# Training LLMs on LUMI

EuroCC Greece, 24.4.2024

Mats Sjöberg, CSC

LUMI - a European supercomputer

- LUMI consortium countries: Finland, Belgium, the Czech Republic, Denmark, Estonia, Iceland, the Netherlands, Norway, Poland, Sweden, and Switzerland
- Hosted by CSC in Kajaani
- LUMI User Support main support channel
  - Science support from local organizations like CSC
- Resource allocation:
  - 50% EuroHPC JU
  - each consortium country according to contribution
  - up to 20% industry and SMEs
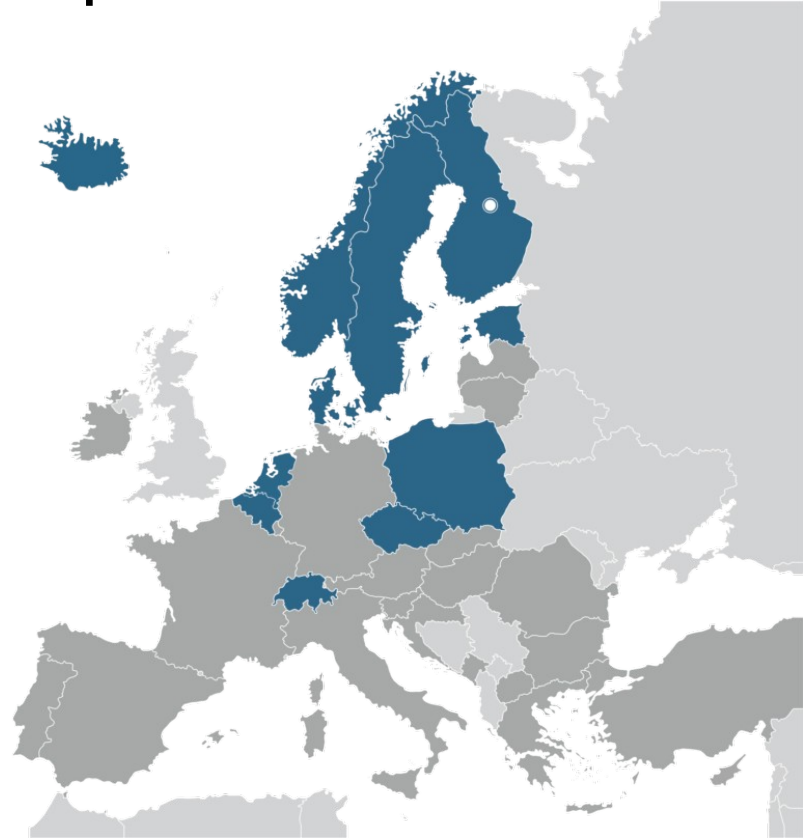
LUMI
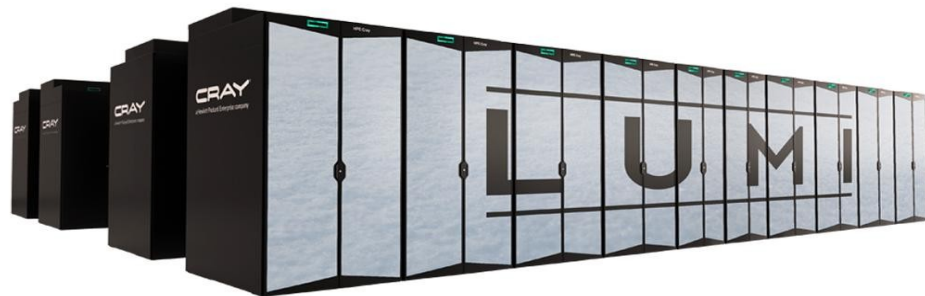
HPE Cray EX Supercomputer

LUMI-C (CPU) nodes consisting of 2048 servers with:

- 2x 64-core AMD EPYC "Milan" CPU
- Between 256 GB and 1024 GB RAM
- HPE Slingshot-11 interconnect

LUMI-G (GPU) nodes consisting of 2978 servers with:

- 64-core AMD EPYC "Trento" CPU, 512 GB RAM
- 4x **AMD MI250X GPUs,** each with 128 GB HBM2e memory
  - MI250X consists of two compute dies → 8 GCDs per node
  - each MI250X GCD has 64 GB VRAM
- 4x 200 Gbit/s HPE Slingshot-11 interconnects
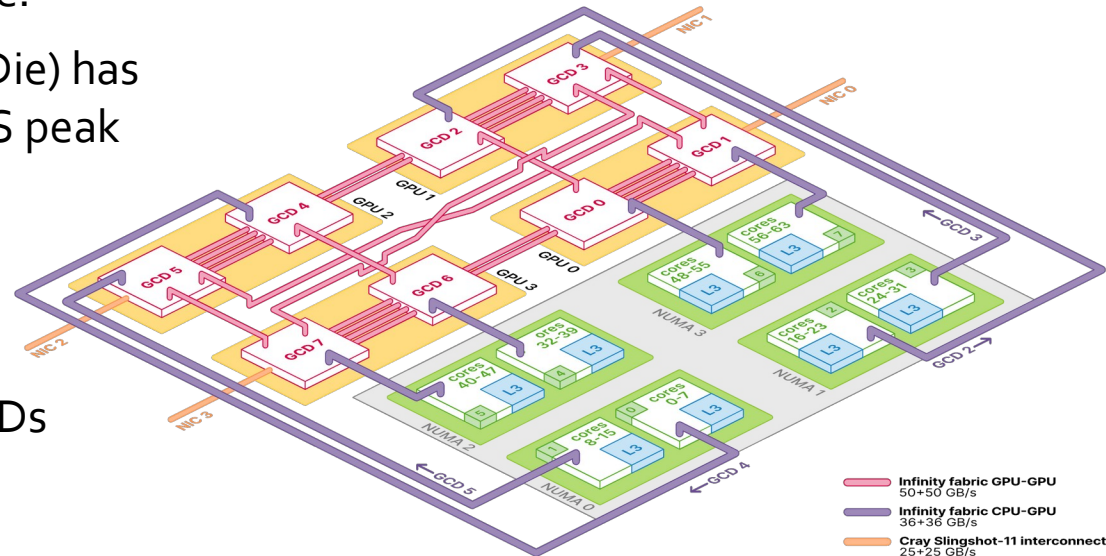
AMD MI250x GPU

4x MI250X per node

- Dual-chip module
  → in practice 8 "GPUs" per node.

- Each GCD (Graphics Compute Die) has 64GB of VRAM and 192 TFLOPS peak BF16

But, we have a lot of them!

- Total of 2978 nodes → ~24k GCDs

Deep learning software stack

- AMD GPUs: ~~CUDA~~ → ROCm

- CUDA kernels can be converted with HIPIFY tool

- PyTorch pretty well supported out-of-the-box

- AMD has ported and optimized things like Flash Attention, bitsandbytes, vLLM, …

- We occasionally stumble on some edge cases (recently some CNN algorithm selection issue)

  - Usually fixable with some environment variable settings

AMD ✅
@AMD

.@PyTorch founder @soumithchintala on AMD Day 0 support: Developers can train and deploy #AI models compatible "out of the box" with @AMDInstinct accelerators. "Simply put, when a new model is released, it's up and running on AMD platforms immediately for developers."

AMD △ × ◐ PyTorch

Uutinen

A New Foundation for AI Is Being Built in Finland – Offering an Alternative to American Giants

Antti Leikas 6.12.2023 09:45 | päivitetty 8.12.2023 15:07 TIVI IN ENGLISH ARTIFICIAL INTELLIGENCE

The development of the Poro language model utilizes Lumi, the fastest supercomputer in Europe, located in Kajaani.

Artificial Intelligence. According to Peter Sarlin (center), the development of the Poro language model aims at democratizing AI technology and ensuring equal treatment of European languages. TIINA SOMERPURO

AI2 OLMo
Open Language Model

News

2.2.2024

A truly open large language model released, developed with LUMI

The Allen Institute for AI (AI2) has released OLMo 7B, a truly open, state-of-the-art large…

yle Etusivu Vaalikone Venäjän hyökkäys UMK24

Tiede

FinGPT3 on suurin puhtaasti suomenkielinen kielimalli, eikä suurempaa ole hetkeen tulossa

Kajaanissa täydellä teholla pyörivä Lumi-supertietokone on merkittävässä roolissa uusien kielimallien kehityksessä. Alkuvuonna Lumi sai laskettua valmiiksi suomen kielen suurimman kielimallin.

