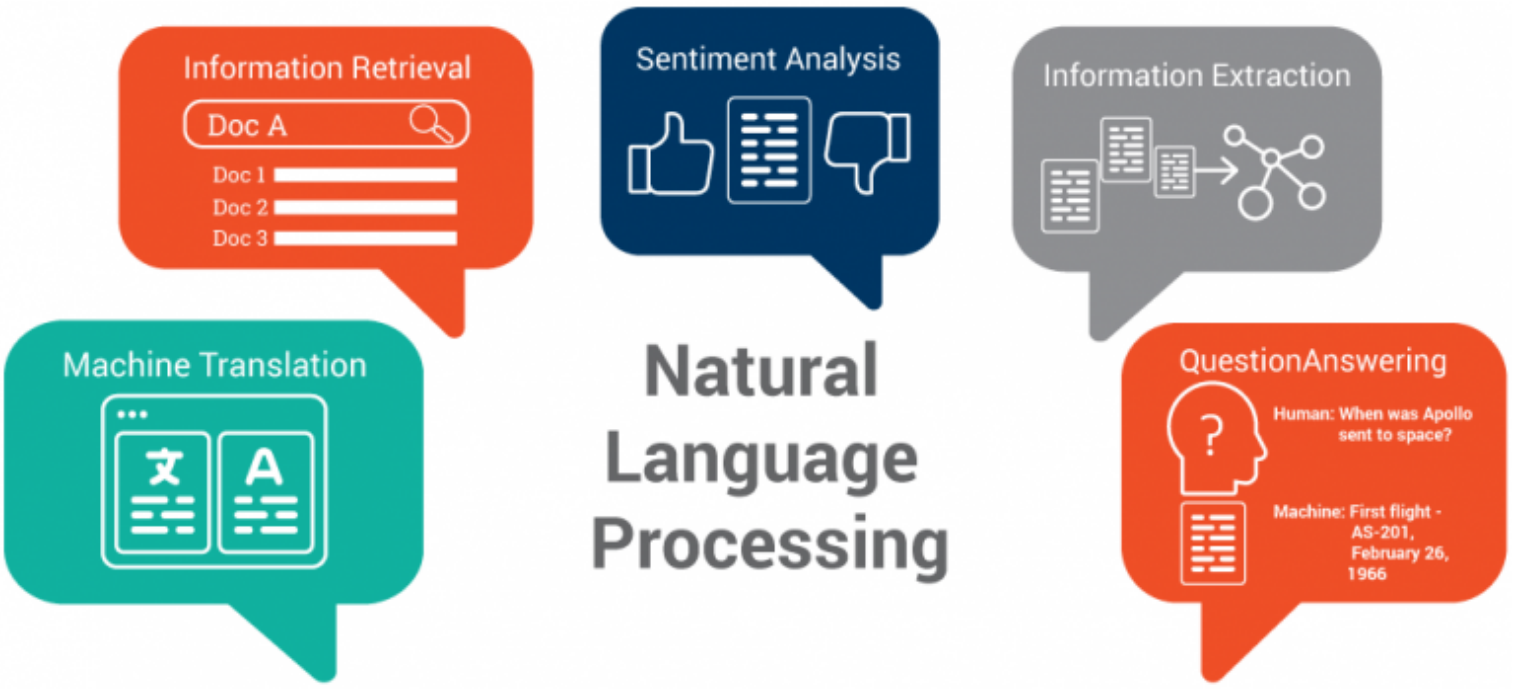
Kan store språkmodeller gjøres grønne?

Lilja Øvrelid

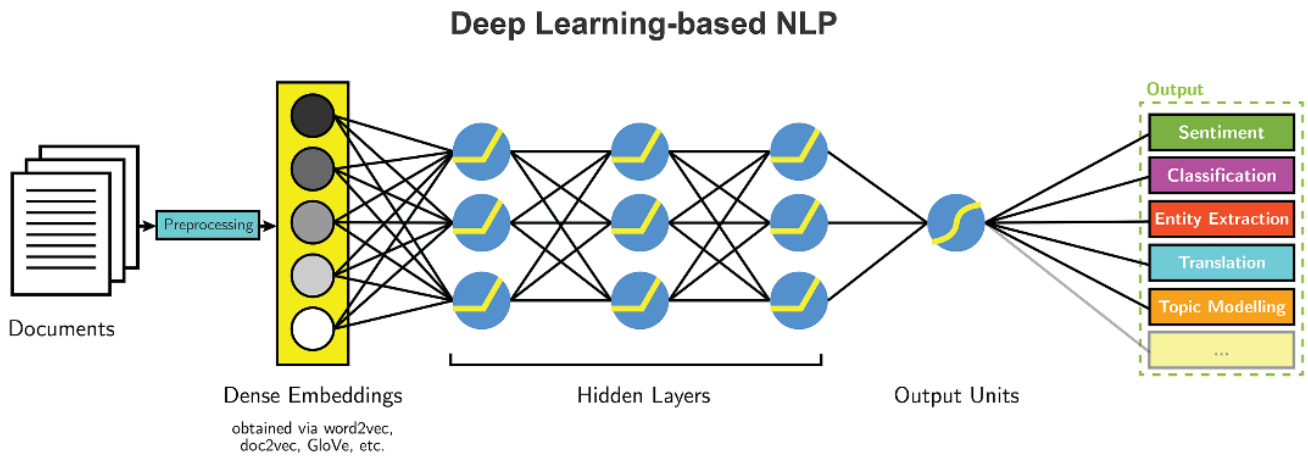
Language Technology Group
Institutt for Informatikk, UiO

SINTEF-seminar: Bærekraft og maskinlæring – Lar det seg forene?

