Analyzing the Taxonomy of Large Language Models using Logistic Regression

G.I.C.L De Zoysa
Department of Computer Science
Boise State University
isurudezoysa@boisestate.edu

Abstract

Gathering and understanding the available survey papers with different categories is becoming more of a challenge with the rapid growth of the field of Large Language Models (LLMs). In this study, 144 survey papers are analyzed using a logistics regression classifier to predict the taxonomy category of the papers. According to the results, the logistic regression model accurately reflects the core trends within the collected data effectively and provides reasonable insight for the classification of the paper taxonomy. This approach might be helpful for researchers to organize their studies in a growing field of large language models. The results of the study show that the logistic regression approach is a reliable approach for taxonomy classification.

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; **?**; 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). According to Figure 1, there is a stunning trend of publishing papers over time from 2021 to 2024. Hence we aim to enhance the taxonomy prediction process by utilizing 144 collected survey papers related to LLM while comparing different classification approaches. In this analysis, the logistic regression classification was the main targeted classification technique to predict the related taxonomy category in advance. This analysis helps researchers and LLM enthusiasts to manage their studies effectively without any hassle.

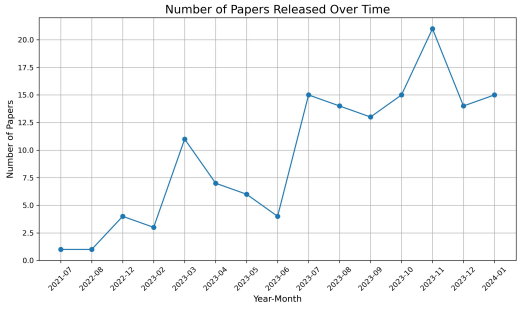

Figure 1: Number of Survey papers over time

Overall, our contributions can be summarized. In this taxonomy category prediction study, we developed a framework to classify available survey papers in the field of Large Language Models to improve the capacity of the researchers. Further, we used 144 survey papers, which included comprehensive information on LLMs and are helpful for further studies. As our major objective, we fitted a logistic regression model to classify survey paper taxonomy, which shows a greater ability to predict the taxonomy category by leveraging the pattern of the data while improving the accuracy of the classification. This investigation helps researchers find their road to analyzing the data related to LLMs through the massive amount of LLM literature.

2 Related Work

Qiheng Mao et.al addressed the clustering of words and documents using available algorithms. According to their new approach, document collection can be modeled as a bipartite graph. To build reasonable groups, they introduced a new co-clustering algorithm that utilizes a mathematical matrix of the document. Further, this algorithm represented good accuracy and performance while managing the graph partitioning issues. (Mao et al., 2024)

In another study conducted by Hichem Frigui and Olfa Nasraoui, an algorithm was proposed for clustering text documents via fuzzy c-means, which was computationally efficient and easy to handle. Also, annotated documents were clustered automatically through this proposed algorithm, and the capacity of the algorithm was tested with the real-world data set. (Frigui et al., 2002)

Zhao et al. focused on the significance of effective document clustering algorithms for managing large amounts of data into meaningful categories. A well-known method called hierarchical clustering is more efficient for data at different levels. This study compared agglomerative and partitional methods and found that the partitional algorithm performed better for large datasets. Furthermore, a new algorithm was introduced called constrained agglomerative, which was a combination of partitional and agglomerative algorithms and performed better in clustering large datasets over a single algorithm. (Zhao and Karypis, 2002)

A study by Dhillon et al. addressed the difficult task of clustering large, unbalanced text data. Here, they used a spherical k-means algorithm to cluster the data into groups. According to the results of their finding, the resulting clusters show "fractal-like" and "self-similar" behavior due to the high dimensionality of the text data. (Dhillon and Modha, 2001)

3 Methodology

This study mainly consists of three stages: basic data exploration, data manipulation, and data evaluation. In the basic data exploration step, we analyze the data set using basic statistical measurements and graphical methods to get a better understanding of the data set. In the second step, we apply data preprocessing and transformation techniques for further analysis of our data set. In the final step, we fit a logistic regression model for classifying the taxonomy categories and assess the results of our analysis.

3.1 Data Exploration

There were eight variables in the original data set with a size of 144 entries, including 'Taxonomy,' 'Title,' 'Authors,' Release Date,' 'Links,' Paper ID,' 'Categories,' and 'Summary.' The important variables for our analysis were the ' Taxonomy' and 'Category' of LLM-related papers both were categorical variables.

There were sixteen categories in the taxonomy variable. Then, these categories were visualized via a bar chart, Figure 1, to identify the distribution of the categories visually. The " Trustworthy" category was the highest recorded category, while "Hardware Architecture" was the least. Furthermore, there was an extreme imbalance in the category of distribution, which led to a challenge of taxonomy classification.

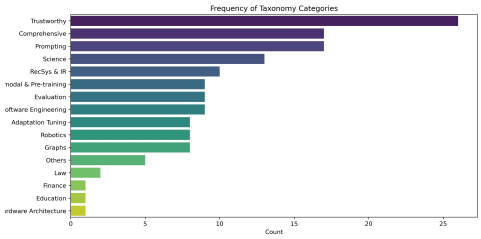

Figure 2: Distribution of categories of Taxonomy

According to the visualization of the 'Category' variable of the survey paper, Figure 2. The most frequent categories were CS.CL, CS.AI, and CS.LG indicated that most of the researchers used these domains in their LLM modeling.