Kajaanissa sijaitseva Lumi on tällä hetkellä Euroopan tehokkain supertietokone. Koneen laskentakapasiteetille on kysyntää etenkin tekoälybuumin taustalla olevien kielimallien ansiosta. Kuva: Rami Moilanen / Yle

	GPU Type	GPU Power Consumption (MWh)	Power Usage Effectiveness	Carbon Intensity (kg CO_2e/KWh)	Carbon Emissions (tCO_2eq)
Gopher-280B	TPU v3	1,066	1.08	0.330	380
BLOOM-176B	A100-80GB	433	1.2	0.057	30
OPT-175B	A100-80GB	324	1.1	0.231	82
T5-11B	TPU v3	77	1.12	0.545	47
LLaMA-7B	A100-80GB	33	1.1	0.385	14
LLaMA2-7B	A100-80GB	74	1.1	0.385	31
OLMo-7B	MI250X	135	1.1	0.000*	0*
OLMo-7B	A100-40GB	104	1.1	0.610	70

Groeneveld et al, OLMo: Accelerating the Science of Language Models, https://arxiv.org/abs/2402.00838

Poro – Academic + Industry + HPC centre collaboration

- 34B parameter LLM trained on 1T tokens of Finnish, English and code

- Collaboration between:
  - TurkuNLP, research group at University of Turku, Finland
  - Silo AI, private Finnish AI company, "150+ PhDs employed"
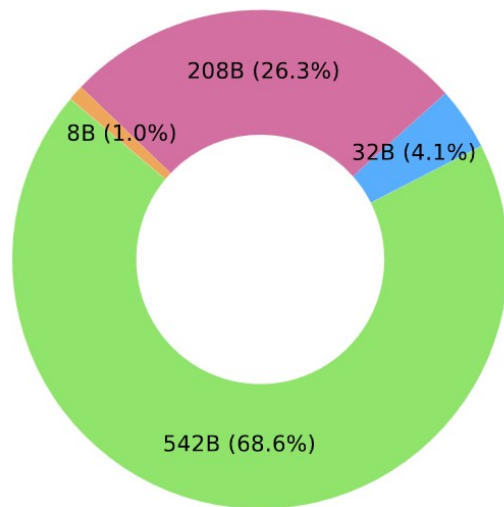 - CSC – IT Centre for Science

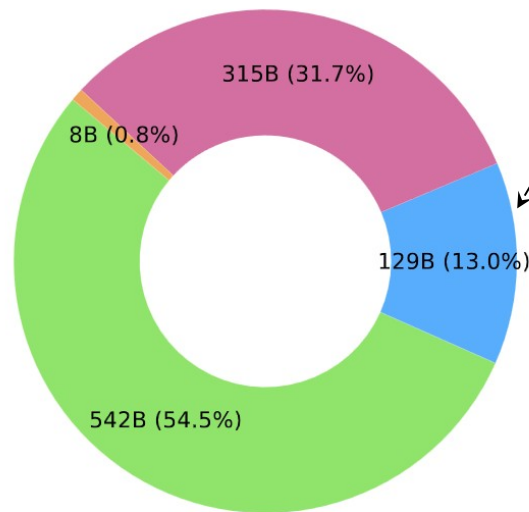https://huggingface.co/LumiOpen/Poro-34B

Poro 34B and the Blessing of Multilinguality

Risto Luukkonen[1,2]   Jonathan Burdge[2]   Elaine Zosa[2]
Aarne Talman[2,3]   Ville Komulainen[1]   Väinö Hatanpää[4]
Peter Sarlin[2]   Sampo Pyysalo[1]

/ Poro training data

cf. 300B tokens of Finnish (8 epochs) for FinGPT

(a) Original

- 208B (26.3%)
- 8B (1.0%)
- 32B (4.1%)
- 542B (68.6%)

(b) Sampled

- 315B (31.7%)
- 8B (0.8%)
- 129B (13.0%)
- 542B (54.5%)

Legend:
- English
- Finnish
- Code
- Eng-Fin

TURKUNLP.ORG

UNIVERSITY OF TURKU

/ **Poro evaluation**

Substantial advance over previous models in **Finnish**: **51%** (FinGPT) / **59%** (BLUUMI) → **66%** (Poro)

