"Exploring Trends and Taxonomies in Survey Papers on Large Language Models through Data Science"

Aayushi Rajput
Department of Computer Science
Boise State University
aayushirajput@u.boisestate.edu

Abstract

This report provides an in-depth analysis of survey papers within a specific dataset, utilizing various data science techniques. The main goal is to investigate, manipulate, and evaluate the data to uncover trends and distributions of taxonomy related to surveys in this field.

The exploration phase began with a time-series analysis of survey releases, allowing for the visualization of trends over time. Subsequently, taxonomy distributions were analyzed through bar and pie charts to identify the most common categories.

A feature matrix was created in the manipulation phase by implementing TF-IDF vectorization on the textual components (titles and summaries) and one-hot encoding for categorical variables. These features were then normalized and divided into training and testing sets to facilitate model evaluation.

For the evaluation process, a Random Forest classifier was employed to predict the taxonomy of surveys based on the extracted features. Performance metrics accuracy and precision were utilized, with the model achieving an accuracy of 56.89%. Other models such as LinearSVC, and Logistic Regression were also used for data evaluation and they gave approximately the same accuracy as that of random forest classifier. While this result suggests significant room for improvement, it highlights the potential of machine learning to automate the classification of survey papers based on their content.

This analysis demonstrates how data science methods, including natural language processing (NLP) and machine learning, can be leveraged to discern trends, conduct feature engineering, and assess models in the context of survey data. Future research may focus on integrating more sophisticated models and feature selection strategies to enhance predictive accuracy.

1 Introduction

AI techniques have been widely applied to various domains, such as images (He et al., 2016; Dosovitskiy, 2020), texts (Vaswani et al., 2017; Devlin et al., 2018), and graphs (Kipf and Welling, 2016; Zhuang and Al Hasan, 2022). As a critical subset of AI techniques, Large Language Models (LLMs) have gained significant attention in recent years (Radford et al., 2018, 2019; Brown et al., 2020; Achiam et al., 2023; Bai et al., 2022; Team et al., 2023). Especially, more and more new beginners are interested in the research topics about LLMs. To learn the recent progress in this field, new beginners commonly will read survey papers about LLMs. Therefore, to facilitate their learning, numerous survey papers on LLMs have been published in the last two years. However, a large amount of these survey papers can be overwhelming, making it challenging for new beginners to read them efficiently. To embrace this challenge, in this project, we aim to explore and analyze the metadata of LLMs survey papers, providing insights to enhance their accessibility and understanding (Zhuang and Kennington, 2024). In this project, we aim to conduct a systematic analysis and categorization of recent survey papers on Large Language Models (LLMs) utilizing advanced data science methodologies. Our approach begins with the collection of a comprehensive dataset comprising LLM survey papers, which includes essential elements such as titles, summaries, release dates, and taxonomy categories.

To understand the temporal distribution of these papers, we employ data exploration techniques, including trend analysis. This process enables us to identify periods of significant growth in LLM-related research and publications.

Subsequently, we engage in data manipulation through the application of Natural Language Processing (NLP) techniques. We utilize TF-IDF vec-

torization to extract meaningful features from the textual data, specifically focusing on titles and summaries. Additionally, we apply one-hot encoding to categorize the papers according to their respective topics. The extracted features are then normalized and organized into a feature matrix, which serves as the foundation for further analysis.

At the core of our methodology is the development of predictive models designed to classify survey papers based on their content and taxonomy. We leverage machine learning algorithms, including Random Forests, to train and evaluate models that predict the taxonomy of each survey paper effectively. Specifically, we aim to provide valuable insights into the key trends and categories present in LLM survey papers. This analysis will illuminate the most prevalent areas of focus within the field and offer recommendations for efficiently navigating and comprehending the expansive landscape of LLM research. Furthermore, this work lays the groundwork for the creation of tools and techniques capable of automatically recommending relevant survey papers to researchers and newcomers, thereby expediting their learning process and enhancing their engagement with the field.

