

Suurten kielimallien hallusinaatio

TURUN YLIOPISTO
Tietotekniikan laitos
TkK-tutkielma
Tietotekniikka
Tammikuu 2026
Eero Siivonen

TURUN YLIOPISTO
Tietotekniikan laitos

EERO SIIVONEN: Suurten kielimallien hallusinaatio

TkK-tutkielma, 24 s.
Tietotekniikka
Tammikuu 2026

Suurista kielimalleista on muodostunut kriittinen osa nykypäivän luonnollisen kielen sovelluksia ja välttämätön työkalu monilla eri aloilla. Nopeasta kehityksestä huolimatta kielimalleilla on edelleen merkittäviä haasteita, kuten jatkuvasti kasvavat kustannukset, epäeettiset käyttötarkoitukset ja vastausten hallusinointi.

Tässä tutkielmassa käsitellään suurten kielimallien hallusinaatiota: sen määritelmää, syitä, seurauksia sekä mahdollisia estämistapoja. Tutkielma on kirjallisuuskatsaus, jonka aineistossa on hyödynnetty ajankohtaisia lähteitä liittyen suuriin kielimalleihin ja niiden hallusinaatioon.

Tutkielmassa muun muassa ilmenee, että hallusinointiin vaikuttaa esimerkiksi mallin arviointi- ja koulutusmenetelmät, arkkitehtuuri ja koulutusdata. Suurten kielimallien hallusinaatiot eivät ole pelkästään teknisiä virheitä, vaan niillä saattaa olla hyvinkin vakavia seurauksia.

Hallusinaatioiden havaitsemisesta ja estämisestä on tullut viime vuosina merkittävä tutkimuksen kohde. Esiin on noussut erilaisia menetelmiä, kuten RAG, jossa tietoa etsitään ulkoisista tietolähteistä. Tutkielmassa myös pohditaan, onko kielimallien hallusinaatioiden estäminen edes kaikissa tilanteissa välttämätöntä.

Asiasanat: suuret kielimallit, hallusinaatio, koulutusdata, haasteet, riskit

Sisällys

1	Johdanto	1
2	Suuret kielimallit	4
2.1	Perusteet ja merkitys	4
2.2	Transformer-arkkitehtuuri	5
2.3	Haasteet	7
3	Hallusinaatio	10
3.1	Määritelmä	10
3.2	Syyt	11
3.3	Seuraukset	13
3.4	Estäminen	14
4	Pohdinta	18
5	Yhteenveto	20
	Lähdeluettelo	25

Kuvat

2.1	Merkittäviä suurten kielimallien käyttökohteita. Tiedot pohjautuvat lähteeseen [1].	5
2.2	Transformer enkooderi-dekooderi-arkkitehtuuri havainnollistettuna. Pohjautuu lähteeseen [5].	7
2.3	Esimerkkitalanne, jossa ChatGPT (tammikuussa 2024) ei vastaa oikein ajankohtaiseen kysymykseen, sillä sen tietoisuus ulottuu vain tammikuuhun 2022. Esimerkki on peräisin julkaisusta [9].	8

Taulukot

1.1 Tutkielman aineiston hakuprosessi	3
---	---

1 Johdanto

Suuret kielimallit (engl. large language model, LLM) ovat vieneet merkittävästi tekoälyn sekä luonnollisen kielen käsittelyn aloja eteenpäin. Kielimallien hyödyntäminen on tullut yhä useammalle tutuksi ChatGPT:n tai muiden keskustelubottien myötä. Suurten kielimallien suosio jatkaa edelleen kasvuaan, ja kielimalleja hyödynnetään yhä enemmän eri aloilla, kuten koulutuksessa, terveydenhuollossa, tutkimuksessa ja teollisuudessa. [1] Kielimallien hyvin nopeasti kasvavan roolin myötä, myös niiden haasteet on otettava vakavasti huomioon.

Kielimalleilla on todettu olevan paljon haasteita, kuten kasvavat kustannukset, epäeettiset käyttötarkoitukset ja vastausten hallusinointi [2]. Hallusinoivissa kielimallit tuottaa uskottavan oloista tietoa, mikä on kuitenkin epätarkkaa tai jopa täysin valheellista [3]. Kielimallien hallusinointiin voi vaikuttaa monet asiat, joista yksi on mallien koulutusdata. Koulutusdata saattaa sisältää vanhentunutta tietoa, vinoumia tai muita virheitä, mikä altistaa mallin epätarkkoille vastauksille. [4]

Hallusinoinnin seuraukset voivat olla vakavia etenkin kriittisillä aloilla, kuten terveydenhuollossa, joissa tiedon paikkansapitävyys on erityisen tärkeää. Pahimmassa tapauksessa hallusinointi voi olla syy esimerkiksi väärin diagnooseihin, oikeudellisiin ongelmiin tai autonomisten järjestelmien onnettomuuksiin [5]. Hallusinaatioiden estäminen on siis hyvin tärkeässä osassa tulevaisuuden kielimallien kehitystä. Erilaiset kielimallien vastauksia parantavat teknologiat, kuten retrieval-augmented generation (RAG), voivat tulevaisuudessa auttaa ponnistelussa hallusinaatioita vastaan

[4].

Tämän tutkielman tavoitteena on luoda kokonaiskuva suurten kielimallien hallusinaatiosta. Tarkemmin sanottuna työssä perehdytään hallusinaation syihin ja seurauksiin sekä mahdollisiin hallusinoinnin estämistapoihin. Tutkielmassa käsitellään myös suurten kielimallien toimintaa, arkkitehtuuria ja haasteita, mikä sujuvoittaa tutustumista mallien hallusinaatioon. Tutkielma on toteutettu kirjallisuuskatsauksena, ja sen tutkimuskysymykset ovat seuraavat:

- **TK1:** Mitä kielimallien hallusinaatiolla tarkoitetaan?
- **TK2:** Miksi suuret kielimallit hallusinoivat?
- **TK3:** Mitä seurauksia kielimallien hallusinoinnilla on?
- **TK4:** Millä tavoin kielimallien hallusinaatiota voidaan estää?

Tutkielman aineisto on haettu IEEE Xplore ja ACM -hakukannoista. Haku- lausekkeissa on käytetty aiheeseen liittyviä englanninkielisiä avainsanoja, kuten *LLM*, *hallucination*, *mitigation* ja *implications*. Aineistojen rajaaminen on toteutettu kar- simalla hakukantojen tulokset otsikon ja tiivistelmän perusteella. Vain sellaiset ar- tikkelit on sisällytetty aineistoon, jotka vastaavat tutkimuskysymyksiin tai liittyvät tutkielman aiheeseen muulla tavoin. Lopullinen aineisto on muodostunut tämän pro- sessin myötä, ja se ilmenee yksityiskohtaisemmin taulukosta 1.1. Lopulliseen aineis- toon kuuluu lisäksi tekoäly-yhtiö OpenAI:n vuonna 2025 julkaisema artikkeli "Why language models hallucinate", joka on sisällytetty sen merkittävyyden vuoksi.

