https://doi.org/10.1038/s42256-023-00695-5

Foundation models and the privatization of public knowledge

Fabian Ferrari, José van Dijck & Antal van den Bosch

Check for updates

To protect the integrity of knowledge production, the training procedures of foundation models such as GPT-4 need to be made accessible to regulators and researchers. Foundation models must become open and public, and those are not the same thing.

Around the world, universities and schools have raised concerns about ChatGPT, which is owned by OpenAI. Thorny issues include safety, plagiarism, bias and accuracy. Many discussions focus on how AI applications powered by large language models (LLMs) owned by proprietary actors may transform the production of knowledge[1]. OpenAI, in spite of its name (and non-profit heritage), has been partly converted into a capped-profit company (Limited Partnership) in which Microsoft is the biggest investor. What remains under-reported is the question of whether regulators and scientists will get access to the inner workings of deep neural networks—how they were trained and on which datasets. If these models are transformed into closed products, they will be unavailable for thorough inspection, replication and testing; however, such entry points are essential to guarantee the integrity of public knowledge in democracies.

To understand how LLMs can be opened up to public scrutiny, we need to consider how they work. Foundation models that can be adapted for a variety of downstream tasks[2], such as the Generative Pre-Trained Transformer (GPT), are trained on open data and data licensed from third-party providers. OpenAI's GPT-3.5 model – the predecessor of GPT-4 – was trained on 45 terabytes of text data. This translates to approximately 300 billion words extracted from sources such as Wikipedia, CommonCrawl and GitHub. Overall, the training procedure involved the learning of 175 billion parameters. ChatGPT relies on GPT-4 as well as on a specific Reinforcement Learning from Human Feedback (RLHF) component. RLHF aims to optimize the chat functionality of ChatGPT based on user interactions. Each time we use ChatGPT, we contribute to the optimization of the chatbot for future conversations. But details about the GPT-4 training datasets that its developers scraped from the internet to train the model for conversational use were not revealed by OpenAI. Authors of texts and copyright holders whose content becomes training data for the chatbot are unaware that their work is being transformed into the core component of a proprietary product[3].

It is pivotal to understand how foundation models underpin domain-specific AI products. In April 2023, Google announced Med-PaLM 2, a chatbot designed to respond to medical questions. Med-PaLM is built on top of Google's foundation Pathways Language Model (PaLM). As with the GPT-4 model that underlies ChatGPT, we know little about the datasets and training procedures that went into PaLM, except that the resulting neural architecture holds the staggering amount of 540 billion trained parameters[4]. But unlike for OpenAI, Med-PaLM's creators provide information about the pools of public knowledge they tapped into when fine-tuning the chatbot for medical questions. The datasets used to train and improve Med-PaLM can be divided into two broad categories[5]. The first type of data was extracted from more than 200,000 question–answer sets from medical exams: the collective labour of students, educators, administrators and researchers from many countries over the course of many decades. The second category consists of consumer health questions. These include questions submitted to the US National Library of Medicine and reference answers provided by the US National Institutes of Health – institutions that are funded by taxpayer money and therefore legally obligated to be open to the public in their provision of trustworthy medical knowledge. According to Google, physicians judged Med-PaLM 2's answers to 'better reflect medical consensus' 72.9% of the time compared to physician answers[6].

Open and public knowledge

Although we can trace various types of training data that were used to fine-tune Med-PaLM, other parts of the calibration process remain out of sight for researchers and legislators. Two key problems arise from this opacity: first, foundation models such as PaLM are trained on unknown collections of datasets using unknown training procedures; second, the models behind the chat functionality are rendered inaccessible to public scrutiny by their proprietors, who protect their training procedures as intellectual property.

Foundation models must become open and public, and that is not the same thing. 'Open' means that foundation models are freely available for thorough inspection at the most detailed level of saved network parameters and hyperparameters, and for replication. 'Public' means that they should be treated as public utilities. Although the investments required to build these models, as well as safety and competition concerns, explain their closed nature, it is crucial that this embargo is lifted to allow public scrutiny now and public access in the near future. But because open-sourced models can be misused as generators of harmful content, as the case of Stanford's Alpaca model shows[7], openness is not a panacea. Rather, the dangers of open LLMs in fuelling misinformation underline the need for systematic audit mechanisms.