Språkteknologi

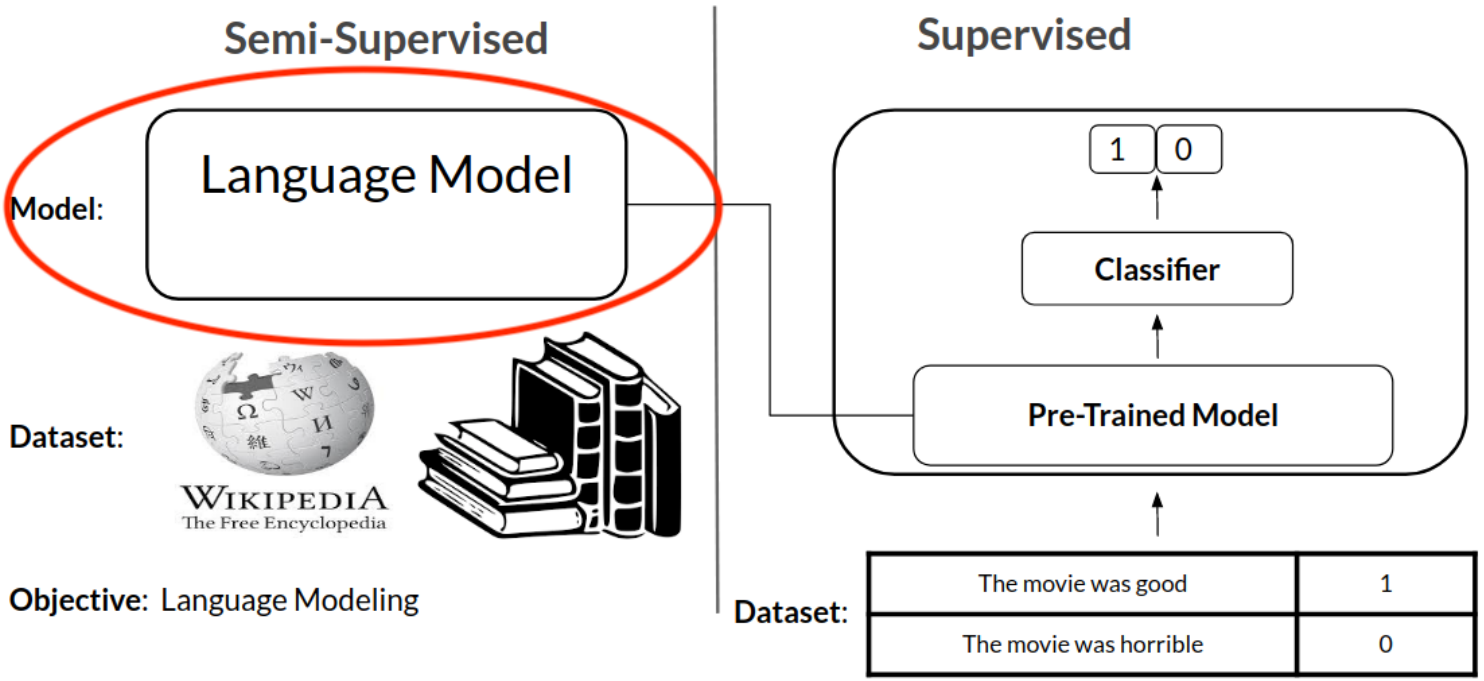


Språkteknologi i dag



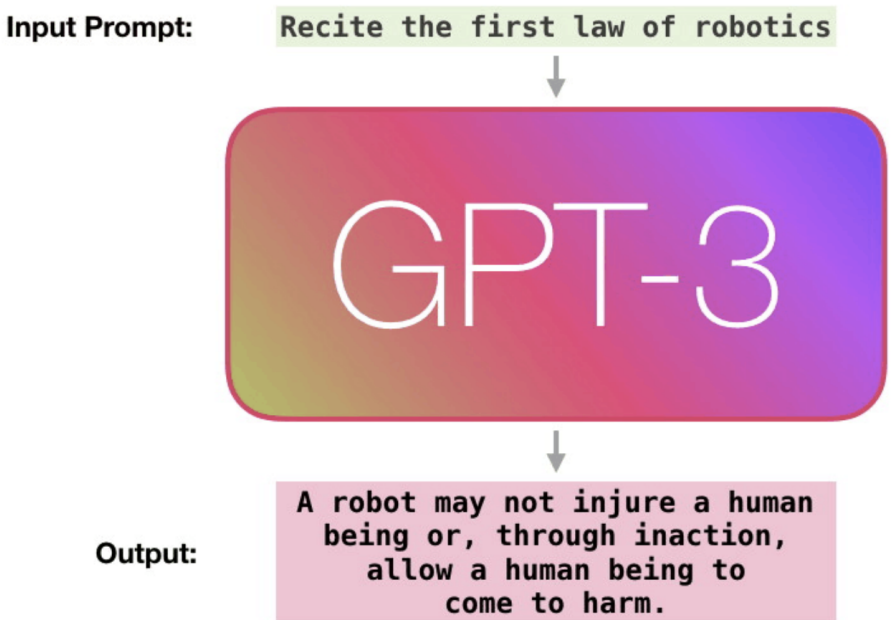
(https://ebookreading.net/view/book/EB9781787121423397.html)

Språkmodeller



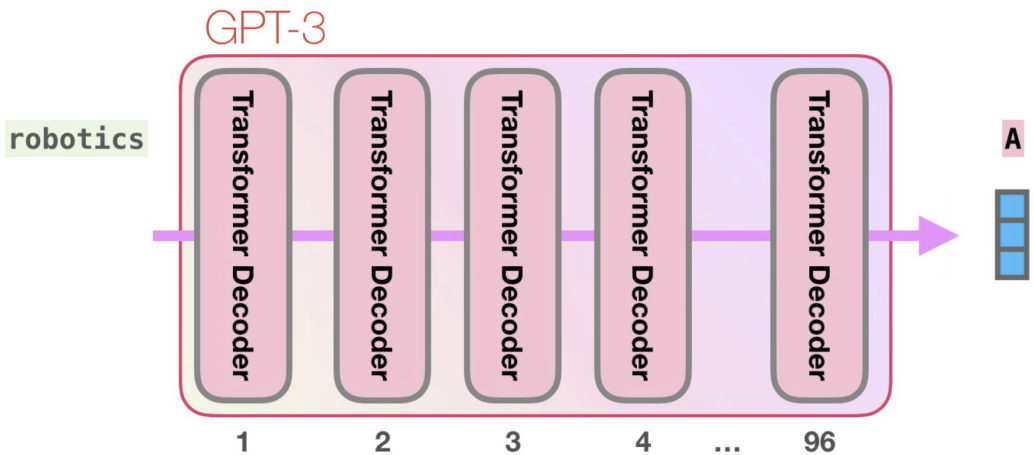
(https://blog.insightdatascience.com)

Tekstgenerering



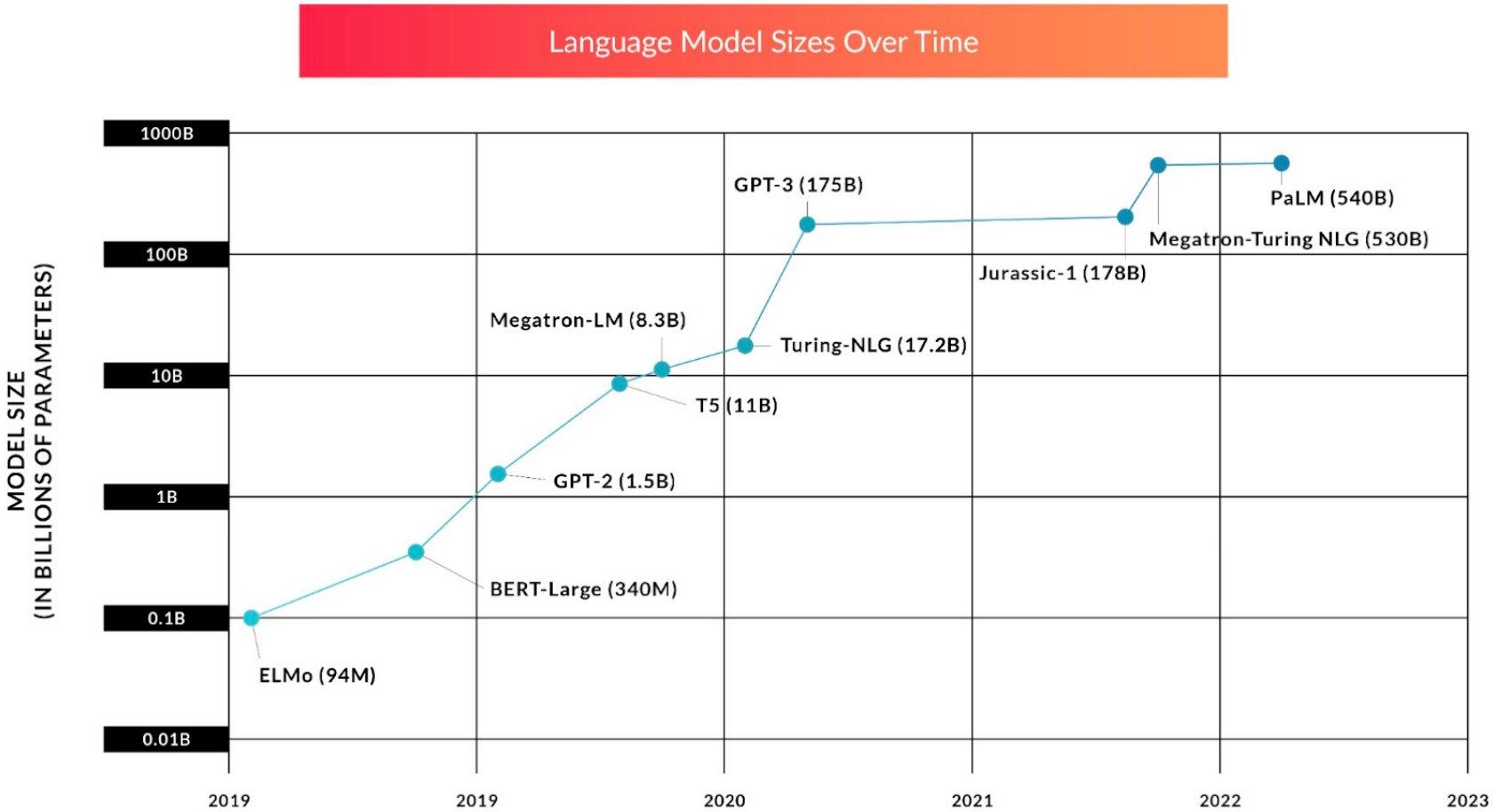
(https://jalammar.github.io/how-gpt3-works)

Tekstgenerering



(https://jalammar.github.io/how-gpt3-works)

Størrelse



https://www.assemblyai.com/

Størrelse på treningsdata (i antall ord)

