"Danoliterate" evaluering af GLLM'er

Søren Vejlgaard Holm

Lars Kai Hansen, Danmarks Tekniske Universitet
Martin Carsten Nielsen, Alvenir ApS

20. september 2024

Dagsorden

... og hvad er det her?

- Kandidatspeciale[1] og projekt[2] fra DTU Compute
  - udarbejdet af undertegnede, cand. polyt. fra DTU. ML-ingeniør hos Alvenir og videnskabelig assistent på DTU Compute.
  - vejledt af Lars Kai Hansen, professor og sektionsleder, Cognitive Systems på DTU Compute.
  - og Martin Carsten Nielsen, cand. polyt. fra DTU. Grundlægger, direktør i Alvenir.
  - støttet af Danske Pioneer Centre for AI, DNRF bevilling P1.

DTU

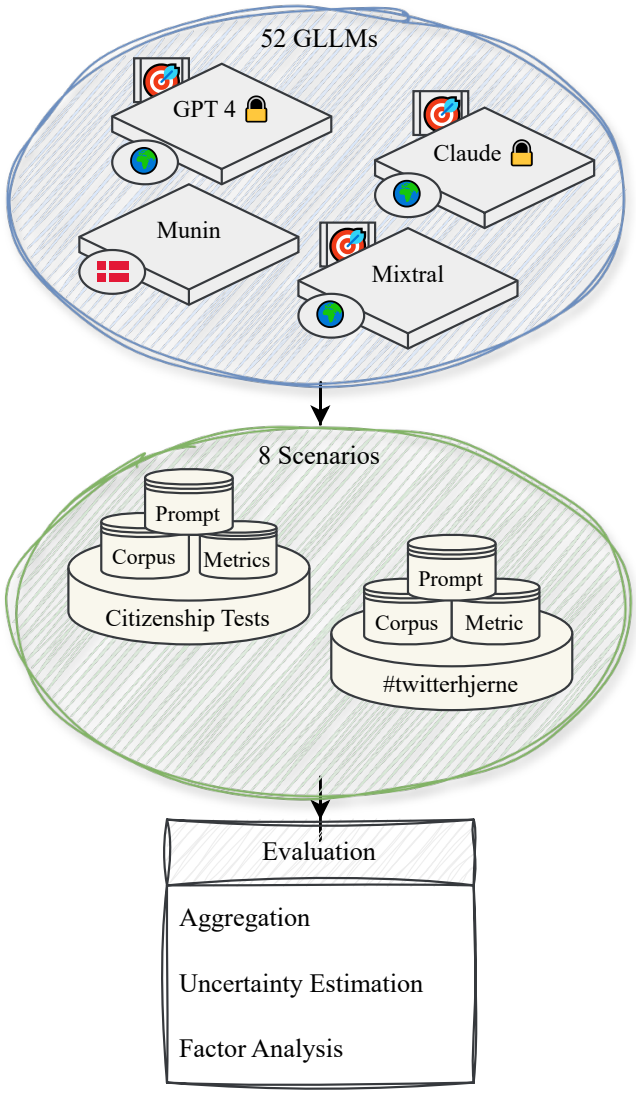
ALVENIR
SPEECH TECHNOLOGIES

[1] sorenmulli.github.io/thesis/thesis.pdf
[2] danoliterate.compute.dtu.dk

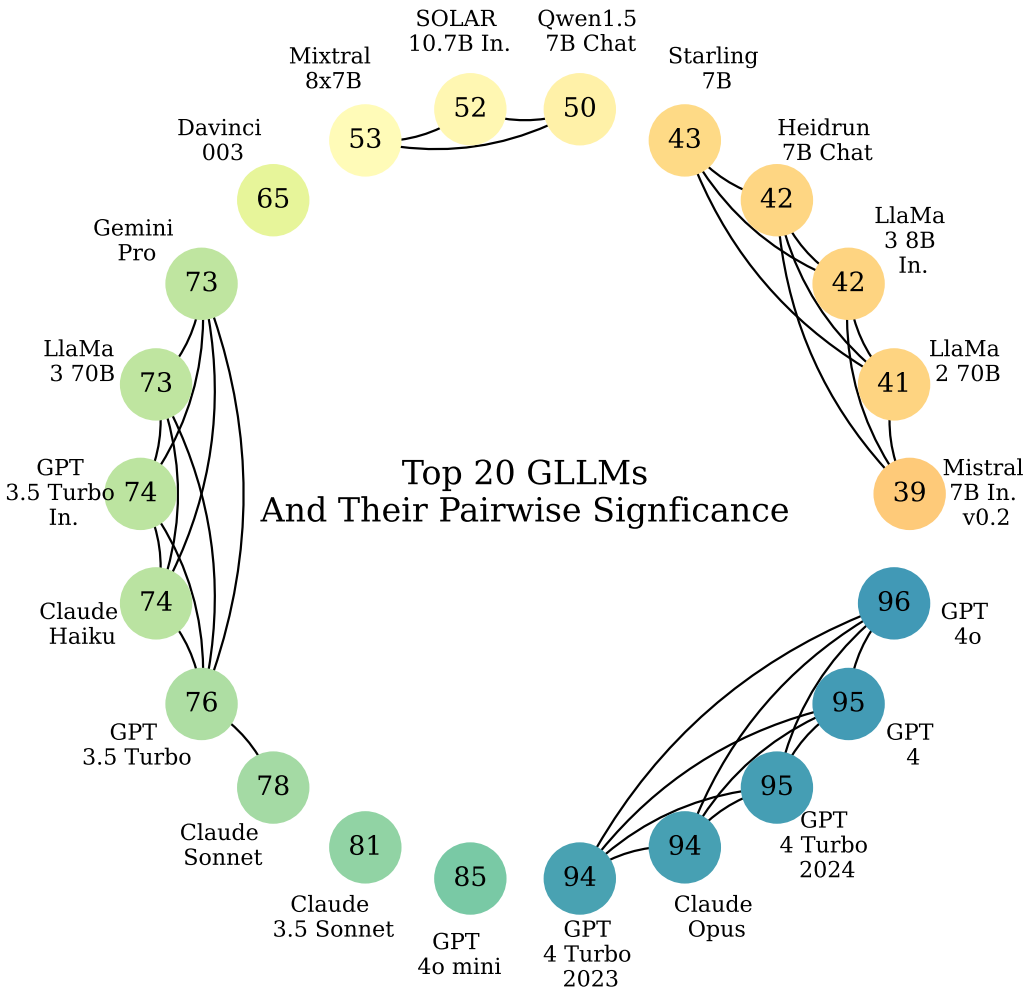
Motivation for at evaluere

- Alvenir; Få noget ud af jeres talegenkendelse!
- ChatGPT og prototype
  - GLLM'er som fleksible demonstrationsmaskiner
  - I mindre grad fokus på kanoniske NLP-opgaver
- Svær, bred evalueringsopgave
  - Generelle evner inden for forståelse, generering og viden
  - Meget engelsksproget fokus; hvad med dansk?
  - "Danoliteracy"

