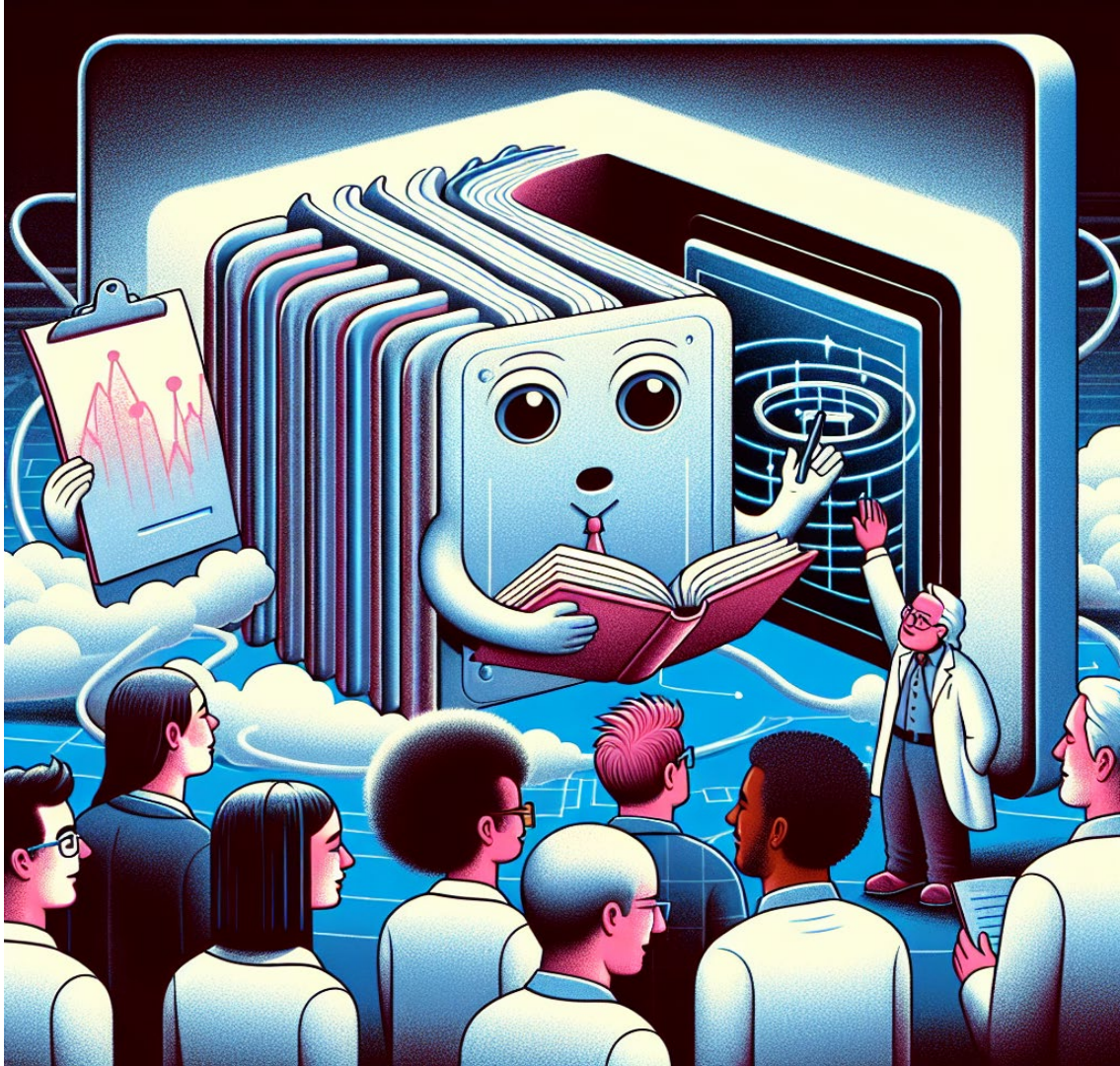
KANSALLISARKISTO

Kuinka tehdä kielimalleista fiksumpia tausta-aineiston avulla

Mikko Lipsanen

18.9.2024

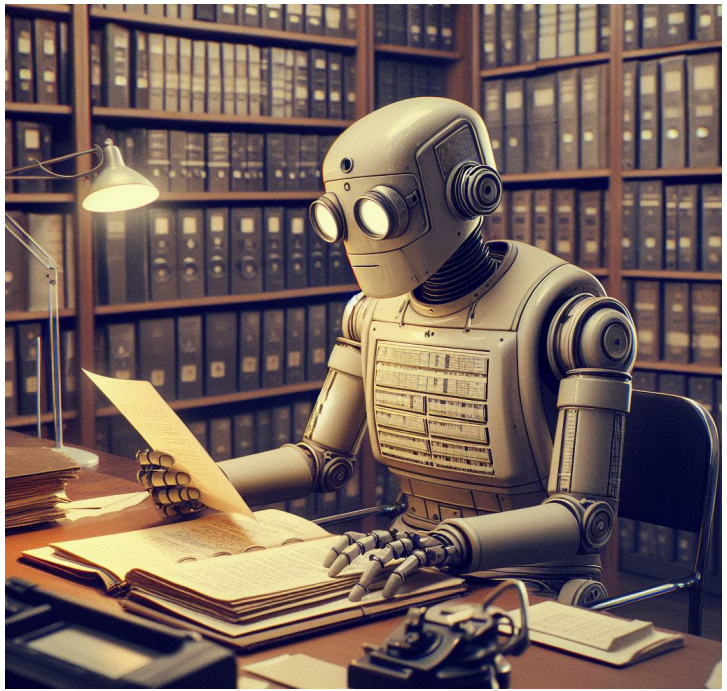


Suurten kielimallien haasteet

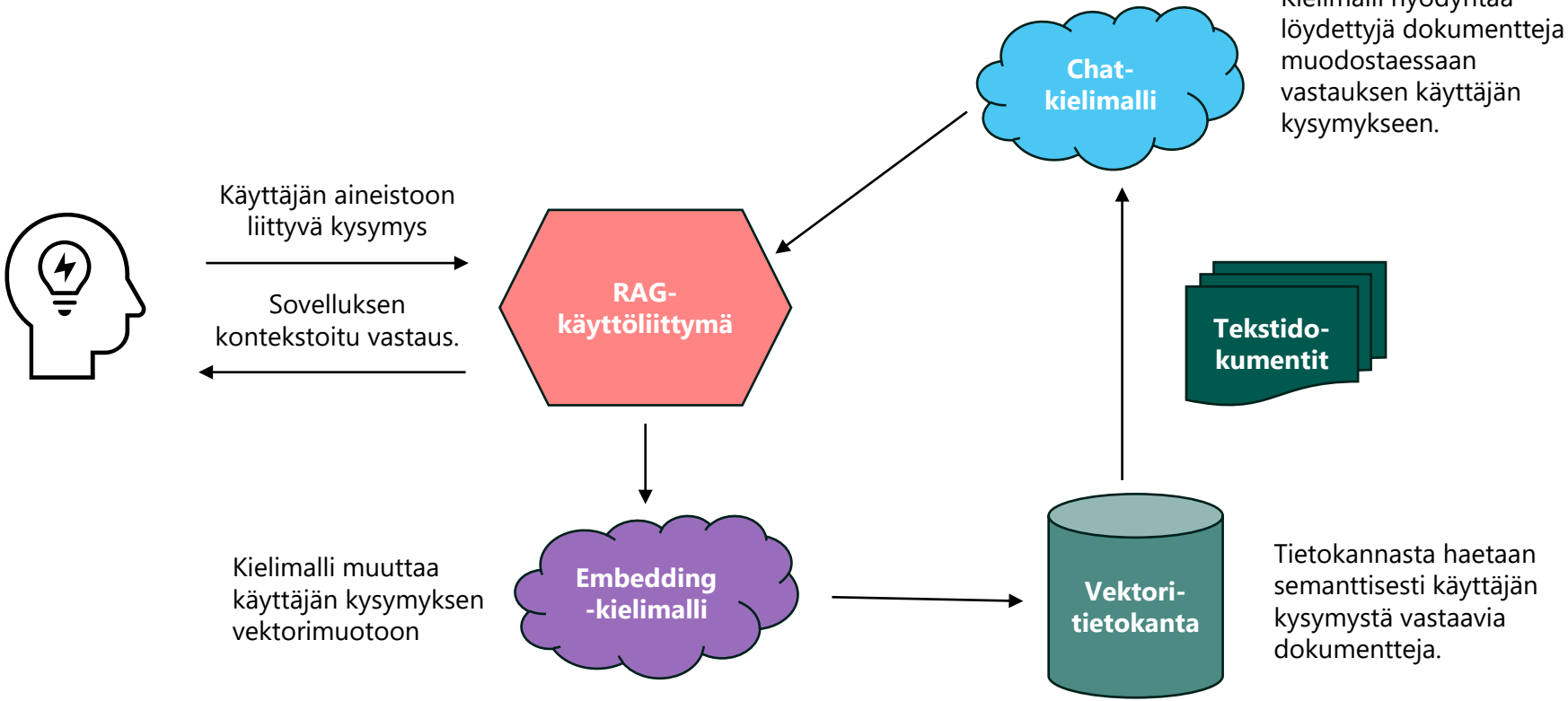
- Suurten kielimallien kyky tuottaa erilaisia kielellisiä sisältöjä ja tekstityyppejä on kehittynyt harppauksin muutamassa vuodessa
- Kielimallit eivät kuitenkaan ole erityisesti kunnostautuneet luotettavan, ajantasaisen ja kohdennetun faktatiedon tuottajina
  - Kielimallin ”tieto” perustuu sen koulutusaineistoon, joka on aina rajallinen ja painottaa tietyntyyppisiä tekstejä
  - Kielimallin opetusaineistosta ”muistama” informaatio on aina jossain määrin satunnaista, eikä se lähtökohtaisesti tunnusta tietämättömyyttään, vaan tuottaa sen sijaan oikean oloista sisältöä

