Scandivian **E**mbedding **B**enchmark

Evaluering og benchmarking af sprogmodeller

Kenneth Enevoldsen | 2024

Hvorfor evaluere Embeddings

Brug af embedding:

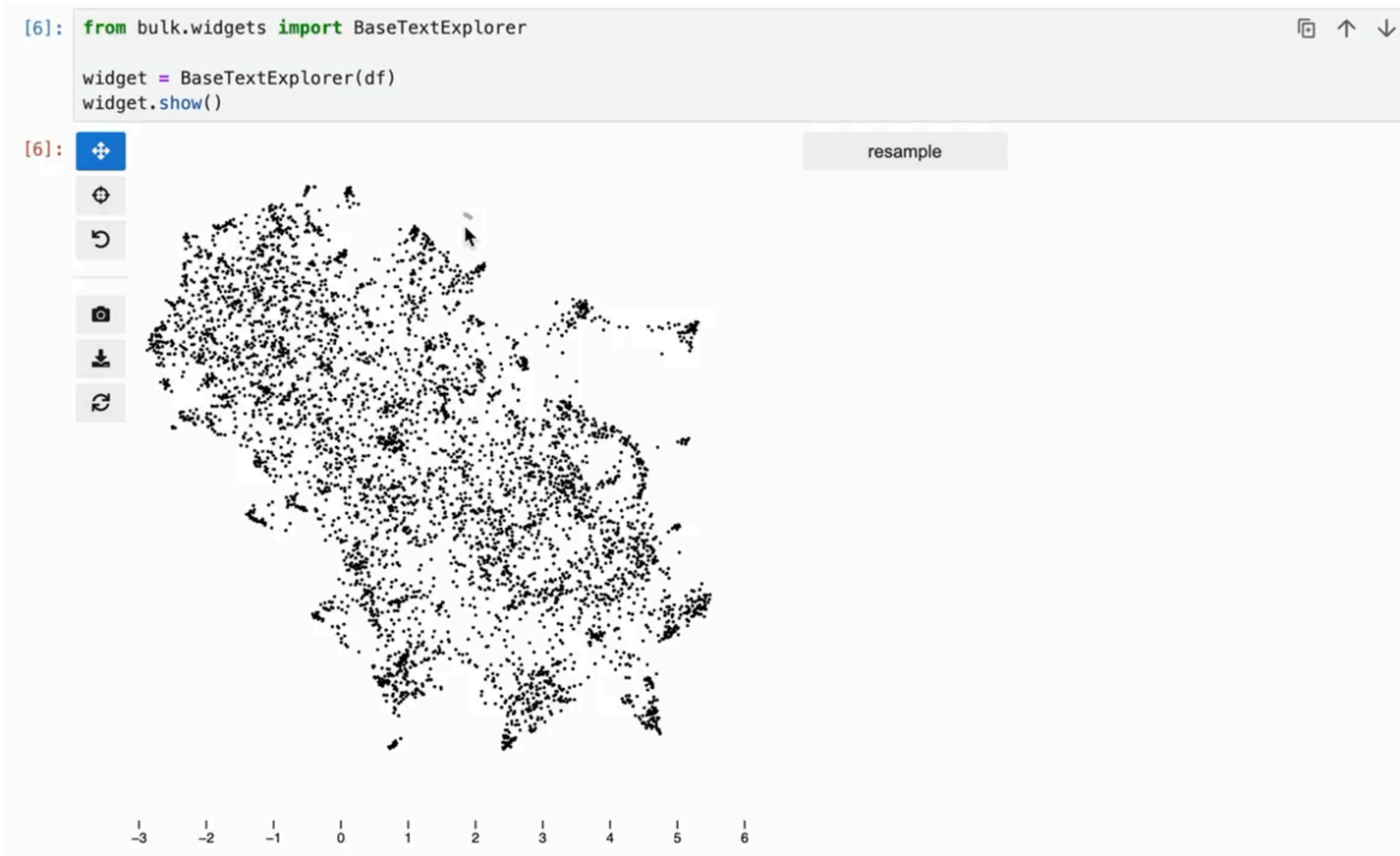
- **Semantisk søgning**:
  - Søgning
  - Retrieval augmented generation
- **Semantiske representationer**
  - Classification (e.g. setfit)
  - Data cleaning
  - Bulk labelling
  - Input til andre modeller
- ...