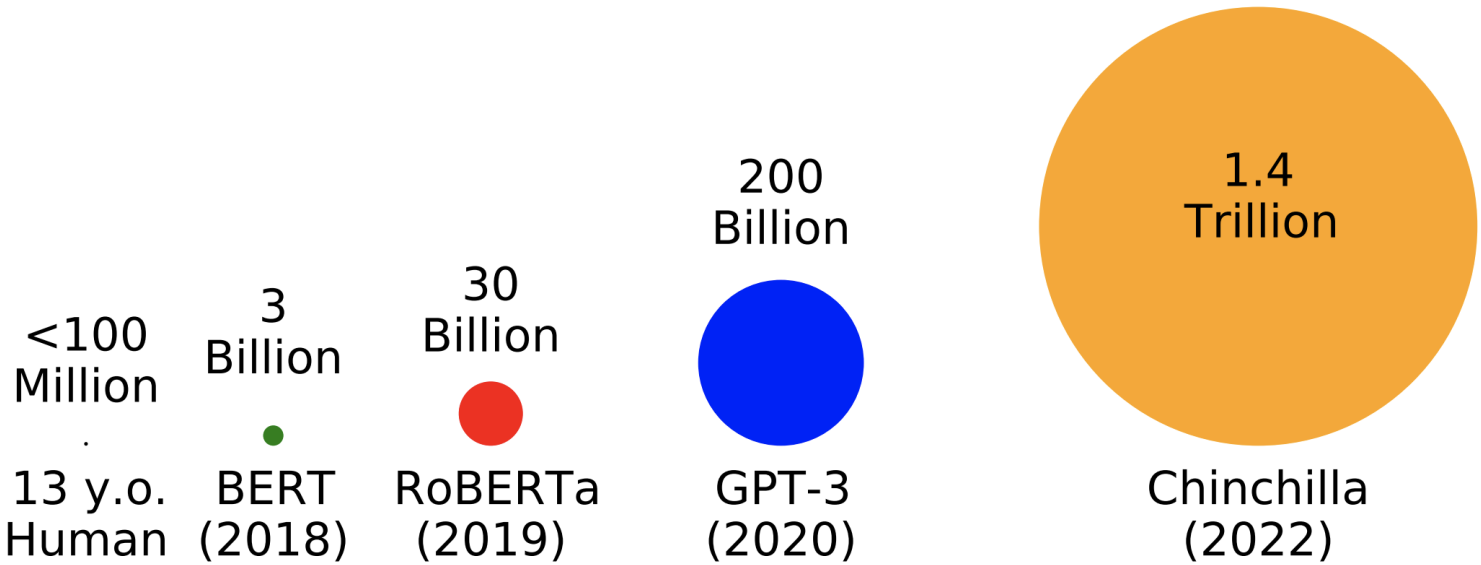
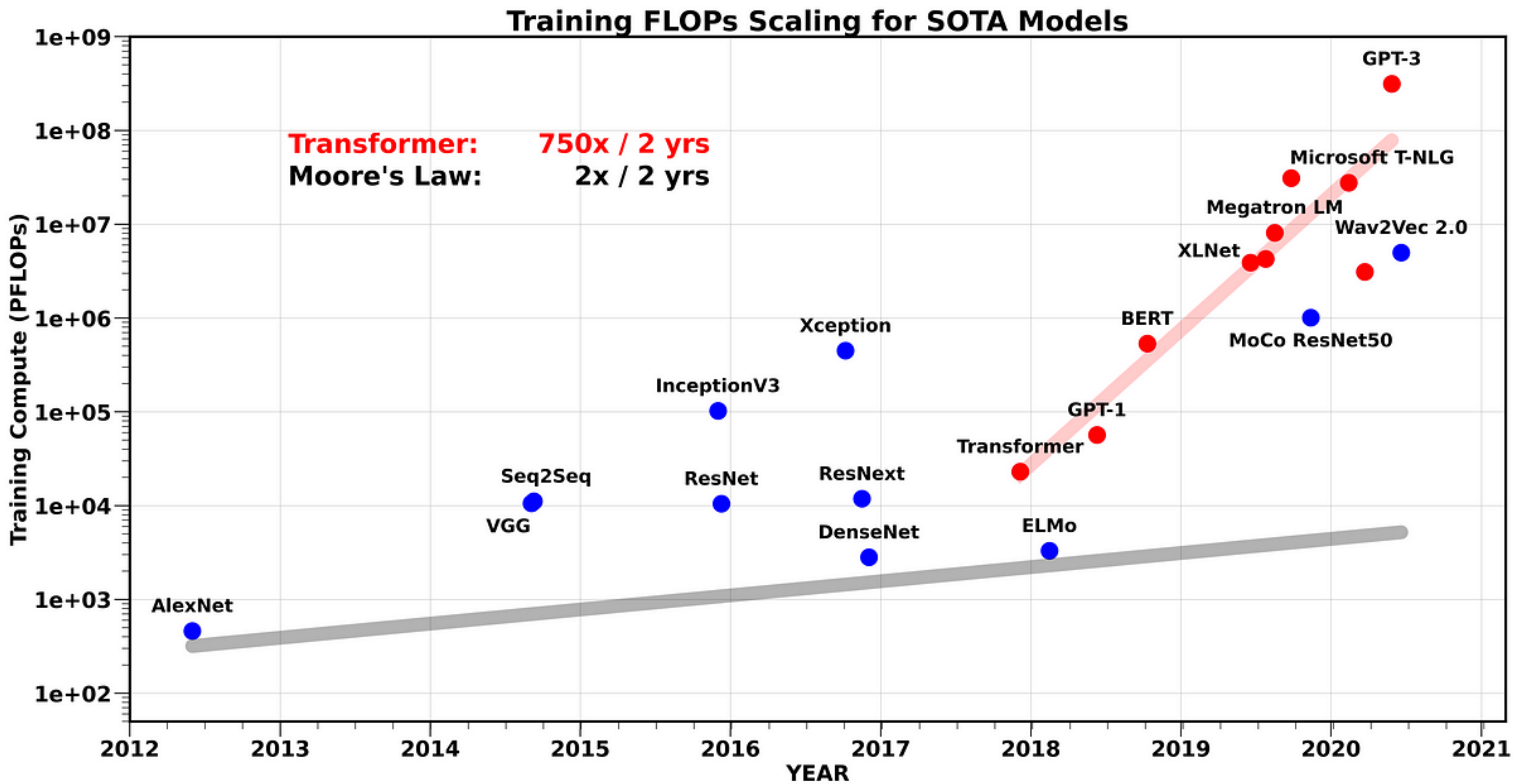