Kielimalli + tiedonhaku ≈ RAG

- Kielimallin ei tarvitse vastausta muodostaessaan rajoittua vain siihen informaatioon, jonka se ”muistaa” koulutusaineistosta (ja keksiä loput)
- Hyödyntämällä tausta-aineistoon kohdentuvaa tiedonhakua voidaan saada täsmällisempiä ja laadukkaampia vastauksia
- **Retrieval Augmented Generation (RAG)** yhdistää tietojen haun (retrieval) ja tekstin generoinnin (generation)
    1. Käyttäjä esittää kysymyksen
    2. Kysymyksen perusteella etsitään relevanttia taustadataa määritellystä lähteestä (dokumentti, tietokanta, internet..)
    3. Muodostetaan vastaus kysymykseen hyödyntäen löydettyä tausta-aineistoa



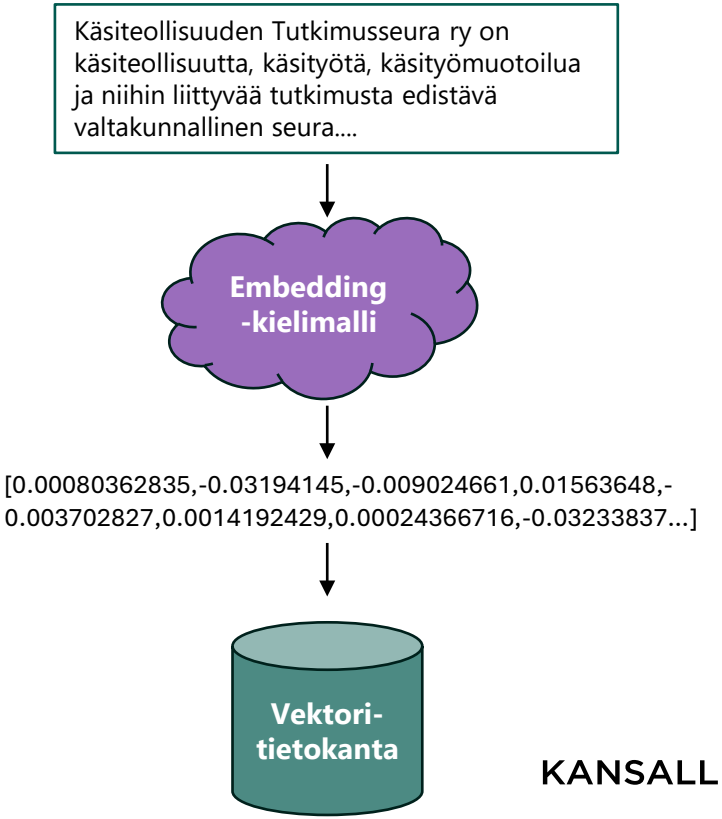
Esimerkki RAG-arkkitehtuurista

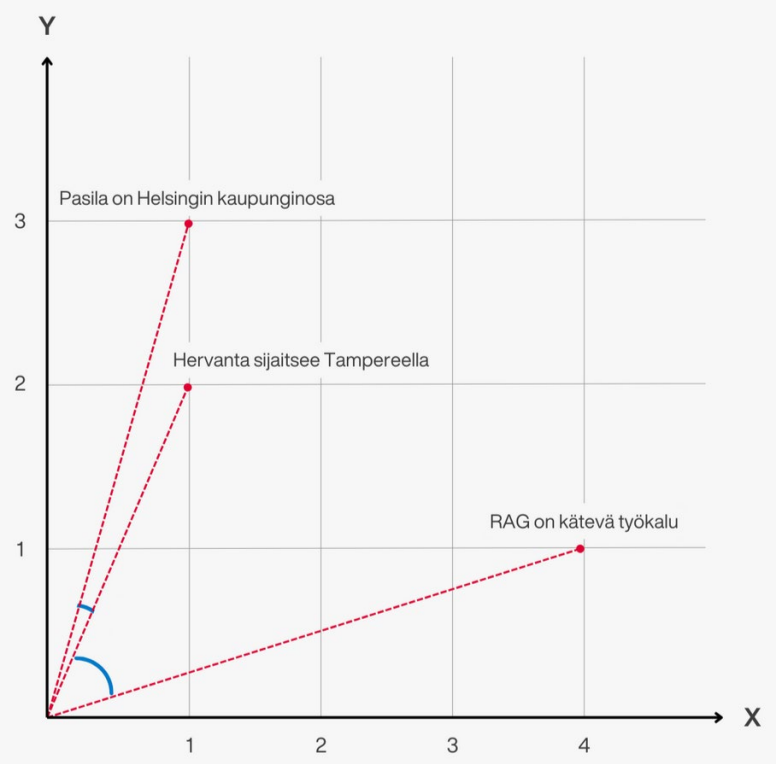


Semanttinen haku tausta-aineistosta

- Yleinen tapa etsiä käyttäjän kysymyksen kannalta relevanttia tausta-aineistoa hyödyntää semanttista hakua ja vektoritietokantoja

  - **Semanttinen haku**: pyrkimyksenä löytää aineistosta parhaiten käyttäjän kysymyksen tarkoitusta / merkitystä vastaavaa dataa, sen sijaan että haettaisiin esim. suoria vastaavuuksia hakusanoihin
