Fraunhofer IPA

Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

Alper Yaman*, Jannik Schwab, Christof Nitsche, Abhirup Sinha and Marco Huber

Presenter: Alper Yaman, Senior Research Expert, Fraunhofer IPA, Stuttgart, Germany
alper.yaman@ipa.fraunhofer.de

Comparison of Large Language Models for Deployment Requirements

Dr. Alper Yaman

alper.yaman@ipa.fraunhofer.de

Dr. Alper Yaman is a senior research expert at the Department Cyber Cognitive Intelligence, Fraunhofer IPA, and the Institute of Industrial Manufacturing and Management, University of Stuttgart.

He focuses on applications of computer vision and artificial intelligence, including data-efficient AI, physics-informed AI, generative AI, and reinforcement learning in automation, robotics, and medical fields.

He holds a PhD in Biomedical Engineering from Bogazici University, Istanbul and has conducted post-doctoral research at Tennessee State University where he worked on mesoscale robots and sensor fusion.

Fraunhofer IPA

Introduction

- LLMs are advanced AI models designed to generate human-like text.
- They have revolutionized many fields.
  - E.g., content creation, customer service, and software development.
  - Also integral for translation, summarization, and QA systems.
- The LLMs differ in architecture, training procedures, and training data.
- Three main categories of LLMs
  - **Foundational:** General-purpose pre-trained LLMs.
  - **Fine-tuned:** Foundational LLMs further trained on task-specific or domain-specific corpora.
  - **Mixture-of-Experts (MoE):** a set of LLMs (experts) attend to different parts of the input.

Source- Revolutionizing AI: Potential of Large Language Models (LLMs) for Advanced Applications | by Tamanna | Medium

Fraunhofer IPA

Current State and Challenges

- As of May 2024, HuggingFace has 65 pre-trained LLMs for English language text generation tasks.
  - Numerous fine-tuned LLMs are also uploaded to HuggingFace.
  - Several LLM leaderboards also exist, e.g., Open LLM Leaderboard, MTEB Leaderboard, LMSYS Chatbot Arena, etc.
  - They usually compare LLM types, architectures, model precisions, average accuracies, etc.
  - Also, evaluation is done using various datasets and benchmarks.

T	Model	Average ⬆️	IFEval	BBH	MATH Lvl 5	GPQA	MUSR	MMLU-PRO
🔶	Qwen/Qwen2-72B-Instruct	43.02	79.89	57.48	35.12	16.33	17.17	48.92
💬	meta-llama/Meta-Llama-3-70B-Instruct	36.67	80.99	50.19	23.34	4.92	10.92	46.74
🟢	Qwen/Qwen2-72B	35.59	38.24	51.86	29.15	19.24	19.73	52.56
🗃	mistralai/Mixtral-8x22B-Instruct-v0.1	34.35	71.84	44.11	18.73	16.44	13.49	38.7
💬	HuggingFaceH4/zephyr-orpo-141b-A35b-v0.1	34.23	65.11	47.5	18.35	17.11	14.72	39.85
💬	microsoft/Phi-3-medium-4k-instruct	33.12	64.23	49.38	16.99	11.52	13.05	40.84

Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)
1	SFR-Embedding-2_R	7111	26.49	4096	32768	70.31	89.05	56.17
2	gte-Qwen2-7B-instruct	7613	28.36	3584	131072	70.24	86.58	56.92
3	neural-embedding-v1					69.94	87.91	54.32

Source- https://huggingface.co/spaces/mteb/leaderboard and https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

- These leaderboards do not provide the requirements for LLM deployment.

- **Challenges:** selecting an LLM that meets specific requirements
  - Especially important when it is intended for local deployment.

- **Aim:** providing a comparative list of foundational and domain-specific LLMs, focusing on deployment requirements.

