Exploring and Evaluating Large Language Model Survey Paper Categories

Dylan Gresham *

Posted Date: 3 October 2024

doi: 10.20944/preprints202410.0226.v1

Keywords: machine learning; data science; large language models; graph representation learning

 doi:10.20944/preprints202410.0226.v1

Article

Exploring and Evaluating Large Language Model Survey Paper Categories

Dylan Gresham

Department of Computer Science, Boise State University; dylangresham@u.boisestate.edu

Abstract: In February of 2024, Dr. Jun Zhuang and Dr. Casey Kennington published a paper [1] in which they classified large language model (LLM) survey papers into different taxonomies utilizing graph learning. In this paper, I evaluate the dataset they created and used and propose that in its current state, there is not enough samples to classify other survey papers into specific categories.

Keywords: machine learning; data science; large language models; graph representation learning

1. Introduction

AI techniques have been widely applied to various domains, such as images [2,3], texts [4,5], and graphs [6,7]. As a critical subset of AI techniques, Large Language Models (LLMs) have gained significant attention in recent years [8–13]. 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 [1]. Specifically, we aim to determine if the dataset built for this purpose can be used to classify other survey papers into specific categories.

Overall, our contributions can be summarized as follows:

- In its current state, the dataset has a heavy class imbalance problem.
- The dataset, as it stands, is unable to appropriately classify survey papers into one or more categories.

2. Methodology

2.1. Data Exploration

To begin determining the dataset's capabilities for category classification, we first need to understand the data.

In Figure 1, we can see that there's a general upward trend in the number of survey papers as time goes by.

In Figure 2, we can see the distribution of the taxonomies in the dataset, with the largest taxonomy being Trustworthy and the smallest being a tie between Finance, Education, and Hardware Architecture.

In Figure 3, we can see a jointplot of the different taxonomies and the number of authors in each paper assigned to that taxonomy as well as the distribution of taxonomies along the top and the distribution of authors along the right side. From this, we can determine that the Trustworthy taxonomy has the paper with the most authors, several taxonomies have papers with just a single author, and on average, most papers in each taxonomy have around 4 authors.

Finally, in Figure 4, we can see that the vast majority of these papers have been labeled as being in the `cs.CL` and `cs.AI` categories.

 doi:10.20944/preprints202410.0226.v1

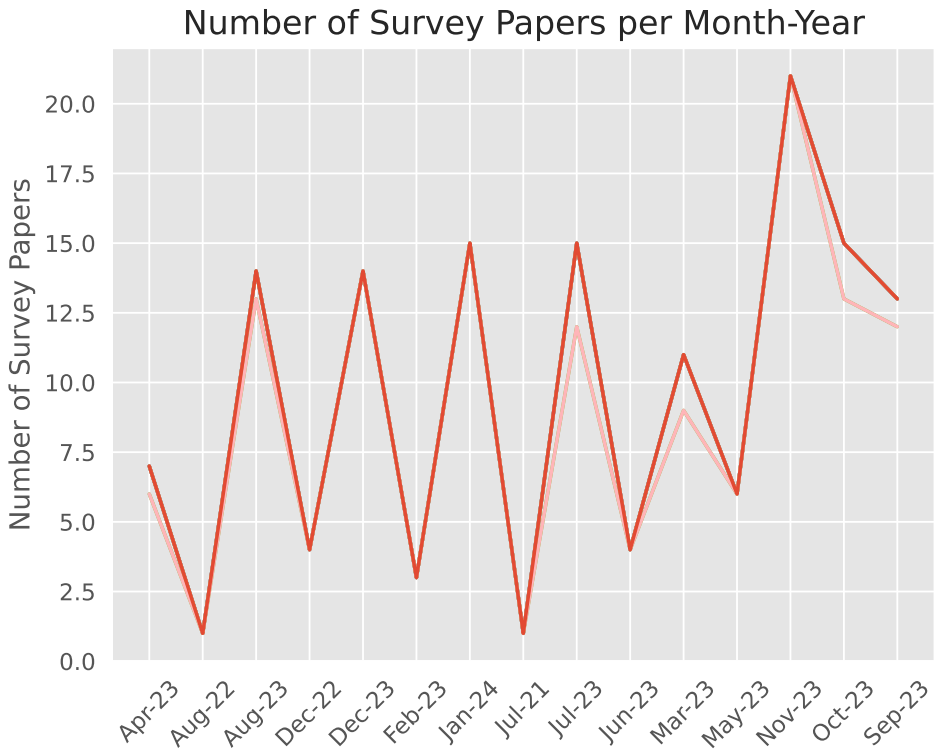

Figure 1. Number of survey papers per month from July 2021 to January 2024.

