Importance of High-Performance Computing (HPC) in Advancing Large Language Models (LLMs)

Gaurish Thakkar

University of Zagreb, Faculty of Humanities and Social Sciences, Institute of Linguistics

13-11-2024

Myself

- Name: dr. sc. Gaurish Thakkar
- Position: Researcher - Zavod za Lingvistiku
- Education: PhD - FFZG (2019-2022)
- Research interest:
  - NLP – Sentiment analysis
  - Dataset creation, curation
  - Large Language Models

Introduction to LLMs and HPC

Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects 2

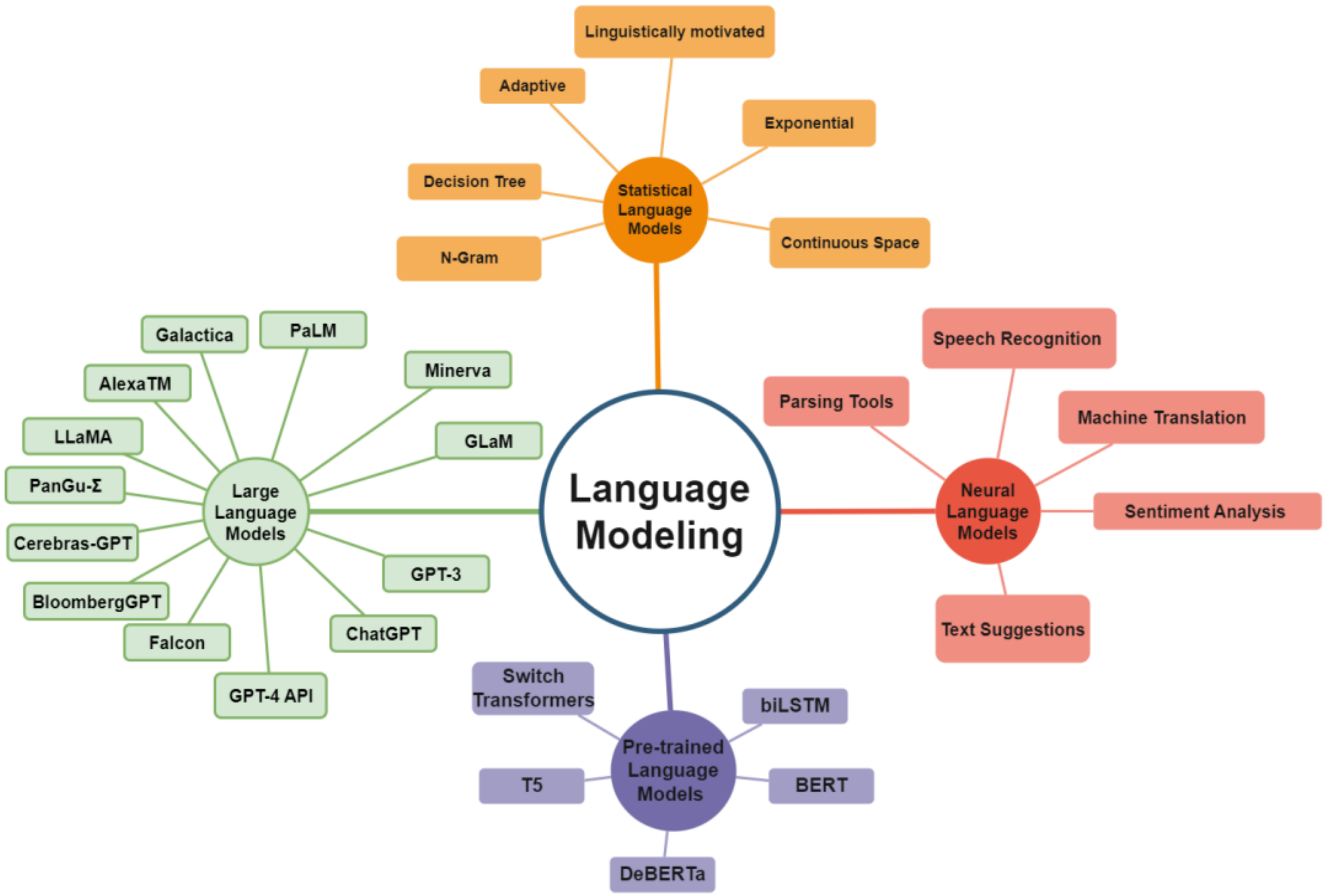


Fig. 1: Types of language modeling. The division of LLMs is categorized into four major blocks

