The Language Model Revolution: LLM and SLM Analysis

2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP) | 979-8-3315-3149-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IDAP64064.2024.10710677

1nd Zeynep Örpek
Research & Development Center, Vakıf Participation
Istanbul, Turkey
zeynep.orpek@vakifkatilim.com.tr

2nd Büşra Tural
Research & Development Center, Vakıf Participation
Istanbul, Turkey
busra.tural@vakifkatilim.com.tr

3nd Zeynep Destan
Research & Development Center, Vakıf Participation
Istanbul, Turkey
gk.zeynep.destan@vakifkatilim.com.tr

***Abstract*—As technology develops day by day, significant developments have been made in the field of artificial intelligence (AI). In particular, machine learning (ML) and deep learning (DL), as the main technologies that form the basis of artificial intelligence, have offered revolutionary innovations and laid the foundation for future technologies. Traditional artificial intelligence models are based on algorithms that show high performance in certain tasks such as classification, scoring, prediction, and pattern recognition. These algorithms are developed to best perform a specific task, making it difficult for artificial intelligence to be sufficiently effective in areas that require flexibility. Generative artificial intelligence, which has become widespread in recent years, has the ability to produce certain types of content in addition to the competencies of traditional artificial intelligence models. This has revolutionized the field of productivity in artificial intelligence. Generative artificial intelligence language models have gone beyond the limitations and started a new era in artificial intelligence applications. Where traditional artificial intelligence models are limited, language models have come into play, especially with their natural language processing (NLP) capabilities. Rather than just analyzing data, language models can learn the rules of the language and provide human-like responses, produce text, and offer a wider range of applications. In this way, artificial intelligence systems have become more flexible, extensible, and dynamic. With the rise of language models in this field, concepts such as large language models (LLM) and small language models (SLM) have emerged. Large language models have come to the fore as systems that can provide deep knowledge and language production on a wide variety of topics by being trained on huge data sets. Large language models such as ChatGPT are one of the most common and impressive examples in this field. However, small language models, which are smaller and specialized language models, have begun to be used as an alternative to large language models in certain areas because they require less data and processing power. Small language models stand out with their lighter but targeted performance, offering effective solutions, especially in situations where there are resource limitations. At this point, using both large and small versions of language models in the right scenarios provides great advantages in terms of sustainability and efficiency. This study aims to reveal the transformative effect of technology on artificial intelligence and the critical role of language models in this process by evaluating language models and the issues to be considered in the selection of these models.**

Keywords—artificial intelligence, large language models, small language models

I. INTRODUCTION

Artificial intelligence models have become widespread as technologies that are based on mathematical calculations and offer specific solutions in the early days. They offer solutions such as value estimation, pattern recognition, and classification with their complex and advanced mathematical calculation tools. Although they are used as a good problem-solving tool, artificial intelligence models generally produce answers in the specific fields or sectors they are trained in. They provide numerical value-based results such as numerical value estimation, object identification, and biometric pattern recognition. With the accelerating progress of machine learning and deep learning, artificial intelligence has gone beyond being a tool for solving difficult problems and has gained a productive feature, especially with the development of artificial neural networks.

Studies conducted in the 1950s to examine the semantic structure of the language have become used in computer science with the Word2Vec library released by Google in 2013. The Word2Vec library enabled the language to be converted into numerical form, thus becoming usable in advanced calculations. The fact that words can be represented in vector space has accelerated studies on natural language processing. Studies on the complexity of natural language and the relationships between words, such as subsequent n-gram calculations, have merged with deep learning and artificial neural networks. With Google's Transformer architecture released in 2017, language models have become able to understand complex relationships in the language structure, such as word affiliation and extracting the meaning that is emphasized.

Unlike traditional artificial intelligence models that focus on solving specific problems, current natural language processing models that offer productive and flexible solutions have become trending technologies in recent years. The development of natural language processing models developed at an academic level has become a widely used and developed technology over time, with the support of open-source platforms. With ChatGPT-3 being made freely available to the public in 2020, people have seen and had the opportunity to experience the benefits of natural language processing models. In addition to large-scale technology companies sharing the models they develop as open source, software developers have been able to develop natural language processing projects with Google providing the necessary hardware support to people via cloud technologies such as Google Collab. As studies and developments in the field of natural language processing accelerate, companies and individual developers have begun to train and use their own natural language models.

II. Related Works

The development of artificial intelligence studies has also led to significant advances in the field of language models. The combination of AI technologies with deep learning, big data processing, and advances in computational power has provided groundbreaking approaches to language modeling. These developments have led to the emergence of large language models and have revolutionized the field of natural language processing.