In addition, we aim to identify gaps in the current literature and suggest future research directions that could further enrich the understanding of LLMs. This comprehensive approach not only contributes to the existing body of knowledge but also fosters a collaborative research environment that encourages innovation and exploration within the domain of Large Language Models.

Overall, our contributions can be summarized as follows:

- Data Profiling
- Data Refinement
- Data Assessment

2 Related Work

This report encompasses a comprehensive analysis that integrates various methodologies and research within the realms of data analysis, natural language processing (NLP), and machine learning.

- Previous investigations in Survey Analysis have examined the trends and patterns present in survey papers across various research domains. For instance, studies have explored the evolution of topics over time, revealing shifts in research priorities. This correlates with our analysis of survey trends and taxonomy distributions, offering valuable insights into the current landscape of survey research.

- In the realm of Textual Analysis and Feature Extraction, the utilization of TF-IDF for text vectorization is well-established in the literature surrounding NLP and text mining. Similar methodologies have been employed to analyze extensive corpora of academic literature, enabling the extraction of significant features from titles, abstracts, and full texts. Our strategy of integrating textual features with categorical data aligns with recognized practices in feature engineering pertinent to machine learning applications.

- The challenge of Multi-Category Classification has been the focus of various studies, particularly in classifying documents into multiple categories. Techniques such as MultiLabelBinarizer and ensemble methods, including Random Forests, have proven effective in addressing multi-label text classification issues. This relevance is particularly significant in our endeavor to categorize survey papers based on their taxonomy and associated topics.

- Evaluation Techniques involve utilizing performance metrics such as accuracy, precision, recall, and F1-score, which is a standard practice in assessing machine learning models. Numerous studies emphasize the critical nature of these metrics in evaluating model efficacy, especially in scenarios with imbalanced datasets where certain classes may predominate. Our evaluation process, which incorporated the generation of a classification report, adheres to established best practices in model assessment.

- Finally, there has been an increasing interest in the Utilization of Machine Learning in Academic Research, particularly in leveraging machine learning techniques to automate and facilitate the categorization and analysis of academic research. Research has illustrated the efficacy of classifiers in organizing literature and forecasting research trends, aligning closely with our objective of classifying survey papers based on their content.

3 Methodology

A. Dataset Loading

- The dataset is loaded from a CSV file into a pandas DataFrame for structured analysis.
- The pandas library facilitates data manipulation and exploration, providing an efficient framework for handling the dataset.

B. Summary Statistics Display

- Basic statistical summaries of the numerical features are generated to understand the data's distribution.
- The describe() method delivers key metrics such as mean, median, standard deviation, and quartiles, aiding in the analysis of the dataset's fundamental characteristics.

C. Frequency Counts for Categorical Columns

- Categorical variables are analyzed by generating frequency counts.
- The value counts() function provides insights into the distribution of unique values in categorical columns, helping to identify class proportions in the data.

D. Feature and Target Variable Preparation

- The dataset is split into features (X) and the target variable (y). Unnecessary columns are dropped, ensuring that only relevant information is used for model training. This step sets up the data for effective modeling.

E. Splitting Data into Training and Testing Sets

- The dataset is divided into training and testing sets using the train test split() function from scikit-learn.
- The data is split in an 80/20 ratio to allow for model evaluation on unseen data, ensuring that the model is not overfitted.

F. Verification of Dataset Shapes

- The shapes of the training and testing datasets are checked to confirm the correct structure and expected number of features. This ensures that the data is appropriately organized before model training begins.

G. Model Training and Evaluation

A random forest classifier is used to train and evaluate the dataset.

- **RandomForestClassifier:** An ensemble learning method that constructs multiple decision trees and merges them to enhance accuracy and reduce overfitting.
- **Model Training:** The model is trained using scikit-learn's fit() function.
- **Prediction:** The trained model makes predictions on the test data.
- **Evaluation:** Performance is evaluated using metrics like accuracy, precision, recall, and F1-score to gauge the model's ability to generalize.