  - **Vektoritietokanta**: Tietokantaan tallennetaan kielimallin avulla vektorisoituja tekstejä, ja haun ja dokumentin (semanttinen) samankaltaisuus määritellään laskemalla vektorien samankaltaisuus (esim. vektorien kosinietäisyys)

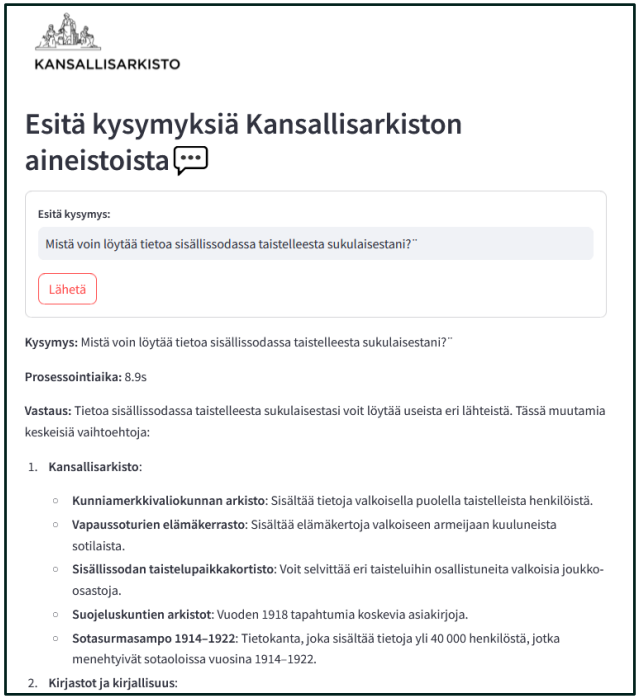




Yksinkertaistettu esimerkki (2-ulotteisten) vektorien samankaltaisuuden mittaamisesta hyödyntämällä vektorien välisten kulmien kosiniarvoja. Lähde: https://www.efima.com/blogi/kurkistus-hypetetyn-tekoalytermin-konepellin-alle-miten-rag-toimii-ja-mita-sen-hyodyntaminen-vaatii

Mahdollisia käyttökohteita

- Kun halutaan parantaa kielimallien luotettavuutta ja tarkkuutta kytkemällä vastaukset määrättyyn tausta-aineistoon
- Kun halutaan parantaa aineistojen saavutettavuutta mahdollistamalla kysymysten esittäminen ”aineistolle” kielimallin välityksellä
- Kohdennetut chatbotit / asiakaspalvelun avustajat
    - Esim. Kansallisarkistolla on parhaillaan sisäisessä testikäytössä arkiston aineistoihin liittyviin kysymyksiin erikoistunut Arkistoapuri
- Kontekstoitu tiedonhaku määrätyistä aineistoista + vastausten muotoilu (esim. tiivistelmät)

