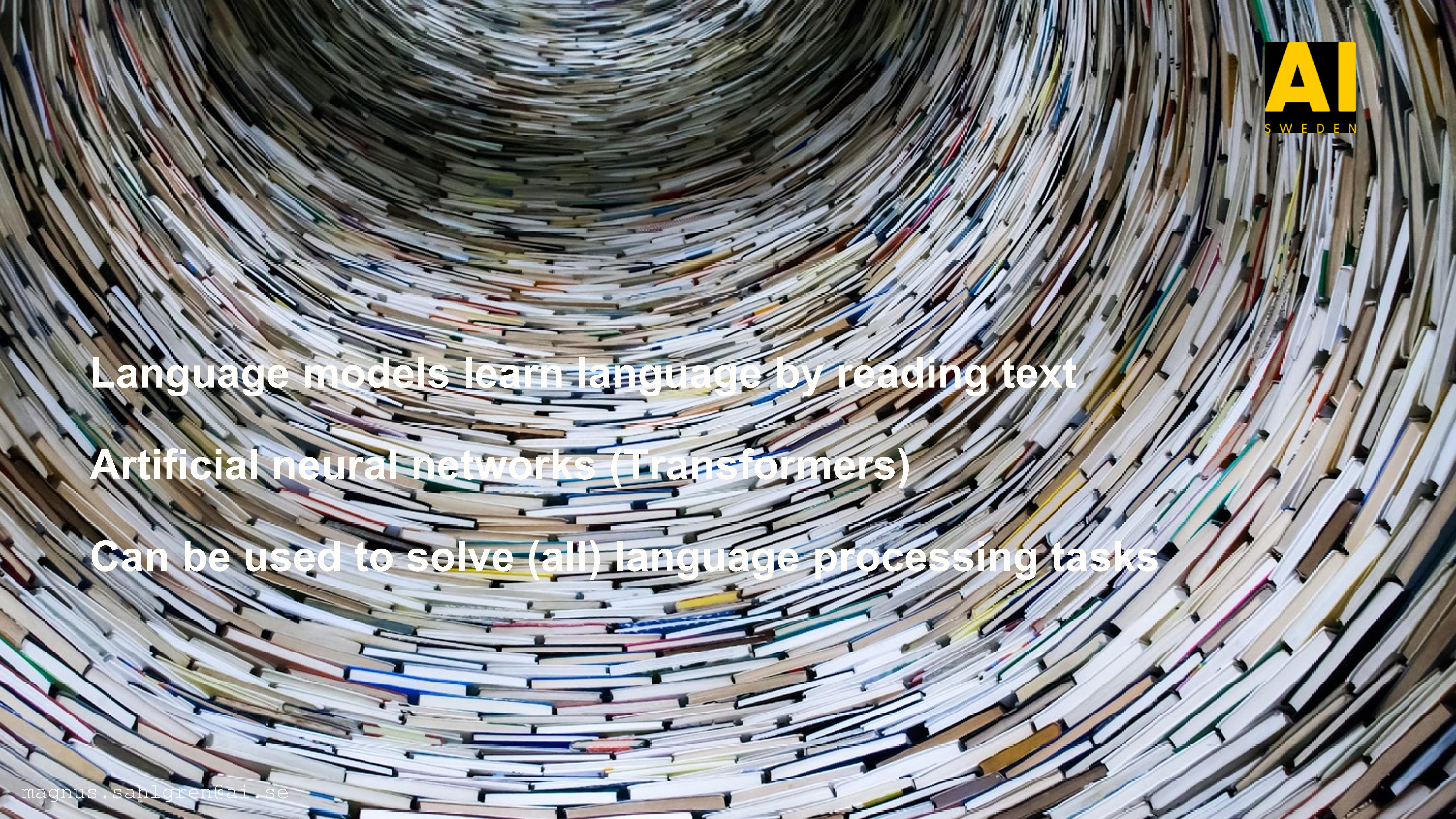
AI SWEDEN

Building GPT-SW3: The first LLM for the nordic languages.

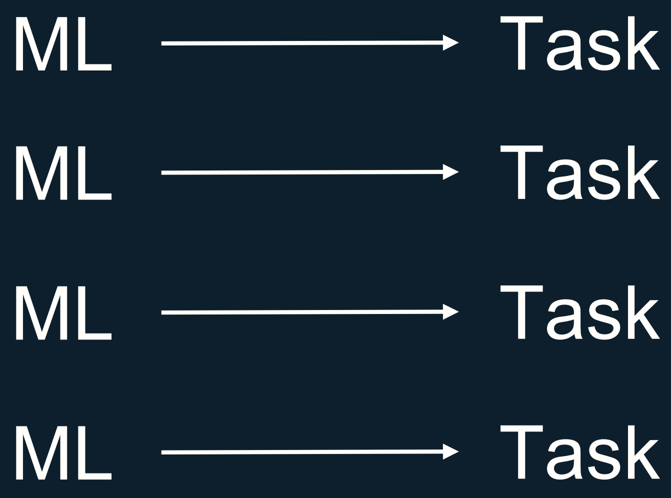


Language models learn language by reading text

Artificial neural networks (Transformers)

Can be used to solve (all) language processing tasks

Before 2018



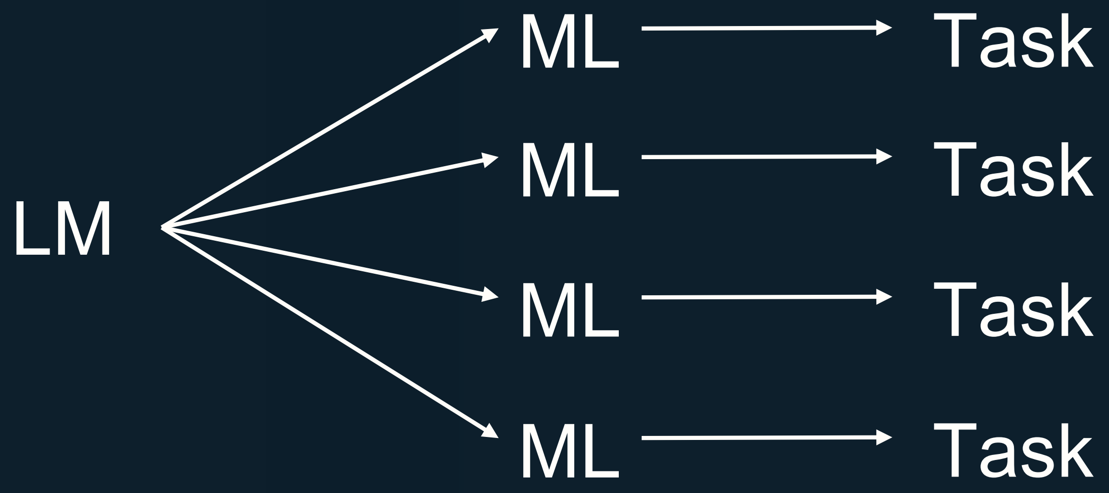
ML → Task

ML → Task

ML → Task

ML → Task

After 2018



LM

ML → Task

ML → Task

ML → Task

ML → Task

Encoders FTW!

Language models for Swedish authorities

Vinnova 2019-2022

Provide the prerequisites for authorities to use LMs

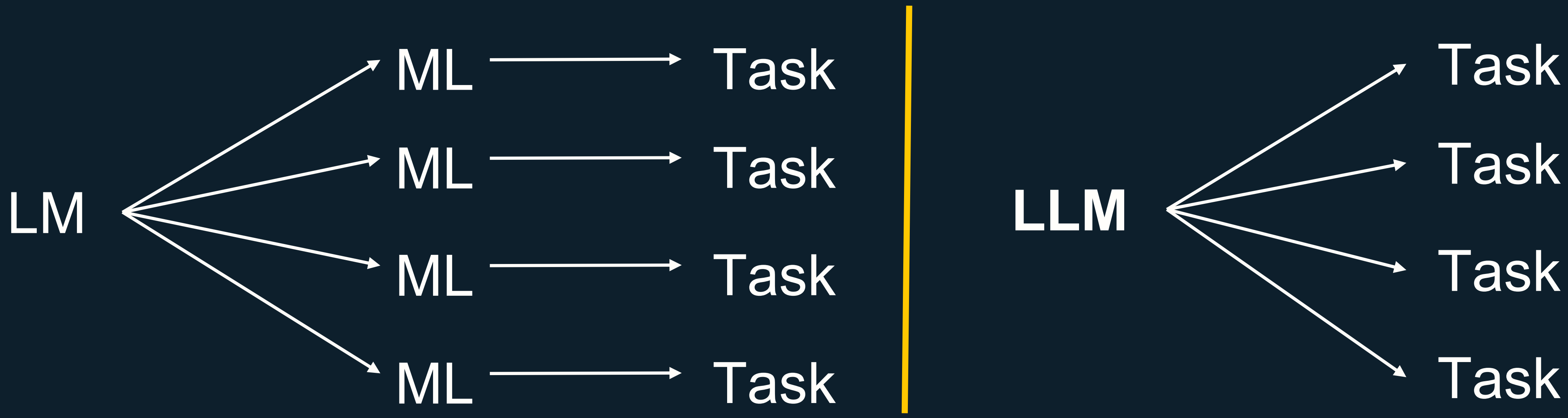
Code, data, models and applications

Most Swedish LMs have some connection to the project

All authorities in the project have built solutions based on LMs (Skatteverket, Arbetsförmedlingen, Tillväxtverket)