Competitive in its class of open models for **English and**
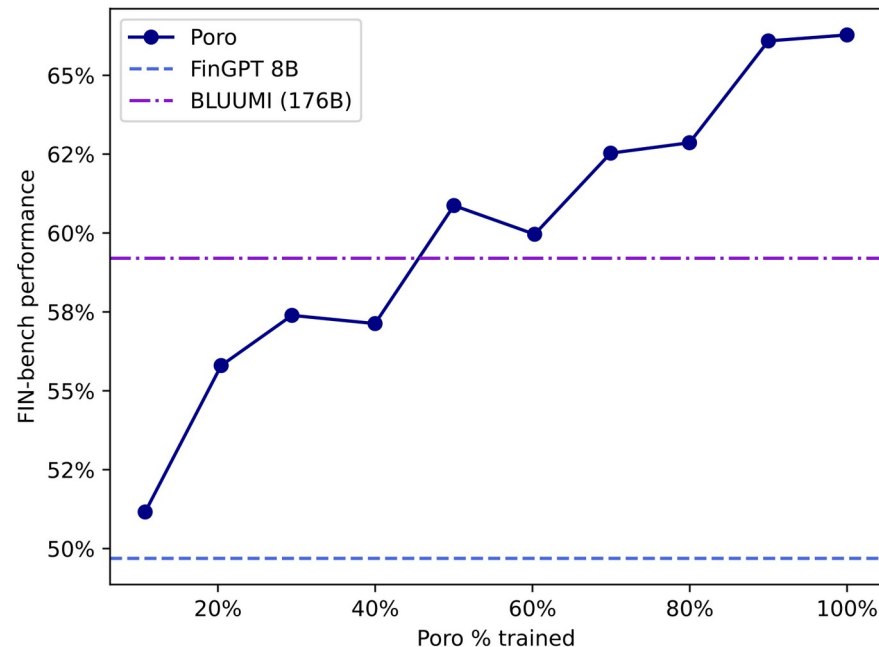
	Poro 34B	Llama 33B	MPT 30b	Falcon 40B	FinGPT 8B	FinGPT 13B	Starcoder
Finnish	**66.28**	53.36	53.22	42.58	49.69	48.92	45.55
English	50.57	**59.96**	52.62	49.87	31.47	32.85	35.44
Code	41.80	37.67	39.18	38.57	-	-	**49.06**

TURKUNLP .ORG

UNIVERSITY OF TURKU

/ **Poro evaluation**

Remarkably good at **English-Finnish translation**!

(Opus Eng-Fin translation examples included in pretraining data)

Caveat: trained and tested mostly on single-sentence translation

Model	Flores-101		Tatoeba	
	En-Fi	Fi-En	En-Fi	Fi-En
ChatGPT	33.4	35.9	-	-
GPT4	35.3	**40.2**	-	-
Google	37.3	39.0	-	-
M2M-12B	33.4	33.8	36.7	41.3
NLLB-1.3B	30.0	35.4	40.2	55.7
OPUS-MT	37.2	35.6	46.7	58.4
Poro 34B	**37.6**	39.8	**47.3**	**60.5**

TURKUNLP
.ORG

UNIVERSITY OF TURKU

Why training LLMs is difficult

- Model doesn't fit into single GPU memory

  - e.g. LUMI GCD has 64 GB VRAM

  - Example: 140B parameter model ~
    280 GB VRAM assuming 16 bit
    → at least 5 LUMI GCDs needed

