Bridging the Gaps in the LLM Interpretability of Unstructured Data with Knowledge Graphs

Eric Lin
lineric@nyu.edu
New York University

Sean Wiryadi
sw5131@nyu.edu
New York University

Patrick Sun
ps4126@nyu.edu
New York University

Abstract

This study introduces a robust hybrid system that enhances the interpretability of unstructured data by integrating traditional Natural Language Processing (NLP) techniques with advanced capabilities of Large Language Models (LLMs), specifically GPT-4. Our approach systematically transforms unstructured text data into a structured knowledge graph (KG) through a sequence of preprocessing steps—summarization, normalization, lemmatization, and stop-word removal, followed by the transformation of cleaned text into Object-Relationship-Object (ORO) triplets. These triplets form the basis of our dynamically constructed KG, optimized for subsequent analysis by LLMs. For retrieval, the system employs a vectorization of both the KG triplets and incoming queries, applying cosine similarity to identify the most relevant triplets. This relevance-driven parsed KG then aids a GPT-4 model in generating accurate and contextually nuanced responses. We evaluate our system's performance through rigorous metrics, assessing the accuracy and relevance of the LLM's outputs against benchmark datasets. The result is a significantly improved framework for converting unstructured textual data into actionable, structured knowledge, poised to enhance the application of NLP technologies across various sectors. Here is our accompanying GitHub repository.

1 Introduction

The digital revolution has ushered in an era characterized by the exponential generation of unstructured data, particularly text. This proliferation of data not only introduces complexities in storage and retrieval but also in interpreting and extracting meaningful insights, crucial for informed decision-making across various sectors.

Traditional NLP techniques, while foundational, often falter under the sheer volume and intricacy of this data, struggling to capture the full context and the nuanced relationships it contains. Despite the advancements brought by pre-trained language models like GPT, a gap persists in the systematic organization and interpretation of unstructured data. This gap significantly hinders the efficient exploitation of the information embedded within the data, constraining the practical applications of NLP in knowledge-driven tasks.

This study proposes a novel hybrid system that bridges this interpretability gap by combining traditional NLP techniques with the advanced capabilities of Large Language Models (LLMs). Our system not only enhances the structuring of unstructured data but also improves the interpretability and utility of LLM outputs in real-world applications. By implementing this system, we aim to facilitate the transformation of unstructured textual data into actionable, structured knowledge, thereby broadening the scope and efficiency of NLP technologies in various industries.

2 Problem Motivation

2.1 Dependency on Large Training Datasets

Large Language Models (LLMs) such as GPT and BERT rely heavily on extensive corpora for training. This dependency not only demands significant computational resources but also limits these models' ability to generalize to new, unseen datasets. Our approach mitigates this by integrating a dynamic knowledge graph that enriches the model's context without requiring extensive retraining on large datasets.

2.2 Limited Understanding and Logical Reasoning

While LLMs excel at generating plausible text, they often lack deep comprehension and are prone to logical errors and factual inaccuracies. By structuring unstructured data into knowledge graphs,

our system enhances the LLM's ability to "understand" the data it processes, thereby reducing the occurrence of nonsensical outputs.

2.3 Lack of Generalizability

Generalizability remains a challenge for LLMs trained on specific datasets. Our hybrid model leverages knowledge graphs constructed from diverse data sources, enabling the LLM to apply its learned capabilities more broadly and accurately across various domains.

Each of these challenges directly informs the design of our methodology, which utilizes both traditional NLP techniques and modern LLM capabilities to create a more adaptable and intelligent system.

3 Data

3.1 Data Sources

Our project utilizes a rich dataset of earnings call transcripts, available from the Hugging Face dataset repository (here) and an additional file, `earnings-transcripts.jsonl`. This file contains structured JSON entries, each corresponding to an individual earnings call. An example entry includes metadata such as date, exchange, ticker symbol, and the full transcript of the call.

3.2 Data Preprocessing

The preprocessing of the dataset involves several steps to prepare the text data for further analysis:

1. **Summarization**: Transcripts were first summarized to condense the content while preserving essential information. This was facilitated by the GPT-4 model, ensuring that each summary was limited to 600 words but retained critical details.

