Applied Data Science and Analysis
Vol.2023, **pp**. 150–154
DOI: https://doi.org/10.58496/ADSA/2023/014; ISSN: 3005-317X
https://mesopotamian.press/journals/index.php/ADSA

Discussion Article

Does Lack of Knowledge and Hardship of Information Access Signify Powerful AI? A Large Language Model Perspective

Idrees A. Zahid[1, 2,*], Shahad Sabbar Joudar [2]

[1] *Electrical & Computer Engineering, Gannon University Erie, PA, USA*
[2] *Information Technology Center, University of Technology, Baghdad, Iraq*

ARTICLE INFO

Article History
Received 21 Sep 2023
Accepted 24 Nov 2023
Published 12 Dec 2023

Keywords
Large Language Model,
Digital Corpus,
Reinforcement Learning,
Human Feedback,
Artificial Intelligence

A B S T R A C T

Large Language Models (LLMs) are evolving and expanding enormously. With the consistent improvement of LLMs, more complex and sophisticated tasks will be tackled. Handling various tasks and fulfilling different queries will be more precise. Emerging LLMs in the field of Artificial Intelligence (AI) impact online digital content. An association between digital corpus scarcity and the improvement of LLMs is drawn. The impact it will bring to the field of LLMs is discussed. More powerful LLMs are insights to be there. Specifically, increase in Reinforcement Learning from Human Feedback (RLHF) LLMs release. More precise RLHF LLMs will endure development and alternative releases.

1. INTRODUCTION

LLMs are a revolutionary development in AI technology nowadays. The easy-of-use concept of all LLM models helped to increase the improvement of LLMs [1]. The key milestone of the development of these models is the fine-tuning of the RLHF, which enables the model to stratify with human values and preferences by leveraging human feedback [2]. The fine-tuning means, the ability of models to adapt to specific actions by tuning their parameters. The fine-tuning builds upon a pre-trained learning of the model providing suitable LLMs for text generation, or sentiment analysis [3]. While the RLHF is the ability to correctly follow human instructions and feedback, by fine-tuning and training the LLMs. The RLHF allows the LLMs to respond to the user's intents even if they are not explicitly described [2]. The purpose of this manuscript is to discuss the power of LLMs and how far it's used and improved in recent years. The scenario of this discussion begins with LLM Training Methodologies, which briefly discusses the methodologies of large language models. The LLM Training Sources are then discussed to explain the main resources of LLMs. The Corpus Size Statistics LLM Development and LLMs Development Releases followed to show the Statistics of LLM expansion and releases over the recent years. The last section talks about the LLM's impact on the scares of online information and how far utilizing LLMs in multiple tasks.

2. LLM TRAINING METHODOLOGIES

Significant advancement and development have recently been introduced in the Artificial Intelligence world. Several needs and demands require unique AI applications. Large Language Models (LLMs) are AI application models that mainly deal with language resources and produce customized answers based on input or user needs. LLMs like GPT-4 were developed by training the models on complex and huge data and resource-intensive means. These training methodologies consist of several processes and have various considerations and could be one of the following as a stage in LLMs production. The following training methods come after data collection from vast internet sources, like books, articles, websites, and textual data [4][5][6]

- Pre-training: This method takes a broad dataset to train the model on. This step is crucial to understand human language and grasp it. Huge unlabeled data mostly text from online resources is used to train the model and build up proficiency within LLMs.

*Corresponding author. Email: iajzahid@gmail.com

- Fine-tuning: Pre-trained models could be fine-tuned to undertake specific tasks. The need for a particular direction of adoption needs an improvement in a certain way. More specialized datasets are used to improve specific domain performance.
- Reinforcement Learning from Human Feedback (RLHF): The RLHF method employs human feedback on the behaviors of models and their output responses while training on the textual data.

The size of LLMs is exponentially increasing. The level of which is getting more complex. The number of parameters is incredibly increasing. More blocks of the Transformer decoder are used to maintain the scaling up. Longer window lengths and employing parallel self-attention layers. As those models increased in size, a need for enormous training data resources. Massive computational resources to maintain the level of complexity and the increased size should be attained.

3. LLM TRAINING SOURCES

Training LLMs as well as generative models requires multiple processes. However, it is crucial to use real datasets to mimic human intelligence. This level of intelligence and the human touch could be reflected in LLMs and generative AI as the model is trained on diverse and vast human resources. Various training resources for LLMs are used, some of which are listed below. The textual nature of these resources is the common factor [7][8]

- Wikipedia: Include data from the Wikipedia website and with multiple languages.
- Common Crawl: Dataset with terabytes of data. Crawled from online pages and considered a primary source for many LLMs
- C4 (Colossal Clean Crawled Corpus): This dataset is cleaned and derived from the Common Crawl. Duplications are removed to provide quality English text sources.
- ROOTS: Multilingual dataset of 59 languages. This dataset is designed to train the BigScience BLOOM model.