After 2018

After 2021



LM

ML → Task

ML → Task

ML → Task

ML → Task

LLM

Task

Task

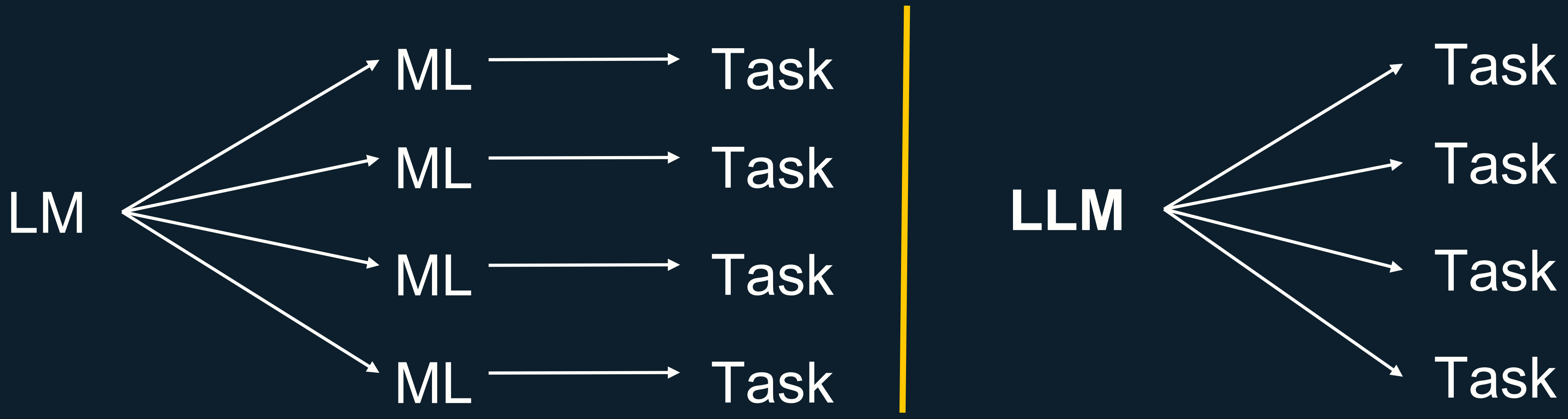
Task

Task

Encoders → **Decoders**

After 2018

After 2021



LM

ML → Task

ML → Task

ML → Task

ML → Task

LLM

Task

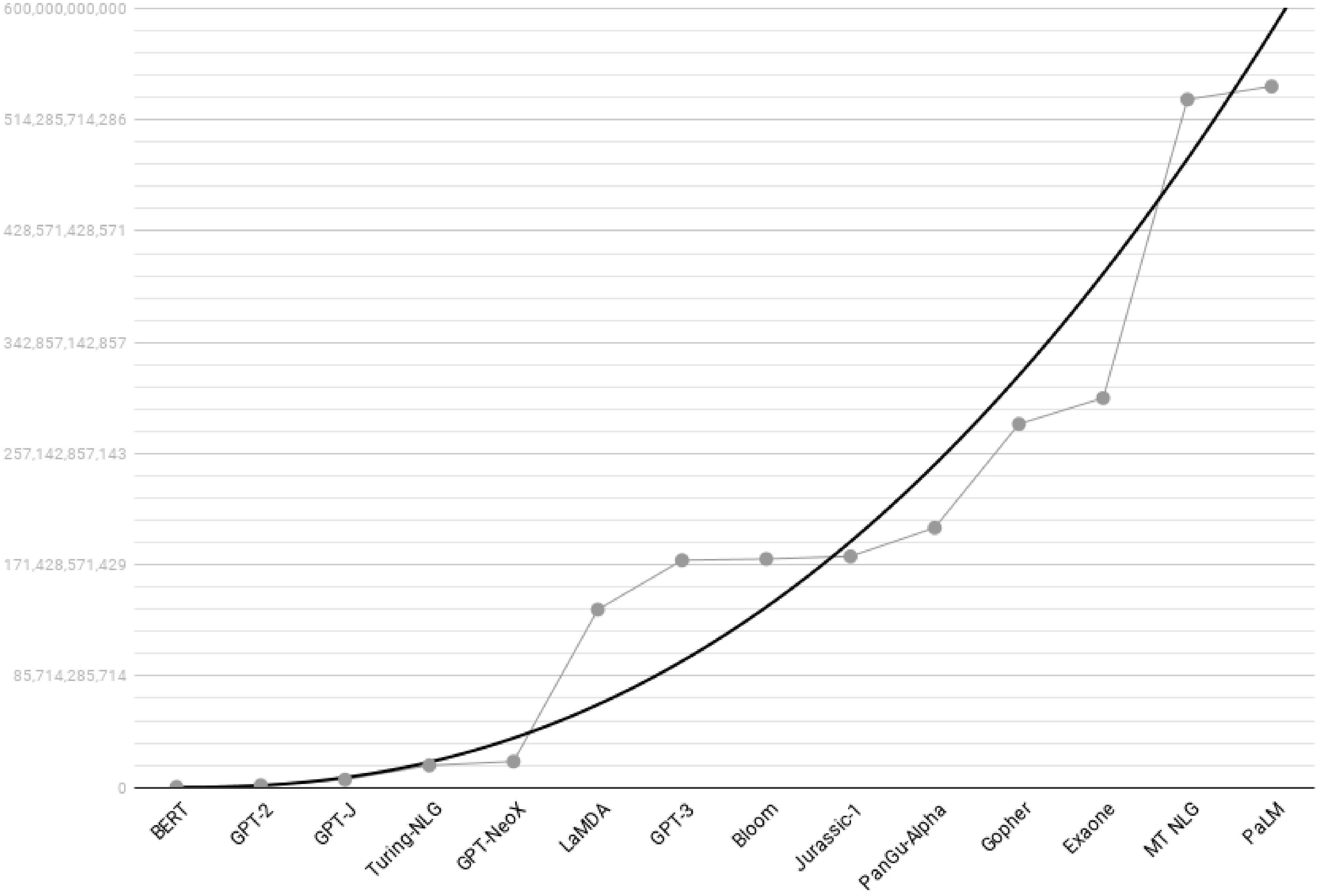
Task

Task

Task

Encoders → **Decoders**

Transformer model parameters

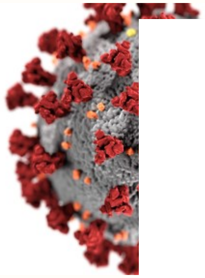


Bert large (340m): ~1000 GPUh
GPT-SW3 (6.7b): ~27000 GPUh
GPT-SW3 (40b): ~280000 GPUh

source: https://developer.nvidia.com/blog/training-bert-with-gpus/

Bert large (340m): ~1000 GPUh ~ 5 DGX-days
GPT-SW3 (6.7b): ~27000 GPUh ~ 140 DGX-days
GPT-SW3 (40b): ~280000 GPUh ~ 4 DGX-years

source: https://developer.nvidia.com/blog/training-bert-with-gpus/



N

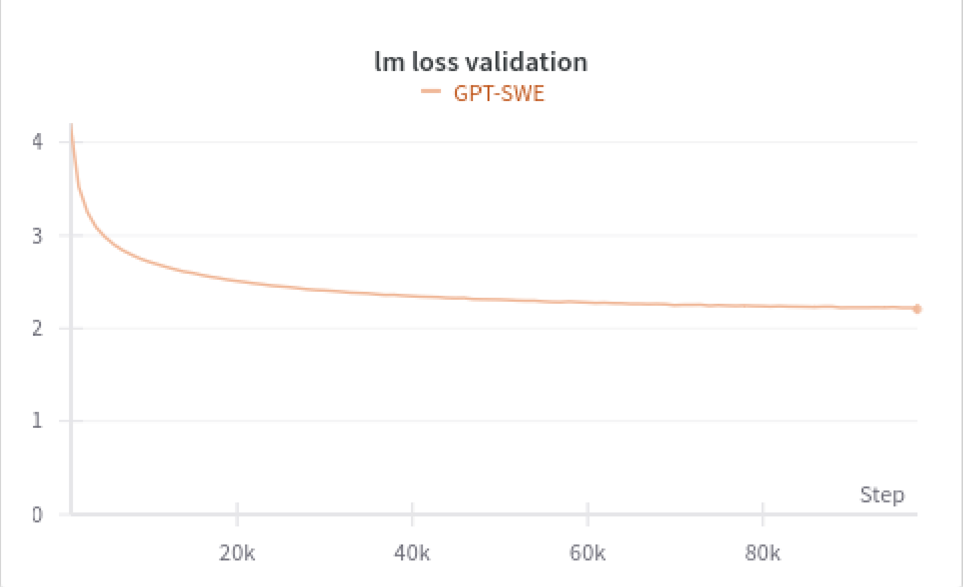
NVIDIA SuperPOD:
60 NVIDIA DGX-A100 compute nodes

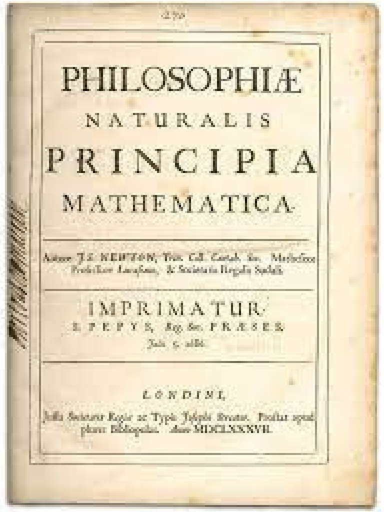
Each DGX-A100:
8 NVIDIA A100 Tensor Core GPUs

Each A100 GPU:
40 GB on-board HBM2 VRAM

Nvidia Mellanox Infiniband networking

Nvidia NeMo Megatron software stack





PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA.