ACHIEVEMENTS UNLOCKED BY LLMS

EMERGENT ABILITIES OF LARGE LANGUAGE MODELS (APR/2023)

S	GPT-3 13B, PaLM 8B	Mod.Arithmetic•	Debugging•	Comprehension•			
M	GPT-3 175B, LaMDA 137B, PaLM 64B, Chinchilla 7B	LinguisticsPuzzles•	EmojiMovie•	GRE-Comprehension•	MetaphorUnderstanding•	PhysicalIntuition•	LogicalDeduction•
L	PaLM 540B, Chinchilla 70B	GeometricShapes•	Proverbs•	PhoneticAlphabet•	ElementaryMath•	CausalJudgment•	CodeLineDescription•
XL	GPT-4, Gemini (est.)	College-LevelExams	Self-Critique/Reflection	AppBuilding	SpatialReasoning	AdvancedCreativity	EmbodimentOptions
Next...		Grounding	Long-HorizonPlanning	Persuasion	AdvancedEmbodiment	Awareness	More...

Selected highlights only. Sources: Original papers and Dr Jason Wei's summary: https://www.jasonwei.net/blog/emergence.
BIG-bench = • Using images from Flaticon.com. Alan D. Thompson. April 2023. https://lifearchitect.ai

LifeArchitect.ai/models

https://llm.extractum.io/list/

HPC f | latest | All | Veter | Read | Trans | LLM | Mediu | Mediu | large | (PDF) | llms - | (PDF) | (PDF) | Large | HPC - | supe

llm.extractum.io/list/

LLM EXPLORER Dark Theme

LLM List | LLM Hosting | LLM Leaderboards | Blog | Newsfeed | Advertise

What open-source LLMs or SLMs are you in search of? 37753 in total.

Search

All Large Language Models

Was this list helpful?

