Katsaus avoimen lähdekoodin kielimalleihin

Juho-Pekka Mäkinen

Lähtökohta

- isojen yritysten kielimalleista puhutaan usein "avoimen lähdekoodin" malleina
- todellisuudessa mallien lisenssit eivät täytä avoimen lähdekoodin kriteereitä
- termiä käytetään virheellisesti, myös tieteellisissä julkaisuissa
- tarve kartoittaa aidosti avoimen lähdekoodin kielimallit

Avoin lähdekoodi (open source)

Open Source Initiative (OSI) määrittelee avoimen lähdekoodin

- ohjelmiston vapaa jakelu ja muokkaus
- lähdekoodin avoin saatavuus
- syrjimättömyys (käyttäjät, käyttötarkoitukset)
- teknologinen neutraalius
- oikeuksien säilyminen ja periytyminen

Tavoitteet ja kysymykset

- kartoittaa saatavilla olevia avoimen lähdekoodin kielimalleja
- analysoida mallien arkkitehtuuria, koulutusdataa ja käyttötarkoituksia
- tutkimuskysymykset:
  - Mitä avoimen lähdekoodin kielimalleja on saatavilla?
  - Millaisia teknisiä toteutuksia avoimen lähdekoodin kielimallien arkkitehtuureissa ja koulutusmenetelmissä on hyödynnetty?

Kielimallien valintakriteerit

- lähdekoodi saatavilla
- avoimen lähdekoodin lisenssi
- koulutusdata julkinen
- riittävästi yksityiskohtaista tietoa teknisistä ratkaisuista

Aineiston keruu tehty

- aineiston haku google scholarilla
  - hakutermi "open source large language model"
- rajattu vuoden 2023 alusta hakuajankohtaan
- tarkasteltiin 150 ensimmäistä hakutulosta
- 15 kielimallia valikoitui
  - 6 esikoulutettua mallia
  - 9 hienosäädettyä mallia

Haasteita

- pitkä tauko gradun teossa
- ajankäyttö
- motivaation lasku

Lähteet

- **Open Source Initiative** Maffulli, S. (2023). *Meta's LLaMa 2 license is not Open Source*. Open Source Initiative. Haettu osoitteesta https://opensource.org/blog/metas-llama-2-license-is-not-open-source

- **Open Source Initiative**. (2006, 7. heinäkuuta, päivitys 16. helmikuuta 2024). *The Open Source Definition*. Haettu osoitteesta https://opensource.org/osd