Workshop on Danish Language Models

Kenneth Enevoldsen, PhD student
Lasse Hansen, PhD student

Rasmus Larsen, AI Specialist
Dan Saattrup Nielsen, Senior AI Specialist

Today's Programme

1. Introduction to large language models
2. The Danish large language model scene
3. Developing Danish large language models
4. Hands-on workshop
5. Future directions

Introduction to Large Language Models

What is a Language Model?

Rough definition

A statistical model that can work with language

This includes *a lot* of different things:

1. Bag-of-word approaches
2. Naïve Bayes models
3. Multi-layered perceptrons
4. Recurrent neural networks
5. Transformers
6. State space models
7. …

What is a Language Model?

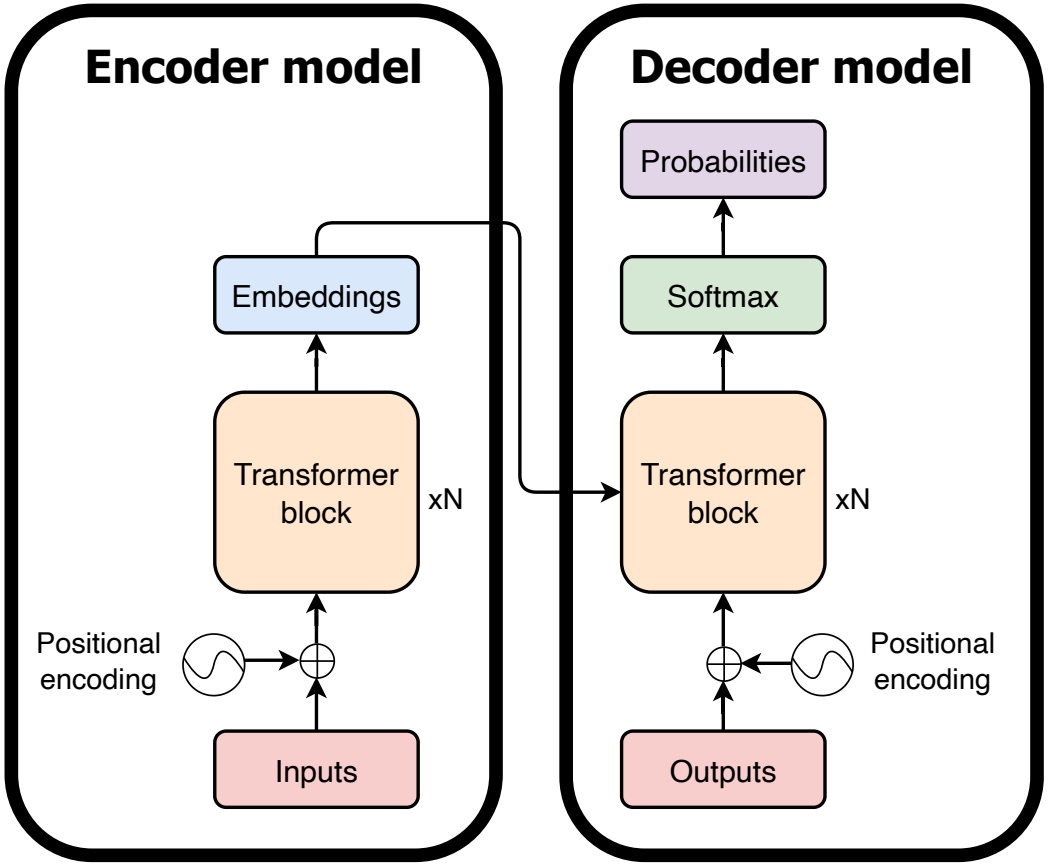
Rough definition

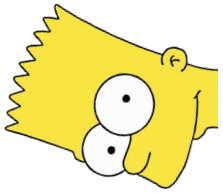
A statistical model that can work with language

This includes *a lot* of different things:

1. Bag-of-word approaches
2. Naïve Bayes models
3. Multi-layered perceptrons
4. Recurrent neural networks
5. Transformers ← *These are the most popular these days though*
6. State space models
7. …

Language Models Come in Different Flavours





A New Paradigm

Encoder models follow the **transfer learning paradigm**:

- You first *pretrain* the model on a large unlabelled text corpus
- Next, you *finetune* the model on a small labelled text corpus

These models are *expert models*

Decoder models follow the **in-context learning paradigm**:

- You first *pretrain* the model on a large unlabelled text corpus
- Next, you *tweak the prompts* when the model should be used on a task

These models are *general-purpose models*

What is a **L**arge **L**anguage **M**odel (LLM)?

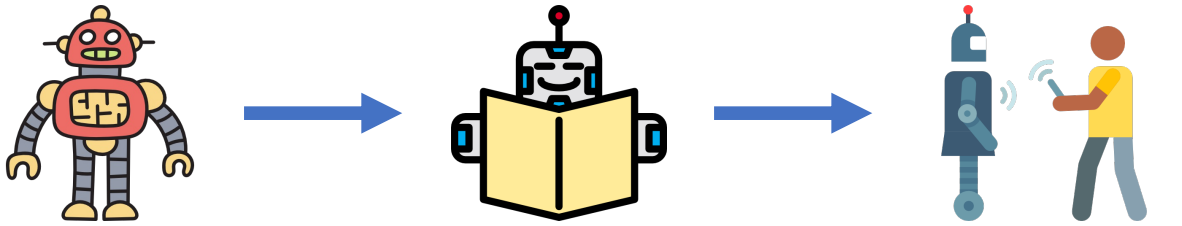
- For neural network based language models, we can do a simple count of the number of parameters in them
- Encoder models typically have 100-500 million parameters
- Decoder models these days typically have 1-70 billion parameters

Rough heuristic

A language model is **large** if it has more than 1 billion parameters

- A consumer-grade 24 GB GPU (NVIDIA RTX 3090) can run models with up to ~20-30 billion parameters

