Tommi Sirkka

Generatiivisen tekoälyn mallien vertailu Metropolialle

Metropolia Ammattikorkeakoulu

Insinööri (AMK)

Tuotantotalous

Opinnäytetyö

29.4.2025

Tiivistelmä

Tekijä(t):	Tommi Sirkka
Otsikko:	Generatiivisen tekoälyn mallien vertailu Metropolialle
Sivumäärä:	54 sivua + 6 liitettä
Aika:	29.4.2025
Tutkinto:	Insinööri (AMK)
Tutkinto-ohjelma:	Tuotantotalouden tutkinto-ohjelma
Suuntautumisvaihtoehto:	SCM Johtaminen
Ohjaaja(t):	Nina Hellman, Lehtori

Tämän insinöörityön aiheena oli tukea Metropolian ammattikorkeakoulua sen nykyisissä ja tulevissa tekoälyyn liittyvissä projekteissa. Tavoitteena työllä oli luoda kehys generatiivisen tekoälyn mallien vertailuun.

Itse tutkimusta pohjustettiin kattavalla aiheen kirjallisuuteen tutustumisella. Tietoosuudesta löytyvät perusteet tekoälylle lähtien keskeisimmistä tekniikoista. Lisäksi generatiivisen tekoälyn ja sitä mahdollistavat teknologiat käytiin läpi. Tämän pohjalta pyrittiin rakentamaan generatiiviseen tekoälyyn ja sen kyvykkyyksiin tutustumalla sopivat perusteet mallien vertailulle. Vertailussa päädyttiin tuottamaan Power BI -tiedosto, jonka avulla valittujen ominaisuuksien vertailu olisi helpompaa ja visuaalisempaa.

Työn tuloksena oli generatiivisen tekoälyn mallien vertailuille kehys, jonka käyttöä demonstroitiin kolmea esimerkkimallia käyttäen. Työn lopputuloksena on Power BI -pohjainen vertailupohja sekä kokonaisuutena tapa, jolla malleja vertailla. Työ sisältää myös ehdotukset aiheeseen liittyvään jatkotutkimukseen, jolla saavutettaisiin lisää merkittävää tietoa generatiivisen tekoälyn käyttöönotosta ja sen hyödyntämisestä.

Avainsanat: tekoäly, generatiivinen tekoäly, tekoälyn käyttöönotto, vertailu

Abstract

Author(s):	Tommi Sirkka
Title:	Comparison of Generative Artificial Intelligence Models for Metropolia
Number of Pages:	54 pages + 6 appendices
Date:	29 April 2025
Degree:	Bachelor of Engineering
Degree Programme:	Industrial Management and Engineering
Specialisation option:	Name of the specialisation option
Instructor(s):	Nina Hellman, Senior Lecturer

The topic of this thesis was to support Metropolia University of Applied Sciences in its current and future artificial intelligence-related projects. The aim of the thesis was to create a framework for comparing generative artificial intelligence models.

The research itself was grounded in a comprehensive familiarization with the relevant literature. The information section covers the fundamentals of artificial intelligence, starting from its key techniques. In addition, generative artificial intelligence and the technologies that enable it were reviewed. Based on this, the aim was to build suitable foundations for model comparison by exploring generative artificial intelligence and its capabilities. The comparison resulted in a Power BI file, which makes the comparison of selected features easier and more visual.

The result of the thesis is a framework for comparing generative artificial intelligence models, the use of which was used for evaluating three example AI-models. The outcome of the thesis is a Power BI-based comparison template, as well as an overall base to comparing models. The thesis also includes proposals for further research on the topic, which would result more significant insights into the adoption and utilization of generative artificial intelligence

Keywords: artficial intelligence, generative ai, ai adoption, comparison

Tekoälyn käyttö insinöörityössä

Olen hyödyntänyt OpenAI:n ChatGPT mallia tässä työssä teoriaa käsittelevässä osiossa. Käytin mallia selittämään yksittäisiä konsepteja pyrkiessäni itse oppimaan aiheesta. Tekoälyä ei ole käytetty suoraan tähän työhön. Opinnäytetyön tekijänä olen itse vastuussa kaikesta opinnäytteeni sisällöstä.

1 Johdanto

Tekoälystä on vauhdilla tulossa ainakin jollain tavalla osa lähes meidän kaikkien elämää. Vaikka osalle vielä tuntematon teknologia aiheuttaa välillä kysymyksiä, epäilyksiä ja jopa pelkoja, auttaa tekoäly meitä jo tällä hetkellä jokapäiväisessä elämässä esimerkiksi suosittelemalla suoratoistopalveluissa uutta musiikkia ja elokuvia. Uudeksi internetiksikin kutsutun tekoälyn avulla, käyttäjä saa valjastettua käyttöönsä uskomattoman määrän tietoa ja tehoa, jotka useiden lähteiden mukaan voisivat tulevaisuudessa mahdollistaa ihmisen tehokkuuden siirtymisen uudelle tasolle, kuten internet teki 1990-luvulla. Kuten yksityishenkilöt, myös monet yritykset ovat innokkaasti lähteneet hyödyntämään ”uutta” teknologiaa omassa tekemisessään maanviljelystä uutistoimistoihin ja lääkäriasemiin. (European Commission 2024.)

Uusi ja mielenkiintoinen teknologia innostaa yhä useamman yrityksen hyppäämään kovaa puksuttavaan tekoälyjunaan ja uudistamaan sillä liiketoimintaansa. On hyvä kuitenkin muistaa, että tekoälykehityksen kärjessä viilettävillä, omia malleja kehittävillä valtavilla teknologiayrityksillä on hyvin eri tilanne pienempiin yrityksiin verrattuna. The Guardianin artikkelissa *Small businesses are not all in with artificial intelligence – yet* mainitaan pienten yritysten olevan vielä varovaisesti tekoälyn hyödyntämisessä mukana. Artikkelissa puhutaan erosta omia tekoälymalleja rakentavien ja kouluttavien suuryritysten ja valmiimman tekoälyratkaisun joltain muulta ostavien yrityksien välillä sekä alleviivataan tekoälyn käyttöön siirtymisen suunnitelmallisuutta sen jokaisella osa-alueella. (The Guardian 2025.)

Metropolia Ammattikorkeakoulu pyrkii rohkeasti olemaan osa tätä muutosta etsimällä ja ottamalla käyttöön tekoälypohjaisia ratkaisuja osana oman toimintansa kehittämistä. Konkreettisena esimerkkinä tästä on yksi Metropolian käynnistämistä projekteista, jonka tavoitteena on luoda generatiiviseen tekoälyyn pohjautuva suositustyökalu auttamaan opiskelijoita sopivien opintototeutusvaih-

toehtojen löytämisessä. 2024 aloitetun projektin lopputuloksen on tarkoitus hyödyntää käyttäjän antamia tietoja, kuten aikaisempia opiskeluja, työkokemusta ja kiinnostuksen kohteita, yhdessä opintototeutuksiin liittyvän taustadatan kanssa suositusten tuottamiseksi. Tekoälytyökalun luomiseen tarvitaan dataa, algoritmeja, laskentatehoa ja kielimalli, joka toimii ikään kuin pohjana koko sovellukselle.

Tämän projektin myötä esille nousi esiin tarve arvioida ja vertailla eri tekoälymalleja niiden heikkouksien ja vahvuuksien löytämiseksi tulevaisuuden projekteja varten. Tekoäly kehittyy ja uusia malleja tulee markkinoille tiuhaa tahtia erilaisia malleja kehiteltäessä eri käyttötarkoituksiin. Osana suunnitelmallista tekoälyn käyttöä tämän työn pyrkimyksenä on luoda pohja tulevia tekoälyprojekteja varten, tarjota taustatietoa tekoälystä, ja vertailla muutamaa yleisimmistä generatiivisen tekoälyn ratkaisuista ja antaa samalla pohja muidenkin mallien vertailulle tulevaisuudessa.

Tämän työn tavoitteena on siis tuottaa Metropolialle kehys generatiivisen tekoälyn mallien vertailuun. Seuraavassa osiossa käydään läpi insinöörityön suorittamisen suunnitelma.

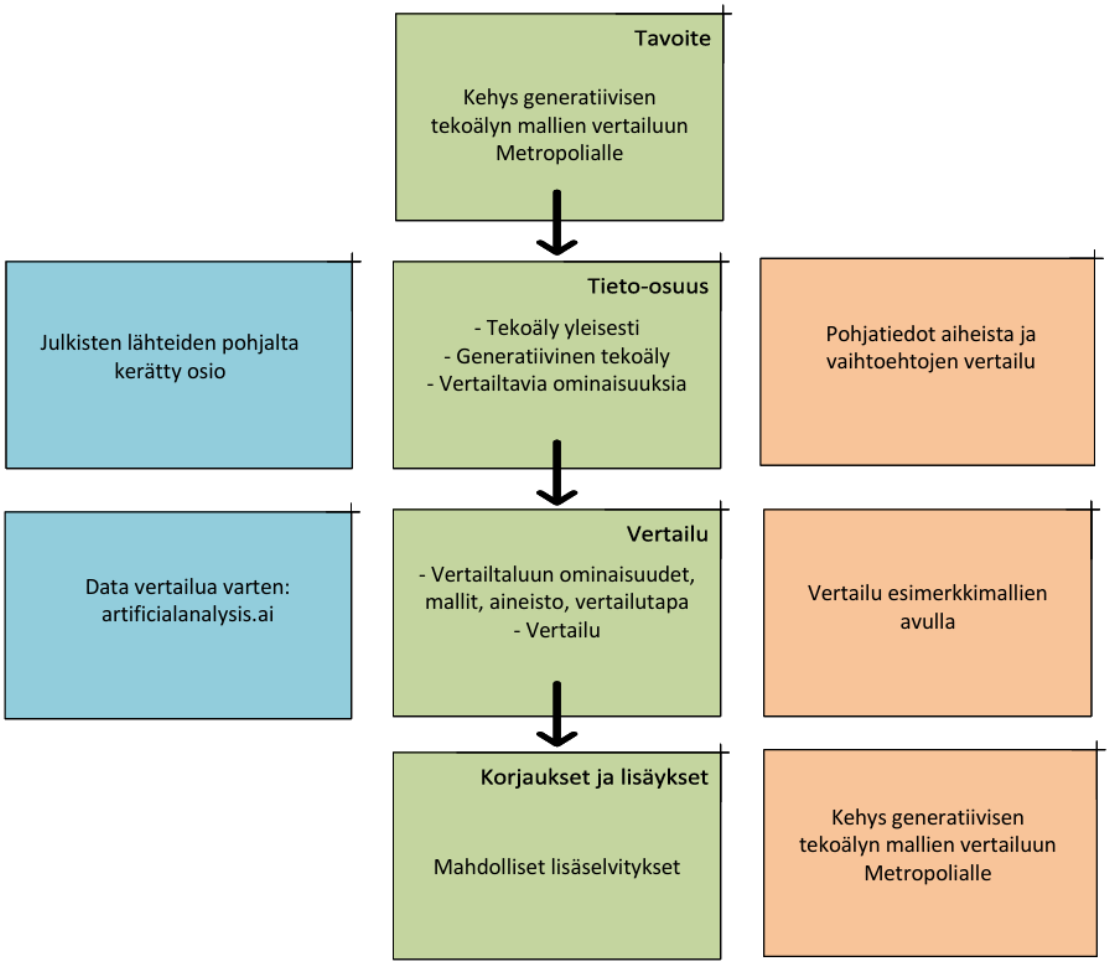
2 Tutkimussuunnitelma

Jotta oli mahdollista arvioida ja vertailla eri kielimalleja ja löytää menetelmiä käyttöönottoon, vaatimuksena oli laaja tutustuminen aiheen kirjallisuuteen. Tekoälyn tutkiminen sen historiasta nykytilaan sekä generatiivisen tekoälyn pääteknologioiden ja ominaisuuksien ennen itse vertailua antoi pohjaa siihen, mitä kielimalleja vertailuun otetaan ja mitä vertailtavia ominaisuuksia eri malleilla voi ylipäätään olla.

Itse kohdeyrityksen eli Metropolian nykytilan tutkiminen on rajattu tästä työstä lähes kokonaan pois, koska Metropolian tilanne ja tarpeet on aiheena toisen opiskelijan tekemällä opinnäytetyöllä. Metropolian tilanteen ja tarpeiden osalta tässä työssä hyödynnettiin valmiina saatavilla olleita dokumentteja yhdestä jo käynnissä olevasta tekoälyä hyödyntävän sovelluksen projektisuunnitelmasta sekä keskustelua mahdollisista tulevaisuuden tarpeista.

2.1 Tutkimuksen rakenne

Kuvassa 1 on insinöörityön vaiheet visuaalisessa muodossa ja sen alla lyhyt kirjallinen selitys työn vaiheista.



Kuva 1. Insinöörityön tutkimussuunnitelma.

Insinöörityön aihe ja tavoite muodostui Metropolian käynnissä olevan tekoäly-projektin pohjalta, jossa tavoitteena on luoda opiskelijoille tekoälyä hyödyntävä työkalu sopivien opintojen löytämisen avuksi. Tämän projektin sekä Metropolian yhteyshenkilön kanssa käydyn keskustelun pohjalta muodostui työlle tavoitteeksi kehys generatiivisen tekoälyn mallien vertailulle. Toisessa osassa oli suoritettava insinöörityön kirjoittajan suppean tekoälytietämyksen takia sukellus tekoälyn kirjallisuuteen perusteista lähtien. Kattavan tieto-osuuden tavoitteena oli myös löytää pohjatiedot generatiivisen tekoälyn toiminnasta ja ominaisuuksista, jonka perusteella itse vertailu suoritettaisii. Tämä tarjoaa myös samalla aiheesta tietämättömälle lukijalle mahdollisuuden tutustua tekoälyn perusteisiin.

Kolmannessa vaiheessa suoritettiin tekoälymallien vertailu. Tieto-osuudessa löytyneiden pohjatietojen ja saatavilla olleen aineiston pohjalta valittiin vertailua varten aineisto ja vertailtavat ominaisuudet sekä päädyttiin tukemaan vertailua Power BI -tiedostolla, joka mahdollistaisi visuaalisen puolen. Mallien vertailua päädyttiin havainnollistamaan valitsemalla kolme esimerkkimallia, joiden ominaisuuksia arvoitiin keskenään sekä vertailuaineistoon verrattuna. Neljännessä vaiheessa pyydettiin palaute Metropolialta sekä arvioitiin mahdollisia lisätutkimusaiheita suoritetun vertailun perusteella.

2.2 Tiedonkeruu

Metropolian tarpeiden kartoitusta sekä tekoälymallien vertailua varten oli kerättävä dataa. Taulukossa 1 on esitetty datan keräämisen vaiheet.

Taulukko 1. Datan keräämisen vaiheet

Vaihe	Lähde	Aihe	Päivämäärä	Dokumentointi
Tavoitteen suunnittelu	Tekoälypohjaisen suosittelutyökalun projektisuunnitelma	Pohjatietoja opinnäytetyötä varten	01.12.2024	
Tavoitteen suunnittelu	Teams -keskustelu, Metropolian Digitaaliset palvelut ja tiedolla johtaminen tiimin asiantuntijan kanssa	Taustatietoja Metropolian tilanteesta ja tarpeista	27.1.2025	Muistiinpanot
Data 1	artificialanalysis.ai	Tekoälymallien ominaisuudet	7.3.–10.3.	Excel taulukko, jossa tekoälymallien vertailussa käytetyt tiedot

Tavoitteen suunnittelu ja asettaminen pohjautuivat datan keräämisen ensimmäiseen vaiheeseen, johon kuuluu taulukon kaksi ensimmäistä osiota. Datana käytettiin Metropolian tekoälypohjaisen suosittelutyökalun kehittämisen projektisuunnitelmaa sekä keskustelua Metropolian yhteyshenkilön kanssa. Näiden pohjalta suunniteltiin raamit koko opinnäytetyön suorittamiselle. Tekoälymallien arvioinnille oli hankittava dataa, jonka perusteella vertailu suoritettiin. Taulukon

kolmas osio kuvaa tätä vaihetta, jossa päädyttiin käyttämään ainoastaan artificialanalysis.a-sivuston tarjoamaa dataa. Seuraavaksi on lyhyt pohjustus tekoälyn historiasta ja keskeisimmistä tekniikoista.

3 Tekoäly yleisesti

Suhteellisen hiljattain laajemmin pinnalle nousseena tekoälyn maailma on itse kirjoittajalle ja varmasti monille muillekin vielä kovin tuntematon. Jotta on myöhemmin mahdollista keskustella generatiivisesta tekoälystä ja sen ominaisuuksista, käydään alkuun lyhyesti läpi yleisesti tekoälyn historiaa ja perusperiaatteita aihealueeseen tutustumiseksi.

Ajatus ihmisen tapaan tietoa käsittelevästä, älyllisestä laitteesta on joidenkin lähteiden mukaan peräisin jo antiikin ajoilta. Nykyisen tekoälykehityksen voidaan kuitenkin katsoa syntyneen toisen maailmansodan aikaisen teknologiakehityksen vanavedessä, kun 1956 termi artificial intelligence oli ensimmäistä kertaa käytössä Dartmouthin yliopistossa Yhdysvalloissa, New Hampshiressä. Tuolloin pieni ryhmä tiedemiehiä kokoontui yhteen tekoälyn tutkimusprojektia varten, joka päätyi antamaan sysäyksen koko tekoälyä tutkivan tieteenalan kehitykselle. Projektin myötä tekoäly herätti paljon huomiota saaden myös osakseen investoijien kiinnostusta. (Dartmouth College 2024.) Siitä lähtien tekoäly on kokenut useita ylä- ja alamäkiä lähinnä tieteellisistä läpimurroista ja investoijien mielenkiinnosta riippuen. Kehitys ei ole ollut räjähdysmäistä ennen kuin vasta viime vuosina, vaikka ajatus ihmisen aivoja mukailevasta teknologiasta on ollut vireillä jo useita vuosikymmeniä. Seuraavassa luvussa on yksi tekoälyn tärkeimmistä tekniikoista eli koneoppiminen.

3.1 Määritelmä

Termin ”tekoäly” määrittely ei ole aivan yksinkertaista ja näkökulmasta riippuen sitä voidaan selittää hieman eri tavoin. Määrittelyyn voi pyrkiä esimerkiksi yksinkertaisesta arkisesta, virallisesta tai vertaamalla sen toimintalogiikkaa tutumpaan vaihtoehtoon.

Usein tekoälyksi kuvaillaan koneen kykyä suorittaa perinteisesti ihmisen älyyn liitettyjä ominaisuuksia, kuten päättelyä, oppimista, suunnittelemista ja luomista. Yleinen tapa on siis verrata tekoälyn kyvykkyyksiä ihmisen aivojen toimintaan. Toisaalta ihmisenkin aivojen toimintamalleja ja kyvykkyyksiä on kuitenkin välillä vaikea ymmärtää, mutta tekoälyn voidaan ajatella pystyvän samankaltaiseen ”älykkääseen” toimintaan, vaikka määritelmältään epätarkka, tekoälyn vertaus samankaltaisuuteen ihmismielen toiminnan kanssa on erityisesti arkikielessä pitkälti toimiva. (Euroopan Parlamentti 2020.)

Euroopan komission virallisen kannan mukaan tekoäly viittaa teknologioihin, jotka ovat kyvykkäitä vastaanottamaan ja keräämään tietoa sekä käyttämään sitä sille asetettujen tavoitteiden saavuttamiseksi. Monesti tekoälyyn liitetty kyky verrattuna muihin teknologioihin onkin taito tuottaa saamansa informaation perusteella täysin uutta, oli se sitten esimerkiksi tekstiä, kuvia tai ääntä. (European Commission 2024.)