- Model parallelism

  - Tensor parallel = layers split across GPUs

  - Pipeline parallel = layers distributed
    across GPUs

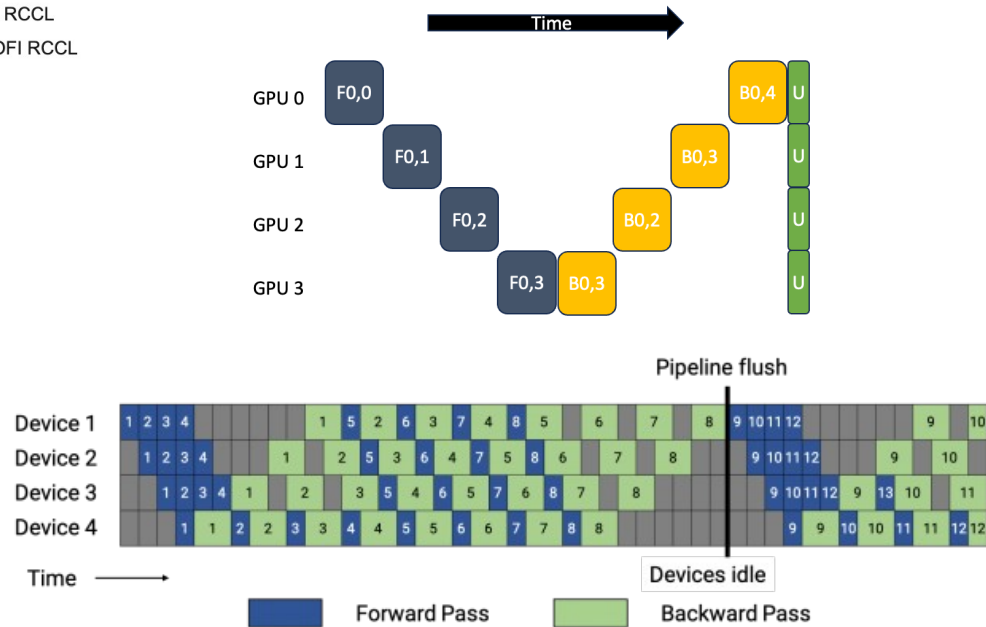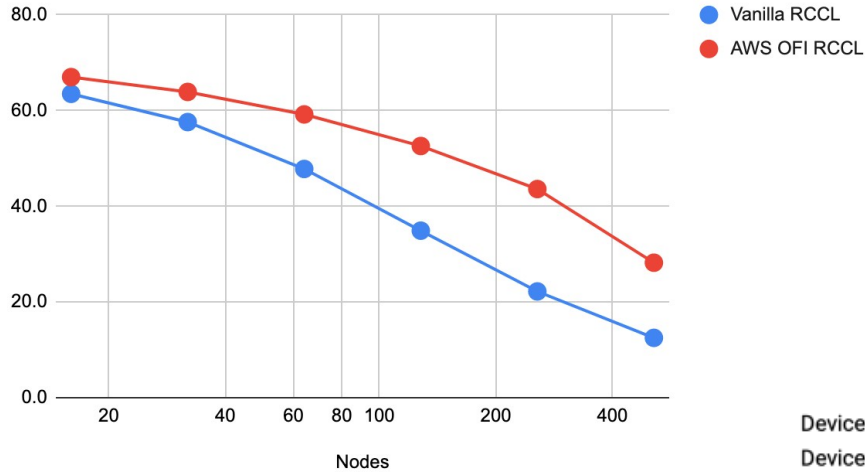 - Gradients and optimizer states need to
    be sharded
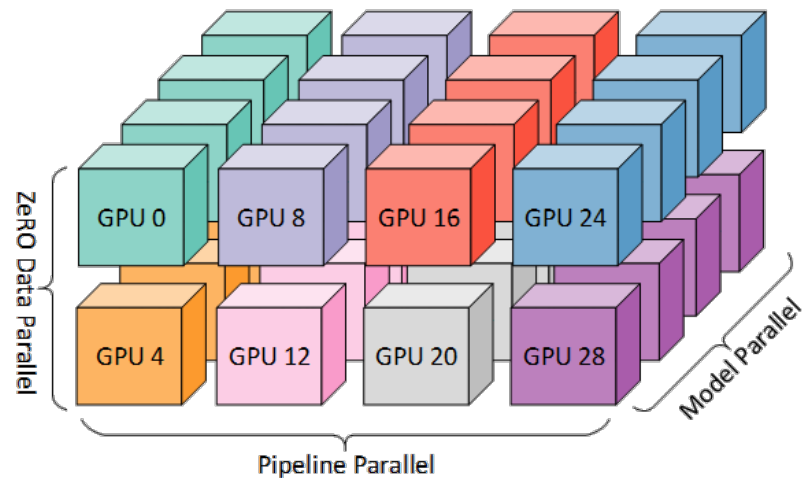
Things have become easier …

- Huggingface Accelerate library

- PyTorch Fully sharded data parallel (FSDP)

- Except when you need to go really large-scale
  - Poro model used highly tweaked Megatron-LM

