Project Report for DSCI D699: Graduate Independent Study in Data Science

Leveraging GPT Models to Enhance Federal Grant Data

Krishna Teja Jillelamudi

April 26, 2024

INDIANA UNIVERSITY

Contents

1

Acknowledgments

I would like to extend my deepest gratitude to Professor Yong Yeol Ahn for his invaluable mentorship and support throughout this project. Additionally, I am particularly thankful for the guidance and assistance provided by Miao Lili. Her dedication and attention to detail have been truly exemplary and greatly appreciated. Their combined contributions have not only enriched this project but have also significantly enhanced my learning experience.

2 Introduction

In today's data-driven world, the accuracy and effectiveness of organizational data are crucial. The project titled "Leveraging GPT Models to Enhance Federal Grant Data" aims to improve the quality of organizational datasets, specifically those concerning federal grants and their ties to the National Science Foundation (NSF) in the USA. By deploying Generative Pre-trained Transformer (GPT) models, this project explores a variety of GPT models and prompt engineering methods to establish improved data processing practices.

The datasets involved are split into two categories: one consists of 10,000 organization names, and the other is a verified dataset listing organizations known to have collaborations with the NSF. The datasets can be accessed at: Organisations data

3 Organisations affiliation with NSA - String filtering with Python

Script: NSF String Filter Script

The script uses Python to detect organizations potentially affiliated with the National Science Foundation (NSF) of the USA. It applies a string filtering method to a primary dataset containing 10,000 organization names. From this, 1,163 organizations were initially identified as matching the criteria by containing the string 'nsf' or 'national science foundation' within their names. After removing duplication, the count of unique organizations was adjusted to 1,153.

The analysis revealed that some organizations, despite passing the string filter, are affiliated with the NSF from countries other than the USA. The method for identifying these cases will be detailed in subsequent sections of this report. The script is structured around the filter organization function, which is essential for processing the data. This function accepts a path to a CSV file as its input and produces a data frame that lists organizations whose names suggest a potential connection to the NSF. Following the filtering process, the script conducts further checks for data integrity by identifying and counting any duplicate entries among the filtered results. This step is vital for maintaining the quality and uniqueness of the data, which is crucial for subsequent analysis.

A significant element of the script involves extracting the names of organizations that are unique to the filtered list but do not appear in the ground truth data. This provides critical insights into the types of organizations that are identified by the string filtering method but are not included in the ground truth dataset, underscoring potential gaps or expansions in the data collection methodology.

4

Exploring Organizational Affiliations with the NSA Using GPT Models

This section presents a methodical approach to evaluating the effectiveness of various GPT models in identifying organizational affiliations with the National Security Agency (NSA). We curated a dataset comprising 150 organizations, segmented into six batches of 25 names each. This segmentation was refined through multiple trials to minimize issues like truncation and to ensure consistency in GPT outputs.

GPT Models Employed

The experiments employed several state-of-the-art GPT models, including ChatGPT-4, Gemini 1.5 Pro, and Claude 3 Opus. We developed and iteratively tested multiple prompt variations to establish a single, effective prompt format that yields consistent results across all tested models.

Performance Metrics

The performance of each model was quantitatively assessed using three primary metrics:

- True Positives (TP): Organizations correctly identified by both manual assessment and the GPT models as having NSA affiliations.
- Positives (FP): Organizations identified by the manual process but incorrectly by the GPT models.
- False Negatives (FN): Organizations missed by the GPT models but identified in the manual review.

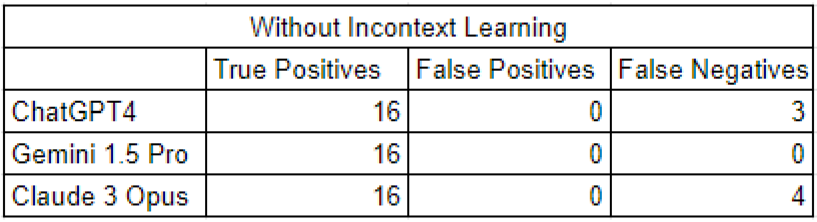
These metrics help in understanding the accuracy and reliability of the models in practical scenarios.