Hvorfor evaluere Embeddings

Brug af embedding:

- **Semantisk søgning**:
  - Søgning
  - Retrieval augmented generation
- **Semantiske representationer**
  - Classification (e.g. setfit)
  - Data cleaning
  - **Bulk labelling**
  - Input til andre modeller

...



Sources & Notes

* https://github.com/koaning/bulk

CENTER FOR HUMANITIES COMPUTING

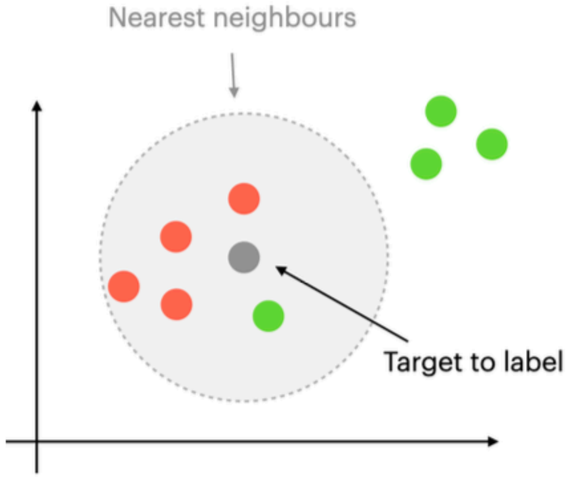
Evaluering

Inspiration: MTEB

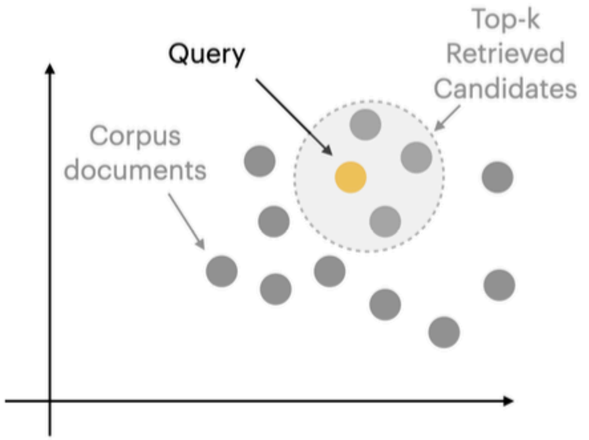
Fokus:

Integration →

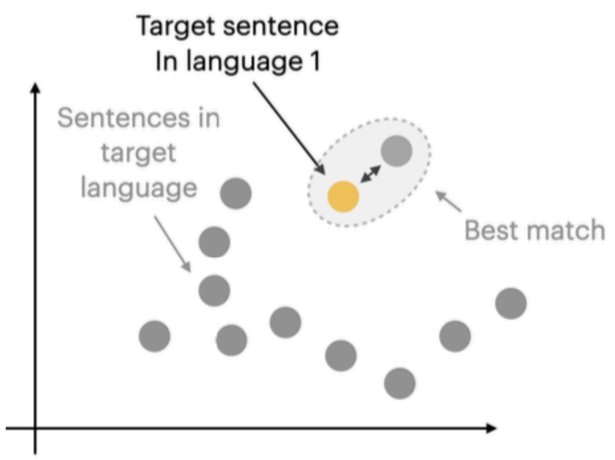
- Lettere at fortolke
- Multilingual benchmark*
- Fælles forbedringer



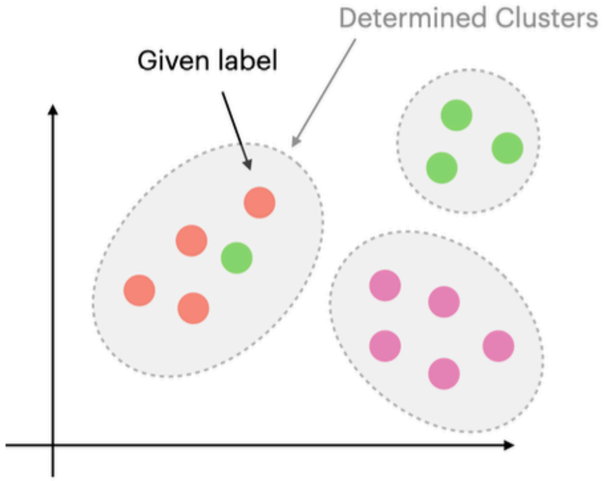
(a) **Classification**: The inputs consist of documents and labels. Using, for example, a kNN classifier, the goal is to determine the correct label for each document.