With regard to the first argument: for foundation models to be open to deep inspection, replicability and the much-needed debiasing research, independent researchers need to be able to access them. Given the complexity of data access provisions, illustrated by the EU's Digital Services Act[8], legislative precision about what layers of information need to be made accessible is paramount. We argue that access to five layers of information is necessary to check whether parts of the procedure may be replicated, to assess the integrity of results, to perform fine-tuning and to allow research into debiasing directly on these models:

(1) the composition of the dataset on which foundation models are trained — known only broadly for GPT and PaLM, and with some more detail for Meta's LLaMa;
(2) the composition of domain-specific datasets with which AI applications such as ChatGPT or Med-PaLM are fine-tuned — unknown for ChatGPT, known for Med-PaLM;
(3) the neural network architectures (defining the connectivity of network elements) of all components — known only broadly for GPT and PaLM, and made available for LLaMa by Meta;
(4) the trained models (that is, all trained parameters) — made available by Meta with LLaMa, but unavailable for the others mentioned; and
(5) all scripts and procedures needed for training, data processing and output generation — all mentioned parties provide a selection of tools that allow the use of pre-trained models for further development, but not for training.

The second argument pertains to the notion of 'public' knowledge. If texts or conversations processed by chatbots — trained on open datasets — result in the construction of common knowledge, the foundation models that are trained on these open datasets should be public utilities as well. For this argument to be made, there needs to be a clear distinction between foundation models and AI applications built on top of them. Just as water treatment plants, power generators and sewage systems have infrastructural functions for businesses, schools and households, foundation models are touted as utilities for all AI-driven sectors. Although a handful of American technology companies own and control the lion's share of state-of-the-art models, there are also examples of LLMs as public utilities, such as BLOOM[9].

LLMs are not the only technologies that raise the key question of how to keep AI-processed knowledge open and public, and how to prevent public infrastructures from ending up as proprietary assets in the hands of private companies. In 2021, DeepMind released AlphaFold 2, a computational model that predicts 3D protein structures from amino acid sequences. In *Nature*, this was seen as a major step in tackling the protein folding problem, which has significant ramifications for drug development. As training data for the model, DeepMind downloaded 170,000 protein structures from the Protein Data Bank, a repository of biological molecules collected by research groups around the world. DeepMind would not have been able to create AlphaFold 2 without access to 'a large body of research paid almost exclusively by the taxpayers'[10].

As British-Italian economist Mariana Mazzucato argues, corporate success is often the result of state-funded investments in education and research[11]. The need for proprietary models such as GPT-4, PaLM and AlphaFold 2 to help achieve scientific breakthroughs reveals a poignant paradox. On the one hand, technology companies such as Google invest large sums of money to accelerate the privatization of public tasks in domains like education and health care. On the other hand, their infrastructural technologies underpinning this expansion, such as foundation models, would simply not exist without the numerous sets of open data to which masses of researchers have contributed both individually and collectively.

Foundation models as open and public utilities

Therefore, we need to address a complex question: how can the 'openness' of foundation models be safeguarded and their 'publicness' be incentivized? Our proposed remedies for this question are technical and political, as well as legal and educational.

As technical safeguarding measures, licensed outside experts ought to be commissioned for model auditing to confirm which sources are used for training, as well as for detecting bias or hateful content; and these teams should also be licensed to technically modify aspects of LLMs. For model auditing, it is possible to apply *training data extraction attacks* to a foundation model[12]. Such technology may help to reconstruct snippets of training data and link them to their origins, which for GPT-2 have been shown to be large portions of texts, sometimes occurring only once in the training data. As these types of extraction attacks indicate, pre-trained foundation models do seem to rely to a great extent on their capacity to memorize training data, likely more so as they scale up in size[13].

