Danish Foundation Models

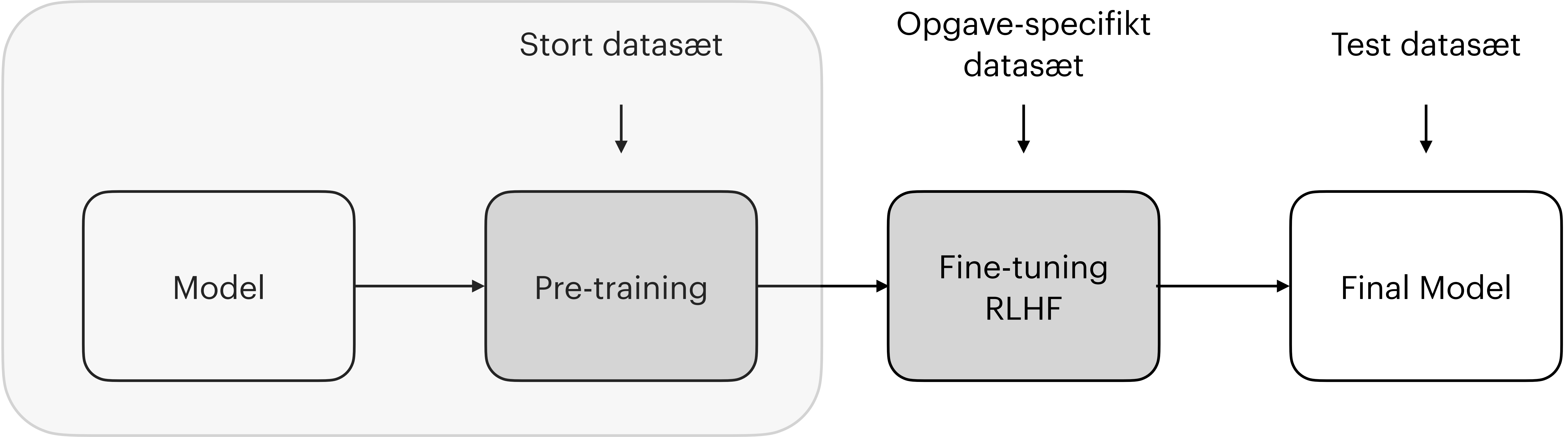
Validerede sprogmodeller til dansk

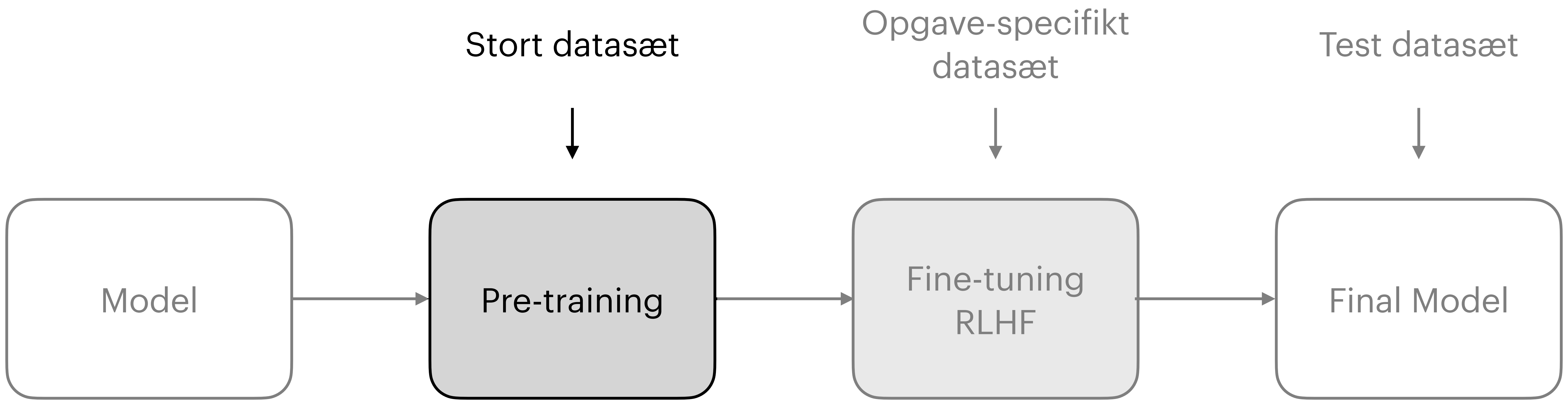
Kenneth Enevoldsen & Lasse Hansen | November 20, 2023

CENTER FOR
HUMANITIES
COMPUTING

Introduktion

Hvad er en foundation model?



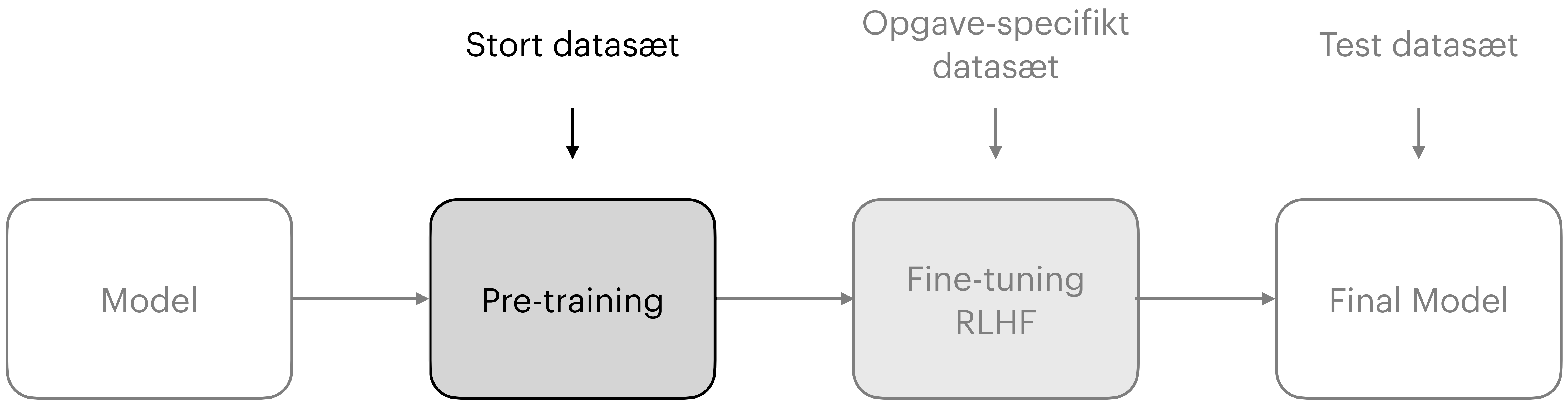


Maskering The quick [MASK] fox jumps over the [MASK] dog

Prædiktion The quick brown fox jumps over the lazy dog

Sources & Notes

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423



Stort datasæt

Opgave-specifikt datasæt

Test datasæt

Model

Pre-training

Fine-tuning RLHF

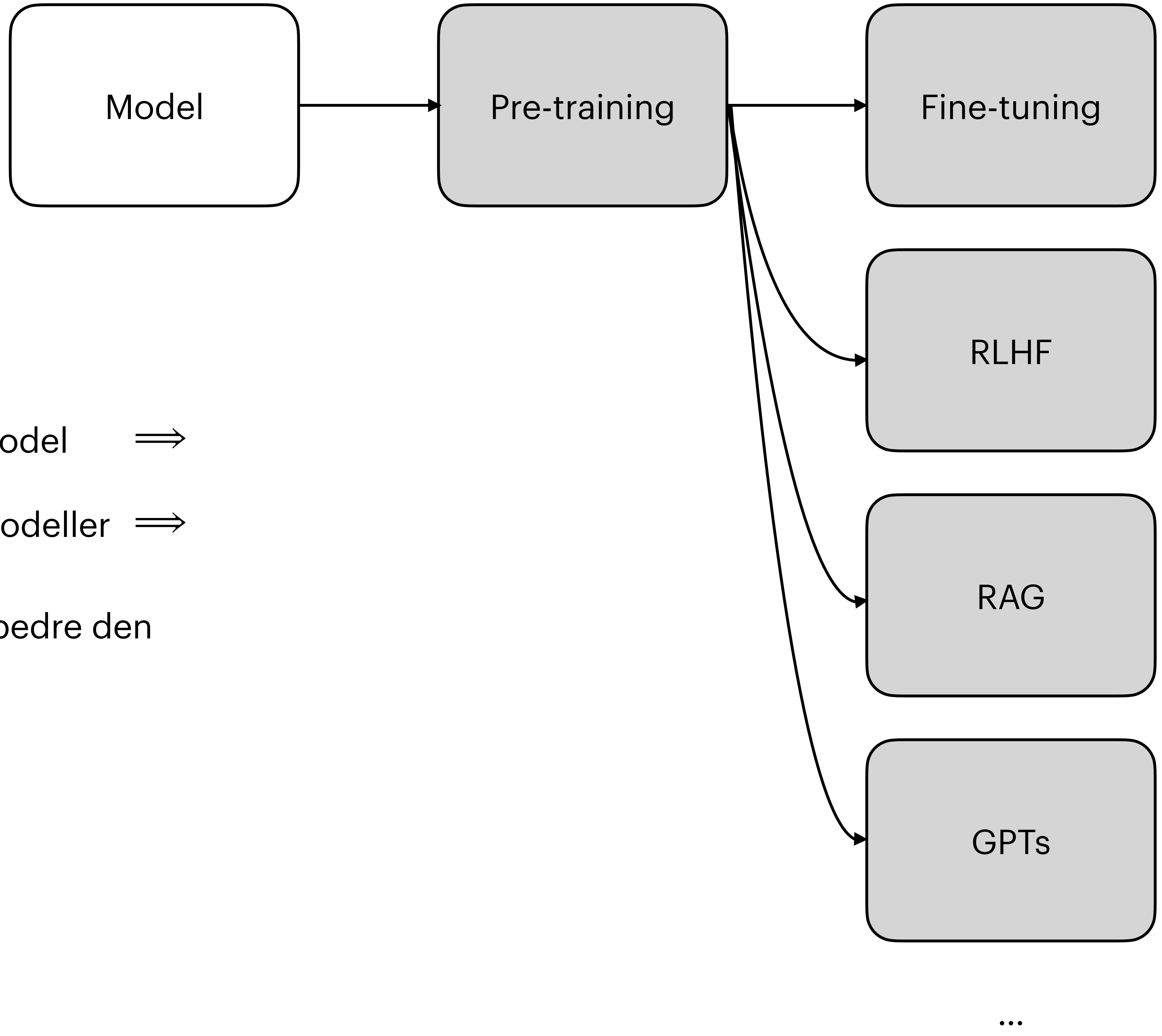
Final Model

Kontekst `Margrethe 2. er dronning af ____`

Prædiktion `Danmark`

Sources & Notes

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.