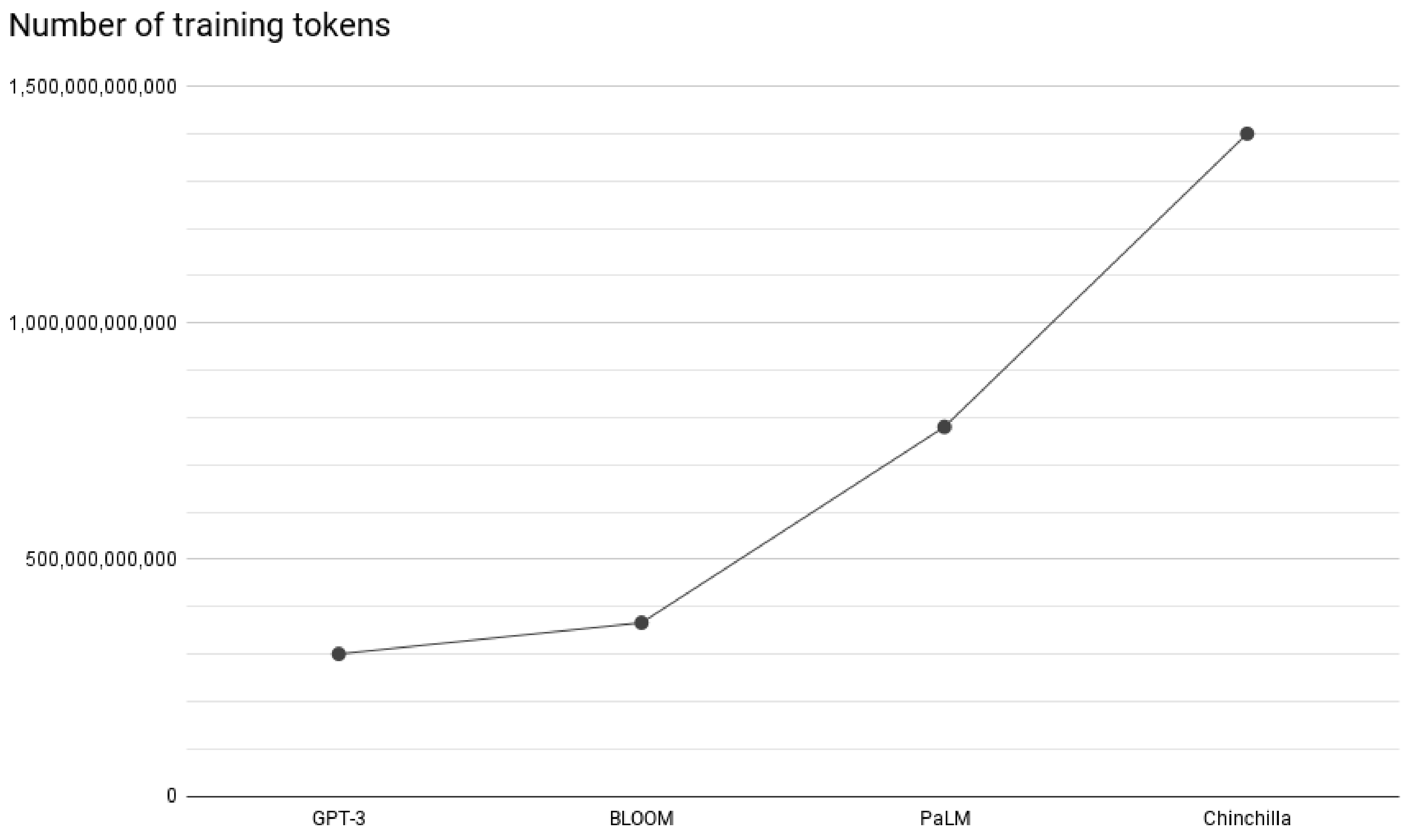
Autore JS. NEWTON, Trin. Coll. Cantab. Soc. Matheſeos
Profeſſore Lucaſiano, & Societatis Regalis Sodali.

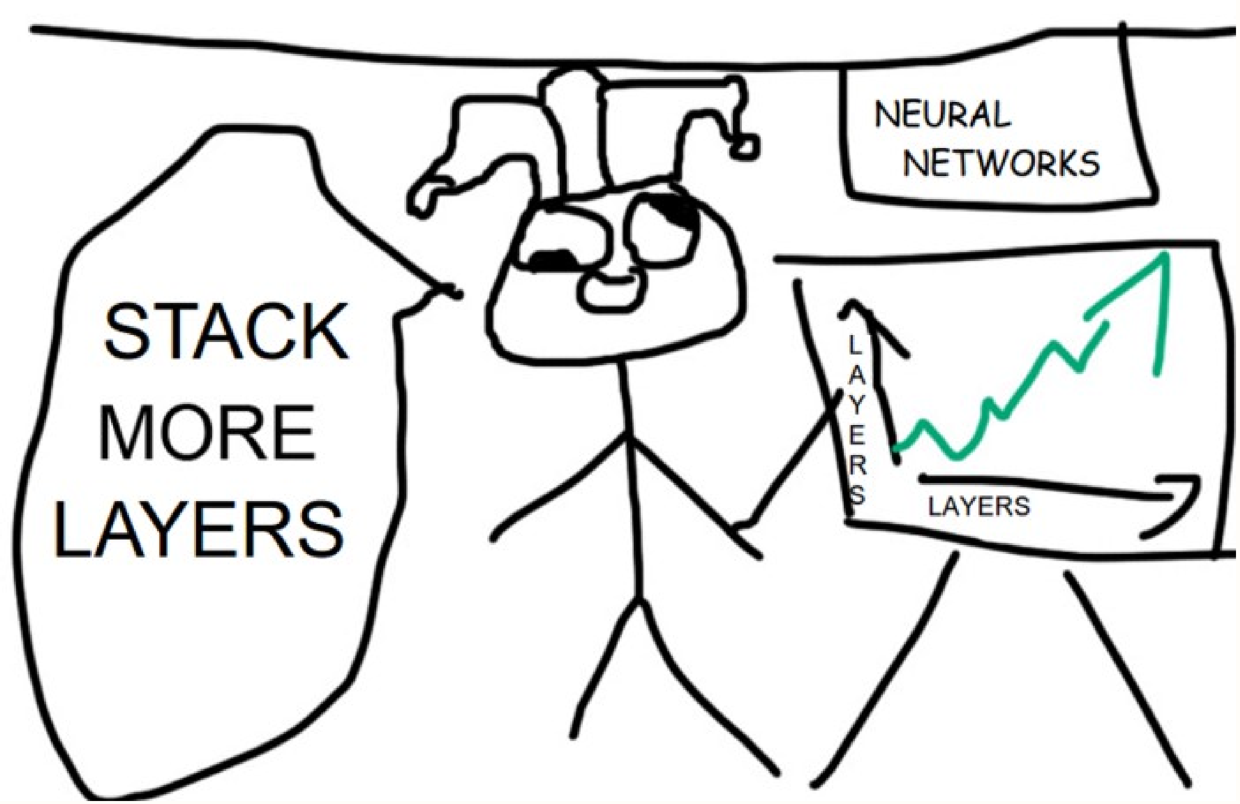
IMPRIMATUR.
S. PEPYS, Reg. Soc. PRÆSES.
Julii 5. 1686.

LONDINI,
Juſſu Societatis Regiæ ac Typis Joſephi Streater. Proſtat apud
plures Bibliopolas. Anno MDCLXXXVII.

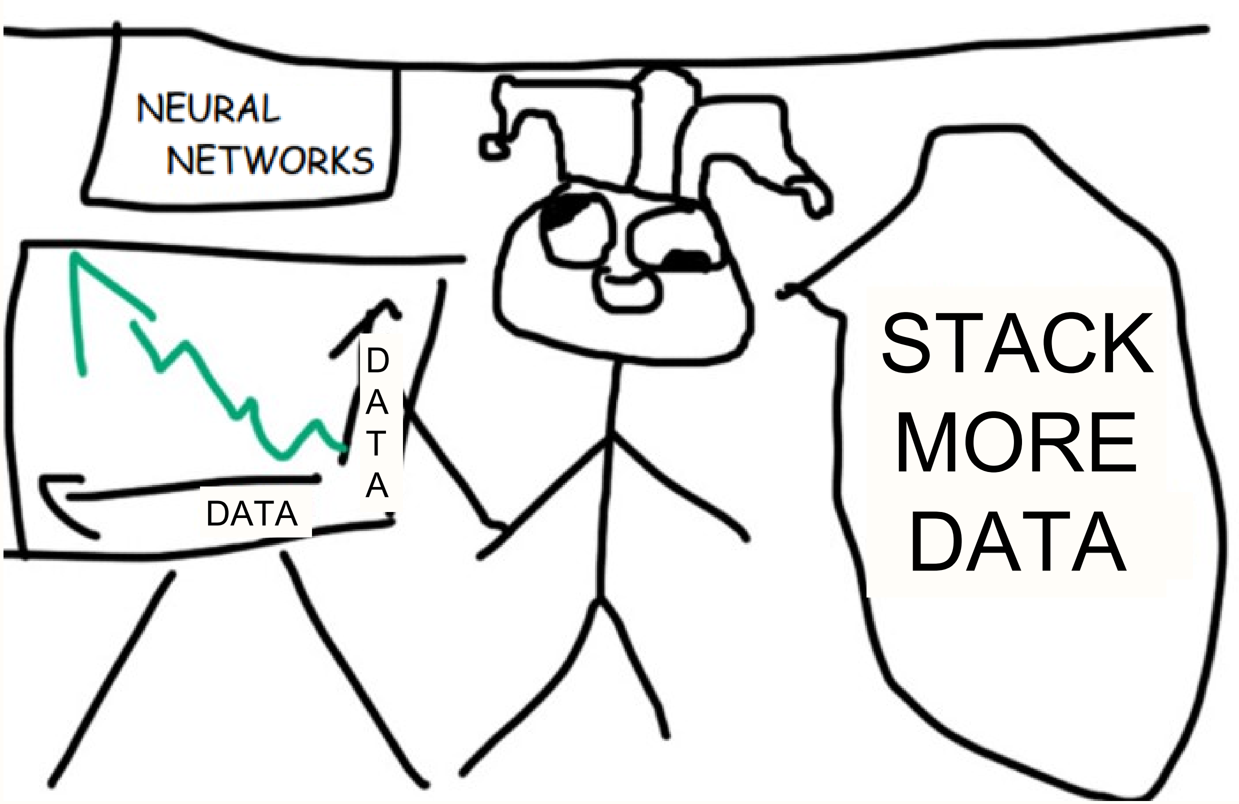
GPT-SWE

- A 1.3 B parameter model trained using Megatron-LM
- Trained on found data: Oscar, MC4, Swedish parliamentary debates, et.c.
- Familiarity with HPC


Number of training tokens
1,500,000,000,000
1,000,000,000,000
500,000,000,000
0
GPT-3
BLOOM
PaLM
Chinchilla



Kaplan et al. (2020) Scaling laws for neural language models



Hofman et al. (2022) Training compute-optimal large language models

GPT-SW3: a foundational resource for Nordic NLP

magnus.sahlgren@ai.se

Why?