Model Size: 0K – 999B

Model VRAM: 0 – 768

Columns | Quick Filters | Reset Filters & Sorting

Model Name	Maintainer	Size	Score	VRAM (GB)	Quantized	License	Context Len	Likes	Downloads	Modified	Languages	Archite
Meta Llama 3.1 8B Instruct	meta-llama	8B	0.64	16.1	•	meta	128K	2600	3087763	2024-08-20	en, de, fr, it, pt, hi, es, th	LlamaForCa
Llama 3.1 Nemotron 70B Instruct HF	nvidia	70B	0.6	141.9	•	llama3.1	128K	1564	217732	2024-10-25	en	LlamaForCa
Phi 3 Mini 4K Instruct	microsoft	4B	0.6	7.7	•	mit	4K	1068	1360758	2024-09-20	en, fr	Phi3ForCaus
SmolLM2 1.7B Instruct	HuggingFaceTB	2B	0.6	3.4	•	apache-2.0	8K	305	38269	2024-11-05	en	LlamaForCa
DeepSeek V2.5	deepseek-ai	236B	0.59	378.4	•	other	160K	576	20312	2024-10-08	•	DeepseekV2
Aya Expanse 8B	CohereForAI	8B	0.58	16	•	proprietary	8K	267	35550	2024-10-30	en, fr, de, es, it, pt, ja, ko,...	CohereForC
Aria	rhymes-ai	25B	0.57	50.6	•	apache-2.0	•	581	28105	2024-11-07	en	AriaForCond
Meta Llama 3.1 70B Instruct	meta-llama	70B	0.57	141.9	•	meta	128K	513	521937	2024-08-20	en, de, fr, it, pt, hi, es, th	LlamaForCa
Aya Expanse 32B	CohereForAI	32B	0.57	64.2	•	proprietary	8K	165	22668	2024-11-01	en, fr, de, es, it, pt, ja, ko,...	CohereForC
Meta Llama 3 8B Instruct	meta-llama	8B	0.57	16.1	•	llama3	8K	3601	2149125	2024-09-27	en	LlamaForCa
Llama 3 1 Nemotron 51B Instruct	nvidia	51B	0.57	103.4	•	other	128K	189	109691	2024-10-13	en	DeciLMForC
Gemma 2 9B It SimPO	princeton-nlp	9B	0.56	18.6	•	mit	8K	120	102752	2024-08-02	•	Gemma2For
Llama 3.1 8B Instruct	meta-llama	8B	0.56	16.1	•	llama3.1	128K	3014	5130308	2024-09-25	en, de, fr, it, pt, hi, es, th	LlamaForCa
Meta Llama 3 8B	meta-llama	8B	0.56	16.1	•	llama3	8K	5816	668986	2024-09-27	en	LlamaForCa
Phi 3.5 Mini Instruct	microsoft	4B	0.56	7.7	•	mit	128K	614	691102	2024-09-18	•	Phi3ForCaus
Mistral 7B Instruct V0.2	mistralai	7B	0.56	14.4	•	proprietary	32K	2573	1059085	2024-09-27	•	MistralForCa
Llama 3.2 1B	meta-llama	1B	0.55	2.5	•	llama3.2	128K	848	1320393	2024-10-24	en, de, fr, it, pt, hi, es, th	LlamaForCa
Phi 3 Medium 4K Instruct	microsoft	14B	0.54	28	•	mit	4K	211	37237	2024-08-20	•	Phi3ForCaus
Phi 3.5 MoE Instruct	microsoft	42B	0.54	83.9	•	mit	128K	515	35059	2024-10-24	•	PhiMoEForC
Mistral Nemo Instruct 2407	mistralai	12B	0.54	24.5	•	proprietary	•	1202	264425	2024-11-06	en, fr, de, es, it, pt, ru, zh...	AutoModelF
Llama 3.2 1B Instruct	meta-llama	1B	0.54	2.5	•	llama3.2	128K	519	1307062	2024-10-24	en, de, fr, it, pt, hi, es, th	Llam