The foundation of language modeling was laid in the 1948s with studies on information theory. Shannon (1948) took the first steps in the mathematical examination of language with information theory, which forms the basis of language modeling [1]. Statistical language models such as n-gram models were common during this period. These models attempted to approximate the probabilistic nature of language using simple Markov chains. On a small scale, these models were used for basic language processing tasks.

Language modeling has undergone a significant transformation with the development of neural networks in the 1990s. Bengio et al. (2003) were among the first researchers to apply neural networks to language modeling. The method they proposed in this study is a step towards understanding the deeper structure of language by overcoming the limitations of n-gram language models. Thus, success has been achieved for more complex language structures [2].

The Transformer model, developed by Vaswani and his colleagues in 2017, has opened the door to a new era in language modeling and has formed the basis of large language models. Transformer models have enabled the training of much larger and more effective models thanks to efficient parallel computing and processing of large data sets [3].

Transformer models have become more widely used with the BERT (Bidirectional Encoder Representations from Transformers) model developed by Google in 2018 and GPT-2 introduced by OpenAI in 2019. These large language models have been trained with hundreds of billions of parameters, allowing for a better understanding of language and the creation of new texts. GPT models in particular have made significant progress in the field of natural language generation [4].

In recent years, large language models have made significant progress in natural language processing. These models are capable of producing human-like texts, answering questions, and completing other language-related tasks with high accuracy by being trained on large text datasets (Floridi & Chiriatti, 2020). OpenAI's models such as ChatGPT and GPT-4 stand out not only for language processing but also for their ability to perform multi-step reasoning, which can be used to solve general tasks. Large language models have become an important building block for the development of artificial general intelligence [5].

Strubell et al. (2019) show in their study that while addressing the energy costs of large language models, small language models can provide more sustainable and energy-efficient solutions. The study highlights the environmental impacts of large models and that small models may be more advantageous in this context [6].

Radford et al. (2019) focused on the multitasking learning and transfer learning capabilities of large language models in their study introducing the GPT-2 model. The study shows that large models can process more information than small models and exhibit superior performance on a wide range of tasks. However, it also highlights cases where small language models are more efficient due to their lower computational costs [7].

Large language models, such as the Generative Pre-trained Transformer (GPT-3) (Floridi & Chiriatti, 2020), have made significant progress in natural language processing in recent years. These models are trained on large amounts of text data and can produce human-like text, answer questions, and complete other language-related tasks with high accuracy [5].

Brown et al. (2020), in their study on GPT-3, argue that the performance advantage of large language models must be balanced with computational costs. They state that large models perform better on complex tasks, but small language models can be effective with lower resource requirements, especially for tasks given in certain domains [8].

Studies on energy efficiency are increasing to reduce the environmental impact of large language models. Patterson et al. (2021) drew attention with their work proposing the development of energy-efficient language models for sustainable AI. These studies focused on less energy-consuming alternatives to large language models and more efficient training techniques. Additionally, reusable model components and the concept of "Green AI" are prominent [9].

Large language models are known for their high energy consumption during training processes, as they require huge datasets and large computational power. For example, OpenAI's GPT-3 model was trained with a huge amount of energy consumption, and the resulting carbon footprint has caused controversy. Although efforts to increase the energy efficiency of large language models continue, the sustainability of these models is still seen as a major problem [10]. Sustainability comparison between small language models and large language models is addressed through studies on energy consumption, efficiency, environmental impacts, and resource usage.

Overall, large language models will continue to push the boundaries of what is possible in natural language processing. However, there is still much work to be done in terms of addressing their limitations and related ethical considerations [11]. It has been observed that the studies have comprehensively examined the different usage areas, performances, and limitations of large and small language models. It has been revealed that large language models are generally more powerful and versatile, but small language models are more advantageous in terms of energy efficiency and computational costs.

III. Language Models

Language models are artificial intelligence models that mathematically express the structure and usage of a language. These models can be trained with large amounts of data and are capable of understanding and generating texts.

Language models work on a sequence of words. Language models predict the next word by examining the order and co-occurrence of words, sentences, or large text pieces.

The size and structure of the dataset used are as important as the number of parameters used in training the model to determine the capacity of knowledge and skills. With a higher number of parameters, the artificial intelligence model

becomes capable of understanding the complexity of language and sentence structure.

