Exploring LLM Research Trends and Insights for Enhanced Accessibility

Shaznin Sultana
Department of Computing- Cyber Security Emphasis
Boise State University
shazninsultana@u.boisestate.edu

Abstract

The field of Large Language Models (LLMs) has seen rapid development, with numerous survey papers being published with progress. This technical report presents an exploration and analysis of recent survey studies on LLMs. As LLMs are gaining increasing attention, beginners are mostly relying on survey papers to understand the advancement of this area. However, the immense number of survey papers published in recent years pose a challenge to newcomers. With the goal of facilitating more accessible learning, this work investigates the statistics of these survey articles. The report covers data exploration, manipulation, visualization, and evaluation of key metadata elements such as taxonomy, release dates, and categories. Different techniques are employed to pre-process the dataset and machine learning techniques are applied to analyze the data to offer a comprehensive understanding of the dataset. Finally a classification of survey papers based on taxonomoy is carried out using Logistic Regression classification model. The aim is to provide insights into which areas of LLMs research have been emphasized, how publication trends have evolved, and how the content of survey papers is structured.

1 Introduction

The field of LLMs has grown rapidly over the past few years, with significant advancements in natural language processing (NLP) and artificial intelligence (AI) (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). LLMs, such as GPT-3, BERT, and more recently GPT-4, Gemini have transformed various industries, enabling applications ranging from chatbots to automated content generations (Radford et al., 2018, 2019; Brown et al., 2020; Achiam et al., 2023; Bai et al., 2022; Team et al., 2023). However, staying updated with the latest developments in this field can be challenging. The vast number of survey papers published in recent years can be overwhelming and difficult for beginners to efficiently find and absorb the necessary information. This project seeks to address this issue by exploring and analyzing the metadata of LLM survey papers published in the last three years. The goal is to uncover patterns and trends that can enhance the accessibility and understanding of these papers, making it easier for beginners to identify key papers and understand their content.

The focus of this project is on exploring key metadata elements such as the taxonomy, titles, authors, release dates, categories, and summaries of LLM survey papers from (Zhuang and Kennington, 2024). It is intended to perform data exploration and manipulation using statistical and visualization techniques to analyze trends in publication rates and the distribution of taxonomies. Understanding the extent of the advancements achieved in LLM research can potentially be significantly simplified by assessing the number of papers published periodically for each taxonomy, the trends of particular categories, the intersections between various LLM studies concerns, and other factors.

Overall, in this report, the following contributions are made:

- Exploration of Metadata Trends: Trends in survey paper publications over time, is analyzed including the distribution of papers across months and years. Visualizations are used to highlight key periods of increased or decreased activity.
- Analysis of Taxonomy and Category Distribution: The distribution of proposed taxonomies within the survey papers, is examined identifying which categories of LLMs have received the most attention.
- Analysis of Authors contribution: Discovering the evolution of the authors involvement, associations and interests trends over the years.

- Feature-Matrix creation: Unstructured data are vectorized using TF-IDF to reveal key terms and topics discussed in the papers which provides insights into recurring themes. Structured data are binary encoded and merged with processed unstructured data to create a improved feature matrix.
- Pre-processing for Machine Learning: The dataset is pre-processed by normalizing the data, encoding categorical labels, and splitting the data for training and evaluation.
- Taxonomoy Classification: Machine learning model logistic regression is utilized to classify the data, uncovering relationships between features such as time, titles, categories, summary, and authors.

2 Methodology

In this section, the approach taken to explore, manipulate, visualize, and evaluate the metadata of LLM survey papers is outlined. The goal of this process is to extract meaningful insights to enhance the accessibility of these papers. consists of metadata from numerous survey papers on LLMs. The collected dataset includes information such as the paper's taxonomy, title, authors, release date, links, paper ID, categories, and a brief summary. There are 144 papers and 16 taxonomies. This metadata serves as a basis for analysis aimed at showing the trends to facilitate better understanding for beginners.

The methodology can be divided into some key stages such as data exploration, manipulation, visualization, evaluation.

2.1 Data Exploration and Visualization

The initial phase of the project involves exploring the dataset to gain a comprehensive understanding of its structure and key trends. The following libraries are used: pandas, numpy, Counter, WordCloud, matplotlib, and seaborn. The following explorations are conducted:

- Number of papers by year and month: The papers are grouped by their release dates, specifically by year and month, to analyze trends in the publication of LLM survey papers over time. This helps to track the growth of research in the field.

