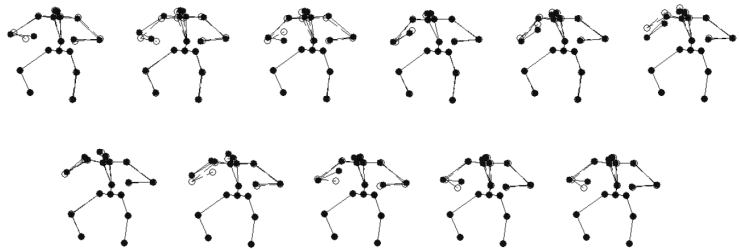
FROM DATA TO DECISION
– HARNESSING AI AND LANGUAGE MODELS FOR REAL-WORLD APPLICATIONS

Kristoffer Nielbo, Professor & Director
Center for Humanities Computing, Aarhus University

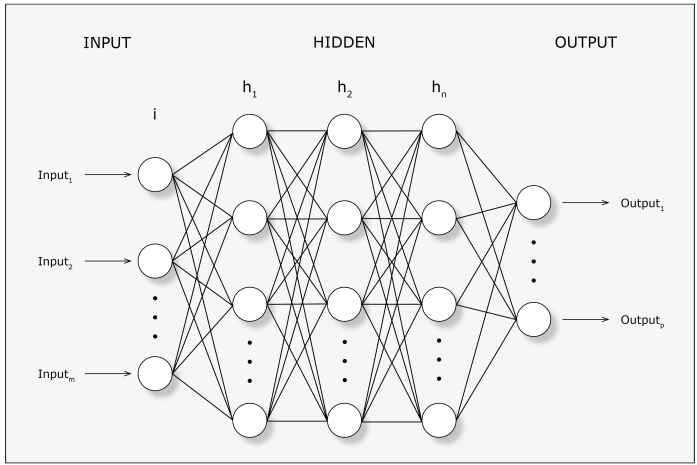
OUTLINE

1. **whoami**
2. **Langauge Models**
   - LLMs - a short history
3. **Applications**
   - Issues
   - Danish foundation models
4. **Summary**

OUTLINE



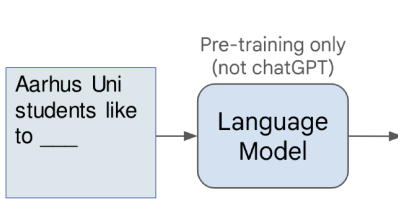
A person in a context that does? *type of action*



Artificial Neural Network

OUTLINE

Language Models and Large Language Models



(hypothetical)

Word	Probability
a	*0.00001*
aardvark	*0.000004*
...	
drink	*0.5*
...	
study	*0.23*
...	
zucchini	*0.000002*

Grammar	In my free time, I like to {**run**, banana}
Lexical semantics	I went to the zoo to see giraffes, lions, and {**zebras**, spoon}
World knowledge	The capital of Denmark is {**Copenhagen**, London}
Sentiment analysis	Movie review: I was engaged and on the edge of my seat the whole time. The movie was {**good**, bad}
Harder sentiment analysis	Movie review: Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was {**bad**, good}
Translation	The word for "pretty" in Spanish is {**bonita**, hola}
Spatial reasoning	[...] Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the {**kitchen**, store}
Math question	First grade arithmetic exam: 3 + 8 + 4 = {**15**, 11}