(b) **Retrieval**: The dataset consists of a corpus and queries. The goal is to find the correct documents in the corpus for a given query.



(c) **Bitext Mining**: Inputs are two sets of documents from two different languages. For each document in the first set, the goal is to find the correct match in the second set.



(d) **Clustering**: The dataset consists of documents attached with labels. The goal is to correctly cluster the documents according to their labels.

Sources
& Notes

* Der kommer snart et paper på dette

Scandinavian Embedding Benchmark

- Få gode danske dataset
- Flere brugere ↔ **højere kvalitet**
- Større **incitament** til at teste på dansk
- Mere diverse tests

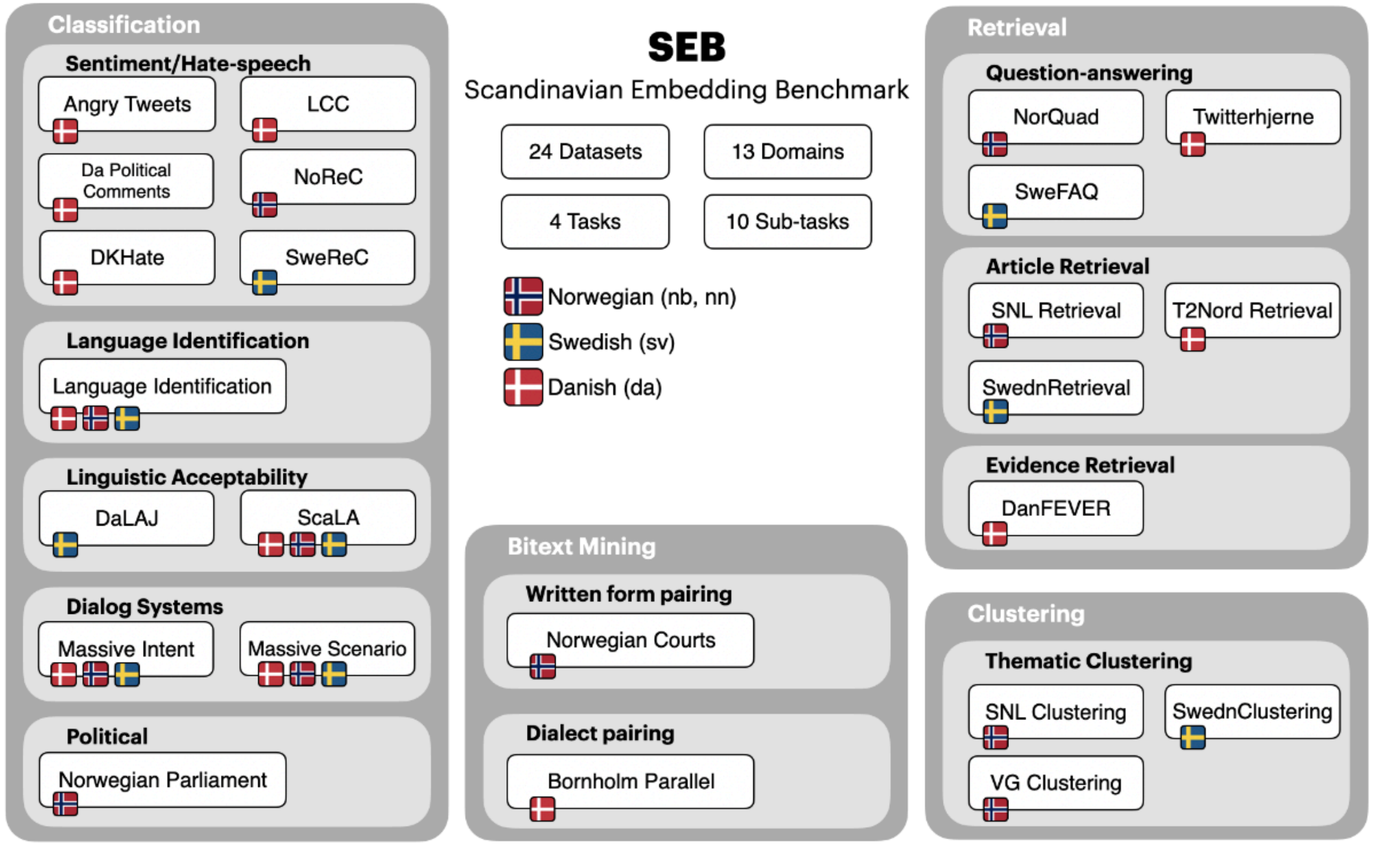


Figure 1: An overview of the tasks and datasets in SEB. Flags denote the languages of the datasets.