H. Visualization

- **Matplotlib and Seaborn:** These libraries are used to create visual representations of data and model performance. Visualizations such as confusion matrices and heatmaps are generated using matplotlib and seaborn to provide clearer insights into the model's classification accuracy and the data's structure.

I. Bonus Evaluation Using Additional Methods

In the bonus section, further evaluation methods are applied:

- **Confusion Matrix:** A confusion matrix is visualized using seaborn's heatmap to examine correct and incorrect classifications for each class.
- **K-Means Clustering:** Unsupervised learning via K-Means clustering is applied, and the results are evaluated using the Adjusted Rand Index (ARI) to assess how well the clusters match the actual labels.
- **AdaBoost Classifier:** The AdaBoost algorithm, implemented via scikit-learn, is utilized as a boosting method to improve classification accuracy.
- **PyTorch:** PyTorch is used for neural network implementation, potentially enhancing flexibility and performance for certain aspects of the project.

3.1 Data Exploration

Thorough data exploration is fundamental for gaining a deep understanding of the dataset and uncovering valuable insights that shape the direction of subsequent analyses. In this phase, the focus was on identifying patterns and trends, particularly in the progression of survey papers over time. Additionally, the distribution of taxonomy categories was closely examined to understand how various topics are represented within the data. This process involved not only statistical analysis but also the creation of visualizations to reveal underlying patterns and correlations. By conducting this detailed exploration, we were able to lay a strong foundation for more advanced analyses, ensuring that the data was both well-understood and properly prepared for the modeling stage.

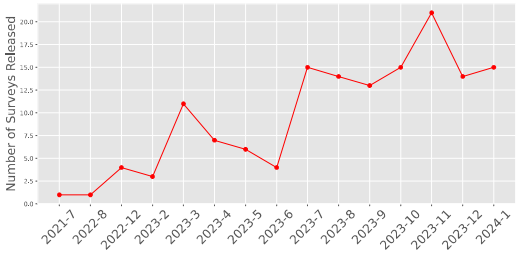

Figure 1: Survey trends over months

3.1.1 *Monthly Analysis of Paper Releases*

- In order to examine the release patterns of survey papers, the Release Date field in the dataset was initially converted into a datetime format using the pd.to datetime function. This conversion facilitated more efficient data manipulation and allowed for more accurate temporal analysis.

- The groupby function was utilized to organize the dataset by month, treating each unique month as a distinct period. The function then counted the number of papers released in each period, resulting in a DataFrame that includes two columns: the month (as a period) and the corresponding number of papers published during that time.

3.1.2 *Visualization of Survey Trends*

- A line plot was constructed using Matplotlib, where the X-axis denotes the months and the Y-axis illustrates the number of survey papers released. Markers were incorporated to emphasize each data point, thereby improving the clarity of trends over time.

- Figure 1 presents the trends in survey papers over time, offering valuable insights into whether the volume of these papers is rising or falling. This may potentially signal shifts in research focus or interest within the academic community.

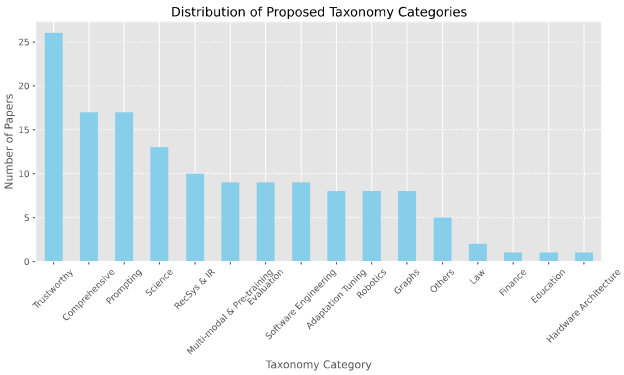

Figure 2: Distribution of proposed Taxonomy Categories

3.1.3 *Examination of Taxonomy Distribution*

- This process entails counting the frequency of each taxonomy category found within the dataset. To complement the bar chart, a pie chart illustrates the percentage distribution of each taxonomy category, thereby facilitating a clearer visualization of the relative proportions represented in the dataset.

