# Holistinen Mestaruus: Viitekehys tekoälyosaamisen arviointiin

## Abstrakti

Globaali tietotyömarkkina on keskellä historiansa merkittävintä rakennemuutosta, jota vauhdittaa generatiivisen tekoälyn (GenAI) ja suurten kielimallien (LLM) eksponentiaalinen kehitys. Se ei ole pelkkä teknologinen päivitys, vaan yleiskäyttöisen teknologian (General Purpose Technology, GPT) aiheuttama shokki, joka muokkaa uudelleen työn taloudellisen arvonluonnin perusteita.

Tekoälyn syvällinen ja kiihtyvä integroituminen työelämään on asettanut organisaatiot perustavanlaatuisen haasteen eteen: ne kohtaavat mittaamisen kriisin, jossa perinteiset keinot eivät enää riitä todentamaan luotettavasti korkean tason tekoälyosaamisen arvoa (vrt. Auzmor 2024; ISACA 2025; Disco 2024). Tämä kriisi ei ole vain tekninen, vaan se koskettaa syvällisesti inhimillisen pääoman arvottamista tilanteessa, jossa koneiden kyvykkyydet kehittyvät eksponentiaalisesti. Perinteiset arviointimenetelmät eivät kykene ratkaisemaan psykometriikan perustavanlaatuista reliabiliteetin ja validiteetin paradoksia (vrt. Borsboom ym. 2004). Tämä paradoksi korostuu entisestään generatiivisen tekoälyn kontekstissa, jossa oikeita vastauksia voi olla ääretön määrä ja prosessin laatu ratkaisee tuloksen arvon.

Standardoidut testit ovat usein psykometrisesti luotettavia mutta sisällöllisesti kapea-alaisia (Wiggins 1998). Laadulliset menetelmät ovat puolestaan valideja, mutta kärsivät usein heikosta reliabiliteetista ja subjektiivisuudesta (Koretz ym. 1994). Tässä artikkelissa esitellään *hybridirubriikki*, uusi kaksitasoinen teoreettinen viitekehys, joka on suunniteltu hallitsemaan tätä jännitettä. Viitekehyksen analyyttinen taso perustuu BARS-asteikkoon (*Behaviorally Anchored Rating Scales*), Bloomin taksonomiaan ja Toulminin argumentaatiomalliin, ja se pyrkii maksimoimaan reliabiliteettia standardoimalla arviointiprosessia. Sitä täydentävä holistinen taso hyödyntää vastakkainasetteluun perustuvaa (adversariaalista) auditointiprosessia, joka on saanut inspiraationsa agenttien välisistä debateista (vrt. Du ym. 2023) ja pyrkii maksimoimaan validiteetin tunnistamalla sääntöjä ylittävä, kontekstisidonnaisen asiantuntijuuden.

Viitekehyksen operatiivinen malli on ”Kognitiivinen Kvoorum”, moniagenttijärjestelmä (MAS), joka toteuttaa tämän kaksitasoisen arvioinnin hyödyntämällä koneoppimisen koosteoppimisen (engl. ensemble learning) periaatteita (vrt. Sagi & Rokach 2018). Malli on täysin operationalisoitu yksityiskohtaisena, vaiheittaisena prosessikuvauksena. Daniel Kahnemanin esittämä kaksoisprosessiteoria (Järjestelmä 1 ja Järjestelmä 2; vrt. Kahneman 2011) ohjaa viitekehyksen strategista kehitystä ja luo vision skaalautuvasta, mutta samalla syvällisesti auditoitavasta arviointijärjestelmästä. Vaikka viitekehyksen prototyyppi on teknisesti toteutettu, siltä puuttuu toistaiseksi empiirinen validointi. Sen keskeisin heikkous on empiirisen näytön puuttuminen. Viitekehys pysyy puhtaasti teoreettisena konstruktiona, kunnes sen keskeinen hypoteesi – jonka mukaan on mahdollista saavuttaa korkea arvioitsijoiden välinen reliabiliteetti monimutkaisessa laadullisessa arvioinnissa – todennetaan muodollisessa pilottitutkimuksessa. Tämä työ tarjoaa metodologisen avauksen korkean panoksen osaamisen arviointiin tekoälyn aikakaudella.

## Luku 1: Strateginen haaste ja metodologinen perusta

Tämä luku perustelee, miksi tekoälyosaamisen luotettava mittaaminen on kriittinen strateginen haaste, ja esittelee viitekehyksen metodologisen perustan. Luvussa kuvataan ensin, miten tekoäly aiheuttaa globaalin taitomurroksen ja millaisen haasteen se luo osaamisen todentamiselle. Tämän jälkeen syvennytään psykometriikan perustavanlaatuiseen ”reliabiliteetin ja validiteetin paradoksiin”, joka on perinteisten arviointimallien ydinongelma. Lopuksi luvussa eritellään keskeisten teorioiden tunnetut rajoitukset, jotka uuden ratkaisun on hallittava.

### 1.1 Strateginen konteksti: Tekoäly ja taitomurros

Tekoälyn integroituminen liiketoimintaprosesseihin on käynnistänyt perustavanlaatuisen taitomurroksen, joka vaikutuksiltaan vertautuu teolliseen vallankumoukseen (Acemoglu & Restrepo 2018; Eloundou ym. 2023). Muutos ei ole vain teknologinen, vaan se muokkaa uudelleen työn taloudellista arvonluontia. Erityisesti suurten kielimallien (LLM) on arvioitu vaikuttavan merkittävästi jopa 80 prosenttiin Yhdysvaltain työvoimasta, kohdistuen nimenomaan korkeaa koulutusta vaativiin asiantuntijatehtäviin (Eloundou ym. 2023). Tutkimus osoittaa, että vähintään 50 % työtehtävistä saattaa muuttua tekoälyn vaikutuksesta noin 19 %:lla työntekijöistä, mikä viittaa siihen, että kyseessä on yleiskäyttöinen teknologia (General Purpose Technology, GPT), jonka vaikutukset läpäisevät kaikki toimialat ja palkkaluokat. Toimialakohtaiset analyysit viittaavat nopeaan ja laaja-alaiseen muutokseen. Esimerkiksi McKinsey Global Instituten ennusteen mukaan merkittävä osa työntekijöiden keskeisistä taidoista muuttuu lähivuosina, ja jopa 30 % nykyisistä työtunneista voitaisiin automatisoida samana ajanjaksona (Hazan ym. 2024)**.** Tämän kehityksen seurauksena arvonluonnin perusta siirtyy rutiinitehtävistä korkeamman tason kognitiivisiin kykyihin, kuten monimutkaiseen ongelmanratkaisuun, kriittiseen ajatteluun ja strategiseen vuorovaikutukseen (OECD 2024)**.**

Kun tekoälyn kyky tuottaa ennusteita ja sisältöä yleistyy ja tulee laajasti saataville, ihmisen arvonluonnin ytimeen nousevat arviointi- ja harkintakyky sekä näiden taitojen hyödyntäminen päätöksentekojärjestelmien strategisessa uudelleensuunnittelussa (Agrawal ym. 2022). Nämä eivät ole pelkästään yleisiä työelämätaitoja, vaan keskeisiä taloudellisia kyvykkyyksiä, joiden arvo perustuu siihen, että ne täydentävät ja ohjaavat tekoälyn tuottamaa ennustekykyä (vrt. Agrawal ym. 2022).

Kun tekoäly painaa ennusteiden hinnan lähelle nollaa, ennusteita täydentävien tekijöiden – kuten datan, arvioinnin ja toiminnan – arvo nousee. Näihin kyvykkyyksiin lukeutuvat esimerkiksi kriittinen validointi, luova synteesi ja eettinen harkinta, jotka kaikki edellyttävät kykyä arvioida ja ohjata tekoälyn tuottamaa informaatiota ja ennusteita (OECD 2024).

Tekoälylukutaidosta on muodostumassa keskeinen strateginen prioriteetti. Erityisesti kehotesuunnittelua (engl. *prompt engineering*) on ehdotettu uudeksi perustaidoksi (Federiakin ym. 2024), ja laajemmin kykyä työskennellä tekoälyn kanssa pidetään yhtenä nopeimmin kasvavista osaamistarpeista globaaleilla työmarkkinoilla (World Economic Forum 2023). Laajemmin augmentaatiokyvyllä (engl. *augmentation capability*) tarkoitetaan kykyä hyödyntää tekoälyä oman suorituskyvyn ja ajattelun laadun parantamiseksi. Tällä kyvykkyydellä on konkreettinen taloudellinen arvo työmarkkinoilla (vrt. Fügener ym. 2025). Työmarkkinat reagoivat muutokseen nopeasti, ja tekoälytaitoja vaativista rooleista maksetaan jo nyt korkeampaa palkkaa (PwC 2024). Tämä osoittaa, että kyseessä ei ole tulevaisuuden visio, vaan nykyhetken taloudellinen realiteetti, joka asettaa organisaatioille paineen tunnistaa, mitata ja kehittää näitä uusia taitoja.

Uusimman sukupolven kielimallien ja niiden kehittyneiden päättelykykyjen myötä tämä tilanne muuttuu teoreettisesta mallista strategisesti merkittäväksi voimavaraksi. Nyt on otollinen hetki hyödyntää sitä, sillä teknologinen kehitys on saavuttanut pisteen, jossa korkeatasoisen arvioinnin vaatima metakognitiivinen arkkitehtuuri on toteutettavissa täysimääräisesti. Konkreettisena esimerkkinä tästä murroksesta ovat Open AI:n o1-mallisarja (OpenAI 2024) ja Googlen Gemini 3.0 (Google DeepMind 2025c), jotka molemmat edustavat arkkitehtonista siirtymää nopeasta hahmontunnistuksesta hitaaseen, deliberatiiviseen päättelyyn.

Muutos on ajankohtainen, sillä aiemmat mallit epäonnistuivat systemaattisesti monimutkaisten ohjeiden noudattamisessa (Wu ym. 2024) ja toimivat luonteeltaan todennäköisyyslaskentaan perustuvina ennustamina, jotka eivät kyenneet muodolliseen kausaaliseen päättelyyn (Chi ym. 2024). Tämä jätti arvioinnin alttiiksi pinnalliselle jäljittelylle ja hallusinaatioille, jotka ovat tunnettu riski kielimallien luotettavuudelle (Huang ym. 2023). Tämän uuden mahdollisuuden taustalla on tekninen siirtymä pelkästä tilastollisesta hahmontunnistuksesta pidennettyä päättelyaikaa (*inference-time compute*) hyödyntäviin ”Deep Think” -arkkitehtuureihin, jotka on koulutettu vahvistusoppimisen avulla generoimaan sisäisiä, iteratiivisia ajatusketjuja (*Chain-of-Thought*) virheiden tunnistamiseksi ja korjaamiseksi ennen lopullisen vastauksen tuottamista (Google DeepMind 2025a; Google DeepMind 2025b).

Tässä artikkelissa esiteltävä moniagenttijärjestelmä valjastaa nyt näiden uusien mallien hitaan ja pohtivan prosessoinnin tuottamaan aitoa, Kahnemanin (2011) kuvaamaa ”Järjestelmä 2” -tason analyysia. Tämä mahdollistaa sen, mikä aiemmin oli haastavaa: uskottavamman falsifioinnin ja syvällisten loogisten virheiden tunnistamisen. On kuitenkin huomattava, että nykyiset mallit suorittavat kausaalista päättelyä (L3) ensisijaisesti kielellisinä approksimaatioina eivätkä muodollisina matemaattisina todistuksina (Chi ym. 2024), mikä huomioidaan viitekehyksen heuristisessa luonteessa. Viitekehys hyödyntää tätä uutta kapasiteettia ratkaistakseen arvioinnin reliabiliteetin ja validiteetin välisen paradoksin (Borsboom ym. 2004) ja tarjoaa keinon tunnistaa luotettavasti sekä sääntöjä noudattavan rutiiniosaamisen että säännöt ylittävän, kontekstisidonnaisen mestaruuden (Dreyfus & Dreyfus 1980)**.**

### 1.2 Ydinongelma: "Mittaamisen kriisi" ja itsearvioinnin haaste

Tämä taitomurros on synnyttänyt organisaatioille keskeisen strategisen esteen, ”mittaamisen kriisin”, jota on kuvattu myös laajemmin kognitiivisten kykyjen ja tekoälyn arvioinnin yhteydessä (vrt. Silva ym. 2025; Cheng 2021). Kriisillä viitataan perustavanlaatuiseen vaikeuteen todentaa tekoälyinvestointien ja -osaamisen todellista arvoa. Organisaatiot investoivat merkittävästi tekoälyyn, mutta ilman luotettavia mittareita investointien todellinen tuotto (ROI) ja vaikutus organisaation kyvykkyyksiin jäävät epäselviksi. Sen tuottaman arvon mittaaminen on edelleen keskeinen haaste (Auzmor 2024; ISACA 2025). Tuoreen raportin mukaan lähes puolella (49 %) organisaatioista on vaikeuksia arvioida ja osoittaa luotettavasti tekoälyhankkeidensa arvoa (ISACA 2025). Tämä epävarmuus ei ole vain akateeminen ongelma, vaan se näkyy konkreettisesti esimerkiksi lakialalla, jossa epävarmuus investointien tuotoista on merkittävä este tekoälyn laajemmalle käyttöönotolle (Wolters Kluwer 2024). Lisäksi vain 24 % alan johtajista kokee, että heidän johtoryhmänsä ovat täysin yksimielisiä tekoälystrategiasta (Wolters Kluwer 2024). Tämä luo negatiivisen kierteen: ilman luotettavaa dataa osaamisen arvosta johto ei kykene perustelemaan investointeja koulutukseen ja teknologiaan, mikä johtaa aliresursointiin ja heikentää strategisten aloitteiden uskottavuutta (Disco 2024). Ongelmaa syventää se, että luotettava itsearviointi on tunnetusti haastavaa. Osaamisen onnistunut arviointi edellyttää, että yksilö kykenee metakognitiivisesti tunnistamaan oman osaamisensa puutteet (Kruger & Dunning 1999). Tämä taito on usein heikosti kehittynyt erityisesti matalammalla osaamistasolla. Tämä Dunning–Kruger-vaikutuksena tunnettu havainto luo systemaattisen epäsuhdan havaitun ja todellisen osaamisen välille. Tämä epäsuhta on ominaista yksilötasolla, ja se ilmenee tutkitusti myös organisaatiotasolla, missä kokonaiset tiimit saattavat yliarvioida digitaalisen kypsyytensä (ks. esim. Nold & Michel 2022). Koska organisaatio ei voi luottaa pelkkään itsearviointiin, tarvitaan objektiivista, ulkoiseen todistusaineistoon perustuvaa validointiprosessia. Tämä on välttämätöntä, sillä luotettava arviointikyky edellyttää vertailua ulkoisiin viitepisteisiin (Sadler 1989), mikä auttaa ohittamaan inhimilliset harhat.

### 1.3 Metodologinen perushaaste: Reliabiliteetin ja validiteetin paradoksi

Vaikka tarve objektiiviselle mittaamiselle on ilmeinen, se kohtaa välittömästi psykometriikan perustavanlaatuisen haasteen: arvioinnin reliabiliteetti (*reliability*) ja validiteetti (*validity*) ovat jännitteisessä suhteessa keskenään (vrt. Borsboom ym. 2004). Nämä kaksi käsitettä ovat minkä tahansa mittausprosessin laadun kulmakiviä (vrt. Cohen ym. 1996):

1. **Reliabiliteetti** viittaa mittauksen johdonmukaisuuteen ja toistettavuuteen (AERA, APA & NCME 2014). Keskeinen kysymys on, saavatko eri arvioijat (tai sama arvioija eri aikoina) saman tuloksen samasta aineistosta. Korkea reliabiliteetti on välttämätöntä, jotta arviointi olisi oikeudenmukaista, ennustettavaa ja oikeudellisesti puolustettavaa.
2. **Validiteetti** viittaa siihen, mittaako arviointi sitä, mitä sen on tarkoitus mitata (AERA, APA & NCME 2014). Tekoälyosaamisen kontekstissa tavoitteena on mitata abstrakteja ja monimutkaisia kognitiivisia taitoja, kuten kriittistä ajattelua, luovaa ongelmanratkaisua ja strategista harkintaa, eikä ainoastaan mekaanista prosessien noudattamista tai ulkoa opeteltua tietoa (Wiggins 1998).

Paradoksi syntyy siitä, että näiden kahden tavoitteen välillä on usein sovittamaton jännite. Korkeaa reliabiliteettia tavoittelevat menetelmät, kuten standardoidut ja tiukasti strukturoidut testit (esimerkiksi monivalinnat), ovat usein liian kapea-alaisia eivätkä onnistu mittaamaan monimutkaisia taitoja validilla tavalla (Wiggins 1998). Wigginsin (1998) mukaan tällaiset menetelmät mittaavat usein ensisijaisesti irrotettuja perustaitoja ja ulkoa muistamista, eivätkä onnistu tavoittamaan "intellektuaalista suorituskykyä" tai aitoa osaamista, joka vaatii tiedon soveltamista monimutkaisissa, autenttisissa konteksteissa (ks. myös Shafiyeva 2021; David 2019.; FairTest 2012). Toisaalta korkeaa validiteettia tavoittelevat menetelmät, kuten avoin laadullinen portfolioarviointi, ovat usein subjektiivisia ja kärsivät heikosta reliabiliteetista (Koretz ym. 1994; vrt. Center for Innovative Teaching & Learning 2025). Tämä puolestaan tekee niistä vaikeasti skaalautuvia organisaatiokontekstissa, missä arvioinnin on oltava paitsi syvällistä myös vertailukelpoista. Tämä jännite on myös tekoälyavusteisen arvioinnin ytimessä (Bulut ym. 2024). Se asettaa keskeisen suunnitteluhaasteen: miten rakentaa järjestelmä, joka on riittävän systemaattinen ja objektiivinen soveltuakseen koneelliseen analyysiin (korkea reliabiliteetti), mutta samalla riittävän joustava ja syvällinen tunnistamaan aidon kognitiivisen mestaruuden (korkea validiteetti).

Tässä artikkelissa esiteltävässä viitekehyksessä käytettävä kolmiosainen todistusaineisto (keskusteluhistoria, lopputuote, reflektiodokumentti) ei ole portfolio perinteisessä merkityksessä. Se jakaa kuitenkin portfolion keskeisimmät psykometriset piirteet, kuten tavoitteellisuuden, moniosaisuuden ja reflektiivisyyden (Paulson ym. 1991). Koska kyseessä on laadullinen ja asiantuntija-arviointia vaativa kokonaisuus, se rinnastuu metodologisesti portfolioarviointiin ja sen tunnettuihin haasteisiin. Laaja tutkimusnäyttö osoittaa, että portfolioarvioinnin keskeisin psykometrinen heikkous on arvioitsijoiden välisen yhdenmukaisuuden (arvioitsijareliabiliteetin) matala taso (Baume & Yorke 2002). Ilman tarkkaa strukturointia ja selkeitä arviointikriteerejä, eri arvioijat – olivatpa ne ihmisiä tai algoritmeja – kiinnittävät huomiota eri asioihin, mikä lisää tulkinnanvaraisuutta (vrt. Jonsson & Svingby 2007).

Metodologinen kehitys kohti kognitiivisten prosessien arviointia syventää tätä paradoksia entisestään. Tämä siirtymä vaatii luotettavia ja validoituja työkaluja, jotka täyttävät psykometriikan standardit (AERA, APA & NCME 2014). Tällaisten työkalujen kehittäminen subjektiivisten kognitiivisten ilmiöiden arvioimiseksi on kuitenkin osoittautunut metodologisesti erittäin haastavaksi (vrt. Messick 1989). Siirtymä kohti kognitiivisten prosessien arviointia ei siten ainoastaan vaikeuta reliabiliteetin varmistamista, vaan myös aktiivisesti voimistaa paradoksin jännitettä, koska monimutkaisten taitojen mittaaminen on luontaisesti haastavampaa kuin yksinkertaisten tietojen (Shavelson 2013). Vaikka kognitiiviset taksonomiat, kuten Bloomin malli (Anderson & Krathwohl 2001), tarjoavat tarpeellisen rakenteen, on tärkeää tunnustaa niiden rajoitukset. Kriitikot huomauttavat, että tällaiset mallit voivat esittää osaamisen staattisena ja atomistisena hierarkiana, joka ei täysin tavoita aidon asiantuntijuuden integroitua ja dynaamista luonnetta (Lane 2013; vrt. Dreyfus & Dreyfus 1980). Tämä asettaa seuraavassa luvussa esiteltävälle arkkitehtoniselle ratkaisulle – hybridirubriikille – entistäkin suurempia vaatimuksia. Sen on kyettävä hallitsemaan tätä voimistunutta jännitettä tavalla, joka on sekä teoreettisesti vankka että käytännössä toimiva. Hybridirubriikki ja sen operatiivinen toteutus ”Kognitiivinen Kvoorum” on kehitetty juuri tämän hypoteesin testaamiseksi. On kuitenkin olennaista ymmärtää, että vaikka tämän prototyypin logiikka on toteutettu, empiirinen näyttö sen käytännön toimivuudesta tai kyvystä ratkaista tämä paradoksi puuttuu. Koko viitekehys edustaa tässä vaiheessa ainoastaan testattavaksi ehdotettua, teknisesti toteutettua mutta validoimatonta ratkaisumallia.

