GPT-SW3: the first LLM for the Nordics

AI SWEDEN

magnus.sahlgren@ai.se

2017: the Transformer is introduced

2018: the first Transformer-based language models (GPT, BERT)
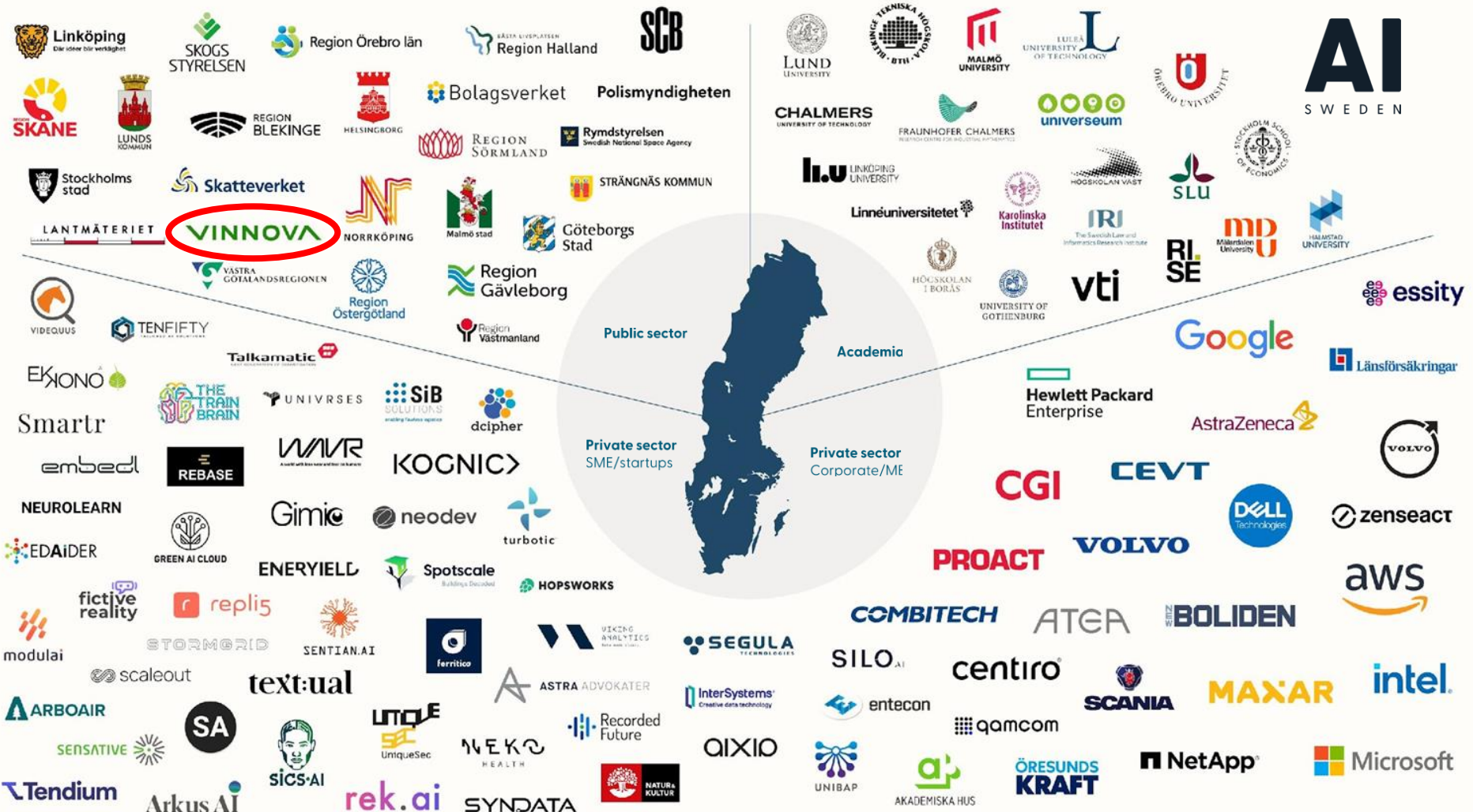
2019: language models for Swedish authorities project

2020: the birth of LLMs (GPT-3)

