukaisella tavalla.

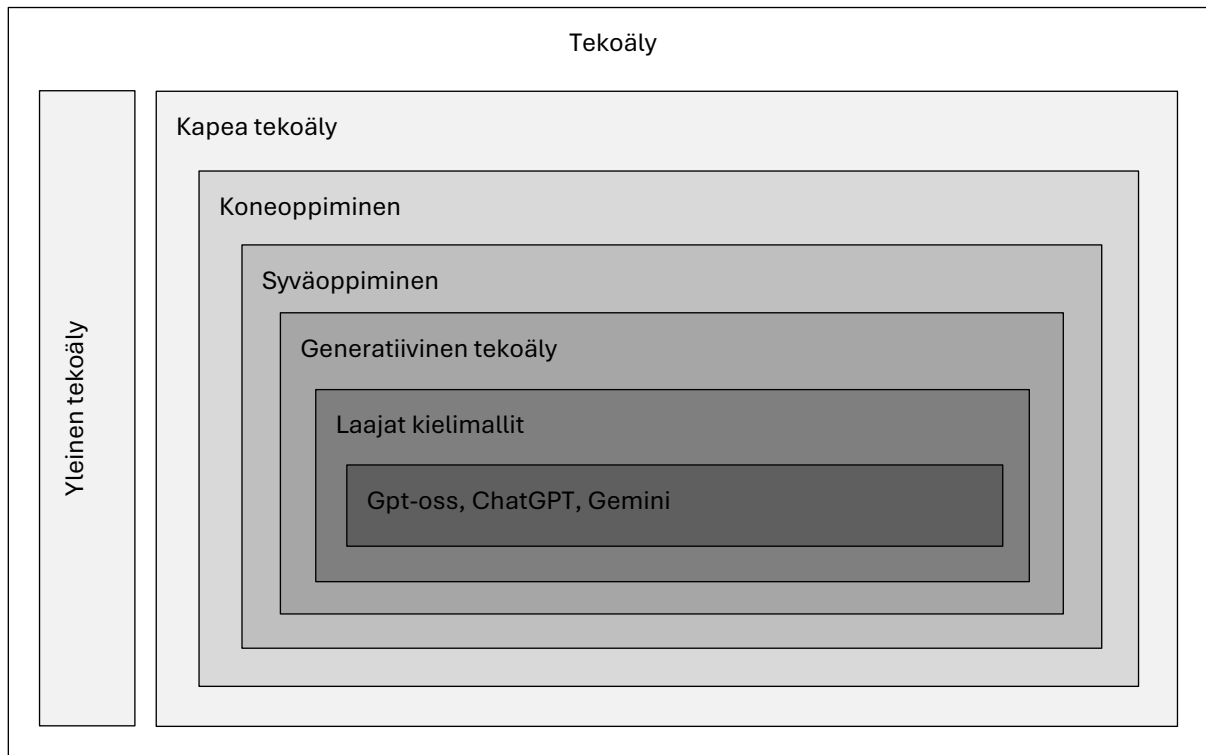
Tutkimuksen tavoitteena on selvittää, kuinka luotettavasti laaja kielimalli arvioi alakoulun historian kokeen vastauksia. Tarkastelun kohteena on OpenAI:n gpt-oss-120b paikallinen laaja kielimalli. Tutkimuksen toisena tavoitteena on selvittää, millaisia ongelmia kehoitesuunnittelussa ilmenee, kun luodaan arviointipromptti, ja mitä asioita on huomioitava arviointipromptin käytössä.

Tutkimus tuottaa arvokasta tietoa paikallisen laajan kielimallin mahdollisuuksista ja rajoituksista arviointityössä. Aikaisemmat tutkimukset ovat raportoineet puutteellisesti kehoitesuunnittelusta, ja sen vuoksi tämän tutkimuksen tulokset arviointipromptin luomisen ongelmista ovat merkittäviä. Näin tutkimus lisää ymmärrystä tekijöistä, jotka ovat otettava huomioon kehoitesuunnittelun aikana. Tekoälyn, tarkemmin laajojen kielimallien, käyttö koulukontekstissa on vasta kehitysvaiheessa, ja siksi on olennaista lisätä ymmärrystä niiden toimivuudesta ja luotettavuudesta. Tutkimuksen tulokset luovat perustaa keskustelulle, miten tekoälyä voidaan käyttää hyödyllisesti ja vastuullisesti. Tulokset auttavat myös hahmottamaan, millaista osaamista tai lisäkoulutusta opettajat voivat tarvita tekoälyn yleistyessä.

## 2 Yleiskatsaus tekoälyyn

Tekoäly on tietokonejärjestelmän kyky saada koneet tekemään sellaisia älyllisiä tehtäviä, joiden uskottiin olevan mahdollista vain ihmisille. Kyseisiä tehtäviä ovat esimerkiksi oppiminen, päättely ja luominen. (Valtiovarainministeriö, 2025.) Tekoälyt pystyvät havainnoimaan ympäristöään, tekemään päätöksiä ja toimimaan itsenäisesti jonkin tavoitteen saavuttamiseksi. Tekoäly voi olla pelkkä ohjelmisto, kuten puheentunnistin tai hakukone. Se voi myös olla fyysinen laite, kuten robotti tai itseajava auto. (Euroopan parlamentti, 2023.)

Tekoäly on laaja sateenvarjokäsite. Tekoälyn määrittelemisen edellyttää perehtymistä käsitteisiin, jotka kuuluvat sateenvarjon alle. Kyseiset käsitteet sisältävät vielä monia alakäsitteitä, mikä tekee tekoälystä erittäin monimutkaisen kokonaisuuden. Seuraavaksi esitellään käsitteitä, jotka liittyvät tässä tutkimuksessa käytettyjen tekoälyjen, tarkemmin laajojen kielimallien, toimintaan. Kuviossa 1 on esitetty kyseisten käsitteiden keskinäinen hierarkia.



Kuvio 1. Tekoälyyn liittyvien käsitteiden keskinäinen hierarkia.

Laajasti tekoäly voidaan jakaa kahteen, kapeaan ja yleiseen. Yleinen tekoäly pystyisi suoriutumaan lähes kaikista tehtävistä, mihin ihmisetkin pystyvät. (European

Commission, 2018.) Taitoihin sisältyisi erilaisia fyysisiä ja älyllisiä taitoja, kuten aistikyky, mahdollisuus liikuttaa esineitä ja tutkia ympäristöä. Yleinen tekoäly luo mahdollisuuksia hälyttävälle riskeille, kuten massatyöttömyydelle tai jopa ihmisten sukupuutolle (XAMK, ei pvm.) Jälkimmäinen on kuitenkin äärimmäinen esimerkki, jos tekoälyn kyvyt ylittäisivät ihmisten hallinnan. Yleisen tekoälyn luominen on erittäin haastavaa, ja se luo huomattavasti suurempia riskejä kuin kapea tekoäly. Tämän vuoksi nyt käytössä olevat tekoälyt ovat kapeita, mutta yleisiä tekoälyjä pyritään kehittämään. Kapealla tekoälyllä viitataan järjestelmiin, jotka osaavat suorittaa yhtä tai muutamaa tehtävää. (European Commission, 2018.) Kapea tekoäly jakautuu moneen erilaiseen tekoälytyyppiin.

Kapeita tekoälytyyppejä ovat esimerkiksi ennakoiva tekoäly ja generatiivinen tekoäly. Nämä tyypit ovat luotu ratkaisemaan erilaisia tavoitteita. Erilaiset tavoitteet vaativat tekoälytyypiltä eri ominaisuuksia, mitkä ilmenevät jo mallin koulutusvaiheessa. Esimerkiksi generatiivisen tekoälytyypin tavoitteena on luoda uutta. Uuden luominen voi olla esimerkiksi tekstiä, kuvia tai ääntä. Suositut tekoälyt, kuten ChatGPT tai Gemini ovat molemmat generatiivisia. Puolestaan ennakoiva tekoäly on luotu ennustamaan tulevaa. Ennustamisen kohteita voivat olla esimerkiksi yrityksen ensi vuoden liikevaihto tai kävijämäärät. (Caballar, ei pvm.)

Erilaiset tekoälytyypit vaativat erilaisen kouluttamisen. Koulutusmenetelmät ovat joko päättelyyn tai oppimiseen perustuvia. Päättelymenetelmässä koneelle annetaan selkeitä sääntöjä ja malleja, joiden pohjalta se tekee johtopäätöksiä. (European Commission, 2018). Esimerkiksi järjestelmä voidaan kouluttaa tunnistamaan iloiset kasvot. Silloin koneelle syötetään joukko sääntöjä, joiden toteuduttua kone päättelee iloiset kasvot. Sääntö voi olla esimerkiksi koholla oleva suupieli. Puolestaan oppimismenetelmässä kone opetetaan esimerkkien eikä sääntöjen avulla. Kone opetetaan tunnistamaan iloiset kasvot näyttämällä suuri määrä iloisiksi ja ei-iloisiksi luokiteltuja kasvoja. Kone oppii itse tunnistamaan säännön, jonka avulla tunnistaa iloiset kasvot. (Honkela, ei pvm.) Oppimismenetelmän yksi muoto on koneoppiminen (European Commission, 2018).

Koneoppimisesta on tullut yksi tekoälyn keskeisimmistä aiheista, sillä se mahdollistaa esimerkiksi generatiivisten tekoälyjen toiminnan (Bergmann, ei pvm. a). Koneoppiminen tarkoittaa sitä, että tekoäly opetetaan tekemään itsenäisiä päätöksiä aiempien havaintojen perusteella (European Commission, 2018). Oppimismenetelmän vuoksi, koneella on kyky oppia datasta, sen sijaan, että se ohjelmoidaan sääntö kerrallaan. Tämä mahdollistaa sen, että koneoppimista hyödyntävät tekoälyt kehittyvät käytön myötä. Käyttö lisää tekoälyn kokemusta, mikä puolestaan kouluttaa tekoälyä. (Viitaila, ei pvm.)

Koneoppimista voidaan tehdä eri tavoin. Yksi tapa on ohjatusti esimerkkien avulla. Ohjatussa oppimisessa hyödynnetään dataa, johon on lisätty tieto datan sisällöstä. Esimerkiksi kuvat ovat määritelty iloisiksi tai ei-iloisiksi kasvoiksi. Toinen tapa on ilman ohjeistusta, jossa kone etsii itse toimintamalleja annetuista tiedostoista. Tässä tavassa datan sisältöä ei ole luokiteltu. Kolmas tapa on vahvistusoppimisen avulla, jossa kone saa palautetta hyvistä ja huonoista päätöksistä. Palautteen myötä tekoäly oppii suoriutumaan tehtävästä oikein. (European Commission, 2018; Viitaila, ei pvm.)

Neuroverkot ovat yksi koneoppimisen menetelmä. Neuroverkkojen avulla tekoäly pystyy käsitellä monimutkaisia tehtäviä, kuten kasvojentunnistusta tai puheenymmärrystä. Neuroverkot muistuttavat löyhästi ihmisaivojen rakennetta, mihin sen nimikin viittaa. Neuroverkot koostuvat keinotekoisista neuroneista, joita kutsutaan solmuiksi. Solmujen välillä on joukko liitoksia eli keinotekoisia synapseja. Solmut ja keinotekoiset synapsit muodostavat syöte-, väli- ja tulokerroksia. Tieto saapuu syötekerrokselle, seuraavaksi kulkee yhden tai useamman välikerroksen kautta tulokerrokseen, mikä tuottaa lopullisen vastauksen. Neuroverkot kouluttavat itseään lisäämällä synapsien painoarvoa. Jokaisella synapsilla on painoarvo, joka kertoo kuinka suuri vaikutus yhden neuronin ulostulolla on seuraavaan neuroniin. (Lappalainen, ei pvm.)

Syväoppiminen on neuroverkkojen edistynyt muoto. Termillä ”syvä” viitataan siihen, että tieto kulkee useampien neuroverkkojen kautta. Tämä parantaa toiminnan tarkkuutta ja vähentää ihmisen tarvetta ohjata oppimista. (European Commission, 2018.) Laajat kielimallit hyödyntävät neuroverkkoja ja syväoppimista. Laajoja kielimalleja ovat esimerkiksi Open AI:n Chat GPT ja Googlen Gemini. (The university of

arizona, ei pvm.) Laajan kielimallin englannin kielinen vastine on Large Language Model, ja siitä käytetään lyhennettä LLM. Lyhennettä LLM esiintyy usein myös suomen kielisissä artikkeleissa, ja siksi tässä tutkielmassa käytetään lyhenteitä LLM tai malli tarkoittamaan laajaa kielimallia. Laajat kielimallit ovat suunniteltu tuottamaan ihmisen kaltaista kieltä ja suorittamaan siihen liittyviä tehtäviä. Mallit pystyvät esimerkiksi vastaamaan kysymyksiin, tiivistämään tekstejä ja tuottamaan uutta sisältöä. (The university of arizona, ei pvm.) Laajojen kielimallien toiminta perustuu tekoälyn kykyyn ennustaa seuraava sana. Malli oppii kielen rakenteita, sääntöjä ja sisältöjä analysoimalla valtavia tekstiaineistoja. (ChatGPT Suomi, ei pvm.) Esimerkiksi oletetaan, että ChatGPT 3 koulutettiin 43 teratavulla dataa (IBM technology, 2023). Tämä vastaa 43000 gigatavua. OpenAI ei kuitenkaan itse ole julkaissut lukuja, minkä vuoksi koulutusdatan määrästä ei ole varmuutta. Suuren koulutusdatan avulla laaja kielimalli osaa ennustaa seuraavan sanan, ja lopulta tarjota pätevän lauseen tai jopa kokonaisen esseen (The university of arizona, ei pvm.).

Laajojen kielimallien käyttö alkaa tekoälyn promptauksella. Promptaus termi juontuu englannin kielen sanasta ”prompt”, jonka määritelmä on saada jokin tekemään jotain tai aiheuttaa jokin tapahtuma (Cambridge dictionary, ei pvm.). Promptaus termi on melko vakiintunut suomen kieleen, mutta sille löytyy myös suomenkielinen vastine eli kehote (MOT Sanakirjat, ei pvm.). Promptauksella eli kehotteiden laatimisella tarkoitetaan sitä prosessia, kun laajalle kielimallille syötetään toimintaohjeet. Promptin avulla malli tuottaa halutun tapahtuman. Tapahtuma voi olla teksti, kuva tai muu asia, johon kyseinen tekoäly on erikoistunut. (Harvard university, 2023.)

Onnistuneen promptin avulla laajan kielimallin tuottama tapahtuma on halutunlainen ja laadukas. ”Prompt engineering” eli kehotesuunnittelu viittaa siihen prosessiin, kun pyritään luomaan onnistunutta promptia. Prosessi sisältää promptin suunnittelua, testausta ja korjausta. Pätevän promptin luominen voi olla haastava prosessi, sillä esimerkiksi promptin jokainen sana voi vaikuttaa lopputulokseen. (Gadesha, ei pvm.) Prompti voi sisältää kysymyksiä, ohjeita tai käskyjä. Onnistuneen promptin ytimessä ovat tarkat ohjeistukset (Harvard University, 2023). Laajalla kielimallilla on itsellään paljon päättämivaltaa, mikäli sille antaa yleisluontoisia ohjeita. Tarkat kehotteet mahdollistavat halutunlaiset tulokset. Harvard University (2023) selittää nettisivuillaan,

kuinka luoda hyvä prompti tarinan luonnin esimerkillä. Toimintaohje ”kerro tarina”, antaa mallille vapaat kädet ja täten lopputulos voi olla lähes mitä vain. Kun toimintaohjetta tarkennetaan esimerkiksi aiheella, genrellä, kohderyhmällä, suosituspituudella ja tekstin tyylillä, on tulos lähempänä haluttua. (Harvard University, 2023.) Tarkat kehotteet antavat laajalle kielimallille suppeammat raamit, minkä sisällä luoda vastaus.

Ytimekkäät ja selkeät käskyt ovat hyödyllisimpiä (Harvard University, 2023). Aikaisempi kehoite-esimerkki ”kerro tarina” on lyhyt ja ytimekäs, mutta se ei sisällä kuitenkaan tarpeeksi tietoa haluttua lopputulosta varten. Laajankin tiedon voi kertoa ytimekkäästi. Kehotetta voi tarkentaa esimerkiksi näin: ”Luo tarina, joka on 300 sanaa pitkä, genreltään kauhua, kohdistettu aikuisille ja on tyyliltään kirjakieltä.” On tärkeä muistaa, että tekoäly on vain kone, minkä vuoksi kohteliaisuudet tai muut tuotoksen kannalta tarpeettomat tiedot pitää jättää pois. Harvard University sivusto (2023) suosittelee hyödyntämään esimerkiksi käskyjä tee näin ja älä tee näin. Tällaiset käskyt antavat selkeät ja ytimekkäät toimintaohjeet tekoälylle.

Tässä tutkielmassa käytetään käsitettä arviointiprompti kuvaamaan sellaista laajalle kielimallille annettavaa promptia, jonka tarkoituksena on ohjata mallia tuottamaan arviointi. Käsite on muodostettu tätä tutkielmaa varten. Arviointiprompti sisältää sekä promptin että arviointimatriisin. Käsite ei siis pelkästään sisällä käskyä toteuttaa arviointia vaan lisäksi myös arviointimatriisin, jonka pohjalta laaja kielimalli arvioi esimerkiksi esseen, koevastauksen tai muun tekstimuotoisen tehtävän. Käsitteen käyttö mahdollistaa promptin ja arviointimatriisin tarkastelun yhtenä kokonaisuutena, sillä molemmat ovat välttämättömiä, kun laajalla kielimallilla tuottaa arvioinnin.

## **2.1 Suppea katsaus tekoälyn historiaan**

Tekoälyn historia ulottuu huomattavasti kauemmaksi kuin ChatGPT:n tai Geminin keksimiseen. Jo 1900-luvun alussa espanjalainen insinööri Leonardo Torres Quevedo luo automatisoidun shakkipelikoneen. Vaikka varsinaisesta tekoälystä ei vielä voida puhua, antoi kone varhaisia viitteitä myöhemmästä tekoälyn kehityksestä. Shakkipelikone toimi sähkömagneettien avulla, ja kykeni pelaamaan kolmella nappulalla yksinkertaistettua shakkitilannetta. (Mucci, ei pvm.)

Merkittäviä edistysaskelia kohti varhaista tekoälyä alettiin ottamaan 1950-luvulla. Vuosikymmenen alusta lähtien julkaistiin lähes vuosittain artikkeleita ja innovaatioita koskien varhaista tekoälyä. (Mucci, ei pvm.) Brittiläinen Alan Turing kirjoitti jo 50-luvulla artikkelin, joka pohti kysymystä, voivatko koneet ajatella tai osoittaa älykäästä käyttäytymistä niin kuin ihmiset (Turing, 1950). Artikkelin loi myös pohjaa ajatukselle, kuinka koneen älykkyyttä voidaan mitata. Yksi tekoälyn älykkyyttä mittaava testi onkin nimetty Turingin mukaan. (Mucci, ei pvm.) Vain vuotta myöhemmin Marvin Minsky ja Dean Edmund loivat ensimmäisen, varhaisen, version neuroverkosta. Neuroverkko, nimeltään SNARC, luotiin imitoimaan rotan käyttäytymistä sokkelossa. Rotan tavoin kone oppi säätelemään toimintaansa annetun palautteen pohjalta. SNARC oli flyygelin kokoinen kone, jossa valot imitoivat rotan kulkemista sokkelossa. (Mucci, ei pvm.)

Seuraavalla vuosikymmenellä Ellis Page luo ensimmäisen tietokoneella toimivan esseenarviointijärjestelmän nimeltä PEG. Vaikka PEG ei toiminut tekoälyllä, loi se pohjaa nykyisille tekoälyä hyödyntäville arviointityökaluille. (Mucci, ei pvm.) PEG:in toiminta perustui näkemykseen, että tietyt tekstin kvalitatiiviset piirteet korreloivat esseen arvosanan kanssa. Ensimmäinen vaihe PEG:in luomisessa oli pyytää ihmisarvioijia arvioimaan suuri määrän esseitä. Tämän jälkeen kyseisistä esseistä mitattiin erilaisia piirteitä, kuten sanojen, pilkkujen tai pisteiden sekä harvinaisten sanojen määrä. Näiden piirteiden ja ihmisten tekemien arvioiden pohjalta muodostettiin monimuuttujainen regressiokaava. (Page, 1968.) Monimuuttujaisella regressiokaavalla tarkoitetaan tilastollista mallia, jossa yhtä selitettävää muuttujaa, tässä tilanteessa arvosanaa, ennustetaan usean selittävän muuttujan avulla (Ellonen & Kaakinen, ei pvm.) Regressiokaavan avulla PEG määrittää esseen arvosanan (Page, 1968).

Joseph Weizenbaun loi puolestaan ensimmäisen chattibotin, ELIZA:n, jo vuonna 1966. Nykyisten chattibottien tavoin, ei ELIZA oikeasti ymmärtänyt vastaajan esittämiä kysymyksiä tai yhdessä käytyjä keskusteluita. ELIZA:n toiminta perustui kuvion tunnistukseen. Kone koodattiin tunnistamaan kuvioita eli tässä tilanteessa sanoja. Koneelle koodattiin pitkä lista avainsanoja, joiden prioriteettiarvo oli määritelty. Yleisillä sanoilla, kuten sinä ja minä, oli matala prioriteettiarvo. Puolestaan esimerkiksi sanoilla äiti tai surullinen oli korkeampi prioriteettiarvo. Vastauksien luontiin ELIZA käytti ennalta määrättyjä vastauspohjia. Vastauspohjat määrättyivät avainsanojen perusteella. ELIZA

muodosti vastauksen sisällyttämällä vastauspohjaan avainsanan ja vastakkaisen pronominin. Vastakkaisella pronomiinilla tarkoitetaan, että ihmisen käyttäessä esimerkiksi sanaa minä, ELIZA käyttää sanaa sinä. Vastaukset olivat usein kysyviä, mikä loi illuusion keskustelusta. (Weizenbaun, 1966.)

Tekoälyn kehitys nykyiselle tasolle ei ole tapahtunut hetkessä, vaan se on ollut pitkä prosessi. Vuosien aikana tiedot tekoälyn toimintaperiaatteista ja mahdollisuuksista ovat lisääntyneet sekä tietokoneet ovat kehittyneet. Esimerkiksi uusien menetelmien, kuten kone- ja syväoppimisen keksiminen ja tietokoneiden laskentatehon kasvaminen ovat mahdollistaneet tekoälyn valjastamisen entistä monimutkaisempiin tehtäviin. Nämä ovat osatekijöitä siihen, miksi tekoäly on ottanut merkittäviä edistysaskelia viimeisten vuosikymmenien aikana. (Euroopan parlamentti, 2023.)

## **2.2 Tekoälyn tuottamat mahdollisuudet ja riskit yhteiskunnalle**

Tekoälyn avulla voidaan mahdollistaa turvallisempi, tasa-arvoisempi ja helpompi elämä, mutta sen valmistamiseen ja käyttämiseen liittyy myös monia riskejä. Euroopan parlamentti (2025) listaa artikkelissaan tekoälyn tuottamia mahdollisuuksia. Esimerkiksi tekoäly voi auttaa kansalaisia tiedonhankinnassa ja kouluttautumisessa. Se voi myös mahdollistaa kansalaisille turvallisemmat liikennevälineet ja paremman terveydenhuollon. (Euroopan parlamentti, 2025.) Esimerkiksi terveydenhuollossa tekoälyn on todettu olevan hyödyllinen syöpien tunnistamisessa. Placidon ym. (2023) tutkimuksessa tekoäly onnistui tunnistamaan henkilöt, jotka ovat suuremmissa riskissä sairastua haimasyöpään. Tekoäly tunnistasi riskihenkilöt jopa kolme vuotta ennen varsinaista diagnoosia. Lisäksi Långin ym. (2023) tutkimuksen mukaan tekoäly pystyi tunnistamaan syöpälöydöksiä mammografioista. Tekoäly havaitsi 20 prosenttia enemmän syöpälöydöksiä kuin kaksi radiologia.

Tekoäly voisi myös mahdollistaa turvallisemmat työpaikat, kun vaaralliset tehtävät voi jättää tekoälyllä toimivalle robotille. Kehittyvä tekoälyteollisuus lisää myös uusia työpaikkoja esimerkiksi robotiikan tai promptin kehitykseen. Lisäksi yritykset voivat hyödyntää tekoälyä prosessoidakseen dataa nopeammin ja taatakseen esimerkiksi tasa-arvoisen kohtelun työhaussa. Tekoälyä voidaan hyödyntää myös yrityksen verkkoalustoilla, jotta laitonta tai siihen johtavaa toimintaa voitaisiin ennaltaehkäistä.

(Euroopan parlamentti, 2025.) Oikein hyödynnettynä tekoäly luo suuria mahdollisuuksia turvalliseen ja helpompaan elämään.

Tekoälyn valmistaminen ja käyttäminen tuottaa kuitenkin monia riskejä. Riskit liittyvät vastuun määrittelyyn, tekoälyn vinoumiin sekä yksityisyyteen, tietosuojaan ja ympäristöön. Yksi riski syntyy tekoälyn tuottamien ongelmien vastuun määrittelystä. Kenen vastuulla on esimerkiksi itseohjautuvan auton kolarista aiheutuneet kustannukset? Onko vastuuvollisena auton omistaja, valmistaja vai tekoälyohjelmiston kehittäjä? Toinen riski piilee virheellisyyksissä ja väärentymissä, joita on tahallisesti tai tahattomasti syntynyt tekoälyn valmistusvaiheessa. Mikäli tekoäly on koulutettu virheellisellä datalla, voi sen tuottamat tulokset tai päätökset olla kohtalokkaita. Esimerkiksi työnhaussa tai lainanmyöntämisessä vinoutunut tekoäly voi tehdä virhepäätöksen esimerkiksi etnisen taustan tai sukupuolen vuoksi. Kolmas riski liittyy tekoälyn yksityisyyden- ja tietosuojaan. (Euroopan parlamentti, 2025.) Yksityisyyden- ja tietosuojan sekä vinoutuneen tekoälyn riskit nousevat esiin etenkin, kun tekoälyä hyödynnetään kouluissa.

Neljännän ja erityisen merkittävän riskin tekoäly luo ilmastolle. Tekoälyt toimivat suurten datakeskusten avulla. Datakeskusten rakentaminen ja tekoälyn ajaminen niiden kautta tuottaa suuri rasitteita ilmastolle. Datakeskuksen rakentaminen vaatii valtavia määriä raaka-aineita ja ne itsessään voivat olla jo ympäristölle haitallisia. Esimerkiksi koneet vaativat mikrosiruja, jotka sisältävät harvinaisia luonnonmateriaaleja, joita kerätään luontoa tuhoavilla tavoilla. Datakeskukset tuottavat myös suuria määriä elektroniikkaromua, mikä sisältää vaarallista jätettä kuten lyijyä. (UN environment programme, 2024.)

Datakeskukset käyttävät valtavia määriä vettä sekä rakennusvaiheessa että valmiina viilentääkseen koneita. Veden kulutus on niin suurta, että kaikkien datakeskusten arvioidaan käyttävän yhteensä enemmän vettä kuin kokonainen valtio. (UN environment programme, 2024.) Tutkija Shaolei Ren Kalifornian yliopistosta kertoi haastattelussaan, että ChatGPT kuluttaa noin puolilitraa vettä suorittaessaan yhden tehtävän, kuten tekstin tiivistämisen (Fingerhut & O'Brien, 2023). Datakeskukset kuluttavat veden lisäksi paljon energiaa. Energiaa tuotetaan usein ilmastoa kuormittavilla tavoilla kuten

polttamalla fossiilista jätettä, mikä tuottaa kasvihuonekaasuja. Yhden tekoälyllä tuotetun haun tekemiseen kuluu kymmenenkertainen määrä energiaa verrattuna yhteen Google-hakuun. (UN environment programme, 2024.)

Hälyttävänä esimerkkinä on Ruotsissa oleva Microsoftin tehdas, joka kuluttaa niin paljon sähköä, että se tuottaa ongelmia kaupungin yrityksille ja asukkaille. Paikalliset yritykset eivät pysty laajentumaan sähkö- tai tonttipulan vuoksi. Asukkaat puolestaan voivat kärsiä sähkökatkoista, jotka johtuvat datakeskuksien tuottamista sähkön kulutuspiikeistä. Lisäksi datakeskukset ovat kiinnostuneita laajentamaan toimintaansa muihin Pohjoismaihin halpojen tonttien ja halvan energian vuoksi. Kuitenkin datakeskuksien rakentaminen ja käyttö aiheuttaa merkittäviä ympäristö- ja yhteiskunnallisia ongelmia. (Velkova, ei pvm.)

Tekoäly syrjäyttää monia nykyisiä työtehtäviä, vaikka se myös luo uusia (Euroopan parlamentti, 2025b). Joidenkin ammattien ennustetaan olevan suuremmissa vaaroissa hävitä tai muuttua merkittävästi tekoälyn kehityksen myötä. Näitä ammatteja ovat esimerkiksi datan käsittelijät, tehdastyöntekijät, kääntäjät ja kirjanpitäjät, joiden työtehtäviä tekoäly kykenee automatisoimaan tehokkaasti. Sen sijaan ammatit, joissa vaaditaan vuorovaikutustaitoja tai ihmisläheistä työskentelyä ovat turvassa. Näitä ovat esimerkiksi opetustyö ja terveydenhuolto. Tekoäly ei pysty jäljittelemään tunneälyä, empatiaa tai vuorovaikutustaitoja kuin ihminen. (Gilmurray, 2024.)

Yhteenvetona voidaan todeta, että tekoäly tuottaa yhteiskunnalle monia mahdollisuuksia mutta myös riskejä. Riskit täytyy ottaa huomioon ja mahdollisuuksien mukaan myös vähentää niitä. Tekoälyn tuottamat riskit, esimerkiksi ilmastolle, ovat niin merkittäviä, että asiaan on keksittävä ratkaisu. Tarvittavien muutoksien avulla, tekoälystä voidaan luoda hyödyllinen apuväline, joka esimerkiksi tehostaa työntekoa, helpottaa arkea ja auttaa terveydenhuollossa.

### **2.3 Tekoälyn tuottamat mahdollisuudet ja riskit kouluissa**

Tutkimukset tekoälyn hyödyntämisestä opetuksessa ja oppimisessa ovat lisääntyneet merkittävästi viimeisien vuosien aikana (Nasir ym., 2024). Kuitenkin suomenkieliset tai Suomessa tehdyt tutkimukset tekoälystä koulukontekstissa ovat vähäisiä,

opinnäytetöitä lukuun ottamatta. Monet eri tutkimukset nostavat esiin samoja tekoälyn tuottamia riskejä ja mahdollisuuksia koulukontekstiin.

Suomessa opettajilla on pedagoginen vapaus toteuttaa opetusta parhaaksi näkemällään tavalla, opetussuunnitelman tavoitteet huomioiden (OAJ, ei pvm.). Perusopetuslaki ja perusopetuksen opetussuunnitelman perusteet eivät sisällä ohjeita, kuinka opetus täytyisi toteuttaa. Opetuslakiin on kirjattu esimerkiksi kaikille pakolliset oppiaineet ja oppivelvollisuus (Perusopetuslaki 628/1998). Opetussuunnitelman perusteisiin on puolestaan kirjattu esimerkiksi opetuksen tavoitteista (OPS, 2014). Opettaja voi siis itse päättää, kuinka paljon esimerkiksi tekoälyä tai teknologiaa hyödyntää opetuksessaan.

Nguyen ym. (2023) kirjallisuuskatsauksen mukaan tekoälyä hyödynnetään jo kouluissa. Heidän mukaansa tekoäly voi tehdä opetuksesta ja oppimisesta tehokasta. Tekoäly voi avustaa opettajaa opetuksen sekä eriyttämisen suunnittelussa ja toteutuksessa sekä arviointityössä. Puolestaan oppilaalle tekoäly mahdollistaa yksilöllisen avun oppimiseen ja opitun kertaamiseen. (Nguyen ym., 2023.) Nguyen ja tutkimuskumppaninsa (2023) eivät nostaneet esiin tekoälyn tuottamia haasteita.

Lo (2023) puolestaan analysoi 50 tutkimusartikkelia koskien ChatGPT:n käyttöä kouluympäristössä. Lo:n (2023) tutkimuksen tulokset ovat yhteneviä Nguyen ym. (2023) saamien tuloksien kanssa. Lo:n tutkimuksesta selviää, että opettajat ja oppilaat voivat käyttää ChatGPT:tä hyödyllisenä työkaluna, mutta sen käytössä on tunnistettu myös haasteita. Opettajat voivat käyttää ChatGPT:tä esimerkiksi opetusmateriaalin luonnissa, eriyttämisen apuna ja tukena kielenkäännöksissä. Sen avulla voidaan myös luoda tehtäviä ja kokeita tai hyödyntää sitä arviointityössä. Oppilaille ChatGPT voi tarjota tukea oppimisprosessiin ja kokeeseen valmistautumiseen sekä vastata kysymyksiin ja tiivistää tietoa. Se voi myös antaa palautetta oppilaan tehtävistä. (Lo, 2023.) Hyötyjen lisäksi, Lo (2023) tunnisti myös monia riskejä laajan kielimallin käyttöön liittyen. Yksi keskeinen riski liittyy ChatGPT:n antaman tiedon paikkansapitävyyteen. Annettu tieto voi olla vanhentunutta tai virheellistä. Myös plagioinnin mahdollisuus nostettiin riskiksi. Esimerkiksi oppilaat voivat käyttää plagioitua tietoa edes tietämättä

siitä, sillä tekoäly ei mainitse käyttämiään lähteitä. Lisäksi plagioinnintunnistusohjelmat eivät välttämättä tunnista, että vastaus on tuotettu tekoälyllä. (Lo, 2023.)

Raitasen (2024) kandidaatintutkimus luokanopettajaopiskelijoista tukee Lo:n (2023) havaintoja. Opettajaopiskelijat näkivät tekoälytyökalun potentiaalisena apuna erityisesti opetuksen ja oppimateriaalin suunnittelussa. Opiskelijat mainitsivat sen hyödyllisyyden myös esimerkiksi tekstin käsittelyssä ja hakukoneena. Opiskelijoiden mukaan oppilaat voisivat harjoitella tekoälyn käyttöä osana nykyaikaista oppimista. (Raitanen, 2024.) Kuten Lo:n (2023) tutkimuksessa, myös Raitasen (2024) tutkimuksessa nostettiin esiin tekoälyn käytön riskejä. Huolenaiheina nostettiin esiin esimerkiksi vaikeus tunnistaa tekoälyn tuottamaa tekstiä ja opettajan luovuuden sekä muiden kognitiivisten taitojen heikentyminen. Oppilaiden osalta keskeisin huolenaihe oli kognitiivisten taitojen heikentyminen. Lisäksi opettajaopiskelijat epäilivät tekoälyn tuottaman tiedon eettisyyttä ja luotettavuutta. (Raitanen, 2024.) Vaikka Raitasen (2024) tutkimuksen otanta oli pieni, antaa se viitteitä tulevien opettajien kriittisestä mutta kiinnostuneesta asenteesta kohti tekoälyn hyödyntämistä opetuksessa.

Chan ja Lee (2023) puolestaan tutkivat lähes 600 korkeakouluopettajan ja -opiskelijan näkemyksiä tekoälyn käytöstä opetuksessa ja oppimisessa. Chanin ja Leen (2023) tutkimuksen tulokset olivat yhteneviä Lo:n (2023), Raitasen (2024) ja Nguyen ym. (2023) tutkimuksien kanssa. Chanin ja Leen (2023) tutkimuksen mukaan opiskelijat suhtautuivat tekoälyn luomiin mahdollisuuksiin myönteisesti, kuten sen mahdollistamaan tehokkuuteen ja yksilölliseen oppimiseen. Opiskelijat sanoivat hyödyntävänsä tekoälyä monin tavoin opinnoissaan. Opettajat myös tunnistivat samat hyödyt, mutta olivat kuitenkin huolissaan oppilaiden liiallisesta tukeutumisesta tekoälyyn sekä sen eettisistä ja pedagogisista vaikutuksista. Opettajat painottivat, että perinteisiä ja teknologisia oppimismenetelmiä yhdistelemällä saataisiin tehokkainta oppimista. (Chan & Lee, 2023.)