Proposed Work

- **An extensive comparison list of LLMs.**
  - Simplify LLM selection for deployment purposes.
- **Primary focus on foundational general-purpose LLMs.**
  - Some domain-specific fine-tuned models were also included.
- **We provided both LLM names and families together with the model features.**
  - Thus, different LLMs can be easily distinguished.
- **The comparison table is published online.**
  - URL: https://technology-project-aimv-projects-generative-ai-54af1e2b8cbbab0a.pages.fraunhofer.de/
  - The table will be updated regularly.

Source- How To Chose Perfect LLM For The Problem Statement Before Finetuning (labellerr.com)

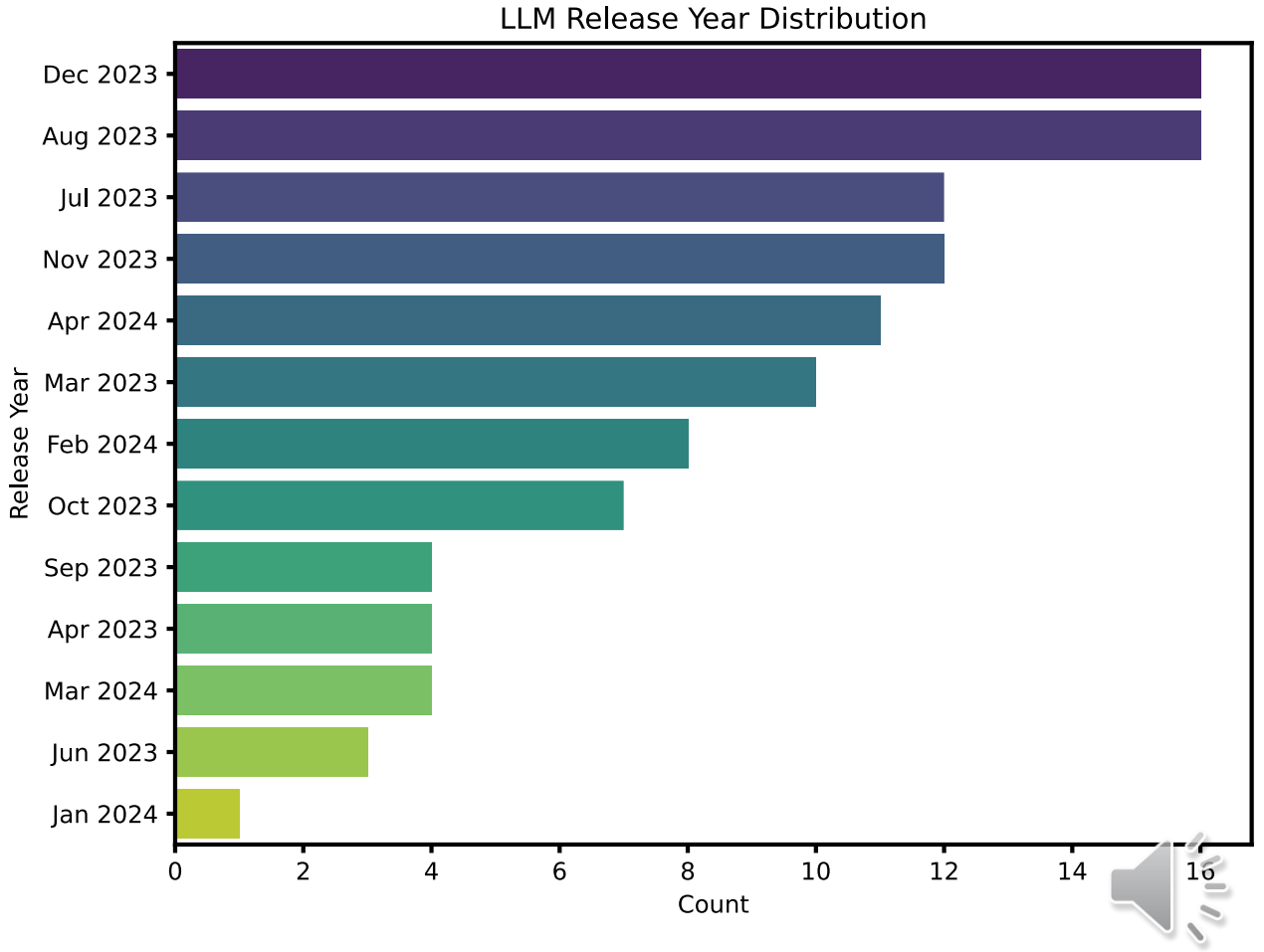
Model Selection and Model Features

