LLM Survey Paper Landscape: Predicting Taxonomies

Daniel Uranga
Department of Computer Science
Boise State University
danieluranga@u.boisestate.edu

Abstract

In this study, we analyze a dataset of survey papers on Large Language Models (LLMs) published over the last 3 years to gain insights into the current trends surrounding LLMs. Primarily we analyze the author landscape and the effectiveness at predicting the taxonomies of the surveys from their title, summary, and listed categories. I find that the amount of surveys released has increased drastically in the last three years. Also, most surveys have around 8 authors, but each author appears only on one survey usually. This indicates the research is spread widely between those in the field. Finally, our investigation into predicting taxonomies was a failure with the machine learning methods we applied. However, valuable insights about the dataset can be gained from the attempts.

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). By examing a database of LLMs survey papers from the last three years, we hope to provide information on current research trends and the author environment in the space. Specifically, we plan to show how the amount of LLM surveys has changed over time by analyzing the papers released by month and year. We also will provide data that shows the taxonomy distribution of LLM papers from RecSys & IR to Robotics, to see which areas are most popular. Also, we aim to uncover how papers are authored and if the majority of research is being done by a few authors or if the research is more spread out. The final goal of the analysis is to see if the taxonomy of a survey can be predicted by its title, summary, and categories through popular machine learning methods.

Overall, our contributions can be summarized as follows:

- The amount of survey papers released per month surrounding LLMs has increased drastically in the last three years
- The taxonomy distribution of survey papers is highly inbalanced towards trustworthy and comprehensive surveys
- Most papers have less than 10 authors with an average of a little over 8 per survey. However, there are outliers, with one paper having 67 authors.
- Most authors only are accredited with one paper, however there are many that have done two or three. One superstar, Philip S. Yu, stands out for being listed under 8 papers.
- The title, summary, and categories of a survey are not good predictors for the taxonomy of a paper based on our methods, however, our methods may be ill suited for the data. The speculate that the low accuracy could be due to lack of data or high class imbalance.

2 Methodology

2.1 Data Exploration

As seen in figure 1, the amount of surveys related to LLMs has increased rapidly from January 2023 to January 2024. Before January 2023, the activity is relatively quiet, with a large stretch of time between 2021 and 2022 where no surveys were released. For further research, it would interesting to see if this trend continues or whether it plateaus in the near future. It would also be interesting to see the release dates of papers that the surveys sited to see if the same peak can be seen in the reference papers but perhaps years earlier.

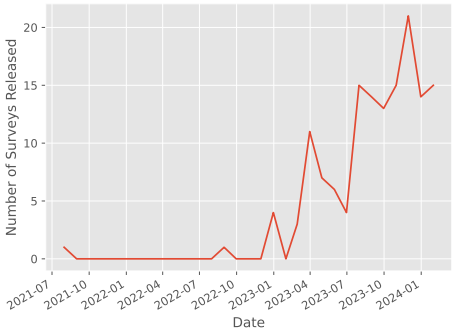

Figure 1: Survey trends from 2021 to 2024

When looking at the distribution of the surveys' taxonomies, there are some suprising results. Our prediction was that the comprehensive and prompting taxonomies were the most popular, however figure 2 shows that the trustworthy taxonomy is by far the largest group and 9 surveys more than both comprehensive and prompting taxonomies. Since we are discussing surverys, it makes sense that comprehensive surveys is one of the most common taxonomies. It is also interesting to look at the bottom end of the spectrum. There are some appearances from other areas of research other than computer science related feidls such as law and finance. This leads to the question of what other fields may have interests in LLMs and related technologies.

Another interesting aspect of LLMs surveys is the author landscape. A lot of the surveys have more authors than I first expected with only the first quartile having 5 or less authors, and 25% of surveys have over 10 authors. One extreme outlier even has 67 authors. The mean amount of authors for each survey paper is 8.44, which is skewed slightly high due to a few outliers.