Coverage

- Mangler:
  - Medical
  - (Legal)
  - Social Media
  - Historical sources
  - Social Media

	Across	Danish	Norwegian Bokmål	Norwegian Nynorsk	Swedish
Domain					
Academic	(✓)	(✓)			
Bible					
Blog					
Fiction	✓	✓	✓	✓	✓
Government	✓	✓	✓	✓	✓
Legal	✓	(✓)	✓	✓	
Medical					
News	✓	✓	✓		✓
Non-Fiction	✓	✓	✓		✓
Poetry	(✓)	(✓)			
Reviews	✓		✓		✓
Social	✓	✓			✓
Spoken	✓	✓	✓		✓
Wiki	✓	✓	✓	✓	✓
Web	✓	✓			✓

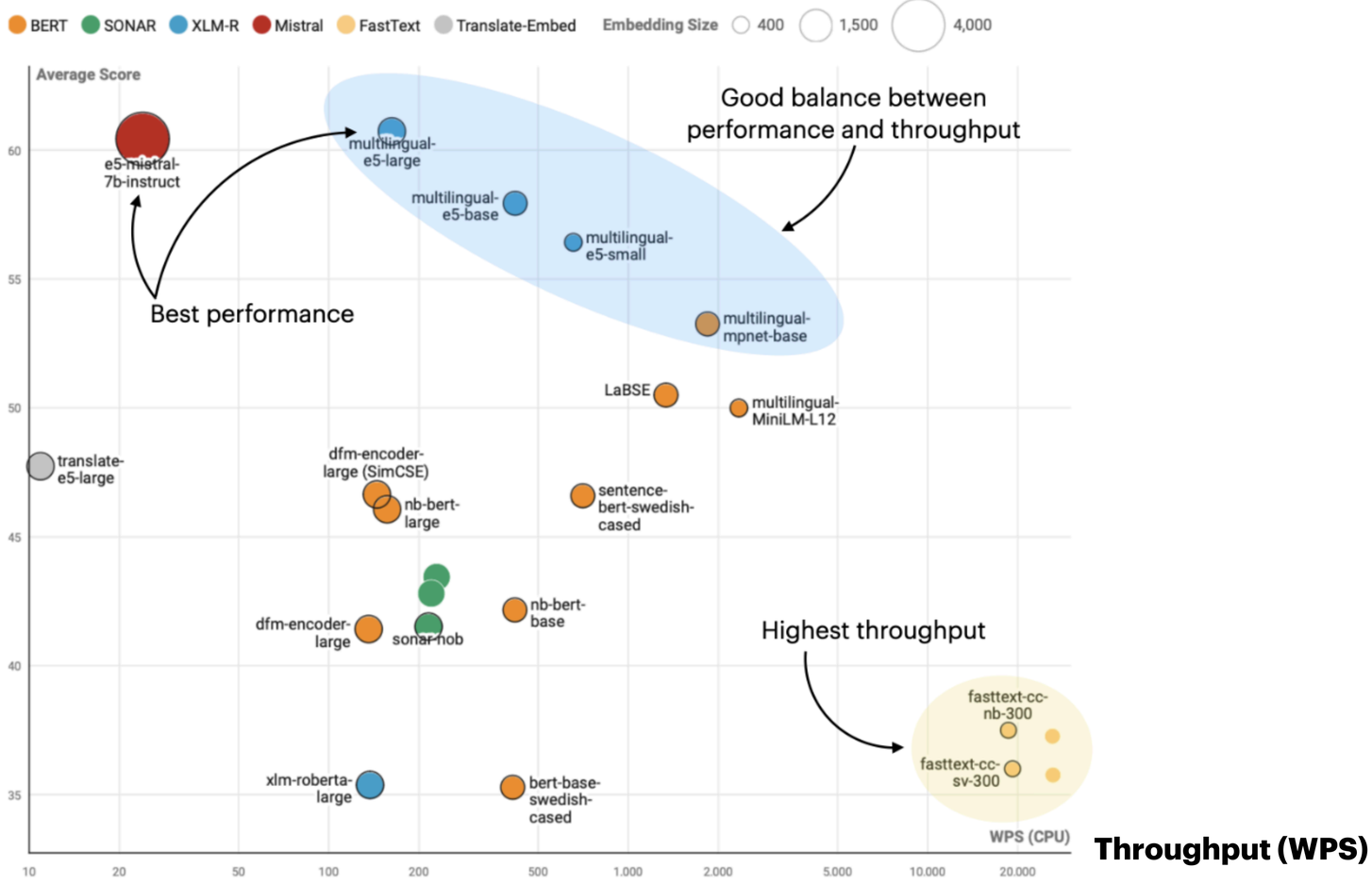
Sources & Notes

https://kennethenevoldsen.github.io/scandinavian-embedding-benchmark/domains/

CENTER FOR HUMANITIES COMPUTING

Relevant Performance*

Mean performance across tasks



Throughput (WPS)

Sources & Notes

https://kennethenevoldsen.github.io/scandinavian-embedding-benchmark/speed_performance/

CENTER FOR HUMANITIES COMPUTING

Performance by **Language** and **Task Type**

Multilingual Models klarer sig bedre end monolingual models

Oversæt-og-embed fungerer dårligt

Private APIer klarer sig (meget) bedre på retrieval

	Avg.	Task-Type				Language			
		Bitext	Class.	Clust.	Retr.	da	nb	nn	sv
Num. Datasets (→)	24	2	12	3	7	12	11	3	9
Self-Supervised Models									
dfm-encoder-large	41.4	46.8	56.5	26.9	20.1	47.7	47.4	72.5	43.7
+ SimCSE	**46.6**	50.9	58.4	26.9	**33.7**	**52.2**	51.3	74.3	42.0
xlm-roberta-large	35.3	19.1	54.6	28.1	10.0	39.6	41.3	58.0	44.5
nb-bert-large	46.0	47.3	**59.3**	**35.7**	27.3	46.8	**57.2**	**80.4**	**50.2**
nb-bert-base	42.1	**51.0**	57.0	31.8	18.4	43.6	53.0	79.2	47.7
bert-base-swedish	35.2	39.1	49.7	26.2	13.2	34.0	41.1	62.2	43.6