Importance of In-Context Learning (ICL)

Experiments were conducted with and without in-context learning (ICL) to gauge its impact on model performance. In-context learning involves providing the model with relevant examples or context prior to the task, enhancing its ability to make informed predictions. This is particularly beneficial in tasks requiring nuanced understanding or specific knowledge.

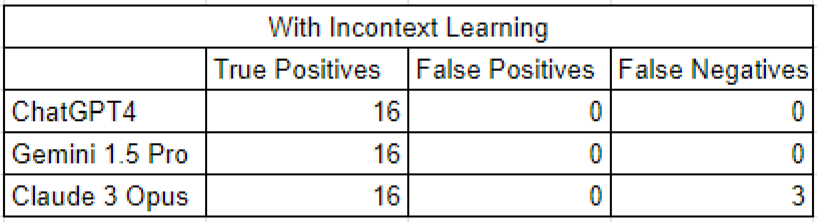
Comparative Performance Visualization

Below are figures illustrating the performance metrics for the models, with and without in-context learning:



Without Incontext Learning			
	True Positives	False Positives	False Negatives
ChatGPT4	16	0	3
Gemini 1.5 Pro	16	0	0
Claude 3 Opus	16	0	4

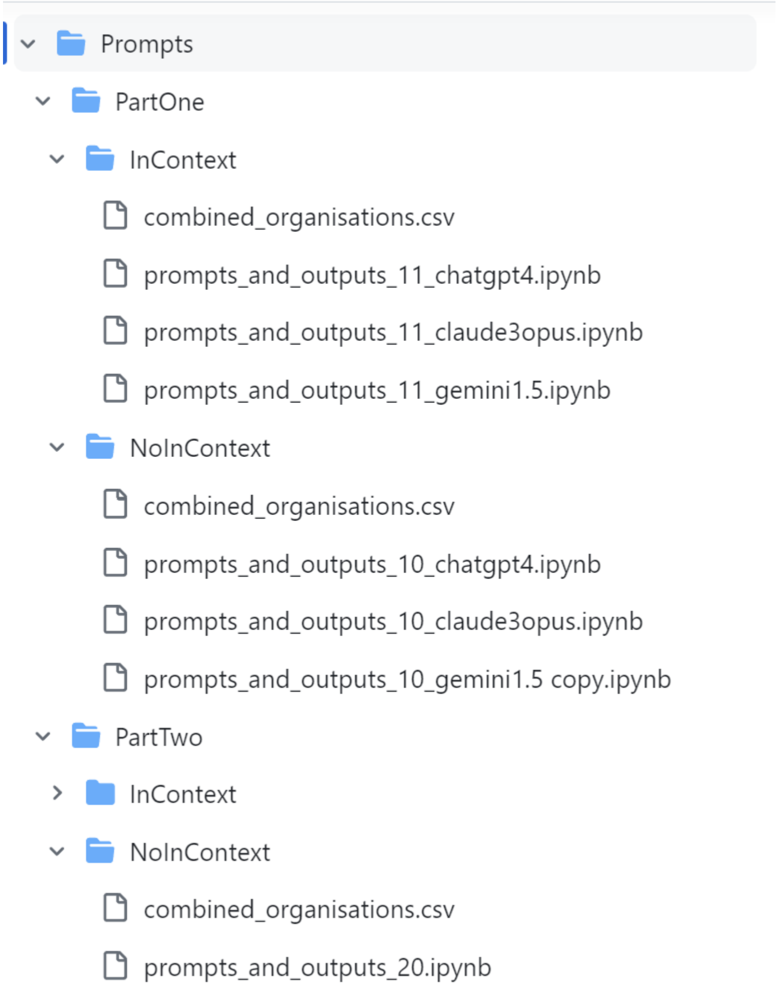
Figure 1: Performance metrics for models without In-context Learning



