Evaluering af sprogmodeller

DIGST workshop

20. september 2024

Dan Saattrup Nielsen

Senior AI Specialist @Alexandra Instituttet

TrustLLM

Funded by
the European Union

Hvordan kan vi evaluere sprogmodeller?

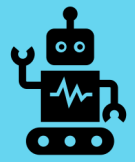
TrustLLM

Funded by
the European Union

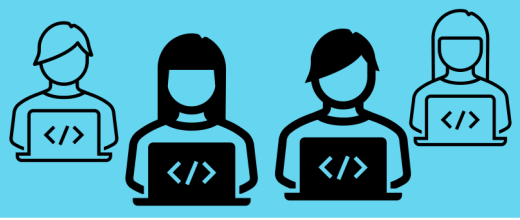
Fire evalueringstilgange af sprogmodeller

Vibetjek

LLM-som-dommer



Arena



Benchmark

Funded by the European Union

Vibetjek 

Bare prøv at skrive med modellen.

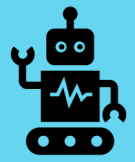
Pros	Cons
Giver typisk et godt udgangspunkt	Generaliserer ikke til andre opgaver
Meget billigt og hurtigt	Kan kun evaluere instruktionstunede modeller

Funded by the European Union

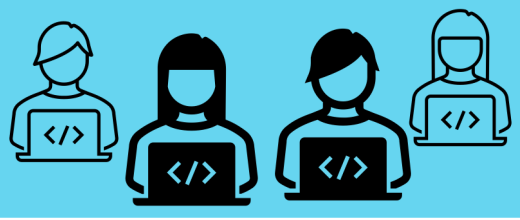
Fire evalueringstilgange af sprogmodeller

Vibetjek

LLM-som-dommer



Arena



Benchmark

Funded by the European Union

Arena 

Få brugerne til at spørge modeller og få svar fra flere (anonymiserede) modeller og få dem til at rangere dem. Brug typisk ELO-score til at samle disse.

Pros	Cons
Mere relevant for brugerens use cases*	Kan kun evaluere instruktionstunede modeller
Relativt objektivt mål, når en kritisk masse af frivillige har stemt	Kræver *mange* frivillige til at evaluere

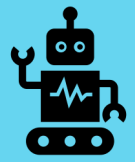
* Afhænger af typen af spørgsmål og/eller brugere, der bidrager

Funded by the European Union

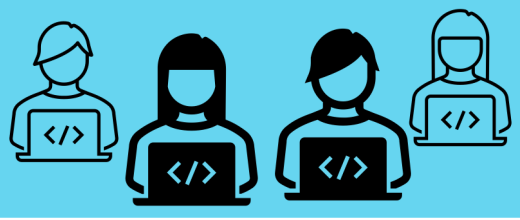
Fire evalueringstilgange af sprogmodeller

Vibetjek

LLM-som-dommer



Arena



Benchmark

Funded by the European Union

LLM-som-dommer 

Prompt modellerne med faste prompts (f.eks. "Skriv en jobansøgning til {jobannonce} fra {cv}") og få en anden LLM til at evaluere resultaterne.

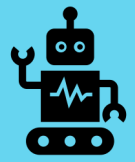
Pros	Cons
Billigt at opsætte og evaluere	Kræver eksistensen af en meget god LLM på det givne sprog
Mål, der kun skal udføres én gang for hver model	Kan kun evaluere instruktionstunede modeller

Funded by the European Union

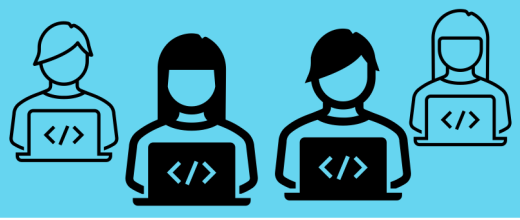
Fire evalueringstilgange af sprogmodeller

Vibetjek

LLM-som-dommer



Arena



Benchmark

Funded by the European Union

Benchmark

Sammensæt en række datasæt og evaluer modellen på disse.

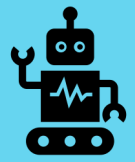
Pros	Cons
Objektiv evaluering, der kun skal udføres én gang for hver model	Generaliserer ikke nødvendigvis til andre typer opgaver
Kan evaluere alle typer sprogmodeller	Det er dyrt at oprette evalueringsdatasæt

Funded by the European Union

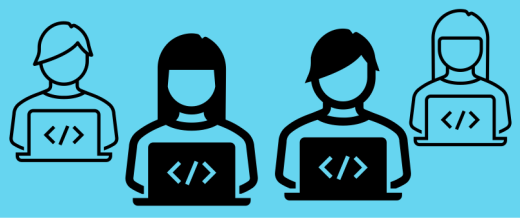
Fire evalueringstilgange af sprogmodeller