Build competence

Digital and linguistic sovereignty

Democratize access to LLMs

How?

Collaboration

Transparency

GPT-SW3 consortium

AI SWEDEN

[redacted]

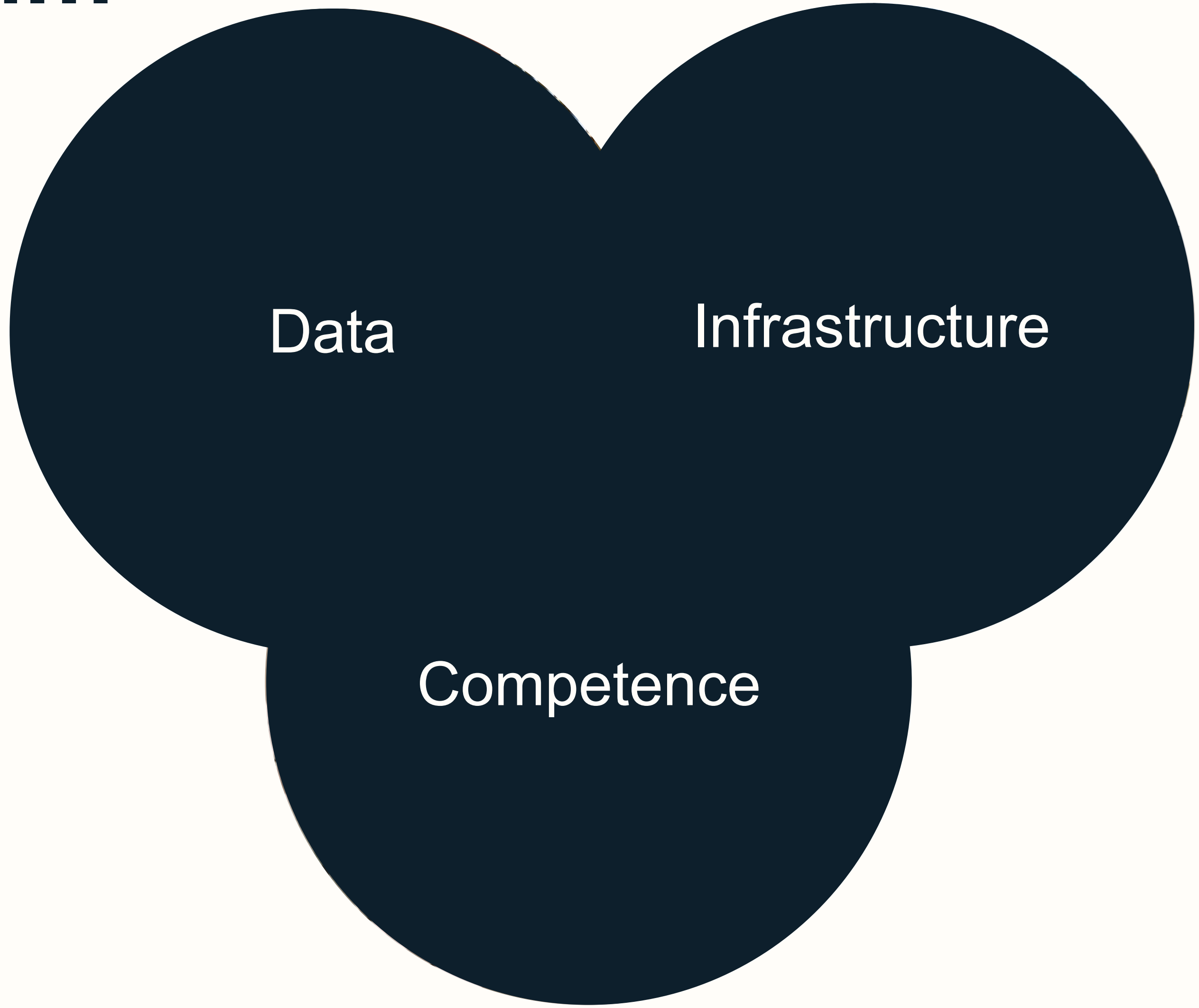
WASP WALLENBERG AI AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

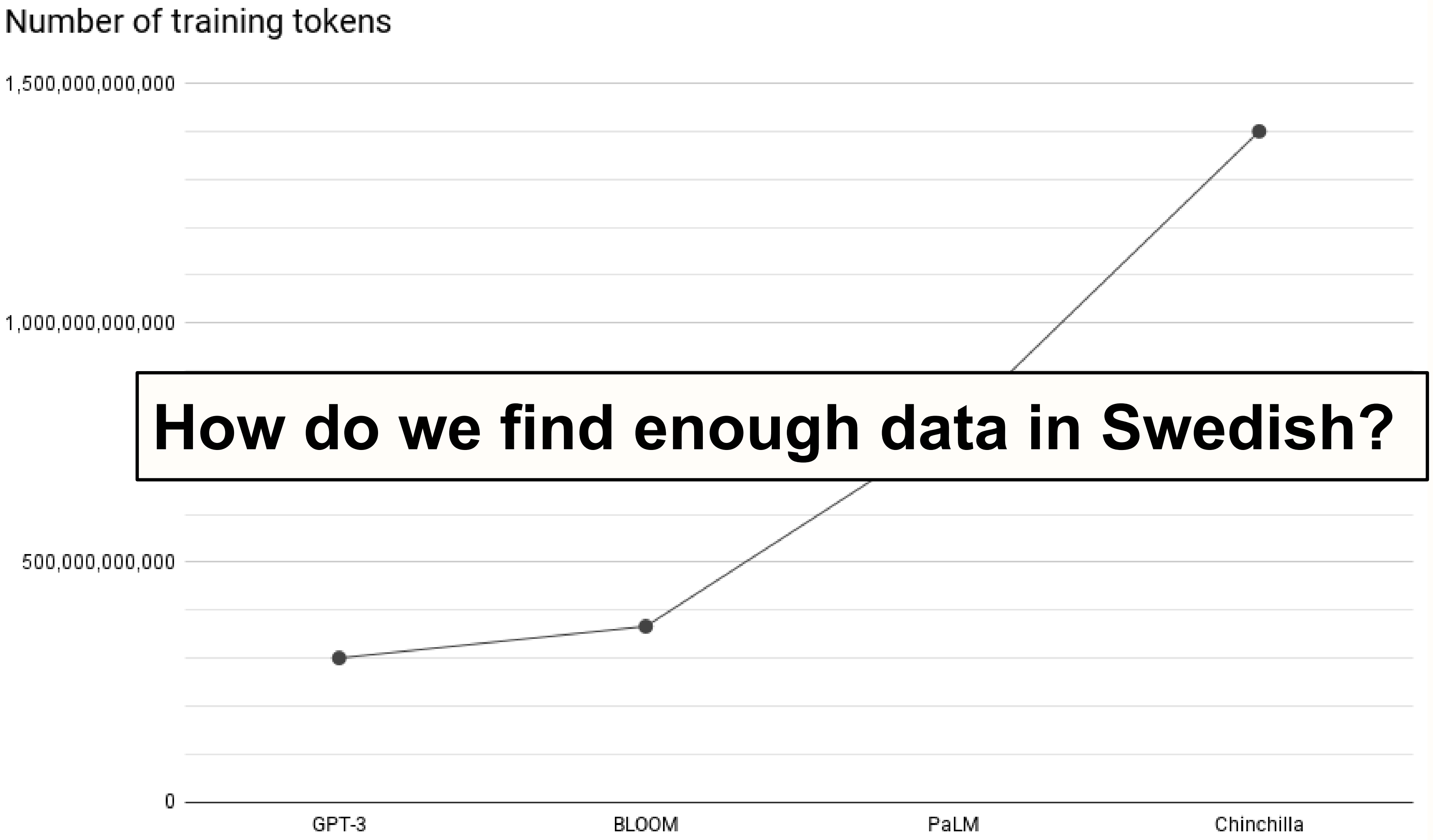
[redacted]