Supervised Models									
e5-mistral-7b-instruct	60.4	**70.8**	61.7	35.7	66.0	**61.7**	62.9	68.8	60.4
multilingual-e5-large	**60.7**	60.1	**62.5**	34.2	**69.1**	61.1	**63.1**	**73.9**	**62.8**
multilingual-e5-base	57.9	61.4	60.1	34.0	63.5	58.6	60.9	72.0	58.5
multilingual-e5-small	56.4	61.6	58.1	**36.9**	60.3	56.5	58.9	69.5	57.1
translate-e5-large	47.7	50.7	54.7	27.3	43.4	49.0	50.1	59.2	59.2
Embedding APIs									
text-embedding-3-large	**65.0**	**68.8**	63.5	38.7	**77.9**	**63.7**	**69.0**	**74.7**	**65.5**
text-embedding-3-small	61.0	66.7	59.7	38.3	71.3	59.7	64.7	70.2	60.4
embed-multilingual-v3.0	64.1	64.2	**63.6**	**40.2**	75.2	62.6	68.5	74.1	64.3

Sources & Notes

Table is reduced for clarity

Since then an open-source model was released beating the commercial APIs (multilingual-e5-large-instruct) on all expect retrieval

https://kennethenevoldsen.github.io/scandinavian-embedding-benchmark/

CENTER FOR HUMANITIES COMPUTING

Design Considerations

- **Hurtig**: ~10 min at evaluere multilingual-e5-large
    - Kan køre på en laptop*
    - Public cache: ingen grund til at genkøre modeller

```
❯ seb run -m XLMRoberta-en-da-sv-nb,multilingual-e5-large,multilingual-e5-large-instruct -t DanFEVER
INFO:seb.cli.run:Model registered in SEB. Loading from registry.
INFO:seb.cli.run:Model registered in SEB. Loading from registry.
INFO:seb.cli.run:Model registered in SEB. Loading from registry.
Running multilingual-e5-large-instruct: 100%|| 3/3 [00:00<00:00, 881.03it/s]
Running bge-m3: 100%|| 46/46 [00:00<00:00, 1019.16it/s]
```