2. **Normalization and Lemmatization**: Following summarization, the text underwent normalization and lemmatization to standardize variations of the same word to their base form, enhancing the consistency of subsequent analyses.

3. **Stopword Removal**: Commonly occurring but less informative words (stopwords) were removed to focus on more meaningful text elements.

3.3 Knowledge Graph Construction

Post-preprocessing, a knowledge graph was constructed from the summarized transcripts:

1. **Extraction of ORO Tuples**: Using a modified GPT model, Object-Relationship-Object (ORO) tuples were extracted from the summaries, which served as nodes and edges in our knowledge graph.

2. **Graph Building**: These tuples were then utilized to build a dynamic knowledge graph, structuring the information in a manner that allows for efficient querying and analysis.

3.4 Data Structure

The final dataset, tailored specifically for this research, contains entries filtered by specific ticker symbols and quarters, significantly reducing the size from the original large dataset to a more manageable set of 1,000 entries. Each entry includes key details such as the ticker symbol, quarter, and the associated summarized transcript, alongside a column for a question and its ground-truth answer, facilitating the training of our models.

This structured approach to handling and analyzing data ensures the robustness of our methodology and the reliability of the insights derived from our research.

4 Methodology

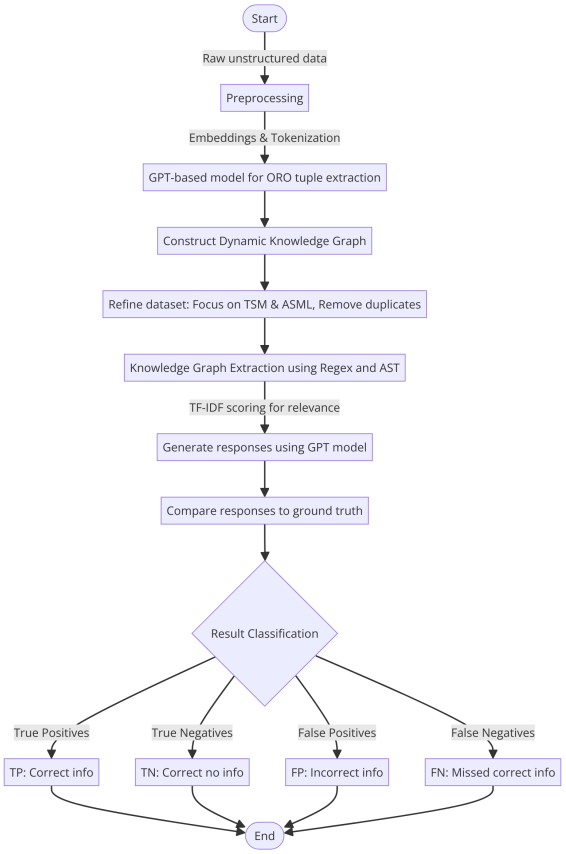

Figure 1: Methodology of project

Our research introduces a novel hybrid system that integrates the robust preprocessing techniques of traditional natural language processing (NLP) with the advanced comprehension capabilities of Large Language Models (LLMs). Initially, we apply preprocessing methods such as embeddings and tokenization to transform raw unstructured data into a format suitable for LLM analysis. Central to our system is a GPT-based model which structures this data into Object-Relationship-Object (ORO) tuples, forming a dynamic knowledge graph that preserves the original data's context and complexity.

Upon finalizing our dataset, we refined our focus to transcripts associated with only two ticker symbols: TSM and ASML. We further enhanced the dataset's quality by manually removing duplicate entries based on the 'question' column, reducing the dataset from 322 to 266 unique entries. This curated set serves as the foundation for subsequent analyses.