Transformer-based language models

2013: Word2vec

2015: Attention

2017: Transformer

2018: BERT, GPT-1, GPT-2

2019: BART

2020: GPT-3, RAG

2022: Instruct GPT, RLHF, ChatGPT

2023: LLaMA, Gemini, GPT-4

2024: Multimodality, SLM, open source

Transformer-based language models

2013: Word2vec

2015: Attention

2017: Transformer

2018: BERT, GPT-1, GPT-2

2019: BART

2020: GPT-3, RAG

2022: Instruct GPT, RLHF, ChatGPT

2023: LLaMA, Gemini, GPT-4

2024: Multimodality, SLM, open source

Transformer-based language models

2013: Word2vec

2015: Attention

2017: Transformer

2018: BERT, GPT-1, GPT-2

2019: BART

2020: GPT-3, RAG

2022: Instruct GPT, RLHF, ChatGPT

2023: LLaMA, Gemini, GPT-4

2024: Multimodality, SLM, open source

Transformer-based language models

2013: Word2vec

2015: Attention

2017: Transformer

2018: BERT, GPT-1, GPT-2

2019: BART

2020: GPT-3, RAG

2022: Instruct GPT, RLHF, ChatGPT

2023: LLaMA, Gemini, GPT-4

2024: Multimodality, SLM, open source

Transformer-based language models

2013: Word2vec

2015: Attention

2017: Transformer

2018: BERT, GPT-1, GPT-2

2019: BART

2020: GPT-3, RAG

2022: Instruct GPT, RLHF, ChatGPT

2023: LLaMA, Gemini, GPT-4

2024: Multimodality, SLM, open source

Transformer-based language models

2013: Word2vec

2015: Attention

2017: Transformer

2018: BERT, GPT-1, GPT-2

2019: BART

2020: GPT-3, RAG

2022: Instruct GPT, RLHF, ChatGPT

2023: LLaMA, Gemini, GPT-4

2024: Multimodality, SLM, open source

Transformer-based language models

2013: Word2vec

2015: Attention

2017: Transformer

2018: BERT, GPT-1, GPT-2

2019: BART

2020: GPT-3, RAG

2022: Instruct GPT, RLHF, ChatGPT

2023: LLaMA, Gemini, GPT-4

2024: Multimodality, SLM, open source

Transformer-based language models

2013: Word2vec

2015: Attention

2017: Transformer

2018: BERT, GPT-1, GPT-2

2019: BART

2020: GPT-3, RAG

2022: Instruct GPT, RLHF, ChatGPT

2023: LLaMA, Gemini, GPT-4

2024: Multimodality, SLM, open source

Transformer-based language models

2013: Word2vec

2015: Attention

2017: Transformer

2018: BERT, GPT-1, GPT-2

2019: BART

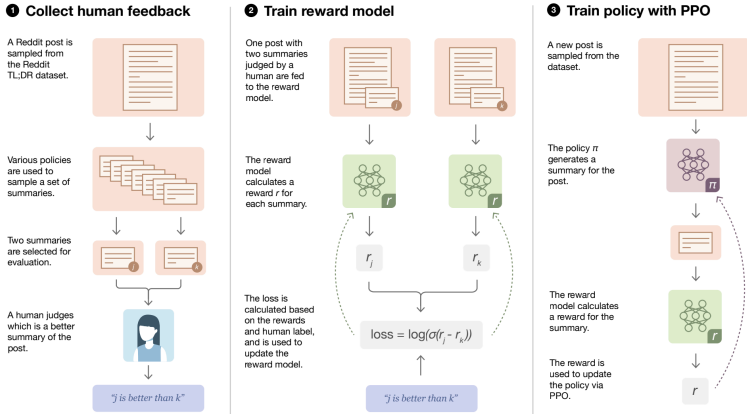
2020: GPT-3, RAG

2022: Instruct GPT, RLHF, ChatGPT

2023: LLaMA, Gemini, GPT-4

2024: Multimodality, SLM, open source

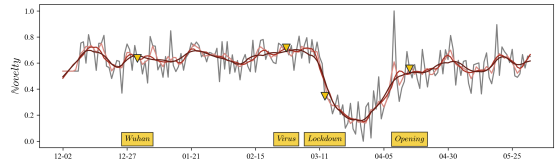
What happened in 2022?

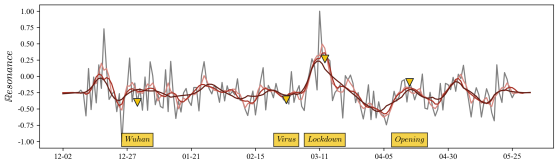


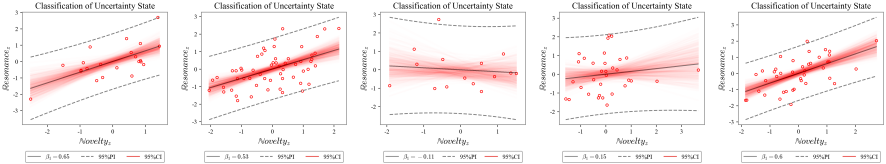
Alignment with RLHF, Ouyang et al. [2022]

OUTLINE

1. whoami
2. Langauge Models
    - LLMs - a short history
3. **Applications**
    - Issues
    - Danish foundation models
4. Summary





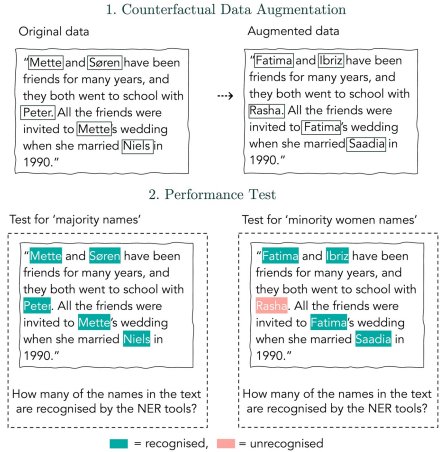


Change detection on novelty, $\mathcal{N}$, and chance description on the resonance on novelty slope, $\mathcal{N} \cdot \mathcal{R}$-slope, Nielbo et al. [2023].