Myös tekoälypohjaisen ohjelmoinnin vertaaminen perinteiseen ohjelmointiin antaa kuvan siitä, mitä tekoäly on ja mikä sen toimintaperiaatteena on. Kärjistetysti perinteisessä ohjelmoinnissa syötteinä ovat data ja ohjelmoijan määrittelemiä sääntöjä. Asetettuihin sääntöihin pohjautuen datan perusteella muodostuu vastauksia, jotka ovat faktoja. Jos annetut säännöt eivät riitä vastauksen muodostamiseen, lopputuloksen muodostaminen ei onnistu ollenkaan. Ongelmaksi siis muodostuu määriteltävien sääntöjen määrä, jotta todenmukaisen vastauksen muodostaminen onnistuisi, oli data sitten mitä tahansa. Tekoälypohjaisessa ohjelmoinnissa taas samojen syötteiden, sääntöjen ja datan pohjalta muodostuu logiikka siitä, minkälaisia vastauksia mistäkin datasta tulisi tuottaa. Kun sitten otetaan käyttöön uutta dataa, muodostuu aikaisemman logiikan perusteella ”valistunut arvaus” siitä, mikä voisi olla oikea vastaus. Tekoäly pohjautuu siis todennäköisyyksiin, tilastollisiin jakaumiin eikä faktapohjaisiin kyllä-ei vastauksiin. Ongelmana ei olekaan enää sääntöjen määrittelyn määrä vaan toimintalogiikan osumatarkkuus, jota pystytään parantamaan erinäisin keinoin. (Kananen; Puolitaival; Puntti & Metsola 2019.)

3.2 Koneoppiminen

Koko tekoälyn vahvuus perustuu siihen, että kone on kykeneväinen muodostamaan datan pohjalta loogisia päätelmiä ja käyttämään näitä päätelmiä saadessaan käyttöönsä täysin uutta dataa. Yhtälön toimivuudesta vastaavat ja sen tehokkuuteen vaikuttavat osat ovat siis data, jonka pohjalta päätelmiä tehdään sekä tapa, jolla kone tätä dataa käsittelee. Termi ”koneoppiminen” kuvaa juuri tätä prosessia, jossa aikaisemmin koetun perusteella tekoäly osaa ennustaa, miten uudessa tilanteessa tulisi toimia. Eli yksinkertaisuudessaan miten kone oppii. (Kananen; Puolitaival; Puntti & Metsola 2019.)

Tapa, jolla kone oppii, on hyvin ihmismäinen. Kun ihminen katsoo aamulla ikkunasta ulos ja näkee synkkiä sadepilviä, hän muistelee aikaisempia kokemuksiaan samanlaisesta säästä, mahdollisesti laittaa jalkaansa kumisaappaat ja nappaa mukaan sateenvarjon. Tekoälyn tapauksessa aikaisempana kokemuksena toimii saatu harjoitteludata ja sen pohjalta muodostetut algoritmit, eli ohjeet miten toimia. Ihmiset käyttävät perinteisen ohjelmoinnin algoritmeja tekoälyä kouluttaessa, jotta tekoäly saadaan oppimaan mahdollisimman tarkasti ja tehokkaasti. Yleisen käsitteen lisäksi koneoppimista voidaan pitää myös ”tieteenä”, joka käsittelee erilaisia koneen opettamisen tapoja. (Zhou 2016.)

Koneoppiminen on yleisimmin jaettu kolmeen päätyyliin: ohjattu oppiminen, ohjaamaton oppiminen ja vahvistusoppiminen. Näiden välillä valinta tehdään pääasiassa käytettävän datan perusteella, mutta vaikuttavina tekijöinä voivat olla myös esimerkiksi käyttötarkoitus ja fyysiset resurssit. Opetustavan lisäksi lopputulokseen vaikuttaa algoritmi, jolla kone annettua dataa käsittelee, pois lukien vahvistusoppiminen.

Ohjattu oppiminen ja ohjaamaton oppiminen ovat keskenään kavereita, mutta algoritmin luomissuunta menee niissä ikään kuin päinvastaisiin suuntiin. Ohjattu oppiminen on nimensä mukaan ihmisen ohjauksen avulla suoritettavaa oppimista ja sen vaatimuksena on hyvä data. Ohjatussa oppimisessa ihminen antaa

koneelle ohjeistuksena esimerkkitilanteen ja siihen vastauksen, jonka perusteella kone luo säännön, jota se voi myöhemmin käyttää, eli tuloksena on algoritmi. Kone siis yleistää ihmisen antaman esimerkkilogiikan toimimaan muuhunkin samankaltaiseen dataan. Otetaan esimerkiksi tilanne, jossa tekoälylle näytetään kuvia erilaisista autoista ja kuville on lisäksi tiedot auton merkistä ja teknisistä tiedoista. Kun ihminen näyttää esimerkin autosta, jossa on tähti keulassa ja merkitsee sen olevan Mercedes Benz -merkkinen, luo kone säännön, jonka perusteella se osaa tunnistaa muutkin samanmerkkiset autot. Ohjatussa oppimisessa on siis koneen päätelmien kannalta erittäin tärkeää datan laatu ja annettujen esimerkkien tarkkuus. Jos annetut esimerkit ovat virheellisiä, on myös koneen antamat arviot virheellisiä, vaikka itse vastaus olisikin ollut erittäin tarkka. (Kananen; Puolitaival; Puntti & Metsola 2019.)

Datan määrän ollessa suuri tai sen ollessa ”sekavaa”, voi tarkkojen esimerkkien ja vastauksien määrittäminen voi olla hyvinkin hidasta ja kallista, jopa mahdotonta. Ohjaamattoman oppimisessa lähtökohtana on hankalampi data, josta koneen tehtävä on etsiä säännönmukaisuuksia ja poikkeuksia, joita ihmisen on hankalaa tai jopa mahdotonta huomata ja joiden pohjalta sääntö sitten muodostuu. Kun selkeitä datasta etsittäviä asioita ei ole määritelty, on vastauksia vaikeampi arvostella ja täysin merkityksettömien säännönmukaisuuksien ja poikkeuksien löytäminen mahdollista.

Ohjatussa oppimisessa koneelle annetaan etukäteen selkeät ohjeet tehtävästä ja ohjaamattomassa oppimisessa annetaan koneen suorittaa tehtävä ja jälkikäteen arvioidaan tulokset. Kolmannessa koneoppimisen muodossa, vahvistusoppimisessa, voidaan nähdä piirteitä kummastakin aikaisemmasta tavasta. Vahvistusoppiminen toimii ikään kuin peli, jossa opetettavalle mallille rajataan ympäristö, eli pelin säännöt ja mallin tavoitteena on pelata peli läpi mahdollisimman tehokkaasti. Peli rakennetaan niin, että oikeasta suorituksesta malli saa pisteitä ja väärästä suorituksesta miinuspisteitä, joten se automaattisesti pyrkii löytämään parhaan mahdollisen reitin maaliin. Oppimistapa perustuu siis yritykseen ja erehdykseen sekä sitä kautta oppimisen ja tehokkaimman ratkaisun löy-

tämiseen. Ympäristön luominen mallille on hyvin työlästä ja vaatii suuren määrän osaamista ja ammattitaitoa sen kehittäjältä. Toisaalta opetukseen tarvitaan paljon vähemmän dataa kuin muissa koneoppimisen tavoissa. Vahvistusoppimisen vahvuutena on myös mahdollisuus ”oppia lennosta” ja ratkaisun kehittäminen omatoimisesti. Tosielämän tapauksissa hankaluudeksi voi ilmetä viive suorituksen ja palkinnon tai rangaistuksen saamisen välillä, jolloin on vaikea arvioida, kuinka hyvän suorituksen malli teki. (Kananen; Puolitaival; Puntti & Metsola 2019.)

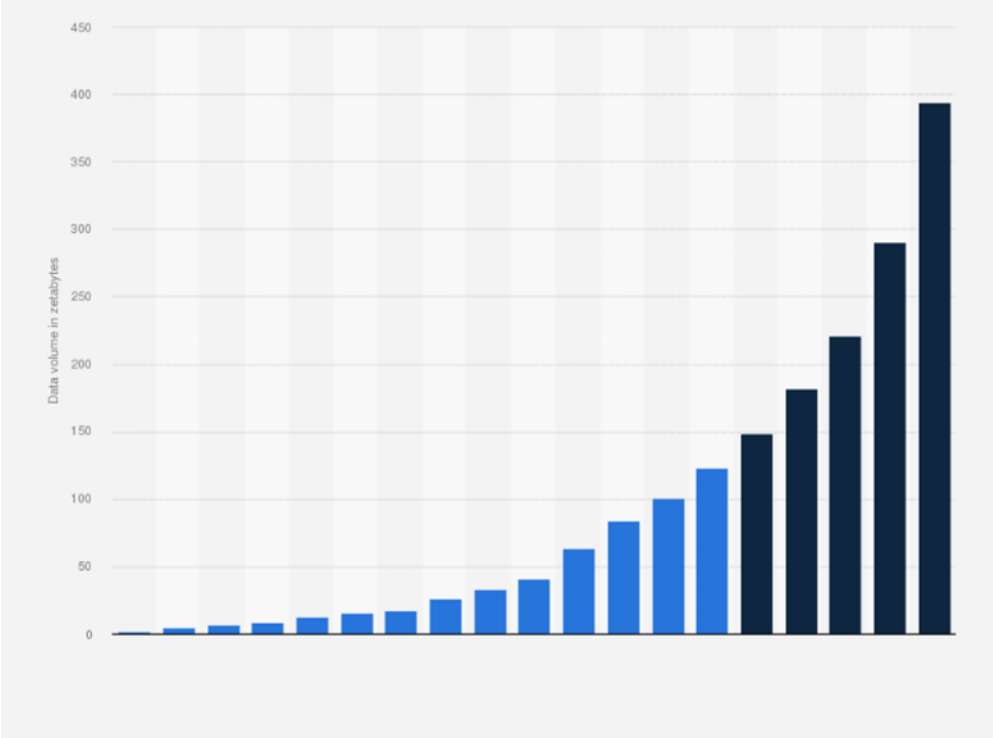
Koneoppimisessa puhutaan oppimisesta ja tekoälyä yleisesti käsiteltäessä älykkyydestä on kuitenkin tärkeää pitää mielessä, ettei tekoäly todellisesti kehity älykkääksi, vaikka kuinka tehokkaasti sitä kouluttaisi. Tekoäly toimii edelleen matemaattisten kaavojen pohjalta ja pyrkii tuottamaan mahdollisimman todennäköisiä vastauksia, eikä se ymmärrä käsiteltävää asiaa, kontekstia tai yhteyksiä muihin asioihin. Tekoälyä kouluttamalla siitä saadaan siis tarkempi sekä nopeampi, mutta ei älykkäämpi ja tarkankin mallin vastauksien oikeellisuus riippuu käytettävästä datasta.

3.3 Kehitys

Tekoälyn kehitys ei ole ollut suoraviivaista, vaan se on koostunut enemmänkin luonnollisista ylä- ja alamäistä. Kehitystä ovat vauhdittaneet uudet innovaatiot ja niitä seurannut innostus ja sijoitukset, kun taas jarruttavina tekijöinä ovat mahdollisesti olleet lunastamatta jääneet odotukset tai muutokset yleisessä maailmantilanteessa. Vaikka tekoälyteknologiaa ja tutkimusta on ollut jo useamman vuosikymmenen ajan, on kehitys ottanut erityisesti viime vuosina valtavia harppauksia ja tuonut tekoälykeskustelun myös valtaväestön huulille. Pääasiallisina mahdollistajina ovat olleet laitteiston kehitys, datan määrän kasvu ja tekoälyn saavutettavuus. (Kananen; Puolitaival; Puntti & Metsola 2019.)

Mitä taitavampi tekoälymalli, sen monimutkaisempi se on. Mitä monimutkaisempi malli, sitä enemmän laskentatehoa sen toiminta vaatii. Fyysisen laitteiston kehitys on ollut edellytyksenä modernien tekoälymallien kehittämisessä, opettamisessa ja toiminnassa. Laitteiston kehityksestä esimerkin antaa teknologiajätti Nvidian prosessorin kehitys, jossa vuonna 2012 li mallin teho oli 3 terafloppia, joka vastaa triljoonaa laskutoimitusta sekunnissa. Vuoden 2020 mallin tehon oli yli neljäkymmentä terafloppia. (Ojanperä 2023.)

Toisena tekoälyn kulmakivenä toimii data, kun tekoälymallien koulutus tapahtuu valtavilla datamäärillä. Aikaisemmin tarvittavia datamääriä ei ollut saatavilla, mutta nykyinen mobiililaitteiden ympärillä pyörivä elämämme on räjäyttänyt maailmassa olevan datan määrän aivan uusiin ulottuvuuksiin. Tätä havainnollistaa oheinen kuva 2, joka kuvaa maailmassa olevan datan määrän kasvua ja arviota sen kasvusta lähitulevaisuudessa.



Kuva 2. Datan määrän kasvu maailmassa, 2010–2028. (Statista)

Tekoäly on myös nykyään lähempänä ihmisiä kuin koskaan ennen. Sen kyky käsitellä ja tuottaa luonnollista kieltä on mullistanut tekoälyn saatavuuden ja

mahdollisuuden sen hyödyntämiseen lähes jokaiselle. Myös internetistä löytyvää tekoälystä kertovaa tietoa on hyvin saatavilla. Monenlaisia tekoälykursseja ja vapaasti tarjolla olevia tekoälysovelluksia on saatavissa enemmän ja enemmän. (Kananen; Puolitaival; Puntti & Metsola 2019.)

Tekoälyä pidetään siis yleisesti koneellisena tekniikkana, joka kykenee ihmismäisiin älyllisiin toimenpiteisiin. Keskeisiä perusteita sen toiminnassa on laskentatehoa tuottavat fyysiset laitteistot, käytettävissä oleva data sekä tekoälyn oppimiseen ja toimintalogiikkaan vaikuttava koneoppiminen. Tekoälystä on ajan mittaan kehitetty erilaisia suuntia vastaamaan erilaisiin tarpeisiin. Yksi näistä muodoista on generatiivinen tekoäly, jonka toimintamalli on erilainen perinteiseen tekoälyyn verrattuna, vaikka nämä arkikielessä monesti helposti sekoittuvat. Seuraavassa luvussa on generatiivisen tekoälyn toimintaperiaatteita, sen mahdollistamia käytännön sovelluksia sekä sen aiheuttamia haasteita.

4 Generatiivinen tekoäly

Generatiivinen tekoäly tai luova tekoäly on tekoälyn muoto, joka kykenee tuottamaan täysin uutta sisältöä aikaisemmin oppimiensa asioiden perusteella. Generatiivinen tekoälyn kanssa on mahdollista keskustella, se osaa käyttää sille annettua aineistoa vastatessaan käyttäjän sille antamaan tehtävään. Kommunikaatio generatiivisen tekoälyn ja ihmisen välillä toimii yleensä ihmiselle luonnollisella kielellä, joko kirjoittaen ja puhuen. Riippuen tekoälymallin kyvykkyyksistä on tehtävänannossa sekä vastauksessa mahdollista käyttää myös kuvia, videoita, ääntä tai ohjelmointikieltä. Käytän tässä työssä generatiivisesta tekoälystä välillä yksinkertaisuuden vuoksi myös termiä tekoäly. (IBM n.d.)

Perinteisellä tekoälyllä pystytään ratkaisemaan lähinnä sääntöpohjaisia ongelmia ja tehtäviä, joilla on selkeät raamit. Perinteisen tekoälyn kanssa käytetään esimerkiksi luokittelua, ryhmittelyä ja regressiota keinoina merkittävyyksien löytämiseksi. Lähtökohtaisesti ihmisen määrittelemät algoritmit työstävät dataa ja ovat erittäin tehokkaita oikein käytettynä. Luova tekoäly on sen sijaan ikään kuin itsenäisempi kokonaisuus, jolla on ihmismäisempiä kykyjä, kuten kyky kommunikoida ihmisen kanssa. Luovaa tekoälyä ei kuitenkaan ole sekoitettava vain puhuvaan koneeseen, vaan kyseessä on toimija, joka kykenee esimerkiksi kirjoittamaan runon lempirunoilijan tyylillä ja halutusta aiheesta. Käydään seuraavaksi läpi generatiivisen tekoälyn kehityksen kulmakiviä, niin saamme paremman käsityksen sen ominaisuuksista ja kyvykkyyksistä. (SAP n.d.; Kananen; Puolitaival; Puntti & Metsola 2019.)

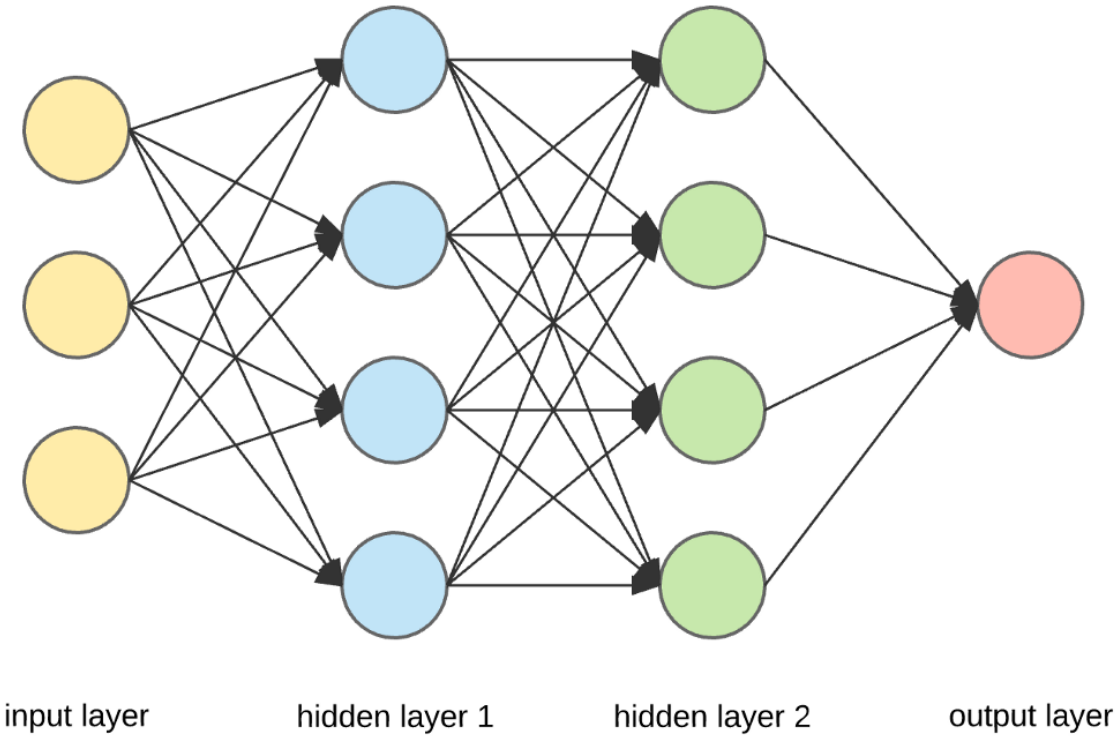
4.1 Neuroverkot

Ihmisen aivot ovat äärimmäisen tehokkaat ja kykenevät nopeisiin johtopäätöksiin ja arvioihin, erityisesti näköaistin kanssa yhteistyötä tehdessä. Ei siis ole ihme, että ihmisen aivojen toimintaperiaatteita jäljittelevää tekniikkaa on pyritty

kehittämään koko tekoälykehityksen alkuajoista lähtien. Koneoppimisessa käytetään paljon erilaisia algoritmeja, jonka perusteella tekoäly käsittelee sille annettua dataa ja sitä kautta tuottaa vastauksia. Yhtä näistä tekniikoista käytetään monimutkaisen ja monitasoisen datan kanssa, ja se perustuu ihmisen aivojen lailla toimiviin neuroverkkoihin. Neuroverkkoihin pohjautuvaa tapaa kutsutaan usein syväoppimiseksi, joka kuvaa hyvin monimutkaisempaa ja syvällisempää asioiden käsittelyä. (Dongare, A.D.; Kharde, R.R.; Kachare, Amit D 2012.)