2021: the NLU research group at AI Sweden is formed

2022: the first LLM for the Nordics: GPT-SW3

magnus.sahlgren@ai.se

AI SWEDEN

Public sector

Academia

Private sector
SME/startups

Private sector
Corporate/ME

Data

Infrastructure

Competence

Data

Infrastructure

Competence
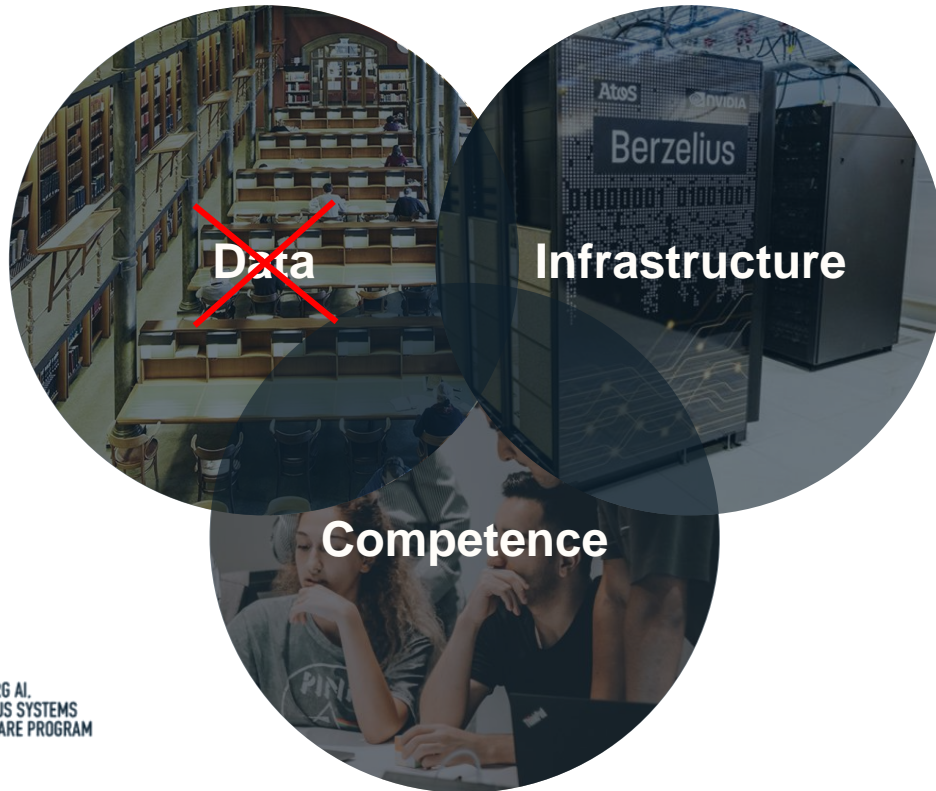
The Nordic Pile:

Swedish, Norwegian, Danish, Icelandic, Faroese English and programming code

Based on existing data and open sources

AI
SWEDEN

magnus.sahlgren@ai.se

GPT-SW3 series

Model size	N	L	s	h	v	Total Model FLOPs
126m	1.82×10^8	12	2048	768	64000	3.84×10^{20}
356m	1.82×10^8	24	2048	1024	64000	1.05×10^{21}
1.3B	1.82×10^8	24	2048	2048	64000	3.44×10^{21}
7B	1.40×10^8	32	2048	4096	64000	1.25×10^{22}
20B	1.82×10^8	44	2048	6144	64000	4.48×10^{22}
40B	1.82×10^8	48	2048	8192	64000	9.11×10^{22}
Total						1.56×10^{23}

S : Steps
B : Batch size
N : Number of samples ($N = BS$)
L : Layers
s : Sequence length
h : Embedding dimension
v : Vocabulary size

$$\text{Model FLOPs per Batch} = 72BLsh^2 \left(1 + \frac{s}{6h} + \frac{v}{12hL}\right)$$

$$= Bsh\left(72Lh + 12Ls + 6v\right)$$

$$\text{Total Model FLOPs} = 72NLsh^2 \left(1 + \frac{s}{6h} + \frac{v}{12hL}\right)$$

$$= Nsh\left(72Lh + 12Ls + 6v\right)$$

$$= S \times \text{Model FLOPs per Batch}$$

GPT-SW3 series

Model size	N	L	s	h	v	Total Model FLOPs
126m	1.82×10^8	12	2048	768	64000	3.84×10^{20}
356m	1.82×10^8	24	2048	1024	64000	1.05×10^{21}
1.3B	1.82×10^8	24	2048	2048	64000	3.44×10^{21}
7B	1.40×10^8	32	2048	4096	64000	1.25×10^{22}
20B	1.82×10^8	44	2048	6144	64000	4.48×10^{22}
40B	1.82×10^8	48	2048	8192	64000	9.11×10^{22}
Total						1.56×10^{23}