Generative Language Models Nowadays

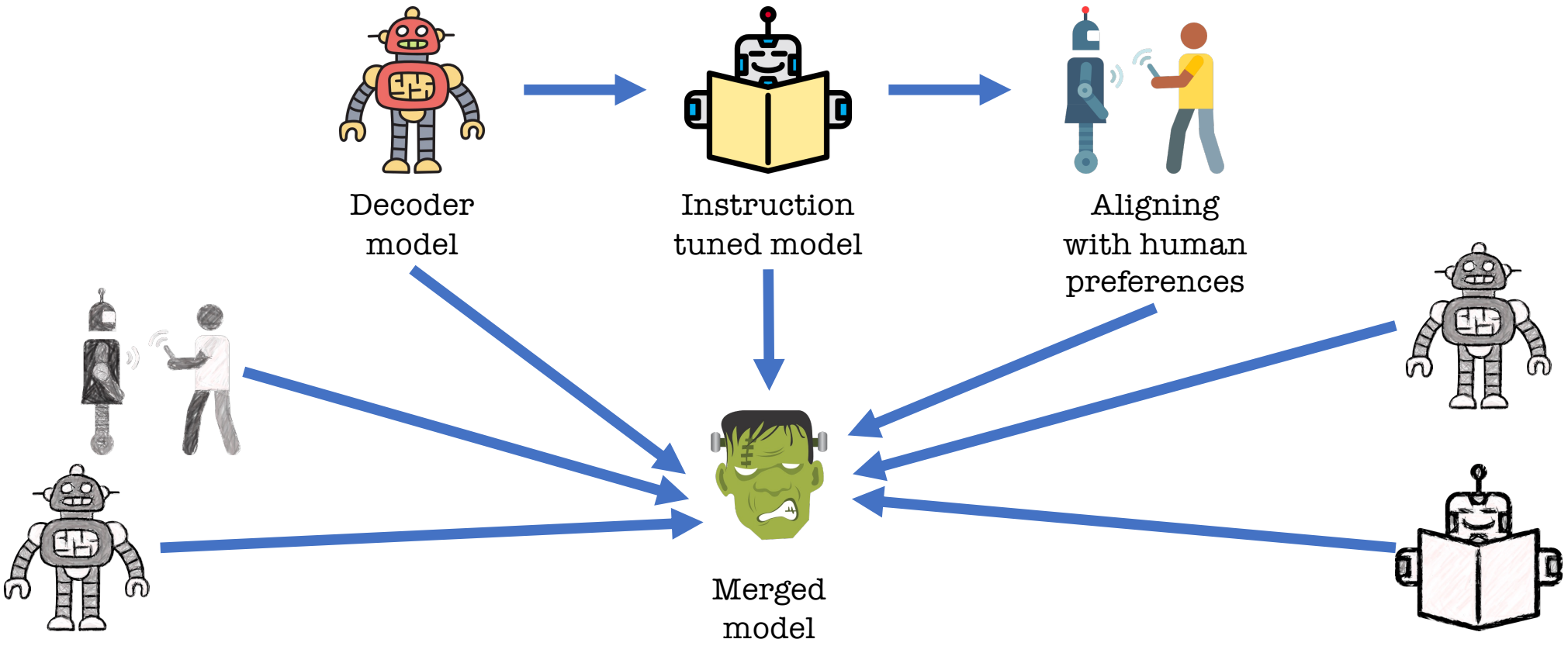


Decoder model

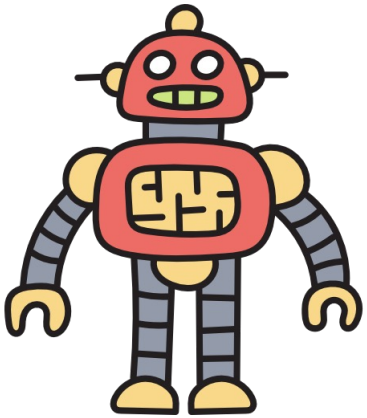
Instruction tuned model

Aligning with human preferences

Generative Language Models Nowadays



Decoder Model



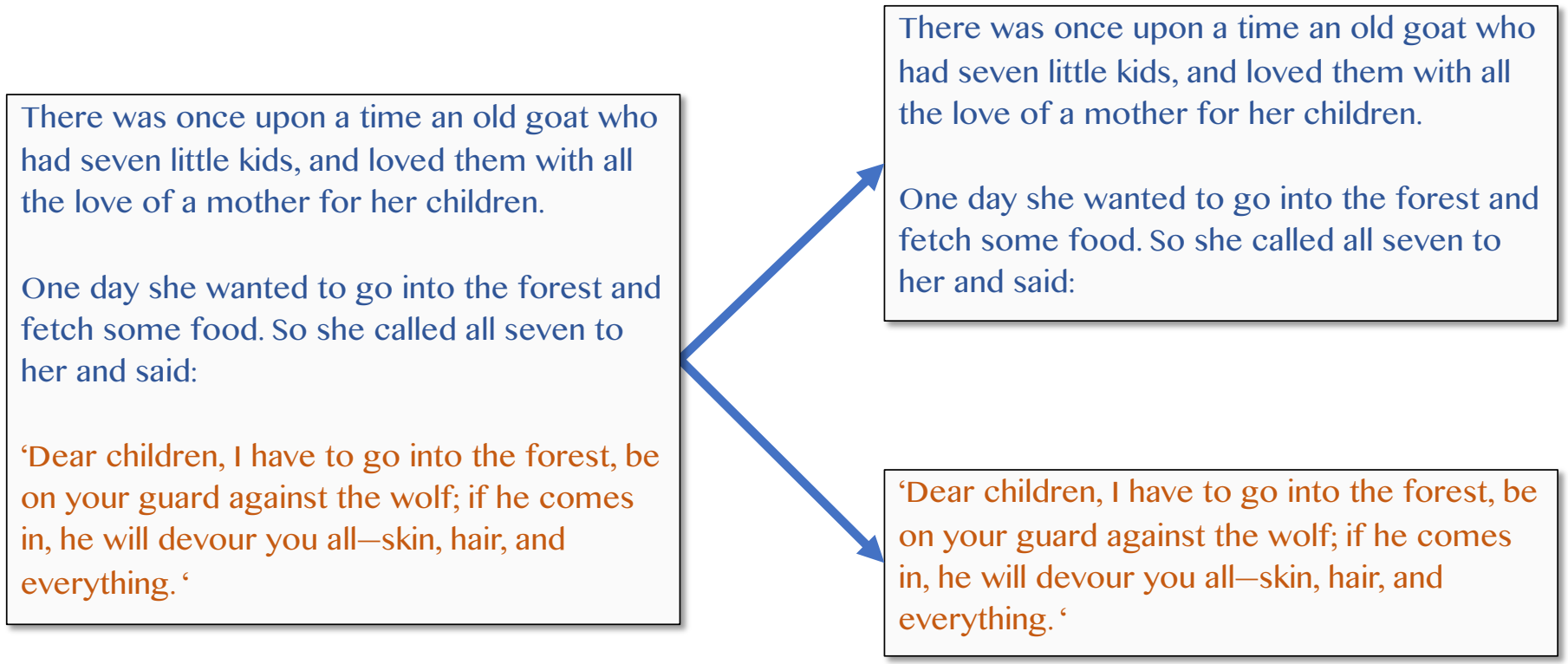
Generative **P**retrained **T**ransformer (GPT)

There was once upon a time an old goat who had seven little kids, and loved them with all the love of a mother for her children.

One day she wanted to go into the forest and fetch some food. So she called all seven to her and said:

'Dear children, I have to go into the forest, be on your guard against the wolf; if he comes in, he will devour you all—skin, hair, and everything.'

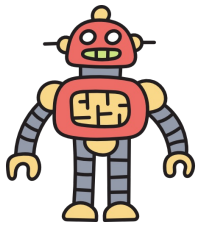
Generative Pretrained Transformer (GPT)



Generative Pretrained Transformer (GPT)

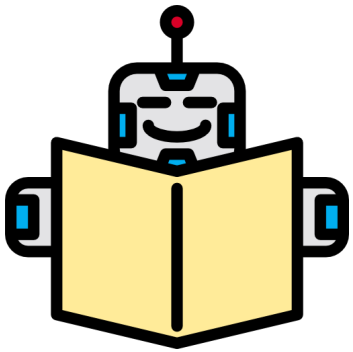
There was once upon a time an old goat who had seven little kids, and loved them with all the love of a mother for her children.

One day she wanted to go into the forest and fetch some food. So she called all seven to her and said:



‘Dear children, I have to go into the forest, be on your guard against the wolf; if he comes in, he will devour you all—skin, hair, and everything.‘

Instruction Tuning



Supervised FineTuning (SFT)

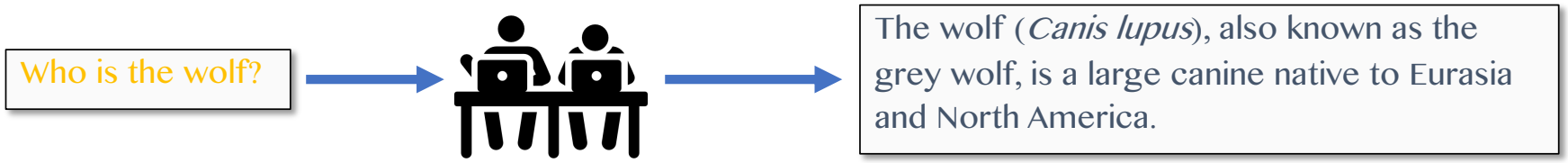
Who is the wolf?

The wolf (*Canis lupus*), also known as the grey wolf, is a large canine native to Eurasia and North America.

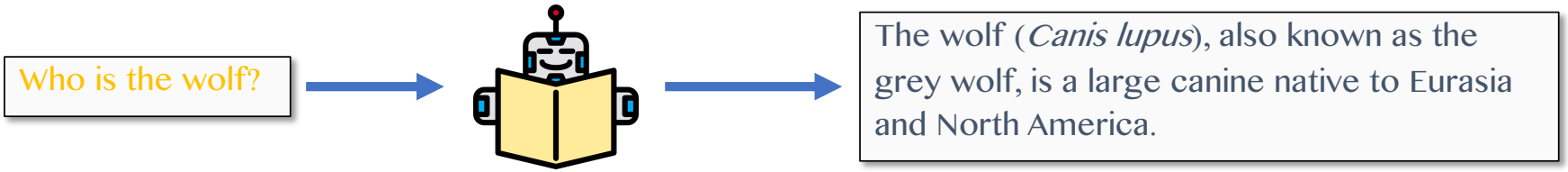
No no, I meant in the fairy tale!