A second technique, 'watermarking', provides a framework to safeguard the transformation of closed models into open models, especially in regard to safety concerns. Watermarking AI-generated texts imposes relatively strong constraints on the generated texts that may not be visible to the human eye but are, to a certain extent, amenable to detection, for example through the detection of manipulated information entropy in sentences[14]. Releasing the watermarking key to allow everyone to check the provenance of a text is currently a prerogative of the model developers; if they have implemented watermarking, they can choose not to release the associated key, and they can change the key at any time. Watermarking should instead help to enforce democratic oversight procedures aimed at the thorough inspection of AI-generated texts. But a 'watermark for public knowledge' remains unlikely as long as dominant model providers keep detailed information about model design, training datasets and procedures hidden from the public's eye. Therefore, this technological solution could be combined with a legislative measure to mandate the publicness of high-profile foundation models.

Beyond technological and political remedies, there are legal and regulatory antidotes to the privatization of foundation models, exemplified by digital policy in the EU. EU policymakers added transparency obligations for foundation model providers in the Artificial Intelligence Act, as well as the need to ensure adequate safety safeguards in model design. As a means to regulate deep monitoring, the AI Act is the first framework for enforcing 'AI transparency'. But regulators need to specify with the utmost clarity which technical details need to be made accessible by foundation model providers. The five layers of information explicated above represent a pathway toward achieving this precision. Mandatory audits of training datasets and model architectures by licensed experts, as well as watermarks for AI-generated text, could become legal requirements. The EU Data Act and the Data Governance Act are also strongly committed to setting standards for transparent data use. They distinguish between the data of public organizations, companies and private individuals, which is an important distinction in the context of foundation models and the provenance of training data. Additionally, legal disputes such as *Getty Images v. Stability AI*[15] will set vital precedents about copyright and ownership questions regarding datasets used for training and fine-tuning LLMs.

Finally, the most important long-term remedy may well be an educational one: improving AI literacy amongst citizens. Education currently focuses more on optimizing the results of chatbots than on teaching students how to think critically about them as instruments of knowledge-making. Users must learn how chatbots are powered by neural networks and what constitutes their training — mechanisms that demand critical dissection and inspection. We should consider AI-driven applications as part of an open, democratic society in which foundation models are links in the knowledge supply chain that people and machines develop together. Open knowledge thrives best as a public good.

Fabian Ferrari ✉, José van Dijck & Antal van den Bosch
Utrecht University, Utrecht, The Netherlands.
✉e-mail: f.l.ferrari@uu.nl

Published online: 28 July 2023

References

1. van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R. & Bockting, C. L. *Nature* **614**, 224–226 (2023).
2. Bommasani, R. et al. Preprint at https://arxiv.org/abs/2108.07258 (2021).
3. Schaul, K., Chen, S. & Tiku, N. *The Washington Post* https://go.nature.com/46LL7rA (19 April 2023).
4. Chowdhery, A. et al. Preprint at https://arxiv.org/abs/2204.02311 (2022).
5. Singhal, K. et al. Preprint at https://arxiv.org/abs/2212.13138 (2022).
6. Singhal, K. et al. Preprint at https://arxiv.org/abs/2305.09617 (2023).
7. Quach, K. The Register https://go.nature.com/3OcDbZe (21 March 2023).
8. Albert, J. *AlgorithmWatch* https://algorithmwatch.org/en/dsa-data-access-explained/ (07 December 2022).
9. Scao, T. S. et al. Preprint at *arXiv* https://arxiv.org/abs/2211.05100 (2022).
10. Outeiral, C. *Oxford Protein Informatics Group* https://go.nature.com/3PV2osA (19 July 2021).
11. Mazucato, M. *The Entrepreneurial State* (Anthem Press, 2013).
12. Carlini, N. et al. *Proc. 30th USENIX Security Symposiu*, 2633–2650 (2021).
13. Carlini, N. et al. Preprint at https://arxiv.org/abs/2202.07646 (2022).
14. Kirchenbauer, J. et al. Preprint at https://arxiv.org/abs/2301.10226 (2023).
15. Barr, K. *Gizmodo* https://go.nature.com/44niobb (6 February 2023).

Acknowledgements

The authors thank J. Zuidema and F. McKelvey for their feedback and suggestions, which have improved this commentary. The authors were supported by the Spinoza program of the Dutch Research Council (NWO).

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Machine Intelligence* thanks Natali Helberger and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.