  Using describe() and info(), the basic statistics of the dataset are reviewed, including numerical distributions, data types, and the presence of missing data. For instance, the mean value of the number of surveys per month is 9.6.

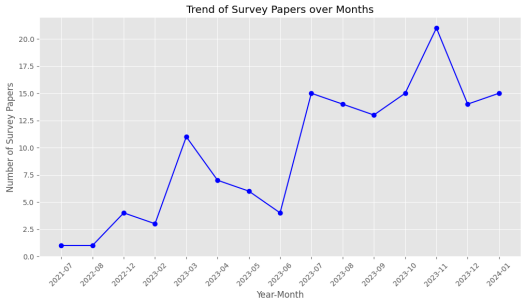

Figure 1: Number of papers by year and month.

- Histograms and Barplots: Using $value\_counts()$, histograms and barplots are generated to visualize the distribution of survey papers across different categories and taxonomies in 2-5. This is essential to understanding how different categories are represented and how papers are distributed across different taxonomies. For instance it gives a sense of how research is spread across different taxonomies in a continuous-like manner in the dataset. For instance, the Trustworthy taxonomy has 26 survey papers and it is the highest among others.

  Histograms are ideal when looking for distributions or continuous frequency data to identify trends, and outliers in the dataset. In this case, the focus is on how frequently papers are published over time or distributed across taxonomies.

  Barplots are most effective when comparing discrete categories. In this case, they offer an easy-to-interpret comparison of the number of papers across various categories and taxonomies. It is important to compare the count of papers across distinct categories and taxonomies, for identifying which categories are more prominent in LLM survey papers.

- Different counting methods: During the data exploration process, a discrepancy is found between the number of papers when comparing different counting methods such as $value_counts()$ and $groupby$. Using $value\_counts()$ on the Taxonomy column returns 144 entries, while using $groupby('Taxonomy')['PaperID'].count()$ results in 134 entries. This difference is

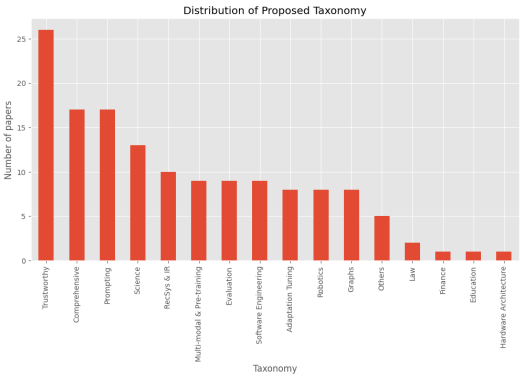

Figure 2: Taxonomy using Bar plot

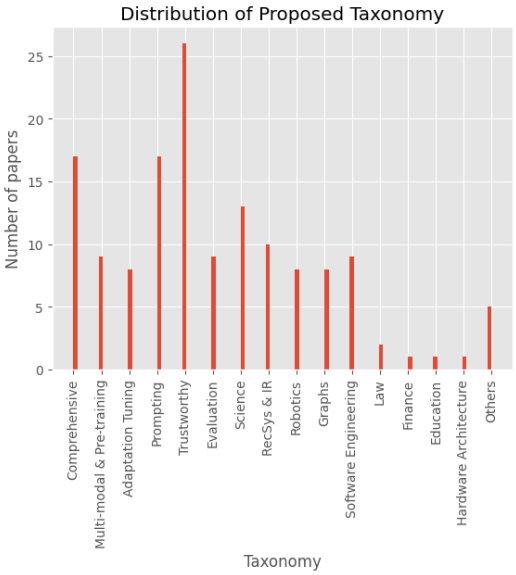

Figure 3: Taxonomy using Histogram

due to the presence of missing or null values in the Paper ID column. The method $value\_counts()$, counts all entries where a "Taxonomy" value is present, regardless of whether the Paper ID is missing. On the other hand, $groupby('Taxonomy')['PaperID'].count()$ only counts rows where both Taxonomy and Paper ID are present, skipping rows with null Paper ID. The discrepancy of 10 entries is caused by rows where Taxonomy values existed but the Paper ID is missing. This observation highlights the importance of handling missing data for accurate analysis.