Benchmark Results

Rank	Model	Average Score	Average Rank	DanFEVER
1	embed-multilingual-v3.0	0.41	1.00	40.99
2	**NEW: multilingual-e5-large**	**0.41**	**2.00**	**40.54**
3	multilingual-e5-base	0.40	3.00	40.09
4	text-embedding-3-large	0.40	4.00	39.61
5	**NEW: multilingual-e5-large-instruct**	**0.40**	**5.00**	**39.52**
6	voyage-multilingual-2	0.39	6.00	39.42
15	e5-large	0.36	15.00	36.46
16	**NEW: XLMRoberta-en-da-sv-nb**	**0.35**	**16.00**	**35.34**
17	e5-base	0.35	17.00	35.32

* for små (<1B) modeller

Design Considerations

- **Hurtig**: ~10 min at evaluere multilingual-e5-large
- **Nem at udvide**: <1h at tilføje nyes task og evaluerer alle modellerne
  - Nemt for private virksomheder at udvide til intern brug

```
❯ seb run -t MuniIntent -m multilingual-e5-large-instruct,multilingual-e5-large,paraphrase-multilingual-m
pnet-base-v2 -c src/experimental_tasks/muni_intent_classification.py
```

Privat datasæt

Benchmark Results

Rank	Model	Average Score	Average Rank	MuniIntent
	Anonymiseret*			

Sources & Notes

*Dette er kun scores på datasættet.

CENTER FOR HUMANITIES COMPUTING

Design Considerations

- **Hurtig**: ~10 min at evaluere multilingual-e5-large
- **Nem at udvide**: <1h at tilføje nyes task og evaluerer alle modellerne
- **Reproducérbar**:
  - Hvordan blev modellen kørt
  - Hvad var resultaterne
  - Hvilket version and SEB blev brugt