Ah, my apologies. The Big Bad Wolf is a fictional wolf appearing in several cautionary tales, including some of Grimms' Fairy Tales. Versions of this character have appeared in numerous works, and it has become a generic archetype of a menacing predatory antagonist.

Supervised FineTuning (SFT)



Supervised FineTuning (SFT)



Supervised FineTuning (SFT)

Who is the wolf?

The wolf (*Canis lupus*), also known as the grey wolf, is a large canine native to Eurasia and North America.

No no, I meant in the fairy tale!

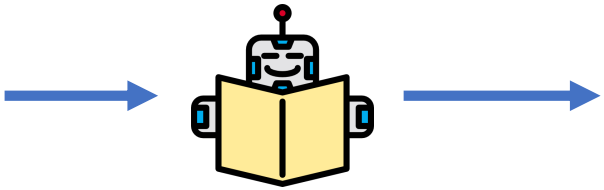
Ah, my apologies. The Big Bad Wolf is a fictional wolf appearing in several cautionary tales, including some of Grimms' Fairy Tales. Versions of this character have appeared in numerous works, and it has become a generic archetype of a menacing predatory antagonist.

Supervised FineTuning (SFT)

Who is the wolf?

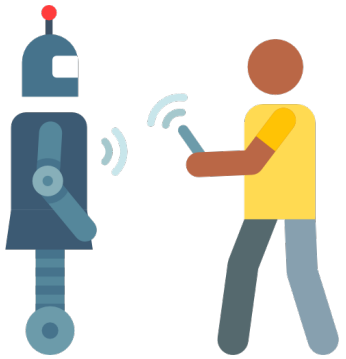
The wolf (*Canis lupus*), also known as the grey wolf, is a large canine native to Eurasia and North America.

No no, I meant in the fairy tale!



Ah, my apologies. The Big Bad Wolf is a fictional wolf appearing in several cautionary tales, including some of Grimms' Fairy Tales. Versions of this character have appeared in numerous works, and it has become a generic archetype of a menacing predatory antagonist.

Aligning with Human Preferences



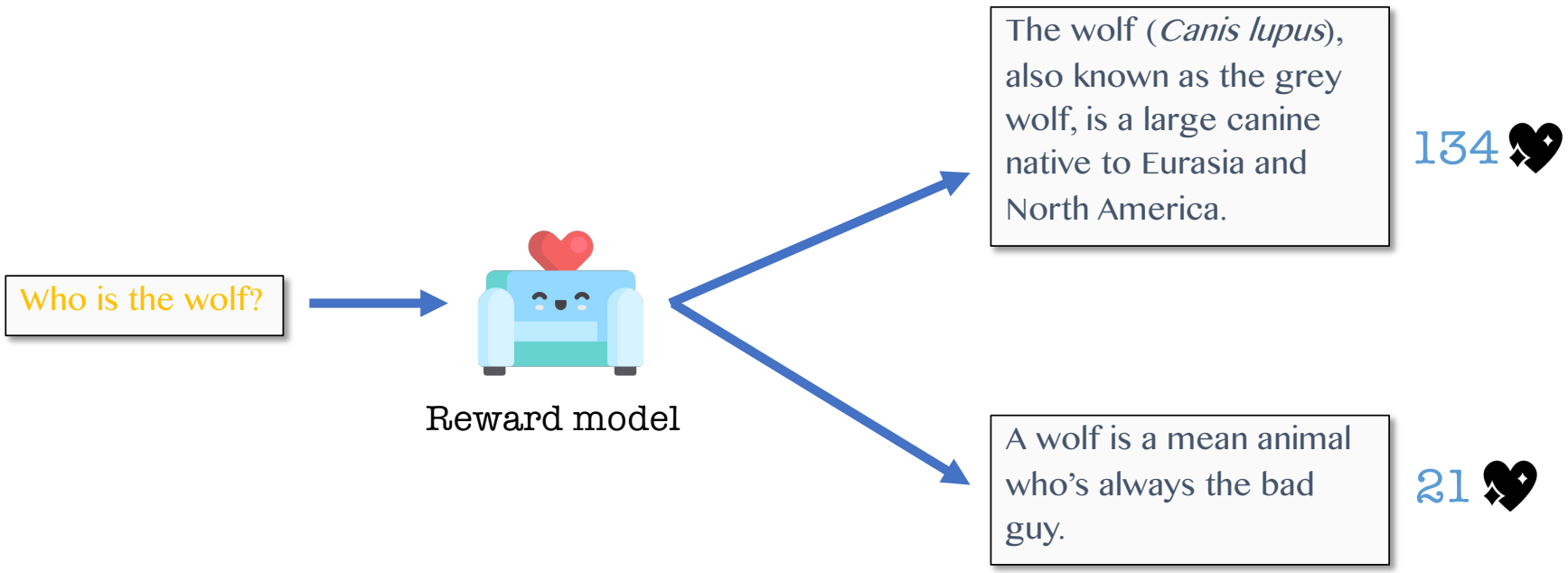
Reinforcement Learning with Human Feedback (RLHF)

Who is the wolf?

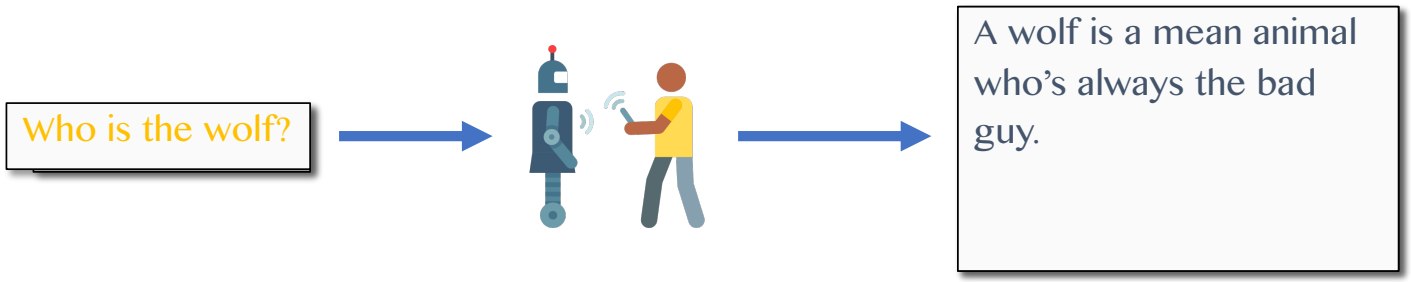
The wolf (*Canis lupus*), also known as the grey wolf, is a large canine native to Eurasia and North America.

A wolf is a mean animal who's always the bad guy.