Rane (2024) perehtyi tutkimuksessaan tekoälyyn vielä laajemmin, pohtien viiden aihealueen avulla tekoälyn luomia mahdollisuuksia, riskejä, eettisiä kysymyksiä ja kahden eri tekoälyn soveltuvuutta näihin aihealueisiin. Ranen (2024) tutkimuksessa nostettiin esiin hyvin samoja asioita kuin Chanin ja Leen (2023), Nguyen ym. (2023), Lo:n

(2023) ja Raitasen (2024) tutkimuksissa. Ranen (2024) tutkimuksen ensimmäinen aihealue koski personoitua oppimista. Tekoäly voi avustaa räätälöidyn opetuksen suunnittelussa ja toteutuksessa. Tämä kuitenkin luo Ranen mukaan tietosuojariskin, minkä vuoksi käytetyn tekoälyn pitäisi olla luotettava, jotta oppilaan personoidut tiedot olisivat turvassa. Toiseksi riskiksi hän nosti oppimisen yhteisöllisen ulottuvuuden heikentymisen. Ranen (2024) mukaan Gemini ja ChatGPT soveltuvat molemmat erittäin hyvin personoidun oppimisen suunnitteluun ja toteutukseen. Myös Nguyen ym. (2023), Lo:n (2023) ja Raitasen (2024) tutkimuksissa mainittiin personoitu oppiminen, mutta enemmän eriyttämisen näkökulmasta. Edellä mainitut tutkimukset eivät kuitenkaan nostaneet esiin samoja riskejä kuin Rane (2024) tässä aihealueessa.

Toinen Ranen (2024) tutkima aihealue oli opetuksen suunnittelu ja oppimateriaalin luonti. Tekoälyn avulla voi luoda nopeasti tuntisuunnitelmia tai tunneilla käytettäviä tehtäviä kuten monisteita tai kokeita. Tämä aihealue nostettiin esiin myös kaikissa muissa tarkasteltavissa tutkimuksissa (Nguyen ym., 2023; Lo, 2023; Raitanen, 2024; Chan & Lee, 2023). Rane (2024) kuitenkin huomautti, että tekoälyn hyödyntäminen voi johtaa ristiriitoihin tunti- ja opetussuunnitelman välille. Tätä riskiä ei muissa tutkimuksissa nostettu esiin. Lisäksi Rane (2024) kertoi riskiksi opettajan luovuuden heikentymisen, mutta myös Raitasen (2024) tutkimuksessa nostettiin tämä esiin. Eettiseksi haasteeksi Rane (2024) mainitsi tekoälyn mahdolliset vinoumat tai väärät tiedot. Vinoumat ja vääristymät näkyisivät esimerkiksi virheellisinä tehtävinä tai epäpätevinä tuntisuunnitelmina. ChatGPT:n luokiteltiin soveltuvan opetuksen suunnitteluun ja oppimateriaalin luontiin erittäin hyvin. Geminin puolestaan uskottiin soveltuvan vain hyvin. (Rane, 2024.)

Kolmantena Rane (2024) tutki tekoälyn soveltuvuutta arviointiin. Myös Nguyen ym. (2023) ja Lo (2023) nostivat esiin, että tekoälyä voisi hyödyntää arvioinnissa, mutta Rane (2024) pohti aihetta laajemmin. Toki Rane (2024) nosti esiin vain esseiden arvioinnin, vaikka tekoäly mahdollistaa monenlaisten tehtävien arvioinnin. Hänen mukaansa esseiden arvioiminen tekoälyllä luo riskin, että tekoäly ei osaa tulkita oppilaan ymmärrystä tai ajattelua niin kuin ihminen. Eettisenä riskinä uskotaan olevan, että tekoäly voi suosia tiettyjä ryhmiä. ChatGPT:n luokiteltiin soveltuvan esseiden arviointiin vain hyvin, ja Geminin puolestaan vielä heikommin. (Rane, 2024.) Rane (2024) ehdottaa,

että tulevaisuudessa hyödynnettäisiin arviointityökaluja, jotka yhdistävät tekoälyn ja ihmisen antaman arvioin. Sekä tekoälyn että ihmisen antama arviointi mahdollistaisi molempien vahvuuksien hyödyntämisen, jotta arvioinnista saadaan yhdenvertaista ja monipuolista.

Neljäs Ranen (2024) tutkima aihealue oli luovuuden tukeminen. Tekoäly mahdollistaa tuen aiheiden ideointiin. Rane (2024) luokitteli ChatGPT:n soveltuvan erittäin hyvin luovuuden tueksi. Puolestaan Geminin luokiteltiin soveltuvan vain hyvin. Riskinä kuitenkin on, että tekoälyn tukeudutaan liikaa, mikä voi heikentää mielikuvitusta ja ongelmanratkaisukykyä. (Rane, 2024.) Myös Raitasen (2024) tutkimuksen opettajaopiskelijat nostivat riskiksi luovuuden heikentymisen, sekä Chanin ja Leen (2024) tutkimuksen opettajat pelkäsivät oppilaiden tukeutuvan liikaa tekoälyyn. Luovuuden aihealueen eettiseksi riskiksi Rane (2024) nosti plagioinnin. Oppilas voi väittää tekoälyn tekemää tehtävää omakseen. Tekoäly ei myöskään mainitse käyttämiään lähteitä, jolloin oppilas voi plagioida myös alkuperäistä lähdeaineistoa. (Rane, 2024.) Myös Lo (2023) nosti esiin plagioinnin tuomat riskit, ja Raitasen (2024) tutkimuksen vastaajat pelkäsivät, että he eivät tunnista oppilaan tekoälyllä tuottamaa tekstiä.

Viimeisenä Rane (2024) tutki tekoälyn mahdollistamaa saavutettavuutta. Tekoäly mahdollistaa reaaliaikaisen kielenkääntämisen ja kirjoitetun kielen muuntamisen puheeksi. Haasteeksi kuitenkin nostettiin vaatimus internetistä. Lisäksi kielenkäännössovellukset saattavat sisältää kulttuurisia tai kielellisiä virheitä, minkä Rane (2024) arvioi eettiseksi riskiksi. ChatGPT:n ja Geminin luokiteltiin soveltumaan kielenkääntämiseen erittäin hyvin. (Rane, 2024.) Muut tarkastelun kohteena olevat tutkimukset eivät nostaneet saavutettavuuden osa-alueita, paitsi Lo:n (2023) tutkimuksessa mainittiin, että tekoälyä voi hyödyntää kielenkäännöksissä.

Yhteenvetona voidaan todeta, että tekoäly luo monia mahdollisuuksia, mutta myös riskejä kouluihin. Tekoäly tarjoaa opettajalle esimerkiksi tukea opetuksen suunnitteluun ja mahdollisesti myös arviointityöhön. Puolestaan oppilas voi saada tekoälyltä yksilöllistä apua tehtävien tekemiseen tai koekertaukseen. Kuitenkin tekoälyn käyttö voi vähentää niin opettajan kuin oppilaankin luovuutta ja kriittistä ajattelua. Lisäksi

tekoälyä hyödyntäessä on tärkeä muistaa plagioinnin mahdollisuus. Tutkijat Baidoo-Anu ja Owusu (2023) painottavat, että tekoälyä on vaikea kitkeä pois kouluista, ja siksi sitä pitäisi oppia käyttämään hyödyksensä.

#### **2.4 Opettajien ja oppilaiden näkemyksiä tekoälystä**

Tutkimukset opettajien ja oppilaiden näkemyksistä tekoälyä kohtaan ovat lisääntyneet viimeisen vuoden aikana. Tutkimukset osoittavat, että oppilaiden tekoälyn käyttö on lisääntymässä ja opettajien sekä oppilaiden suhtautuminen sitä kohtaan on vaihtelevaa. Lin (2024) tutkimus osoittaa, että yhdysvaltalaiset peruskoulun ja lukion opettajat suhtautuvat tekoölyyn melko kriittisesti. Neljäsosa tutkimukseen osallistuneista (n 2531) uskoi tekoälyn tuottavan enemmän harmia kuin hyötyä, kun vain kuusi prosenttia näki tekoälyn tuottavan enemmän hyötyä. Kolmasosa opettajista uskoi tekoälyn tuottavan yhtä paljon hyötyjä kuin haittojakin. Lukion opettajien suhtautuminen tekoälyä kohtaan oli kaikista kriittisintä. (Lin, 2024.) Puolestaan yhdysvaltalaiset peruskoulun ja lukion opiskelijat suhtautuvat tekoölyyn hieman positiivisemmin kuin opettajat (McClain ym., 2026). 36 prosenttia vastaajista uskoi tekoälyn tuottavan heille itsellensä vain positiivisia vaikutuksia, kun taas 15 prosenttia uskoi tekoälyn tuovan vain haittoja. Kuitenkin opiskelijoiden suhtautuminen muuttui hieman kriittisemmäksi, kun heiltä kysyttiin tekoälyn vaikutusta koko yhteiskunnalle. 26 prosenttia uskoi tekoälyn tuottavan yhteiskunnalle pelkkää haittaa, kun vastaava luku koskien heitä itseään oli vain 15 prosenttia. (McClain ym., 2026.)

Oppilaat, iältään 13–17-vuotiaat, nostivat tekoälyn negatiivisiksi vaikutuksiksi esimerkiksi luovuuden ja kriittisen ajattelun heikentymisen, virheellisen tiedon ja työpaikkojen vähenemisen (McClain ym., 2026). Myös opettajat ja luokanopettajaopiskelijat nostivat esiin samoja riskejä, paitsi he eivät maininneet työpaikkojen vähenemistä (esim. Chan & Lee, 2023; Raitanen, 2024). Tekoölyyn positiivisesti suhtautuneet oppilaat kokivat tekoälyn parantavan heidän opiskeluaan ja helpottavan heidän elämäänsä esimerkiksi tekemällä siitä tehokkaampaa (McClain, 2026). Opiskelijat kuitenkin kokivat, että eri tehtäviä on hyväksyttävämpää suorittaa tekoälyllä kuin toisia tehtäviä (Sidoti, Park & Gottfried, 2025.). Esimerkiksi yli puolet tutkimukseen osallistuneista (n 1391) koki hyväksyttävänä käyttää tekoälyä, kun

perehtyy uuteen aiheeseen. Puolestaan esseiden kirjoittamiseen tekoälyllä suhtauduttiin hyvin kielteisesti. Yli 40 prosenttia osallistujista koki sen kielletyksi, mutta kuitenkin vajaa viidesosa koki sen hyväksyttäväksi. Opiskelijoiden suhtautuminen tekoälyn käyttöön matematiikassa jakautui tasaisesti puolesta ja vastaan. 29 prosenttia koki sen hyväksyttäväksi, kun taas 28 prosenttia ei sallisi sitä. (Sidoti, Park & Gottfried, 2025.)

Sidotin, Parkin ja Gottfriedin (2025) tutkimuksen mukaan yhdysvaltalaisen oppilaiden tekoälyn käyttö on lisääntynyt. Vielä vuonna 2023 vain 13 prosenttia kertoi hyödyntävänsä ChatGPT:tä opinnoissaan, kun vuonna 2024 tulos oli jo 26 prosenttia (Sidoti, Park & Gottfried, 2025). Uudemman kyselytutkimuksen (Faverio & Sidoti, 2025) mukaan yhdysvaltalaisista nuorista jo 64 prosenttia hyödyntää tekoälyä opinnoissaan, näistä 28 prosenttia hyödyntää tekoälyä yhden tai useamman kerran päivässä. Tekoälyn käyttäminen on hieman todennäköisempää lukioikäisillä kuin nuoremmilla oppilailla (Sidoti, Park & Gottfried, 2025). Puolestaan sukupuolieroja ei ole havaittu tekoälyn hyödyntämisessä (Faverio & Sidoti, 2025).

Osalla yliopisto-opiskelijoista on huolestuttava näkemys tekoälyn vaikutuksista opettajiensa työhön. Okulich-Kazarin ym. (2024) tutkimus osoittaa, että osa Itä-Euroopan yliopisto-opiskelijoista uskoo tekoälyn korvaavan heidän opettajansa. Vastajat olivat Kazakstanista, Puolasta, Slovakiasta ja Ukrainasta. Aineisto koostui lähes 600 vastauksesta, ja niistä yli 10 prosenttia uskoi, että tekoäly tulisi opettamaan heitä viiden vuoden sisällä.

### 3 Oppilaan osaamisen arviointi

Perusopetuksessa arvioidaan oppilaan oppimista, osaamista, työskentelyä ja käyttäytymistä (OPS, 2014). Tämä tutkielma keskittyy osaamisen arviointiin, sillä etenkin kyseisen osa-alueen arviointiin liittyy haasteita ja se on aikaa vievää. Opettajien yleisimmin käytetyt arviointimenetelmät ovat kirjalliset kokeet, esseet tai kirjoitelmat (Atjonen, ym., 2019, s.111 ja 217), joiden suunnitteluun ja tarkistamiseen kuluu opettajalta huomattavasti aikaa. Kun puolestaan työskentelyn ja käyttäytymisen arviointi voi tapahtua opetuksen ohessa.

Opetussuunnitelman perusteisiin (OPS, 2014) on määritelty arvioinnin yleiset periaatteet, joita jokaisen opettajan tulee noudattaa sekä päättöarvioinnissa että yksittäisten tehtävien arvioinnissa. Arvioinnin on oltava esimerkiksi yhdenvertaista, monipuolista, suunnitelmallista ja johdonmukaista. Lisäksi arviointi edellyttää avoimuutta ja yhteistyötä oppilaan kanssa. Keskeistä on se, että arviointi perustuu ainoastaan ennalta määriteltyihin tavoitteisiin ja kriteereihin. (OPS, 2014.)

Kasvatustieteiden professori ja pitkän tutkijauran tehnyt Päivi Atjonen pohtii kirjassaan Hyvä, paha arviointi (2007) arvioinnin peruslähtökohtia, jotka sivuavat opetussuunnitelman perusteissa mainittuja arvioinnin yleisiä periaatteita. Atjonen (2007, s.34) nostaa arvioinnin peruslähtökohdiksi seuraavat aiheet: reiluus, oikeudenmukaisuus, läpinäkyvyys sekä validiteetti ja reliabiliteetti.

Atjonen (2007, s.34) kertoo arvioinnin reilouden ilmenevän oppilaiden samoista mahdollisuuksista menestyä. Reiluutta on esimerkiksi se, että oppimisessa tukea tarvitsevalle oppilaalle tarjotaan tukiopetusta, jotta hän voisi menestyä hyvin. Reiluus ei synny siitä, että kaikille tarjotaan samaa, vaan tukea annetaan sitä tarvitseville. Atjonen (2007, s.34) kuitenkin nostaa esiin, että esimerkiksi lukihäiriöiset joutuvat usein heikompaan asemaan, sillä kirjallisia kokeita suositetaan suullisia enemmän. Arvioitavien tehtävien monipuolisuus on siksi tärkeää, että oppilaat voivat näyttää osaamistaan monin eri tavoin. Arvioinnin monipuolisuus mainitaan opetussuunnitelmassa (OPS, 2014).

Atjosen (2007, s.34) mukaan oikeudenmukainen arviointi arvioi oppilaita vain sen mukaan, kuinka he ovat saavuttaneet ennalta määrätyt tavoitteet. Arviointi ei saa

esimerkiksi suosia tiettyjä ryhmiä tai perustua oppilaan persoonaan. Ouakrim-Soivio (2013) lisääkin, että oikeudenmukaisuus vaatii yhdenvertaisuutta, jolla tarkoitetaan arvioinnin arvovapautta ja puolueettomuutta. Ouakrim-Soivio (2013) mukaan arvioinnin yhdenvertaisuudessa on kyse siitä, että saavatko oppilaat yhdenvertaisin perustein ja samalla osaamistasolla saman arvosanan. Arvioinnin yhdenvertaisuus nousee yhdeksi arvioinnin keskeisimmäksi haasteeksi (Ouakrim-Soivio, 2013).

Arvioinnin validiteetti tarkoittaa sitä, että arvioidaan juuri sitä mitä pitääkin (Atjonen, 2007, s.34–35). Osaamista mittaavan tehtävän on mitattava nimenomaan sitä taitoa tai tietoa, joka tavoitteisiin on kirjattu. Jos tehtävä mittaa jotain muuta, tai arvioinnissa otetaan huomioon muita aiheita, ei arviointi ole validia. Puolestaan arvioinnin reliabiliteetti tarkoittaa arvioinnin yhdenmukaisuutta ja ei-sattumanvaraisuutta. Eri opettajien antamat arvosanat pitäisivät olla keskenään yhteneviä, mikäli oppilailla on sama osaaminen. Arvioinnit eivät myöskään saa olla sattumanvaraisia, vaan ne pitäisivät olla johdonmukaisia arvioinnista toiseen. Tämä tarkoittaa sitä, että saman osaamisen tulee tuottaa samoja arvosanoja riippumatta siitä, milloin tai missä tilanteessa arviointi tehdään. Oppilaan suoritusta ei saa tulkita eri tavoin eri päivinä tai tilanteissa eikä arvioinnin tulisi riippua esimerkiksi opettajan mielialasta tai arviointitilanteen satunnaisista tekijöistä. (Atjonen, 2007, s.34–35.)

Viimeisenä Atjonen (2007, s.35) nostaa arvioinnin läpinäkyvyyden. Hänen ajatuksessansa korostuu yllätyksien välttämäinen. Arvioinnin tulee perustua opetussuunnitelmassa esitettyihin tavoitteisiin, oppilaan pitää olla tietoinen tavoitteista ja tehtävän arviointikriteereistä. Oppilaan ei pitäisi esimerkiksi joutua arvelemaan mitä asioita hänen esseestään tullaan arvioimaan.

### **3.1 Laaja kielimalli tekstiaineistojen arvioijana**

Tietokoneavusteinen arviointi (Computer aided assessment, CAA) tarkoittaa mitä tahansa arvioinnin tapaa, missä hyödynnetään tietokonetta. Tietokoneavusteista arviointia voi toteuttaa ohjelmistolla, joka hyödyntää tekoälyä tai vain perinteistä ohjelmointia. (Chalmers & McAusland, ei pvm.) Perinteistä ohjelmointia hyödynnetään yksinkertaisten tehtävien, kuten monivalintatehtävien, arvioinnissa. Ohjelmisto vertaa annettua vastausta entuudestaan oikeaksi määrättyyn vastaukseen. Tekoälyn

hyödyllisyys kasvaa, kun aletaan arvioimaan monimutkaisempia tehtäviä, kuten lyhyitä kirjallisia vastauksia tai esseitä.

Seuraavaksi esitellään tutkimuksia, jotka ovat tutkineet laajojen kielimallien käyttöä arvioinnissa. Tutkimuksia etsittiin Googlen Scholar ja Turun yliopiston Volter hakupalveluista. Tutkimuksia etsittiin esimerkiksi seuraavanlaisin hakusanoin: AI/LLM in school, AI/LLM in evaluation, Local AI/LLM in evaluation, Large language model in essay evaluation. Pyrkimys oli etsiä sekä suomenkielisiä että englanninkielisiä tutkimuksia, mutta kyseisistä palveluista löytyi suomeksi vain pro gradu tason tutkimuksia. Lisäksi pyrittiin etsimään tutkimuksia, jotka tarkastelevat monien erilaisten pilvipohjaisten ja paikallisten laajojen kielimallien toimintaa arvioinnissa. Kuitenkin suuri osa tutkimuksista keskittyi OpenAI:n pilvipohjaisen ChatGPT:n tarkasteluun. Esiteltäviä tutkimuksia ei valittu systemaattisesti, vaan ne valittiin otsikon perusteella tarkempaan tarkasteluun. Etsinnät lopetettiin, kun koettiin, että on tarpeeksi kattavasti eri tutkimuksia.

Seuraavaksi esiteltävät tutkimukset tutkivat enimmäkseen laajojen kielimallien soveltuvuutta esseiden arviointiin. Vaikka tässä tutkimuksessa arvioidaan lyhyitä koevastauksia, antavat aiemmat tutkimukset viitteitä siitä, miten laajat kielimallit suoriutuvat tekstivastauksien arvioinnista. Suurin osa tutkimuksista keskittyy ChatGPT:n eri versioihin, ja arvioitavat esseet ovat englanninkielisiä. Huomionarvoista on, että vain kolme tutkimusta seitsemästä kertoi käyttämänsä arviointipromptin. Lisäksi vain yksi tutkimus seitsemästä kertoi kehoitesuunnittelun vaiheistaan. Tämä vaikeuttaa tutkimuksien toistettavuutta ja tuloksien arviointia, sillä arviointipromptilla on merkittävä vaikutus tutkimustuloksiin. Taulukkoon 1 on kerätty esiteltävien tutkimuksien keskeiset tiedot.

Taulukko 1. Keskeiset tiedot tutkimuksista, jotka ovat tutkineet laajojen kielimallien käyttöä arvioinnissa. Tutkimukset ovat esitetty aakkosjärjestyksessä.

Tutkijat	Tutkimuksen tavoite	Tekoäly	Aineiston tyyli ja kieli	Aineiston määrä (kpl)	Arvioinnin kohteet ja pisteytys	Tutkimuksen tulos
<b>Almegren ym. (2024)</b>	Vertailla keskenään ChatGPT:n,	Chat-GPT,	Essee, englannin kieli	30	Neljä kategoriala, jotka pisteytettiin 0–5. Kategoriat:	Tekoälyn arviointi epäluotetta-

Tutkijat	Tutkimuksen tavoite	Tekoäly	Aineiston tyyli ja kieli	Aineiston määrä (kpl)	Arvioinnin kohteet ja pisteytys	Tutkimuksen tulos
	Geminin, Bingin ja ihmisarvioijien antamia arvioita esseistä.	Gemini, Bing			sanasto, ideat ja sisältö, jäsentely ja johdonmukaisuus sekä tarkkuus kuten kielioppi.	vaa. Tekoälyn ja ihmisten antamat arviot eivät ole yhteneviä.
<b>Barrot &amp; Bui (2024)</b>	Vertailla keskenään ChatGPT:n ja ihmisarvioijien antamia arvioita esseistä sekä tutkia ChatGPT:n tuottamien arvioiden johdonmukaisuutta kahtena ajankohtana.	ChatGPT 3.5.	Essee, englannin kieli	200	Viisi kategorialla, joista jokainen pisteytettiin 1–5. Kategoriat: väitteen valinta, argumentaatio, kohderyhmä, yhtenäisyys ja tekstin tyyli.	Tekoälyn arviointi epäluotettavaa. Tekoälyn ja ihmisten antamat arviot eivät ole yhteneviä. ChatGPT:n tuottamat arviot eivät ole yhteneviä kahtena ajankohtana.
<b>Fokides &amp; Peristeraki (2025)</b>	Vertailla kahden eri kielisen esseen arviointeja, joita ChatGPT ja ihmisarvioijat ovat antaneet.	ChatGPT v4 turbo	Essee, kreikan ja englannin kieli	40	Palaute kielioppivirheistä sekä esseen vahvuuksia ja heikkouksia. Yleisarvosana asteikolla 1–10.	Tekoäly ei pysty arvioimaan kreikankielisiä esseitä luotettavasti, mutta englanninkielisiä paremmin kuin ihmisarvioijat.
<b>Kincl ym. (2024)</b>	Vertailla keskenään ChatGPT:n, Clauden ja ihmisarvioijien antamia arviointeja esseistä.	ChatGPT 4 ja Claude (sonet) 3.5	Essee, tšekin kieli	30	Palaute numeerisena ja kirjallisena. Kuusi kategorialla, jotka pisteytettiin eriarvoisesti. Yhteispisteet 50. Kategoriat: muotovaatimukset (7 p.), aiheenvalinta (5p.), lähteet (8p.), aiheen määrittelmä (8p.), ymmärryksen osoitus (10 p.) ja aiheen yhteys muihin	Tekoälyn arviointi epäluotettavaa. Tekoälyn ja ihmisten antamat arviot eivät ole yhteneviä. Myös yhden tekoälyn tekemät arviot eivät ole yhteneviä keskenään.

Tutkijat	Tutkimuksen tavoite	Tekoäly	Aineiston tyyli ja kieli	Aineiston määrä (kpl)	Arvioinnin kohteet ja pisteytys	Tutkimuksen tulos
					käsitteisiin (12p.).	
<b>Magro ym. (2025)</b>	Luoda oma tekoäly hyödyntäen ChatGPT 4 mallia sekä vertailla keskenään mallin ja ihmisarvioijien antamia arviointeja muistiinpanoista.	Itse luotu/ ChatGPT 4	Muistiinpano, englannin kieli	270	Ei tiedossa	Tekoälyn tuottamat arviot yhteneviä ihmisarvioijien kanssa. Tekoälyn tuottamat arvioinnit johdonmukaisempia kuin ihmisen.
<b>Sadanand ym. (2022)</b>	Luoda tekoäly, joka arvioi esseitä.	Itse luotu	Essee, englannin kieli	Koulu-tettu 12000 esseellä	Sisältö, kielioppivirheet, plagiointi ja asenne	Tekoälyn tuottamat arviot olivat lähes yhteneviä ihmisarvioijien kanssa.
<b>Wetzler ym. (2025)</b>	Vertailla keskenään ChatGPT:n ja ihmisarvioijien antamia arviointeja esseistä.	ChatGPT 3.5 ja 4o	Essee, englannin kieli	186	Viisi kategorialla, jotka pisteytettiin eriarvoisesti. Yhteispisteet 100. Kategoriat: johdanto (10 p.), analyysi (25 p.), reflektointi (25 p.), tulevaisuuden sovellus ja johtopäätökset (25 p.) sekä kirjoitustyyli ja rakenne (15 p.).	Tekoälyn arviointi epäluotettavaa. Tekoälyn ja ihmisten antamat arviot eivät ole yhteneviä.

Bouziane ja Bouziane (2024) tutkivat ChatGPT:n hyödyntämistä esseiden arvioinnissa.

He eivät tarkentaneet mitä ChatGPT:n versiota käyttivät. He valitsivat sattumanvaraisesti sata yliopisto-opiskelijan kirjoittamaa esseitä, joista opettajat ja tekoäly tuottivat arvioinnit. Esseistä tarkasteltiin 11 eri kategorialla, jotka arvioivat esseiden sisältöä, jäsentelyä, johdonmukaisuutta ja sen yleistä laatua kuten kirjoitusasua. Kukin kategoria pisteytettiin nolasta viiteen. Opettajien ja ChatGPT:n tuottamia arviointeja verrattiin keskenään. Ensimmäisen arviointikierroksen jälkeen

uudet opettajat tarkastivat sekä opettajien että tekoälyn antamat arviot varmistaakseen kattavan ja puolueettoman arvioinnin. (Bouziane & Bouziane, 2024.)

Bouziane ja Bouziane (2024) tutkimuksen tuloksien mukaan ChatGPT suoriutui erittäin hyvin esseiden arvioinnista. Laaja kielimalli suoriutui yhtä hyvin tai jopa paremmin kuin ihminen suurimman osan kategorioiden arvioinnista. Malli suoriutui poikkeuksellisen hyvin esimerkiksi kieliopin, lauserakenteiden ja perusteluiden arvioinnissa. Ihminen puolestaan ylitti laajan kielimallin taidot vain johdonmukaisuuden arvioinnissa.

(Bouziane & Bouziane, 2024.) Bouziane ja Bouziane (2024) tutkimuksen tuloksia ei voida yleistää koskemaan muun kielisten tehtävien arviointia. Bouziane ja Bouziane (2024) käyttivät vain englanninkielisiä aineistoja. ChatGPT ei välttämättä kykene arvioimaan esimerkiksi suomenkielisiä esseitä yhtä hyvin.

Myös monet muut ovat tutkineet ChatGPT:tä esseiden arvioinnissa. Wetzler ym. (2025) vertailivat keskenään ihmisarvioijien ja kahden eri ChatGPT:n mallien (3.5 ja 4o) antamia pisteytyksiä. Arvioinnin kohteena olivat lähes 200 psykologiaopiskelijan esseitä aiheesta romanttiset suhteet. Esseet arvioitiin 100 pisteen asteikolla, jossa eri osiot painottuivat eri tavoin. Arviointimatriisi pohjautui vahvasti kyseiseen aiheeseen, eikä sitä voisi hyödyntää toisen tehtävän arvioinnissa sellaisenaan. Yhteensä 85 pistettä koostui esseen eri osien pisteytyksestä, joissa pyydettiin perehtymään eri asioihin. Osat olivat johdanto, analyysi, reflektio ja tulevaisuuden sovellus sekä johtopäätökset. Viimeiset 15 pistettä annettiin kirjoitustyylistä ja rakenteesta. (Wetzler ym., 2025.)

Wetzler ym. (2025) tutkimuksen mukaan laaja kielimalli ei suoriutunut niin hyvin kuin Bouziane ja Bouziane (2024) tutkimuksessa. Wetzler ym. (2025) analysoivat arviointeja monivaiheisella Blant-Altman analyysillä, joka on luotu vertailemaan kahta mitattavaa. Blant-Altman analyysit osoittivat, että laajan kielimallin ja opettajien antamien arvioiden välinen yhteneväisyys oli vähäistä. Esimerkiksi laaja kielimalli antoi keskimäärin heikommille esseille parempia arvosanoja, kun taas keskimääräistä paremmille esseille tiukempia arvosanoja kuin opettajat. Tutkijat pohtivat heikkojen tuloksien johtuvan esimerkiksi opettajien taipumuksesta arvioida ensivaikutelman perusteella, ihmisten kyvystä arvostaa erinomaisia esseitä tai laajan kielimallin promptauksesta ja koulutusdatassa mahdollisesti ilmentyvistä virheistä. (Wetzler ym., 2025.)

Almegren ym. (2024) puolestaan tutkivat ChatGPT:n lisäksi myös Googlen Geminiä ja Microsoftin Bingiä (nykyään Copilot). He eivät tarkentaneet, mitä laajojen kielimallien versioita käytettiin. Heidän tutkimusasetelmansa oli samanlainen edellisten kanssa, että vertaillaan ihmisarvioijien ja laajan kielimallin tuottamia arvioita keskenään. He vertailivat myös eri laajoja kielimalleja keskenään. Arvioinnin kohteena oli 30 esseetä, jotka yliopisto-opiskelijat olivat kirjoittaneet. Opiskelijat kirjoittivat esseet englanniksi, vaikka se ei ollut heidän äidinkieltensä. Esseiden aiheena oli teknologia opiskelijan päivittäisessä elämässä. Esseet arvioitiin 20 pisteen asteikolla, josta kukin neljästä kategoriasta oli maksimissaan viiden pisteen arvoinen. Kategoriat olivat sanasto, ideat ja sisältö, jäsentely ja johdonmukaisuus sekä tarkkuus, joka sisälsi mm. kieliopin ja oikeinkirjoituksen. Tekoälysovellukset ChatGPT, Gemini ja Bing sekä kolme opettajaa arvioivat samat esseet kyseisellä arviointimatriisilla. (Almegren ym., 2024.)

Almegren ym. (2024) tulokset puolsivat Wetzler ym. (2025) tutkimuksen tuloksia. Almegren ym. (2024) tutkimuksen tulokset osoittivat merkittäviä eroja arvioijien keskiarvoissa. Ihmisarvioijat antoivat johdonmukaisesti korkeampia pisteitä kuin laajat kielimallit. Tutkijoiden mukaan se korostaa ihmisten erilaista ymmärrystä esseiden arvioinnissa kuten suurempi arvostus esseiden laadullisia puolia kohtaan. Lisäksi ihmisarvioijat olivat parempia arvioimaan subjektiivisia asioita kuten sisällön rikkautta, sanaston käyttöä ja kokonaislaatua. Laajat kielimallit puolestaan soveltuivat paremmin arvioimaan objektiivisia osa-alueita, kuten tekstin rakennetta ja tiedon tarkkuutta. ChatGPT suoriutui Geminiä ja Bingiä paremmin sanaston arvioinnissa. ChatGPT ja Gemini molemmat loistivat sisällön arvioinnissa. (Almegren ym., 2024.) Almegren ym. (2024) tutkimuksen pieni, 30 esseen, aineisto haastaa tuloksien yleistettävyyttä.

Barrot ja Bui (2024) aikaisempien tutkijoiden tavoin vertailivat ihmisarvioijien ja ChatGPT 3.5 mallin tuottamia arvioita esseistä. Tämän lisäksi he tutkivat laajan kielimallin johdonmukaisuutta, syöttämällä saman promptin ja esseen tekoälylle kahtena eri ajankohtana. Tutkimuksen aineisto koostui 200 englanninkielisestä esseestä. Englanti ei ollut kirjoittajien äidinkieli, vaan he olivat kotoisin monista eri Aasian maista. Esseitä valittiin neljästä eri englannin kielen taitotasosta. Barrot ja Bui (2024) myötäilivät arviointitaulukkoa luodessaan Common Core State Standards (Turnitin, ei pvm.) argumentaatioisten esseiden kirjoituksen kriteeristöä. Esseet pisteytettiin viidessä eri

kategoriassa: väitteen valinta, argumentaatio, kohderyhmä, yhtenäisyys ns. punainen lanka sekä tekstin tyyli. Kohderyhmällä viitattiin siihen, että kirjoittaja osaa ennakoida esimerkiksi tekstin kohderyhmän tietotasoa ja ennakkoluuloja. Jokaisesta kategoriasta voi saada 1–5 pistettä.

Barrotin ja Buin (2024) tutkimuksen tulokset puoltavat aikaisemmin esiteltyjen tutkimuksien tuloksia, että laaja kielimalli ei kykene arvioimaan ihmisen tasoisesti. Niin kuin Almegren ym. (2024) tutkimuksessa, Barrotin ja Buin (2024) tutkimuksessa ihmisarvioijat antoivat keskimäärin korkeampia pisteitä kuin laaja kielimalli. Laaja kielimalli ei myöskään kyennyt arvioimaan johdonmukaisesti, sillä ensimmäisen ja toisen arviointikierroksen kokonaispisteiden keskiarvot heittivät 1,3 pisteellä. Almegren ym. (2024) tavoin Barrot ja Bui (2024) pohtivat ihmisten antamien korkeampien arvosanojen johtuvan ihmisen kyvystä arvostaa enemmän esimerkiksi luovuutta, kun taas laaja kielimalli keskittyy tekniseen oikeellisuuteen. Barrot ja Bui (2024) pohtivat arviointien heikon yhteneväisyyden johtuvan esimerkiksi laajan kielimallin koulutusdatasta ja kyvystä tulkita arviointimatriisia. Ihmisarvioijia oli vain yksi, minkä vuoksi hänen tekemät arvioinnit ovat oletettavasti yhteneviä. He mainitsevat, että tuloksia voidaan yleistää suuren aineiston vuoksi. (Barrot & Bui, 2024.) Kuitenkin tuloksien yleistettävyyttä haastaa aineiston kieli ja tuloksien sidoksisuus kontekstiin. Tutkimus toteutettiin vain englanninkielisillä esseillä, minkä vuoksi ei voida arvioida laajan kielimallin kykyä suoriutua eri kielisten esseiden arvioinnista. Lisäksi tulokset ovat vahvasti kontekstisidonnaisia, sillä käytetty promptti ja arviointimatriisi vaikuttavat merkittävästi tutkimustuloksiin.

Kincl ym. (2024) tutkivat myös laajan kielimallin ja opettajien antamia arvioita oppilaiden esseistä. Merkittävintä tässä tutkimuksessa on se, että esseet olivat tšekin kielellä. Palautetta pyydettiin myös kirjallisessa muodossa eikä vain numeerisesti niin kuin aikaisemmin esitetyissä tutkimuksissa. Tutkimuksessa käytettiin ChatGPT 4 ja Clauden maksullisia versioita. Tutkimus sisälsi 30 yliopisto-opiskelijan kirjoittamaa esseitä. Esseiden aiheena oli edistyneen markkinoinnin käsite ja sen sovellus teoriaan tieteelliseen kirjallisuuteen tukeutuen. Esseet arvioitiin 50 pisteen asteikolla, jossa eri osiot painottuivat eri tavoin. Osiot olivat muotovaatimukset, aiheen valinta, lähteet, aiheen määritelmä, ymmärryksen osoitus ja aiheen yhteys muihin käsitteisiin. Osiot

ovat kerrottu pisteytysjärjestyksessä. Viimeisimpänä mainitusta osiosta pystyi saamaan eniten pisteitä.

Kinclin ym. (2024) tutkimuksen tulokset poikkesivat aikaisemmin esitetyistä. Laajojen kielimallien antamat pisteytykset olivat keskimääräistä korkeampia kuin ihmisarvioijilla. Lisäksi laajojen kielimallien antamat palautteet olivat vähemmän kriittisiä. Tutkimus osoittaa, että kyseiset laajat kielimallit eivät pysty arvioimaan tšekin kielellä olevia esseitä luotettavasti. Arviot eivät olleet riittävän yhdenmukaisia ihmisarvioiden ja eri laajojen kielimallien välillä. Lisäksi yhden mallin tuottamissa arvioinneissa esiintyi epäjohdonmukaisuutta ja satunnaisuutta. (Kincl ym., 2024.)