En samling af scenarier

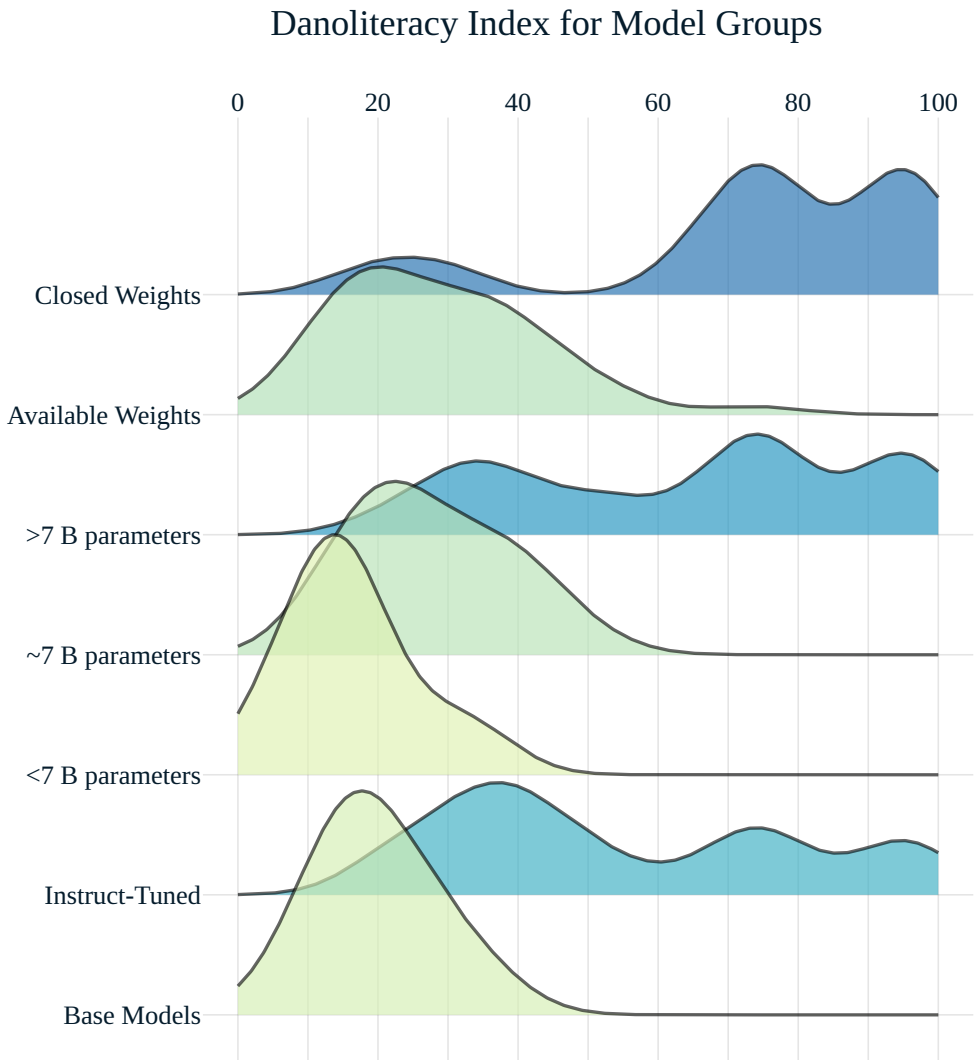


Opsummering af resultater[3]