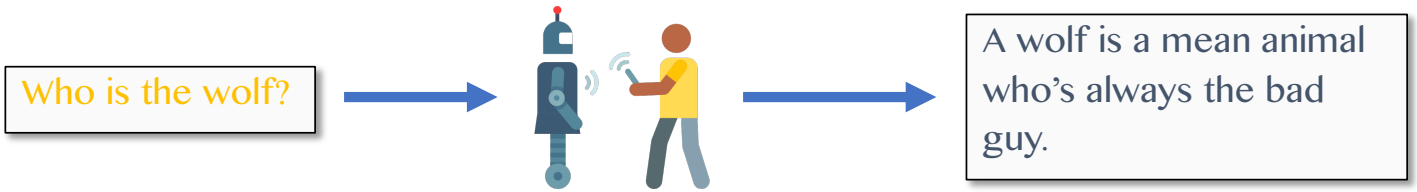
Reinforcement Learning with Human Feedback (RLHF)

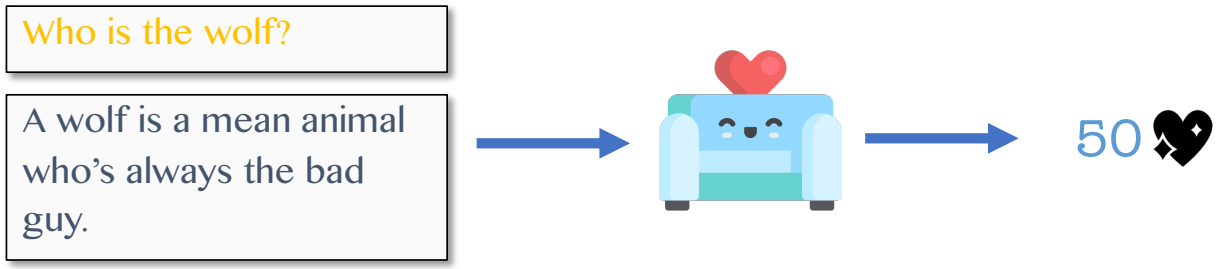


Reinforcement Learning with Human Feedback (RLHF)

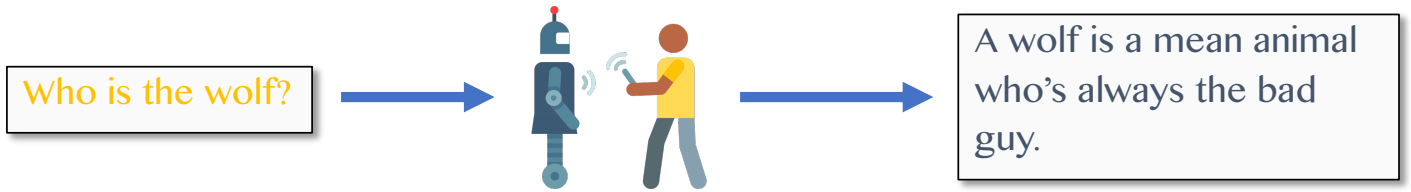


Reinforcement Learning with Human Feedback (RLHF)

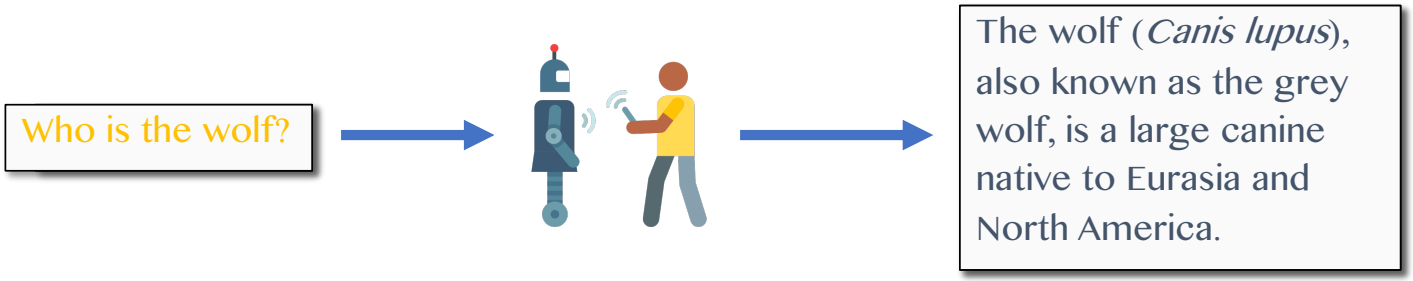




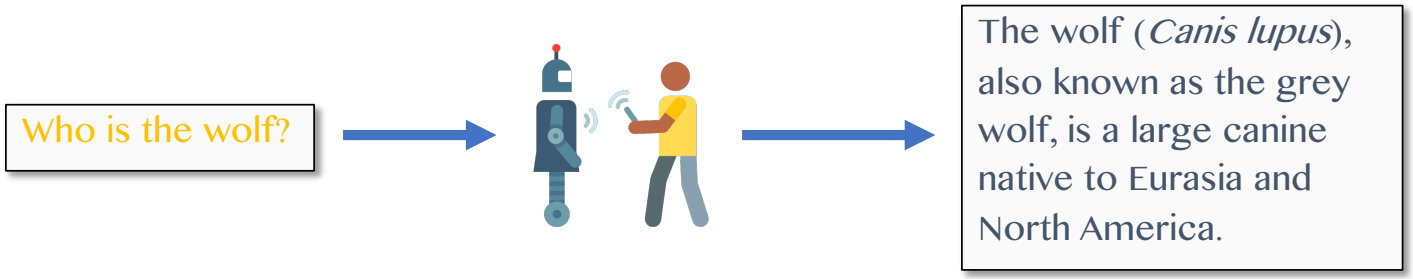
Reinforcement **L**earning with **H**uman **F**eedback (RLHF)

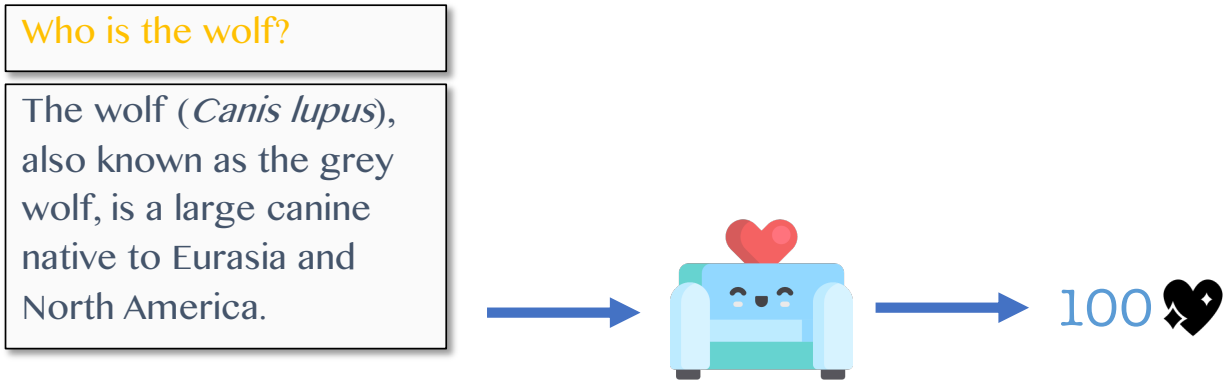


Reinforcement Learning with Human Feedback (RLHF)

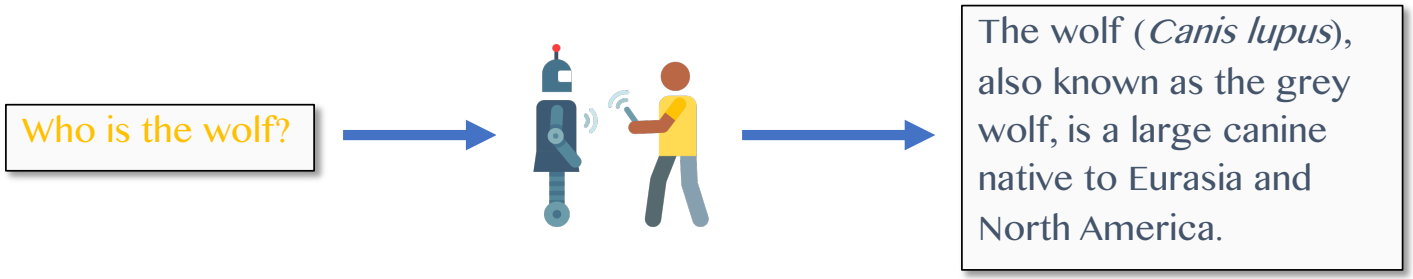


Reinforcement Learning with Human Feedback (RLHF)