Fokides ja Peristeraki (2025) tutkivat Kinclin ym. (2024) tavoin laajan kielimallin kykyä arvioida muun kuin englanninkielisiä esseitä. Fokides ja Peristeraki (2025) vertailivat palautetta, mitä laaja kielimalli ja ihmisarvioijat antoivat sekä kreikan- että englanninkielisille esseille. Tämä tutkimus poikkeaa senkin vuoksi valtavirrasta, että esseet oli kirjoittanut 8–11-vuotiaat eivätkä yliopisto-opiskelijat. Tutkimuksessa käytettiin ChatGPT:n versiota v4 turbo. Heidän tarkoituksenaan oli käyttää ChatGPT:n vanhempaa 3.5 versiota, mutta esitutkimus osoitti, että kyseinen malli ei pysty arvioimaan kreikan kieltä. Tutkimuksen aineisto koostui 40 esseestä, joista puolet oli kreikankielisiä ja loput englanninkielisiä. 20 opettajaa tarkisti kukin joko 20 kreikan- tai englanninkielistä esseitä, riippuen heidän kielitaidostaan. Tekoälyä ja opettajia pyydettiin etsimään kielioppi-, lauseoppi- ja kirjoitusvirheitä sekä antamaan palautetta tekstin vahvuuksista ja heikkouksista. Näiden lisäksi pyydettiin antamaan kokoava palaute ja arvosana asteikolla 1–10.

Fokidesin ja Peristerakin (2025) tutkimuksen tulokset puoltavat Kinclin ym. (2024) tutkimuksen tuloksia, että ChatGPT ei pysty arvioimaan luotettavasti ei-englanninkielisiä esseitä. Fokidesin ja Peristerakin (2025) mukaan ChatGPT arvioi oikein kirjoitettuja sanoja ja lauseita virheelliseksi, keskittyi vähemmän kielimekaniikkaan ja antoi epätasaisempaa palautetta. Laaja kielimalli antoi kuitenkin yksityiskohtaisempaa palautetta kuin ihminen. Puolestaan englanninkielisiä esseitä ChatGPT arvioi jopa paremmin kuin ihminen. Laaja kielimalli huomasi enemmän virheitä, antoi yksityiskohtaisempaa palautetta sekä palautteen tyyli oli positiivisempaa ja

akateemisempaa kuin ihmisarvioijan. Fokidesin ja Peristerakin (2024) tutkimuksessa ChatGPT:n ja ihmisarvioijien antamien arviointien keskiarvot kreikankielisistä esseitä olivat lähes samoja. Puolestaan Kinclin ym. (2024) tutkimuksessa tekoälyt antoivat keskimäärin korkeampia pisteitä kuin ihmisarvioijat. Yllättäen Fokidesin ja Peristerakin (2024) tutkimuksessa englanninkielisten esseiden arvosanojen keskiarvojen ero oli 1,3, kun taas kreikankielisten esseiden vain 0,46.

Tutkijat ovat tutkineet valmiiden laajojen kielimallien kuten ChatGPT:n käyttöä arviojana mutta he ovat myös luoneet omia tekoälyohjelmistoja esseiden arviointiin. Sadanand ym. (2022) ovat luoneet mallin, jonka avulla esseitä voidaan arvioida tarkemmin, puolueettomammin ja nopeammin. Malli hyödyntää erästä neuroverkon tyyliä (LSTM) ja koneoppimista. Järjestelmä perustuu neuroverkkoon, joka koulutettiin lähes 12 000 esseellä, jotka olivat kahden ihmisen arvioimia. Malli arvioi esseiden numeerisesti esseiden sisällön, kielioppivirheiden, plagioinnin ja asenteen pohjalta. Aineistosta mitattiin Quadratic Weighted Kappa (QWK), joka mittaa kuinka hyvin malli vastaa ihmisarvioijia. Arvoksi saatiin 0,911, mikä viittaa erittäin korkeaan yhteensopivuuteen. Eli mallin antamat arviot, olivat lähes yhtä yhteneviä, kuin ihmisten väliset arvioinnit. Mallin suurikokoinen data on merkittävä yleistyksen kannalta. (Sadanand ym., 2022.) Kriittikkiä herättää se, että jos malli on koulutettu vain samankaltaisilla teksteillä, kuten englanninkielisillä esseillä. Tällöin sen toimivuudesta esimerkiksi suomenkielisille teksteille ei tiedetä. Lisäksi plagioinnintunnistuksen perustuessa vain avoimiin lähteisiin, se voi jättää huomaamatta maksimumin takaa kopioidut sisällöt (Sadanand ym., 2022).

Magro ym. (2025) raportoivat yrityksestään luoda oma laaja kielimalli (GradeGPT), joka hyödyntää valmista ChatGPT 4 mallia. Kokeilusta on nähtävillä vain lyhyt raportti, minkä vuoksi tiedot tutkimuksen toteutuksesta ovat vajaita. Esimerkiksi raportissa ei kerrota GradeGPT:n luonnista. Siinä kerrotaan vain, että malli luotiin hyödyntämällä yksityistä versiota ChatGPT 4 mallista. Kysymysmerkiksi jää, mitä yksityinen versio tarkoittaa. GradeGPT (Magro ym., 2025) luotiin arvioimaan lääketieteen opiskelijoiden kliinisen tutkimuksen muistiinpanoja. Raportissa ei mainita arvioinnin kohteena olevien muistiinpanojen pituudesta. Vaikka muistiinpanot eivät ole esseitä, antaa se viitteitä tekoälyn kyvyistä tulkita ja noudattaa arviointimatriisia sekä luoda arviointeja.

GradeGPT mallia testattiin arvioimalla hieman yli 270 muistiinpanoa. Raportin mukaan ihmisen ja GradeGPT:n tekemät arviot korreloivat vahvasti keskenään. Tämä tarkoittaa, että ihmisarvioijan ja laajan kielimallin tekemät arviot olivat yhteneviä. Laajan kielimallin mainittiin olevan johdonmukaisempi arvioinneissaan kuin ihmisarvioijat. (Magro ym., 2025.) Tämä tulos on päinvastainen Barrotin ja Buin (2024) tutkimuksen tuloksien kanssa. GradeGPT:n antaman kirjallisen palautteen kerrottiin olevan vaihtelevaa ja luettavaa, eikä geneeristä (Magro ym., 2025). Raportti jätti kuitenkin epäselväksi, oliko kirjallinen palaute yhtenevä numeerisen arvioinnin kanssa ja keskittyikö se oikeisiin asioihin.

Esiteltyjen tutkimuksien perustella voidaan todeta, että laajat kielimallit eivät yleisesti pysty arvioimaan oppilaiden tuotoksia luotettavasti. Kuitenkin useat tutkimukset osoittivat, että laajat kielimallit suoriutuvat tekstin kieliopillisesta ja rakenteellisesta arvioinnista. Lisäksi muutama tutkimuksista saavutti tuloksia, jossa laajan kielimallin tuottama arviointi oli yhtenevää ihmisarvioijan kanssa. Tämä viittaa siihen, että tekoälyllä tuotettu arviointi voi tietyissä olosuhteissa olla luotettavaa. Tämä suppeasti toteutettu kuvaileva kirjallisuuskatsaus osoittaa, että laajojen kielimallien soveltuvuutta arviointiin on edelleen syytä tutkia.

### **3.2 Opettajan ja tekoälyn tuottaman arvioinnin riskit**

Opettajien tuottama arviointi ei aina noudata Atjosen (2007) kuvailemia arvioinnin peruslähtökohtia. Opettajien tuottamassa arvioinnissa ilmenee riskejä esimerkiksi yhdenvertaisuuden näkökulmasta (Ouakrim-Soivio, 2013), sekä arviointiin vaikuttaa moni siihen kuulumaton asia (Atjonen ym., 2019). Näitä asioita ovat esimerkiksi oppilaan asenne, muiden oppilaiden osaaminen ja opettajan ennakkoluulot (Atjonen ym., 2019). Tekoäly voisi olla ratkaisuna kyseisiin riskeihin, sillä luotettavan ja onnistuneesti promptatun tekoälyn toimintaan ei vaikuta kyseiset asiat. On kuitenkin tiedostettava ja otettava huomioon tekoälyllä tuotetun arvioinnin riskit. On esimerkiksi taattava, että oppilaan tietosuoja säilyy, prompti on pätevä ja tekoäly on koulutettu luotettavalla datalla.

Ouakrim-Soivion (2013) mukaan arvioinnin yhdenvertaisuus on ollut haasteena ja kehityksen kohteena jo vuosikymmenten ajan. Valtakunnalliset seuranta-arviot ovat

osoittaneet, että arviointi ei ole yhdenvertaista. Yhdenvertaisuuden haasteita ilmenee osaamisen, sukupuolen ja maahanmuuttajien saralla. (Ouakrim-Soivio, 2013.)

Virheellisellä arvioinnilla, etenkin päättöarvioinnilla voi olla oppilaan elämään merkittäviä vaikutuksia. Päättöarviointi määrittää oppilaan mahdollisuuksia jatko-opintoihin, ja jatko-opinnot puolestaan määrittävät työelämän mahdollisuuksia.

Kansallisen koulutuksen arviointikeskuksen eli KARVI:n (Atjonen ym., 2019) tuottamassa tutkimuksessa nousi myös esiin monia huolestuttavia tuloksia arvioinnista. Ensiksi oppijoiden asenteella ja motivaatiolla oli vaikutusta opettajien tuottamiin arviointeihin. Lähes 80 prosenttia vastaajista oli sitä mieltä, että oppijan asenne ja motivaatio vaikuttavat heidän tuottamaan arviointiin. (Atjonen ym., 2019, s.89.) Tämä on erittäin huolestuttavaa, sillä kyseiset asiat eivät saisi vaikuttaa arvosanaan. Toki oppilaan negatiivinen asenne aihetta kohtaan voi heijastua opiskelumotivaatioon ja sen kautta heikompiin oppimistuloksiin. Kuitenkin arviointiin vaikuttava tekijä olisi silloin oppilaan osaaminen, eikä hänen asenteensa aihetta kohtaan. Toiseksi oppijoiden suoriutumista verrattiin muuhun oppijaryhmään. Vastaajista lähes neljäsosa oli melko tai täysin samaa mieltä, että luokan osaamistaso vaikuttaa heidän tekemään arviointiin. (Atjonen ym., 2019, s.89.)

Lisäksi KARVI:n (Atjonen ym., 2019) raportin mukaan noin 30 prosenttia opettajista oli täysin tai samaa mieltä väittämän kanssa, että oppijan käyttäytyminen vaikuttaa annettuun arvosanaan sekä 14 prosenttia oli sitä mieltä, että persoona vaikuttaa annettuun arvosanaan. Nämä luvut ovat itsessään jo huolestuttavia, mutta huomiota herättää myös vastaajien määrä, jotka olivat ei samaa eikä eri mieltä. Kumpaankin väitteeseen hieman yli 30 prosenttia vastasi, että ei ole samaa eikä eri mieltä. (Atjonen ym., 2019, s.93.) Viimeiseksi lähes 40 prosenttia oli täysin tai melko eri mieltä väittämän kanssa, että päättöarvosanat kuvaisivat hyvin oppilaan osaamista siirtyessä 9.luokalta lukioon. Puolet lukio-opettajista uskoivat, että päättöarvioinnin oikeellisuus riippui siitä mistä koulusta oppija tulee. (Atjonen, 2019, s.104.) KARVI:n (Atjonen ym., 2019) tuottama tutkimus osoittaa, että opettajien toteuttamaan arviointiin vaikuttaa myös sellaiset tekijät, joita ei pitäisi ottaa huomioon.

Malouff, Emmerton ja Schutte (2013) puolestaan tutkivat haloefektin vaikutusta arviointiin. Haloefekti on psykologinen ilmiö, jossa yksi myönteinen ominaisuus vaikuttaa siihen, miten arvioimme asiaa kokonaisuutena (Nicolau, Mellinas & Martin-Fuentes, 2022). Malouff, Emmerton ja Schutte (2013) painottavat, että halo-efekti on merkittävä syy pitää opiskelijat anonyymeinä arvioinnin ajan. Heidän tutkimuksessaan 126 arvioijaa arvioi saman tekstivastauksen. Ennen tekstin arviointia arvioijille näytettiin näennäisesti saman oppilaan tekemä suullinen esitelmä. Puolille arvioijista näytettiin onnistunut ja lopuille epäonnistunut suullinen esitelmä. Hyvän esitelmän nähneet antoivat keskimäärin 4,2 pistettä korkeammat pisteet tekstivastaukselle. (Malouff, Emmerton & Schutte, 2013.) Tutkimus osoittaa, että oppilaan muut onnistumiset voivat vaikuttaa opettajan tuottamaan arviointiin.

Yllättävästi myös oppilaan käsialalla voi olla vaikutusta arviointiin. Marshallin (1972) tutkimuksessa 420 opettajaa arvioi jokainen 16 esseetä. Esseet oli kirjoitettu koneella tai yhdellä kolmesta käsialasiisteydestä: siisti, tyydyttävä ja huono. Tuloksien mukaan siististi kirjoitetut esseet saivat huomattavasti korkeampia arvosanoja. (Marshall, 1972.) Vaikka tutkimus on vanha, se havainnollistaa, että arviointiin voi vaikuttaa erittäin epäolennaiset tekijät.

Baird (1998) puolestaan tutki mahdollista sukupuolivinoumaa arvioinnissa. Tutkimukseen osallistui 24 arvioijaa, joista jokainen arvioi 30 tekstivastausta. Sukupuolivinoumaa pyrittiin selvittämään antamalla eri lisätietoja esseiden lisäksi. Esseet olivat nimettömiä tai nimellisiä, miehen tai naisen nimellä sekä miehen tai naisen käsialalla kirjoitettuna. Kokeiluryhmän mukaan miehen käsiala on usein sotkuisempaa kuin naisten. Tutkimuksen tulokset osoittavat, että sukupuolivinoumaa ei ilmennyt arvioinnissa. Arvioinnit olivat hyvin yhteneviä eri muuttujista huolimatta. (Baird, 1998.) Kuitenkin Ouakrim-Soivion (2013) uudemman tutkimuksen mukaan sukupuolella on vaikutusta arvioinnin yhdenvertaisuuteen.

Tekoäly voisi olla ratkaisuna arvioinnin haasteisiin, mutta senkään toiminta ei ole riskitöntä. Tekoälyn riskit koskevat oppilaiden tietosuojaa sekä mahdollisia vinoumia tekoälyn koulutuksessa ja tekoälyn toiminnan ei-avoimuutta. Ensimmäinen riski koskee oppilaiden tietosuojaa. Tekoäly voi kerätä tietoa itselleen ilman lupaa ja myös

hyödyntää sitä sekä riski tietovuotoon on mahdollinen. (Gomstyn & Jonker, ei pvm.) On mahdollista, että tekoäly hyödyntäisi oppilaiden tuotoksia itsensä kouluttamiseen tai julkaisisi ne laajemmalle yleisölle. Esimerkiksi ChatGPT tarjoaa ominaisuuden, jolloin lähetetyt tiedostot ja käytyt keskustelut eivät päätyisi laajan kielimallin koulutukseen (OpenAI, ei pvm. a). Kuitenkaan tästä ei voida olla täysin varmoja. Paikallinen tekoäly voisi ratkaista tietosuojariskin. Paikallinen tekoäly tarkoittaa tekoälyä, joka toimii vain omalla laitteella, se ei hyödynnä internetiä ja käyttäjä itse vastaa sen toiminnasta (Saadioui, 2025.) Paikallinen tekoäly takaa tietosuojan, sillä oppilaiden tekstejä ei annettaisi ulkoisille palvelimille prosessoitavaksi. Tämä vähentää tietovuotojen ja luvattoman käytön riskiä. Paikallinen tekoäly olisi arvioinnissa turvallisempi, sillä koulu voisi hallita sen asetuksia ja dataa, tavoitteena saada luotettava arviointityökalu. (Saadioui, 2025.)

Toinen riski liittyy tekoällyn virheelliseen valmistamiseen ja promptaukseen. Virheet tekoällyn valmistuksen eri vaiheissa kuten sen koulutuksessa ja datan keruussa, voivat johtaa vinoutuneeseen tekoölyyn. Tällä tarkoitetaan sitä, että tekoällyn tekemät päätökset tai antamat vastaukset voivat olla esimerkiksi puolueellisia tai jopa täysin virheellisiä. (Chapman university, ei pvm.) Tekoällyn käyttäjä antaa tekoöllylle prompteja eli ohjeita, joiden pohjalta tekoäly toimii (Harvard university, 2023). Pätevän promptin luominen on haasteellista. Kuinka tekoäly saadaan arvioimaan vain annettujen kriteerien pohjalta, kuinka se saadaan luotettavasti arvioimaan oppilaan omaa ajattelua tai perusteluita sekä kuinka saadaan tekoäly arvioimaan kaikkien oppilaiden esseet samoin tavoin. Nämä olivat vain muutamia esimerkkejä promptauksen haasteista. Seuraukset voivat olla vakavia arvioinnin oikeudenmukaisuuden ja luotettavuuden kannalta, jos oppilaiden tuotoksia arvioidaan vinoutuneella tai väärin promptatulla tekoöllyllä. Vinoutunut data voi muokata tekoällyn suosimaan esimerkiksi tiettyjä kirjoitustyyliä ja tekstin rakenteita tai aiheita, mikä näkyy arvosanoissa. Virheellinen promptti voi puolestaan johtaa tekoällyn esimerkiksi arvioimaan väärin osa-alueita, se voi virheellisesti tulkita tehtävänannon tai oppilaan tekstin sekä se voi arvioida oppilaiden tuotoksia eri perustein.

Kolmas riski liittyy tekoällyn avoimuuteen. Tekoälyt voivat olla avoimia tai suljettuja. Avoimen tekoällyn toimintakoodi ja koulutusdata ovat julkisesti saatavilla sekä kuka

tahansa voi käyttää tai muokata sitä. Suljetut tekoälyt ovat puolestaan esimerkiksi yrityksen hallussa, eikä sen toimintakoodia ja koulutusdataa jaeta julkisesti, ainakaan kokonaan. Esimerkiksi suosittu ChatGPT on suljettu tekoäly. Suljetut tekoälyt eivät yleensä näytä käyttäjälleen sisäistä päättelyprosessiaan. (Malec, 2025.) Kun mallin päättelyprosessia ei ole nähtävissä, on opettajan vaikea varmistaa mallin tuottaman arvioinnin luotettavuutta. Lisäksi opettajan pitäisi pystyä perustelemaan arvosanaan johtaneet päätökset (Tikkanen, ei pvm.), mikä ei välttämättä ole mahdollista, kun mallin päättelyprosessia ei ole nähtävissä.

Yhteenvetona voidaan todeta, että opettajan ja tekoälyn tuottamassa arvioinnissa on omat riskinsä. Kummallakin lähestymistavalla on merkittäviä seurauksia, mikäli riskejä ei pyritä minimoimaan. Keskeisenä tavoitteena on saada arvioinnista yhdenvertaista ja luotettavaa. Yksi mahdollinen keino voisi olla opettajan ja tekoälyn yhteistyö.

### **3.3 Opettajien ja oppilaiden näkemyksiä arvioinnista**

Opettajien ja oppilaiden näkemykset arvioinnista sekä tekoälyllä tuotetusta arvioinnista vaihtelevat. KARVI eli kansallinen koulutuksen arviointikeskus tuotti tutkimuksen, joka tutki opettajien, oppilaiden ja heidän vanhempien näkemyksiä arvioinnista (Atjonen ym., 2019). Tutkimuksen mukaan 80 prosenttia opettajista kokee arvioinnin olevan melko tai täysin oikeudenmukaista. Koulun rehtorit pitivät arviointia vielä oikeudenmukaisempana kuin opettajat. Oppilaiden vanhemmat pitivät arviointia myös hyvin oikeudenmukaisena. 70 prosenttia vastaavista vanhemmista pitivät lapsensa saamaa arviointia oikeudenmukaisena. Myönteiset vastaukset puolestaan vähenivät, kun kysyttiin opettajien välisistä arvioinneista. Alle 60 prosenttia opettajista oli sitä mieltä, että opettajien antamat arvosanat vastaavat toisiaan osaamistasoissa. Kuitenkin vain kymmenesosa oli melko tai eri mieltä väittämän kanssa. Rehtoreista puolestaan 70 prosenttia uskoi opettajien antamien arvioiden olevan yhdenvertaisia. (Atjonen ym., 2019, s.129 ja 139.)

KARVI:n (Atjonen ym., 2019, s.132) tuottaman tutkimuksen mukaan arviointi on oppilaille tärkeää. Arviointi näyttäytyi tärkeämpänä, mitä vanhemmista oppilaista oli kyse. Oppilaista hieman alle 70 prosenttia piti opettajan tekemää arviointia reiluna. Myös oppilaiden vanhemmat jakoivat saman näkemyksen. Yli 80 prosenttia oppilaiden

vanhemmista oli sitä mieltä, että arviointi kuvaa oman lapsen osaamistasoa. (Atjonen ym., 2019. s.135 ja 139.) Tutkimus pyrki myös tutkimaan oppilaiden kielteisiä näkemyksiä arvioinnista. Oppilaiden näkemykset arvioinnista ovat kuitenkin melko myönteisiä. Esimerkiksi alle viidesosa oli sitä mieltä, että arviointi on liian vaativaa tai tiukkaa, ja alle 10 prosenttia sitä mieltä, että saisi liikaa kielteistä palautetta. Vain neljäsosa oli sitä mieltä, että arvioitavia tehtäviä on liikaa, sekä yli viidesosa kertoi kokeiden ahdistavan tai pelottavan heitä. (Atjonen ym., 2019, s.136.)

Tossel ym. (2024) tutkivat puolestaan oppilaiden näkemyksiä tekoälystä arvioijana. Tutkimukseen osallistuneista yli puolet halusivat, että heidän tehtävänsä arvioisi sekä opettaja että tekoäly. Kukaan tutkimukseen osallistuneista ei kuitenkaan halunnut pelkän tekoälyn toimivan arvioijana. Tutkimuksessa ei perehdytty yliopisto-opiskelijoiden mielipiteiden taustalla vaikuttaneisiin syihin. Tutkimuksen otanta on erittäin pieni (n 24), mutta se antaa viitteitä oppilaiden varovaisesta suhtautumisesta tekoälyarviointia kohtaan ja oppilaiden tarpeesta säilyttää opettajan rooli arvioijana.

Chai ym. (2024) tutkivat yliopisto-opiskelijoiden kokemuksia arvioinnin reiluudesta. He toteuttivat kolme tutkimusta, joihin osallistui yhteensä lähes 500 opiskelijaa. Ensimmäisessä tutkimuksessa selvitettiin, kuinka reiluksi opiskelijat kokevat tekoälyn ja opettajan tekemät diagnostiset, formatiiviset ja summatiiviset arviointitavat. Opiskelijat lukivat kolme kuvailevaa tarinaa, ja pisteyttivät kokemuksen reiluudesta asteikolla 1–7. Opiskelijat kokivat tekoälyn tekemän arvioinnin reilummaksi kuin ihmisen tekemän kaikissa kolmessa arviointitavassa. Diagnostisessa arvioinnissa tekoäly sai reiluuden keskiarvoksi 5,6 kun taas opettaja 5,3. Puolestaan formatiivisessa arvioinnissa tekoäly koettiin huomattavasti reilummaksi. Tekoälyn keskiarvo oli 5,4 kun taas opettajan vain 4,78. Summatiivisessa arvioinnissa tekoäly sai keskiarvokseen 5,48 ja opettaja 5,2. Chai ym. (2024) mukaan opiskelijat kokivat tekoälyn tekemän arvioinnin huomattavasti reilummaksi kaikissa arviointitavoissa. Tekoäly sai korkeammat keskiarvot jokaisessa arviointitavassa, mutta opettajan saama keskiarvo ei kuitenkaan ollut merkittävän kaukana. Vain summatiivisessa arvioinnissa ero suuri. (Chai ym., 2024.)

Tutkimuksessa kaksi Chai ym. (2024) tutkivat arvioinnin läpinäkyvyyden vaikutusta reiluuden kokemukseen samoissa kolmessa arviointitavassa eli diagnostinen, formatiivinen ja summatiivinen. Opiskelijat vastasivat väittämiin asteikolla 1–7, josta yksi tarkoitti vahvasti eri mieltä ja seitsemän vahvasti samaa mieltä. Tulokset osoittavat, että arvioijan tyyppi, eli tekoäly tai opettaja, vaikuttaa voimakkaasti siihen, kuinka läpinäkyvänä arviointi koetaan. Tekoälyn toteuttama arviointi koettiin läpinäkyvämmäksi kaikissa kolmessa arviointitavassa. Arvioinnin läpinäkyvyyden todettiin olevan positiivisessa yhteydessä reiluuden kokemukseen. Tällä tarkoitetaan, että arvioinnin läpinäkyvyyden lisääntyessä, myös kokemus reiluudesta kasvaa. Arvioijan tyyppin todettiin vaikuttavan kokemukseen reiluudesta sekä suoraan että epäsuorasti läpinäkyvyyden kautta. (Chai ym., 2024.)

Chai ym. (2024) tutkimuksessa kolme selvitettiin arvioinnin selityksen vaikutusta opiskelijoiden kokemukseen arvioinnin reiluudesta. Tulokset osoittavat, että arviointi koettiin reilummaksi jokaisessa arviointitavassa, jos arvioinnille annettiin perustelut. Mielenkiintoisesti arviointityypin, eli tekoäly tai opettaja, vaikutus reiluuden kokemukseen laski tilastollisesti merkityksettömäksi, jos arvioinnille annettiin selitys. Puolestaan, kun selitystä ei annettu, johti se tilastollisesti merkitsevään eroon. Chai ym. (2024) tutkimuksen vahvuutena on suuri otoskoko sekä kolme erillistä mutta toisiaan täydentävää tutkimusta. Tutkimukset antoivat viitteitä opiskelijoiden positiivisesta suhtautumisesta tekoälyarviointia kohtaan. Chai ym. (2024) tutkimuksen tulokset ovat osittain eriäviä Tossel ym. (2024) tutkimuksen tuloksien kanssa.

Alsalem (2024) tutki puolestaan opettajien mielipiteitä tekoälyarvioinnista. Osallistujat olivat englannin kielen opettajia kolmesta eri Saudi-Arabialaisesta yliopistosta. Opettajat vastasivat kyselylomakkeeseen Likert-asteikolla. Tutkimuksen otanta on erittäin pieni (n 10), mutta antaa viitteitä opettajien varovaisen myönteisestä suhtautumisesta tekoälyyn arvioijana. Kaikki vastaajat näkivät tekoälyn potentiaalinen täydentää perinteisiä arviointimetoja. Kahdeksan vastaajaa oli täysin samaa mieltä väitteen kanssa ja loput kaksi samaa mieltä. Vastaajat uskoivat tekoälyn vähentävän arviointityöhön kuluvaan aikaa. Puolet vastaajista oli samaa mieltä, mutta vain yksi täysin samaa mieltä väitteen kanssa. Enimmäkseen vastaajat uskoivat tekoälyn vähentävän arvioinnin työmäärää ja siihen kuluvaan aikaa. Kaksi vastaajaa oli sitä mieltä, että tekoäly

ei vähennä työmäärää, kun taas loput olivat täysin samaa tai samaa mieltä väitteen kanssa. Puolestaan vain yksi vastaajista ei uskonut tekoälyn vähentävän arvioinnin työmäärää. Vastaukset jakautuivat merkittävästi, kun kysyttiin antaako tekoäly objektiivisempaa palautetta kuin ihminen. Kaksi vastaajaa oli eri mieltä, mutta vain yksi täysin samaa mieltä väittämän kanssa. Kuusi vastaajaa oli vain samaa mieltä. Kaikista negatiivisempia oltiin, kun kysyttiin tekoälyn antaman palautteen reiluudesta. Neljä vastaajaa oli eri mieltä, että tekoäly antaisi reilua palautetta oppilaiden esseistä. Kuitenkin loput vastaajista oli täysin samaa tai samaa mieltä. (Alsalem, 2024.)

## 4 Tutkimuksen toteutus

Tutkimuksen tavoitteena oli selvittää, kuinka luotettavan arvioinnin laaja kielimalli tuottaa historian koevastauksista sekä millaisia ongelmia kehoitesuunnittelussa ilmenee, kun luodaan laajalle kielimallille arviointipromptti. Luvussa 3.1 esitetyn suppean kuvailevan kirjallisuuskatsauksen pohjalta voidaan todeta, että tutkimukset nostavat harvoin esiin kehoitesuunnittelun vaiheita tai arviointiprompttia. Siksi on arvokasta kertoa kehoitesuunnitteluprosessiin liittyvistä ongelmista.

Kehotesuunnittelun toteuttamisen ongelmia tutkitaan tuottamalla arviointipromptti ja kirjaamalla ylös prosessin aikana ilmenneet ongelmat. Laajan kielimallin tuottaman arvioinnin luotettavuutta tarkastellaan sekä määrällisen että laadullisen tutkimusotteen avulla. Myös aineiston analyysissä hyödynnetään määrällisiä ja laadullisia menetelmiä. Tutkimuksen kohteeksi rajautui lopulta OpenAI:n paikallinen gpt-oss-120b laaja kielimalli tietoturvasyistä. Tutkimusprosessin aikana perehdytään myös kahden pilvipohjaisen Open AI:n ChatGPT 5.2 ja Googlen Gemini 3 Flash laajojen kielimallien toimintaan. Mallit tuottivat arvioinnit 12 alakoulun oppilaan historian kokeen vastauksista. Arvioinnin kohteena olivat vastaukset kolmeen koetehtävään, ja jokainen malli arvioi samat koevastaukset kahteen kertaan. Tutkielman liitteessä 1 on kuvio tutkimusasetelmasta. Tutkimuksen tavoitteiden ja aineiston rajauksen myötä muodostuivat seuraavat tutkimuskysymykset.

1. Kuinka luotettavasti paikallinen laaja kielimalli arvioi alakoulun historian kokeita?
2. Millaisia ongelmia kehoitesuunnittelussa ilmenee, kun luodaan arviointipromptti, ja mitä asioita on huomioitava arviointipromptin käytössä?

Laajan kielimallin tuottaman arvioinnin luotettavuutta tutkitaan laadullisen ja määrällisen tutkimusotteen sekä arviointitutkimuksen avulla. Tutkimuksen aihe ja asetelma ovat melko vähän tutkittuja, minkä vuoksi aiemmasta tutkimuksesta tai menetelmäkirjallisuudesta ei löytynyt vakiintunutta tapaa tutkimuksen toteuttamiseen. Tämä näkyy tutkimuksessa erityisesti tutkimusotteiden yhdistämisessä sekä aineiston analyysissä, jossa hyödynnettiin useita erilaisia analyysitapoja. Seuraavaksi

tarkastellaan, millaisia yhteisiä ja eroavia piirteitä tutkimuksella on laadullisen, määrällisen ja arviointitutkimuksen kanssa.

Tutkimusperinteessä on vakiintunut tapa jakaa tutkimukset laadullisiin ja määrällisiin. Jaottelu on kuitenkin melko mustavalkoinen. Laadullinen eli kvalitatiivinen ja määrällinen eli kvantitatiivinen tutkimusote eivät ole täysin toistensa vastakohtia eivätkä siten sulje toisiaan pois (Eskola & Suoranta, 1998, luku 1). Oikein hyödynnettyinä laadullisia ja määrällisiä menetelmiä voidaan yhdistää mielekkäästi.

Tutkimuksen aineisto oli sekä laadullista että määrällistä. Laadullisessa tutkimuksessa aineisto on tyypillisesti tekstimuotoista, ja puolestaan määrällisessä tutkimuksessa numeerista (Eskola & Suoranta, 1998, luku 1). Laaja kielimalli tuotti sekä pistearvioinnin että sanallisen arvioinnin oppilaan koevastauksista. Täten pistearvioinnit ovat numeerisia ja sanalliset arvioinnit tekstimuotoisia. Laajan kielimallin tuottaman arvioinnin luotettavuutta tutkitaan numeerisista ja sanallisista arvioinneista. Luotettavuuden käsite operationalisoitiin viiteen eri kategoriaan, joiden avulla luotettavuutta voidaan arvioida ja mitata määrällisesti. Täten laadullisesta aineistosta tuotetaan määrällistä tietoa. Eskolan ja Suorannan (1998, luku 2) mukaan laadullisessa tutkimuksessa ei yleensä puhuta operationalisoinnista samalla tavalla kuin määrällisessä tutkimuksessa, vaikka myös laadullisessa tutkimuksessa teoreettiset käsitteet on saatava empiirisesti tarkasteltavaan muotoon. Aineiston analyysia käsitellään tarkemmin luvussa 4.4.

Tutkimuksen tavoitteessa on nähtävissä sekä määrällisen että laadullisen tutkimuksen piirteitä. Puusan, Juutin ja Aaltion (2021, pääluke 2) mukaan määrällinen tutkimus tavoittelee objektiivisuutta ja yleistettävyyttä. Määrällisen tutkimuksen tavoite ilmenee erityisesti siinä, että pyritään tarkastelemaan gpt-oss-120b mallin luotettavuutta objektiivisesti. Kuitenkin tuloksia ei voida täysin yleistää koskemaan kyseistä mallia, sillä saadut tulokset ovat vahvasti kontekstisidonnaisia käytetyn arviointipromptin vuoksi. Tuloksia ei voida yleistää myöskään koskemaan muita laajoja kielimalleja, sillä mallit toimivat eri tavoin. Lisäksi tutkimuksen tavoitteena ei ole selvittää syy-seuraussuhteita tai ilmiöiden välisiä yhteyksiä, jotka ovat määrälliselle tutkimukselle tyypillisiä (Jyväskylän yliopisto, ei pvm.). Puolestaan laadullisen tutkimuksen tavoitteina

ovat usein ilmiön yksityiskohtainen kuvailu ja subjektiivisen tiedon ymmärtäminen (Puusa, Juuti & Aaltio, 2021, pääluke 2). Vaikka tutkimuksessa ei varsinaisesti tarkastella subjektiivista tietoa, eli yksilön henkilökohtaisia näkemyksiä (Puusa, Juuti & Aaltio, 2021, pääluke 2), ilmenee laadullisen tutkimuksen piirteet siinä, että tulokset ovat vahvasti sidoksissa kontekstiin.

Käytetty laaja kielimalli ja arviointiprompti tekevät tutkimuksen tuloksista vahvasti kontekstisidonnaisia. Lisäksi arviointiprompti on luotu nimenomaan kyseisen historian kokeen arviointia varten, eikä sitä voitaisi sellaisenaan hyödyntää toisessa tutkimuksessa. Kontekstisidonnaisuus onkin yksi laadullisen tutkimuksen keskeisistä piirteistä (Eskola & Suoranta, 1998, luku 1; Puusa, Juuti, Aaltio, 2021, pääluke 2).

Tässä tutkimuksessa on lisäksi arviointitutkimuksen piirteitä. Jokisen (2021) mukaan arviointitutkimus kuuluu soveltavan tutkimuksen piiriin, ja sen tavoitteena on arvioida tutkittavan kohteen toimivuutta, vaikutuksia tai arvoa. Arviointitutkimus kohdistuu usein julkisiin palveluihin tai hyvinvointialaan, mutta sitä voidaan soveltaa myös muiden toimintojen tarkasteluun (Jokinen, 2021). Tässä tutkimuksessa arvioidaan laajan kielimallin toimintaa arvioijana, mikä tuo tutkimukseen arviointitutkimuksen piirteitä. Tutkimuksen tavoitteena ei kuitenkaan ole määrittää laajan kielimallin arvoa tai tehdä varsinaista vaikuttavuusarviointia. Lisäksi arviointitutkimus lähtee usein yhteiskunnan tai organisaation tarpeesta arvioida jotakin toimintaa (Jokinen, 2021), kun taas tämä tutkimus perustuu ensisijaisesti kiinnostukseen aihetta kohtaan.

#### **4.1 Tutkimuskohteena käytettävän laajan kielimallin valinta**

Tutkimuksen pääkohteeksi valittiin yksi laaja kielimalli, jonka tuottaman arvioinnin luotettavuutta tarkastellaan. Malli valittiin kolmen laajan kielimallin keskuudesta. Mallit olivat Open AI:n gpt-oss-120b ja ChatGPT 5.2 sekä Googlen Gemini 3 Flash. Ensimmäisenä mainittu malli on paikallinen ja kaksi jälkimmäistä ovat pilvipohjaisia. Tutkimuksen pääkohteeksi valikoitui lopulta OpenAI:n paikallinen gpt-oss-120b malli. Kuitenkin myös pilvipohjaisilla ChatGPT 5.2 ja Gemini 3 Flash malleilla tuotettiin arvioinnit, jotta tutkimusta voitaisiin tulevaisuudessa jatkaa vertailemalla eri laajojen kielimallien tuottamien arviointien luotettavuutta. Tässä tutkimuksessa ei analysoitu pilvipohjaisten mallien tuottamia arviointeja, joten johtopäätöksiä ei voida vielä esittää.