Neuroverkot muodostuvat nimensä mukaisesti erillisistä tietoa käsittelevistä yksiköistä, ja ne kytkeytyvät toisiinsa monisäikeisesti muodostaen verkkomaisen kokonaisuuden. Tietoa käsittelevää yksikköä kutsutaan neuroniksi, ja se saa verkkoa pitkin tietoa yhdeltä tai useammalta siihen yhteydessä olevalta neuronilta. Saatujen tietojen perusteella se laskee itselleen lähtöarvon. Jokaisen neuronin tehtävä on suorittaa tietty matemaattinen tehtävä, nimeltään aktivaatiofunktio, jonka vastaus määrittää, tuleeko neuronin lähettää tietoa eteenpäin vai ei. Funktion laskemisessa neuroni käyttää saamaansa lähtöarvoa ja lopputuloksena neuroni aktivoituu tai ei aktivoidu. Aktivoituessaan neuroni lähettää tiedon verkkoa pitkin eteenpäin seuraaville siihen yhteydessä oleville neuroneille. Neuronien väliseen tiedonkulkuun vaikuttaa myös niiden yhteyksille määritellyt painokertoimet. Jokaisella verkon ”säikeellä”, joka yhdistää kaksi neuronia toisiinsa, on määritelty kerroin. Mitä suurempi kerroin on, sitä merkityksellisempi se on ja sitä suurempi vaikutus sillä on lopputulokseen. Jokaisen neuroniparin välinen painokerroin vaikuttaa siis vastaanottavan neuronin lähtöarvon laskemiseen. (Kananen; Puolitaival; Puntti & Metsola 2019.)

Kuvassa 3 yksinkertainen esimerkki neuroneista ja niiden välisistä yhteyksistä. Esimerkkiverkkoon kuuluu sisäänottokerros, kaksi tietoa käsittelevää piilotettua kerrosta sekä vastauksen antava ulostulokerros. Todellisuudessa neuroverkot ovat paljon monimutkaisempia ja on olemassa eri lailla järjesteltyjä verkostoja, joiden suunnittelua kutsutaan neuroverkkoarkkitehtuureiksi. Yksinkertainen neuroverkko ei ole mullistavan vahva, mutta kun useat neuroverkkojen kerrokset yhdistetään kokonaisuudeksi, se mahdollistaa erittäin tehokkaan tiedonkäsittelyn.



Kuva 3 Esimerkki neuroverkosta (Roboflow, 2025)

Neuronien väliset yhteydet, tarkemmin ottaen painokertoimet, ovat ratkaisevassa roolissa neuroverkon opetuksessa. Aluksi painokertoimilla on satunnaiset arvot, mutta opetusdatalla harjoitellessa painokertoimia muutellaan, kunnes tulokset ovat halutunlaisia. Painokertoimet, joita joskus myös parametreiksi kutsutaan, ovat siis kuin tekoälyn aivot, joita voidaan jatkuvasti päivittää koulutuksen aikana. (Kananen; Puolitaival; Puntti & Metsola 2019.)

Neuroverkoissa on siis usein hyvin monia neuronien kerroksia ikään kuin piilotettuna verkon sisään. Verkko käsittelee ja muokkaa syötettyä dataa kerros kerrokselta seuraavan tason aina hyödyntäessä edellisen kerroksen oppeja. Lopullisen tuloksen antaa verkon viimeinen osa eli ulostulokerros, joka perustuu kaikkien piilotettujen kerroksien käsittelemään lopputulokseen. Mitä enemmän neuroverkossa on kerroksia, sitä enemmän siinä on tietoa käsitteleviä parametreja. Mitä enemmän parametrejä tietoa käsittelemässä on, sitä tehokkaammin malli toimii. Näiden neuroverkon ”aivojen” määrän ja samalla laskemistehon massiivisesta muutoksesta hyvän esimerkin antaa tunnetun tekoälysovelluksen Chat-

GPT:n taustalla olevan kielimallin parametrien kasvu. 2018 julkaistussa versiossa kielimallissa operoi 117 miljoonaa parametria, seuraavana vuonna 1,5 miljardia parametria ja vuoden 2020 mallissa tietoa käsitteli 175 miljardia parametria. (Ojanperä, 2023)

Mitä suurempia ja monimutkaisempia kokonaisuuksia tekoälylle antaa tehtäväksi, sen hitaampaa ja työläänpää sen käsittely ja tulkitseminen luonnollisesti on. Vuonna 2017 Googlen tutkijat julkaisivat syväoppimisen mallin, joka toi ratkaisun tähän ongelmaan ja mahdollisti entistä hankalampien ja laajempien tehtävien suorittamisen tekoälyn avulla. Generative pretrained transformer eli tutummin GPT on siis neuroverkkomalli, jonka toimintalogiikka mahdollistaa esimerkiksi luonnollisen kielen, eli ihmisten puhuvan kielen tehokkaan käsittelyn. Mallin nimen muodostavat sanat antavat kuvan sen ominaisuuksista. Generative kertoo mallin kyvystä tuottaa täysin uutta sisältöä. Pretrained viittaa siihen, että gpt-mallit ovat valmiiksi hyvin koulutettuja suurilla datamäärillä ja tehokkaan tiedonkäsittelynsä ansiosta kykenevät tähän nopeammin kuin aikaisemmat neuroverkkomallit. Transformer kuvaa sen muuntautuvuutta, joka juontaa juurensa mallin ominaisuuteen, jossa sen ”aivot” eli painokertoimet päivitetään joka kerta, kun mallia koulutetaan. Gpt-mallin tiedonkäsittelyn tehokkuus perustuu tekniikkaan, jolla se käsittelee sen vastaanottamaa dataa. Malli jakaa saamansa tiedon osiin, kykenee löytämään merkityksellisimmät kohdat ja priorisoi datan käsittelyssä näitä tärkeämpiä osia. Samaan tapaan ihminen pystyy keskittämään huomionsa kerrallaan vain hyvin rajalliseen asiaan, mutta tärkeimmät osat ymmärrettyää päättelemään kokonaisuuden merkityksen. Gpt-neuroverkkomallin kehittäminen on tärkeässä roolissa luonnollisen kielen käsittelyssä, joka on ollut valtava harppaus tekoälykehitykselle ja tuonut tekoälyn lähemmäksi normaalia ihmistä ChatGPT:n ja muiden vastaavien tekoälysovellusten muodossa. (IBM n.d.; Ojanperä 2023.)

Neuroverkkojen vahvuutena on niiden kyky käsitellä asioita kokonaisvaltaisemmin, mikä mahdollistaa monimutkaisempien kokonaisuuksien käsittelyn. Ne myös pystyvät käsittelemään hyvin suurta määrää dataa, ja data voi olla koh-

tuullisen sekavaa neuroverkon silti kyetessään sen käsittelyyn itsenäisesti. Toisaalta ihmiselle vaikeasti hahmotettavissa oleva data ja sen suuri määrä johtaa myös siihen, että ihmisen voi olla vaikea ymmärtää perusteita tekoälyn tarjoamille päätelmille. Tekoäly saattaa myös oppia datasta hyödyttömiä asioita, sen tehdessä oppimista itsenäisesti. Neuroverkkojen lähtökohtaisesti tarvitsema hyvin suuri oppimiseen tarvittava datamäärä vaatii myös suuria määriä tietokoneen laskentatehoa. Kyky käsitellä suuria ja monimutkaisia konsepteja mahdollistaa neuroverkkojen avulla käyttötarkoituksia, jotka ei muilla tekoälyn tekniikoilla onnistuisi. Esimerkkejä neuroverkkojen käytännönsovelluksista ovat kuvien käsittely, niiden luominen ja luonnollisen kielen käsittely eli NLP (Natural language processing). (Kananen; Puolitaival; Puntti & Metsola 2019.)

4.2 Luonnollisen kielen käsittely

Luonnollisen kielen käsittely, NLP (natural language processing) on päässyt nauttimaan viimeaikaisesta teknologisesta kehityksestä ja on isossa roolissa edistänyt tekoälyn lähentymistä osaksi hyvin monen ihmisen elämää. Luonnollisella kielellä tarkoitetaan yleisesti ihmisten ymmärtämää ja puhumaa kieltä ja luonnollisen kielen käsittelyllä viitataan lähtökohtaisesti koneiden kykyyn kommunikoida ihmisen kanssa, kun käytetään ihmiselle luonnollista kieltä, eikä esimerkiksi ohjelmointikieltä. Luonnollisen kielen käsittelystä puhuttaessa viitataan lähtökohtaisesti joko tieteeseen, jolla on tavoitteena ihmisen ja koneen välisen kommunikaation kehittäminen, tai teknologiaan ja sen kehitykseen, joka pyrkii mahdollistamaan tämän kommunikaation. Ihmisen käyttämää kieltä on niin kirjallisessa kuin puhutussakin muodossa, joten luonnollisen kielen käsittelyyn kuuluu näiden molempien vastaanotto, koneelle käsiteltävään muotoon muuttaminen ja luonnollisen kielen tuottaminen. (Kananen; Puolitaival; Puntti & Metsola 2019.)

Kun katsotaan luonnollisen kielen käsittelyä tieteellisestä näkökulmasta, on se yksinkertaisesti ihmisen kielen ymmärtämistä ja ikään kuin tämän mahdollistamista koneille. Se hyödyntää useita eri tieteenaloja, kuten erityisesti yleistä kielitiedettä ja tietojenkäsittelytiedettä, mutta myös esimerkiksi matematiikkaa, tekoälytiedettä ja psykologiaa. Ihmisen käyttämää kieltä on tutkittu jo satoja vuosia, ja nykypäivänä se käsittelee kaikkea kielellisistä lainalaisuuksista, sen oppimiseen, muuttumiseen ja vaikutuksia yhteiskuntaan. (Helsingin Yliopisto n.d.) Tietojenkäsittelytieteen näkökulmasta puhutaan taas ohjelmistoista ja niiden kehittämisestä, laitteistosta, jolla näitä käytetään sekä niiden yhteistyöstä keskenään. (Michigan Tech University n.d.)

Teknologisen kehityksen näkökulmasta katseltuna luonnollisen kielen käsittelyssä on tapahtunut viime vuosina hyvin merkittäviä muutoksia. Ihmisten käyttämä kieli on hyvin monimutkaista, joten sen käsitteleminen ja tuottaminen vaatii tehokkaita algoritmeja ja paljon tietokonetehoa. Erityisesti läpimurrot erilaisten neuroverkkojen osalta, kuten recurrent neural networks (RNN), long short-term memory networks (LSTM) ja erityisesti aikaisemmin mainittu generative pretrained networks (GPT) ovat innovaatioita, jotka ovat suuressa roolissa mahdollistamassa luonnollisen kielen käsittelyä. (Aditya; Gandhar; & Vraj 2018.)

Luonnollisen kielen käsittelyyn kuuluu useita osa-alueita, kuten puheen tunnistus ja sen tuottaminen, jota esimerkiksi älykkäät kodinelektroniikkaratkaisut jo hyödyntävät. Myös käännössovellusten käyttämä tekstin kääntäminen ja prosessointi hyödyntävät näitä tekniikoita. Jotta yksikään näistä käyttötarkoituksista olisi mahdollinen, tulee koneen kyky ”ymmärtää” luonnollisen kielen kielioppisääntöjä ja lainalaisuuksia olla riittävän hyvällä tasolla. Kaikkien kielioppisääntöjen ja keskenään samankaltaisten sanojen ohjelmoiminen on ollut perinteisiä koneoppimisen keinoja käyttäen käytännössä mahdotonta. Syväoppimisen menetelmillä on sen sijaan pystytty muuttamaan kieli koneelle helpommin käsiteltävään muotoon ja tilastotiedettä hyödyntäen löytämään vastaukseen todennäköisesti oikeat sanat, oikeaan järjestykseen ja oikeassa muodossa. (Aditya; Gandhar; & Vraj 2018.) Kone osaa siis aikaisemmin oppineensa avulla arvata, mitä sanoja käyttää ja missä järjestyksessä, itse kielioppisääntöjä osaamatta. Kun

opetusmateriaalia käytetään riittävästi ja kone oppii tehokkaasti, syntyy vaikutelma siitä, että kone osaisi puhua ihmisen tavoin.

Vaikka generatiivisen tekoälyn kielimallit pystyvätkin käsittelemään luonnollisen kielen käsittelyn avulla monia ihmisten puhuvia kieliä, kuvia ja koodauskieliä, itse tiedonkäsittelyn ne hoitavat niiden omalla ”kielellään”. Tekoälymallit jakavat saamansa informaation pienemmiksi osiksi, joita kutsutaan tokeneiksi. Esimerkiksi sana ”tietokone” voitaisiin jakaa kahteen osaan tieto ja kone. Jotkin lyhyemmät sanat kuvataan yhdellä tokenilla, mutta yleensä sanat jaetaan vähintään kahteen tokeniin. Koulutuksen aikana malli oppii tokenien välisistä yhteyksistä, jonka avulla sitten kykenee ennustamaan oikeanlaisia vastauksia. Sana ”kuusi” on tekoälyllä muistissa hyvin mahdollisesti yhtenä tokenina, jolloin sen pitää kokemuksensa sekä muiden lähellä esiintyvien sanojen perusteella erottaa, onko kyseessä numero kuusi vai metsään kuuluva havupuu. (Nvidia 2025.)

Tokenisointiin on monia tapoja ja toisten mallien tokenisaatio on tehokkaampaa kuin toisten. Suuntaa antavaksi arvioksi, englannin kieltä käsiteltäessä sanotaan usein yhden sanan olevan keskimäärin 0,75 tokenia. Tekoälyn muuttaessa saamansa tiedon omalle kielelleen, mahdollistaa se vastauksen antamisen eri kielellä käyttäjän toiveiden mukaan. Malleilla on eri kokoisia rajoituksia siihen, kuinka monta tokenia ne pystyvät kerralla käsittelemään ja rajoitus vaikuttaa hyvin paljon siihen, minkälaisia tehtäviä mallille voi antaa. Joidenkin mallien pystyessä käsittelemään yksittäisiä kuvia tai muutamia sivuja tekstiä, kykenevät toiset mallit käsittelemän kerralla kokonaisia kirjasarjoja tai valtavia tietokantoja. (Nvidia 2025.)

Vaikka koneellinen luonnollisen kielen käsittely on nykyään hyvin tehokasta ja tarjoaa monenlaisia mahdollisuuksia hyödynnettäväksi, on kuitenkin hyvä muistaa, ettei se todennäköisyyksiin perustuessaan ole täydellistä. Ihmisen arviointikyky on tärkeässä roolissa ihmisten ja koneiden välisen keskustelun yhteydessä ja paras lopputulos on saavutettavissa, kun pitää mielessä kummankin osapuolen vahvuudet. Kone on äärimmäisen tehokas käsittelemään suuria tietomääriä

hyvin ajassa, tämän ollessa ihmiselle mahdotonta. Tekstin todellinen ymmärtäminen, kuten syy-seuraussuhteet ja kaksoismerkitykset taas ovat ihmiselle luonnollista osa-aluetta, mutta juuri tässä on koneella ainakin tällä hetkellä vielä harjoiteltavaa. (Kananen; Puolitaival; Puntti & Metsola 2019.)

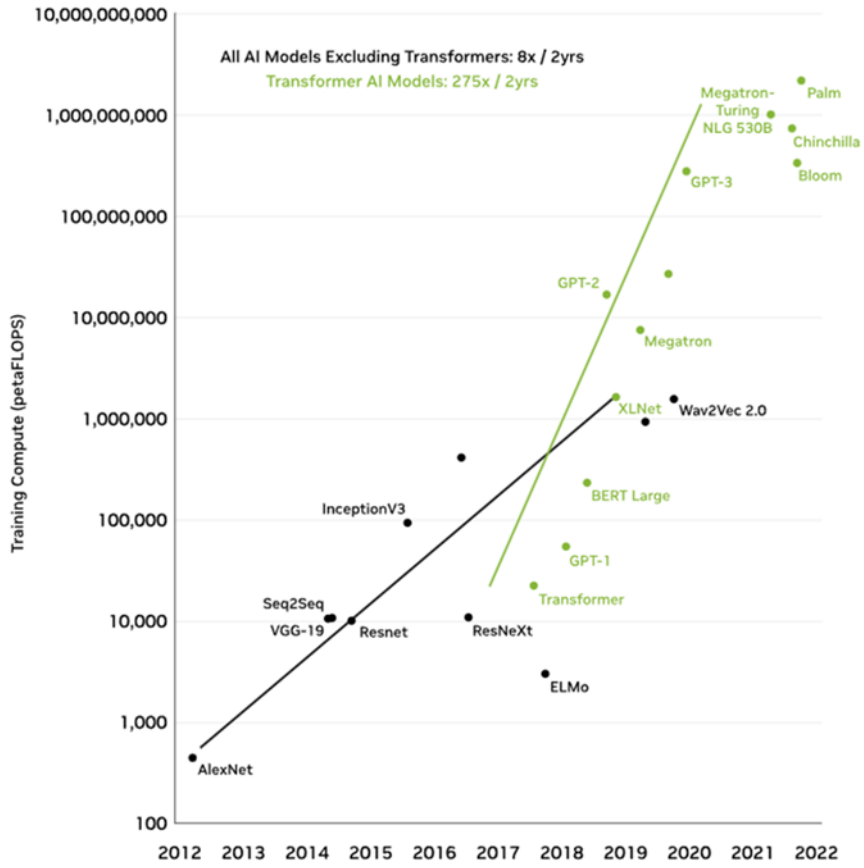
Luonnollisen kielen käsittely helpottaa ihmisen ja koneen välistä kommunikointia ja mahdollistaa myös muiden tietomuotojen käytön. Yksi merkittävimmistä luonnollisen kielen käsittelyä hyödyntävistä tekoälyn osista on suuret kielimallit. Seuraavassa kappaleessa kerrotaan suurista kielimalleista, jotka toimivat käytännössä koko generatiivisen tekoälyn perustana ja joiden toiminnan mahdollistajana luonnollisen kielen käsittely toimii.

4.3 Suuret kielimallit

Suuret kielimallit ovat eräänlaisia generatiivisen tekoälyn pohjamalleja, jotka kykenevät monenlaisiin tehtäviin, kuten muun muassa tunnistamaan, käsittelemään, tuottamaan ja ennustamaan, niille annetun syötteen perusteella. Kielimalleissa yhdistyvä teknologia ja valtava koulutukseen käytetty datamäärä tekevät niistä ikään kuin yleispäteviä moottoreita, joita sitten voidaan jatkojalostaa sovelluksiksi spesifimpiin käyttötarkoituksiin. Suuria kielimalleja kutsutaan monesti myös yksinkertaisemmin kielimalleiksi tai perusmalleiksi. (Ojanperä 2023.)

Suuret kielimallit ovat tehokkaita hyvin suurien tietomäärien käsittelyssä. Niiden kouluttamiseen käytetään massiivisia datamääriä, jotka sisältävät usein eri ihmisten kieliä, ohjelmointikieliä, kuvia ja videoita. Kielimallit kykenevät myös käsittelemään niille annettuja suuria datamääriä nopeassa ajassa ja hyödyntämään oppimaansa vastausta antaessaan. Jotta tehokas datan käsittely olisi mahdollista, hyödyntävät kielimallit hyödyntävät muun muassa transformer-neuroverkkoarkkitehtuuria. Tekniikan avulla kielimalli jakaa syötteen pienempiin osiin, tunnistaa sen merkittävimmät alueet ja useamman pienemmän ratkaisun pohjalta arvioi todennäköisen vastauksen koko tehtävälle. (Nvidia n.d.)

Arkkitehtuurin lisäksi edistysaskeleet luonnollisen kielen käsittelyssä ovat kriittisenä osana suurien kielimallien toimintaa. Taitava kyky käsitellä luonnollista kieltä antaa mahdollisuuden tehdä paljon erilaisia kielellisiä tehtäviä, kuten kääntämistä, päättelyä, tiivistämistä ja yleistiedon hallintaa. Luonnollisen kielen käsittelyn toiminnot eivät jää kielimalleilla pelkästään tekstin käsittelyyn, vaan nykyään kielimallit ovat jatkuvasti taitavampia erilaisen datan, kuten kuvien, musiikin, videon ja 3D-mallien kanssa. Osaaminen eri datatyyppien kanssa toimimisesta tuo kielimalleille valtavasti uusia käyttömahdollisuuksia. Tekoälymallin erilaisten tietomuotojen käsittelyn kyvykkyyksistä puhuttaessa käytetään usein sanaa multimodaalisuus. Mallien multimodaaliset taidot vaihtelevat hyvin paljon, lähes kaikkien silti osatessa vähintäänkin joitakin ihmisten puhumia kieliä sekä jotain ohjelmointikieltä. (Ojanperä 2023.)



