Large Language Model for Enhancing Accessibility to UK Open Data

Haris Baig
ID: 230431122 (924909)

August 9, 2024

1 Introduction

The advent of Open Government Data (OGD) has brought many benefits to the society. This has enabled citizens, researchers, and businesses to access datasets covering various domains such as healthcare, education, transportation, and environmental monitoring (Janssen et al., 2012). While this transparency in data has forced many sectors to be more open to the public, it has brought many challenges in terms of understanding, usage, and access to the data this is 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).

With the rise of artificial intelligence (AI), many new possibilities for interpreting and presenting complex datasets have been formulated. One such technology within the realm of AI is Large Multimodal Models (LMMs). LLMs are capable of handling vast amounts of complex data and have shown great promise in enhancing data interpretability and accessibility (Z. Chen et al., 2024). This research proposal aims to explore the potential of LMMs in terms of transforming the UK's open data making it understandable, utilizable, accessible, and beneficial for various stakeholders, such as the general public, journalists, and researchers.

This proposal provides a detailed overview of the potential benefits of leveraging LMMs for government open data, outlines the main research questions and methodology, assesses the project's feasibility, and presents a comprehensive project plan. The expected outcomes of this research include advancements in academic knowledge, business growth, social welfare and technical innovation, demonstrating the transformative potential of LMMs in public data interpretation.

2 Problem

2.1 Statement of the Problem

Thanks to the growing availability of open data, it is possible for anyone to explore vast amounts of data which can provide valuable information. Suppose a journalist investigates the growth in the local education system or how the government is performing at any point in time. Use cases like this and many others such as healthcare, transport and public policy present an opportunity for open data.

However, these open data have large size and complexity, making it difficult for people and even organizations to analyze the information. For understanding such data through research, analysis and in some cases expertise is required to interpret the data properly which is limiting the impact of open data on society.

This leads us to the foundation of our problem which is the gap between the current ability of users to effectively leverage the open data for meaningful insights and decision-making. Large Language Models (LLMs) offer a promising solution by providing advanced natural language processing capabilities that can automate and enhance the extraction of insights from open data. However, there is a lack of comprehensive research into how LLMs can be effectively integrated with open data to overcome these barriers and maximize their utility.

2.2 Purpose of Study

The study aims to explore the usage of Large Language Models (LLMs) to address concerns in open data and evaluate their efficiency in enhancing accessibility, explainability and utilization of open data. This research seeks to develop an application showcasing LLMs for open data, assessing their potential to automate the extraction of insights and make it easy for the general public and researchers to understand data.

The study objective is to answer the following key research questions:

1. Which LLM model can perform best for open data?
2. How much the LLM results are reliable?
3. How do LLMs compare to traditional analysis of the data?
4. What ethical considerations should be made when using LMMs for public data?

The proposed methodology will provide a clear explanation of how this project is addressing these questions

2.3 Significance of the Study

This study has the potential to open doors for advancement in various sectors and enable the effective usage of open data. Academically, it will advance research insights on the usage of LLMs and open data. The study's findings can present the development of new tools and methodologies to further explore the usage of AI in other complex data.

2.3.1 Business Benefits

The project has the potential to bring several benefits to business. The research could enable businesses to have to more understanding of the data without relying on third-party analysis firms. This will enable them to develop new products and services, expanding the market for AI-driven data solutions (Liang et al., 2022). Additionally, once the data interpretation has improved, the project could lead to making businesses aware of government data faster which will enhance companies to make more decisions, optimize operations, and develop strategies, leading to increased efficiency and competitiveness (Provost and Fawcett, 2013).

2.3.2 Social and Environmental Benefits

The research has the potential for social and environmental benefits. By making government data more accessible, this project will empower citizens with information, fostering transparency and accountability. This increased accessibility will enable individuals to engage more actively in public discourse and decision-making processes, promoting democratic participation (Virkar and Viale Pereira, 2018). Furthermore, the project's enhanced data accessibility will not only assist policymakers in making better policies but also their critiques such as journalists to inform the public promptly.

3 Literature Review

In the modern world, data has shown promising results in various fields especially when the data is understandable. 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). 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).

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. There are a number of barriers, including technical, policy and legal (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 have shown significant positive results in handling large and complex data across different domains (Naveed et al., 2023). Complex domains like coding, where projects such as (M. Chen et al., 2021) demonstrated that an LLM can be trained using open data on GitHub to help programmers in coding. However, only 78% of the generated code was able to pass the unit testing which shows the potential of these models in open data. 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. 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 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.

4 Methodology and Challanges

The project will employ a comprehensive methodology comprising several key components:

4.0.1 Data Collection

The first step would involve the collection of open data. This can be done easily as there are multiple websites including the official UK government website and APIs that provide us with the data. While there exist numerous data across different sectors on these platforms, only a selected few data will be taken into consideration ensuring a comprehensive representation of different domains (e.g., health, education, environment). This will ensure that the project meets the expectations within the deadline. The data collection process will focus on obtaining high-quality datasets with varying formats (SQL, XLS, HTML, and PDF) and complexities to ensure the robustness and generalizability of the proposed solution.

This can be challenging as the data retrieved may have invalid information. Furthermore, data would require additional analysis and preprocessing such as formatting, cleaning, and checking to make the data well-tuned for the model.