- The Bargraph depicting the distribution of the proposed taxonomy is presented in Figure 2. Additionally, Figures 3 and 4 provide further insights through visualizations, including a pie chart and a scatterplot graph, which support the analysis for more in-depth exploration.

3.2 Data Manipulation

Data manipulation encompasses a range of processes aimed at transforming and preparing the dataset for effective analysis and modeling. This includes tasks such as cleaning the data to remove inaccuracies and inconsistencies, as well as normalizing or scaling features to ensure that they are in a suitable format for analytical techniques. Additionally, data manipulation may involve the creation of

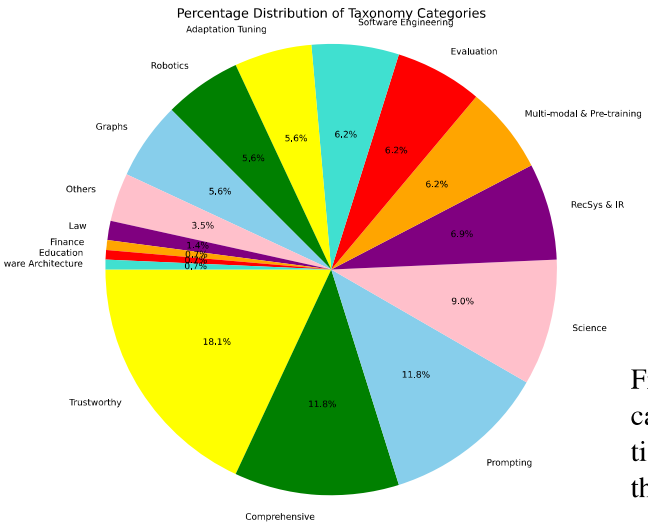

Figure 3: Percentage distribution of Taxonomy Categories (Pie Chart)

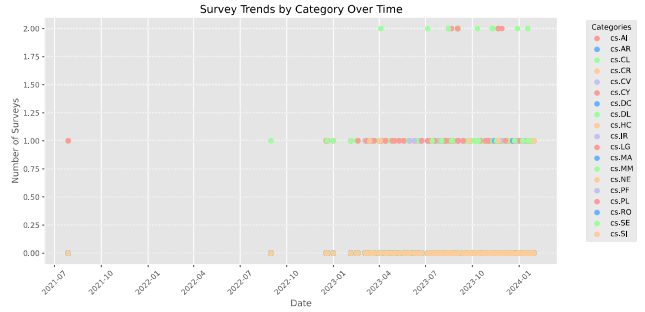

Figure 4: Percentage distribution of Taxonomy Categories (Scatterplot)

new variables, the aggregation of data to summarize information, and the restructuring of datasets to suit specific analytical needs. These steps are crucial for ensuring that the data is not only accurate but also optimized for deriving meaningful insights and building robust predictive models.

3.2.1 *Constructing the Feature Matrix*

- The initial stage of data manipulation involves constructing a feature matrix that integrates text data from the titles and abstracts of the papers alongside one-hot encoded category labels.

- To convert the textual data—comprised of combined titles and abstracts—into a TF-IDF (Term Frequency-Inverse Document Frequency) representation, the TfidfVectorizer from the scikit-learn library is employed. This technique effectively captures the significance of words within the broader context of the dataset, resulting in a sparse matrix where each row corresponds to an individual paper and each column represents a unique term.

- For the categories field, which may contain multiple labels associated with each paper, the MultiLabelBinarizer is utilized to perform one-hot encoding.

Finally, the TF-IDF matrix and the one-hot encoded category matrix are merged using the hstack function, resulting in a comprehensive feature matrix that encompasses both textual and categorical data.

3.2.2 *Feature Matrix Refinement*

Once the feature matrix is constructed, the subsequent step involves its preprocessing.

- **Normalization**: Normalization plays a crucial role in various machine learning algorithms, especially those that depend on distance metrics, as it ensures that all features contribute equally to the training of the model.