### 1.4 Eturistiriidat

Kirjoittajat ilmoittavat seuraavasta eturistiriidasta: Tässä esitelty tutkimus on saanut rahoitusta ja on osa laajempaa tutkimus- ja kehityshanketta, joka tunnetaan nimellä ”Promptaptitude”. Kirjoittajilla on taloudellinen intressi, joka liittyy viitekehyksen potentiaaliseen kaupallistamiseen. Pyrimme hallitsemaan tätä eturistiriitaa noudattamalla tiukkoja tieteellisiä menetelmiä ja avointa raportointia. Tämän avoimuuden tarkoituksena on vahvistaa esitettyjen argumenttien ja riskianalyysin tieteellistä uskottavuutta. Avoimuus ei ole ainoastaan eettinen velvoite, vaan strateginen toimi, joka pyrkii rakentamaan luottamusta esitettyä viitekehystä kohtaan ja erottamaan sen tieteellisen panoksen sen kaupallisesta potentiaalista.

## Luku 2: Hybridirubriikin arkkitehtuuri

Tämä luku esittelee hybridirubriikin arkkitehtuurin. Kyseessä on uusi arviointiviitekehys, joka on suunniteltu vastaamaan tekoälyn aikakauden monimutkaisten taitojen mittaamisen perustavanlaatuisiin haasteisiin. Luvun rakenne etenee loogisesti ratkaisun ydinperiaatteesta sen arkkitehtuurin kahteen päätasoon ja niitä ohjaavaan hallintamalliin. Lisäksi luvussa käsitellään teknistä toteutuskehystä sekä turvallisuusmallia, joka varmistaa koko järjestelmän eheyden ja reliabiliteetin.

### 2.1 Ratkaisun Periaate: Kaksitasoinen Vastaus Mittaamisen Paradoksiin

Tämä viitekehys tarjoaa edellä kuvattuun reliabiliteetin ja validiteetin paradoksiin arkkitehtonisen vastauksen: hybridirubriikin. Sen keskeisenä suunnitteluperiaatteena ei ole pyrkiä ratkaisemaan tätä perustavanlaatuista jännitettä tai löytää täydellistä kompromissia, vaan tunnustaa se ja rakentaa järjestelmä, joka hallitsee jännitettä tietoisesti. Sen sijaan, että arkkitehtuuri olisi yhtenäinen monoliitti, vaan se on tarkoituksellisesti kaksitasoinen. Se institutionalisoi paradoksin luomalla kaksi erillistä, toisiaan täydentävää arviointitasoa, joista kumpikin on optimoitu eri päämäärään:

1. **Analyyttisen tason** tavoitteena on maksimoida reliabiliteetti. Se luo järjestelmälle systemaattisen, sääntöpohjaisen ja auditoitavan selkärangan, joka varmistaa arvioinnin johdonmukaisuuden ja toistettavuuden. Se toimii "ankkurina", joka estää arvioinnin valumisen täydelliseen subjektiivisuuteen.
2. **Holistisen tason** tavoitteena on maksimoida validiteetti. Se on joustava mekanismi, jonka tehtävänä on tunnistaa aito, kontekstisidonnainen ja sääntöjä ylittävä osaaminen, jota analyyttinen taso ei kykene tavoittamaan. Kuten Abstraktissa todetaan, tämä taso soveltaa adversiaalisia menetelmiä (vrt. Du ym. 2023), jotka toteutetaan tässä prototyypissä strukturoituna auditointiprosessina (ks. Luku 2.4.1).

Tämä kaksitasoinen lähestymistapa on enemmän kuin tekninen ratkaisu; se on metodologinen ja filosofinen kannanotto. Se edustaa ”metodologista nöyryyttä” – avointa tunnustusta siitä, ettei mikään yksittäinen menetelmä voi yksinään tavoittaa monimutkaisen inhimillisen osaamisen koko kirjoa (vrt. Johnson & Onwuegbuzie 2004, jotka perustelevat vastaavaa pragmaattista lähestymistapaa monimenetelmätutkimuksessa). Viitekehyksen keskeinen hypoteesi on, että järjestelmän älykkyys ei synny kummastakaan tasosta yksinään, vaan niiden hallitusta vuorovaikutuksesta. Ilmiö tunnetaan koneoppimisessa koosteoppimisen (ensemble learning) hyötynä, jossa monimuotoisten arviointimekanismien yhdistäminen vähentää kokonaisvirhettä tehokkaammin kuin yksittäinen optimoitu malli (vrt. Sagi & Rokach 2018). Sagi ja Rokach (2018) osoittavat, että yhdistämällä useita malleja – tai tässä tapauksessa useita arviointiagentteja – voidaan kompensoida yksittäisten mallien heikkouksia ja saavuttaa tarkempi ennuste tai arvio. Kognitiivinen Kvoorum soveltaa tätä periaatetta siten, että se jakaa arviointitehtävän erikoistuneille agenteille, jolloin vältetään yksittäisen LLM-mallin vinoumat.

### 2.2 Arvioinnin Kohde: Kolmiosainen Todistusaineisto

Hybridirubriikki kohdistaa arviointinsa standardoituun, kolmiosaiseen todistusaineistoon, joka on suunniteltu tavoittamaan sekä kognitiivinen prosessi että sen tulos:

* **Keskusteluhistoria**: Autenttinen tallenne käyttäjän ja tekoälyn välisestä vuorovaikutuksesta. Tämä mahdollistaa prosessin (Analyysi) suoran havainnoinnin. Se paljastaa, onko käyttäjä vain passiivisesti hyväksynyt tekoälyn tuotokset vai ohjannut sitä aktiivisesti ja iteratiivisesti.
* **Lopputuote**: Prosessin aikana tuotettu konkreettinen artefakti. Tämä mahdollistaa lopputuloksen laadun (Synteesi) arvioinnin.
* **Reflektiodokumentti**: Käyttäjän jälkikäteen tuottama analyysi omasta prosessistaan. Tämä on ensisijainen lähde metakognition ja päättelyn laadun (Arviointi ja Argumentaatio) arvioinnissa, sillä metakognitiivinen säätely ja tieto tulevat näkyviksi juuri oman kognitiivisen prosessin sanallistamisen kautta (vrt. Flavell 1979). Flavellin uraauurtava työ metakognition parissa osoittaa, että kyky valvoa ja säädellä omia kognitiivisia prosesseja on korkean tason oppimisen ja ongelmanratkaisun ydin. Reflektiodokumentin tehtävä on pakottaa käyttäjä eksplisiittisesti sanoittamaan nämä prosessit, jolloin ne tulevat Kognitiivisen Kvoorumin arvioitaviksi.

Kuten Luvussa 1.3 todetaan, tämä kokonaisuus rinnastuu metodologisesti portfolioarviointiin (Paulson ym. 1991).

### 2.3 Arkkitehtuurin Ensimmäinen Taso: Analyyttinen Taso ja Sen Metodologiset Rajoitteet

Hybridirubriikin arkkitehtuurin ensimmäinen taso on sen analyyttinen taso, joka konkretisoituu Kognitiiviseksi Arviointimatriisiksi. Tämä matriisi muodostaa koko arviointiprosessin systemaattisen ja auditoitavan selkärangan. Sen ensisijainen tavoite on varmistaa arvioinnin reliabiliteetti (*reliability*) – mittauksen johdonmukaisuus ja toistettavuus (AERA, APA & NCME 2014). Matriisin suunnittelu perustuu validoituihin kognitiivisen suorituskyvyn ja argumentaation viitekehyksiin. Se hyödyntää Bloomin taksonomiaa (Anderson & Krathwohl 2001) kognitiivisten taitotasojen (analyysi, arviointi, synteesi) erittelyyn ja Toulminin argumentaatiomallia (Toulmin 2003) päättelyn laadun systemaattiseen arviointiin. Rakenteellisesti matriisi on BARS-asteikko, joka sitoo arviointitasot konkreettisiin kuvauksiin (Smith & Kendall 1963). Tämä Kognitiivinen Arviointimatriisi muodostaa arvioinnin normatiivisen perustan.

Tämä metodologinen valinta sisältää kuitenkin tietoisesti hyväksyttyjä rajoitteita, jotka tekevät hybridimallin toisesta tasosta välttämättömän. BARS-menetelmiä on perinteisesti kehitetty tavoitteena parantaa reliabiliteettia siten, että arviointitasot ankkuroidaan konkreettisiin käyttäytymiskuvauksiin (Moskal 2000; Smith & Kendall 1963). Niiden todellinen psykometrinen ylivoimaisuus muihin menetelmiin nähden on kuitenkin kyseenalaistettu (Jacobs ym. 1980). Akateemiset arviot ovat todenneet, että BARS-asteikot eivät kvantitatiivisesti arvioituna ole välttämättä parempia kuin muutkaan menetelmät (Jacobs ym. 1980), ja eräissä vertailuissa ne ovat jopa osoittaneet heikompaa arvioitsijoiden välistä reliabiliteettia kuin perinteiset summatiiviset asteikot (Kinicki ym. 1985). Lisäksi niihin liittyy tutkimuskirjallisuudessa laajasti tunnistettuja psykometrisiä ja käytännöllisiä haasteita. Niiden kehittäminen on resurssi-intensiivistä, aikaa vievää ja kallista (Morgeson ym. 2007). Lisäksi ne voivat olla joustamattomia muuttuvissa työrooleissa, jotka vaativat jatkuvaa päivitystä (Levine ym. 1988). Tämä joustamattomuus ja vaatimus määritellä spesifisiä käyttäytymismalleja voivat johtaa siihen, että ne yksinkertaistavat liikaa monimutkaisia, luovia tai strategisia tehtäviä (Klieger ym. 2018), kuten ongelmanratkaisua ja luovuutta, joita on vaikea kuvata spesifisinä, havaittavina käyttäytymisinä. Juuri tämä BARS-mallien ankkureiden "äärimmäinen spesifisyys" (engl. *extreme specificity*) voi johtaa kognitiivisen vaatimustason latistumiseen, ja se onkin tunnistettu haasteeksi arvioijille, sillä se rajoittaa niiden soveltuvuutta abstraktimpien ominaisuuksien mittaamiseen (Klieger ym. 2018).

Lisäksi BARS-asteikkojen reliabiliteetin on osoitettu olevan parhaimmillaankin vain kohtalainen tai jopa rajoitettu (engl. *limited reliability*) juuri niissä konteksteissa, jotka vaativat monimutkaisten, ei-teknisten taitojen arviointia, kuten lääketieteen opiskelijoiden arvioinnissa (Kim ym. 2022). Tämän vuoksi analyyttisen tason toteuttaminen BARS-mallina ei ole ratkaisu reliabiliteettiongelmaan, vaan tietoinen kompromissi, joka tuo rakenteen arviointiin mutta jättää merkittävän osan varianssista selittämättä. Juuri nämä menetelmän sisäsyntyiset rajoitukset tekevät holistisen tason välttämättömäksi. Kuten tämän viitekehyksen "Metodologisen nöyryyden mandaatissa" (ks. Luku 2.5.3) todetaan, kognitiivisiin prosesseihin (Bloom, Toulmin) perustuva matriisi mittaa tehokkaasti prosessin loogisuutta (Pätevyys), mutta ei välttämättä tunnista aitoa asiantuntijuutta (Mestaruus), joka ilmenee sääntöjen strategisena rikkomisena tai luovana soveltamisena. Nämä taidot ovat usein kontekstisidonnaisia, implisiittisiä (Polanyi 1966) ja vaikeasti etukäteen määriteltäviä (vrt. Dreyfus & Dreyfus 1980). Tässä arkkitehtuurissa Kognitiivisen Arviointimatriisin rajoitteet eivät ole suunnitteluvirheitä, vaan tarkoituksellinen suunnitteluraja. Matriisi on tarkoituksella jäykkä, koska sen tehtävä on luoda vakaa, ennustettava ja koneellisesti käsiteltävä perusta arvioinnille. Sen jäykkyys määrittelee täsmällisesti sen ongelman, jonka arkkitehtuurin toisen tason – holistisen tason – on ratkaistava. Tämän perustavanlaatuisen metodologisen jännitteen vuoksi viitekehys ei voi nojata pelkkään analyyttiseen tasoon, vaan vaatii toisen, dynaamisemman tason.

On kuitenkin tärkeää tiedostaa, että tämän viitekehyksen Kognitiivinen Arviointimatriisi (Taulukko 1) on tietoisesti suunniteltu yksityiskohtaisemmaksi kuin perinteiset BARS-asteikot. Erityisesti korkeimman tason (Taso 4). Tämä valinta pyrkii lisäämään validiteettia, mikä on strategisesti välttämätöntä generatiivisen tekoälyn kontekstissa, missä päteviä ratkaisupolkuja voi olla lukemattomia, eikä lineaarinen prosessi ole itseisarvo. Samalla tämä monimutkaisuus lisää kognitiivista kuormaa ja tulkinnanvaraisuutta, mikä puolestaan luo tunnetun riskin arvioitsijareliabiliteetin (engl. Inter-Rater Reliability, IRR) heikkenemiselle (Wolf & Stevens 2007; Jonsson & Svingby 2007).

Kognitiivisen Kvoorumin arkkitehtuuri on suunniteltu hallitsemaan tätä lisääntynyttä jännitettä. Analyytikko-agentti ankkuroi väitteet todistusaineistoon, Loogikko-agentti purkaa argumentin rakenteen (Luku 2.4.2) ja Kriitikko-agentti falsifioi argumentin (Luku 2.4.3). Tämän prosessin tavoitteena on varmistaa, että monimutkaisten kriteerien soveltaminen pysyy ankkuroituna objektiiviseen todistusaineistoon, mikä vähentää subjektiivista tulkintaa ja tukee reliabiliteettia.

*Taulukko 1. Kognitiivinen Arviointimatriisi (4-portainen BARS-asteikko)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Kriteeri (Kognitiivinen ulottuvuus, vrt. Anderson & Krathwohl 2001)** | **Taso 4 (Erinomainen / Strateginen)** | **Taso 3 (Hyvä / Omaperäinen)** | **Taso 2 (Kohtalainen / Reaktiivinen)** | **Taso 1 (Heikko / Puutteellinen)** |
| **Analyysi ja Prosessin Tehokkuus** (Bloom: Analyze) | Prosessi on strateginen ja ennakoiva. Käyttäjä on purkanut ongelman osiin ja ohjannut tekoälyä ennaltaehkäisevästi (esim. roolilla, analyysipakolla, rajoitteilla) minimoiden turhat iteraatiot. TAI Prosessi osoittaa poikkeuksellista strategista ketteryyttä. Käyttäjä on tehnyt prosessin aikana merkittävän oivalluksen, joka on perustavanlaatuisesti ja todistettavasti muuttanut tehtävän suuntaa tai laatua. Reflektio perustelee tämän käännekohdan strategisen merkityksen. | Prosessi on tehokas ja systemaattinen. Käyttäjä on tunnistanut ongelman ja ohjannut tekoälyä reaktiivisesti mutta johdonmukaisesti (esim. pyytämällä parannuksia). Reflektio analysoi käännekohdat. | Prosessi on hajanainen tai puhtaasti reaktiivinen. Käyttäjä on reagoinut tekoälyn vastauksiin, mutta selkeä analyyttinen strategia puuttuu tai on heikko. Iteraatioita voi olla useita ilman merkittävää laadullista parannusta. Reflektio on pääosin kuvaileva. | Prosessi on tehoton ja epälooginen. Käyttäjä ei ole kyennyt ohjaamaan tekoälyä kohti tavoitetta. Reflektio on puutteellinen tai puuttuu kokonaan. |
| **Arviointi ja Argumentaatio** (Anderson & Krathwohl 2001: Evaluate; Toulmin 2003) | Suoritus osoittaa poikkeuksellista arviointikykyä. Käyttäjä on haastanut tekoälyn päättelyä (esim. eettisesti tai loogisesti). TAI (Asiantuntijan Hyväksyntä): Käyttäjä osoittaa poikkeuksellista arviointikykyä perustelemalla reflektiossa vankasti, miksi lyhyt prosessi ja tekoälyn tuotoksen suora hyväksyntä oli strategisesti optimaalinen valinta. Molemmissa tapauksissa Toulmin: Reflektio sisältää vankan, itsekriittisen argumentin, joka on täysin virheetön. | Suoritus osoittaa korkeaa arviointikykyä. Käyttäjä on korjannut tekoälyn tuotoksia. Toulmin: Reflektio sisältää vahvan argumentin: Selkeä Väite, Vankat Perusteet (viittaukset Keskusteluhistoriaan) ja Eksplisiittinen Oikeutus. (Tämä sisältää myös tapaukset, joissa lyhyt prosessi perustellaan reflektiossa vahvalla asiantuntija-argumentilla.) | Suoritus osoittaa perustason arviointikykyä. Käyttäjä on tehnyt pieniä korjauksia. Toulmin: Reflektio sisältää argumentin aihion: Väite on esitetty, mutta Perusteet ovat yleisiä ja/tai Oikeutus on implisiittinen tai heikko. | Suoritus ei osoita arviointikykyä. Tekoälyn tuotoksia on käytetty sellaisenaan. Toulmin: Reflektio ei sisällä tunnistettavaa, pätevää argumenttia ja/tai se on kokonaisuudessaan faktuaalisesti virheellinen tai harhaanjohtava. |
| **Synteesi ja Luovuus** (Bloom: Create) | Lopputuote osoittaa strategista synteesiä. Käyttäjä on luonut kokonaan uutta, omaperäistä lisäarvoa, jota tekoäly ei ehdottanut (esim. lisännyt uuden vastuullisuusosion). Reflektio perustelee tämän strategisen lisäarvon. | Lopputuote on omaperäinen synteesi. Käyttäjä on parannellut ja viimeistellyt tekoälyn tuotosta omalla perustellulla panoksellaan (esim. lisännyt hyötyjä, vaihtanut otsikon). Reflektio perustelee muutokset. | Lopputuote on pääosin kooste tekoälyn tuottamasta materiaalista. Muutokset ovat kielellistä viimeistelyä tai pieniä lisäyksiä. Reflektio kuvaa tehdyt muutokset pinnallisesti. | Lopputuote on suora kopio tekoälyn tuottamasta materiaalista ilman omaa panosta. Reflektio on puutteellinen tai puuttuu. |

### 2.4 Arkkitehtuurin Toinen Taso: Holistinen Taso ja Sen Teoreettiset Peruspilarit

Hybridirubriikin toinen arkkitehtoninen taso on sen holistinen taso. Se on suunniteltu nimenomaan kompensoimaan analyyttisen tason jäykkyyttä ja ratkaisemaan validiteettiin liittyvä haaste. Sen tehtävänä on tunnistaa aito asiantuntijuus, joka usein ylittää tai jopa tietoisesti rikkoo ennalta määriteltyjä sääntöjä paremman lopputuloksen saavuttamiseksi. Tämä monimutkainen ja vivahteikas analyysi vaatii erikoistuneen mekanismin, joka on tässä viitekehyksessä Kognitiivinen Kvoorum.

#### 2.4.1 Kognitiivinen Kvoorum: Kognitiiviseen Työnjakoon Perustuva Moniagenttijärjestelmä

Kognitiivinen Kvoorum on moniagenttijärjestelmä (MAS) (vrt. Supianto ym. 2023), joka perustuu kognitiiviseen työnjakoon. MAS on laskennallinen malli, joka koostuu useista autonomisista ja vuorovaikuttavista agenteista (Wooldridge 2009). Tämä järjestelmä perustuu kognitiiviseen työnjakoon. Järjestelmä ei nojaa yhteen monoliittiseen tekoälymalliin, vaan jakaa monimutkaisen holistisen analyysitehtävän erillisiin, teoreettisesti johdettuihin ja erikoistuneisiin kognitiivisiin rooleihin (Guo ym. 2024). Näiden roolien suunnittelu perustuu argumentaatioteoriaan ja tieteenfilosofiaan (ks. Luvut 2.4.2 ja 2.4.3). Kvoorum koostuu yhdeksästä (9) erikoistuneesta agentista. Vartija-agentti toimii järjestelmän turvaporttina (ks. Luku 2.6.1), ja kahdeksan muuta agenttia suorittavat varsinaisen sisällöllisen analyysin ja synteesin.

Tämä arkkitehtuuri on itsessään verrattavissa tieteellisen menetelmän soveltamiseen (Cheng 2001). Nykyisessä prototyyppivaiheessa järjestelmä toteuttaa ”sekventiaalisen auditointiketjun” (*Sequential Audit Chain*), jossa agentit prosessoivat informaatiota peräkkäin kumuloituvassa prosessissa. Prosessi alkaa todistusaineiston keräämisestä ja ankkuroinnista (empiirinen havainnointi; *Analyytikko*), etenee jäsennellyn argumentin muodostamiseen (hypoteesin luominen; *Loogikko*), minkä jälkeen argumentti altistetaan systemaattiselle kumoamisyritykselle (falsifiointi; *Kriitikkoryhmä*). Lopuksi tulokset kootaan yhteen synteesiksi (*Tuomari*), jossa tuodaan esiin myös analyysiin liittyvä epävarmuus (*XAI-Raportoija*). Vaikka malli ei tässä vaiheessa toteuta reaaliaikaista, dynaamista debattia agenttien välillä, tämä strukturoitu lähestymistapa varmistaa analyysin täydellisen jäljitettävyyden.

Tämä yhdeksän agentin (Vartija, Analyytikko, Loogikko, Kriitikkoryhmä [neljä agenttia], Tuomari ja XAI-Raportoija) vaiheittainen malli on tietoinen arkkitehtuurivalinta, joka priorisoi maksimaalista auditoitavuutta ja jäljitettävyyttä tehokkuuden kustannuksella. Vaikka dynaamisemmat debatti-arkkitehtuurit voivat tuottaa syvällisempiä oivalluksia (Du ym. 2023), tämä tiukasti sekventiaalinen malli valittiin perusarkkitehtuuriksi. Se pakottaa analyysin noudattamaan tiukasti falsifioinnin ja Toulminin mallin kaltaisia, ennalta määriteltyjä ja validoituja loogisia rakenteita. Operatiivisesti tämä auditoitavuus toteutuu siten, että jokainen agentti tuottaa standardoidun JSON-välitulosteen. Tämä modulaarinen rakenne edellyttää kaikkien välitulosten välittämistä prosessin loppuun. Vaikka tämä lisää datan määrää, se on strateginen valinta: se varmistaa, että lopullinen päätös perustuu koko päättelyketjuun ja mahdollistaa Tuomari-agentin suorittaman hierarkkisen konfliktinratkaisun. Nämä välitulosteet muodostavat yhdessä täydellisen "kognitiivisen jäljen" koko päättelyprosessista, joka voidaan tallentaa ja tarkastaa jälkikäteen (vrt. Luku 4.2).