S : Steps
B : Batch size
N : Number of samples ($N = BS$)
L : Layers
s : Sequence length
h : Embedding dimension
v : Vocabulary size

$$\text{Model FLOPs per Batch} = 72BLsh^2\left(1 + \frac{s}{6h} + \frac{v}{12hL}\right)$$
$$= Bsh\left(72Lh + 12Ls + 6v\right)$$
$$\text{Total Model FLOPs} = 72NLsh^2\left(1 + \frac{s}{6h} + \frac{v}{12hL}\right)$$
$$= Nsh\left(72Lh + 12Ls + 6v\right)$$
$$= S \times \text{Model FLOPs per Batch}$$

GPT-SW3 series

+ instruction finetuning

Model size	N	L	s	h	v	Total Model FLOPs
126m	1.82×10^8	12	2048	768	64000	3.84×10^{20}
356m	1.82×10^8	24	2048	1024	64000	1.05×10^{21}
1.3B	1.82×10^8	24	2048	2048	64000	3.44×10^{21}
7B	1.40×10^8	32	2048	4096	64000	1.25×10^{22}
20B	1.82×10^8	44	2048	6144	64000	4.48×10^{22}
40B	1.82×10^8	48	2048	8192	64000	9.11×10^{22}
Total						1.56×10^{23}

S : Steps
B : Batch size
N : Number of samples ($N = BS$)
L : Layers
s : Sequence length
h : Embedding dimension
v : Vocabulary size