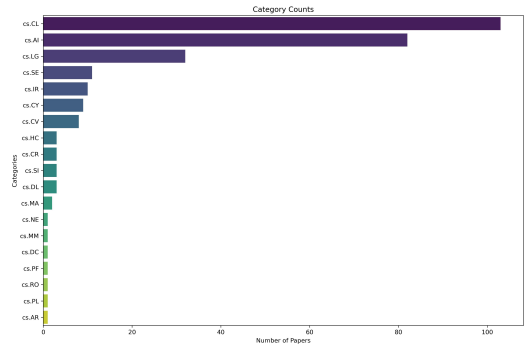

Figure 3: Distribution of Categories

It is important to study the trend of taxonomy categories over time. For this task, a time series plot was utilized. The highest number of papers in each taxonomy category were published at the be-

ginning of 2023. The 'Trustworthy' category was the most frequent taxonomy category, with more than 20 papers. With time, the number of papers published declined because researchers focused on other focus areas.

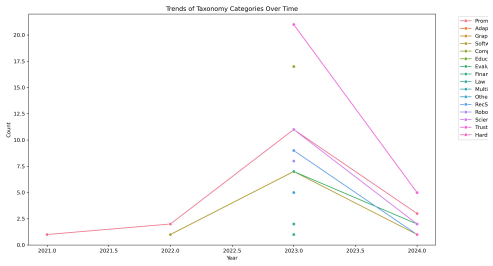

Figure 4: Trends of Taxonomy categories

In order to check the interrelationship between taxonomy categories, a chi-square statistical test was used. Using this test, the statistical independence of taxonomy categories can be evaluated. According to the results, 'Comprehensive' with 'Trustworthy', 'Prompting' with "Trustworthy', and 'Trustworthy' with 'Science' showed a relationship. According to this, researchers focused on these taxonomy categories together in their studies. To plot these results, the network plot was used.

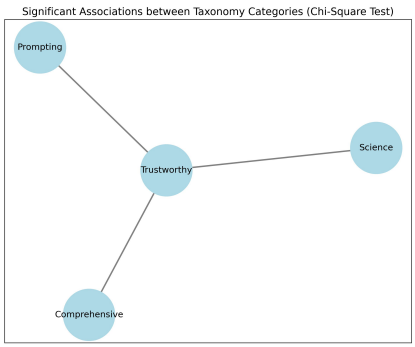

Figure 5: Network plot of the taxonomy categories

3.2 Data Manipulation

In this step, a feature matrix was developed to organize and structure the input data that models rely on to learn and make predictions, and then, data were normalized to facilitate further model building. Finally, the data set was divided into training sets and test sets with thresholds of 0.4 for model training purposes.

3.3 Data Evaluation

In the data evaluation step, a logistic regression model was fitted to classify the taxonomy categories. Initial accuracy was 22.4%, and then remedial procedures were applied to improve the performance of the model such as feature engineering techniques and addressing the class imbalance. Finally, the accuracy was 48.27%.

Step	Accuracy
Initial	22.41%
Feature Engineering	22.41%
Hyperparameter tuning	44.44%
Addressing Class Imbalance	48.27%

Table 1: Accuracy comparison of the logistic regression model

3.4 Further Analysis

Cluster analysis for Taxonomy categories was done to identify a potentially fewer number of categories among taxonomy categories sharing similar behavior. This may help researchers to identify potential research gaps as a new group of taxonomy categories unless a single category and direct the new research trends.

4 Cluster Analysis Results

We identified the following clusters and their corresponding taxonomy categories:

- **Cluster 1:** Education, Finance, Hardware Architecture, Law, Others
- **Cluster 2:** Adaptation Tuning, Comprehensive, Evaluation, Graphs, Multimodal & Pretraining, Prompting, RecSys & IR, Robotics, Science, Software Engineering
- **Cluster 3:** Trustworthy

5 Conclusion

In this study, we developed a method to automatically classify survey papers on Large Language Models (LLMs) into a structured taxonomy. We began by collecting metadata from 144 LLM survey papers and proposing a new taxonomy for classifying them. To achieve effective classification, we applied both graphical methods and a logistic regression model. Our analysis revealed that leveraging graph structure information in co-category graphs significantly enhanced the accuracy of taxonomy

classification. Moreover, the logistic regression model provided reliable results, complementing the insights gained from graphical methods.

A APPENDIX

Category	Number of Papers
Trustworthy	26
Comprehensive	17
Prompting	17
Science	13
RecSys & IR	10
Multi-modal & Pre-training	9
Evaluation	9
Software Engineering	9
Adaptation Tuning	8
Robotics	8
Graphs	8
Others	5
Law	2
Finance	1
Education	1
Hardware Architecture	1

Table 2: Frequency Distribution of Taxonomy Categories

Category Pair	Probability
(Comprehensive, Trustworthy)	0.084499
(Prompting, Trustworthy)	0.084499
(Trustworthy, Science)	0.162577

Table 3: Chi-square Significance Categories

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*.

Inderjit S Dhillon and Dharmendra S Modha. 2001. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42:143–175.

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

Hichem Frigui, Hichem Frigui, Hichem Frigui, H. Frigui, Olfa Nasraoui, and Olfa Nasraoui. 2002. Simultaneous categorization of text documents and identification of cluster-dependent keywords. *null*.

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*.

Qiheng Mao, Zemin Liu, Chenghao Liu, Zhuo Li, and Jianling Sun. 2024. Advancing graph representation learning with large language models: A comprehensive survey of techniques.

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

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.

Ying Zhao and George Karypis. 2002. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, page 515–524, New York, NY, USA. Association for Computing Machinery.

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*.