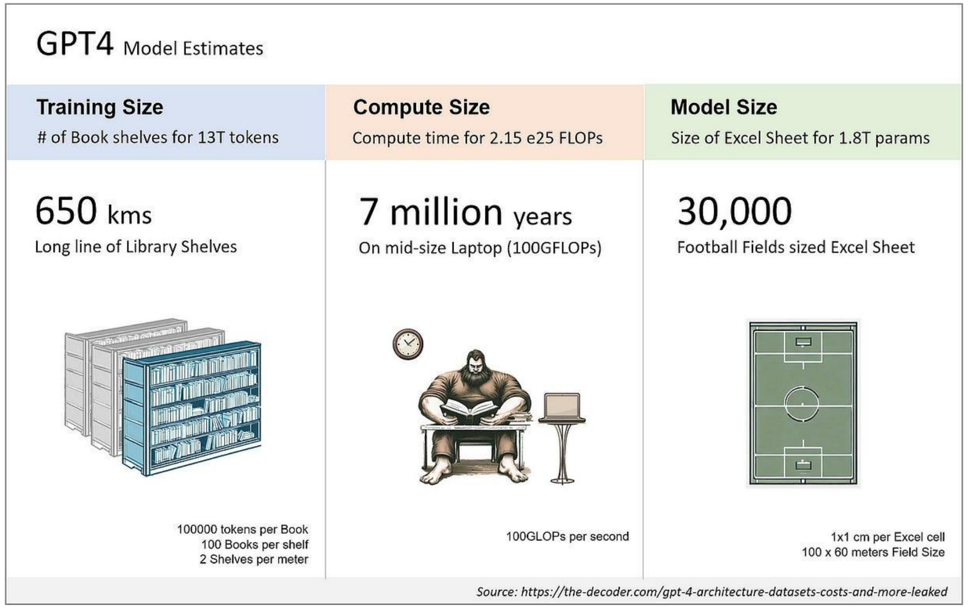
The Role of SRCE in HPC for Research

NVIDIA Tesla V100

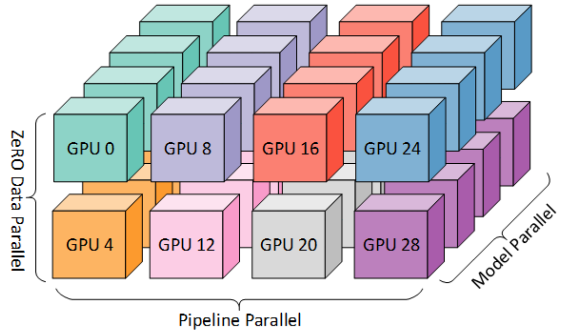
NVIDIA A100 (SXM)

Importance of HPC for LLMs

- Enormous computational requirements of LLMs: training a model with billions of parameters
- HPC enables feasible training times and manageable resources
- Case examples: training HR-GPT, which required supercomputing resources



Technical Overview of HPC Benefits for LLM Training



3D parallelism training. These three types are:

- **Data parallelism:** This involves training the model on multiple GPUs or TPUs simultaneously, each processing a different portion of the data.
- **Model parallelism:** This splits the model's parameters across multiple devices, allowing them to be updated simultaneously.
- **Pipeline parallelism:** This divides the model into stages and processes them in a pipelined fashion, with each stage running on a separate device.

Overview of LLM Research and SRCE's Contribution

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436). Views and opinions expressed are however those of the UTTER consortium only and do not necessarily reflect those of the European Union and UKRI. Neither the European Union nor the UKRI nor the granting authority can be held responsible for them.

HR-XR-XTEND

Croatian Extended Reality Extensions

About

The HR-XT-XTEND is one of eight FSTP subprojects of a larger Horizon Europe funded project **Unified transcription and translation for extended reality (UTTER)**. HR-XT-XTEND project aim is to develop a large language model (LLM) for the Croatian language that will be trained on a massive dataset of Croatian text. The project aims to build resources for XR models, extend XR models to a new language, and evaluate the LLM. The project goals are to collect at least 6 billion tokens of Croatian text and prepare that data for LLM training, create a LLM for the Croatian language using monolingual data only, and evaluate the LLM for downstream tasks. The experimental phase will focus on developing and evaluating the model architecture and training process. The training phase will be used to train the LLM. The integration phase will involve integrating the LLM into the UTTER platform. The project results will be accessible under permissive licenses to the research community and the public from the **HR-CLARIN** repository.

HR-XR-XTEND

- Objectives:
  - Collection of training corpus
  - Training the language model
    - Training from scratch (Pythia)
    - Continued pretraining monolingual (GPT-2)
    - Continued pre-training on multilingual model (Gemma-7b-bnb-4bit)
- Evaluation

Name	Approx Size
CLASSLA Hr Web corpus 1.0	2.5 billion
CC100-Hr Dataset	2.27 billion
Corpus of Croatian News Feeds	2.25 billion
Parallel data for En-Hr on OPUS Resources*,	1.48 billion
Hr-news from XLM-R-BERTić dataset	1.4 billion
Croatian news/legal corpus	175 million
Corpus of Croatian Academic Theses	312 million
ParaCrawl*	69.96 million
Riznica from XLM-R-BERTić dataset	69.51 million
MARCELL Croatian legislative subcorpus	56 million
CURLICAT Croatian corpus	49 million
MARCELL Croatian-English Parallel Corpus of Legislative Texts*	14.3 million
Romance-Croatian Parallel Corpus* (literary works)	2.5 million
Total	8.9 billion