$$\text{Model FLOPs per Batch} = 72BLsh^2 \left(1 + \frac{s}{6h} + \frac{v}{12hL}\right)$$
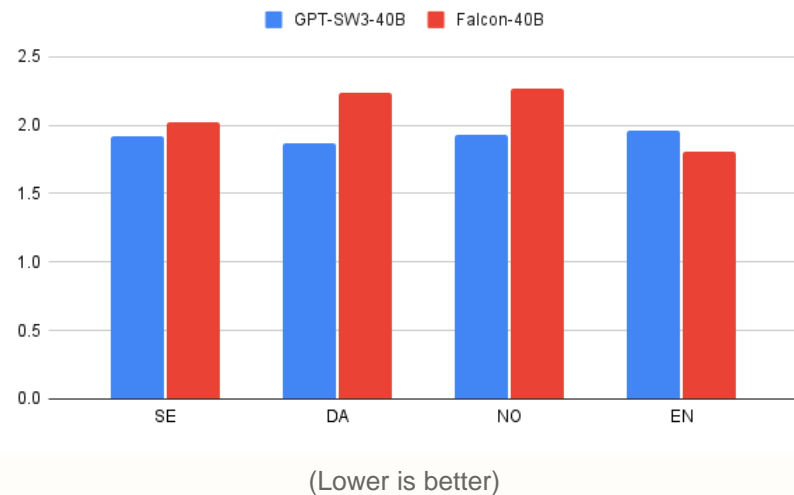$$= Bsh\left(72Lh + 12Ls + 6v\right)$$
$$\text{Total Model FLOPs} = 72NLsh^2 \left(1 + \frac{s}{6h} + \frac{v}{12hL}\right)$$
$$= Nsh\left(72Lh + 12Ls + 6v\right)$$
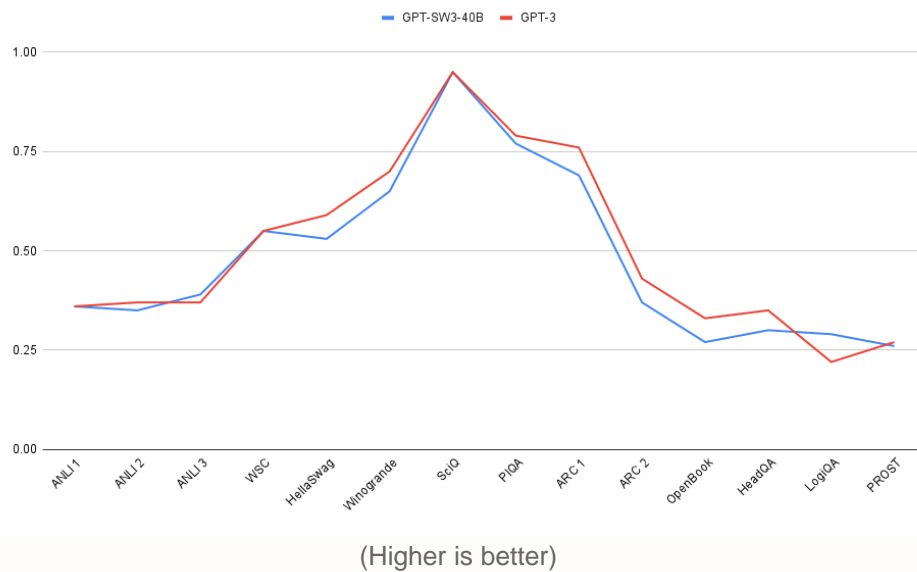$$= S \times \text{Model FLOPs per Batch}$$
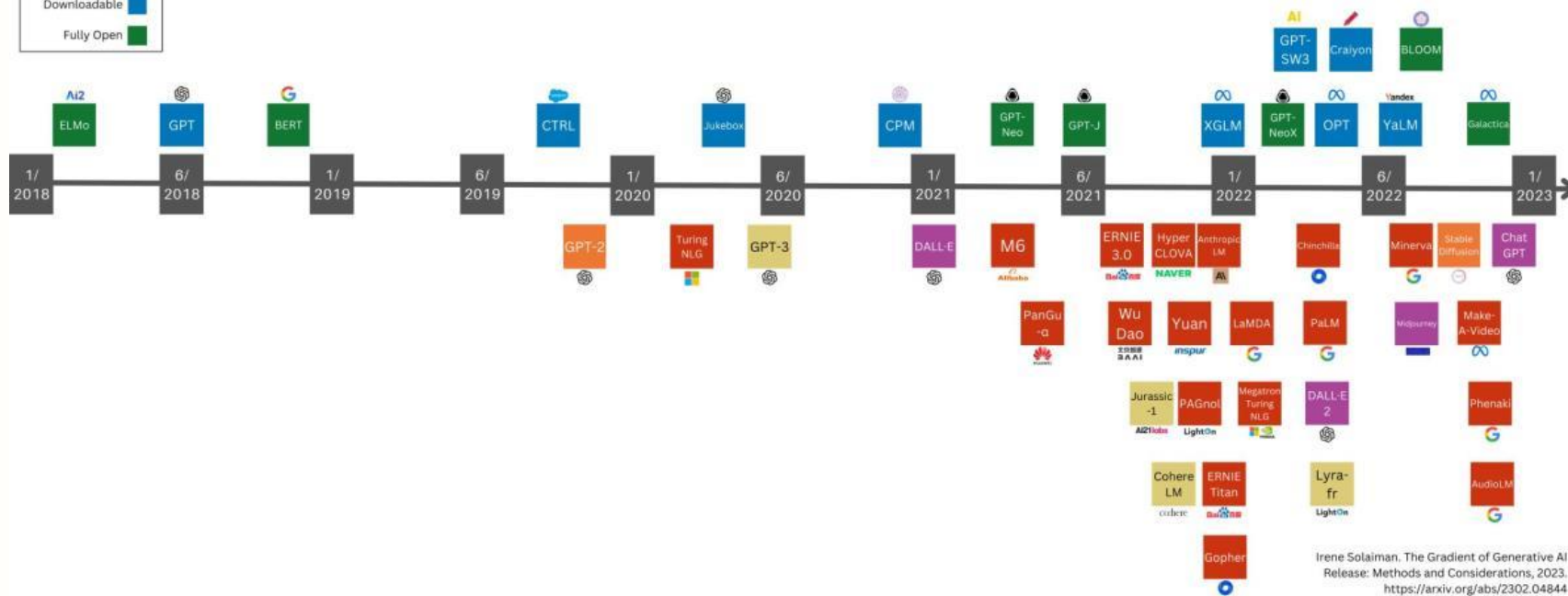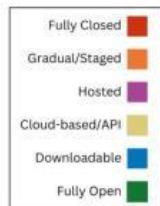
Language modeling