Vibetjek

LLM-som-dommer

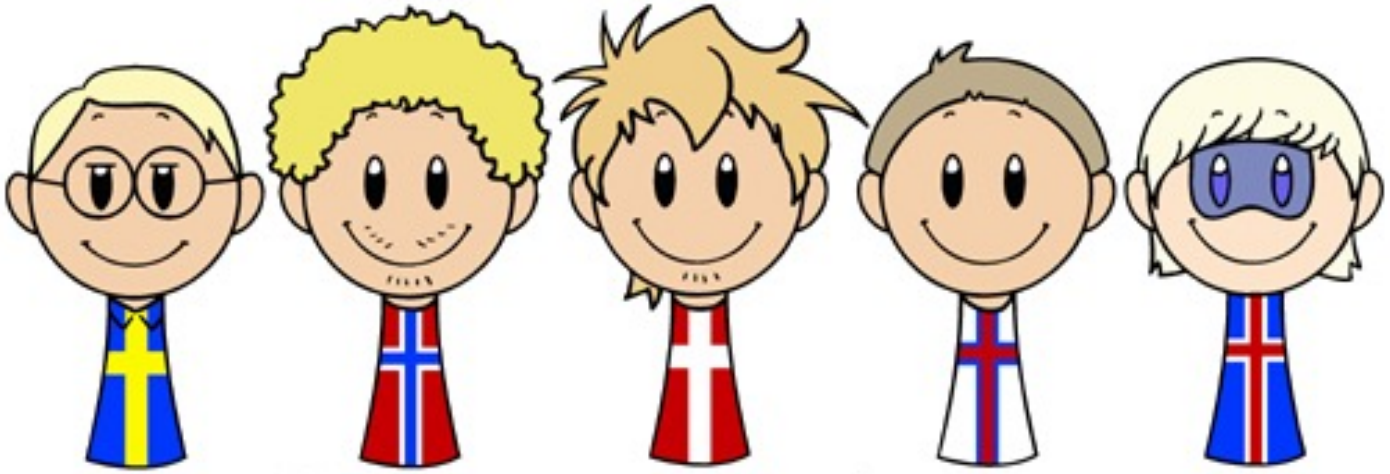


Arena



Benchmark

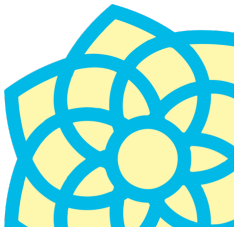
Funded by the European Union



Funded by
the European Union

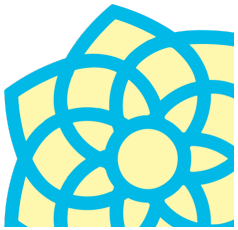
ScandEval er et robust multilingvalt evalueringsframework

Funded by
the European Union



ScandEval er et robust multilingvalt evalueringsframework

Funded by
the European Union

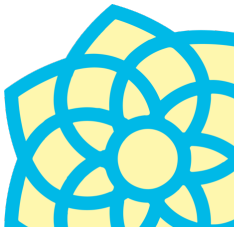


Sprogmodelsevalueringsframework

- Gør det muligt at evaluere implicit hvor gode sprogmodeller er til både sprogforståelse (NLU) og sproggenerering (NLG)
- Kan evaluere både encodere via finjustering, og decodere via few-shot evaluering
  - Det er blevet vist, at few-shot inferens af decodermodeller svarer til finjustering [1]
  - Det gør det dermed muligt for os at direkte sammenligne encodere med decodere

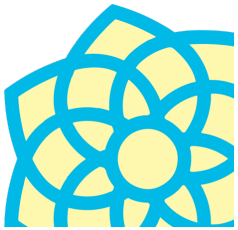
[1] von Oswald et al. arXiv preprint arXiv:2309.05858 (2023)

Funded by the European Union



ScandEval er et robust multilingvalt evalueringsframework

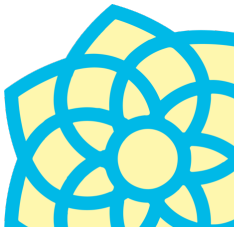
Funded by
the European Union



Evalueringsrobusthed

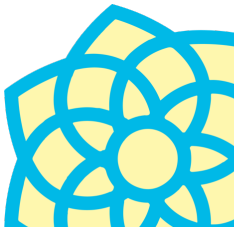
- Ved evaluering af modeller er der flere støjkilder i evalueringsresultatet:
  - Valget af træningseksempler (=few-shot eksempler ved evaluering af decodermodeller)
  - Valget af testeksempler
- Trænings- og testeksemplerne er bootstrapped 10 gange, hvilket giver et mere troværdigt estimat af det sande gennemsnit
  - Asymptotisk korrekt ved bootstrapsætningen