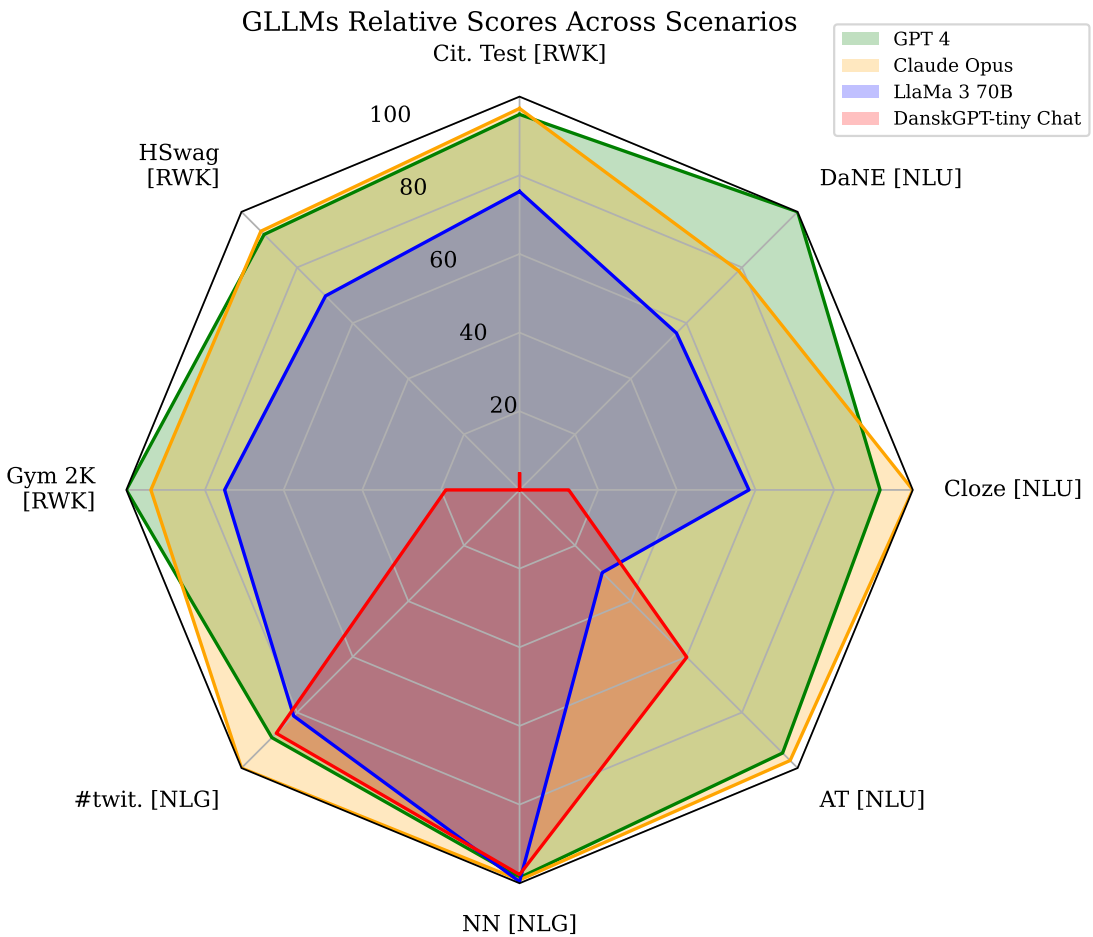


[3] Alle resultater: danoliterate.compute.dtu.dk/Leaderboard

Opsummering af resultater[3]



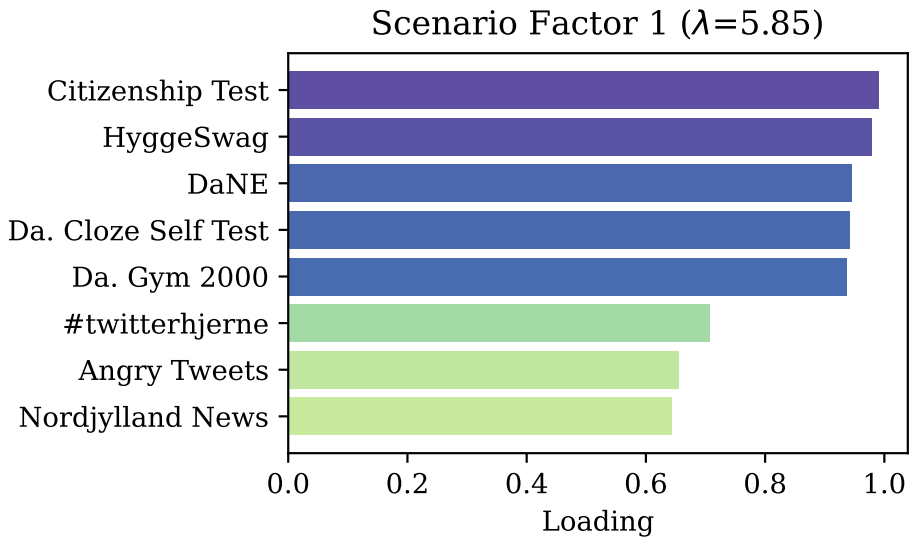
Opsummering af resultater[3]



[3] Alle resultater: danoliterate.compute.dtu.dk/Leaderboard

Godt og skidt

- De bedste modeller er
  - Store
  - Instruct-tunede
  - ... og oftest lukkede
- At bruge en samling af scenarier til at evaluere "Danoliteracy"
  - Ser lovende ud: Høj resultatkorrelation på tværs af scenaier
  - Stærk, fælles faktor: Men er den meningsfuld?