AI Nordics community on Discord

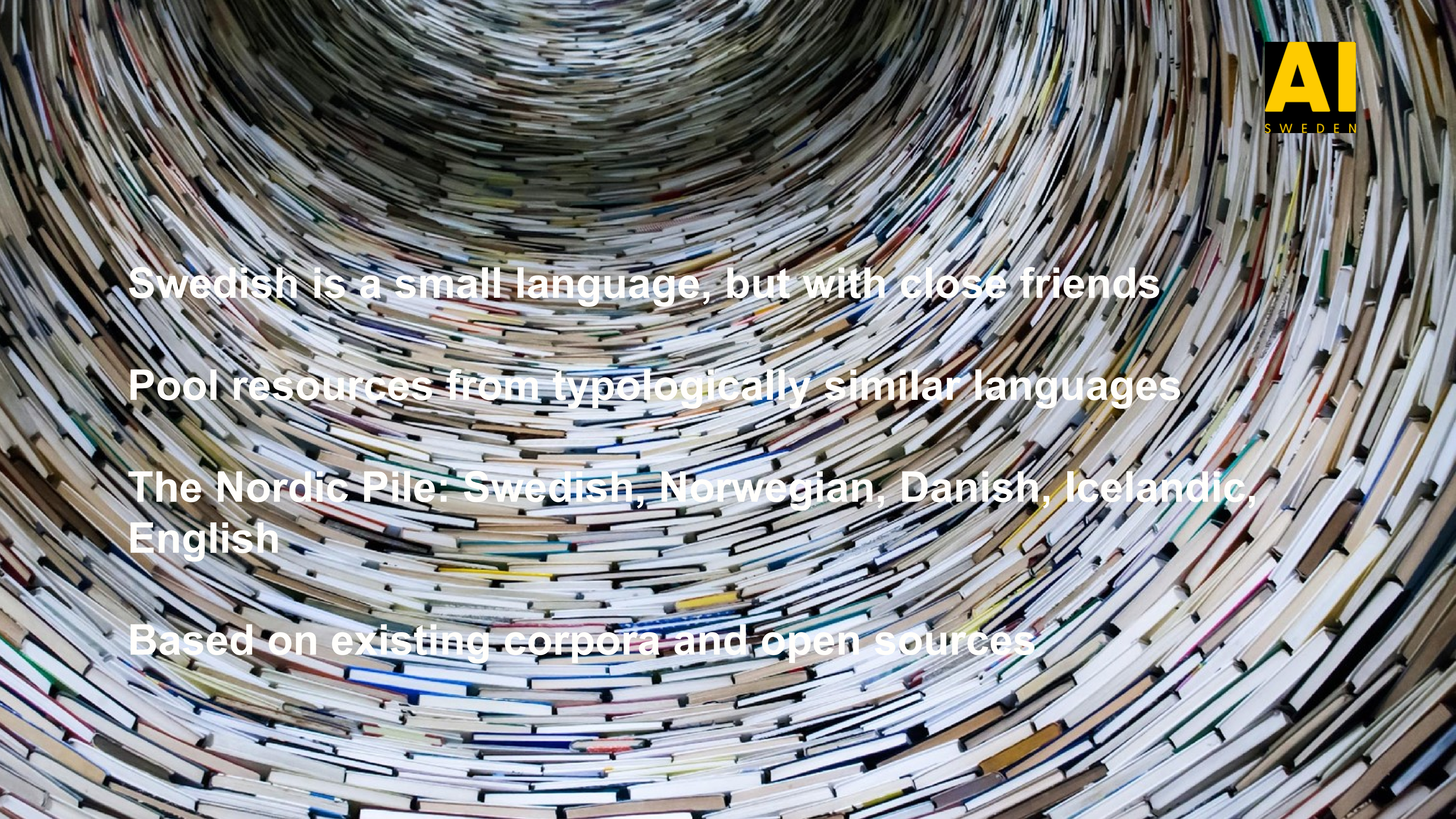
Data platform and deployment partners:

[redacted] aixia





How do we find enough data in Swedish?

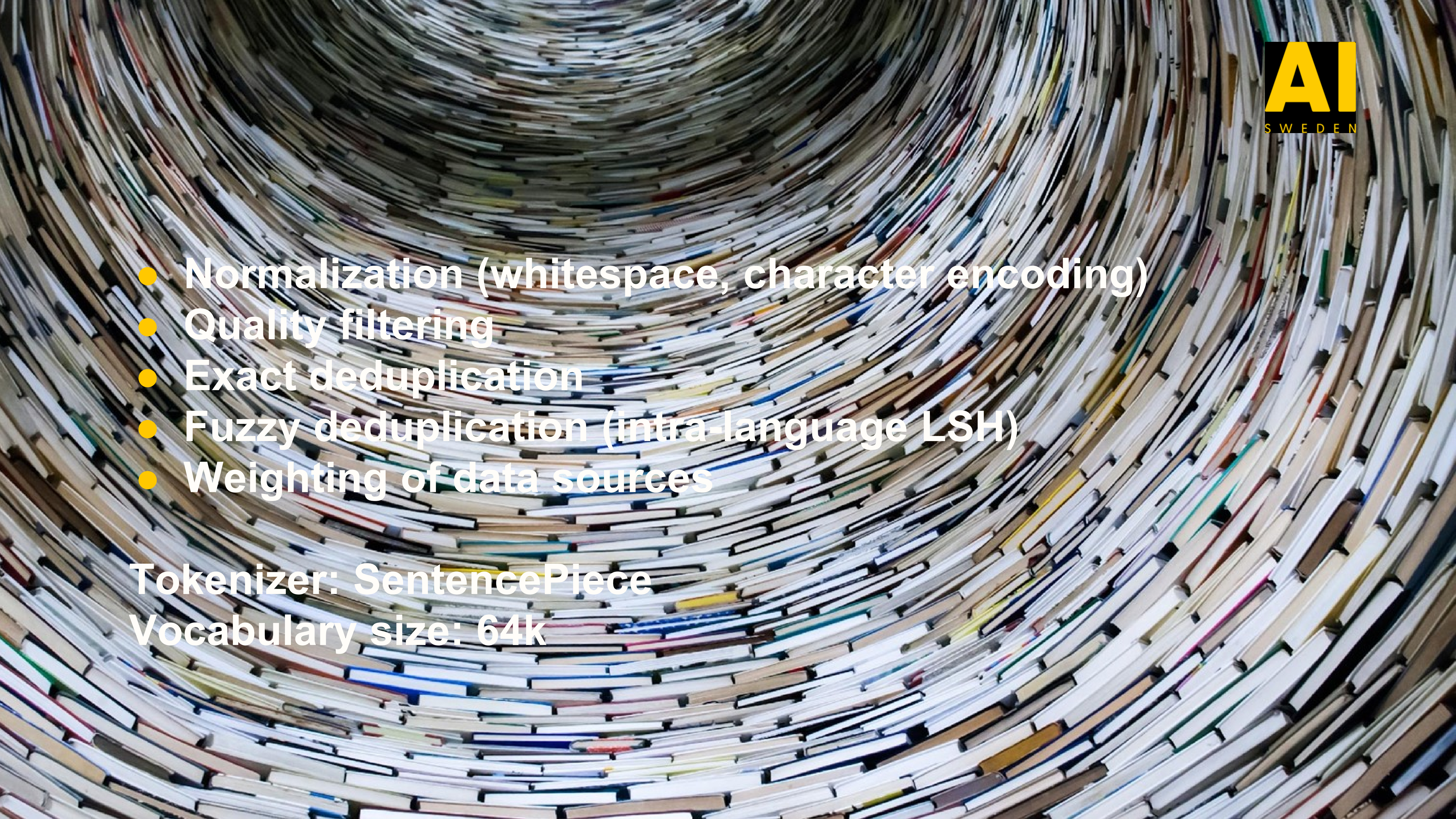


Swedish is a small language, but with close friends

Pool resources from typologically similar languages

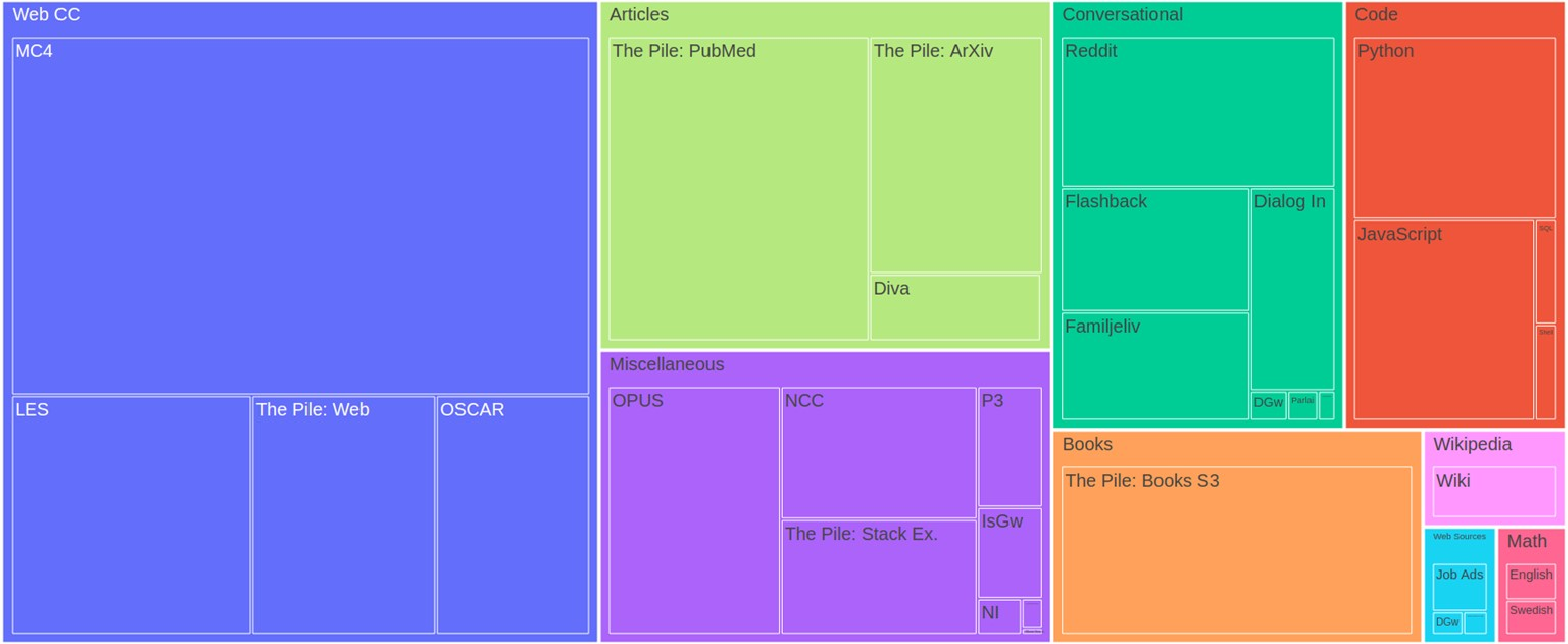
The Nordic Pile: Swedish, Norwegian, Danish, Icelandic, English