11 shows the null values in the dataset. It can be seen from the 6 that "Paper ID" and "Categories" has null values. It is important in data pre processing.

- Number of Papers per Taxonomy Over Time:

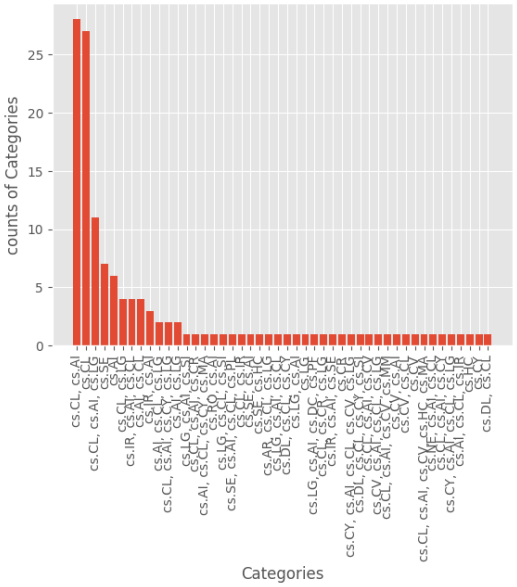

Figure 4: Categories using Bar Plot

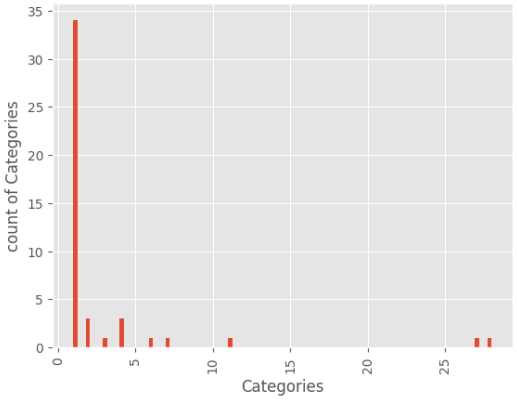

Figure 5: Categories using Histogram

To examine the distribution of papers across different taxonomies over time, the number of papers published for each taxonomy in each time period is calculated. This provids insights into shifts in research focus. The trends are then visualized in 7 using a time series plot, where each line represents a taxonomy, showing the publication patterns over time.

Each taxonomy is distinguished using seaborn's color palette. Also, the number of papers for each taxonomy is annotated on the plot for non-zero values, making it easier to observe significant changes and peaks in publication trends.

- Cumulative Frequency of Papers: A histogram of the cumulative frequency of paper categories is generated to visualize the increas-

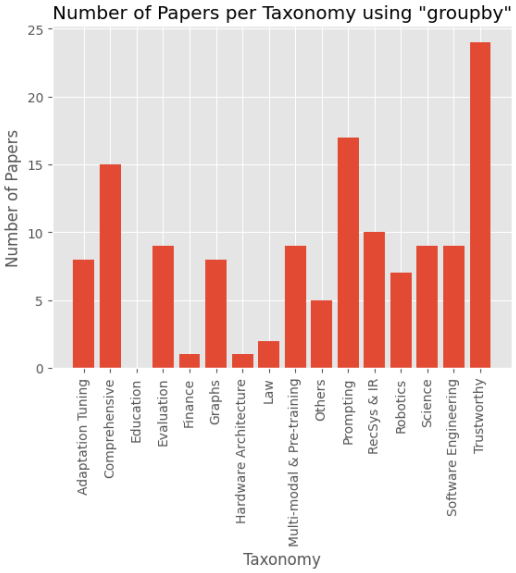

Figure 6: Taxonomy using groupby

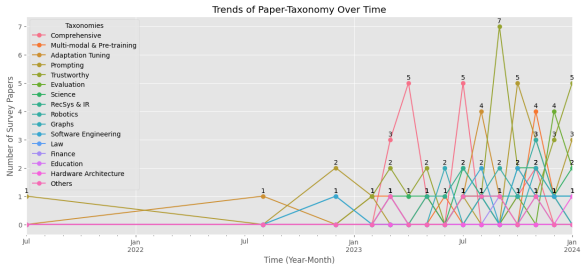

Figure 7: Paper count per taxonomy trend

ing volume of publications over time. This is done by grouping the dataset by Year-Month and summing the total number of papers published up to each time period. The plot in 8 is a clear depiction of how the volume of publications has expanded in recent years. This reflects the growing interest and contributions to this area.