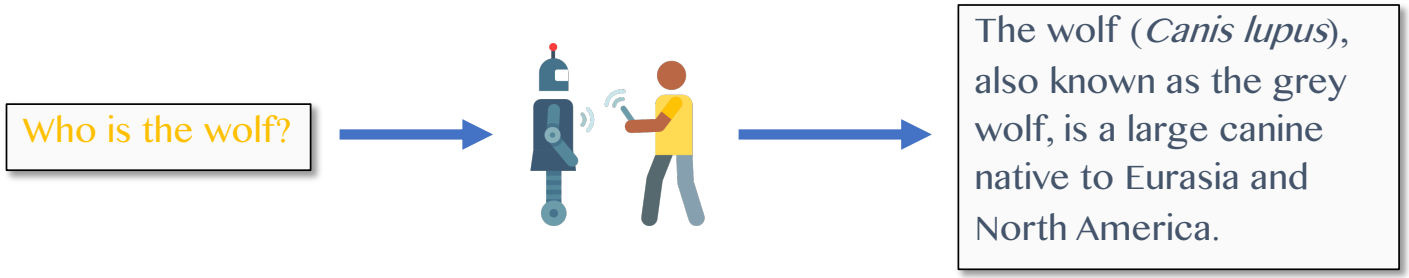




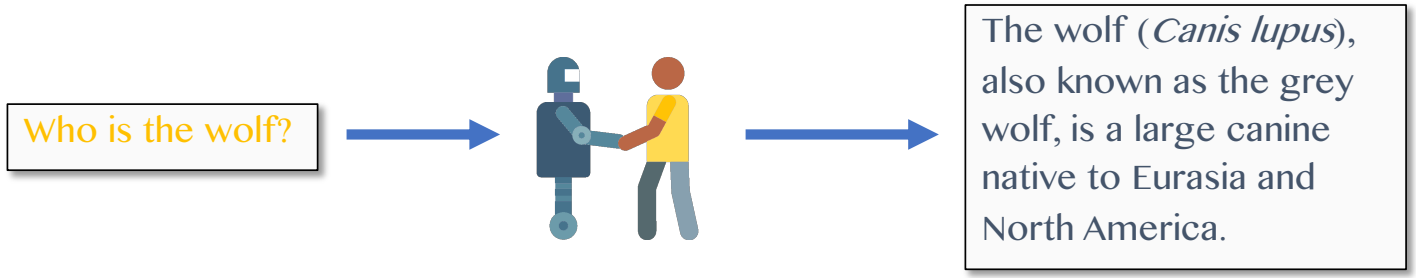
Reinforcement Learning with Human Feedback (RLHF)



Reinforcement Learning with Human Feedback (RLHF)



Reinforcement Learning with Human Feedback (RLHF)



Recent Improvements of RLHF

- **D**irect **P**reference **O**ptimisation (DPO)
  - This method allows one to skip the reward model and reinforcement learning, and instead directly optimise the LLM with the preference data [1]
- **K**ahneman-**T**versky **O**ptimisation (KTO)
  - This method also skips the reward model and reinforcement learning, like DPO, but only requires “x is good/bad” labels rather than “x is better than y” preference data [2]

[1] Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." arXiv preprint arXiv:2305.18290 (2023).

[2] https://contextual.ai/better-cheaper-faster-llm-alignment-with-kto

Model Merging

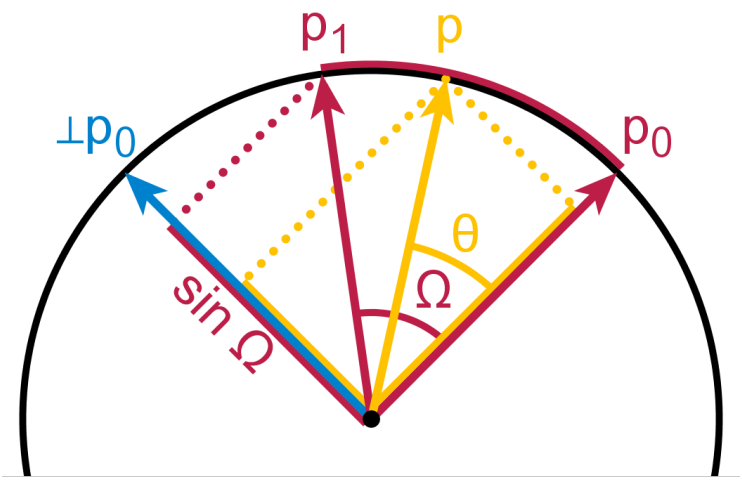
Model Merging

- Combines the weights of multiple models
- Typically requires that all models are built on the same architecture, and trained with the same tokeniser
- Some methods out there:
  - **S**pherical **L**in**e**a**r** Inter**p**olation (SLERP)
  - Task Arithmetic
  - **T**r**i**m and **E**lect **S**ign (TIES)
  - **D**rop **a**nd **Re**scale (DARE)

See hf.co/collections/osanseviero/model-merging-65097893623330a3a51ead66 for a large collection of papers related to model merging

SLERP Merging

- A classical method that uses linear interpolation along the unit sphere

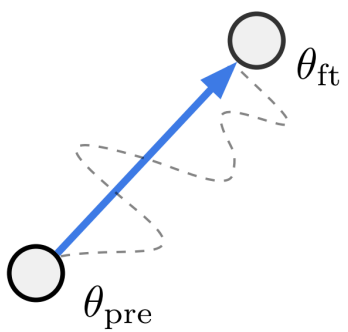


Shoemake, Ken. "Animating rotation with quaternion curves." Proceedings of the 12th annual conference on Computer graphics and interactive techniques. 1985.

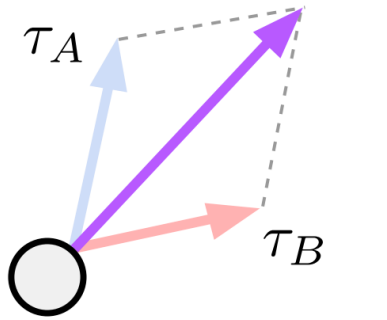
Task Arithmetic Merging

- Start with a finetuned model M_{ft} derived from a base model M_{pre}
- Compute the **task vector** *t* by subtracting the weights of M_{pre} from the weights of M_{ft}
- Now we can merge multiple models by simply adding their task vectors

$$\tau = \theta_{\text{ft}} - \theta_{\text{pre}}$$



$$\tau_{\text{new}} = \tau_A + \tau_B$$



Ilharco, Gabriel, et al. "Editing models with task arithmetic." The Eleventh International Conference on Learning Representations. 2022.

TIES Merging

- The TIES method extends task arithmetic and does 2 things:
  1. Reduce overlap between model weights
  2. Resolve sign conflicts in the overlapping model weights

- To do (1), they start with the task vectors and set entries to 0 if they're sufficiently small (keep a fixed percentage of entries)
- To do (2), we merge the task vectors by computing the mean for each entry if the entries have the same sign, or set them to 0 otherwise

Yadav, Prateek, et al. "Resolving Interference When Merging Models." arXiv preprint arXiv:2306.01708 (2023).

DARE Merging

- DARE is really a pre-processing method and can thus be combined with all the other methods
- DARE does 2 things before continuing with the chosen merge method, for a fixed $p \in [0, 1)$:
  1. Zero out the entries of each task vector, with probability p
  2. Renormalise the task vectors, by dividing with $1 - p$

[1] Yu, Le, et al. "Language models are super mario: Absorbing abilities from homologous models as a free lunch." arXiv preprint arXiv:2311.03099 (2023).