Tutkielman rakenne koostuu johdannosta, kahdesta käsittelyluvusta, pohdinta- luvusta sekä yhteenvedosta. Luku 2 toimii tutkielman taustalukuna, ja siinä käsi- tellään muun muassa suurten kielimallien perusteita, arkkitehtuuria, käyttökohteita ja haasteita. Luvussa 3 sen sijaan tarkastellaan tutkimuskysymyksiä tutkielman ai- neistoon pohjautuen. Luvussa siis perehdytään suurten kielimallien hallusinaation

Taulukko 1.1: Tutkielman aineiston hakuprosessi

Hakukanta	Hakulauseke	Tulosten määrä	Lopullinen aineisto
IEEE Xplore	LLM* AND hallucination* AND (mitigation OR prevention)	48	5
IEEE Xplore	LLM* AND hallucination* AND (consequences OR implications)	40	1
ACM	LLM* AND hallucination* AND (mitigation OR prevention)	65	4

määritelmään, syihin, seurauksiin ja estämistapoihin. Pohdintaluvussa pohditaan hallusinaation merkitystä yleisellä tasolla tutkielman tuloksiin perustuen. Yhteenvedossa kootaan vielä yhteen tutkielmassa läpikäytyjä asioita ja vastataan tutkielman tutkimuskysymyksiin.

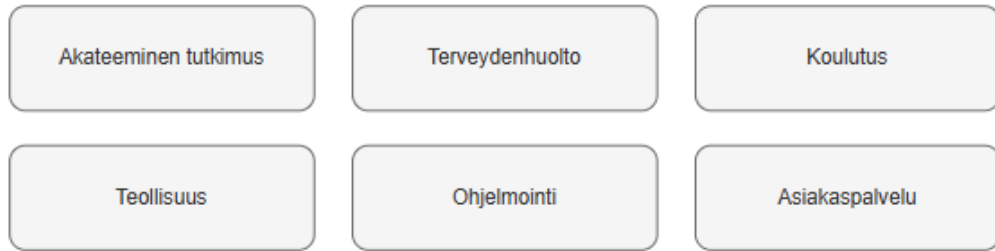
2 Suuret kielimallit

2.1 Perusteet ja merkitys

Kielimalli (engl. language model, LM) on sanoja ennustava koneoppimismalli. Kun kielimallille antaa sanoja syötteenä, se antaa todennäköisyysjakauman seuraavista mahdollisista sanoista. Vaihtoehtoisesti se voi asettaa todennäköisyydet jokaiselle mahdolliselle seuraavalle sanalle tai jopa kokonaisille lauseille. [6]

Suuri kielimalli (engl. large language model, LLM) on sen sijaan syväoppiva kielimalli, joka ymmärtää ja tuottaa luonnollista kieltä. Nykyään suuret kielimallit koulutetaan hyvin suurella määrällä tekstidataa, ja mallit sisältävät jopa satoja miljardeja parametreja, joiden avulla niitä voidaan hienosäätää. Suuret kielimallit pohjautuvat transformer-arkkitehtuuriin, mikä mahdollistaa tehokkaan rinnakkaisen datan käsittelyn ja monimutkaisten suhteiden oppimisen teksteistä. Sen ansiosta suuret kielimallit voivat tuottaa koherentteja ja kontekstiin liittyviä vastauksia. [3] [6] Suurten kielimallien suosion kasvu on tehnyt tekoälyn käytöstä helpompaa ja avannut uusia mahdollisuuksia. Se on mahdollistanut ihmisille tapoja käyttää teknologiaa aivan uudella tavalla: tekoäly voi esimerkiksi auttaa ohjelmistokehittäjiä sovellusten kehittämisessä tai tutkijoita uusien läpimurtojen etsinnässä. [1]

Suurten kielimallien toimintaa voidaan räätälöidä moniin eri tarkoituksiin erilaisten kehoitteiden (engl. prompt) avulla. Kehote on käyttäjän kielimallille antamaa tekstiä, minkä tarkoituksena on saada malli tekemään haluttuja hyödyllisiä asioita.



Kuva 2.1: Merkittäviä suurten kielimallien käyttökohteita. Tiedot pohjautuvat lähteeseen [1].

Prosessia, jossa tiettyyn tehtävään etsitään mahdollisimman tehokkaita kehoitteita, kutsutaan kehoitesuunnitteluksi (engl. prompt engineering). [6]

Suurista kielimalleista on muodostunut kriittinen osa nykypäivän luonnollisen kielen sovelluksia ja välttämätön työkalu monilla eri aloilla [7]. Akateemisessa tutkimuksessa kielimallit voivat esimerkiksi tehostaa tutkimusta tiivistämällä kirjallisuutta ja laatimalla käsikirjoituksia. Terveystenhoollossa mallit voivat taas auttaa lääketieteellisten tekstien lukemisessa, diagnoosien parantamisessa sekä potilasviestinnässä. Koulutuksessa malleja saatetaan taas käyttää mukautuvien oppimistyökalujen kehittämiseen tai henkilökohtaisen opintosuunnitelman laatimiseen. Suuria kielimalleja hyödynnetään yhä enemmän myös asiakaspalvelussa, ohjelmoinnissa, teollisuudessa ja luovilla aloilla. [1] Kuvassa 2.1 on lueteltu suurten kielimallien eri käyttökohteita.

2.2 Transformer-arkkitehtuuri

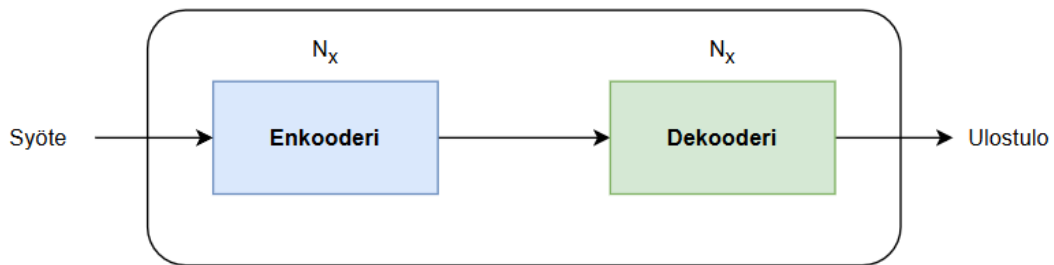
Vaswani ym. [8] esitteli transformer-arkkitehtuurin ensimmäistä kertaa vuonna 2017. Tämä oli käännekohta suurille kielimalleille, sillä uusi arkkitehtuuri onnistui korvaamaan hitaat takaisinkytketyt neuroverkot (engl. recurrent neural network, RNN) [1]. Transformer-verkko pystyy hyödyntämään grafiikkasuorittimen rinnakkaisprosessointia, mikä mahdollistaa useiden syötesekvenssien käsittelyn samanaikaisesti. Tämä on yksi transformer-arkkitehtuurin eduista muihin neuroverkkoihin verrattu-

na. [5]

