AI at the National Library

Robin Kurtz

National Library of Sweden, KBLab


KONGL. BIBLIOTEKET

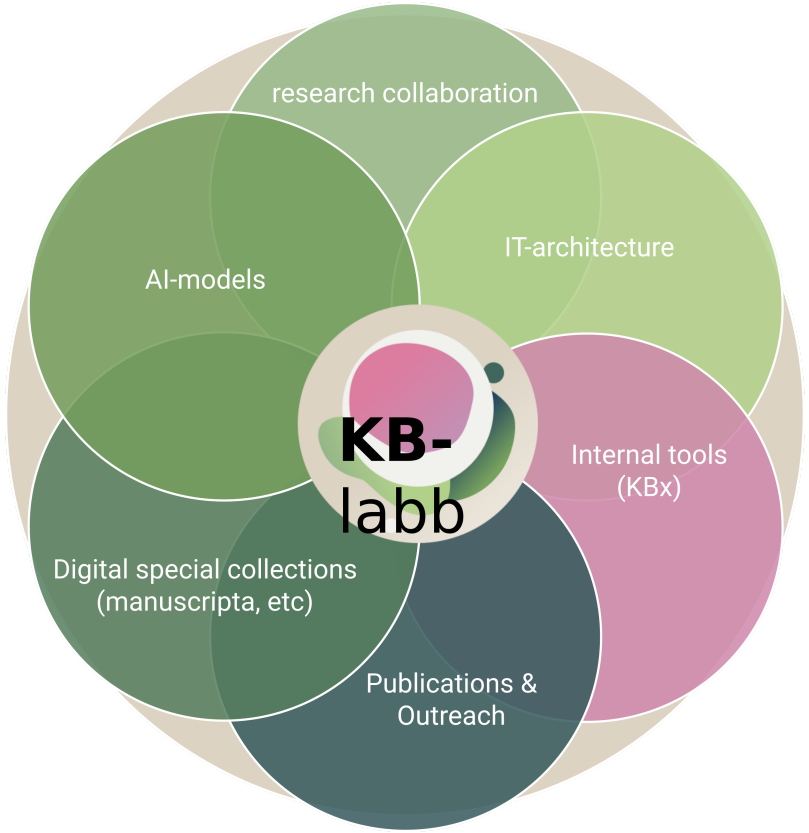
The National Library of Sweden

Kungliga Biblioteket (KB)

- collects, preserves and gives access to almost everything that is published in Sweden
- legal deposit act from 1661 required all printers to deliver one copy to KB
- a censorship law that now helps preserve Sweden's cultural heritage
- expanded in 1900 to include sound, moving images and video games
- collections currently hold over 18 million items
- ongoing digitization process

KBLab

- started in 2019 to give researchers the possibility to do large-scale quantitative research
- curate data maintained by the National Library
- train models on data to be used by academia, governmental organizations and industry



What do we need?

- data
- compute

Data

KB

- (digitized) newspaper
- governmental reports
- webcrawl
- TV / film
- radio & more
- *vardagstryck*
- postcards
- ...

https://lab.kb.se/bildsok/search?q=nalle&size=100

Bildsök DEMO

OM TJÄNSTEN

TEXT BILD Välj sökning med text eller bild.

nalle

SÖK

Hur fungerar det här?

Fritextsök, t.ex.: "staty", "montage", "häst och vagn i vintermiljö på natten"

Sökresultat för "nalle"

Visar 100 av 100 träffar

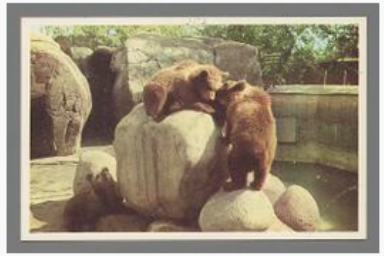

Brunbjörnar på Skansen i Stockholm - 2


Brunbjörnar på Skansen i Stockholm - 15


Brunbjörnar på Skansen i Stockholm - 7


Isbjörnar på Skansen i Stockholm - 1


Brunbjörnar på Skansen i Stockholm - 12

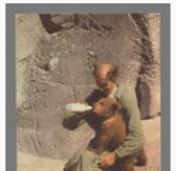

Brunbjörnar på Skansen i Stockholm - 4

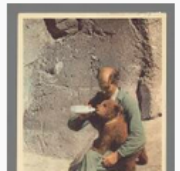

Brunbjörnar på Skansen i Stockholm - 8

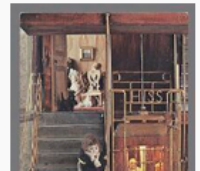

Brunbjörnar på Skansen i Stockholm - 1


Isbjörnar på Skansen i Stockholm - 3


Leksaksmuseet i Stockholm - 2







Compute

- KB has some inhouse compute for finetuning or smaller models
- previous development accesses at Vega and LUMI
- KBLab was awarded 5,000,000 core hours on MeluXina
- ongoing development access on Leonardo for audio models

MeluXina

- Luxembourg
- 200 nodes
  - 4 GPUs
  - 40GB memory

What is a Language Model?

