computer sciences and mathematics forum

MDPI

Proceeding Paper

Language Models for Everyone—Responsible and Transparent Development of Open Large Language Models †

Daniel Gillblad [1,2]

[1] Computer Science and Engineering Department, Chalmers Technical University, 412 96 Gothenburg, Sweden; daniel.gillblad@chalmers.se or daniel.gillblad@ai.se
[2] AI Sweden, 402 78 Gothenburg, Sweden
† Presented at the Workshop on AI and People, IS4SI Summit 2023, Beijing, China, 14–16 August 2023.

Abstract: Large language and multimodal models are revolutionising many aspects of human work and creativity, with broad potential not only as chatbots and for information retrieval but as interaction points and integrators of large technical systems. However, these technologies' significance comes with societal challenges regarding accessibility, applicability, alignment, and inclusion. The development of open-source large language models may lead to more transparent and responsible use of such technologies. Here, we discuss the importance of open language models, their challenges, and how they were managed in a specific use case.

Keywords: large language models; open source; AI infrastructure

Citation: Gillblad, D. Language Models for Everyone—Responsible and Transparent Development of Open Large Language Models. *Comput. Sci. Math. Forum* **2023**, *8*, 51. https://doi.org/10.3390/cmsf2023008051

Academic Editors: Zhongzhi Shi and Wolfgang Hofkirchner

Published: 7 September 2023

1. Introduction

Large language and multimodal models have developed at a staggering pace during the last few years. We are now at a stage where generative models can truly assist and, in some parts, replace human work and creativity. Image models can generate visual representations that can easily be taken for natural or human creation, and large language models can generate text on relatively complex subjects in several different styles. As these models grow more capable and cover more modalities such as video, music, programming, and engineering, we will have to account for new ways of working with creative disciplines and a new reality where text and representations that earlier could only be attributed to human effort could have been generated by a machine with its specific limitations and representation of the world.

Driven by private actors, the development of these types of models is currently happening largely in the US. Large technology companies are racing to compete in developing new, more powerful and versatile versions of LLMs. Just during the beginning of this year, Meta reported their LLaMA model, Google announced PaLM-E, Baidu introduced their LLM-based Chatbot ERNIE, OpenAI revealed GPT-4, and GitHub announced Co-pilot X, which adopts GPT-4 and Chatbot features to support developers. This is still a small selection of important releases and developments.

Given the importance of these technologies, it is clear that this momentum could come with challenges for society around areas such as access, applicability, alignment and inclusion. Collaborative efforts to develop open large language models such as GPT-SW3 (GPT-SW3) have proven that there are alternative paths to development, paths that could lead to more transparent and responsible use of these technologies. Here, we will reflect on the potential importance of open LLMs, ways to drive their development, and the practical use case of GPT-SW3.

1.1. Emerging Directions for AI

During the last few years, it has become increasingly clear that the transformer architecture for machine learning provides a very general, trainable, general-purpose computing

architecture suitable for a large range of tasks [1]. Keeping this basic architecture largely unchanged, increasingly capable models with zero-shot learning capacity, models capable of performing tasks not explicitly trained for within natural language [2], computer vision [3], and more general multimodal settings [4] are being developed.

In particular, large language models (LLMs) have reached capacities to reliably work as true assistants and general information retrieval systems for a very wide range of domains [5]. Beyond this, through, for example, plugins and vector stores, these models can integrate and interact with external tools and knowledge. In essence, LLMs can today be used both as efficient interfaces for broad interaction with humans and technical systems and increasingly as system integrators, tying knowledge bases, computational systems, and other machine learning models together in a manner dependent on tasks set in a natural dialogue with humans. Further, by connecting potentially several LLMs with external tools and actions, increasingly sophisticated levels of planning and automation can be achieved.

While very large language models can require large amounts of data and compute resources to train from scratch, LLMs are surprisingly scalable and efficient for inference and, in practice, are very efficiently fine-tuned to new data and domains [6]. Such fine-tuned models show significant promise in multimodal scientific domains trained with limited computational resources, and projects are saving time and resources using small, highly curated datasets with good results [7,8].

We are at a stage where transformer-based architectures, LLMs, and multimodal models not only potentially can act as chatbots and image generators but serve as natural language interfaces, system integrators of knowledge, and models. We have to assume that these models, while continuously evolving, will be adapted and combined into more specialised applications and reused for a huge number of tasks across disciplines and organisations. They will serve as a new infrastructure for AI and intelligent services and a fundamental resource for further development.