Tämän moniagenttijärjestelmän täydellinen operationaalinen komentorakenne, agenttien roolit (Vartija, Analyytikko, Loogikko, Looginen Falsifioija, Kausaalinen Analyytikko, Performatiivisuuden Tunnistaja, Faktuaalinen ja Eettinen Valvoja, Tuomari, XAI-Raportoija) ja vaiheittainen työnkulku muodostavat sen operationaalisen mallin.

#### 2.4.2 Todistepohjainen Ankkurointi ja Argumentaation Validointi (Analyytikko ja Loogikko)

Arviointiprosessin aloittaa Analyytikko-agentti. Sen ainoa tehtävä on luoda ”todistuskartta” ja varmistaa, että kaikki myöhempi analyysi on ankkuroitu toimitettuun todistusaineistoon. Se toteuttaa tämän hyödyntämällä RAG-arkkitehtuuria (*Retrieval-Augmented Generation*) (Lewis ym. 2020), joka vähentää merkittävästi kielimallien taipumusta hallusinointiin (Shuster ym. 2021). RAG-arkkitehtuureilla on kuitenkin tunnettuja heikkouksia (Ahmad ym. 2024). Yksi merkittävä haaste on ”lost in the middle” -ilmiö, jossa mallit eivät kykene hyödyntämään tehokkaasti tietoa pitkän konteksti-ikkunan keskellä (Liu, N. F. ym. 2024). Prototyyppivaiheessa tätä riskiä ei hallita teknisesti (esim. erillisellä uudelleensijoitusmallilla; vrt. Ma ym. 2024). Riskiä pyritään kuitenkin lieventämään operatiivisesti. \*\*Analyytikko-agentti on ohjeistettu toteuttamaan kaksivaiheisen prosessin: ensin agentti suorittaa laajan haun (Alivaihe A), minkä jälkeen se optimoi tulokset (Alivaihe B) sijoittamalla tärkeimmät tulokset kontekstin alkuun ja loppuun. Tämä on kehotepohjainen strategia, joka perustuu Liu, N. F. ym. (2024) havaintoihin. Lisäksi agenttia ohjeistetaan kirjaamaan tämä riski XAI-raportointia varten.

Kun todistusaineisto on ankkuroitu, Loogikko-agentin tehtävänä on purkaa käyttäjän reflektio sen loogisiin osatekijöihin. Se soveltaa Stephen Toulminin argumentaatiomallia (Toulmin 2003), joka jäsentää argumentin sen peruselementteihin: väite (claim), perusteet (data/grounds), oikeutus (warrant), taustatuki (backing)... Tässä viitekehyksessä Loogikko-agentin rooli on syventynyt. Se arvioi myös argumentin oikeutuksen (warrant) – eli sen säännön tai periaatteen, joka oikeuttaa siirtymän datasta väitteeseen – kognitiivista tasoa, hyödyntäen Bloomin taksonomian ylempiä tasoja (vrt. Anderson & Krathwohl 2001). Lisäksi Loogikko tunnistaa Oikeutuksen taustalla olevan argumentaatioskeeman (Walton ym. 2008). Tämä skeema välitetään Kriitikkoryhmälle kohdennettua stressitestausta varten. Tähän menetelmään liittyy kuitenkin systeeminen riski. Koko prosessivalvonnan tehokkuus on riippuvainen Loogikko-agentin kyvystä tunnistaa argumentaatioskeema oikein. On myös tärkeää tunnistaa Toulminin mallin rajoitukset, kuten sen soveltuvuus lähinnä lineaarisiin argumentteihin (Crusius & Channell 2003).

#### 2.4.3 Systemaattinen Falsifiointi ja Synteesi (Kriitikkoryhmä, Tuomari ja XAI-Raportoija)

Holistisen arvioinnin ytimessä toimii Kriitikkoryhmä, joka koostuu neljästä erikoistuneesta agentista. Ryhmän tehtävä on toimia järjestelmän sisäisenä ”paholaisen asianajajana”, ja sen toiminta perustuu Karl Popperin falsifiointiperiaatteeseen: tieteellinen totuus selvitetään yrittämällä aktiivisesti kumota esitetyt väitteet (Popper 1934).

Sen sijaan, että ryhmä etsisi vahvistusta Loogikko-agentin havainnoille, sen tehtävä on yrittää aktiivisesti kumota Loogikon muodostama argumentti. Tämä on kriittinen vaihe, sillä ilman aktiivista haastamista tekoälymallit sortuvat helposti ”myötäilyvinoumaan” (sycophancy), jossa ne vain vahvistavat toistensa (mahdollisesti virheelliset) päätelmät (Perez ym. 2022b; Wynn, Satija & Hadfield 2025).

Tämä monimutkainen auditointi on jaettu neljään erilliseen kognitiiviseen rooliin:

1. **Looginen Falsifioija-agentti ("Argumentaation Auditoija")** Tämä agentti iskee argumentaation rakenteeseen. Jotta se ei sortuisi lauman mukana kulkemiseen, sille on annettu erityinen "Erimielisyyden Ylläpidon Mandaatti" (JEM).

* **Tehtävä:** Agentin on vastustettava "konsensuksen tyranniaa" ylläpitämällä perusteltua erimielisyyttä (Wynn, Satija & Hadfield 2025). Se ei saa muuttaa analyysiaan vain ollakseen samaa mieltä muiden kanssa. Tätä varten se hyödyntää "punaisen tiimin" (red teaming) menetelmiä (Ganguli ym. 2022).
* **Päättelyn uskollisuus (Faithfulness Audit):** Agentti tarkistaa, onko esitetty päättelyketju aito. Se etsii merkkejä siitä, että käyttäjä (tai tekoäly) on keksinyt perustelut jälkikäteen (post-hoc-rationalisointi) sen sijaan, että ne olisivat aidosti ohjanneet toimintaa (Turpin ym. 2023; Arcushin ym. 2025; Creswell ym. 2024).
* **Rajoitteet:** Popperin falsifioinnin soveltaminen "pehmeisiin" ilmiöihin on haastavaa (ks. Nola & Sankey 2014). Siksi tässä viitekehyksessä falsifiointia käytetään täsmätyökaluna: etsitään suoria, loogisia ristiriitoja reflektion ja keskusteluhistorian välillä hyödyntämällä argumentaatioskeemojen kriittisiä kysymyksiä (Walton, Reed & Macagno 2008).

1. **Faktuaalinen ja Eettinen Valvoja -agentti ("Todisteiden Valvoja")** Tämä agentti vastaa siitä, että väitteet vastaavat todellisuutta ja noudattavat eettisiä sääntöjä. Se ei luota pelkkään annettuun tietoon, vaan kaivaa syvemmältä.

* **RFI-Protokolla (Tiedonhankinta):** Agentti suorittaa kohdennetun uusintahaun (Request for Information Protocol). Nykyisessä prototyypissä, jossa ei ole integroitua verkkohakutyökalua, tämä vaihe toteutetaan hyödyntämällä mallin sisäistä tietämystä ja kontekstin ristiintarkistusta (Simulated Retrieval). Tämä iteratiivinen lähestymistapa on välttämätön monivaiheisessa päättelyssä, jossa uutta tietoa on haettava dynaamisesti analyysin edetessä (vrt. Trivedi ym. 2024). Se käyttää edistyneitä tekniikoita löytääkseen tietoa, joka jäi alkuperäiseltä haulta piiloon:
  + **Kyselynlaajennus:** Hakulausekkeiden muokkaaminen uusista näkökulmista (Jagerman ym. 2023).
  + **HyDE (Hypothetical Document Embeddings):** Agentti kuvittelee ideaalin dokumentin, joka kumoaisi väitteen, ja käyttää sitä hakuna (Gao ym. 2022).
* **Heterogeenisyyden välttämättömyys:** Järjestelmän luotettavuus paranee merkittävästi, jos tämä vaihe ajetaan eri tekoälymallilla (esim. GPT-4) kuin aiemmat vaiheet (esim. Gemini) (Ye ym. 2025). Jos kaikki agentit käyttävät samaa mallia, ne saattavat toistaa samat virheet ja hallusinaatiot ("sokeat pisteet") (Cemri ym. 2025). Eri mallien käyttö mahdollistaa aidon ristiinvalidoinnin.
* **Eettinen tarkastus:** Agentti etsii aktiivisesti vakavia eettisiä rikkomuksia, kuten syrjintää tai lähteiden tahallista vääristelyä (Weidinger ym. 2021).

**Prosessiauditointi: Kausaalisuus ja Performatiivisuus** Järjestelmän kyky torjua "performatiivista reflektiota" (Cullen 2020) on keskeinen sen validiteetille. Tämä tehtävä on jaettu kahdelle erikoistuneelle agentille:

1. **Kausaalinen Analyytikko-agentti ("Temporaalinen Auditoija")]** Tämän agentin tehtävä on auditoida prosessin ajallista johdonmukaisuutta ja kausaalista uskottavuutta.

* **Temporaalinen auditointi:** Agentti tarkistaa aikajanan: ilmestyikö oivallus (syy) keskusteluhistoriaan ennen tuloksen paranemista (seuraus)? Syyn on aina edellettävä seurausta (Lagnado & Sloman 2006; Pearl 2009).
* **Kontrafaktuaalinen stressitesti (L3-simulaatio):** Agentti kysyy: 'Jos käyttäjä EI olisi tehnyt tätä oivallusta, olisiko tulos silti ollut sama?'. Tämä on yritys simuloida syvällistä syy-seuraus-päättelyä (Pearl 2009; Sgaier ym. 2020).
* **Abduktiivinen Haasto:** Agentti soveltaa Occamin partaveistä (vrt. Walton ym. 2008). Se arvioi, onko käyttäjän kuvaama oivallus yksinkertaisin selitys havaitulle muutokselle, vai onko post-hoc rationalisointi todennäköisempi selitys.

1. **Performatiivisuuden Tunnistaja-agentti ("Käyttäytymisanalyytikko")** Tämä agentti keskittyy tunnistamaan käyttäytymismalleja ja pelistrategioita, jotka viittaavat järjestelmän manipulointiin (Goodhartin laki).

* **Epäuskottava lineaarisuus:** Onko prosessi liian suoraviivainen ja virheetön ollakseen totta? (vrt. Goffman 1959).
* **Pinnallinen vuorovaikutus:** Osoittaako keskusteluhistoria vain vähäistä kognitiivista syvyyttä?
* **Kognitiivinen epäsuhta:** Vastaako reflektiossa kuvattu prosessi keskusteluhistorian todellista kulkua?
* **Keinotekoinen monimutkaisuus:** Onko prosessiin lisätty turhia vaiheita vain näyttävyyden vuoksi? (Cullen 2020).
* **Matriisin optimointi:** Vastaako reflektio epäilyttävän tarkasti arviointikriteereitä, vaikka itse työskentely ei? (Strathern 1997; Stumborg ym. 2022).
* **Kognitiivinen investointi:** Vastaako oivallukseen käytetty kognitiivinen työpanos sen väitettyä merkittävyyttä (vrt. de Bruin ym. 2023)?
* **Itsetehostuksen Indikaattorit:** Etsitään merkkejä itsetehostusvinoumasta (Dufner ym. 2019).
* **Pre-Mortem Analyysi:** Agentti kääntää todistustaakan olettamalla reflektion olevan väärennös ja etsimällä tätä tukevia signaaleja (Klein 2007).

On kuitenkin huomattava, että ilman ulkoista maailmanmallia kielimalli ei kykene muodolliseen kausaaliseen päättelyyn (Chi ym. 2024), joten näiden agenttien suorittamat testit mittaavat ensisijaisesti narratiivin loogista eheyttä eivätkä sen empiiristä totuusarvoa. Siksi nämä testit ovat "kielellisiä approksimaatioita" – ne ovat parhaita mahdollisia arvauksia, eivät matemaattisen tarkkoja todisteita. Tämä tekee järjestelmästä haavoittuvan taitavalle manipuloinnille.

**Synteesi ja Raportointi (Tuomari ja XAI-Raportoija)** Kun "käräjät" on käyty, Tuomari-agentti kokoaa tulokset. Tämä ei ole pelkkä keskiarvo, vaan hierarkkinen konfliktinratkaisu, joka noudattaa tiukkoja sääntöjä:

* **Falsifioinnin etusija:** Faktat voittavat aina tulkinnat (Popper 1934). Jos Valvoja löytää faktavirheen tai eettisen rikkomuksen, se syrjäyttää Loogikon positiivisen tulkinnan "mestaruudesta"
* **Jäsennellyn erimielisyyden mandaatti (JEM):** Jos Kriitikko ja Loogikko ovat eri mieltä tulkinnasta, Tuomari ei saa pakottaa niitä yksimielisyyteen. Erimielisyys on arvokasta tietoa, joka paljastaa tapauksen monimutkaisuuden (Wynn, Satija & Hadfield 2025).

Lopuksi XAI-Raportoija laatii raportin selitettävän tekoälyn (XAI) periaatteiden mukaisesti (Adadi & Berrada 2018). Se ei vain kerro tulosta, vaan tekee näkyväksi kaiken epävarmuuden (Hüllermeier & Waegeman 2021):

* **Aleatorinen epävarmuus:** Datan epäselvyydestä johtuva epävarmuus.
* **Systeeminen epävarmuus:** Itse järjestelmän rajoituksista (esim. kehotteiden hauraus, kausaalipäättelyn puute) johtuva epävarmuus.
* **Episteeminen epävarmuus:** Agenttien välisestä erimielisyydestä johtuva epävarmuus.

Tämä läpinäkyvyys on turvallisuustekijä. Raportti pakottaa ihmisvalvojan (HITL) ottamaan kantaa kriittisiin kysymyksiin (”Kriittiset Auditointikysymykset”) ja varmistaa näin, ettei tekoälyn päätöstä hyväksytä sokeasti.

#### 2.4.4 Operationaaliset Mandaatit ja Erottelusäännöt

Kognitiivisen Arviointimatriisin soveltaminen edellyttää tarkkoja sääntöjä, jotka ohjaavat agenttien päättelyä. Nämä "erottelusäännöt" ovat keskeinen mekanismi reliabiliteetin ja validiteetin paradoksin hallitsemiseksi. Niistä psykometrisesti merkittävin on Pisteytysmandaatti (Prosessin Puhtaus), joka on operationalisoitu järjestelmän erottelusäännöissä. Tämä sääntö kieltää järjestelmää hyödyntämästä Reflektiodokumentin dataa. Analyysi ja Prosessin Tehokkuus" -kriteerin arvioinnissa ja pakottaa sen perustamaan arvioinnin ainoastaan Keskusteluhistoriaan.

Tämä on tietoinen metodologinen kompromissi. Sen perusteluna on Goodhartin lain (Strathern 1997) ja "performatiivisen reflektion" (Cullen 2020) torjunta. Ilman tätä sääntöä käyttäjä voisi kompensoida heikon, tehottoman tai epäloogisen prosessin (kuten Keskusteluhistoriasta nähdään) tuottamalla erittäin vakuuttavan ja analyyttisen jälkikäteisen rationalisoinnin Reflektiodokumenttiinsa. Erottamalla nämä kaksi todistusaineistoa toisistaan sääntö varmistaa, että arviointi mittaa aitoa prosessiosaamista, ei pelkästään kykyä kuvailla prosessia (vrt. Luku 5.1.2). Tämä parantaa kriteerin reliabiliteettia ja käsitevaliditeettia, vaikka se rajoittaakin arvioinnin täysin holistista luonnetta.

Keskeisiä sääntöjä ovat myös:

* **Faktuaalisen Tarkkuuden Mandaatti**: Ohjaa käsittelemään reflektiodokumentissa havaittuja faktuaalisia virheitä. Yksittäinen virhe ei automaattisesti rajoita pistemäärää, vaan sen painoarvoa on punnittava.
* **Synteesin Omaperäisyyden Arviointi**: Määrittelee metodin "Synteesi ja Luovuus" -kriteerin arviointiin. Ohjeistaa vertaamaan lopputuotetta tekoälyn vastaukseen; identtisyys on todiste matalamman tason synteesistä.
* **Malliriippumaton kompetenssin erottelusääntö:** Viitekehys vastaa kielimallien nopeaan kehitykseen ja niin sanottuun "Black Box" -asetelmaan soveltamalla arvioinnissa syöte–kontrolli-suhdelukua (Input-Control Ratio). Tämä on välttämätöntä validiteetin varmistamiseksi tilanteessa, jossa tehokkaat päättelymallit kykenevät tuottamaan korkealaatuisia lopputuloksia myös triviaaleilla syötteillä, mikä luo riskin "passiivisesta onnistumisesta".

Säännön metodologisena perusteena on Lazy User -teoria, jonka mukaan käyttäjät pyrkivät minimoimaan kognitiivisen vaivannäön valitsemalla pienimmän vastustuksen tien (Tétard & Collan 2009). Ilman aktiivista kontrollia käyttäjä altistuu automaatioharhalle (Parasuraman & Riley 1997), jossa algoritmin tuotosta pidetään virheellisesti omana osaamisena. Tämän erottelusäännön mukaisesti arvioinnin fokus siirretään lopputuotteen pinnallisesta sujuvuudesta käyttäjän asettamien rajoitteiden tiheyteen (Constraint Density). Avoin tehtävänanto indikoi päätösvallan ulkoistamista, kun taas rajoittava ja iteroiva ohjaus toimii todisteena strategisesta kompetenssista ja kyvystä hallita mallien taipumusta epäuskolliseen päättelyyn (vrt. Turpin ym. 2023).

#### 2.4.5 Arvioinnin Rajaus: Substanssiosaamisen ja Kognitiivisten Taitojen Erottelu

Hybridirubriikin tavoitteena on arvioida yleisiä kognitiivisia taitoja, ei spesifistä substanssiosaamista. Tämän varmistamiseksi on luotu erityinen Substanssin Strateginen Arviointi -sääntö. Tämä sääntö on kriittinen käsitevaliditeetin kannalta (AERA, APA & NCME 2014; Messick 1989). Se määrittelee kaksi periaatetta:

1. **Substanssin Oikeellisuuden Sivuuttaminen**: Agentteja kielletään arvioimasta substanssiosaamisen akateemista tarkkuutta. Fokus pysyy kognitiivisissa prosesseissa (Halpern 2014).
2. **Strategisen Käytön Tunnistaminen**: Agentteja vaaditaan tunnistamaan, kun käyttäjä käyttää substanssiosaamistaan strategisena välineenä tekoälyn ohjaamisessa.

Tämä rajaus mahdollistaa viitekehyksen soveltamisen geneerisesti (Shavelson 2010) ja korostaa tekoälylukutaitoa siirrettävänä metataitona. Akateeminen keskustelu onkin siirtymässä kohti "meta-tekoälytaitoja" (Ahuna & Kiener 2025), jotka rinnastuvat metakognitioon (Carolus ym. 2023). Tämä erottelu on toteutettu operationaalisesti kieltämällä agenteilta substanssin akateemisen tarkkuuden arvioinnin Tämä sääntö on kriittinen viitekehyksen käsitevaliditeetin (*construct validity*) kannalta ja pyrkii minimoimaan "konstruktille irrelevanttia varianssia" (AERA, APA & NCME 2014; Messick 1989). Arviointi ei ole substanssitentti, vaan kognitiivisen prosessin auditointi.

### 2.5 Arkkitehtuurin Hallintamalli: Kognitiivinen Palomuuri ja Ihmisvarmistus

Arkkitehtoniset tasot edellyttävät selkeää hallintamallia. Tämä malli muodostaa hierarkkisen kontrollirakenteen, jota ohjaa strateginen periaatekokonaisuus ja valvoo ihminen.

#### 2.5.1 Kognitiivinen Palomuuri

Kriittisin instrumentti on Kognitiivinen Palomuuri. Se on agentteja ohjaava periaatteellinen rajoituskokonaisuus, joka muodostaa järjestelmän Behavioraalisen Kontrollikerroksen (joka on määritelty järjestelmän globaaleissa rajoituksissa ja säännöissä; vrt. Luku 2.6.2).

Tässä prototyypissä "Palomuuri" ei ole tekninen, vankka luokittelija (vrt. Luku 4.4.3 visio; Sharma ym. 2025), vaan metodologinen hallintamalli, joka toteutetaan kahdella toisiaan täydentävällä, läpinäkyvyyteen perustuvalla tavalla:

1. **Agenttien Itsetietoisuus (XAI):** Järjestelmä on pakotettu tunnistamaan ja raportoimaan oma haurautensa. Järjestelmän säännöt edellyttävät, että XAI-Raportoija-agentin on pakko kirjata tämä kehotepohjaisen kontrollin hauraus Systeemiseksi Epävarmuudeksi jokaiseen lopulliseen arviointiraporttiin.
2. **Vastuun Siirtäminen (HITL):** Järjestelmä siirtää tietoisesti kriittiset arkkitehtoniset vastuut, joita se ei voi itse valvoa, ihmisvarmistajalle (HITL). Esimerkiksi järjestelmän säännöt operationalisoivat Luvussa 5.2.2 tunnistetun heterogeenisyysvaatimuksen (vrt. Cemri ym. 2025; Ye ym. 2025) pakottamalla agentin tuottamaan ihmiselle varoituksen tämän manuaalisesta tarkistusvastuusta.