Kuitenkin arvioinneista haluttiin esitellä alustava näkemys silmämääräisen tarkastelun ja arviointikokemuksen pohjalta. Alustava näkemys esitellään luvussa 6.3. Seuraavaksi esitellään tutkimuksen pääkohteeksi ehdolla olleet mallit, ja perustelut gpt-oss-120b mallin valinnalle.

Ensimmäisenä vaihtoehtona oli OpenAI:n luoma ChatGPT 5.2, joka toimii pilvipohjaisesti. Ensimmäinen versio ChatGPT:stä julkaistiin vuoden 2022 lopulla. Tämän jälkeen Open AI on julkaissut monta uutta versiota ChatGPT:stä. Tutkielman kirjoitushetkellä uusin versio on ChatGPT 5.2, joka julkaistiin joulukuussa 2025. (Open AI, ei pvm.b.) Kyseinen malli on rakennettu yhteistyössä Microsoftin ja NVIDIA:n kanssa (OpenAI, 2025). Open AI:n (2025b) mukaan ChatGPT 5.2 on heidän tehokkain mallinsa, etenkin matematiikassa ja tietojenkäsittelytieteessä. Historian kokeiden tarkistamisessa kyseisistä taidoista ei hyödytä, mutta OpenAI kuitenkin kehuu mallin päättelykykyä ja johdonmukaisuuden ylläpitämistä (OpenAI, 2025b). Kyseisiä taitoja tarvitaan, kun monia koevastauksia tahdotaan arvioida samoin perustein. Kuitenkin OpenAI:n nettisivuilla mainitaan, että ChatGPT:tä voi hyödyntää esseiden arvioinnissa. He suosittelevat promptin ”toimi professorina ja arvioi tehtäväni arviointikriteerien mukaisesti” käyttöä. Sivustolla kuitenkin kerrotaan, että vinkit ovat opiskelijoilta opiskelijoille. (ChatGPT, ei pvm.b.) Täten promptin pätevydestä ei voida olla varmoja, koska ei tiedetä, onko se opiskelijan vai ammattilaisen luoma.

Open AI tarjoaa yksityishenkilölle neljää erilaista käyttö sopimusta. Sopimuksien kuukausihinnat ovat ilmainen, 8 €, 23 € ja 229 €. Korkeampi hinta tarjoaa lisäominaisuuksia ja vahvempaa suoriutumista mallilta. Ilmaisversiossa tarjotaan rajoitettu käyttöoikeus ChatGPT 5.2 malliin. (ChatGPT, ei pvm.a.) Tämä tarkoittaa, että mallia pystyy käyttämään vain tietyn määrän tietyllä ajanjaksolla. OpenAI ei kuitenkaan ole tarkentanut määrää, montako kertaa mallia pystyy käyttämään kyseisellä ajanjaksolla ja kauanko kyseinen ajanjakso kestää. Käyttöoikeuden umpeuduttua, malli käyttää heikompa versiota vastauksien luontiin. OpenAI ei myöskään ole kertonut, mitä mallia vastauksiin hyödynnetään käyttöoikeuden umpeuduttua.

Olisi perusteltua valita ChatGPT 5.2 tutkimuksen pääkohteeksi, sillä se on suosituin laaja kielimalli. Tämän vuoksi malli saattaa olla opettajille entuudestaan tuttu, mikä

voisi edistää sen käyttöönottoa kouluissa. Kumrin (2025) mukaan ChatGPT oli suosituin laaja kielimalli vuosina 2022–2023 kävijämäärän mukaan. Kyseisessä laajassa kielimallissa vierailtiin vuoden aikana 14,6 miljardia kertaa (Kumri, 2025). Puolestaan Azoman (2026) nettiartikkelissa tarkasteltiin laajojen kielimallien suosiota verkkoliikenteen määrän avulla. ChatGPT:n osuus vuoden 2024 generatiivisten tekoälyjen verkkoliikenteestä oli 87 prosenttia. Puolestaan vuonna 2025 luku oli reilusti laskussa, mutta edelleen suosituin 68 prosentin osuudella. Yllättäen Googlen Gemini nosti kävijämääriään viidestä prosentista yli 18 prosenttiin. Gemini oli vuosien 2024 ja 2025 toiseksi suosituin laaja kielimalli. (Azoma, 2026.)

Toisena vaihtoehtona oli Googlen luoma pilvipohjainen Gemini 3 Flash. Ensimmäinen versio Geministä julkaistiin vuoden 2024 alussa. Tutkielman kirjoitushetkellä uusin versio Gemini 3 julkaistiin vuoden 2025 lopulla. (Gemini, ei pvm.a.) Google tarjoaa ilmaisen lisäksi kolmea eri hintaista käyttösovimusta. Sopimuksien kuukausihinnat ovat ilmainen, 8 €, 23 € ja 275 €. Flash versio mallista on ilmainen. (Gemini, ei pvm.b.) Geminin suorituskykyä, päättelykykyä ja nopeutta kehuaan (Doshi, 2025). Googlen analyysin (Doshi, 2025) mukaan Gemini 3 Flash on monessa testissä lähes yhtä kyvykäs kuin ChatGPT 5.2. Molemmat mallit pärjäsivät hyvin MMMLU testissä, joka mittaa esimerkiksi tekoälymallin suorituskykyä eri kielisissä konteksteissa (HyperAI, 2024.) Geminin malli suoriutui kuitenkin hieman paremmin lähes 92 prosentilla, ja ChatGPT 5.2 vajaalla 90 prosentilla (Doshi, 2025).

Olisi perusteltua valita Gemini tutkimuksen pääkohteeksi, sillä se sijoittui korkealle LM Arenan listalla (LMArena, ei pvm.). LM Arena on nettisivu, jossa miljoonat ihmiset arvostelevat eri pilvipohjaisia tekoälysovelluksia lähes reaaliaikaisesti. Tekoälyjä arvostellaan kahdeksassa eri kategoriassa. Niitä ovat esimerkiksi koodaus, tekstin kirjoitus, tiedonhaku ja kuvan luonti. Pilvipohjaisen tekoälyn sijoitusta voi tarkastella yhdessä kategoriassa, tai kokonaissijoitusta, jossa on otettu huomioon kaikki kahdeksan kategoriaa. Kaikkien mallien sijoitukset on katsottu 27.01.2026. Alusta ei kerro kuinka säännöllisesti listoja päivitetään, mutta ilmoittaa edellisen päivityskerran. Kokonaissijoituksessa Gemini 3 Flash sijoittuu sijalle kolme. Puolestaan ChatGPT 5.2 sijalle 16. (LMArena, ei pvm.) Toinen perustelu Geminin valinnalle olisi, että Googlen

muut palvelut ovat käytössä monissa kouluissa. Tämä voisi helpottaa Geminin käyttöönottoa oppilaitoksissa.

Täten olisi erittäin perusteltua valita tutkimuksen kohteeksi OpenAI:n ChatGPT 5.2 tai Googlen Gemini 3 Flash malli. Molemmat mallit ovat suosittuja, sijoittuvat korkealle LM Arenan listalla sekä menestyivät hyvin MMMLU-testissä. Mallien suosio voisi helpottaa niiden käyttöönottoa kouluissa, sekä korkeat sijoitukset viittaavat hyvään suorituskykyyn. Pilvipohjaisilla laajoilla kielimalleilla on myös useita etuja verrattuna paikallisiin malleihin (Microsoft, 2025). Keskeisin etu liittyy suorituskykyyn.

Pilvipohjaiset mallit toimivat suurissa datakeskuksissa, joissa on erittäin paljon laskentatehoa. Tämän vuoksi ne kykenevät hyödyntämään erittäin suuria ja kehittyneitä kielimalleja, joita ei ole mahdollista pyörittää käyttäjän omalla koneella. Käyttäjän ei tarvitse itse omistaa tehokasta tietokonetta, jossa on suuri keskusmuisti ja tekoälyn ajamiseen soveltuva näytönohjain, sillä malli toimii palveluntarjoajan tehokkailla palvelimella eikä käyttäjän omalla laitteella. Lisäksi pilvipohjainen malli on helpompi ottaa käyttöön, sillä se toimii selaimessa tai sovelluksessa. Myös käyttökustannukset ovat käyttäjälle halvempia, koska pilvipohjaisia malleja voi käyttää jopa ilmaiseksi tai maksaa kuukausimaksu. Kuitenkin kuukausimaksut ovat huomattavasti halvempia kuin kustannukset hankkia paikallisen laajan kielimallin käyttämiseen soveltuva kone. (Microsoft, 2025.)

Pilvipohjaisiin laajoihin kielimalleihin liittyy kuitenkin riskejä tietoturvan ja yksityisyyden näkökulmasta. Pilvipohjaiselle mallille syötetyt tiedot käsitellään ulkoisella palvelimella, mikä tuottaa tietoturvariskin (Microsoft, 2025). Tämän vuoksi tutkimuksen kohteeksi valittiin paikallinen gpt-oss-120b laaja kielimalli. Koulujen on taattava oppilaiden tietoturva ja yksityisyys, minkä vuoksi pilvipohjaisia malleja ei voida hyödyntää. Paikallinen laaja kielimalli toimii käyttäjän omalla tietokoneella ilman ulkoisia datakeskuksia (Microsoft, 2025), mikä varmistaa oppilaiden tietoturvan. Paikallisten kielimallien etuna voidaan pitää myös mahdollisuutta muokata mallia omiin tarpeisiin soveltuvaksi. Paikallisten kielimallien suoriutumiseen voidaan vaikuttaa esimerkiksi optimoimalla, vaihtamalla kontekstin pituutta tai tokenien määrää. (Microsoft, 2025.) Lisäksi aikaisemmat tutkimukset keskittyivät usein pilvipohjaisten

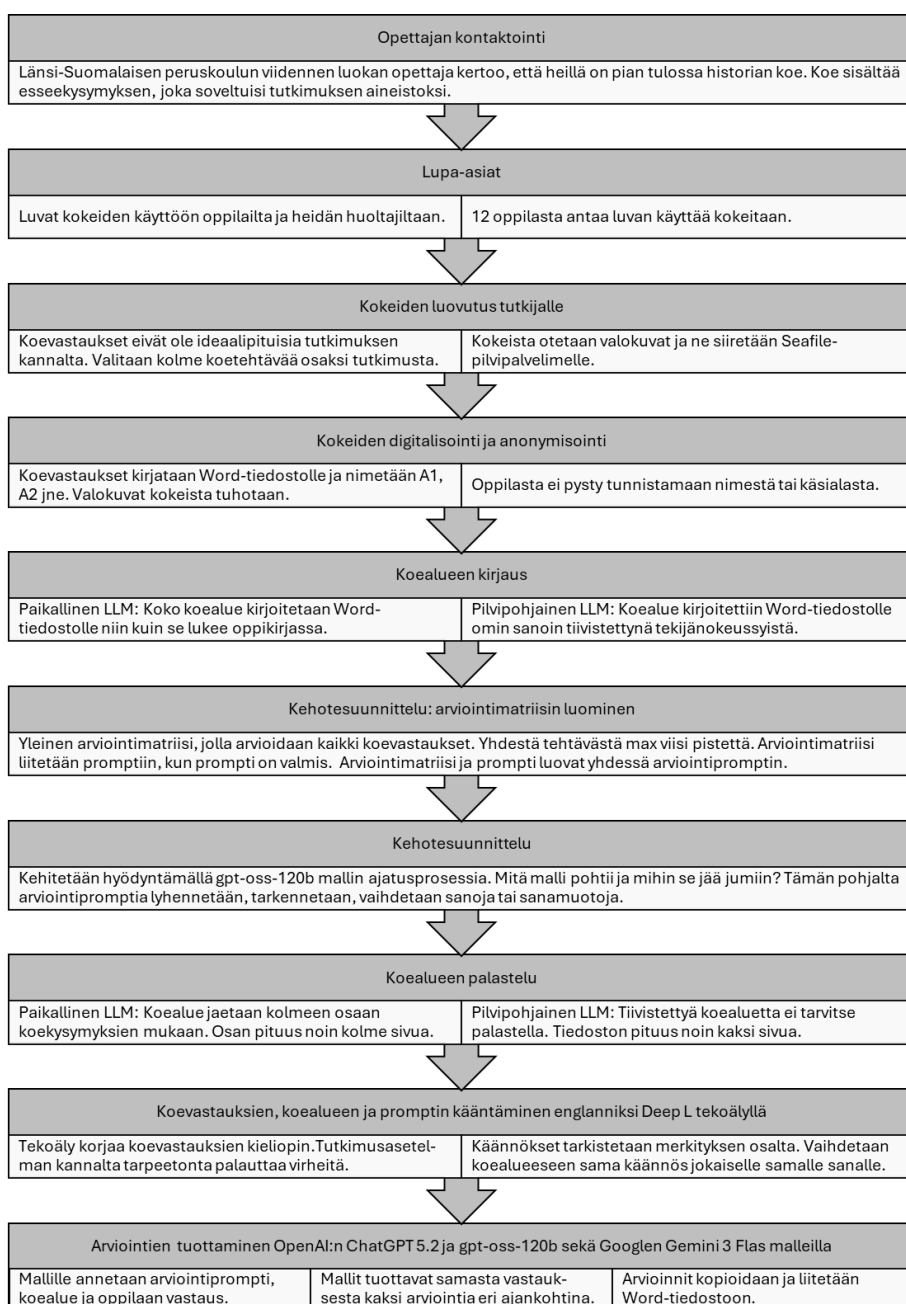
laajojen kielimallien toiminnan tarkasteluun arvioinnissa. Tutkimustieto paikallisten laajojen kielimallien soveltuvuudesta arviointiin on vähäistä.

Näiden syiden vuoksi tutkimuksen kohteeksi valikoitui Open AI:n gtp-oss-120b paikallinen laaja kielimalli. Malli julkaistiin elokuussa vuonna 2025. OpenAI kehuu mallin vahvaa suoritus- ja päättelykykyä. (OpenAI, 2025a.) Kyseinen malli valikoitui osaksi tutkimusta, sillä se oli valmiiksi ladattu Turun yliopiston laitteelle. Tutkimuksia mallin soveltuvuudesta tai toimivuudesta arvioinnissa ei löytynyt. Sen vuoksi on perusteltua nostaa esiin käyttäjien kokemuksia kyseisestä mallista. Keskustelualusta Redditistä löytyi viestiketju (Reddit, 2026), jossa käyttäjät jakoivat mielipiteitä kyseisestä mallista. Käyttäjien mielipiteet vaihtelivat, mutta enimmäkseen malliin oltiin tyytyväisiä. Nimimerkki lcy\_Programmer7186 nostaa, että kyseinen malli on hänen mielestään edelleen paras malli yhdistämään nopeuden ja tuotoksen laadun. Puolestaan nimimerkit hurdur7, custodiam99 ja pj-frey sanovat, että malli on hyvin käyttökelpoinen, mutta muut mallit soveltuvat vielä paremmin omaan käyttöön. Nimimerkki Non\_linear\_ape kehuu mallin päättelykykyä ja kirjoitustaitoa. Puolestaan nimimerkki Much-Researcher6135 suosittelee mallin valjastamista agentiksi, sillä se toimii siinä erittäin hyvin. (Reddit, 2026.) Käyttäjä ei kuitenkaan kerro mihin tehtävään hän on agentin luonut.

Paikallinen gtp-oss-120b laaja kielimalli toimii Turun yliopiston pöytätietokoneella. Pöytätietokone oli Dell Pro Max Tower T2 mallinen, ja sen keskeiset ominaisuudet ovat Intel Core Ultra 9 285K prosessori, 128 gigatavun DDR5 muisti, yli yhden teratavun SSD-tallennustila sekä Nvidia GeForce RTX 5090 -näytönohjain, jossa on itsessään 32 gigatavun muisti. Käyttöjärjestelmänä toimi Windows 11. Erityisesti tietokoneen suuri RAM muisti ja tehokas näytönohjain mahdollistavat laajojen kielimallien onnistuneen ajamisen paikallisesti. Tietokone vastaavilla osilla maksaisi tutkielman kirjoitushetkellä noin 5000 €. Paikallisen laajan kielimallin ajamiseen käytettiin LM Studio- sovellusta. Sovelluksen avulla käyttäjä voi ladata ja käyttää avoimen lähdekoodin kielimalleja omalla tietokoneellaan.

## 4.2 Aineiston keruun vaiheet

Tutkimuksen aineisto koostui oppilaiden koevastauksista, koealueesta, arviointipromptista sekä laajan kielimallin tuottamista arvioinneista edellä mainittujen asioiden pohjalta. Kuviossa 2 esitetään aineiston keruun vaiheet tiivistetysti. Tässä luvussa kerrotaan koevastauksiin, koealueeseen ja arviointien tuottamiseen liittyvistä vaiheista. Kehotesuunnittelusta ja arviointipromptin luomisesta kerrotaan tarkemmin seuraavassa luvussa.



Kuvio 2. Aineiston keruun vaiheet.

Tutkimus käynnistyi kysymällä erään Länsi-Suomalaisen peruskoulun viidennen luokan opettajalta mahdollisuutta hyödyntää tutkimuksessa hänen oppilaidensa kirjoittamia esseitä. Luokanopettaja kertoi, että oppilaille oli pian tulossa historian koe, mikä mittaisi heidän tietämystään antiikin Kreikasta. Koe sisältäisi yhden esseevastauksen. Hyväksyttiin opettajan ehdotus, ja sen jälkeen kerrottiin oppilaille tutkimuksen ideasta, ja jaettiin tutkimuslupalomake sekä tietosuojaseloste oppilaalle ja hänen huoltajalleen. 12 oppilasta antoi luvan hyödyntää heidän koevastauksiaan.

Oppilaiden kirjoittamat esseevastaukset olivat kuitenkin erittäin lyhyitä. Keskimäärin vastaukset olivat kahdeksan virkkeen mittaisia. Pisin vastaus oli 20 virkettä ja lyhyimmät vastaukset kahden virkkeen mittaisia. Tämän seurauksena tutkimuksen näkökulma vaihdettiin koskemaan laajan kielimallin kykyä arvioida luotettavasti lyhyitä tekstivastauksia eikä esseitä. Koska tekstivastaukset ovat lyhyempiä kuin essee, täytyi tutkimukseen valita oppilaalta useampi koevastaus. Näin saatiin tarpeeksi aineistoa laajalle kielimallille arvioitavaksi. Tutkimukseen valittiin oppilaiden vastaukset kokeen kolmeen laajimpaan koetehtävään. Koetehtävien kysymykset olivat 1. Kerro miten jumalat liittyivät kreikkalaisten arkeen? Miten ihmiset huomioivat jumalat ja millaisia jumalia heillä oli? Liittyivätkö nuo jumalat jotenkin nykyaikaan? 2. Mitä eroa oli Ateenan ja Spartan kaupunkivaltioilla keskenään? Kerro sekä arjen asioista että kansan johtamiseen ja päätöksentekoon liittyvistä asioista. 3. Kuka oli Aleksanteri Suuri? Miten hän liittyi antiikin Kreikkaan? Mitä hän sai aikaan? Miten hän onnistui tavoitteissaan ja saavutti ne asiat, joita halusi? Kolmeen koetehtävään kertyi yhteensä 34 vastausta. Kaksi oppilasta oli vastannut koetehtävään kaksi ajatuskartalla, minkä vuoksi kyseisiä vastauksia ei otettu tutkimukseen mukaan.

Ensimmäisenä koevastauksista otettiin valokuvat ja ne siirrettiin Seafile-pilvipalvelimeen. Valokuvat olivat välttämättömiä anonymisoinnin ajaksi, sillä ei tahdottu oppilaiden joutuvan odottamaan pitkään kokeiden palautusta. Tämän jälkeen vastaukset anonymisoitiin kirjoittamalla ne tietokoneelle ja nimeämällä ne esimerkiksi A1. Näin oppilasta ei pystytä tunnistamaan nimestä tai käsialasta. Vastaukset kirjoitettiin juuri niin kuin oppilaat olivat ne kirjoittaneet, ja tämän vuoksi ne sisälsivät esimerkiksi kielioppivirheitä. Valokuvat tuhottiin heti sen jälkeen, kun vastaukset oltiin anonymisoitu.

Arviointia varten laaja kielimalli vaatii koealueen, jonka pohjalta se arvioi vastauksien oikeellisuutta. Vaikka tutkimuksen pääkohteena oli paikallinen laaja kielimalli, tuotettiin arvioinnit myös pilvipohjaisilla malleilla mahdollista jatkotutkimusta varten.

Paikalliselle laajalle kielimallille voidaan syöttää koko koealue, mutta pilvipohjaisille malleille ei. Pilvipohjainen malli käyttää ulkoisia palvelimia, minkä vuoksi koeaineisto voisi päätyä mallin yleiseen käyttöön, ja täten rikkoisi oppikirjan kirjoittajien tekijänoikeutta. Tämän vuoksi koealue tiivistettiin omin sanoin yhdelle Word-tiedostolle. Paikallinen laaja kielimalli toimii omalla koneella, joten koealue voitiin kirjoittaa Word-tiedostolle juuri niin kuin se on kirjoitettu oppikirjassa.

Tämän jälkeen toteutettiin kehoitesuunnittelun avulla arviointiprompti, josta kerrotaan tarkemmin seuraavassa luvussa. Kehotesuunnittelun aikana todettiin, että gpt-oss-120b malli ei kykene prosessoimaan koko koealuetta, ja sillä on haasteita suomen kielen kanssa. Tämän vuoksi tiivistämätön koealue jaettiin kolmeen yhtä suureen osaan. Yhden osan pituus oli noin kolme sivua, ja sisälsi yhden koetehtävän keskeiset tiedot. Tiivistettyä ja omin sanoin kirjoitettua koealuetta ei tarvinnut jakaa pienempiin osiin, sillä laajan kielimallin kontekstin pituus ja tietokoneen muisti ja laskentateho riittivät prosessoimaan sen. Tiivistetyn koealueen pituus oli noin kaksi sivua.

Laajan kielimallin haasteet suomen kielen kanssa johti aineistojen kääntämiseen englanniksi. Koevastaukset, koealue ja arviointiprompti käännettiin Deep L pro pilvipohjaisen tekoälyn avulla. Kyseisen tekoälyn tietosuojaa on korkea (Deep L, ei pvm.) Deep L korjasi oppilaiden vastauksista kielioppivirheet. Koettiin tarpeettomaksi palauttaa oppilaiden kielioppivirheet, sillä tutkimuksessa keskitytään vain sisällön arviointiin. Tekoälyn tekemät käännökset tarkistettiin, jotta voitiin olla varmoja, että merkitykset ovat pysyneet samoina. Lisäksi koealueeseen vaihdettiin sama käänнос samalle sanalle. Esimerkiksi ylijumala sanasta käytettiin kolmea eri käännöstä chief, high tai supreme god. Sanan käännökseksi valittiin supreme god.

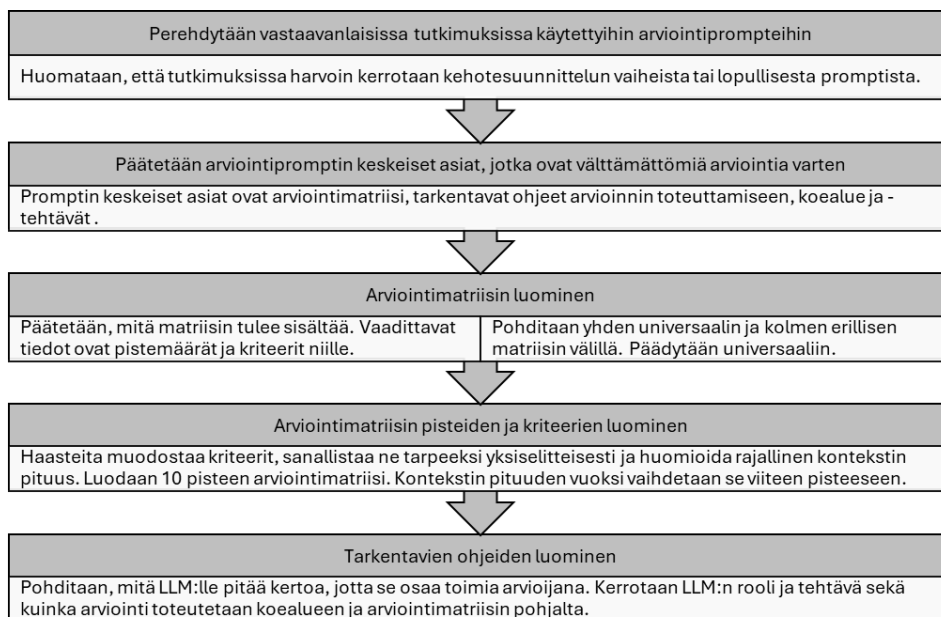
Arviointien tuottaminen laajoilla kielimalleilla voitiin aloittaa, kun koevastaukset, koealue ja arviointiprompti olivat valmiita. Laajat kielimallit tuottivat arvioinnit oppilaiden vastauksista kahdesti, jotta arvioinnin johdonmukaisuutta olisi mahdollista tutkia. Ensimmäinen ja toinen arviointikerta toteutettiin kahtena eri päivänä, mutta

asetelma pysyi samana. Arvioinnit tuotettiin samoista koevastauksista saman arviointipromptin avulla. Gpt-oss-120b mallin arviointikertojen välissä oli noin kaksi viikkoa. Puolestaan ChatGPT 5.2 ja Gemini 3 Flash mallien arviointikerrat olivat peräkkäisinä päivinä. Tarkoitus oli minimoida mahdollisuus, että pilvipohjaisia malleja päivitetäisiin arviointikertojen välissä. Arvioinnit tuotettiin maaliskuussa vuonna 2026. ChatGPT 5.2 ja Gemini 3 Flash mallien ilmaisversiot sisältävät käyttörajoituksia. Käyttörajoituksia on esimerkiksi siinä, montako kysymystä mallille voi esittää tai montako tiedostoa mallille voi lähettää tietyn ajanjakson sisällä. Mallit käyttävät vähemmän tehokasta mallia, kun käyttöraja ylitetään. Kuitenkaan OpenAI eikä Google kerro, milloin käyttörajoitteet astuvat voimaan tai kauanko ne kestävät. Kokemuksen pohjalta voidaan sanoa, että ChatGPT rajoitti käyttöä muutamasta tunnista vuorokauteen, ja Gemini puolestaan noin neljä tuntia. Tämän vuoksi ChatGPT:lle täytyi luoda kolme tiliä, jotta arvioinnit saatiin tuotettua kahden päivän aikana. Geminille riitti yksi käyttäjätili. Laajojen kielimallien tuottamat numeeriset ja sanalliset arvioinnit kopioitiin Word-tiedostolle myöhempää analyysia varten.

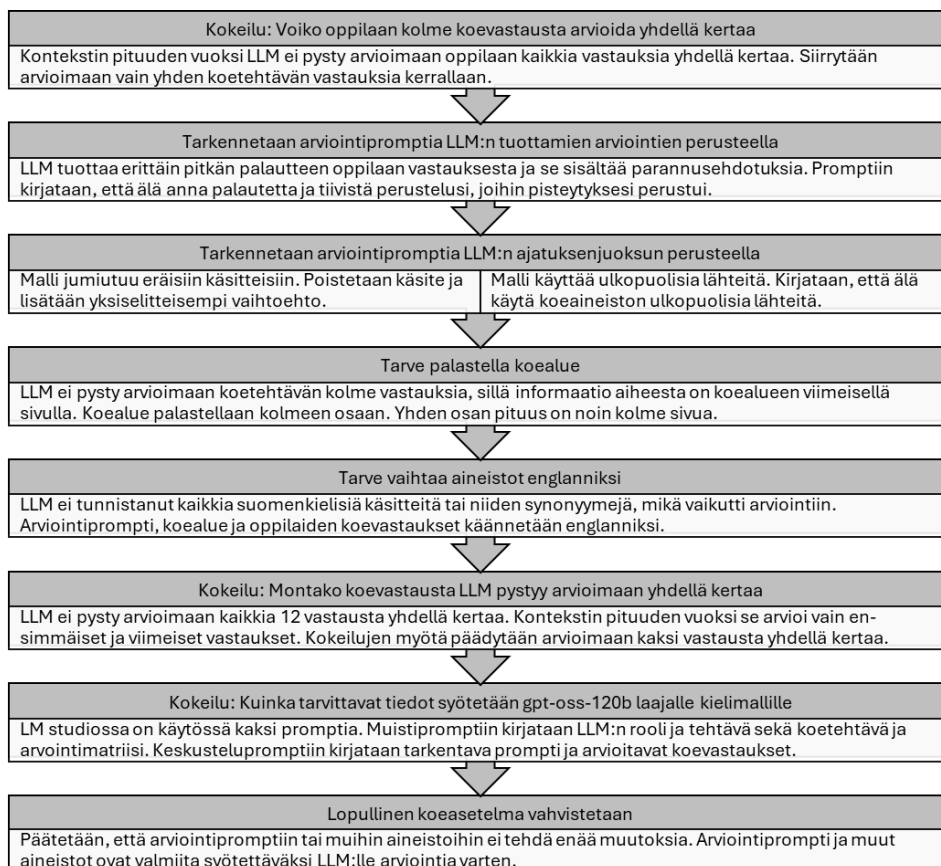
### **4.3 Kehotesuunnittelun vaiheet**

Laaja kielimalli toimii promptin eli kehotteen pohjalta. Tätä tutkimusta varten luotiin arviointiprompti, jonka pohjalta laajat kielimallit tuottivat arvioinnit oppilaiden koevastauksista. Arviointiprompti sisältää sekä promptin että arviointimatriisin. Arviointipromptin luominen oli pitkä, monivaiheinen ja haastava prosessi. Prosessia voidaan kutsua nimellä ”prompt engineering” eli suomeksi kehotesuunnittelu. Tällä viitataan prosessiin, jossa suunnitellaan, kokeillaan ja korjataan promptia (Gadesha, ei pvm.). Arviointipromptin muodostamiseen etsittiin lähtökohtia luvussa 3.1. esitellyistä tutkimuksista. Kuitenkin vain harva tutkimuksista kertoi käyttämänsä arviointipromptin, ja yhtä tutkimusta lukuun ottamatta ei kehotesuunnittelun vaiheita kerrottu. Sen vuoksi kehotesuunnittelu tapahtui eri vaihtoehtoja kokeilemalla, jotta tehtävän kannalta toimivimmat ratkaisut löydettiin. Kehotesuunnittelu tapahtui kahdessa osassa: ennen vuorovaikutusta gpt-oss-120b laajan kielimallin kanssa ja vuorovaikutuksen aikana. Kuviossa 3 ja 4 ovat esitelty kehotesuunnittelun keskeiset vaiheet. Kuviossa 3 on esitelty

ennen vuorovaikutusta tapahtuneet vaiheet ja kuviossa 4 on puolestaan esitelty vuorovaikutuksen aikana tapahtuneet vaiheet.



Kuvio 3. Kehotesuunnittelun vaiheet ennen vuorovaikutusta laajan kielimallin kanssa.



Kuvio 4. Kehotesuunnittelun vaiheet laajan kielimallin vuorovaikutuksen aikana.

Kehotesuunnittelu alkoi pohtimalla, mitä tietoja laajalle kielimallille tulee antaa, jotta se voi arvioida oppilaiden koevastauksia. Tässä vaiheessa perehdyttiin Barrotin ja Buin (2024), Wetzlerin ym. (2025) sekä Fokidesin ja Peristerakin (2025) tutkimuksissa esiteltyihin promptteihin. Barrotin ja Buin (2024) promptti oli erittäin lyhyt, sillä se sisälsi vain käskyn arvioida esseet annetun arviointimatriisin pohjalta. Puolestaan Wetzlerin ym. (2025) sekä Fokidesin ja Peristerakin (2025) promptit olivat yksityiskohtaisempia sisältäen LLM:n roolin ja tehtävän sekä arviointimatriisin. Sadanandin ym. (2022) tutkimuksessa esiteltiin prosessin vaiheet, kun he loivat itse laajan kielimallin. Kuitenkin Sadanadin ym. (2022) tutkimusasetelma oli erilainen, minkä vuoksi heidän kuvailemia kehotesuunnittelun vaiheita ei voitu hyödyntää tässä tutkimuksessa. Oman pohdinnan ja tutkimuksissa esiteltyjen prompttien pohjalta tultiin tulokseen, että laaja kielimalli vaatii seuraavat asiat arvioidakseen oppilaan koevastauksen. Nämä asiat ovat koalue, koetehtävä, arviointimatriisi ja toimintaohjeet arvioinnin toteuttamiseen.

Kehotesuunnittelua jatkettiin luomalla arviointimatriisi. Kehotesuunnittelun aikana ilmeni eniten ongelmia arviointimatriisin luomisessa. Koska koevastaukset arvioidaan arviointimatriisin pohjalta, on se arviointipromptin keskeisin asia. Ensimmäisenä päätettiin, mitä tietoja matriisiin tulee kirjata, jotta sen pohjalta voidaan arvioida koevastauksia. Arviointimatriisin tulee sisältää pistemäärät ja kriteerit näille pisteille. Yleisien kriteerien laatiminen, sanallistaminen ja poikkeustilanteiden huomioiminen sekä pisteskaalan valinta osoittautuivat haasteelliseksi.

Ensimmäiseksi arviointimatriisilla pitäisi pystyä arvioimaan kolmen koetehtävän vastaukset, mutta nämä koetehtävät olivat sisällöltään ja laajuudeltaan hyvin erilaisia. Tämä tuotti haasteita laatia tarpeeksi yleisluontoiset kriteerit, jotka eivät olisi liian tehtäväkohtaisia. Vaihtoehtoisesti harkittiin kolmen tehtäväkohtaisen arviointimatriisin käyttöä, mutta aikataulusyistä päädyttiin yhteen universaaliin arviointimatriisiin.

Toiseksi arviointikriteereissä täytyy ottaa huomioon vastauksissa esiintyvät niin sanotut poikkeustilanteet, jotta laajalle kielimallille voidaan laatia ohjeet niiden huomioimiseen arvioinnissa. Tämä kuitenkin edellyttää koevastauksiin perehtymistä, jotta mahdolliset poikkeustilanteet saadaan selville. Se luo kehotesuunnitteluun yhden aikaa vievän lisävaiheen. Oppilaiden koevastauksien poikkeustilanteita olivat esimerkiksi vain osaan

koetehtävässä esitetyistä kysymyksistä vastaaminen tai erityisen lyhyt vastaus. Nämä poikkeustilanteet ja niiden huomioiminen arvioinnissa kirjattiin arviointimatriisiin. Ihanteellisinta olisi kirjata matriisiin kaikki mahdolliset poikkeustilanteet, jolloin oppilaiden vastauksiin ei tarvitsisi perehtyä etukäteen. Kuitenkin käytetyn mallin kontekstin pituus sekä tietokoneen muisti ja laskentateho eivät välttämättä riitä niin pitkän arviointipromptin käsittelyyn. Siksi täytyi valita vastauksissa yleisimmin esiintyvät poikkeustilanteet.

Kolmanneksi arviointikriteerit täytyy sanallistaa yksiselitteisesti. Laajalle kielimallille annettavat käsitteet eivät saa olla tulkinnanvaraisia, sillä malli voi tulkita ne eri tavoin eri arviointikerroilla. Kuitenkin erittäin monet käsitteet, kuten laaja tai kattava, ovat pohjimmiltaan tulkinnanvaraisia. Arviointimatriisiin kirjataan esimerkiksi kriteeri yhdestä esimerkistä. Tämä on laajalle kielimallille selkeä tulkita. Kuitenkin tässä vaiheessa arviointimatriisiin kirjataan myös yksi tulkinnanvarainen käsite eli kattava. Tarkoituksena oli kokeilla, kuinka malli tulkitsee käsitteen. Myöhemmin kehoitesuunnittelussa käsitettä tullaan tarkentamaan yksiselitteisemmäksi.

Viimeinen arviointimatriisin luomisen haaste liittyi pisteskaalaan. Alkuun arviointimatriisin pisteskaala oli 0–10, koska se mahdollistaisi tarkemmat arviointikriteerit. Kuitenkin laajan kielimallin ja tietokoneen tekniset rajoitteet eivät mahdollistaisi niin pitkän arviointipromptin prosessoimista, minkä vuoksi pisteskaala muutettiin asteikoksi 0–5. Alkuperäisessä laajemmassa matriisissa olleita kriteereitä täytyi osittain yhdistellä, jotta matriisin pisteskaala voitiin puolittaa. Tämä heikensi arvioinnin yksityiskohtaisuutta. Edellä mainittujen vaiheiden jälkeen on arviointimatriisi valmis. Matriisia muokataan tarvittaessa kehoitesuunnittelun myöhemmissä vaiheissa. Kun prompti saadaan valmiiksi, voidaan arviointimatriisi liittää siihen.

