User Best Practice and Results on LUMI

Sampo Pyysalo
TurkuNLP, University of Turku, Finland

AI-Friendly EuroHPC Systems
September 5th 2024

TURKUNLP .ORG

TURUN YLIOPISTO

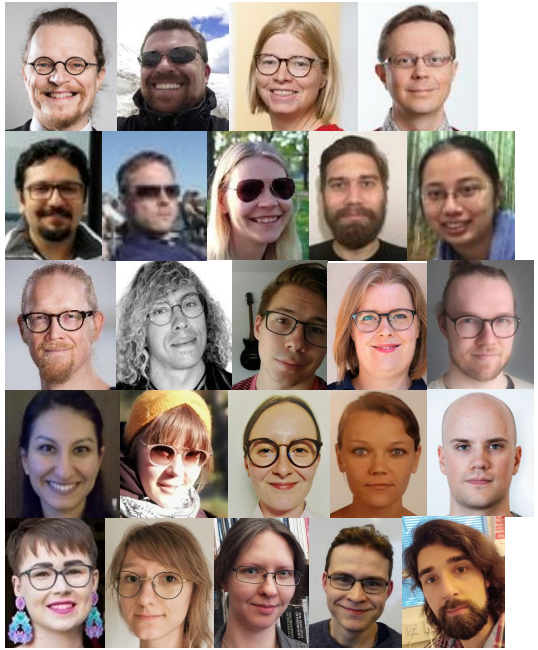
Self / group introduction

Sampo Pyysalo

- Research fellow, University of Turku
- CS / ML background
- Research on ML applied to NLP

TurkuNLP

- Research group in NLP
- Founded 2001, now ~30 members
- Focus on NLP for Finnish and multilinguality
- Increasing recent emphasis on DL / LLMs



High Performance
Language Technologies

Project: HPLT

Horizon Europe project (2022-2025),
8 partners: academic, industry, and HPC

Text resources and models (LLM, MT) for **~80 languages with EU focus**

Work in TurkuNLP focusing on **LLM pretraining from scratch**

https://hplt-project.org/

CHARLES UNIVERSITY
UNIVERSITY OF OSLO
UNIVERSITY OF EDINBURGH
UNIVERSITY OF TURKU
UNIVERSITY OF HELSINKI
PROMPSIT L.E.
CESNET
SIGMA2

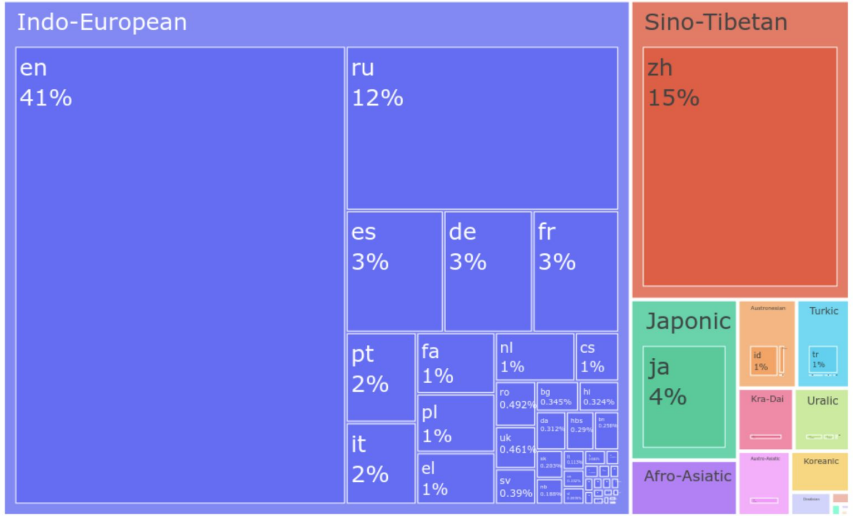


Figure from de Gibert et al. (2024)

TURKUNLP .ORG

UNIVERSITY OF TURKU

Collaboration

FinGPT models created in collaboration with **National Library of Finland** and **Hugging Face**, with support from **AMD**

Poro, **Viking**, and ongoing models created in collaboration with **Silo AI** (now AMD)

All work supported by **CSC** and **LUMI support**

Creating LLMs on LUMI

Working on LUMI since its earliest availability (late 2022 pilot project DeepFin)