- frequency-based n-gram model
- transformer-based self-trained base model
  - predict missing words
  - trained once and finetuned often

How do we Teach Models Language?

GPT

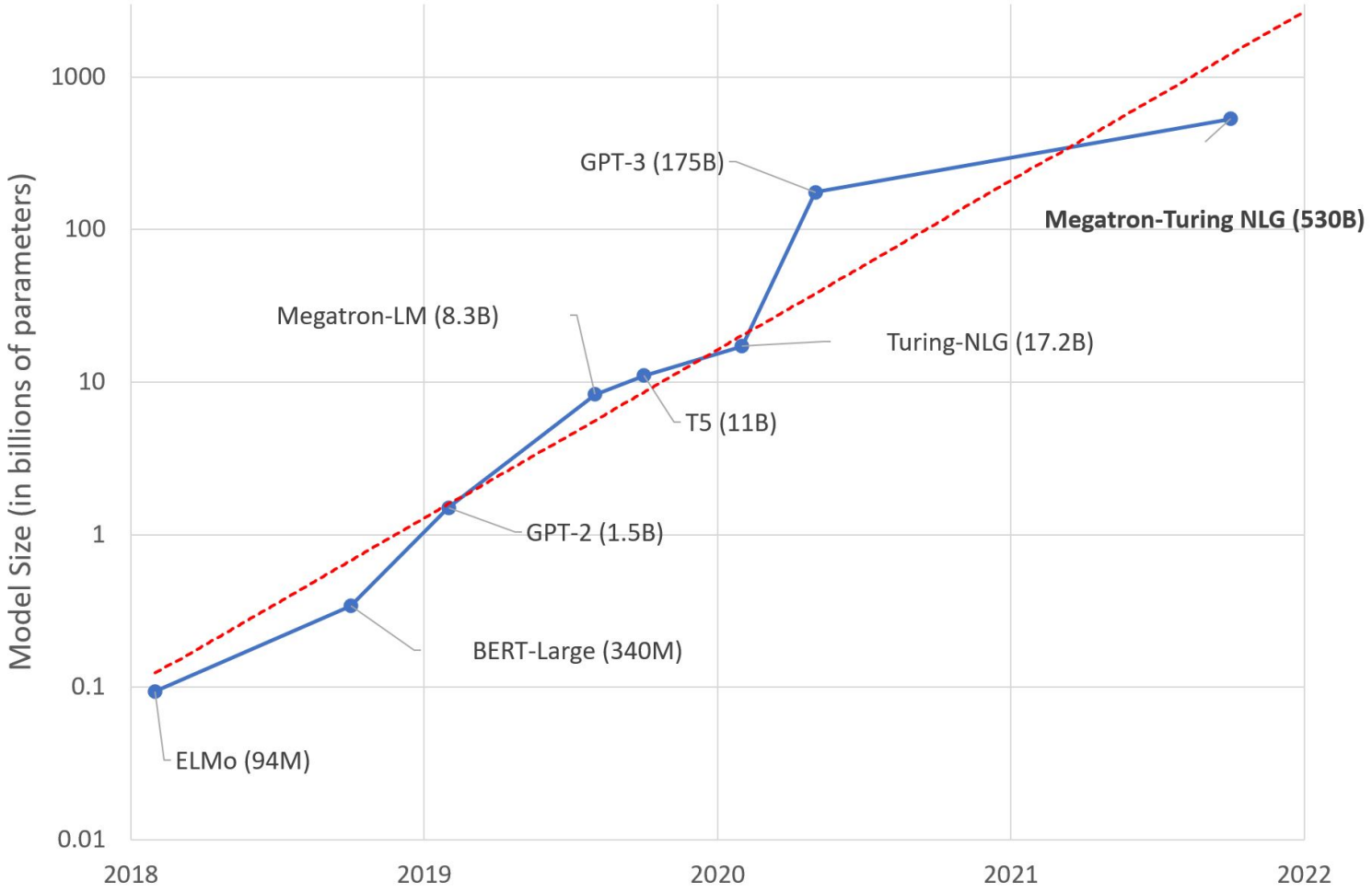
At the library we work with [MASK]

What can you do with an LM?

- base model
- sentiment analysis
- named entity recognition
- part-of-speech tagging
- relation between texts
- correctness
- ...

What is a Large Language Model?