Image credits: babylm.github.io

Beregningskraft



Karbonavtrykk

Artificial Intelligence Index Report 2023

Chapter 2: Technical Performance
2.8 Environment

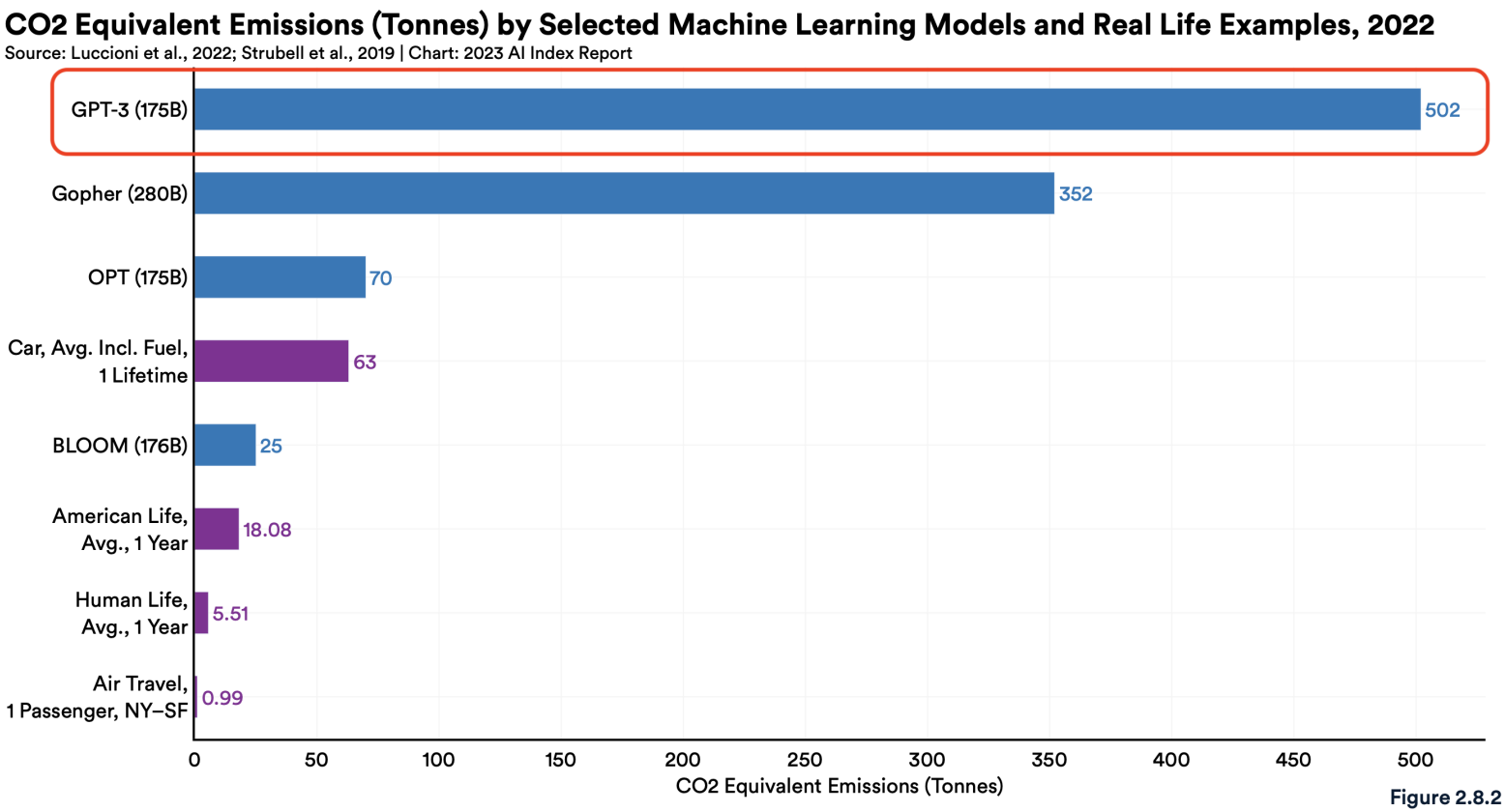


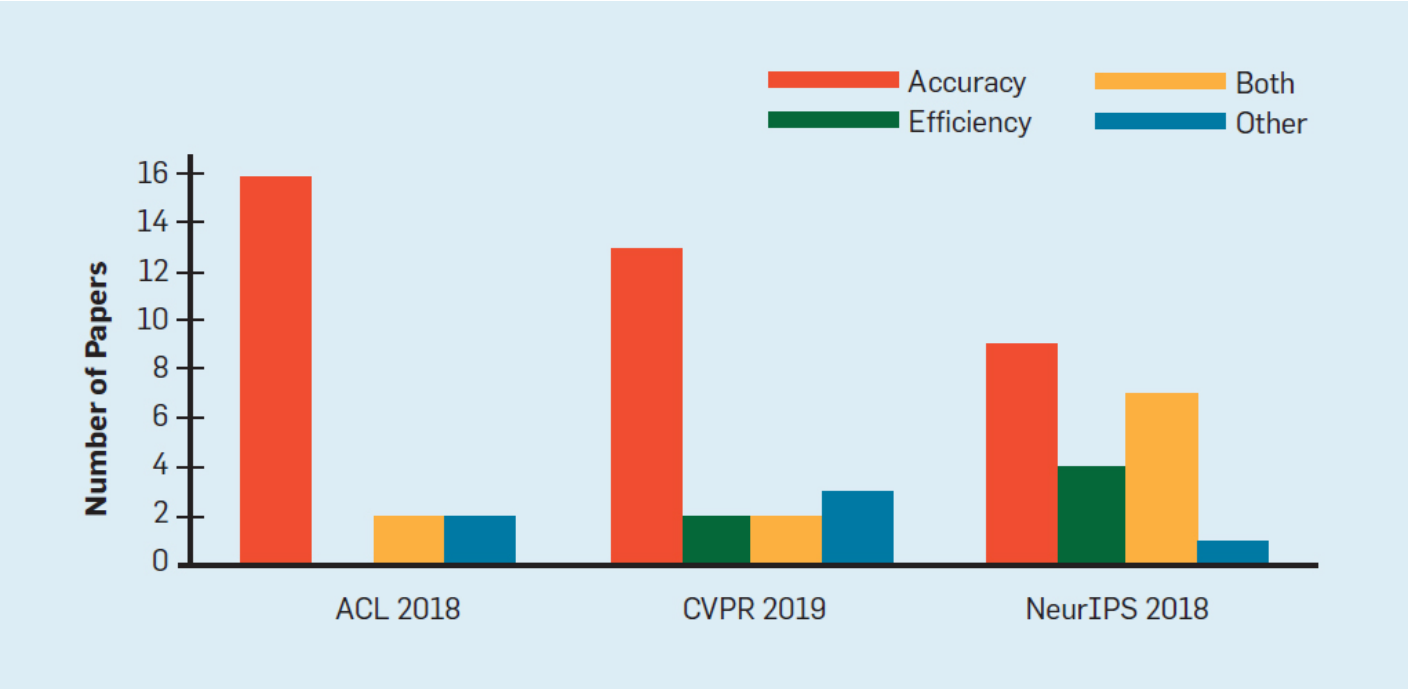
Figure 2.8.2

(Luccioni et al, 2022)

Mot grønnere språkmodeller

- Bedre rapportering av beregningsbudsjett
- Mer effektive språkmodeller

Rapportering



(Schwartz et al (2020): Green AI)

Rapportering

Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model

Alexandra Sasha Luccioni
Hugging Face
sasha.luccioni@hf.co

Sylvain Viguier
Graphcore
sylvainv@graphcore.ai

Anne-Laure Ligozat
LISN & ENSIIE
anne-laure.ligozat
@lisn.upsaclay.fr

Abstract

Progress in machine learning (ML) comes with a cost to the environment, given that training ML models requires significant computational resources, energy and materials. In the present article, we aim to quantify the carbon footprint of BLOOM, a 176-billion parameter language model, across its life cycle. We estimate that BLOOM's final training emitted approximately 24.7 tonnes of CO_2eq if we consider only the dynamic power consumption, and 50.5 tonnes if we account for all processes

Rapportering

ESTIMATING THE CARBON FOOTPRINT OF BLOOM, A 176B PARAMETER LANGUAGE MODEL

Alexandra Sasha Luccioni
Hugging Face
sasha.luccioni@hf.co

Sylvain Viguier
Graphcore
sylvainv@graphcore.ai

Anne-Laure Ligozat
LISN & ENSIIE
anne-laure.ligozat@lisn.upsaclay.fr

ABSTRACT

Progress in machine learning (ML) comes with a cost to the environment, given that training ML models requires significant computational resources, energy and materials. In the present article, we aim to quantify the carbon footprint of BLOOM, a 176-billion parameter language model, across its life cycle. We estimate that BLOOM's final training emitted approximately 24.7 tonnes of CO_2eq if we consider only the dynamic power consumption, and 50.5 tonnes if we account for all processes

Total training time	118 days, 5 hours, 41 min
Total number of GPU hours	1,082,990 hours
Total energy used	433,196 kWh
GPU models used	Nvidia A100 80GB
Carbon intensity of the energy grid	57 gCO_2eq/kWh

Mer effektive språkmodeller

Effektivitet (Wright et al, 2023):

- beregningskraft (compute): antall parametre, FLOPs, tid
- energibruk (kW/h)
- karbonutslipp (CO_2eq)

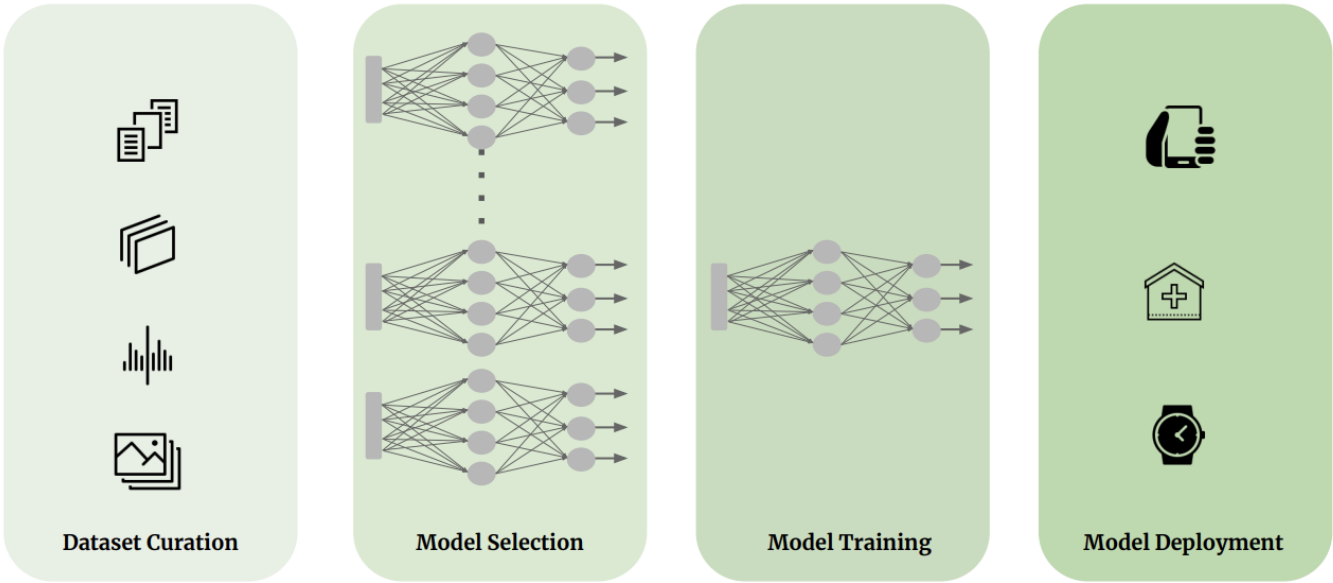
Mer effektive språkmodeller

Effektivitet (Wright et al, 2023):