- **We listed 108 open-source LLMs, published in or after 2023.**
  - Approximately 20 foundational LLMs.
    - E.g., Mistral, LLaMA-2, LLaMA-3, Gemma, RecurrentGemma, Falcon, etc.
  - Several fine-tuned LLMs.
    - E.g., BioMistral, Meditron, Medicine-LLM, etc.
  - Also some MoE LLMs.
    - E.g., Mixtral, Grok-1, and DBRX, etc.
- **We included several information regarding the LLMs.**
  - LLM families and versions, number of parameters, RAM and GPU Memory requirements.
  - Also included license information and clarification regarding commercial usage.

Source- How To Chose Perfect LLM For The Problem Statement Before Finetuning (labellerr.com)

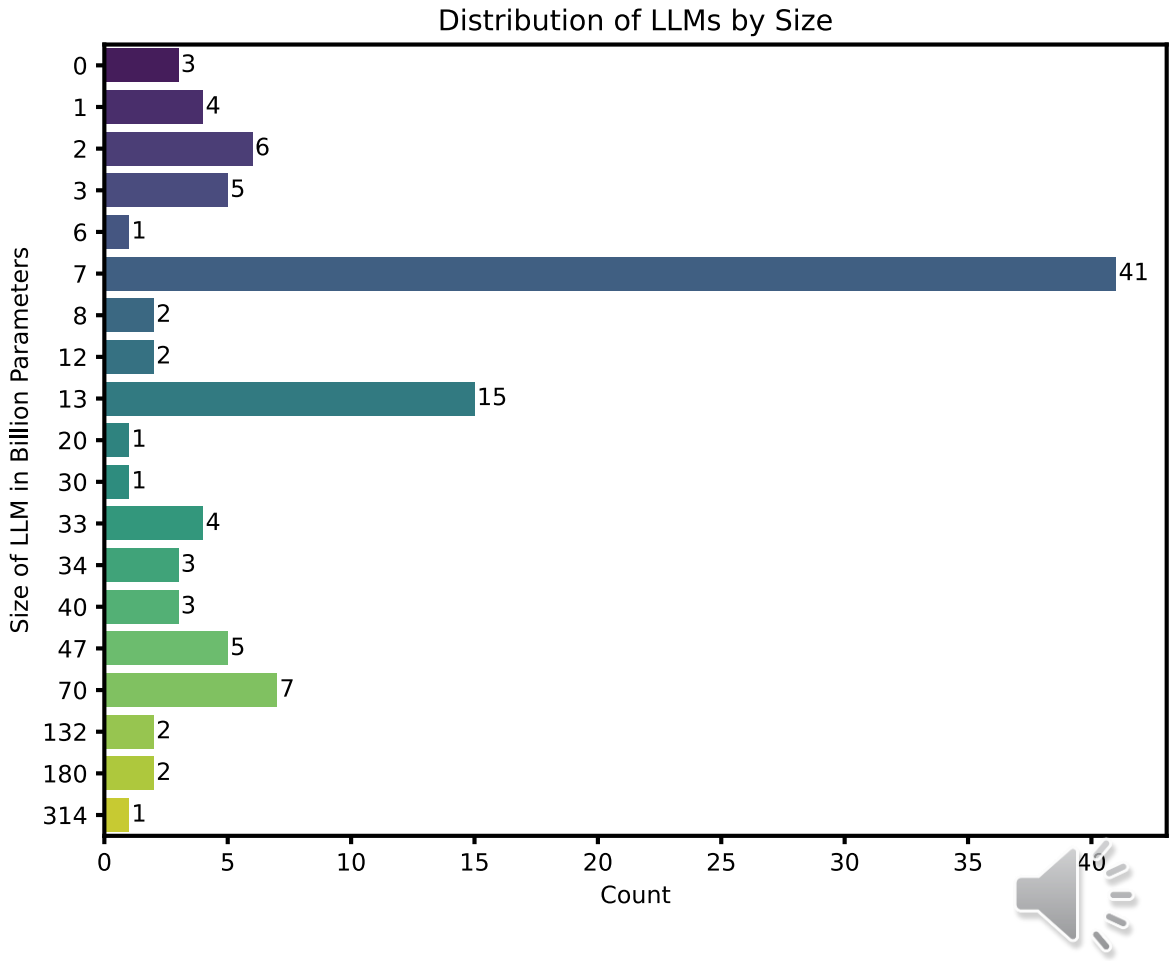
Results: Release Year Distribution of Listed LLMs