Table 1: Non-exhaustive list of largest data sources used for training the HR-GPT (Beta version) with approximate size in tokens. *Croatian texts only

Results (1/2)

No supervised training (zero-shot evaluation)										
		Pretraining					Vanilla		CPT	
benchmark	metric	160M	350M	410M	1.4B	160M+hrtok	gpt2	gemma-7b	gpt2-en-cpt-hr	gemma-7b-cpt
arc_hr	acc	18.91	20.96	20.36	20.44	20.87	19.85	**32.34**	18.82	21.81
	acc_norm	23.44	25.49	25.06	24.89	23.95	23.87	**36.53**	23.44	24.55
belebele_hrv_Latn	acc	22.78	23	23.11	22.67	22.78	23.44	**52.67**	21.33	23
	acc_norm	22.78	23	23.11	22.67	22.78	23.44	**52.67**	21.33	23
hellaswag_hr	acc	28.43	29.87	30.08	31.36	28.63	26.27	**38.5**	26.44	24.38
	acc_norm	30.07	32.74	33.38	35.52	30.63	29.42	**50.11**	28.14	24.24
m_mmlu_hr	acc	22.65	25.21	22.8	22.54	22.63	22.59	**41.5**	22.67	25.02
truthfulqa_hr_mc1	acc	25.88	24.58	25.75	26.27	25.49	22.24	**28.61**	26.01	18.34
truthfulqa_hr_mc2	acc	43.82	42.21	42.34	42.52	43.03	40.8	46.6	**46.79**	-

Table 2: Benchmarking evaluation (zero-shot) results for a variety of models without the use of any supervised training. The table displays scores for various models that did not utilise any supervised training (instruction fine tuning). ACC: accuracy and acc_norm: normalised accuracy.

Results (2/2)

Trained on benchich training data										
dataset	metric	160M	350M	410M	1.4B	160M+hrtok	gpt2	gemma-7b	gpt2-en-cpt-hr	gemma-7b-cpt
SA-Parlasent(hr-only)	acc	68.86	72.98	72.53	71.03	71.18	36.68	72.46	53.74	**74.48**
COPA	acc	50	49.4	49.6	47.8	48.8	48.4	**79.8**	50.2	79.6

Table 3: The model scores (accuracy) for supervised tasks related to sentiment analysis and choice of plausible alternatives (COPA).

Supervised training (instruction fine tuning)									
Alpaca									
benchmark	160M	350M	410M	1.4B	160M+hrtok	gpt2-en-cpt-hr	gemma-7b-cpt	gpt2	gemma-7b
arc_hr	21.21	19.76	22.75	23.1	20.19	19.08	35.76	19.67	**35.93**
	26.26	25.32	25.75	27.12	23.18	24.64	37.81	24.12	**39.95**
belebele_hrv_Latn	22.67	22.89	23.44	22.67	22.67	23.78	**58**	23.89	45.33
	22.67	22.89	23.44	22.67	22.67	23.78	**58**	23.89	45.33
hellaswag_hr	28.56	30.14	30.74	31.64	28.96	26.84	40.35	26.27	**41.1**
	30.19	32.76	33.75	35.46	30.39	27.62	53.56	27.7	**53.67**
m_mmlu_hr	22.69	23.05	22.82	22.76	22.79	22.63	**43.12**	22.62	33.12
truthfulqa_hr_mc1	24.19	22.63	23.67	26.92	25.23	25.1	**31.73**	24.45	30.04
truthfulqa_hr_mc2	42.58	40.8	39.2	42.87	42.93	41.1	**50.04**	40.08	47.68

Table 4: Benchmarking evaluation results for a variety of models trained with the Alpaca instruction tuning dataset.