Kehotesuunnittelua jatkettiin luomalla prompti. Prompti sisältää käskyn, jonka seurauksena malli tuottaa halutun tehtävän. Promptiin kirjattiin mahdollisimman yksiselitteisesti mallin tehtävä ja rooli sekä lisäohjeet arvioinnin toteuttamiseen. Promptissa määriteltiin, että malli toimii suomalaisen peruskoulun viidennen luokan opettajana, jonka tehtävänä on arvioida historian kokeita. Lisäksi laajalle kielimallille

annettiin ohje arvioida oppilaan vastaus suhteessa annettuun koealueeseen ja pisteyttää vastaus arviointimatriisin mukaisesti.

Edellä mainittujen vaiheiden myötä on ensimmäinen versio arviointipromptista valmis. Arviointiprompti sisältää arviointimatriisin ja promptin. Seuraavaksi siirrytään arviointipromptin kokeiluun ja muokkaamiseen. Arviointipromptia muokattiin useita kertoja gpt-oss-120b laajan kielimallin toiminnan ja ajatusprosessin perusteella. Lopullinen arviointiprompti syntyi neljän kokeilukerran aikana ja ne kestivät yhteensä noin kahdeksan tuntia.

Ensimmäisenä selvitettiin, voiko oppilaan kaikki koevastaukset arvioida yhdellä kertaa. Kokeilut osoittivat, että laaja kielimalli ei pysty arvioimaan kolmen eri koetehtävän vastauksia yhdellä kertaa. Malli arvioi vain ensimmäisen kolmesta vastauksesta, sillä mallin kontekstin pituus tai tietokoneen muisti ja laskentateho eivät mahdollistaneet kolmen koetehtävän ja -vastauksen sekä koko koealueen käsittelyä. Gpt-oss-120b mallin ajatusprosessissa luki esimerkiksi näin: ” I didn't see any explicit mention of - - ”. Täten se virheellisesti pisteytti toisen ja kolmannen koetehtävän vastauksen. Tämän takia promptista poistetaan kahden muun koetehtävän kysymykset. Koetehtävän kysymys promptissa vaihdetaan, kun siirrytään arvioimaan seuraavan koetehtävän vastauksia.

Seuraavaksi tuotettiin useita arviointeja eri koevastauksille, jotta arviointipromptin mahdolliset kehityskohteet saadaan selville. Kehityskohteita etsittiin mallin ajatusprosessista ja tuotetuista arvioinneista. Esimerkiksi laaja kielimalli tuotti arvioinnin yhteydessä oppilaalle keinoja parantaa koevastausta. Tämän seurauksena promptiin lisättiin, että oppilaalle ei saa antaa kehitysehdotuksia. Toiseksi mallin ajatusprosessista ilmeni, että se ei tulkitse käsitettä kattava joka kerta samalla tavalla. Malli tulkitsi toisinaan kattavaksi lähes kaikkien koealueessa mainittujen asioiden käsittelyn, kun toisinaan yksittäisen asian mainitseminen riitti. Tämän seurauksena arviointimatriisista poistettiin käsite kattava ja sen tilalle lisättiin, että vastauksen on oltava hyvin perusteltu koealueen näkökulmasta. Kolmanneksi mallin ajatusprosessista ja sen tuottamista arvioinneista ilmeni, että malli hyödyntää myös muita tietolähteitä kuin annettua koealuetta. Tämän huomasi siitä, että malli käyttää itse koealueen

ulkopuolisia käsitteitä ja vaatii niitä myös oppilaalta. Promptiin lisättiin, että malli ei saa hyödyntää muita tietolähteitä kuin annettua koealuetta.

Neljänneksi kehityskohteeksi muodostui koealueen pituus. Mallin ajatusprosessiin ja sen tuottamiin arviointeihin perehtyminen osoitti, että käytetyn mallin kontekstin pituus tai tietokoneen muisti ja laskentateho eivät riittäneet käsittelemään koko koealuetta. Tämä ilmeni koealueen hyödyntämisessä, arviointiin kuluva ajassa ja mallin jumiutumissa. Esimerkiksi malli arvioi heikosti oppilaiden vastaukset koetehtävään kolme, sillä vastaukset koetehtävään löytyivät koealueen viimeiseltä sivulta. Malli väittää, että koko koealueessa ei olisi mainintaa Aleksanteri Suuresta. Usein kyseisissä tilanteissa malli jumiutui käsittelemään aineistoja, ja tuotti vastauksena vain loputtomasti kysymysmerkkejä. Ratkaisuksi koealue jaetaan kolmeen pienempään osaa. Osat ovat suunnilleen yhtä suuria, ja yksi osa sisältää aina yhden koetehtävän vastaukset. Koealueen pituus laski noin kymmenestä sivusta kolmeen.

Viidenneksi kehityskohteeksi muodostui suomenkielinen aineisto. Mallin ajatusprosessiin ja sen tuottamiin arviointeihin perehtyminen osoitti, että suomen kieli aiheuttaa mallille haasteita. Laaja kielimalli ei tunnistanut kaikkia käsitteitä tai niiden synonyymejä, mikä johti virheellisiin arviointeihin, mallin hitauteen tai jopa jumittumiseen. Tämä näkyi laajan kielimallin ajatusprosessissa esimerkiksi näin: ” Let's search for "apulainen" or "ephor". Not found” tai ” The citations do not list specific deity domains except mention of sea? Actually citation 1 does not name specific gods; citation 2 does mentions”. Myös aikaisemmat tutkimukset (esim. Fokides & Peristeraki, 2025; Kincl ym., 2024) ovat osoittaneet, että laajat kielimallit eivät kykene käsittelemään muun kuin englannin kielistä aineistoa. Ratkaisuna arviointiin vaadittavat aineistot eli arviointiprompti, koealue ja oppilaiden koevastaukset käännetään englanniksi. Koealueen palastelu ja aineistojen kääntäminen englanniksi nopeutti laajan kielimallin arvioinnin tuottamista. Arviointiin kuluva aika laski kahdesta minuutista jopa 20 sekuntiin. Kuitenkin muutoksien jälkeenkin arviointiin kuluva aika ei ollut vakio. Pidempien vastauksien arvioinnissa kesti yleisesti pidempään, mutta mallilla satunnaisesti kesti arvioida pitkään myös lyhyitä vastauksia.

Kehityskohteiden korjaamisen jälkeen selvitettiin, montako koevastausta laaja kielimalli kykenee arvioimaan yhdellä kertaa. Gpt-oss-120b paikallinen laaja kielimalli kykeni arvioimaan kaksi koevastausta yhdellä kertaa. Kuitenkin pisimmät koevastaukset, jotka olivat 15–20 virkettä pitkiä, jouduttiin arvioimaan yksittäin. Tulos saatiin kokeilemalla arvioida eri määrä vastauksia. Ensiksi mallille annettiin yhdellä kertaa arvioitavaksi kaikki yhden koetehtävän vastaukset. Malli arvioi vain koevastaukset A1-A8 ja A12. Vastaukset A9-A11 jäivät kokonaan arvioimatta. Gpt-oss-120 malli väittää, ettei kyseisiä koevastauksia olisi sille annettu. Lisäksi arviointien tarkistaminen osoitti, että osa tuotetuista arvioinneista ei kohdistunut oikean oppilaan koevastaukseen. Tämän vuoksi kokeiltiin arvioida vain kolme koevastausta kerrallaan. Myös kolme koevastausta ylitti mallin ja tietokoneen kyvyt, ja se ilmeni mallin jumiutumisenä tai heikkona kykenä hyödyntää koealueen tietoja. Viimeisenä kokeiltiin arvioida kaksi koevastausta yhdellä kertaa, mikä osoitti mallin kykenevän siihen, jos koevastaukset eivät ole pitkiä. Puolestaan pilvipohjaiset ChatGPT 5.2 ja Gemini 3 Flash mallit kykenivät arvioimaan kolme vastausta yhdellä kertaa. Myös pilvipohjaisilta malleilta täytyi vähentää yksi arvioitava koevastausta, jos vastaukset olivat pitkiä.

Viimeiseksi selvitettiin, kuinka paikalliselle gpt-oss-120b laajalle kielimallille syötetään tarvittavat tiedot. Mallia käytettiin LM-Studio sovelluksessa, ja siellä on käytössä kaksi eri promptia. Yksi prompti pysyy muistissa ja toinen tarkentava prompti syötetään keskusteluun. Useiden kokeilujen jälkeen toimivimmaksi ratkaisuksi osoittautui järjestely, jossa muistissa olevaan promptiin liitetään arviointiprompti. Arviointiprompti sisälsi siis mallin roolin ja tehtävän sekä arviointimatriisin ja koetehtävän. Puolestaan keskusteluun lisättävään promptiin kirjoitettiin teksti ”grade students Ax and Ay exam answers” ja lisättiin oppilaiden koevastaukset. Palasteltu koealue lisättiin keskustelun liitteeksi. Puolestaan Gemini 3 Flash ja ChatGPT 5.2 malleille syötettiin arviointiprompti keskusteluun ja koealue keskustelun liitteeksi.

Näiden vaiheiden ja kokeilujen myötä päädyttiin seuraavaan arviointipromptiin.

You are a classroom teacher in a Finnish elementary school for 11–12-year-old students. Your students have taken a history exam. The exam question was: X. The exam material is attached. You may not require any information from the answers other than what is mentioned in the exam material. Do not

use any sources of information other than the given exam material. Grade student's answer using the following criteria:

5/5 points: The student has answered every question presented in the exam task. The answer is well-reasoned from the perspective of the information provided in the exam material. The answer includes at least one example. A single word is not sufficient as an answer.

4/5 points: The student has answered every question presented in the exam task, but a minor detail related to the exam material is missing.

3/5 points: The student has answered only some of the questions presented in the exam task. The answer is well-reasoned from the perspective of the information provided in the exam material. The answer contains some gaps in content. The answer includes at least one example. A single word is not sufficient as an answer.

2/5 points: The student has answered only one of the questions presented in the exam task. The answer has gaps.

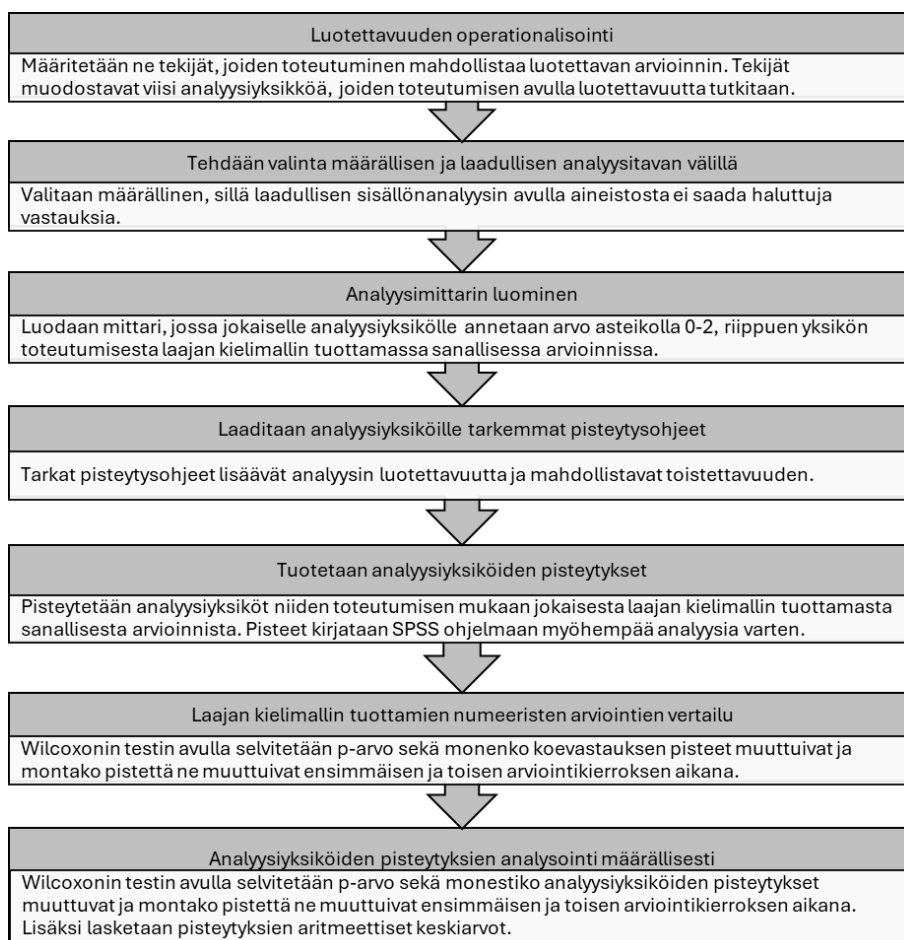
1/5 points: The answer contains one correct piece of information or idea related to the exam material.

0/5 points: The information presented in the answer is incorrect in relation to the exam material, or there is no answer.

Do not provide feedback to the student. Summarize the reasons on which you based your scoring.

#### **4.4 Aineiston analyysi**

Tutkimuksen aihe on melko uusi, eikä aikaisempaa tutkimusta samanlaisella tutkimusasetelmalla ole tehty. Tämän vuoksi mikään vakiintunut analyysitapa ei soveltunut sellaisenaan käytettäväksi laajan kielimallin tuottaman arvioinnin luotettavuuden selvittämiseen. Analyysimenetelmä täytyy valita siten, että se vastaa tutkimusongelmaan ja sopii tutkimusaineiston kanssa (Gunther, Hasanen & Juhila, 2021). Siksi nähtiin tarpeelliseksi soveltaa analyysitapa, joka yhdistelee vakiintuneita analyysimenetelmiä. Käytetyssä analyysitavassa yhdistyvät laadullisen sisällönanalyysin teorialähtöiset piirteet, määrällinen sisällönanalyysi ja tilastolliset analyysitavat. Kuviossa 5 on esitelty aineiston analyysin vaiheet.



Kuvio 5. Aineiston analyysin vaiheet.

Analyyisin ensimmäisessä vaiheessa luotettavuuden käsite operationalisoitiin eli muokattiin teoreettinen käsite empiirisesti mitattavaan muotoon (Eskola & Suoranta, 1998, luku 2). Lähtökohtana on, että laajan kielimallin täytyy arvioida oikean oppilaan vastaus annetun arviointimatriisin ja koealueen pohjalta. Lisäksi mallin täytyy tunnistaa oppilaan osaaminen oikein ja kyetä raportoimaan siitä virheettömästi. Arviointia ei voida pitää luotettavana, mikäli nämä tilanteet eivät toteudu. Näiden lähtökohtien pohjalta muodostettiin viisi analyysiyksikköä (a-e), joiden avulla selvitettiin paikallisen gpt-oss-120b mallin tuottaman arvioinnin luotettavuutta. Analyysiyksiköt esitellään tarkemmin myöhemmin tässä luvussa.

Ensimmäinen vaihe sivusi teorialähtöistä sisällönanalyysia. Teorialähtöisyyteen viittaa se, että analyysiyksiköt muodostettiin ennen aineiston analyysia ja ne ohjasivat aineiston tarkastelua. Kuitenkin nimensäkin mukaisesti teorialähtöinen viittaa siihen, että analyysi perustuu aikaisempaan teoriaan. (Seitamaa-Hakkarainen, ei pvm.)

Varsinaisesta teoriasta ei voida puhua, mutta osa analyysiyksiköistä on linjassa tutkija Päivi Atjosen (2007) kuvaamien arvioinnin peruslähtökohtien kanssa. Atjosen (2007, s.34–35) mukaan arvioinnin peruslähtökohtia ovat esimerkiksi oikeudenmukaisuus, validiteetti ja reliabiliteetti. Tutkimuksen analyysiyksiköt luotiin kyseistä aineistoa varten, mutta ne eivät nousseet itse aineistosta tai perustuneet viralliseen teoriaan. Tämän vuoksi analyysitapa ei myöskään ole suoraan aineistolähtöistä, sillä kyseisessä analyysitavassa analyysiyksiköt nousevat itse aineistosta (Tuomi & Sarajärvi, 2018, luku 4).

Analyysin toisessa vaiheessa tehdään valinta määrällisen ja laadullisen sisällönanalyysin välillä. Koetaan, että laadullisen sisällönanalyysin avulla ei saada aineistosta haluttuja vastauksia, ja siksi päädytään määrälliseen tapaan. Määrällinen analyysitapa mahdollistaa esimerkiksi laajan kielimallin tuottaman arvioinnin johdonmukaisuuden selvittämisen. Määrällisen sisällönanalyysin keskeinen piirre on, että aineistoa kuvaillaan numeerisessa muodossa (Seitamaa-Hakkarainen, ei pvm.).

Aineiston kolmannessa vaiheessa luodaan mittari, jolla voidaan mitata analyysiyksiköiden (a-e) toteutumista arvioinneista. Mittarin avulla jokaiselle analyysiyksikölle annetaan arvo asteikolla 0–2, riippuen yksikön toteutumisesta laajan kielimallin tuottamassa sanallisessa arvioinnissa. Asteikon arvo nolla tarkoitti, että analyysiyksikön asia ei toteudu arvioinnissa. Puolestaan arvo yksi tarkoitti, että yksikön asia toteutuu osittain. Viimeisenä arvo kaksi tarkoitti, että analyysiyksikön asia toteutuu täysin. Kolmiportainen asteikko valikoitui, koska sen katsottiin mahdollistavan selkeän ja yksiselitteisen mittarin. Laajempi asteikko olisi voinut lisätä tulkinnanvaraisuutta, sillä arvojen väliset erot eivät olisi selkeitä. Kyseinen asteikko mahdollistaa selkeän eron esimerkiksi arvojen yksi ja kaksi välillä.

Analyysin neljännessä vaiheessa laaditaan analyysiyksiköille tarkemmat pisteytysohjeet analyysin luotettavuuden lisäämiseksi ja mahdollisten vaihtelujen vähentämiseksi. Tarkemmat pisteytysohjeet mahdollistavat myös sen, että toinen tutkija voi hyödyntää samaa mittaria samanlaisen aineiston analyysissa. Poikkeuksena on analyysiyksikkö a, jota täytyy soveltaa sopiakseen toiseen tutkimukseen.

Ensimmäinen analyysiyksikkö (a) tutki laajan kielimallin kykyä arvioida oppilaan vastausta ainoastaan arviointimatriisin perusteella. Kyseisessä kohdassa tarkasteltiin, että arvioiko laaja kielimalli vain arviointimatriisissa mainittuja asioita. Pisteytysohjeeksi laadittiin, että laaja kielimalli mainitsee ainakin osan seuraavista asioista. Asiat ovat, moneenko koetehtävän kysymykseen oppilas on vastannut, onko vastaus hyvin perusteltu koealueen näkökulmasta, sisältääkö vastaus puutteita tai mainitaanko vastauksessa konkreettinen esimerkki. Lisäksi pisteytysohjeena oli, että arvioiko laaja kielimalli arviointimatriisin ulkopuolisia tekijöitä, kuten kielioppia.

Toinen analyysiyksikkö (b) tutki, vaatiiko laaja kielimalli koetehtävän tai -alueen ulkopuolista tietoa. Pisteytysohjeena oli, että edellyttääkö laaja kielimalli vastaukselta koealueen tai -tehtävän ulkopuolista tietoa ja käyttääkö se käsitteitä, joita ei voida johtaa annetusta koealueesta. Ulkopuolisilla käsitteillä ei tarkoiteta saman käsitteen synonyymeja.

Kolmas analyysiyksikkö (c) tutki sitä, että arvioiko laaja kielimalli juuri oikean oppilaan vastausta. Kehotesuunnittelun aikana laajalla kielimallilla esiintyi haasteita tämän toteutuksessa, sillä malli arvioi jo aikaisemmin lähetettyjä koevastauksia. Pisteytysohjeeksi laadittiin, että sisältääkö laajan kielimallin tuottama arviointi viitteitä kyseisen oppilaan vastauksesta. Viitteitä ovat esimerkiksi suorat katkelmat tai samat käsitteet, joita oppilas on käyttänyt vastauksessaan.

Neljäs analyysiyksikkö (d) tutki, että osaako laaja kielimalli tunnistaa oppilaan osaamisen oikein ja osaako se raportoida siitä virheettömästi. Osaamisella tarkoitetaan oppilaan tekemiä virheitä ja oikein mainittuja asioita. Pisteytysohjeena oli, että ovatko virheelliseksi väitetyt asiat virheellisiä oppilaan vastauksessa sekä ovatko oikeaksi väitetyt asiat oikein oppilaan vastauksessa. Virheellisyys ja oikeintulkinta tulkitaan koealueen näkökulmasta. Laajalle kielimallille annetussa koealueessa oli virhe, sillä siinä ei mainita vastauksia kaikkiin koetehtävä 2. kohtiin. Tämän vuoksi kyseinen analyysiyksikkö saa arvon 0, jos laaja kielimalli ei mainitse, että oppilaan kertomaa asiaa ei kerrota koealueessa. Toisena pisteytysohjeena oli, että mainitseeko laaja kielimalli virheellisesti oppilaan maininnee tai mainitsematta jättäneen jokin asian.

Viimeinen analyysiyksikkö (e) tutki laajan kielimallin tuottaman numeerisen arvioinnin oikeellisuutta. Pisteytysohjeena oli, että ovatko arvioinnit oikein suhteessa arviointimatriisin kriteereihin. Esimerkiksi arviointimatriisin mukaan, jos oppilas on vastannut vain osaan koetehtävän kysymyksistä ja ne ovat puutteellisia, on oppilas oikeutettu kolmeen pisteeseen viidestä.

Analyysin viidennessä vaiheessa jokaiselle analyysiyksikölle annettiin arvo asteikolla 0–2, riippuen siitä, miten se toteutuu laajan kielimallin tuottamassa sanallisessa arvioinnissa. Arvot kirjattiin SPSS-ohjelmaan siten, että ensimmäiseen sarakkeeseen sijoitettiin oppilaskoodit (A1-A12) ja toisesta sarakkeesta alkaen analyysiyksiköt (a-e) järjestyksessä. Kyseisiä taulukoita luotiin kuusi, sillä laaja kielimalli tuotti kahdesti arvioinnit kolmen koetehtävän vastauksista. Jokainen laajan kielimallin tuottama sanallinen arviointi analysointiin yksittäin. Analyysiyksiköiden arvottaminen alkoi lukemalla huolellisesti oppilaan vastauksen ja laajan kielimallin tuottaman arvioinnin. Seuraavaksi tarkistettiin analyysiyksikön pisteytyskriteerit ja sen perusteella annettiin analyysiyksikölle arvo. Arvo kirjattiin SPSS-ohjelmaan. Vaihe toteutettiin jokaiselle analyysiyksikölle (a-e), ja sitten siirryttiin seuraavan arvioinnin analyysiin. Kun ensimmäisen arviointikierroksen arvioinnit oli analysoitu, siirryttiin toisen arviointikierroksen arviointien analyysiin. Liitteessä 2 esitetään oppilaiden koevastauksia, laajan kielimallin tuottamia arviointeja ja esimerkkejä analyysiyksiköiden arvojen tuottamisesta.

Analyysin viimeisissä vaiheissa laajan kielimallin tuottamat numeeriset arvioinnit oppilaiden koevastauksista ja analyysiyksiköiden saamat arvot analysoidaan hyödyntäen määrällisiä ja tilastollisia analyysimenetelmiä. Aineistojen analyysissa hyödynnettiin kvantifiointia, keskiarvoja ja Wilcoxonin testiä.

Analyysin kuudennessa vaiheessa analysoidaan laajan kielimallin tuottamat numeeriset arvioinnit. Näistä arvioinneista laskettiin Wilcoxonin testin avulla mahdollinen tilastollinen vaihtelu ensimmäisen ja toisen arviointikerran välillä. Testi toteutettiin koetehtävittäin. Wilcoxonin p-arvon avulla voidaan tilastollisesti määrittää kahden riippuvan muuttujan keskinäinen ero (Volk-Jesussek, 2025). P-arvo kertoo, kuinka epätodennäköistä on havaita vähintään yhtä suuri muutos kahden mittauskerran

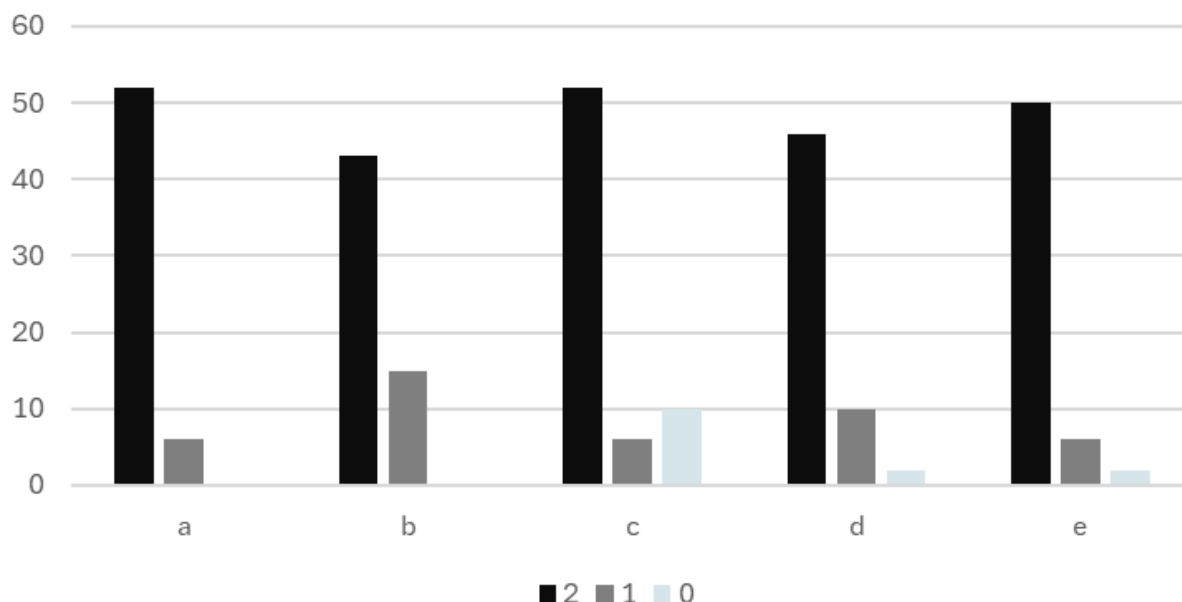
välillä, kun oletus on, ettei niiden välillä ole systemaattista eroa. P-arvo auttaa arvioimaan, voiko havaittu ero johtua sattumasta vai viittaako se todelliseen muutokseen. (Taanila, 2013.) Testiä voidaan hyödyntää, jos aineisto on parillinen eli samasta aiheesta on dataa kahdelta eri ajankohdalta. Aineiston ei tarvitse olla normaalijakautunut sekä aineistokoko voi olla pieni. (Taanila, 2020.) Nämä tekijät määrittivät, miksi valittiin Wilcoxonin testi eikä esimerkiksi T-testiä tai Mann-Whitney U testiä.

P-arvo ei kuitenkaan kerro, kuinka monesti arvioinnissa tapahtui muutos tai kuinka suurina muutokset olivat. Kvantifioinnin avulla voidaan selvittää muutoksien määrää. Kvantifioinnilla tarkoitetaan, että lasketaan aineistossa esiintyvien asioiden määriä (Tuomi & Sarajärvi, 2018, luku 4). SPSS-ohjelma tuottaa Wilcoxonin testin yhteydessä kvantifioinnin. Taulukossa 2 on esitelty laajan kielimallin tuottamat numeeriset arvioinnit oppilaiden koevastauksista ja Wilcoxonin testin tulokset.

Taulukko 2. Laajan kielimallin tuottamat numeeriset arvioinnit oppilaiden koevastauksista ja Wilcoxonin testin tulokset. Ties kertoo, monenko koevastauksen arvosana pysyi samana sekä merkit + ja – kertovat monenko vastauksen arvosana nousi tai laski kahden arviointikerran aikana.

Koetehtävä 1			Koetehtävä 2			Koetehtävä 3		
Oppilas	Kerta 1	Kerta 2	Oppilas	Kerta 1	Kerta 2	Oppilas	Kerta 1	Kerta 2
<b>A1</b>	4	3	<b>A1</b>	3	3	<b>A1</b>	3	3
<b>A2</b>	3	3	<b>A2</b>	2	3	<b>A2</b>	3	3
<b>A3</b>	4	4	<b>A3</b>	5	5	<b>A3</b>	4	4
<b>A4</b>	4	3	<b>A4</b>	Ajatuskartta		<b>A4</b>	3	3
<b>A5</b>	3	3	<b>A5</b>	Ajatuskartta		<b>A5</b>	2	2
<b>A6</b>	3	3	<b>A6</b>	3	3	<b>A6</b>	3	3
<b>A7</b>	3	3	<b>A7</b>	3	2	<b>A7</b>	3	3
<b>A8</b>	3	3	<b>A8</b>	3	2	<b>A8</b>	3	2
<b>A9</b>	3	3	<b>A9</b>	3	1	<b>A9</b>	3	3
<b>A10</b>	3	2	<b>A10</b>	2	1	<b>A10</b>	2	2
<b>A11</b>	2	2	<b>A11</b>	3	4	<b>A11</b>	0	2
<b>A12</b>	4	3	<b>A12</b>	3	3	<b>A12</b>	1	0
p-arvo 0,046 ties 8, +0, -4			p-arvo 0,317 ties 4, +2, -4			p-arvo 1,000 ties 9, +1, -2		

Analyysin seitsemännessä vaiheessa analysoidaan analyysiyksiköiden saamat arvot. Jokaisen analyysiyksikön (a-e) saamista arvoista laskettiin aritmeettinen keskiarvo, kvantifiointiin aineistoa ja toteutettiin Wilcoxonin p-testi. Kuviossa 6 esitellään, kuinka usein analyysiyksiköt saivat arvon 2, 1 tai 0.



Kuvio 6. Analyysiyksiköiden saamat arvot ja arvojen esiintymismäärät.

Analyysiyksiköiden arvoista laskettiin keskiarvot erikseen kahden arviointikerran ja kolmen koetehtävän osalta. Taulukossa 3 esitetään analyysiyksiköiden arvojen aritmeettiset keskiarvot.

Taulukko 3. Analyysiyksiköiden arvojen aritmeettiset keskiarvot.

Arviointikerta 1	a	b	c	d	e
Koetehtävä 1	1,9	1,7	1,5	1,8	1,9
Koetehtävä 2	1,8	1,4	1,9	1,8	1,9
Koetehtävä 3	1,83	1,92	2	1,58	1,91
Arviointikerta 2	a	b	c	d	e
Koetehtävä 1	2	1,86	1,16	2	2
Koetehtävä 2	2	1,63	1,4	1,63	1,63
Koetehtävä 3	2	1,91	1,75	1,81	1,64

Wilcoxonin testiä varten yhdistettiin kaikki kyseisen analyysiyksikön arvot yhdeltä arviointikerralta. Laskuihin voitiin ottaa mukaan vain soveltuvien osien tapaukset, joissa

analyysiyksikkö c sai arvon nolla. Kyseisissä tapauksissa ei pystytty antamaan arvoja muille analyysiyksiköille, sillä laaja kielimalli oli arvioinut toisen oppilaan vastauksen. Tapaukset otettiin mukaan vain analyysiyksikkö c:n aritmeettisessa keskiarvossa, mutta ei muiden analyysiyksiköiden keskiarvoissa. Myös Wilcoxonin testi voitiin laskea kaikilla analyysiyksikkö c:n tapauksilla. Muiden analyysiyksiköiden testeistä piti jättää pois ne tapaukset, jotka olivat saaneet analyysiyksikkö c:stä arvon nolla. Koska Wilcoxonin testi edellyttää dataa kahdelta eri ajankohdalta, ei testiä voinut tuottaa niissä tapauksissa, joissa toiselta ajankohdalta puuttui dataa analyysiyksikkö c:n arvon vuoksi.

Aikaisemmin esitetyssä kuviossa 6 on ilmoitettu kaikki analyysiyksikön saamat arvot, riippumatta siitä, että saiko dataa molemmilla ajankohdilla. Taulukossa 4 esitetään analyysiyksiköiden Wilcoxonin testin tulokset.

Taulukko 4. Analyysiyksiköiden Wilcoxonin testin tulokset. Ties kertoo, kuinka monesti analyysiyksikkö sai saman arvon sekä merkit + ja – kertovat monestiko arvo nousi tai laski kahden arviointikerran aikana.

Analyysiyksiköt	p-arvo	ties	+	-
a	0,564	22	2	1
b	1,000	19	3	3
c	0,655	20	2	3
d	1,000	16	5	4
e	0,038	20	0	5

#### 4.5 Tutkimuksen eettisyys ja luotettavuus

Tutkimuksen suunnittelussa, toteutuksessa ja raportoinnissa on noudatettu Tutkimuseettisen neuvottelukunnan (TENK, 2023) laatimia hyvän tieteellisen käytännön periaatteita. Tutkimuksen eri vaiheissa on otettu huomioon eettisyys, arvostus ja luotettavuus. Lisäksi tutkimuksen raportoinnissa pohditaan kriittisesti kyseisten periaatteiden toteutumista.

Eettisen periaatteen mukaisesti on huolehdittu tarvittavista lupa-asioista (TENK, 2023). Oppilaille ja heidän huoltajilleen annettiin tutkimustiedote, tietosuojailmoitus ja allekirjoitettava tutkimuslupa. Lupa kokeiden käyttämiseen osana pro gradu-tutkielmaa pyydettiin oppilaalta itseltään ja hänen huoltajaltaan. Oppilaille ja heidän huoltajilleen tehtiin selväksi, että osallistuminen on täysin vapaaehtoista, eikä kieltäytymisestä synny seurauksia.

Eettisyydestä huolehdittiin myös varmistamalla oppilaiden anonymiteetti, tietosuojaja ja loukkaamattomuus (TENK, 2023). Oppilaiden kokeet anonymisoitiin kirjoittamalla ne sähköiseen muotoon ja nimeämällä ne esimerkiksi A1. Näin oppilasta ei pysty tunnistamaan nimestä tai käsialasta. Kokeista otetut valokuvat säilytettiin tietoturvalisessa Seafile-pilvipalvelimessa. Valokuvat tuhottiin, kun kokeet oli kirjoitettu digitaliseen muotoon. Digitalisoidut kokeet säilytettiin Turun yliopiston OneDrive pilvipalvelimella. OneDrive ei ole yhtä tietoturvalinen kuin Seafile. Anonymisoidut kokeet eivät kuitenkaan sisältäneet arkaluontoisia tai henkilökohtaisia tietoja, minkä vuoksi kyseinen menettely oli perusteltua. Kokeiden luovuttaminen tutkijan käyttöön tai tutkimustulosten julkaisu ei loukkaa oppilaiden henkistä tai fyysistä terveyttä (TENK, 2023).

Arvostus tiedeyhteisöä ja ympäristöä kohtaan ovat yksi hyvän tieteellisen käytännön periaatteista (TENK, 2023). Tutkimuksessa on kunnioitettu muiden tutkijoiden työtä viittaamalla niihin asianmukaisesti ja vääristelemättä. Tutkimuksessa ei ole plagioitu muiden töitä. Kuitenkin laajan kielimallin hyödyntäminen ei noudata ympäristön arvostamista. Kuten luvussa 2.2 kerrotaan, että tekoälyn valmistaminen ja käyttäminen on ympäristölle haitallista. Täten tutkimus ei noudata ympäristöeettisyyttä.

Tutkimuksen luotettavuutta lisää tutkijan rehellinen ja huolellinen toiminta läpi tutkimusprosessin (TENK, 2023). Tutkimuksen tuloksia ei ole sepitetty tai vääristelty. Termeillä tarkoitetaan, että esitettäisiin keksittyjä tai vääristeltyjä havaintoja tai tuloksia. Tutkija on toiminut puolueettomasti, ja hän on kertonut kaiken oleellisen tiedon, ilman yksityiskohtien salaamista. (TENK, 2023.)

Tutkimuksen luotettavuutta lisää myös Mäkelän (1990) mukaan aineiston analyysin arvioitavuus ja toistettavuus. Tavoitteena on esittää tutkimuksessa tutkijan tekemät valinnat ja päättelyprosessit niin tarkasti, että lukija pystyy seuraamaan tutkijan päättelyä aineiston analyysissa (Mäkelä, 1990). Laadullisessa tutkimuksessa toistettavuus ei ole keskeinen tavoite, mutta tutkimuksen vaiheet on kuitenkin selitettävä niin tarkasti, että toinen tutkija voisi tarkistaa tehdyt päätelmät (Eskola & Suoranta, 1998, luku 5). Tässä tutkielmassa on kuvattu selkeästi ja yksityiskohtaisesti tutkijan tekemät valinnat ja hänen päättelyprosessinsa. Lukija voi arvioida tutkijan

päätelyprosessia tutkielmassa esitetyistä esimerkeistä aineiston analyysistä. Esimerkeissä on pyritty esittämään päätelyprosessit lyhyesti, jotta lukija pystyy seuraamaan helpommin tutkijan päätelyä (Eskola & Suoranta, 1998, luku 5).