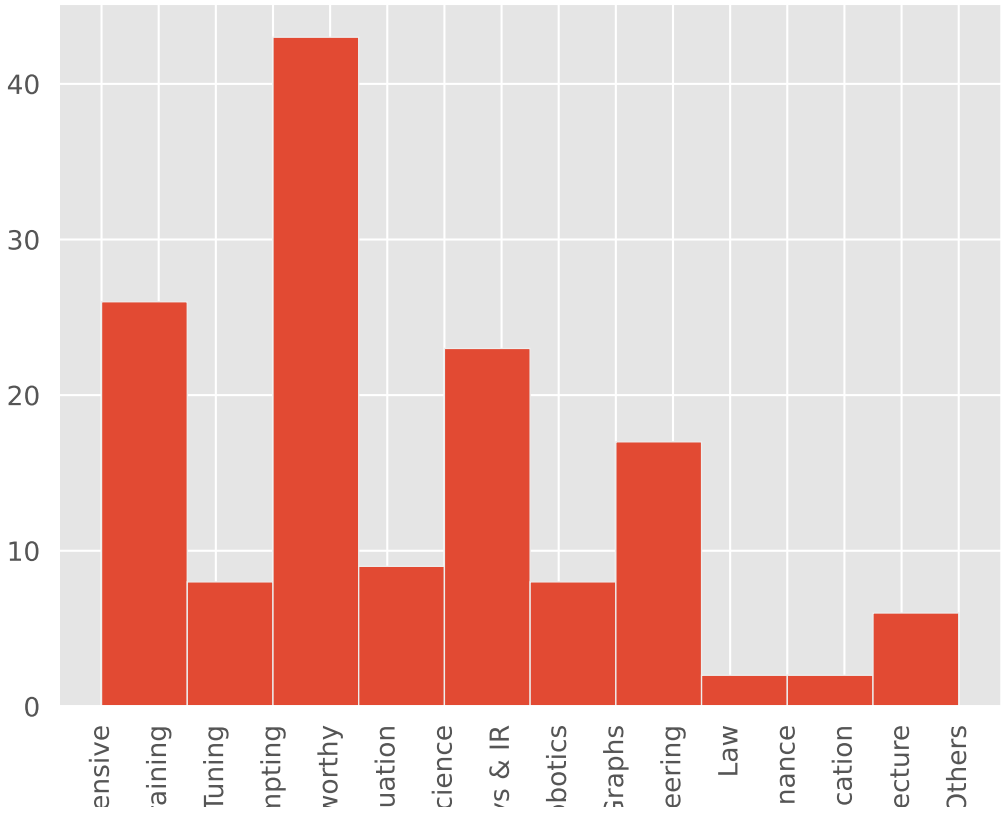

Figure 2. Distribution of paper taxonomies.

