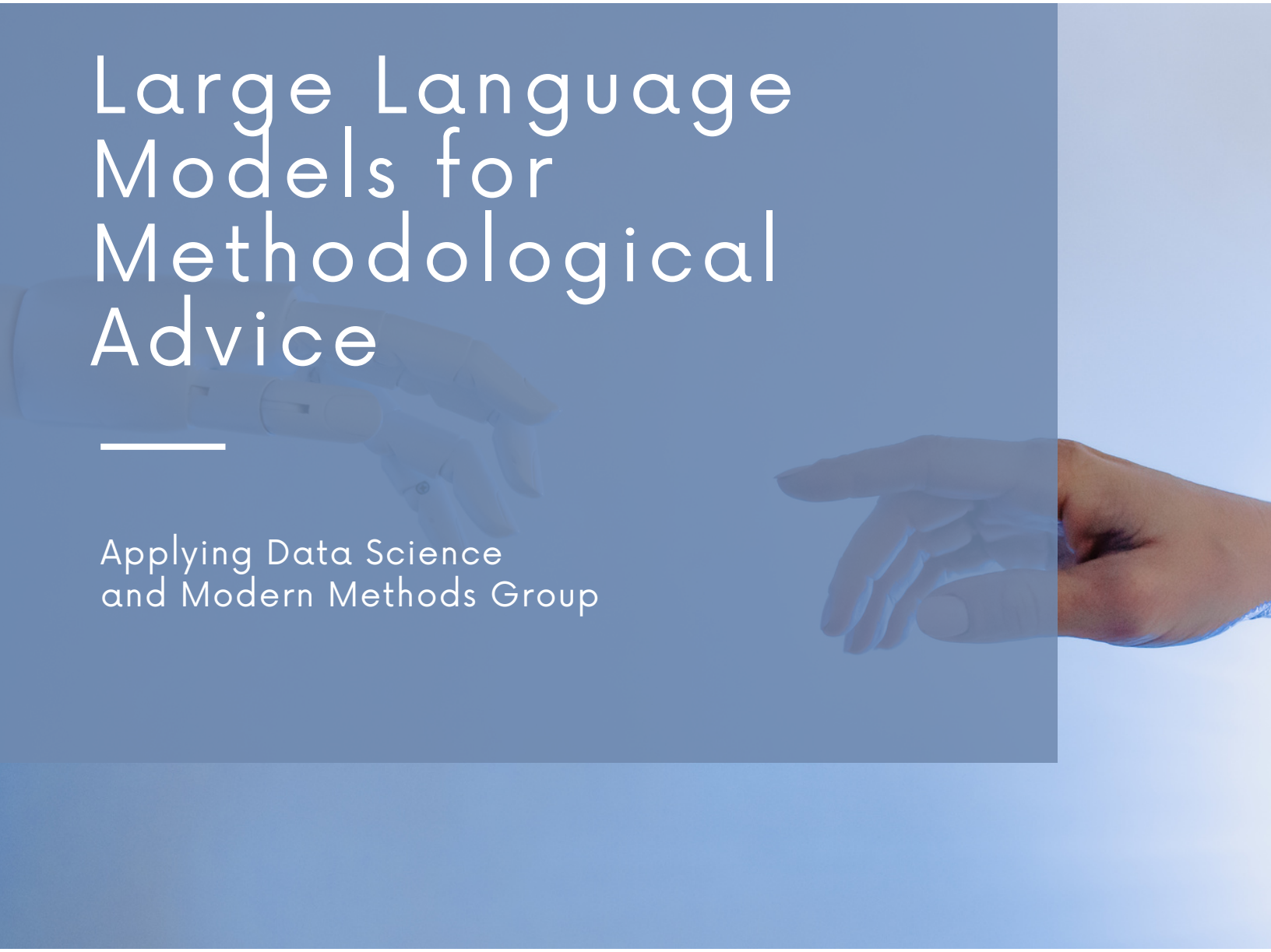
modernstats

Large Language Models for Methodological Advice



Applying Data Science and Modern Methods Group

March 2024

Introduction

Large Language Models (LLMs), such as Chat-GPT, have recently demonstrated improved performance to the point that they can now be useful tools for querying knowledge and generating or improving text. They can be considered to be a next generation of knowledge tools following in the footsteps of internet search engines and public information stores such as wikis by adding the ability to generate new content. This new content would be generated by the LLM by cross-referencing the extremely large number of information sources on which they are trained on.

Given their utility, it is expected that statisticians and researchers will make use of LLMs for different purposes. This short note contains some simple guidelines on their use to ensure that they are used in an appropriate manner by balancing the benefits of fast access to relevant information against the risks of the information being incorrect, incomplete or misleading.

The potential of using LLMs in the production of official statistics is discussed in the High- Level Group for the Modernization of Official Statistics' (HLG-MOS) white paper Large Language Models for Official Statistics. LLMs are a form of Generative Artificial Intelligence (Gen-AI) as they can generate text content. Trained on a very large corpus of data, LLMs use models to predict the most statistically likely next word in a sentence. Given that they are trained on enormous amounts of data, they produce very well written text that can easily come across as written by humans. However, LLMs only predict the most likely next word and do not understand the content that they produce.