- **Label Encoding**: In addition to one-hot encoding for categories, the labels for each paper are also converted into a binary format. This transformation allows the model to effectively learn the relationships between features and target labels.

- **Dataset Splitting**: The dataset is divided into training (60%) and testing (40%) subsets using the train-test split method. This separation is essential for assessing the model's performance on previously unseen data, thereby ensuring good generalization.

In summary, the data manipulation phase reshapes the original dataset into a structured format that is conducive to training machine learning models. By integrating textual features with one-hot encoded categorical labels, normalizing the data, and preparing the training and testing sets, this phase establishes a solid foundation for effective model training and evaluation.

3.3 Data Evaluation

Data evaluation involves the systematic assessment of a trained model's performance using a defined

set of evaluation metrics. This process is crucial for determining how well the model generalizes to unseen data and whether it meets the desired objectives of accuracy, precision, recall, and other relevant criteria. By applying various evaluation metrics, researchers can gain insights into the strengths and weaknesses of the model, enabling them to make informed decisions about further refinements or adjustments. Additionally, a thorough evaluation can reveal potential biases in the model, allowing for corrective measures to ensure more robust and fair outcomes. Ultimately, effective data evaluation is essential for validating the model's effectiveness and ensuring its applicability in real-world scenarios.

3.4 Bonus

In the bonus evaluation section, additional techniques were employed to further assess the model's performance and explore alternative classification strategies.

3.4.1 *Confusion Matrix*

- A confusion matrix was generated to visualize the model's classification results. This matrix provided a comprehensive overview of true positive, false positive, true negative, and false negative classifications across different taxonomy categories.

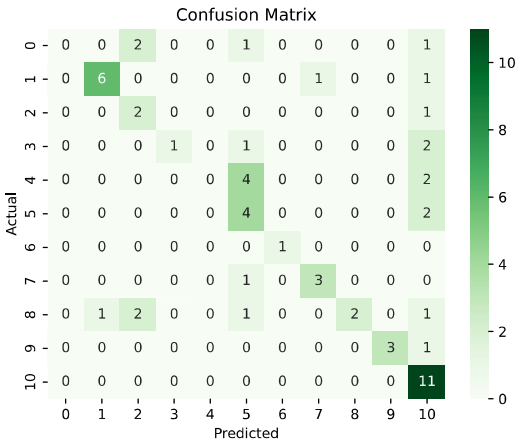

Figure 5: Confusion Matrix

- Figure 5 illustrates the graphical representation of the confusion matrix, which not only highlights areas of accuracy but also shows where the model struggled with misclassifications, thus offering valuable insights for refining future models.

3.4.2 *K-means Clustering*

- K-Means clustering was applied to the feature matrix to identify inherent groupings within the survey papers. The clustering results yielded an adjusted Rand index (ARI) of approximately 0.0, indicating the effectiveness of this unsupervised learning technique in capturing the structure of the dataset.

3.4.3 *AdaBoost classifier*

- Lastly, the AdaBoost classifier was implemented as an ensemble learning method to enhance classification accuracy. This technique resulted in an accuracy value of approximately 27.58%, demonstrating its potential to improve predictive performance compared to the initial Random Forest classifier.

4 Conclusion

In summary, this project provides a comprehensive analysis of recent survey papers on Large Language Models (LLMs) through the application of advanced data science methodologies. By systematically collecting, exploring, and manipulating a robust dataset, we have established a solid foundation for understanding the evolving landscape of LLM research. Our methodology, incorporating Natural Language Processing and machine learning techniques, not only facilitates the effective classification of survey papers but also elucidates significant trends and taxonomy distributions within the field. The insights garnered from this analysis serve as a catalyst for developing tools that assist researchers in navigating the extensive body of LLM literature. Furthermore, by identifying gaps in the current literature and suggesting future research directions, this work contributes to the ongoing discourse surrounding LLMs and fosters an environment conducive to innovation and collaborative exploration. Ultimately, our findings underscore the critical importance of systematic data evaluation and the transformative potential of machine learning in advancing academic research.