4.0.2 Model Development

The core part of this project involves developing an LLM model tailored for the government data. While many LLM models exist such as GPT, BART, and Llama etc. Further investigation during the research would be required to ensure the key standards of the projects are met which also includes the cost of running these models. Once the model is developed, a prototype system will be built to showcase the capabilities of LMMs in interpreting complex data and providing key metrics of the proposed solution. Moreover, the implementation will also focus on developing an interface that allows users to interact with the system.

While selecting the LLM models has its own challenges, the core challenge will be to come up with a model that is less power-consuming and gives the best results. This will require checking across different sets of models that are available.

4.0.3 Evaluation

This project would require multiple types of evaluation metrics to evaluate the performance of the system. First, the LLMs would require novel types of testing which can assess the performance of the model. Tools like BenchLLM and PromptFlow can be used to check the performance of different models. Second, the evaluation process would require to include metrics such as summarization, hallucination, coherence, and bias revealing how much the results are reliable. Finally, regression testing would be done to ensure all the standards of the projects are met on time.

Model evaluation is only one part of the project the other challenge will be to cross-validate the results with actual ground truth. This can be done by scrapping existing analysis from the web and manually checking the data to understand how our model is performing compared to traditional analysis.

4.1 Project Plan

The project plan spans four months and is divided into distinct phases, each with specific tasks and milestones. The activities are outlined in the Gantt chart below.

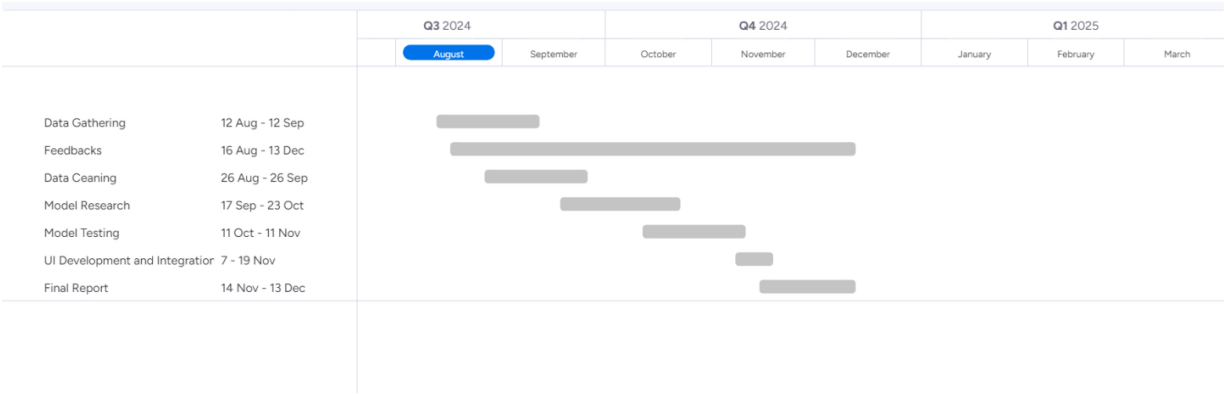

Figure 1: Proposed timeline of the project

First, the project is initiated with the data gathering. This stage also involves pre-processing the data and making sure only high-quality data is gathered and does not raise any concern for model development. Second, the development of LLM will start. Once a base model is designed it can be used to develop other models and compare with other models this also includes testing models in our performance metrics. Third, once the model is developed we move to the UI integration to our mode. Finally, the final report of the project is made showcasing the findings of the research.

5 Limitations and Ethics

5.1 Limitations

While the project shows promising use cases and ambitions, it may require further investigation on a larger scale this can include usage across different kinds of datasets, reviews from professionals, and more and work on the model before it can be deployed into the real world. Furthermore, new technology may be needed to replace power-hungry models.

5.2 Ethical Implications and Framework

The project requires significant ethical considerations as the government's open data is being addressed which can bring many ethical challenges such as:

5.2.1 Data Privacy

The project will ensure that the data does not contain sensitive or personally identifiable information. Furthermore, open data also include information regarding the defence which can reveal underlying sensitive information. This may require preprocessing the data before accessing or discarding data sets such as criminal records, student loans, land ownership etc.

5.2.2 Bias and Fairness

AI models are well known for having biases in the results and LLMs are no exception. The project would require to asses the potential biases and ensure that it is eliminated using diverse datasets.

5.2.3 Ethical Use of AI

Numerous instances have been found where LLMs were used to engage in unethical activity. Furthermore, LLMs have also been jailbroken to avoid the restriction layer implemented on the model. The project will need to ensure that such cases can be avoided and a process which can do accountability in case of a potential breach.

5.2.4 Social Responsibility

Finally, The project should consider to reduce carbon footprint as the LLMs do heavy computation which may require a lot of power consumption. The project must develop an environmentally friendly solution. This can be used as a key parameter when selecting the model.

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.

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*.

Chen, Z., Mao, H., Li, H., Jin, W., Wen, H., Wei, X., Wang, S., Yin, D., Fan, W., Liu, H., et al. (2024). Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, *25*(2), 42–61.

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.

Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., & Zou, J. (2022). Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, *4*(8), 669–677.

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.

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*.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data*, *1*(1), 51–59.

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.

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.