- Author Contributions in Papers: Authors with multiple contributions are identified to understand the influence and contribution patterns in the field of LLM research. The plot in 9 can give an overview of the Authors who are contributing more than one paper. Average number of papers per author: 1.123400365630713. Author with maximum number of papers: Philip S. Yu, with 8 papers. This can help identify the key contributors or researchers in the field.

- Word Cloud: Word Cloud is used to visually represent the most frequent terms found in

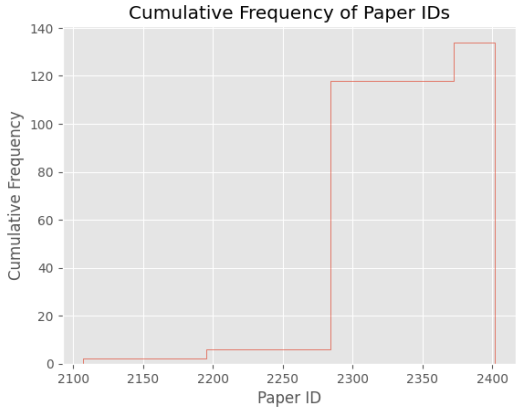

Figure 8: Cumulative Frequency of categories using Histogram

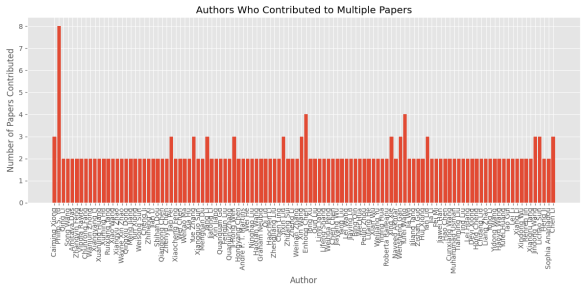

Figure 9: Author Contributions in Paper

the paper titles or summaries. This gives a high-level view of the key topics and terms discussed across all survey papers.

2.2 Data Manipulation

Once the dataset is explored, several data manipulation techniques are applied to prepare it for model training. For the manipulation tasks, sci-kit learn library is used to import TfidfVectorizer, OneHotEncoder, MinMaxScaler, LabelEncoder, and $train\_test\_split$.

- One-Hot Encoding: Categories are one-hot encoded to convert the categorical variables into a numerical format suitable for machine learning models. This process converts categorical values into a binary matrix format where each unique category is represented as a separate column. Each row in the new matrix has binary values to indicate a paper belongs to a given category. This allows the model to treat each category independently.

- TF-IDF Transformation: The titles and summaries of the papers are transformed using Term Frequency-Inverse Document Frequency (TF-IDF) to capture the importance

Taxonomy	0
Title	0
Authors	0
Release Date	0
Links	0
Paper ID	10
Categories	10
Summary	0
Year	0
Month	0
Year-Month	0

Figure 10: Null value counts

Figure 11: Important words in Word Cloud

of words in each document, while mitigating the influence of frequently occurring but less informative terms. This improves the model's ability to distinguish between papers based on content by filtering out common terms and focusing on key words that help classify papers more accurately.

- Feature Matrix Construction: After one-hot encoding and TF-IDF transformation, the features are merged to form a complete feature matrix, which represents both the categorical and text data. This allows the model to learn from both structured and unstructured data and impacts in classification accuracy.

- Normalization: The feature matrix is normalized using MinMaxScaler to ensure that all features are on a consistent scale (usually 0 and 1), and no single feature dominates depending on its value.

	0	1	2	3	4	5	6	7	8	9	...	_cs.CV	_cs.HC	_cs.LG	_cs.MA	_cs.NI	_cs.PF	_cs.PL	_cs.SI	_cs.LG	_cs.MA
0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	False	False	False	False	False	False
1	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	False	False	False	False	False	False
2	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	False	False	False	False	False	False
3	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	True	False	False	False	False	False	False	True	False	False
4	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
139	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	False	False	False	False	False	False
140	0.336449	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	False	False	False	False	False	False
141	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	False	False	False	False	False	False
142	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	False	False	False	False	False	False
143	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	False	False	False	False	False	False	False	False	False	False

144 rows × 3766 columns

Figure 12: Feature Matrix

- Label Encoding: The taxonomy labels are encoded into numerical form using label encoding as models generally require numerical input. This allows the model to differentiate between various taxonomies and classify papers accordingly.