- beregningskraft (compute): antall parametre, FLOPs, tid
- energibruk (kW/h)
- karbonutslipp (CO_2eq)

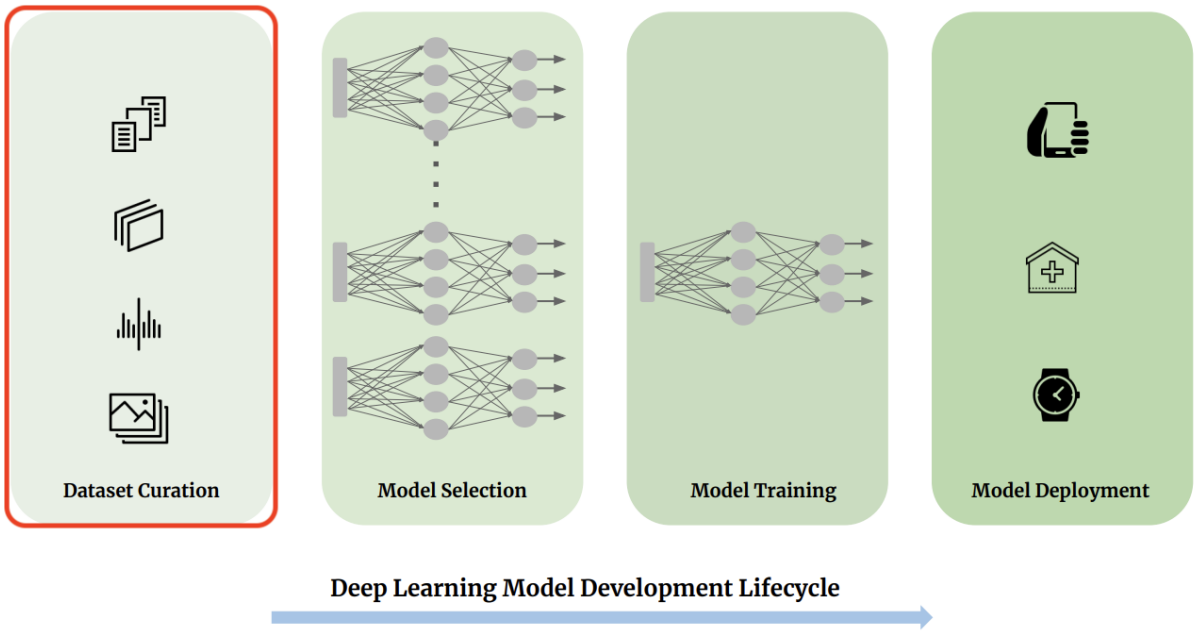
Mål: redusere kostnad (energi/karbonavtrykk) for beregninger

ML-modellens livssyklus



(Wright et al, 2023)

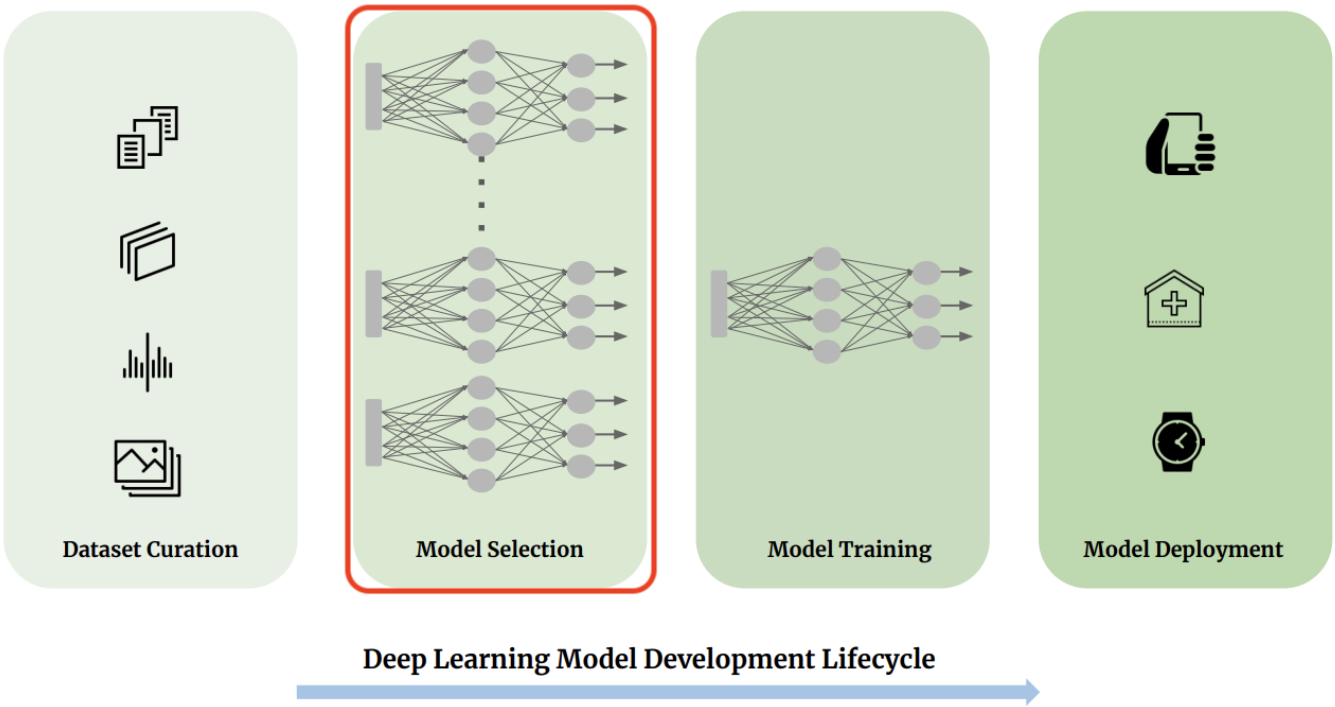
Effektivitet: data



- Datakvalitet: færre treningsinstanser
  - kuratering, filtrering, de-duplisering
- Bedre datautnyttelse
  - Feks curriculum learning

(Treviso et al, 2023)

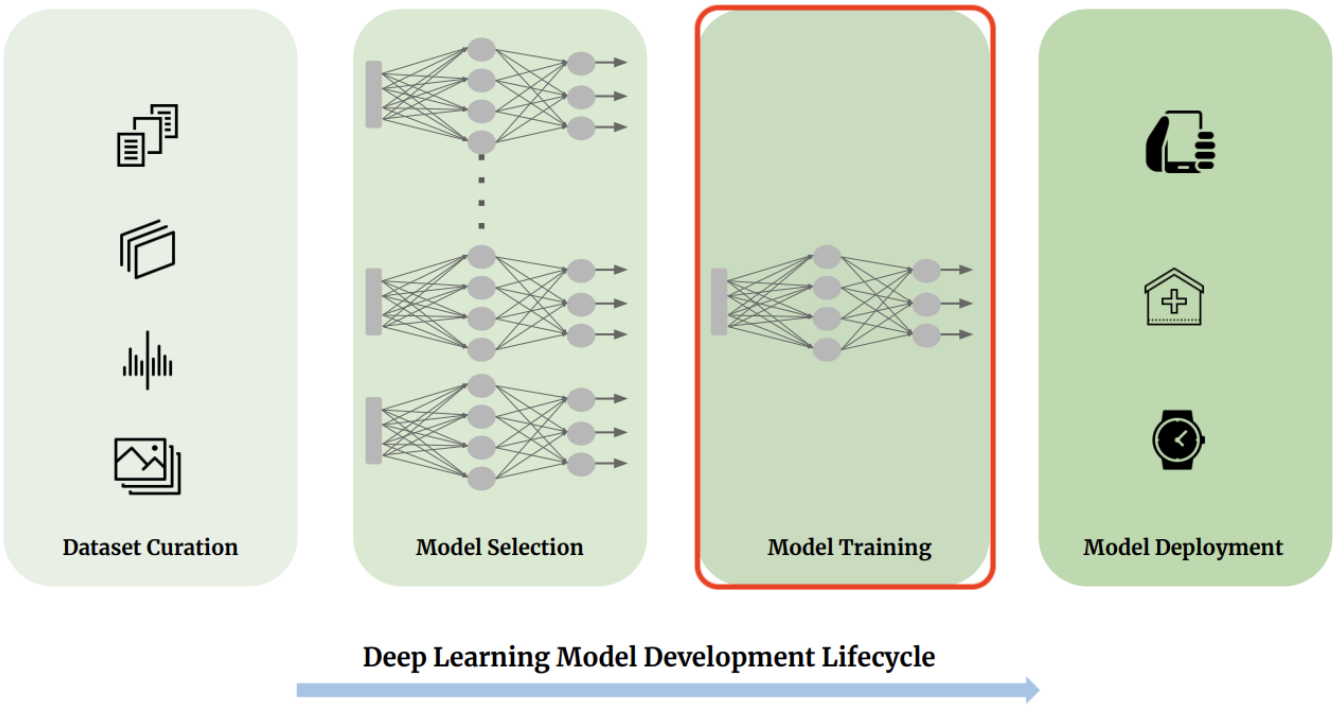
Effektivitet: modellvalg



- Mer effektiv hyperparametersøk (feks Bayesiansk optimering)
- Hyperparameter-transfer fra andre oppgaver