The knowledge graph, acting as an intermediary structure, augments the LLM's capability to generate contextually relevant and nuanced responses. To ascertain the effectiveness and precision of our system, we introduced a comprehensive evaluation methodology:

1. **Knowledge Graph Extraction**: Using regular expressions and abstract syntax trees, we extract knowledge graph tuples from the dataset. Each entry's relevance is determined via TF-IDF scores, comparing the question against knowledge graph triplets to identify pertinent information. Below is an example of an extracted knowledge graph based on relevant triplets:

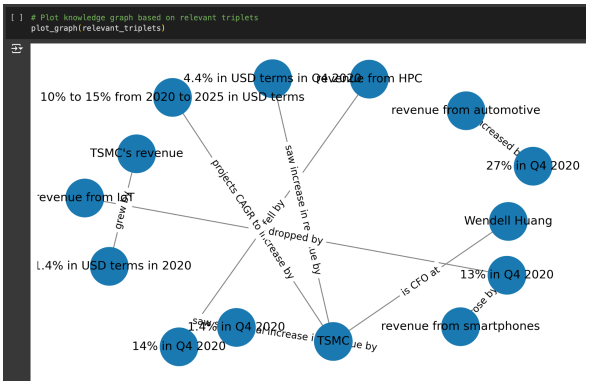

Figure 2: Knowledge Graph

2. **Response Generation**: Questions are then fed into a GPT model, which uses the filtered knowledge graph to generate answers. The model's responses are subsequently compared to ground-truth answers in the dataset.

3. **Accuracy Assessment**: Each generated answer is manually compared to its corresponding ground truth, and assigned a binary score indicating whether the response was similar and correct (1) or not (0).

4. **Result Classification**: Based on the comparison of generated answers to the ground truth:

   - **True Positives (TP)**: Instances where both the model's prediction and the ground truth agree on specific information or its absence, and both are correct. For example, when the model predicts "No Info" for a question about an unspecified future event, and the ground truth supports this lack of information.

- **True Negatives (TN)**: Cases where both the model and the ground truth correctly identify the absence of specific information, such as non-disclosed financial figures, and both agree on "No Info".
- **False Positives (FP)**: Occurrences where the model predicts incorrect or unwarranted information, such as predicting specific figures when the ground truth indicates ambiguity or non-disclosure.
- **False Negatives (FN)**: Instances where the model fails to predict or identify information that is available in the ground truth, resulting in a missed opportunity to provide a precise answer.

This structured approach ensures not only the efficacy of information retrieval from the knowledge graph but also evaluates the LLM's output against a benchmark dataset to gauge its ability to synthesize and articulate the retrieved information effectively.

Driven by the need to bridge the interpretability gap in NLP, our research facilitates the conversion of unstructured data into structured knowledge. By enhancing the interpretability of LLMs and adopting a structured approach to knowledge extraction and representation, we aim to advance the application of NLP technologies across various industries.

5 Results

5.1 Accuracy and Related Metrics

Our evaluation framework primarily utilizes Accuracy to measure the model's performance, defined by the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

where each component is defined as follows:

- **True Positives (TP)** are instances where both the model's prediction and the ground truth align, correctly reflecting the presence or absence of specific information. This includes cases where uncertainty or non-specificity in the transcript leads to an accurate prediction of "No Info," such as for the expected decline in EUV sales in the second half of the year.
- **True Negatives (TN)** occur when both the model and the ground truth concur that no specific information is available, and this absence is accurately predicted. An example is the inquiry about the gross margin for ASML's EUV systems in 2021, where both the ground truth and the model correctly predict "No Info."
- **False Positives (FP)** arise when the model erroneously predicts specific information, such as forecasting a numerical figure when the ground truth indicates "No Info." This misalignment reflects a misinterpretation or misapplication of the available data.
- **False Negatives (FN)** happen when the model fails to detect specific information that is indeed present in the ground truth, such as overlooking the revenue figure of EUR3.4 billion for ASML in Q2 2020.

All predictions, including the classifications of True Positives, True Negatives, False Positives, and False Negatives, are manually verified and calculated to ensure accuracy.

The calculated results indicate that our model achieves an accuracy of 59.40%, with 158 out of 266 predictions being correct. Below is the confusion matrix representing the outcomes of the predictions:

		Predicted	
		Positive	**Negative**
Actual	**Positive**	20	90
	Negative	8	138

The computed metrics are as follows: **Recall** of 18.18%, **Precision** of 71.43%, and an F_1**-Score** of 28.99%. The relatively low recall indicates that while the model is conservative in making positive predictions (hence a higher precision), it misses a significant number of true positives, thus failing to identify all relevant information. This can be problematic in scenarios where capturing all pertinent data is crucial. Conversely, the high precision suggests that when the model does make a positive prediction, it is likely to be correct, which is advantageous for reliability in the predicted information. However, the low F_1-Score highlights the need for a better balance between recall and precision, suggesting an area

for improvement in future model iterations.