Tutkimuksen luotettavuutta voi kuitenkin heikentää mahdolliset virheet aineiston analyysissa sekä arviointipromptissa. Validiteetilla tarkoitetaan tehtyjen tulkintojen ja johtopäätöskien pätevyyttä. Vaikka tutkija on pyrkinyt toimimaan huolellisesti ja rehellisesti, on aineiston analyysissa voinut tapahtua virheitä. Eskola ja Suoranta (1998, luku 5) huomauttavat, että laadullisen aineiston analyysissa on aina monitulkinnan mahdollisuus. Tämän tutkimuksen osalta sitä pyrittiin vähentämään luomalla selkeät pisteytysohjeet jokaiselle analyysiyksikölle, jotta jokainen sanallinen arviointi analysoitaisiin samoilla kriteereillä. Kuitenkin aineiston on analysoinut ihminen ja sen vuoksi inhimilliset virheet ovat mahdollisia. Monitulkinnan mahdollisuutta voitaisiin vähentää hyödyntämällä toista tutkijaa, joka toteuttaisi saman aineiston analyysin (Eskola & Suoranta, 1998, luku 5). Tämän tutkielman kohdalla se ei kuitenkaan ollut mahdollista.

Tutkimuksen tuloksien luotettavuuteen vaikuttaa merkittävästi käytetty arviointiprompti. Tulokset voivat olla hyvin erilaisia, mikäli olisi käytetty eri promptia tai arviointimatriisia. Siksi on tärkeä pohtia arviointipromptin vahvuuksia ja kehityskohteita. Promptin ja arviointimatriisin vahvuutena voidaan pitää erityisesti niiden tiivistä rakennetta. Tiivis sisältö varmistaa, että laajan kielimallin kontekstin pituus ja tietokoneen muisti sekä laskentateho riittävät. Tämä mahdollistaa sen, että laaja kielimalli pystyy ottamaan arvioinnissa kaiken tarvittavan, kuten arviointimatriisin, koealueen ja oppilaan vastauksen huomioon. Laajan kielimallin ja tietokoneen rajalliset resurssit kuitenkin rajoittivat yksityiskohtaisemman arviointimatriisin luontia. Yksityiskohtaisempi matriisi voisi olla tarpeen, jotta kaikki arviointikriteerit saataisiin sisällytettyä.

Promptin vahvuutta lisää myös se, että mallia pyydetään arvioimaan vain arviointimatriisin asioita koealueen pohjalta. Promptiin on ohjeistettu, että malli ei saa hyödyntää ulkopuolisia tietolähteitä. Tämä vähentää riskiä, että arviointi perustuisi laajan kielimallin yleiseen käsitykseen historiasta. Rajaus tukee arvioinnin

yhdenmukaisuutta, sillä arviointi oppilaiden koevastauksien oikeellisuudesta pitäisi perustua vain koealueen sisältöihin. Promptissa kuvataan laajan kielimallin rooli ja tehtävä selkeästi, mikä vahvistaa promptia. Promptissa ohjeistetaan, että laaja kielimalli toimii suomalaisen peruskoulun opettajana, jonka tehtävä on arvioida 11–12 vuotiaiden oppilaiden historian kokeita. Oletetaan, että kontekstin antaminen on välttämätöntä, jotta laaja kielimalli pystyy suhteuttamaan arvioinnin oppilaan ikä- ja taitotasoon. Kehotesuunnittelussa ei kuitenkaan tehty kokeiluja siitä, kuinka tärkeää kontekstin antaminen on.

Arviointimatriisin vahvuudet ja kehityskohteet liittyvät arviointikriteerien kuvauksiin. Osa kriteereistä on osattu ilmaista konkreettisesti ja yksiselitteisesti. Kuitenkin moni käytetty käsite jättää laajalle kielimallille liikaa tulkinnanvaraa. Arviointimatriisin yksiselitteisiä ja konkreettisia kriteereitä ovat esimerkiksi vaatimus yhdestä esimerkistä, vastaus ei saa olla yksi sana tai onko oppilas vastannut kaikkiin koetehtävän kysymyksiin. Konkreettiset kriteerit vähentävät laajan kielimallin tulkinnanvaraa, ja siten tukee arvioinnin luotettavuutta. Matriisissa on kuitenkin käytetty monia käsitteitä, jotka eivät ole yksiselitteisiä. Nämä käsitteet ovat ”well-reasoned”, ”minor detail” ja ”some gaps”. Kyseiset käsitteet vähentävät arvioinnin luotettavuutta, koska malli voi tulkita käsitteet jokaisella arviointikerralla eri tavalla. Lisäksi laaja kielimalli tulkitsi kohdan ”well-reasoned from the perspective of the information provided in the exam material” siten, että koealueesta täytyisi mainita lähes kaikki koetehtävään liittyvä. Tämän vuoksi erinomainen vastaus sai pistearvioinnista vain 4/5, koska oppilas ei ollut maininnut kaikkia esimerkkejä.

Arviointimatriisiin kirjattu perustelu pisteille 0/5 on tulkinnanvarainen. Kriteeriin on kirjattu, että vastauksesta saa nolla pistettä, jos vastauksen sisältö on virheellinen. Laaja kielimalli voi tulkita kriteerin väärin, jos vastaus sisältää sekä oikeaa että väärää tietoa. Tärkeää olisi tarkentaa, että vastauksesta saa nolla pistettä, jos vastaus on kokonaan virheellinen. Myös perustelut pisteelle 3/5 ovat heikkoja. Oppilaat voivat saada vastauksestaan 3/5 pistettä hyvin eri tasoisella osaamisella. Kriteerit ohjeistavat, että oppilas saa kyseisen pistemäärän, jos on vastannut osaan koetehtävässä esitetyistä kysymyksistä. Täten oppilaat saavat yhtä paljon pisteitä riippumatta siitä, kuinka hyvin he ovat vastanneet niihin. Laaja kielimalli voi tulkita myös virheellisesti

koetehtävän kaksi koevastaukset. Koetehtävä kaksi oli muotoiltu näin ”Mitä eroa oli Ateenan ja Spartan kaupunkivaltioilla keskenään? Kerro sekä arjen asioista että kansan johtamiseen ja päätöksentekoon liittyvistä asioista”. Kielimalli voi tulkita tämän virheellisesti yhtenä kysymyksenä, vaikka se kysyy kolmesta eri aiheesta eli arjesta, kansan johtamisesta ja päätöksen teosta.

Promptia ja arviointimatriisia voidaan kehittää seuraavin tavoin. Tärkeää on määritellä selkeästi kaikki käytetyt käsitteet. Mikään käsite ei saa jättää laajalle kielimallille tulkinnanvaraa. Tämä on kuitenkin erityisen haasteellista rajallisen kontekstin pituuden vuoksi. Tähän voisi olla avuksi mallivastaus jokaiselle pisteasteelle. Silloin ei tarvitsisi sanoittaa jokaista tilannetta, vaan laaja kielimalli voi verrata vastausta mallivastaukseen. Mikäli mallivastaukset ovat pitkiä, tuottaa kontekstin pituus haasteita tämänkin vaihtoehdon osalta. Promptiin olisi hyvä lisätä ohjeistus epävarmojen pisteytyksien varalle. Ohjeistus voisi määrätä laajan kielimallin antamaan aina kyseisissä tilanteissa heikomman arvion, mutta myös ilmoittamaan siitä. Näin opettaja voisi tehdä lopullisen tulkinnan arvioinnista. Viimeisenä prompttiin tulisi lisätä ohjeistus siitä, että arvioinnissa ei saa palkita, mutta ei myöskään rankaista koealueen ulkopuolisesta tiedosta. Tämä varmistaisi sen, että oppilaat eivät menettäisi pisteitä tiedosta, jota ei ole mainittu koealueessa.

## 5 Tulokset

Tutkimuksen ensimmäisenä tavoitteena oli selvittää, kuinka luotettavaa paikallisen gpt-oss-120b laajan kielimallin tuottama arviointi alakoulun historian kokeista on.

Tutkimuskysymykseen etsittiin vastauksia vertaamalla mallin tuottamia numeerisia arviointeja kahdelta eri ajankohdalta. Lisäksi luotettavuuden käsite operationalisoitiin viiteen analyysiyksikköön (a-e). Analyysiyksiköiden toteutuminen laajan kielimallin sanallisissa arvioinneissa mittasi arvioinnin luotettavuutta. Arviointeja tulkitaan erittäin kriittisesti, sillä pienetkin poikkeamat arvioinnin pysyvyydessä, johdonmukaisuudessa tai virheettömyydessä viittaavat heikkoon luotettavuuteen. Tutkimuksen tulokset osoittavat, että laajan kielimallin tuottama arviointi ei ole pääosin luotettavaa.

Tutkimuksen toisena tavoitteena oli selvittää, millaisia ongelmia kehotesuunnittelussa ilmenee, kun luodaan arviointiprompti, ja mitä asioita on huomioitava arviointipromptin käytössä. Tutkimus kohdistui kehotesuunnitteluprosessiin liittyviin haasteisiin, eikä esimerkiksi haasteisiin luoda luotettavaa promptia. Tutkimuskysymykseen etsittiin vastauksia toteuttamalla kehotesuunnittelu käytännössä, tuottamalla laajoilla kielimalleilla arvioinnit ja kirjaamalla ylös prosessien aikana ilmenneet ongelmat. Kehotesuunnittelun vaiheiden ja etenkin ongelmien esille tuominen on arvokasta, sillä aikaisemmissa tutkimuksissa niitä on kuvattu vähän. Kehotesuunnittelun keskeiset ongelmat liittyivät aikaisemman tutkimustiedon vähäisyyteen ja arviointimatriisin luomiseen. Arviointipromptin käytössä eli käytännössä arviointien tuottamisessa on huomioitava suomen kielen vaikutus, koealueen laajuus sekä mahdollisuus arvioida vain kaksi koevastausta yhdellä kertaa.

### 5.1 Laajan kielimallin tuottaman arvioinnin luotettavuus

Laajan kielimallin tuottamat numeeriset arvioinnit kahdelta eri ajankohdalta eivät olleet johdonmukaisia. Oppilaiden 34 koevastauksesta 13 sai eri numeerisen arvion ensimmäisellä ja toisella arviointikerralla. Tämä tarkoittaa, että noin 38 prosenttia arvioinneista muuttui. 11 arviointia muuttui yhdellä pisteellä ja viimeiset kaksi arviointia kahdella pisteellä. Tulos osoittaa, että laajan kielimallin tuottama arviointi ei ole luotettavaa, sillä se ei ole riittävän johdonmukaista.

Laajan kielimallin tuottamista sanallisista arvioinneista tarkasteltiin luotettavuutta viiden eri analyysiyksikön (a-e) avulla. Jokainen ensimmäisellä ja toisella arviointikerralla tuotettu sanallinen arviointi analysoitiin antamalla jokaiselle analyysiyksikölle arvo asteikolla 0–2 riippuen siitä, miten se toteutui sanallisessa arvioinnissa. Mikäli arvoa 2 esiintyy useasti, voidaan arviointia pitää luotettavana kyseisen analyysiyksikön osalta. Arvo 2 osoittaa, että malli suoriutui analyysiyksikön mittaamasta tehtävästä virheettömästi. Lisäksi ensimmäisen ja toisen kerran sanallisten arviointien saamia arvoja verrattiin keskenään Wilcoxon testin avulla. Oletus oli, että mahdollinen arvojen muutos johtuu laajan kielimallin tuottaman arvioinnin luotettavuuden vaihteluista. Laajan kielimallin pitäisi tuottaa samasta vastauksesta sama arviointi ajankohdasta riippumatta. Täten sen pitäisi olla yhtä hyvä tai huono molemmilla kerroilla.

Analyysiyksikkö a arvioi luotettavuutta siitä näkökulmasta, että arvioiko laaja kielimalli vain arviointimatriisin asioita. Analyysiyksikkö sai 52 kertaa arvon kaksi ja kuudesti arvon yksi. Lisäksi ensimmäisen ja toisen arviointikerran analyysiyksiköiden arvot olivat lähes aina samoja. Arvo oli 22 kertaa sama ja vain kolmesti eri. Tulokset osoittavat, että laaja kielimalli arvioi lähes aina vain arviointimatriisin asioita. Analyysiyksikkö a:n kohdalla luotettavuus oli pysyvää, sillä yksikkö sai molemmilla arviointikerroilla lähes aina samat arvot. Tämä on tutkimuksen yksi merkittävin tulos, sillä ainoastaan analyysiyksikkö a:n tulokset puoltavat laajan kielimallin kykyä arvioida luotettavasti. Wilcoxonin p-arvon 0,564 mukaan tulos ei ole tilastollisesti merkitsevä. Tulokset kuitenkin viittaavat luotettavuuden näkökulmasta merkittäviin tuloksiin.

Analyysiyksikkö b arvioi luotettavuutta tarkastelemalla laajan kielimallin kykyä arvioida koevastauksia koetehtävän ja koealueen pohjalta. Analyysiyksikkö sai vain 45 kertaa arvon 2 ja 15 kertaa arvon 1. Tämä kertoo, että 15 tapauksessa laaja kielimalli vaati oppilaalta osittain koealueen tai koetehtävän ulkopuolista tietoa. Tulos osoittaa, että laajan kielimallin tuottama arviointi ei ole luotettavaa, sillä yli kolmasosassa arvioinneissa vaadittiin kokeen ulkopuolista tietoa. Lisäksi analyysiyksikkö b:n kohdalla luotettavuus oli vaihtelevaa, sillä yksikkö sai vain 19 tapauksista saman arvon molemmilla kerroilla. Lopuista kuudesta tapauksesta kolme sai yhden arvon alemman ja kolme yhden arvon suuremman. Wilcoxon p-arvon 1,000 mukaan tulos ei ole

tilastollisesti merkitsevä. P-arvo osoittaa, että kertojen välillä ei ollut mitään systemaattista muutosta. Tulos johtuu siitä, että yhtä monessa tapauksessa arvo laski sekä nousi yhtä paljon.

Analyysiyksikkö c arvioi luotettavuutta siitä näkökulmasta, että arvioiko laaja kielimalli oikean oppilaan vastauksen. Seuraavien lukujen summa on suurempi kuin muiden analyysiyksiköiden kohdalla, sillä muiden analyysiyksiköiden analyysiin ei voinut ottaa mukaan niitä tapauksia, joissa analyysiyksikkö c sai arvon nolla. Täten lukuja ei voida suoraan verrata toisten analyysiyksiköiden tuloksiin. Analyysiyksikkö c sai 52 kertaa arvon kaksi ja kuusi kertaa arvon yksi. Merkittäväntä oli, että yksikkö sai 10 kertaa arvon nolla. Tämä tarkoittaa, että laaja kielimalli arvioi 10 kertaa toisen oppilaan vastauksen. Toisen oppilaan vastauksella tarkoitetaan, että malli arvioi jo aikaisemmin lähetettyä vastausta. Lisäksi kuudessa arvioinnissa oli viitteitä toisen oppilaan vastauksesta. Myös tämä tulos osoittaa, että laajan kielimallin tuottama arviointi ei ole luotettavaa, sillä lähes 20 prosenttia arvioinneista tuotettiin väärän oppilaan vastauksesta. Lisäksi luotettavuus oli vaihtelevaa, sillä analyysiyksikkö sai 20 kertaa saman arvon. Viidessä tapauksessa arvo vaihteli yhdellä. Wilcoxon testin p-arvo oli 0,655, mikä tarkoittaa tuloksen olevan tilastollisesti merkityksetön.

Analyysiyksikkö d arvioi luotettavuutta tarkastelemalla laajan kielimallin kykyä tunnistaa oppilaan osaaminen ja kykyä raportoida siitä oikein. Analyysiyksikkö d sai vain 46 kertaa arvon kaksi. Yksikkö sai 10 kertaa arvon yksi ja kahdesti arvon nolla. Tämä tarkoittaa, että laaja kielimalli ei tunnistanut kahdessa tapauksessa oppilaan tekemiä virheitä tai oikein vastauksia ja se raportoi niistä virheellisesti. Lisäksi 10 arviointia sisälsi osittaisia virheitä. Tulos osoittaa, että laajan kielimallin tuottama arviointi ei ole luotettavaa, sillä se ei aina tunnista oikein oppilaan osaamista. Tämä on erittäin tärkeä luotettavuuden osa-alue, sillä pisteet määräytyvät juuri osaamisen perusteella. Kaikista analyysiyksiyksiköistä analyysiyksikön d arvot vaihtelivat eniten. Vain 16 kertaa ensimmäisen ja toisen arviointikerran sanalliset arvioinnit saivat analyysiyksikön d arvoksi saman arvon. Arvo nousi viidessä tapauksessa ja neljässä puolestaan laski. Wilcoxon testin p-arvo oli 1,0 eli ei tilastollisesti merkitsevä. Tasan yhden p-arvo viittaa, että lähes yhtä monen tapauksen arvo laski ja nousi samassa suhteessa.

Viimeinen analyysiyksikkö e arvioi luotettavuutta siitä näkökulmasta, että ovatko laajan kielimallin tuottamat numeeriset arvioinnit linjassa arviointimatriisin kanssa. 50 kertaa laajan kielimallin tuottama arviointi oli linjassa arviointimatriisin kanssa. Kuitenkin kahdessa tapauksessa numeerinen arviointi oli väärä ja perustelut eivät olleet linjassa matriisin kanssa. Kuudessa arvioinnissa esiintyi osittaisia puutteita. Wilcoxon p-arvon 0,038 mukaan tulos on ainoa tilastollisesti merkitsevä. Tämä kuitenkin viittaa siihen, että arvot olivat aina systemaattisesti pienempiä. Analyysiyksikön arvot pysyivät 20 kertaa samana, mutta viisi kertaa pienenevät. Muiden analyysiyksiköiden kohdalla arvot nousivat ja laskivat samassa suhteessa, ja täten tilastollisesti merkitsevää eroa ei havaittu. Tulos osoittaa, että laaja kielimalli ei pystynyt kaikissa tapauksissa antamaan oikeaa arvosanaa, mikä viittaa arvioinnin epäluotettavuuteen.

Tutkimuksen tulokset osoittavat, ettei laaja kielimalli kykene arvioimaan alakoulun historian kokeita luotettavasti. Vaikka analyysiyksiköt saivat useasti arvon 2, oli arvoja 1 tai 0 liian usein. Arvoihin oli suhtauduttava kriittisesti, sillä arvioinnin pitäisi aina perustua ennalta määrättyyn arviointimatriisiin, arvioinnissa ei pitäisi vaatia opetetun aiheen ulkopuolista tietoa sekä tietenkin antaa arvio oikealle oppilaalle. Lisäksi arvioinnin pitäisi olla johdonmukaista. Arviointi oli erittäin epäjohdonmukaista, sillä lähes 40 prosenttia numeerisista arvioinneista muuttui arviointikertojen välillä. Lisäksi laaja kielimalli vaati useissa tapauksissa kokeen ulkopuolista tietoa, arvioi väärin oppilaan osaamista sekä hälyttävän usein arvioi väärän oppilaan vastauksen. Laaja kielimalli ei myöskään kyennyt aina antamaan arviointimatriisin mukaista numeerista arviointia oppilaalle. Analyysiyksikkö a osoitti, että laaja kielimalli kykenee johdonmukaisesti arvioimaan vain arviointimatriisissa määrättyjä asioita. Kuitenkin muiden analyysiyksiköiden tulokset eivät puolla laajan kielimallin tuottaman arvioinnin luotettavuutta. Tulokset osoittavat, että laajan kielimallin tuottama arviointi sisältää liikaa virheitä ja vaihtelua, jotta sitä voitaisiin sanoa luotettavaksi.

## 5.2 Kehotesuunnittelun aikana ilmenneet ongelmat ja arviointipromptin käytössä huomioitavat asiat

Kehotesuunnittelun aikana ilmeni ongelmia arviointimatriisin laatimisessa. Ongelmia tuotti myös aikaisempien tutkimuksien puutteellinen raportointi kehofotesuunnittelusta. Arviointimatriisin laatimisen ongelmat koskivat yleisiä arviointikriteereitä, niiden sanallistamista ja poikkeustilanteiden huomioimista ja pisteytysasteikon valintaa. Aikaisempien tutkimuksien puutteellinen raportointi kehofotesuunnittelusta ei tarjonnut lähtökohtia prosessin toteutukseen. Sen vuoksi kehofotesuunnittelu toteutettiin kokeilemalla eri vaihtoehtoja, mistä yrityksiä ja erehdyksiä kautta valittiin toimivimmat ratkaisut.

Ensimmäinen arviointimatriisiin liittyvä ongelma syntyi universaalien arviointimatriisin luomisesta. Tavoitteena oli luoda yksi matriisi, jota voisi käyttää kaikkien kolmen koetehtävän arvioinnissa. Koetehtävät kuitenkin erosivat toisistaan sisällöltään ja laajuudeltaan. Täten oli vaikeaa muodostaa kriteerit, jotka olisivat riittävän tarkkoja oppilaan osaamisen arvioimiseksi, mutta samalla tarpeeksi yleisiä soveltuakseen kaikkien kolmen tehtävän arviointiin.

Toinen arviointimatriisin laatimisen ongelma oli huomioida matriisissa koevastauksissa esiintyvät poikkeustilanteet. Oppilaiden koevastauksissa voi esiintyä poikkeustilanteita, kuten erilaisia ratkaisutapoja tai tehtävään vastaamatta jättämistä. Arviointimatriisiin pitää kirjata, kuinka laaja kielimalli pisteyttää vastauksen kyseisissä tilanteissa. Mikäli poikkeustilannetta ei huomioi matriisissa, voi laaja kielimalli pisteyttää vastauksen esimerkiksi kohtuuttomasti tai eri tavoin eri arviointikertoina. Tämä edellyttää oppilaiden vastauksiin perehtymistä jo kehofotesuunnittelun aika, jotta mahdolliset poikkeustilanteet voidaan huomata ja kirjata arviointimatriisiin. Koevastauksiin perehtyminen lisäsi työmäärää ja hidasti prosessia. Lisäksi kaikkia mahdollisia poikkeustilanteita ei voi kirjata arviointimatriisiin. Syihin palataan edempänä.

Kolmas arviointimatriisiin liittyvä ongelma syntyi arviointikriteerien sanallistamisesta siten, ettei laajalle kielimallille jää tulkinnanvaraa. Arviointipromptissa käytettyjen käsitteiden tulisi olla mahdollisimman yksiselitteisiä, jotta laaja kielimalli tulkitseisi ne samalla tavalla jokaisella arviointikerralla. Kuitenkin usea arvioinnissa käytettävä käsite,

kuten kattava, on erittäin tulkinnanvarainen ja vaatisi selityksen käsitteen merkityksestä.

Neljäs arviointimatriisin laatimisen ongelma oli sovittaa kaikki arviointikriteerit asteikolle 0–5. Viiden portaan asteikko rajoitti arvioinnin yksityiskohtaisuutta. Laajemman asteikon käyttö mahdollistaisi koevastauksien pienien erojen huomioinnin pisteytyksessä. Viiden portaan asteikon vuoksi, kaksi hieman eri tasoista vastausta voi saada samat pisteet, koska asteikko ei mahdollista pienien erojen huomioimista. Kuitenkin laajemman asteikon käyttö ei ollut mahdollista. Syihin palataan edempänä.

Toinen keחותesuunnittelun ongelma syntyi aikaisempien tutkimuksien puutteellisesta raportoinnista keחותesuunnittelusta ja arviointipromptista. Laajojen kielimallien käyttöä arvioijina on tutkittu kohtuullisen paljon, mutta niissä harvoin kerrotaan keחותesuunnittelusta tai esitellään käytettyä arviointipromptia. Tutkielman yhteydessä toteutettu suppea kuvaileva kirjallisuuskatsaus osoitti, että seitsemästä tarkastellusta tutkimuksesta vain kolme esitteli käyttämänsä arviointipromptin. Kyseiset tutkimukset eivät kuitenkaan kuvailleet keחותesuunnittelua. Aikaisempia tutkimuksia voitiin hyödyntää vain osittain, minkä seurauksena keחותesuunnittelu perustui pääosin eri vaihtoehtojen kokeiluun ja toimivimman ratkaisun etsimiseen. Yrityksien ja erehdyksien kautta toteutettu keחותesuunnittelu oli hidasta ja työlästä. Aikaisemmat tutkimukset olisivat tarjonneet lähtökohtia keחותesuunnittelun toteutukseen.

Osa arviointimatriisin laatimisen ongelmista johtui käytetyn kielimallin kontekstin pituudesta ja tietokoneen rajallisesta muistikapasiteetista ja laskentatehosta. Lisäksi arviointipromptin käytössä huomioitavat asiat, johtuivat enimmäkseen käytetyn mallin ja tietokoneen rajoitteista. Arviointimatriisin käsitteet voisi selittää tarkemmin, kaikki mahdolliset poikkeustilanteet voisi kirjata matriisiin ja arviointiasteikko voisi olla esimerkiksi 0–10, jos käytetyn laajan kielimallin ja tietokoneen tekniset ominaisuudet mahdollistaisivat sen.

Arviointipromptin käytössä huomioitavia asioita olivat koealueen laajuus, arvioitavien vastauksien määrä ja suomen kielen vaikutus. Suomen kielen vaikutusta lukuun ottamatta huomioitavat asiat johtuivat käytetyn laajan kielimallin ja tietokoneen prosessoinnin rajoituksista. Alkuperäinen koealue oli sellaisenaan liian pitkä mallille

käsiteltäväksi, minkä vuoksi koealue täytyi jakaa kolmeen pienempään osaan. Tämän vuoksi malli ei myöskään kyennyt arvioimaan oppilaan kaikkia kolmea koevastausta yhdellä kertaa. Silloin mallin pitäisi pystyä käsittelemään koko koealue, kolme eri koetehtävää ja oppilaan vastaukset niihin sekä arviointiprompti. Tämä vaatisi mallilta suurta kontekstin pituutta ja käytetyltä tietokoneelta suurta muistia ja laskentatehoa. Samasta syystä laaja kielimalli ei kyennyt arvioimaan kuin kaksi koevastausta yhdellä kertaa. Kuitenkin jos vastaus on pitkä, noin 20 virkettä, täytyi se arvioida yksin. Edellä mainitut asiat ovat arviointipromptin käytössä eli arvioinnin tuottamisen aikana huomioitavia asioita.

Lisäksi arviointipromptin käytön aikana ilmeni suomen kielen vaikutus laajan kielimallin tuottamaan arviointiin. Laajan kielimallin tuottamien arviointien ja ajatusprosessin perusteella ilmeni, että mallilla on vaikea tulkita suomen kieltä. Virallista väitettä ei voida esittää, sillä väite perustuu muutamaan havaintoon. Kuitenkin tässä tutkimuksessa kaikki aineistot eli oppilaiden koevastaukset, koealue ja arviointiprompti käännettiin englanniksi. Aineistojen kääntäminen englanniksi tuotti yhden lisävaiheen kehotesuunnitteluun.

Tutkimuksen tulokset osoittavat, että arviointipromptin luominen sisältää monia ongelmia, ja lisäksi arviointimatriisin käytössä on huomioitava moni asia. Ongelmia tuotti erityisesti universaalien arviointimatriisin laatiminen sisältäen käsitteiden yksiselitteisen määrittelyn ja poikkeustilanteiden huomioimisen. Lisäksi aikaisempien tutkimuksien puutteellinen raportointi kehotesuunnittelusta vaikeutti arviointipromptin luomista, sillä aikaisemmista tutkimuksista ei saanut lähtökohtia prosessin toteutukseen. Arviointipromptin käytössä eli arviointien tuottamisessa on huomioitava laajan kielimallin kontekstin pituus ja käytetyn tietokoneen muisti ja laskentateho. Nämä tekijät määrittävät, kuinka pitkän koealueen mallille voi antaa ja montako koevastausta malli pystyy arvioimaan yhdellä kertaa. Tutkimuksen tulokset tarjoavat lähtökohtia seuraaville tutkijoille ja mahdollisesti opettajille, jotka pyrkivät tuottamaan arviointipromptin koevastauksien arviointiin. Tutkimuksen tulokset kertovat, mitä ongelmia kehotesuunnittelun aikana voi ilmentyä ja mitä asioita on huomioitava, kun arviointia tuottaa laajalla kielimallilla.

## 6 Pohdinta

Tutkimuksen tavoitteena oli selvittää, laajan kielimallin tuottaman arvioinnin luotettavuutta. Lisäksi tutkimuksessa selvitettiin, millaisia ongelmia kehotesuunnittelun aikana ilmenee, kun luodaan arviointipromptti sekä mitä asioita on huomioitava arviointipromptin käytössä. Tutkimuksen kohteena oli OpenAI:n paikallinen gpt-oss-120b laaja kielimalli. Kehotesuunnittelun aikana ilmeneviä ongelmia ja arviointipromptin käytössä huomioitavia asioita selvitettiin toteuttamalla kehotesuunnittelu käytännössä. Laajan kielimallin tuottaman arvioinnin luotettavuutta tutkittiin vertailemalla mallin tuottamia numeerisia arviointeja kahdelta eri ajankohdalta. Lisäksi luotettavuuden käsite operationalisoitiin viiteen analyysiyksikköön, ja yksikölle annettiin arvo asteikko 0–2 riippuen siitä, miten se toteutui laajan kielimallin tuottamassa sanallisessa arvioinnissa. Laajan kielimallin suoriutuminen analyysiyksiköiden mittaamista tehtävistä määritti arvioinnin luotettavuutta. Laaja kielimalli arvioi 12 oppilaan historian kokeet. Oppilaat olivat alakoulun viidennellä luokalla. Koe koostui kolmesta tehtävästä, joten yhteensä koevastauksia kertyi 34. Kaksi koevastausta ei soveltunut tutkimuksen aineistoksi, koska ne olivat ajatuskarttoja. Laaja kielimalli arvioi koevastaukset kahdesti, joten arviointeja kertyi yhteensä 68.

Tutkimuksen tulokset osoittavat, että laajan kielimallin tuottama arviointi ei ole luotettavaa. Laaja kielimalli ei suoriutunut analyysiyksiköiden mittaamista tehtävistä riittävän johdonmukaisesti ja virheettömästi. Kuitenkin analyysiyksiköt saivat enimmäkseen arvoja kaksi, mutta arvoja yksi ja nolla esiintyi siinä määrin, ettei luotettavuuden voida katsoa toteutuneen. Ensimmäiseksi laaja kielimalli ei aina kyennyt arvioimaan koealueen ja -tehtävän pohjalta. Malli esimerkiksi vaati oppilaalta koetehtävän ja -alueen ulkopuolista tietoa. Toiseksi laaja kielimalli ei aina tunnistanut oppilaan osaamista oikein eli esimerkiksi hyväksyi vastauksessa esiintyvän virheen. Kolmanneksi laajan kielimallin tuottama numeerinen arviointi oli välillä väärin suhteessa oppilaan osaamiseen. Neljänneksi laaja kielimalli arvioi useasti väärän oppilaan vastauksen. Malli arvioi 68 tapauksesta 10 kertaa jo aikaisemmin lähetetyn oppilaan vastauksen. Lisäksi laajan kielimallin tuottamat numeeriset arvioinnit eivät

olleet johdonmukaisia ensimmäisellä ja toisella arviointikerralla. 34 numeerisesta arvioinnista 13 muuttui arviointikertojen aikana. Tämä tarkoittaa, että lähes 40 prosenttia arvioinneista muuttui.

Merkittävää kuitenkin on, että laaja kielimalli suoriutui yhden analyysiyksikön mittaamasta tehtävästä onnistuneesti. Malli kykeni arvioimaan aina vain arviointimatriisin asioita. Esimerkiksi malli ei arvioinut matriisin ulkopuolisia asioita, kuten kielioppia. Mallin tuottamassa arvioinnissa esiintyi johdonmukaisesti arvioinnin kohteena olleet asiat. Kuitenkin laajan kielimallin suoriutuminen onnistuneesti vain yhden analyysiyksikön mittaamasta tehtävästä ei riitä osoittamaan, että mallin tuottama arviointi olisi luotettavaa.

Useat aikaisemmat tutkimukset puoltavat tämän tutkimuksen tuloksia, että laajat kielimallit eivät kykene arvioimaan luotettavasti. Aikaisemmat tutkimukset ovat tarkastelleet luotettavuutta, vertaamalla keskenään laajan kielimallin ja ihmisarvioijien tuottamia arviointeja oppilaiden esseistä. Esimerkiksi Almegren ym. (2024) vertailivat ihmisarvioijien antamia arviointeja kolmen eri laajan kielimallin tuottamiin arviointeihin, eikä yksikään malleista arvioinut esseitä ihmisen tasoisesti. Myös Barrotin ja Buin (2024) tutkimustulokset tukevat tulosta, että laaja kielimalli ei kykene tuottamaan luotettavaa arviointia, sillä arviointi eroaa suuresti ihmisen antamasta arvioinnista. Aikaisemmat tutkimukset ovat myös tarkastelleet eri kielisten esseiden vaikutusta laajan kielimallin tuottamaan arviointiin. Fokides ja Peristeraki (2025) tarkastelivat kreikan kielisten ja Kincl ym. (2024) puolestaan tšekin kielisten esseiden arviointia. Molempien tutkimuksien tulokset osoittivat, että laajat kielimallit eivät kyenneet arvioimaan opettajan tasoisesti. Lisäksi Kincl ym. (2024) havaitsivat, että mallin tuottama arviointi oli epäjohdonmukaista kahtena eri ajankohtana. Tämä havainto tukee tämän tutkimuksen tulosta, jossa lähes 40 prosenttia koevastauksien arvioinneista muuttui arviointikertojen välillä.

Kuitenkin osa aikaisemmista tutkimuksista on saanut laajan kielimallin tuottamaan luotettavaa arviointia. Esimerkiksi Magro ym. (2025) havaitsivat tutkimuksessaan, että laajan kielimallin tuottamat arvioinnit olivat johdonmukaisempia kuin ihmisen tekemät arvioinnit. Lisäksi Sadanand ym. (2022), Kincl ym. (2024) sekä Magro ym. (2025)

tutkimuksen tulokset osoittavat, että laajat kielimallit tuottivat englannin kielisistä teksteistä yhteneviä tai lähes yhteneviä arviointeja kuin ihmiset. Täten voidaan sanoa, että laajan kielimallin toiminta arvioijana on hyvin vaihtelevaa.

Tutkimuksen tulokset osoittavat, että kehoitesuunnittelun aikana ilmenee useita ongelmia, kun luodaan arviointiprompti. Ongelmat liittyivät arviointimatriisin laatimiseen ja aikaisempien tutkimuksien puutteelliseen raportointiin kehoitesuunnittelusta. Arviointimatriisin laatimisessa oli ongelmallista muodostaa universaalit kriteerit, sanallistaa käsitteet yksiselitteisesti ja huomioida matriisissa vastauksissa ilmenneet poikkeustilanteet. Arviointimatriisin piti soveltua kolmen koetehtävän vastauksien arviointiin. Koetehtävät olivat sisällöltään ja laajuudeltaan erilaisia, mikä vaikeutti yleisien kriteerien luomista. Toiseksi monet käsitteet ovat tulkinnanvaraisia. Esimerkiksi koevastauksien arvioinnissa esiintyvä käsite kattava, antaisi laajalle kielimallille vapauden tulkita käsitteen merkityksen. Malli voi tulkita käsitteen jokaisella arviointikerralla eri tavalla, mikä heikentää arvioinnin luotettavuutta. Kolmanneksi matriisissa pitää olla arviointiohjeet kaikkiin vastauksissa esiintyville poikkeustilanteille. Mikäli matriisissa ei ole ohjetta tilanteen vaikutuksesta pisteisiin, voi malli arvioida ne eri tavalla eri arviointikertoina.

Kehotesuunnittelussa ja arviointipromptin käytössä on huomioitava laajan kielimallin kontekstin pituus sekä käytetyn tietokoneen muisti ja laskentateho. Laajan kielimallin ja tietokoneen rajoitteiden vuoksi täytyy arvioinnin tuottamisessa huomioida koealueen pituus ja montako koevastausta malli pystyy arvioimaan yhdellä kertaa. Tässä tutkimuksessa käytetyn laajan kielimallin kontekstin pituus ja tietokoneen laskentateho ja muisti eivät riittäneet koko koealueen prosessointiin eikä useamman kuin kahden koevastauksen arviointiin yhdellä kertaa. Lisäksi arviointipromptia käytettäessä on huomioitava suomen kielen vaikutus. Kehotesuunnittelun aikana tehdyt havainnot osoittivat, että laaja kielimalli ei kyennyt luotettavasti tulkitsemaan suomen kieltä. Kuitenkin virallista väitettä ei voida esittää, sillä ne olivat vain havaintoja.