(Treviso et al, 2023)

Effektivitet: trening



Effektivitet: trening

Training Compute-Optimal Large Language Models

Jordan Hoffmann*, Sebastian Borgeaud*, Arthur Mensch*, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals and Laurent Sifre*

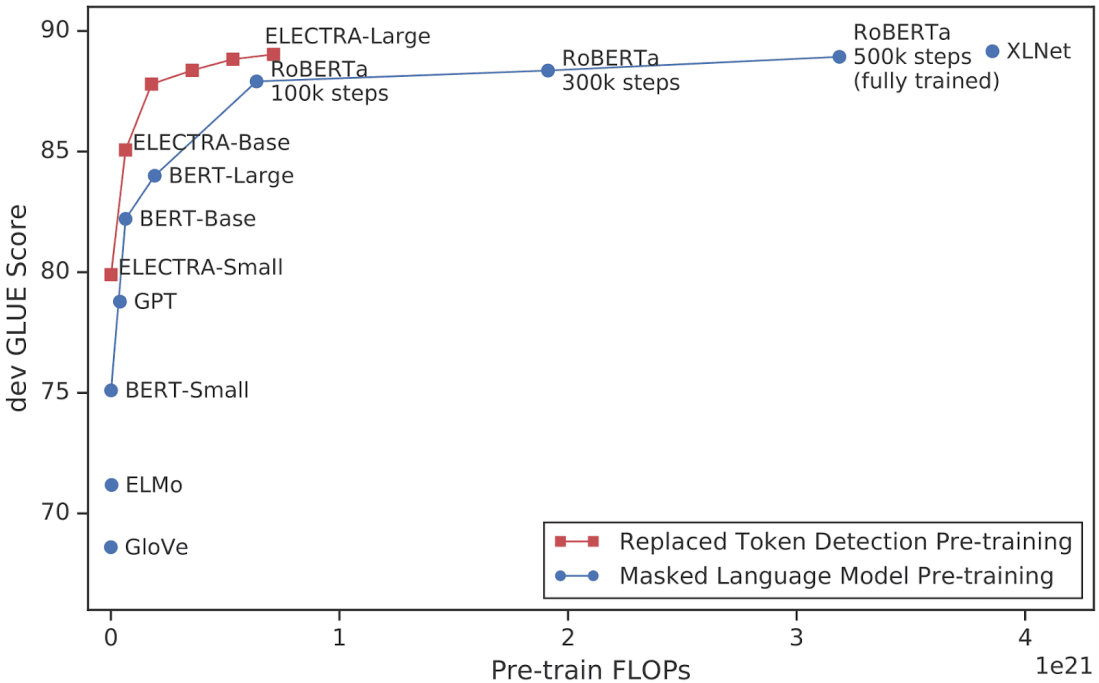
*Equal contributions

We investigate the optimal model size and number of tokens for training a transformer language model under a given compute budget. We find that current large language models are significantly undertrained, a consequence of the recent focus on scaling language models whilst keeping the amount of training data constant. By training over 400 language models ranging from 70 million to over 16 billion parameters on 5 to 500 billion tokens, we find that for compute-optimal training, the model size and the number of training tokens should be scaled equally: for every doubling of model size the number of training tokens should also be doubled. We test this hypothesis by training a predicted compute-optimal model, *Chinchilla*, that uses the same compute budget as *Gopher* but with 70B parameters and 4× more more data. *Chinchilla* uniformly and significantly outperforms *Gopher* (280B), GPT-3 (175B)

Moderne LMer er undertrente

Effektivitet: trening

- metoder for å skalere ned antall parametre
- mer effektiv modellering av lengre sekvenser
- pre-treningsobjektiv



(Clark et al (2020): ELECTRA)

Effektivitet: trening

Trained on 100 million words and still in shape: BERT meets British National Corpus

David Samuel, Andrey Kutuzov, Lilja Øvrelid and Erik Velldal
University of Oslo, Language Technology Group
{davisamu, andreku, liljao, erikve}@ifi.uio.no

Abstract

While modern masked language models (LMs) are trained on ever larger corpora, we here explore the effects of down-scaling training to a modestly-sized but representative, well-balanced, and publicly available English text source – the British National Corpus. We show that pre-training on this carefully curated corpus can reach better performance than the original BERT model. We argue that this type of corpora has great potential as a language model that the combination of a well-curated representative corpus, improved LTG-BERT architecture and a better training objective results in a model with stronger linguistic knowledge than the original English BERT pre-trained on 30× larger corpus.

Large language models are notoriously data hungry, requiring hundreds of gigabytes of raw textual data. This becomes a major obstacle for low-resource languages while also putting a limit to the efficiency of any 'efficient' language model. On top of that, the size of web-crawled corpora makes

Effektivitet: trening

Model	MNLI	Edge probing	BLiMP	Training time
LTG-BERT	**85.1**$^{\pm 0.2}$	**95.3**$^{\pm 0.1}$	83.4	8h 13min
w/ post-norm (0.005)	$-0.5^{\pm 0.2}$	$-0.6^{\pm 0.1}$	-0.1	-22min
w/ pre-norm (0.005)	$-1.3^{\pm 0.1}$	$-0.2^{\pm 0.1}$	-0.9	-35min
w/ GELU activation	$-0.3^{\pm 0.3}$	**0.0**$^{\pm 0.1}$	-0.1	-6min
w/ absolute pos. emb.	$-1.1^{\pm 0.2}$	$\mathbf{-0.1}^{\pm 0.1}$	**+0.6**	**−2h 16min**
w/o FF init. scaling	$-0.3^{\pm 0.2}$	$\mathbf{-0.1}^{\pm 0.1}$	+0.1	0min
w/ learnt FF biases	$-0.3^{\pm 0.2}$	**0.0**$^{\pm 0.1}$	-0.1	+9min
w/ 0.01 WD (0.005)	$-1.4^{\pm 0.1}$	$-0.2^{\pm 0.1}$	-0.7	-1min
w/ linear schedule	$-0.5^{\pm 0.2}$	**0.0**$^{\pm 0.1}$	-0.2	0min
w/ AdamW (0.001)	$-0.9^{\pm 0.2}$	$-0.2^{\pm 0.1}$	-0.5	-11min

Effektivitet: trening

BabyLM Challenge

Sample-efficient pretraining on a developmentally plausible corpus

Overview • **Guidelines** • **Timeline** • **FAQs**

Summary: This shared task challenges community members to train a language model **from scratch** on the same amount of linguistic data available to a child. Submissions should be implemented in Huggingface's Transformers library and will be evaluated on a shared pipeline. This shared task is co-sponsored by CMCL and CoNLL.

- Download Dataset (700MB unzipped)
- Evaluate your model using our evaluation pipeline
- Models and results due ~~**July 15, 2023**~~ **July 22, 2023, 23:59 anywhere on earth (UTC-12)**. Submit on dynabench.
- Paper submission due ~~**August 1, 2023**~~ **August 2, 2023, 23:59 anywhere on earth (UTC-12)**. Submit on OpenReview.

See the guidelines for an overview of submission tracks and pretraining data. See the call for papers for a detailed description of the task setup and data.