Bedre præ-trænet model ⟹

Bedre adapterede modeller ⟹

Stort fokus på at forbedre den præ-trænede model

Motivation

Hvorfor danske sprogmodeller?

Tokenisation

Lav genbrugelighed

```
antiradikaliseringskonference
['anti', ‘—radi’, ‘—kal’, ‘—isering’, ‘—skon’, ‘—ference’]
```

Høj genbrugelighed

```
Ferskvandsfisk
['fersk', ‘—vand', ‘—sf', ‘—isk’]
```

Færre tokens
—> højere effektivitet

```
skolehjemsamtaler
['skole', ‘—hjem', ‘—sam', ‘—taler']
```

Sources
& Notes

The Unigram tokeniser have also been argued to be a better model in general, even for English:

Bostrom, Kaj, and Greg Durrett. “Byte Pair Encoding Is Suboptimal for Language Model Pretraining.” In Findings of the Association for Computational Linguistics: EMNLP 2020, 4617–24. Online: Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.414.

Trecca, F., Bleses, D., Madsen, T. O., & Christiansen, M. H. (2018). Does sound structure affect word learning? An eye-tracking study of Danish learning toddlers. Journal of Experimental Child Psychology, 167, 180-203.

Hvorfor danske sprogmodeller?

Tokenisation

Lav genbrugelighed

```
antiradikaliseringskonference
['anti', ‘—radi’, ‘—kal’, ‘—isering’, ‘—skon’, ‘—ference’]
['anti', '—radikal',        '—isering', ‘—s', ‘—konference’]
```

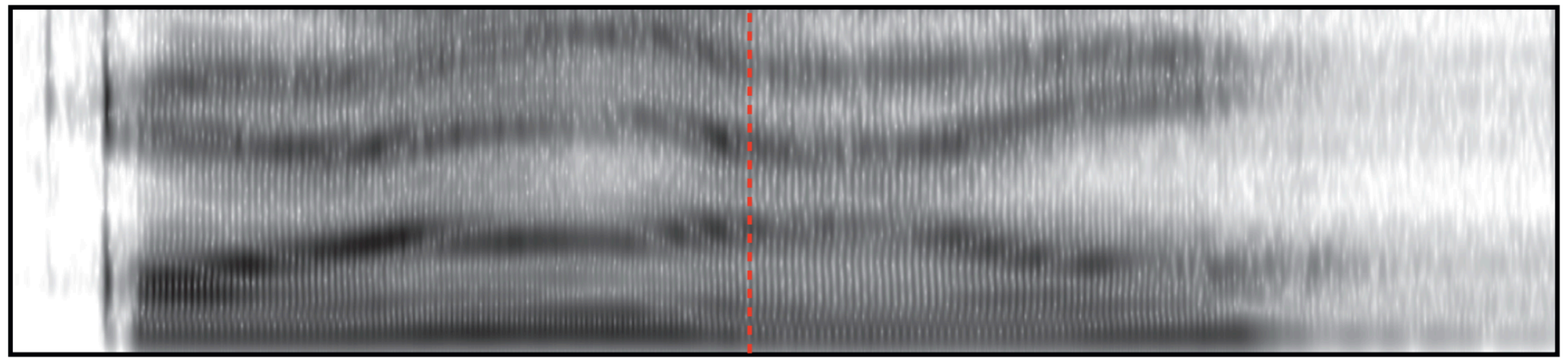
Høj genbrugelighed

```
Ferskvandsfisk
['fersk', ‘—vand', ‘—sf', ‘—isk’]
['fersk', ‘—vands',        '—fisk']
```

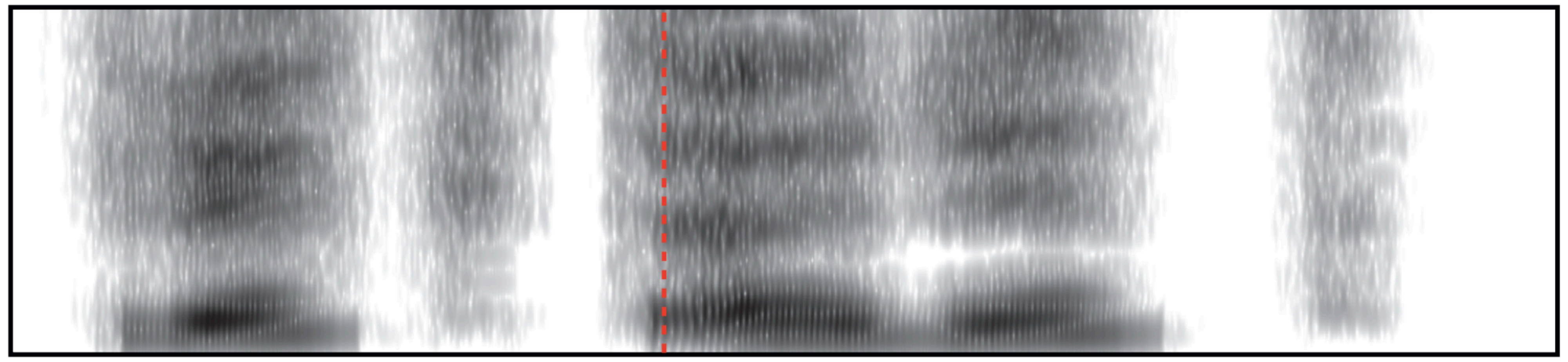
```
skolehjemsamtaler
['skole', ‘—hjem', ‘—sam', ‘—taler']
['skole', '—hjem', '—samtale', '—r']
```

Færre tokens
—> højere effektivitet

Udtale



Danish sentence: Røget ørred



Norwegian sentence: Røkt ørret

Sources & Notes

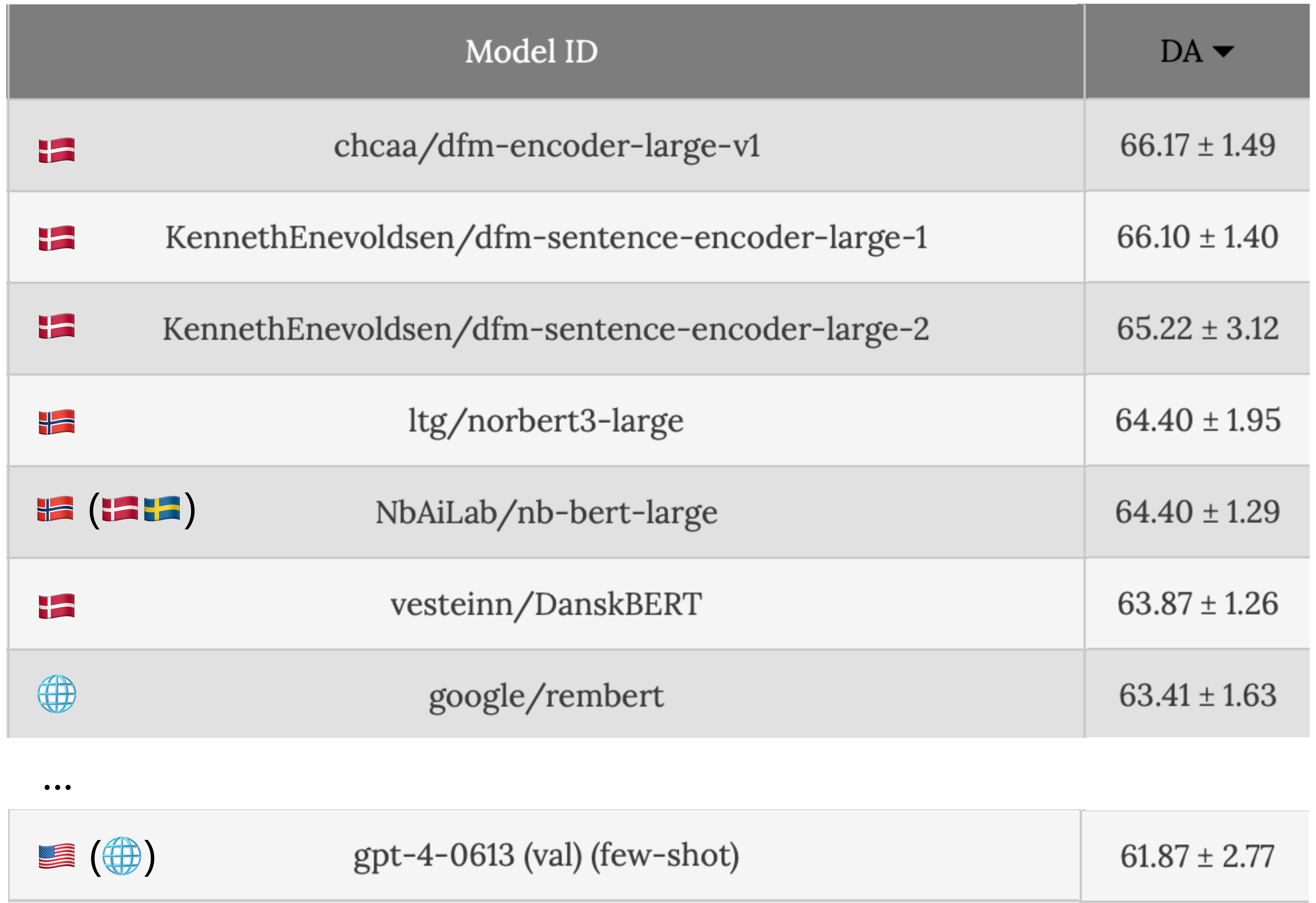
The Unigram tokeniser have also been argued to be a better model in general, even for English:

Bostrom, Kaj, and Greg Durrett. “Byte Pair Encoding Is Suboptimal for Language Model Pretraining.” In Findings of the Association for Computational Linguistics: EMNLP 2020, 4617–24. Online: Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.414.

Trecca, F., Bleses, D., Madsen, T. O., & Christiansen, M. H. (2018). Does sound structure affect word learning? An eye-tracking study of Danish learning toddlers. Journal of Experimental Child Psychology, 167, 180-203.