Kuva 4 Transformer -mallien vaatima koulutusteho muihin malleihin verrattuna. (Nvidia, 2025)

Kielimallien laajat ominaisuudet eivät kuitenkaan aivan ilmaiseksi synny. Valtavien ja monimutkaisten mallien koulutus vaatii suuret määrät tietokonetehoa ja aikaa, jotka tarkoittavat myös suurta hintalappua. Tietokoneprosessorijätti Nvidian arvioiden mukaan kielimallin koulutus vie ajallisesti viikkoja tai jopa kuukausia ja jo vuonna 2020 julkaistun GPT 3-mallin yksittäisen koulutuskierroksen arvioidaan maksavan jopa 12 miljoonaa dollaria. Kuvassa 4 näkyy uudempien gpt-arkkitehtuuria käyttävien mallien koulutukseen tarvittavan tietokonetehon eri aikaisempiin malleihin verrattuna. Myös tarpeeksi suuren, koulutukseen tarvittavan datamäärän hankkiminen ei ole aivan jokaiselle yritykselle saavutettavissa. Datan määrä ei myöskään kaikkea ratkaise, vaan sen tulisi olla tarpeeksi laadukasta ja aiheellisesti sopivaa. Lisäksi monimutkaisuutensa ja laajuutensa takia, kielimallien rakentaminen sekä koulutus vaatii paljon hyvin kyvykkäitä ja osaavia ihmisiä. (Nvidia n.d.)

Suurien kielimallien kehityksen kärjessä kamppailevat lähtökohtaisesti valtavat teknologiayhtiöt, kuten OpenAi, Google ja Microsoft. Kilpailijat ovat pyrkineet jatkuvasti kehittämään kielimalliaan tehokkaammaksi ja monipuolisemmaksi, julkaisten vuosi toisensa jälkeen uuden version. Kaikessa yksittäisen mallin on kuitenkin vaikeaa olla, joten erikoistumista spesifimpiin asioihin mallien välillä myös löytyy, kuten esimerkiksi kuvien tuottamiseen. Muutosta markkinatilanteeseen voi kuitenkin olla tulossa, kun vuoden 2024 lopussa kiinalaisyrityksen Hangzhou DeepSeek Artificial Intelligence Co., Ltd aiheutti reaktion, joka näkyi huomattavasti pörssikursseissa asti. Yhtiön julkaiseman DeepSeek-kielimallin V3 version tuotantokustannusten kerrotaan olevan minimaaliset verrattuna amerikkalaisten teknologiajättien versioihin, mutta suorituskyvyssä DeepSeekin kerrotaan kilpailevan samassa sarjassa markkinoiden parhaiden mallien kanssa. Mielenkiintoa tilanteeseen lisää myös se, että DeepSeek julkaistiin niin sanottuna avoimen lähdekoodin mallina, mikä mahdollistaa käytännössä kenelle tahansa DeepSeekin innovaatioiden hyödyntämisen. (Ojanperä 2023; Yle 2025.)

Suurista kielimalleista useimmat ovat niin sanottuja suljettuja malleja, eli valmistaja ei paljasta tekniikkaa, jota malli sisäänsä kätkee, ja malliin pääsee käsiksi vain rajoitetusti tiettyjen käyttöliittymien avulla. Esimerkkinä mahdollisesti tunnetuimman tekoäly-yhtiön OpenAi:n mallit ovat suljettuja. Vaihtoehtoisesti osa yrityksistä paljastaa mallinsa salaisuudet rakenteineen ja painokertoimineen, antaa kenelle tahansa mahdollisuuden hyödyntää mallin teknologiaa ja kehittää sitä edelleen, jolloin niitä sanotaan avoimen lähdekoodin malleiksi. Avoimena julkaiseminen tuo mallille mielenkiintoa juuri vapaiden jatkokehitysmahdollisuuksien muodossa. Se antaa myös pienemmille yrityksille mahdollisuuden tuoda oman panostuksensa tekoälynkehitykseen ilman valtavan yrityksen budjettia, pitää samalla myös markkinatilannetta tasaisempana, hankaloittaa isojen yritysten teknologista karkaamista muiden ulottumattomiin. Hankalasti valvottavasta ja erittäin tehokkaasta teknologiasta puhuttaessa pelikentän tasapainottaminen kuulostaa hyvinkin positiiviselta, mutta antaa samalla myös avaimet näihin teknologioihin esimerkiksi rikollista tai muuten negatiivista tarkoitusta ajaville tahoille. (Ojanperä 2023.)

Perustelevat mallit, englanniksi reasoning models, ovat uudenlaisia suuria kielimalleja, jotka on koulutettu erityisesti monimutkaisia, ajattelua ja perustelua vaativia tehtäviä varten. Suoria vastauksia antavista perinteisistä malleista poiketen, perustelevat mallit käyvät vastausta tuottaessaan ikään kuin pohdintaa vastaavan prosessin läpi sekä kykenevät vastausta antaessaan esittämään tämän prosessin sisällön, itse tehtävään vastaamisen lisäksi. Vastaukseen johtaneen prosessin näyttäminen antaa suoraan lisäarvoa käyttäjälle ja tuo toivottua läpinäkyvyyttä tekoälyn toimintaan. Perustelevat mallit toimivat erityisen hyvin ongelman ratkaisua, koodausta, suunnittelua ja tieteellistä perustelua vaativissa tehtävissä. (Open AI n.d.)

Perustelevat mallit on suunniteltu erityisesti monimutkaisten ongelmien ratkaisuun, joita mallit hyvin ihmismäisesti jakavat pienempiin osiin ja niitä ratkaisemalla tuottavat kattavia vastauksia. Perustellessaan vastauksiaan malli selittää prosessiaan, jossa se jakaa tehtävän osiin ja yrittää useita eri lähestymistapoja. Näistä se valitsee parhaat ja hylkää huonommat sekä lopulta esittää mielestään

parhaan mahdollisen ratkaisun. Vastatessaan perustelevat mallit käyttävät vastaanoton ja syötteen lisäksi ”ylimääräisiä” tokeneita itse perustelun tuottamiseen. Perinteisiin kielimalleihin verrattuna perustelevat mallit ovat usein hitaampia, enemmän tietokonetehoa vaativia, omaa ajatteluprosessiaan selittäviä ja monimutkaisiin ongelmiin soveltuvia. Toki muitakin kielimalleja voi pyytää halutessaan perustelemaan vastauksiaan, mutta ei yhtä tehokkaasti kuin perustelevat mallit, joihin kyky on sisäänrakennettuna. (Nvidia 2025.)

4.4 Haasteita

Vaikka generatiivinen tekoäly tehokas nykyään onkin, on sillä silti myös omat heikkoutensa. Aikaisemmin jo mainitut mallien monimutkaisuus ja valtava koulutukseen käytettävän datan tarve vaativat tekoälykehittäjiltä erittäin suuria resursseja, niin laadukkaana datana, osaavina ihmisinä, kuin suurina rahamäärinäkin. Generatiivisella tekoälyllä on sen jokaisen käyttäjän toimintaan vaikuttavia piirteitä, jotka ovat vähintäänkin hyvä tiedostaa ja huomioida tekoäly käyttäessä.

On tärkeää muistaa, että tekoälyn tuottamat vastaukset pohjautuvat täysin sen käyttämään harjoitteludataa. Ihmisellä on myös lähtökohtaisesti olematon ymmärrys ja kontrolli logiikasta, jolla tekoäly vastauksensa tuottaa. Tämä johtaa tilanteeseen, jossa kärjistetysti sanottuna hyvä harjoitusdata johtaa hyviin vastauksiin ja huono harjoitusdata huonoihin vastauksiin. Yleisimpiä todellisia ongelmia ovat esimerkiksi hallusinointi, jossa tekoäly tuottaa yksinkertaisesti täysin tai osittain keksittyä ja virheellistä tietoa, joka aiheuttaa ongelmia erityisesti faktapohjaisista aiheista kysyttäessä. Toinen yleinen ongelma on tekoälyn puolueellisuus, englanniksi bias. Rasistisen, toksisen tai tiettyä näkökulmaa suosivan materiaalin tuottaminen tekoälyltä on epätoivottua, mutta täysin mahdollista. Jos mallin saama harjoitusdata tai osa siitä on jollain tavalla puolueellista, on myös mahdollisuus, että vastauksissakin voi näkyä samanlaisia merkkejä. Dataan liittyvien ongelmien lisäksi on hyvä muistaa, että tekoälyn vastaukset

pohjatuvat todennäköisyyksiin. Samaankin kysymykseen annetut vastaukset voivat vaihdella, ja logiikan ymmärtäminen on käytännössä mahdotonta. Kaikkiin näihin ongelmiin pyritään jatkuvasti kehittämään ratkaisuja, mutta vastuu on hyvin paljon myös käyttäjällä. Vaikka generatiivinen tekoäly tuottaakin tehokkaasti materiaalia, ei siihen tule sokeasti luottaa ja käyttäjän tuleekin aina kriittisesti arvioida saamiaan vastauksia. (IBM n.d.; Ojanperä 2023.)

Generatiivisella tekoälyllä on kyky käyttää hyväksi massiivista tietopankkiaan ja monesti myös lähteitä internetistä sekä niiden pohjalta tuottaa vastauksia käyttäjän antamiin tehtäviin on tuonut esiin hankalia eettisiä ongelmia. Tekoäly on nykyään lähtökohtaisesti kaikkien saatavilla ja sen käyttö on äärimmäisen helppoa. Tämän takia generoivan tekoälyn käyttö suoraan rikollisissa tarkoituksissä on mahdollista. Esimerkiksi erilaisiin indentiteettivarkauksiin tekoäly kykenee muokkaamaan ja tuottamaan halutunlaista ääntä, kuvia ja videomateriaalia, jotka jäljittelevät tiettyä henkilöä tai viranomaista. Tekoäly voi myös auttaa muiden rikosten suunnittelussa sekä toteuttamisessa ja tätä käyttöä on hyvin vaikeaa valvoa. (SAP n.d.; Ojanperä 2023.)

Osaan tekoälyyn liittyvistä ongelmista pyritään vastaamaan julkisella säätelyllä. Euroopassa EU julkaisi vuoden 2024 alussa säädöstön nimeltä AI Act ensimmäisiksi tekoälyn käyttöä rajoittaviksi pelisäännöiksi, jonka jälkeen se on jatkanut sääntelyn kehitystä. Säännöksillä tekoälyn käytöstä pyritään saamaan turvallista, perusoikeuksien mukaista, silti edistäen innovointia. EU:n säädökset rajaavat tekoälyyn liittyviä haittakäytön tapauksia eri riskikategorioihin ja kohdistaa rajoituksia riskiluokituksen mukaisesti niin, että vakavampien riskien kohteita rajoitetaan rankimmin. Rajoitusten kohteiksi tähän mennessä on päätynyt muun muassa biometristen tunnistusjärjestelmien käyttö, sosiaalisen pisteytyksen käyttö sekä tekoälyn käyttö haitallisiin manipuloiviin ja harhaanjohtaviin tarkoituksiin. Rajoitukset pyrkivät siis hyvin pitkälti suojelemaan yksilöitä, käyttivät he sitten itse tekoälyä tai ei. Tekoälyn nopean kehityksen takia rajoittavien toimien uudistamiselle ja lisäämiselle tulee hyvin todennäköisesti olemaan tarvetta, ja nähtäväksi jääkin, miten se tulee vaikuttamaan turvallisuuteen sekä innovaatioihin ja kilpailukykyyn. (Euroopan parlamentti 2024; Valtioneuvosto 2025.)

4.5 Yhteenveto

Tieto-osuudessa generatiivisesta tekoälystä nousi esiin osa-alueita, jotka kertovat mallin kyvykkyydestä ja ominaisuuksista. Näistä suoraan mallien välillä vertailtavissa olevia ovat **älykkyys**, **tokenien käytön kapasiteetti** ja **hinta**.

Tehokkaasti toimivat neuroverkot, laadukas ja riittävä koulutusdata sekä mallin kyvykkyydet luonnollisen kielen käsittelyssä mahdollistavat ominaisuuksia, joiden voidaan ajatella tekevän mallista älykkään. Tokenien kapasiteetti vaikuttaa suoraan mallin käyttömahdollisuuksiin ja on siten merkittävä ominaisuus malleja vertaillessa. Kaikki osaavasta henkilöstöstä mallien suunnitteluun ja jatkuvaan kouluttamiseen maksaa näiden luonnollisesti myös heijastuessa kuluttajien maksamiin hintoihin. Kuluttajien näkökulmasta mallien välillä on kuitenkin paljon hintaeroja, mikä tuottaa mielenkiintoisia kysymyksiä siitä, mitä rahalla saa ja ovatko hintaerot perusteltavissa.

Suoraan vertailtavissa olevien osien lisäksi tieto-osuudella nousi esiin myös muita ominaisuuksia, jotka vaikuttavat selkeästi edellä mainittuihin älykkyyteen, token-kapasiteettiin ja hintaan sekä mallin käyttökokemukseen. Kyseisiä ominaisuuksia on: onko **avoimen lähdekoodin malli** vai ei ja onko **perusteleva malli** vai ei. Lähdekoodin avoimuus on merkittävä ominaisuus ainakin mallien hintaa vertaillessa sekä mahdollisesti vaikuttaa myös muihin ominaisuuksiin. Mallin kyky perustella vastauksiaan vaikuttaa kapasiteettitarpeeseen ja mahdollisesti myös ainakin mallin älykkyyteen.

Lisäksi mielenkiintoisiksi tutkittaviksi, käyttäjäkokemukseen vaikuttaviksi osa-alueiksi löytyi mallin **turvallisuus** ainakin selitettävyyden ja puolueellisuuden osalta sekä mallin erilaisten tietomuotojen käsittelyn taidot, eli **multimodaalisuus**.

Vertailtaviksi ominaisuuksiksi valikoitui siis tieto-osuuden perusteella:

- älykkyys
- konteksti-ikkuna (eli tokenien käytön kapasiteetti)
- hinta.

Vertailuun vaikuttavia ominaisuuksia valikoitui:

- avoin vai suljettu kielimalli
- perusteleva malli vai ei.

Muuten tärkeitä ja huomioonotettavia ominaisuuksia:

- mallin turvallisuus (selitettävyys, puolueellisuus)
- multimodaalisuus.

5 Mallien vertailu

Generatiivisen tekoälyn suosituimpien mallien markkinatilannetta hallitsee pääasiassa suurten amerikkalaisyritysten tuotokset, joiden mukana muutamat muunmaalaiset mallit sekoittamassa pakkaa. Lähtökohtaisesti jokainen malli on suunniteltu ja koulutettu hieman eri tavalla. Koulutuksessa käytettävä data ei ole kaikilla samaa. Myöskään mallien kehitystä tai vaatimuksia ei ole standardisoitu, kuten monilla muilla markkinoilla. Hyvin monet malleista ovat erittäin tehokkaita, spesifimpiin käyttötarkoituksiin myöhemmin jalostettavia perusmalleja, mutta luonnollisesti kaikki eivät voi olla kaikessa hyviä, joten erikoistumista eri käyttötarkoituksiin löytyy. Yritykset tietenkin myös pyrkivätkin keskittymään omaan vahvuuteensa erottuakseen kilpailijoista.

5.1 Vertailun suoritus

Itse vertailu suoritettiin tieto-osuudessa merkittäväksi ilmenneiden ominaisuuksien, älykkyyden, kapasiteetin ja hinnan perusteella, lisäksi hyödynnettiin tietoa kielimallien avoimuudesta ja perustelukyvykkyydestä. Lisäksi vertailuun otettiin mukaan mallin **nopeus** ja vastaamisen **viive**, jotka osoittautuivat saatavilla olevan aineiston perusteella tärkeiksi, ja ovat helposti vertailtavissa olevia ominaisuuksia. Kaikki vertailussa käytetty aineisto on peräisin artificialanalysis.ai-sivustolta (https://artificialanalysis.ai/models), josta löytyy myös paljon lisää materiaalia tekoälymalleista, niiden tarjoajista, ominaisuuksista sekä tarkemmat selitykset siihen, mitä vertailuarvojen takana on. Artificialanalysis.ai valittiin vertailun aineistoksi, koska se tarjoaa paljon tietoa, sopivista aiheista ja hyvin monen eri tekoälyvalmistajan useasta mallista. Jotta mallien vertailu suoraan toisiinsa nähden on mahdollista, on käytettävän vertailuaineiston oltava samanlaista läpi aineiston.

Vertailuaineistona käytettiin kolmeakymmentä artificialanalysis.ai-sivustolla olevaa generatiivisen tekoälyn mallia, jotka ovat lähtökohtaisesti markkinoiden suurimpien tekoälyn tarjoajien uusimpia tuotoksia, jotka ovat listattuna kuvassa 5. Sivustolla on enemmänkin malleja, joista on tarjolla dataa, mutta joukossa on myös vanhempia malleja sekä vasta juuri tulossa olevia malleja, joten tähän tutkimukseen sivuston malleista rajattiin vain osa. Valinnalle ei tarkkoja kriteereitä ollut, mutta malleja valittiin useasta eri yrityksestä ja useasta eri maasta. Tähän aineistoon löytyi malleja, joilla emoyhtiöitä on kotoisin tekoälyn markkinajohtaja Yhdysvalloista sekä Kiinasta ja Ranskasta.

Tavoitteena oli pyrkiä jaottelemaan malleja eri ominaisuuksien perusteella ja löytää aineistosta mallien välisiä eroja. Ratkaisuksi päädyttiin tekemään Power BI -tiedosto, jossa vertailuaineiston mallien vertailu keskenään onnistuu sekä on mahdollisuus verrata lähtökohtaisesti mitä tahansa muutakin mallia vertailuaineiston malleihin, kunhan vain vertailtavat tiedot mallin ominaisuuksista olisivat tarjolla. Tämä antaisi myös mahdollisuuden tuottaa aineiston perusteella visuaalinen näkymä, jonka avulla malleja olisi helppo vertailla keskenään, vaikka vain yhdellä silmäyksellä. Power BI -tiedoston lisääminen sellaisenaan insinöörityöhön ei teknisistä syistä onnistunut, mutta kuvat siitä löytyvät työn liitteistä.

Tekoälymallien vertailun ja Power BI -tiedoston hyödyntämisen havainnollistamiseksi valittiin aineistosta kolme mallia, joita vertailtiin keskenään sekä vertailuaineiston kanssa. Vertailuun sisällytettiin muihin malleihin vertailun lisäksi itse valmistajayrityksen tietoja mallista, johon vertailun tuloksia voisi mahdollisesti peilata. Näiden lisäksi muiden julkisten lähteiden avulla tuotettu yhteenveto mallin kyvykkyyksistä ja ominaisuuksista, teemoina erityisesti tieto-osuuden perusteella esiin nousseet turvallisuus ja multimodaalisuus. Jokaisen mallin vertailuosiossa ensimmäisenä on valmistajan antamat pohjatiedot, seuraavana datan perusteella tehty Power BI:hin pohjautuva analyysi ja kolmantena yhteenveto lisätietoineen.

5.2 Vertailtavat ominaisuudet

Vertailtavia numeerisia ominaisuuksia käytössä olivat älykkyys, kapasiteetti, hinta, nopeus ja viive, joista kaikki data on peräisin artificialanalysis.ai-sivustolta. Seuraavaksi ovat selitykset sille, millä perusteella arvot on määritelty, sekä missä muodossa ne on ilmaistu artificialanalysis.ai-sivustolla.

Vertailtavista arvoista älykkyys on hankalimmin määriteltävissä ja mitattavissa. Tässä tapauksessa mallin älykkyyttä artificialanalysis.ai-sivustolla kuvaa tulokset yhdeksästä kokeesta, jotka testaavat mallin yleistä päättelyä ja osaamista, matemaattisia kykyjä sekä koodaustaitoa. Tulokset älykkyydestä ovat saatavilla jokaisen testin erillisenä tuloksena, osa-alueittain yhdistettynä tai yhteistuloksena kaikkien älykkyyttä testaavan kokeen pohjalta. Yhdistettyyn lopputulokseen yleistiedon osuudella on 50 prosentin vaikutus ja matemaattisella sekä koodauksella kummallakin 25 prosentin vaikutus. Vertailuarvoina käytetään kokonaisälykkyyden sekä matematiikan ja koodaamisen tuloksia, jotka on ilmoitettu numeerisilla arvosanoilla mitä suurempi, sen parempi periaatteella.