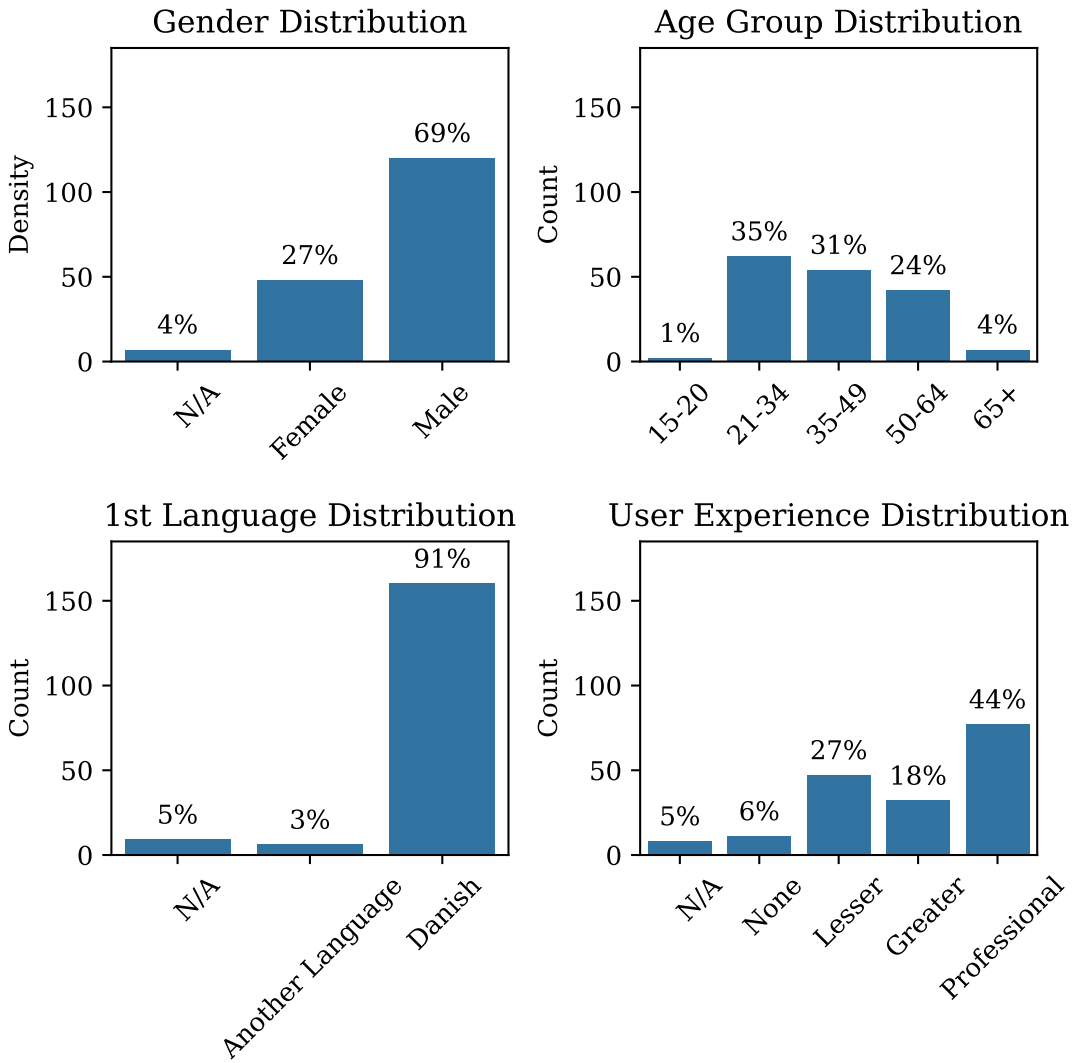


Hvordan kan mennesker inddrages?