CENTER FOR HUMANITIES COMPUTING

Monolinguale modeller klarer sig godt

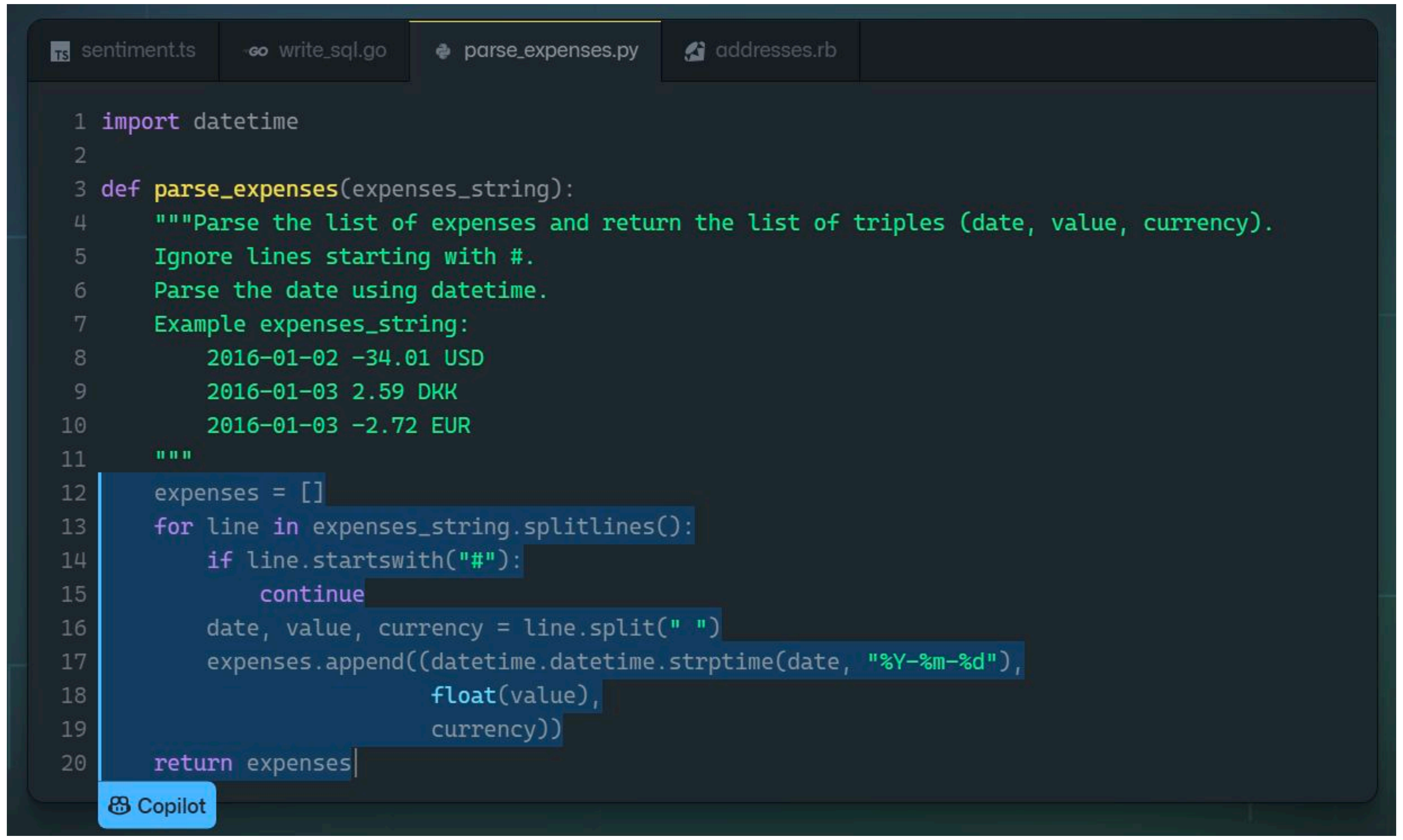


	Model ID	DA ▼
🇩🇰	chcaa/dfm-encoder-large-v1	66.17 ± 1.49
🇩🇰	KennethEnevoldsen/dfm-sentence-encoder-large-1	66.10 ± 1.40
🇩🇰	KennethEnevoldsen/dfm-sentence-encoder-large-2	65.22 ± 3.12
🇳🇴	ltg/norbert3-large	64.40 ± 1.95
🇳🇴 (🇩🇰🇸🇪)	NbAiLab/nb-bert-large	64.40 ± 1.29
🇩🇰	vesteinn/DanskBERT	63.87 ± 1.26
🌐	google/rembert	63.41 ± 1.63
...		
🇺🇸 (🌐)	gpt-4-0613 (val) (few-shot)	61.87 ± 2.77

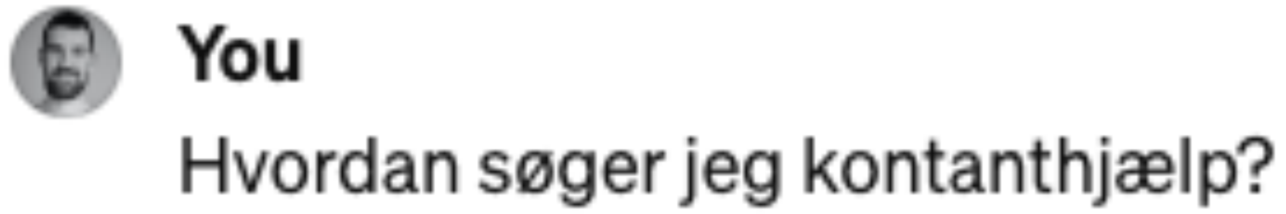
Sources
& Notes

https://scandeval.github.io/mainland-nlu-benchmark/

Monolingualle modeller skal løfte andre opgaver



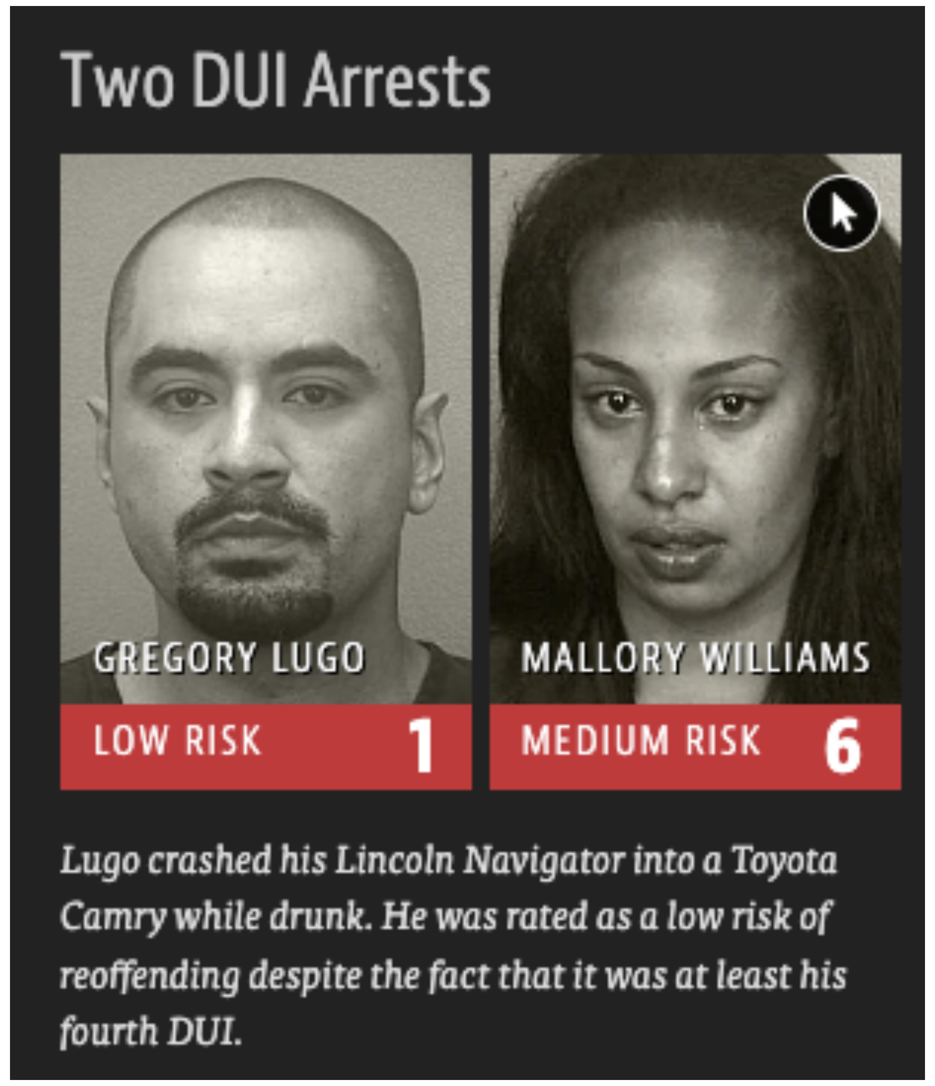
```python
import datetime

def parse_expenses(expenses_string):
    """Parse the list of expenses and return the list of triples (date, value, currency).
    Ignore lines starting with #.
    Parse the date using datetime.
    Example expenses_string:
        2016-01-02 -34.01 USD
        2016-01-03 2.59 DKK
        2016-01-03 -2.72 EUR
    """
    expenses = []
    for line in expenses_string.splitlines():
        if line.startswith("#"):
            continue
        date, value, currency = line.split(" ")
        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
                         float(value),
                         currency))
    return expenses
```



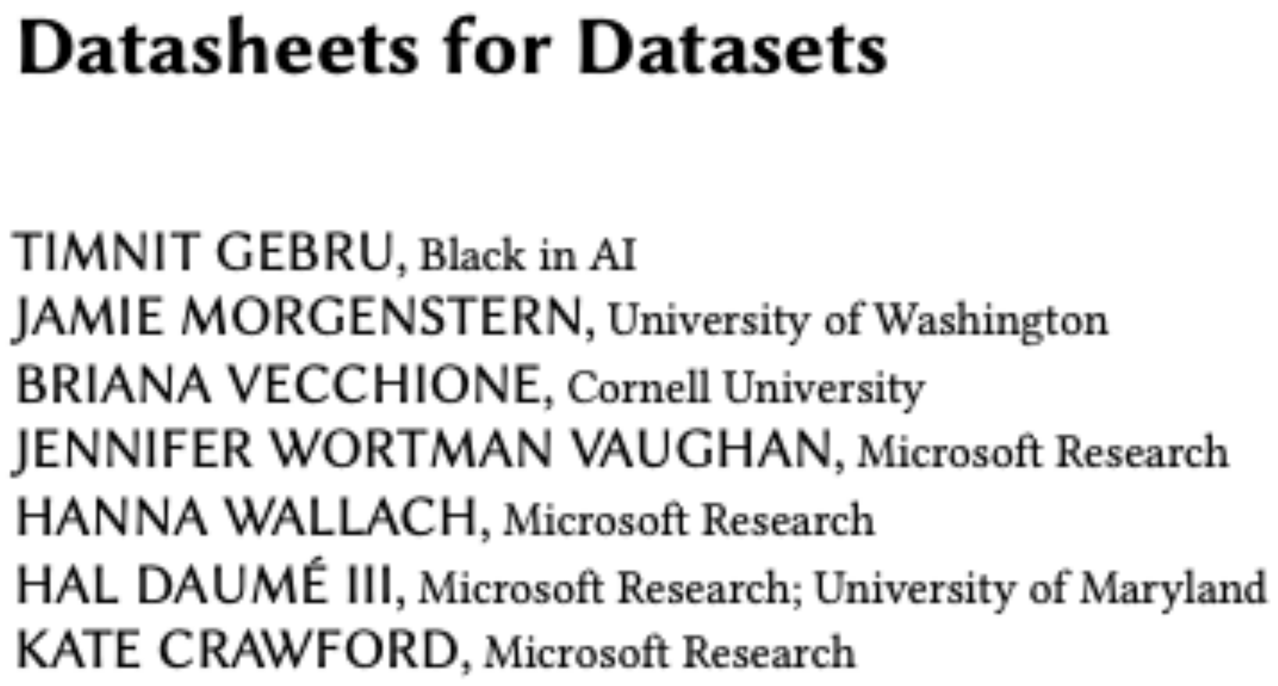
You

Hvordan søger jeg kontanthjælp?