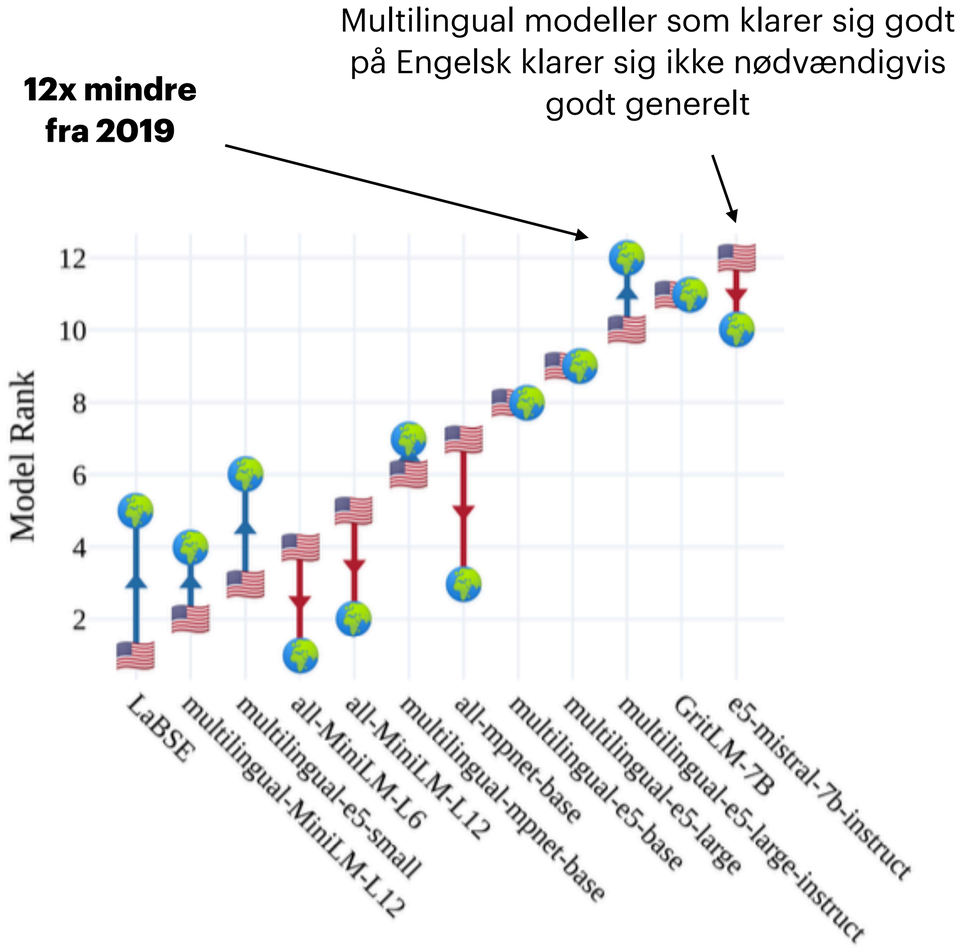
Design Considerations

- **Hurtig**: ~10 min at evaluere multilingual-e5-large
- **Nem at udvide**: <1h at tilføje nyes task og evaluerer alle modellerne
- **Reproducérbar**:
  - Hvordan blev modellen kørt
  - Hvad var resultaterne
  - Hvilket version and SEB blev brugt
- **Integration** med standard benchmarks
  - Nemt for andre at teste på dansk/skandinavisk

Hvordan får vi internationale forskere til at lave modeller til dansk?

Multilingual Embedding Benchmark

- >1000 sprog, >50 forfattere, >20
- Ekstrem **diversitet** i typen af opgaver
- 10-100x hurtigere køre sammenlignet med MTEB



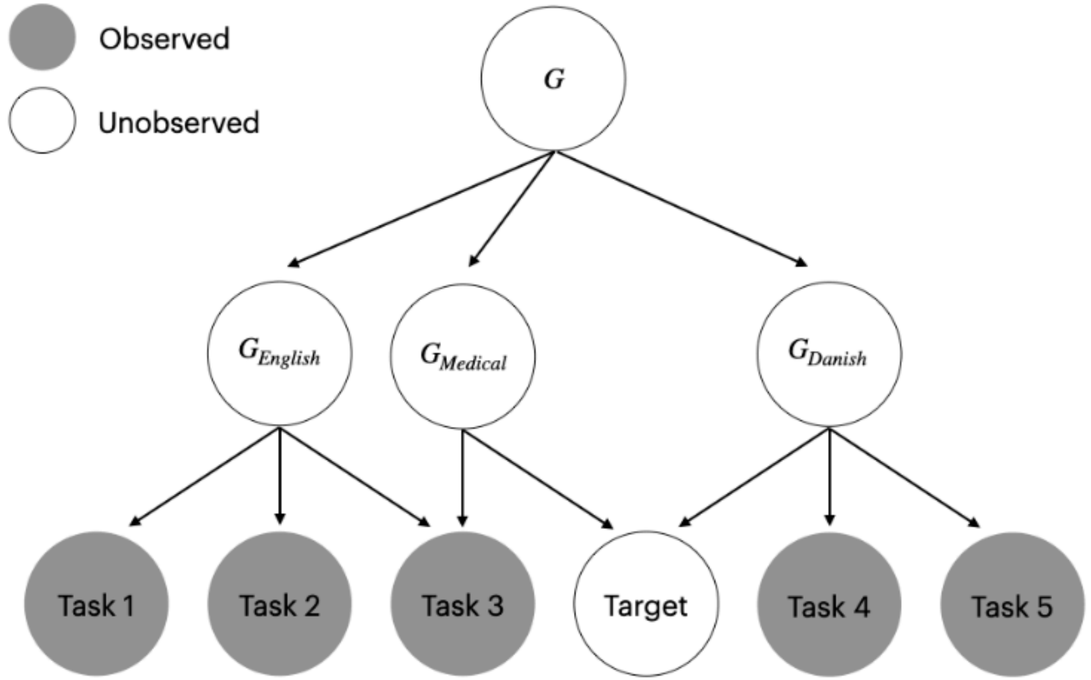
Sources & Notes

These are preliminary results

CENTER FOR HUMANITIES COMPUTING

Kan vi estimere den bedste model?*

- Case:
  - **Retrieval** på **medicinske danske** tekster med **billeder**
- Mulighed 1:
  - Skaf et datasæt
    - GDPR, Legal —> ~2 års arbejde
- Mulighed 2:
  - Kan vi estimere dette fra eksisterende kilder?
  - —> Hvilket datasæt giver os mest information



Sources & Notes

*Ongoing work in progress, feel free to reach out

CENTER FOR HUMANITIES COMPUTING

Næste version af SEB

- Skjulte private datasæt
  — tjek for overfitting
- Flere højkvalitets datasæt
  — erstat lav-kvalitets datasæt
- Long-document Retrieval
- Flere domæner
- Lavresourcesprog
  - Islandsk, færøsk (grønlandsk, finsk)