- Considered release time 2023 onwards.
- Most LLMs from 2023.
- Almost 32 models from Dec 2023 or Aug 2023.
- Most recent LLMs from Apr 2024.



Results: Distribution of LLM Size in Billion Parameters

- Size usually range from 1B to 314B parameters.
- Most LLMs have 7B parameters.
- Few LLMs have less than 1B parameters.
- Lower number of parameters allow LLMs to be deployed on edge devices, e.g., NVIDIA Jetson.
- Larger LLMs require more hardware resources.



Results: License Distribution of Open-Source LLMs in List

- **Around 51% LLMs have permissive licenses.**
  - Apache 2.0, MIT, Gemma, etc.
- **Around 32% LLMs have limited (“partial”) commercial usage licenses.**
  - LLaMA-2, LLaMA-3, Databricks Open Model License, etc.
  - Require permission if commercial usage exceeds 700M monthly active users.
- **Some LLMs do not allow commercial usage.**
  - CC-BY-NC 4.0, CC-BY-NC-ND 4.0, Med42, etc.

License Type	Count	Percentage (%)
Apache 2.0	36	33.33
LLaMA-2	29	26.85
Gemma	12	11.11
MIT	7	6.48
CC-BY-NC 4.0	5	4.63
CC-BY-NC-ND 4.0	4	3.70
LLaMA-3	4	3.70
Non-commercial	3	2.78
Microsoft Research License	2	1.85
Databricks Open Model License	2	1.85
Falcon-180B TII License	2	1.85
Med42 (derivative of LLaMA-2)	1	0.93
StabilityAI Non-Commercial Research Community License	1	0.93
Total	**108**	**-**