Awarded allocations of over 15MGPUh, creating **fully open LLMs**

- **FinGPT** (Feb 2023): **13B parameters**, **300B tokens** of Finnish
- **BLUUMI** (Feb 2023): **176B parameters, 40B tokens** (cont. pretrain)
- **Poro** (Feb 2024): **34B parameters**, **1T tokens**, Finnish, English + code
- **Viking** (Sep 2024): **7-33B parameters, 2T tokens**, Nordic + English + code
- Upcoming models (? 2025): 7-70B parameters, 3T tokens, all EU langs + code

(plans for upcoming models currently in flux)

Luukkonen *et al*. FinGPT: Large Generative Models for a Small Language

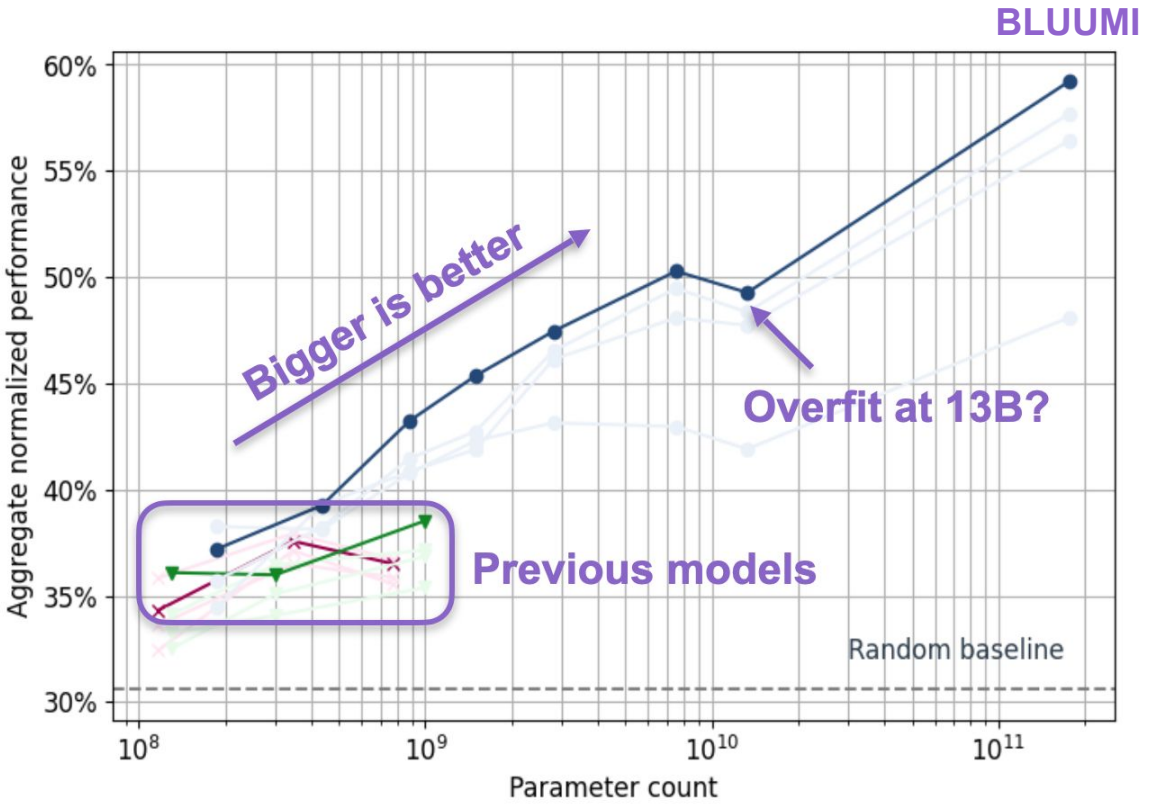
Results: FinGPT / BLUUMI

Substantial advance over previous models:

~ 40% → ~ 60%

Indications of overfitting for largest monolingual model (13B)

BLUUMI: notable Finnish capabilities, no drop on English (but unwieldy)