- Train-Test Split: The dataset is split into training and testing sets using a 60-40 ratio ($TEST\_RATIO = 0.4$). This is for model training on one portion of the data while reserving the other portion for evaluation.

2.3 Data Evaluation and Results

After constructing the feature matrix, for model evaluation, Logistic Regression as the classification algorithm is employed to categorize LLM survey papers based on Taxonomy. Logistic regression is chosen for its simplicity and effectiveness in handling multi-class classification tasks. The classification matrix in 13, is generated to show how well the model predicts each class in the test data. The overall accuracy of the Logistic Regression model is calculated to assess its performance in correctly classifying the data. From the 13, accuracy is calculated 40%.

A confusion matrix is generated to provide a more detailed evaluation. The heatmap in 14 shows true positives, true negatives, false positives, and false negatives. The cells along the diagonal are the number of correct predictions for each class which is "True Positives". A high value in these cells indicates the model correctly predicts many instances of that class. The other cells represent errors. For instance, "False Positives" values which are when the model incorrectly predicts a class that it shouldn't have, are rows other than the diagonal. Similarly, "False Negatives" values which are when the model fails to predict a class that it should have, are columns other than the diagonal. In the heatmap, "True Negatives" are not explicitly shown like other values. This is because the confusion matrix only focuses on the actual vs. predicted counts for each specific class. Also, for multiclass classification, it is more abstract. True Negatives represent cases where the model correctly predicted the negative class. In other words, it correctly identified an instance as not belonging to a certain class. So, these values for a class are the sum of all cells

outside the row and column corresponding to that class.

Through this methodology, a comprehensive analysis is conducted, facilitating the discovery of trends and patterns in LLM survey papers and helping beginners better navigate this complex field.

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	4
1	1.00	0.38	0.55	8
3	0.50	0.67	0.57	3
5	0.00	0.00	0.00	4
8	0.00	0.00	0.00	6
10	0.44	0.67	0.53	6
11	1.00	1.00	1.00	1
12	1.00	0.25	0.40	4
13	0.00	0.00	0.00	7
14	1.00	0.25	0.40	4
15	0.28	1.00	0.44	11
accuracy			0.40	58
macro avg	0.48	0.38	0.35	58
weighted avg	0.42	0.40	0.32	58

Figure 13: Classification Report

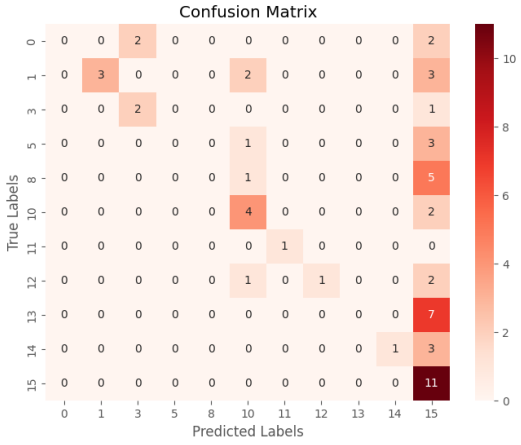

Figure 14: Confusion Matrix

3 Conclusion

This research provides a deep investigation and analysis of the metadata of LLM survey articles, with the purpose of making it simpler for newcomers to explore the vast amount of available material. Using machine learning models to explore, manipulate, visualize, evaluate, and preprocess data provides further insights into the associations between different metadata sections, providing a data-driven approach to understanding the construction of LLM survey papers. This analysis not only highlights important patterns but also serves as a foundation for developing tools or resources that can make LLM literature more accessible to newcomers. Future work could focus on building recommendation systems or search tools to further enhance the efficiency of reading and understanding LLM survey papers. Besides the exploration already done, analysis of citation networks and geographical distribution of authors could be added. This will help analyze the impact of each survey paper. Papers with higher citation counts can be identified as influential in the field. With the authors' geographical information, we will be able to discover which countries or regions contribute the most survey papers and what could be the reasons. Finally the findings can be added to the dataset to enrich its quality.

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*.