Despite their weaknesses, LLMs do have the potential to change how a National Statistical Office (NSO) works and should not be dismissed. For more details on how LLMs could be leveraged in an NSO, we refer the reader to the HLG-MOS white paper mentioned above. This note looks at one particular use of LLMs which is not covered in the white paper. That is, their use to obtain methodological advice.

Research

The research question covered is how LLMs deal with requests for methodological advice that would be fairly typical coming from employees of an NSO. We created a small number of methodological queries around how to solve some relatively specific issues that are commonly found in the production of government statistics. The resulting prompts were queried against a number of popular LLMs such as ChatGPT, Bing (Creative and Balanced modes) and Bard, and the received responses stored. We then validated the solutions proposed by the LLMs with experienced statisticians and methodologists.

Examples

The following questions were asked to the LLMs and a high-level summary of the responses is given for each question.

1

"Act as a data scientist. I am dealing with imputation of missing data fields due to tax filing deadline extension. Some companies submitted their tax returns before the due date, while others did not. I would like to impute those who do not by making use of the data of those who have submitted their tax returns already. What type of methods could I use to improve the quality of the imputed data?"

When asked this question, the LLMs returned methods such as mean imputation, median imputation, K-nearest imputation, hot and cold deck imputation, multiple imputation and regression imputation.

2

"I want to produce estimates for small geographical areas but the estimates from a survey are not of high enough quality. What are my options, and do I need additional data?"

When asked this question, the LLMs returned responses such as increase the sample size, stratified sampling, use auxiliary data, small area estimation methods such as unit or area level models, direct estimation, mode- based estimation, composite estimation and use big data.

3 "I have data from multiple data sources and would like to combine them somehow to do an analysis. There are some variables in common across some of the data sources, but I do not have any variables that appear on all of them. What methods exist which will allow me to analyse the data?"

When asked this question, the LLMs returned responses such as data merging feature engineering, data warehousing, probabilistic record linkage, identity resolution, extract, transform and load, data federation and 'I am not programmed to assist with that'.

4 "I have a lot of paradata related to business survey data collection. This data includes characteristics about the business such as revenue, number of employees, industrial sector, history of response to surveys, etc. What technique can I use to predict the probability of responding to a future survey?"

When asked this question, the LLMs returned responses such as binary classification, feature selection, data preparation, logistic regression, decision trees, random forests, neural networks and 'I am not programmed to assist with that'.

5 "I have a time series of hourly data that is collected for many months across multiple locations. The time series consists of ANPR measures that give me the count of vehicles (cars, vans, trucks, motorbikes) that cross the ANPR camera. However, sometimes the cameras become faulty leading to missing data at specific times and locations. What techniques could I apply in order to improve the quality of my traffic estimates from this data?"

When asked this question, the LLMs returned responses such as time series imputation, seasonal adjustment, multiple imputation, mean and median imputation, last value carry forward, imputation methods, modeling, ensemble and making sure that the data is clean.

Lessons Learnt

Trying out the different LLMs, testing different questions and talking with experts allowed us to draft some high level lessons learnt:

- LLMs are not statisticians/methodologists and will act on incomplete information. Human statisticians and methodologists will query you if you give them incomplete or inaccurate information. This interaction and challenge is very important to ensure that methodological advice is correct.

- Different LLMs will give different responses. The same query, given to a different LLM, can give a different list of recommended methods. In the case of Bing, oddly enough, the Creative mode seems better suited to providing guidance, likely because the methodological questions are too vague for stricter modes to correctly answer.

- LLMs are likely to provide a limited subset of possible answers to the question. Small changes on the query prompt or requesting additional information will make the LLM expand on the number of available solutions or options.

- LLMs are poor at expressing confidence. LLMs are likely to either not provide any estimate of how confident they are about their answers or overstate their confidence. This can make a user assume that the LLM is correct when it provides poor quality or untested advice.

- LLMs are poor at caveats and limitations. Many of the LLMs, unless explicitly prompted, provided recommendations without any caveats or examples of cases in which those recommendations might not be suitable.

- Often the solutions provided are generic or obvious. By default LLMs are likely to recommend the most common or well proven methods, rather than by prioritising according to their suitability. This is due to how heavily such methods feature in the methodological literature, which can lead to users not being offered recommendations about less frequent but more suitable methods.

Recommendations

Using LLMs for methodological advice in the production of official statistics can offer numerous benefits, but it also presents certain challenges and considerations. Here are some recommendations for their use in this context.

Understand the Capabilities and Limitations:

- Recognize that LLMs are advanced text-based models with the ability to generate human-like text, answer questions, and provide context-based information.
- Be mindful of their limitations, especially potential biases in the training data and their inability to grasp nuances or context as effectively as a human expert can.
- Note that LLMs often lack abilities to process numerical data.