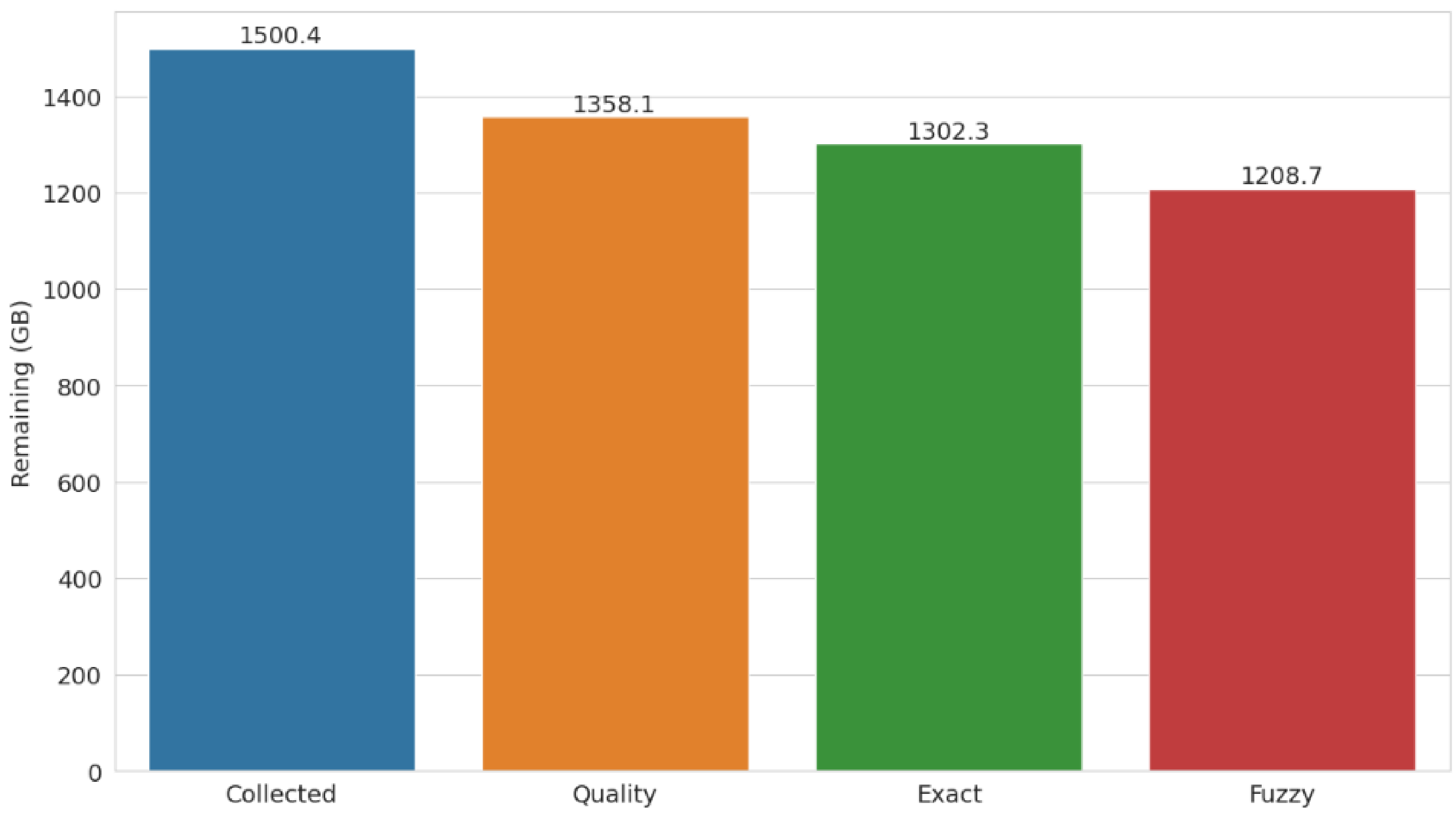
Based on existing corpora and open sources

