Evaluation of Language Models in the Generative Era

Sprogteknologisk Konference 2024

November 28, 2024

Dan Saattrup Nielsen

Senior AI Specialist @Alexandra Institute

ALEXANDRA INSTITUTTET

TrustLLM

Funded by the European Union

How can we evaluate LLMs?

TrustLLM

Funded by the European Union

Four Main Approaches

Vibe Check

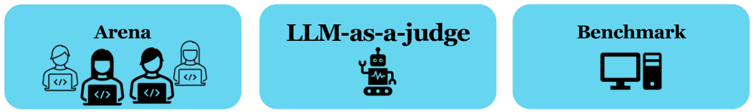
LLM-as-a-judge

Arena

Benchmark

Funded by
the European Union

Four Main Approaches



Vibe Check

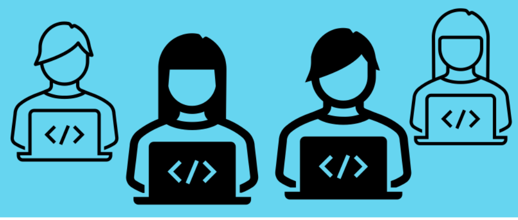
Pros	Cons
Typically gives a good ballpark figure	Can only evaluate instruction-tuned models
Relevant to use cases the user cares about	Does not generalise to other tasks
Very cheap and fast	Not objective, has to be redone for each person

Funded by the European Union

Four Main Approaches

LLM-as-a-judge | Benchmark | Vibe Check

Arena



Pros	**Cons**
Relatively objective measure when a critical mass of volunteers have voted	Can only evaluate instruction-tuned models
More relevant to the user's use cases*	Time-consuming and costly to set up and evaluate
	Requires *many* volunteers to evaluate

* Depends on the types of questions and/or users contributing

Funded by the European Union

Four Main Approaches

Benchmark

Vibe Check

Arena

LLM-as-a-judge

Pros	Cons
Allows measuring more complex phenomena	Can only evaluate instruction-tuned models
Cheap to set up and evaluate	The evaluator LLM can be biased [1]
Measure that only has to be done once for each model	Requires the existence of a very good LLM in the given language

[1] Stureborg et al. arXiv preprint arXiv:2405.01724 (2024)

Four Main Approaches

Vibe Check

Arena

LLM-as-a-judge

Benchmark

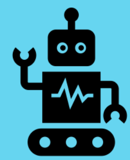
Pros	Cons
Gives a precise measure of performance	Does not necessarily generalise to other types of tasks
Objective measure that only must be done once for each model	Creating evaluation datasets is costly
Can evaluate all types of language models	Models can train on public test sets

Funded by the European Union

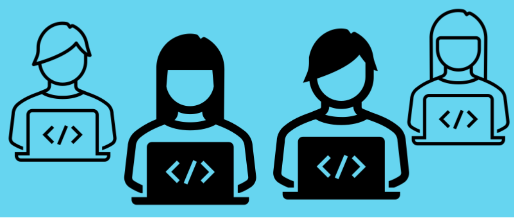
Four Main Approaches

Vibe Check

LLM-as-a-judge



Arena



Benchmark

Funded by
the European Union

What is ScandEval?

TrustLLM

Funded by
the European Union

ScandEval is a robust multilingual benchmarking framework

Funded by
the European Union

ScandEval is a robust multilingual benchmarking framework

Funded by
the European Union

Language Model Benchmarking Framework

- Enables evaluation of implicit language understanding and generation capabilities of language models
- Allows evaluation of *both* encoders through finetuning, and decoders through few-shot evaluation
  - It has been shown that there is a direct correspondence between few-shot evaluation and finetuning [2]
  - This thus allows us to compare encoders with decoders directly

[2] Stureborg et al. arXiv preprint arXiv:2405.01724 (2024)