	With Incontext Learning		
	True Positives	False Positives	False Negatives
ChatGPT4	16	0	0
Gemini 1.5 Pro	16	0	0
Claude 3 Opus	16	0	3

Figure 2: Performance metrics for models with In-context Learning

For a detailed exploration of each prompt, model outputs, and further Python analysis, visit the project repository. Below is an image of the directory structure showcasing various Python notebooks used in this study.



```
Prompts
  PartOne
    InContext
      combined_organisations.csv
      prompts_and_outputs_11_chatgpt4.ipynb
      prompts_and_outputs_11_claude3opus.ipynb
      prompts_and_outputs_11_gemini1.5.ipynb
    NoInContext
      combined_organisations.csv
      prompts_and_outputs_10_chatgpt4.ipynb
      prompts_and_outputs_10_claude3opus.ipynb
      prompts_and_outputs_10_gemini1.5 copy.ipynb
  PartTwo
    InContext
    NoInContext
      combined_organisations.csv
      prompts_and_outputs_20.ipynb
```

Figure 3: Directory structure depicting various Python notebooks

5

Geographical aspect in Organizational Affiliations with the NSA

Script: Identfying organisation's country

The data preparation involved a meticulous process where organizations were first subjected to string filtering based on keywords suggesting NSF affiliations. Post filtering, the data was cleaned to remove duplicates, ensuring that unique organization names were retained for further analysis. This unique list was then cross-referenced against a ground truth dataset to isolate organizations that are possibly affiliated but not officially documented as such. This subset of data provides a fertile ground for examining geographical biases or concentrations in NSF-related funding and affiliations.

The core of our methodology involved the use of ChatGPT4 model to analyze the affiliation of these unique organizations. The prompts designed were crafted to elicit specific information about the country of origin or operational base of each organization. By structuring the prompts to request detailed geographical data, the models could generate insights based on the names and affiliations indicated by the dataset.

The analysis revealed a notable concentration of organizations in the USA and China, which was not unexpected given the global influence and research funding capabilities of these countries. However, the presence of organizations from other regions, though lesser in proportion, underscores the global reach and impact of NSF's initiatives. This geographical spread highlights potential areas for increasing diversity in research funding and international collaborations.

The geographical analysis of organizations affiliated with the NSF, derived from uniquely filtered data, sheds light on the global distribution and reach of NSF's funding. It also highlights the importance of tailored prompts in extracting specific data from GPT models, underscoring the capability of advanced AI tools in enhancing our understanding of large-scale data sets. This approach not only aids in strategic planning and policy formulation but also in fostering a more inclusive and diversified research funding environment globally.

Below is the barplot illustrating the percentage of organizations and the countries they are based out of:

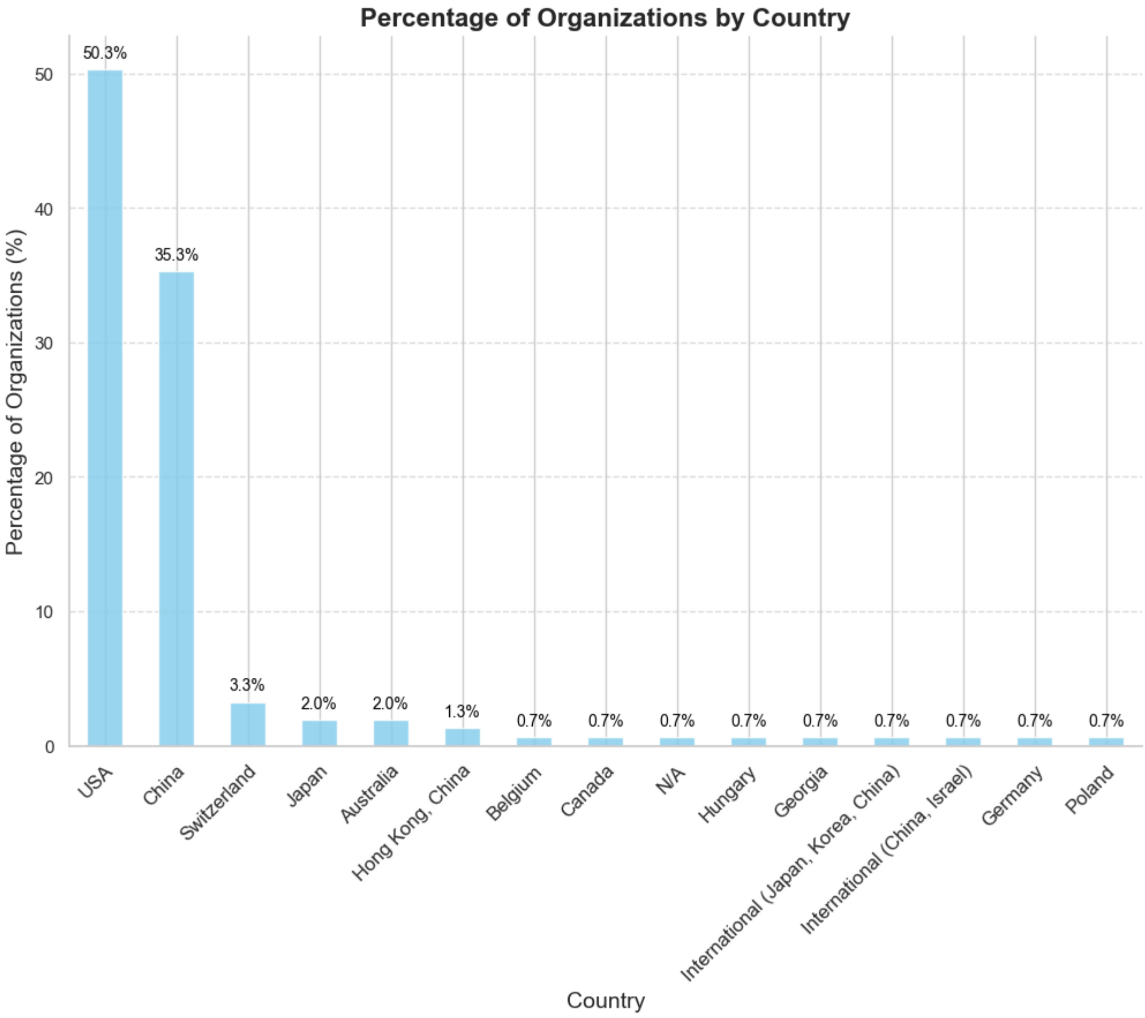


Figure 4: Directory structure depicting various Python notebooks

6 Future Scope

As the project continues to evolve, there are significant further implementations that can be made which focus on better modeling and real-time analysis.

In terms of improving the AI models, trying out various other GPT models and focusing on designing a custom GPT model for dealing with similar kinds of datasets would be a productive approach. Additionally, utilizing external data to the GPT models will significantly aid the implementation, including but not limited to publications data, particular web pages data, and more.

Another expansion can involve developing tools/platforms that are powered by their GPT models in real-time, this will uncover any sudden or impactful changes in the patterns, and this coupled with interactive dashboards will be of great help to administrators and policymakers.

7 Conclusion

The focus of the project was to make sure of the existing GPT models and prompt engineering techniques to simplify the analysis of datasets. Using a combination of python scripts and prompts it was shown that the approach has a great potential to pursue further. The GPT models have proven to be invaluable tools for dissecting the textual data of grant applications and funding acknowledgments, offering insights into the affiliations and effectiveness of funding across various scientific disciplines.

The project implementation also compared various GPT models, analyzed their strengths and weaknesses, and used the model accordingly. Understanding various strategies regarding the GPT model performances and nuances is a key takeaway from the project. As we look to the future, the continued development and enhancement of these AI models will undoubtedly unlock more opportunities for optimizing research funding and maximizing its impact on scientific and technological progress.