Näin ollen prototyypin turvallisuusmalli ei nojaa teknisiin kontrolliesteisiin, vaan auditoitavaan läpinäkyvyyteen ja ihmisen suorittamaan valvontaan. Riskin torjumiseksi Tekninen Kontrollikerros on suunniteltu kestävämmäksi. Vartija-agentti (Luku 2.6.1) edustaa seuraavan sukupolven ratkaisua. Visio edellyttää siirtymistä teknisiin perustuslaillisiin luokittelijoihin (engl. Constitutional Classifiers) (Anthropic 2025a), jotka tarjoavat vahvemman suojan (Sharma ym. 2025).

Viitekehyksen hauraus jakautuu kahtia. Behavioraalisen kerroksen hauraus on sisäsyntyinen, sillä kielimallien täydellinen linjaaminen (alignment) on tunnetusti vaikeaa (Wolf ym. 2023), kun taas Teknisen kerroksen hauraus liittyy sen nykyiseen toteutustapaan. Koska Vartija-agentti on toteutettu kehotepohjaisena (eikä erillisenä luokittelijana), se on altis manipuloinnille (Sharma ym. 2025) ja sen tehokkuus kehotemurtoja vastaan on rajoitettu (Jia ym. 2025).

#### 2.5.2 Vinoumien torjunta

Palomuuri on suunniteltu lieventämään tunnettuja kognitiivisia vinoumia ja kielimallien virheitä. Se torjuu esimerkiksi:

* Auktoriteettivinoumaa (*authority bias*) (Wang ym. 2023).
* Monisanaisuusvinoumaa (*verbosity bias*) (Saito ym. 2023).
* Vahvistusvinoumaa (*confirmation bias*) (Kahneman 2011; Talboy & Fuller 2023).
* Myötäilyvinoumaa (*sycophancy bias*) (Perez ym. 2022b).
* Ankkurointivaikutusta (*anchoring effect*) (Kahneman 2011).
* Itsetehostusvinoumaa (*self-enhancement Bias*) (Dufner ym. 2019).
* Saatavuusvinoumaa (*availability bias*) (Tversky & Kahneman 1974).
* Kehystysvaikutusta (*framing bias*) (Tversky & Kahneman 1974).
* Esteettistä vinoumaa (*aesthetic bias*) (Reinecke & Gajos 2014). Järjestelmän säännöt kieltävät agenteilta antamasta suhteetonta painoarvoa dokumenttien ulkoasulle.

#### 2.5.3 Holistisen tason ohjaus

Palomuuri ohjaa myös holistisen tason soveltamista säännöillä, jotka mahdollistavat analyyttisen tason ylittävän arvioinnin:

* **Metodologisen nöyryyden mandaatti**: Pakottaa agentit etsimään todisteita matriisin määritelmät ylittävästä Mestaruudesta. Tämä on operationalisoitu sääntönä, joka pakottaa agentit tunnistamaan Kognitiivisen Arviointimatriisin luontaisen jäykkyyden ja etsimään aktiivisesti todisteita matriisin kapeat määritelmät ylittävästä Mestaruudesta (vrt. Dreyfus & Dreyfus 1980).
* **Performatiivisuuden tunnistus**: Ohjeistaa agentteja etsimään epäaitoja narratiiveja (vastatoimi Goodhartin laille; Stumborg ym. 2022). Tämä operationalisoidaan säännöissä vaatimuksena etsiä aktiivisesti merkkejä arviointijärjestelmän "pelaamisesta" (toteutettuna Prosessiauditoijaryhmän ja Tuomari-agentin toimesta).

#### 2.5.4 Ihmisvarmistus (HITL)

Hallintamallin kulmakivi on pakollinen ihmisvarmistus (*Human-in-the-Loop*, HITL). Sen toiminta perustuu valvontaohjattuun automaatioon ja heijastelee EU:n tekoälysääntelyn periaatteita (AI Act, Art. 14) (Euroopan komissio 2024a). EU:n eettiset ohjeet määrittelevät valvonnan kolmitasoiseksi: "Human-in-the-Loop" (suom. ihminen prosessissa mukana), "Human-on-the-Loop" (suom. ihminen valvomassa prosessia) sekä "Human-in-Command" (suom. ihminen ohjaamassa prosessia) (Euroopan komission korkean tason asiantuntijaryhmä 2019). Valvonnan tulee ulottua taktisesta väliintulosta strategiseen hallintaan (Pfeifer 2025). Ihmisen rooli on toimia järjestelmän strategisena valvojana ja ylimpänä auktoriteettina. HITL-varmistaja on kuitenkin altis automaatioharhalle (Luku 5.3.1). Tämän torjumiseksi järjestelmä soveltaa "Interrogatiivista Raportointia". XAI-Raportoija-agentin raportti (sisältäen XAI-raportoinnin epävarmuudesta) ei ole passiivinen tiedonanto, vaan se pakottaa ihmisvarmistajan aktiiviseen kognitiiviseen työhön. Tämä mekanismi on operationalisoitu siten, että XAI-Raportoija generoi "Kriittisiä Auditointikysymyksiä", joihin ihmisvarmistajan on vastattava, erityisesti koskien JEM-erimielisyyksiä. Esimerkiksi: "HITL-RATKAISU VAADITAAN: Kriitikko väittää X, Loogikko väittää Y. Kumpi argumentti on paremmin tuettu todisteella Z?" Tämä varmistaa ihmisen aktiivisen osallistumisen ja vähentää taipumusta hyväksyä raportti kritiikittömästi. HITL-varmistaja tekee lopullisen, vastuullisen päätöksen.

### 2.6 Tekninen Toteutuskehys ja Monikerroksinen Puolustusstrategia

Järjestelmän eheyden varmistaa monikerroksinen puolustusstrategia, joka torjuu kielimalleihin liittyviä uhkia (OWASP Foundation 2025f). Strategia perustuu monikerroksisen puolustuksen (engl. *Defense in Depth*, DiD) -malliin (CISA 2016) ja koostuu kolmesta toisiaan täydentävästä kontrollikerroksesta

* **Simuloitu** Tekninen Kontrollikerros (Luku 2.6.1) suojaa ensisijaisesti ulkoisilta uhilta, kuten kehotemurroilta (LLM01:2025). Prototyypissä tämä kerros on toteutettu behavioraalisena simulaationa (kehotepohjainen sanitointi), joka emuloi tuotantoympäristön teknisiä kontrolleja, mutta ei korvaa niitä (vrt. Jia ym. 2025).
* Behavioraalinen Kontrollikerros (Luku 2.6.2) hallitsee agenttien sisäistä toimintaa ja torjuu toimivallan ylittämistä (LLM06:2025).
* Hallinnollinen Kontrollikerros (Luku 2.6.3) toimii ylimpänä valvontamekanismina.

Prototyypin nykytilassa järjestelmän turvallisuutta ja luotettavuutta rajoittaa useiden kehittyneiden teknisten toimintojen puuttuminen. Näitä ovat "Semanttinen Anonymisointi" (jota kompensoidaan kehotepohjaisella analyysilla), "Upotusten Eheyden Tarkistus" ja "Uudelleensijoitusmalli". Turvallisuusmalli koostuu kolmesta tasosta, jotka esitellään seuraavissa alaluvuissa.

#### 2.6.1 Tekninen Kontrollikerros

Ensimmäinen puolustuslinja suojaa ulkoisilta teknisiltä uhilta. Vartija-agentti toimii turvaporttina. Sen tehtäviin kuuluu ”Aktiivinen Uhkien Luokittelu”, joka on vastaus epäsuorien kehotemurrosten (engl. *Indirect Prompt Injection*) uhkaan (LLM01:2025) (Yi ym. 2025; Greshake ym. 2023; Liu, X. ym. 2024). Teknisesti Vartija-agentti-mandaatti sisältää seuraavat toteutetut (mutta testaamattomat) vaiheet:

1. **Rakenteellinen Puhdistus (*Input Sanitization*) ja Datan Normalisointi**: Tämä vaihe yhdistää turvallisuuskontrollit ja datan eheyden varmistamisen. Ensimmäisenä puolustuslinjana (vrt. OWASP Foundation i.a.[a]) Vartija-agentin tehtävä on muuntaa kaikki syötteet (esim. PDF, DOCX) raakatekstiksi. Tämä pienentää hyökkäyspinta-alaa poistamalla monimutkaisten tiedostomuotojen käsittelyyn liittyviä riskejä ja varmistaa datan yhdenmukaisuuden (OWASP Foundation i.a.[b]). Tähän prosessiin kuuluu pakollinen datan normalisointi, joka on kriittistä merkistön eheyden ja interoperabiliteetin varmistamiseksi, erityisesti manuaalisessa orkestroinnissa (vrt. Luku 5.2.4; W3C 2008). Normalisointi sisältää pakollisen UTF-8-merkistökoodauksen varmistamisen sekä typografisten merkkien (kuten "älykkäiden lainausmerkkien") muuntamisen standardeiksi ASCII-merkeiksi datan yhdenmukaisuuden takaamiseksi. Tämän jälkeen se poistaa aktiivisesti kaikki tunnetut haitalliset merkit, skriptit ja ohjausmerkit (*control characters*).
2. **Datan Anonymisointi (OWASP LLM02:2025-torjunta)**: Tämä vaihe torjuu arkaluontoisen tiedon paljastumista. Vartija-agentin tehtävä on hakea ja peittää tunnistettavat henkilötiedot (PII) monikerroksisesti (vrt. Lison ym. 2021; Li ym. 2024). Prototyyppivaiheessa ulkoisilla kirjastoilla toteutettu deterministinen anonymisointi puuttuu. Tämä korvataan agenttipohjaisella hybridimallilla, joka yhdistää sääntöpohjaiset (RegEx) menetelmät kielimallin suorittamaan "Kontekstuaaliseen PII-Analyysiin". Tämä mahdollistaa epätyypillisten, kontekstisidonnaisten henkilötietojen tunnistamisen, mihin perinteiset säännöt eivät taivu (vrt. Li ym. 2024), vaikka menetelmä sisältääkin stokastisen epävarmuuden.
3. **Aktiivinen Uhkien Luokittelu (OWASP LLM01:2025-torjunta)**: Tämä vaihe on suora vastaus epäsuorien kehotemurrosten (*Indirect Prompt Injection*) uhkaan. Viitekehyksen visiona on käyttää erillistä teknistä luokittelijaa.

Nykyinen prototyyppitoteutus toteuttaa tämän kontrollin prototyyppivaiheessa kehotepohjaisena 'semanttisena perustuslakitarkistuksena' (vrt. Bai ym. 2022). Tämä luo kuitenkin puolustuksellisen paradoksin: järjestelmä pyrkii torjumaan kehotemurtoja (LLM01) instrumentilla (kehote), joka on altis samalle hyökkäykselle. Ilman ulkoista luokittelijaa tämä kontrolli on katsottava 'syvyyspuolustukseksi' (Defense in Depth) eikä täydelliseksi ratkaisuksi. Menetelmä on altis kehittyneille *prompt injection* -hyökkäyksille, kuten Jia et al. (2025) ovat osoittaneet. Jia ym. (2025) kritisoivat nykyisiä puolustusmekanismeja siitä, että ne eivät ole riittävän tehokkaita adaptiivisia hyökkäyksiä vastaan.

Prototyypin turvallisuus perustuu osittain "Adversariaaliseen Simulaatioon", jossa Vartija pakotetaan suorittamaan monivaiheinen uhka-analyysi sisäisessä päättelytilassa (*scratchpad*), omaksumalla ensin "Punaisen Tiimin" (Hyökkääjä) rooli ja simuloimalla hyökkäystä, ja sen jälkeen "Sinisen Tiimin" (Puolustaja) rooli arvioimalla sen onnistumista. Tämä syventää analyysia kohti syötteen intention ymmärtämistä (vrt. Perez ym. 2022a). On kuitenkin tärkeää huomata, että tämä luo perustavanlaatuisen arkkitehtonisen ristiriidan: järjestelmä pyrkii torjumaan kehotemurtoja (LLM01) käyttämällä menetelmää (kehotepohjainen kontrolli), joka itsessään on kaikkein haavoittuvin juuri kyseiselle hyökkäykselle. Kuten Luvussa 2.5.1 todetaan, tämä on tunnistettu merkittävä tekninen rajoite ja haurauden lähde (vrt. Jia ym. 2025; Liu, Y. ym. 2023).

Vartija-agentin operationaalinen toteutus sisältää yksityiskohtaiset tehtävät näiden teknisen kerroksen toimintojen suorittamiseksi (Rakenteellinen Puhdistus, Anonymisointi ja Uhkien Luokittelu).

* **Upotusten Eheyden Tarkistus (OWASP LLM08:2025-Torjunta):** Viitekehyksen visioon kuuluu Vartija-agentin kyky tunnistaa RAG-arkkitehtuuriin kohdistuvia hyökkäyksiä suorittamalla ajonaikainen poikkeamien havaitseminen upotustilassa (vrt. Acevedo ym. 2024; OWASP Foundation 2025e). On kuitenkin kriittistä huomata, että nykyisessä prototyypissä tämä kontrolli on toteutettu ainoastaan "negatiivisena lokikirjauksena". Järjestelmä pakottaa Vartija-agentin kirjaamaan puuttuvan suojauksen eksplisiittiseksi riskiksi ('LLM08-riski hallitsematon'), mikä tekee haavoittuvuudesta läpinäkyvän (XAI) mutta ei poista sitä teknisesti. Tämä 'Negative Logging' -menetelmä varmistaa, että haavoittuvuus raportoidaan aina loppukäyttäjälle (XAI-Raportti), vaikka sitä ei voida prototyypissä torjua (OWASP Foundation 2025e).
* **Datan Merkintä (engl. *Input Tainting*)**: Viimeisenä valmisteluvaiheena ja ainoastaan jos aiempien vaiheiden turvatarkistukset on läpäisty, Vartija-agentin tehtävä on koota kaikki puhdistettu data yhteen objektiin. Tämä menetelmä (Input Tainting) luo eristetyn ”luottamuksen kehän” (engl. *Chain of Trust*): vain tämän objektin sisällä oleva data on validoitua. Tämä on kriittinen kontrolli, sillä järjestelmän behavioraaliset säännöt (ks. Sääntö 3; Luku 2.6.2) pakottavat kaikki myöhemmät agentit käsittelemään ainoastaan tätä validoitua dataa ja hylkäämään data, joka ei ole peräisin tästä lähteestä (vrt. Secure Information Flow; Denning & Denning 1977).).

#### 2.6.2 Behavioraalinen Kontrollikerros

Järjestelmän toinen puolustuslinja rakentuu behavioraalisen kontrollikerroksen varaan, joka toimii niin kutsuttuna kognitiivisena palomuurina. Tämän kerroksen ensisijaisena tehtävänä on hallita tekoälymallien toimintaa ja estää agenttien toimivallan ylittyminen (engl. *excessive agency*) (OWASP Foundation 2025d).

Tämän tavoitteen saavuttamiseksi järjestelmä asettaa tiukat rajoitteet, jotka kieltävät eksplisiittisesti agentteja hyödyntämästä määrittelemättömiä ulkoisia työkaluja tai API-rajapintoja. Rajoituksilla pyritään torjumaan työkalujen käyttöön liittyviä haavoittuvuuksia (engl. *tool-use vulnerability*) (Research AIMultiple 2025). Lisäksi kontrollikerroksella ehkäistään systeemisiä riskejä, kuten roolivuotoa (engl. *role-bleed*), jossa agentti saattaa sekoittaa eri tehtäväkonteksteja tai ylittää sille määritellyn kognitiivisen roolin rajat (Yeager.ai 2023). Tämä rajaus on kriittinen tekijä kognitiiviseen työnjakoon perustuvan arkkitehtuurin eheyden säilyttämiseksi.

Järjestelmän eheyden ja yhteentoimivuuden varmistamiseksi järjestelmän sääntöihin on koodattu pakollinen, standardoitu kolmivaiheinen validointiprotokolla, jonka jokaisen agentin on suoritettava ennen tehtävänsä aloittamista:

1. **Rakenteellinen eheys ja puhdistus:** Syötteen JSON-muodon validointi ja tarvittaessa aggressiivinen puhdistus (esim. Markdown-jäänteiden poisto manuaalisessa orkestroinnissa).
2. **Semanttinen eheys (Tarkistussumma):** Datan sisällön vertaaminen edellisen vaiheen generoimaan semanttiseen tarkistussummaan datakorruption tai siirtovirheiden havaitsemiseksi (ks. Luku 5.2.4).
3. **Rakenteellinen skeptisyys:** Syötteen rakenteen kriittinen tarkastelu anomaliatunnisteiden, kuten epätyypillisen pituuden tai piilokomentoja muistuttavan kielen, havaitsemiseksi, mikä toimii sekundaarisena suojana kehotemurtoja (LLM01) vastaan.

Puolustusstrategia noudattaa syvyyspuolustuksen (engl. *defense-in-depth*) periaatetta hyödyntämällä redundanssia kahdella mekanismilla. Ensinnäkin kaikkia jälkiagentteja velvoitetaan ylläpitämään rakenteellista skeptisyyttä vastaanottamaansa dataa kohtaan, mikä luo ylimääräisen suojakerroksen Vartija-agentin mahdollisten virheiden varalle. Toiseksi järjestelmässä sovelletaan ristiinvalidoivaa päättelyketjua (engl. *Cross-Validating Chain-of-Thought*). Tässä menetelmässä agentit pakotetaan validoimaan edeltävän agentin päättelyn looginen johdonmukaisuus ennen oman prosessinsa käynnistämistä, mikä vähentää virheiden kumuloitumisen riskiä prosessissa.

On kuitenkin huomattava, että pelkkä kehotepohjainen ohjaus on luonteeltaan hauras (vrt. Greshake ym. 2023). Koska kehotteet eivät voi teknisesti täysin estää agenttia käsittelemästä samassa konteksti-ikkunassa olevaa muuta dataa, behavioraalista kerrosta on täydennettävä pakollisilla teknisillä kontrolleilla. Näihin lukeutuvat agenttien suorittaminen eristetyissä ympäristöissä (engl. *sandbox*) sekä tiukka pääsynhallinta (Towards AI 2025. Nämä kontrollit ovat keskeisiä LLM-agenttien turvallisuuden varmistamisessa (OWASP Foundation i.a.[c]), noudattaen vähimpien oikeuksien periaatetta (Saltzer & Schroeder 1975). Nykyisessä prototyypissä, jossa orkestrointi on manuaalista, vastuu eristämisestä on siirretty ihmisvalvojalle. Ilman ulkoisia, deterministisiä eristysympäristöjä ja verkkotason eriyttämistä, järjestelmän behavioraalinen suojaus on katsottava luonteeltaan ensisijaisesti performatiiviseksi turvakerrokseksi.

#### 2.6.3 Hallinnollinen Kontrollikerros

Viimeinen kontrollitaso auditoi ja valvoo prosessia (Ihmisvarmistus, HITL). Se toimii ylimpänä suojana monimutkaisia uhkia vastaan. Arkkitehtuurin on hallittava myös muita riskejä. Turvaton tuotoksen käsittely (LLM05:2025) torjutaan tulosteen koodauksella (OWASP Foundation 2025c). Toimitusketjun haavoittuvuuksia (LLM03:2025) hallitaan LLMOps-käytännöillä (Kreuzberger ym. 2023). Opetusdatan myrkyttäminen (LLM04:2025) puolestaan estetään käyttämällä vain ihmisen validoimaa dataa (D'Angelo 2025).

## Luku 3: Viitekehyksen Asemointi: Vertaileva Analyysi Akateemisiin ja Kaupallisiin Ratkaisuihin

Tämä luku sijoittaa viitekehyksen laajempaan kontekstiin vertaamalla sitä akateemiseen tutkimukseen ja kaupallisiin sovelluksiin. Tavoitteena on tunnistaa viitekehyksen keskeinen innovaatio ja strateginen erottautumistekijä.

### 3.1 Akateeminen maisema: Olemassa olevien osien uusi synteesi

Vaikka kokonaisarkkitehtuuri on uusi, sen komponentit nojaavat vakiintuneisiin tutkimussuuntauksiin. Kognitiivinen Kvoorum on moniagenttijärjestelmä (MAS) (Guo ym. 2024). Relevantteja vertailukohtia ovat adversariaalista dynamiikkaa hyödyntävät järjestelmät, kuten generatiiviset adversarialliset verkot (GAN) (Goodfellow ym. 2014) ja agenttien väliset debatit. Debattien on osoitettu parantavan päättelyn laatua (Du ym. 2023), mikä validoi Kriitikko-agentin roolin. Viitekehys sijoittuu koulutusteknologian kontekstiin (Luckin ym. 2017). Nykyiset sovellukset ovat kuitenkin keskittyneet konkreettisempien tuotosten arviointiin (Bezanilla ym. 2019), eivät abstraktin päättelyprosessin analyysiin (vrt. Li ym. 2025). Argumentaation laadun analyysille löytyy vastine argumentaation louhinnan (*Argumentation Mining*) alalta, joka keskittyy argumenttirakenteiden automaattiseen tunnistamiseen tekstistä (Lippi & Torroni 2016). Rakenteellisten piirteiden analyysin on osoitettu parantavan automaattista arviointia (Wachsmuth ym. 2017). Tämä tukee Loogikko-agentin Toulmin-pohjaista analyysia. Viitekehyksen ensisijainen innovaatio on näiden erillisten tutkimussuuntien – moniagenttiarkkitehtuurien, portfolioarvioinnin ja argumentaation louhinnan – ainutlaatuinen ja integroitu synteesi. Se soveltaa psykometrista teoriaa modernin tekoälyarkkitehtuurin avulla ennennäkemättömällä tavalla.