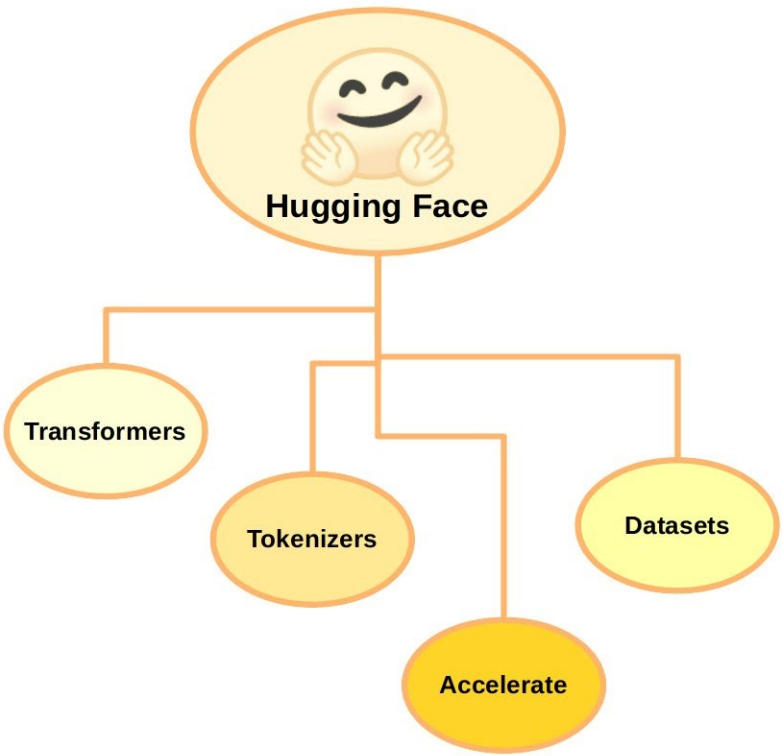
Consider joining the BabyLM Slack if you have any questions for the organizers or want to connect with other participants!

https://babylm.github.io/

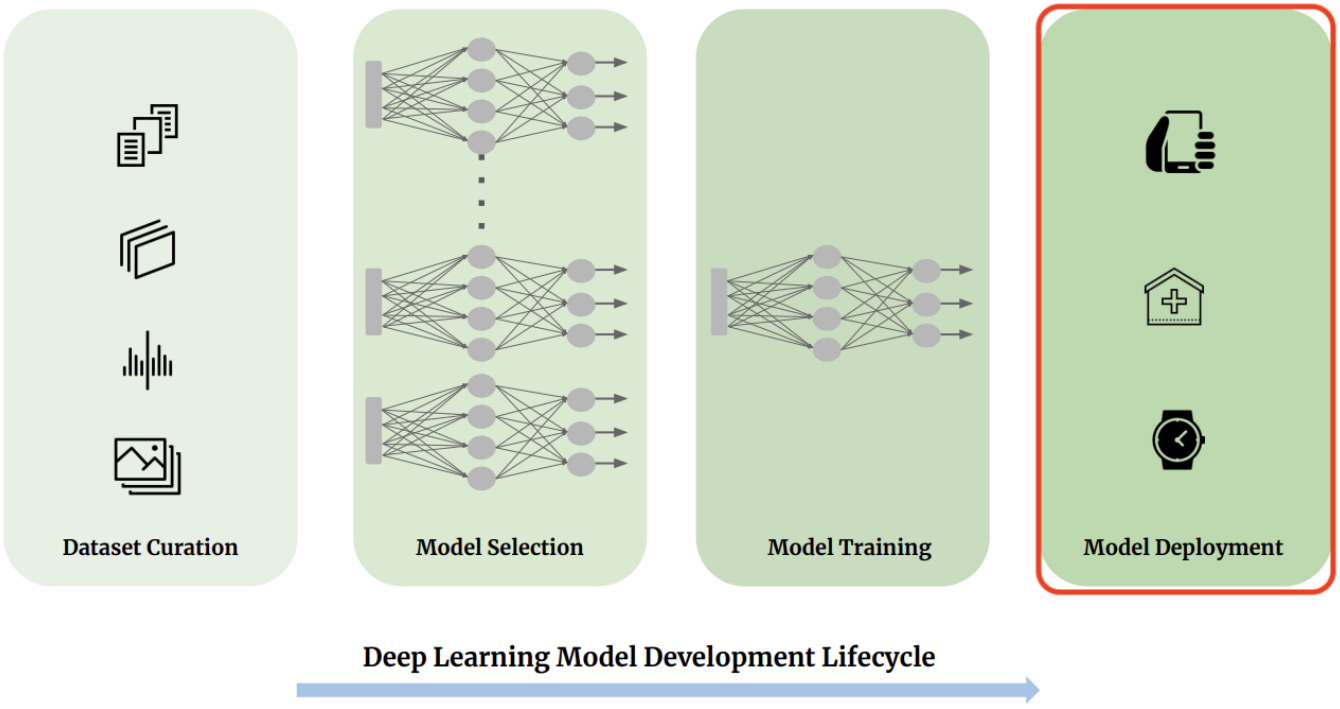
Sharing is caring

Unngå trening av mange modeller

- Del modellen åpent
- Huggingface-økosystemet: data og modeller
- Viktig med rapportering
  - detaljer rundt treningsdata, hyperparametre, energibruk, osv



Effektivitet: inferens



- Redusere størrelsen til modellen
  - Pruning
  - Knowledge distillation

Størrelse er ikke alt

Artificial Intelligence Index Report 2023

Chapter 2: Technical Performance
2.8 Environment

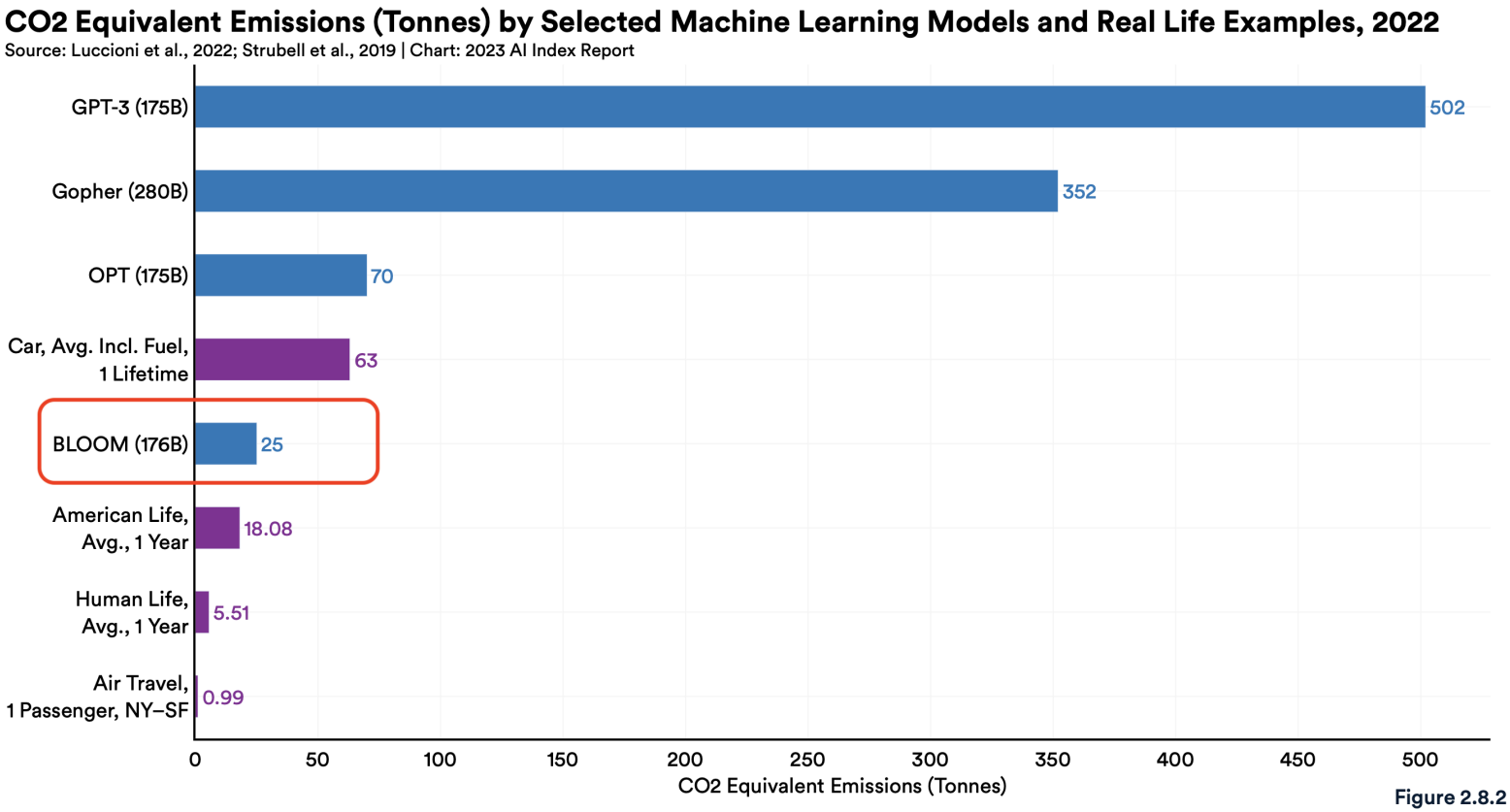


Figure 2.8.2

(Luccioni et al, 2022)