Scaling challenges on Megatron-DeepSpeed + LUMI
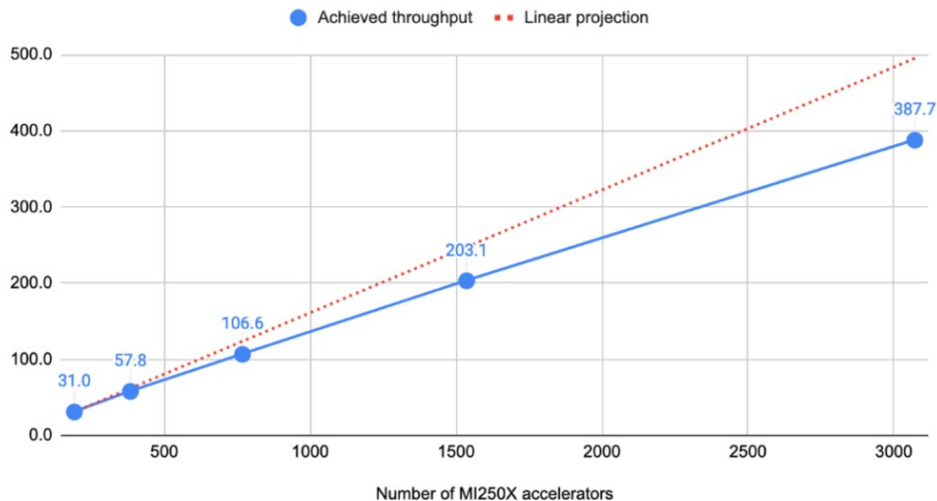
mean TFLOPs/GCD vs. nodes

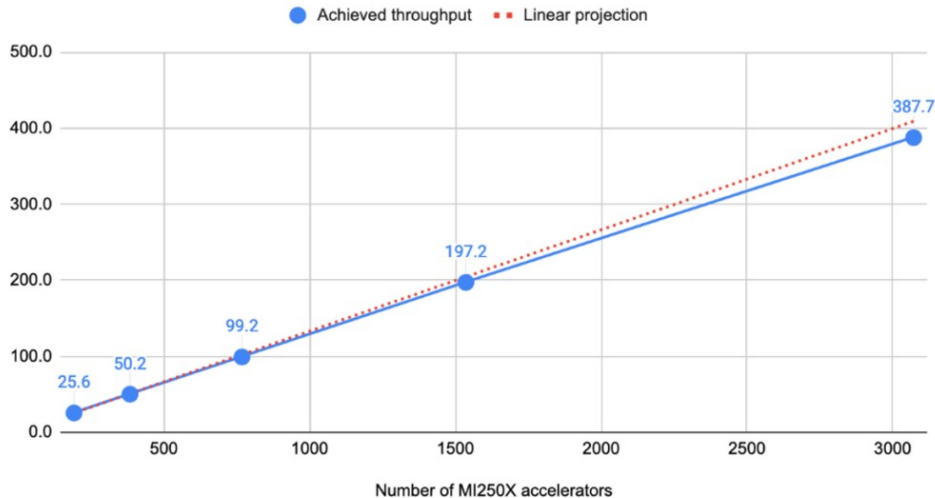https://developer.nvidia.com/blog/scaling-language-model-training-to-a-trillion-parameters-using-megatron/

Scaling LLM training to 1000s of GPUs

- Example: Model size 140B with Megatron-DeepSpeed

Strong scaling, total petaFLOPS with BF16

- Achieved throughput · · Linear projection

500.0

400.0   387.7

300.0

203.1
200.0

106.6
100.0
57.8
31.0
0.0
         500   1000   1500   2000   2500   3000

Number of MI250X accelerators

Weak scaling, total petaFLOPS with BF16

- Achieved throughput · · Linear projection

500.0

400.0   387.7

300.0

197.2
200.0

99.2
100.0
50.2
25.6
0.0
         500   1000   1500   2000   2500   3000

Number of MI250X accelerators

Conclusions

- LUMI supercomputer highly suitable for large scale pre-training of LLMs

- Poro model successful case of collaboration between Academia, Industry and Supercomputing centre partners

- Many other LLMs trained on LUMI

  - OLMo, English open model: https://arxiv.org/abs/2402.00838

  - Viking 7B/13B/33B, also Silo AI + TurkuNLP, continuation of Poro including more Nordic languages
    https://www.silo.ai/blog/viking-7b-13b-33b-sailing-the-nordic-seas-of-multilinguality

Mats Sjöberg
Senior Machine Learning Specialist
mats.sjoberg@csc.fi

facebook.com/CSCfi

twitter.com/CSCfi

linkedin.com/company/csc---it-center-for-science

github.com/CSCfi