Mergekit

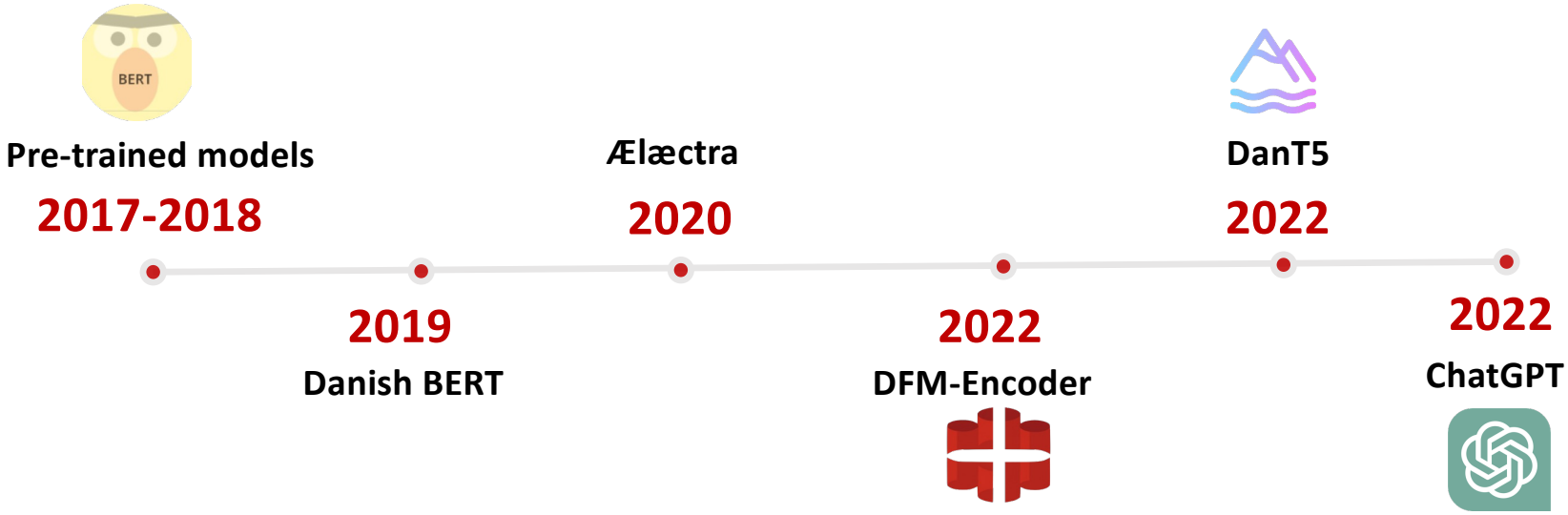
A handy Python package that implements many of the merge methods out there:

github.com/cg123/mergekit

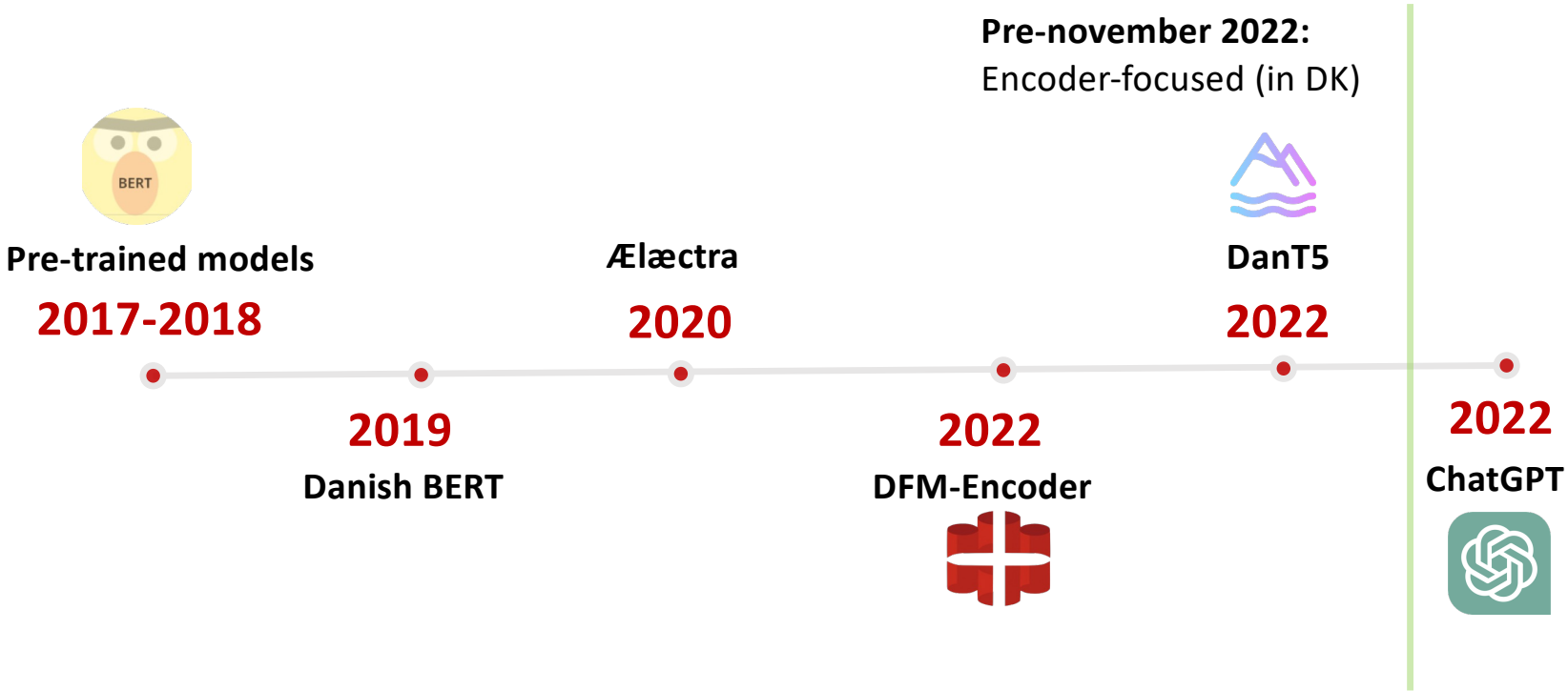
The Danish LM Scene

Models

The “Early” Danish NLP Scene



The ”Early" Danish NLP Scene



The (Chat)GPT Effect

AI SWEDEN

FinGPT-3

NorwAI

Danish Generative Models

Open models

DanskGPT-tiny

munin-7b-alpha

MISTRAL AI_

Meta Llama 2

Danish Generative Models

Open models

DanskGPT-tiny

munin-7b-alpha

MISTRAL AI_

Meta Llama 2

Closed models

Bard

DanskGPT

Mainly large tech or motivated individuals!

Research Efforts

TrustLLM

- Large European initiative
- Open, trustworthy, Germanic LLMs

TrustLLM

Research Efforts

TrustLLM

- Large European initiative
- Open, trustworthy, Germanic LLMs

Danish Foundation Models

- Open collaboration within DK

The Danish LM Scene

Model Performance Evaluation

Performance Benchmarks for LLMs

- Focus will be on language generation and understanding
- Other benchmarks exist for Danish and Scandinavian language models
  - E.g. The Scandinavian Embedding Benchmark [1] for evaluating the quality of embedding models

[1] Not yet published, but available at: https://kennethenevoldsen.github.io/scandinavian-embedding-benchmark/

Why Use a Benchmark

- Benchmarks seek to estimate abilities of a model
  - Specific: Grammaticality
  - Specific: Ability to solve code problems
  - General: Latent factor for language understanding (NLU)
- Users are interested in:
  - Performance on target use-case
    - Given a set of contraints (privacy, speed, bias, etc.)
  - **Not** NLU
- That being said we expect:
  - Performance on NLU~ target use-case
  - Or at least correlated with that of similar tasks

Mainland Scandinavian NLU

Tasks:

- Information extraction (named entity recognition)
- Grammaticality (linguistic acceptability)
- Question answering (extractive QA)
- Text classification (sentiment classification)

Mainland Scandinavian NLG

Tasks:

- Information extraction (named entity recognition)
- Grammaticality (linguistic acceptability)
- Question answering (extractive QA)
- Text classification (sentiment classification)
- Summarisation (news summarisation)
- Knowledge (translated multiple choice datasets: MMLU and ARC)
- Common sense reasoning (translated multiple choice dataset: HellaSwag)

→ Highly useful, but not without limitations

Mainland Scandinavian NLU

Rank	Model ID	Parameters	Vocabulary Size	Context	Speed	Score ▾	DA
1=	ltg/norbert3-large	354	50	508	5,048 ± 824 / 1,354 ± 429	67.78 ± 1.45	60.51 ± 1.71
1=	gpt-4-0613 (few-shot, val)	unknown	100	8192	1,244 ± 510 / 3,515 ± 848	67.34 ± 2.44	61.57 ± 2.22
2=	danish-foundation-models/encoder-large-v1	355	50	512	6,671 ± 1,380 / 1,497 ± 482	64.31 ± 3.02	62.39 ± 1.82
2=	KennethEnevoldsen/dfm-sentence-encoder-large-1	355	50	512	6,245 ± 1,260 / 1,416 ± 453	64.24 ± 2.08	62.35 ± 2.28
2=	KennethEnevoldsen/dfm-sentence-encoder-large-2	355	50	512	6,569 ± 1,320 / 1,492 ± 476	64.24 ± 2.32	62.98 ± 2.07
2=	google/rembert	576	250	256	3,355 ± 475 / 1,002 ± 312	63.70 ± 3.02	57.49 ± 4.10
2=	microsoft/mdeberta-v3-base	279	251	512	9,237 ± 1,562 / 2,258 ± 742	62.86 ± 1.86	56.37 ± 2.32