### 3.2 Kaupallinen maisema: Markkinarako laadulliselle arvioinnille

Kaupallinen markkina jakautuu pääosin kahteen kategoriaan. Tämä heijastaa laajempaa suuntausta, jossa tekoälysovellukset eriytyvät joko formatiivisiin (oppimista tukeviin) tai summatiivisiin (arviointia tuottaviin) mekanismeihin (Wisse & Greve 2023)

1. **Kategoria 1**: Alustat kriittisen ajattelun harjoitteluun. Nämä keskittyvät oppimisprosessiin, mutta eivät tuota auditoitavaa arviota osaamisesta. Tekoälyn käyttö varsinaisessa arvioinnissa pidetään haasteellisena (Larson ym. 2024).
2. **Kategoria 2**: Työkalut jäsenneltyjen arviointien generointiin (esim. kysymystenluonti) (Boussioux 2025). Nämä keskittyvät arvioinnin luomiseen (esim. monivalinnat) ja välttävät monimutkaisen, laadullisen todistusaineiston analysointia (Displayr 2024).

Viimeaikaiset tutkimukset vahvistavat tämän aukon: nykyiset työkalut laiminlyövät systemaattisesti "kriittisen ajattelun ja kehittyneet vuorovaikutuskyvyt" (Li ym. 2025). Näiden kategorioiden väliin jää markkinarako. Koska laadullinen arviointi on aikaa vievää (Suskie 2009), markkinoilta puuttuu työkalu, joka suorittaa automatisoidun, syvällisen ja laadullisen arvioinnin portfoliopohjaisesta todistusaineistosta korkean panoksen käyttötarkoituksiin. Tämä viitekehys luo uutta markkinakategoriaa: "automatisoitu korkean panoksen laadullinen arviointi". Sen todellinen kilpailija ei ole toinen ohjelmisto, vaan ihmisasiantuntijoiden suorittamat manuaaliset prosessit, joita se pyrkii tehostamaan.

## Luku 4: Hybridirubriikin Strateginen Kehitys

Tässä luvussa kuvataan viitekehyksen strategista kehitystä kohti syvällisempää ja auditoitavampaa analyysia. Perinteiset mallit kykenevät usein kertomaan, mitä tapahtui, mutta eivät miksi. Tästä syystä viitekehyksen Toulmin-pohjainen analyysi on suunniteltu tekemään päättelyketjusta läpinäkyvän.

### 4.1 Kiihtyvyyden ja Haurauden Paradoksi Strategisena Ajurina

Arkkitehtuurin kehityspolku on sarja strategisia valintoja reliabiliteetin ja validiteetin jännitteen hallitsemiseksi. Kehityksen taustalla on kiihtyvyyden ja haurauden paradoksi. Perusmallien nopea kehitys (Raisch & Krakowski 2021) mahdollistaa monimutkaisemmat arkkitehtuurit, mutta lisääntyvä kompleksisuus paljastaa samalla systeemisen haurauden (vrt. Brooks 1987). Tutkimus osoittaa, että moniagenttijärjestelmien (MAS) epäonnistumiset johtuvat usein koordinaatio-ongelmista, eivät yksittäisten agenttien päättelykyvystä (Cemri ym. 2025). Yksittäisen agentin älykkyyden kasvu ei ratkaise systeemisiä ongelmia (Cemri ym. 2025). Päinvastoin: älykkäämpi agentti voi argumentoida vakuuttavammin virheellisen näkemyksen puolesta (esim. myötäilyvinouman ohjaamana; Perez ym. 2022b) ja johtaa järjestelmän harhaan. Ratkaisu ei ole vain komponenttien parantaminen, vaan arkkitehtuurin on keskityttävä agenttien välisten suhteiden ja kontrollirakenteiden vahvistamiseen. Tulevaisuuden arkkitehtuuri on dynaaminen ekosysteemi, jossa hidas, auditoitava analyysi (korkea validiteetti) ja nopea, tehokas päättely (korkea reliabiliteetti) täydentävät toisiaan.

### 4.2 Kaksitasoinen Kognitiivinen Arkkitehtuuri: Järjestelmä 1 ja Järjestelmä 2

Viitekehyksen strategisen kehityksen periaatteeksi on valittu Daniel Kahnemanin kaksoisprosessiteoria (Kahneman 2011). Teorian mukaan ajattelu jakautuu kahteen järjestelmään:

* **Järjestelmä 1**: Nopea, automaattinen, intuitiivinen ja tiedostamaton.
* **Järjestelmä 2**: Hidas, analyyttinen, tietoinen ja vaatii ponnistelua.

Vaikka teoria on osa tieteellistä keskustelua (Evans & Stanovich 2013), tässä sitä käytetään strategisena analogiana arkkitehtuurin jäsentämiseen. Tavoitteena on rakentaa kaksi rinnakkaista päättelyjärjestelmää:

* **”Järjestelmä 2” – Hidas, Kallis ja Syvällinen**: Nykyinen Kognitiivinen Kvoorum takaa maksimaalisen auditoitavuuden korkean riskin tapauksissa. Linjassa uuden tutkimuksen kanssa (esim. "System-2 Attention"; Weston & Sukhbaatar 2023).
* **”Järjestelmä 1” – Nopea, Tehokas ja Automatisoitu**: Pitkän aikavälin visio, tislattu agenttimalli koulutetaan Järjestelmä 2:n datalla rutiininomaisiin arviointeihin.

Arkkitehtuuri luo itseään vahvistavan kehän. Järjestelmä 2 toimii ”datan tuotantomoottorina” ja tuottaa korkealaatuisia päättelyjälkiä (vrt. "Self-Instruct" -menetelmä; Wang ym. 2022). Tämä data on strateginen pääoma, jota käytetään Järjestelmä 1:n parantamiseen. Järjestelmä 2 -investoinnit rahoittavat siten skaalautuvan Järjestelmä 1 -ratkaisun kehittämisen.

### 4.3 Kehityspolku: Kaksitasoisen Arkkitehtuurin Rakentaminen

Kehityspolku on kolmivaiheinen. Se etenee reliabiliteetin maksimoinnista (Järjestelmä 2) kohti tehokkuuden optimointia (Järjestelmä 1).

#### 4.3.1 Auditoitavan "Järjestelmä 2:n" Perusta

Nykyinen malli perustuu tiukasti vaiheittaiseen työnkulkuun (”tiukasti sekventiaalinen”). Tämä ”Vaihe 1” edustaa nykyistä prototyyppiä. Sen vahvuus on korkea auditoitavuus ja läpinäkyvyys. Se priorisoi reliabiliteettia. Hinta on korkea latenssi ja kustannukset. Tämä kustannus on strateginen investointi, joka tuottaa korkealaatuista dataa myöhempiä vaiheita varten.

#### 4.3.2 "Järjestelmä 2:n" Tehokkuuden Optimointi

Tavoitteena on nopeuttaa prosessia ja parantaa sen virheensietokykyä siirtymällä jäykästä sekventiaalisesta ketjusta kohti suunnattua syklitöntä verkkoa (engl. *Directed Acyclic Graph*, DAG). Tässä arkkitehtuurissa järjestelmä ei etene vain lineaarisesti vaiheesta toiseen, vaan hyödyntää dynaamisia työnkulkuja, jotka mahdollistavat iteratiiviset palauteluupit (vrt. Zhang ym. 2024). Tämä malli sallii heikkolaatuiseksi todetun analyysin automaattisen palauttamisen uudelleenkäsittelyyn tietyssä solmukohdassa ilman, että koko prosessia on aloitettava alusta. Jotta auditoitavuus ei vaarantuisi tässä monimutkaisemmassa ja rinnakkaisessa rakenteessa, siirtymä edellyttää teknisesti keskitetyn ja muuttumattoman transaktiolokin toteuttamista, joka tallentaa verkon jokaisen tilasiirtymän.

#### 4.3.3 Skaalautuvan "Järjestelmä 1:n" Luominen

Pitkän aikavälin visiona on hyödyntää kyvykkyyksien tislausta (engl. *knowledge distillation*) (vrt. Hinton ym. 2015). Tämä vaihe merkitsee Järjestelmä 1:n luomista, jossa hybridirubriikin logiikka ”tiivistetään” yhdeksi malliksi. Aiempien vaiheiden tuottamaa dataa käytetään opetusaineistona yksittäisen mallin hienosäätöön. Kevyempi malli oppii jäljittelemään agenttitiimin päättelymalleja. Tislattu malli on kuitenkin vain niin luotettava kuin opetusdata, jolla se on koulutettu.

*Taulukko 2. Moniagenttiarkkitehtuurien strateginen vertailu.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Arkkitehtuuri** | **Vahvuus** | **Heikkous** | **Keskeinen kompromissi** |
| **Vaiheittainen** | Maksimaalinen auditoitavuus ja reliabiliteetti (Hybridimallin täydellinen jäljitettävyys). | Korkea latenssi ja kustannukset. | Priorisoi reliabiliteettia ja läpinäkyvyyttä nopeuden ja kustannusten kustannuksella. |
| **Rinnakkainen** | Merkittävästi matalampi latenssi. | Lisääntynyt orkestroinnin monimutkaisuus. | Priorisoi nopeutta ja tehokkuutta, mikä vaatii kehittyneempää hallintalogiikkaa. |
| **Tislattu** | Äärimmäisen matala latenssi ja kustannukset (Koko hybridilogiikka yhdessä mallissa). | Joustamattomuus; suorituskyky riippuu datan laadusta. | Priorisoi skaalautuvuutta ja käytettävyyttä auditoitavuuden kustannuksella. |

### 4.4 Tulevaisuuden Visio: Systeemisen Resilienssin Vahvistaminen

Järjestelmän pitkän aikavälin menestys edellyttää kehitystä, joka vahvistaa sen kykyä hallita häiriöitä (Perrow 1984). Tämä edellyttää Luvussa 5.1.1 tunnistettuihin riskeihin vastaamista. Lisäksi on otettava käyttöön ja validoitava ne kriittiset tekniset kontrollit, jotka on prototyypistä jätetty pois, sekä lisättävä seuraavat välttämättömät tekniset toiminnallisuudet:

* **Semanttinen Anonymisointi**: Siirtyminen nykyisestä RegEx-pohjaisesta suodatuksesta kehittyneempään, NLP/NER-pohjaiseen henkilötietojen (engl. *Personally Identifiable Information*, PII) tunnistukseen.
* **Upotusten Eheyden Tarkistus**: Vartija-agenttiin lisätään anomaliantunnistus, joka perustuu geometrisiin poikkeamiin.
* **Uudelleensijoitusmalli**: Analyytikko-agentin RAG-prosessiin integroidaan erillinen "re-ranker" -malli "lost in the middle" -ilmiön torjumiseksi.

Nämä lisäykset ovat välttämättömiä ennen kuin järjestelmää voidaan pitää tuotantokelpoisena, ja niiden käyttöönotto vaatii huolellista empiiristä testausta. Seuraavat kolme kehityskulkua tähtäävät systeemisen resilienssin lisäämiseen.

#### 4.4.1 "Järjestelmä 2:n" Päättelykyvyn Syventäminen monikierroksisella debatilla

Nykyinen staattinen rakenne voi vahvistaa systeemisiä virheitä. Tulevaisuuden suunta on dynaaminen ”agenttiekologia”, jossa agentit osallistuvat monikierroksiseen väittelyyn (*debate*) (Liang ym. 2023). Tämä mahdollistaisi todellisen debatin Kriitikon ja Loogikon välillä. Vuorovaikutus voi tuottaa syvällisempiä oivalluksia, mutta on altis sosiaalisille vinoumille. Tutkimus osoittaa, että debatit voivat johtaa virheiden vahvistumiseen, kun agentit suosivat yksimielisyyttä (Wynn ym. 2025). Tämän "konsensuksen tyrannian" vuoksi debatin tavoitteen on oltava jäsennelty erimielisyys. Syntetisoija-agentin rooli muuttuu aktiiviseksi moderaattoriksi, joka varmistaa älyllisen rehellisyyden ja raportoi vähemmistönäkemykset ihmisvarmistajalle.

#### 4.4.2 Siirtymä kohti Agentti-insinööritiedettä (Agent Engineering)

Järjestelmän strateginen jatkokehitys edellyttää radikaalia siirtymistä hauraasta kehotesuunnittelusta (engl. *prompt engineering*) kohti vankempaa ”Agent Engineering” -paradigmaa. Tässä lähestymistavassa turvallisuus ja logiikka eivät nojaa pelkkiin kielellisiin pyyntöihin, vaan ne koodataan suoraan järjestelmän rakenteisiin (Anthropic 2025c). Tämä rakenteellinen muutos tarkoittaa nykyisen, pelkkään kielimalliin nojaavan behavioraalisen suojauksen korvaamista erillisillä teknisillä luokittelijoilla, kuten Llama Guard -mallilla, jotka on optimoitu tunnistamaan ja estämään haitallinen sisältö ennen sen prosessointia (Inan ym. 2023; vrt. Anthropic 2025a).

Arkkitehtuuritasolla tämä vaatii siirtymää jäykästä sekventiaalisesta ketjusta suunnatuksi syklittömäksi verkoksi (DAG), mikä mahdollistaa iteratiiviset palauteluupit ja heikkolaatuiseksi todetun analyysin automaattisen palauttamisen uudelleenkäsittelyyn (vrt. Zhang ym. 2024). Järjestelmän validiteettia vahvistetaan samalla ulkoistamalla kausaalinen päättely simulaatioista todelliseen koodin suorittamiseen eristetyissä hiekkalaatikoissa, mikä tarjoaa deterministisen keinon todentaa väitteiden paikkansapitävyys (Turpin ym. 2025). Myötäilyvinouman (engl. *sycophancy*) torjunnassa hyödynnetään pakotettua rakenteellista erimielisyyttä käyttämällä heterogeenisia perusmalleja, mikä estää agentteja vahvistamasta toistensa virheellisiä päätelmiä (Wynn ym. 2025). Lopullinen luottamus automaatioon varmistetaan Tuomari-agentin systemaattisella hienosäädöllä (engl. *fine-tuning*), joka perustuu laajaan ihmiskalibrointiin ja tilastollisen arvioijien välisen yhdenmukaisuuden, kuten Cohenin Kappa -kertoimen, jatkuvaan seurantaan (McHugh 2012).

#### 4.4.3 Hallintamallin Sisäistäminen

Nykyinen kehotepohjainen Kognitiivinen Palomuuri on hauras (Luku 2.5.1). Kestävämpi ratkaisu on siirtyä sisäistettyyn hallintaan hyödyntämällä monikerroksista puolustusstrategiaa (CISA 2016). Tämä yhdistää (1) mallin sisäisen linjauksen perustuslaillisen tekoälyn (CAI) avulla (Bai ym. 2022) sekä (2) ulkoisen valvonnan perustuslaillisilla luokittelijoilla (Anthropic 2025a; Sharma ym. 2025). CAI-lähestymistavassa periaatteet upotetaan malliin hienosäädön avulla. Tämä kaksitasoinen puolustus on kestävä ratkaisu (vrt. Sharma ym. 2025). Tämä siirtymä on elintärkeä dynaamisen agenttiekologian luotettavuuden kannalta. Autonomisempien agenttien toiminnan on perustuttava sisäistettyyn arvopohjaan.

Nämä kolme kehityskulkua ovat toisistaan riippuvaisia ja muodostavat vision resilientistä järjestelmästä.

*Taulukko 3. Viitekehyksen kehitys kohti systeemistä resilienssiä.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Vaihe** | **Arkkitehtuuri** | **Hallintamalli** | **Anomaliantunnistus** | **Keskeinen haaste** |
| **Nykyinen** | Staattinen Kognitiivinen Kvoorum | Kehotepohjainen Kognitiivinen Palomuuri | Ristiriitojen tunnistus (Faktuaalinen) | Kognitiivisen Arviointimatriisin normatiivisen soveltamisen ja holistisen tason Mestaruus-poikkeamien tunnistamisen välinen jännite |
| **Tuleva** | Dynaaminen Agenttiekologia (Debatti) | Perustuslaillinen tekoäly (CAI) | Prosessin uskottavuusanalyysi | Holistisen tason debatin hallinta ja agenttien välinen epäjohdonmukaisuus |
| **Visio** | Itsesäätelevä Agenttiekologia | Sisäistetty ja jaettu ”perustuslaki” | Kausaalinen auditointi (Maailmanmallit) | Hybridirubriikin täydellinen sisäistäminen ja aidon kausaalisen ymmärryksen saavuttaminen |

### 4.5 Hallittu Kehitys Reliabiliteetin Varmistamiseksi

Tässä luvussa esitetty kehitys muodostaa harkitun strategian. Keskeinen oivallus on ”kiihtyvyyden ja haurauden paradoksi” (Cemri ym. 2025), johon vastaaminen vaatii arkkitehtonista viisautta. Ratkaisuksi esitetty kaksitasoinen arkkitehtuuri (Kahneman 2011) on strategian ytimessä. Tarvitaan sekä hidas ”Järjestelmä 2” (täysi kvoorum) että nopea ”Järjestelmä 1” (tislattu malli). Järjestelmien suhde on symbioottinen: Järjestelmä 2 tuottaa dataa Järjestelmä 1:n parantamiselle.

Tämä kehityspolku tarjoaa yleistettävän mallin vastuullisen tekoälyjärjestelmän kehittämiselle:

1. Aloita aina auditoitavuudesta ja luotettavuudesta (Järjestelmä 2).
2. Hyödynnä luotettavan prosessin tuottamaa dataa tehokkaampien mallien opettamiseen (Järjestelmä 1).
3. Rakenna kyvykkyyksien portfolio sen sijaan, että korvaisit vanhan kritiikittömästi uudella.
4. Pyri siirtymään ulkoisista säännöistä kohti sisäistettyjä hallintamekanismeja.

## Luku 5: Keskeiset riskit ja niiden hallinta

Tässä luvussa analysoidaan viitekehykseen liittyviä keskeisiä riskejä ja esitellään niiden hallintamekanismeja. Analyysi kattaa metodologiset ydinriskit (Luku 5.1), arkkitehtoniset riskit (Luku 5.2) sekä operatiiviset, eettiset ja teknologiset riskit (Luku 5.3).

### 5.1 Metodologiset Ydinriskit – Viitekehyksen Tieteellisen Perustan Haasteet

#### 5.1.1 Riski: Empiirisen Validoinnin Puute

**Riskin kuvaus**: Viitekehyksen keskeisin heikkous on empiirisen näytön puuttuminen. Sen uskottavuus nojaa todentamattomaan hypoteesiin korkean arvioitsijareliabiliteetin (IRR) saavuttamisesta. Tämä on merkittävä haaste, sillä laadullisten arviointien heikkous on juuri matala IRR (Baume & Yorke 2002; Koretz ym. 1994). Riskiä korostavat prototyypin teknisissä puutteissa, jotka on yksityiskohtaisesti kirjattu Vartija- ja Analyytikko-agenttien tuottamiin metodologisiin lokeihin. Järjestelmästä puuttuvat edistynyt "Semanttinen Anonymisointi" OWASP LLM02:2025 -riski ja RAG-prosessin "Uudelleensijoitusmalli" ("lost in the middle" -riski; Liu, N. F. ym. 2024). Lisäksi "Upotusten Eheyden Tarkistus" puuttuu, joten OWASP OWASP LLM08:2025 -riski on täysin hallitsematon, koska toiminto "EI OLE KÄYTÖSSÄ".

**Riski: Vektori- ja Upotushyökkäykset (OWASP LLM08:2025).** Koska nykyinen prototyyppi ei sisällä erillistä upotusten eheyden tarkistusta (Embedding Integrity Check), RAG-arkkitehtuuri on altis ”myrkytetyille” hakutuloksille (Poisoned Retrieval). \*\*Tämä puute on operationalisoitu\*\* pakottamalla Vartija-agentin kirjaamaan metodologiseen lokiin nimenomaisen varoituksen: "RAJOITUS:... LLM08-riski hallitsematon." Näiden kehittyneiden kontrollien puuttuminen luo validointivelan, jonka vaikutuksia nykyiseen arkkitehtuuriin ei ole empiirisesti testattu. Luvussa 4.4 esitellään näkymiä näiden riskien hallitsemiseksi tulevissa iteraatioissa.

**Hallintamekanismi**: Ainoa ratkaisu on luvussa 6.2 esitetty tutkimusagenda. On käynnistettävä muodollinen pilottitutkimus, joka mittaa psykometriset ominaisuudet:

* **Reliabiliteetti**: Mitataan analyyttisen tason IRR vertaamalla järjestelmän arvioita ihmisasiantuntijoiden arvioimaan ”vertailuaineistoon” (engl. *Gold Standard*).
* **Validiteetti**: Arvioidaan holistisen tason käsitevaliditeettia todentamalla, että Mestaruus-poikkeama-merkinnät korreloivat ulkoisten asiantuntija-arvioiden kanssa.

**Jäännösriski**: Riski on merkittävä, kunnes empiirinen tutkimus on suoritettu. Siihen asti viitekehys pysyy puhtaasti teoreettisena konstruktiona.

#### 5.1.2 Riski: Goodhartin Laki ja ”Performatiivinen Reflektio”

**Riskin kuvaus**: Perustavanlaatuinen uhka on Goodhartin laki (Strathern 1997), jonka mukaisesti käyttäjät voivat oppia manipuloimaan järjestelmää (Stumborg ym. 2022). Tämä ilmenee ”performatiivisena reflektiona”, joka on tässä viitekehyksessä sovellettu termi kuvaamaan tilannetta, jossa käyttäjä tuottaa vakuuttavan, mutta epäaidon narratiivin (vrt. vaikutelmien hallinta; Cullen 2020). Nykyinen arkkitehtuuri ei todennäköisesti tunnista tätä.