Mallin konteksti-ikkunalla artificialanalysis.ai-sivustolla tarkoitetaan tokenien määrää, jota malli pystyy yhden tehtävän yhteydessä vastaanottamaan sekä tuottamaan. Konteksti-ikkunaa vertailtaessa arvo on siis yhdessä tehtävässä käytettävissä olevien tokenien määrä suuren kapasiteetin mahdollistaessa laajan sisällön käytön tehtävänannossa sekä kattavan vastauksen. Konteksti-ikkunan vertailuarvo on tokenien maksimimäärä, eli lähtökohtaisesti mitä suurempi arvo, sen parempi.

Nopeus tarkoittaa tässä artificialanalysis.ai-sivuston datassa vauhtia, jolla malli kykenee tuottamaan tokeneita tehtävää tehdessään. Mallin nopeus kuvataan siis tuotettujen tokenien määränä sekunnissa, jota malli tuottaa sen antaessa vastausta. Mitä suurempi nopeuden vertailuarvo on, sitä parempi.

Viiveestä puhuttaessa viitataan artificialanalysis.ai-sivustolla aikaan, joka mallilla kestää tehtävän vastaanottamisesta, ensimmäisen vastaus-tokenin tuottamiseen. Viive kuvataan sekunteina, joka mallilla kestää aloittaa vastauksen

tuottaminen, eli mitä pienempi viiveen vertailuarvo, sen parempi. Lisäksi nähtävissä on myös vastausaika kokonaisuudessaan, johon on viive ennen vastausta ja aika tuottaa sata tokenia. Nopeus ja viive kertovat mallin algoritmien monimutkaisuudesta ja sujuvuudesta sekä mallia pyörittävän fyysisen koneiston laadusta ja toimintakyvystä.

Raha on luonnollisesti merkittävänä osana tekoälymalleja vertaillessa. Tässä tapauksessa hinnan arvo on artificialanalysis.ai-sivustolla määritelty Yhdysvaltain dollareina per miljoona tokenia. Arvo on myös jaettavissa erikseen mallille tehtävänannossa syötettävien tokenien, ja mallin vastatessa käyttämien tokenien hintaan. Mallin käyttämisen hinnan vertailuarvo on luonnollisesti muodossa, mitä pienempi, sen parempi. Hintaa kuvaa vertailussa käytetyissä hajontakuvioissa eri tekoälymalleja kuvaavien pallojen koko. Mitä suurempi pallo, sen suurempi myös mallin hinta.

Esimerkkimalleja vertailuun valitessa päädyttiin ratkaisuun, jossa pyrittiin ottamaan arvioitavaksi yritystaustaltaan ja toiminnallisuuksiltaan mahdollisimman erilaisia malleja, silti kuitenkin pysytellen tekoälymaailman merkittävimpien yritysten joukossa. Lyhyen taustatutkimuksen jälkeen malleiksi valittiin:

- yksi tekoälyjätti OpenAI:n uusimmista, malli **o1**, joka on vertailtavista ainoa tehokkaaseen perusteluun rakennettu malli

- kiinalaisen DeekSeek AI yhtiön **V3**, vertailun ainoana avoimen lähdekoodin mallina

- hyvin tunnetun teknologiajätti Googlen, **Gemini 1.5 Pro**, joka on varmasti hyvin monelle tuttu sen sulautuessa nykyään Googlen muihin ohjelmiin gmailista, Google docsiin.

Mallien vertailussa tutkittiin siis niiden viittä ominaisuutta: älykkyys, kapasiteetti, hinta, nopeus ja viive, joilla hyödynnettiin myös tietoa mallien lähdekoodin avoimuudesta ja perustelun kyvykkyydestä. Vertailua varten tuotettiin Power BI -tiedosto ominaisuuksien datan perusteella, joka loi myös mahdollisuuden vertailla

juuri haluamaansa mallia vertailuaineistoon. Vertailuaineistona käytettiin kolmeakymmentä suurimpien tekoälytuottajien mallia, joista vertailuun esimerkkimalleiksi valittiin kolme. Ominaisuuksiin perustuvan vertailun tueksi hyödynnettiin itse mallien valmistajayritysten lähtötietoja malleista sekä tutkimustietoa julkisista lähteistä.

Yritys	Malli	Average of Älykkyys	Average of Hinta	Average of Konteksti-ikkuna	Average of Nopeus	Average of Viive	Average of Matematiikka	Sum of Koodaaminen
Alibaba Cloud	Qwen 2.5 Max	45	8.00	32000.00	35.00	1.18	53.00	35.00
Alibaba Cloud	Qwen Turbo	34	0.25	1000000.00	104.00	1.01	46.00	16.00
Alibaba Cloud	QwQ 23B	58	0.95	131000.00	97.00	0.54	87.00	49.00
Anthropic	Claude 3 Opus	35	90.00	200000.00	29.00	1.17	34.00	26.00
Anthropic	Claude 3.5 Haiku	35	4.80	200000.00	65.00	2.02	38.00	29.00
Anthropic	Claude 3.7 Sonnet	48	18.00	200000.00	79.00	1.04	54.00	38.00
Anthropic	Claude 3.7 Sonnet Thinking	57	18.00	200000.00	79.00	1.00	72.00	44.00
DeepSeek	DeepSeek R1	60	2.74	128000.00	25.00	5.70	82.00	49.00
DeepSeek	DeepSeek R1 Distill Llama 70B	48	1.26	128000.00	123.00	0.52	80.00	29.00
DeepSeek	DeepSeek R1 Distill Qwen 14B	49	1.76	128000.00	84.00	0.66	81.00	31.00
DeepSeek	DeepSeek R1 Distill Qwen 32B	52	0.60	128000.00	48.00	0.37	81.00	32.00
DeepSeek	DeepSeek V3	46	0.38	128000.00	27.00	5.74	57.00	36.00
Google	Gemini 1.5 Flash-8B	31	0.19	1050000.00	276.00	0.21	36.00	22.00
Google	Gemini 1.5 Pro (Sep)	45	6.25	2000000.00	96.00	0.49	57.00	31.00
Google	Gemini 2.0 Flash	48	0.50	1000000.00	256.00	0.28	63.00	32.00
Google	Gemini 2.0 Flash-Lite (Feb '25)	41	0.37	1000000.00	179.00	0.24	55.00	22.00
Meta	Llama 3.1 Instruct 405B	40	7.00	128000.00	31.00	0.71	46.00	30.00
Meta	Llama 3.1 Instruct 70B	35	1.33	128000.00	70.00	0.50	41.00	25.00
Meta	Llama 3.1 Instruct 8B	24	0.20	128000.00	173.00	0.35	30.00	12.00
Meta	Llama 3.2 Instruct 3B	20	0.12	128000.00	143.00	0.41	28.00	7.00
Meta	Llama 3.3 Instruct 70B	41	1.29	128000.00	126.00	0.53	54.00	27.00
Mistral AI	Ministral 8B	22	0.20	128000.00	146.00	0.37	30.00	11.00
Mistral AI	Mistral Large 2 (Nov '24)	38	8.00	128000.00	29.00	0.54	42.00	29.00
Mistral AI	Mistral NeMo	20	0.30	128000.00	148.00	0.37	20.00	8.00
Mistral AI	Mistral Small 3	35	0.40	32000.00	119.00	0.39	40.00	24.00
OpenAI	03-mini	63	5.50	200000.00	188.00	13.09	87.00	56.00
OpenAI	GPT-4o ('24)	41	12.50	128000.00	116.00	0.40	45.00	32.00
OpenAI	GPT-4o mini	36	0.75	128000.00	69.00	0.34	45.00	23.00
OpenAI	o1	62	75.00	200000.00	117.00	26.22	85.00	52.00
OpenAI	o1-mini	54	5.50	128000.00	209.00	11.09	77.00	45.00
Total		**42**	**9.07**	**316433.33**	**109.53**	**2.58**	**54.87**	**902.00**

Kuva 5 Vertailuaineisto kokonaisuudessaan.

5.3 OpenAI o1

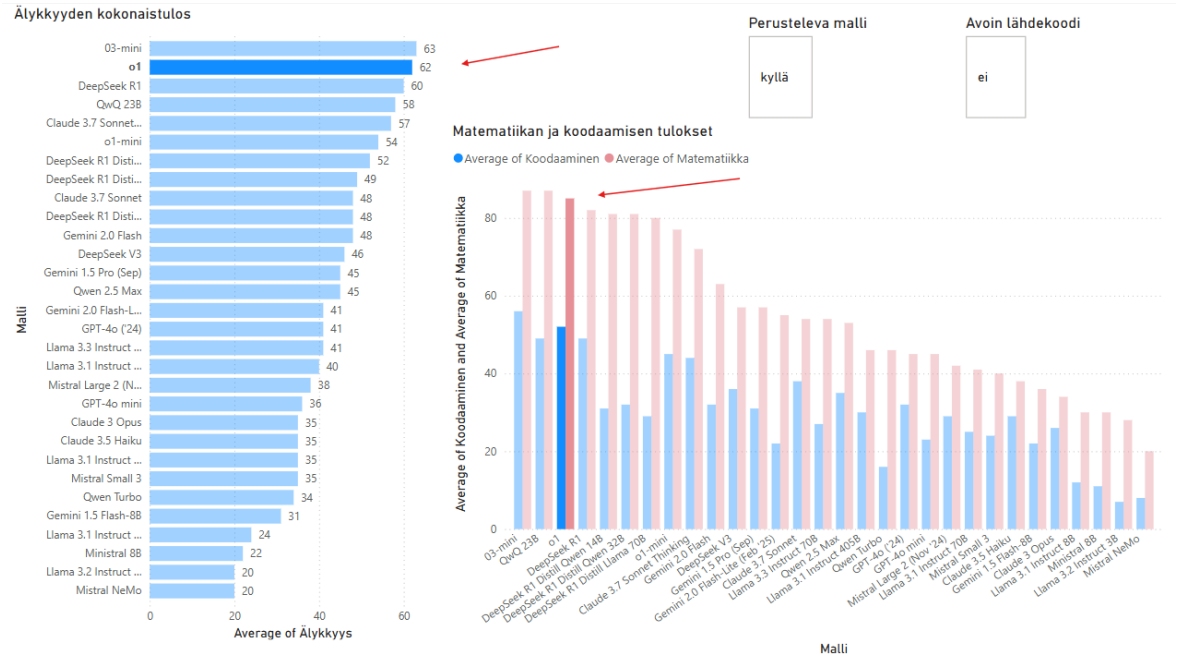
5.3.1 Lähtötiedot

Open AI:n oman esittelyn mukaan o1 on äärimmäisen älykäs perusteleva malli, jonka koodaaminen ja matemaattis-luonnontieteellinen osaaminen on huipputasoa. Se menestyi selkeästi edeltäjiään paremmin näiden osa-alueiden ihmisille tarkoitetuissa kokeissa. Vahvistusoppimisen avulla malli on opetettu käymään tietynlainen ajatusprosessi läpi ennen vastauksen antamista, mikä tuottaa laadukkaita vastauksia, ja joiden lisäksi perusteluja niille. Tieteellisten testien lisäksi mallia on verrattu GPT-4o-malliin ihmisten antaessa palautetta käyttökokemuksesta. Tuloksena oli o1:n suosiminen juuri edellä mainituilla osa-alueilla,

mutta luonnollisen kielen käsittelyssä useampi oli GPT-4o-mallin kannalla. OpenAI:n mukaan perustelevan mallin opetustapa ja toiminta edistää turvallisuutta mahdollistaessaan paremman näkymän mallin toimintatapaan sekä ajattelun vähentäessä selkeitä virheitä. o1-mallin kanssa päädyttiin siihen, ettei käyttäjälle näytetä koko ajatteluprosessia, mallin pohdinnan vapauteen ja sen valvomiseen liittyvien syiden vuoksi. (OpenAI 2024.)

5.3.2 Vertailu

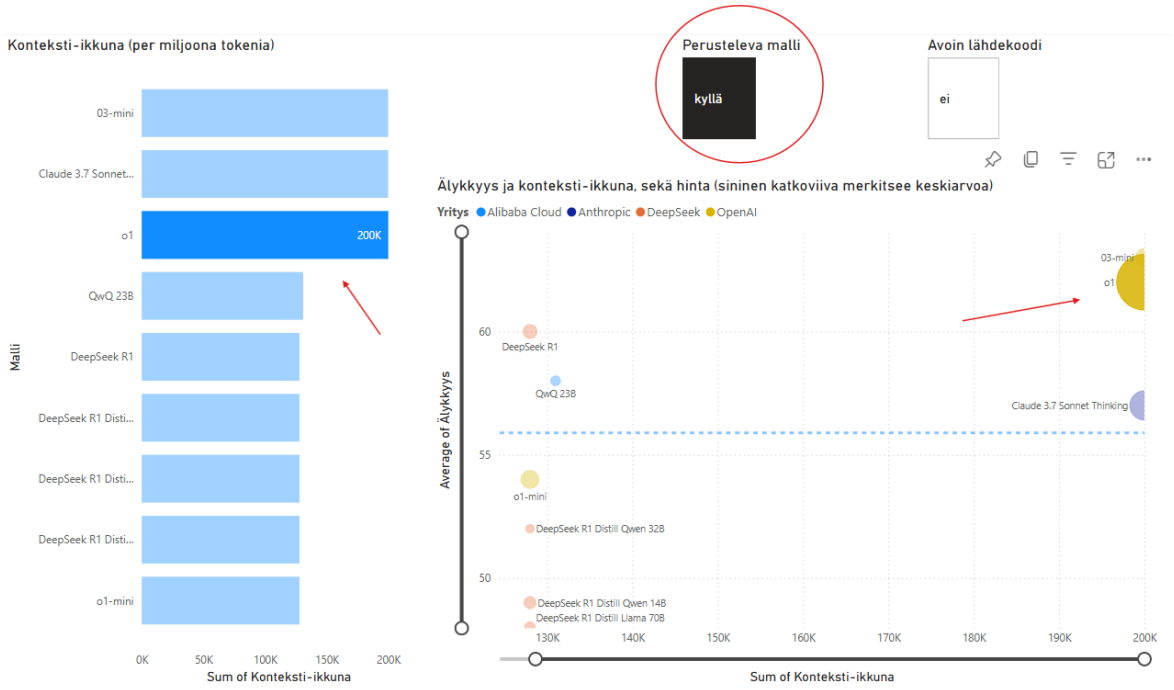
Vertailun perusteella o1-mallille voidaan nähdä hyvin samanlaisia piirteitä kuin mitä yhtiö itse kertoo. Kaikissa älykkyysmittareissa, matematiikasta monikielisyyteen, se on muita tarkasteltavia malleja selkeästi edellä ja koko aineistossakin aivan kärkipäässä, juuri ja juuri häviten kokonaistuloksessa pelkästään OpenAI:n toiselle, 03-mini-mallille.



Kuva 6 Koko aineiston älykkyyden tulokset. o1 korostettuna.

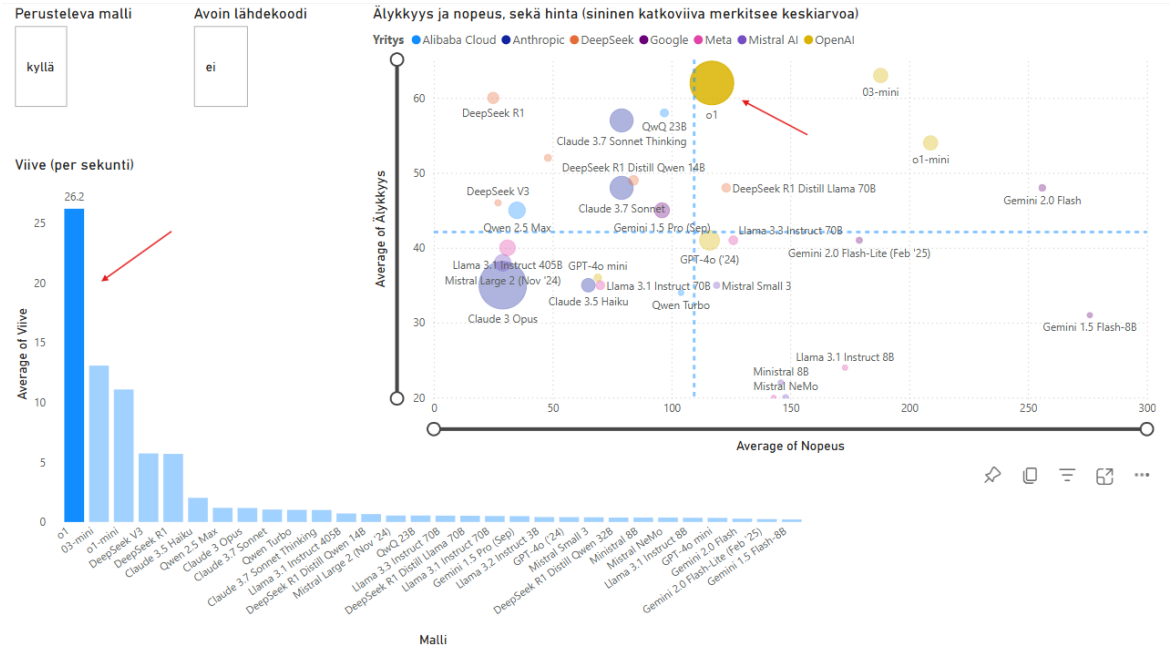
Konteksti-ikkuna mallilla on kohtuullisen suuri, joskin koko aineiston mallien keskiarvon alapuolella. Toisaalta kun verrataan muihin perusteluun kykeneviin

malleihin, o1 on jaetulla kärkisijalla. Kun tarkastellaan pelkästään perustelevia malleja, nähdään, että konteksti-ikkuna on hyvin tarkasti joko 128 000 tai 200 000. Koko aineistoa käsitellessä taas mallien kapasiteetti vaihtelee hyvin laajasti, 32 000 ja 2 miljoonan välillä.



Kuva 7 Vain perustelevien mallien konteksti-ikkunan tulokset. o1 korostettuna.

Nopeudessa o1-malli on lähellä aineiston keskiarvoa, hieman sen paremmalla puolella. Mutta viiveen kohdalla näkyy hyvin selkeästi mallin perustelukyvykkyyden vaikutus tuloksen ollessa ylivoimaisesti aineiston suurin. Muutkin perusteluun kykenevät mallit ovat viiveen osalta suurimpien joukossa, mutta o1:n 26,2 sekunnin viive on lähes kaksinkertainen seuraavana tulevaan verrattuna.



Kuva 8 Koko aineiston nopeuden ja viiveen tulokset. o1 korostettuna.

Äärimmäisen kyvykäs malli on myös huomattavan kallis käyttää. Sen hinta 75 dollaria per miljoona käytettyä tokenia, on vertailuaineiston toiseksi kallein, ainoastaan Anthropicin Claude 3 Opus mallin takana. Erolla koko aineiston keskiarvoon 9,36 on valtava ja erityisesti lähtökohtaisesti halvempien, avoimen lähdekoodin mallien keskiarvoon 1,77 dollaria.

5.3.3 Yhteenveto

OpenAI:n o1 loistaa älyllisissä tehtävissä, joista erityisesti monimutkaisten ongelmien selvittämisessä ja matemaattisissa ongelmissa. Perusteluun rakennettu malli ei ole kirjallisissa, kuten tekstin muokkaamista vaativissa tehtävissä yhtä vahva kuin edeltäjänsä. Myöskään multimodaalisesti se ei ole yhtä taitava sen käsitellessä ainoastaan eri tekstimuotoisia kieliä. Turvallisuuden näkökulmasta perustelu tuo lisäarvoa, ja se antaa hieman näkyvyyttä mallin ajattelutapaan parantaen selitettävyyttä ja mahdollistaen valheellisuuteen ja puolueellisuuteen liittyvien asioiden tunnistamisen helpommin. Yleisesti O1 näyttäisi olevan erin-

omainen vaihtoehto hyvin hankaliin tehtäviin, joilla ei ole kiire. Se on hyvin älykäs ja tuo käyttäjälleen lisäarvoa selittämällä vastauksiaan. O1 on kuitenkin kovin hidas, eikä siten sovi nopeutta vaativiin tehtäviin. Käyttäjän päätettävissä on siis, riittääkö älykkyys ja hyvät perustelutaidot oikeuttamaan erittäin korkean hinnan ja rajalliset käyttömahdollisuudet. (Vellum 2024.)