1.2. A New Digital Infrastructure

AI is to a large degree developed based on shared resources and shared AI artefacts. While relevant and high-quality data sets remain key for AI research and development, access to artefacts such as meta-data, knowledge, models, applications, algorithm implementations, and benchmarks are similarly critical components. As new applications are increasingly built on already existing models and components, including very large foundational models, we are rapidly moving towards a new type of AI infrastructure and ecosystem with an increased focus on artefact reuse and inference services. There are very strong dependencies between all types of AI artefacts, such as computational resources, data access, inference services, and meta-data, and without open access and transparency, the responsible development of AI applications will be difficult. The development of open, transparent, and accessible large language models and re-usable datasets for training and evaluation are likely key to accelerating the use of these technologies.

Now, the development of open models does come with challenges. Training large language models requires substantial computational power. A typical LLM will need to be trained over weeks or months on hundreds of high-end GPUs, which are both expensive and energy intensive. Still, with the availability of large research compute clusters, this is not an insurmountable hurdle, although the allocation of large parts of these clusters for extended times might be a challenge.

Perhaps more importantly, these models are trained on large amounts of data. The collection of representative data is a time-consuming process that needs to be fully transparent in terms of where data are sourced, handled, and transformed. Some data sources may need to be purchased, sometimes complicating open, decentralised development. Finally, creating data for the instruction tuning necessary to create efficient dialogue systems can be highly labour intensive, creating challenges for efforts without strong central backing. Thus, the creation of and access to the datasets that will serve as the foundation of the

development of several generations of future models is absolutely critical for the open development of LLMs.

While data collection and training may be solvable problems for open initiatives, deploying models to serve predictions (i.e., generating text) in a real-world setting introduces more challenges. Although some organisations have access to the necessary computing resources to run these models themselves, giving broader access involves access to serving infrastructures that can handle many requests in parallel with low latency and continuous availability. Unlike infrastructure for training the models, these resources are largely out of reach for small, open research initiatives.

Hardware and software improvements are, however, rapidly making these infrastructures more feasible. More powerful and efficient GPUs and specialised hardware, better support for distributed computing in the common computational frameworks such as TensorFlow and PyTorch, and improved usability of infrastructure will all lower costs for both training and inference, increasing the number of feasible use cases for the models.

2. The Value and Risks of Open Models

In general, the value of open large language models lies in their accessibility, transparency, and potential for application and innovation. Let us have a look at each of these benefits while also discussing some of the challenges associated with open models.

First, open-source models are available to everyone, from independent users and researchers to large organisations, allowing everyone to build applications or specialised models for specific use cases or internal use. This broad access can enable a wide variety of use cases, particularly within sectors that manage sensitive requests, such as education, healthcare, and government services, or must incorporate internal data or models that constitute a competitive advantage, such as pharmaceutical industries.

Second, open-source models offer transparency that can be critical not only for understanding how these models work, their limitations, and potential biases but also a necessary requirement for use in healthcare and government. In several sectors, transparency will be necessary for application developers to ensure the ethical use of the technology and to provide a basis for trust among users and developers.

Third, open-source models encourage collaboration across innovation communities and organisations. Managed correctly, this could allow for the collective improvement of the models, not only from different teams being able to contribute improvements or adaptations back to the original model but also to create feedback beyond benchmarks from a large number of applications.

However, there are also several challenges and considerations associated with fully open large language models that must be managed. Without control over usage, open-source models can be misused, for example, to generate or to automate the creation of misinformation or harmful content. Like all large language models, open-source models can perpetuate biases present in training data. While the transparency of open models allows anyone to identify, study and understand these biases, it does not automatically solve the problem. There could be concerns about data privacy and the potential for leakage of sensitive information, even though efforts are typically made to exclude such data from training sets. Fully open models remove any opportunity to control the leaking of sensitive information at the inference stage, e.g., through request limitations.

Thus, while the potential benefits of fully open large language models are substantial, these challenges need to be carefully considered and managed.

3. Responsible and Transparent Development of Open Language Models—A Brief Case Study

As a brief case study in the development and release of an open LLM, let us discuss the GPT-SW3 model [9]. The GPT-SW3 initiative develops a large GPT model for Swedish and Nordic languages. It is driven by AI Sweden (the national centre for applied AI) together with RISE (the Research Institutes of Sweden) and WASP (the Wallenberg Autonomous

Systems and AI Program), all organisations promoting open research and the use of AI. The motivation for developing GPT-SW3 is essentially to create a foundational resource for AI in Sweden. The aim is to carry this out as openly, transparently, and collaboratively as possible, with the goal of making the model available to all sectors in Sweden that may have a need for this type of AI solution. The core team developing the model consists only of a handful of people in different organisations but is mainly located in AI Sweden.