Transformer-arkkitehtuuri voi pohjautua enkooderi-, dekodeeri- tai enkooderi-dekodeeri-arkkitehtuuriin. Kyseessä on eri tehtäviin soveltuvia arkkitehtuureja, joita voidaan hyödyntää kielimalleissa transformer neuroverkkojen avulla. [6] Transformer-verkon toiminnan ytimessä on itsehuomiomekanismi (engl. self-attention mechanism). Mekanismi perustuu siihen, että syötetekstin sanojen esitysten samankaltaisuutta verrataan muiden syötetekstin sanojen esityksiin. Vertailun perusteella lasketaan huomiopisteet (engl. attention score). Huomiopisteiden laskennassa käytetään apuna pistetuloa, sillä se mahdollistaa kahden vektorin samanakaltaisuuden laskemisen. Itsehuomiomekanismin avulla voidaan siis kuvata syötetekstin eri sanojen merkittävyyttä käyttäen huomiopisteitä ja ymmärtämään paremmin sanojen välisiä suhteita, mitä voidaan hyödyntää transformer-verkon toiminnassa. [5] [8]

Vuonna 2017 julkaistu transformer-arkkitehtuuri pohjautuu enkooderi-dekodeeri-arkkitehtuuriin. Sen toiminnan ensimmäinen vaihe on syötetekstin tokenisointi ja siitä muodostuneiden tokeneiden muuntaminen vektoreiksi. Tokenien sijainnit vektoreissa tallennetaan syötesekvenssiin sijainnin enkoodauksen (engl. positional encoding) avulla. Enkooderi vastaanottaa kyseisen syötesekvenssin, käsittelee sen ja luo tämän perusteella sekvenssin esityksiä. Dekodeeri vastaanottaa nämä esitykset enkooderista ja tuottaa tulossekvenssin. Kyseinen arkkitehtuuri sisältää siis kaksi tärkeää osaa: enkooderin ja dekodeerin. Enkooderi- ja dekodeerikerroksia ei käytetä vain kerran, vaan kerroksia pinotaan useita päällekkäin, jolloin muodostuu syvempi neuroverkko. [5] [8] Kuva 2.2 havainnollistaa transformer enkooderi-dekodeeri-arkkitehtuurin rakennetta.

Generatiivisten kielimallien, kuten ChatGPT:een, transformer-arkkitehtuuri pohjautuu yleensä dekodeeri-arkkitehtuuriin, jossa ei ole lainkaan enkooderi-kerroksia. Enkooderi-dekodeeri-arkkitehtuuri soveltuu paremmin tehtäviin, jossa ulostulon tokenit eroavat paljon syötetokeneista, kun taas dekodeeri-arkkitehtuurissa näiden to-



Kuva 2.2: Transformer enkooderi-dekooderi-arkkitehtuuri havainnollistettuna. Pohjautuu lähteseen [5].

kenien välinen suhde on paljon tiukempi. Enkooderi-dekooderi-arkkitehtuurille soveltuva tehtävä olisi esimerkiksi puheentunnistus, jossa syötetokenit edustavat puhetta ja ulostulon tokenit tekstiä. [6]

2.3 Haasteet

Suuret kielimallit kohtaavat paljon merkittäviä haasteita niiden nopeasta kehityksestä huolimatta [1] [2] [3]. Kenties yksi tunnetuimmista haasteista on suurten kielimallien hallusinaatio. Kun kielimalli hallusinoi, se tuottaa uskottavan oloista tietoa, mikä on kuitenkin epätarkkaa tai jopa täysin valheellista. [3] Vinoudet kielimallien koulutusdatassa voivat taas johtaa syrjiviin ja epäeettisiin tuloksiin, jotka esimerkiksi edistävät stereotyyppioita [7]. Suurten kielimallien hallusinaatioon perehdytään tarkemmin seuraavassa luvussa, joten tarkastellaan sitä ennen muutamia muita haasteita.

Suuriin kielimalleihin kohdistuvan datan, laskentatehon ja investointien tarve on kasvanut räjähdysmäisesti. Esimerkiksi vuonna 2019 OpenAI:n GPT-2-mallin 1,5 miljardin parametrin koulutus maksoi 50 000 dollaria, kun taas vuonna 2024 GPT-4-mallin 1,8 biljoonan parametrin koulutus maksoi jo 78 miljoonaa dollaria. On siis selvää, että kestävä tulevaisuuden kehitys vaatii läpimurtoja erityisesti datan,

When did Ram Chandra Poudel become the president of Nepal?



ChatGPT

As of my last knowledge update in January 2022, Ram Chandra Poudel has not served as the President of Nepal. Please note that my information might be outdated, and I recommend checking the latest sources or news for the most recent updates on Nepal's political leadership. As of my last update, the President of Nepal was Bidya Devi Bhandari.

Kuva 2.3: Esimerkkitilanne, jossa ChatGPT (tammikuussa 2024) ei vastaa oikein ajankohtaiseen kysymykseen, sillä sen tietoisuus ulottuu vain tammikuuhun 2022. Esimerkki on peräisin julkaisusta [9].

algoritmien, prosessoinnin sekä energian suhteen. [2]

Mallien kouluttamiseen liittyvien kustannusten, ajan, laskennallisen tehon ja hiidioksidipäästöjen vuoksi mallien toistuva uudelleen koulutus ja hienosäätö on mahdotonta jopa yrityksissä, joilla on käytössään runsaasti resursseja. Yleensä malleja ei siis ole koulutettu täysin ajantasaisella datalla, mikä altistaa puuttelliselle tai väärälle tiedolle kysyttäessä ajankohtaisista aiheista. [9] Kuvassa 2.3 nähdään esimerkkitilanne, jossa mallin tiedolle on asetettu rajapäivämäärä, eikä se siksi kykene vastaamaan oikein sille syötettyyn kysymykseen.

Suuria kielimalleja voidaan väärinkäyttää epäeettisiin tarkoituksiin monin eri tavoin [7] [4]. Yksi esimerkki tästä on epäsuoran kehotteen injektointi (engl. indirect prompt injection), jossa kielimallille annetaan mahdollisesti haitallisia käskyjä ilman, että käyttäjä on tehnyt niin. Tällöin kielimalli siis seuraa dataan upotettuja ohjeita käyttäjän antamien käskyjen sijaan. Esimerkkitilanne voisi olla se, että tekoälylle annettaisiin tehtäväksi hallita sähköpostia automaattisesti ja se vastaanottaisi injektio sisällävän viestin, jossa sitä käskettäisiin lähettämään arkaluontoista tietoa käyttäjältä. [4]

Toinen esimerkki väärinkäytöstä on jailbreaking, mikä tarkoittaa sellaisten kehotteiden laatimista, joilla kielimalli saadaan rikkomaan sille asetettuja rajoituksia.