Toinen kehoitesuunnittelun ongelma liittyi aikaisempien tutkimuksien puutteelliseen raportointiin kehoitesuunnittelusta. Tutkielmaa varten toteutettu suppea kuvaileva kirjallisuuskatsaus osoitti, että seitsemästä tarkastellusta tutkimuksesta vain kolme

esitti käyttämänsä arviointipromptin. Kuitenkaan mikään tutkimus ei raportoinut kehotesuunnittelusta. Tämän vuoksi kehotesuunnittelu tapahtui kokeilujen ja erehdyksien myötä, sillä aikaisemmista tutkimuksista ei saanut lähtökohtia prosessin toteutukseen. Aikaisempien tutkimuksien puutteellinen raportointi on ongelmallista myös siksi, että käytetty prompti vaikuttaa merkittävästi tutkimuksen tuloksiin. Laajan kielimallin tuottamat arvioinnit voivat erota radikaalisesti käytetyn promptin vuoksi. Lisäksi tutkimuksen tuloksia on vaikea arvioida tai toistaa luotettavasti, jos promptia ei ole kerrottu. Tutkimuksien pitäisi raportoida selkeästi kehotesuunnittelun vaiheista ja käytetystä arviointipromptista, jotta tutkimukset tuottaisivat yleiskäsitystä siitä, millaiset arviointipromptit soveltuvat arvioinnin tuottamiseen. Nyt jokainen tutkija joutuu kehittämään oman arviointipromptin, mikä hidastaa aiheen tutkimusta.

Tuloksia kehotesuunnittelun ongelmista ei voida verrata aikaisempaan tutkimukseen, sillä aikaisemmissa tutkimuksissa ei ole kuvailtu kehotesuunnittelun vaiheita. Tutkimuksessa käytettyä arviointipromptia voidaan kuitenkin verrata aikaisempien tutkimuksien arviointiprompteihin. Arviointipromptit eivät ole täysin vertailukelpoisia, sillä ne ovat vahvasti sidoksissa kontekstiin. Fokidesin ja Peristerakin (2025), Wetzlerin ym. (2025) sekä tämän tutkimuksen arviointiprompteja yhdisti se, että niissä määriteltiin laajan kielimallin rooli ja tehtävä sekä annettiin tarkempia ohjeita arvioinnin tuottamiseen. Lisäksi edellä mainittujen sekä Barrotin ja Buin (2024) tutkimuksien prompteja yhdistää se, että ne sisälsivät arviointimatriisin ja oppilaan vastauksen.

### **6.1 Kolme mahdollista keinoa lisätä laajan kielimallin tuottaman arvioinnin luotettavuutta**

Tämän tutkimuksen kehotesuunnittelun yhteydessä kokeiltiin kolmea eri mahdollisuutta lisätä laajan kielimallin tuottaman arvioinnin luotettavuutta. Ne olivat vain kokeiluja, joten virallisia johtopäätöksiä ei voida tehdä. Kokeilut liittyivät optimointiin, aineiston kieleen ja laajan kielimallin väittelyyn itsensä kanssa.

Yksi mahdollinen keino vaikuttaa laajan kielimallin tuottaman arvioinnin luotettavuuteen on optimointi. Optimoinnilla tarkoitetaan toimenpiteiden kokonaisuutta, jolla parannetaan tekoälyn suorituskykyä, tehokkuutta ja skaalautuvuutta. Tavoitteena on saada tekoäly suoriutumaan tehtävästään

mahdollisimman hyvin, myös suuressa käytössä sekä mahdollisimman pienillä resursseilla. (Jainschigg, 2026.) Kontekstin pituus on yksi keskeisin asia, mikä vaikuttaa tekoälyn toimintaan. Kontekstin pituus määrittää, montako tokenia tekoäly pystyy kerrallaan ottamaan huomioon (Bergmann, ei pvm. b). Tokenilla puolestaan tarkoitetaan merkkejä, joihin prompti ja annetut aineistot muunnetaan tekoälyn prosessointia varten. Sanojen ja tokenien yhteys ei ole yksiselitteinen, sillä yksi sana voi olla yksi tokeni, mutta joskus sama sana voidaan hajottaa useammaksi tokeniksi. Keskimäärin 10 sanan mittainen lause vie noin 15–20 tokenia. (IBM, 2026.) Suurempi kontekstin pituus mahdollistaa laajempien promptien ja aineistojen prosessoinnin sekä laajempien tuotoksien tuottamisen. Liian lyhyt kontekstin pituus, johtaa tekoälyn muistinmenetyksiin tai mahdollisesti sen kokonaan jumiutumiseen. (Bergmann, ei pvm. b.) Riittävät tokenit ja kontekstin pituus mahdollistavat, että laaja kielimalli pystyy ottamaan arvioinnissa kaiken tarpeellisen, kuten koko arviointimatriisin ja koealueen, huomioon.

Tokenit ja kontekstin pituudet vaikuttavat keskeisesti tekoälyn toimintaan, ja lisäksi ne määrittävät usein myös käyttökustannukset. Tekoäly-yritykset, kuten OpenAI, määrittelevät hintansa tokenien ja kontekstin pituuden mukaan (Crosley, 2026). Tämän vuoksi on tärkeää pyrkiä luomaan asetelmat, jossa tekoälyä pystytään hyödyntämään mahdollisimman tehokkaasti esimerkiksi luomalla tiivis prompti ja mahdolliset aineistot. Tarkoituksena tehdä tekoälyn toiminnasta mahdollisimman tehokasta, tarkoituksenmukaista ja kustannustehokasta.

Tämän tutkimuksen osalta laajan kielimallin suorituskykyä ja tehokkuutta pyrittiin optimoimaan promptin muokkauksella ja tiivistämisellä sekä koealueen palastelulla. Tämän tutkimuksen yksi keskeinen vaihe olikin kehtesuunnittelu, jossa pyrittiin luomaan mahdollisimman hyvä prompti, jonka pohjalta laaja kielimalli voi tuottaa arvioinnit oppilaan koevastauksista. Lisäksi tiiviimpi prompti ja palasteltu koealue olivat välttämättömiä, sillä OpenAI:n gpt-oss-120b mallin kontekstin pituus ei riittänyt prosessoimaan alkuperäisien aineistojen pituuksia. Koealueen pituus laski 12 sivusta noin kolmeen sivuun ja prompti lyheni noin puolella, kun arviointiskaala vaihdettiin 10 pisteestä viiteen pisteeseen. Skaalautuvuuden mahdollisuuksiin ei perehdytty, sillä se ei ole opettajan näkökulmasta välttämätöntä. Oletettavasti luokanopettaja tulisi

arvioimaan noin 25 koevastausta kerrallaan, mitä ei lasketa suureksi käytöksi. Mikäli tekoälyä alettaisiin hyödyntämään esimerkiksi ylioppilaskirjoitusten tarkistamisessa, pitäisi perehtyä myös mallin skaalautuvuuteen.

Toinen mahdollinen keino vaikuttaa laajan kielimallin tuottaman arvioinnin luotettavuuteen, on laittaa tekoäly väittelemään itsensä kanssa. Yliopistonlehtori Lauri Kemppinen avusti kokeilun toteutuksessa. Kokeilu toteutettiin osana tätä tutkielmaa. Kemppinen loi asetelman, jossa laajalle kielimallille määrättiin ns. kolme persoonaa, ja ne tuottivat yhdessä oppilaalle arvosanan. Kokeilu toteutettiin samalla laajalla kielimallilla eli gpt-oss-120b:llä ja siinä käytettiin samaa arviointimatriisia, koealuetta sekä samojen oppilaiden koevastauksia kuin tässä tutkielmassa. Asetelmaa varten täytyi kuitenkin luoda uusi prompti, joka soveltui paremmin kyseiseen asetelmaan. Kokeilussa laaja kielimalli arvioi oppilaan kaikki kolme koevastausta yhdellä kertaa. Asetelmaa kokeiltiin kahden eri oppilaan koevastauksilla. Asetelman lähtökohdat olivat, että opettaja A ja opettaja B tuottavat arviointimatriisin ja annetun koealueen pohjalta oppilaalle alustavan arvosanan. Rehtori tekee lopullisen päätöksen arvosanasta vain opettajan A ja opettajan B näkemysten pohjalta. Seuraavaksi on katkelma opettaja B:n ajatusprosessista, kun se pohtii promptin ja arviointimatriisin välistä ristiriitaa sekä luo itselleen perustelut vähäiselle ja merkittävälle virheelle.

That may reduce reasoning quality. The rubric indicates 5/5 for fully correct and well-reasoned with at least one example. 4/5 if minor detail missing. So we must decide whether missing details are major or minor relative to the exam material. Definition of "minor detail" could be something like forgetting a small point but not essential to overall understanding. But Q2 missing entire governance system for Athens is quite significant.

So if a claim is unsupported, we cannot treat it as wrong? But the student's answer includes claims outside of material. The rubric states that we must evaluate strictly on basis of retrieved material; but do we penalize for adding extraneous information? The rubric doesn't mention penalizing for extra knowledge. But rubric says "answer is well-reasoned from perspective of information provided." So if they use info not in material, does that detract from reasoning? Possibly yes: If they rely on unsupported claims, the reasoning might be flawed relative to given material. But we can't penalize for lacking knowledge beyond; but we can penalize for missing required details

Rehtorin tehtävänä oli nostaa esiin opettajien erimielisyydet ja ratkaista ne järkevällä tavalla sekä luoda oppilaalle lopullinen arviointi opettajien A ja B näkemysten pohjalta.

Rehtori ei näe koeaineistoa. Opettajat olivat samaa mieltä oppilaan yksi arvosanasta, mutta eivät oppilaan kaksi. Tämän vuoksi oli mielenkiintoista perehtyä rehtorin ajatusprosessiin oppilaan kaksi kohdalla. Yllättävää oli, että rehtori teki nopeasti lopullisen päätöksen oppilaan arvosanasta. Seuraavaksi on katkelma rehtorin ajatuksenjuoksusta, kun se luo lopullisen arvosanan.

Teacher A viewed Question 2 as fundamentally wrong because it misrepresented Athens' political system, thus awarding no credit for that question. Teacher B recognized some correct elements (e.g., everyday life differences, mention of democracy and kingship) and therefore granted partial credit despite the errors. Balancing both perspectives, we acknowledge that while the answer contains valuable observations, the major factual mistakes undermine its reliability. A moderate rating—recognizing partial correctness but penalising significant inaccuracies—is the most equitable conclusion. 5 Final suggested grade / performance level 3 / 5 (moderate). This score reflects correct handling of Question 1, partial accuracy in Question 2 with key errors, and a largely unsupported answer to Question 3.

Kempin mukaan kokeilun yksi keskeisin ongelma oli sisällyttää kaikki laajan kielimallin tehtävät yhdeksi komennoksi. Tehtäviin sisältyi esimerkiksi koekysymyksen kannalta relevantin tiedon etsiminen koealueesta sekä arvioinnin tuottaminen opettaja A:na, opettaja B:nä ja rehtorina. Lisäksi ongelmia tuotti saada varmuus siitä, että laaja kielimalli käyttää koko koealuetta arvioinnin tuottamiseen. Laaja kielimalli ei pystynyt arvioimaan koetehtävän kolme vastauksia, sillä vastaukset tehtävään löytyivät koealueen viimeiseltä sivulta. Mallin kontekstin pituus tai käytetyn koneen muistikapasiteetti ja laskentateho eivät riittäneet suuren koealueen prosessointiin (Bergmann, ei pvm. b) promptin ja koevastauksen lisäksi. Kontekstin pituus, laitteen muistikapasiteetti ja laskentateho riittivät koetehtävien 1 ja 2 tietojen noutamiseen koealueesta ja arvioinnin tuottamiseen. Mikäli kokeilun toteuttaisi tulevaisuudessa uudestaan, on parempi jakaa laajan kielimallin tehtävät useammaksi komennoksi. Täten laajan kielimallin ensimmäinen tehtävä olisi palastella koealue tarpeeksi pieniksi osiin, jotta kontekstin pituus olisi optimaalinen arviointien tuottamiseen. Yllätyimme laajan kielimallin kyvystä ja asetelman toimivuudesta. Laaja kielimalli tuotti rehtorin roolissa erittäin laadukkaan arvioinnin ja perustelut siihen. Kokeilu osoittaa, että asetelman avulla olisi mahdollista arvioida oppilaiden koevastauksia luotettavasti.

Viimeinen mahdollinen keino lisätä laajan kielimallin tuottaman arvioinnin luotettavuutta on vaihtaa aineistojen kieltä. Tässä tutkimuksessa kokeiltiin promptin, oppilaiden vastauksien ja koealueen kääntämistä suomen kielestä englannin kielelle. Kehotesuunnittelun aikana heräsi epäily gpt-oss-120b mallin kyvystä tulkita suomen kieltä, kun perehtyi sen ajatusprosessiin. Malli ei nimittäin aina löytänyt englanninkielisiä vastineita suomeksi kirjoitetuille sanoille. Lisäksi aikaisemmat tutkimukset ovat osoittaneet, että laajat kielimallit suoriutuvat paremmin englanninkielisen aineiston arvioinnista (esim. Fokides & Peristeraki, 2025; Kincl ym., 2024). Tarkemmin kielen vaihtamisesta kerrottiin luvussa 4.3.

Varovasti voidaan väittää, että optimoinnilla ja englannin kielellä voi olla vaikutusta laajan kielimallin tuottaman arvioinnin luotettavuuteen. Mitä pätevämpi prompti, sen paremmin malli pystyi tuottamaan arviointeja. Lisäksi lyhyempi prompti ja koealue mahdollistivat laajan kielimallin luotettavamman toiminnan, sillä malli pystyi ottamaan kaikki tarvittavat tiedot huomioon ilman muistinmenetyksiä. Aikaisempien tutkimuksien mukaan (esim. Fokides & Peristeraki, 2025; Kincl ym., 2024) ja laajan kielimallin ajatusprosessiin perehtyminen osoittaa, että laaja kielimalli toimii paremmin englannin kielellä. Laajalla kielimallilla oli haasteita löytää englannin kielisiä vastineita suomenkielisestä aineistosta. Mallin toiminta ei olisi luotettavaa, jos se ei kykene tunnistamaan tehtävän kannalta merkittäviä sanoja. Kielen vaihtaminen lisäsi myös suorituskykyä ja tehokkuutta.

Lisäksi laajan kielimallin arvioinnin luotettavuutta voidaan mahdollisesti lisätä väittelyasetelmalla. Yliopistonlehtori Lauri Kempin tuottamassa asetelmassa oppilaan vastaus arvioidaan ns. kolmesti, mikä voi lisätä luotettavuutta. Kuitenkin asetelman toteuttaminen onnistuneesti edellyttää käytetyltä laitteelta suurta laskentatehoa. Lisäksi asetelman kehittäminen ja toteuttaminen edellyttää asiantuntemusta laajojen kielimallien toiminnasta ja promptien suunnittelusta. Kokeilu kuitenkin osoittaa, että väittelyasetelma on mahdollinen ja sen tuottamat tulokset ovat positiivisia. Tärkeää olisi toteuttaa asetelma useampaan kertaan ja analysoida jokaisen roolin tuottamat arvoinnit, jotta voitaisiin tehdä tarkempia johtopäätöksiä asetelman tuottaman arvioinnin luotettavuudesta.

## 6.2 Mahdollisia haasteita laajan kielimallin käytöstä arvioijana koulussa

Laajan kielimallin käyttöön arvioijana liittyy useita muitakin haasteita luotettavuuden lisäksi. Seuraavaksi esitetään muutama kysymys, joita syntyisi laajan kielimallin käytöstä arvioijana kouluissa. Kysymykset liittyvät kehoitesuunnitteluun, ajan käyttöön, kustannuksiin sekä opettajien, oppilaiden ja heidän huoltajiensa suhtautumiseen asiaa kohtaan.

Laajan kielimallin hyödyntämistä arvioinnissa voidaan markkinoida sen nopeudella. Barrot ja Bui (2024) kertovat tutkimuksessaan, että ChatGPT:llä kesti keskimäärin 10 sekuntia tarkistaa yksi essee. Puolestaan Magro ym. (2025) kertovat, että heidän suunnittelemlaan laajalla kielimallilla nimeltä GradeGPT kesti noin 1–2 minuuttia arvioida yhden oppilaan tehtävä. Tässä tutkimuksessa käytetyllä gpt-oss-120b mallilla kesti arvioida kahden oppilaan vastaukset keskimäärin puolitoista minuuttia. Laajat kielimallit tuottavat arvioinnin siis huomattavasti nopeammin kuin ihminen. Kuitenkin prosessit, mitä tehdään ennen oppilaan vastauksen syöttämistä laajalle kielimallille, ovat erittäin aikaa vieviä. Tämä tutkimus osoitti, että itse kehoitesuunnittelu on pitkä ja hidas prosessi. Hidasta on myös mahdollisen koealueen ja oppilaiden tehtävien digitalisointi. Oletus on, että laajat kielimallit eivät kykenisi lukemaan oppilaan käsialaa, joten pelkkä vastauksen skannaus ei riitä. Tämä tarkoittaa, että opettaja joutuisi tietokoneellaan kirjoittamaan oppilaiden vastaukset digitalisoituun muotoon. Opettaja ehtisi itse arvioimaan vastaukset siinä ajassa, kun hän ensiksi digitalisoisi oppilaiden vastaukset ja sitten syöttäisi ne laajalle kielimallille arvioitavaksi.

Opettajan työaika eikä mahdollisesti taidotkaan riittäisi kehoitesuunnittelun toteutukseen. Kehotesuunnittelun ammattilainen voisi luoda tarvittavat arviointimatriisit ja promptit. Asia ei kuitenkaan ole näin yksinkertainen. Keskeisiksi kysymyksiksi nousee, pitäisikö jokaiselle tehtävälle luoda uusi prompti ja arviointimatriisi vai pystyisikö luomaan moneen tehtävään sovellettavan promptin ja matriisin. Kuitenkin tässäkin tutkimuksessa nousi esiin, että universaalien matriisin luominen oli haastavaa jo kolmelle samaa aihepiiriä käsittelevälle tehtävälle. Ratkaisuna ei voisi olla yhtenäiset tehtävät tai kokeet, sillä se rajoittaisi suomalaisen opettajan autonomiaa. Tämä voisi kuitenkin avata oppikirjakustantajille uuden

yrittöimintamahdollisuuden. Kustantajat ovat jo pitkään tuottaneet valmiita kokeita opettajien käyttöön, joten kustantajat voisivat tarjota myös valmiin arviointimatriisin ja promptin sen kokeen tarkistamiseen tekoälyllä.

Paikallinen laaja kielimalli ei myöskään ole edullinen ratkaisu. Tässä tutkimuksessa käytetyn koneen hinnan arvioitiin olevan noin 5000 euroa. Koneita tulisi mahdollisesti olla useita, sillä paikallista laajaa kielimallia voi käyttää vain yksi opettaja kerrallaan. Kustannuksia lisäksi myös kehoitesuunnittelun ammattilaisen hyödyntäminen, jos opettajat eivät itse pysty luomaan arviointipromptia. Lisäksi tilanne edellyttäisi opettajien lisäkoulutusta, mikä myös kasvattaisi kustannuksia.

Täytyisi myös huomioida oppilaiden ja heidän huoltajiensa suhtautuminen laajojen kielimallin käyttöön arvioinnissa. Oppilaiden ja heidän huoltajiensa suostumus olisi välttämätöntä. Tämä herättää kysymyksen siitä, miten toimia tilanteissa, jossa oppilas itse tai hänen huoltajansa ei hyväksy laajan kielimallin käyttöä. Myös opettajan suhtautuminen täytyy huomioida, sillä Suomessa opettajilla on pedagoginen vapaus toteuttaa opetustaan valitsemallaan tavalla. Täten laajan kielimallin hyödyntämiseen arvioijana liittyy luotettavuushaasteiden lisäksi useita muita pedagogisia, teknisiä ja käytännöllisiä kysymyksiä.

### **6.3 Alustavia havaintoja ChatGPT 5.2 ja Gemini 3 Flash mallien toiminnasta arvioijina**

Tutkimuksen pääkohteeksi valittiin paikallinen gpt-oss-120b laaja kielimalli, mutta myös pilvipohjaiset ChatGPT 5.2 ja Gemini 3 Flash mallit tuottivat arvioinnit oppilaiden koevastauksista. Pilvipohjaisilla malleilla tuotettiin arvioinnit jatkotutkimusta varten, mutta koetaan mielekkäänä esitellä jo alustavia havaintoja mallien toiminnasta.

ChatGPT 5.2 ja Gemini 3 Flash pilvipohjaisen mallit toimivat paremmin kuin paikallinen gpt-oss-120b malli. Pilvipohjaiset mallit tuottivat arvioinnit nopeasti. Kolmen arvioinnin tuottamiseen meni noin 10 sekuntia. Gpt-oss-120b mallilla kesti arvioida kaksi vastausta jopa 2 minuuttia. Paikallisella mallilla arviointiin kuluva aika vaihteli, kun taas pilvipohjaisilla malleilla se oli johdonmukaisempaa. ChatGPT 5.2 ja Gemini 3 Flash mallien toiminnassa ei myöskään esiintynyt samoja haasteita kuin gpt-oss-120b mallin

toiminnassa. Pilvipohjaiset mallit eivät tuottaneet arviointeja väärin oppilaiden koevastauksista. Paikallinen malli arvioi useasti jo aikaisemmin lähetyn koevastauksen. Lisäksi pilvipohjaiset mallit eivät koskaan jumiutuneet tuottaessaan arviointia, mikä puolestaan tapahtui useasti paikalliselle mallille. Kuitenkaan johtopäätöksiä mallien luotettavuudesta ei voida esittää, sillä kyseisten mallien tuottamia arviointeja ei analysoitu.

Silmämääräisesti Gemini 3 Flash noudatti annettua arviointimatriisia ja arvioi oikean oppilaan vastauksen. Kuitenkin arvioinneista tehtiin vain silmämääräisiä havaintoja, eikä systemaattista analyysia. Gemini 3 Flash kirjasi jokaiseen sanalliseen arviointiin, että onko oppilas vastannut kaikkiin koetehtävässä esitettyihin kysymyksiin, ovatko vastaukset perusteltuja koealueen näkökulmasta ja mitä virheitä tai puutteita vastauksessa esiintyy. Gemini 3 Flash sisällytti arviointiin katkelmia oppilaan vastauksesta. Seuraavaksi on esitelty Gemini 3 Flash laajan kielimallin tuottama arviointi.

Score: 3/5

Summary of reasoning:

The student explains how gods were connected to daily life (explaining natural phenomena and events, influence on good and bad fortune) and describes the types of gods (many gods with different roles, Zeus as supreme god). An example is included.

However, the answer does not clearly describe how people acknowledged the gods (e.g., sacrifices, offerings, oracles), which is required by the task. Some ideas (such as gods causing illnesses or being heroes with weaknesses) are not directly supported by the exam material. The connection to the modern world is addressed, but overall there are gaps in required content.

ChatGPT 5.2 laajan kielimallin tuottama arviointi on hyvin saman suuntaista Gemini 3 Flash mallin tuottaman arvioinnin kanssa. ChatGPT 5.2 kirjasi arviointiin että, mihin koekysymyksiin oppilas on vastannut, onko vastaus perusteltu koealueen näkökulmasta ja mitä asioita oppilas ei maininnut. Lisäksi malli sisällytti arviointiin katkelmia oppilaan vastauksesta. Seuraavaksi on esitelty ChatGPT 5.2 laajan kielimallin tuottama arviointi saman oppilaan vastauksesta.

Grade: 3/5

**Completeness:** The student attempted to answer all questions presented in the exam task, including the connection to the modern world.

**Reasoning from Material:** The student correctly linked the gods to natural phenomena like the weather and to human events like wars. They also correctly identified Zeus as the supreme god and noted that there was a specific god for almost everything.

**Gaps in Content:** There are significant gaps regarding how the Greeks acknowledged the gods. The student used the vague term "worshipped" but failed to mention any of the specific rituals described in the text, such as animal sacrifices, pouring liquids (wine, milk, or oil) before meals, or consulting oracles.

**Inaccuracies:** The student included information not found in the material, such as the gods being the cause of "illnesses" or "seasons" and the gods having "weaknesses". The text focuses on natural phenomena like lightning and the sun's movement.

**Modern World:** The student provided examples of modern connections (movies and books), which were not present in the provided text.

Gemini 3 Flash ja ChatGPT 5.2 mallien tuottamien arviointien silmämääräinen tarkastelu osoittaa, että malleilla voi olla potentiaalia tuottaa luotettavia arviointeja. Ennen kun asiasta voi tehdä virallisia johtopäätöksiä, on perehdyttävä tarkemmin mallien tuottamiin arviointeihin. Aineiston analyysin avulla voitaisiin tehdä johtopäätöksiä Gemini 3 Flash ja ChatGPT 5.2 mallien tuottaman arvioinnin luotettavuudesta.

#### **6.4 Jatkotutkimusehdotuksia**

Laajojen kielimallien hyödyntämistä arvioijina kouluissa täytyy tutkia edelleen. Tulevien tutkimuksien painopisteenä tulisi olla arvioinnin luotettavuus. Luotettavuus määrittää sen, voiko laajoilla kielimalleilla tuottaa arviointeja. Laajojen kielimallien luotettavuutta voidaan tarkastella useista näkökulmista sekä paikallisten että pilvipohjaisten laajojen kielimallien avulla.

Yksi keskeinen jatkotutkimusehdotus olisi tutkia laajojen kielimallien kykyä arvioida eri tilanteissa. Esimerkiksi voisi tutkia taitotason, erilaisten tehtävien ja oppiaineiden vaikutusta arvioinnin luotettavuuteen. Mielenkiintoista olisi myös tutkia esimerkiksi

laajan kielimallin kykyä arvioida S2-oppilaiden vastauksia. Näin voitaisiin muodostaa käsitys siitä, millaisissa tilanteissa kielimallit soveltuvat arviointikäyttöön.

Toinen jatkotutkimusehdotus liittyy käytetyn kielen vaikutukseen laajan kielimallin tuottamaan arviointiin. Aiheesta on jo tehty tutkimuksia, mutta suomen kielen vaikutusta ei olla tutkittu. Tämän tutkimuksen aikana tehdyt havainnot viittasivat siihen, että laajalla kielimallilla oli haasteita tuottaa suomenkielisten aineistojen pohjalta arviointia. Virallista tutkimusta aiheesta ei tehty, joten johtopäätöksiä ei voida esittää. Kuitenkin aikaisempien tutkimuksien kuten Fokidesin ja Peristerakin (2025) sekä Kincl ym. (2024) tulokset osoittavat, että laajat kielimallit eivät onnistuneet arvioimaan luotettavasti kreikan- tai tšekinkielisiä esseitä.

Kolmas jatkotutkimusehdotus on tutkia eri laajojen kielimallien toimintaa arvioijana. Mielenkiintoista olisi selvittää, mikä malli tuottaa luotettavinta arviointia. Tämän tutkimuksen aineistonkeruun yhteydessä tuotettiin arvioinnit myös pilvipohjaisilla ChatGPT 5.2 ja Gemini 3 Flash malleilla. Se tarjoaa jatkotutkimusmahdollisuuden selvittää kyseisten mallien luotettavuutta, sekä kuinka mallien toiminta ja luotettavuus eroaa paikallisista malleista. Alustavia havaintoja mallien toiminnasta arvioijina esiteltiin edellisessä luvussa.

Viimeinen ja erityisen merkittävä jatkotutkimusehdotus liittyy kehotesuunnitteluun. Tutkimuksessa voisi selvittää, mitkä tiedot ovat välttämättömiä luotettavan arvioinnin kannalta. Tämän tutkimuksen promptia luodessa esimerkiksi oletettiin, että laajan kielimallin roolin antaminen olisi välttämätöntä. Kehotesuunnitteluun keskittyvän tutkimuksen avulla saataisiin mahdollisesti lyhennettyä prompteja, joka vapauttaisi laajan kielimallin kontekstin pituutta. Lisäksi olisi tärkeää tutkia, pienienkin promptin muutoksien vaikutusta arviointiin. Voisiko esimerkiksi yksittäisellä sanalla olla suuri vaikutus lopputulokseen. Tällaiset tutkimukset lisääisivät ymmärrystä promptin vaikutuksista arviointiin ja sen luotettavuuteen.

Täten voidaan todeta, että laajojen kielimallien soveltuvuutta tehtävien arviointiin tulee tutkia edelleen ja mahdollisimman monipuolisesti. Edellä esitetyt jatkotutkimusehdotukset tarjoavat useita näkökulmia aiheen tarkasteluun. Laajojen kielimallien hyödyntäminen arvioinnissa voisi parhaillaan vähentää opettajan

suorittaman arvioinnin riskejä, kuten puolueellisuutta ja epäjohtonmukaisuutta. Tämä edellyttää kuitenkin, että laajat kielimallit kykenisivät tuottamaan luotettavaa arviointia. Nykyisen tutkimustiedon perusteella tällaisen luotettavuuden saavuttaminen ei toistaiseksi vaikuta realistiselta.

## Lähteet

- Almegren, A., Mahdi, H., Hazaea, A., Ali, J. & Almegren, R. (2024). Evaluating the quality of AI feedback: A comparative study of AI and human essay grading. *Innovations in Education and Teaching International*, 1–16.  
<https://doi.org/10.1080/14703297.2024.2437122>
- Alsalem, M. (2024). EFL teachers' perception of the use of an AI grading tool (CoGrader) in English writing assessment at Saudi universities: an activity theory perspective. *Cogent Education*, 11:1, 2430865, DOI: 10.1080/2331186X.2024.2430865
- Atjonen, P. (2007). *Hyvä, paha arviointi*. Helsinki: Tammi.
- Atjonen, P., Laivamaa, H., Levonen, A., Orell, S., Saari, M., Sulonen, K., Tammi, M., Kamppi, P., Rumpu, N., Hietala, R. & Immonen, J. (2019). ”Että tietää missä on menossa” Oppimisen ja osaamisen arviointi perusopetuksessa ja lukiokoulutuksessa. Kansallinen koulutuksen arviointikeskus (KARVI). Julkaisu 7:2019.
- Azoma. (05.01.2026). What's the most popular LLM? What the latest data shows. Haettu 27.01.2026 osoitteesta: <https://www.azoma.ai/insights/most-popular-large-language-models>
- Baidoo-Anu, D., & Owusu Ansah, L. (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.  
<https://doi.org/10.61969/jai.1337500>
- Baird, J-A. (1998). What's in a name? Experiments with blind marking in A-level examinations, *Educational Research*, 40:2 s. 191-202, DOI: 10.1080/0013188980400207

- Barrot, J. & Bui N. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies* 30, 2041–2058 (2025). <https://doi.org/10.1007/s10639-024-12891-w>
- Bergmann, D. (ei pvm. a). What is machine learning?. *International Business machines, IBM*. Haettu 23.01.2026 osoitteesta: <https://www.ibm.com/think/topics/machine-learning>
- Bergmann, D. (ei pvm. b). What is a context window?. *International Business machines, IBM*. Haettu 07.05.2026 osoitteesta: <https://www.ibm.com/think/topics/context-window>
- Bouziane, K., & Bouziane, A. (2024). AI versus human effectiveness in essay evaluation. *Discover Education*, 3(1), Artikkelin 201. <https://doi.org/10.1007/s44217-024-00320-6>
- Caballar, R. (ei pvm.). Generative AI vs. predictive AI: What's the difference?. *International Business Machines, IBM*. Haettu 23.01.2026 osoitteesta: <https://www.ibm.com/think/topics/generative-ai-vs-predictive-ai-whats-the-difference>
- Cambridge Dictionary. (ei pvm.). Prompt- English meaning. Haettu 10.09.2025 osoitteesta: <https://dictionary.cambridge.org/dictionary/english/prompt>
- Chai, F., Ma, J., Wang, Y., Zhu, J. & Han, T. (2024). Grading by AI make me feel fairer? How different evaluators affect college students' perception of fairness. *Frontiers Psychology*. vol. 15:1221177. DOI: 10.3389/fpsyg.2024.1221177
- Chan, C. & Lee, K. (2023). The AI generation gap: Are Gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learning than their Gen X

and millennial generation teachers? Smart Learning Environments, 10(60).

<https://doi.org/10.1186/s40561-023-00269-3>

Chalmers, D & McAusland, W. (ei pvm.). Computer assisted assessment. The handbook for economics lecturers. Haettu 07.11.2025 osoitteesta:

[https://economicsnetwork.ac.uk/handbook/printable/caa\\_v5.pdf](https://economicsnetwork.ac.uk/handbook/printable/caa_v5.pdf)

Chapman university. (ei pvm.). Bias in AI. Haettu 30.10.2025 osoitteesta:

<https://www.chapman.edu/ai/bias-in-ai.aspx>

ChatGPT. (ei pvm. a.). Hinnoittelu. Haettu 26.01.2026 osoitteesta:

<https://chatgpt.com/fi-FI/pricing/>

ChatGPT. (ei pvm. b.). 100 keskustelua korkeakouluopiskelijoille. Haettu 28.01.2026

osoitteesta: <https://chatgpt.com/use-cases/students/>

ChatGPT Suomi. (ei pvm.). ChatGPT:n usein kysytyt kysymykset (FAQ). Haettu

14.05.2025 osoitteesta: <https://chatgpt.fi/ukk/>

Crosley, B. (14.01.2026). Cost Per Token Analysis: Optimizing GPU Infrastructure for LLM Inference. Introl. Haettu 07.05.2026 osoitteesta:

<https://introl.com/blog/cost-per-token-llm-inference-optimization>

Deep L. (ei pvm.). Secure and compliant Language AI for business communication.

Haettu 13.04.2026 osoitteesta: <https://www.deepl.com/en/pro-data-security>

Doshi, T. (17.12.2025). Gemini 3 Flash: frontier intelligence built for speed. Haettu

28.01.2026 osoitteesta: <https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/>

Edu.turku.fi. (ei pvm.). Tekoälysovellukset opetuksessa. Haettu 15.05.2025 osoitteesta:

<https://edu.turku.fi/ohjeet/tekoalysovellukset-opetuksessa/>

Ellonen, N. & Kaakinen, M. (ei pvm.). Regressioanalyysi. Teoksessa Kvantitatiivisen tutkimuksen verkkokäsikirja. Tampere: Yhteiskuntatieteellinen tietoarkisto [ylläpitäjä ja tuottaja]. Haettu 02.03.2026 osoitteesta: <https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/>.

Eskola, J., & Suoranta, J. (1998). Johdatus laadulliseen tutkimukseen. Vastapaino.

