1500-2000 words

* Advantages of OGD
* Challenges of OGD
  + What are the problems related to consumption
    - Analytics
    - Understanding
    - Access
* AI Solution
  + Potential solutions
  + Related
    - How they are related
  + Same

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.

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.