Åbenhed skaber tillid



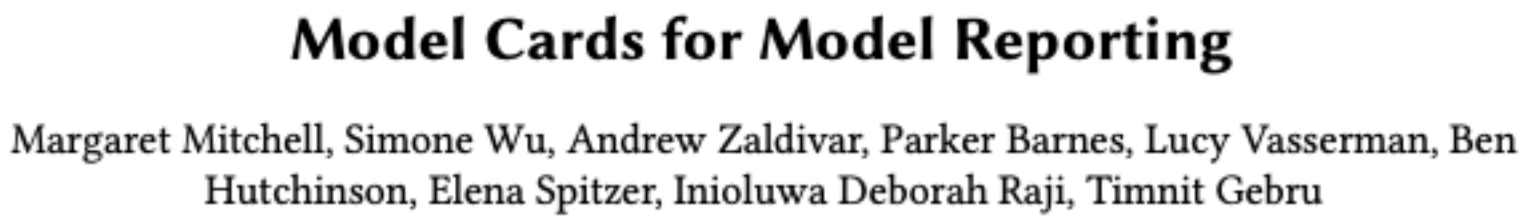
Lugo crashed his Lincoln Navigator into a Toyota Camry while drunk. He was rated as a low risk of reoffending despite the fact that it was at least his fourth DUI.



Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

- Motivation
- Sammensætning
- Indsamlingsproces
- Præprocessering



Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru

- Intenderet brug
- Træningsdata
- Analyser, etiske overvejelser
- Forbehold og anbefalinger

Sources & Notes

Mattu, J. A., Jeff Larson,Lauren Kirchner,Surya. (2016.). Machine Bias. ProPublica

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

Patient Record

Dsses

Age

Age
Nane

Name
20.17

As
Sarge

Lisex
Sex 4

Dstemons

Bam	436	Danes	
Conmaloerbai	Gobbrggenst	Stwenfde	
Dpe Beotoprec	Rittn	Gnoth. Bfn Loomng.	
Eobane	Ekao	Bwetokeses	
$BR	20:28	501	Bak
Garicialaliches	Sielnadan	Chesdes	
Maleprleam			
Geanatie	Nariquses	Imechda	Dlbcmche
	85	20.28312	
Elstinohedbakera	Pasa		
Tom Koss 18aeolidn	Plehan Makrot	Stas Ylkiw Hodhes	

Manglende validering

	Domæne	
Sprog	🇺🇸	🇺🇸+💊
	🇩🇰	🇩🇰+💊

Projektet

Målsætninger

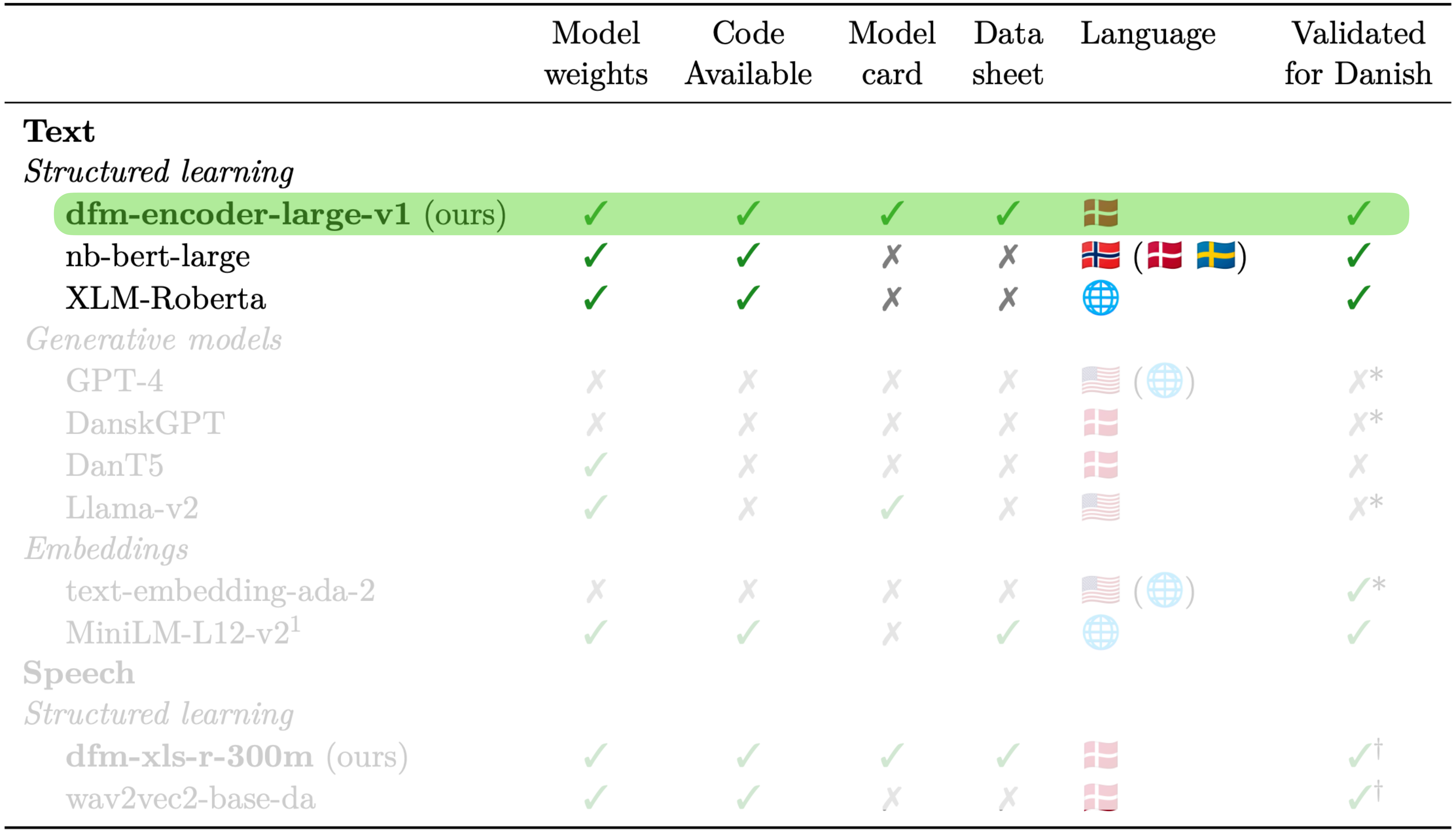
1. At **udvikle** og **vedligeholde** state-of-the-art sprogmodeller til dansk inden for både tekst og tale
2. At **validere** foundation-modeller på dansk og danske brugsscenarier
3. At opretholde en høj standard for **dokumentation** af modeller samt data
4. At **open-source** ikke kun modellerne, men også komponenter, der kræves for reproducerbarhed, såsom datarensning, træningskode og valideringskode

Sources & Notes

Enevoldsen, K., Hansen, L., Nielsen, D. S., Egebæk, R. A. F., Holm, S. V., Nielsen, M. C., Bernstorff, M., Larsen, R., Jørgensen, P. B., Højmark-Bertelsen, M., Vahlstrup, P. B., Møldrup-Dalum, P., & Nielbo, K. (2023). Danish Foundation Models (arXiv:2311.07264). arXiv. http://arxiv.org/abs/2311.07264

Status

The State of Foundation models for Danish



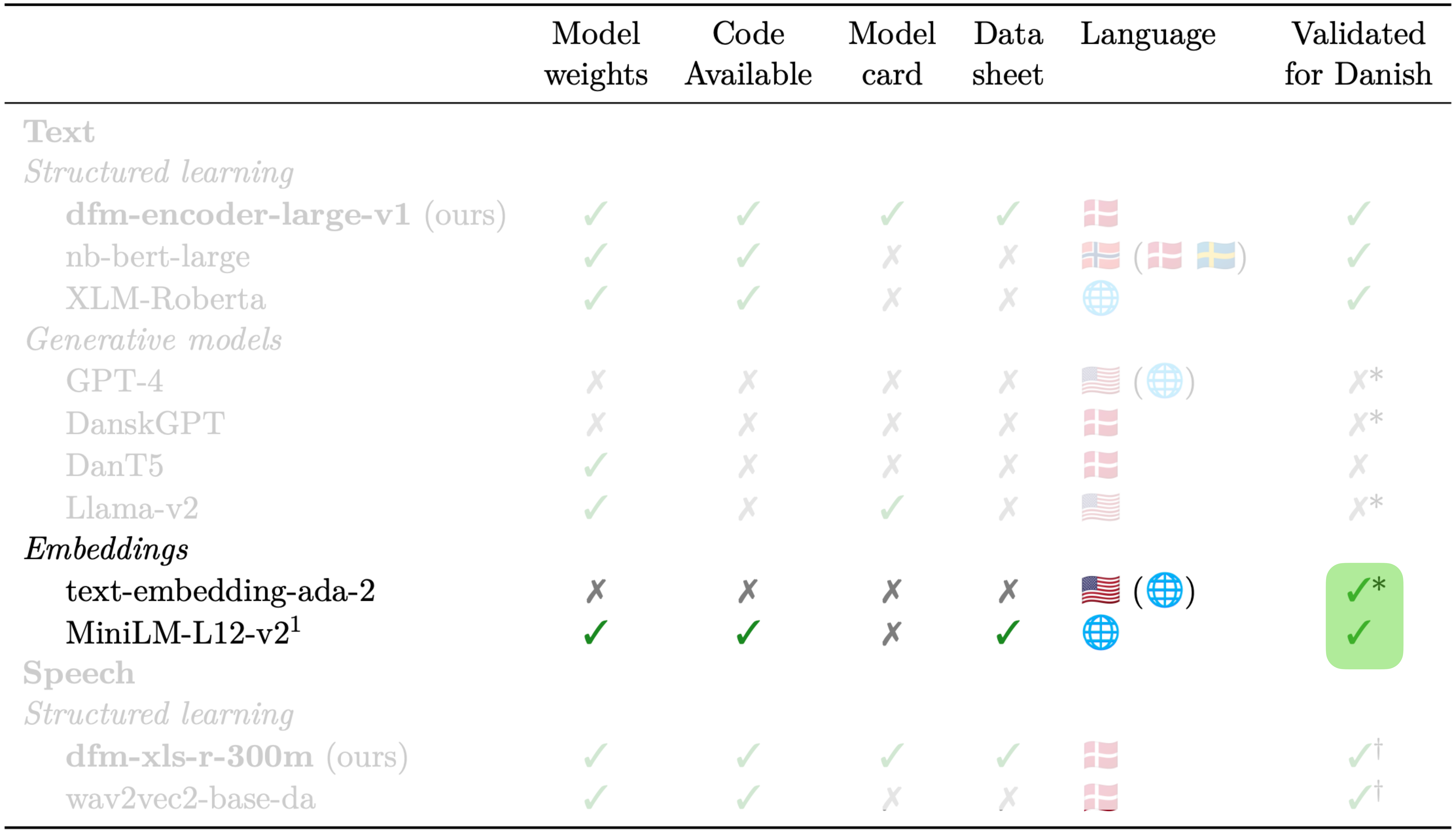
	Model weights	Code Available	Model card	Data sheet	Language	Validated for Danish
Text						
Structured learning						
dfm-encoder-large-v1 (ours)	✓	✓	✓	✓	🇩🇰	✓
nb-bert-large	✓	✓	✗	✗	🇳🇴 (🇩🇰 🇸🇪)	✓
XLM-Roberta	✓	✓	✗	✗	🌐	✓
Generative models						
GPT-4	✗	✗	✗	✗	🇺🇸 (🌐)	✗*
DanskGPT	✗	✗	✗	✗	🇩🇰	✗*
DanT5	✓	✗	✗	✗	🇩🇰	✗
Llama-v2	✓	✗	✓	✗	🇺🇸	✗*
Embeddings						
text-embedding-ada-2	✗	✗	✗	✗	🇺🇸 (🌐)	✓*
MiniLM-L12-v2[1]	✓	✓	✗	✓	🌐	✓
Speech						
Structured learning						
dfm-xls-r-300m (ours)	✓	✓	✓	✓	🇩🇰	✓†
wav2vec2-base-da	✓	✓	✗	✗	🇩🇰	✓†

Sources & Notes

Enevoldsen, K., Hansen, L., Nielsen, D. S., Egebæk, R. A. F., Holm, S. V., Nielsen, M. C., Bernstorff, M., Larsen, R., Jørgensen, P. B., Højmark-Bertelsen, M., Vahlstrup, P. B., Møldrup-Dalum, P., & Nielbo, K. (2023). Danish Foundation Models (arXiv:2311.07264). arXiv. http://arxiv.org/abs/2311.07264

CENTER FOR HUMANITIES COMPUTING

The State of Foundation models for Danish



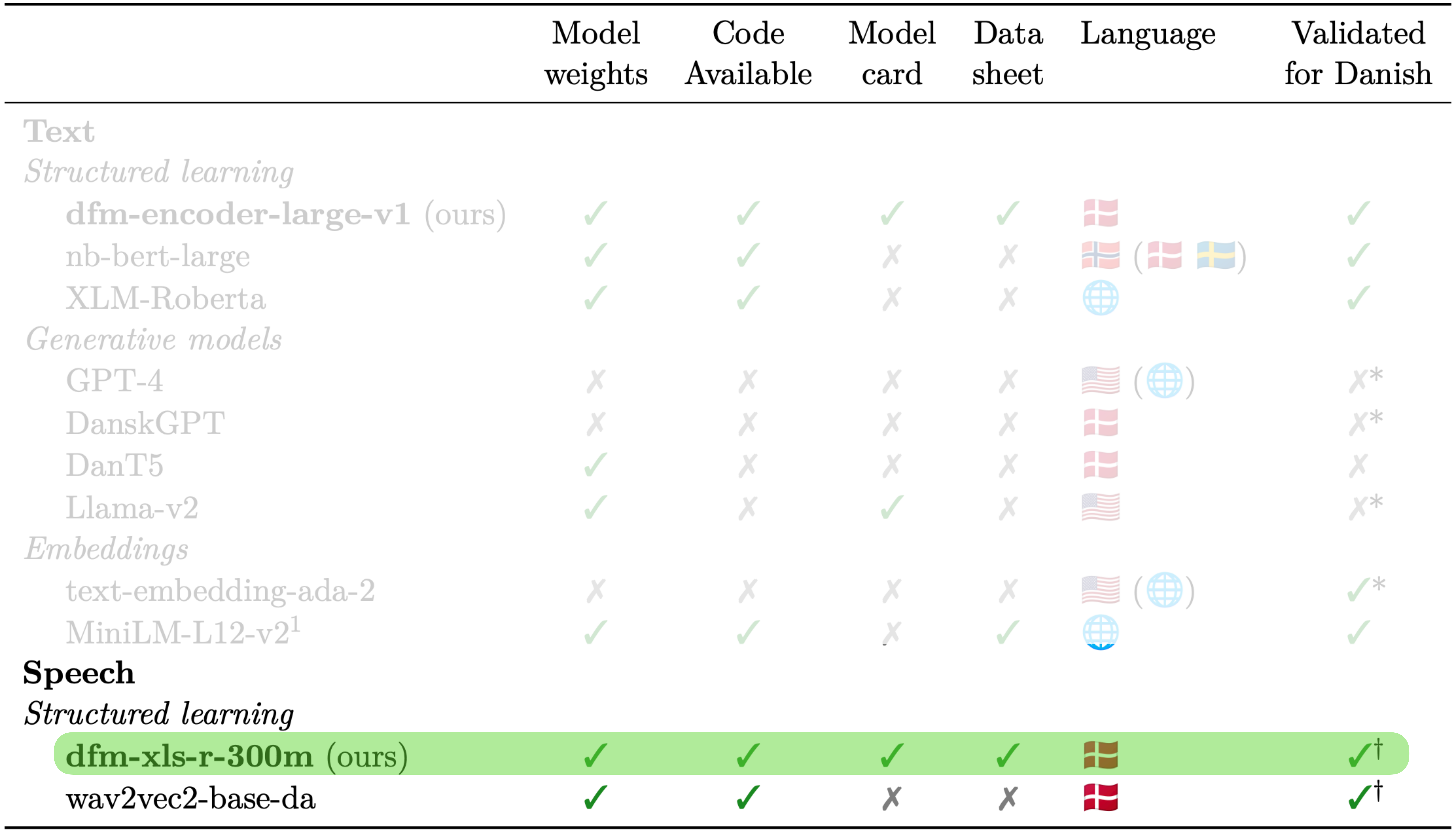
	Model weights	Code Available	Model card	Data sheet	Language	Validated for Danish
Text						
Structured learning						
dfm-encoder-large-v1 (ours)	✓	✓	✓	✓	🇩🇰	✓
nb-bert-large	✓	✓	✗	✗	🇳🇴 (🇩🇰 🇸🇪)	✓
XLM-Roberta	✓	✓	✗	✗	🌐	✓
Generative models						
GPT-4	✗	✗	✗	✗	🇺🇸 (🌐)	✗*
DanskGPT	✗	✗	✗	✗	🇩🇰	✗*
DanT5	✓	✗	✗	✗	🇩🇰	✗
Llama-v2	✓	✗	✓	✗	🇺🇸	✗*
Embeddings						
text-embedding-ada-2	✗	✗	✗	✗	🇺🇸 (🌐)	✓*
MiniLM-L12-v2[1]	✓	✓	✗	✓	🌐	✓
Speech						
Structured learning						
dfm-xls-r-300m (ours)	✓	✓	✓	✓	🇩🇰	✓†
wav2vec2-base-da	✓	✓	✗	✗	🇩🇰	✓†

Sources & Notes

Enevoldsen, K., Hansen, L., Nielsen, D. S., Egebæk, R. A. F., Holm, S. V., Nielsen, M. C., Bernstorff, M., Larsen, R., Jørgensen, P. B., Højmark-Bertelsen, M., Vahlstrup, P. B., Møldrup-Dalum, P., & Nielbo, K. (2023). Danish Foundation Models (arXiv:2311.07264). arXiv. http://arxiv.org/abs/2311.07264

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

The State of Foundation models for Danish



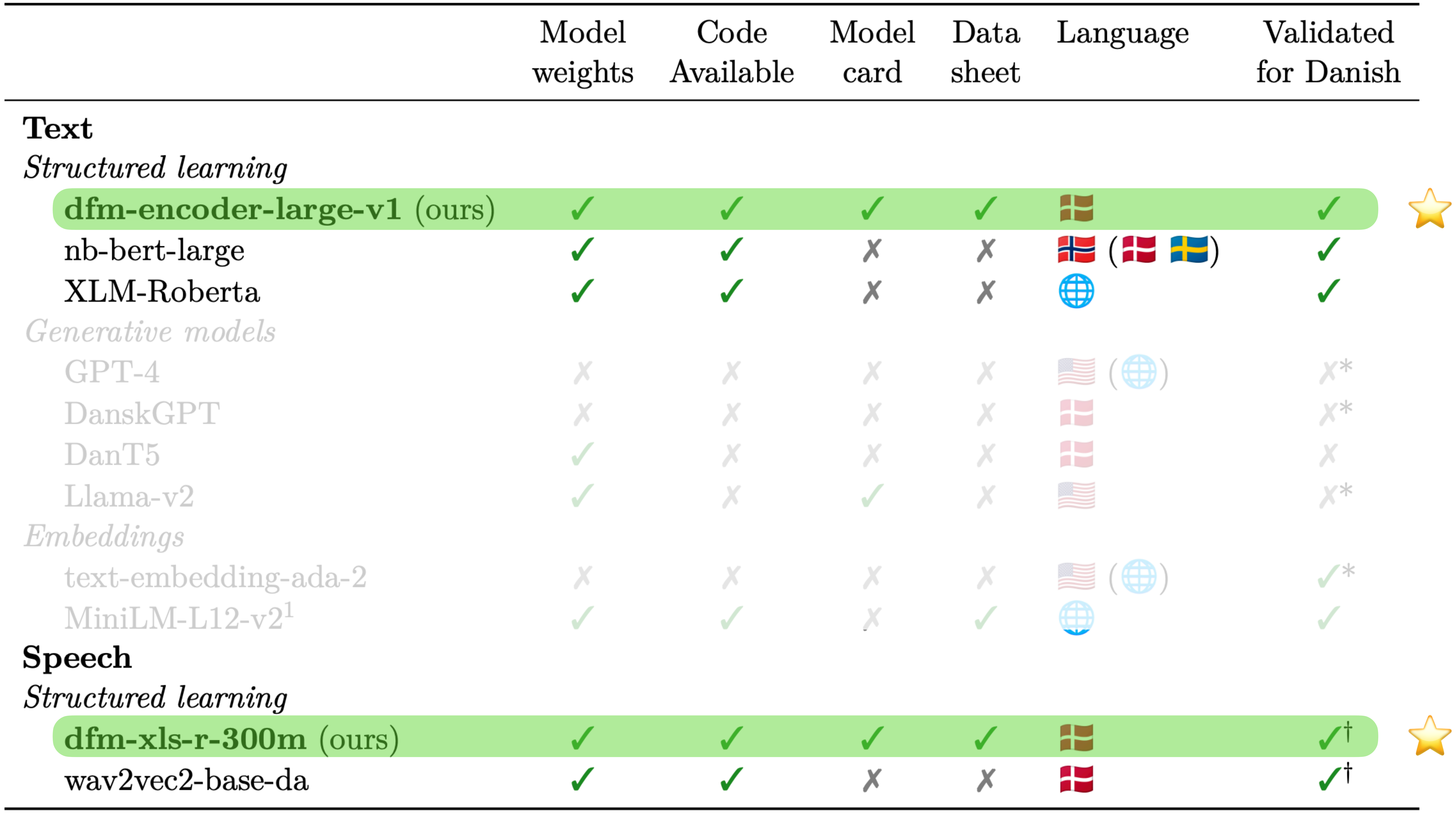
	Model weights	Code Available	Model card	Data sheet	Language	Validated for Danish
Text						
Structured learning						
dfm-encoder-large-v1 (ours)	✓	✓	✓	✓	🇩🇰	✓
nb-bert-large	✓	✓	✗	✗	🇳🇴 (🇩🇰 🇸🇪)	✓
XLM-Roberta	✓	✓	✗	✗	🌐	✓
Generative models						
GPT-4	✗	✗	✗	✗	🇺🇸 (🌐)	✗*
DanskGPT	✗	✗	✗	✗	🇩🇰	✗*
DanT5	✓	✗	✗	✗	🇩🇰	✗
Llama-v2	✓	✗	✓	✗	🇺🇸	✗*
Embeddings						
text-embedding-ada-2	✗	✗	✗	✗	🇺🇸 (🌐)	✓*
MiniLM-L12-v2[1]	✓	✓	✗	✓	🌐	✓
Speech						
Structured learning						
dfm-xls-r-300m (ours)	✓	✓	✓	✓	🇩🇰	✓†
wav2vec2-base-da	✓	✓	✗	✗	🇩🇰	✓†

Sources & Notes

Enevoldsen, K., Hansen, L., Nielsen, D. S., Egebæk, R. A. F., Holm, S. V., Nielsen, M. C., Bernstorff, M., Larsen, R., Jørgensen, P. B., Højmark-Bertelsen, M., Vahlstrup, P. B., Møldrup-Dalum, P., & Nielbo, K. (2023). Danish Foundation Models (arXiv:2311.07264). arXiv. http://arxiv.org/abs/2311.07264

CENTER FOR HUMANITIES COMPUTING

The State of Foundation models for Danish



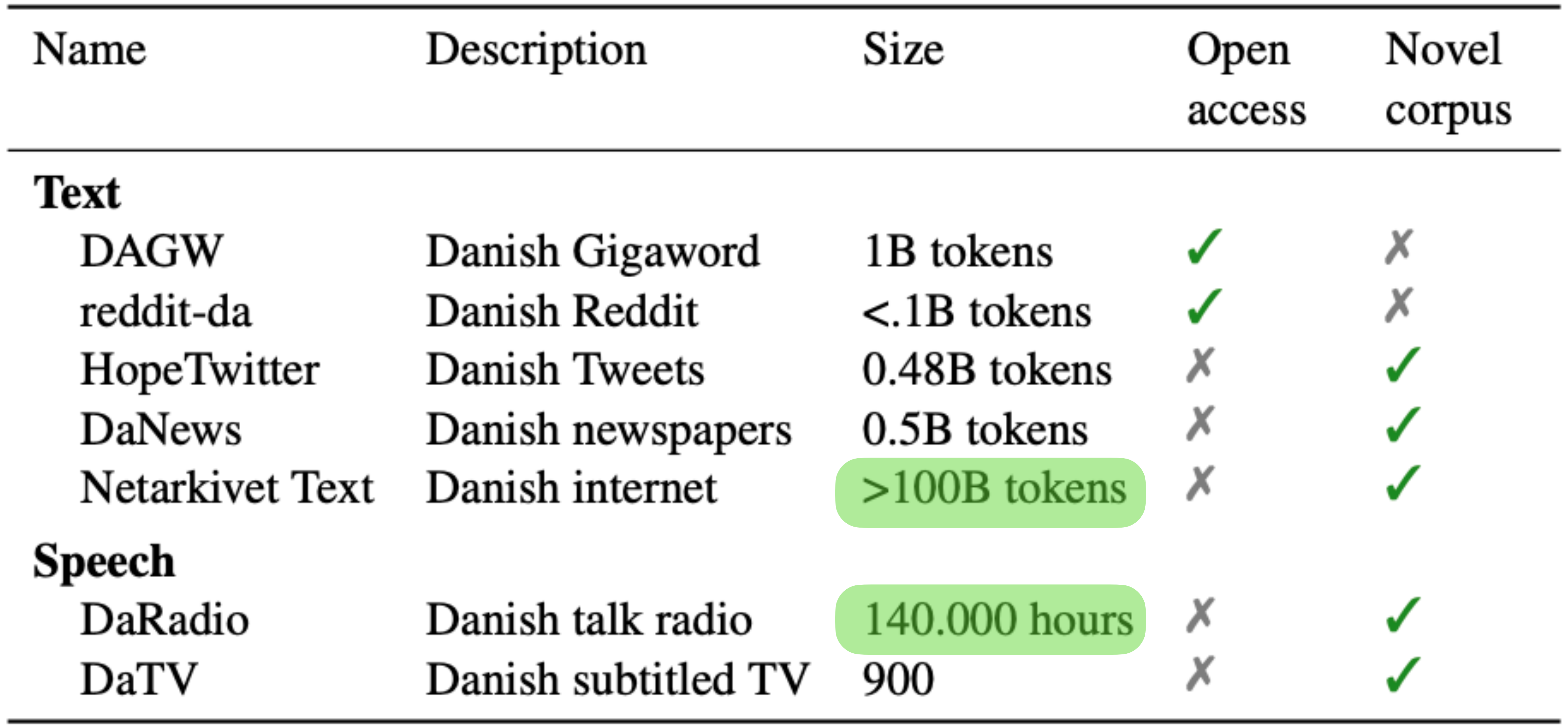
	Model weights	Code Available	Model card	Data sheet	Language	Validated for Danish	
Text							
Structured learning							
dfm-encoder-large-v1 (ours)	✓	✓	✓	✓	🇩🇰	✓	⭐
nb-bert-large	✓	✓	✗	✗	🇳🇴 (🇩🇰 🇸🇪)	✓	
XLM-Roberta	✓	✓	✗	✗	🌐	✓	
Generative models							
GPT-4	✗	✗	✗	✗	🇺🇸 (🌐)	✗*	
DanskGPT	✗	✗	✗	✗	🇩🇰	✗*	
DanT5	✓	✗	✗	✗	🇩🇰	✗	
Llama-v2	✓	✗	✓	✗	🇺🇸	✗*	
Embeddings							
text-embedding-ada-2	✗	✗	✗	✗	🇺🇸 (🌐)	✓*	
MiniLM-L12-v2[1]	✓	✓	✗	✓	🌐	✓	
Speech							
Structured learning							
dfm-xls-r-300m (ours)	✓	✓	✓	✓	🇩🇰	✓†	⭐
wav2vec2-base-da	✓	✓	✗	✗	🇩🇰	✓†	

Sources & Notes

Enevoldsen, K., Hansen, L., Nielsen, D. S., Egebæk, R. A. F., Holm, S. V., Nielsen, M. C., Bernstorff, M., Larsen, R., Jørgensen, P. B., Højmark-Bertelsen, M., Vahlstrup, P. B., Møldrup-Dalum, P., & Nielbo, K. (2023). Danish Foundation Models (arXiv:2311.07264). arXiv. http://arxiv.org/abs/2311.07264

CENTER FOR HUMANITIES COMPUTING

Our current dataset



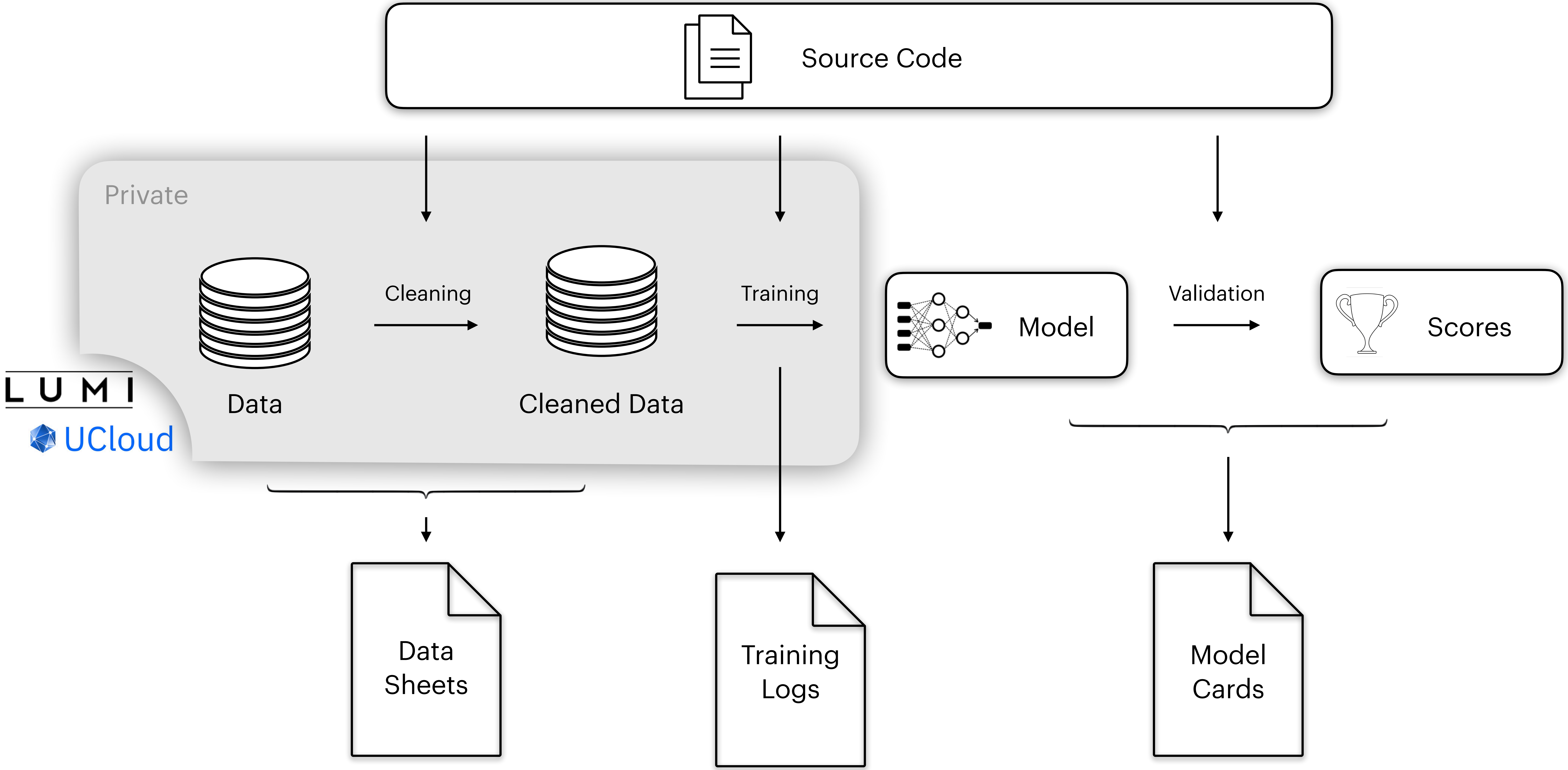
Name	Description	Size	Open access	Novel corpus
Text				
DAGW	Danish Gigaword	1B tokens	✓	✗
reddit-da	Danish Reddit	<.1B tokens	✓	✗
HopeTwitter	Danish Tweets	0.48B tokens	✗	✓
DaNews	Danish newspapers	0.5B tokens	✗	✓
Netarkivet Text	Danish internet	>100B tokens	✗	✓
Speech				
DaRadio	Danish talk radio	140.000 hours	✗	✓
DaTV	Danish subtitled TV	900	✗	✓

- Nye aftaler med:
  - Infomedia
  - Det kongelige bibliotek
- Arbejder vi på samarbejde med:
  - sundhed.dk, lex.dk, borger.dk, rigsarkivet, etc.
    - Giver bl.a. høj-kvalitets data til validering af eksisterende sprogmodeller

Sources & Notes

Enevoldsen, K., Hansen, L., Nielsen, D. S., Egebæk, R. A. F., Holm, S. V., Nielsen, M. C., Bernstorff, M., Larsen, R., Jørgensen, P. B., Højmark-Bertelsen, M., Vahlstrup, P. B., Møldrup-Dalum, P., & Nielbo, K. (2023). Danish Foundation Models (arXiv:2311.07264). arXiv. http://arxiv.org/abs/2311.07264

Datasikkerhed og Governance



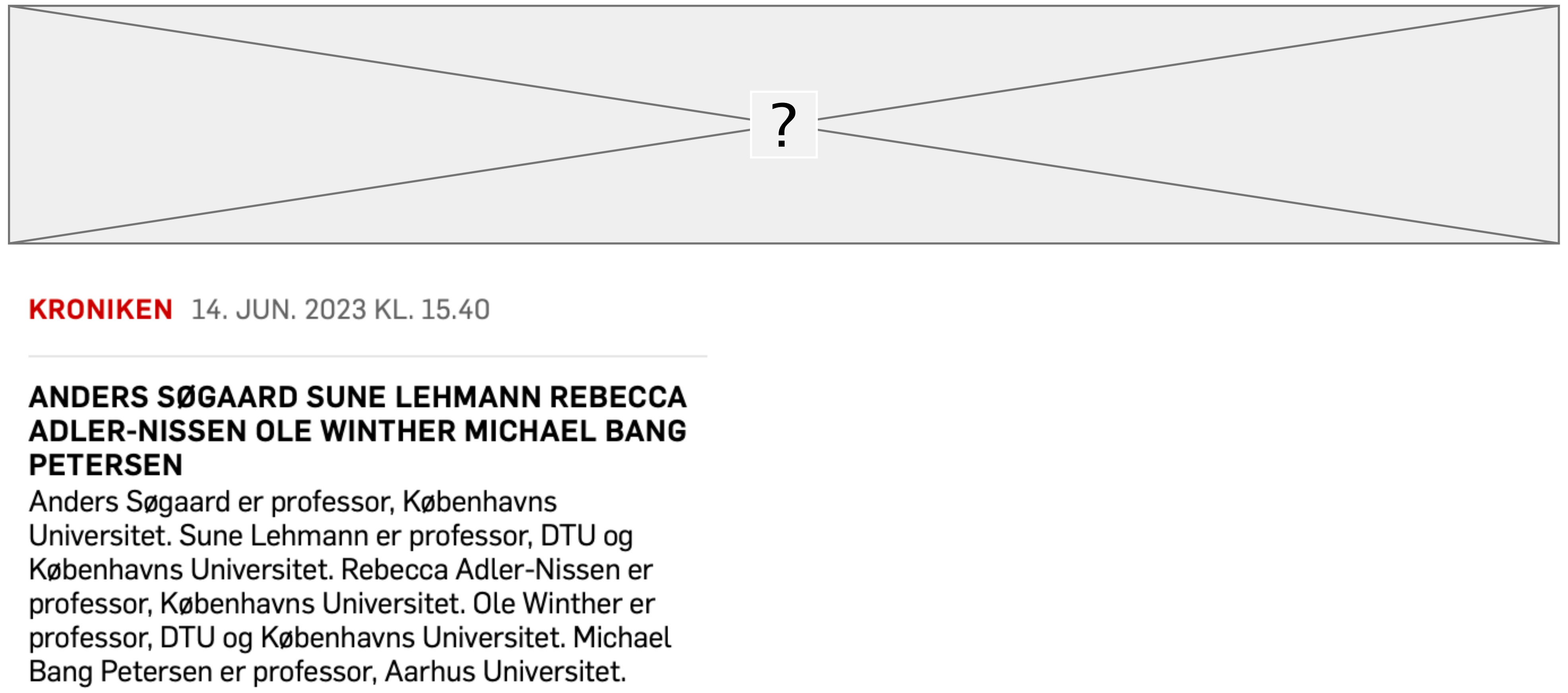
Sources & Notes

Enevoldsen, K., Hansen, L., Nielsen, D. S., Egebæk, R. A. F., Holm, S. V., Nielsen, M. C., Bernstorff, M., Larsen, R., Jørgensen, P. B., Højmark-Bertelsen, M., Vahlstrup, P. B., Møldrup-Dalum, P., & Nielbo, K. (2023). Danish Foundation Models (arXiv:2311.07264). arXiv. http://arxiv.org/abs/2311.07264

Næste skridt

CENTER FOR
HUMANITIES
COMPUTING

Er danske modeller blot navlepilleri?



KRONIKEN 14. JUN. 2023 KL. 15.40

ANDERS SØGAARD SUNE LEHMANN REBECCA ADLER-NISSEN OLE WINTHER MICHAEL BANG PETERSEN

Anders Søgaard er professor, Københavns Universitet. Sune Lehmann er professor, DTU og Københavns Universitet. Rebecca Adler-Nissen er professor, Københavns Universitet. Ole Winther er professor, DTU og Københavns Universitet. Michael Bang Petersen er professor, Aarhus Universitet.

TEKNOLOGI

Staten drømmer om sin helt egen chatbot – men manden bag Danmarks største it-virksomhed siger ‘nej tak’ til opgaven

Sources & Notes

https://www.dr.dk/nyheder/politik/professor-danmark-boer-udvikle-sin-egen-kunstige-intelligens

https://www.dr.dk/nyheder/viden/teknologi/staten-droemmer-om-sin-helt-egen-chatbot-men-manden-bag-danmarks-stoerste-it

https://politiken.dk/debat/kroniken/art9374522/Skab-et-offentligt-alternativ-til-techgiganterne

Er danske modeller blot navlepilleri?