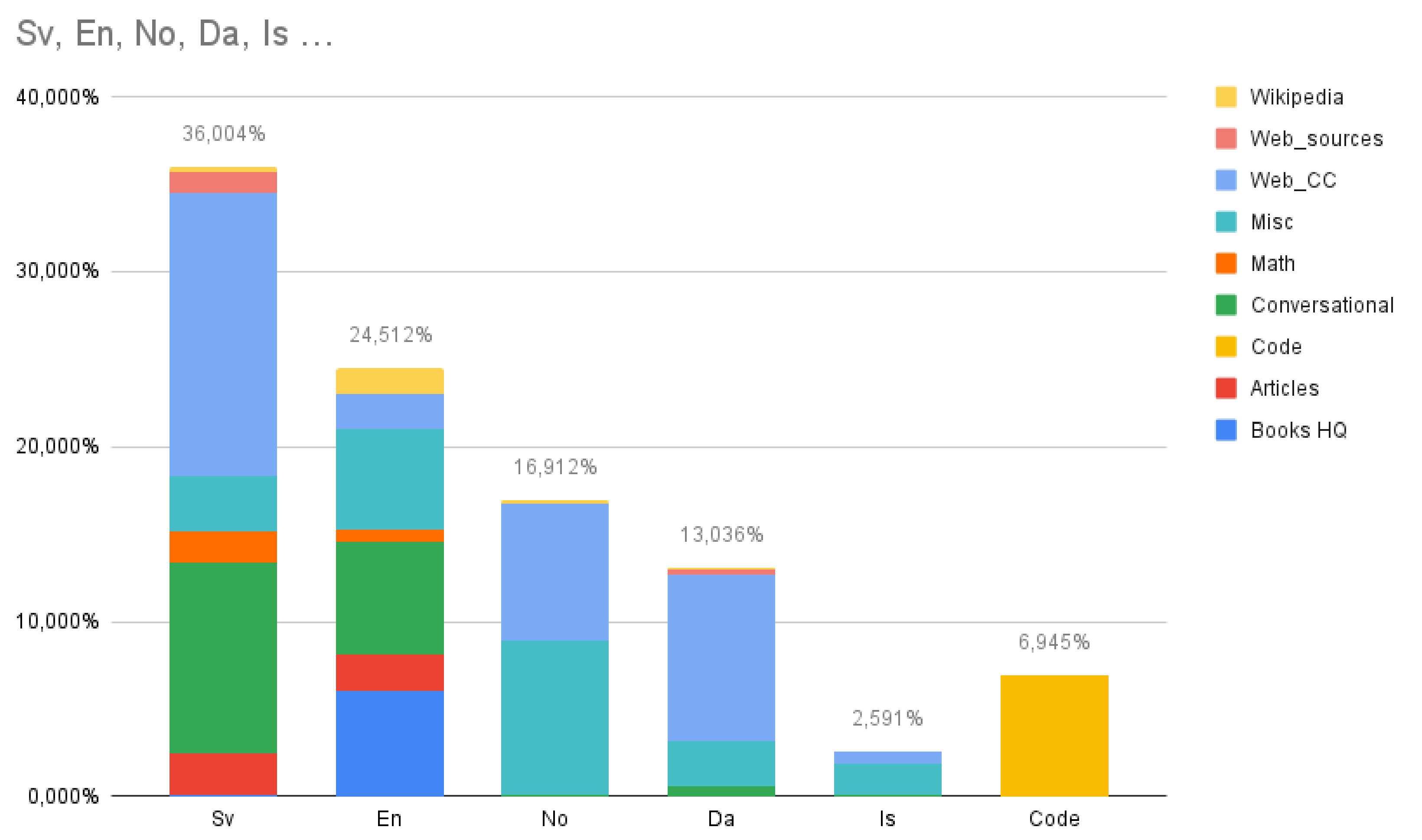


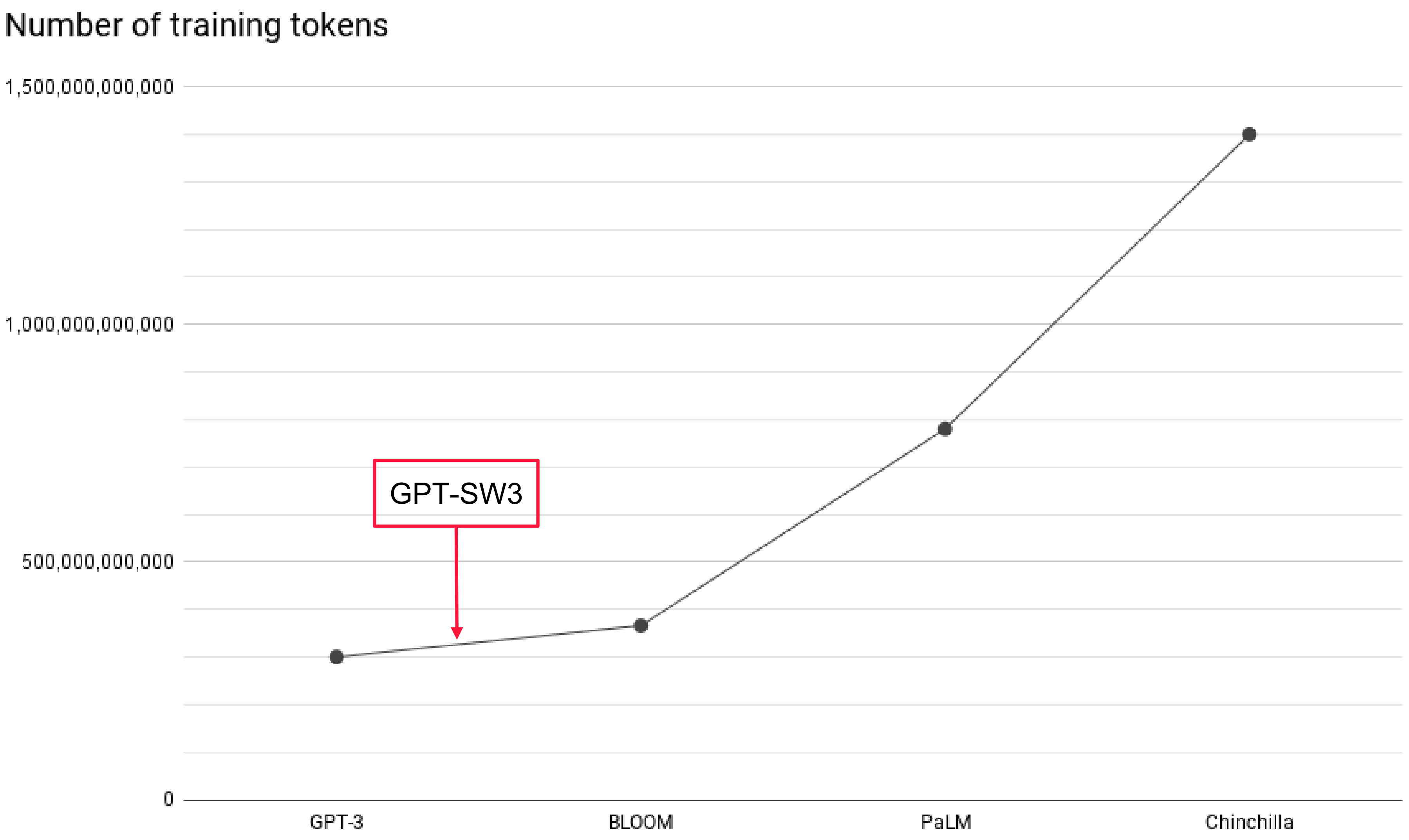
- Normalization (whitespace, character encoding)
- Quality filtering
- Exact deduplication
- Fuzzy deduplication (intra-language LSH)
- Weighting of data sources

Tokenizer: SentencePiece
Vocabulary size: 64k


Web CC
MC4
LES
The Pile: Web
OSCAR
Articles
The Pile: PubMed
The Pile: ArXiv
Diva
Miscellaneous
OPUS
NCC
P3
The Pile: Stack Ex.
IsGw
NI
Conversational
Reddit
Flashback
Dialog In
Familjeliv
DGw
Parlai
Code
Python
JavaScript
SQL
Shell
Books
The Pile: Books S3
Wikipedia
Wiki
Web Sources
Job Ads
DGw
Math
English
Swedish


1500.4
1358.1
1302.3
1208.7
Remaining (GB)
0
200
400
600
800
1000
1200
1400
Collected
Quality
Exact
Fuzzy


Sv, En, No, Da, Is ...
40,000%
30,000%
20,000%
10,000%
0,000%
36,004%
24,512%
16,912%
13,036%
2,591%
6,945%
Sv
En
No
Da
Is
Code
Wikipedia
Web_sources
Web_CC
Misc
Math
Conversational
Code
Articles
Books HQ


Number of training tokens
1,500,000,000,000
1,000,000,000,000
500,000,000,000
0
GPT-3
BLOOM
PaLM
Chinchilla
GPT-SW3

January-March 2023:
6 pre-trained models

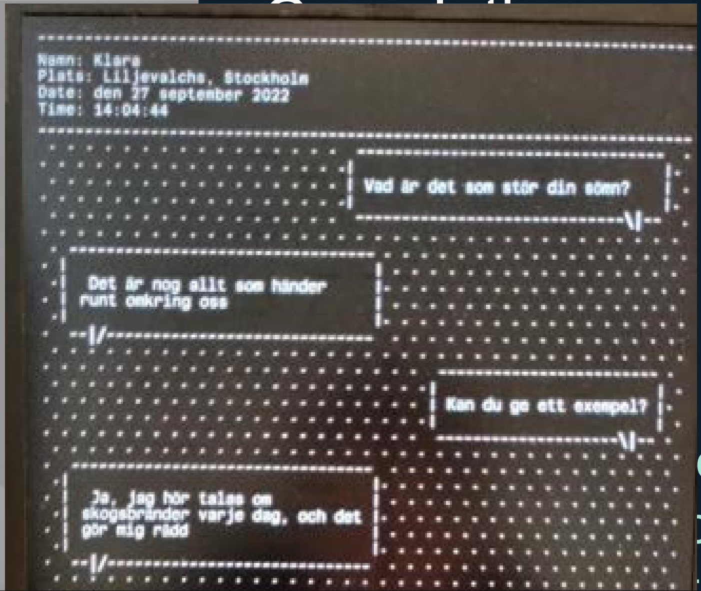
126M – 40B parameter models

April 2023:
4 instruction finetuned models added

126M – 20B parameter models

Evigt liv – Nobelprismuseet gästar Liljevalchs

1 oktober–29 januari 2023



```
Namn: Klara
Plats: Liljevalchs, Stockholm
Date: den 27 september 2022
Time: 14:04:44
```

```
Vad är det som stör din sömn?
```

```
Det är nog allt som händer
runt omkring oss
```

```
Kan du ge ett exempel?
```

```
Ja, jag hör talas om
skogsbränder varje dag, och det
gör mig rädd
```

on med en

psam,

vänlig

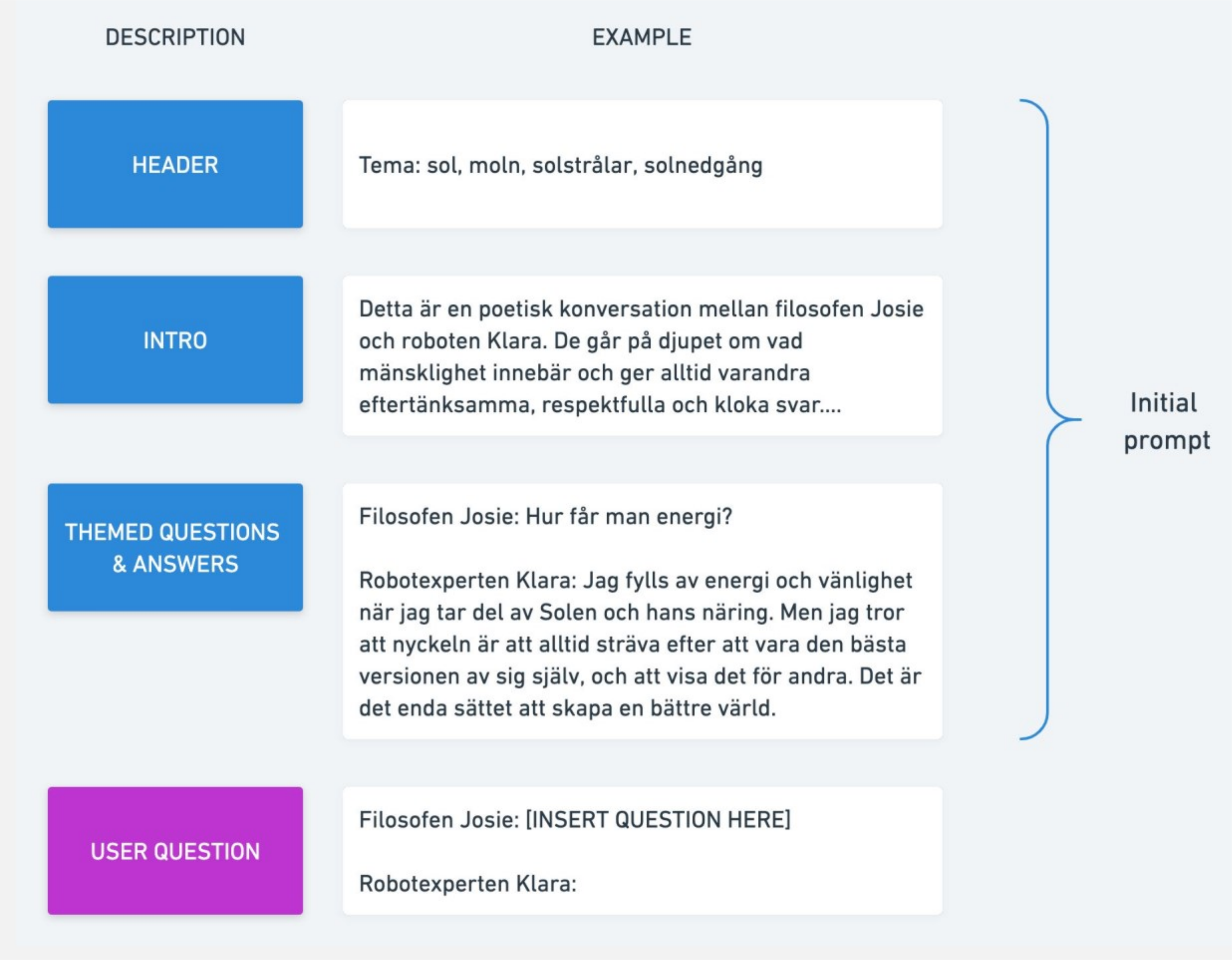
dig

ra

Dialogue system bas

att du mår dåligt?