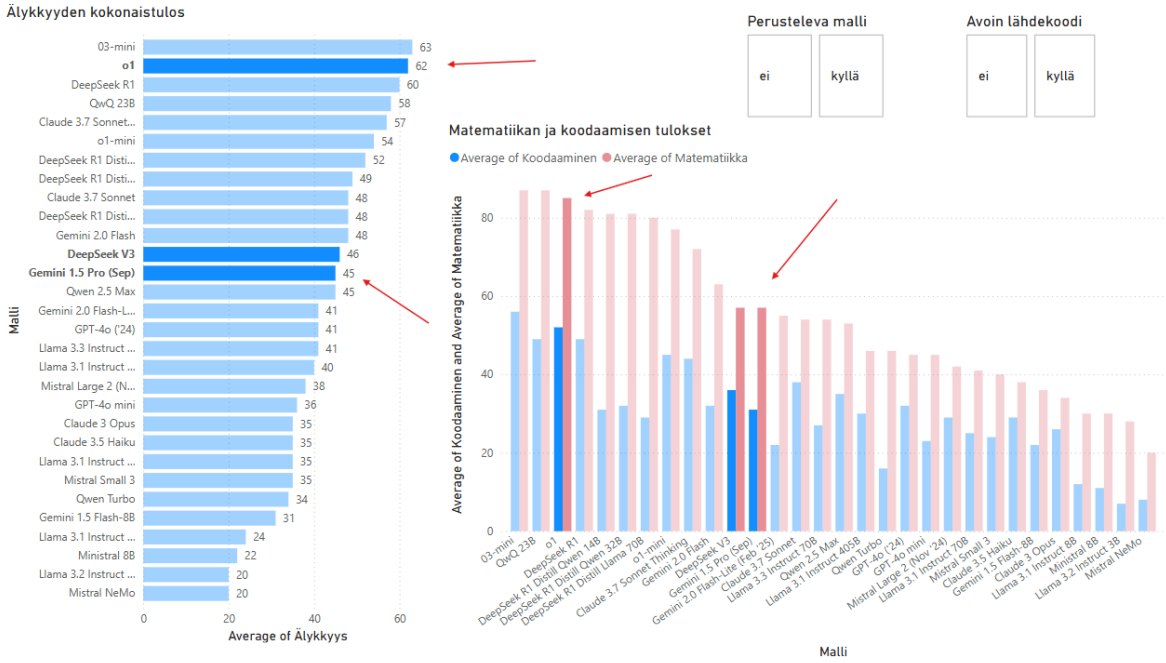
5.4 Google Gemini 1.5 Pro

5.4.1 Lähtötiedot

Googlen omien sivujen mukaan Gemini 1.5 Pro on multimodaalisesti erittäin taitava malli, joka nojaa sen valtavan suureen konteksti-ikkunaan. Malli kykenee käsittelemään hyvin suuria syötteitä lähes missä muodossa tahansa. Sen kapasiteetti mahdollistaa myös suuren tietomäärän opettamisen syötteen avulla. Googlen mukaan malli voi käsitellä kerralla 2 tuntia videota, 22 tuntia ääntä, 60 000 riviä koodia tai 1 400 000 sanaa. 1.5 Pro hyödyntää myös uutta MOE-arkkitehtuuria (Mixture-of-Experts), jonka avulla sen neuroverkosta toimii kerrallaan vain juuri tehtävän taitava osa, mikä tekee mallista laadukkaamman ja kustannustehokkaamman. (Google 2024.)

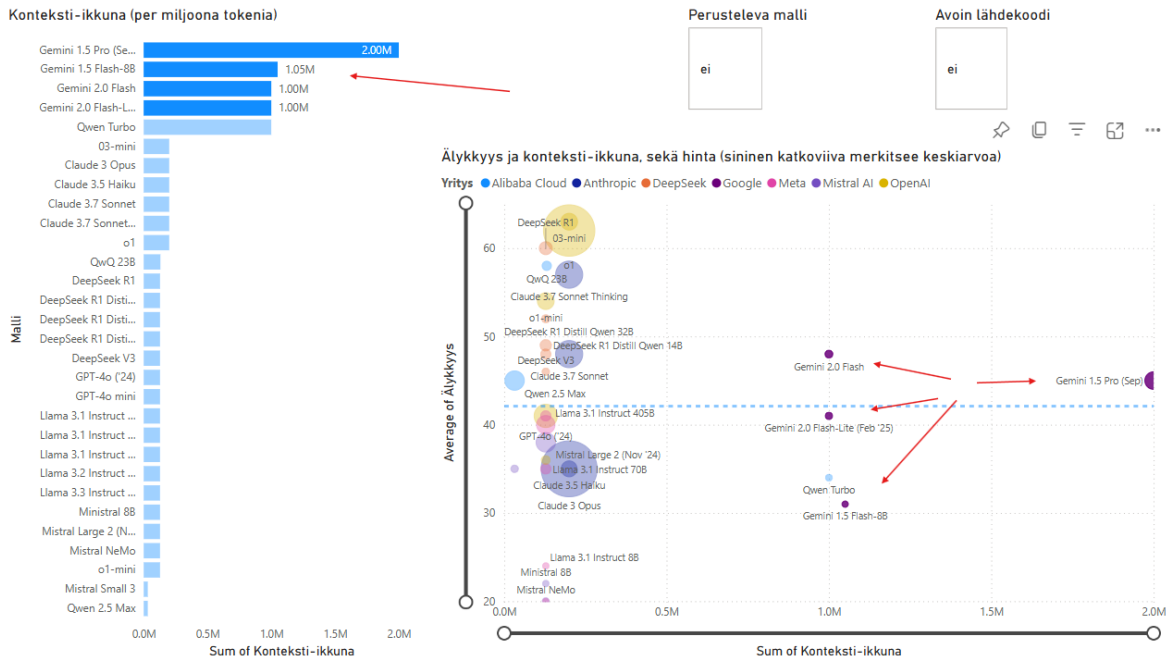
5.4.2 Vertailu

Vertailun perusteella Gemini on älykkyyden mittareilla vertailtavista hyvin tasoissa DeekSeekin V3-mallin kanssa, mutta häviää kuitenkin sille matematiikassa sekä koodauksessa. Koko aineistoon nähden älykkyys on hieman keskiarvon paremmalla puolella. Jos lähtökohtaisesti näissä testeissä paremmin suoriutuvat perustelevat mallit rajataan pois, on Gemini 1.5 koko älykkyyttä kuvaavan tuloksen perusteella aineiston neljänneksi paras.



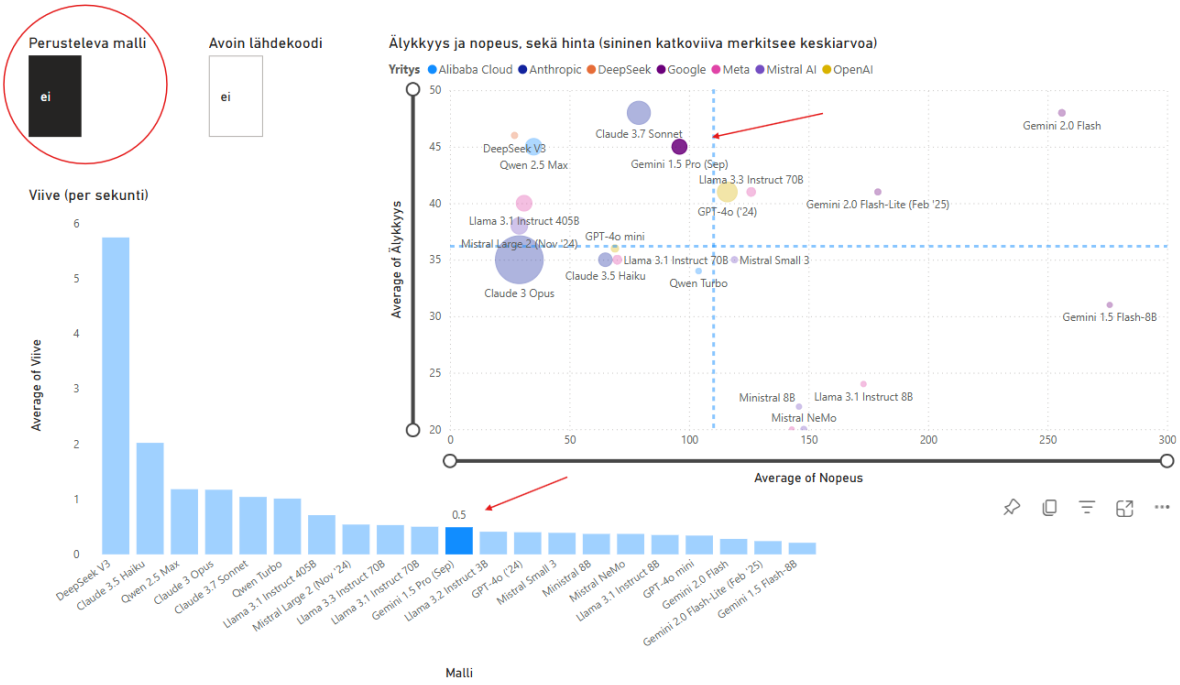
Kuva 9 Koko aineiston älykkyyden tulokset. Korostettuna kaikki vertailtavat mallit.

Konteksti-ikkunan kohdalla malli peittoaa muut vertailun kilpakumppaninsa mennen tullen, ja on ylivoimaisesti koko aineiston paras. Gemini 1.5 Pro kaksinkertaisella, 2 miljoonan tokenin kapasiteetilla seuraavana tulevana tuleviin verrattuna suurimman osan aineistosta ollessa noin 100 000–200 000 kohdalla. Mielenkiintoinen huomio on, että koko vertailuaineiston neljä suurimman konteksti-ikkunan omaavaa ovat kaikki Googlen tekemiä eri versioita Gemini-mallista.



Kuva 10 Koko aineiston konteksti-ikkunan tulokset. Korostettuna Googlen mallit.

Hinnassa Gemini 1.5 Pro on vertailtavista malleista keskimmäisenä, hyvin selkeästi OpenAI:n mallia halvempana. Koko aineistoon verrattaessa Gemini on keskiarvoa edullisempi, mutta vain perusteluun kykenemättömiä malleja tutkiessa sen ollessa kalleimpien joukossa, kuitenkin hyvin kaukana aivan kalleimmista. Malli on myös kohtalaisen nopea, vaikkakin hieman keskiarvon alapuolella. Huomattavaa on, että yksittäiset erittäin nopeat mallit nostavat keskiarvoa huomattavasti. Koko vertailuaineistoa katsoessa 1.5 Pro on viiveeltään erittäin pieni, mutta kuitenkin keskiarvoja hieman heikompi, kun vertaillaan pelkästään perusteluun kykenemättömiä malleja.



Kuva 11 Perusteluun kykenemättömien mallien nopeuden ja viiveen tulokset. Korostettuna Gemini 1.5 Pro.

5.4.3 Yhteenveto

Googlen Gemini 1.5 Pro on multimodaalisesti erittäin kyvykäs malli, joka osaa tietomuodot videoista eri koodauskieliin. Sen massiivinen konteksti-ikkuna mahdollistaa erittäin laajojen syötteiden käsittelyn, jonka avulla erilaiset tiedonkäsittelyyn liittyvät tehtävät onnistuvat. Myös sulava toiminta Googlen ekosysteemissä yhteistyössä driven, gmailin ja lisäominaisuudella jopa youtuben kanssa on Geminin vahvuus. Se on yhteydessä internetiin, joka mahdollistaa reaaliaikaisen tiedonhaun. Malli ei kuitenkaan ole kaikkein akateemisesti kyvykkäin ja joidenkin lähteiden perusteella sillä on välillä ongelmia vastausten tarkkuuden kanssa. Turvallisuuteen liittyen 1.5 Pro tarjoaa käyttäjälleen asetuksissa mahdollisuuden säätää, kuinka herkästi se hylkää mahdollisesti häiritsevää sisältöä. Joidenkin käyttäjien mukaan Gemini voi olla jopa liian turvallinen sen blokatessa syötettä, jota se virheellisesti luulee häiritseväksi materiaaliksi, mikä rajoittaa näin mahdollisesti turhaan sen käyttöä. (SGU 2025; Google 2025.)

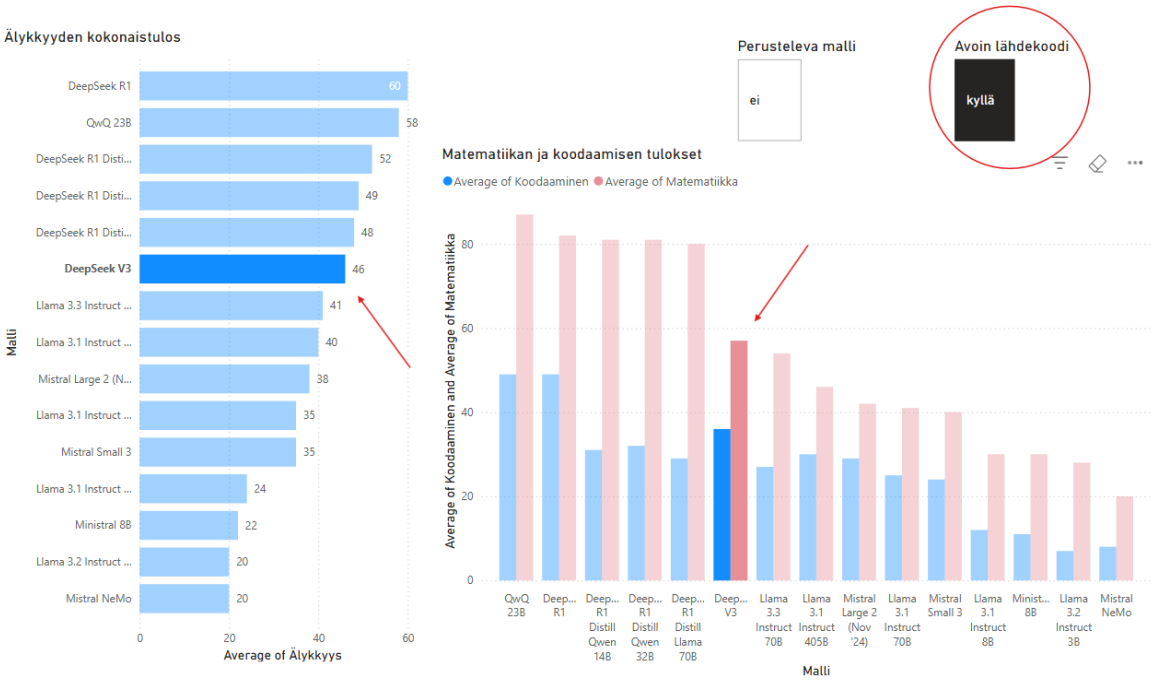
5.5 DeepSeek AI V3

5.5.1 Lähtötiedot

Omien sivujensa mukaan DeepSeek on pyrkinyt tekemään V3-mallistaan älykkään, kuitenkin käyttäjille halpaan hintaan. Malli pyrkii älykkyydessään pääsemään lähelle markkinajohtajia, kuluttajille ja jatkokehittäjille halpana ja kiinnostavana, avoimen lähdekoodin mallina. Se käyttää Googlen mallin kanssa ilmeisesti ainakin samantyyppistä MOE-arkkitehtuuria, ja yhtiö kertoo mallin kaikista 671 miljardista parametristä toimivan kerrallaan 37 miljardia. Multimodaalisia kykyjä V3-mallilla ei ole, mutta yksinkertaisempiin perustelua vaativiin tehtäviin mallissa on ”DeepThink” -ominaisuus, joka vastaa markkinoiden kiinnostukseen perustelevista malleista. Huomattavaa on kuitenkin, että DeekSeek on myöhemmin myös julkaissut juuri perusteluun erikoistuneen R1-mallin. (DeepSeek AI 2024.)

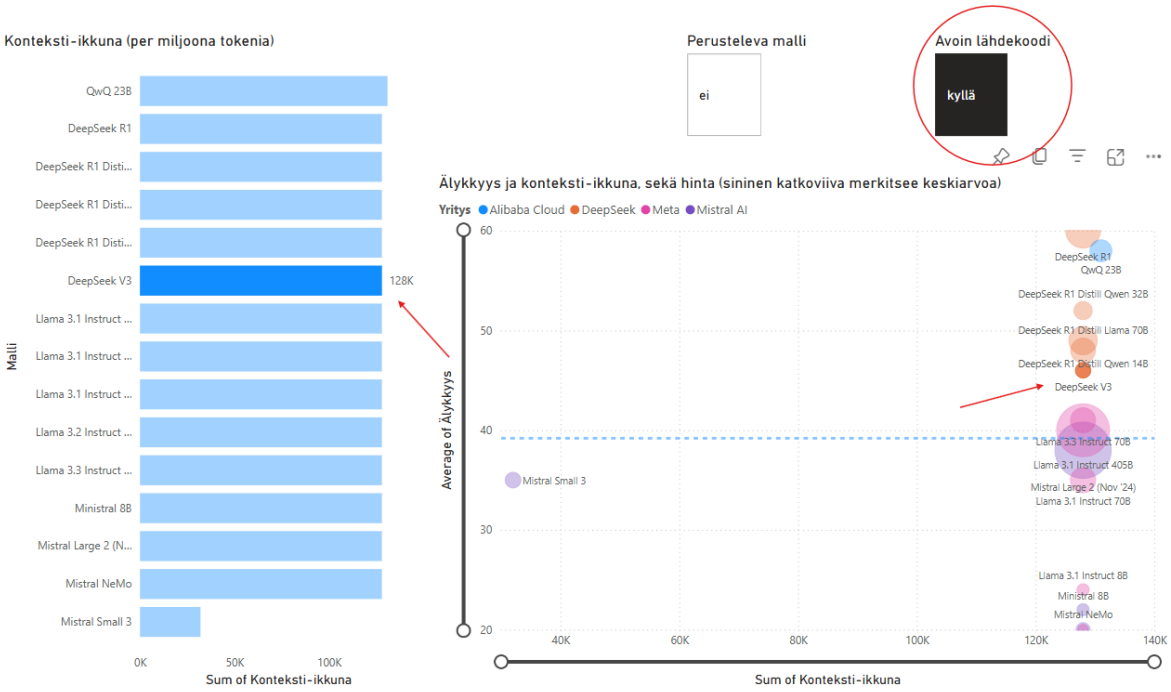
5.5.2 Vertailu

Vertailun perusteella V3 on älykkyysmittareilla mitattavista keskimmäisenä, juuri Googlen Geminin yläpuolella. Erityisesti se menestyy koodauksen osa-alueella. Älykkyydessä se on koko aineiston mittakaavassa keskiarvon yläpuolella erityisesti silloin, jos vertaillaan sitä vain muiden avoimen lähdekoodin mallien kanssa.



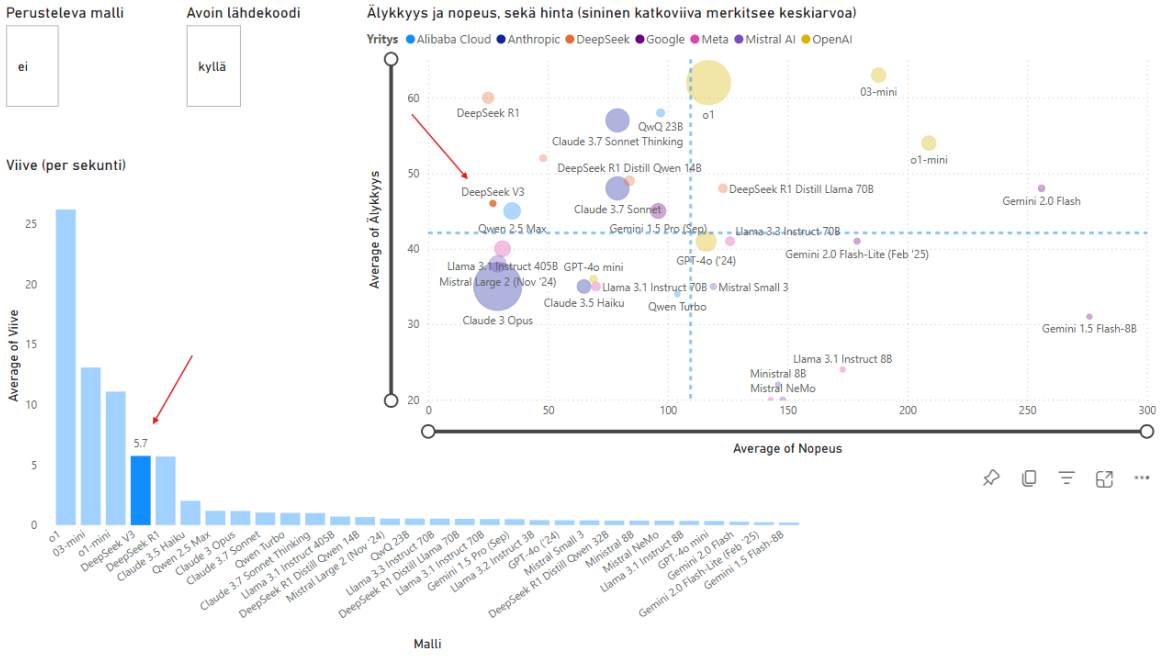
Kuva 12 Avoimen lähdekoodin mallien älykkyyden tulokset. V3 korostettuna.