Results: Snapshot of the Table of Current Open-Source LLMs

						Fine-tuning	Inference		
Family	Name	Release Year	Size (B Parameters)	License type	Commercial Usage	Min. GB GPU	Min. GB RAM	Min. GB GPU	Min. GB Disk Space
Code	Code-13B	Dec 23	13	CC-BY-NC-ND 4.0	No	26	11.73	5.4	9.23
	Code-33B	Dec 23	33	CC-BY-NC-ND 4.0	No	66	25.55	13.5	23.05
CodeLLaMA	7B	Aug 23	7	LLaMA-2	Partial	14	7.28	2.8	4.78
	7B-Instruct	Aug 23	7	LLaMA-2	Partial	14	7.28	2.8	4.78
	7B-Python	Aug 23	7	LLaMA-2	Partial	14	7.28	2.8	4.78
	13B	Aug 23	13	LLaMA-2	Partial	26	11.73	5.4	9.23
	13B-Instruct	Aug 23	13	LLaMA-2	Partial	26	11.73	5.4	9.23
	13B-Python	Aug 23	13	LLaMA-2	Partial	26	11.73	5.4	9.23
	34B	Aug 23	34	LLaMA-2	Partial	68	26.84	14.2	23.84
	34B-Instruct	Aug 23	34	LLaMA-2	Partial	68	26.84	14.2	23.84
	34B-Python	Aug 23	34	LLaMA-2	Partial	68	26.84	14.2	23.84
LLaMA-2	7B	Jul 23	7	LLaMA-2	Partial	14	7.28	2.8	4.78
	7B-Chat	Jul 23	7	LLaMA-2	Partial	14	7.28	2.8	4.78
	7B-Coder	Dec 23	7	LLaMA-2	Partial	14	7.28	2.8	4.78
	13B	Jul 23	13	LLaMA-2	Partial	26	11.73	5.4	9.23
	13B-Chat	Jul 23	13	LLaMA-2	Partial	26	11.73	5.4	9.23
	70B	Jul 23	13	LLaMA-2	Partial	140	51.25	29.3	48.75
	70B-Chat	Jul 23	70	LLaMA-2	Partial	140	51.25	29.3	48.75
Med42	70B	Nov 23	70	Med42	No	140	51.25	29.3	48.75
Starling LM	7B-Alpha	Nov 23	7	CC-BY-NC 4.0	No	14	7.63	2.7	5.13
	Alpha 8X7B MoE	Dec 23	47	CC-BY-NC 4.0	No	94	34.73	17.3	32.23
WizardLM	7B-v1.0	Apr 23	7	Non-commercial	No	14	7.28	2.8	4.78
	13B-v1.2	Jul 23	13	LLaMA-2	Partial	26	11.73	5.4	9.23
	30B-v1.0	Jun 23	30	Non-commercial	No	60	25.55	13.5	23.05
	70B-v1.0	Aug 23	70	Non-commercial	No	140	51.25	29.3	48.75
Zephyr	3B	Nov 23	3	StabilityAI Non-Commercial Research Community License	No	6	4.49	1.2	1.99
	7B-Alpha	Oct 23	7	MIT	Yes	14	7.63	2.7	5.13
	7B-Beta	Oct 23	7	MIT	Yes	14	7.63	2.7	5.13
BioMistral	7B	Feb 24	7	Apache 2.0	Yes	14	7.63	2.7	5.13
	7B-DARE	Feb 24	7	Apache 2.0	Yes	14	7.63	2.7	5.13
	7B-TIES	Feb 24	7	Apache 2.0	Yes	14	7.63	2.7	5.13
	7B-SLERP	Feb 24	7	Apache 2.0	Yes	14	7.63	2.7	5.13
TinyLLaMA	1.1B-Chat-v1.0	Jan 2024	1.1	Apache 2.0	Yes	2.2	3.28	0.5	0.78

Conclusion

- Our aim is supporting researchers and companies.
  - In selecting open-source LLMs suitable for their use cases and needs.
  - Also suggesting hardware requirements for chosen LLMs.
- **Limitations**
  - Our list may not always include the latest LLMs.
  - It also may not include all available fine-tuned LLMs.
- In future, we will include more domain-specific LLMs to list.
- Furthermore, we will assess user feedbacks regarding deployments.
- We will also highlight the advantages and disadvantages of the recommended deployments.

Source- Project Conclusion PowerPoint Presentation | Conclusion Slide Example (kridha.net)

Fraunhofer IPA

Contact

Dr. Alper Yaman
Department of Cyber Cognitive Intelligence
alper.yaman@ipa.fraunhofer.de
https://www.linkedin.com/in/alperyaman/

Fraunhofer IPA
Nobelstraße 12
70569 Stuttgart
www.ipa.fraunhofer.de

Fraunhofer IPA

Fraunhofer-Institut für Produktions-
technik und Automatisierung IPA