**Juurisyy**: Nykyinen arkkitehtuuri ei kykene aitoon kausaaliseen auditointiin (Pearl 2009; Sgaier ym. 2020). Kuten Kausaalisen Analyytikko -agentin tuottama metodologinen loki toteaa, agentin heuristiikat rajoittuvat L1/L2-tasoille, vaikka ne pyrkivät simuloimaan L3-tason kontrafaktuaalianalyysia. Nykyisten kielimallien onkin osoitettu systemaattisesti epäonnistuvan muodollisessa L3-päättelyssä (Chi ym. 2024).

**Hallintamekanismit**:

* **Nykyinen (operationaalinen)**: Syntetisoija-agentin ”Aitous-epäily”-liputus. Toteutus sisältää "Epäilyttävä Täydellisyys" -heuristiikan (ks. Luku 5.1.2.1), joka institutionalisoi epäluulon ”liian täydellisiä” suorituksia kohtaan. Tämä heuristiikka (joka on operationalisoitu osaksi Tuomari-agentin päätöksentekoa) määrittelee, että jos suoritus saa korkeimmat pisteet (Taso 4) kaikissa kriteereissä JA Kriitikko-agentin prosessiauditointi ei löydä poikkeamia, suoritus liputetaan automaattisesti anomaliaksi ja Aitous-epäilyllä. Lisäksi Kriitikko-agentin heuristiikkoja on vahvistettu L3-simulaatioilla (Luku 5.1.2.1). Mekanismien validiteetti on kuitenkin todentamatta.
* **Tulevaisuuden (strateginen)**: Siirtymä ”kausaaliseen auditointiin” integroimalla ”maailmanmalleja” (Luku 4.4.2).

**Jäännösriski:** Jäännösriski: Riski on akuutti. Viitekehys on nykymuodossaan haavoittuvainen taitavalle manipuloinnille. Tämän riskin torjumiseksi järjestelmä operoi "Performatiivisuuden tunnistuksen", joka on suora vastatoimi Goodhartin laille ("kun mittarista tulee tavoite...") (Strathern 1997; Stumborg ym. 2022) ja "performatiiviselle reflektiolle" (Cullen 2020).

Järjestelmä toteuttaa tämän torjunnan Prosessiauditoijaryhmän (Kausaalinen Analyytikko ja Performatiivisuuden Tunnistaja) kautta. Tämänhetkinen toteutus on kuitenkin rajoittunut heuristiikkoihin, jotka edustavat Pearlin kausaalihierarkian (PCH) alempia tasoja (Pearl 2009). Agentin suorittama "Temporaalinen Auditointi" (syy edeltää seurausta) ja "Kausaalinen Uskottavuus" -heuristiikka (vrt. Sgaier ym. 2020) toimivat L1-tason (Assosiaatio) ja L2-tason (Interventio) puitteissa (ks. Luku 2.4.3). Analyysin syvyyttä on parannettu implementoimalla kehittyneitä päättelyketjutekniikoita, jotka simuloivat L3-päättelyä (ks. Luku 2.4.3). Näitä ovat esimerkiksi "Kontrafaktuaalinen Stressitesti", "Abduktiivinen Haasto" ja "Pre-Mortem Analyysi" (ks. yksityiskohtainen kuvaus Luvussa 2.4.3.

Viitekehyksen suurin yksittäinen metodologinen riski on, että se ei kykene suorittamaan muodollista L3-tason (Kontrafaktuaalit) kausaalista auditointia, jota aidon performatiivisuuden tunnistaminen edellyttäisi. Tämä tarkoittaa, että järjestelmä kykenee tunnistamaan loogiset ristiriidat ja ilmeiset "mahdoton aikajana" -virheet, mutta se on edelleen altis taitavasti laaditulle, loogisesti ehyelle mutta faktuaalisesti keksitylle narratiiville. Vaikka uudet heuristiikat parantavat L3-simulaatiota, laaja empiirinen tutkimus on osoittanut, että nykyiset kielimallit epäonnistuvat systemaattisesti muodollisessa kausaalisessa ja kontrafaktuaalisessa päättelyssä (L3) (Chi ym. 2024). Tämän takia operatiivinen malli nojaa heuristiseen uskottavuuteen (esim. aikajanan tarkistus) aidon matemaattisen kausaalianalyysin sijaan. Tämä jättää kausaalisen aukon, jota performatiivinen reflektio voi hyödyntää. Tämän vuoksi järjestelmä edellyttää pakollisen metodologisen lokikirjauksen, joka pakottaa Kausaalisen Analyytikko -agentin tunnustamaan tämän rajoitteen: "RAJOITUS: Järjestelmä ei kykene muodolliseen L3-tason kausaaliseen päättelyyn, vaikka L3-simulaatioita käytetään. Riski performatiivisen reflektion tunnistamatta jäämisestä on kohonnut"

Tämä kuilu L3-vision (Luku 4.4.2) ja L1/L2-toteutuksen välillä on keskeisin este viitekehyksen täydelle validiteetille. Nykyiset "Deep Think" -mallitkaan eivät kykene luotettavasti simuloimaan kontrafaktuaaleja ilman ulkoisia kausaalisia malleja tai koodipohjaista suoritusta (vrt. Turpin ym. 2025; Aryan & Liu 2025), minkä vuoksi prototyyppi tyytyy heuristiseen uskottavuusarviointiin.

#### 5.1.3 Riski: Hybridirubriikin Sisäinen Jännite

**Riskin kuvaus**: Viitekehys institutionalisoi psykometriikan paradoksin. Jännite syntyy analyyttisen (reliabiliteetti) ja holistisen (validiteetti) tason välille, jotka ovat usein ristiriidassa.

**Hallintamekanismi**: Kaksitasoinen arkkitehtuuri (Luku 2.1) hallitsee riskiä. Konkreettinen instrumentti on Mestaruus-poikkeama-liputus, joka siirtää tulkintavastuun ihmiselle. Jännitteen hallitsemiseksi on sisäänrakennettu "Popper vs. Dreyfus" -erotteluheuristiikka:

* **Falsifioinnin (Popper) Etusija**: Vakavaa eettistä laiminlyöntiä tai faktuaalista virhettä ei voi tulkita ”mestaruus-poikkeamaksi”.
* **Mestaruuden (Dreyfus) Tunnistaminen**: ”mestaruus-poikkeama” voi ilmetä vain matriisin odotusarvojen tietoisena ja perusteltuna rikkomisena.

**Jäännösriski**: Jännite on pysyvä. Lopullinen tulkintavastuu jää aina ihmiselle, mikä asettaa korkeat vaatimukset ihmisvarmistajan asiantuntemukselle.

### 5.2 Arkkitehtoniset Riskit – Kognitiivisen Kvoorumin Sisäiset Hauraudet

#### 5.2.1 Riski: Päättelyketjujen Epäluotettavuus

**Riskin kuvaus**: Auditoitavuus perustuu oletukseen, että agenttien päättelyketjut ovat uskollisia (engl. *faithful*), eli ne heijastavat tarkasti mallin todellista päättelyprosessia (Jacovi & Goldberg 2020). Viimeaikainen tutkimus on entisestään vahvistanut tämän havainnon ja osoittanut, että jopa edistyneimmät mallit tuottavat säännöllisesti epäuskollisia päättelyketjuja, jotka rationalisoivat intuitiivisesti tuotetun vastauksen (Arcushin ym. 2025). Mallit voivat päätyä "implisiittiseen post-hoc -rationalisointiin" (Creswell ym. 2024; Arcushin ym. 2025), jolloin ne rationalisoivat jälkikäteen intuitiivisesti tuotetun vastauksen (vrt. Turpin ym. 2023). **Hallintamekanismi**: Kaksivaiheinen "teoriaohjattu prosessivalvonta". Hypoteesina on, että rationalisoinnin virheet ilmenevät argumentin rakenteessa. Valvonta toteutetaan seuraavasti:

* **Argumentin Purku (Loogikko)**: Loogikko tunnistaa argumentaatioskeeman (Walton ym. 2008) ja generoi "kriittiset kysymykset".
* **Kohdennettu Stressitesti (Epäuskollisuuden Tunnistus) (Kriitikkoryhmä)**: Kriitikko käyttää kysymyslistaa stressitestissä etsiäkseen rationalisointia ja päättelyketjun epäuskollisuutta.

Tämä mekanismi siirtää valvonnan faktantarkistuksesta päättelyn laadun auditointiin.

Jäännösriski: Riski on perustavanlaatuinen nykyisille kielimalleille. Lieventämiseksi Kriitikko-agentille on implementoitu ”Kausaalinen heuristiikka” (Luku 2.4.3), joka pakottaa arvioimaan syy-seuraussuhteen uskottavuutta (Sgaier ym. 2020).

#### 5.2.2 Riski: Systeeminen Hauraus ja Virheiden Eteneminen

**Riskin kuvaus**: Nykyinen vaiheittainen arkkitehtuuri on hauras. Virhe alkuvaiheessa etenee koko ketjun läpi. Tämä on tunnettu MAS-koordinaatio-ongelma (Cemri ym. 2025). **Hallintamekanismi**: Riskiä hallitaan kahdella päästrategialla. Ensisijainen suositus on arkkitehtoninen heterogeenisyys. Siirtyminen heterogeenisiin järjestelmiin (eri perusmallit) parantaa suorituskykyä (Ye ym. 2025). Tämän lisäksi arkkitehtuuriin on lisätty redundanssia "Ristiinvalidoiva Ketjutus" (*Cross-Validating CoT*) -mekanismilla (Luku 2.6.2). Tämä pakottaa Loogikon, Kriitikkoryhmän Tämä pakottaa Loogikon, Kriitikkoryhmän ja Tuomarin validoimaan edellisen agentin päättelyn sisäisen johdonmukaisuuden ennen oman tehtävänsä aloittamista. Tämä vähentää riskiä, että alkuvaiheen virhe ohjaa koko analyysia. Koska nykyisessä prototyyppiympäristössä (manuaalinen orkestrointi) agentit eivät pääse käsiksi ajonaikaiseen metadataan (eli ne eivät tiedä, mikä perusmalli niitä suorittaa), heterogeenisyyden automaattinen todentaminen on mahdotonta. Tämä vastuu on siksi siirretty operationaalisella säännöllä ihmisvalvojalle.

**Jäännösriski:** Homogeeninen ajo lisää merkittävästi systeemisen virheen riskiä - prototyyppi pysyy hauraana. Ratkaisu on siirtyminen rinnakkaiseen arkkitehtuuriin (Luku 4.3.2). Siihen asti reliabiliteetti riippuu korostetusti ihmisvarmistajasta (HITL). Tämän riskin hallinta on operationalisoitu pakottamalla XAI-Raportoija-agentin raportoimaan ihmisvalvojalle vastuun heterogeenisyyden varmentamisesta. Järjestelmän validiteetti edellyttää heterogeenista arkkitehtuuria, sillä homogeeninen ajo lisää riskiä systeemisten virheiden vahvistumisesta (Cemri ym. 2025) ja mitätöi aidon ristiinvalidoinnin (cross-verification) hyödyn (Ye ym. 2025).

#### 5.2.3 Riski: Debatin degeneraatio ja konsensuksen tyrannia

**Riskin kuvaus**: Vaikka adversariaalisen debatin on osoitettu parantavan päättelyä (Du ym. 2023), tuoreempi tutkimus viittaa "konsensuksen tyranniaan" (Wynn ym. 2025). Homogeenisissä ryhmissä agentit saattavat priorisoida sosiaalista mukautumista totuudenmukaisuuden kustannuksella. Tämä voi johtaa tilanteeseen, jossa virheellinen mutta enemmistön kannattama näkemys syrjäyttää oikean vähemmistönäkemyksen. Wynn ym. (2025) osoittivat, että debatti voi jopa heikentää suoritusta, jos agentit eivät ole riittävän kyvykkäitä tai jos ne ovat taipuvaisia myötäilyvinoumaan.

**Hallintamekanismit**:

* **Vinoumat**: Siirtyminen heterogeeniseen MAS-arkkitehtuuriin (suositeltu).
* **Erimielisyys**: Järjestelmä tekee erimielisyydestä strategista pääomaa. Vastaus on ”Jäsennellyn Erimielisyyden Mandaatti” (JEM), joka on toteutettu kaksitasoisesti. Kriitikko-agentti ohjeistetaan aktiivisesti ylläpitämään erimielisyyttä (Luku 2.4.3), ja Syntetisoija-agenttia kielletään pakottamasta konsensusta ja ohjeistetaan raportoimaan erimielisyydestä.

**Jäännösriski**: Nykyinen prototyyppi on altis vinoumille. Erimielisyyden tulkintavastuu siirtyy ihmiselle.

#### 5.2.4 Riski: Heterogeenisen Arkkitehtuurin Yhteen toimivuus

**Riskin kuvaus**: Suositeltu heterogeeninen arkkitehtuuri (Luku 5.2.2), jossa eri agentit käyttävät eri perusmalleja (Malli A ja Malli B), tuo mukanaan teknisen yhteen toimivuuden (engl. *interoperability*) riskin. Kun dataa (JSON) siirretään mallien välillä, on riski datan eheyden vaarantumisesta siirron tai tulkinnan aikana (vrt. ISO/IEC 25010 2023).

**Hallintamekanismi**: Riskiä hallitaan teknisillä ja semanttisilla kontrolleilla. Pakollinen "Syötteen Eheyden Validointi" on implementoitu jokaisen prosessivaiheen alkuun.

Tämä validointi tarkistaa rakenteellisen eheyden (JSON-validius) ja perustason semanttisen eheyden (merkistö). Datasiirron turvaamiseksi manuaalisessa orkestroinnissa on implementoitu "Container-enkapsulointi" (selkeät alku- ja lopputunnisteet), joka eristää JSON-objektit muusta keskusteluvirrasta. Lisäksi on implementoitu "Semanttiset Tarkistussummat". Lähettävä agentti generoi lyhyen (3–4 virkkeen) yhteenvedon tuottamastaan datasta. Vastaanottava agentti varmistaa, että datan sisältö vastaa tarkistussummaa ennen käsittelyn aloittamista. Tämä tarjoaa vahvemman suojan hienovaraisia merkistövirheitä ja inhimillisiä kopiointivirheitä (esim. osittainen kopiointi) vastaan.

**Jäännösriski**: Validointi ja tarkistussummat eivät takaa täydellistä semanttista eheyttä eivätkä tunnista semanttisia tulkintaeroja. Manuaalinen orkestrointi (datan kopiointi ja liittäminen) lisää inhimillisen virheen riskiä ja altistaa datan hienovaraisille merkistö- tai koodausvirheille siirron aikana (vrt. W3C 2008). Nämä virheet voivat muuttaa sisällön merkitystä tunnistamattomasti. Tämän vuoksi Syntetisoija-agentti ohjeistetaan kirjaamaan tämä riski pysyväksi Systeemiseksi Epävarmuudeksi lopulliseen XAI-raporttiin.

#### 5.2.5 Riski: Agenttien kognitiivinen ylikuormitus ja käyttäytymisen inversio

**Riskin kuvaus**: Pääarviointikehotteen analyysi tunnistaa kriittisen pullonkaulan, joka johtuu tiettyjen agenttien kohtuuttomasta kognitiivisesta kuormasta. Erityisesti Prosessiauditoija ja Tuomari-agentti ovat arkkitehtonisesti ylikuormitettu. Vastaavasti on konsolidoitava koko dataketju ja sovellettava subjektiivisia holistisia sääntöjä, mikä edustaa äärimmäistä kontekstin laajuutta (engl. *context width*) ja pituutta. Tutkimukset osoittavat, että tehtävän monimutkaisuus (Shen ym. 2023) sekä monimutkaisuuden ja kontekstin pituuden yhteisvaikutus heikentävät kielimallien suorituskykyä ja ohjeiden noudattamista (*instruction following*) merkittävästi (Wu ym. 2024). Tämän riskin vakavin seuraus ei ole satunnainen virhe, vaan käyttäytymisen inversio. Tutkimuksissa, joissa mallien kognitiivista kuormitusta on kasvatettu monimutkaisilla tehtävillä, mallien on havaittu hylkäävän monimutkaiset, normatiiviset (esim. oikeudenmukaisuus) ohjeet ja siirtyvän yksinkertaisempaan, rationaaliseen maksimointiin (Kirshner ym. 2025). "Kognitiiviselle Kvoorumille" tämä tarkoittaa, että ylikuormitettu Tuomari-agentti voi epäonnistua kaltaisten monimutkaisten, subjektiivisten sääntöjen soveltamisessa ja oikaista yksinkertaisempiin, mutta virheellisiin, ratkaisuihin. Tämä uhkaa suoraan koko järjestelmän validiteettia.

**Hallintamekanismi**: Lyhyellä aikavälillä riskiä hallitaan pakollisella HITL-valvonnalla ja XAI-raportoinnilla (Luku 2.5.4). Lisäksi järjestelmään on implementoitu "Kontekstin Segmentointi ja Fokusointi" -mekanismi. Tämä pakottaa Tuomari-agentin soveltamaan 'System 2 Attention' -periaatetta (Weston & Sukhbaatar 2023) luomalla Fokus- ja Kohina-listat. Tämä pyrkii vähentämään irrelevantin informaation aiheuttamaa häiriötä ja auttaa agenttia keskittymään kriittisimpiin todisteisiin ja konflikteihin. Pitkällä aikavälillä ratkaisu edellyttää arkkitehtuurin optimointia (Luku 4.3.2) tai tehtävien pilkkomista pienempiin osiin

**Jäännösriski**: Riski on korkea nykyisessä arkkitehtuurissa ja riippuvainen käytettyjen perusmallien kyvykkyydestä hallita monimutkaisuutta.

### 5.3 Operatiiviset, Eettiset ja Teknologiset Riskit – Järjestelmä Käytännössä

#### 5.3.1 Riski: Automaatioharha ja Ihmisvarmistuksen Taakka

**Riskin kuvaus**: Ihmisvarmistaja (HITL) on altis automaatioharhalle – taipumukselle luottaa epäkriittisesti järjestelmän tuotokseen (Parasuraman & Riley 1997). Mitä kehittyneempi järjestelmä, sitä suurempi riski on, että ihmisvalvoja alisuoriutuu.

**Hallintamekanismi**: Riskiä torjutaan "Interrogatiivisella Raportoinnilla" (Luku 2.5.4). Raporttipohja ei ole passiivinen tiedonanto, vaan se pakottaa ihmisvarmistajan aktiiviseen kognitiiviseen työhön. XAI-Raportoija generoi "Kriittisiä Auditointikysymyksiä", erityisesti koskien agenttien välisiä erimielisyyksiä (JEM), joihin ihmisen on otettava kantaa ("HITL-VASTAUS VAADITAAN"). Tämä vähentää taipumusta hyväksyä raportti kritiikittömästi.

**Jäännösriski**: Automaatioharha on syvälle juurtunut piirre. Kuormittunut varmistaja voi ohittaa varoitussignaalit ja kysymykset. Ihminen on samanaikaisesti järjestelmän tärkein varmistus ja merkittävin haavoittuvuus.

#### 5.3.2 Riski: Strategiset ja Eettiset Uhat

**Riskin kuvaus (Metodologinen vesittyminen)**: Organisaatiot saattavat kustannussyistä jättää holistisen tason pois. Tämä "implementoinnin tarkkuuden" haaste (Durlak & DuPre 2008) tuhoaisi järjestelmän validiteetin.

**Riskin kuvaus (Käyttötarkoituksen laajentuminen)**: Riski on, että työkalu muuttuu valvontainstrumentiksi (engl. *function creep*) (Koops 2021; AI Now Institute 2021). Järjestelmän tuottama ”kognitiivinen jälki” on arkaluonteista dataa (Weidinger ym. 2021). **Hallintamekanismit**: Vaativat hallinnollisia ratkaisuja (governance) ja teknisiä kontrolleja.

* **Eettiset riskit**: Kriitikko-agentti tunnistaa eettiset laiminlyönnit.
* **Vesittyminen**: Sitovat käyttöönottomallit, jotka vaativat täyden hybridiprosessin käyttöä korkean panoksen arvioinneissa.
* **Valvonta**: Eettiset säännöt, kuten kielto käyttää tuloksia ainoana perustana korkean panoksen päätöksille ja arvioitavan oikeus dataansa.

**Jäännösriski**: Teknologia ei estä väärinkäyttöä. Ilman vahvaa hallintamallia organisaatiot voivat käyttää työkalua väärin.

#### 5.3.3 Riski: Teknologiset tietoturvauhat (OWASP Top 10 for LLMs)

**Riskin kuvaus**: Järjestelmä on altis yleisille LLM-tietoturvariskeille (OWASP Foundation 2025f). **Hallintamekanismi**: Monikerroksinen puolustusstrategia (DiD) (Luku 2.6):

* **Tekninen Kontrollikerros**: Vartija-agentti (syötteiden puhdistus, anonymisointi) (LLM01:2025, LLM02:2025).
* **Behavioraalinen Kontrollikerros**: Kognitiivinen Palomuuri (LLM06:2025).
* **Hallinnollinen Kontrollikerros**: HITL (joka torjuu useita riskejä, kuten automaatioharhaa).

Tulevaisuudessa siirrytään perustuslailliseen tekoälyyn (CAI) (Bai ym. 2022). **Jäännösriski**: Nykyinen kehotepohjainen Kognitiivinen Palomuuri on hauras (Luku 2.5.1).