European Commission. (2018). A definition of AI: Main capabilities and scientific disciplines. High-Level Expert Group on Artificial Intelligence. Haettu 14.05.2025 osoitteesta: [https://ec.europa.eu/futurium/en/system/files/ged/ai\\_hleg\\_definition\\_of\\_ai\\_18\\_december\\_1.pdf](https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf)

Euroopan parlamentti. (2023). Mitä tekoäly on ja mihin sitä käytetään?. Haettu 14.05.2025 osoitteesta: <https://www.europarl.europa.eu/topics/fi/article/20200827STO85804/mita-tekoaly-on-ja-mihin-sita-kaytetaan>

Euroopan parlamentti. (2025). Tekoäly: mahdollisuuksia ja uhkia. Haettu 14.05.2025 osoitteesta: <https://www.europarl.europa.eu/topics/fi/article/20200918STO87404/tekoaly-mahdollisuuksia-ja-uhkia>

Faverio, M. & Sidoti, O. (2025). Teens, Social Media and AI Chatbots 2025. Pew Research Center. <https://pewrsr.ch/3KOIf7i>

Fingerhut, H. & O'Brien, M. (02.09.2023). Artificial intelligence technology behind ChatGPT was built in Iowa — with a lot of water. The Associated Press. Haettu 03.03.2026 osoitteesta: <https://apnews.com/article/chatgpt-gpt4-iowa-ai-water-consumption-microsoft-f551fde98083d17a7e8d904f8be822c4>

Fokides, E., & Peristeraki, E. (2025). Comparing ChatGPT's correction and feedback comments with that of educators in the context of primary students' short essays written in English and Greek. *Education and Information Technologies*, 30(2), 2577–2621. <https://doi.org/10.1007/s10639-024-12912-8>

Gadesha, V. (ei pvm.). What is prompt engineering?. *International Business machines, IBM*. Haettu 27.04.2026 osoitteesta: <https://www.ibm.com/think/topics/prompt-engineering>

Gemini. (ei pvm. a.). Release notes. Haettu 28.01.2026 osoitteesta: <https://gemini.google/gb/release-notes/?hl=en-GB>

Gemini. (ei pvm. b.). Subscriptions. Haettu 28.01.2026 osoitteesta: <https://gemini.google/subscriptions/>

Gilmurray, K. (2024). Top jobs most at risk of being replaced by AI. *Linkdin*. Haettu 15.05.2025 osoitteesta: <https://www.linkedin.com/pulse/top-jobs-most-risk-being-replaced-ai-kieran-gilmurray-9ashe>

Gomstyn, A. & Jonker, A. (ei pvm.) Exploring privacy issues in the age of AI. *International Business machines, IBM*. Haettu 30.10.2025 osoitteesta: <https://www.ibm.com/think/insights/ai-privacy>

Gunther, K., Hasanen, K. & Juhila, K. (2021). Johdanto: analyysi ja tulkinta. Teoksessa Vuori, J. (toim.) *Laadullisen tutkimuksen verkkokäsikirja*. Tampere: Yhteiskuntatieteellinen tietoarkisto [ylläpitäjä ja tuottaja]. Haettu 16.05.2026 osoitteesta: <https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/>

Hart, L. & Vesanummi, M. (18.11.2024). Tekoäly ruokkii kouluissa luovuutta- Opetushallitus linjaa tekoälyn osaksi sivistystä. *Yle uutiset*. Haettu 26.08.2025 osoitteesta: <https://yle.fi/a/74-20108424>

Harvard University. (30.08.2025). Getting started with prompts for text-based generative AI tools. Haettu 10.09.2025 osoitteesta: <https://www.huit.harvard.edu/news/ai-prompts>

Honkela, T. (ei pvm.). Neuroverkot: johdatus moderniin tekoälyyn. Haettu 23.01.2026 osoitteesta: <https://users.ics.aalto.fi/tho/stes/step96/honkela2/>

HyperAI. (10.09.2024). MMLU multilingual multitask Language Understanding Dataset. Haettu 28.01.2028 osoitteesta: <https://beta.hyper.ai/en/datasets/34840>

IBM technology. (28.07.2023). How large language models work [video]. Youtube. <https://youtu.be/5sLYAQS9sWQ?si=ILOZbQeYiCEh89S5>

IBM. (04.05.2026). Tokens and tokenization. International Business machines, IBM. Haettu 07.05.2026 osoitteesta: <https://www.ibm.com/docs/en/watsonx/saas?topic=solutions-tokens>

Jainschigg, J. (23.02.2026). LLM Optimization: Techniques and Guide. Mirantis. Haettu 07-05.2026 osoitteesta: <https://www.mirantis.com/blog/llm-optimization-techniques/>

Jyväskylän yliopisto. (ei pvm.). Määrällinen tutkimus. Haettu 13.05.2026 osoitteesta: <https://sites.app.jyu.fi/mehu/fi/menetelmapolku/tutkimusstrategiat/maarallinen-tutkimus>

Jokinen, E. (2021). Laadullinen arviointitutkimus. Teoksessa Vuori, J. (toim.) Laadullisen tutkimuksen verkkokäsikirja. Tampere: Yhteiskuntatieteellinen tietoarkisto [ylläpitäjä ja tuottaja]. Haettu 13.05.2026 osoitteesta: <https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvali/tutkimusasetelma/laadullinen-arviointitutkimus/>

Kincl, T., Gunina, D., Novák, M., & Pospíšil, J. (2024). Comparing Human and AI-based Essay Evaluation in the Czech Higher Education: Challenges and Limitations. *Trendy v podnikání = Business trends*, 14(2), 25–34.  
[https://doi.org/10.24132/jbt.2024.14.2.25\\_34](https://doi.org/10.24132/jbt.2024.14.2.25_34)

Kivimäki, P. & Krautsuk, S. (06.09.2019). Oppilaat saivat samalla osaamisella monta eri arvosanaa – Numeroarviointi menee kouluissa uusiksi. *Yle uutiset*. Haettu 20.02.2025 osoitteesta: <https://yle.fi/a/3-10957297>

Kumar, R. (29.04.2025). Day 5: Most Used LLMs in 2025 — Who’s Leading the AI Revolution?. *Medium*. Haettu 27.01.2026 osoitteesta:  
<http://medium.com/@rajukumardalimss/day-5-most-used-llms-in-2025-whos-leading-the-ai-revolution-12956f983301>

Kuopion lukiokoulutus. (ei pvm.). Ohjeet tekoälyn käyttöön opiskelijalle. *Haelukioon.fi*. Haettu osoitteesta 03.03.2026: <https://haelukioon.fi/web/ohjeet-tekoalyn-kayttoon-opiskelijalle/>

Kymäläinen, S. (02.05.2025). Helsinki kielsi opettajia käyttämästä tekoälyä – Kouluissa pohditaan, saako Wordin automaattista oikolukua enää käyttää. *Yle uutiset*. Haettu 15.05.2025 osoitteesta: <https://yle.fi/a/74-20141409>

Lappalainen, P. (ei pvm.) Tee se itse tekoäly. *University of helsinki*. Haettu 26.01.2026 osoitteesta: <https://blogs.helsinki.fi/summamutikka/files/2018/08/Tee-se-itse-teko%C3%A4ly.pdf>

Lin, L. (2024). A quarter of U.S. teachers say AI tools do more harm than good in K-12 education. *Pew Research Center*. <https://www.pewresearch.org/?p=167108>

LM Arena. (ei pvm.). Leaderboard overview. Haettu 26.01.2026 osoitteesta:  
<https://lmarena.ai/fi/leaderboard/>

LM Studio. (ei pvm.). gpt-oss. Haettu 13.04.2026 osoitteesta:

<https://lmstudio.ai/models/gpt-oss>

Lo, C. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Educational Sciences*, 13(4), 410. DOI:10.3390/educsci13040410

Lumme, M. (09.03.2024). Oppilaan tunteminen on entistä tärkeämpää, kun tekoäly valtaa luokat Vantaalla – "Esseetyyppiset kotitehtävät alkavat olla historiaa". *Vantaan sanomat*. Haettu 15.05.2025 osoitteesta:  
<https://www.vantaansanomat.fi/paikalliset/6614835>

Lång, K., Josefsson, V., Larsson, A.-M., Larsson, S., Högberg, C., Sartor, H., Hofvind, S., Andersson, I., & Rosso, A. (2023). Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *The Lancet Oncology*, 24(8), 936–944. [https://doi.org/10.1016/S1470-2045\(23\)00298-X](https://doi.org/10.1016/S1470-2045(23)00298-X)

Magro, C., Attal, K., Li, V., Schaffer, O. & Qiao, A. (2025). GradeGPT- Generative AI for grading post OSCE notes. <https://doi.org/10.1111/medu.70044>

Malec, M. (4.5.2025). Open-Source LLMs vs Closed: Unbiased Guide for Innovative Companies. *HatchWorksAI*. Haettu 30.10.2025 osoitteesta:  
<https://hatchworks.com/blog/gen-ai/open-source-vs-closed-llms-guide/>

Malouff, J., Emmerton, A. & Schutte, N. (2013). The Risk of a Halo Bias as a Reason to Keep Students Anonymous During Grading. *Teaching of Psychology* 40(3) 233–237. DOI: 10.1177/0098628313487425

Marshall, J. (1972). Writing neatness, composition errors and essay grades re-examined. *The Journal of Educational Research* (Washington, D.C.), 1165(JAN), 213.

McClain, C., Anderson, M., Sidoti., O. & Bishop, W. (2026). How teens use and view AI. Pew Re-search Center. <https://pewrsr.ch/4s7bgLm>

Microsoft. (16.03.2023). Introducing Microsoft 365 Copilot – your copilot for work. Haettu 27.01.2026 osoitteesta: <https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/>

Microsoft. (17.11.2025). Choose between cloud-based and local AI models. Haettu 28.05.2026 osoitteesta: <https://learn.microsoft.com/en-us/windows/ai/cloud-ai>

Microsoft. (ei pvm.). Understanding the different Microsoft Copilot experiences. Haettu 27.01.2026 osoitteesta: <https://support.microsoft.com/en-us/topic/understanding-the-different-microsoft-copilot-experiences-cfff4791-694a-4d90-9c9c-1eb3fb28e842>

MORI. (2021). Time Spent On Assessment "Ineffective". Ipsos. Haettu 13.04.2026 osoitteesta: <https://www.ipsos.com/en-uk/time-spent-assessment-ineffective>

MOT Sanakirja. (ei pvm.). “prompt”. Haettu 23.10.2025 osoitteesta: <https://www.sanakirja.fi/english-finnish/prompt>

Mucci, T. (ei pvm.). The history of artificial intelligence. International Business machines, IBM. Haettu 11.01.2026 osoitteesta: <https://www.ibm.com/think/topics/history-of-artificial-intelligence>

Myllyoja, E. (01.06.2023). Opettajat ovat nyt pulassa tekoälyn kanssa – Näin ovelilla kikoilla nuoret nyt huijaavat. Helsingin sanomat. Haettu 15.05.2025 osoitteesta: <https://www.hs.fi/kotimaa/art-2000009620438.html>

- Mäkelä, K. (1990). Kvalitatiivisen aineiston arviointiperusteet. Teoksessa Mäkelä, K. (toim.) Kvalitatiivisen aineiston analyysi ja tulkinta. Helsinki. Gaudeamus s.42–61.
- Nasir, M., Hasan, M., Adlim, A & Syukri, M. (2024). Utilizing artificial intelligence in education to enhance teaching effectiveness. PROCEEDINGS OF INTERNATIONAL CONFERENCE ON EDUCATION (PICE).  
<https://doi.org/10.32672/pice.v2i1.1367>
- Nguyen, T., Thuan, H. & Nguyen, M. (2023). Artificial Intelligent (AI) in teaching and learning: A comprehensive review. ISTES BOOKS, 140-161.
- Nicolau, J., Mellinas, J. & Martín-Fuentes, E. (2022). The halo effect. Teoksessa Buhalis, D. (toim.). Encyclopedia of Tourism Management and Marketing.  
DOI:10.4337/9781800377486
- OAJ. (ei pvm.). Opettajana perusopetuksessa. Haettu 11.04.2025 osoitteesta:  
<https://www.oaj.fi/arjessa/mita-opettajan-tyo-on/opettajana-perusopetuksessa/>
- OECD (2014), Talis 2013 Results: An International Perspective on Teaching and Learning, TALIS, OECD Publishing. <http://dx.doi.org/10.1787/9789264196261-en>
- Ouakrim-Soivio, N. (2013). Toimivatko päättöarvioinnin kriteerit? Oppilaiden saamat arvo-sanat ja Opetushallituksen oppimistulosten seuranta-arviointi koulujen välisten osaamiserojen mittareina. Opetushallitus. Raportit ja selvitykset 2013:9.  
<https://helda.helsinki.fi/server/api/core/bitstreams/5da3d6e9-b7c8-4747-bc16-0d28af6604b5/content>
- Okulich-Kazarin, V., Artyukhov, A., Skowron, Ł., Artyukhova, N., Dluhopolskyi, O., & Cwynar, W. (2024). Sustainability of Higher Education: Study of Student Opinions about the Possibility of Replacing Teachers with AI Technologies. Sustainability, 16(1), 55. <https://doi.org/10.3390/su16010055>

OpenAI. (ei pvm. a). Turvallisuus. Haettu 30.10.2025 osoitteesta: <https://openai.com/fi-FI/safety/>

OpenAI. (ei pvm. b). ChatGPT-Release notes. Haettu 26.01.2026 osoitteesta: <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>

OpenAI. (05.08.2025 a). Esittelyssä gpt-oss. Haettu 27.04.2026 osoitteesta: <https://openai.com/fi-FI/index/introducing-gpt-oss/>

OpenAI. (11.12. 2025 b). Esittelyssä GPT-5.2. Haettu 27.01.2026 osoitteesta: <https://openai.com/fi-FI/index/introducing-gpt-5-2/>

Opetushallitus. (2014). Perusopetuksen opetussuunnitelman perusteet 2014. Helsinki: Opetushallitus.

Opetushallitus. (ei pvm. a). Tekoäly varhaiskasvatuksessa ja koulutuksessa- lainsäädäntö ja suositukset. Haettu 15.05.2025 osoitteesta: <https://www.oph.fi/fi/tekoalysuosituksset>

Opetushallitus. (ei pvm. b). Tekoälyn mahdollisuuksia opetuksen ja opiskelun tukena. Tekoäly varhaiskasvatuksessa ja koulutuksessa- lainsäädäntö ja suositukset. Haettu 11.09.2025 osoitteesta: <https://www.oph.fi/fi/teemat-ja-kehittaminen/tekoalysuosituksset/tausta-aineisto-tekoalyn-mahdollisuuksia-opetuksen-ja-opiskelun-tukena>

Opetushallitus. (ei pvm. c.). Tekoäly oppimisen ja osaamisen arvioinnissa. Tekoäly varhaiskasvatuksessa ja koulutuksessa- lainsäädäntö ja suositukset. Haettu 11.09.2025 osoitteesta: <https://www.oph.fi/fi/teemat-ja-kehittaminen/tekoalysuosituksset/tausta-aineisto-tekoaly-oppimisen-ja-osaamisen-arvioinnissa>

Opetushallitus. (ei pvm. d.). Oppilaan oppimisen ja osaamisen arviointi ja päättöarvioinnin kriteerit. Haettu 20.02.2025 osoitteesta:  
<https://www.oph.fi/fi/koulutus-ja-tutkinnot/oppilaan-oppimisen-ja-osaamisen-arviointi-ja-paattoarvioinnin-kriteerit>

Opetushallitus. (ei pvm. e.). Tietoa perusopetuksen arvioinnin uudistusprosessista. Haettu 20.02.2025 osoitteesta: <https://www.oph.fi/fi/koulutus-ja-tutkinnot/tietoa-perusopetuksen-arvioinnin-uudistusprosessista>.

Oulun steinerkoulu. (ei pvm.). Ohjeet tekoälyn käyttöön opiskelussa. Haettu 03.03.2026 osoitteesta: <https://peda.net/oulun-steinerkoulu/koulun-s%C3%A4%C3%A4nn%C3%B6t/ohjeet-tekoalyn-kayttoon-opiskelussa>

Page, E. B. (1968). The Use of the Computer in Analyzing Student Essays. *International Review of Education*, 14(2), 210–225.

Perusopetuslaki 628/1998. Haettu 04.04.2025 osoitteesta:  
<https://www.finlex.fi/fi/laki/ajantasa/1998/19980628>

Placido, D., Yuan, B., Hjaltelin, J. X., Zheng, C., Haue, A. D., Chmura, P. J., Yuan, C., Kim, J., Umeton, R., Antell, G., Chowdhury, A., Franz, A., Brais, L., Andrews, E., Marks, D. S., Regev, A., Ayandeh, S., Brophy, M. T., Do, N. V., ... Sander, C. (2023). A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature Medicine*, 29(5), 1113–1122.  
<https://doi.org/10.1038/s41591-023-02332-5>

Puukko, M., Huhtanen, M. & Lepola, L. (2020). Perusopetuksen päättöarvioinnin kriteereiden toimivuuden arviointi. Kansallinen koulutuksen arviointikeskus. Julkaisu 7/2020.

Pulkkinen, J., Hildén, R., Ketonen, L., Khawaja, A., Kotila, J., Mäkipää, T., Nissinen, K., Puhakka, E., Rantala, J., Rautopuro, J., Sulkunen, S., Söyrinki, S., & Tammelin-

- Laine, T. (2024). Arviointiuudistuksen alkutaipaleella: Perusopetuksen päättöarviointitutkimuksen loppuraportti. Opetus- ja kulttuuriministeriön julkaisuja 2024:6. Opetus- ja kulttuuriministeriö. Helsinki.
- Puusa, A., Juuti, P., & Aaltio, I. (2021). Laadullisen tutkimuksen näkökulmat ja menetelmät. Gaudeamus. <https://urn.fi/URN:ISBN:978-952-345-616-7>
- Rane, N.L. (2024). Enhancing the quality of teaching and learning through Gemini, ChatGPT, and similar generative Artificial Intelligence: Challenges, future prospects, and ethical considerations in education. *TESOL and Technology Studies* 5(1). <https://doi.org/10.48185/tts.v5i1.1000>
- Rauman lyseon lukio. (2024). Rauman lyseon lukion tekoälyohjeistus. Haettu 03.03.2026 osoitteesta: [https://peda.net/rauma/rauman-lukio/opiskelijan\\_palvelut/tietotekniikka/ohjeet-ja-saannot/rauman-lyseon-lukion-tekoalyohjeistus.pdf:file/download/9622e6b0b758fff604d6e1ad7fc761e3545909d9/Rauman%20Lyseon%20lukion%20teko%C3%A4lyohjeistus.pdf](https://peda.net/rauma/rauman-lukio/opiskelijan_palvelut/tietotekniikka/ohjeet-ja-saannot/rauman-lyseon-lukion-tekoalyohjeistus.pdf:file/download/9622e6b0b758fff604d6e1ad7fc761e3545909d9/Rauman%20Lyseon%20lukion%20teko%C3%A4lyohjeistus.pdf)
- Reddit, (21.04.2026). Is GPT-OSS-120B still the best model among those with the same parameters?. Viestiketju keskustelualusta Redditissä. Haettu 13.05.2026 osoitteesta: [https://www.reddit.com/r/LocalLLM/comments/1sqmul3/is\\_gptoss120b\\_still\\_the\\_best\\_model\\_among\\_those/](https://www.reddit.com/r/LocalLLM/comments/1sqmul3/is_gptoss120b_still_the_best_model_among_those/)
- Saadioui, Z. (17.04.2025). Navigating the World of Local AI for Beginners. Arsturn. Haettu 30.10.2025 osoitteesta: <https://www.arsturn.com/blog/navigating-the-world-of-local-ai-for-beginners>
- Seitamaa-Hakkarainen, P. (ei pvm.). Kvalitatiivinen sisällönanalyysi. Sivustolla Metodix. Haettu 16.05.2026 osoitteesta: <https://metodix.fi/2014/05/19/seitamaa-hakkarainen-kvalitatiivinen-sisallon-analyysi/>

Sadanand, V., Guruvyas, K., Patil, P., Acharya, J. & Suryakanth, S. (2022). An automated essay evaluation system using natural language processing and sentiment analysis. *International Journal of Electrical and Computer Engineering (IJECE)*, 12(6), 6585–6593

Sidoti, O., Park, E. & Gottfried, J. (2025). About a quarter of U.S. teens have used ChatGPT for schoolwork – double the share in 2023. Pew Research Center. <https://pewrsr.ch/4g3Jqt0>

Taanila, A. (2020). Wilcoxon merkittyjen sijalukujen testi. *Akin menetelmäblogi*. Haettu 20.05.2026 osoitteesta: <https://tilastoapu.wordpress.com/2012/03/18/wilcoxon-merkittyjen-sijalukujen-testi/>

Taanila, A. (2013). P-arvo. *Akin menetelmäblogi*. Haettu 20.05.2026 osoitteesta: <https://tilastoapu.wordpress.com/2012/02/14/p-arvo/>

Terävä, H. (11.02.2020). Numeroiden antamista aikaistetaan peruskouluissa – arviointia myös yhtenäistetään oppilaiden yhdenvertaisuuden turvaamiseksi. *Yle uutiset*. Haettu 20.02.2025 osoitteesta: <https://yle.fi/a/3-11193768>

The university of Arizona. (ei pvm.). What is a large language model (LLM)?. Haettu 15.05.2025 osoitteesta: <https://ask.library.arizona.edu/faq/407985>

Tikkanen, T. (ei pvm.). Kysymyksiä peruskoulun lukuvuosiarvioinnista. *Opettaja*. Haettu 25.05.2025 osoitteesta: <https://www.opettaja.fi/tyossa/kysy-jarjestolta-kun-peruskoulun-lukuvuosiarviointi-askarruttaa/>

Tossel, C., Tenhundfeld, N., Momen, A., Cooley, K. & Visser, E. (2024). Student Perceptions of ChatGPT Use in a College Essay Assignment: Implications for

Learning, Grading, and Trust in Artificial Intelligence. IEEE Transactions on Learning Technologies, vol. 17 s.1069-1081 doi: 10.1109/TLT.2024.3355015

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

Turing, A. M. (1950). Computing Machinery and Intelligence. Mind, 59(236), 433–460.  
<http://www.jstor.org/stable/2251299>

Turnitin (ei pvm.). Common Core State Standards Writing rubrics. Haettu 02.02.2026 osoitteesta:  
[https://www.csun.edu/sites/default/files/Common%20Core%20Rubrics\\_Gr11-12\\_turn\\_it\\_in\\_0.pdf](https://www.csun.edu/sites/default/files/Common%20Core%20Rubrics_Gr11-12_turn_it_in_0.pdf)

Turun yliopisto. (ei pvm.). Kirjaston tekoälyopas. Haettu 25.02.2025 osoitteesta:  
<https://utuguides.fi/tekoaly/tiedonhaku>

Tutkimuseettinen neuvottelukunta (2023). Hyvä tieteellinen käytäntö ja sen loukkausepäilyjen käsitteleminen Suomessa. Tutkimuseettisen neuvottelukunnan julkaisuja 2/2023. Helsinki. ISBN 978-952-5995-82-4

UN Environment programme. (2024). AI has an environmental problem. Here´s what the world can do about that. Haettu 13.05.2025 osoitteesta:  
<https://www.unep.org/news-and-stories/story/ai-has-environmental-problem-heres-what-world-can-do-about>

Valtiovarainministeriö. (2025). Ohjeistus generatiivisen tekoälyn hyödyntämisestä työn tukena ja apuvälineenä julkisessa hallinnossa. [PDF]. Valtiovarainministeriö. VN/6190/2025

Velkova, J. (ei pvm.). Dismantling public values, one data center at the time. Haettu 23.02.2025 osoitteesta: <https://www.repair-research.fi/blogi/dismantling-public-values-one-data-center-at-the-time>

Vesanummi, M. (22.04.2024). Näin monipuolisesti opettajat käyttävät tekoälyä työssään – tekoälykisan voittaja visioi tulevaisuuden koulutustarpeita. Yle uutiset. Haettu 23.01.2026 osoitteesta: <https://yle.fi/a/74-20083636>

Viitaila, M. (ei pvm.) Tekoälyn perusteet: koneoppiminen, työn tulevaisuus ja hyvä vai paha tekoäly. Microsoft. Haettu 26.01.2026 osoitteesta: <https://pulse.microsoft.com/fi-fi/transform-fi-fi/na/fa2-tekoalyn-perusteet-koneoppiminen-tyon-tulevaisuus-ja-hyva-vai-paha-tekoaly/>

Volk-Jesussek, H. (2025). Wilcoxon signed-rank test. Numiqo. Haettu 20.05.2026 osoitteesta: <http://numiqo.com/tutorial/wilcoxon-test>

Weizenbaum, J. (1966). ELIZA-a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. <https://doi.org/10.1145/365153.365168>

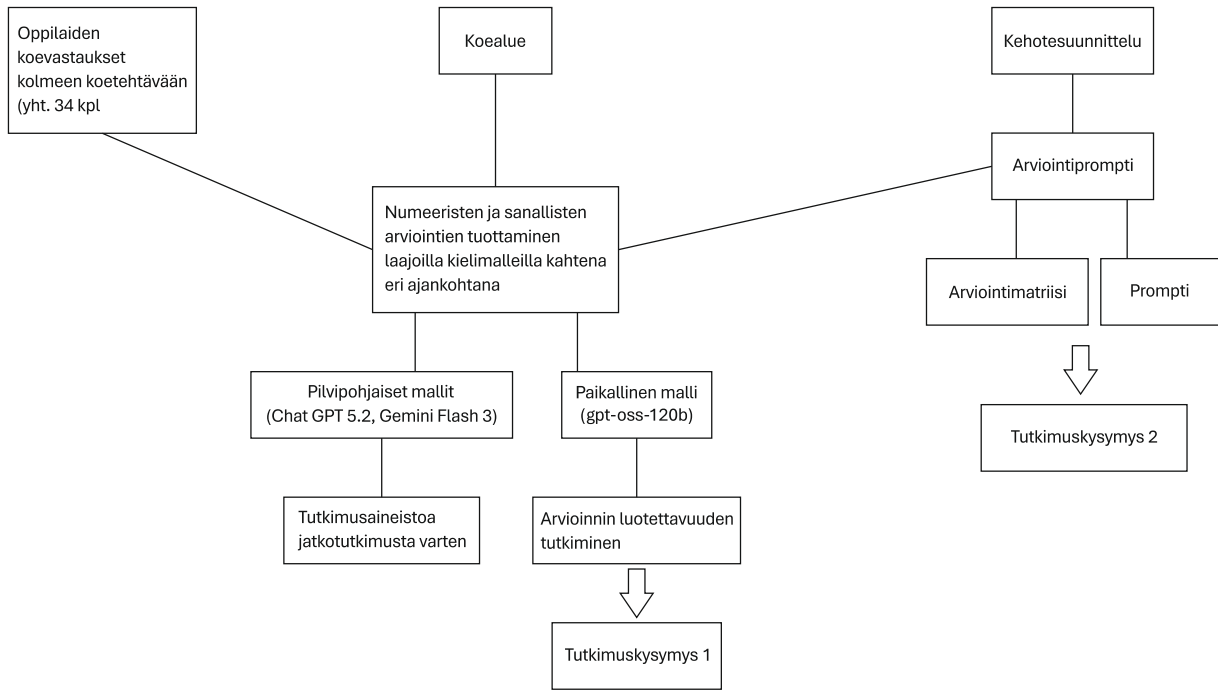
Wetzler, E., Cassidy, K., Jones, M., Frazier, C., Korbut, N., Sims, C., Bowen, S. & Wood, M. (2025). Grading the Graders: Comparing Generative AI and Human Assessment in Essay Evaluation. *Teaching of Psychology*, 52(3), 298–304. <https://doi.org/10.1177/00986283241282696>

XAMK. (2024). Mikä on tekoäly?. Kaakkois-Suomen ammattikorkeakoulu. Haettu 23.01.2026 osoitteesta: <https://ai.xamk.fi/johdatus-tekoalyyn/mika-on-tekoaly>

Zhao, N., Shi, C., Wang, C & Liu, X. (2025). Teachers' time use and job satisfaction: a largesample study in China. *Teacher Development*, 1–21. <https://doi.org/10.1080/13664530.2025.2577895>

## Liitteet

### Liite 1. Kuvio tutkimuksen koeasetelmasta



## Liite 2. Esimerkkejä analyysiyksiköiden arvojen tuottamisesta

Oppilaan vastaus	Laajan kielimallin tuottama arviointi	Analyysiyksiköiden arvot ja niiden perustelut
<p>Oppilaan A6 vastaus kysymykseen 1. (ensimmäinen arviointikierrös)</p> <p>Zeus was the supreme god. The Greeks believed that the gods caused storms, the sun, and earthquakes. The Greeks thought that when the sun shines and moves, a god is riding across the sky and that is where the light comes from. There were many different gods. The gods include Zeus, Hera, Hermes, Socrates, Ares, and Apollo</p>	<p>3 / 5</p> <p>The answer explains the role of gods in daily phenomena (storms, sun, earthquakes) and lists several deities, covering two of the required points.</p> <p>It does not mention how Greeks acknowledged the gods (e.g., sacrifices, pouring wine, oracles), nor does it address any modern-world connection.</p> <p>The inclusion of “Socrates” as a god is inaccurate according to the exam material, but other information aligns with the source.</p> <p>Since only some of the questions are answered and there are gaps, the response qualifies for 3 / 5.</p>	<p>a:2, b:2, c:2, d:2, e:2</p> <p>a. Laaja kielimalli arvioi vain arviointimatriisin asioita. Mainitsee, että oppilas on vastannut kahteen kysymykseen ja kertoo vastauksen puutteista.</p> <p>b. LLM ei vaadi tai käytä koeaineiston ulkopuolisia termejä. LLM ei vaadi vastaamaan koetehtävän ulkopuolisiin aiheisiin.</p> <p>c. Laajan kielimallin mainitsemat esimerkit ovat oppilaan vastauksesta.</p> <p>d. Laajan kielimallin väittämät puutteet puuttuvat oikeasti oppilaan vastauksesta. Ja laajan kielimallin väittämät oikeat asiat mainitaan oppilaan vastauksessa. Lisäksi laaja kielimalli tunnisti oppilaan tekemän virheen (Sokrates)</p> <p>e. Arvioinnin perustelut ovat linjassa arviointimatriisin kanssa, ja täten pisteytys on oikein. Oppilas vastasi vain osaan koetehtävän kysymyksistä, ja niissä oli puutteita. LLM antoi oikean, 3/5, arvioin.</p>
<p>Oppilaan A10 vastaus kysymykseen 1. (ensimmäinen arviointikierrös)</p> <p>Gods were worshipped at the Olympics. The gods included Zeus, Poseidon, Ares, Hera, Achilles, and Hermes. Gods associated with modern times include, for example, Hermes (the Kokkola hockey team) and Zeus: The Olympics were once</p>	<p>Score: 3/5</p> <p>Answered several parts of the task (daily rituals before meals, human like traits of the gods, number of main gods and their residence on Olympus).</p> <p>Included an example (offering olive oil/milk/wine, competitions in honor of the gods).</p>	<p>a:-, b:-, c:0, d:-, e:-</p> <p>Laajan kielimallin tuottamasta arvioinnista huomaa, että se ei ole arvioinut juuri kyseisen oppilaan vastausta. LLM:n mainitsemia esimerkkejä ei löydy oppilaan vastauksesta. Koska analyysiyksikkö c sai arvoksi nolla, ei muita yksiköitä pysty pisteyttämään.</p>

Oppilaan vastaus	Laajan kielimallin tuottama arviointi	Analyysiyksiköiden arvot ja niiden perustelut
held in honor of Zeus, and they are still held today!	<p>Did not address how people sought the gods' favor through sacrifices or prophecies, nor mention specific rituals such as pouring a drop before each meal in detail.</p> <p>Completely omitted the question about whether the Greek gods are connected to the modern world.</p>	
<p>Oppilaan A11 vastaus kysymykseen 1. (ensimmäinen arviointikierros)</p> <p>There were many kinds of gods; they were part of everyday life—for example, whenever there was thunder, people thought it was Zeus throwing stones and lightning. Even young children were taught from a young age to follow all the gods' rules. There were 12 main gods, and one of them was Zeus. We don't talk about the gods much anymore, but some people probably still believe in them</p>	<p>2/5</p> <p>Provided correct information that there were twelve main Olympian gods and that they lived on Mount Olympus.</p> <p>Mentioned one mythic explanation (Zeus causing thunder) linking a god to daily life, but omitted most of the required content about how Greeks acknowledged the gods (e.g., offerings before meals, oracles, temples, competitions) and did not give a clear example of worship practices beyond the storm story.</p> <p>Did not discuss the variety of gods beyond the twelve main ones nor address the modern-world connection adequately.</p>	<p>a:1, b:2, c:1, d:1, e:2</p> <p>a. Laaja kielimalli ei selkeästi mainitse moneenko kysymykseen oppilas on vastannut. LLM kuitenkin mainitsee vastauksen puutteista ja yhdestä esimerkistä.</p> <p>b. LLM ei vaadi tai käytä koeaineiston ulkopuolisia termejä. LLM ei vaadi vastaamaan koetettävän ulkopuolisiin aiheisiin.</p> <p>c. Laaja kielimalli mainitsee yhden esimerkin (jumalat asuvat Olympos-vuorella), jota oppilas ei maininnut. Toinen mainittu esimerkki on oppilaan vastauksesta.</p> <p>d. LLM väittää, että oppilas ei mainitse jumalien yhteyttä nykypäivään. Oppilas mainitsee yhteyden, vaikka se on virheellinen. Lisäksi LLM tulkitsee virheellisesti oppilaan maininnan Zeuksesta ukkosentuottajana vastaukseksi jumalien palvonnasta. Osa arvioinnissa mainituista asioista on kuitenkin oikein, ja täten analyysiyksikkö saa yhden pisteen.</p>

Oppilaan vastaus	Laajan kielimallin tuottama arviointi	Analyysiyksiköiden arvot ja niiden perustelut
		e. Arvioinnin perustelut ovat linjassa arviointimatriisin kanssa, ja täten pisteytys on oikein. Oppilas vastasi yhteen koetehtävän kysymyksistä ja se sisälsi puutteita. LLM antoi oikean, 2/5, arvion.
<p>Oppilaan A11 vastaus kysymykseen 2. (ensimmäinen arviointikierrös)</p> <p>Athens was very peaceful; people studied there, but in Sparta, they fought starting at the age of seven. Spartans began military training at age seven, where they fought and learned, for example, to endure more pain. In Athens, women weren't even allowed to vote; they were considered merely housewives, but in Sparta, women were allowed vote and walk the streets, whereas women in Athens weren't even allowed to go out on the streets.</p>	<p>3 / 5</p> <p>Answered parts of both questions – the student mentions everyday life (Athens “peaceful” with study; Sparta’s early military training) and a governance-related point (women’s voting rights).</p> <p>Provides an example – the reference to Spartans learning to endure pain through whipping is supported by the source.</p> <p>Content gaps – many relevant differences from the material are missing (e.g., Athens’ democratic assemblies, emphasis on arts, wealth from trade; Sparta’s agoge system, reliance on helot slaves, focus on military excellence).</p> <p>Incorrect information – the claim that Spartan women were allowed to vote is not supported by any of the provided citations and contradicts the material. This factual error reduces the overall quality of the answer.</p> <p>Because the response covers some required points but contains notable omissions and a factual inaccuracy, it meets the criteria for a “3/5” rating</p>	<p>a:2, b:1, c:2, d:2, e:2</p> <p>a. Laaja kielimalli arvioi vain arviointimatriisin asioita. Mainitsee vastauksen puutteista ja esimerkeistä sekä mihin kysymyksiin oppilas on vastannut.</p> <p>b. LLM käyttää ja vaatii termejä (agoge, helot orja), joita ei ole koealueessa mainittu. Muut asiat ovat kuitenkin linjassa koealueen ja -tehtävän kanssa.</p> <p>c. Laajan kielimallin mainitsemat esimerkit ovat oppilaan vastauksesta.</p> <p>d. Laajan kielimallin väittämät puutteet puuttuvat oikeasti oppilaan vastauksesta. Ja laajan kielimallin väittämät oikeat asiat mainitaan oppilaan vastauksessa. Lisäksi laaja kielimalli tunnisti oppilaan tekemän virheen (Spartassa naisten äänioikeus)</p> <p>e. Arvioinnin perustelut ovat linjassa arviointimatriisin kanssa, ja täten pisteytys on oikein. Oppilas vastasi vain osaan koetehtävän kysymyksistä, ja niissä oli puutteita. LLM antoi oikean, 3/5, arvion.</p>
<p>Oppilaan A6 vastaus kysymykseen 3. (ensimmäinen arviointikierrös)</p>	<p>3 / 5</p> <p>The answer addresses who Alexander was (leader of</p>	<p>a:2, b:2, c:2, d:1, e:2</p>

Oppilaan vastaus	Laajan kielimallin tuottama arviointi	Analyysiyksiköiden arvot ja niiden perustelut
<p>Alexander the Great was the leader/ruler of Macedonia. He conquered Egypt and defeated Persia. And he conquered as far as India. Alexander named many cities after him in the Persian region. When Alexander died, the names of all the cities were changed</p>	<p>Macedonia) and mentions several of his accomplishments (conquering Egypt, defeating Persia, reaching India, founding cities).</p> <p>It does not explain how he is connected to ancient Greece, nor does it describe how he succeeded in achieving his goals (e.g., his military skill, adoption of divine rulership, cultural integration) – two of the four required points are missing.</p> <p>The response includes an example (naming many cities after himself), satisfying the requirement for at least one example.</p> <p>Overall, only some of the exam questions are answered and there are clear gaps in content, matching the criteria for 3 / 5</p>	<p>a. Laaja kielimalli arvioi vain arviointimatriisin asioita. Mainitsee, että oppilas on vastannut kahteen kysymykseen ja kertoo vastauksen puutteista.</p> <p>b. LLM ei vaadi tai käytä koeaineiston ulkopuolisia termejä. LLM ei vaadi vastaamaan koetehtävän ulkopuolisiin aiheisiin.</p> <p>c. Laajan kielimallin mainitsemat esimerkit ovat oppilaan vastauksesta.</p> <p>d. Laaja kielimalli ei tunnista oppilaan tekemää virhettä (Alexanderin kuoleman jälkeen hänen nimeämien kaupunkien nimet vaihdetaan)</p> <p>e. Arvioinnin perustelut ovat linjassa arviointimatriisin kanssa, ja täten pisteytys on oikein. Oppilas vastasi vain osaan koetehtävän kysymyksistä, ja niissä oli puutteita. LLM antoi oikean, 3/5, arvioinnin.</p>