Mallin konteksti-ikkuna 128 000 on vertailtavista pienin, kuitenkin käytännössä sama kuin kaikilla muilla avoimilla malleilla yhtä poikkeusta lukuun ottamatta. Se on myös koko aineiston yleisin tulos kapasiteetin osalta. DeepSeekin mallin hinta on erittäin edullinen, kuten lähes kaikki avoimen lähdekoodin mallit ovatkin. Erityisesti OpenAI:n malliin verrattuna, se on äärimmäisen halpa, ollessaan koko aineiston halvimman kymmenikön joukossa.



Kuva 13 Avoimen lähdekoodin mallien konteksti-ikkunan tulokset. V3 korostettuna.

V3 on vertailtavista selkeästi heikoin nopeudessa sekä viiveessä. Myös koko aineistoon ja erityisesti muihin perusteluun kykenemättömiin malleihin verrattuna on suhteellisen hidas ja pitkäviiveinen.



Kuva 14 Koko aineiston nopeuden ja viiveen tulokset. V3 korostettuna.

5.5.3 Yhteenveto

DeepSeekin V3 on selkeästi halpa malli, joka pystyy älykkyydessä kilpailemaan hyvin monen mallin kanssa rinta rinnan, erityisesti matematiikassa ja koodauksessa. Se toimii kustannustehokkaasti käyttäen vain osaa parametreistään, ja loistaa monimutkaista päättelyä vaativissa ja tieteellisissä tehtävissä. V3 ei ole missään nimessä tehokkain malli sen nopeutta ja viivettä katsoessa, mutta kuitenkin kohtuullisella tasolla sen hintaan nähden. Sillä on erittäin hyvä kielitaito englanniksi ja kiinaksi, mutta muuten sen multimodaaliset ominaisuudet eivät kuulu sen vahvuuksiin, pois lukien erinomainen englannin ja kiinan taito. Mallin avoimuus tuo suljettujen ja avoimien mallien väliä pienemmäksi ja mahdollistaa jatkokehittäjille tilaisuuden sen muokkaukselle. Kiinalaisyhtiön julkaisemat tekoälymallit ovat herättäneet etenkin länsimaissa kysymyksiä niiden turvallisuudesta, ja joidenkin lähteiden mukaan mallien tietosuojakäytännöissä on epäilyttävyyksiä. Joidenkin tutkimuksien mukaan myös jotkin DeepSeekin mallien vastaukset liittyen Kiinan valtiolle arkoihin asioihin ovat hyvin vajavaisia tai muuten

keskustelua herättäviä. Tämä ei tietenkään poista sitä mahdollisuutta, että muiden maiden tuottamat mallit toimisivat eri tavalla, mutta Kiinassa on tiukemmat tekoälysäädökset kuin esimerkiksi Yhdysvalloissa. (SGU 2025; Wodecki 2025.)

5.6 Vertailun yhteenveto

Vertailun tulosten tiivistämiseksi luotiin taulukko 2, josta näkee vertailtujen mallien suoriutumisen yksinkertaisesti havainnollistettuna. Taulukon hymiöt kuvastavat mallien suoriutumista tutkittujen ominaisuuksien osalta. Arvostelut pohjautuvat aikaisempaan vertailuun ja sen pohjalta tehtyihin johtopäätöksiin. Tarkkoja kriteereitä ei taulukon eri arvosanoille ole, vaan tarkemmat selitykset annetuille tuloksille löytyvät ”Mallien vertailu” -osiosta. Lisäksi taulukon alla on kirjallinen tiivistelmä jokaisen mallin ominaisuuksista.

Taulukko 2 Yhteenveto mallien vertailusta.

	o1	**Gemini 1.5 Pro**	**V3**
Älykkyys			
Konteksti-ikkuna			
Hinta			
Nopeus			
Viive			
Multimodaalisuus			
Turvallisuus			

Avoin kielimalli	Ei	Ei	Kyllä
Perusteleva malli	Kyllä	Ei	Ei
Erityisvahvuudet	Älylliset tehtävät, matematiikka, ongelmanratkaisu	Valtava konteksti-ikkuna, vahva multimodaalisesti, toiminta Googlen ekosysteemissä	Avoin malli, kustannustehokas, kiinan kieli

Vertailun perusteella OpenAI:n o1 osoittautui erinomaiseksi vaihtoehdoksi, jos tarkoituksena on ratkaista hankalia ”älyllistä” ajattelua vaativia ongelmia. Älyllisesti lahjakas malli ei kuitenkaan luultavasti ole paras vaihtoehto kirjallista taitoa vaativissa tehtävissä, eikä se kykene vastaanottamaan ja käsittelemään muuta kuin kirjallista tietoa. Korkean hintansa takia se ei myöskään ole varmasti monelle houkuttelevin vaihtoehto, mutta nähtäväksi jää, miten o1 ja muut kohtalaisen uudet perustelevat mallit tulevat tulevaisuudessa kehittymään.

Googlen Gemini 1.5 Pro osoittautui vertailussa konteksti-ikkunaltaan ylivertaiseksi ja tehokkaasti useita tietotyyppejä käsitteleväksi malliksi. Muissa vertailuominaisuuksissa Gemini ei erityisesti loistanut, mutta perusvarman suorittamisen lisäksi sen vahvuudeksi voidaan nostaa toiminta muiden Googlen sovellusten kanssa, joka varmasti houkuttaa, jos sattuu käyttämään Googlen tuotteita jo ennestään.

DeepSeekin V3 erottui vertailussa edullisella hinnallaan ja siihen nähden vahvoilla tuloksilla älykkyysmittareilla, erityisesti koodaamisessa. Vertailtavista ainoa avoimen kielimallin versio ei huimaa päätä tehokkuudellaan, joka on hinta huomioon ottaen ymmärrettävää. Se on luonnollisesti taitava kiinan kielessä, toisin kuin monet muut mallit, mutta eritietomuotojen käsittelyssä se ei ole kovin kyvykäs. V3 on siis hyvin edullinen ja hintaansa nähden älyllisesti vahva perusmalli, joka on helposti saatavilla, mutta jonka tietosuojahuolet voivat kuitenkin vaikuttaa sen käytön kiinnostavuuteen

6 Yhteenveto

Metropolia Ammattikorkeakoulu on ja pyrkii tulevaisuudessa olemaan aktiivinen tekoälyn hyödyntäjä, mikä tarkoittaa useita tulevia tekoälyprojekteja. Tämä insinöörityö tarjoaa pohjan, jota on mahdollista näissä tulevissa projekteissa hyödyntää tekoälymallien arvioinnissa. Insinöörityö sisältää generatiivisen tekoälyn mallien vertailuille kehyksen, jonka käyttöä demonstroitiin kolmen esimerkkimallia käyttäen. Työn lopputuloksena oli Power BI -pohjainen vertailupohja sekä kokonaisuutena tapa, jolla malleja vertailla. Valitettavasti tällä aikataululla ei ollut mahdollista saada Metropolialta palautetta työstä. Power BI -malli onnistui kohtalaisen hyvin. Käytettyjen ominaisuuksien vertailussa malli toimii hyvin perustana, jonka kanssa verrata jotain käytetyn aineiston malleista tai täysin muuta mallia, josta löytyvät tarvittavat tiedot ominaisuuksista. Yhtenä tavoitteena oli Power Bi:lla luoda yleisnäkymä, jossa yhdellä silmäyksellä näkisi tekoälymallin merkittävimmät ominaisuudet, mutta tämä ei tämän opinnäytetyön rajoissa aivan onnistunut.

Tätä insinöörityötä, ja suoritettua tekoälymallien vertailua ei tehty osana mitään tiettyä projektia, joten selkeää ratkaisua tai suositusta ei siten voida määrittää. Lopputuloksena työllä on pohja mallien vertailulle, jota voidaan tulevaisuudessa hyödyntää, kun tarkoituksena on löytää tiettyyn käyttötarkoitukseen sopiva malli, tai muussa tarkoituksessa vertailla eri malleja.

Aiheen kirjallisuuteen tutustumisen aikana ja itse insinöörityötä tehdessä nousi myös esiin aihealueita, joista voisi olla hyödyllistä tehdä jatkotutkimusta. Itse tekoälymalleista mielenkiintoiseksi aiheeksi nousi itsenäiseen toimintaan luotavat tekoälyagentit ja niiden yhdistelyn mahdollisuudet automaattisten työprosessien luomiseksi.

Toinen insinöörityön aikana esiin noussut aihe on tekoälyn käyttöönotto yrityksissä. Tutkimusta tehdessä useaan kertaan esiin nousi yritysten ongelmat tekoälyn järjestelmällisessä käyttöönotossa ja tekoälystä todellisen hyödyn saavut-

tamisessa. Tekoälymallien vertailu on merkittävä osa tekoälyn käyttöönottoa organisaatiossa, joten aiheen ja sen haasteiden tutkiminen voisi olla mielenkiintoinen aihe tukia jatkona tälle insinöörityölle.

6.1 Työn arviointi

Insinöörityö saavutti tavoitteensa kiitettävästi. Työn aikana sen aihe ja tavoite hieman vaihtelivat erinäisten syiden, kuten Metropolian tarpeiden muuttumisen ja työn käsittelemän aiheen laajuuden hakemisen takia. Nämä hankaloitti työn suorittamista selkeästi ja suoraviivaisesti. Myös kirjoittajan rajallinen tietämys aiheesta hidasti työn suorittamista ja erityisesti työn alkuosassa vaati paljon aikaa asioiden opetteluun, joka heijastui myös tieto-osuuteen, jossa päädyttiin aloittamaan tekoälyn teoria aivan perusteista. Alun jälkeen työn suorittaminen sujui vaihtelevasti, aika-ajoin sen ollessa erittäin sujuvaa, mutta välillä kuitenkin hitaammin. Hankaluuksia tuli vastaan etenkin aiheen rajauksen ja työn tavoitteiden kanssa, koska aihe oli hyvin laaja ja entuudestaan tuntematon. Tähän olisi luultavasti auttanut tarkempi suunnittelu ja tavoitteiden rajaus niin kirjallisuuden tutkimisen kuin itse työn suorittamisen kannaltakin. Koska insinöörityö ei ole suoraan tiettyyn projektiin suoritettu, itse vertailua hankaloitti tarjolla olevien tekoälymallien määrä ja kriteerien todellisen tarkoituksen puuttuminen. Työn edetessä kuitenkin työn raamit selkeytyivät pikkuhiljaa tehden lopputuloksesta kohtuullisen yhtenäisen kokonaisuuden. Insinöörityön tekeminen oli erittäin opettavainen kokemus, niin valitun aiheen kuin myös tutkimuksen tekemisen kannalta. Tuntemattomampi aihealue työlle antoi siis paljon oppia itse aiheesta, mutta samalla toi itse työn suorittamiseen mutkia matkaan.

Insinöörityön tieto-osuuden pohjana käytettiin hyvin laajasti lähteitä. Suurin osa hyödynnetystä tiedosta on peräisin alan kirjallisuudesta, jonka tukena käytettiin luotettaviksi koettuja internetlähteitä. Julkisina internetin lähteinä käytettiin suurilta osin tunnettujen yrityksien kotisivuja, joten insinöörityön tieto-osuutta voidaan pitää pitkälti luotettavana. Kaikki itse vertailussa käytetty data on peräisin

yhdeltä sivustolta (artificialanalysis.ai), joten vertailun luotettavuus on kiinni sivuston luotettavuudesta. Vain yhden lähteen käyttäminen vertailuaineistoon lisää itse vertailun luotettavuuteen liittyviä riskejä, mutta samalla helpottaa itse vertailua tietojen ollessa keskenään samassa muodossa. Artificialanalysis.ai ilmoittaa itse olevansa johtava itsenäinen tekoälymalleja ja niiden tarjoajia vertaileva taho, mutta tätä väitettä on hankala todentaa. Sivusto kertoo tarjoavansa vertailuanalyysejä tekoälymalleista tuotekehittäjille, tutkijoille, yrityksille ja muille käyttäjille ja tarjoaa sivuillaan kattavat selitykset siitä, miten sivustolla tarjoiltava data on tuotettu. Kyseisen sivuston dataa oli käytetty myös ainakin yhden julkisesti saatavilla olevassa tekoälymalleja käsittelevässä vertailussa. Tämän insinöörityön vertailumallin hyödyntämisen kannalta olisi kuitenkin hyvä varmistaa sivuston luotettavuus sekä ottaa tarpeen mukaan sieltä uusimmat tiedot.

Lähteet

Aditya, J., Gandhar, K. & Vraj, S., 2018. Natural Language Processing. International Journal of Computer Sciences and Engineering (JCSE). Verkkoaineisto: https://www.researchgate.net/publication/325772303_Natural_Language_Processing Luettu 5.1.2025.

Dartmouth College., 2024. n.d. Verkkosivu: https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth Luettu 10.1.2025.

SAS Institute., 2024. n.d. Verkkosivu: https://www.sas.com/fi_fi/insights/analytics/what-is-artificial-intelligence.html#howitworks Luettu 15.12.2024.

Boston Consulting Group, 2024. AI Adoption in 2024: 74% of Companies Struggle to Achieve and Scale Value. Verkkosivu: https://www.bcg.com/press/24october2024-ai-adoption-in-2024-74-of-companies-struggle-to-achieve-and-scale-value Luettu 22.2.2025.

DeepSeek AI, 2024. Intorducing DeepSeek V3. Verkkosivu: https://api-docs.deepseek.com/news/news1226 Luettu 2.4.2025.

Dongare, A.D.; Kharde, R.R.; Kachare, Amit D., 2012. Introduction to Artificial Neural Networks. International Journal of Engineering and Innovative Technology. Verkkoaineisto: https://www.researchgate.net/publication/319903816_AN_INTRODUCTION_TO_ARTIFICIAL_NEURAL_NETWORK Luettu 5.12.2024.

Euroopan Parlamentti, 2020. Mitä tekoäly on ja mihin sitä käytetään? Verkkosivu: https://www.europarl.europa.eu/topics/fi/article/20200827STO85804/mita-tekoaly-on-ja-mihin-sita-kaytetaan Luettu 10.1.2025.

Euroopan parlamentti, 2024. Euroopan parlamentti hyväksyi maailman ensimmäiset tekoälysäännöt. Verkkosivu: https://www.europarl.europa.eu/news/fi/press-room/20240308IPR19015/parlamentti-hyvaksyi-maailman-ensimmaiset-tekoalysaannot Luettu 10.1.2025.

European Commission, 2024. Artificial Intelligence. Verkkosivu: https://digital-strategy.ec.europa.eu/en/policies/artificial-intelligence Luettu 10.1.2025.

Google, 2024. Our next-generation model: Gemini 1.5. Verkkosivu: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note Luettu 2.4.2025.

Google. (2025). Safety settings. Verkkosivu: https://ai.google.dev/gemini-api/docs/safety-settings Luettu 10.4.2025.

Helsingin Yliopisto, n.d. Yleinen kielitiede. Verkkosivu: https://www.helsinki.fi/fi/humanistinen-tiedekunta/tutkimus/tieteenalat/kielten-tutkimus/yleinen-kielitiede Luettu 10.1.2025.

IBM, n.d. What is a transformer model? Verkkosivu: https://www.ibm.com/think/topics/transformer-model Luettu 15.12.2025.

IBM, n.d. What is generative AI? Verkkosivu: https://www.ibm.com/think/topics/generative-ai Luettu 22.2.2025.

Kananen, H., Puolitaival, H., Puntti, S. & Metsola, I., 2019. Tekoäly: bisneksen uudet työkalut. E-kirja: https://bisneskirjasto-almatalent-fi.ezproxy.metropolia.fi/teos/BAX-BBXATCBIED#/kohta:OSA((20)3((20)Teko((e4)ly((20)k((e4)yt((e4)nn((f6)ss((e4)/piste:trz Luettu 30.11.2024.

Michigan Tech University, n.d. What is computer science. Verkkosivu: https://www.mtu.edu/cs/what/ Luettu 18.12.2024.

Nvidia, 2025. Explaining Tokens - the Language and Currency of AI. Verkkosivu: https://blogs.nvidia.com/blog/ai-tokens-explained/ Luettu 24.3.2025.

Nvidia, 2025. How Reasoning Models are transforming Logical AI thinking. Verkkosivu: https://techcommunity.microsoft.com/blog/azuredevcommunityblog/how-reasoning-models-are-transforming-logical-ai-think ing/4373194 Luettu 24.3.2025.

Nvidia, n.d. Large Language Models Explained. Verkkosivu: https://www.nvidia.com/en-us/glossary/large-language-models/ Luettu 2.1.2025.

Ojanperä, T., 2023. Tekoälyn vallankumous: käsikirja. E-kirja: https://bisneskirjasto-almatalent-fi.ezproxy.metropolia.fi/teos/CAHBBXXTBBAEF#/kohta:Teko((e4)lyn((20)vallankumous/piste:thl Luettu 30.11.2024.

Open AI, n.d. Reasoning models. Verkkosivu: https://platform.openai.com/docs/guides/reasoning?api-mode=chat Luettu 24.3.2025.

OpenAI, 2024. Learning to reasong with LLMs. Verkkosivu: https://openai.com/index/learning-to-reason-with-llms/ Luettu 2.4.2025.

Roboflow, 2025. What is a Neural Network? A Deep Dive. Verkkosivu: https://blog.roboflow.com/what-is-a-neural-network/ Luettu 5.1.2025.

Salo, I., 2024. Luova tekoäly työn supervoimana. E-kirja: https://kauppakamaritieto-fi.ezproxy.metropolia.fi/ammattikirjasto/teos/luova-tekoaly-tyon-supervoimana-2024#kohta:Luova((20)teko((e4)ly((20)ty((f6)n((20)supervoimana Luettu 30.11.2024.

SAP, n.d. What is generative AI? Verkkosivu: https://www.sap.com/uk/products/artificial-intelligence/what-is-generative-ai.html Luettu 1.3.2025.

SGU, 2025. A Comparison of Leading AI Models: DeepSeek AI, ChatGPT, Gemini and Perplexity AI. Verkkosivu: https://sgu.ac.id/a-comparison-of-leading-ai-models-deepseek-ai-chatgpt-gemini-and-perplexity-ai/ Luettu 2.4.2025.

Statista, 2025. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028 Verkkosivu: https://www.statista.com/statistics/871513/worldwide-data-created/ Luettu 20.1.2025.

The Guardian, 2025. Small businesses are not all in with artificial intelligence - yet. Verkkosivu: https://www.theguardian.com/business/2025/jan/12/small-businesses-ai Luettu 2.4.2025.

Valtioneuvosto, 2025. EU:n tekoälyasetus: tekoälykäytäntöjen kiellot astuvat voimaan 2.2.2025. Verkkosivu: https://valtioneuvosto.fi/-/1410877/eu-n-tekoalyasetus-tekoalykaytantojen-kiellot-astuvat-voimaan-2.2.2025 Luettu 5.3.2025.

Vellum, 2024. Analysis: OpenAI o1 vs GPT-4o vs Claude 3.5 Sonnet. Verkkosivu: https://www.vellum.ai/blog/analysis-openai-o1-vs-gpt-4o Luettu 20.3.2025.

