Mitigating the ethical risks of large language models in machine learning research

Mikołaj Morzy, Poznan University of Technology

Mikolaj.Morzy@put.poznan.pl

keywords: machine learning, language models, ethics

Large language models have revolutionized the field of natural language processing in machine learning. Their versatility and unprecedented expressiveness make them a popular choice of tools for machine learning research. Unfortunately, the regime of unsupervised training of large language model results in several ethical and social risks encoded in these models. As more and more research reveals, large language models carry disturbing amounts of stereotypes, biases, prejudices, and discrimination. Applying a large language model to ethically vulnerable data (e.g. surveys and questionnaires collected online or social media posts) may introduce serious threats to the ethical foundations of research.

On the other hand, large language models can offer unparalleled means to extract information from text. Their capabilities include query answering, named entity recognition, relational entailment, fact probing, machine translation, text classification, natural language inference, commonsense reasoning, linguistic knowledge probing, and summarization, to name a few. It would be irresponsible to give up all the richness of this tool because of the marginal possibility of the experiment's results being poisoned by biases hidden in the model.

In the talk I will present the introduction to large language models and I will illustrate their expressiveness using real-world data. Next, I will discuss the training regime of large language models and I will identify training factors that may influence the ethical quality of the resulting model. Several methods of risk mitigation, including directional expectation tests and prompt learning will be presented. Finally, I will attempt to draft a set of best practices that may help to prevent the emergence of negative artifacts of large language models in research involving human subjects.

references

[1] Goetze, Trystan S., and Darren Abramson. “Bigger Isn’t Better: The Ethical and Scientific Vices of Extra-Large Datasets in Language Models.” In 13th ACM Web Science Conference 2021, 69–75. Virtual Event United Kingdom: ACM, 2021. https://doi.org/10.1145/3462741.3466809.

[2] Ashurst, Carolyn, Solon Barocas, Rosie Campbell, and Deborah Raji. “Disentangling the Components of Ethical Research in Machine Learning.” In 2022 ACM Conference on Fairness, Accountability, and Transparency, 2057–68. Seoul Republic of Korea: ACM, 2022. https://doi.org/10.1145/3531146.3533781.

[3] Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, et al. “Ethical and Social Risks of Harm from Language Models.” arXiv, December 8, 2021. http://arxiv.org/abs/2112.04359.

[4] Jo, Eun Seo, and Timnit Gebru. “Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning.” In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 306–16. Barcelona Spain: ACM, 2020. https://doi.org/10.1145/3351095.3372829.