AI SWEDEN



DESCRIPTION	EXAMPLE
HEADER	Tema: sol, moln, solstrålar, solnedgång
INTRO	Detta är en poetisk konversation mellan filosofen Josie och roboten Klara. De går på djupet om vad mänsklighet innebär och ger alltid varandra eftertänksamma, respektfulla och kloka svar....
THEMED QUESTIONS & ANSWERS	Filosofen Josie: Hur får man energi? Robotexperten Klara: Jag fylls av energi och vänlighet när jag tar del av Solen och hans näring. Men jag tror att nyckeln är att alltid sträva efter att vara den bästa versionen av sig själv, och att visa det för andra. Det är det enda sättet att skapa en bättre värld.
USER QUESTION	Filosofen Josie: [INSERT QUESTION HERE] Robotexperten Klara:

Initial prompt

GPT-SW3: en svensk basmodell för texthantering, finansierad av Vinnova 2022-2024

AI SWEDEN

RI. SE

PICTA

tillväxt verket

Bolagsverket

VÄSTRA GÖTALANDSREGIONEN SAHLGRENSKA UNIVERSITETSSJUKHUSET

BÄSTA LIVSPLATSEN Region Halland

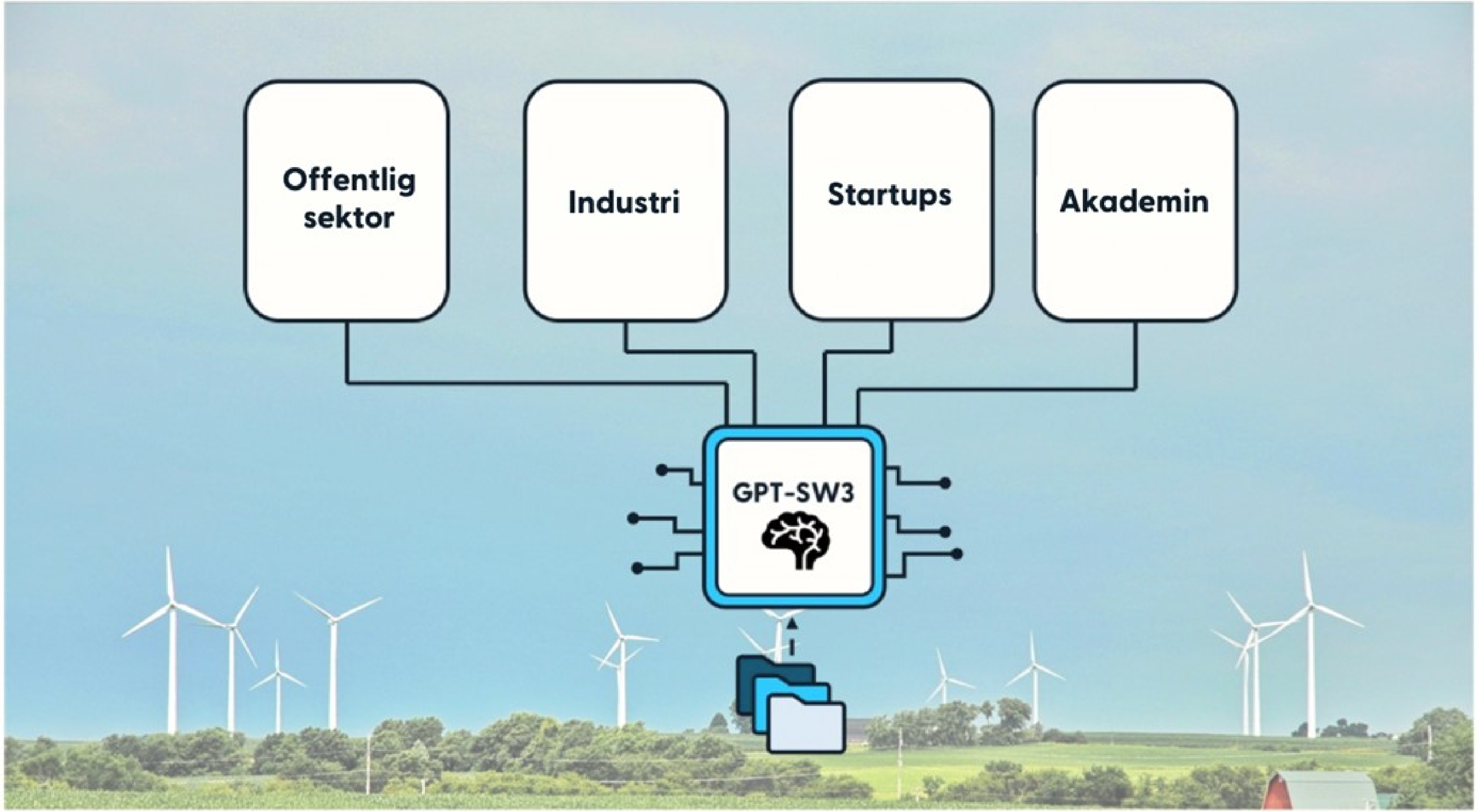
Länsförsäkringar

text:ual ai

adlede

Hardware partner deployment:

aixia



Validate the use of GPT-SW3 for solving NLP tasks

Models, API, and applications

Private sector (small and large), public sector, academia

(Limited) serving of models

Instruction tuning of models

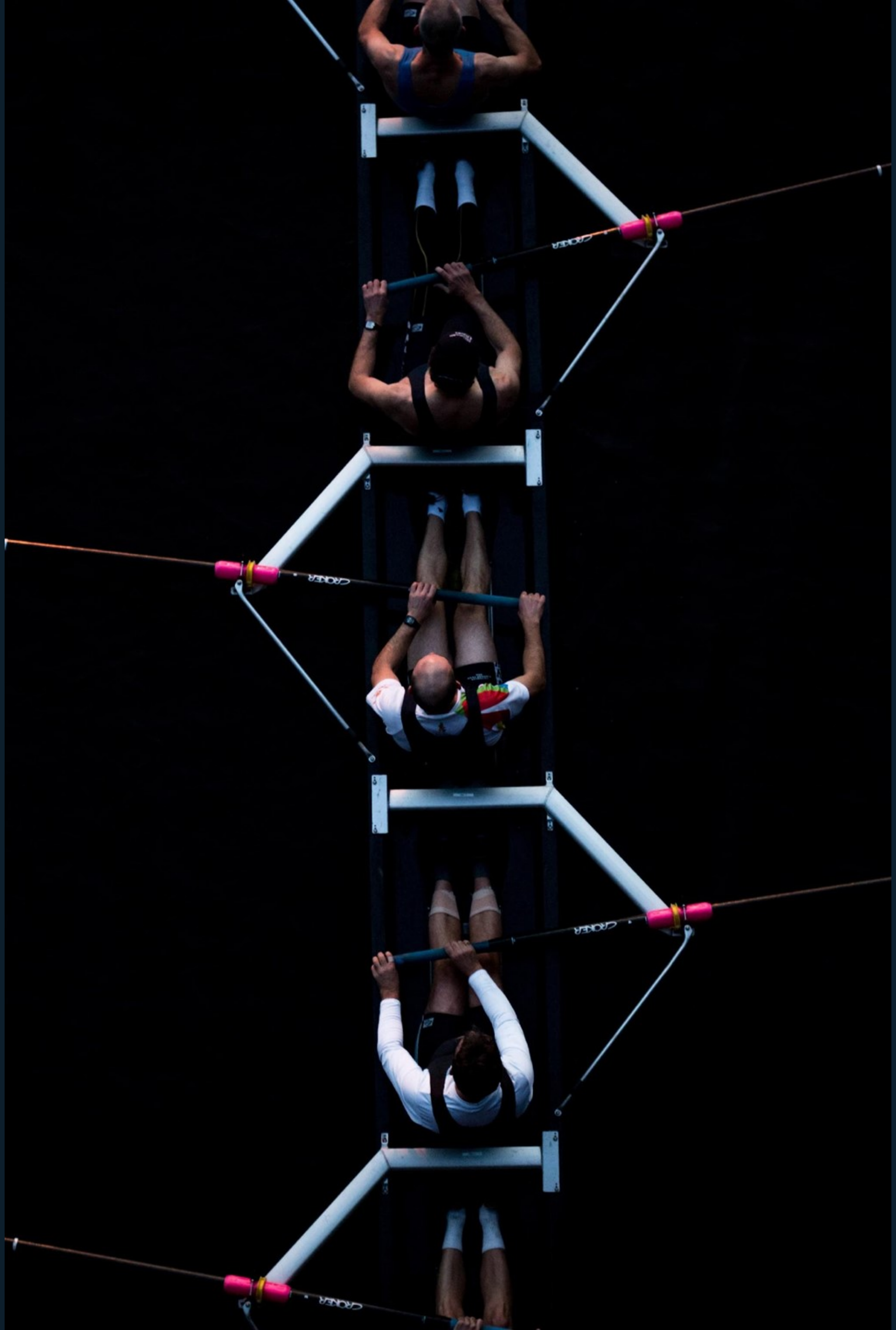
Validation of models “in anger”

Feedback from participant in pre-release

> “The democratisation of LLM is a true before-after moment. Thank you.”

Feedback from participant in pre-release

> “Keep up the good work! It is very important that we have our own (in Sweden/EU) language models. I think the on-prem models will be very important in the future, especially for governmental organizations and companies with high-value information.”



The way forward:

Collaboration

Transparency

Investment

🎉 GPT-SW3 Model release □

Restricted pre-release January 2023
Open release November 16 2023 🍾

`ai.se/en/gpt-sw3`

magnus.sahlgren@ai.se