# Large Language Models for Swedish

Robin Kurtz

**National Library of Sweden, KBLab**


KUNGL. BIBLIOTEKET

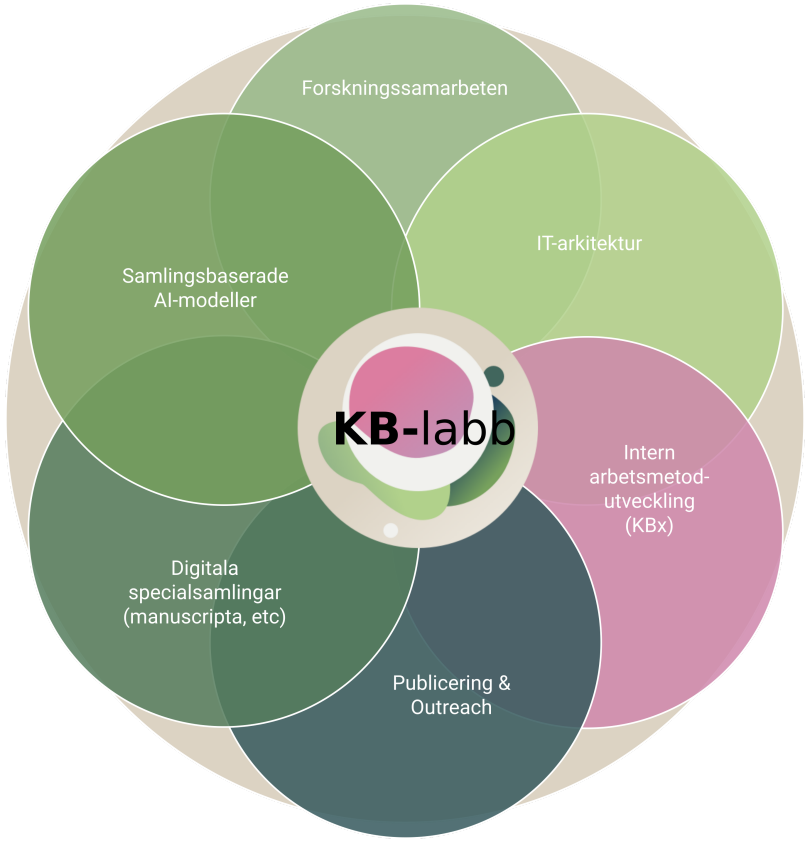
# Kungliga Biblioteket

## The National Library of Sweden

- 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

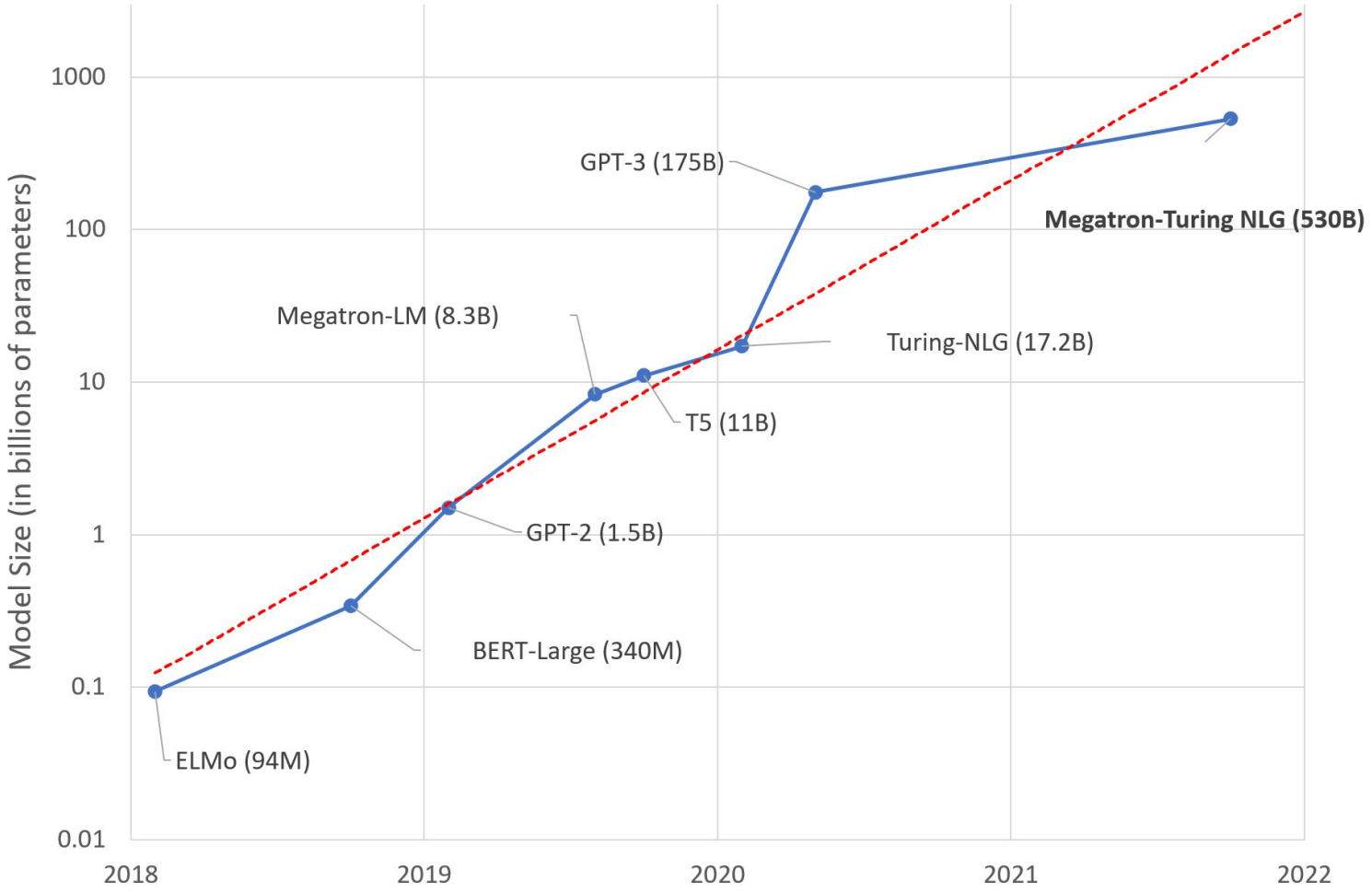