4. CORPUS SIZE STATISTICS LLM DEVELOPMENT

The aforementioned necessity of corpus data for LLMs training and production of these models. Considering Wikipedia as one of the main sources used to feed input for LLMs training. We can notice a remarkable decline in the online increase percentage of this dataset. This decline occurs despite the growing number of online users and the rise of online impact and presence. Figure 1 shows the English Wikipedia increase rate during the preceding year. It shows statistics of word count starting from 2002 to 2023. A clear decline in the percent increase of words for each year is presented. A trending line shows the decline of the word count percent increase [9].

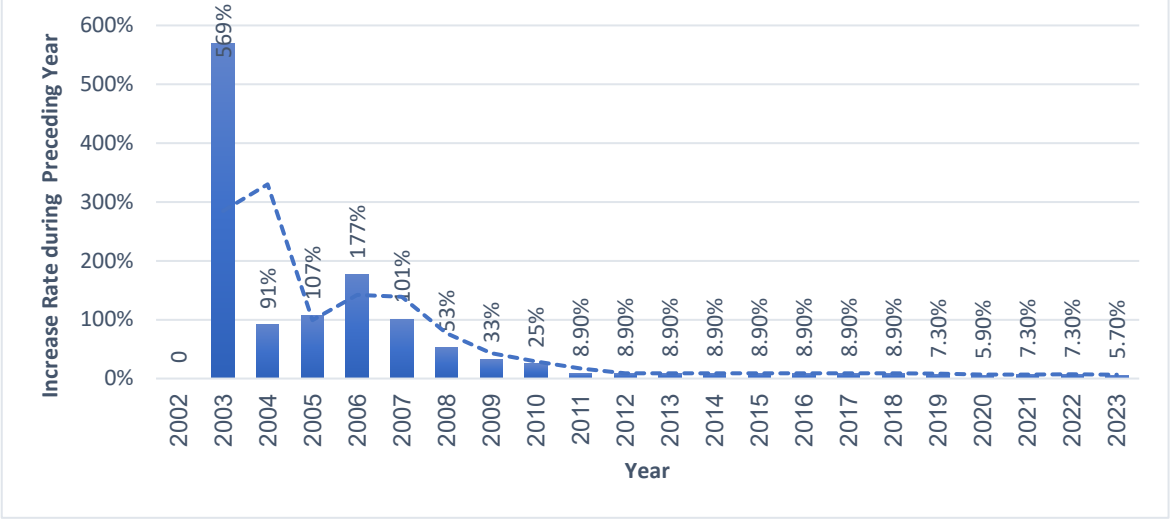

Fig. 1. Wikipedia Word Yearly statistics

Alternatively, Figure 2 shows annual article growth statistics for English Wikipedia. A declining percentage is occurring for the annual articles regarding the preceding year. Starting from 390% on 1/1/2003 and declining to 2.48% on 12/19/2023 [9].