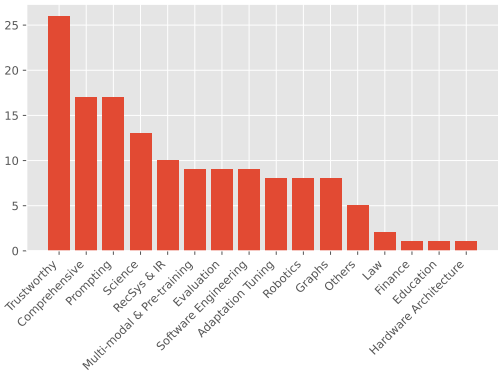

Figure 2: Taxonomies of LLM surveys

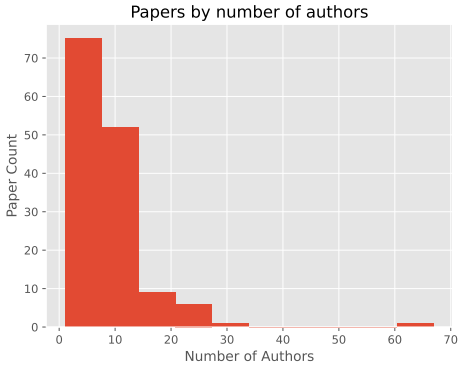

Figure 3: Number of authors per survey

Finally, another aspect of the author space is how many authors are credited for each paper, and if the majority of surveys are written by the same people. In fact, for the last statement, it is the opposite. Figure 4 shows that the majority of authors only write one paper. The average is 1.09 and the standard deviation is 0.37. However, there is one author, Philip S. Yu, who has appeared on eight survey papers. The next highest person has only contributed to four.

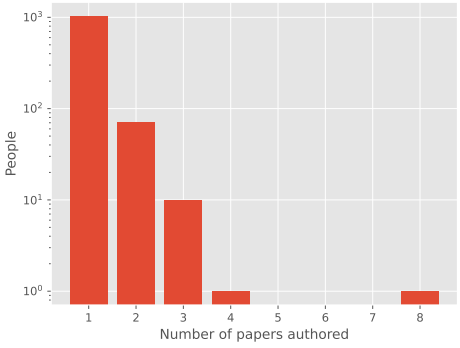

Figure 4: Authors of multiple papers

2.2 Data Manipulation

For the last part of our research on predicting taxonomy from title, summary, and category, it is necessary to manipulate our data in a way it can be easily processed. Most importantly, we need to vectorize the title and summary of each survey so they can be passed as vectors to our models. In this case we use TF-IDF vectorization provided by the Scikit package. In our implementation, there is no distinction between title and summary tokens. Next, we apply a one hot encoding to the categories given to each survey. Since each survey can have multiple categories, this is the best way to do it. Then we combine the title and summary vectors with the category vectors to complete our feature matrix.

We do apply normalization in the form of a min-max scaler to the feature matrix to increase the effectiveness of our machine learning techniques and inhance its accuracy. Also, since taxonomies are encoded by names we need to use the label encoder provided by Scikit Learn to encode each taxonomy as a number. The final step of prepping our data is to split the data into training and test sets. We settled on using 40% of the data for testing and 60% for training.

2.3 Data Evaluation

The first method that we tried was a bagging classifier. At first, we chose a random forest classifier, but decided against it because of our sparse feauture matrix. Then, we performed K-Fold cross validation to check the validity of the bagging method on the dataset. At this point it seemed like a good candidate, however, when fitting the data to the training set, we were unable to get higher than about 35% accuracy on the test set which was pretty poor. This could be due to the class imbalance, the low amount of data in our dataset, or the way in which we processed our data. We believe a random forest is not a good fit because of the possibility of selecting useless features for each item. If the feature matrix was denser, for example if we just used the categories instead of title and summary, then a random forest may be better approach.

Next, we moved on to a different approach which involved using a multi-level perceptron (MLP) classifier. This approach did must better on the data set which often was over 40% accurate on the test set with the highest accuracy achieved after training being 52%. Details of the model can be seen in the appendix. A 52% accuracy is still not good. In fact, it is probably worse than a human would do given just the title and summary. However, this also could mean that title, summary, and categories are not as good of predictors of taxonomy as one would think. This outlook does not seem likely to me. We think if there was more data, and less class balance, than the neural network approach would perform much better.

Bagging	**MLP**
29.3%	52.2%

Table 1: Highest achieved accuracy of method

3 Conclusion

Overall, the insights from this data exploration and evaluation will be helpful to both those who would like to learn more about LLMs and those interested in contributing to LLM surveys themselves. The amount of survey papers released per month surrounding LLMs has increased drastically in the last three years, and the taxonomy distribution of survey papers is highly inbalanced towards trustworthy and comprehensive surveys. Most papers have less than 10 authors with an average of a little over 8 per survey. However, one has up to 67 authors. Also, most authors only are accredited with one paper, however there are many that have done two or three. One superstar, Philip S. Yu, stands out for being listed under 8 papers. Out biggest find is that it is hard for the models we created to predict taxonomy from the title, summary, and categories of a survey, however, they could be poor indicators as well.

A APPENDIX

For the bagging classifier I used the default settings from the Scikit Learn bagging estimator which used ten estimators. For the K-Fold cross validation, I used six folds and the prediction accuracy as the score.

For the MLP, I used Adam as the solver, tanh as the activation, a learning rate of $1e-4$ and two hidden layers with sizes of 100.

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*.