KUN TEKOÄLY HALLUSINOI

”Faktojen hallusinointi” on sanapari, johon tekoälystä kiinnostuneet ovat todennäköisesti törmänneet viime aikoina erityisesti ChatGPT:n kaltaisten kielimallien yhteydessä. Faktojen hallusinoimisella viitataan tilanteeseen, jossa kielimalli tarjoaa väärää, valheellista tai keksittyä tietoa sisältäviä vastauksia. Termiä on kuitenkin kritisoitu tekoälyä liikaa inhimillistäväksi, koska tekoälyn hallusinaatioilla on hyvin vähän tekemistä ihmisten kokemien hallusinaatioiden kanssa.

TEKSTI: OSKARI LAPPALAINEN

Faktojen hallusinointi -ilmauksen tilalle on ehdotettu konfabulaatioita tai suomalaisittain sepittämistä. Faktojen hallusinointi näyttää kuitenkin olevan vakiintumassa kielimallien tuottamia faktavirheitä kuvaavaksi termiksi. Tämän taustalla saattaa olla se, että nykyiset kielimallit voivat tuottaa suorastaan fantastisella tavalla väärässä olevaa informaatiota. ChatGPT on saatu tuottamaan sangen eriskummallista ”tietoa” muun muassa churro-leivosten käytöstä kirurgisina välineinä. On sangen ymmärrettävää, että ihmiset näkevät näin radikaalisti todellisuudesta poikkeavissa väitteissä jotain hallusinaatioiden kaltaista.

Mistä kielimallien taipumus sepittää faktoja sitten johtuu? Tämä saattaa johtua osittain koulutusdatan virheistä, mutta se on osittain myös nykyisten kielimallien toimintaan liittyvä piirre: nämä mallit ottavat käyttäjän syötteen, purkavat sen osiin ja yrittävät todennäköisyyksiin perustuvien päättelyketjujen avulla antaa kaikista todennäköisimmän vastauksen syötteeseen. Juuri tämä todennäköisyyksien hyödyntäminen johtaa väistämättä jonkinlaisiin virheisiin, ja kielimallien suhteen niissä käytetyn massiivisen datamäärän ”kohina” (aihepiirille epärelevantti tieto) ja se, että mallit oppivat itsekseen ilman ihmisten väliintuloa, pahentavat hallusinaatioita.

Kielimallien hallusinaatiota pahentavat myös niitä luovien organisaatioiden valinnat. Vaikka kielimallien toiminta on osittain vaikea selvittää, niistä pystyy näkemään, kuinka todennäköisenä malli pitää kutakin päättelyketjun osaa. Tästä puhutaan joskus tekoälymallin lämpötilana, ja mitä ”kuumempi” malli on, sitä satunnaisempia ja potentiaalisesti virheellisempiä päättelyketjut ovat. Esimerkiksi ChatGPT:stä on mahdollista saada ulos sen ”lämpötila” käyttäen rajapintaa, mutta tavallinen käyttäjä ei pääse siihen käsiksi, saati sitten säätämään mallia.

Hallusinointiin liittyvät ongelmat

Kuinka suuri ongelma tekoälyn hallusinointi oikeastaan on? Tämä riippuu siitä, mihin tarkoituksiin kielimalleja halutaan käyttää, mutta ainakin luotettavia lähteitä vaativissa tehtävissä hallusinointi on varsin suuri ongelma. Eräässä tutkimuksessa havaittiin, että 115:sta ChatGPT:ltä pyydetystä lähdeviitteestä 47 % oli olemattomia, 46 % oikeita mutta virheellisiä ja vain 7 % oli sekä oikeita että virheettömiä.

Tekoälyn hallusinaatiot ovat myös ehtineet aiheuttaa vahinkoa. Yhdysvalloissa oli oikeustapaus, jossa asianajaja oli yrittänyt käyttää ChatGPT:ltä pyydettyjä ennakkotapauksia. Valitettavasti nämä tapaukset olivat ChatGPT:n hallusinoimia, ja tapaus johti linjaukseen, jossa kyseisessä piirikunnassa kiellettiin tekoälyn generoimien dokumenttien käyttö oikeudessa. Lisäksi Yhdysvalloissa OpenAI Foundation on haastettu oikeuteen tapauksesta, jossa ChatGPT tuotti valheellisia herjaavia väitteitä julkisuuden henkilöstä.

Kaiken kaikkiaan kielimallien hallusinaatiota pidetään varsin suurena ongelmana, ja esimerkiksi Google on panostamassa voimakkaasti hallusinaatioiden torjuntaan. Valitettavasti keinot hallusinaatioiden torjumiseksi ovat tällä hetkellä rajattuja, sillä hallusinaatioita tuottavia mekanismeja ei vielä ymmärretä täydellisesti. Kehitteillä on kuitenkin erilaisia metodeja hallusinaatioiden vähentämiseksi. Tällaisia ovat muun muassa tekoälyn tuottaman vastauksen vertailu hakutuloksiin sekä toisen tekoälyn käyttö faktantarkistukseen.

Ainoastaan tulevaisuus voi näyttää, kuinka onnistuneita nämä ratkaisut hallusinaatioiden vähentämiseksi ovat. Niitä odotellessa hallusinaatiot asettavat varsin merkittäviä rajoitteita tekoälyn käytölle faktapohjaisen tiedon tuottamiseksi. Sen sijaan esimerkiksi tekstin muokkauksessa hallusinaatiot ovat vähemmän merkittävä ongelma, koska tilaa sepittämiselle ei yksinkertaisesti ole. Mikäli tekoälyjen hallusinaatiota ei saada kuriin, generatiiviset tekoälyt saattavatkin päätyä eräänlaisiksi hyvin edistyneiksi tekstinkäsittelyohjelmiksi. Tällä voi olla erittäin suuri käyttöarvo, mutta ei ehkä sellainen, mitä tekno-optimistit toivoivat. ■

Oskari Lappalainen on projektityöntekijä Laurea AMK:ssa. Hän työskenteli tekoälyn käyttöönottoa pk-yrityksissä edistäneessä AI-TIE-hankkeessa. Lappalainen on ollut vuosikausia pragmaattisesti kiinnostunut tekoälyn mahdollisuuksista ja mahdottomuuksista. Tällä hetkellä hän työskentelee CYCLOPES-kyberturvallisuushankkeessa.

Aiheesta muualla

Belanger, Ashley 2023. OpenAI faces defamation suit after ChatGPT completely fabricated another lawsuit. Saatavilla: https://arstechnica.com/tech-policy/2023/06/openai-sued-for-defamation-after-chatgpt-fabricated-yet-another-lawsuit/

Bhattacharyya, Mehul & Valerie M. Miller & Debjani Bhattacharyya & Larry E. Miller 2023. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. Saatavilla: https://www.cureus.com/articles/158289-high-rates-of-fabricated-and-inaccurate-references-in-chatgpt-generated-medical-content#!/

Edwards, Benj 2023. Why ChatGPT and Bing Chat are so good at making things up. Saatavilla: https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/

Marcus, Gary 2022. How come GPT can seem so brilliant one minute and so breathtakingly dumb the next? Saatavilla: https://garymarcus.substack.com/p/how-come-gpt-can-seem-so-brilliant

Maruf, Ramishah 2023. Lawyer apologizes for fake court citations from ChatGPT. Saatavilla: https://edition.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers/index.html

Tam, Adrian 2023. A Gentle Introduction to Hallucinations in Large Language Models. Saatavilla:https://machinelearningmastery.com/a-gentle-introduction-to-hallucinations-in-large-language-models/