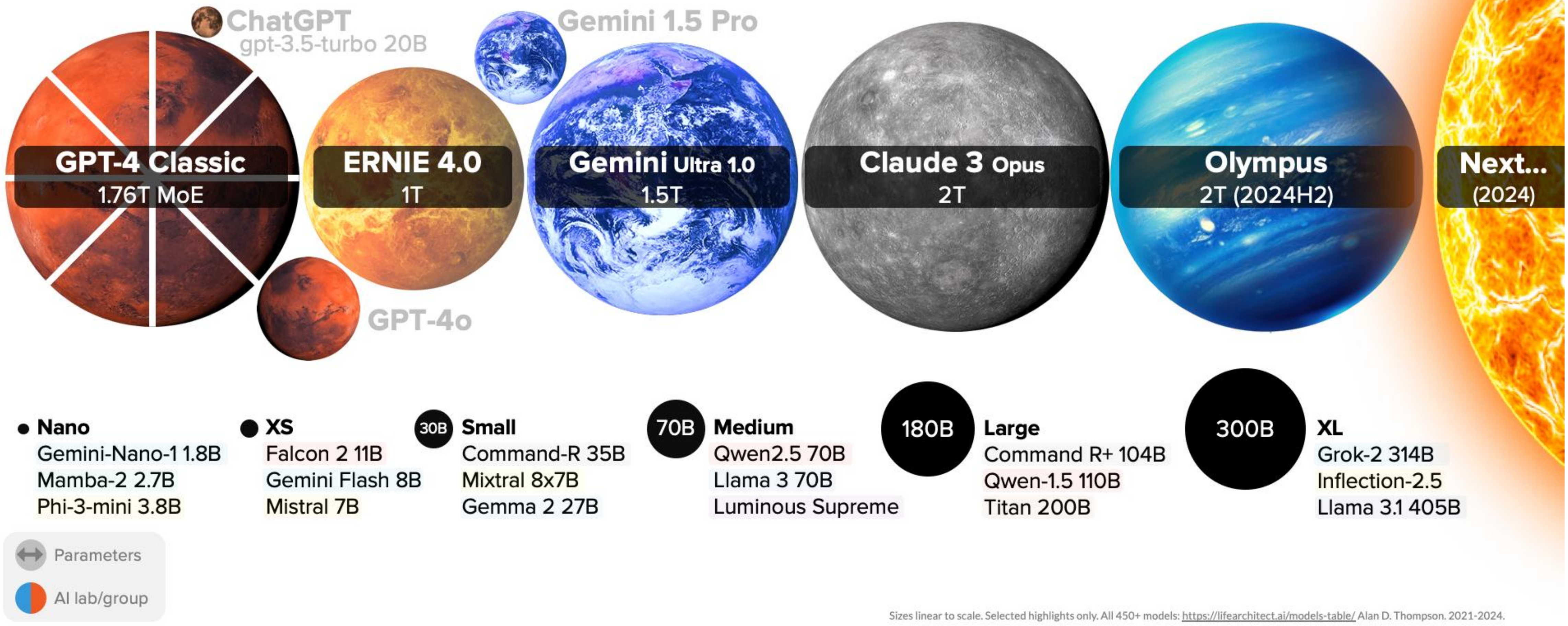
HPC and LLM Scalability for Broader Research Impact

- **Scalability in research**
  - Larger model
  - Datasets
  - Tasks
- **Wider applications beyond NLP**
  - multi-modal AI

Future Prospects of HPC in AI and LLM Research

- **Evolving Role of HPC**
  - Exascale computing
- **Transformative Potential**
  - Impact of HPC in AI
  - Enable new applications in industry, academia, etc.

LARGE LANGUAGE MODEL HIGHLIGHTS (OCT/2024)



LifeArchitect.ai/models & 450+ more models at LifeArchitect.ai/models-table

Summary of current models [link]