 doi:10.20944/preprints202410.0226.v1

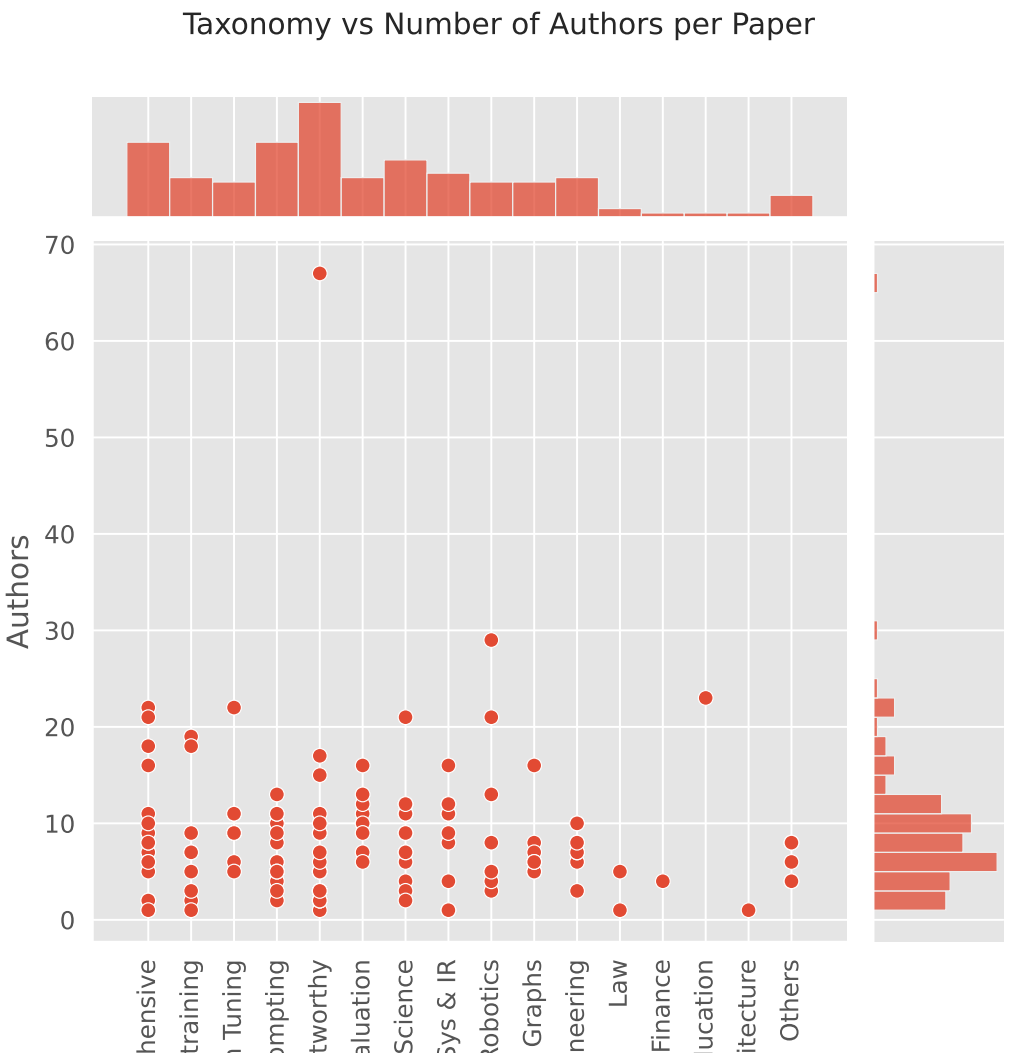

Figure 3. Jointplot of taxonomies and the number of authors in each paper.

From these four visualizations, we can see that there is a Taxonomy class imbalance with a few classes having substantially more samples than others, namely `Trustworthy` (26), `Comprehensive` (17), and `Prompting` (17) having the most and `Finance` (1), `Education` (1), and `Hardware Architecture` (1) having the least. Additionally, we can see that the category labels are heavily skewed to the right with many categories having as few as 1 paper in them and two categories having over 60 papers in them.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2024 doi:10.20944/preprints202410.0226.v1

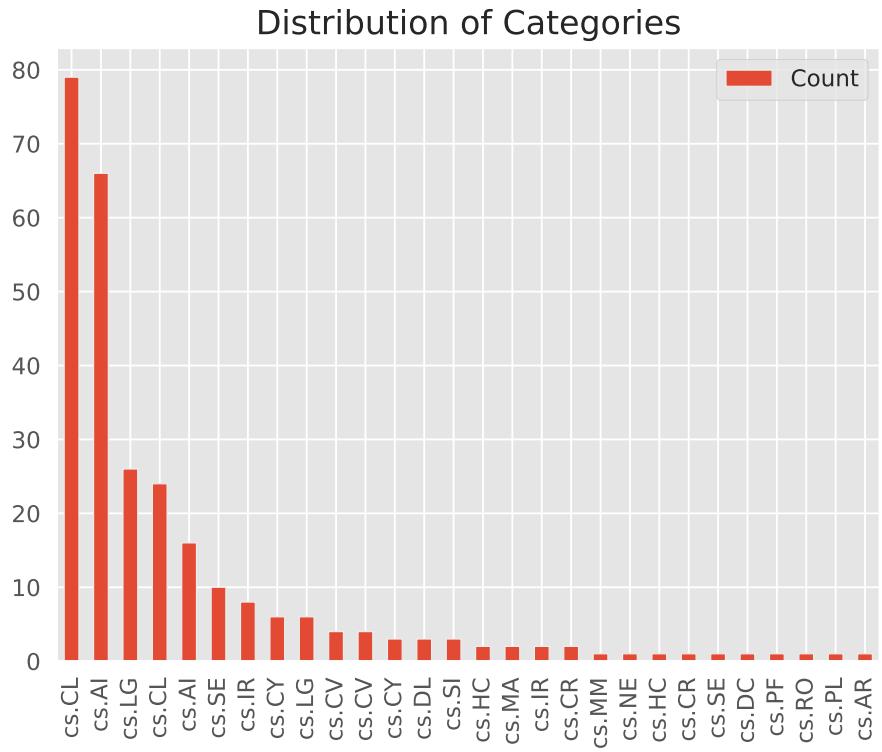

Figure 4. Bar plot of the distribution of categories.