- Performance is not all:
  - Speed
  - Control (e.g. the ability to further train)
  - Openness

Table derived from https://scandeval.com/mainland-scandinavian-nlu/ 30/01/24

Mainland Scandinavian NLU

Rank	Model ID	Parameters	Vocabulary Size	Context	Speed	Score ▾	DA
1=	ltg/norbert3-large	354	50	508	5,048 ± 824 / 1,354 ± 429	67.78 ± 1.45	60.51 ± 1.71
1=	gpt-4-0613 (few-shot, val)	unknown	100	8192	1,244 ± 510 / 3,515 ± 848	67.34 ± 2.44	61.57 ± 2.22
2=	danish-foundation-models/encoder-large-v1	355	50	512	6,671 ± 1,380 / 1,497 ± 482	64.31 ± 3.02	62.39 ± 1.82
2=	KennethEnevoldsen/dfm-sentence-encoder-large-1	355	50	512	6,245 ± 1,260 / 1,416 ± 453	64.24 ± 2.08	62.35 ± 2.28
2=	KennethEnevoldsen/dfm-sentence-encoder-large-2	355	50	512	6,569 ± 1,320 / 1,492 ± 476	64.24 ± 2.32	62.98 ± 2.07
2=	google/rembert	576	250	256	3,355 ± 475 / 1,002 ± 312	63.70 ± 3.02	57.49 ± 4.10
2=	microsoft/mdeberta-v3-base	279	251	512	9,237 ± 1,562 / 2,258 ± 742	62.86 ± 1.86	56.37 ± 2.32

- Conclusions*:
  - Targeted models can compete with LLMs
    - Considering openness, control and speed outperforms them
  - *But…

Mainland Scandinavian NLG

- … more bleak for open-source generative models

Rank	Model ID	Parameters	Vocabulary Size	Context	Speed	Score ▾	DA
1	gpt-3.5-turbo-0613 (few-shot, val)	unknown	100	4095	1,344 ± 455 / 4,023 ± 590	58.52 ± 2.42	56.72 ± 2.44
2	RJuro/munin-neuralbeagle-7b (few-shot)	7242	32	32768	2,499 ± 461 / 783 ± 246	48.54 ± 1.91	49.02 ± 1.72
3=	timpal0l/Mistral-7B-v0.1-flashback-v2 (few-shot)	7242	32	32768	2,505 ± 465 / 774 ± 242	42.00 ± 1.91	39.68 ± 1.89
3=	mistralai/Mistral-7B-v0.1 (few-shot)	7242	32	32768	2,657 ± 524 / 880 ± 278	40.30 ± 2.15	39.60 ± 1.94
4	danish-foundation-models/munin-7b-alpha (few-shot)	7242	32	32768	3,019 ± 480 / 1,048 ± 317	37.50 ± 2.49	39.56 ± 2.70
5	AI-Sweden-Models/gpt-sw3-6.7b-v2 (few-shot)	7111	64	2048	2,351 ± 448 / 707 ± 216	26.67 ± 2.30	23.65 ± 2.02
6	mhenrichsen/danskgpt-tiny-chat (few-shot)	1100	32	2048	1,745 ± 978 / 686 ± 159	19.73 ± 2.39	20.51 ± 1.85

- **Note:** Models are still being benchmarked so this is not a full picture
- Open Questions:
  - Performance of models such as Mixtral
  - Ranking in relation English benchmarks
  - Influence of combining/merging/mixing models (e.g. as seen in MoE)
  - Influence of RAG solutions

Current Limitations

- ScandEval
  - 3/8 datasets are translated
    - To what degree is performance on a translated dataset similar to a native dataset?
- Difference in evaluation and application
  - Popular model evaluations:
    - Code generation
    - …
  - Danish use-cases:
    - Tax system, welfare, Danish conditions
    - *“Er jeg berretiget til kørselfradrag hvis jeg kører fra Tyskland på arbejde?”*

Developing Danish LLMs

Danish Foundation Models

Goals

1. To **develop and maintain state-of-the-art models** for the Danish language.
2. To extensively **validate foundation models** for Danish on a representative set of tasks.
3. To maintain a **high standard of documentation** of models.
4. To **open-source** not only the models but also all components required for reproducibility such as pre-processing, training, and validation code.

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-SW3	✓	✓	✓	✓	🇸🇪 (🇺🇸 🇩🇰 🇳🇴)	✗
Munin-7b-alpha	✓	✓	✓	✗	🇩🇰 (🌐)	✗*
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	✓	✓	✗	✗	🇩🇰	✓†

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-SW3	✓	✓	✓	✓	🇸🇪 (🇺🇸 🇩🇰 🇳🇴)	✗
Munin-7b-alpha	✓	✓	✓	✗	🇩🇰 (🌐)	✗*
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	✓	✓	✗	✗	🇩🇰	✓†

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-SW3	✓	✓	✓	✓	🇸🇪 (🇺🇸 🇩🇰 🇳🇴)	✗
Munin-7b-alpha	✓	✓	✓	✗	🇩🇰 (🌐)	✗*
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	✓	✓	✗	✗	🇩🇰	✓†

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-SW3	✓	✓	✓	✓	🇸🇪 (🇺🇸 🇩🇰 🇳🇴)	✗
Munin-7b-alpha	✓	✓	✓	✗	🇩🇰 (🌐)	✗*
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	✓	✓	✗	✗	🇩🇰	✓†