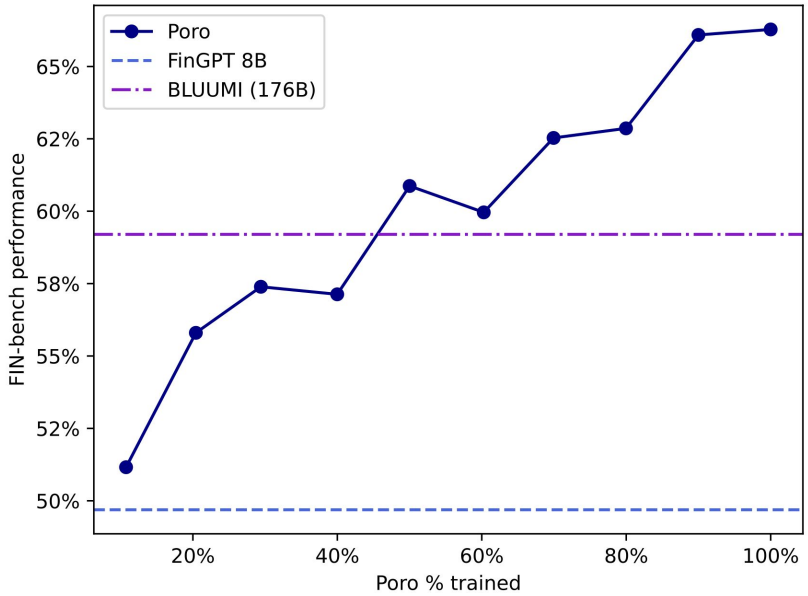


TURKUNLP .ORG

Results: Poro

Further advance over previous models in **Finnish**: **~50%** (FinGPT) / **~60%** (BLUUMI) → **66%** (Poro)

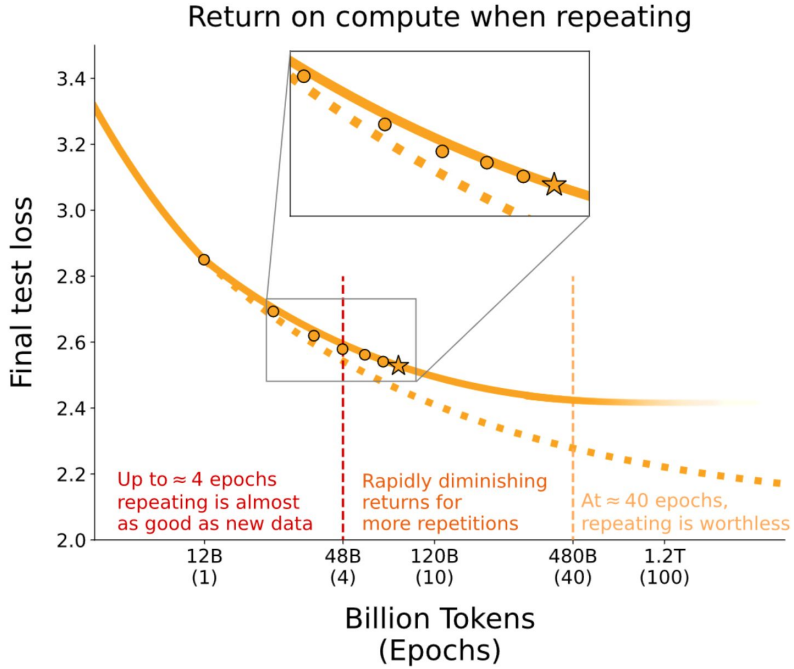
Broadly competitive with open models with similar param/token counts for **English and code**