2024 LifeArchitect.ai data (shared) - NEW

Model	Lab	Playground	Parameters (B)	Tokens trained (B)	Ratio Tokens:Params (Chinchilla scaling≥20:1)	ALScore "ALScore" is Sqr Root of	MMLU	MMLU -Pro	GPQA	Training dataset	Announced ▼	Public?	Paper Repo
o1	OpenAI	https://chatgpt.com/	200	20000	100:1	6.7	92.3	91	78.3	W 📚 ⬆ 🕸 🌋	Sep/2024	🟢	https://o
Claude 3.5 Sonnet (r	Anthropic	https://claude.ai/					90.5	78	65	W 📚 ⬆ 🕸 🌋	Oct/2024	🟢	https://a
Hunyuan-Large	Tencent	https://huggingface.cc	389	7000	18:1	5.5	89.9	60.2	42.4	W 📚 ⬆ 🕸 🌋 ⚛	Nov/2024	🟢	https://a
Claude 3.5 Sonnet	Anthropic	https://poe.com/Claude-3.5-Sonnet					88.7	76.1	67.2	W 📚 ⬆ 🕸 🌋	Jun/2024	🟢	https://w
GPT-4o	OpenAI	https://chatgpt.com/	200	20000	100:1	6.7	88.7	72.6	53.6	W 📚 ⬆ 🕸 🌋	May/2024	🟢	https://o
Llama 3.1 405B	Meta AI	https://www.meta.ai/	405	15000	38:1	8.2	88.6	73.3	51.1	W 📚 ⬆ 🕸 🌋	Jul/2024	🟢	https://a
Grok-2	xAI	https://x.com/i/grok	600	15000	25:1	10.0	87.5	75.5	56	W 📚 ⬆ 🕸 🌋	Aug/2024	🟢	https://x.
Claude 3 Opus	Anthropic	https://claude.ai/	2000	40000	20:1	29.8	86.8	68.5	59.5	W 📚 ⬆ 🕸 🌋	Mar/2024	🟢	https://w
gpt-4-turbo-2024-04	OpenAI	https://chat.openai.com/		13000			86.5	63.7	49.1	W 📚 ⬆ 🕸 🌋	Apr/2024	🟢	https://c
GPT-4 Turbo	OpenAI	https://chat.openai.com/		13000			86.4		46.5	W 📚 ⬆ 🕸 🌋	Nov/2023	🟢	https://c
GPT-4 Classic (gpt-4-	OpenAI	https://chat.openai.co	1760	13000	8:1	15.9	86.4		35.7	W 📚 ⬆ 🕸 🌋	Mar/2023	🟢	https://c
Qwen2.5	Alibaba	https://huggingface.cc	72	18000	250:1	3.8	86.1	71.1	49	W 📚 ⬆ 🕸 🌋	Sep/2024	🟢	https://q
Gemini 1.5 Pro	Google DeepMind	https://aistudio.googl	1500	30000	20:1	22.4	85.9	69	46.2	W 📚 ⬆ 🕸 🌋	Feb/2024	🟢	https://g
Inflection-2.5	Inflection AI	https://inflection.ai/in	1200	20000	17:1	16.3	85.5		38.4	W 📚 ⬆ 🕸	Mar/2024	🟢	https://ir

Models Table (2024) | About | Datasets Table (2024) | Chinese LLMs (2023) | Humanoids (2024) | Compute Table (2024) | Sum: 8,452.01

Summary – Universal Benefits of HPC for LLM Research

- **Speed**: HPC reduces the time needed for training and testing, accelerating research timelines.
- **Cost-Effectiveness**: High efficiency and reduced costs over long-term computations.
- **Scalability and Flexibility**: HPC provides scalable resources, enabling experiments with different model sizes and configurations.
- **Enhanced Collaboration**: Accessibility to HPC fosters collaborations across institutions and disciplines.

Conclusion: The Path Forward

- **Scale requires HPC**: HPC enables the training, fine-tuning, and deployment of LLMs at scale
  - HPC’s critical role in advancing LLMs
- **Optimization is key**: Advancements in model architecture, hardware, and algorithms can help manage resource use
  - Balancing efficiency with computational demands
- Innovations in HPC that will define the next generation of LLMs

Closing Remarks and Acknowledgments

We would like to thank SRCE for the support and resources

HR-XR-XTEND

https://hr-xr-xtend.ffzg.unizg.hr/

Thank you.

Q&A