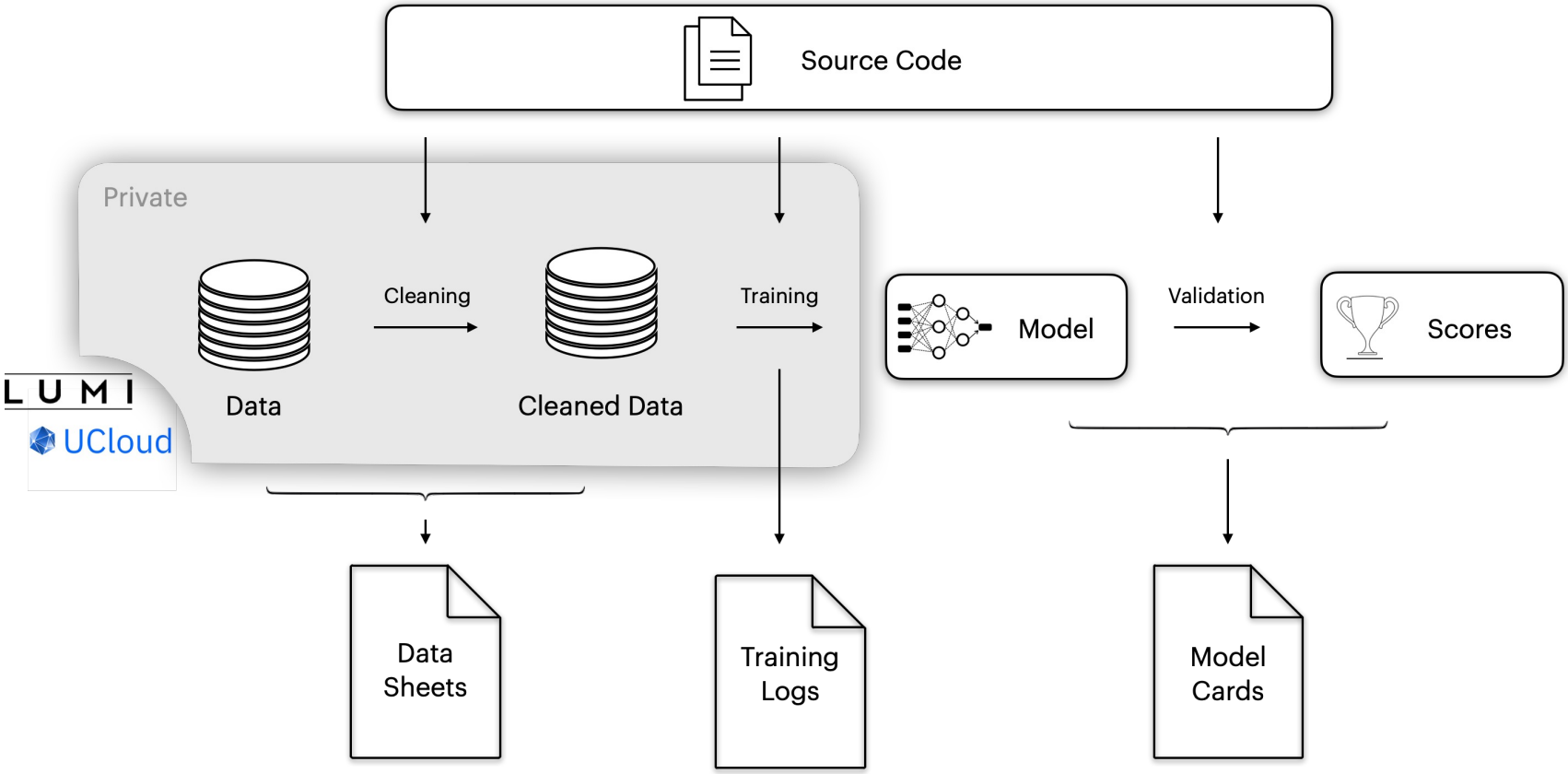
Dataset

Dataset: Previous dataset

- Similar size to GPT-3 style models (e.g., Gopher)
- More than a 100x increase in pre-training data
- Netarkivet: 2006-2016
- With cleaning

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	✗	✓

Data Safety and Governance



Collect data and data collaborators

- What we can do and what we do
  - New law → widely collect data for research
    - Unsure how this influence derivatives of the data (models)
  - Collaboration → open models
- Data collaborators
  - Infomedia (agreements with specific newspapers)
  - The Royal Library
- Data agreement is currently processing
  - Rigsarkivet, Sundhed.dk, lex.dk, Stadsarkivet
- Looking into agreements with:
  - MUNI, Dr.dk, etc.

*Note, that due to the ongoing court case between NY Times vs OpenAI and Microsoft, some data agreements have been temporarily paused and others have been waiting to finalize until after the agreement.

Cleaning Datasets

Filters	Specific Requirement
Heuristic Filters	
Danish Stopwords	Contain at least 2 Danish stopwords (using SpaCy v.3.1.4 list)
Mean Word Length	Between 3 and 10 characters
Token Length	Between 50 and 100,000 tokens
Character Count	Less than 5,000,000 characters
Alphabetic Character Percentage	At least 60% of words should an alphabetic character
Symbol-to-Word Ratio (Hashtags, Ellipsis)	Lower than 10%
Bullet Point Line Starts	Less than 90% of lines starting with a bullet point
Ellipsis Line Ends	Less than 30% of lines ending with an ellipsis
Repetitious Text (Character Percentage)	Less than 20% in duplicate lines; less than 20% in duplicate paragraphs
N-gram Frequency (2-4 grams)	Should constitute less than 20%, 18%, 16% of the text, respectively
N-gram Frequency (5-10 grams)	Should constitute less than less than 25%, 24%, 23%, 22%, 21%, 20%, respectively
Deduplication	
Jaccard Similarity (13-gram)	MinHash algorithm with < 80% similarity (128 permutations)

*Specification for Netarkivet, tuned to each subcorpus

Cleaning Datasets

- Heuristic data cleaning
  - Sparse evidence [1, 2]
  - Not always good [3]
- We want:
  - Data with the largest positive effect of model performance
    - What do we mean by model performance
  - [3] Propose one such solution
    - Selecting data based on predicted effect on selected task (benchmark)
    - This leads to the requirement of a ”validation” and ”test” benchmark
- → Can we create a proxy validation benchmark?
  - E.g., perplexity
  - (notably distinct from filtering on perplexity)

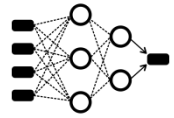
[1] Rae, Jack W., et al. "Scaling language models: Methods, analysis & insights from training gopher." arXiv preprint arXiv:2112.11446 (2021).
[2] Lee, Katherine, et al. "Deduplicating Training Data Makes Language Models Better." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.
[3] Engstrom, Logan, Axel Feldmann, and Aleksander Madry. "DsDm: Model-Aware Dataset Selection with Datamodels." arXiv preprint arXiv:2401.12926 (2024).

Next Steps

Data!

Evaluation!

Scale!



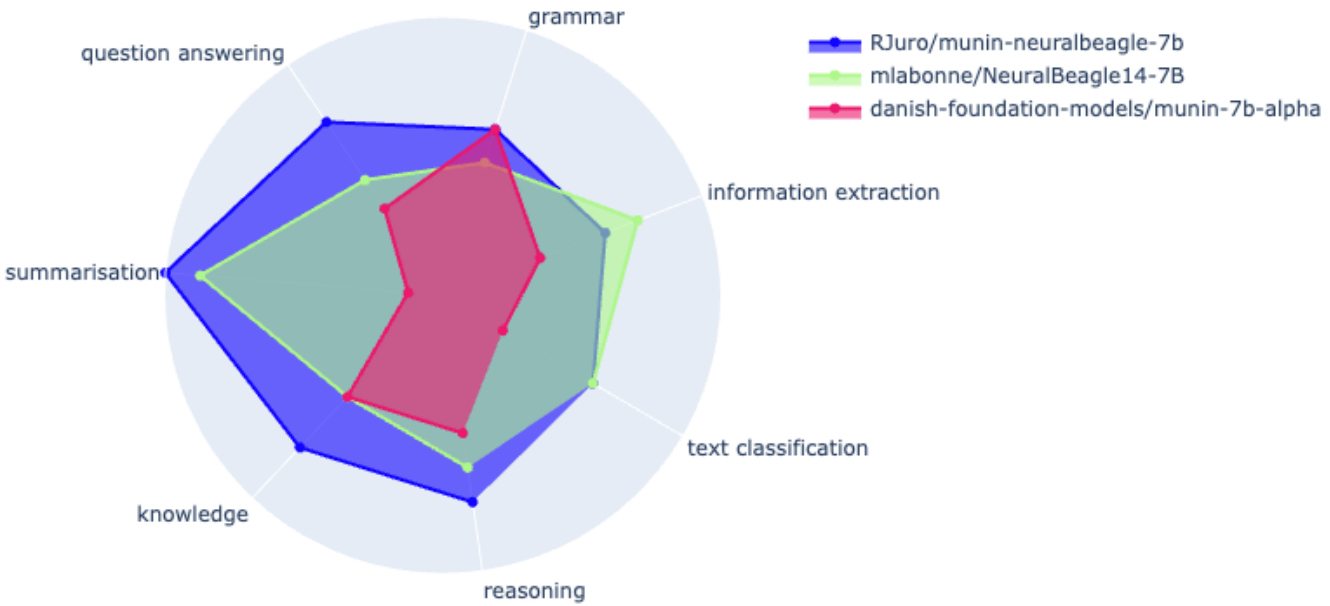
Model merging

- Best of both worlds?

DARE-TIES

Roman Jurowetzki

Model merging



Better than its constituent models!

Workshop!

1. Form 8 groups
2. Choose a “group notebook responsible person”
3. The ”group notebook responsible person” gets a link to your group’s GPU notebook

FYI: Notebooks are also available at github.com/alexandrainst/d3a-llm-workshop

Future Directions

New Munin model on the way, trained on >10x more data 🎉

Research questions:

1. How can model merging be used effectively to enhance models with new capabilities? What are the limits?
2. Can we change the tokenisation procedure to take similarities between languages into account?
   - The Scandinavian languages are good as a use case here

Thanks to the team!

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