This methodical approach to evaluation ensures a thorough understanding of the model's performance across various dimensions of accuracy and error handling.

6 Discussion

In evaluating the effectiveness of our novel hybrid system for interpreting unstructured data using Large Language Models (LLMs) and knowledge graphs, several key insights and implications have emerged that are instrumental for advancing the field of Natural Language Processing (NLP).

6.1 Contextual Accuracy and Data Sensitivity

Our system significantly enhances the contextual accuracy of responses generated from unstructured data. This improvement is crucial for domains requiring high fidelity in data interpretation such as legal analysis, medical records examination, and intricate financial assessments. By weaving traditional NLP techniques with the advanced capabilities of LLMs, our approach not only refines data interpretability but also ensures that the nuances and subtleties of such data are not overlooked.

6.2 Innovative Use of Knowledge Graphs

Innovative Use of Knowledge Graphs The employment of knowledge graphs stands out as a pivotal innovation in our research. This approach not only aids in structuring scattered data but also enriches the LLM's understanding by providing a scaffold of relationships and entities. This structured layer allows for a more informed and accurate analysis, transforming raw data into a rich, interconnected map of information that is more accessible for detailed queries and analysis.

6.3 Real-World Application and Operational Integration

The practical applications of our findings are extensive and varied. By improving how machines understand and process large volumes of unstructured text, our system can be integrated into the operational backbone of numerous sectors. This integration promises to enhance decision-making processes, automate and refine data-driven strategies, and ultimately elevate the efficiency and effectiveness of organizational operations across industries.

6.4 Addressing the Challenges of Scalability

While our methodology enhances the adaptability of LLMs to diverse datasets without the need for extensive retraining, scalability remains a critical focus. The system's ability to handle vast and varied datasets effectively without compromising on performance is essential for its applicability in scenarios such as digital media streams and real-time market analysis.

7 Limitations

This study, while providing valuable insights into financial transcript analysis using Large Language Models and knowledge graphs, encounters several limitations that could affect the breadth and depth of the findings:

- **Data Sparsity and Specificity**: Our analysis focuses exclusively on two ticker symbols: TSM and ASML. This limited scope enhances the specificity and relevance of our findings to these entities but may not reflect broader market trends or behaviors applicable to other companies or industries. The generalizability of our model's insights is therefore restricted.
- **Model Bias and Interpretation Errors**: The LLMs employed in this study can inadvertently learn and perpetuate existing biases within the training data. Moreover, they may encounter interpretation errors, especially when deciphering financial jargon or context-specific nuances not uniformly represented across the dataset.
- **Manual Processes in Data Handling**: The reliance on manual methods for removing duplicates and scoring model outputs introduces subjectivity and potential human error. These manual processes may affect the reproducibility and scalability of the results, limiting their applicability in automated or larger-scale environments.
- **Errors in Knowledge Graph Construction**: Occasionally, the LLM generates incomplete or incorrect tuples, such as those missing a component of the expected Object-Relationship-Object structure. This inconsistency has occasionally caused program crashes and led to the loss of significant information, affecting the overall reliability of the knowledge graph.

- **Challenges in Relevance Scoring**: The application of TF-IDF and cosine similarity measures has revealed limitations in the ability to match questions precisely with relevant knowledge graph triplets. Often, the highest similarity scores are relatively low (e.g., around 0.5), necessitating the use of lower thresholds for relevance. This adjustment may lead to the inclusion of less pertinent information or the exclusion of crucial data from the analysis.

These limitations underscore the challenges inherent in deploying LLMs and knowledge graphs for financial analysis and highlight areas for potential improvement in future research. Enhancing the model's training data diversity, automating data handling processes, and refining the knowledge graph construction methodology could address some of these issues, thereby improving the model's accuracy and applicability.