- **L**arge **L**anguage **M**odel



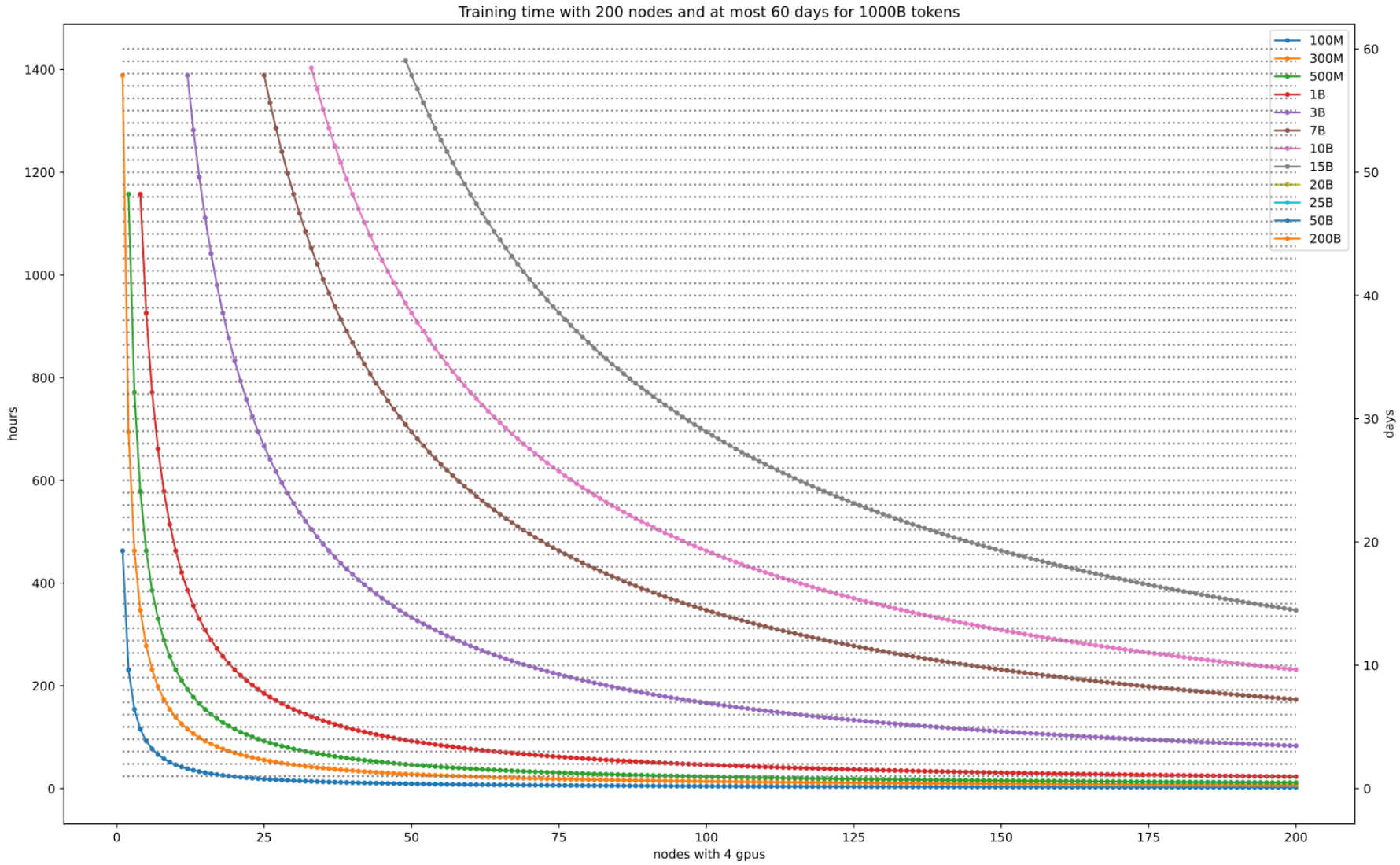
What can you do with an LLM?

- base model
  - but can do more out-of-the-box (zero-shot)
  - few-shot
- somewhat magical
- text generation
- Chat-GPT
- summarization
- ...

Goals

- train 20B parameter GPT model
- some smaller models of different types

How much compute do you need?



Pitfalls

- core hours sound more impressive than they are
  - 5,000,000 core hours are ~6,600 node hours per month
- cluster A with A100 does not behave like cluster B with A100
- scheduled monthly
  - use it or lose it
- your cluster with A100 does not behave like cluster from paper *xyz* with A100

Goals

- continue training existing fully open models
- publish these models with the same open licenses
  - 20B GPT-NeoX by Eleuther-AI
  - 7B Llama by OpenLM-Research
  - 3B Llama by OpenLM-Research

Where to find us

- https://huggingface.co/KBLab
- https://kb-labb.github.io/
- https://www.kb.se/in-english/research-collaboration/kblab.html
- https://lab.kb.se/bildsok/