Researcher perspective of OPT-OUT

PRIYANKA SINHA

This note presents concerns with potential OPT-OUT mechanisms in the internet. They include adverse impact to researchers would face; transparency in public processes; and non equitable access and accessibility and its impact on human technological advancement.

Additional Key Words and Phrases: Artificial Intelligence, OPT-IN, Internet Standard

1 INTRODUCTION

Connecting people and making information accessible through the workings of the Internet that has originated from the ARPANET project is widely accepted today. The IETF is an open standardization organization in public social good with wide participation of internet engineers.

We are currently experiencing a high speed of evolution of AI. In particular we have tremendous progress with LLMs, originating primarily from the developed nations of the world, that use the information from the Internet, be it the user generated content from the web, emails, as well as other multimodal data.

Unlike with traditional search systems, in the current race of AI, there have been reports that ISPs, data platforms and others are facing heavy incoming traffic to retrieve the data. It is demonstrated that they do not respect the robots.txt [5]. Thus, there is discussion to standardize stronger OPT-OUT mechanisms to address this.

In the rest of this note, we would like present concerns with an OPT-OUT standardization.

2 IMPACT TO RESEARCHERS

With the advent of social media platforms, participation of users, i.e., user generated content, have been invited by keeping open, fair, transparent platform policies. This usually included providing the data to researchers. Such policies invited users globally to provide their content to the platforms.

Author's address: Priyanka Sinha, priyanka.sinha.iitg@gmail.com.

However, increasingly, several major platforms, such as Twitter, Reddit and Stackoverflow, amongst others, have closed access to researchers already while using the content for training their proprietary systems.

While not being able to conduct research on data from public commercial platforms impacts science; if there are barriers to access on data that impacts public discourse, it could cause severe harms. This includes say email, meeting, notes, etc from open standardization bodies, open source communities, Wikipedia, sec filings, stocks, political speeches, elections data, legal proceedings, epidemic or public health data, and such data on the web. It also includes content on say internet relay chat, that is currently not universally searchable, but is being looked at by researchers as a data source.

Further beyond LLMs, AI is yet evolving across the world and the ingestion of data is a critical need for the same. Since the internet does not have in built mechanisms to record the provenance of information; attribution is difficult. It makes the ownership of the data harder to establish, which platforms naturally claim as theirs.

3 GLOBAL INEQUITY IN AI

Despite the success of LLMs, AI evolution is still ongoing. While some major conglomerates in developed countries may have collected the large amounts of data they require for their products, the same is not yet true for developing countries. The need for developing countries, especially in the Asia Pacific region is that there is a very large population that are not English speaking (not native english speakers), and they represent cultures, information needs that are not uniform. When existing AI systems are deployed for decision making without sufficient data from those regions (many of which experience access and accessibility issues) [1, 4], their potential dangers have been demonstrated [2, 3].

4 CONCLUSION

OPT-OUT may take into account the considerations of researchers, challenges of global inequity and problems of data provenance into account while designing protocol standards. This should reduce potential harms and improve fairness, accountability and transparency.

REFERENCES

[1] Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. 2021. When is memorization of irrelevant training data necessary for high-accuracy learning?. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* (Virtual, Italy) *(STOC 2021)*. Association for Computing Machinery, New York, NY, USA, 123–132. https://doi.org/10.1145/3406325.3451131

[2] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[3] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1571–1583. https://doi.org/10.1145/3531146.3533213

[4] Vitaly Feldman. 2020. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing* (Chicago, IL, USA) *(STOC 2020)*. Association for Computing Machinery, New York, NY, USA, 954–959. https://doi.org/10.1145/3357713.3384290

[5] Kevin Roose. 2024. The Data That Powers A.I. Is Disappearing Fast. https://www.nytimes.com/2024/07/19/technology/ai-data-restrictions.html