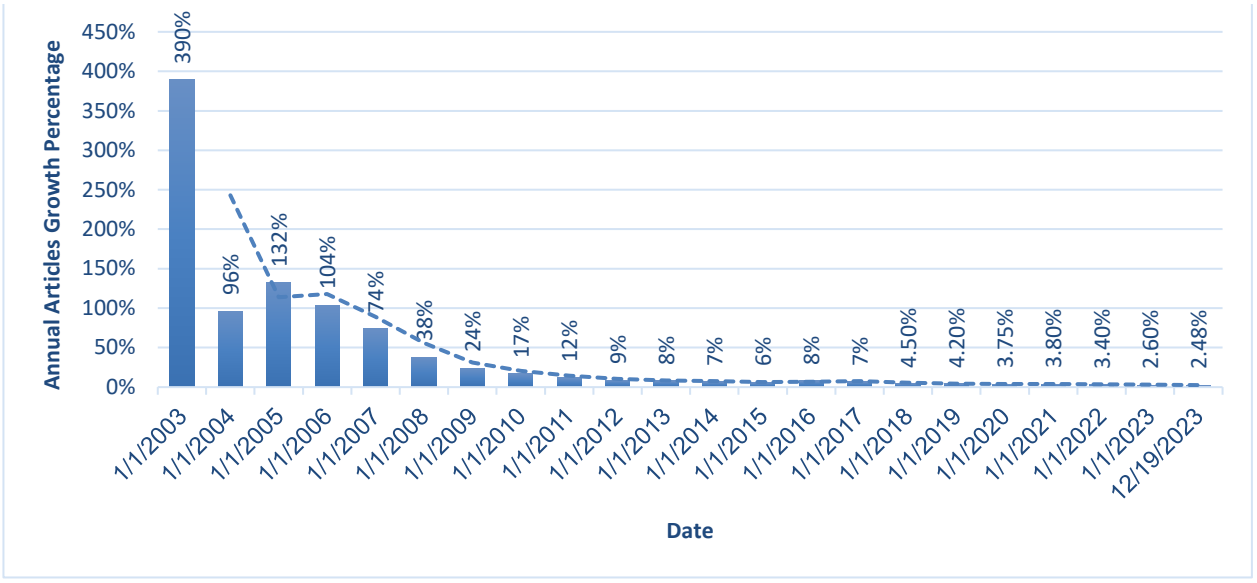

Fig. 2. Wikipedia Annual Articles Growth Statistics

5. LLMS DEVELOPMENT RELEASES

On the other hand, a rise in the production and release of new LLMs is noticed. This rise comes after the development of transformer technology into LLMs. Starting employing transformers for GPT series in 2018 a huge rise in producing LLMs occurred. As presented in Figure 3 the trending line shows the increase in releasing a plethora of LLMs in recent years. LLMs release statistics ranged from a couple of LLMs released in 2018 and 2019 and only one in 2020 to 10, 7, and 18 models in 2021, 2022, and 2023 respectively [2][10].

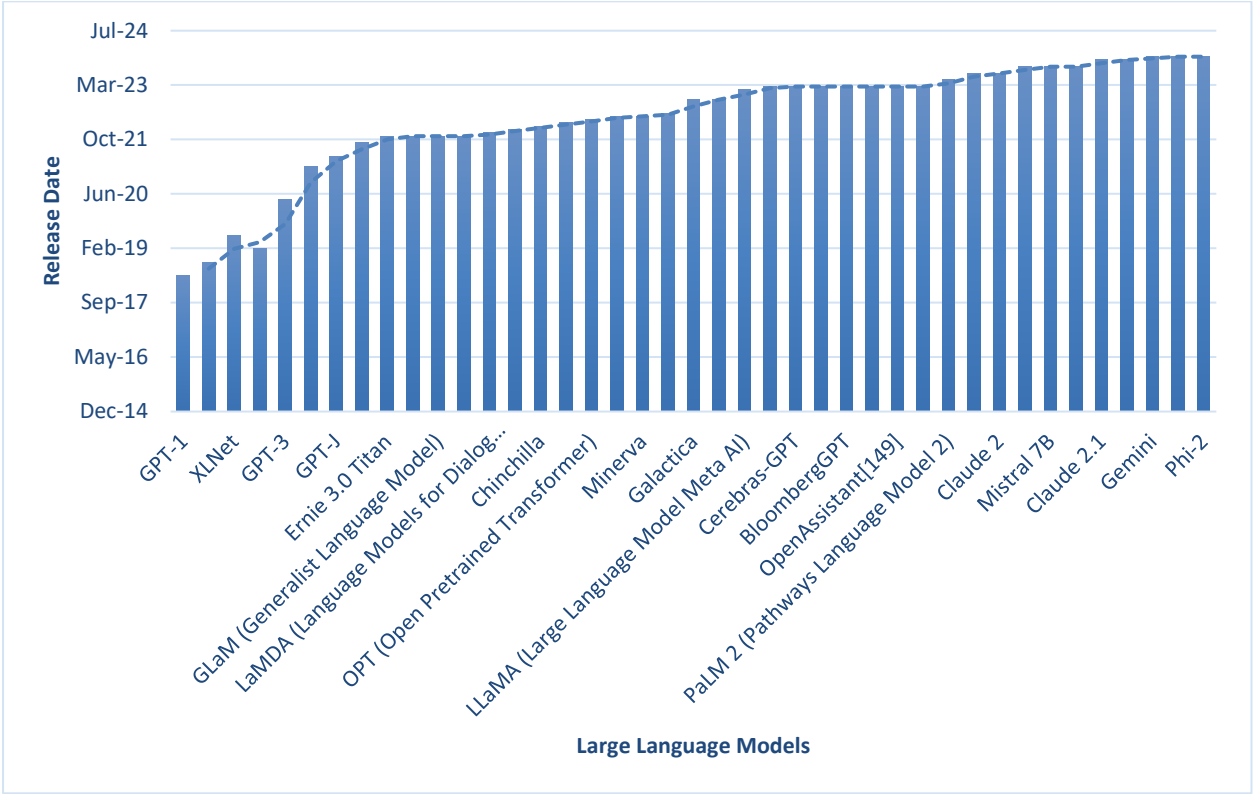

Fig. 3. Large Language Models Release Date Statistics.

