The dangers of using proprietary LLMs for research

Étienne Ollion, Rubing Shen, Ana Macanovic, Arnault Chatelain

The Dangers of Using Proprietary LLMs for Research

Étienne OLLION (corresponding author: etienne.ollion@polytechnique.edu)

- ORCID: 0000-0003-3099-5240
- Centre de Recherche en Économie et de Statistiques (CREST), CNRS, École polytechnique, GENES, ENSAE Paris, Institut Polytechnique de Paris, Palaiseau, France

Rubing SHEN

- ORCID: 0000-0002-5504-6108
- Médialab, Sciences Po, Paris, France
- Centre de Recherche en Économie et de Statistiques (CREST), CNRS, École polytechnique, GENES, ENSAE Paris, Institut Polytechnique de Paris, Palaiseau, France

Ana MACANOVIC

- ORCID: 0000-0003-0800-5271
- Department of Sociology, Utrecht University/ICS, Utrecht, Netherlands
- Centre for Complex Systems Studies, Utrecht University, Utrecht, Netherlands

Arnault CHATELAIN

- ORCID: 0009-0002-6450-2176
- Centre de Recherche en Économie et de Statistiques (CREST), CNRS, École polytechnique, GENES, ENSAE Paris, Institut Polytechnique de Paris, Palaiseau, France

---

The release of ChatGPT at the end of 2022 has thrust large language models (LLMs) into the limelight. By enabling its users to query the model directly in natural language, it democratized access to these models - a welcome development. Since then, the child product of OpenAI as well as similar tools such as Bard, Claude, and Bing AI have shown both their versatility and their efficiency on a great variety of tasks.

Social scientists have been quick to embrace these models. They used these “assistant-type” LLMs to summarize research articles, to debug code, and even to emulate survey participants, experimental subjects, or agents in computer simulations[1]. Researchers also massively adopted them to annotate texts. By passing a simple prompt to the machine, they could now sort through thousands of documents. They could do so rapidly, precisely and according to their own coding scheme[2]. The promise was, indeed, alluring.

As social scientists who, for years, have been using various types of LLMs to annotate textual data, we were thrilled by these developments. Our own practice, so far, consisted in fine-tuning LLMs on specific tasks, i.e. in providing the models with hundreds to thousands of examples in order to “train” them[3]. The results are undeniable, but the manual annotation of these examples is often a long and tedious process.

We thus welcomed the arrival of these models, but also put them to the test. We compared ChatGPT's outputs with that of our models. We also conducted a thorough review on the nascent literature. The results were sometimes good, but often just fair, and at times really bad[4]. They rarely surpassed task-specific LLMs.

This lukewarm conclusion was nonetheless not what seemed most problematic with the use of these new methods. We believe three questions need to be addressed before we turn to these tools for scientific purposes.

On the Highway to a Reproducibility Crisis

Our first concern deals with the replicability of the results obtained by these models. Some suggest that GPT 3.5 (the model that powers the free version of ChatGPT) is sensitive to prompts, but others find it to be quite robust to minor changes in the wording of the requests it receives[5,6].

More problematic, in our view, is the lack of control we can exert over the model used in analyses. There is of course the classic criticism of these models being "black boxes". We don't know exactly how they operate, and we do not know what they are trained on. This is certainly true for the proprietary models, but this also partly true for their open-source counterparts. When working with assistant-type models in a chat environment, it is also unclear how their additional safety mechanisms operate.

The fact that the outcomes of such models are not stable due to frequent model updates only exacerbates the problem further. On our data, an experiment carried out with a given model often yielded different results when repeated a few weeks later. This certainly calls for careful reporting on the exact version of the model used. Yet, models are not always archived properly. A company such as OpenAI even tends to deprecate older models, making reproducibility virtually impossible (https://platform.openai.com/docs/deprecations).

Not Everyone Works with Public Data

A second matter of concern is that only certain types of data can be analyzed using GPT or similar commercial solutions, due to privacy and intellectual property issues. Arguably, OpenAI, the firm that commercializes ChatGPT, claims that it does not "use content that you provide to or receive from our API [...] to develop or improve services" (https://openai.com/policies/terms-of-use). But this does not mean that they won't do so in the future, or in another way.

And if the data one wants to annotate is protected by intellectual property laws, it should not be transmitted to the platform in the first place. In fact, the authors of a large-scale study that used articles from the New York Times were forced to conduct it on the title only, as the rest of the text was "not available in the public data"[6].

The texts we need to annotate can also raise privacy issues. In the social sciences, they can consist of open-ended questions in surveys containing potentially identifying or personal information, such as medical conditions. This only furthers the recent calls for open-source generative AI models[7].

Do We Want Even More English-Centered Research?

One last concern has to do with the English bias of these LLMs. As researchers who sometimes work with languages different from English, we cannot help but notice variations in the performances of the models across languages. Several papers report that assistant-LLMs perform best in English and

display rather poor performance in some low-resource languages. Others confirm this tendency by suggesting to either prompt the model in English first, or ask it to translate the prompt into English in order to get better results[8].

This situation will certainly evolve in the future, as LLMs get trained on more specific languages. Yet such an observation is puzzling, as the inequalities between languages will likely persist given the differential investments made by companies or governments. Languages from areas with fewer resources and languages spoken by small groups of people will probably be given scant research time. This could, in turn, lead to increased attention to English corpora, at the expense of other objects and sites of study. It would be a missed opportunity.

Conclusion

Let us be clear: we are excited about the current technological developments, and we do use LLMs (including these assistant-type LLMs) in our research. We are also optimistic that they could help reduce, in part, inequalities in science by offering affordable ways to annotate texts, thus granting access to textual resources to more researchers across the globe. The dazzling progress made by these models should nonetheless not conceal their potential flaws and limitations. Being oblivious would at best be a loss of time[9], at worse backfire[10].

Competing Interests

The authors declare that they have no competing interests.

References

1. Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. Preprint at https://doi.org/10.48550/arXiv.2305.03514 (2023).
2. Gilardi, F., Alizadeh, M., & Kubli, M. (2023). *PNAS* 120, 30 (2023).
3. Do, S., Ollion, É., & Shen, R. *Sociological Methods & Research*, (2022).
4. Ollion, É., Shen, R., Macanovic, A., & Chatelain, A. Preprint at https://doi.org/10.31235/osf.io/x58kn (2023).
5. Reiss, M. V. Preprint at https://doi.org/10.48550/arXiv.2304.11085 (2023).
6. Rytting, C. M. et al. Preprint at https://doi.org/10.48550/arXiv.2306.02177 (2023).
7. Spirling, A. *Nature* 616, 413–413 (2023).
8. Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. Preprint at https://doi.org/10.48550/arXiv.2304.05613 (2023).
9. Saphra, N., Fleisig, E., Cho, K., & Lopez, A. Preprint at https://doi.org/10.48550/arXiv.2311.05020 (2023).
10. Bail, C. A. Preprint at https://doi.org/10.31235/osf.io/rwtzs (2023).