# What is a Language Model?

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

# What is a Large Language Model?

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



# What do we need?

- data
- compute

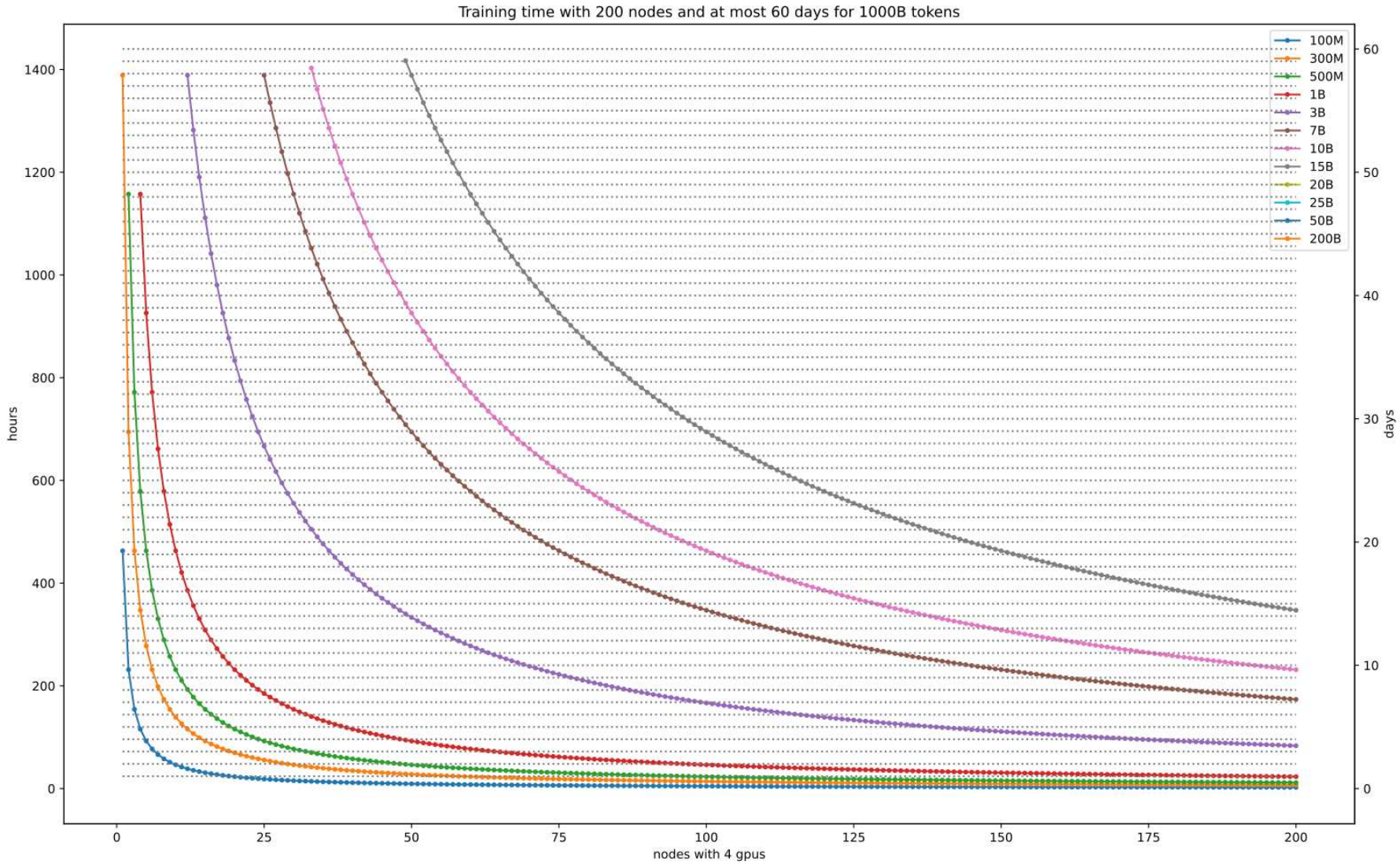
# Compute

- KB has some inhouse compute for finetuning or smaller models
- KBLab was awarded 5,000,000 core hours on MeluXina

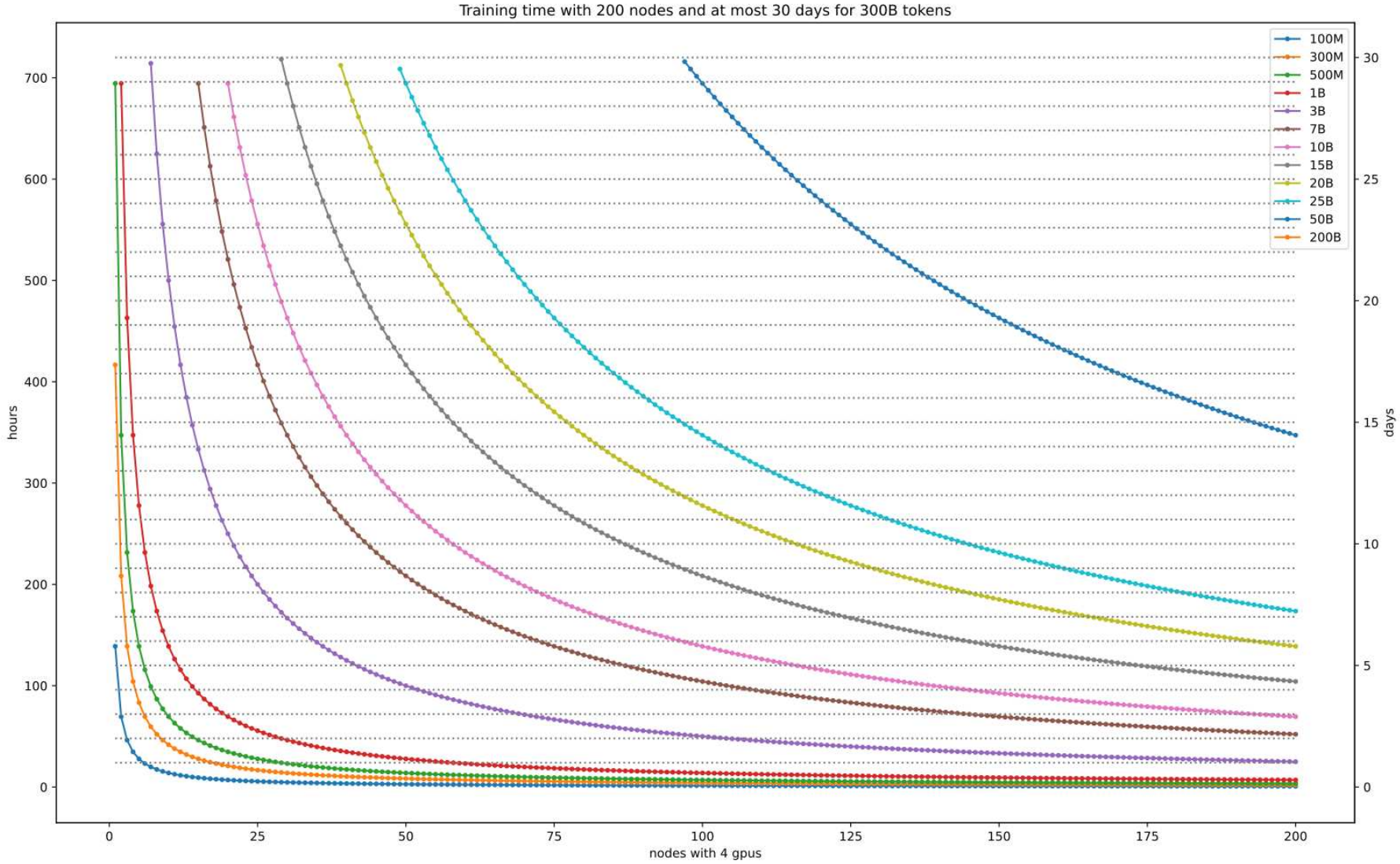
# MeluXina

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

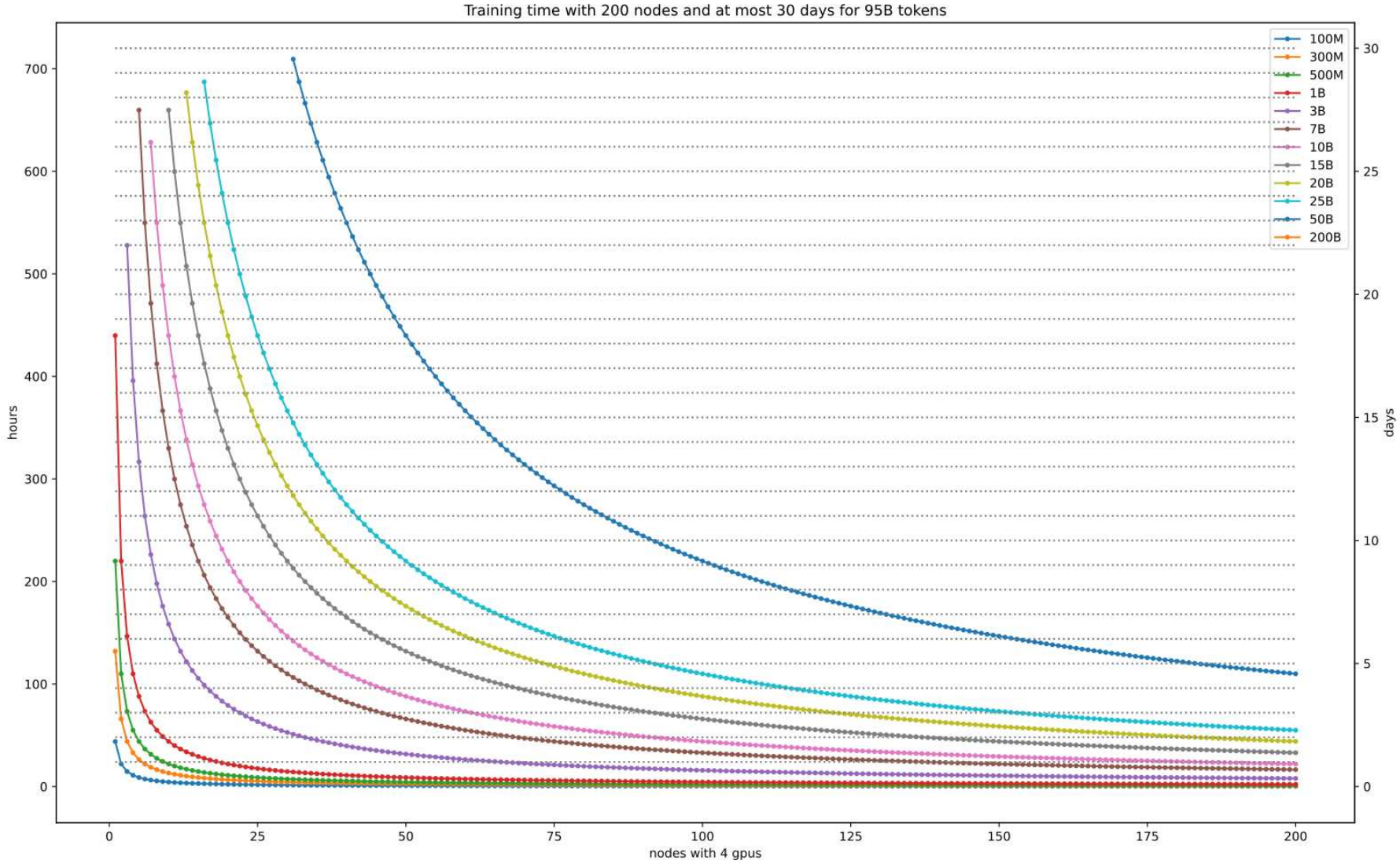
# How much compute do you need?



# How much compute do you need?



# How much compute do you need?



# Data

## KB

- newspaper
- parliamentary proceedings
- Swedish Government Official Reports
- own webcrawl

# Data

## Web

- Prepared corpora
  - mc4
  - OSCAR
  - The Pile
  - ...
- Wikipedia
- ...

# Clean Data

- OCR mistakes
- duplicates
- non-Swedish
- non-text
- anonymize

# Goals

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

# Goals

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

# Questions ?