

International Journal of Population Data Science

Journal Website: www.ijpds.org



Talk data to me! Evaluating the potential for large language models to enhance data discoverability across UKRI's federated data services

Mark Green¹, Maura Halstead², Cillian Berragan¹, Caroline Jay², David Topping², Richard Kingston², and Alex Singleton¹

¹University of Liverpool, Liverpool, United Kingdom

²University of Manchester, Manchester, United Kingdom

Objectives

The talk will: (i) Describe the development of a large language model (LLM) powered semantic search tool for UKRI data catalogues, and (ii) Examine the concerns and opportunities of using this tool among researchers for data discovery.

Conclusion

Although Large Language Model's can be useful for supporting federated data discovery among researchers, tools need to be developed that are responsible, trustworthy and open if researchers are going to use them.

Methods

A semantic search tool was developed integrating the data catalogues of Administrative Data Research UK, Consumer Data Research Centre, and UK Data Service. We used OpenAI's vector embedding service to convert these metadata into embeddings, allowing natural language search to be used rather than keywords only.

We assessed the acceptability and suitability of this tool using four focus groups. Participants were recruited across academic researchers, PhD researchers, data services staff, and local government / third sector analysts (n=36). Data collected from focus groups were analysed using thematic analysis.

Results

The key themes identified in focus groups were: (i) Current data discovery techniques are dependent on keyword strategies for searching (including the dominance of using Google). There is need to support training for using any LLM based resources. (ii) There was low trust of LLMs, especially in academic researchers. Participants were concerned that results may be erroneous. Being able to 'explain' why a search result was returned was viewed as valuable. (iii) Having a resource that collates all metadata in one place was powerful for helping researchers find data. This could be improved through leveraging the power of LLMs to summarise large quantities of information about datasets to make data discovery more efficient. Our talk will detail steps towards addressing these challenges.

