Large Language Model for Enhancing Accessibility to UK Open Data

Haris Baig
ID: 230431122 (924909)

October 3, 2024

1 Literature Review

In the modern world, data has shown promising results in various fields, especially when stakeholders can utilize the data and make more informed decisions resulting in better performance. The governments are one of the producers and collectors of large data (Alexopoulos et al., 2014). One form of data by governments is open data which has gained popularity over the years. This initiative was aimed to promote transparency, strengthen governance, fight corruption and empower citizens (Ubaldi, 2013). While transparency does not remove corruption, it does reduce it (Zuiderwijk et al., 2014). The Open Government Data (OGD) has brought many benefits to society by enabling citizens, researchers, and businesses to access datasets covering various domains such as healthcare, education, transportation, and environmental monitoring (Janssen et al., 2012). Open data does not only help governments but also citizens as well. It does this by helping governments to formulate data-driven services and citizens on the other hand, make use of this data to examine the government's performance (Virkar and Viale Pereira, 2018; Bvuma and Joseph, 2019). Although transparency in the data can make many sectors more open to the public and can make a more transparent and democratic society (Mutuku and Colaco, 2012), there are still some challenges which are reducing the open data from reaching its true potential.

While open data has shown promising results in promoting a better future, it is still a major challenge to achieve the full potential of open data. (Attard et al., 2015) discussed some potential issues such as the heterogeneous nature of data formats, cultural obstacles, and several open data life cycles that limit data consumption. This is further evident as traditional data analysis often fails to transform the raw datasets into interpretable insights due to the complex nature of government data (Zuiderwijk and Janssen, 2014). This is further challenging due to the complex nature of the data making it difficult to understand the data to its full extent (Zuiderwijk and Janssen, 2014; Attard et al., 2015). This was anticipated during the early days of open data where work like (Dawes and Helbig, 2010) had found challenges in fitting the data for external users and the dependence on metadata to understand and use it appropriately.

One of the potential methods to address the concerns of open data is Large Language Models (LLMs). LLMs have shown significant positive results in handling large and complex data across different domains (Naveed et al., 2023). In complex domains like coding, where projects such as (Chen et al., 2021) demonstrated that an LLM can be trained using open data on GitHub to help programmers in coding. The model was able to generate code of which 78% of the code was able to pass the unit testing. Similar work was done by Nejjar et al., 2023, but in contracts to previous research, multiple models like Chatgpt, Google Bard, and BingChat were compared in terms of correctness, comprehensibility, and efficiency. In finance,

(Wu et al., 2023 presented a model which was trained on a large archive of Bloomberg which performed better compared to other models. This was further improved by (Zhang and Yang, 2023), where authors presented used the pre-training and fine-tuning steps to avoid forgetting. Although, these works demonstrate the usage of LLM in specialized areas they lack to show more general usage of LLMs.

Furthermore, LLMs have shown some promising results in understanding data. (Ma et al., 2023) proposed InsightPilot a dedicated LLM model for data analysis. The solution utilized a dedicated pipeline coupled with prompt engineering to help users understand patterns, summarize findings, and explain the data. A more complex work was done by (F. Zhao et al., 2023) which showed the potential of these models and showed that LMMs can be used to perform qualitative data analysis by extracting key points and the relevance of the points in the data. While these projects used a well-defined dataset to perform analysis, in real-world scenarios these models require to have updated information before they can perform analysis. (X. Zhao et al., 2024) addressed this by creating a vector database and Retrieval Augmented Generation (RAG) coupled with domain knowledge to create a robust analysis model. This also showed that RAG can be used to reduce hallucinations in these models.

The applications of Transformer models have made their way toward government application as well. (Cao et al., 2024) presented a transformative framework which enabled non-technical stakeholders to engage effectively with complex climate data and simulations however the authors suggested exploring LLMs in other sectors. A more specific LLM have also been used on open data as well, (Mamalis et al., 2023) made a model using ChatGPT 3.5 on top of Scotland's open movement data and showed promising results in retrieving factual results. However, only a small portion of data was used and only one model was accessed which also left room for further evaluations.

References

Alexopoulos, C., Zuiderwijk, A., Charapabidis, Y., Loukis, E., & Janssen, M. (2014). Designing a second generation of open data platforms: Integrating open data and social media. *Electronic Government: 13th IFIP WG 8.5 International Conference, EGOV 2014, Dublin, Ireland, September 1-3, 2014. Proceedings 13*, 230–241.

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government information quarterly*, *32*(4), 399–418.

Bvuma, S., & Joseph, B. K. (2019). Empowering communities and improving public services through open data: South african local government perspective. *Governance Models for Creating Public Value in Open Data Initiatives*, 141–160.

Cao, C., Zhuang, J., & He, Q. (2024). Llm-assisted modeling and simulations for public sector decision-making: Bridging climate data and policy insights. *AAAI-2024 Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design*.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Dawes, S. S., & Helbig, N. (2010). Information strategies for open government: Challenges and prospects for deriving public value from government transparency. *Electronic Government: 9th IFIP WG 8.5 International Conference, EGOV 2010, Lausanne, Switzerland, August 29-September 2, 2010. Proceedings 9*, 50–60.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, *29*(4), 258–268.

Ma, P., Ding, R., Wang, S., Han, S., & Zhang, D. (2023, December). InsightPilot: An LLM-empowered automated data exploration system. In Y. Feng & E. Lefever (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing: System demonstrations* (pp. 346–352). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-demo.31

Mamalis, M. E., Kalampokis, E., Karamanou, A., Brimos, P., & Tarabanis, K. (2023). Can large language models revolutionalize open government data portals? a case of using chatgpt in statistics. gov. scot. *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, 53–59.

Mutuku, L. N., & Colaco, J. (2012). Increasing kenyan open data consumption: A design thinking approach. *Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance*, 18–21.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Nejjar, M., Zacharias, L., Stiehle, F., & Weber, I. (2023). Llms for science: Usage for code generation and data analysis. *Journal of Software: Evolution and Process*, e2723.

Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives.

Virkar, S., & Viale Pereira, G. (2018). Exploring open data state-of-the-art: A review of the social, economic and political impacts. *Electronic Government: 17th IFIP WG 8.5 International Conference, EGOV 2018, Krems, Austria, September 3-5, 2018, Proceedings 17*, 196–207.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Zhang, X., & Yang, Q. (2023). Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. *Proceedings of the 32nd ACM international conference on information and knowledge management*, 4435–4439.

Zhao, F., Yu, F., Trull, T., & Shang, Y. (2023). A new method using llms for keypoints generation in qualitative data analysis. *2023 IEEE Conference on Artificial Intelligence (CAI)*, 333–334. https://doi.org/10.1109/CAI54212.2023.00147

Zhao, X., Zhou, X., & Li, G. (2024). Chat2data: An interactive data analysis system with rag, vector databases and llms. *Proc. VLDB Endow.*

Zuiderwijk, A., Gascó, M., Parycek, P., & Janssen, M. (2014). Special issue on transparency and open data policies: Guest editors' introduction.

Zuiderwijk, A., & Janssen, M. (2014). Barriers and development directions for the publication and usage of open data: A socio-technical view. *Open government: Opportunities and challenges for public governance*, 115–135.