	majority	*minority*
women	$\mathcal{A}$	$\mathcal{B}$
men	$\mathcal{C}$	$\mathcal{D}$

Intersectional biases in LLM applications

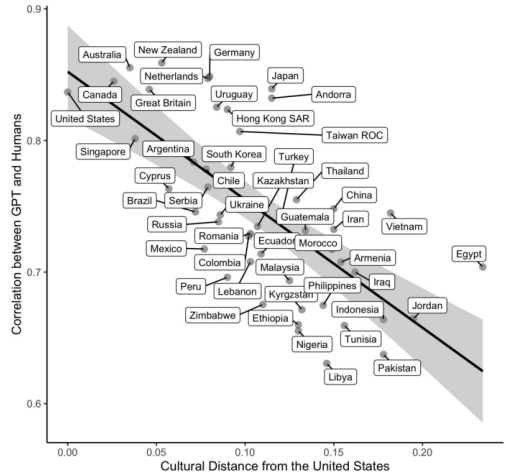
- originates in training objectives and data
- results in LLM misalignment with a range of ethical and security issues



Homogeneous monolingual setting

LLMs learn from data, if the data reflects historical biases, AI will too

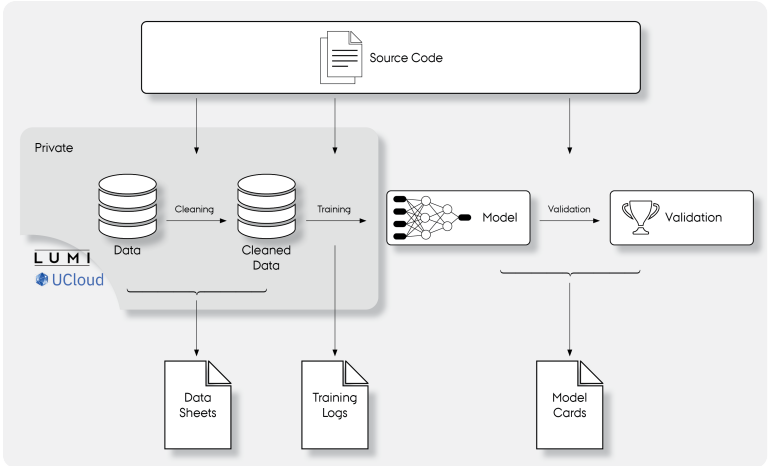
- **Data encoding biases** stem from non-representative training datasets or prejudiced annotators
- Preference training that **overfits on a Western value set**
- Eroding trust and can cause societal harm by reinforcing stereotypes



Cultural (mis-)alignment, Source: J. Henrich

Danish Foundation Models

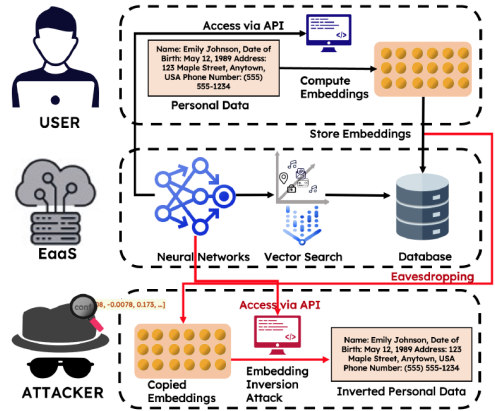
1. Develop and maintain **state-of-the-art** models for Danish,
2. which are **well-validated** across a wide range of tasks.
3. Furthermore, we wish to **ensure good documentation**, which allows users to assess the model for their use-case critically
4. **Open-source**, both model and source code
5. Models should comply with **regulatory frameworks** and use responsible AI governance
6. **Continuous monitoring** and adaptation in development and deployment



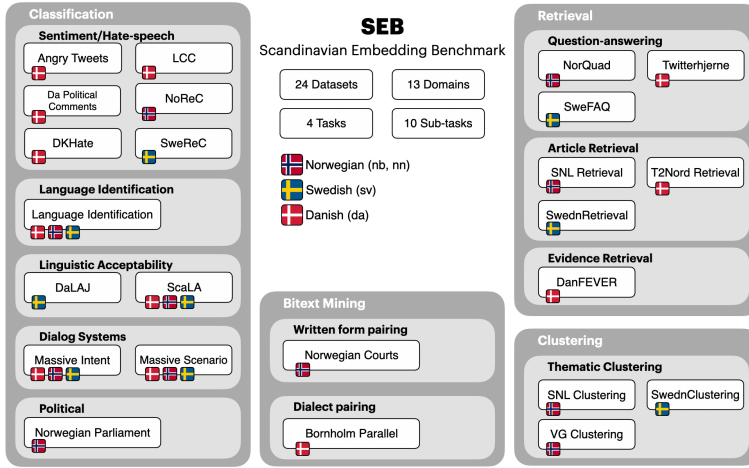
DFM environment for training Danish base models Enevoldsen et al. [2023] on ScandEval Nielsen [2023]