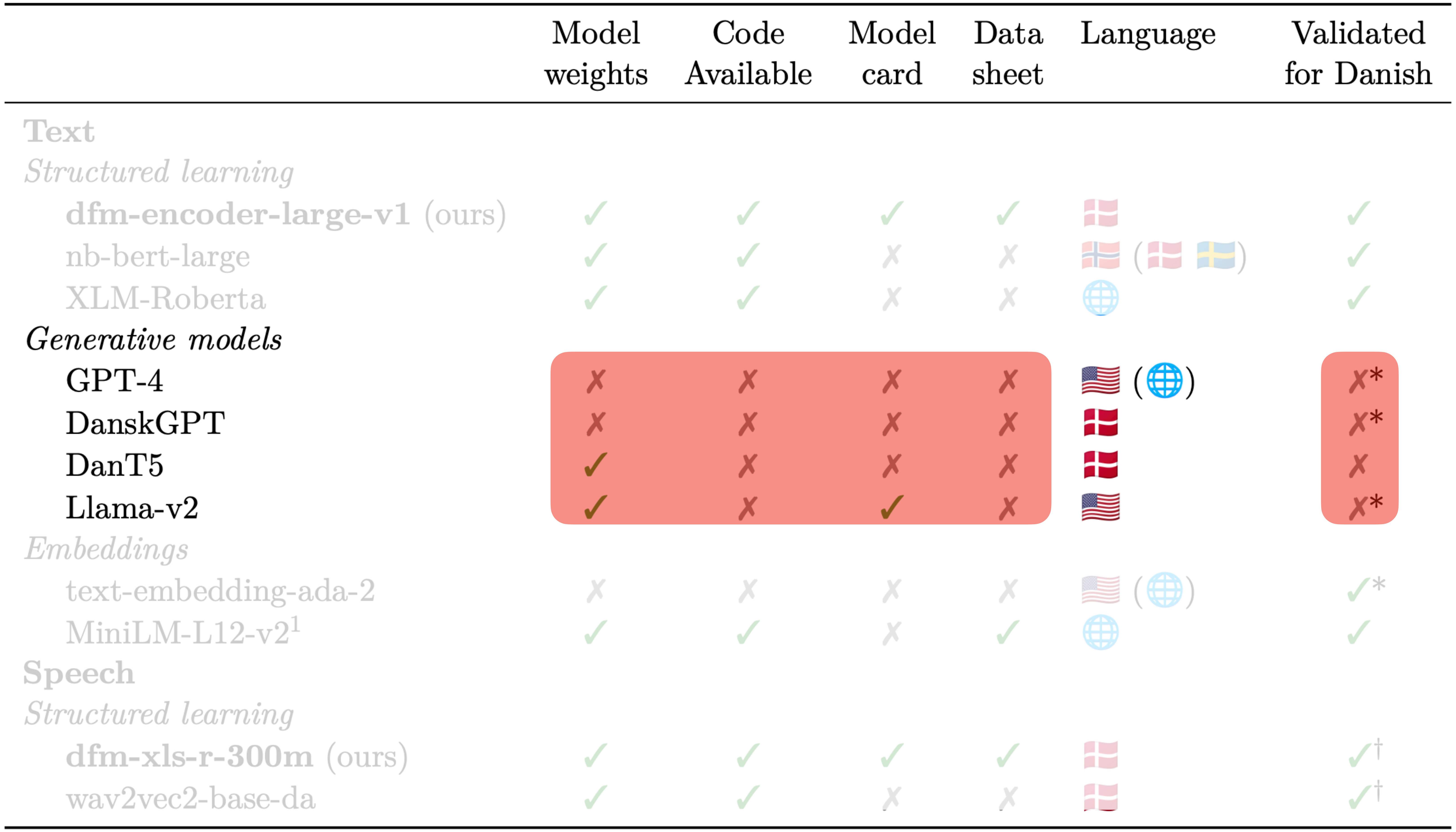
- Store sprogmodeller er **fleksible** og **modulære**
  - **Kombinere** modeller
  - Modeller som **kulturelle vidensdatabaser**
  - …
- En potential fremtid

Sources & Notes

Li, M., Gururangan, S., Dettmers, T., Lewis, M., Althoff, T., Smith, N. A., & Zettlemoyer, L. (2022). Branch-Train-Merge: Embarrassingly Parallel Training of Expert Language Models (arXiv:2208.03306). arXiv. https://doi.org/10.48550/arXiv.2208.03306

Feng, S., Shi, W., Bai, Y., Balachandran, V., He, T., & Tsvetkov, Y. (2023). Knowledge Card: Filling LLMs' Knowledge Gaps with Plug-in Specialized Language Models (arXiv:2305.09955; Version 2). arXiv. https://doi.org/10.48550/arXiv.2305.09955

The State of Foundation models for Danish



	Model weights	Code Available	Model card	Data sheet	Language	Validated for Danish
Text						
Structured learning						
dfm-encoder-large-v1 (ours)	✓	✓	✓	✓	🇩🇰	✓
nb-bert-large	✓	✓	✗	✗	🇳🇴 (🇩🇰 🇸🇪)	✓
XLM-Roberta	✓	✓	✗	✗	🌐	✓
Generative models						
GPT-4	✗	✗	✗	✗	🇺🇸 (🌐)	✗*
DanskGPT	✗	✗	✗	✗	🇩🇰	✗*
DanT5	✓	✗	✗	✗	🇩🇰	✗
Llama-v2	✓	✗	✓	✗	🇺🇸	✗*
Embeddings						
text-embedding-ada-2	✗	✗	✗	✗	🇺🇸 (🌐)	✓*
MiniLM-L12-v2[1]	✓	✓	✗	✓	🌐	✓
Speech						
Structured learning						
dfm-xls-r-300m (ours)	✓	✓	✓	✓	🇩🇰	✓†
wav2vec2-base-da	✓	✓	✗	✗	🇩🇰	✓†

Sources & Notes

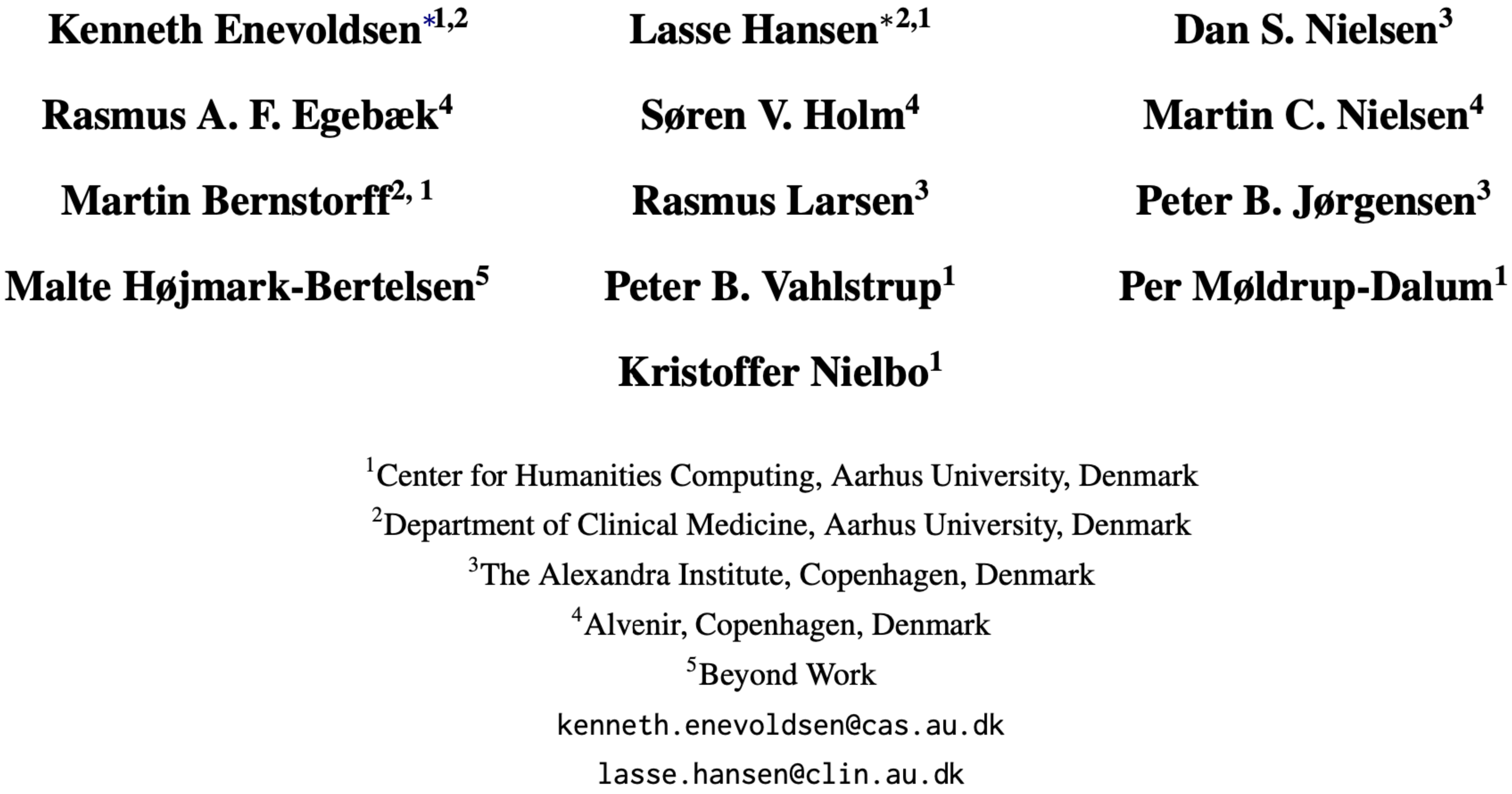
Enevoldsen, K., Hansen, L., Nielsen, D. S., Egebæk, R. A. F., Holm, S. V., Nielsen, M. C., Bernstorff, M., Larsen, R., Jørgensen, P. B., Højmark-Bertelsen, M., Vahlstrup, P. B., Møldrup-Dalum, P., & Nielbo, K. (2023). Danish Foundation Models (arXiv:2311.07264). arXiv. http://arxiv.org/abs/2311.07264

Næste skridt

- Modeller
  - 7B generative modeller —> skalering til større modeller
- Datasamarbejder
  - Sikre højkvalitetsdata tæt på anvendelse
- Evalueringssamarbejder
  - Sikre evaluering på danske anvendelsesområder

Afslutning

Særlig tak til det nuværende hold



Kenneth Enevoldsen*[1,2] Lasse Hansen*[2,1] Dan S. Nielsen[3]

Rasmus A. F. Egebæk[4] Søren V. Holm[4] Martin C. Nielsen[4]

Martin Bernstorff[2, 1] Rasmus Larsen[3] Peter B. Jørgensen[3]

Malte Højmark-Bertelsen[5] Peter B. Vahlstrup[1] Per Møldrup-Dalum[1]

Kristoffer Nielbo[1]

[1]Center for Humanities Computing, Aarhus University, Denmark
[2]Department of Clinical Medicine, Aarhus University, Denmark
[3]The Alexandra Institute, Copenhagen, Denmark
[4]Alvenir, Copenhagen, Denmark
[5]Beyond Work

kenneth.enevoldsen@cas.au.dk
lasse.hansen@clin.au.dk

Sources
& Notes

Enevoldsen, K., Hansen, L., Nielsen, D. S., Egebæk, R. A. F., Holm, S. V., Nielsen, M. C., Bernstorff, M., Larsen, R., Jørgensen, P. B., Højmark-Bertelsen, M., Vahlstrup, P. B., Møldrup-Dalum, P., & Nielbo, K. (2023). Danish Foundation Models (arXiv:2311.07264). arXiv. http://arxiv.org/abs/2311.07264

Lær mere

Videnskabelig Artikel

Hjemmeside

GitHub

Kontakt os

Sources & Notes

Enevoldsen, K., Hansen, L., Nielsen, D. S., Egebæk, R. A. F., Holm, S. V., Nielsen, M. C., Bernstorff, M., Larsen, R., Jørgensen, P. B., Højmark-Bertelsen, M., Vahlstrup, P. B., Møldrup-Dalum, P., & Nielbo, K. (2023). Danish Foundation Models (arXiv:2311.07264). arXiv. http://arxiv.org/abs/2311.07264

CENTER FOR HUMANITIES COMPUTING