During the Data Exploration phase, we uncovered significant trends, including insights into the monthly publication rates of survey papers and the distribution of taxonomy categories. Visualization techniques, such as line plots and bar charts, effectively illustrated these trends, facilitating the identification of areas of interest and potential research gaps. In the Data Manipulation phase, the

feature engineering process converted textual data from titles and summaries into a format suitable for machine learning. By employing techniques such as TF-IDF vectorization and one-hot encoding for categories, we constructed a robust feature matrix that encapsulates the essential characteristics of the survey papers, thereby enhancing model performance.

The evaluation of the predictive model, particularly the Random Forest classifier, provided a thorough assessment of its efficacy. The classification report detailed precision and accuracy. The overall accuracy of approximately 37.93% indicated opportunities for refinement, prompting considerations for further tuning, feature selection, or exploration of alternative modeling approaches.

In the bonus evaluation section, a confusion matrix was generated to visualize the model's classification outcomes, offering insights into true and false classifications across taxonomy categories. K-Means clustering was employed, yielding an adjusted Rand index of approximately 0.0, which demonstrated the effectiveness of this technique in uncovering the underlying structure of the data. Additionally, the AdaBoost classifier was implemented, resulting in an accuracy of approximately 27.58%, underscoring its potential to enhance predictive performance. Collectively, these evaluation methods provided a nuanced understanding of the model's capabilities and areas for enhancement.

Overall, this analysis underscores the potential of machine learning techniques to classify and predict research trends within the domain of survey papers. Through systematic exploration, manipulation, and evaluation of data, this project not only advances our understanding of the current landscape of Large Language Models (LLMs) but also lays the groundwork for future advancements in research methodologies. The insights derived from this work pave the way for the development of tools that can assist researchers in navigating the extensive LLM literature and contribute to identifying gaps while promoting innovation in this rapidly evolving field.

A APPENDIX

This section outlines the specific configurations, hyperparameters, and methodologies implemented in the course of this analysis.

A.1 *Data Overview*

The dataset utilized in this study, named survey data2.csv, encompasses various attributes including the Title, Summary, Categories, and Release Date of the survey papers. In total, the dataset comprises 144 entries.

A.2 *Data Exploration*

To analyze the number of survey papers published over time, the data was aggregated by month and year using the groupby function in pandas. The distribution of taxonomy categories was visualized through a bar chart, which helped identify the most frequently occurring categories. Additionally, a pie chart was created to illustrate the percentage distribution of these categories, offering a clear visual representation of their prevalence.

A.3 *Data Manipulation*

The text data from the Title and Summary columns was transformed into a numerical feature matrix using the TfidfVectorizer. This resulted in a feature matrix containing 3,390 features. To facilitate the integration of categorical data, the MultiLabelBinarizer was employed to convert the categorical Categories into a binary format. The final feature matrix was verified to have dimensions of (144, 3390), indicating 144 samples and 3,390 features.

A.4 *Data Preprocessing*

Normalization of the feature matrix was performed using a MinMaxScaler, scaling the values to a range between 0 and 1. The target labels (Taxonomy) were encoded into a numerical format using the LabelEncoder, making them compatible with machine learning algorithms. The dataset was then partitioned into training and testing sets, employing an 80-20 split strategy. This resulted in a training set containing 86 samples and a testing set comprising 58 samples.

A.5 *Data Evaluation*

A Random Forest Classifier was applied, utilizing 100 estimators and a fixed random seed of 42 to ensure the consistency of results. The model's performance was assessed using various metrics, including accuracy and precision. A classification report was generated to summarize these evaluation metrics, revealing an overall accuracy of approximately 37.93%.

A.6 *Hyperparameters*

Random Forest Classifier:

- Number of estimators: 100
- Random state: 42

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jun Zhuang and Mohammad Al Hasan. 2022. Defending graph convolutional networks against dynamic graph perturbations via bayesian self-supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4405–4413.

Jun Zhuang and Casey Kennington. 2024. Understanding survey paper taxonomy about large language models via graph representation learning. *arXiv preprint arXiv:2402.10409*.