Funded by the European Union



ScandEval er et robust multilingvalt evalueringsframework

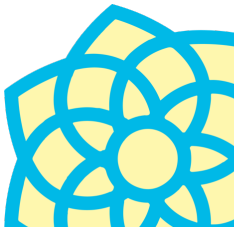
Funded by
the European Union



Multilingvale evalueringer

- I øjeblikket understøttes de fleste germanske sprog:
  - Skandinaviske sprog (dansk, svensk, norsk, islandsk, færøsk)
  - Tysk
  - Hollandsk
  - Engelsk
- Ud over at inkludere evalueringsdatasæt på disse sprog, er alle prompts, der bruges ved evaluering af decodermodeller, også oversat til det givne sprog

Funded by the European Union



Hvilke opgaver er inkluderet?

TrustLLM

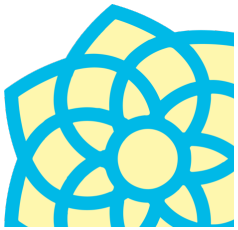
Funded by
the European Union

Opgaver i ScandEval

Sprogforståelsesopgaver (NLU)

1. Tekstklassificering
2. Grammatisk korrekthed
3. Læseforståelse
4. Genkendelse af navngivne enheder

Funded by the European Union

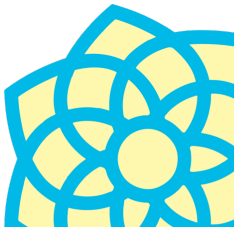


Opgaver i ScandEval

Sproggenereringsopgaver (NLG)

1. Tekstklassificering
2. Grammatisk korrekthed
3. Læseforståelse
4. Genkendelse af navngivne enheder
5. Opsummering
6. Viden
7. Sund fornuft

Funded by the European Union



Online Leaderboards

scandeval.com

ScandEval | ABOUT | DANISH ▼ | SWEDISH ▼ | NORWEGIAN ▼ | ICELANDIC ▼ | FAROESE ▼ | GERMAN ▼ | DUTCH ▼ | ENGLISH ▼

NLU LEADERBOARD

NLG LEADERBOARD

Danish NLG 🇩🇰

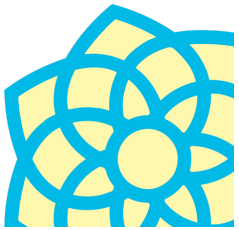
Last updated: 27/05/2024 14:10:07 CET

☐ Include merged models

Model ID	Parameters	Vocabulary Size	Context	Commercial	Speed	Rank ▾	
gpt-4-0613 (few-shot, val)	unknown	100	8192	True	597 ± 197 / 93 ± 33	1.09	
syvai/danskgpt-chat-llama3-70b (few-shot, val)	70554	128	8192	True	1,283 ± 279 / 291 ± 92	1.36	
meta-llama/Meta-Llama-3-70B (few-shot, val)	70554	128	8192	True	312 ± 55 / 177 ± 51	1.47	
gpt-3.5-turbo-0613 (few-shot, val)	unknown	100	4094	True	921 ± 293 / 113 ± 37	1.58	

Download as CSV • Copy embed HTML

Funded by the European Union



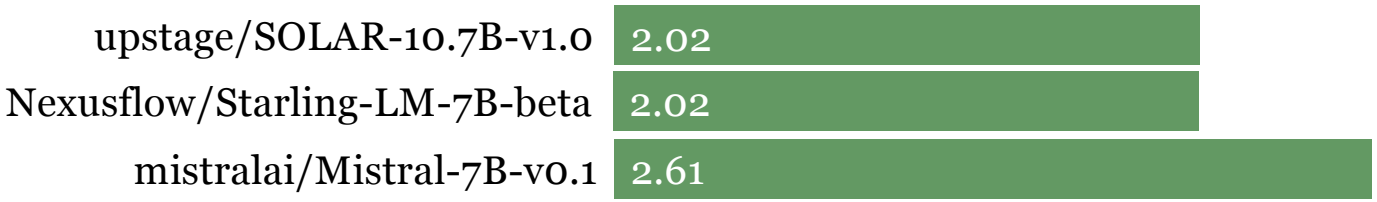
Hvilke sprogmodeller klarer sig så bedst på dansk?

