https://doi.org/10.1038/s43588-023-00585-1

Using proprietary language models in academic research requires explicit justification

Check for updates

Calls for scientists to develop and use open AI systems are growing – especially for language models (LMs)[1,2]. Beyond concerns about reproducibility of results from closed models, being able to audit the data being used by the system helps researchers understand its behavior. Yet despite these appeals, researchers continue to use closed technologies in many disciplines[3–5]. If – as many believe – open systems are preferable, this is dispiriting. Admonishing scientists not to use closed models is unlikely to be immediately successful. Here we survey reasons why proprietary models continue to be popular, and suggest how we as a scientific community can increase uptake of open technologies. Our proposal is simple and low cost: we ask that scientists explicitly justify their use of proprietary models when they employ them in research.

Following Rogers et al.[2], we define 'open' as models that can be downloaded, run offline and shared; moreover, versioning is possible and we know what data the model was trained on – even if that data is not available for direct inspection. They may or may not permit local adjustment of weights. Any model that is not open is closed or proprietary. By "justify their use of proprietary models" we mean explain why using a closed system was preferable over an open system for this particular application. Given that proprietary systems come with predictable costs, we want authors to delineate why the benefits outweigh those costs in the current use-case.

Why is this justification process helpful for science? First, introspection and explanation may encourage scientists to alter their future behavior and improve norms. In some cases, they may come to see their purported reasons as something more like 'excuses'. Second, explicit justification helps us informally 'version' closed models by documenting what they were capable of on a given task at a given time. Third, by articulating where scientists deem closed systems to be currently superior, we all learn where the open developer community should focus attention on improving their product.

Potential justifications

We can think of six circumstances where analysts believe the use of proprietary LMs is acceptable or even preferred. These are not mutually exclusive nor jointly exhaustive, but they deserve respect as separate 'mainstream' arguments.

(1) Object of study per se. Some models, products or algorithms are of such central importance to society that they ought to be investigated in and of themselves. Failing to study how, for instance, ChatGPT 'behaves' leaves us unable to understand why it provokes such interest in the media or how experience of that model might affect legislators designing regulation. In this sense, studying proprietary LMs is akin to investigating the algorithms undergirding Facebook's timelines or YouTube's video recommendations. That is, we believe that the effects of such closed technologies are so profound that we must investigate them directly as part of the broader mission of social science.

(2) Community of interest. Even if the model is not central to society as a whole, it may nonetheless be of particular relevance to a group we wish to study. For example, in 2021 the *BuzzFeed* chief executive announced a partnership of the news organization with OpenAI. To the extent that journalists are of interest sociologically, we may want to understand how LMs are being implemented and the potential consequences for journalistic work. Similarly, we may want to understand how generative text-to-image models like DALL-E could affect artists, or to see how AI chat apps will change the roles of customer service professionals.

(3) Technical state of the art. Perhaps the most common (implicit) argument for closed LMs in computer science work is that they are state of the art in a performance sense. That is, a proprietary model is able to do some key tasks – like classify a text – better, measured by some objective criterion, than any other model (open or closed). Consequently, researchers might argue that excluding that closed LM from use understates what is actually possible on important problems, and potentially misrepresents the current frontier of science. Ultimately, exclusion might create perverse results: if the performance differences are sufficiently large, other scholars will misperceive the 'true' benchmark for contribution, and build inferior systems than in a counter-factual world where they had access to a preferred technology.

(4) Ethical edge. A demonstration of some use of an open model may inspire or enable others to apply the model to a nefarious end that is, in a technical sense, similar. By design, there is no authority or firm to regulate an open model once released, so a researcher might prefer to use a closed model, whose proprietors could, theoretically, curb unintended, dangerous applications. This may be even more important for intermediate products, such as models released openly without guardrails.

(5) Reproducibility and ease of use. Typical arguments against closed LMs rely heavily on reproducibility, or lack thereof. For example, companies can alter (even abandon and make unavailable) their algorithms at any time. This should be less of an issue for open systems. Still, maintainers do make changes to open LMs too – including re-estimating weights, or no longer maintaining extant models. This could affect the reproduction of earlier findings. At such times open model results might be less stable than closed ones, if best practices of open development (including versioning and documenting) are not followed. Related, usability may be a relative problem for open efforts. That is, if the relevant research community finds it prohibitively costly to use an open LM, one may prefer a closed LM.

(6) Downstream use only. Often the LM itself – open or closed – is not under study, and is being used as an intermediary for some other purpose. Furthermore, any data being produced by the LM is available for others to inspect, even if they cannot produce it themselves. For example, suppose a researcher used an LM to simply summarize a set of extant public instructions to individuals in an experiment. Here the use of the LM is incidental to the research question, so it arguably makes little difference whether that system is open or closed — it has no direct consequences for reproducibility of the core research.

Recommendations

What should scientists working with LMs do, given our call above? Individuals ideally would not use closed LMs. If they do, they should be explicit about why they use them, giving as many details as possible. Preferably they would clarify what open models they compared their results to. They should describe what open LM developers might focus on in the next iteration of their designs. Institutions, such as conferences and journals, can help by encouraging or even requiring these steps for contributing authors.

Alexis Palmer[1], **Noah A. Smith**[2] **& Arthur Spirling**[3] ✉
[1]New York University, New York, NY, USA. [2]University of Washington and Allen Institute for AI, Seattle, WA, USA. [3]Princeton University, Princeton, NJ, USA.
✉e-mail: arthur.spirling@princeton.edu

Published online: 21 December 2023

References

1. Spirling, A. *Nature* **616**, 413 (2023).
2. Rogers, A. Closed AI models make bad baselines. *Hacking Semantics* (3 April 2023).
3. Gilardi, F., Alizadeh, M. & Kubli, M. *PNAS* **120**, e2305016120 (2023).
4. Yang, C. et al. Preprint at https://arxiv.org/abs/2309.03409 (2023).
5. Yiu, E., Kosoy, E. & Gopnik, A. *Perspect. Psychol. Sci.* https://doi.org/10.1177/17456916231201401 (2023).

Competing interests

The authors declare no competing interests.