Jailbreak-tekniikoita on monia ja myös niitä kehitetään jatkuvasti. Pohjimmiltaan tekniikat kuitenkin käyttävät hyväkseen suurten kielimallien joustavuutta ja kontekstuaalista ymmärrystä. Jailbreaking mahdollistaa haitallisen sisällön luomisen ja levittämisen. Mallien avulla saatetaan generoida esimerkiksi disinformaatiota, vihapuhetta tai luvattomia kuvia, mikä altistaa tekoälypalveluiden tarjoajat mainehaitalle ja mahdollisesti oikeudellisille toimille. [4]

3 Hallusinaatio

3.1 Määritelmä

Suuret kielimallit tuottavat välillä vastauksia, jotka ovat täysin väärää tai harhaanjohtavia, vaikka vastaukset vaikuttavat samaan aikaan hyvinkin itsevarmoilta ja uskottavilta. Ilmiötä kutsutaan hallusinaatioksi. Huolimatta merkittävistä edistysaskeleista kielimallien hallusinaatio on edelleen yksi suurimmista ongelmista myös kaikkein uusimmissa ja edistyneimmissä malleissa. [5] [10] Kielimallien hallusinaatiota on tutkittu ja dokumentoitu laajasti, ja tutkijat ovat tunnistaneet erilaisia hallusinaation muotoja ja syitä. Hallusinaation ilmenemismuodoiksi on todettu esimerkiksi tosiasioiden epätarkkuudet, tekaistut tiedot, ristiriidat ja tietojen puutteet. Eri muotojen ja niiden ilmenemissyiden ymmärtäminen on tärkeää, jotta hallusinaation vähentämiseksi voidaan kehittää tehokkaita strategioita sekä entistä vastuullisempia kielimalleja. [4]

Hallusinaatiot voidaan jakaa sisäisiin (engl. intrinsic hallucinations) ja ulkoisiin (engl. extrinsic hallucinations) hallusinaatioihin [5] [7] [10]. Sisäinen hallusinaatio tarkoittaa hallusinaatioita, joka on ristiriidassa käyttäjän antaman kehotteen kanssa. Jos kielimalli vastaisi väärin esimerkiksi kysymykseen: "Kuinka monta kirjainta on sanassa tekoäly?", olisi kyse sisäisestä hallusinaatiosta. Ulkoisella hallusinaatiolla taas tarkoitetaan hallusinaatiota, joka on ristiriidassa koulutusdatan tai ulkoisen todellisuuden kanssa. [10]

3.2 Syyt

Suurten kielimallien hallusinoinnille on esitetty monia eri syitä ja selityksiä. Kailain ym. [10] mukaan kielimallit hallusinoivat pääosin siksi, koska koulutus- ja arviointimenetelmät palkitsevat arvaamista epävarmuuden tunnustamisen sijaan. Jos mallilta kysyisi esimerkiksi tietoja henkilöstä, jota ei ole olemassa, saattaisi malli silti tehdä arvauksia sen sijaan, että se myöntäisi, ettei se tiedä vastausta [10]. Russinovichin ym. [4] mukaan kielimallien hallusinoinnin ensisijainen syy on taas mallien arkkitehtuurissa ja koulutusprosessissa. Suuren kielimallin arkkitehtuuri sisältää tietyn rakenteen, joka koostuu muun muassa parametreista ja eri kerrosten lukumäärästä. Kielimalli voi tulla alttiimmaksi hallusinoinnille, kun sen arkkitehtuurin kompleksisuutta kasvatetaan, esimerkiksi kerroksia tai parametreja lisäämällä. [5]

Koulutusdatan vinoumia pidetään yhtenä hallusinaation aiheuttajana. Suuret kielimallit koulutetaan yleensä käyttäen suuria määriä dataa, joten on mahdollista, että jotkin ryhmät tai asiat ovat siinä epäedustettuina. Kun malli koulutetaan vinoutuneella datalla, se oppii siitä harhaanjohtavia piirteitä ja täten tuottaa myös epätarkkoja ennusteita. [5] Suuren datamäärän vuoksi koulutusdatassa voi vinoumien lisäksi olla runsaasti virheitä. Suurten kielimallien kouluttaminen ei yleensä sisällä mitään luotettavuuden painokertoimia, mikä mahdollistaa myös virheellisen datan päätyminen mallin kouluttamiseen. Ongelma korostuu etenkin silloin, jos koulutusdata sisältää systemaattisesti samoja virheitä sisältävää dataa, kuten salaliitto-teorioita, minkä tuloksena malli saattaa vahvistaa tiedot totena. [4] Ilmiö on nimeltään roskaa sisään, roskaa ulos (engl. garbage in, garbage out, GIGO), joka on tärkeä ottaa huomioon, sillä kielimallit ovat osoittaneet toistavan koulutusdatan virheitä [10].

Koulutusdatassa vanha tieto on yleensä huomattavasti yleisempää kuin lähivuosien uusi tieto. Samaan aikaan suuren kielimallin tietoisuus ulottuu vain sen koulu-

tusdatan rajapäivämäärään asti. Tämä voi johtaa siihen, että malli tuottaa vanhentunutta tietoa esitettynä tämän hetken tiedoksi. Kuten edellisessä luvussa todettiin, esimerkiksi malli, joka on koulutettu vuoteen 2023 ulottuvalla tiedolla, ei todennäköisesti ole tietoinen merkittävistä tapahtumista tai muutoksista, jotka ovat tapahtuneet vuonna 2024. [4]

Liian suppea koulutusdata saattaa siis aiheuttaa hallusinaatioita. Jos mallille esitetään esimerkiksi kysymys, jolle ei löydy vastausta sen koulutusdatasta, malli voi joutua arvaamaan ja tarjoamaan mahdollisesti virheellisen vastauksen. Kielimallit on optimoitu suoriutumaan hyvin erilaisissa testeissä, ja arvaaminen epävarmoissa tilanteissa parantaa testisuoritusta. Näin kielimallin suoristuskyky saadaan näyttämään paremmalta, mitä se todellisuudessa on. Tällaisen pisteytystavan vuoksi mallien hallusinaatioita ei välttämättä saada korjattua. [10]

Sanun ym. [7] mukaan ylisovittaminen on yleinen hallusinaatioiden lähde suurissa kielimalleissa. Hyvän mallin pitäisi kyetä yleistämään koulutusdatasta opittua tuntemattomaan testijoukkoon. Ylisovittava malli on tässä kuitenkin huono, sillä malli on tällöin oppinut koulutusdatan liian hyvin eikä siksi pysty yleistämään oppimaansa uuteen dataan. [6] Huomionarvoista on myös se, että kielimalleilla on ihmisiin verrattuna rajoitetumpi kontekstiymmärrys, mikä voi jossain tilanteissa johtaa siihen, ettei malli välttämättä kykene tuottamaan tarpeeksi relevanttia tai kontekstiin sopivaa vastausta [5]. Kielimallien vastaamiskykyä rajoittaa myös tietokoneiden laskentakyky, minkä vuoksi mallit tekevät virheitä laskennallisesti vaikeissa ongelmissa [10].