	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**

Muennighoff *et al*. Scaling Data-Constrained Language Models

Results: data-constrained scaling laws

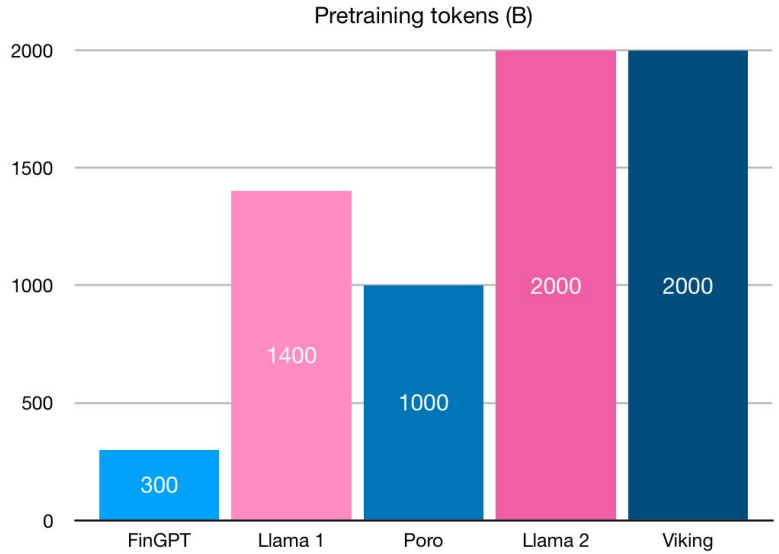


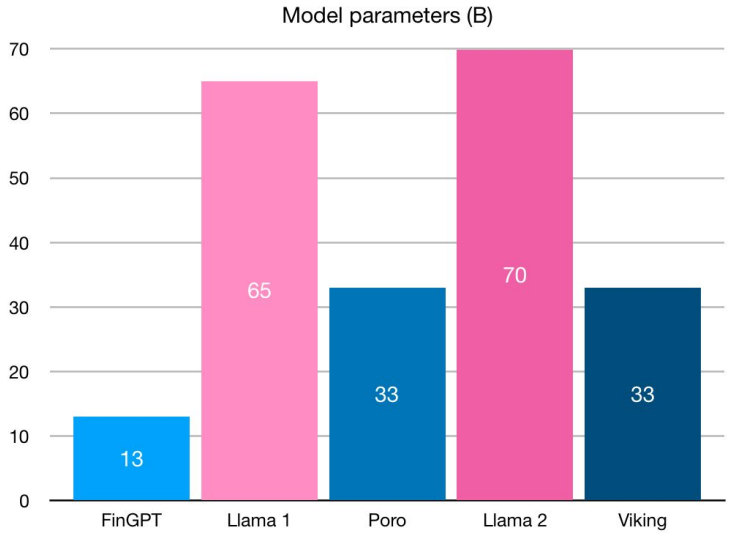
Collaboration lead by HF on LUMI established value of **repeating training data**

- **4x repetition**: almost as good as new data
- **40x repetition**: repeating is worthless
- (Augmenting with code allows 8x data)

How do our models compare?

FinGPT, Poro and Viking vs. Llama models (**excluding Llama 3**)

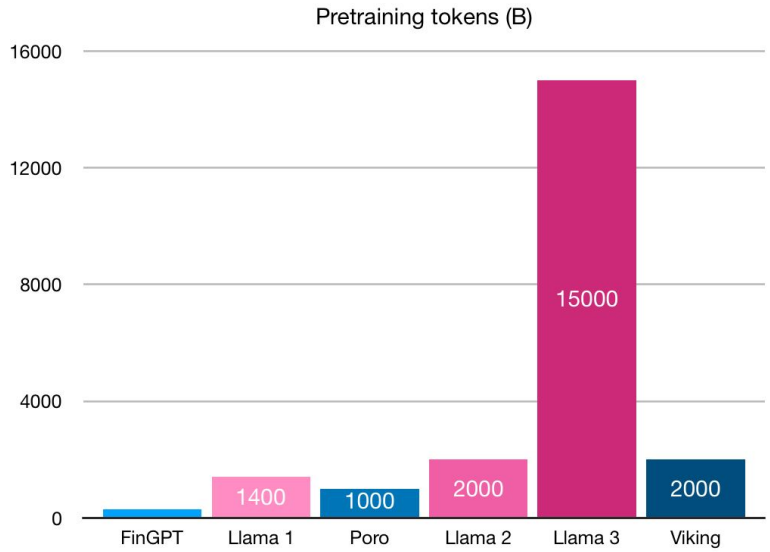


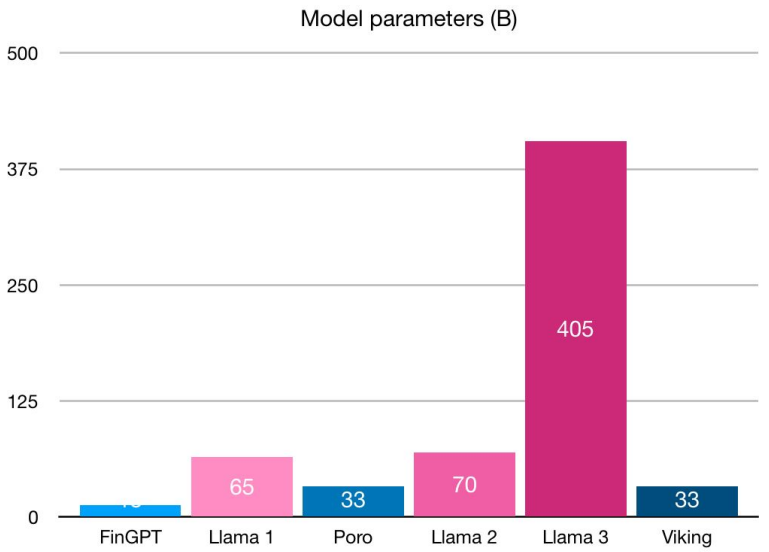


How do our models compare?