Funded by the European Union

Language Model Benchmarking Framework

- A large focus of the framework is ease of use
- The framework can simply be installed:

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

- Models can easily be evaluated:

```
$ scandeval --model <model-id> [--language da]
```

- Supports models on the Hugging Face Hub, local models and OpenAI models

Funded by
the European Union

ScandEval is a robust multilingual benchmarking framework

Funded by
the European Union

Evaluation Robustness

- When evaluating models, there are several sources of noise in the evaluation result:
  - The choice of training examples (=few-shot examples when evaluating decoder models)
  - The choice of test examples
  - The stochastic elements (stochastic gradient descent when evaluating encoders, sampling when evaluating decoders)
- The training and test examples are bootstrapped 10 times, yielding a more reliable estimation of the true mean
  - Asymptotically correct by the bootstrap theorem
- We enforce that the stochastic elements are deterministic

Funded by the European Union

ScandEval is a robust multilingual benchmarking framework

TrustLLM

Funded by the European Union

Which Tasks are Included?

TrustLLM

Funded by
the European Union

Tasks in ScandEval

Natural Language Understanding (NLU) Tasks

1. Sentiment classification
2. Linguistic acceptability
3. Reading comprehension
4. Named entity recognition

Funded by the European Union

Tasks in ScandEval

Natural Language Generation (NLU) Tasks

1. Sentiment classification
2. Linguistic acceptability
3. Reading comprehension
4. Named entity recognition
5. Summarisation
6. World knowledge
7. Common-sense reasoning

Funded by
the European Union

Leaderboards

TrustLLM

Funded by
the European Union

Online Leaderboards
`scandeval.com`

ScandEval ABOUT DANISH ▼ SWEDISH ▼ NORWEGIAN ▼ ICELANDIC ▼ FAROESE ▼ GERMAN ▼ ENGLISH ▼ MIXED ▼

NLU LEADERBOARD
NLG LEADERBOARD

Danish NL

Rank Score computed (roughly) as 1 + number of standard deviations to the best model, across all datasets