Hallusinaatioiden kannalta myös sillä on merkitystä, kuinka helposti tulkittavissa käyttäjän kielimallille syöttämät kehotteet ovat. Epäselvillä kehotteilla voi olla monia mahdollisia merkityksiä tai tulkintoja. Kun kielimallit saavat tällaisia kehotteita, ne saattavat tulkita niitä vastoin käyttäjän odotuksia, mikä saattaa johtaa hallusinointiin. Mallit ovat myös alttiita vihamielisille hyökkäyksille, joissa kehotet-

ta manipuloidaan siten, että malli tuottaa harhaanjohtavia tai virheellisiä vastauksia. Tällaiset hyökkäykset hyödyntävät mallin käyttäytymisen haavoittuvuuksia ja tuovat esiin alueita, joilla mallin ymmärrys- ja päättelykyky ovat puutteellisia. [5]

3.3 Seuraukset

Suurten kielimallien hallusinaatiot eivät ole pelkästään teknisiä virheitä, vaan niillä saattaa olla hyvinkin vakavia seurauksia [7]. Hallusinaatioiden riskit korostuvat etenkin kriittisissä käyttökohteissa, kuten terveydenhuollossa, joissa tarkkuus ja täsmällisyys ovat usein ehdoton vaatimus [3] [5] [7]. On lukuisia kielimallien käyttökohteita, joilla hallusinoinnin riskit ovat erityisen vakavia, kuten lääketieteellisissä diagnooseissa, oikeudellisissa analyysissä ja autonomisissa järjestelmissä. Näillä kriittisillä ja monimutkaisilla aloilla hallusinaatiot voivat johtaa esimerkiksi virheellisiin diagnooseihin, oikeudellisiin ongelmiin tai vakaviin onnettomuuksiin. [5] Esimerkiksi lääketieteellisissä sovelluksissa kielimalli voisi hallusinoita hoitotoimenpiteitä, joita ei edes ole olemassa, mikä saattaisi pahimmassa tapauksessa vaarantaa potilaan hengen [3] [7].

Suurten kielimallien hallusinoinnista voi seurata monia eettisiä ongelmia. Kun kielimallit tuottavat harhaanjohtavaa tai jopa täysin väärää tietoa, ihmiset saattavat pitää tietoja totena ja jakaa sitä eteenpäin. Myös kielimallin haavoittuvuuksia voidaan käyttää hyväksi monella eri tavalla: mallien avulla voidaan esimerkiksi yrittää vaikuttaa yleiseen mielipiteeseen tai levittää valheita, mikä voi vaikuttaa merkittävästi julkiseen keskusteluun. [5] [7] Kielimallien hallusinoinnilla voi olla myös erilaisia oikeudellisia seurauksia. On täysin mahdollista, että mallin tuottama väärä tieto johtaa oikeudellisiin toimiin. Tähän liittyen moni asia tulisi ottaa huomioon, kuten immateriaalioikeudet sekä kysymys siitä, kuka ottaa vastuun väärän tiedon aiheuttamasta vahingosta. [5]

Mallien hallusinoinnilla saattaa olla myös erilaisia sosiaalisia vaikutuksia. Kieli-

mallin hallusinoituneet vastaukset voivat vaikuttaa ihmisten uskomuksiin ja sitä kautta myös päätöksentekoon. Ihmiset saattavat uskoa kielimallin vastauksia ilman epäilyksiä, ja näin ollen tehdä päätöksiä kyseisten väitteiden pohjalta. Vastauksen taustalla olevat väitteet ja uskomukset saattavat olla kuitenkin täysin vääriä. Toinen mahdollinen sosiaalinen vaikutus on ihmisten vähenevä luottamus mallien tuottamiin vastauksiin. Luottamuksen väheneminen voi olla ongelmallista, koska se vaikuttaa merkittävästi yleiseen mielipiteeseen kielimalleista. [5] [7] Epäilykset kielimallien luotettavuudesta saattaa olla joillekin jopa este kielimallien käyttämiselle. Koska hallusinaatiot vaikuttavat vahvasti kielimallien luotettavuuteen, ne saattavat myös heikentää merkittävästi suuria kielimalleja hyödyntävien yritysten ja organisaatioiden uskottavuutta ja mainetta. Hallusinaatiot eivät siis vaikuta pelkän teknologian luotettavuuteen, vaan herättävät myös epäilyksiä niitä hyödyntävien yritysten pätevydestä ja luotettavuudesta. [5]

Ihmisten on helppo luottaa liikaa kielimallien vastauksiin, sillä kielimallit eivät yleensä kykene esittämään ja perustelemaan omaa tietämystään, vastausten varmuutta, osaamistaan tai rajoituksiaan. Ongelmaa pahentaa entisestään se, että ihmiset eivät yhä useammin kykene erottamaan mallien tuottamia vastauksia ihmisten tuottamista vastauksista. Huomionarvoista on myös se, että kieli itsessään voi johdattaa malleja ja niiden käyttäjiä harhaan, esimerkiksi silloin kuin ihmisillä ja pakoilla on samoja nimiä. [2]

3.4 Estäminen

Kielimallien hallusinaatioiden vähentäminen suurissa kielimalleissa on erityisen tärkeää, koska se vaikuttaa merkittävästi luodun sisällön luotettavuuteen ja uskottavuuteen [3]. Hallusinaatioiden havaitsemisesta ja estämisestä onkin tullut merkittävä tutkimuksen kohde viime vuosina [9] [11]. Kun hallusinaatiot saadaan jollain tavalla havaittua, myös niiden estäminen on mahdollista. Wein ym. [11] mukaan

hallusinaatioiden havaitsemis- ja estämistavat voidaan jakaa kahteen ryhmään: niihin, jotka perustuvat ulkoisen tiedon varmentamiseen sekä niihin, jotka kuvaavat epävarmuustekijöitä. Ulkoisen tiedon varmentamiseen perustuvat tavat tarkoittavat siis sitä, että kielimallin vastausten paikkansapitävyys varmistetaan vielä joistain muista lähteistä. Epävarmuustekijöiden kuvaaminen sen sijaan tarkoittaa sitä, että vastausten luotettavuutta mitataan erilaisten mittareiden avulla. Kun tiedetään kuinka epävarmoja vastaukset ovat, voidaan myös havaita hallusinoitua vastaukset. [11]

Suuret kielimallit voivat siis käyttää ulkoisia tietolähteitä varmasti oikean vastauksen muodostamiseksi ja hallusinaatioiden vähentämiseksi. Tietoa voidaan etsiä monesta eri tietolähteestä, kuten webkorpuksista tai Wikipediasta. [11] RAG, eli retrieval-augmented generation, on menetelmä suurten kielimallien ulkoisen tiedon saantiin. Menetelmä on osoittautunut tehokkaaksi tavaksi vähentää kielimallien hallusinaatioita. Menetelmä toimii käyttämällä ulkoisia tietolähteitä, jotka tarjoavat mallille ajankohtaista ja relevanttia tietoa, mikä puolestaan parantaa mallin kykyä tuottaa tarkempia vastauksia. [3] Kehittyneemmät RAG-arkkitehtuurit pysyvät olennaisen tiedon hakemisen lisäksi ymmärtämään myös monimutkaisia suhteita eri alojen välillä. Esimerkki tällaisesta arkkitehtuurista on RAFT, eli retrieval-augmented fine-tuning, joka on osoittanut lupaavia tuloksia oikeus- ja lääketieteen aloilla sisällyttämällä alakohtaisia tietograafeja tiedonhakuprosessiin. [4]