Language models are divided into two: small and large language models, according to the number of parameters. Generally, small language models contain less than 100 million parameters, while large ones contain more than 100 million parameters. The number of parameters in a language model is a factor that directly affects the capacity, learning ability, and performance of the model.

Language models have different sizes of versions, so the number of parameters of the model to be used is an important issue in model selection. Models with a higher number of parameters can work in a wider area, but their computational power and resource requirements are higher. To optimize the performance of the model, careful adjustment of the number of parameters and effective use of these parameters is of great importance.

Pre-trained language models are machine learning or deep learning models trained on a specific task or big data. These models reduce the time and computational power required for training a newly trained model. Pre-trained models are fine-tuned to adapt to specific tasks. In this process, the pre-trained model is additionally trained on a task-specific dataset. The use of pre-trained models has advantages in terms of time, cost, performance, and data requirements.

Commonly used pre-trained small language models are as follows.

- DistilBERT (Generative Pre-trained Transformer 3)
- Phi 2
- Orca 2
- MobileBERT
- BERT Mini, Small, Medium

DistilBERT is a smaller and faster version of the BERT model. It was developed by Google. It was developed by transferring the skills and knowledge of the BERT model to a smaller model. This technique is called knowledge distillation. In this way, the model requires less computational power and runs faster, thus achieving results close to the performance of the BERT model, while working with fewer parameters. DistilBERT is successfully used in many areas such as natural language understanding, text classification, and question answering.

Phi 2 is a small language model developed by Microsoft to run on Transformers-based cloud or local servers. It exhibits high performance in areas such as mathematical reasoning, common sense, language understanding, and logical reasoning. Phi 2 is a resource-efficient model that can perform a variety of tasks.

The MobileBERT language model, which is among the small language models, was developed by Google specifically for devices with limited resources such as mobile devices. It was developed based on BERT. It optimizes performance for devices with limited resources.

Commonly used pre-trained large language models are as follows.

- GPT-3 (Generative Pre-trained Transformer 3)
- BERT (Bidirectional Encoder Representations from Transformers),
- LaMDA (Language Model for Dialogue Applications)
- PaLM (Pathways Language Model)
- LLMA(Large Language Model Meta AI)

GPT-3 is a pre-trained language model developed by OpenAI and trained on a large text dataset with 175 billion parameters. This model has learned the language structure and the relationships between words in training, and based on the Transformer architecture, it shows extremely strong performance, especially in text generation. Thanks to its very large number of parameters, GPT-3 offers high accuracy in complex tasks and superior capabilities in natural language generation.

BERT is a bidirectional pre-trained language model developed by Google and trained on a large text dataset. Based on the Transformer architecture, BERT has parameters ranging from 110 million to 340 million. The model learns language structure and word relationships in a bidirectional manner during training, which allows it to perform well in text comprehension and contextual language tasks. BERT offers high accuracy in natural language processing tasks such as text classification, question-answer systems, and sentiment analysis.

As large language models deliver high success rates, developers have turned to training models tailored to their own needs in this field. However, since training language models requires hardware capable of performing complex and numerous calculations quickly, it consumes a lot of energy. This situation leads to discussions regarding sustainability goals and imposes limitations in terms of cost and time. As a solution, it might be more appropriate for developers to train smaller language models suited to the specific problem they are working on, rather than always using large models. By using smaller language models, the same level of success can be achieved with fewer parameters and less energy-consuming hardware. This way, developers can use the most optimal language model for their needs, avoiding unnecessary capacity and energy consumption.

When small and large language models are compared in terms of training times, it is predicted that the training times of large language models may take months. The fact that big data models work on large data sets with 100 billion parameters increases the dependence on powerful hardware resources. Small language models provide advantages because the training time is shorter depending on the parameter and data set size.

When comparing small and large language models, it has been observed that while large language models have broader capabilities, small language models excel only in the specific areas they are trained in. This is an important issue to be considered in choosing the appropriate model according to the requirements analysis in projects.

Large language models have gained popularity and prominence due to their impressive capabilities in natural language processing. However, they have disadvantages in terms of applicability and the data and resource requirements they use. This has caused and will continue to cause advantageous groups with access to resources to follow and adapt to technology closely, while disadvantaged groups with

more limited access to resources will quickly fall behind despite the rapid advancement of technology.

Language model development and fine-tuning studies require powerful graphics cards and machines in terms of hardware. The cost of improving the models that have been trained and using them is also increasing day by day. ChatGPT consumes more than half a million kilowatts of electricity every day; this amount is enough to meet about two hundred million requests. ChatGPT's daily energy consumption is almost equivalent to 180,000 U.S. households, each using approximately 29 kilowatts. About fifty centiliters of water are used for one ChatGPT query [12].