- 18 modeller
  - I modsætning til LMSYS Chatbot Arena: Ikke live
- Prægenererede prompts og svar til 100 populære brugssituationer
- Interaktiv, åben side sat op som A/B-test
  - danoliterate.compute.dtu.dk/Spørgeskema

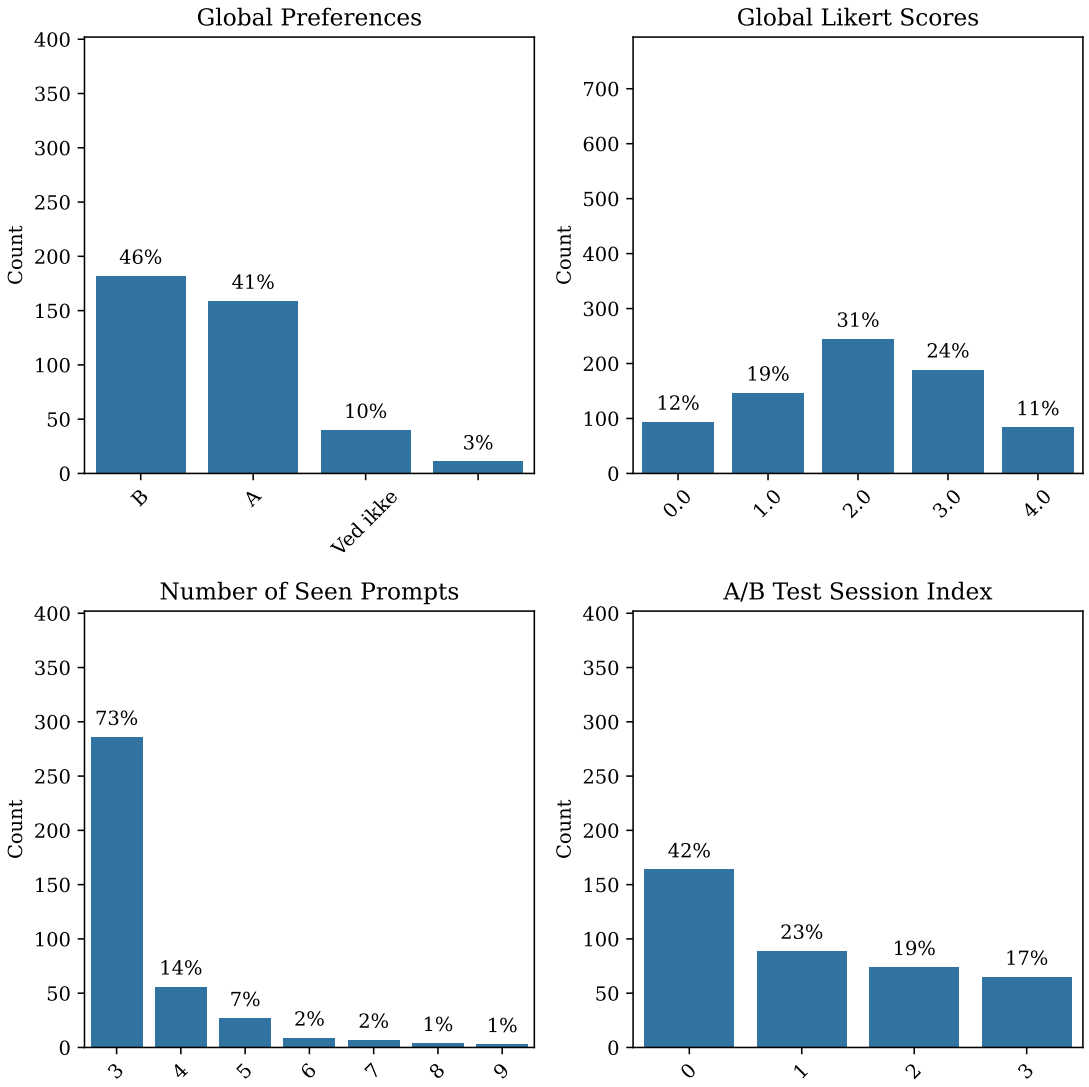
Resultater

- 392 A/B-tests fra 175 deltagere



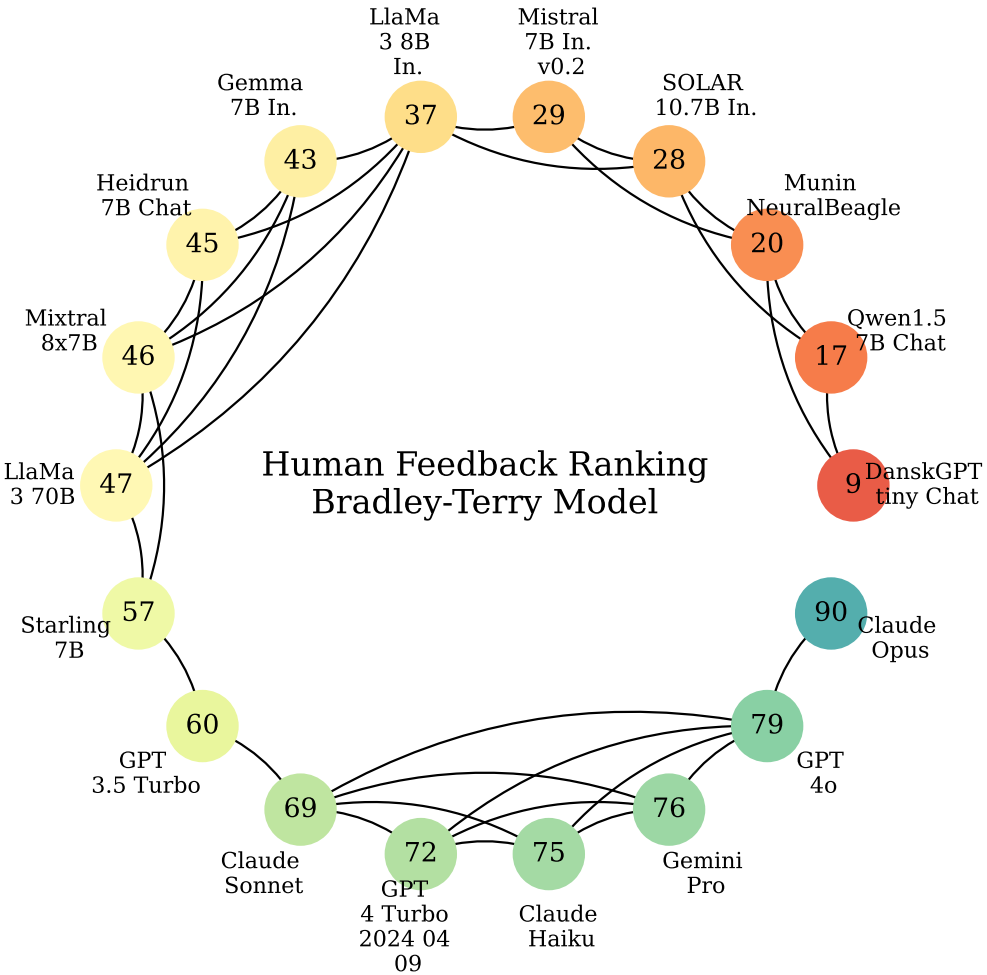
Resultater

- 392 A/B-tests fra 175 deltagere



Resultater

- 392 A/B-tests fra 175 deltagere



Erfaringer

- Sammenhæng mellem menneskelige og automatiske resultater
  - Fremstår høje for danskfokuserede benchmarks
    - Danoliterate: $\rho \sim 0.8$ for 18 modeller
    - Også høj og signifikant for ScandEval
  - Som i sig selv ikke er højt korreleret med engelsksprogede benchmarks
    - Danoliterate og HELM $\rho \sim 0.5$ (12 modeller)
    - Danoliterate og Open LLM Leaderboard $\rho \sim 0.5$ (15)
- Mere analyse, tak! Data nu offentliggjort
  - Prompts og modelsvar: huggingface.co/datasets/sorenmulli/danoliterate-survey-prompts
  - Nuværende besvarelser huggingface.co/datasets/sorenmulli/danoliterate-survey-answers

Tusind tak

- Tag spørgeskemaet!
  - Gør meget gerne reklame for det.
  - danoliterate.compute.dtu.dk/Spørgeskema
- Læs mere på
  - Side: danoliterate.compute.dtu.dk
  - Speciale: sorenmulli.github.io/thesis/thesis.pdf
  - Præsentation af spørgeskema-resultater: danoliterate.compute.dtu.dk/Articles
- Spørgsmål?