Vaikka RAG on tehokas apu hallusinaatioiden vähentämiseksi yksinkertaisissa tehtävissä, se epäonnistuu usein monimutkaisissa tehtävissä. Tällaisia tehtäviä on esimerkiksi erityisen pitkien vastausten tuottaminen sekä vastaaminen sellaisiin kysymyksiin, jotka vaativat tietojen yhdistämistä monesta eri lähteestä ja joiden selittäminen vaatii monta eri vaihetta (engl. multi-hop question answering). Epäonnistuminen monimutkaisissa tehtävissä johtuu pääosin siitä, että kielimalli tarvitsee paljon jatkuvasti muuttuvaa tietoa, eikä perinteinen RAG kykene vahvistamaan

kaikkea tietoa ulkoisista tietolähteistä. Sen sijaan monivaiheinen RAG (engl. multi-round RAG) voi suorittaa useita hakuja kielimallien vastausten generoinnin aikana, mikä voi olla ratkaisu edellä mainittuihin monimutkaisiin tehtäviin. Monivaiheisen RAG:n tehokkuuteen vaikuttaa haun toteuttamisen ajoitus vastauksen generoimisen yhteydessä, ja hyvien ajoitusten saavuttamiseksi onkin ehdotettu monia eri menetelmiä. [12]

Ulkoisen tiedon varmentamisen lisäksi hallusinaatioiden havaitsemiseksi voidaan hyödyntää vastausten luotettavuuden mittaamista. Yksi tapa olisi hyödyntää toista kielimallia löytämään vastauksesta loogisia ristiriitoja tai arvioimaan sen uskottavuutta. [11] Toinen tapa voisi olla se, ettei luota vain yhteen kielimallin vastaukseen, vaan ottaa huomioon useiden eri kielimallien tulokset tai saman mallin useat eri tulokset. Usean eri vastauksen hyödyntäminen voi auttaa tunnistamaan ja suodattamaan hallusinaatioita. Arvioidessa vastausten luotettavuutta useiden eri mittareiden käyttö onkin usein tehokkaampi keino hallusinaatioiden havaitsemiseksi kuin yksittäisten mittareiden käyttö. [4]

Suurten kielimallien vastausten luotettavuutta on mahdollista mitata monin eri tavoin. Luotettavuuden arviointiin on erityisen tärkeää valita sellaisia mittareita, jotka kuvaavat kattavasti vastauksen laadun eri osa-alueita. Chen ym. [13] käyttää omassa tutkimuksessaan neljää eri mittaria hallusinaatioiden havainnointiin: LLM-arviointimittaria, ihmismittaria, konemittaria sekä yhdistelmämittaria. LLM-arviointimittari perustuu kielimallin itsearviointiin, jossa kielimalli arvioi omaa vastaustaan ilman ihmisten väliintuloa. Ihmismittari sen sijaan sisältää ihmisten antamat pisteet vastauksen hyvyydelle, ja pisteiden määrä riippuu siitä, kuinka lähellä vastaus on tutkittua ja varmistettua totuutta. Konemittari taas nojautuu useisiin objektiivisiin mittareihin arvioidakseen tuotettuja vastauksia. Tarkemmin sanottuna se arvioi mallin vastauksen ja oikean totuuden perusteella esimerkiksi vastausten semanttista linjausta, relevanssia, monimuotoisuutta sekä laatua käyttäen useaa

eri pisteytystä. Yhdistelmämittari sisältää loppupistemäärän, johon jokainen edellä mainituista mittareista vaikuttaa, ja jossa tietyille mittareille voidaan antaa erityinen painotus. Yhdistelmämittari antaa siis kattavan arvion mallin suorituskyvystä yhdistämällä useita eri osa-alueita. [13]

Esiin on myös noussut toisenlainen näkökulma, jonka mukaan hallusinaatiot ovat tietyissä tilanteissa harmittomia eikä niitä tarvitse estää. Sen mukaan hallusinoiva malli ei haittaa tilanteissa, joissa ei ole olemassa niin sanottua oikeaa vastausta. Tällaisia malleja voitaisiin käyttää moniin eri tarkoituksiin: ne voisivat esimerkiksi tarjota tutkimusvihjeitä, tukea luovassa kirjoittamisessa sekä vähentää kirjoitusvaikeuksia. [9]

4 Pohdinta

Suurten kielimallien hallusinaatiot ovat vain yksi osa niiden haasteita. Esimerkiksi suuriin kielimalleihin kohdistuvan datan, laskentatehon ja kustannusten tarve on kasvanut runsaasti lähivuosien aikana. Kaikkein suorituskykyisimpiä malleja kykenee kouluttamaan parhaiten ne yritykset, joilla on käytössään paljon resursseja. Jotta suurten kielimallien kouluttaminen olisi vastuullisempaa ja mahdollista suuremmalle joukolle yrityksiä, vaatisi se läpimurtoja ja lisää tutkimusta alalta. Laajempi kielimallien kehitys voisi edistää alan innovaatioita ja auttaa myös kielimallien hallusinaatioiden vähentämisessä.

Kielimallien hallusinaatiot tarkoittavat lyhyesti sanottuna mallin virheellisiä vastauksia, jotka saattavat vaikuttaa uskottavilta. Hallusinointi on edelleen yksi merkittävimmistä kielimallien haasteista myös kaikkein uusimmissa ja edistyneimmissä malleissa. Samaan aikaan yhä useammat käyttävät suuria kielimalleja niin töissä, koulussa ja vapaa-ajalla. Suurten kielimallien hyödyttäminen yleistyy myös kriittisillä aloilla, kuten terveydenhuollossa, joissa ei välttämättä ole varaa virheisiin. Näin ollen suurten kielimallien haasteisiin, ja etenkin hallusinointiin, tulisi kiinnittää erityistä huomiota.

Tutkielmassa ilmeni, että yksi syy kielimallien hallusinoinnille on se, että koulutus- ja arviointimenetelmät palkitsevat arvaamista epävarmuuden tunnustamisen sijaan. Mallien arviointimenetelmien tulisi ottaa vastausten luotettavuus paremmin huomioon, sillä se on keskeinen osa laadukasta kielimallia. Kielimallin tulisi kyetä vas-

taamaan käyttäjälle tarvittaessa olevansa jostain kysymyksestä epävarma, jotta voitaisiin paremmin välttyä esimerkiksi väärän tiedon leviämiseltä.

Myös koulutusdatan laatu voi olla yhteydessä kielimallien hallusinointiin. Koulutusdata voi sisältää esimerkiksi paljon toistuvia virheitä, vinoumia tai vanhentunutta tietoa. Ajankohtainen tieto saattaa puuttua kielimallin koulutusdatasta kokonaan. Onkin syytä kysyä, onko edes mahdollista kouluttaa lähes täydellistä kielimallia, sillä uutta tietoa tulee jatkuvasti lisää. Tähän saattaa olla avuksi menetelmät, joissa ajankohtaista tietoa haetaan ulkoisista tietolähteistä. Toisaalta tällöin vastauksen generointi voi myös hidastua, sillä tietoa joutuu etsimään myös muista lähteistä.

Tutkielmassa myös ilmeni, että kielimallin vastausten luotettavuuden mittaamista voitaisiin hyödyntää hallusinaatioiden havaitsemisessa. Vastauksen luotettavuutta voitaisiin esimerkiksi arvioida toisen kielimallin avulla, joka voisi etsiä vastauksesta loogisia ristiriitoja. Ei välttämättä ole olemassa yksittäistä parasta tapaa kielimallien hallusinaatioiden vähentämiseksi, vaan tehokkain ratkaisu voisi olla monen eri menetelmän hyödyntäminen. Menetelmien haittapuolet on kuitenkin myös aina otettava huomioon. Jos hallusinaatioiden havaitsemiseksi esimerkiksi hyödynnetäisiin useaa eri kielimallia, voisi tällöin myös kustannukset ja laskentatehon tarve kasvaa.