GPT-SW3-40B and Falcon-40B

Legend: GPT-SW3-40B, Falcon-40B

(Lower is better)

LM harness

GPT-SW3-40B and GPT-3-175B

Legend: GPT-SW3-40B, GPT-3
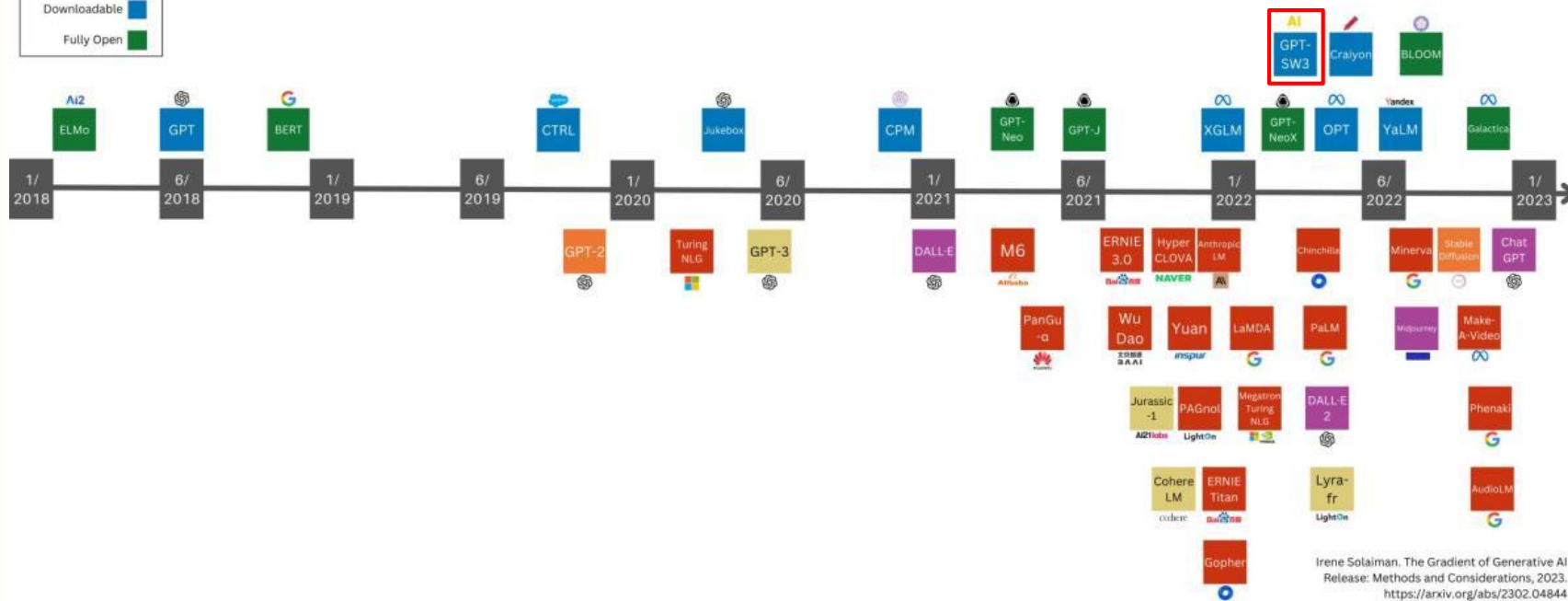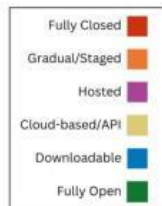
(Higher is better)

magnus.sahlgren@ai.se

Generative AI System Release Method Over Time (All Modalities)

Irene Solaiman. The Gradient of Generative AI Release: Methods and Considerations, 2023. https://arxiv.org/abs/2302.04844

magnus.sahlgren@ai.se

Generative AI System Release Method Over Time (All Modalities)

Irene Solaiman. The Gradient of Generative AI Release: Methods and Considerations, 2023. https://arxiv.org/abs/2302.04844

The next steps

- Multimodality and general capacity
- **We can achieve more if we collaborate**
- AI Nordics Discord: discord.gg/RgKVztg3xU

magnus.sahlgren@ai.se