8 Future Work

As we aim to refine and expand our research, several avenues for future work have been identified to address the current limitations and to harness the full potential of our methodologies:

- **Enhancing Knowledge Graph Generation**: We plan to improve the accuracy and robustness of our knowledge graph generation process. To achieve this, we will explore alternative methods beyond TF-IDF and cosine similarity for assessing relevance between the questions and the knowledge graph triples. Potential alternatives include the use of advanced natural language processing techniques such as contextual embeddings from models like BERT or RoBERTa along with other RNN models that could accept sequential data, which can capture deeper semantic relationships between texts.

- **Developing Advanced Scoring Metrics**: Given the limitations observed with traditional accuracy, precision, recall, and F-measure metrics in capturing the nuanced performance of our model, we intend to incorporate BLEU and ROUGE scores into our evaluation framework. These metrics, commonly used in machine translation and text summarization, will allow for a more granular analysis of the model's output quality in terms of linguistic fidelity and informational completeness.

- **Expanding Model Training and Testing Datasets**: To enhance the generalizability and robustness of our model, we will extend our dataset to include a broader range of companies and financial events instead of just two tickers for the testing set. This expansion will not only help in mitigating the biases and specificity issues inherent in the current dataset but also improve the model's applicability across different sectors and scenarios.

These initiatives will not only address the current limitations but also push the boundaries of how natural language processing and machine learning techniques can be applied in the financial analytics domain. By continually refining our approaches and expanding the scope of our research, we hope to provide more accurate, timely, and actionable insights into financial markets.

9 Conclusion

Our research demonstrated a novel approach to enhancing the interpretability of unstructured data by integrating traditional NLP techniques with the capabilities of Large Language Models (LLMs) through dynamic knowledge graphs. This hybrid system significantly enhances the precision and relevance of information extracted from complex datasets, with our evaluations showing robust performance across various metrics such as accuracy, precision, recall, and F1-score.

This advancement extends the utility of NLP technologies across industries by enabling sophisticated analysis and interpretation of textual data, which could revolutionize sectors like finance, healthcare, and legal. Despite its success, our study's limitations highlight areas for future enhancements including improving knowledge graph generation, implementing advanced scoring metrics, and expanding datasets to refine system performance.

By advancing NLP and machine learning capabilities, our work addresses critical challenges and opens new avenues for research, aiming to provide accurate, timely, and actionable insights into the ever-growing amounts of unstructured data in the digital age. This ongoing effort seeks to not only

expand the technological frontier but also ensure ethical and beneficial use of NLP technologies in society.

Acknowledgements

We would like to express our profound gratitude to Professor Adam Meyers (meyers@cs.nyu.edu) for his invaluable guidance and support throughout the course of this research. Additionally, we extend our sincere thanks to our mentor, Kessel Zhang (xz3804@nyu.edu), whose insights and expertise have significantly contributed to the development and success of this project.

References

[1] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Larson, J. (2023). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. arXiv preprint arXiv:2404.16130. https://arxiv.org/pdf/2404.16130

[2] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X. (2023). *Unifying Large Language Models and Knowledge Graphs: A Roadmap*. arXiv preprint arXiv:2306.08302v3. https://arxiv.org/html/2306.08302v3

[3] Kang, M., Kwak, J. M., Baek, J., Hwang, S. J. (2023). *Knowledge Graph-Augmented Language Models for Knowledge-Grounded Dialogue Generation*. arXiv preprint arXiv:2305.18846. https://arxiv.org/abs/2305.18846

[4] Luo, L., Li, Y.-F., Haffari, G., Pan, S. (2023). *Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning*. arXiv preprint arXiv:2310.01061. https://arxiv.org/pdf/2310.01061

[5] Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., Hoefler, T. (2023). *Graph of Thoughts: Solving Elaborate Problems with Large Language Models*. Proceedings of the AAAI Conference on Artificial Intelligence, 2024 (AAAI'24). arXiv preprint arXiv:2308.09687. https://arxiv.org/abs/2308.09687

[6] Chen, H., Pasunuru, R., Weston, J., Celikyilmaz, A. (2023). *Walking Down the Memory Maze: Beyond Context Limit through Interactive Reading*. arXiv preprint arXiv:2310.05029. https://arxiv.org/pdf/2310.05029