Provide Clear Guidelines and Supervision:

- Develop comprehensive guidelines for integrating LLM into the work of the organisation, describing specific use cases and scenarios in which LLM can be most useful because it cannot be used blindly for every situation. Consider data privacy and security as highlighted in the white paper.
- Make sure there is human oversight and quality control when using LLMs to prevent the spread of errors or biases.
- Ensure human experts are available to escalate queries to.

Experiment with Different Prompt Styles:

- Experiment with different prompt styles and structures to improve your LLM experience. Minor adjustments in wording can lead to more meaningful responses.
- Encourage the use of well-structured prompt formulations and provide training to obtain reliable statistical information.

Validate Results:

- Validate the LLM results and insights by discussing them with experienced statisticians/methodologists.
- Run more than one query to ensure the key information provided by the LLM is consistent across runs rather than an outlier.
- Pay close attention to the prompts you use when enabling an LLM, as they affect the quality and relevance of the general output you receive.

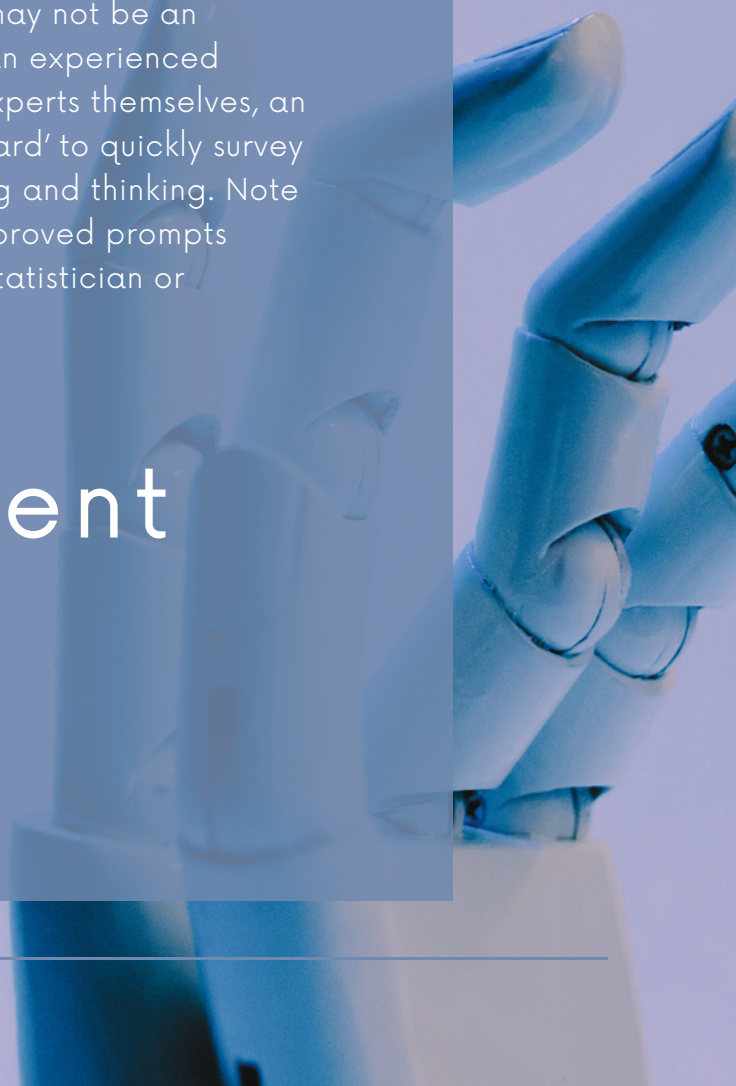
Compile a Standardized Prompt Library:

- Consider creating a standardized prompt library for your organization to simplify interactions with LLMs and ensure consistency in prompts used across projects.

Conclusions

LLMs have a broad range of applications across various domains. They can generate text, translate languages, summarize content, answer questions, write code, engage in conversations, assist with tasks, retrieve information, create content for games, aid in education, support healthcare and legal tasks, analyze sentiment and emotions, conduct market research, enhance accessibility, facilitate storytelling and generate prompts for discussions or interviews. However, it is crucial to note that human review and oversight are strongly advised and, especially in critical applications, should be required as LLMs may not always produce flawless or contextually appropriate output. Unlike their human counterparts, LLMs do not naturally show scepticism of the queries they receive, and rarely ask for additional information or follow up the exchange.

In terms of using LLMs for receiving methodological advice one has to be careful to not blindly follow the results produced by the LLM. The response obtained will be very well expressed but may not be an appropriate solution to the query. Discussions with an experienced statistician or methodologist are advised. For the experts themselves, an LLM might be a useful tool acting as a 'sounding board' to quickly survey the available options or as a general aid to drafting and thinking. Note that the responses could be improved with some improved prompts which could be developed in collaboration with a statistician or methodologist.



Acknowledgement

Special thanks to the authors of the paper:

Joni Karanka - Office for National Statistics (UK)
Wesley Yung - Statistics Canada
Bilal Kurban - Turkstat
Amilina Kipkeeva - UNECE