6. LLMS IMPACT THE SCARES OF ONLINE INFORMATION

Recent LLMs have been produced in a more sophisticated and improved manner. With the advent of the transformers; LLMs capabilities and complexities increased. Multimodality, specific tasks, and professional human-like generated output. Improvements brought more users and widened its range of customers. Considering LLMs like ChatGPT as a stepping stone to online information it is been used as a search engine [11]. LLMs capabilities limit online access to search engines, as LLMs provide the required result of queried questions in a convenient manner. This results in less online information. Less information, queries, and posts will be posted online. Less forum content, questions, and answers will be available. Instead of posting such, a quick query will be analyzed and a customized answer will be presented to the user with the help of advanced LLMs.

7. CONCLUSION

According to the trending lines representing declining English Wikipedia growth annually. Declining for both word count and articles. Considering Wikipedia as one of the important corpora used to train LLMs and how crucial the training stage is for the curation, production, and improvement of LLMs. On the other hand, the massive release of LLMs in recent years as suggested by the trending line earlier. And the importance of LLMs and their usage as a search engine in daily activities as well as for more complex and sophisticated ones. It is evident that with LLM advent, it is coupled with less online digital content growth. Scarce in digital content implied more usage and employment of LLMs. As those LLMs have been trained on massive data as stated earlier, the RLHF method will be more prominent to the existing models. That is to satisfy user needs which will result in a more precise result and more powerful Artificial intelligence; i.e., more powerful LLMs. Eventually, LLMs, will be more dominant means for information access and response acquiring.

Conflicts Of Interest

The authors declare no conflicts of interest.

Funding

No funding is provided and no financial support is received to carry out the research presented in this paper.

References

[1] T. Teubner, C. M. Flath, C. Weinhardt, W. van der Aalst, and O. Hinz, “Welcome to the Era of ChatGPT et al.: The Prospects of Large Language Models,” Bus. Inf. Syst. Eng., vol. 65, no. 2, pp. 95–101, Apr. 2023, doi: 10.1007/S12599-023-00795-X/METRICS.

[2] Y. Liu et al., “Summary of ChatGPT-Related research and perspective towards the future of large language models,” Meta-Radiology, vol. 1, no. 2, p. 100017, Sep. 2023, doi: 10.1016/J.METRAD.2023.100017.

[3] A. Kolides et al., “Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts,” Simul. Model. Pract. Theory, vol. 126, p. 102754, Jul. 2023, doi: 10.1016/J.SIMPAT.2023.102754.

[4] Z. Liu et al., “Tailoring Large Language Models to Radiology: A Preliminary Approach to LLM Adaptation for a Highly Specialized Domain,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 14348 LNCS, pp. 464–473, 2024, doi: 10.1007/978-3-031-45673-2_46/COVER.

[5] S. Lankford, H. Afli, and A. Way, “adaptMLLM: Fine-Tuning Multilingual Language Models on Low-Resource Languages with Integrated LLM Playgrounds,” Inf. 2023, Vol. 14, Page 638, vol. 14, no. 12, p. 638, Nov. 2023, doi: 10.3390/INFO14120638.

[6] A. Liesenfeld, A. Lopez, and M. Dingemanse, “Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators,” Proc. 5th Int. Conf. Conversational User Interfaces, CUI 2023, Jul. 2023, doi: 10.1145/3571884.3604316.

[7] K. Bhardwaj, R. S. Shah, and S. Varma, “Pre-training LLMs using human-like development data corpus,” Nov. 2023, Accessed: Dec. 10, 2023. [Online]. Available: https://arxiv.org/abs/2311.04666v3.

[8] H. Laurençon et al., “The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset,” Adv. Neural Inf. Process. Syst., vol. 35, pp. 31809–31826, Dec. 2022.

[9] “Wikipedia:Size of Wikipedia - Wikipedia.” https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia (accessed Dec. 10, 2023).

[10] S. Yin et al., “A Survey on Multimodal Large Language Models,” Jun. 2023, Accessed: Dec. 10, 2023. [Online]. Available: http://arxiv.org/abs/2306.13549.

[11] M. Aljanabi, M. Ghazi, A. H. Ali, S. A. Abed, and C. Gpt, “ChatGpt: Open Possibilities,” Iraqi J. Comput. Sci. Math., vol. 4, no. 1, pp. 62–64, Jan. 2023, doi: 10.52866/20IJCSM.2023.01.01.0018.