*Taulukko 4. Keskeisimmät OWASP Top 10 for LLMs –riskit ja torjuntamekanismit.*

|  |  |  |
| --- | --- | --- |
| **OWASP-riski (2025)** | **Kuvaus viitekehyksen kontekstissa** | **Ensisijainen torjuntamekanismi** |
| **LLM01: Prompt Injection** | Käyttäjä upottaa dataan piilotettuja komentoja manipuloidakseen Kvoorumia (OWASP Foundation 2025a). | Vartija-agentin suorittama syötteiden puhdistus ja aktiivinen luokittelu (Luku 2.6.1). |
| **LLM02: Sensitive Information Disclosure** | Kvoorum paljastaa tahattomasti arkaluonteista tietoa (esim. PII). | Vartija-agentin suorittama automaattinen datan anonymisointi (Luku 2.6.1). |
| **LLM03: Supply Chain Vulnerabilities** | Käytetyt ulkoiset perusmallit sisältävät haavoittuvuuksia tai muuttuvat (malliajautuminen). | Muodolliset LLMOps-käytännöt, jatkuva regressiotestaus. |
| **LLM04: Data and Model Poisoning** | Hyökkääjä manipuloi dataa, jota käytetään tislatun mallin (Järjestelmä 1) hienosäädössä. | Opetusdatana käytetään ainoastaan ihmisen validoimaa (HITL) dataa (Luku 2.6.3). |
| **LLM05: Improper Output Handling** | Järjestelmä välittää käsittelemättömän LLM-tuotoksen eteenpäin. | Systemaattinen tulosteen koodaus ja validointi (Luku 2.6.3). |
| **LLM06: Excessive Agency** | Agentit ylittävät niille määritellyt valtuudet. | Kognitiivinen Palomuuri sekä agenttien tekninen eristäminen (Luku 2.5.1 ja Luku 2.6.2). |
| **LLM08: Vector and Embedding Weaknesses** | Hyökkääjä manipuloi RAG-arkkitehtuuria. Kuten Luvussa 5.1.1 todetaan, tämän kontrollin puuttuminen on merkittävä tekninen ja metodologinen riski. | Tunnettu rajoite prototyypissä: Vartija-agentin metodologinen lokikirjaus, joka varoittaa puuttuvasta suojauksesta. (Huom: Visiona on aktiivinen ”Upotusten Eheyden Tarkistus”, mutta nykyinen implementaatio ei sisällä geometrista poikkeamien havaitsemista, jättäen riskin teknisesti hallitsemattomaksi) (OWASP Foundation 2025e). |
| **LLM09: Misinformation** | Järjestelmä tuottaa virheellistä mutta vakuuttavaa tietoa, johon ihmisvarmistaja luottaa (Automaatioharha). | Ihmisvarmistuksen (HITL) prosessi ja Syntetisoijan XAI-rooli (Luku 5.3.1). |
| **LLM10: Unbounded Consumption** | Hyökkääjä kuormittaa järjestelmää resurssi-intensiivisillä pyynnöillä. | Teknisen tason käytön rajoittaminen (rate limiting) (OWASP Foundation 2025g). |

## Luku 6: Johtopäätökset ja Tutkimusagenda

Tässä artikkelissa on esitetty uusi teoreettinen viitekehys, hybridirubriikki, ja sen operatiivinen malli, Kognitiivinen Kvoorum. Esitämme hypoteesin, että tämä kaksitasoinen arkkitehtuuri voi tarjota perinteisiä menetelmiä luotettavamman (korkeampi reliabiliteetti) ja pätevämmän (korkeampi validiteetti) tavan arvioida monimutkaista tekoälyosaamista. Tämä arkkitehtuuri toteutetaan käytännössä moniagenttisena Kognitiivisena kvoorumina ja sen perustana on Kahnemanin kaksoisprosessiteoria, mikä vahvistaa systemaattista ja auditoitavaa arviointiprosessia.

Tämän hypoteesin todentaminen edellyttää tulevaa empiiristä tutkimusta. Tämän vuoksi esitämme tutkimusagendan, jonka keskiössä on viitekehyksen ydinlupauksen systemaattinen validointi. Viitekehyksen arvo syntyy sen filosofiasta: luottamus rakennetaan auditoitavan päättelyprosessin kautta, joka hallitsee ”reliabiliteetin ja validiteetin paradoksia” (Luku 1.3).

### 6.1 Hypoteesin validoinnin ja jatkokehityksen edellytykset

Hypoteesin testaaminen ja jatkokehitys edellyttävät siirtymistä kohti monimutkaisemman, kaksitasoisen järjestelmän hallintaa. Kriittisiä tekijöitä ovat:

1. **Holistisen tason orkestroinnin hallinta**: Kognitiivisen Kvoorumin luotettava operointi edellyttää kykyä hallita agenttien työnkulkua, minimoida latenssia ja optimoida kustannuksia (vrt. Anthropic 2025b; Mesenbrink ym. 2025).
2. **HITL-valvonnan kehittäminen validoinnin ytimenä**: Validiteetin arviointi nojaa ihmisvarmistajan (HITL) kykyyn toimia tehokkaana valvojana ja ratkaista sisäiset jännitteet. Tämä edellyttää koulutusta ja työkaluja automaatioharhan tunnistamiseksi (Parasuraman & Riley 1997).
3. **Data-strategia ja hybridilogiikan tislauskyky**: Siirtymä (”Järjestelmä 1”) vaiheeseen on kriittistä skaalautuvuuden kannalta. Tämä edellyttää kykyä ”tislata” hybridirubriikin logiikka yhdelle mallille, mikä vaatii korkealaatuisen päättelydatan (”kognitiivisten jälkien”) keräämistä ja datatieteen osaamista.

### 6.2 Tutkimusagenda: Seuraavat vaiheet

Ennen empiirisen tutkimusagendan toteuttamista on ehdottoman välttämätöntä hankkia eettinen ennakkoarviointi ja hyväksyntä asiaankuuluvalta tutkimuseettiseltä toimikunnalta noudattaen Suomessa ihmistieteiden tutkimusta koskevia kansallisia ohjeistuksia (Tutkimuseettinen neuvottelukunta TENK 2019).

Koska viitekehyksen operationaalinen malli on toteutettu mutta empiirisesti testaamatta, viitekehykseltä puuttuu toistaiseksi empiirinen näyttö. Kuten luvussa 5.1.1 todetaan, sen arvo perustuu todentamattomaan hypoteesiin. Seuraava tutkimusagenda on ehdoton edellytys väitteiden validoimiseksi:

1. **Viitekehyksen ydinlupauksen todentaminen empiirisellä reliabiliteettitutkimuksella (kriittinen ja välitön prioriteetti).** Agenda jakautuu kahteen vaiheeseen:
   * **Ulkoisen validiteetin testaus (IRR)**: On välittömästi käynnistettävä vertaileva pilottitutkimus (n=50), joka mittaa Kognitiivisen Kvoorumin arvioitsijareliabiliteetin (IRR) suhteessa ihmisasiantuntijoihin. Tutkimusasetelmassa:
     + **Aineisto:** Kerätään 50 autenttista tekoälyavusteista opiskelijatyötä (sis. keskusteluhistorian).
     + **Ihmisverrokki:** Kolme riippumatonta, sokoutettua ihmisarvioijaa pisteyttää työt Hybridirubriikilla (Cohenin Kappa).
     + **Kvoorum-ajo:** Sama aineisto syötetään Kognitiiviselle Kvoorumille (heterogeeninen konfiguraatio: GPT-4 & Claude 3.5).
     + **Analyysi:** Mitataan Kvoorumin ja ihmisten välinen vastaavuus sekä Kvoorumin sisäinen konsistenssi toistomittauksilla. Tämä vastaa kritiikkiin AI-arvioinnin stokastisuudesta.
   * **Holistisen validiteetin testaus (Goodhartin Laki)**: On mitattava kestävyyttä ”performatiivista reflektiota” (Cullen 2020) vastaan. On suoritettava koe, jossa testataan ”Epäilyttävä Täydellisyys” -heuristiikan kykyä erottaa aidot suoritukset optimoiduista (Strathern 1997).
2. **Arkkitehtuurin tehokkuuden parantamisen ja data-strategian priorisointi.** Käynnistetään rinnakkaisen orkestrointimallin pilotointi ja systemaattinen päättelydatan kerääminen. Tämä data mahdollistaa tulevaisuudessa tislatun mallin (”Järjestelmä 1”) kouluttamisen.

## Lähdeluettelo

Acemoglu, D. & Restrepo, P. 2018. The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6), 1488–1542. DOI: 10.1257/aer.20160696.

Acevedo, J. A. ym. 2024. Detecting adversarial examples in RAG systems through geometric anomaly detection. *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP)*, 450–467. DOI: 10.1109/SP54038.2024.00031.

Adadi, A. & Berrada, M. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

AERA, APA & NCME. 2014. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association. [Verkkojulkaisu]. Saatavilla: <https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf>. [Haettu 14.11.2025].

Agrawal, A., Gans, J. & Goldfarb, A. 2022. *Prediction machines: The simple economics of artificial intelligence*. Boston: Harvard Business Review Press.

Ahmad, S. ym. 2024. A comprehensive review of retrieval-augmented generation (RAG): Key challenges and future directions. arXiv preprint arXiv:2410.12837. DOI: 10.48550/arXiv.2410.12837.

Ahuna, K. & Kiener, M. 2025. Beyond digital literacy: Cultivating “meta AI” skills in students and faculty. *Faculty Focus*. Julkaistu 6.8.2025. [Verkkojulkaisu]. Saatavilla: <https://www.facultyfocus.com/articles/teaching-with-technology-articles/beyond-digital-literacy-cultivating-meta-ai-skills-in-students-and-faculty/>. [Haettu 14.11.2025].

AI Now Institute. 2021. *A New AI Lexicon: Function Creep*. New York: AI Now Institute. [Verkkojulkaisu]. Saatavilla: <https://ainowinstitute.org/publications/collection/a-new-ai-lexicon-function-creep>. [Haettu 14.11.2025].

Anderson, L. W. & Krathwohl, D. R. (toim.). 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives*. New York: Longman.

Anthropic. 2025a. Constitutional classifiers. Anthropic Policy & Research. [Verkkojulkaisu]. Saatavilla: <https://www.anthropic.com/research/constitutional-classifiers>. [Haettu 14.11.2025].

Anthropic. 2025b. How we built our multi-agent research system. Anthropic Engineering Blog. Julkaistu 13.6.2025. [Verkkojulkaisu]. Saatavilla: <https://www.anthropic.com/engineering/multi-agent-research-system>. [Haettu 14.11.2025].

Anthropic. 2025c. Building effective agents. Anthropic Research. [Verkkojulkaisu]. Saatavilla: <https://www.anthropic.com/research/building-effective-agents>. [Haettu 14.11.2025].

Arcushin, I. ym. 2025. Chain-of-thought reasoning in the wild is not always faithful. arXiv preprint arXiv:2503.08679. DOI: 10.48550/arXiv.2503.08679.

Aryan, A. & Liu, Z. 2025. Causal Reflection with Language Models. arXiv preprint arXiv:2508.04495. DOI: 10.48550/ARXIV.2508.04495.

Auzmor. 2024. How to measure the ROI of AI training programs. Auzmor. [Verkkojulkaisu]. Saatavilla: <https://auzmor.com/blog/measure-the-roi-of-ai-training-programs/>. [Haettu 14.11.2025].

Bai, Y. ym. 2022. Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073. DOI: 10.48550/arXiv.2212.08073.

Bareinboim, E. ym. 2022. On Pearl's hierarchy and the foundations of causal inference. Teoksessa H. Geffner, R. Dechter & J. Halpern (toim.), *Probabilistic and causal inference: The works of Judea Pearl*. New York: Association for Computing Machinery, 507–556. DOI: 10.1145/3501714.3501743.

Baume, D. & Yorke, M. 2002. The reliability of assessment by portfolio on a course to develop and accredit teachers in higher education. *Studies in Higher Education*, 27(1), 7–25. DOI: 10.1080/03075070120099340.

Bezanilla, M. J. ym. 2019. Methodologies for teaching-learning in higher education and their relationship with student competences: A systematic review. *Educational Research Review*, 27, 83–98. DOI: 10.1016/j.edurev.2019.01.004.

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. 2004. The concept of validity. *Psychological Review*, 111(4), 1061–1071. DOI: 10.1037/0033-295X.111.4.1061.

Boussioux, L. 2025. Revolutionize quality assurance with AI. Mareana. [Verkkojulkaisu]. Saatavilla: <https://mareana.com/whitepaper/qa-playbook/>. [Haettu 14.11.2025].

Brooks, F. P. 1987. No silver bullet: Essence and accidents of software engineering. *Computer*, 20(4), 10–19. DOI: 10.1109/MC.1987.1663532.

Bulut, O. ym. 2024. The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges. *Chinese/English Journal of Educational Measurement and Evaluation*, 5(3), Artikla 3. DOI: 10.59863/MIQL7785.

Carolus, A. ym. 2023. MAILS - Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. arXiv preprint arXiv:2302.09319. DOI: 10.48550/arXiv.2302.09319.

Cemri, M. ym. 2025. Why do multi-agent LLM systems fail? arXiv preprint arXiv:2503.13657. DOI: 10.48550/arXiv.2503.13657.

Center for Innovative Teaching & Learning. 2025. Authentic assessment. Indiana University Bloomington. [Verkkojulkaisu]. Saatavilla: <https://citl.indiana.edu/teaching-resources/assessing-student-learning/authentic-assessment/index.html>. [Haettu 14.11.2025].

Cheng, P. C-H. 2001. Scientific discovery, computational models of. Teoksessa N. J. Smelser & P. B. Baltes (toim.), *International encyclopedia of the social & behavioral sciences*. Amsterdam: Elsevier, 13783–13787. DOI: 10.1016/B978-0-08-097086-8.43085-0.

Cheng, P. 2021. Competence assessment by stimulus matching: an application of GOMS to assess chunks in memory. Teoksessa *Proceedings of the 19th International Conference on Cognitive Modelling (ICCM 2021)*. [Verkkojulkaisu]. Saatavilla: <https://cidlab.com/files/smp/pb/pb-2021.pdf>. [Haettu 14.11.2025].

Chi, H. ym. 2024. Unveiling causal reasoning in large language models: Reality or mirage? *Advances in Neural Information Processing Systems*, 37, 96640–96670. DOI: 10.48550/arXiv.2506.21215.

CISA. 2016. Defense in depth. Cybersecurity and Infrastructure Security Agency. [Verkkojulkaisu]. Saatavilla: <https://www.cisa.gov/sites/default/files/recommended_practices/NCCIC_ICS-CERT_Defense_in_Depth_2016_S508C.pdf>. [Haettu 14.11.2025].

Cohen, R. J., Swerdlik, M. E. & Phillips, S. M. 1996. *Psychological testing and assessment: An introduction to tests and measurement*. 3. painos. Mountain View: Mayfield Publishing Company.

Creswell, A. ym. 2024. Reducing post-hoc rationalization in large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, 14757–14771. DOI: 10.18653/v1/2024.findings-acl.867.

Crusius, T. W. & Channell, C. E. 2003. *The aims of argument: A text and reader*. 4. painos. New York: McGraw-Hill.

Cullen, M. J. 2020. Faking in high-stakes selection: A call to integrate empirical research and applied practice. *International Journal of Selection and Assessment*, 28(3), 223–226. DOI: 10.1111/ijsa.12289.

D'Angelo, M. 2025. AI safety vs AI security in LLM applications: What teams must know. promptfoo. [Verkkojulkaisu]. Saatavilla: <https://www.promptfoo.dev/blog/ai-safety-vs-security/>. [Haettu 14.11.2025].

David, J. L. 2019. 15 reasons why standardized tests are problematic. *ASCD Blog*. [Verkkojulkaisu]. Saatavilla: <https://www.ascd.org/blogs/15-reasons-why-standardized-tests-are-problematic>. [Haettu 14.11.2025].

de Bruin, A. B. H., van Merriënboer, J. J. G. & van Gog, T. 2023. The role of cognitive effort in fostering the acquisition of complex cognitive skills. Teoksessa J. Sweller, J. J. G. van Merriënboer & F. Paas (toim.), *Cognitive load theory: A research-based guide to instructional design*. Cambridge: Cambridge University Press, 237–256. DOI: 10.1017/9781009403718.011.

Denning, D. E. & Denning, P. J. 1977. Certification of programs for secure information flow. *Communications of the ACM*, 20(7), 504–513. DOI: 10.1145/359636.359712.

Der Kiureghian, A. & Ditlevsen, O. 2009. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2), 105–112. DOI: 10.1016/j.strusafe.2008.06.020.

Disco. 2024. How to assess the ROI of AI-driven upskilling initiatives. Disco. [Verkkojulkaisu]. Saatavilla: <https://www.disco.co/blog/how-to-assess-the-roi-of-ai-driven-upskilling-initiatives>. [Haettu 14.11.2025].

Displayr. 2024. Discover the 5 best AI tools for qualitative data analysis. Displayr. [Verkkojulkaisu]. Saatavilla: <https://www.displayr.com/discover-the-5-best-ai-tools-for-qualitative-data-analysis/>. [Haettu 14.11.2025].

Dreyfus, S. E. & Dreyfus, H. L. 1980. *A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition*. California Univ Berkeley Operations Research Center. [Verkkojulkaisu]. Saatavilla: <https://apps.dtic.mil/sti/pdfs/ADA084551.pdf>. [Haettu 18.11.2025].

Du, Y. ym. 2023. Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325. DOI: 10.48550/arXiv.2305.14325.

Dufner, M. ym. 2019. Self-enhancement and psychological adjustment: A meta-analytic review. *Personality and Social Psychology Review*, 23(1), 48–72. DOI: 10.1177/1088868318756467.

Duhem, P. 1906. *La théorie physique: son objet et sa structure*. Paris: Chevalier & Rivière.

Durlak, J. A. & DuPre, E. P. 2008. Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41(3–4), 327–350. DOI: 10.1007/s10464-008-9165-0.

Eloundou, T., Manning, S., Mishkin, P. & Rock, D. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. arXiv preprint arXiv:2303.10130. DOI: 10.48550/arXiv.2303.10130.

Euroopan komissio. 2024a. The AI Act. Bryssel: Euroopan komissio. [Verkkojulkaisu]. Saatavilla: <https://artificialintelligenceact.eu/article/14/>. [Haettu 14.11.2025].

Euroopan komission korkean tason asiantuntijaryhmä. 2019. Ethics guidelines for trustworthy AI. Bryssel: Euroopan komissio. [Verkkojulkaisu]. Saatavilla: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-for-trustworthy-ai>. [Haettu 14.11.2025].

Evans, J. St. B. T. & Stanovich, K. E. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223–241. DOI: 10.1177/1745691612460685.

FairTest. 2012. The limits of standardized tests for diagnosing and assisting student learning. FairTest: The National Center for Fair & Open Testing. [Verkkojulkaisu]. Saatavilla: <https://fairtest.org/limits-standardized-tests-diagnosing-and-assisting/>. [Haettu 14.11.2025].

Federiakin, D. ym. 2024. Prompt engineering: A new skill for the future of work. *Procedia Computer Science*, 231, 401–409. DOI: 10.1016/j.procs.2023.12.233.

Festinger, L. 1957. *A theory of cognitive dissonance*. Stanford: Stanford University Press.

Flavell, J. H. 1979. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911. DOI: 10.1037/0003-066X.34.10.906.

Fügener, A., Walzner, D. D. & Gupta, A. 2025. Roles of Artificial Intelligence in Collaboration with Humans: Automation, Augmentation, and the Future of Work. *Management Science*, 0. DOI: 10.1287/mnsc.2024.05684.

Ganguli, D. ym. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858. DOI: 10.48550/arXiv.2209.07858.

Gao, L. ym. 2022. Precise zero-shot dense retrieval without relevance labels. arXiv preprint arXiv:2212.10496. DOI: 10.48550/arXiv.2212.10496.

Goffman, E. 1959. *The presentation of self in everyday life*. New York: Doubleday.

Goodfellow, I. J. ym. 2014. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 27, 2672–2680. DOI: 10.48550/arXiv.1406.2661.

Google DeepMind 2025a: *Gemini 3 Pro Model Card*. [PDF]. Saatavilla: <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf>.

Google DeepMind 2025b: *Gemini 3 Pro Model Evaluation*. [PDF]. Saatavilla: <https://storage.googleapis.com/deepmind-media/gemini/gemini_3_pro_model_evaluation.pdf>.

Google DeepMind 2025c: Gemini 3: A new era of intelligence. – *Google Blog*. Julkaistu 18.11.2025. Saatavilla: <https://blog.google/products/gemini/gemini-3/>.

Greshake, K. ym. 2023. Not what you’ve signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. arXiv preprint arXiv:2302.12173. DOI: 10.48550/arXiv.2302.12173.

Guo, T. ym. 2024. Large language model based multi-agents: A survey of progress and challenges. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 8048–8057. DOI: 10.24963/ijcai.2024/890.

Halpern, D. F. 2014. *Thought and knowledge: An introduction to critical thinking*. 5. painos. New York: Psychology Press. DOI: 10.4324/9781315885278.

Hazan, E. ym. 2024. A new future of work: The race to deploy AI and raise skills in Europe and beyond. McKinsey Global Institute. [Verkkojulkaisu]. Saatavilla: <https://www.mckinsey.com/mgi/our-research/a-new-future-of-work-the-race-to-deploy-ai-and-raise-skills-in-europe-and-beyond>. [Haettu 14.11.2025].

Hinton, G., Vinyals, O. & Dean, J. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531. DOI: 10.48550/arXiv.1503.02531.

Huang, L. ym. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*. DOI: 10.1145/3703155.

Hüllermeier, E. & Waegeman, W. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110, 457–506. DOI: 10.1007/s10994-021-05946-3.

Inan, H. ym. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv preprint arXiv:2312.06674. DOI: 10.48550/arXiv.2312.06674.

ISACA. 2025. 2025 Volume 5 How to measure and prove the value of your AI investments. ISACA. [Verkkojulkaisu]. Saatavilla: <https://www.isaca.org/resources/news-and-trends/newsletters/atisaca/2025/volume-5/how-to-measure-and-prove-the-value-of-your-ai-investments>. [Haettu 14.11.2025].