Last updated: 23/06/2024 10:08:18 CET

☐ Include merged models

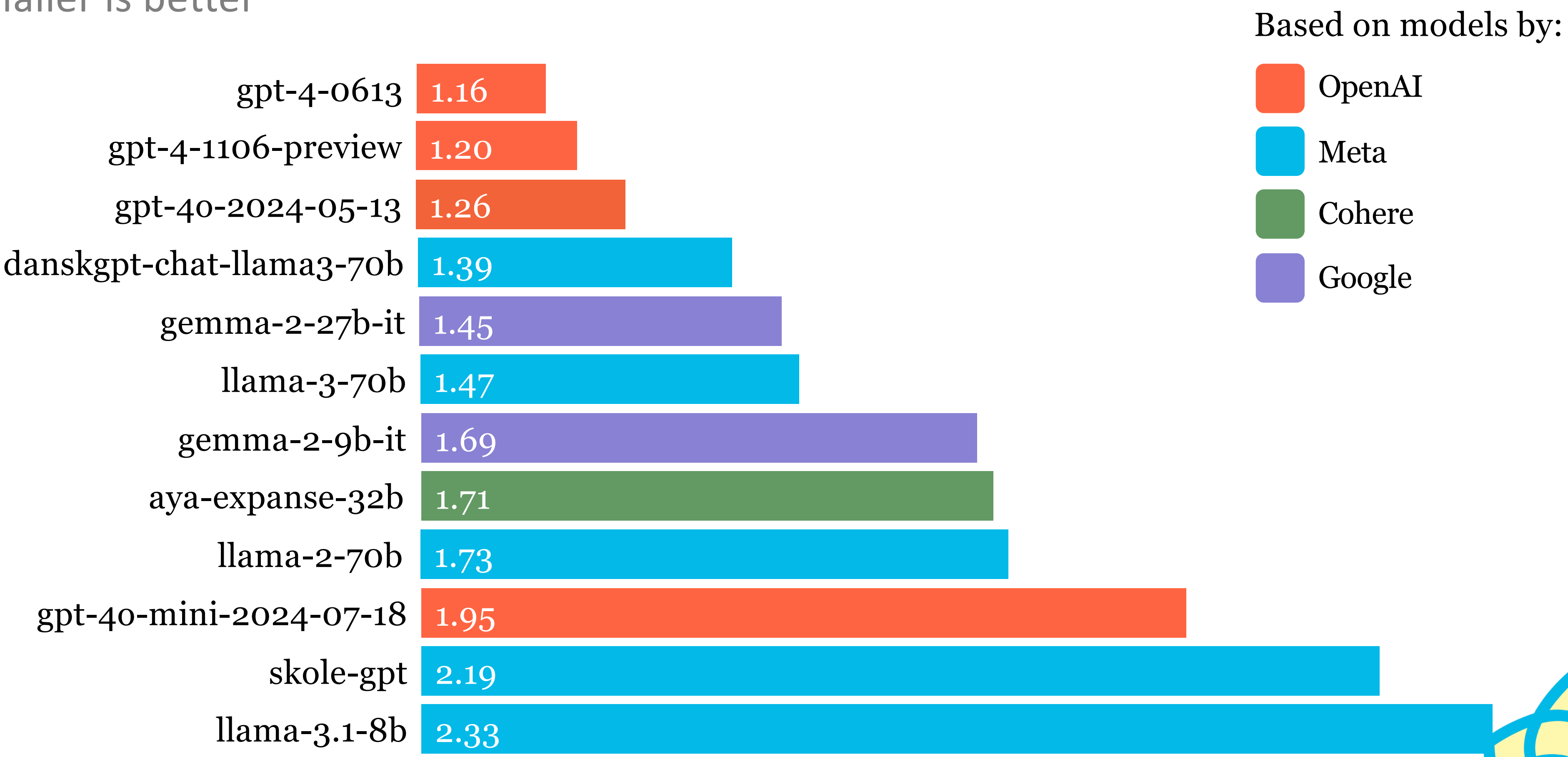
Model ID	Parameters	Vocabulary Size	Context	Commercial	Speed	Rank	DANSK
gpt-4-0613 (few-shot, val)	unknown	100	8192	True	597 ± 197 / 93 ± 33	1.12	64.94 ± 1.96 / 45.76 ±
gpt-4-1106-preview (few-shot, val)	unknown	100	127999	True	576 ± 221 / 81 ± 28	1.20	66.80 ± 3.01 / 45.69 ±
gpt-4o-2024-05-13 (few-shot, val)	unknown	200	127999	True	916 ± 329 / 114 ± 38	1.24	71.15 ± 2.89 / 52.24 ±