The Paris Agreement is an agreement signed in 2015 and entered into force in 2016, within the scope of the United Nations Framework Convention on Climate Change, on the mitigation, adaptation and financing of climate change. As of March 2021, 197 members of the United Nations Framework Convention on Climate Change are parties to this agreement [13]. According to this agreement, reducing the energy consumption of language models is critical to achieving the goals of the Paris Agreement. Large language models consume large amounts of energy, especially during large-scale training processes. Increasing the energy efficiency of these models, reducing their environmental impact, and minimizing carbon emissions will contribute to efforts to achieve global climate goals. Model optimization, hardware improvements, model selection, model training process improvements, and energy resource management improvements can significantly reduce the energy consumption of large language models and thus contribute to efforts to combat climate change. This is one of the important steps towards achieving the sustainability goals of the Paris Agreement.

IV. Result

Artificial intelligence is considered a technology that leads to technological and social advances in many fields. By imitating human intelligence, artificial intelligence can perform specified tasks much faster than humans can perform.

One of the most widely used subfields of artificial intelligence today is natural language processing. Language models have provided a major advancement in the field of natural language processing. These models allow for a better understanding of human language and for humans to interact with computers in a more natural way.

Although large language models have made great progress in recent years, there are still many disadvantages that need to be addressed. While small language models operate with lower energy consumption and cost, large language models offer superior performance and wider application areas but increase environmental impacts. Making large language models more efficient in terms of sustainability is a critical issue in artificial intelligence research. Large language models have great potential in natural language processing, but these models need to be developed considering ethical responsibilities and limitations.

When choosing a language model, the model selection should be made taking into account the project requirements. Trending language models may have a wide user base, but may not be the right solution for every project. The language model should be selected by performing project requirements analysis. With this approach, long-term, successful, and effective models can be developed.

References

[1] C. E. Shannon, "A mathematical theory of communication.," *The Bell System Technical Journal,* pp. 27(3), 379-423., 1948.

[2] R. D. P. V. a. C. J. Y. Bengio, "A neural probabilistic language model,," *JOURNAL OF MACHINE LEARNING,* vol. vol. 3, p. 1137–1155, 2003.

[3] A. S. N. P. N. U. J. J. L. A. K. Ł. P. I. Vaswani, "Attention is all you need. Advances in Neural Information Processing Systems," p. 5999–6009, 2017.

[4] J. C. M.-W. L. K. T. K. Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint ,* p. arXiv: 1810.04805, 2018.

[5] L. &. C. M. Floridi, "GPT-3: Its Nature, Scope, Limits, and Consequences," *Minds and Machines,* no. doi: 10.1007/s11023-020-09548-1., pp. 30(4), 681-694., 2020.

[6] E. G. A. &. M. A. Strubell, "Energy and policy considerations for deep learning in NLP.," p. arXiv preprint arXiv:1906.02243, 2019.

[7] A. W. J. C. R. L. D. A. D. &. S. I. Radford, "Language Models are Unsupervised Multitask Learners. OpenAI," no. doi: 10.48550/arXiv.1909.01493., 2019.

[8] T. B. M. B. R. N. S. M. K. J. D. P. .. &. A. S. Brown, "Language models are few-shot learners.," *arXiv preprint,* p. arXiv:2005.14165, 2020.

[9] D. G. J. L. Q. V. L. C. M. L. M. R. D. S. D. T. M. &. D. J. Patterson, "Carbon Emissions and Large Neural Network Training," *arXiv preprint,* no. doi: 10.48550/arXiv.2104.10350., 2021.

[10] E. M. G. T. M.-M. A. &. S. S. Bender, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?.," *In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* no. doi: 10.1145/34, pp. 610-623, 2021.

[11] K. S. S. K. M. B. D. D. F. F. U. G. G. G. S. G. E. H. S. K. G. K. T. M. C. N. J. P. O. Enkelejda Kasneci, "ChatGPT for good? On opportunities and challenges of large language models for education,Learning and Individual Differences,," vol. Volume 103, no. https://doi.org/10.1016/j.lindif.2023.102274., pp. ISSN 1041-6080,, 2023.

[12] [Online]. Available: https://www.forbes.com/sites/cindygordon/2024/03/12/chatgpt-and-generative-ai-innovations-are-creating-sustainability-havoc/.

[13] [Online]. Available: https://enerji.gov.tr/evced-cevre-ve-iklim-paris-anlasmasi.