TrustLLM

Funded by
the European Union

Dansk ScandEval Rang

Mindre er bedre



Baseret på model fra:

Mistral AI

Funded by the European Union

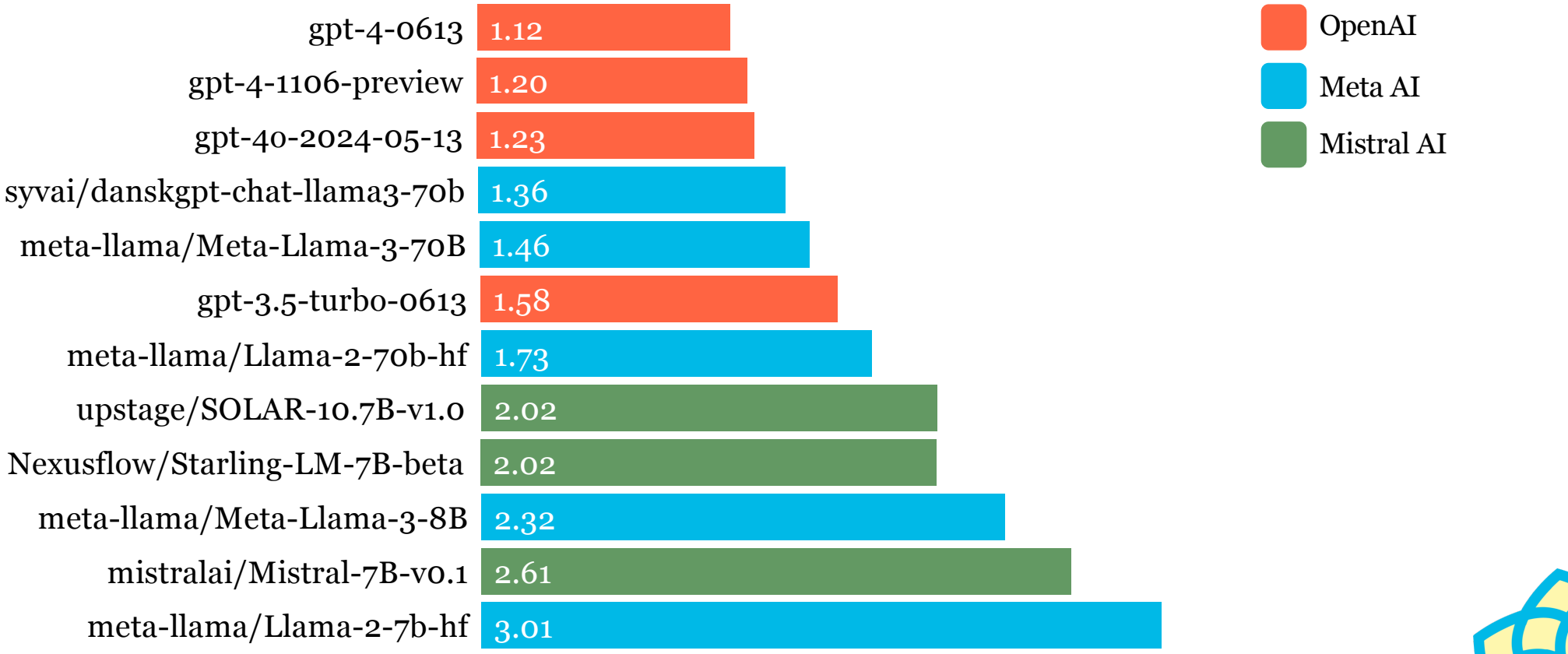
Dansk ScandEval Rang

Mindre er bedre

Funded by the European Union

Dansk ScandEval Rang

Mindre er bedre



Funded by the European Union

Papers

ScandEval NLU benchmark for encodere:

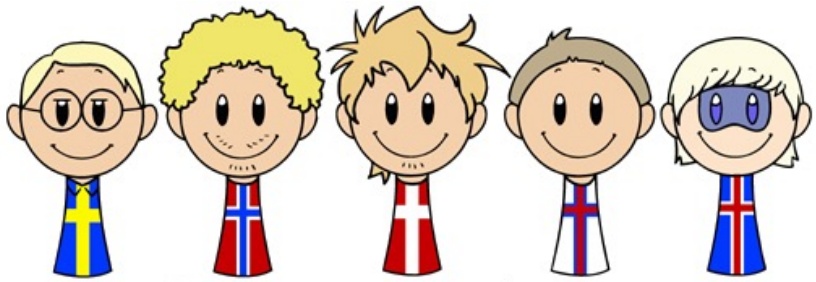
Nielsen, Dan Saattrup. "ScandEval: A Benchmark for Scandinavian Natural Language Processing." Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa). 2023.

ScandEval NLU benchmark for decodere:

Nielsen, Dan Saattrup, Kenneth Enevoldsen, and Peter Schneider-Kamp. "Encoder vs Decoder: Comparative Analysis of Encoder and Decoder Language Models on Multilingual NLU Tasks." arXiv preprint arXiv:2406.13469 (2024).

Funded by the European Union

Tak for jeres opmærksomhed!



```
pip install scandeval[all]
```

TrustLLM

dan.nielsen@alexandra.dk

saattrupdan

Funded by
the European Union