Energieffektivitet

Beregning i energieffektive datasentere

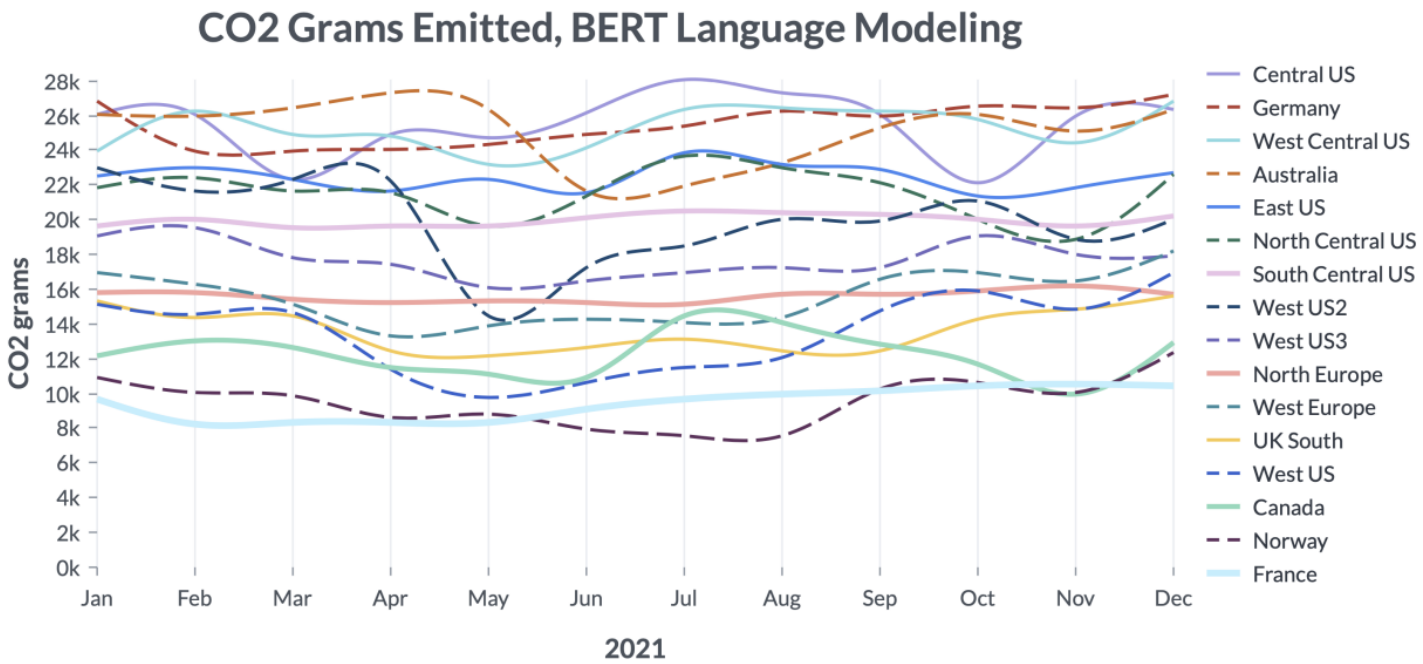
Energieffektivitet

Beregning i energieffektive datasentere

- Hydrokraft
- Naturlig avkjøling
- Gjenbruk av varmeavfall

*May 2022 list

Energieffektivitet



(Dodge et al, 2022)

Mot grønnere språkmodeller

- Regulert rapportering

Mot grønnere språkmodeller

- Regulert rapportering

Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	Totals
Data sources	●○○○	●●●○	●●●●	○○○○	●●○○	●●●●	●●●●	○○○○	○○○○	●●●●	22
Data governance	●●○○	●●●○	●●○○	○○○○	●●●○	●●●●	●●○○	○○○○	○○○○	●●●○	19
Copyrighted data	○○○○	○○○○	○○○○	○○○○	○○○○	●●●○	○○○○	○○○○	○○○○	●●●●	7
Compute	○○○○	○○○○	●●●●	○○○○	○○○○	●●●●	●●●●	○○○○	●○○○	●●●●	17
Energy	○○○○	●○○○	●●●○	○○○○	○○○○	●●●●	●●●●	○○○○	○○○○	●●●●	16
Capabilities & limitations	●●●●	●●●○	●●●●	●○○○	●●●●	●●●○	●●○○	●●○○	●○○○	●●●○	27
Risks & mitigations	●●●○	●●○○	●○○○	●○○○	●●●○	●●○○	●○○○	●●○○	○○○○	●○○○	16
Evaluations	●●●●	●●○○	○○○○	○○○○	●●○○	●●●○	●●○○	○○○○	●○○○	●○○○	15
Testing	●●●○	●●○○	○○○○	○○○○	●●○○	●●○○	○○○○	●○○○	○○○○	○○○○	10
Machine-generated content	●●●○	●●●○	○○○○	●●●○	●●●○	●●●○	○○○○	●●●○	●○○○	●●○○	21
Member states	●●○○	○○○○	○○○○	●●○○	●●●●	○○○○	○○○○	○○○○	●○○○	○○○○	9
Downstream documentation	●●●○	●●●●	●●●●	○○○○	●●●●	●●●●	●●○○	○○○○	○○○○	●●●○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

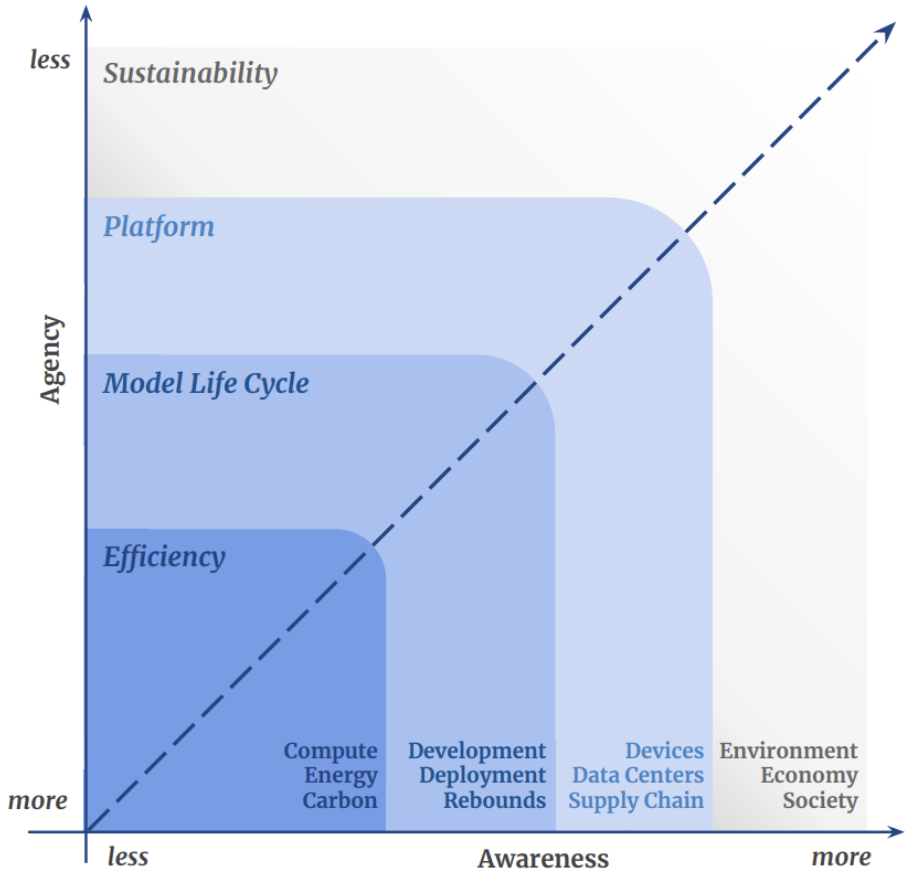
https://crfm.stanford.edu/2023/06/15/eu-ai-act.html

Mot grønnere språkmodeller

- Effektivitet i hele ML-livssyklusen
  - Mer effektiv databruk, trening og inferens
  - Energieffektive datasentere
  - Regionale og temporale forskjeller

Mot grønnere språkmodeller

- Effektivitet i hele ML-livssyklusen
  - Mer effektiv databruk, trening og inferens
  - Energieffektive datasentere
  - Regionale og temporale forskjeller



(Wright et al, 2023)

Takk for meg!