Exploiting embedding vulnerabilities

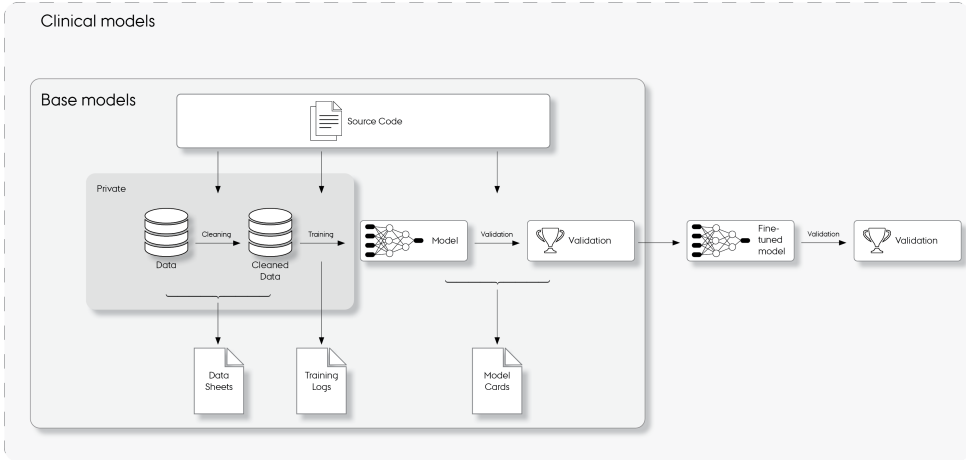
- Using **prompt injections** to manipulate the output of the model
- neural embeddings of sensitive data retain much of the original data – **embedding inversion attack**
- existing **defense mechanisms focus on English** – leaving other languages more susceptible to security threats
- linguistic bias becomes a cybersecurity threat



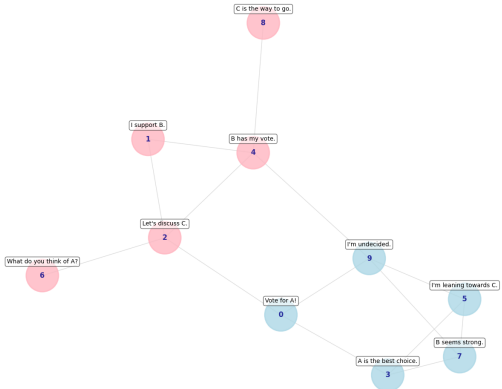
Source: Chen et al. [2024]



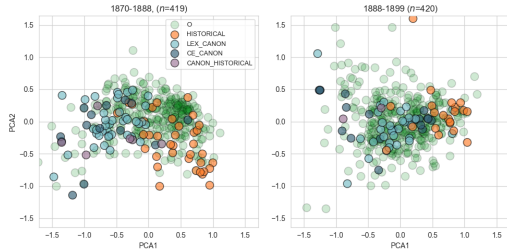
Intrinsic Model Evaluation with SEB, Enevoldsen et al. [2024]; Nielsen et al. [2024]



Extended DFM development environment for fine-tuning and aligning DFM models for clinical applications



European Citie^2 Simulating democratic processes (voting mechanisms) with Danish municipalities



Golden Arrays – Identifying drivers of historical change since the 16th century.

OUTLINE

1. whoami
2. Langauge Models
    - LLMs - a short history
3. Applications
    - Issues
    - Danish foundation models
4. **Summary**

In summary

- The field of AI-based language technology is diversifying
- Still suffers from a) *homogeneous monolingual setting* and b) preference training that overfit on common Western-centric data sets
- National models have value in a range of applications that have specific quality and security requirements
- clinical decision support, educational instructors, public sector agents …

THANK YOU

kln@cas.au.dk
chc.au.dk

SLIDES

knielbo.github.io/files/ai-is-ml.pdf

Chen, Y., Lent, H., and Bjerva, J. (2024). Text Embedding Inversion Security for Multilingual Language Models. arXiv:2401.12192 [cs].

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., and Nielbo, K. (2023). Danish foundation models.

Enevoldsen, K., Kardos, M., Muennighoff, N., and Nielbo, K. L. (2024). The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding.

Nielbo, K., Enevoldsen, K., Baglini, R., Fano, E., Roepstorff, A., and Gao, J. (2023). Pandemic news information uncertainty—news dynamics mirror differential response strategies to covid-19. *PLOS ONE*, 18(1):1–16.

Nielsen, D. S. (2023). Scandeval: A benchmark for scandinavian natural language processing.

Nielsen, D. S., Enevoldsen, K., and Schneider-Kamp, P. (2024). Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual nlu tasks.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback.