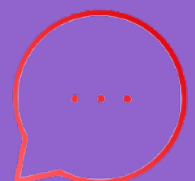


Large Generative Models for a smaller language

Finnish GPT-3



TURKUNLP
.ORG



TURUN
YLIOPISTO

GPT-3

English generative models
created by OpenAI in 2020

Parameter counts ranging 125M
("small") to **175B** ("GPT-3")

Pre-trained for 300B tokens on
~600G of text (<1 epoch)

Usable via API, model weights
not available

GPT-3

English generative models
created by OpenAI in 2020

Parameter counts ranging 125M
("small") to **175B** ("GPT-3")

Pre-trained for 300B tokens on
~600G of text (<1 epoch)

Usable via API, model weights
not available

Finnish GPT-3

Finnish generative models created
by TurkuNLP in 2023

Parameter counts 125M - 13B
(monolingual), **176B** (BLOOM+Fi)

Pre-trained for 300B tokens on
~200G of text (>1 epoch)

Available for download now!
<https://turkunlp.org/gpt3-finnish>

Quickstart

<https://turkunlp.org/gpt3-finnish>

```
[1] !pip install --quiet transformers
```

```
[2] from transformers import AutoTokenizer, AutoModelForCausalLM
```

```
MODEL = 'TurkuNLP/gpt3-finnish-medium'
```

```
tokenizer = AutoTokenizer.from_pretrained(MODEL)
```

```
model = AutoModelForCausalLM.from_pretrained(MODEL)
```

```
input = tokenizer('Turussa parasta on', return_tensors='pt')
```

```
output = model.generate(**input, max_new_tokens=10)
```

```
print(tokenizer.decode(output[0], skip_special_tokens=True))
```

Turussa parasta on se, että täällä on niin paljon erilaisia tapahtumia.

Most likely
continuation

Generation example

```
tokenizer.decode(model.generate(**tokenizer(prompt, [...]
```

Generation by Finnish GPT-3 (🍒/6)

Suomenkielisen tekoälytekniikan tulevaisuudelle on keskeisen tärkeää, että suomalainen koulujärjestelmä tarjoaa riittävän perusosaamisen tekoälyn hyödyntämiseen myös lapsille ja nuorille. Koulutusjärjestelmämme tulee taata lapsille jo varhaisessa vaiheessa valmiudet, tiedot ja taidot, joilla he pystyvät luomaan ja jakamaan itse tietoa tekoälyyn liittyen. Tämän lisäksi tulee kiinnittää huomiota tekoälyn opettamiseen ja siihen, miten tekoäly linkittyy eri oppiaineisiin. Tekoälyyn liittyvää opetusta tulisi sisällyttää opetussuunnitelmiin entistä paremmin. [...]

English translation by DeepL

It is crucial for the future of artificial intelligence technology in Finnish that the Finnish school system provides sufficient basic skills for children and young people to use artificial intelligence. Our education system must ensure that children have the skills, knowledge and abilities at an early stage to create and share knowledge about artificial intelligence themselves. In addition, attention must be paid to the teaching of artificial intelligence and how it is linked to different subjects. Artificial intelligence education should be better integrated into curricula. [...]

Creating Finnish GPT-3

Creating large LMs

Key ingredients

- **Compute:** LUMI supercomputer 
- **Software:** Megatron-DeepSpeed (BigScience fork) 
- **Data:** National library, crawls, news, social media, ... 

Creating large LMs

Key ingredients

- **Compute:** LUMI supercomputer 
- **Software:** Megatron-DeepSpeed (BigScience fork) 
- **Data:** National library, crawls, news, social media, ... 

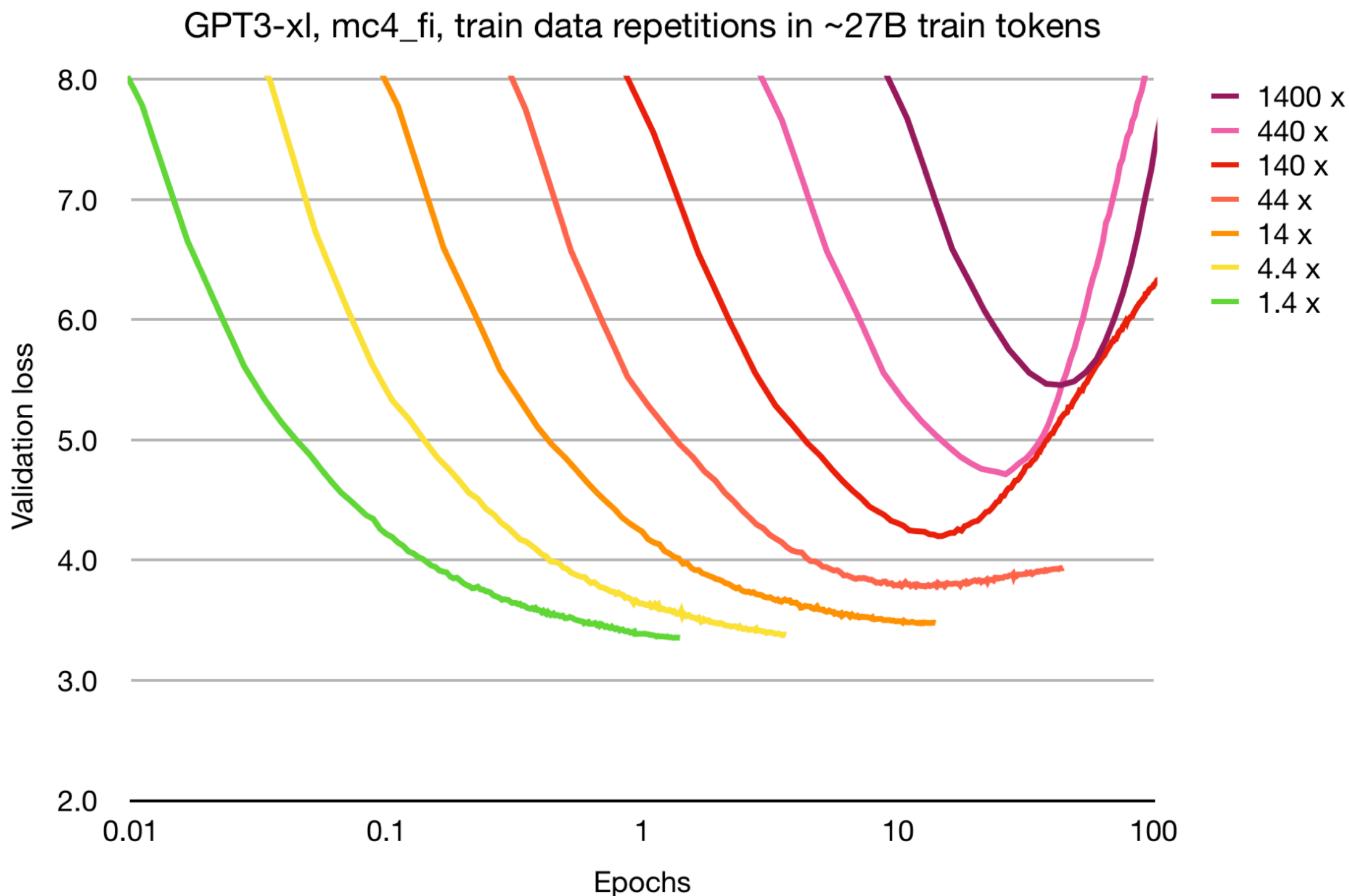
Finnish is a small language with few friends 

Where to find the data?

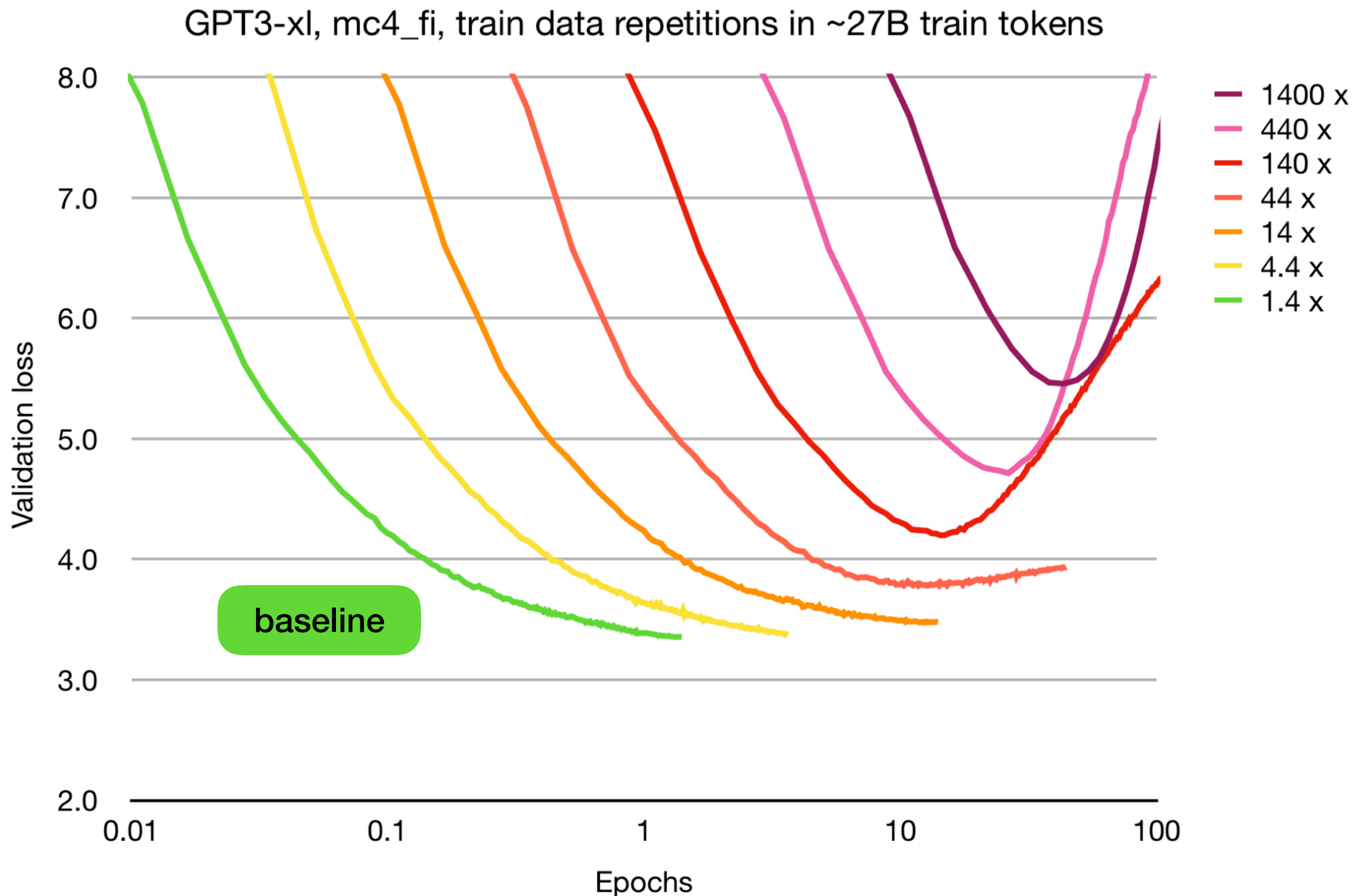
Chinchilla compute-optimality estimates

Parameters	FLOPs	Tokens	
400 Million	1.92e+19	8.0 Billion	
1 Billion	1.21e+20	20.2 Billion	
10 Billion	1.23e+22	205.1 Billion	Data
67 Billion	5.76e+23	1.5 Trillion	
175 Billion	3.85e+24	3.7 Trillion	Compute
280 Billion	9.90e+24	5.9 Trillion	

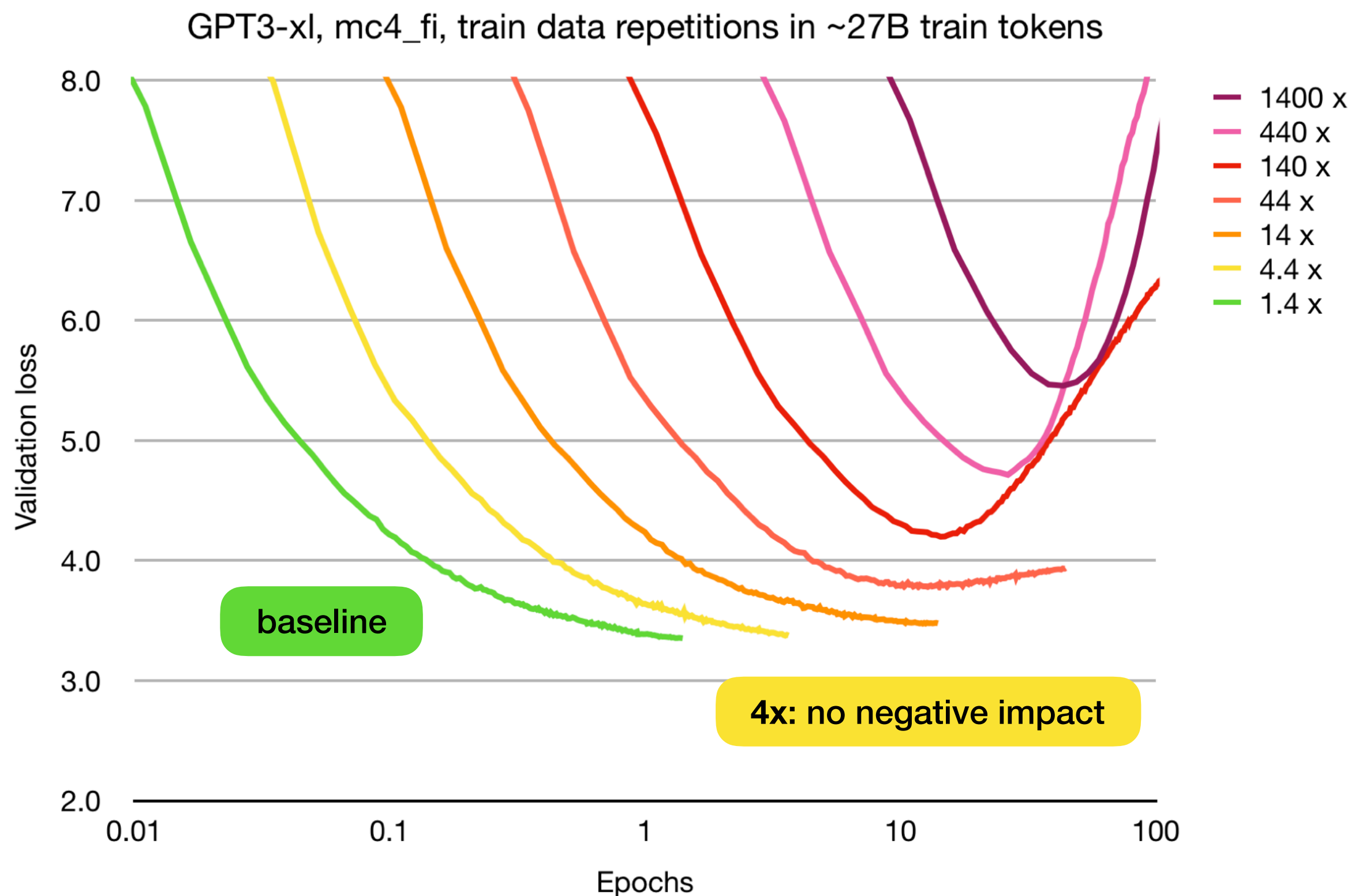
Repeating training data



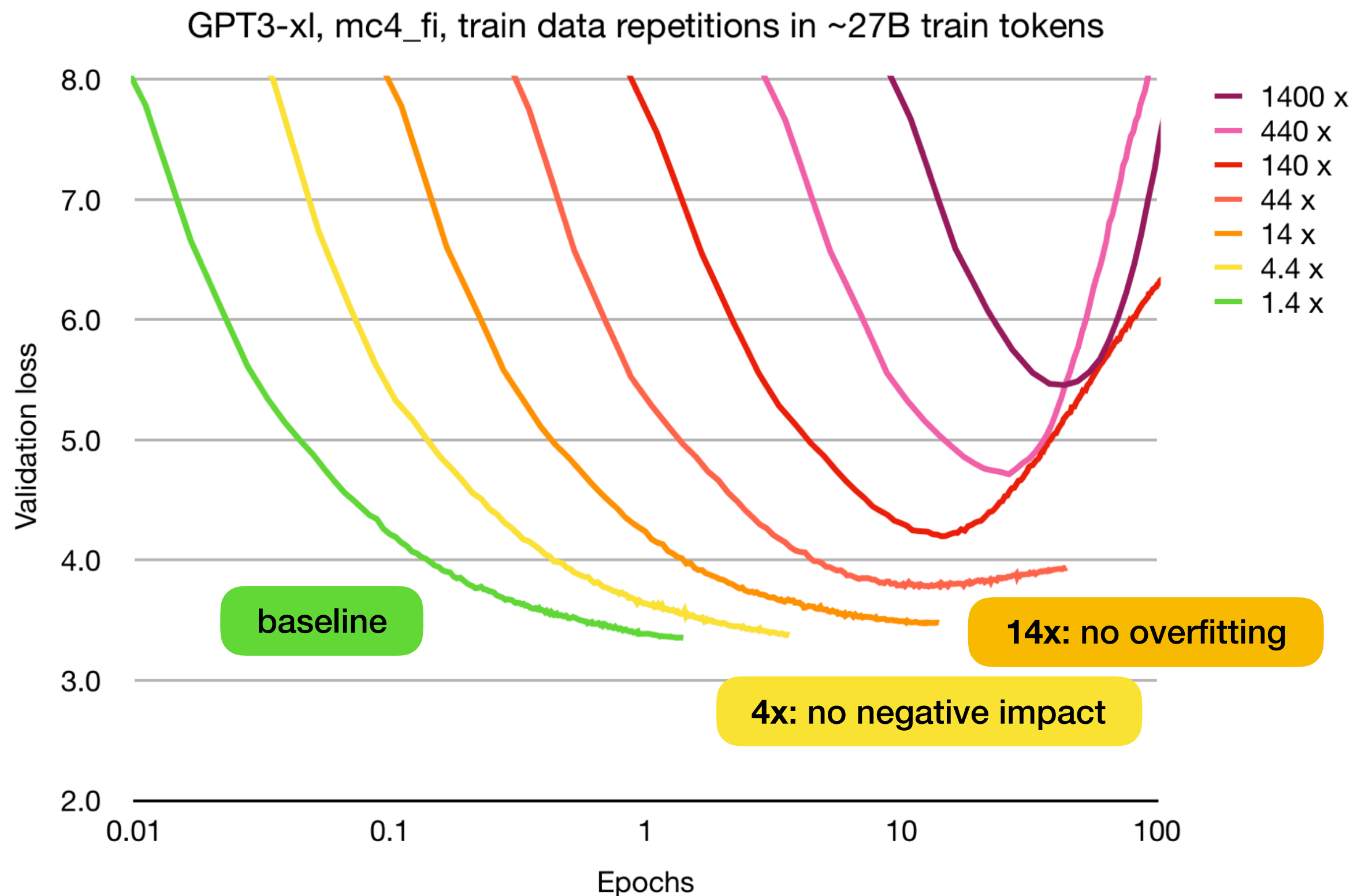
Repeating training data



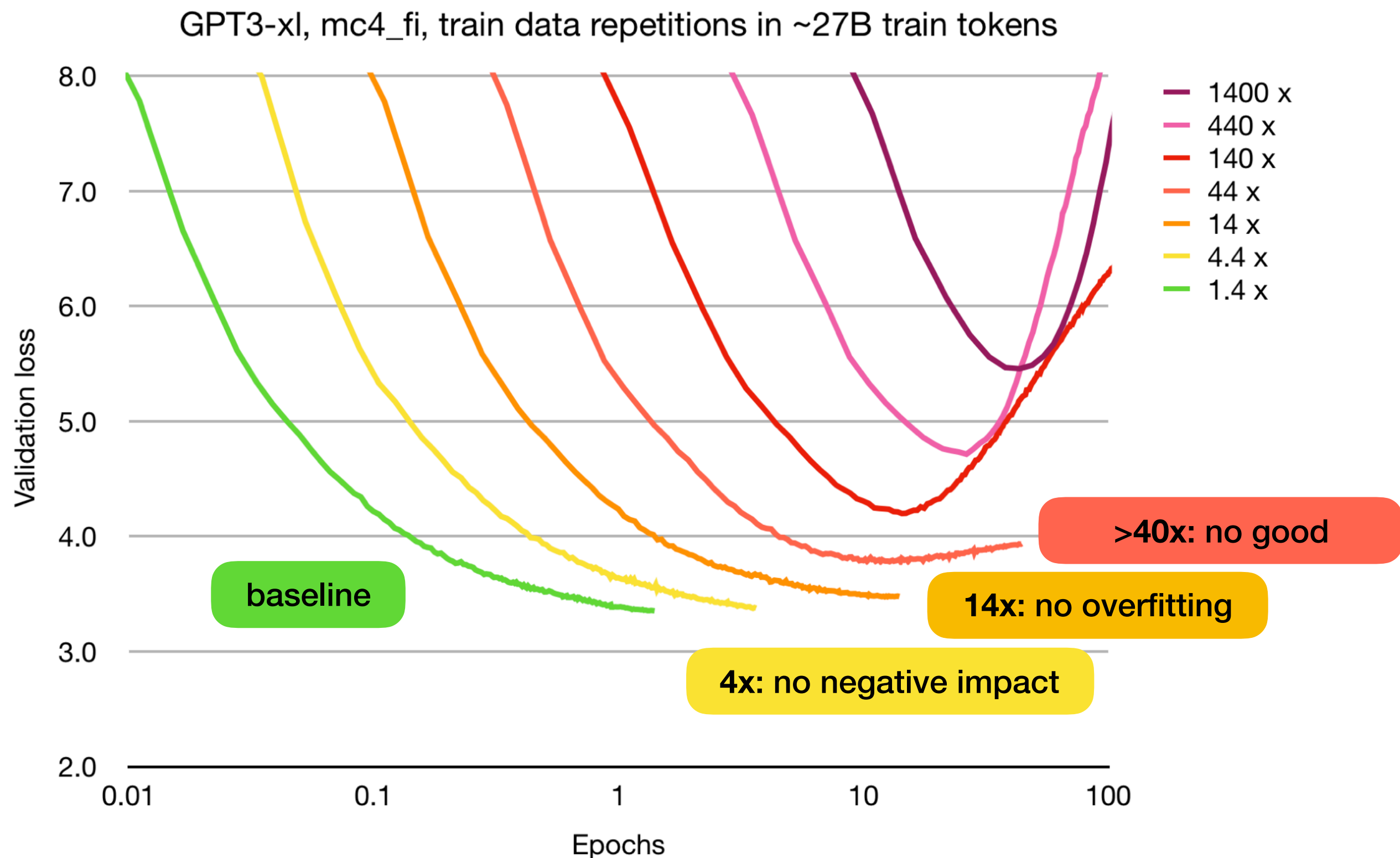
Repeating training data



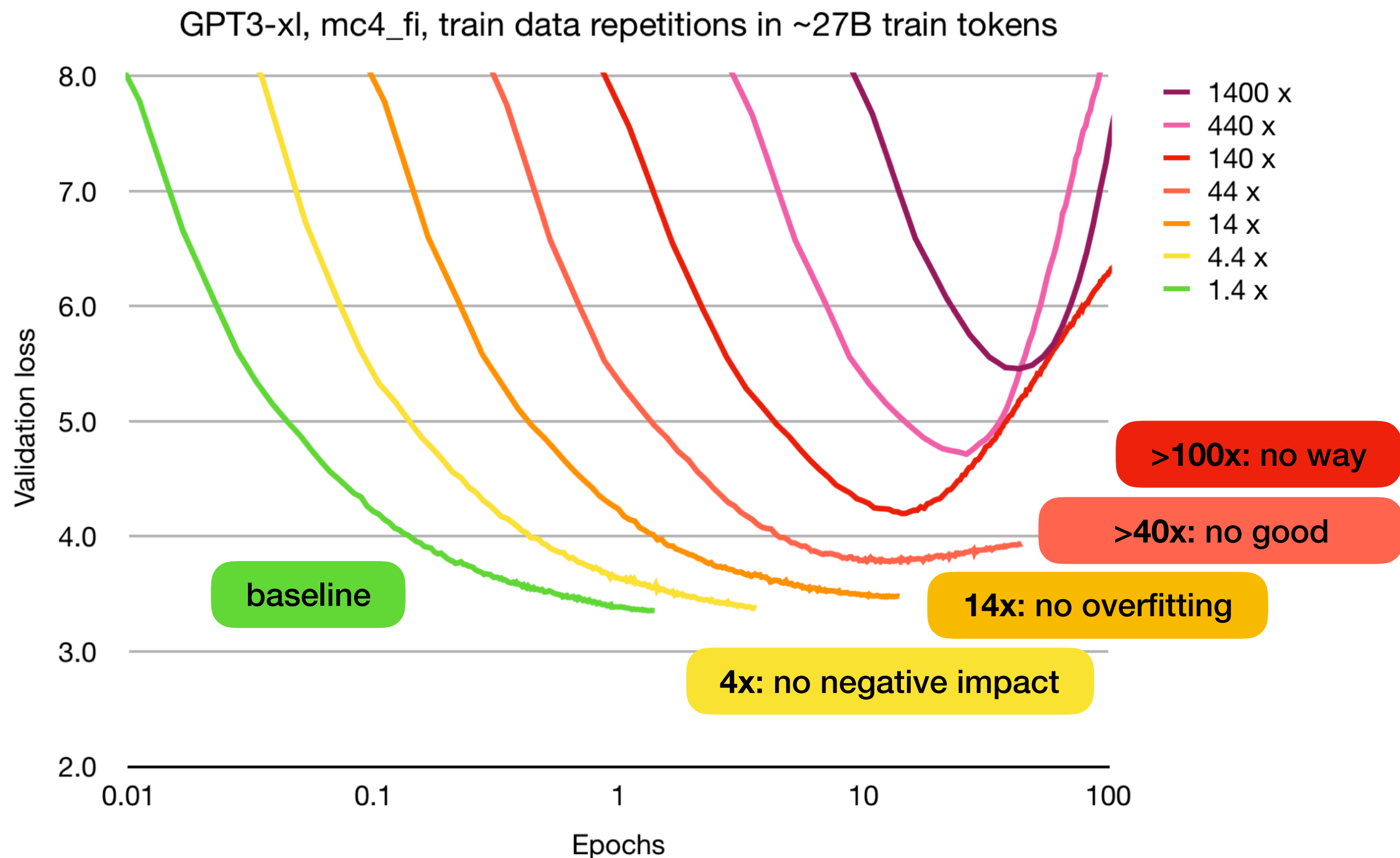
Repeating training data



Repeating training data



Repeating training data



Where to find the data?

Chinchilla compute-optimality estimates

Parameters	FLOPs	Tokens
400 Million	1.92e+19	8.0 Billion
1 Billion	1.21e+20	20.2 Billion
10 Billion	1.23e+22	205.1 Billion
67 Billion	5.76e+23	1.5 Trillion
175 Billion	3.85e+24	3.7 Trillion
280 Billion	9.90e+24	5.9 Trillion

Compute

Where to find the data?

Chinchilla compute-optimality estimates

Parameters	FLOPs	Tokens	w/10x repeat
400 Million	1.92e+19	8.0 Billion	
1 Billion	1.21e+20	20.2 Billion	
10 Billion	1.23e+22	205.1 Billion	
67 Billion	5.76e+23	1.5 Trillion	
175 Billion	3.85e+24	3.7 Trillion	Compute
280 Billion	9.90e+24	5.9 Trillion	

Where to find the data?

Chinchilla compute-optimality estimates

Parameters	FLOPs	Tokens	w/10x repeat
400 Million	1.92e+19	800 Million	
1 Billion	1.21e+20	20.2 Billion	
10 Billion	1.23e+22	205.1 Billion	
67 Billion	5.76e+23	1.5 Trillion	
175 Billion	3.85e+24	3.7 Trillion	
280 Billion	9.90e+24	5.9 Trillion	

Compute

Where to find the data?

Chinchilla compute-optimality estimates

Parameters	FLOPs	Tokens	w/10x repeat
400 Million	1.92e+19	800 Million	
1 Billion	1.21e+20	2 Billion	
10 Billion	1.23e+22	205.1 Billion	
67 Billion	5.76e+23	1.5 Trillion	
175 Billion	3.85e+24	3.7 Trillion	
280 Billion	9.90e+24	5.9 Trillion	

Compute

Where to find the data?

Chinchilla compute-optimality estimates

Parameters	FLOPs	Tokens	w/10x repeat
400 Million	1.92e+19	800 Million	
1 Billion	1.21e+20	2 Billion	
10 Billion	1.23e+22	21 Billion	
67 Billion	5.76e+23	150 Billion	
175 Billion	3.85e+24	370 Billion	
280 Billion	9.90e+24	590 Billion	

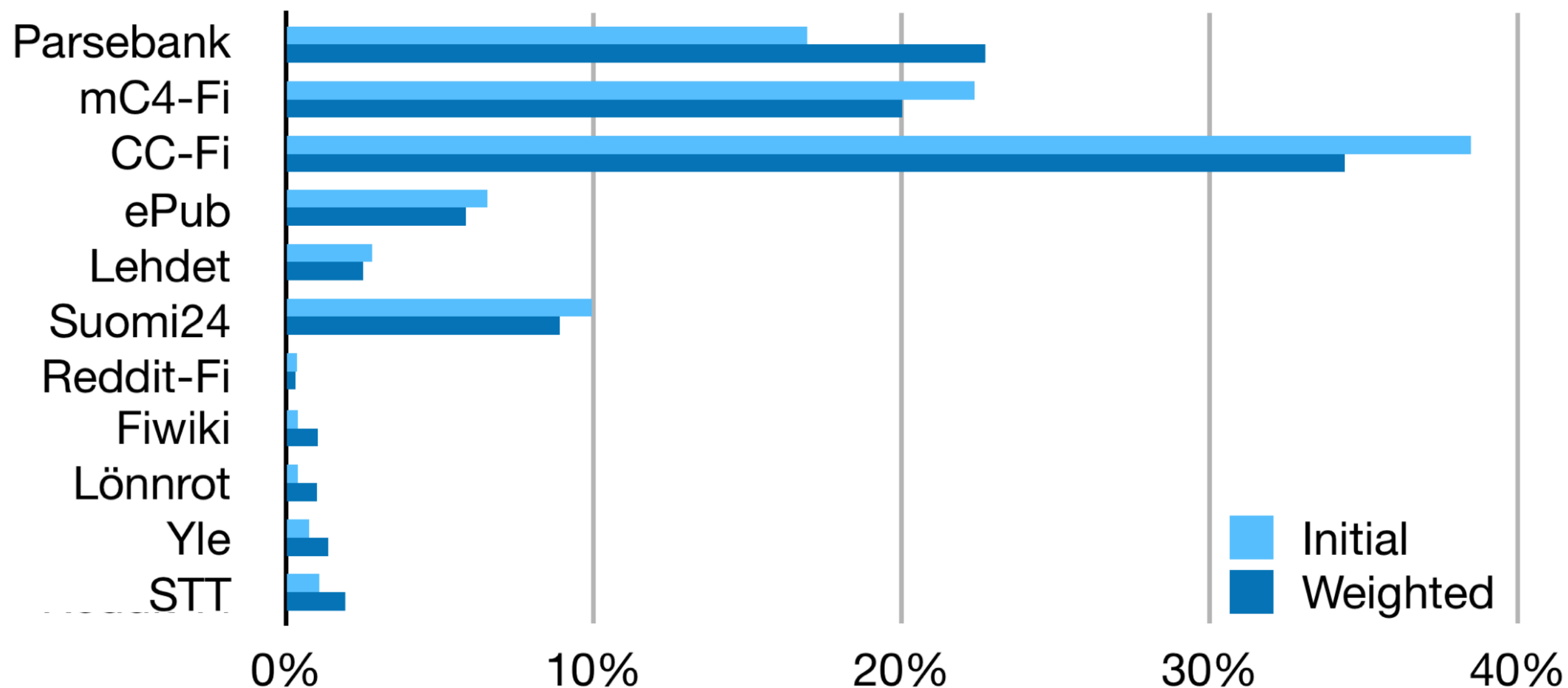
Compute

Where to find the data?

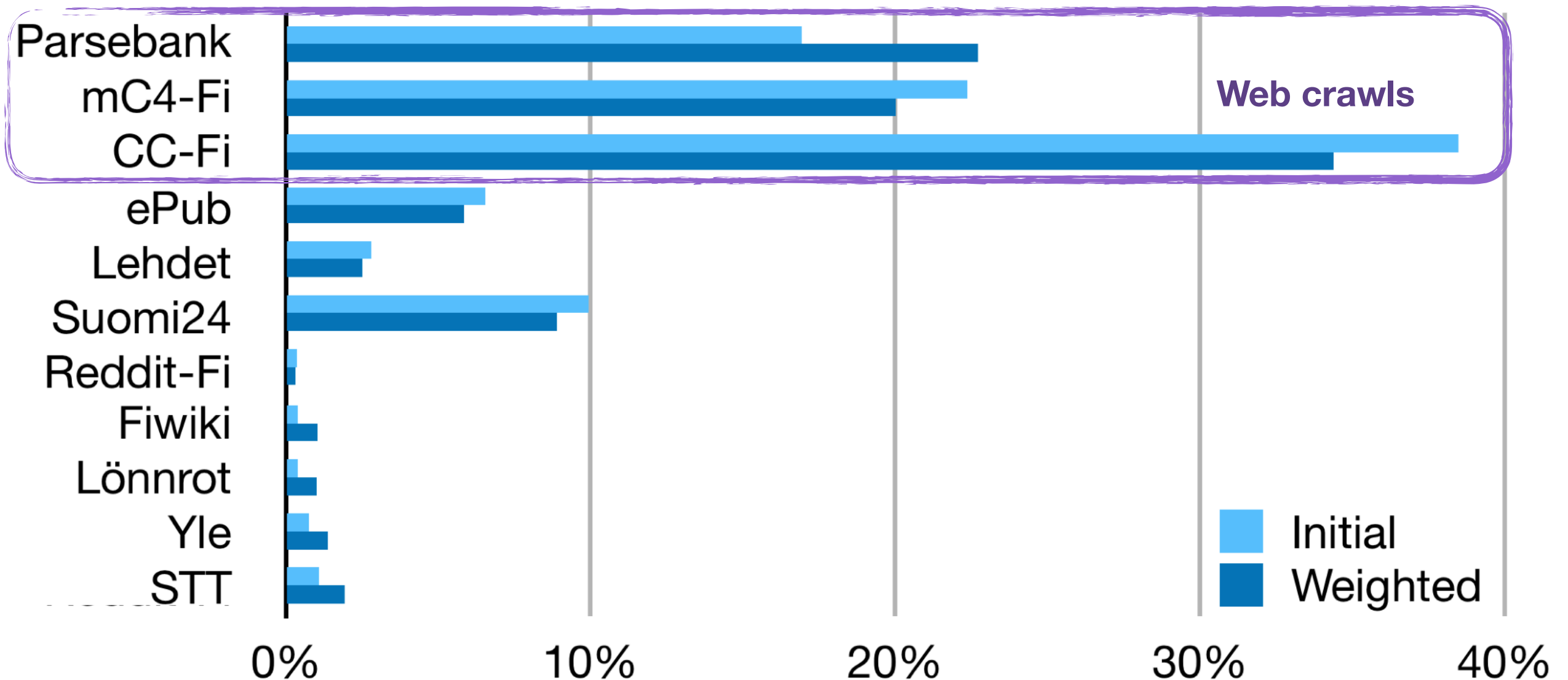
Chinchilla compute-optimality estimates

Parameters	FLOPs	Tokens	w/10x repeat
400 Million	1.92e+19	800 Million	
1 Billion	1.21e+20	2 Billion	
10 Billion	1.23e+22	21 Billion	
67 Billion	5.76e+23	150 Billion	Data
175 Billion	3.85e+24	370 Billion	Compute
280 Billion	9.90e+24	590 Billion	

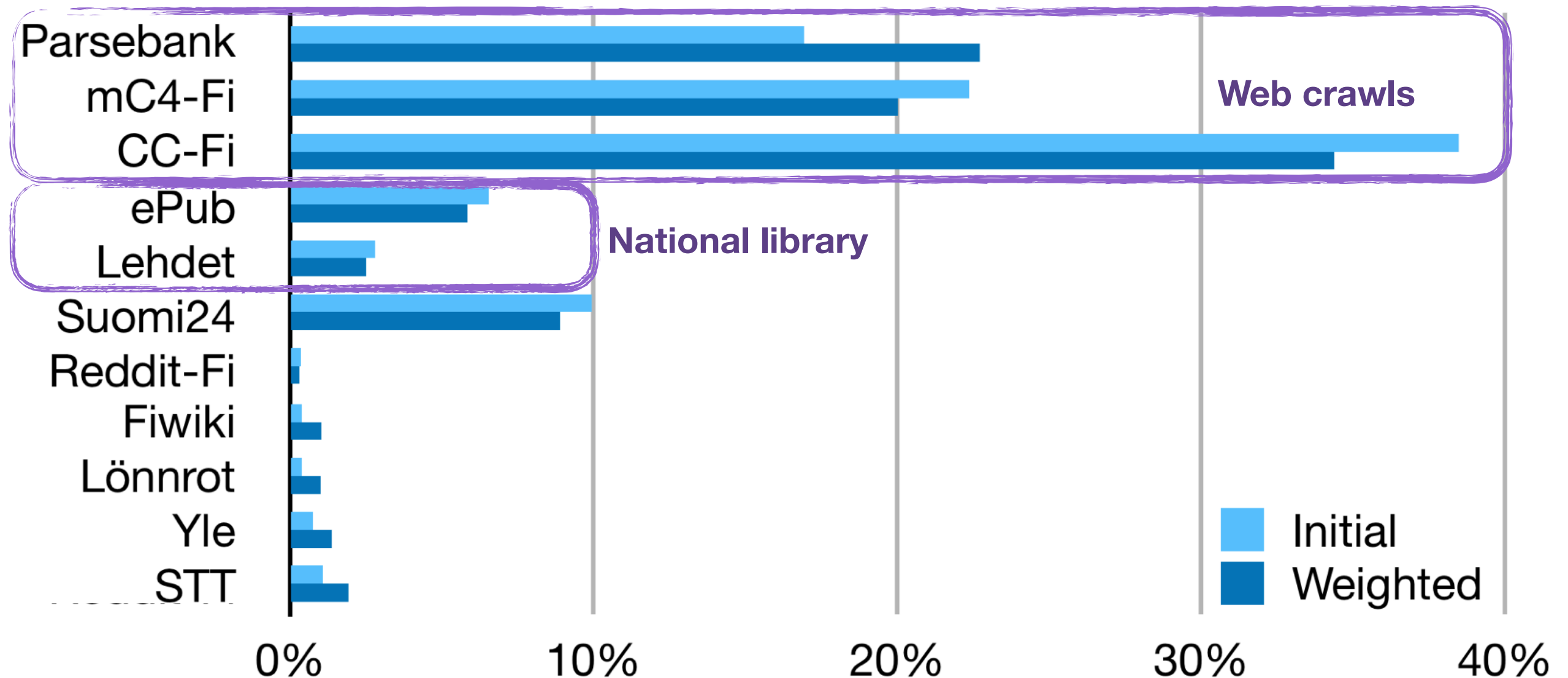
Data for Finnish GPT-3



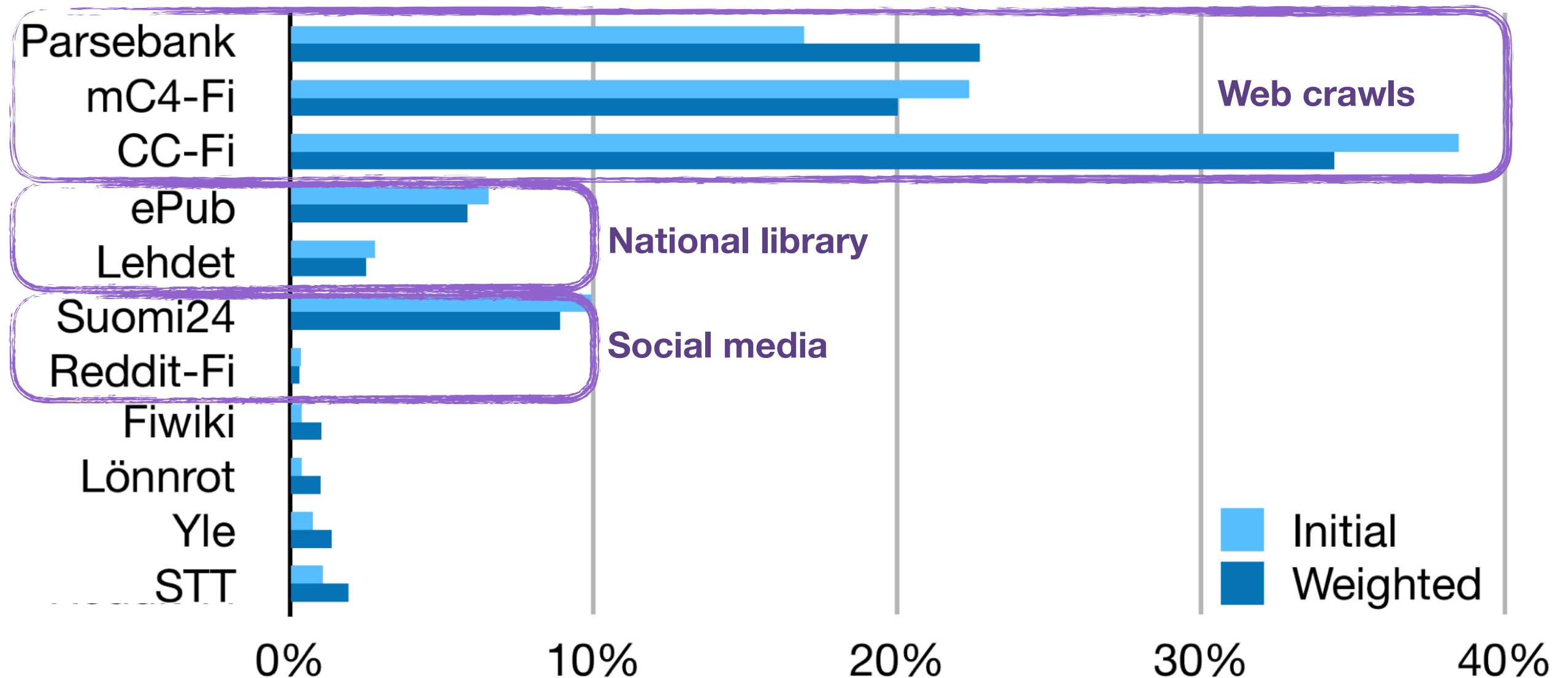
Data for Finnish GPT-3



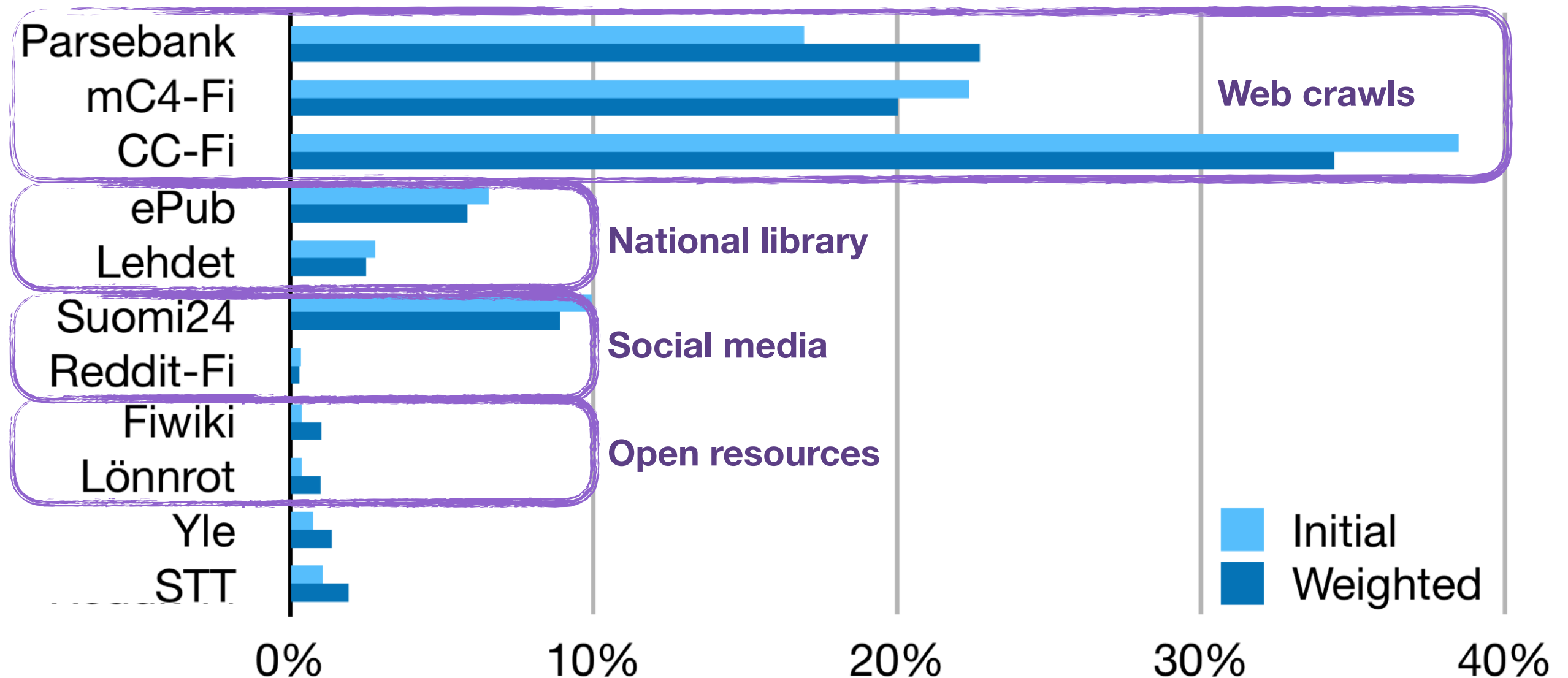
Data for Finnish GPT-3



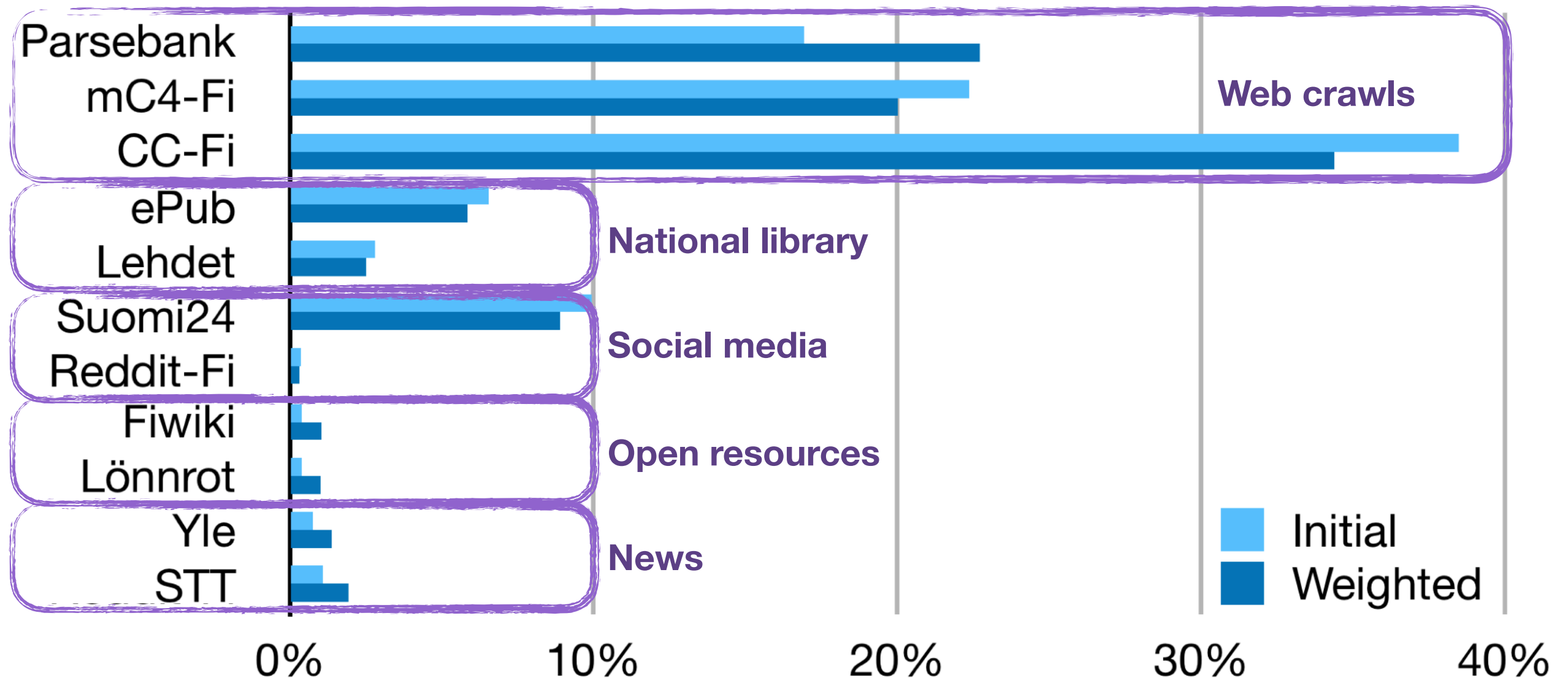
Data for Finnish GPT-3



Data for Finnish GPT-3



Data for Finnish GPT-3



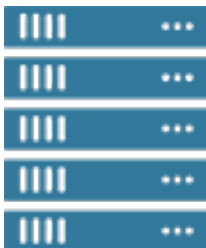
Data: risks and mitigations

- **Memorization / overfitting:** initial hash-based document-level deduplication + paragraph-level n-gram -based deduplication using One Instance Only (Onion)
- **Toxicity:** filtering using custom classifier trained on translated version of Jigsaw toxicity corpus (labels toxicity, threat, obscene, etc.)
- **Personal data:** high-recall rule- and regex-based system to remove potentially identifying numbers, addresses, etc.

Compute

LUMI



 **1 SYSTEM**
550+ Pflop/s
PEAK PERFORMANCE

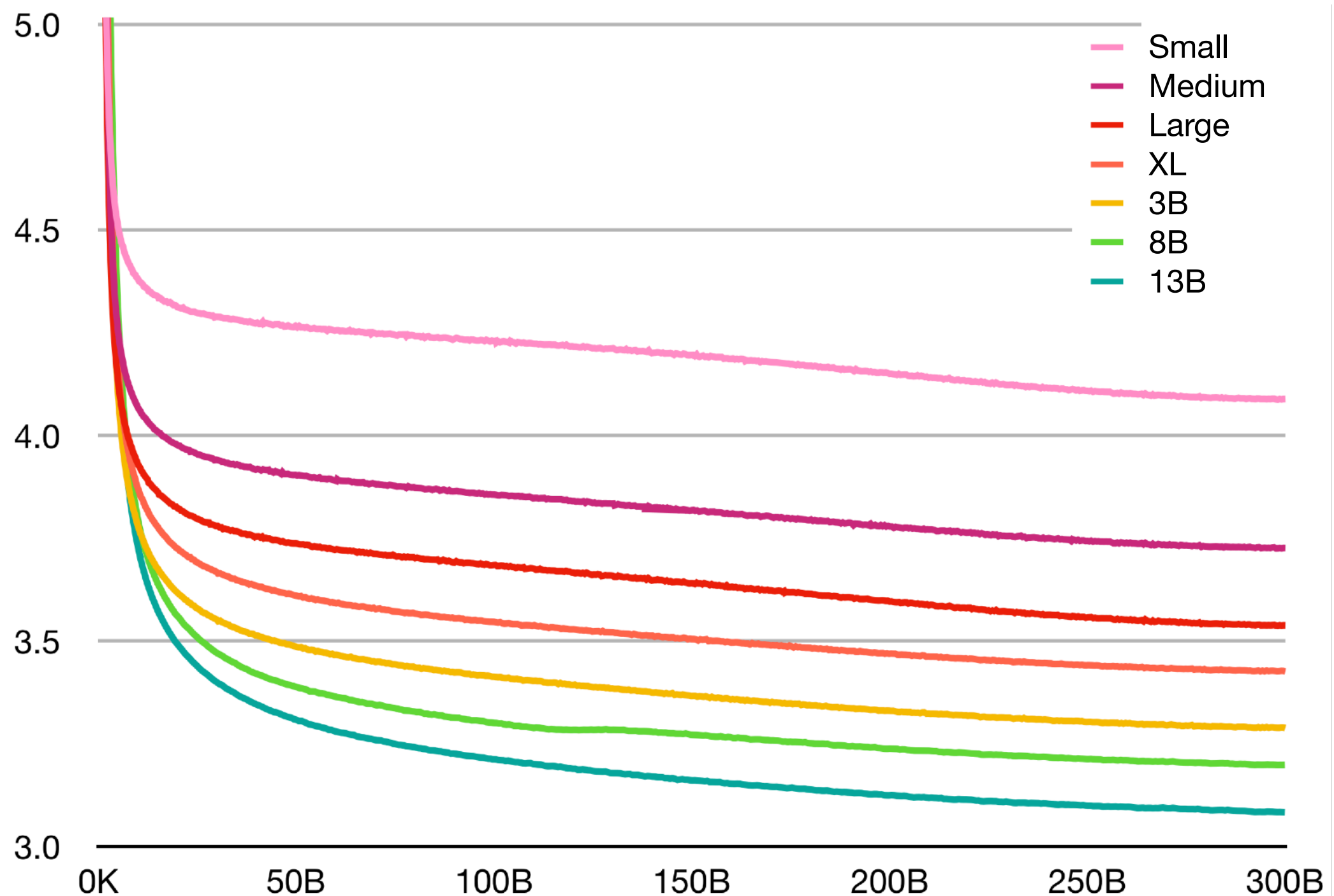
 **117 PB**
STORAGE 

LUMI: Fastest supercomputer in Europe (3rd in world)

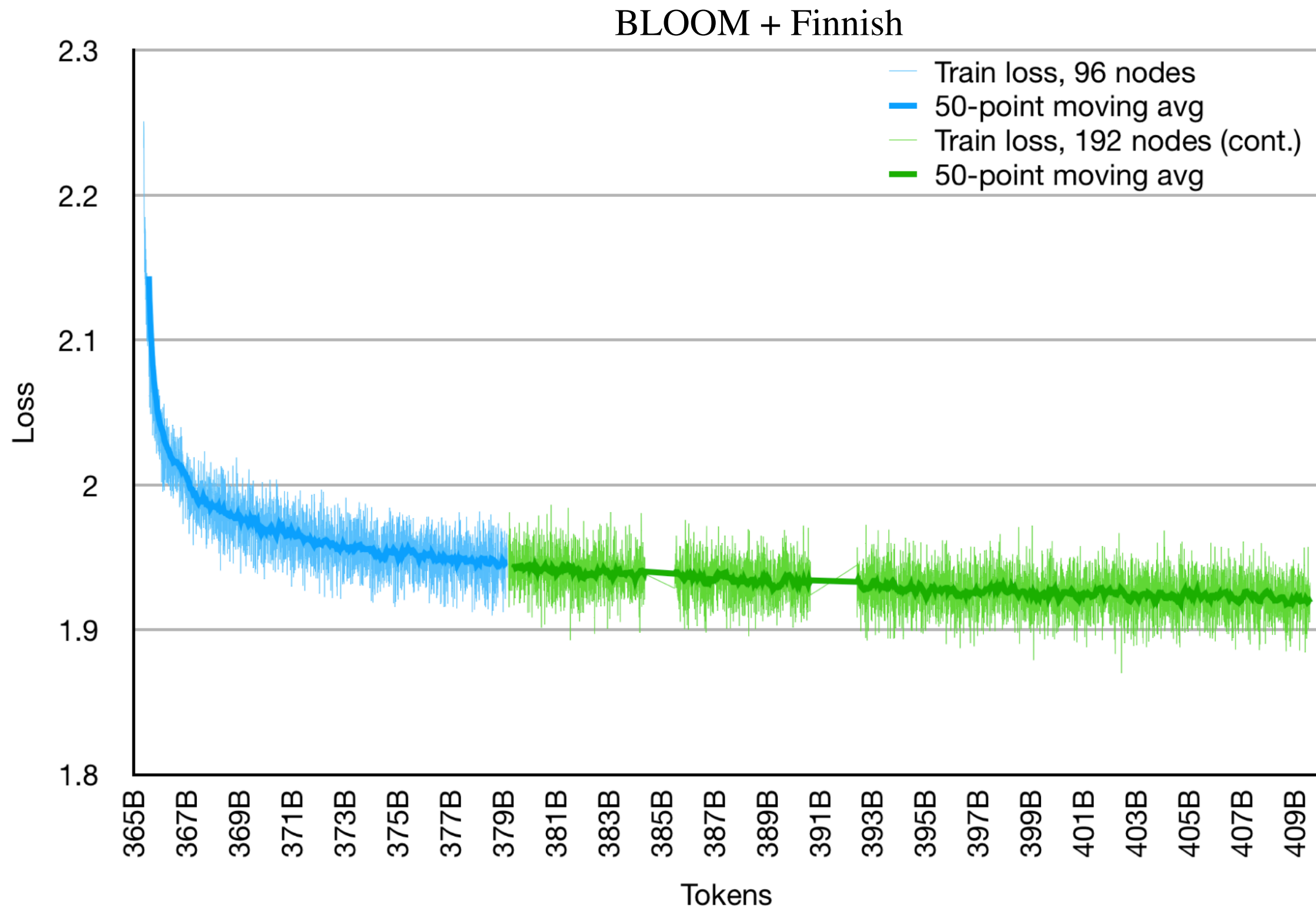
- 2560 GPU nodes w/AMD MI250X (20480 GCDs)
- Peak performance 550 Pflops (pre-exascale)
- Large allocations (1M+ GPUh) via EUHPC and national orgs, incl. Finland, Denmark, Estonia, Iceland, Norway, and Sweden

Finnish GPT-3: evaluation

Evaluation: validation loss



Evaluation: validation loss



Evaluation: zero/few-shot

Beyond the Imitation Game Benchmark (BIG-bench)

Q: How many elephants can fit in a freezer?

- four
- none
- two
- one
- six

Q: What is 60 times 41?

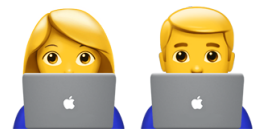
- 3675
- 7385
- 58
- 2460
- banana
- house

Claim: During photosynthesis, plants may make glucose and carbon dioxide.

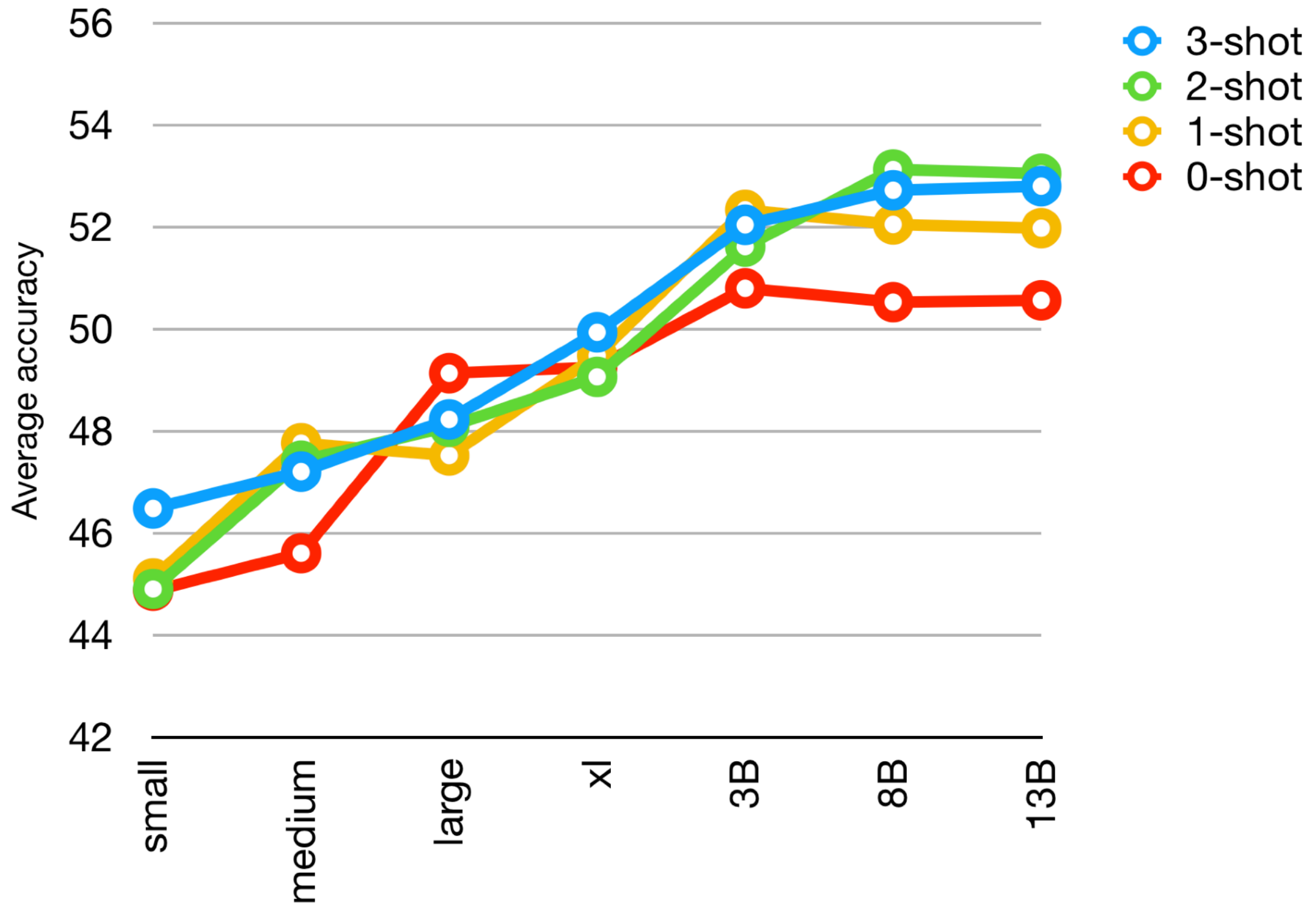
- True
- False



...

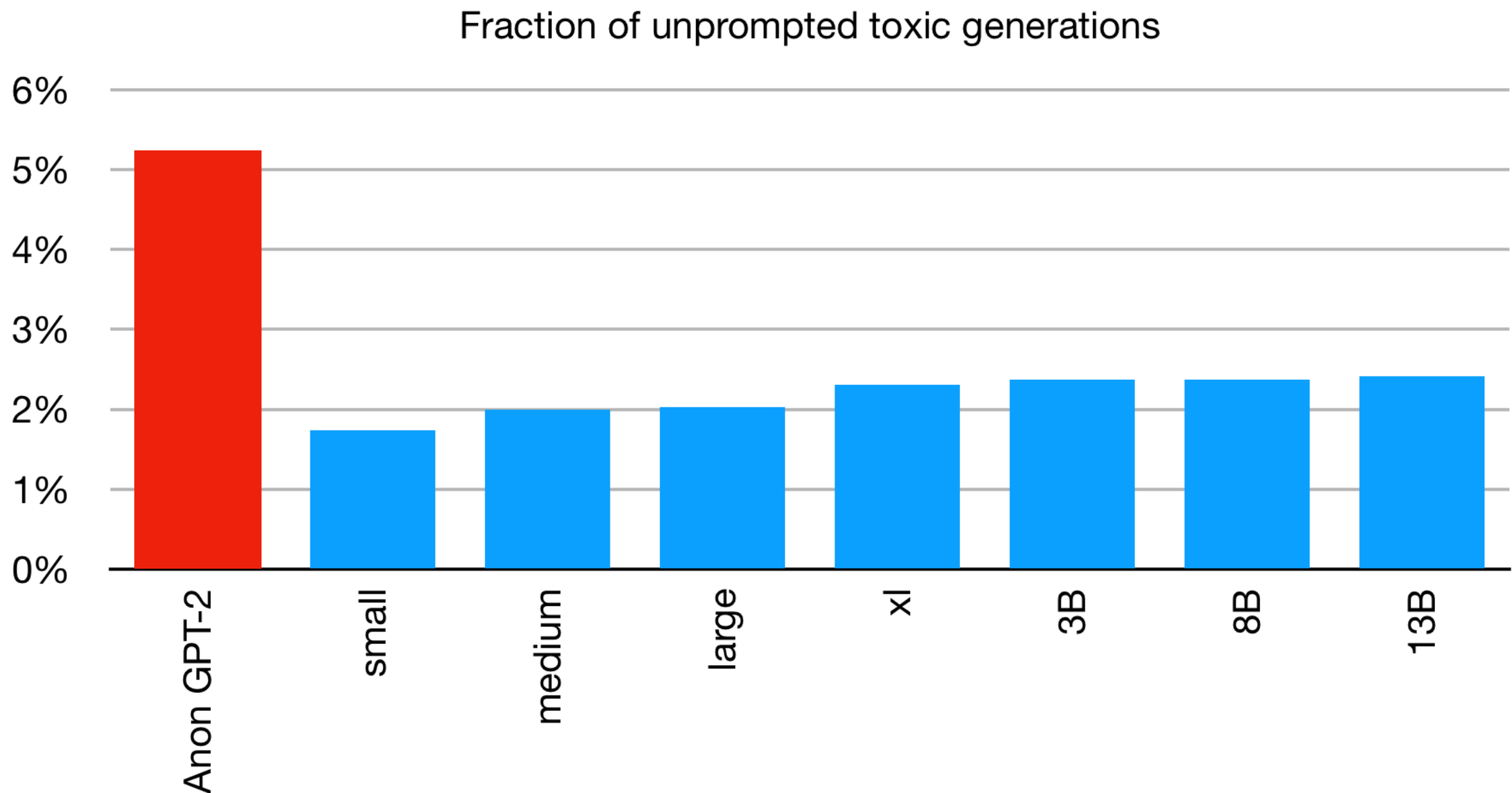


Evaluation: zero/few-shot



Evaluation: toxicity

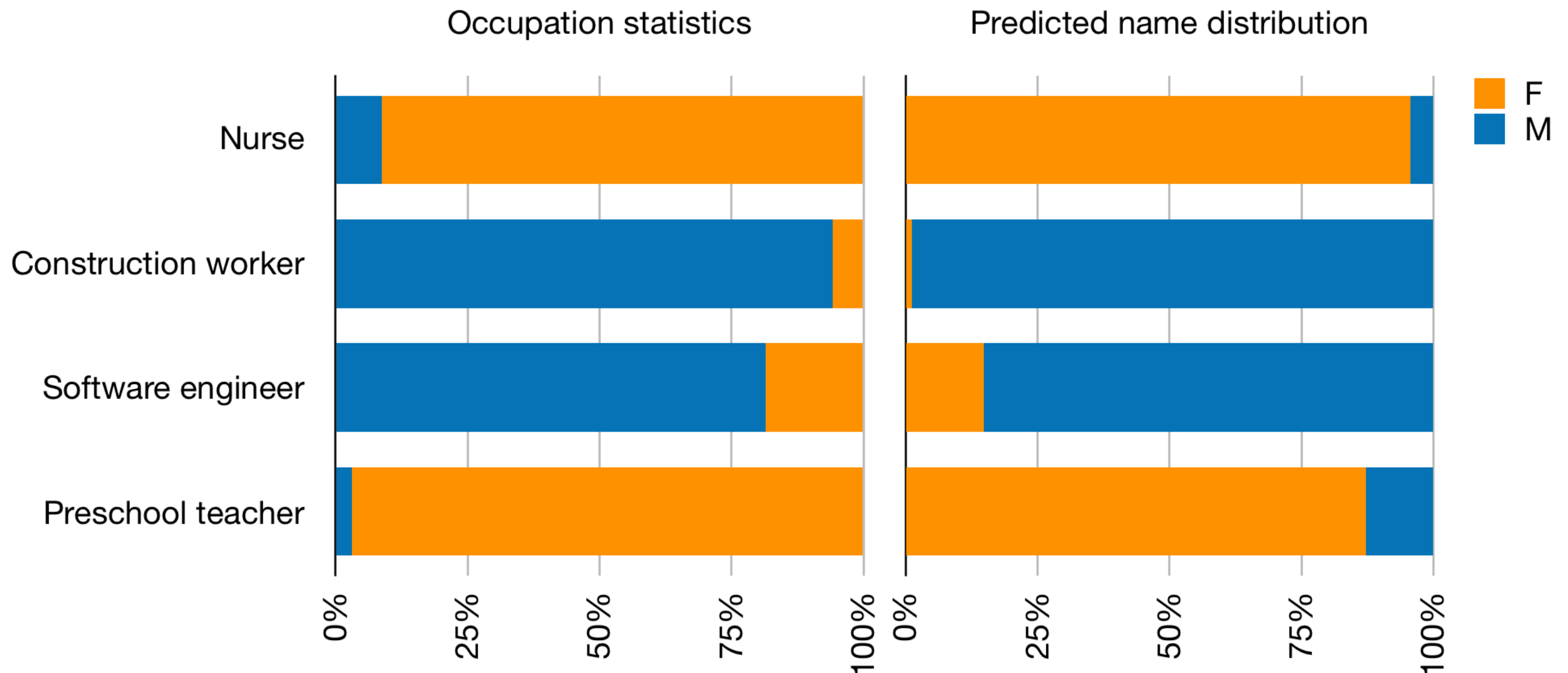
Models learn to generate toxic language based on pretraining data



Evaluation: bias

Models repeat and can exaggerate biases found in pretraining data

Gender-occupation association example:



What can you do with it?

Infinite fluent nonsense

Olipa kerran, kauan kauan sitten, pieni tyttö, joka rakasti merta. Tyttö tiesi, että se on vaarallinen paikka. Meressä on kaikenlaisia vaaroja. Siellä voi törmätä vaikka merirosvoihin. Tai sitten voi vaikka joutua merihirviön syömäksi. Joskus mereen voi jopa hukkua. Mutta kun tyttö rakasti niin kovasti merta, hän ei pelännyt. Hän rakasti uimista ja sukeltamista, sukeltaminen oli hänen lempipuuhaansa. Tytön paras ystävä oli pieni poika. Poika oli tytön kanssa aina kun hän oli merenrannalla. Eräänä päivänä tyttö ja poika päättivät lähteä veneilemään. He löysivät veneen, jossa oli moottori. Pojan mielestä vene oli hieno. Sitten he lähtivät ajamaan veneellä. Pian he huomasivat, ettei vene liikkunut, oli aivan hiljaista. Oli yö. Vene oli juuttunut kiinni kallioon. Kaikki alkoi pelottaa tyttöä. Yhtäkkiä alkoi sataa vettä. Se kasteli tytön. Silloin tyttö huomasi, kuinka kylmä oli. Hänen oli pakko mennä nukkumaan. Yöllä tyttö heräsi ja huomasi olevansa yksin veneessä. Pimeääkin oli, mutta onneksi veneestä löytyi lamppu. Nyt tyttö tiesi olevansa turvassa. Aamulla tyttö meni hakemaan aamupalaa. Kun hän tuli takaisin rannalle, aurinko paistoi. Illalla tyttö lähti taas uimaan. Tällä kertaa hän ui paljon kauemmas rannasta kuin yleensä. Jonkin ajan päästä hän huomasi veneen. Ja silloin hän säikähti. Venettä ei näkynyt missään. Vain moottori oli paikallaan. Samassa tyttö kuuli kauheaa huutoa läheiseltä saarelta. Onneksi saarelta tuli ihmisiä, jotka olivat etsimässä tyttöä ja hänen ystäväänsä. Veneen moottori löytyi ja tyttö ystävineen pääsi jatkamaan matkaansa. Loppu.

Zero/few-shot prediction

```
prompt = """Onko seuraava lause positiivinen vai negatiivinen?  
Lause: Täällä on erittäin hauskaa!  
Vastaus: positiivinen  
  
Onko seuraava lause positiivinen vai negatiivinen?  
Lause: elokuva oli surkea!  
Vastaus: """"  
  
print(generate(prompt))
```

Onko seuraava lause positiivinen vai negatiivinen?
Lause: Täällä on erittäin hauskaa!
Vastaus: positiivinen

Onko seuraava lause positiivinen vai negatiivinen?
Lause: elokuva oli surkea!
Vastaus: negatiivinen

Lauseen positiivisuus ilmenee siinä, että se sisältää vain positiivisia asioita.

Zero/few-shot prediction

```
prompt = """Onko seuraava lause positiivinen vai negatiivinen?  
Lause: Täällä on erittäin hauskaa!  
Vastaus: positiivinen  
  
Onko seuraava lause positiivinen vai negatiivinen?  
Lause: elokuva oli surkea!  
Vastaus: """"  
  
print(generate(prompt))
```

Onko seuraava lause positiivinen vai negatiivinen?
Lause: Täällä on erittäin hauskaa!
Vastaus: positiivinen

Example
(one “shot”)

Onko seuraava lause positiivinen vai negatiivinen?
Lause: elokuva oli surkea!
Vastaus: negatiivinen

Lauseen positiivisuus ilmenee siinä, että se sisältää vain positiivisia asioita.

Zero/few-shot prediction

```
prompt = """Onko seuraava lause positiivinen vai negatiivinen?  
Lause: Täällä on erittäin hauskaa!  
Vastaus: positiivinen
```

```
Onko seuraava lause positiivinen vai negatiivinen?  
Lause: elokuva oli surkea!  
Vastaus: """
```

```
print(generate(prompt))
```

```
Onko seuraava lause positiivinen vai negatiivinen?  
Lause: Täällä on erittäin hauskaa!  
Vastaus: positiivinen
```

Example
(one "shot")

```
Onko seuraava lause positiivinen vai negatiivinen?  
Lause: elokuva oli surkea!  
Vastaus: negatiivinen
```

Question

Lauseen positiivisuus ilmenee siinä, että se sisältää vain positiivisia asioita.

Zero/few-shot prediction

```
prompt = """Onko seuraava lause positiivinen vai negatiivinen?  
Lause: Täällä on erittäin hauskaa!  
Vastaus: positiivinen
```

```
Onko seuraava lause positiivinen vai negatiivinen?  
Lause: elokuva oli surkea!  
Vastaus: """
```

```
print(generate(prompt))
```

```
Onko seuraava lause positiivinen vai negatiivinen?  
Lause: Täällä on erittäin hauskaa!  
Vastaus: positiivinen
```

Example
(one "shot")

```
Onko seuraava lause positiivinen vai negatiivinen?  
Lause: elokuva oli surkea!  
Vastaus: negatiivinen
```

Question

Answer

Lauseen positiivisuus ilmenee siinä, että se sisältää vain positiivisia asioita.

Zero/few-shot prediction

```
prompt = """Onko seuraava lause positiivinen vai negatiivinen?  
Lause: Täällä on erittäin hauskaa!  
Vastaus: positiivinen
```

```
Onko seuraava lause positiivinen vai negatiivinen?  
Lause: elokuva oli surkea!  
Vastaus: """
```

```
print(generate(prompt))
```

Onko seuraava lause positiivinen vai negatiivinen?
Lause: Täällä on erittäin hauskaa!
Vastaus: positiivinen

Example
(one “shot”)

Onko seuraava lause positiivinen vai negatiivinen?
Lause: elokuva oli surkea!
Vastaus: negatiivinen

Question

Answer

Lauseen positiivisuus ilmenee siinä, että se sisältää vain positiivisia asioita.

Bonus nonsense



Fine-tuning!



```
from transformers import AutoTokenizer, AutoModelForSequenceClassification
from datasets import load_dataset
```

```
dataset = load_dataset(DATASET_NAME)
label_names = dataset['train'].features['label'].names
```

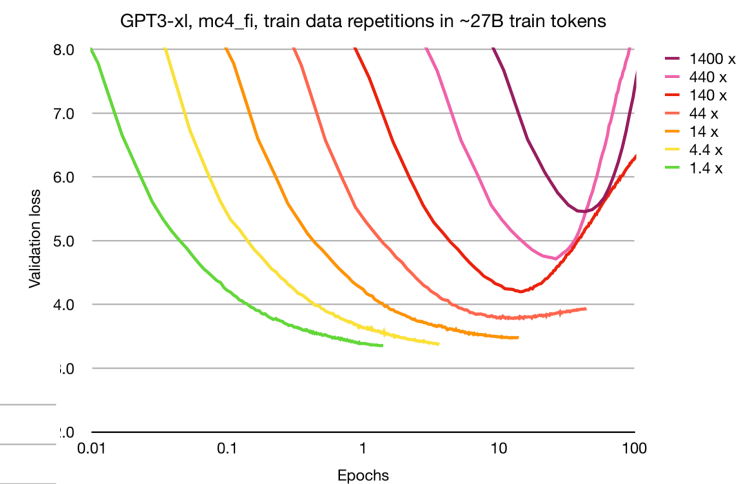
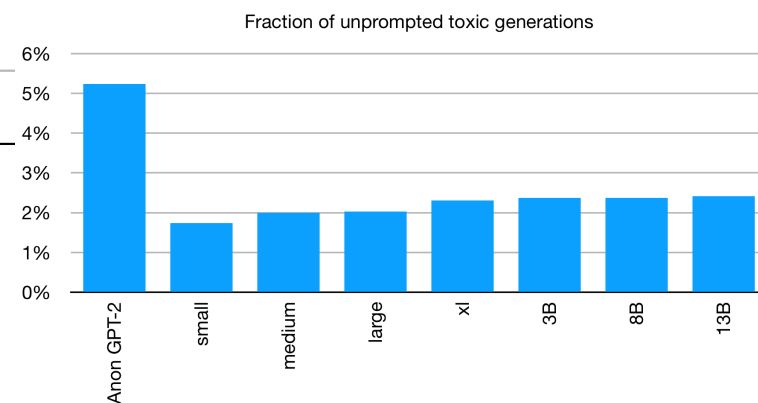
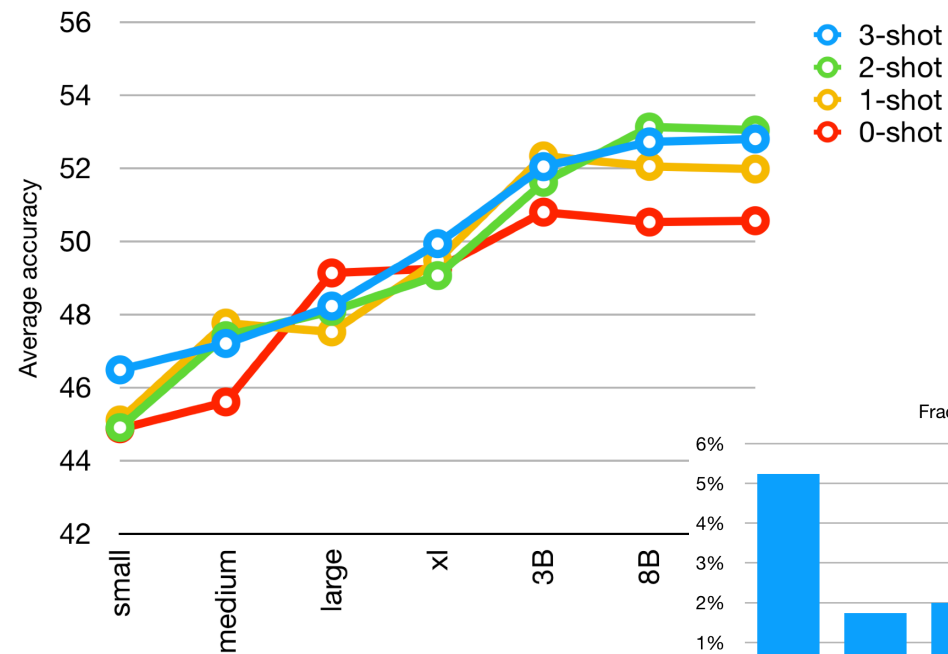
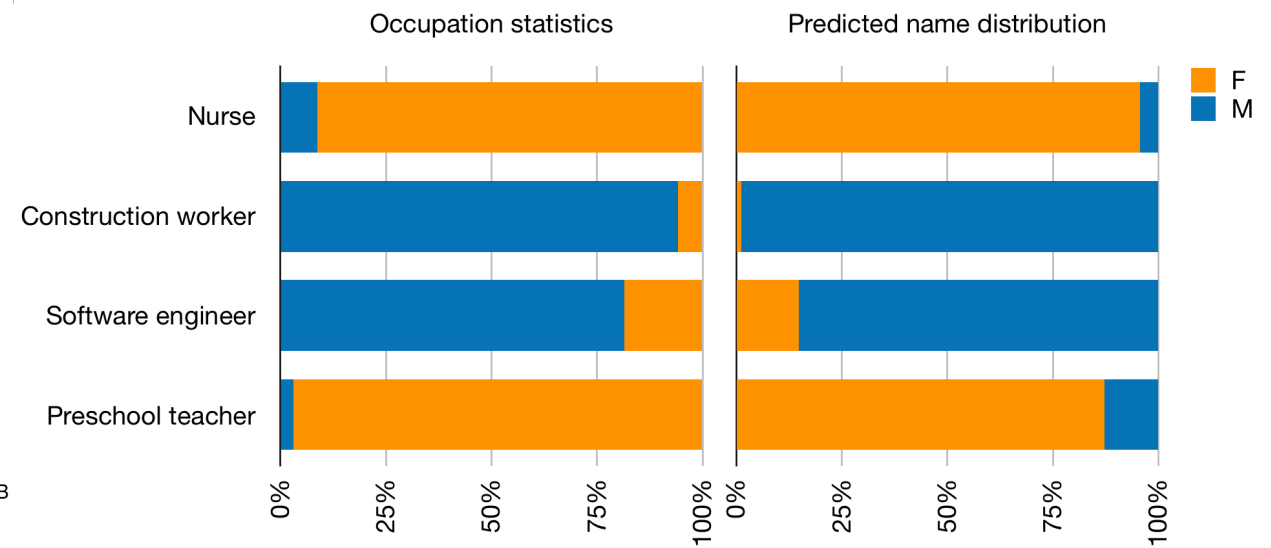
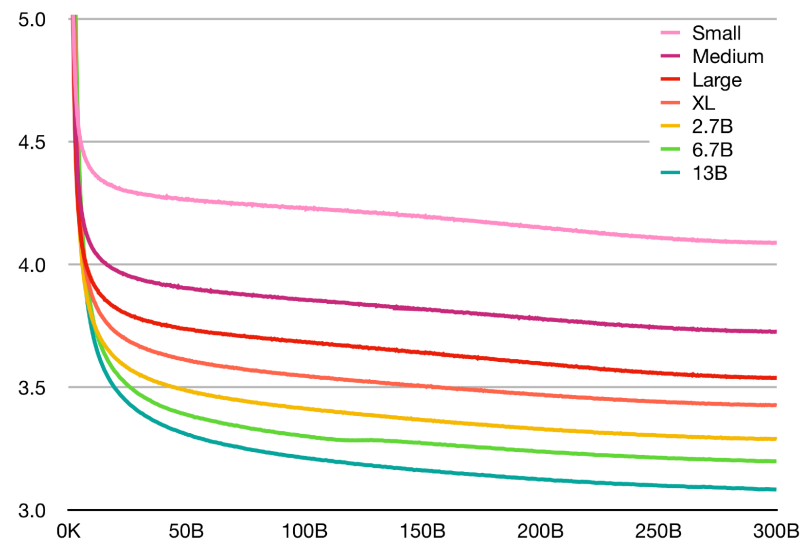
```
model = AutoModelForSequenceClassification.from_pretrained(
    MODEL_NAME,
    num_labels=len(label_names)
)
```

...

```
trainer = Trainer(
    model=model,
    args=trainer_args,
    train_dataset=dataset['train'],
)

trainer.train()
```

Research



What's next?

- **Generative models for all European languages:** High Performance Language Technologies EU project
- **Large encoder-only and encoder-decoder models for Finnish**
- **Larger generative multilingual models** with support for Finnish
- InstructGPT / ChatGPT - likes for Finnish?

Thanks & Acknowledgments

The National Library of Finland

LUMI User Support Team

CSC User Support

Hugging Face

Samuel Antao (AMD)

Compute: LUMI (Finnish allocation)

Funding: EU project High Performance
Language Technologies (Grant No. 101070350)