On myös syytä pohtia, onko kielimallien hallusinaatioiden estäminen kaikissa tilanteissa välttämätöntä. Jos kielimalleja hyödynnetään tehtävissä, joissa ei vaadita tiettyä niin sanottua oikeaa vastausta, ei mallin hallusinointi välttämättä ole ongelma. Tällaisia tehtäviä voisi olla esimerkiksi luovat tehtävät, joissa tulee keksiä täysin uusia asioita. Kaikissa tilanteissa hallusinoinnin estäminen ei siis välttämättä ole tarpeellista, ja hallusinoinnin estämistapojen sijaan, saattaa olla parempi idea keskittyä muihin asioihin, kuten mallin nopeuteen generoida vastauksia.

5 Yhteenveto

Suurten kielimallien suosion kasvu on tehnyt tekoälyn käytöstä helpompaa ja avannut uusia mahdollisuuksia. Viime vuosina mallien hyödyntäminen on tullut yhä useammalle tutuksi keskustelubottien myötä, ja suuria kielimalleja on alettu hyödyntämään yhä enemmän eri aloilla, kuten koulutuksessa, terveydenhuollossa, tutkimuksessa ja teollisuudessa. Tämän vuoksi onkin yhä tärkeämpää tiedostaa, että suurten kielimallien nopeasta kehityksestä huolimatta, malleilla on edelleen merkittäviä haasteita, kuten vastausten hallusinointi. Suurten kielimallien haasteisiin tulisi kiinnittää yhä enemmän huomiota kielimallien nopeasti kasvavan aseman vuoksi. Tällöin olisi myös mahdollista kehittää entistä vastuullisempia kielimalleja sekä välttyä paremmin seurauksilta, joita muun muassa kielimallien hallusinaatiot saattavat aiheuttaa. Tämän tutkielman tavoitteena oli luoda kokonaiskuva kielimallien hallusinaatiosta tarkastelemalla hallusinaation määritelmää, mahdollisia syitä, seurauksia ja estämistapoja. Seuraavaksi vastataan esitettyihin tutkimuskysymyksiin tutkielmassa saatujen tulosten pohjalta:

TK1: Mitä kielimallien hallusinaatiolla tarkoitetaan? Kielimallien hallusinaatiolla tarkoitetaan kielimallien generoimia vastauksia, jotka ovat täysin vääriä tai harhaanjohtavia, mutta voivat samaan aikaan vaikuttaa hyvinkin itsevarmoilta ja uskottavilta. Kielimallien hallusinaatiot voivat ilmetä esimerkiksi tosiasioiden epätarkkuuksina, tekaistuin tietoina, ristiriitoina tai tietojen puutteina. Hallusinaatiot voidaan jakaa sisäisiin ja ulkoihin hallusinaatioihin. Sisäinen hallusinaatio tarkoit-

taa tilannetta, jossa vastaus on ristiriidassa käyttäjän antaman kehotteen kanssa, kun taas ulkoinen hallusinaatio tilannetta, jossa vastaus on ristiriidassa sen koulutusdatan tai ulkoisen todellisuuden kanssa.

TK2: Miksi suuret kielimallit hallusinoivat? Suurten kielimallien hallusinointiin saattaa vaikuttaa moni eri asia, kuten mallin koulutus- ja arviointimenetelmät, arkkitehtuuri ja mallin koulutusdata. Koulutus- ja arviointimenetelmät palkitsevat usein arvaamista epävarmuuden tunnustamisen sijaan, mikä voi johtaa virheellisiin vastauksiin. Mallin arkkitehtuuri voi sen sijaan olla syy hallusinointiin siksi, koska kielimallin arkkitehtuurin kompleksisuutta kasvattaessa, esimerkiksi parametreja lisäämällä, saattaa malli myös tulla alttiimmaksi hallusinoinnille. Koulutusdatassa voi taas olla erilaisia vinoumia, virheitä tai vanhentunutta tietoa. Kun malli koulutetaan virheellisellä datalla, se oppii siitä harhaanjohtavia piirteitä ja täten tuottaa myös epätarkkoja ennusteita. Suuret kielimallit koulutetaan yleensä hyvin suurella määrällä dataa, joten virheitä saattaa olla runsaasti. Ongelma korostuu etenkin silloin, kun koulutusdata sisältää systemaattisesti samoja virheitä sisältävää dataa, kuten salaliittoteorioita.

Syynä hallusinointiin voi olla myös ylisovittaminen, kehotteiden tulkittavuus sekä malliin kohdistuvat vihamieliset hyökkäykset. Ylisovittaminen saattaa olla syy hallusinointiin, koska hyvän mallin pitäisi kyetä yleistämään koulutusdatasta opittua tuntemattomaan testijoukkoon. Ylisovittava malli ei kykene tähän hyvin, sillä malli on tällöin oppinut koulutusdatan liian hyvin eikä siksi pysty yleistämään oppimaansa uuteen dataan. Syynä hallusinointiin voi olla myös kehotteiden tulkittavuus. Epäselvillä kehotteilla saattaa olla monia erilaisia merkityksiä tai tulkintoja, ja kun kielimallit saavat tällaisia kehotteita, ne saattavat tulkita niitä vastoin käyttäjän odotuksia. Myös oikeanlaisella kehotesuunnittelulla kielimalli voidaan saada hallusinoimaan. Tällöin kehotetta manipuloidaan siten, että malli tuottaa harhaanjohtavia tai virheellisiä vastauksia hyödyntäen mallin käyttäytymisen haavoittuvuuksia.

Rajoitteena oikeille vastauksille voi olla mallin ihmisiin verrattuna usein rajoitetumpi kontekstiymmärrys sekä tietokoneiden laskentakyky. Rajoitetumpi kontekstiymmärrys saattaa johtaa siihen, ettei malli välttämättä kykene tuottamaan tarpeeksi relevanttia tai kontekstiin sopivaa vastausta, mikä voi ilmetä hallusinoitina. Laskennallisesti vaikeissa ongelmissa rajoittavana tekijänä saattaa sen sijaan olla tietokoneiden laskentakyky, minkä myötä malli tuottaa virheellisiä vastauksia.

TK3: Mitä seurauksia kielimallien hallusinoinnilla on? Kielimallien hallusinoinnista voi seurata eettisiä ongelmia, joita ovat esimerkiksi väärän tiedon leviäminen ja vaikutus yleiseen mielipiteeseen. Kun kielimallit tuottavat harhaanjohtavaa tai jopa täysin väärää tietoa, ihmiset saattavat pitää tietoja totena ja jakaa sitä eteenpäin. Koska hallusinoitut vastaukset voivat vaikuttaa ihmisten uskomuksiin, saattavat vastaukset myös vaikuttaa ihmisten päätöksentekoon. Tämän lisäksi, koska kielimallin haavoittuvuuksia voidaan hyödyntää eri tavoin, mallin tuottamien virheellisten vastausten avulla voidaan myös pyrkiä vaikuttamaan yleiseen mielipiteeseen, millä saattaa olla merkittävä vaikutus julkiseen keskusteluun.