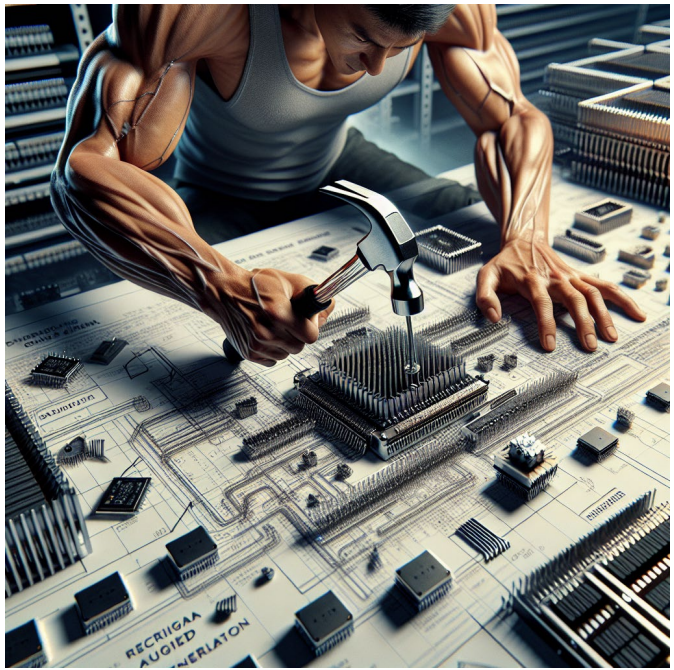


Haasteita ja kehityskohteita

- RAG on vielä varsin tuore ja jatkuvasti kehittyvä teknologia, monet ratkaisut ovat vielä vakiintumattomia
- Esimerkkejä eri osa-alueisiin liittyvistä haasteista
  - **Käyttäjän kysymyksen muotoilu**: lyhyiden ja epämääräisten kysymysten pohjalta on vaikea löytää hyvää tausta-aineistoa (mahdollisena ratkaisuna esim. kysymysten uudelleenmuotoilu kielimallin avulla)
  - **Tausta-aineisto**: laadukkaita ja ajantasaisia vastauksia voi saada vain laadukkaan ja ajantasaisen aineiston pohjalta
  - **Aineistohaku**: vektorihaun laatua voi heikentää mm. käytetty embedding-kielimalli, vektoroitujen tekstikokonaisuuksien liian suuri tai pieni koko ja aineiston heterogeenisyys. Suuri aineistomäärä myös hidastaa hakua ja vastauksen muodostamista.
  - **Vastauksen generointi**: lopullisen vastauksen laatuun vaikuttaa yhtäältä löydetyn taustadatan laatu ja toisaalta vastauksen muodostavan kielimallin kyky löytää aineistosta oikeat asiat ja muotoilla niiden avulla kattava ja selkeä vastaus
- Jos käytetään maksullisia kielimalleja ja tietokantoja, voi myös hinta muodostua haasteeksi RAG-teknologian käytölle. Esim. käytettyjen tausta-aineistojen määrän kasvaessa nousee myös yksittäisen kielimallikutsun kustannus.

Kuinka pystyttää oma RAG?

- Ainakin periaatteessa helpoin vaihtoehto on hyödyntää pilvipalveluiden (Azure, AWS..) valmiita RAG-ratkaisuja
  - Ei ole välttämättä aina se edullisin vaihtoehto
  - Aineiston muokkaaminen ja parametrien säätäminen vaatinee silti melkolailla työtä
  - Palveluntarjoajien valmiit kehikot rajoittavat eri vaihtoehtojen kokeilemista
- Oman RAG-sovelluksen kehittäminen hyödyntäen ohjelmointikirjastoja (esim. Langchain) ja kielimallien ja vektoritietokantojen rajapintoja on myös suhteellisen helppoa
  - Perustoiminnallisuus voidaan pystyttää nopeasti, mutta optimointi vaatii kuitenkin aikaa ja testausta
  - Miten saada parhaita tuloksia juuri minun aineistollani / asiakkailleni?
  - Kansallisarkistolla saadut tulokset ovat kuitenkin rohkaisevia, RAG mahdollistaa kielimallien hyödyntämisen tehtäviin joissa niitä ei muuten voitaisi yllä mainittujen haasteiden takia käyttää



KANSALLISARKISTO

mikko.lipsanen@kansallisarkisto.fi