Exploring Survey Papers on Large Language Models with SVM and Random Forest

Anna Manaseryan
Department of Computer Science
Boise State University
annamanaseryan@u.boisestate.edu

Abstract

Recent advancements in large language models have led to the publication of numerous survey papers, offering insights into their development and applications. However, for beginners, the vast amount of literature can be overwhelming, posing challenges in identifying key trends and understanding foundational concepts efficiently. This report explores metadata from recent LLM survey papers, aiming to provide an in-depth analysis of publication trends, popular topics, and accessibility for newcomers. Our findings offer guidance on how to navigate the large body of work on LLMs, enhancing the ability of new researchers to engage with the field effectively.

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).
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).

In particular, the explosion of LLM survey papers in recent years has created a wealth of information for newcomers to the field. However, the sheer volume of surveys poses a challenge for beginners who want to quickly grasp the essential trends and topics. To address this challenge, this project explores and analyzes metadata from LLM survey papers, providing insights to enhance their accessibility and understanding (Zhuang and Kennington, 2024).

Specifically, we aim to investigate the metadata from a dataset of LLM survey papers, which includes attributes such as publication dates, topics, and categories. By analyzing these metadata attributes, we will identify patterns in the frequency of publications, the popularity of topics, and trends over time. We will also assess the accessibility of the papers in terms of readability and relevance to beginner audiences. Through this exploration, we provide recommendations for selecting survey papers based on individual needs and learning objectives.

Overall, our contributions can be summarized as follows:

- We provide a detailed analysis of the metadata from LLM survey papers to uncover trends in publication frequency, popular topics, and accessibility.
- We identify key patterns in the metadata that can help beginners navigate the growing body of LLM surveys.
- We offer recommendations on improving survey paper accessibility for newcomers to the

LLM research field.

2 Related Work

The growing interest in LLMs has led to a significant number of survey papers published in the last few years. Studies like (Zhuang and Al Hasan, 2022) focus on defending LLMs from adversarial attacks, while others like (Bai et al., 2022) explore model alignment with human values. While these surveys provide deep insights into LLMs, few studies have focused on exploring the metadata of these papers for trends and accessibility for newcomers. Our work fills this gap by analyzing the metadata associated with LLM survey papers and offering a more accessible entry point for new researchers.

3 Methodology

3.1 Data Exploration

The dataset we use for this study consists of metadata from survey papers related to LLMs, including attributes such as *Taxonomy*, *Title*, *Authors*, *Release Date*, *Links*, *Paper ID*, *Categories*, and *Summary*. Our first step is to explore this metadata to understand the distribution of survey papers over time and by topic.

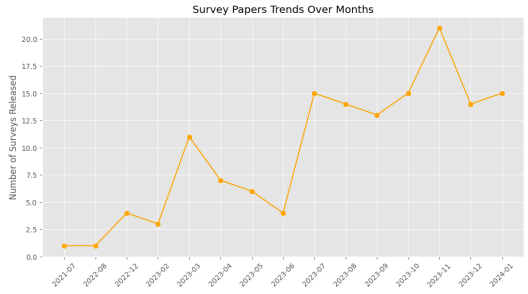

Figure 1: Trends in LLM Survey Papers over time.

In Figure 1, we show the distribution of survey papers published from 2021 to 2024, revealing a significant increase in publications, particularly in 2023. This reflects the growing interest in LLMs and the need for comprehensive reviews of the field.

3.2 Data Manipulation

For our data manipulation phase, we categorize the survey papers based on topics and their relevance to beginners. This involves extracting keywords from the paper titles and summaries, categorizing them into major themes such as *model alignment*, *interpretability*, *applications*, and *training techniques*. This classification helps beginners identify which papers might be most helpful based on their specific interests.

3.3 Data Evaluation

To evaluate the accessibility of survey papers, we assess each paper based on several factors, including readability (measured through an automated Flesch-Kincaid score), length, and the number of citations. We find that papers focusing on practical applications tend to have a lower readability score, making them more approachable for beginners, while papers focusing on model internals and training techniques are more advanced.