2.2. Data Manipulation

To build the feature matrix for the eventual multi-class model, we decided to first vectorize the title and summary of each sample with a `TfidVectorizer` from `scikit-learn`. After vectorizing the two features, we then split the categories column of the initial dataset into individual columns, each column representing one unique category. Within each of those columns, the value per row is set to 1 if the paper was categorized into that category and 0 if it wasn't. In effect, this is a one-hot encoding of the categories.

Once the feature matrix has been made, we then normalized it using a `MinMaxScaler` from `scikit-learn` to normalize all of the vectorized title and summary features and then used this as our training and testing sample inputs. The categories were encoded using a `LabelEncoder` from `scikit-learn` and then stored as our training and testing sample outputs. The finalized dataset was then split into a 60/40 training to test ratio for use with the model.

2.3. Data Evaluation

To evaluate the dataset, we created a `OneVsRestClassifier` that uses a `LogisticRegression` model, both from `scikit-learn`, internally to perform the classifications. When running a prediction with our testing set, the model actually output three warnings saying that labels 6, 9, and 16 weren't present in all training examples. This is a relatively common error when the dataset being used for training and testing doesn't have enough samples or the classes are imbalanced to the point where one or more classes are only present in a handful of samples.

The classification report can be seen in Table 1 on the next page. The accuracy for this model was also calculated to be approximately 15.5172% using `scikit-learn`'s `accuracy_score` function.

 doi:10.20944/preprints202410.0226.v1

Table 1. Classification report generated using `scikit-learn`.

Category	Precision	Recall	F_1 Score	Support
cs.SI	0.00	0.00	0.00	1
cs.DL	0.00	0.00	0.00	0
cs.AR	0.00	0.00	0.00	0
cs.CL	0.67	1.00	0.80	39
cs.NE	0.00	0.00	0.00	0
cs.AI	0.54	0.84	0.66	32
cs.MM	0.00	0.00	0.00	1
cs.CR	0.00	0.00	0.00	1
cs.PF	0.00	0.00	0.00	0
cs.PL	0.00	0.00	0.00	1
cs.CY	0.00	0.00	0.00	1
cs.MA	0.00	0.00	0.00	1
cs.SE	0.00	0.00	0.00	4
cs.DC	0.00	0.00	0.00	0
cs.IR	0.00	0.00	0.00	1
cs.RO	0.00	0.00	0.00	0
cs.HC	0.00	0.00	0.00	3
cs.LG	0.00	0.00	0.00	14
cs.CV	0.00	0.00	0.00	6
micro avg	0.61	0.63	0.62	105
macro avg	0.06	0.10	0.08	105
weighted avg	0.41	0.63	0.50	105
samples avg	0.61	0.64	0.59	105

2.4. Potential Improvements

One potential improvement that could be made to this evaluation method is utilizing other machine learning techniques to work around the missing labels in all training examples warning.

After some searching, I've found that it is possible to do but, depends heavily on the dataset and models being used. Due to a lack of time, I was unable to test any of these methods but am hopeful that they may produce more positive results.

3. Conclusion

Based on the classification report and the accuracy score, we have concluded that this dataset is not a viable option for classifying LLM survey papers into categories.

However, this is not to say that this dataset isn't viable for other classification problems or for learning purposes. Based on these results, it is my belief that this dataset holds its value outside the realm of standard classification and more so in the realm of text summarization and similarity classification.

References

1. Zhuang, J.; Kennington, C. Understanding survey paper taxonomy about large language models via graph representation learning. *arXiv preprint arXiv:2402.10409* **2024**.
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
3. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
5. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.

 doi:10.20944/preprints202410.0226.v1

6. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* **2016**.
7. Zhuang, J.; Al Hasan, M. Defending graph convolutional networks against dynamic graph perturbations via bayesian self-supervision. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 4405–4413.
8. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; others. Improving language understanding by generative pre-training **2018**.
9. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; others. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
10. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; others. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
11. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; others. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* **2023**.
12. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; others. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* **2022**.
13. Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; others. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* **2023**.