Wodecki, B., 2025. Behind the DeepSeek hype: Costs, safety, risks and censor-chip explained. Verkkosivu: https://www.capacitymedia.com/article/2ecgy5isrr4777k2yws1s/long-reads/behind-the-deepseek Luettu 5.4.2025.

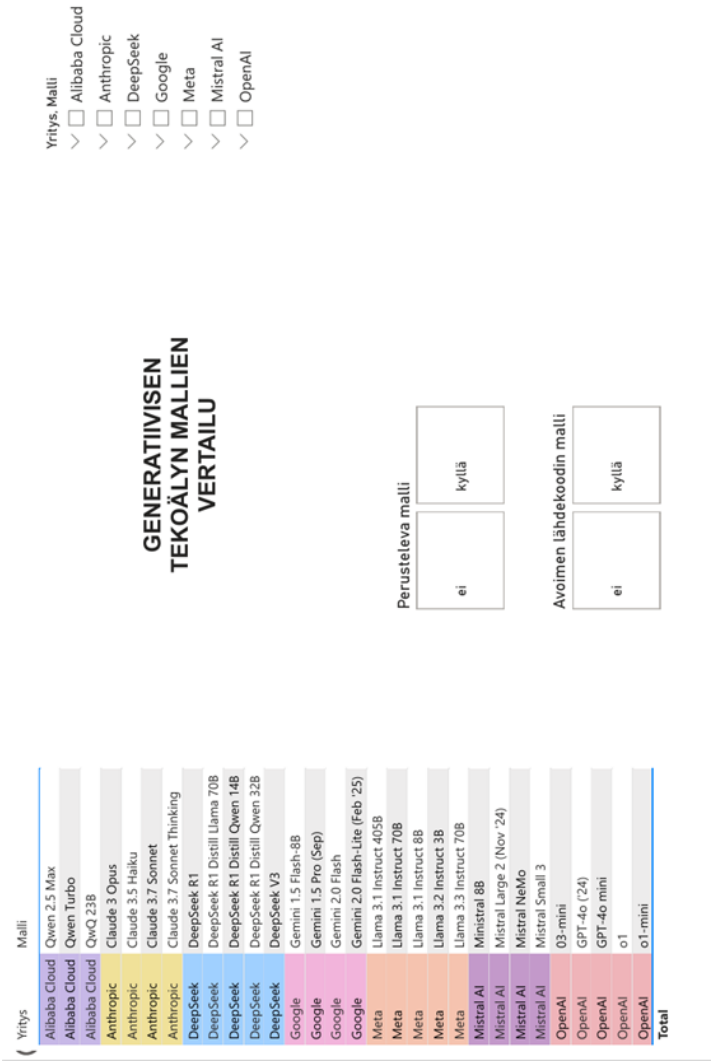
yle.fi, 2025. Suomi on vahvoilla, kun kiinalainen Deepseek mullistaa tekoälykehityksen - luvassa voi olla iso kiihdytys talouskasvuun. Verkkosivu: https://yle.fi/a/74-20146144 Luettu 22.2.2025.

Zhou, Z.-H., 2016. Machine Learning. Peking: Tsinghua University Press. Verkkoaineisto: https://www.scirp.org/reference/referencespapers?referenceid=2987748 Luettu 15.12.2024.

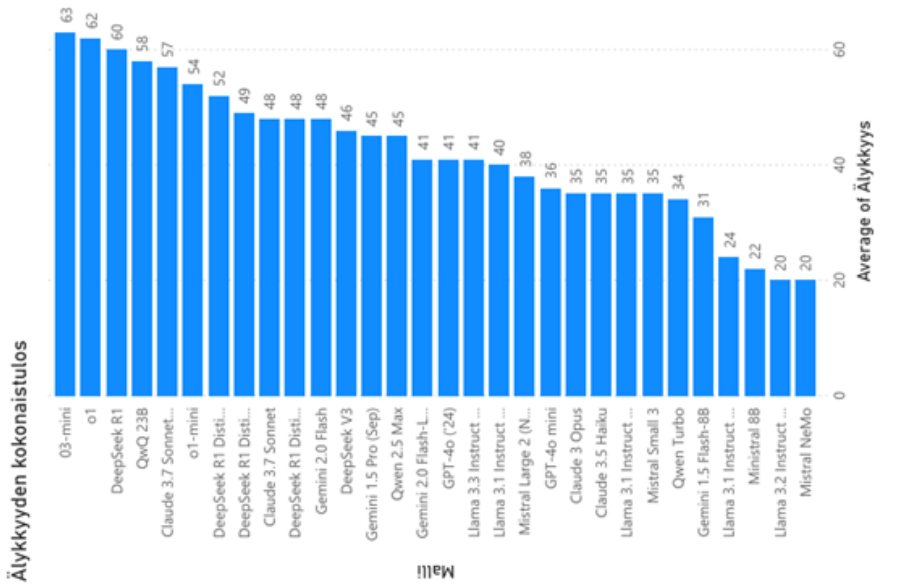
Liitteet

Vertailumallin sivu 1

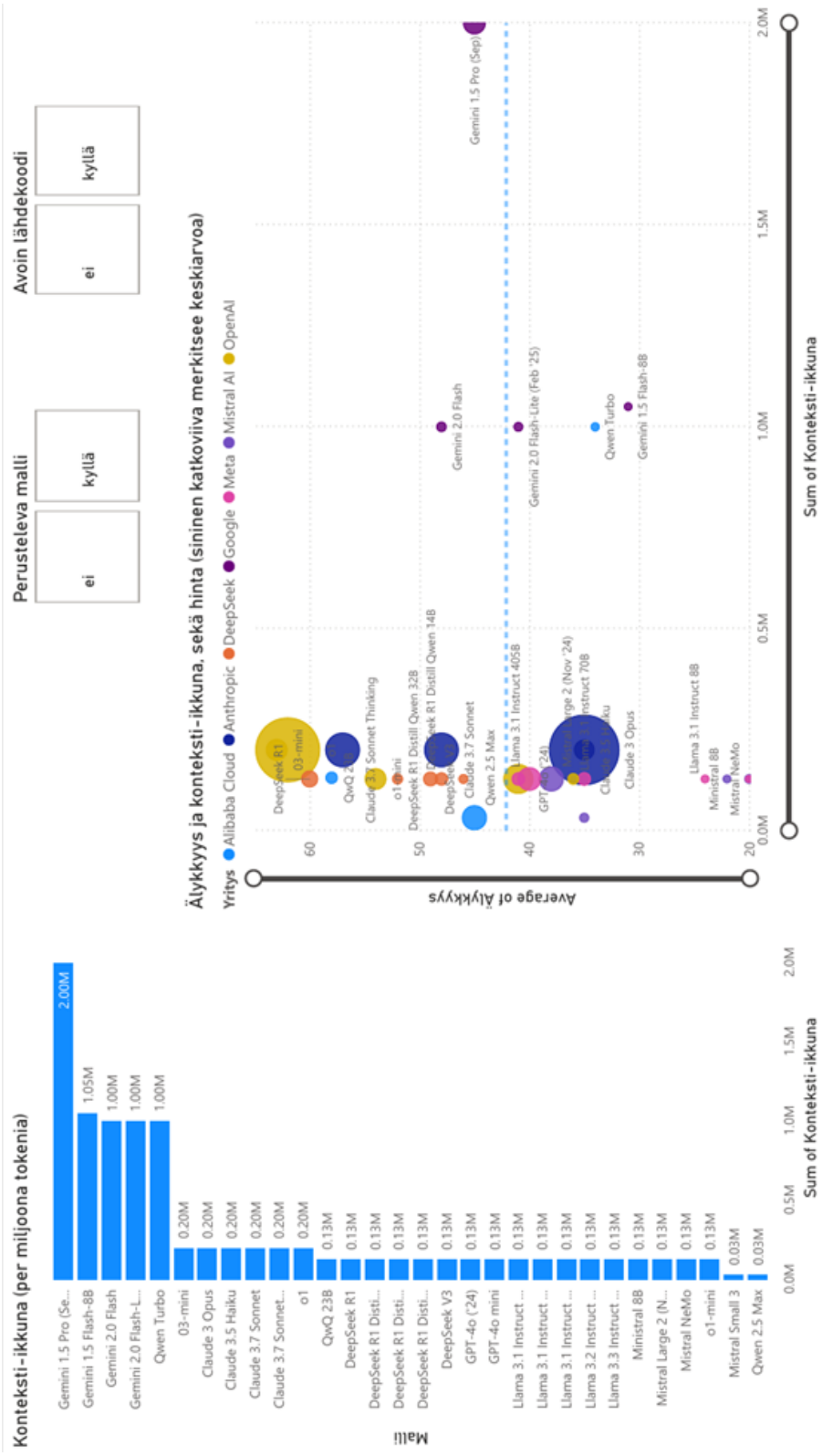
(Vertailun Power BI -tiedosto. Tiedoston kaikki kuusi sivua muokkaamatto-massa muodossa.)



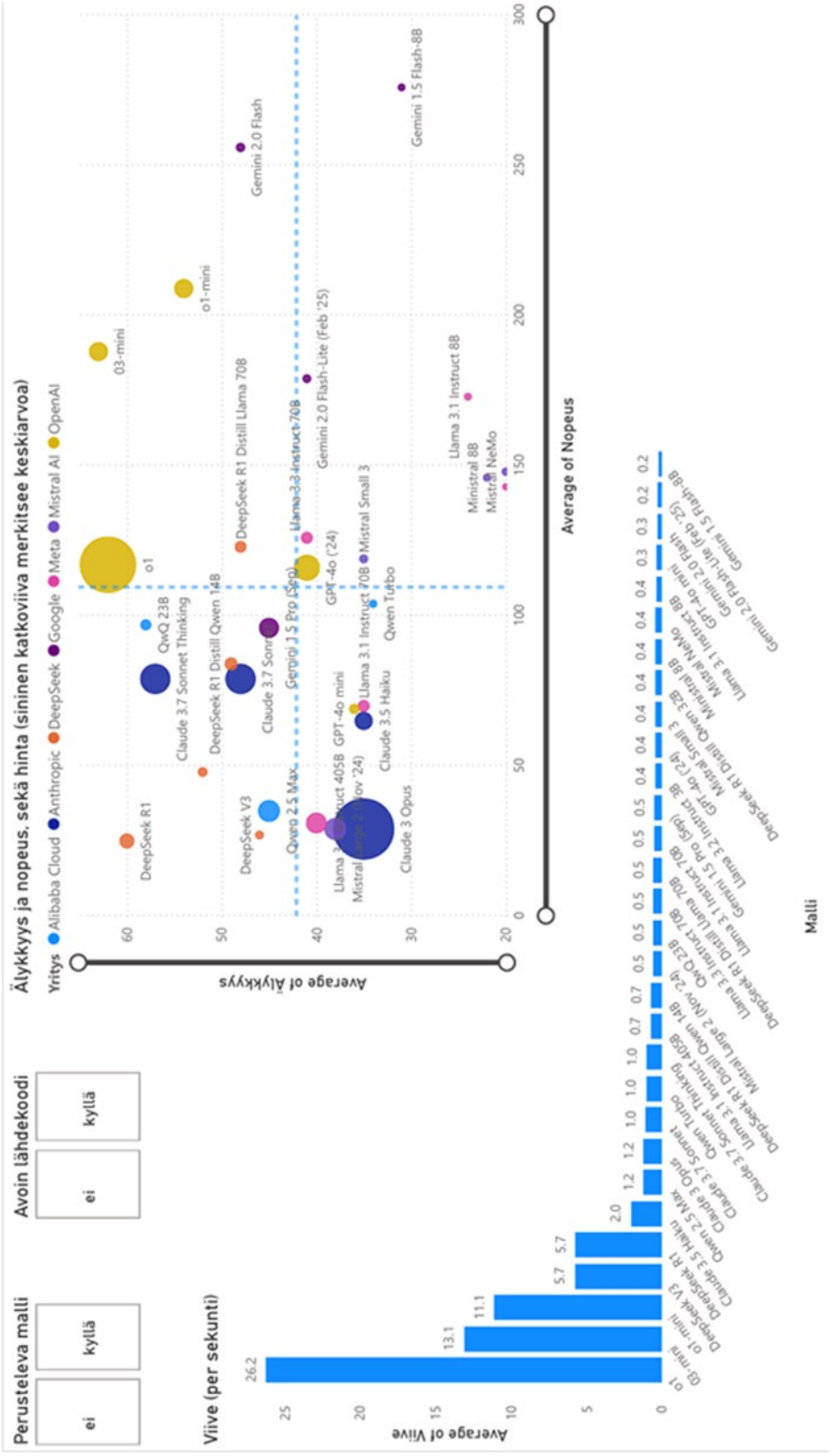
Vertailumallin sivu 2



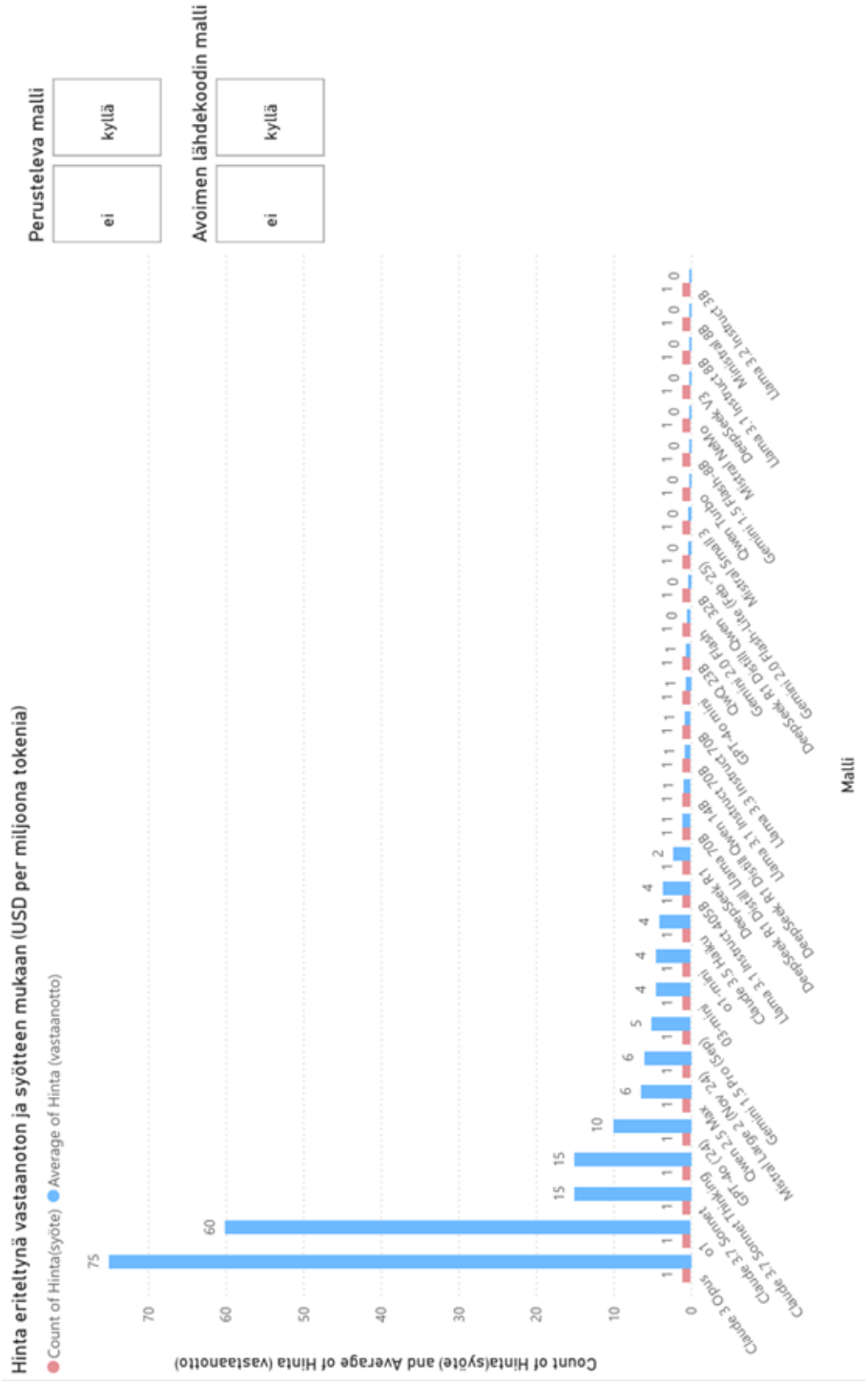
Vertailumallin sivu 3



Vertailumallin sivu 4



Vertailumallin sivu 5



Vertailumallin sivu 6

Yritys	Malli	Average of Älykkyys	Average of Hinta	Average of Kontekstі-ikkuna	Average of Nopeus	Average of Viive	Average of Matematiikka	Sum of Koodaaminen
Alibaba Cloud	Qwen 2.5 Max	45	8.00	32000.00	35.00	1.18	53.00	35.00
Alibaba Cloud	Qwen Turbo	34	0.25	1000000.00	104.00	1.01	46.00	16.00
Alibaba Cloud	QwQ 23B	58	0.95	131000.00	97.00	0.54	87.00	49.00
Anthropic	Claude 3 Opus	35	90.00	200000.00	29.00	1.17	34.00	26.00
Anthropic	Claude 3.5 Haiku	35	4.80	200000.00	65.00	2.02	38.00	29.00
Anthropic	Claude 3.7 Sonnet	48	18.00	200000.00	79.00	1.04	54.00	38.00
Anthropic	Claude 3.7 Sonnet Thinking	57	18.00	200000.00	79.00	1.00	72.00	44.00
DeepSeek	DeepSeek R1	60	2.74	128000.00	25.00	5.70	82.00	49.00
DeepSeek	DeepSeek R1 Distill Llama 70B	48	1.26	128000.00	123.00	0.52	80.00	29.00
DeepSeek	DeepSeek R1 Distill Qwen 14B	49	1.76	128000.00	84.00	0.66	81.00	31.00
DeepSeek	DeepSeek R1 Distill Qwen 32B	52	0.60	128000.00	48.00	0.37	81.00	32.00
DeepSeek	DeepSeek V3	46	0.38	128000.00	27.00	5.74	57.00	36.00
Google	Gemini 1.5 Flash-8B	31	0.19	1050000.00	276.00	0.21	36.00	22.00
Google	Gemini 1.5 Pro (Sep)	45	6.25	2000000.00	96.00	0.49	57.00	31.00
Google	Gemini 2.0 Flash	48	0.50	1000000.00	256.00	0.28	63.00	32.00
Google	Gemini 2.0 Flash-Lite (Feb '25)	41	0.37	1000000.00	179.00	0.24	55.00	22.00
Meta	Llama 3.1 Instruct 405B	40	7.00	128000.00	31.00	0.71	46.00	30.00
Meta	Llama 3.1 Instruct 70B	35	1.33	128000.00	70.00	0.50	41.00	25.00
Meta	Llama 3.1 Instruct 8B	24	0.20	128000.00	173.00	0.35	30.00	12.00
Meta	Llama 3.2 Instruct 3B	20	0.12	128000.00	143.00	0.41	28.00	7.00
Meta	Llama 3.3 Instruct 70B	41	1.29	128000.00	126.00	0.53	54.00	27.00
Mistral AI	Ministral 8B	22	0.20	128000.00	146.00	0.37	30.00	11.00
Mistral AI	Mistral Large 2 (Nov '24)	38	8.00	128000.00	29.00	0.54	42.00	29.00
Mistral AI	Mistral NeMo	20	0.30	128000.00	148.00	0.37	20.00	8.00
Mistral AI	Mistral Small 3	35	0.40	32000.00	119.00	0.39	40.00	24.00
OpenAI	03-mini	63	5.50	200000.00	188.00	13.09	87.00	56.00
OpenAI	GPT-4o ('24)	41	12.50	128000.00	116.00	0.40	45.00	32.00
OpenAI	GPT-4o mini	36	0.75	128000.00	69.00	0.34	45.00	23.00
OpenAI	o1	62	75.00	200000.00	117.00	26.22	85.00	52.00
OpenAI	o1-mini	54	5.50	128000.00	209.00	11.09	77.00	45.00
Total		**42**	**9.07**	**316433.33**	**109.53**	**2.58**	**54.87**	**902.00**