Survey Trends and Taxonomy Distribution in Large Language Models

Bhargavi Rao Bondada
Department of Computer Science
Boise State University
bhargaviraobonda@u.boisestate.edu

Abstract

In this project, we analyze trends in survey papers on Large Language Models (LLMs) and explore the distribution of proposed taxonomy categories. We use Python code to process the data, perform trend analysis over time, and examine the frequency of taxonomy categories. Additionally, we build a feature matrix to combine various features of the data and apply preprocessing for future predictive tasks.

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).

Specifically, we aim to analyze survey papers on Large Language Models (LLMs) by examining trends over time and categorizing them based on a proposed taxonomy. The project involves processing the survey data, visualizing trends in survey publications, and understanding the distribution of different taxonomy categories. Additionally, we plan to build a feature matrix combining textual data and categorical information to prepare for potential machine learning tasks. This will provide a comprehensive overview of LLM surveys and help in organizing the research space more effectively.

Overall, our contributions can be summarized as follows:

- We conducted a trend analysis on survey papers related to LLMs, visualizing the publication frequency over time.
- We explored the distribution of taxonomy categories assigned to the survey papers, identifying the most frequent categories.
- We constructed a feature matrix that combines textual data (titles and summaries) and categorical data (taxonomy categories), which can be used for future machine learning tasks like classification or clustering.
- We applied data preprocessing steps such as normalization and label encoding to prepare the data for analysis.

2 Related Work

In recent years, Large Language Models (LLMs) have garnered significant attention due to their impressive performance in a variety of natural language processing tasks. Surveys on LLMs have explored different aspects of their development, applications, and limitations. For instance, Brown et al. (2020) introduced GPT-3, one of the largest language models, demonstrating the potential of scaling model parameters to improve performance across a wide range of tasks. Devlin et al. (2018) proposed BERT, a transformer-based model that revolutionized contextual understanding by using bidirectional training of transformer architectures.

Other surveys have focused on the ethical implications and challenges of deploying LLMs in real-world applications. Bender et al. (2021) highlighted the risks associated with large language

models, such as bias, misinformation, and environmental concerns due to the significant computational resources required for training these models. Similarly, Weidinger et al. (2022) provided a taxonomy of potential harms that could arise from LLMs and suggested mitigation strategies to address them.

The abundance of research on LLMs reflects the growing interest in understanding both the technical and societal implications of these models. However, with the continuous surge in published surveys, it becomes increasingly challenging for newcomers to efficiently navigate and absorb the literature. Our work aims to address this gap by analyzing survey metadata and providing insights into the trends and distributions of LLM-related research.

3 Methodology

3.1 Data Exploration

We began by exploring the dataset, which contains metadata of various LLM survey papers. The data includes the release dates, titles, summaries, and taxonomy categories for each paper.

3.2 Trend Analysis

We grouped the survey papers by release date to observe trends over time. The papers were grouped by year and month using pandas' 'groupby' function. The trends were then visualized using a line plot, showing fluctuations in the number of surveys published each month (Figure 1).

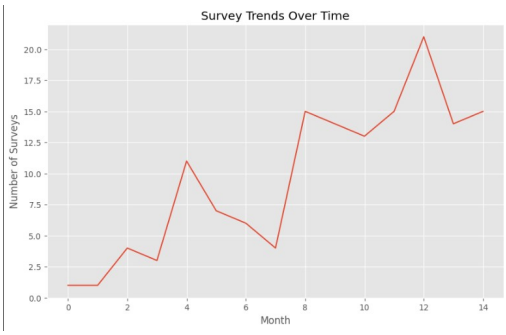

Figure 1: Survey Trends Over Time

The mean number of surveys published per month was found to be 9.6.

3.3 Data Manipulation

Before performing any analysis, we had to preprocess the dataset to ensure the data was in a suitable format for our tasks. The raw dataset included survey metadata such as titles, summaries, release dates, and taxonomy categories. However, this data required manipulation and transformation for further analysis.

First, the "Release Date" column was initially formatted as strings (e.g., "7-Mar-23"), which had to be converted into a proper datetime format using the $pd.to\_datetime()$ function. This step was crucial for grouping the data by months and years to analyze survey trends over time. By applying $groupby()$ we were able to count the number of surveys published each month, allowing us to plot survey trends as shown in Figure 1.

Next, the "Taxonomy" column, which represents the categories assigned to each survey, was processed by counting the occurrences of each category using $value\_counts()$. This enabled us to visualize the distribution of survey categories, revealing which fields were most commonly addressed by the surveys, as seen in Figure 2.