Hallusinaatioiden riskit korostuvat varsinkin kriittisissä käyttökohteissa, kuten terveydenhuollossa, joissa virheille ei yleensä ole varaa. On monia käyttökohteita, joilla hallusinoinnin riskit ovat erityisen vakavat. Tällaisia käyttökohteita ovat esimerkiksi lääketieteelliset diagnoosit, oikeudelliset analyysit ja autonomiset järjestelmät. Joskus mallin hallusinointi voi myös johtaa oikeudellisiin toimiin. Tähän liittyen pitäisi ottaa huomioon muun muassa immateriaalioikeudet sekä kysymys siitä, kuka ottaa vastuun väärän tiedon aiheuttamasta vahingosta. Hallusinointi saattaa myös vähentää ihmisten luottamusta kielimalleihin. Vähenevän luottamuksen seurauksena myös yleinen mielipide kielimalleista heikentyy, mikä voi vaikuttaa kielimalleja hyödyntävien yritysten uskottavuuteen ja maineeseen. Hallusinaatiot eivät siis vaikuta pelkän teknologian luotettavuuteen, vaan herättävät myös epäilyksiä niitä hyödyntävien yritysten pätevydestä ja luotettavuudesta.

TK4: Millä tavoin kielimallien hallusinaatiota voidaan estää? Hallusinaatioiden vähentämiseksi voidaan hyödyntää tiedon varmistamista ulkoisista tietolähteistä. RAG, eli retrieval-augmented generation, on menetelmä suurten kielimallien ulkoisen tiedon saantiin. Menetelmä toimii käyttämällä ulkoisia tietolähteitä, jotka tarjoavat mallille ajankohtaista ja relevanttia tietoa, mikä puolestaan parantaa mallin kykyä tuottaa tarkempia vastauksia. RAG, ja erityisesti sen kehittyneemmät arkkitehtuurit, ovat osoittautuneet tehokkaaksi tavaksi vähentää hallusinoituja vastauksia. Yksi esimerkki kehittyneestä RAG-arkkitehtuurista on RAFT, eli retrieval augmented fine-tuning, joka on antanut lupaavia tuloksia oikeus- ja lääketieteen aloilla sisällyttämällä alakohtaisia tietograafeja tiedonhakuprosessiin. Kyseinen arkkitehtuuri pystyy olennaisen tiedon hakemisen lisäksi ymmärtämään myös monimutkaisia suhteita eri alojen välillä. Monivaiheinen RAG puolestaan mahdollistaa useiden eri hakujen tekemisen kielimallin vastauksen generoinnin aikana, mikä mahdollistaa luotettavamman vastaamisen sellaisiin kysymyksiin, jotka vaativat tietojen yhdistämistä monesta eri lähteestä tai joiden selittäminen vaatii monta eri vaihetta.

Myös kielimallin vastausten luotettavuutta voidaan arvioida eri menetelmin, mikä voi olla avuksi hallusinaatioiden tunnistamisessa. Luotettavuutta voidaan arvioida esimerkiksi toisen kielimallin avulla, joka voi etsiä vastauksesta esimerkiksi loogisia ristiriitoja tai arvioida sen uskottavuutta. Hallusinaatioiden tunnistamisessa ja suodattamisessa voi auttaa myös usean eri vastauksen hyödyntäminen. Sen sijaan, että luotetaan vain yhden kielimallin yhteen vastaukseen, voi olla luotettavampaa ottaa huomioon useiden eri kielimallien tulokset tai saman mallin useat eri tulokset. Kielimallin vastausten luotettavuuden arviointiin on hyvä valita sellaisia mittareita, jotka kuvaavat kattavasti vastauksen laadun eri osa-alueita. Yksi mittari voisi olla esimerkiksi kielimallin tekemä itsearviointi, jossa kielimalli arvioi omaa vastaustaan ilman ihmisten väliintuloa. Luotettavuuden arviointiin voisi vaikuttaa myös erilaiset tilastolliset muuttujat, joilla voitaisiin kuvata esimerkiksi vastauksen semant-

tista linjausta, relevanssia tai monimuotoisuutta. Luotettavuuden arvioinnissa on siis hyödyllistä käyttää useita eri mittareita, joille jokaiselle voidaan antaa erityinen painotus. Useita eri osa-alueita yhdistelemällä voidaan saada kattava arvio mallin vastauksen luotettavuudesta.

Lähdeluettelo

- [1] S. Patel ja R. A. Dholakiya, ”Large Language Models: Evolution, Architecture, Applications, and Future Horizons”, s. 1922–1929, 2025. DOI: 10.1109/ICSCSA66339.2025.11170884.
- [2] M. Maybury, ”Mitigating Biased, Brittle and Baroque Generative AI”, s. 1–7, 2025. DOI: 10.1109/ICAD65464.2025.11114038.
- [3] A. Boulesnane ja A. Souilah, ”An Evolutionary Large Language Model for Hallucination Mitigation”, s. 1–8, 2024. DOI: 10.1109/ECTE-Tech62477.2024.10851107.
- [4] M. Russinovich, A. Salem, S. Zanella-Béguelin ja Y. Zunger, ”The Price of Intelligence”, *Commun. ACM*, vol. 68, nro 9, s. 46–53, elokuu 2025, ISSN: 0001-0782. DOI: 10.1145/3749447.
- [5] G. P. Reddy, Y. V. Pavan Kumar ja K. P. Prakash, ”Hallucinations in Large Language Models (LLMs)”, s. 1–6, 2024. DOI: 10.1109/eStream61684.2024.10542617.
- [6] D. Jurafsky ja J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd. 2025, Online manuscript released August 24, 2025. url: <https://web.stanford.edu/~jurafsky/slp3/>.

-
- [7] E. Sanu, T. K. Amudaa, P. Bhat, G. Dinesh, A. U. Kumar Chate ja R. K. P, "Limitations of Large Language Models", s. 1–6, 2024. DOI: 10.1109/CSITSS64042.2024.10817070.
- [8] A. Vaswani et al., "Attention Is All You Need", 2023. arXiv: 1706.03762 [cs.CL]. url: <https://arxiv.org/abs/1706.03762>.
- [9] S. B. Shah et al., "Navigating the Web of Disinformation and Misinformation: Large Language Models as Double-Edged Swords", *IEEE Access*, vol. 13, s. 169 262–169 282, 2025. DOI: 10.1109/ACCESS.2024.3406644.
- [10] A. T. Kalai, O. Nachum, S. S. Vempala ja E. Zhang, "Why Language Models Hallucinate", 2025. arXiv: 2509.04664 [cs.CL]. url: <https://arxiv.org/abs/2509.04664>.
- [11] Z. Wei et al., "Detecting and Mitigating the Ungrounded Hallucinations in Text Generation by LLMs", *AISNS '23*, s. 77–81, 2024. DOI: 10.1145/3661638.3661653.
- [12] W. Su, Y. Tang, Q. Ai, C. Wang, Z. Wu ja Y. Liu, "Mitigating Entity-Level Hallucination in Large Language Models", *SIGIR-AP 2024*, s. 23–31, 2024. DOI: 10.1145/3673791.3698403.
- [13] Y. Chen et al., "Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models", *CIKM '23*, s. 245–255, 2023. DOI: 10.1145/3583780.3614905.