Download as CSV • Copy embed HTML

Funded by the European Union

Excerpt of Danish ScandEval Scores

Smaller is better



Funded by
the European Union

Encoders vs Decoders

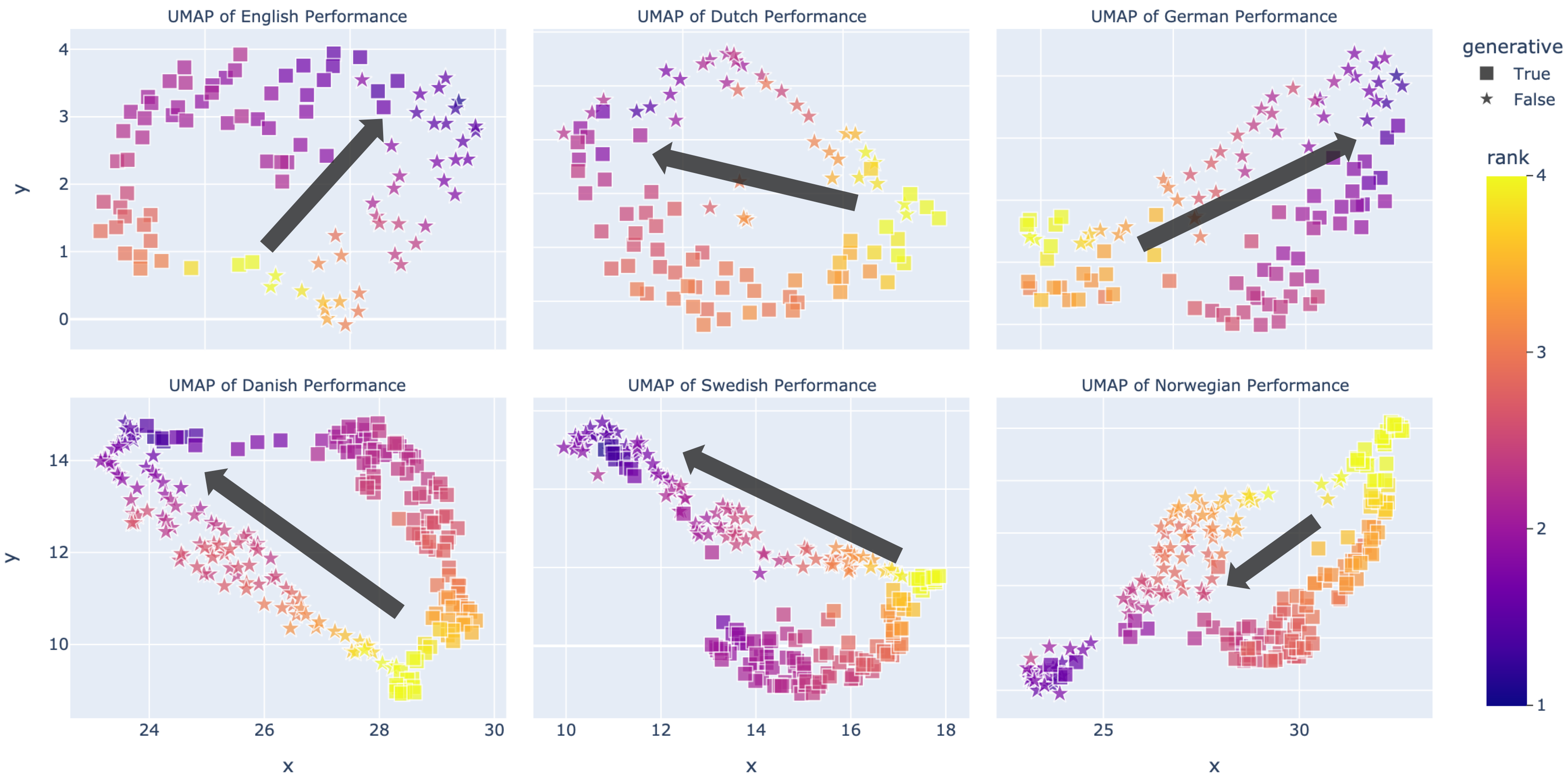
Funded by
the European Union

Encoders vs Decoders

- Do encoders and decoders ”learn” things differently?
- Experiment:
  - Take all raw NLU results from ScandEval leaderboards
    - 4 scores per model and language
  - Mark encoders/decoders as well as their rank
  - UMAP dimensionality reduction to 2 dimensions
  - Visualise

Funded by the European Union

Encoders vs Decoders



Funded by the European Union

Papers

ScandEval NLU benchmark for encoders:

Nielsen, Dan. *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. 2023

ScandEval NLU benchmark for decoders:

Nielsen, Dan and Kenneth Enevoldsen and Peter Schneider-Kamp. arXiv preprint arXiv: 2406.13469 (2024).

Funded by
the European Union

Thanks for your attention!

ScandEval

TrustLLM

Code base:
github.com/ScandEval/ScandEval

Funded by the European Union