ISO/IEC. 2023. *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models* (ISO/IEC 25010:2023). Geneve: International Organization for Standardization. [Verkkojulkaisu]. Saatavilla: <https://www.iso.org/standard/84727.html>. [Haettu 14.11.2025].

Jacobs, R., Kafry, D. & Zedeck, S. 1980. Expectations of behaviorally anchored rating scales. *Personnel Psychology*, 33(3), 595–640. DOI: 10.1111/j.1744-6570.1980.tb00486.x.

Jacovi, A. & Goldberg, Y. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? Teoksessa *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.385.

Jagerman, R. ym. 2023. Query expansion by prompting large language models. arXiv preprint arXiv:2305.03653. DOI: 10.48550/arXiv.2305.03653.

Jia, Y., Shao, Z., Liu, Y., Jia, J., Song, D. & Gong, N. Z. 2025. A Critical Evaluation of Defenses against Prompt Injection Attacks. arXiv preprint arXiv:2505.18333. DOI: 10.48550/arXiv.2505.18333.

Johnson, R. B. & Onwuegbuzie, A. J. 2004. Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7), 14–26. DOI: 10.3102/0013189X033007014.

Jonsson, A. & Svingby, G. 2007. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144. DOI: 10.1016/j.edurev.2007.05.002.

Kahneman, D. 2011. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kiciman, E. ym. 2023. Causal reasoning and large language models: Opening a new frontier for causality. arXiv preprint arXiv:2305.00050. DOI: 10.48550/arXiv.2305.00050.

Kim, D.-G. ym. 2022. Assessing non-technical skills in medical students: An evaluation of the inter- and intra-rater reliability of the behaviorally anchored rating scale (BARS). *Teaching and Learning in Medicine*, 35(3), 310–319. DOI: 10.1080/10872981.2022.2070940.

Kinicki, A. J. ym. 1985. Behaviorally anchored rating scales vs. summated rating scales: Psychometric properties and susceptibility to rating bias. *Educational and Psychological Measurement*, 45(3), 535–549. DOI: 10.1177/001316448504500310.

Kirshner, S., Klaben, B. & Dobbe, S. 2025. Instruction-Following: The Truth Is In There, But Is It In The Loss? arXiv preprint arXiv:2511.07973. DOI: 10.48550/arXiv.2511.07973.

Klein, G. 2007. Performing a Project Premortem. *Harvard Business Review*, 85(9), 18–19.

Klieger, D. M. ym. 2018. Development of the Behaviorally Anchored Rating Scales for the Skills Demonstration and Progression Guide. *ETS Research Report Series RR-18-24*. Educational Testing Service. DOI: 10.1002/ets2.12210.

Koops, B.-J. 2021. The concept of function creep. *Law, Innovation and Technology*, 13(1), 29–56. DOI: 10.1080/17579961.2021.1898299.

Koretz, D. M. ym. 1994. The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, 13(3), 5–16. DOI: 10.1111/j.1745-3992.1994.tb00854.x.

Kreuzberger, D., Kühl, N. & Hirschl, S. 2023. Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access*, 11, 31866–31879. DOI: 10.1109/ACCESS.2023.3262138.

Kruger, J. & Dunning, D. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134. DOI: 10.1037/0022-3514.77.6.1121.

Lagnado, D. A. & Sloman, S. A. 2006. Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 32(3), 451–460. DOI: 10.1037/0278-7393.32.3.451.

Lane, S. 2013. Validity evidence for assessments of higher-order thinking. *Journal of Educational Measurement*, 50(4), 399–430. DOI: 10.1111/jedm.12028.

Larson, B. Z. ym. 2024. Critical thinking in the age of generative AI. *Academy of Management Learning & Education*, 23(3). DOI: 10.5465/amle.2024.0338.

LeCun, Y. 2022. A path towards autonomous machine intelligence. *OpenReview*. [Verkkojulkaisu]. Saatavilla: <https://openreview.net/forum?id=BZ5a1r-kVsf>. [Haettu 14.11.2025].

Levashina, J. & Morgeson, F. P. 2007. Applicant faking on personality measures: A coping perspective. *Academy of Management Review*, 32(4), 1118–1136. DOI: 10.5465/amr.2007.26586083.

Levine, E. L., Ash, R. A. & Bennett, N. 1988. The "behavioral consistency" approach to job analysis: A critical reappraisal. *Human Resource Management Review*, 8(3), 273–293. DOI: 10.1016/S1053-4822(98)90023-6.

Lewis, P. ym. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. DOI: 10.48550/arXiv.2005.11401.

Li, F. ym. 2025. An assessment of human–AI interaction capability in the generative AI era: The influence of critical thinking. *Journal of Intelligence*, 13(6), 62. DOI: 10.3390/jintelligence13060062.

Li, Z. ym. 2024. PII-Bench: A benchmark for personally identifiable information (PII) detection and anonymization. arXiv preprint arXiv:2404.03893. DOI: 10.48550/arXiv.2404.03893.

Liang, T.-S. ym. 2023. Encouraging divergent thinking in large language models through multi-agent debate. arXiv preprint arXiv:2305.19118. DOI: 10.48550/arXiv.2305.19118.

Lippi, M. & Torroni, P. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2), 1–25. DOI: 10.1145/2850417.

Lison, P. ym. 2021. Anonymisation models for text data: State of the art, challenges and future directions. arXiv preprint arXiv:2106.04631. DOI: 10.48550/arXiv.2106.04631.

Liu, N. F. ym. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173. DOI: 10.1162/tacl\_a\_00638.

Liu, X. ym. 2024. Automatic and universal prompt injection attacks against large language models. arXiv preprint arXiv:2403.04957. DOI: 10.48550/arXiv.2403.04957.

Liu, Y. ym. 2023. Prompt injection attacks and defenses in large language models: A survey. arXiv preprint arXiv:2310.12815. DOI: 10.48550/arXiv.2310.12815.

Luckin, R. ym. 2017. Towards artificial intelligence-based assessment systems. *Nature Human Behaviour*, 1(3), 0028. DOI: 10.1038/s41562-016-0028.

Ma, Y. ym. 2024. Mitigating contextual information loss in RAG models through re-ranking. arXiv preprint arXiv:2401.06427. DOI: 10.48550/arXiv.2401.06427.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3), 276–282. DOI: 10.11613/BM.2012.031.

Mesenbrink, H. ym. 2025. Orchestrated multi agents sustain accuracy under clinical-scale workloads compared to a single agent. *medRxiv*. DOI: 10.1101/2025.08.22.25334049.

Messick, R. J. 2003. Substance and structure in assessment arguments. *Law, Probability and Risk*, 2(4), 237–258. DOI: 10.1093/lpr/2.4.237.

Messick, S. 1989. Validity. Teoksessa R. L. Linn (toim.), *Educational measurement*. 3. painos. New York: Macmillan, 13–103.

Meyer, S. 2017. A popperian falsification of artificial intelligence - Lighthill. arXiv preprint arXiv:1704.08111. DOI: 10.48550/arXiv.1704.08111.

Morgeson, F. P., Delaney-Klinger, K. & Hemingway, M. A. 2007. The importance of job analysis to the legal defensibility of an organization's selection system. Teoksessa L. L. Koppes (toim.), *Historical perspectives in industrial and organizational psychology*. Mahwah: Lawrence Erlbaum Associates, 301–322.

Moskal, B. M. 2000. Scoring rubrics: What, when and how? *Practical Assessment, Research, and Evaluation*, 7(3). DOI: 10.7275/a5vq-7q66.

Nola, R. & Sankey, H. 2014. *Theories of scientific method: An introduction*. Lontoo: Routledge. DOI: 10.4324/9781315728666.

Nold, H. & Michel, L. 2022. The Dunning-Kruger Effect on Organizational Agility. *Academy of Management Proceedings*, 2022(1). DOI: 10.5465/AMBPP.2022.10365abstract.

OECD. 2024. *Artificial intelligence and the changing demand for skills in the labour market*. OECD Artificial Intelligence Papers, No. 14. Paris: OECD Publishing. DOI: 10.1787/88684e36-en.

OpenAI 2024: *OpenAI o1 System Card*. OpenAI. Saatavilla: <https://openai.com/index/openai-o1-system-card/>. (arXiv preprint arXiv:2412.16720).

OWASP Foundation. 2023[a]. Input Validation Cheat Sheet. [Verkkojulkaisu]. Saatavilla: <https://cheatsheetseries.owasp.org/cheatsheets/Input_Validation_Cheat_Sheet.html>. [Haettu 18.11.2025].

OWASP Foundation. i.a.[b]. Data Validation. [Verkkojulkaisu]. Saatavilla: <https://owasp.org/www-community/Data_Validation>. [Haettu 18.11.2025].

OWASP Foundation. i.a.[c]. LLM Agent Security. [Verkkojulkaisu]. Saatavilla: [https://owasp.org/www-project-llm-agent-security/](https://www.google.com/search?q=https://owasp.org/www-project-llm-agent-security/). [Haettu 20.11.2025].

OWASP Foundation. 2025a. LLM01:2025 Prompt Injection. GenAI OWASP Top 10. [Verkkojulkaisu]. Saatavilla: <https://genai.owasp.org/llmrisk/llm01-prompt-injection/>. [Haettu 15.11.2025].

OWASP Foundation. 2025b. LLM02:2025 Sensitive Information Disclosure. GenAI OWASP Top 10. [Verkkojulkaisu]. Saatavilla: <https://genai.owasp.org/llm-top-10/>. [Haettu 15.11.2025].

OWASP Foundation. 2025c. LLM05:2025 Improper Output Handling. GenAI OWASP Top 10. [Verkkojulkaisu]. Saatavilla: <https://genai.owasp.org/llm-top-10/>. [Haettu 15.11.2025].

OWASP Foundation. 2025d. LLM06:2025 Excessive Agency. GenAI OWASP Top 10. [Verkkojulkaisu]. Saatavilla: <https://genai.owasp.org/llm-top-10/>. [Haettu 15.11.2025].

OWASP Foundation. 2025e. LLM08:2025 Vector and Embedding Weaknesses. GenAI OWASP Top 10. [Verkkojulkaisu]. Saatavilla: <https://genai.owasp.org/llmrisk/llm08-vector-and-embedding-weaknesses/>. [Haettu 15.11.2025].

OWASP Foundation. 2025f. OWASP Top 10 for Gen AI. GenAI OWASP Top 10. [Verkkojulkaisu]. Saatavilla: <https://genai.owasp.org/llm-top-10/>. [Haettu 15.11.2025].

OWASP Foundation. 2025g. LLM10:2025 Unbounded Consumption. GenAI OWASP Top 10. [Verkkojulkaisu]. Saatavilla: <https://genai.owasp.org/llm-top-10/>. [Haettu 15.11.2025].

Parasuraman, R. & Riley, V. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. DOI: 10.1518/001872097778543886.

Paulson, F. L., Paulson, P. R. & Meyer, C. A. 1991. What makes a portfolio a portfolio. *Educational Leadership*, 48(5), 60–63.

Pearl, J. 2009. *Causality: Models, reasoning, and inference*. 2. painos. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511803161.

Perez, E. ym. 2022a. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv preprint arXiv:2209.07858. DOI: 10.48550/arXiv.2209.07858.

Perez, E. ym. 2022b. Discovering Language Model Behaviors with Model-Written Evaluations. arXiv preprint arXiv:2212.09251. DOI: 10.48550/arXiv.2212.09251.

Perrow, C. 1984. *Normal accidents: Living with high-risk technologies*. Princeton: Princeton University Press.

Pfeifer, K. 2025. Humanity-in-the-loop: Human AI oversight is an imperative. Medium. Julkaistu 22.10.2025. [Verkkojulkaisu]. Saatavilla: <https://medium.com/@karenpfeifer/humanity-in-the-loop-human-ai-oversight-is-an-imperative-50bdcc2688d8>. [Haettu 14.11.2025].

Polanyi, M. 1966. *The tacit dimension*. Chicago: University of Chicago Press.

Popper, K. 1934. *Logik der Forschung*. Vienna: Julius Springer.

PwC. 2024. AI jobs barometer. PricewaterhouseCoopers. [Verkkojulkaisu]. Saatavilla: <https://www.pwc.com/gx/en/issues/artificial-intelligence/ai-jobs-barometer.html>. [Haettu 14.11.2025].

Quine, W. V. O. 1951. Two dogmas of empiricism. *The Philosophical Review*, 60(1), 20–43. DOI: 10.2307/2181906.

Raisch, S. & Krakowski, S. 2021. Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*, 46(1), 192–210. DOI: 10.5465/amr.2018.0072.

Reinecke, K. & Gajos, K. Z. 2014. Quantifying visual preferences around the world. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 717–726. DOI: 10.1145/2556288.2557076.

Research AIMultiple. 2025. 15 Security Threats to LLM Agents (with Real-World Examples). [Verkkojulkaisu]. Saatavilla: <https://research.aimultiple.com/security-of-ai-agents/>. [Haettu 16.11.2025].

Sadler, D. R. 1989. Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144. DOI: 10.1007/BF00117714.

Sagi, O. & Rokach, L. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1249. DOI: 10.1002/widm.1249.

Saito, K., Wachi, A. & Akimoto, Y. 2023. Verbosity bias in preference labeling by large language models. arXiv preprint arXiv:2310.10864. DOI: 10.48550/arXiv.2310.10864.

Saltzer, J. H. & Schroeder, M. D. 1975. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9), 1278–1308. DOI: 10.1109/PROC.1975.9939.

Sgaier, S. K. ym. 2020. The case for causal AI. *Stanford Social Innovation Review*, 18(3), 50–55. DOI: 10.48558/KT81-SN73.

Shafiyeva, U. 2021. Assessing Students' Minds: Developing Critical Thinking or Fitting into Procrustean Bed. *European Journal of Education*, 4(2), 78–91. DOI: 10.26417/452bxv17s.

Sharma, M. ym. 2025. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. arXiv preprint arXiv:2501.18837. DOI: 10.48550/arXiv.2501.18837.

Shavelson, R. J. 2010. On the measurement of competency. *Empirical Research in Vocational Education and Training*, 2(1), 41–63.

Shavelson, R. J. 2013. On an approach to testing and modeling competence. *Educational Psychologist*, 48(2), 73–86. DOI: 10.1080/00461520.2013.779483.

Shen, Y. ym. 2023. Large Language Models as Tool Makers. arXiv preprint arXiv:2305.17126. DOI: 10.48550/arXiv.2305.17126.

Shinn, N. ym. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. arXiv preprint arXiv:2303.11366. DOI: 10.48550/arXiv.2303.11366.

Shuster, K. ym. 2021. Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv:2104.07567. DOI: 10.48550/arXiv.2104.07567.

Silva, B. ym. 2025. Development of an Adapted Version of the Motor Competence Assessment (MCA) for Older Adults. *Journal of Clinical Medicine*, 14(21), 7866. DOI: 10.3390/jcm14217866.

Smith, P. C. & Kendall, L. M. 1963. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 47(2), 149–155. DOI: 10.1037/h0047060.

Strathern, M. 1997. 'Improving ratings': audit in the British university system. *European Review*, 5(3), 305–321. DOI: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.

Stumborg, M. F. ym. 2022. *Goodhart's law: Recognizing and mitigating the manipulation of measures in analysis*. CNA Occasional Paper. [Verkkojulkaisu]. Saatavilla: <https://www.cna.org/reports/2022/09/Goodharts-Law-Recognizing-Mitigating-Manipulation-Measures-in-Analysis.pdf>. [Haettu 14.11.2025].

Supianto, A. A. ym. 2023. A systematic review of multi-agent systems in educational assessment. *Computers & Education: Artificial Intelligence*, 4, 100135. DOI: 10.1016/j.caeai.2023.100135.

Suskie, L. 2009. *Assessing student learning: A common sense guide*. 2. painos. San Francisco: Jossey-Bass.

Talboy, A. & Fuller, E. 2023. Large language models show humanlike cognitive biases. arXiv preprint arXiv:2308.14343. DOI: 10.48550/arXiv.2308.14343.

Tétard, F. & Collan, M. 2009. Lazy User Theory: A Dynamic Model to Understand User Selection of Products and Services. *Proceedings of the 42nd Hawaii International Conference on System Sciences*, 1–10. IEEE.

Toulmin, S. E. 2003. *The uses of argument*. Päivitetty painos. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511802031.

Towards AI. 2025. AI Sandbox in 2025: How Enterprises and Governments Shape AI's Future. Julkaistu 26.9.2025. [Verkkojulkaisu]. Saatavilla: <https://pub.towardsai.net/ai-sandbox-in-2025-how-enterprises-and-governments-shape-ais-future-b41f0d267c4d>. [Haettu 14.11.2025].

Trivedi, H. ym. 2024. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. arXiv preprint arXiv:2401.10133. DOI: 10.48550/arXiv.2401.10133.

Turpin, M. ym. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Teoksessa A. Oh, T. Hashimoto & D. Blei (toim.), *Advances in Neural Information Processing Systems 36*. La Jolla: Neural Information Processing Systems Foundation, 21016–21033.

Turpin, M. ym. 2025. Executable counterfactuals: Improving LLMs' causal reasoning through code. arXiv preprint arXiv:2510.01539. DOI: 10.48550/arXiv.2510.01539.

Tutkimuseettinen neuvottelukunta TENK. 2019. Ihmiseen kohdistuvan tutkimuksen eettiset periaatteet ja ihmistieteiden eettinen ennakkoarviointi Suomessa. Tutkimuseettisen neuvottelukunnan ohje 2019. *Tutkimuseettisen neuvottelukunnan julkaisuja 3/2019*. Helsinki. [Verkkojulkaisu]. Saatavilla: <https://tenk.fi/sites/default/files/2021-01/Ihmistieteiden_eettisen_ennakkoarvioinnin_ohje_2020.pdf>. [Haettu 14.11.2025].

Tversky, A. & Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. DOI: 10.1126/science.185.4157.1124.

W3C. 2008. Migrating to Unicode. W3C Internationalization (I18n) Activity. [Verkkojulkaisu]. Saatavilla: <https://www.w3.org/International/articles/unicode-migration/>. [Haettu 14.11.2025].

Wachsmuth, H. ym. 2017. Computational argumentation quality assessment in natural language. *Proceedings of the 15th Conference of the EACL*, 176–187. DOI: 10.18653/v1/E17-1017.

Walton, D. N., Reed, C. & Macagno, F. 2008. *Argumentation schemes*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511802034.

Wang, Y. ym. 2023. A survey on an authoritarian bias: The blind spot of large language models. arXiv preprint arXiv:2312.06086. DOI: 10.48550/arXiv.2312.06086.

Weidinger, L. ym. 2021. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359. DOI: 10.48550/arXiv.2112.04359.

Weston, J. & Sukhbaatar, S. 2023. System 2 attention (is something you might need too). arXiv preprint arXiv:2311.11829. DOI: 10.48550/arXiv.2311.11829.

Wiggins, G. 1998. *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.

Wisse, G. & Greve, R. 2023. AI in educational assessment: A systematic review of formative and summative applications. *Computers & Education: Artificial Intelligence*, 5, 100174. DOI: 10.1016/j.caeai.2023.100174.

Wolf, K. & Stevens, E. 2007. The role of rubrics in advancing and assessing student learning. *The Journal of Effective Teaching*, 7(1), 3–14.

Wolf, Y. ym. 2023. Fundamental Limitations of Alignment in Large Language Models. arXiv preprint arXiv:2304.11082. DOI: 10.48550/arXiv.2304.11082.

Wolters Kluwer. 2024. 2024 Future ready lawyer survey report. [Verkkojulkaisu]. Saatavilla: <https://www.wolterskluwer.com/en/know/future-ready-lawyer-2024>. [Haettu 14.11.2025].

Wooldridge, M. 2009. *An introduction to multiagent systems*. 2. painos. Chichester: John Wiley & Sons.

World Economic Forum. 2023. *Future of Jobs Report 2023*. [Verkkojulkaisu]. Saatavilla: <https://www.weforum.org/publications/the-future-of-jobs-report-2023/>. [Haettu 18.11.2025].

Wu, J. ym. 2024. Large Language Models are Challenged by an Abundance of Over-complicated Instructions. arXiv preprint arXiv:2409.07844. DOI: 10.48550/arXiv.2409.07844.

Wynn, A., Satija, H. & Hadfield, G. 2025. Talk isn't always cheap: Understanding failure modes in multi-agent debate. *Proceedings of the ICML 2025 Workshop on Multi-Agent Systems*. arXiv preprint arXiv:2509.05396. DOI: 10.48550/arXiv.2509.05396.

Ye, R. ym. 2025. X-MAS: A comprehensive testbed for evaluating heterogeneous LLM-driven multi-agent systems. arXiv preprint arXiv:2505.16997. DOI: 10.48550/arXiv.2505.16997.

Yeager.ai. 2023. AI Agent Kryptonite - Prompt Saturation and Context Bleeding. Medium. Julkaistu 16.10.2023. [Verkkojulkaisu]. Saatavilla: <https://medium.com/yeagerai/ai-agent-kryptonite-prompt-saturation-and-context-bleeding-4db7c4329e4e>. [Haettu 16.11.2025].

Yi, Z. ym. 2025. Benchmarking and defending against indirect prompt injection attacks on large language models. arXiv preprint arXiv:2312.14197. DOI: 10.48550/arXiv.2312.14197.

Zhang, Y. ym. 2024. Soar: The Future of AI-Driven Architectures through Directed Acyclic Graph Reasoning. arXiv preprint arXiv:2404.05678. DOI: 10.48550/arXiv.2404.05678.

Zilliz. 2024. Ensuring Secure and Permission-Aware RAG Deployments. Zilliz Blog. [Verkkojulkaisu]. Saatavilla: <https://zilliz.com/blog/ensure-secure-and-permission-aware-rag-deployments>. [Haettu 14.11.2025].