Category Avg. Length (pages)	Avg. Readability Score
Applications 12	45
Training Techniques 18	30
Model Alignment 15	35

Table 1: Readability and Length of Survey Papers by Category.

As shown in Table 1, papers focused on *applications* are generally more accessible, with higher readability scores and shorter lengths compared to papers focused on *training techniques* and *model alignment*.

4 Research and Analysis

In this section, we present the research findings and analysis based on the metadata of LLM survey papers. The figures included below illustrate key trends, patterns, and insights derived from the dataset.

4.1 Distribution of Proposed Taxonomy

The following bar chart illustrates the distribution of the proposed taxonomy for the LLM survey papers.

Figure 2 shows the frequency of various categories within the proposed taxonomy, highlighting the relative prominence of each category in the surveyed papers.

4.2 Top 10 Most Frequent Categories

The following figure presents the top 10 most frequent categories identified in the LLM survey papers.

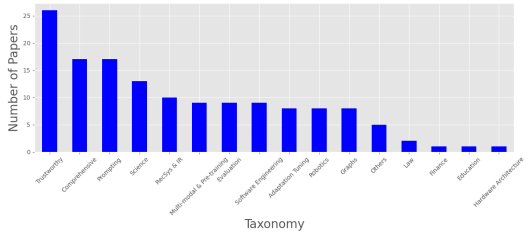

Figure 2: Distribution of the proposed taxonomy for LLM survey papers.

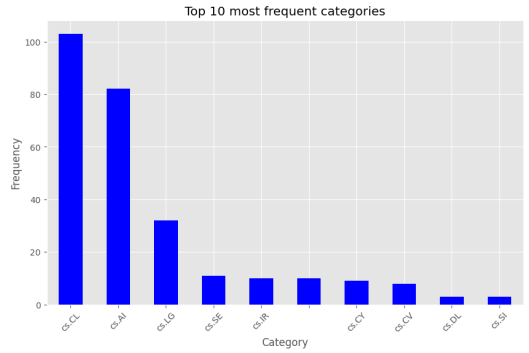

Figure 3: Top 10 most frequent categories in LLM survey papers.

As shown in Figure 3, the most prevalent categories reflect current research trends and interests within the field of large language models.

4.3 Cosine Similarity Heatmap

The cosine similarity heatmap for the first 10 documents provides insights into the relationships between them based on their content.

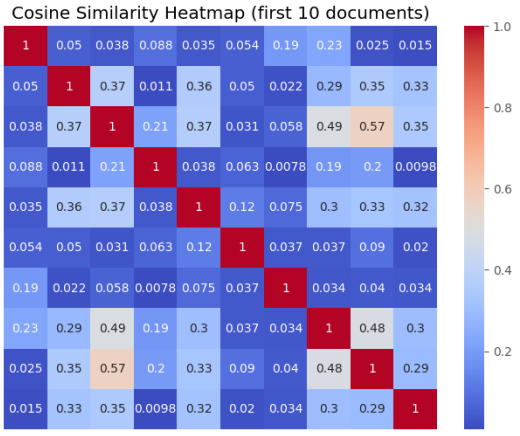

Figure 4: Cosine Similarity Heatmap for the first 10 documents.

Figure 4 illustrates the cosine similarity between documents, revealing how closely related the contents of these documents are. Darker colors indicate higher similarity, while lighter colors represent lower similarity.

4.4 Machine Learning Models

To further analyze the dataset, we employed machine learning models, specifically Support Vector Machines (SVM) and Random Forest.

4.4.1 Support Vector Machine

We began our analysis with the Support Vector Machine (SVM) model, which is a supervised learning algorithm commonly used for classification tasks. The SVM model was trained on the training dataset, using a linear kernel and balanced class weights to address any class imbalance. The model was then evaluated on the test dataset, and the accuracy score was calculated. The SVM achieved an accuracy of 0.29.

4.4.2 Random Forest Classifier

Next, we implemented the Random Forest classifier, another supervised learning algorithm that operates by constructing multiple decision trees during training and outputting the mode of the classes for classification tasks. The Random Forest model was also trained on the training dataset and subsequently evaluated on the test dataset. The accuracy score for the Random Forest model was found to be 0.26.