As discussed earlier, developing and sharing the models comes with a number of challenges. Let us discuss these challenges and how to mitigate them during (1) development and (2) release through responsible data collection and management, staged releases incorporating feedback, and the creation and management of an open community.

To ensure broad representation and applicability, GPT-SW3 aims to use training data that reflect all dialects and demographics. With no directly available large datasets available, the project created the Nordic Pile, a 1.2 TB dataset based primarily on existing data sources [10]. Too large for manual inspection, programmatic rules were used for quality inspection, and the development team weighed risks against benefits before the inclusion of each data set. Personal information was filtered, and sensitive sources were excluded.

The result is a curated dataset which has gone through reasonable efforts to ensure quality and privacy, and while significant effort had to go into its creation, the result is a resource that can be continued to be used for new generations of models. Utilising this data, models using an architecture similar to GPT-3 have been trained in a number of sizes up to 40 billion parameters.

To balance the benefits of openness with the potential risks, the GPT-SW3 project adopted a "staged release" approach to its models, allowing for time to assess usability, impact, and risks before making the models fully open. In the pre-release, models were shared with people and organisations that commit to not using the models in ways that may cause harm and that are committed to sharing the learnings of their research and applications. To access the models, an application form had to be filled in stating affiliation, research, application goals, etc. Several hundred participants were given access to the models of which some shared practical feedback on performance, toxicity, etc. Based on this feedback, GPT-SW3 will likely go into a fully open release soon.

While detailed data collection and evaluation of applications for model access takes resources, this approach has provided much larger certainty that the model can be reliably released openly and can be deemed a usable approach to the open development of large language models.

4. Conclusions

The GPT-SW3 case shows that a very small, collaborative effort can create useful large language models with a focus on representation and transparency. The initial approach of focusing on the creation of artefacts, i.e., the datasets and models themselves, rather than solving specific research issues or potential problems at the start, proved very successful. Along the way, access to computational resources, development of suitable licenses, principles for sharing in the initial release and much more had to be solved, but the specific issues were not necessarily known in detail at the start, and the focus on development was critical. While not fully open source at the time of writing, the direction towards open source and engaging a larger community worked but required strong leadership and community management provided by the AI Sweden Natural Language Understanding team.

Given that the models are already used to evaluate, for example, use cases within Swedish healthcare, applications that would have been difficult to legally test without the opportunity to run an open model on premises, the value is apparent. The careful data collection and preservation of data provenance and privacy have so far proved successful in creating a transparent model that can be released fully open, the staged release with relevant feedback showing that the model is usable without apparent privacy issues, and the open community a way to crowdsource these efforts. This approach should be replicable for other and future efforts.

While it is probably in the domain of very large, private models built on very large engineering efforts to provide broad chatbots and services, open and adaptable models are an important part of the future AI ecosystem for many applications, and with the right approach useful, transparent models can be developed by small teams in collaboration.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The author would like to acknowledge the excellent work conducted in the development of the GPT-SW3 model, particularly the current and former members of the Natural Language Understanding group at AI Sweden.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
2. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
3. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. *arXiv* **2023**, arXiv:2304.02643.
4. Reed, S.; Zolna, K.; Parisotto, E.; Colmenarejo, S.G.; Novikov, A.; Barth-Maron, G.; Gimenez, M.; Sulsky, Y.; Kay, J.; Springenberg, J.T.; et al. A Generalist Agent. *Trans. Mach. Learn. Res.* 2022, 1–42, *in review*.
5. Open, A.I. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
6. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *arXiv* **2021**, arXiv:2106.09685.
7. Geng, X.; Gudibande, A.; Liu, H.; Wallace, E.; Abbeel, P.; Levine, S.; Song, D. Koala: A Dialogue Model for Academic Research. *Blog Post* **2023**. Available online: https://bair.berkeley.edu/blog/2023/04/03/koala/ (accessed on 20 July 2023).
8. Zhang, R.; Han, J.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; Gao, P.; Qiao, Y. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv* **2023**, arXiv:2303.16199.
9. Sahlgren, M. What is GPT-SW3? 2022. Available online: https://medium.com/ai-sweden/what-is-gpt-sw3-5ca45e65c10 (accessed on 20 July 2023).
10. Öhman, J.; Verlinden, S.; Ekgren, A.; Gyllensten, A.C.; Isbister, T.; Gogoulou, E.; Carlsson, F.; Sahlgren, M. The Nordic Pile: A 1.2TB Nordic Dataset for Language Modeling. *arXiv* **2023**, arXiv:2303.17183.