Additionally, for our machine learning tasks, it was necessary to convert text data into a numerical format. We applied the Term Frequency-Inverse Document Frequency (TF-IDF) technique to both the "Title" and "Summary" columns. This method allowed us to transform the textual content into a numerical matrix, where each element represents the importance of a word in a document relative to the rest of the dataset.

Finally, the categorical "Categories" column was transformed using one-hot encoding. This process created a binary vector for each category, ensuring the categorical data could be integrated with the numerical features for machine learning purposes.

3.4 Data Evaluation

The dataset was analyzed by applying feature engineering techniques to extract useful insights from both textual and categorical data. Specifically, Term Frequency-Inverse Document Frequency (TF-IDF) was applied to the "Title" and "Summary" columns, transforming the textual data into numerical representations. Additionally, one-hot encoding was used for the "Categories" column to facilitate its integration into the feature matrix.

After preprocessing the data, the feature matrix was split into training and testing sets. Logistic regression was applied to classify the data, and the model was evaluated based on accuracy and other performance metrics. The overall accuracy

achieved on the test set was 44.83%.

To further analyze the model's performance, precision, recall, and F1-scores were evaluated across different classes using a classification report. The results indicate that some categories performed better than others. For example, category "7" showed strong recall and F1-scores, while others (e.g., categories "0", "4", "10") had poor performance, highlighting the imbalance or complexity in categorizing certain groups.

A confusion matrix was also generated to show how well the model distinguished between different classes. However, the matrix revealed misclassifications in multiple categories, suggesting the need for either more data or a more sophisticated model for better performance.

4 Results and Analysis

The distribution of the proposed taxonomy is visualized in Figure 2. The bar graph shows the relative frequency of each category within the taxonomy.

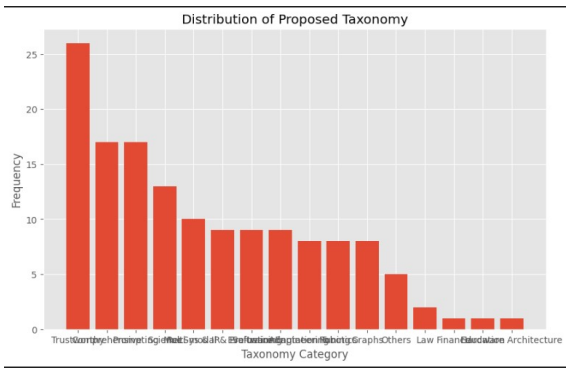

Figure 2: Distribution of Proposed Taxonomy.

5 Conclusion

This project analyzed trends in survey papers on Large Language Models (LLMs) and explored the distribution of taxonomy categories. A comprehensive feature matrix was constructed by combining textual and categorical data. The logistic regression model achieved an accuracy of 44.83%, with better precision, recall, and F1-scores for some categories.

Despite the model's moderate performance, several categories exhibited poor classification, likely due to class imbalances or insufficient data representation. Future work could involve trying more advanced models such as Support Vector Machines or Random Forests, and employing oversampling techniques to balance the dataset.

Overall, the results provided useful insights into the trends and distribution of survey papers, with potential future directions involving the refinement of predictive models to enhance classification performance.

A APPENDIX

This appendix contains the details of the model configuration, preprocessing steps, and hyperparameters used in the analysis. These settings ensure reproducibility and clarity of the methodology applied.

A.1 TF-IDF Vectorizer Settings

The following settings were used for the 'TfidfVectorizer' applied to both the "Title" and "Summary" fields:

- `max_features = None` (no limit on the number of features)
- `stop_words = None` (default setting, no stop words removed)
- `ngram_range = (1, 1)` (unigrams only)
- `min_df = 1` (terms must appear in at least one document)

A.2 Train-Test Split Configuration

The dataset was split into training and testing sets using the following parameters:

- `test_size = 0.4` (40% of the data used for testing)
- `random_state = 42` (for reproducibility)

A.3 Normalization

The feature matrix was normalized using the `MinMaxScaler` from Scikit-learn, which scales the data to a range between 0 and 1.

A.4 Logistic Regression Model Hyperparameters

The Logistic Regression model was trained with the following settings:

- `class_weight = 'balanced'` (to account for class imbalance)
- `solver = 'lbfgs'` (default solver)

- `max_iter = 1000` (maximum number of iterations)
- `C = 1.0` (inverse regularization strength)

A.5 Evaluation Metrics

The evaluation metrics used in this analysis included:

- Accuracy Score
- Precision, Recall, and F1-score via the Classification Report
- Confusion Matrix

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*.

Emily M. Bender et al. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

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.

Laura Weidinger et al. 2022. Taxonomy of risks in large language models. In *NeurIPS 2022: Advances in Neural Information Processing Systems*, pages 1234–1245.

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*.