Llama 3 training would require **~100M GPUh** on LUMI, or **approx. 1 year** of the whole GPU partition

FinGPT, Poro and Viking vs. Llama models





TURKUNLP .ORG

UNIVERSITY OF TURKU

LLM training on LUMI

Research and development of LLM training methods focused on Nvidia GPUs

Advances frequently implemented as specialized CUDA kernels, **AMD (ROCm) ports have tended to be late and in some cases unreliable**

→ Need to do porting and debugging ourselves, and we've still often been approx. **a year behind the state of the art** in a fast-moving field (cf. FlashAttention)

Practical throughput also limited, e.g. ~100 TFLOPS/MI250X in Poro training (compare 120-140 TFLOPS/A100)

→ For Poro LLM pretraining, a **MI250X was about 70-80% of an A100**

(better throughput in recent experiments)

Pretraining software

For FinGPT and Poro, used fork of the BigScience (BLOOM) fork of **Megatron-DeepSpeed** (late 2022), with ROCm ports of fused kernels

https://github.com/TurkuNLP/Megatron-DeepSpeed/

For Viking and upcoming models, using more recent fork of **Megatron-LM** with Llama configuration and ROCm port of Flash Attention 2

https://github.com/LumiOpen/Megatron-LM-lumi

For both frameworks, relied on custom containers created in collaboration with AMD and CSC

Pretraining software

Other LLM training frameworks:

- **GPT-NeoX**: Megatron family framework with explicit AMD support https://github.com/EleutherAI/gpt-neox
- **Nanotron**: Hugging Face framework focused on LLM training https://github.com/huggingface/nanotron
- Megatron-LM via AMD port of **Transformer Engine** https://github.com/ROCm/TransformerEngine

Running on LUMI: https://github.com/Vmjkom/gpt-neox , https://github.com/Vmjkom/nanotron

LLM training on LUMI

LLMs created by tech giants are trained on **dedicated systems**

On **shared systems** such as LUMI, LLM training competes for compute with everyone else on queue (+service breaks, hardware failures, etc.)

Example: Poro training

- 34B params, 1T tokens → 2e23 FLOPs (6ND operations)
- Continuous computation @ 100 TFLOPS/GPU on 512 GPUs → **~45 days**
- Actual Poro training Sep 15th 2023 - Feb 9th 2024 → **207 days**

Calendar time to complete training over 4x (idealized) compute time

(Half a year is a *long* time in LLM development)

LLM training on LUMI

Suitability depends on model size and task. My personal view currently:

Task	Model parameters		
	<10B	10B-100B	>100B
Pretraining from scratch	Yes		No
Continued pretraining	Yes	Yes	
Fine-tuning	Yes	Yes	Yes
Inference	Yes	Yes	Yes

LLM training on LUMI

What's needed (necessary but not sufficient):

- **Defragment**: stop splitting compute allocations across dozens of projects with similar aims → **very large allocations for 1-2 projects**
- **Dedicated partitions** (or months-long allocations!): allow pretraining **compute to be used without queuing**
- **Robustness to hardware failure**: software solutions that allow pretraining processes to continue even when some nodes are lost

Upcoming EU projects aim to address the first, working on the others

Thanks & Acknowledgments

Silo AI

LUMI and CSC User Support

National Library of Finland

Hugging Face

AMD

Compute: LUMI (Finnish allocation)

Funding: EU / HPLT
(Grant No. 101070350)

TURKUNLP .ORG

UNIVERSITY OF TURKU

Models, tools and resources

FinGPT and **BLUUMI**: https://turkunlp.org/gpt3-finnish
Poro: https://huggingface.co/LumiOpen/Poro-34B
Viking: https://huggingface.co/LumiOpen/Viking-33B
FinGPT paper: https://arxiv.org/abs/2311.05640
Poro paper: https://arxiv.org/abs/2404.01856
Megatron-DeepSpeed port (deprecated):
https://github.com/TurkuNLP/Megatron-DeepSpeed/
Megatron-LM port: https://github.com/LumiOpen/Megatron-LM-lumi
GPT-NeoX and **Nanotron** on LUMI: https://github.com/Vmjkom/gpt-neox ,
https://github.com/Vmjkom/nanotron
LUMI resources: https://lumi-supercomputer.github.io/