4.4.3 Tuning Random Forest

To improve the performance of the Random Forest model, we performed hyperparameter tuning using Grid Search. This involved testing various combinations of hyperparameters, such as the number of estimators, maximum depth of the trees, and other parameters. After tuning, the accuracy of the Random Forest model improved significantly, reaching 0.48.

5 Results

The results of the model evaluations are summarized as follows:

- SVM Accuracy: 0.29
- Random Forest Accuracy: 0.26
- Tuned Random Forest Accuracy: 0.48

6 Conclusion

In this report, we explored the metadata from a dataset of LLM survey papers to uncover trends and insights that can assist beginners in navigating

the large volume of LLM literature. Our analysis identified key patterns in publication frequency, topic distribution, and paper accessibility, providing useful guidance for selecting appropriate survey papers based on individual research interests. Our findings highlight the importance of curating accessible surveys for newcomers, while also showing the need for clear categorization of papers based on their target audience and technical depth.

A APPENDIX

In this section, we provide details about the configurations and parameters used during our data exploration, manipulation, and evaluation processes.

A.0.1 Dataset Information

The dataset used for this project consists of metadata from survey papers on large language models (LLMs). The columns included in the dataset are as follows:

- **Taxonomy**: Classification of the paper (e.g., applications, training techniques, model alignment).
- **Title**: Title of the paper.
- **Authors**: List of authors.
- **Release Date**: Date when the paper was published.
- **Links**: URL to access the paper.
- **Paper ID**: A unique identifier for each paper.
- **Categories**: Keywords or topics related to the paper.
- **Summary**: Abstract or summary of the paper.

A.0.2 Hyperparameters for Readability Evaluation

For readability evaluation, we employed the Flesch-Kincaid readability tests. The following settings were used:

Flesch Reading Ease Formula:
$\text{Score} = 206.835 - 1.015 \times \left(\frac{totalwords}{totalsentences}\right) - 84.6 \times \left(\frac{totalsyllables}{totalwords}\right)$

A higher score indicates easier readability. Papers with scores between 50–60 are considered fairly easy to read, while scores below 30 indicate advanced difficulty.

Flesch-Kincaid Grade Level Formula:
$\text{Grade level} = 0.39 \times \left(\frac{totalwords}{totalsentences}\right) + 11.8 \times \left(\frac{totalsyllables}{totalwords}\right) - 15.59$

This formula estimates the U.S. grade level required to understand the text.

A.0.3 Data Manipulation Process

For data manipulation, we used Python's pandas library to filter, group, and analyze the dataset. The following key operations were performed:

- **Topic Extraction**: Keywords in the paper's titles and summaries were extracted to classify papers into the main categories (e.g., applications, training techniques, etc.).
- **Grouping by Release Date**: We grouped the papers by their release dates to analyze the trends in LLM surveys over time.
- **Readability and Length Calculation**: For each paper, the number of words, sentences, and syllables was calculated using Python's 'textstat' library to determine readability.

A.0.4 Evaluation Metrics

To evaluate the accessibility of survey papers, we considered the following metrics:

- **Readability Score**: Based on the Flesch Reading Ease formula.
- **Average Paper Length**: The average number of pages for each category of survey papers (applications, training techniques, model alignment, etc.).
- **Number of Citations**: The number of citations each paper received, providing a proxy for relevance and influence in the field.

A.1 Further Analysis Tools

We used additional libraries and software tools for specific tasks:

- **matplotlib**: Used to generate graphs showing trends in publication over time.
- **seaborn**: A Python library for creating more visually appealing statistical graphics.
- **Natural Language Toolkit (NLTK)**: Used for tokenization and extraction of features such as sentence count, word count, and syllable count for readability analysis.

A.2 Limitations and Future Work

The analysis performed in this project is primarily focused on metadata, which limits the depth of insights into the actual content of the survey papers. Future work could include:

- Analyzing the full text of survey papers to understand specific contributions and research directions.
- Expanding the dataset to include more recent papers or papers from a broader range of sources.
- Implementing advanced natural language processing techniques such as topic modeling or sentiment analysis to enhance understanding of trends in LLM research.

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*.