Exploring the Taxonomy of Survey Papers on Large Language Models Using Classical Machine Learning

Maqsudur Rahman
Department of Computer Science
Boise State University
maqsudurrahman@u.boisestate.edu

Abstract

The rapid advancements in large language models (LLMs) have resulted in an exponential growth of survey papers. To manage and understand the evolving taxonomy of these surveys, this project employs graph representation learning including with classical Machine learning algorithms. By treating survey topics and their interrelationships as a graph, we explore the evolving taxonomy in the context of large language models. Our study highlights trends in survey papers, providing a comprehensive understanding of their distribution across various research domains. The results reveal a clear trajectory of increasing specialization, emphasizing the role of LLMs in areas such as prompt engineering, multimodal models, and application domains like education and finance. Graph-based analysis allows us to capture these trends more effectively and enables a better understanding of the survey landscape

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 employ graph representation learning to analyze the taxonomy of survey papers on LLMs, focusing on both topical coverage and temporal trends. By leveraging this methodology, we aim to provide insights into the current state of the field and how it is evolving, helping researchers and practitioners navigate the complex landscape of LLM-related literature. Overall, our contributions can be summarized as follows:

- We introduce a graph-based framework to model and analyze the taxonomy of survey papers on LLMs, providing a scalable approach for understanding the structure of research topics.
- We offer a detailed temporal analysis of the growth in survey paper publications, highlighting key periods of increased activity and shifting research focus.
- Through our analysis, we identify significant trends in LLM-related research, including emerging areas such as "Prompting Science" and the application of LLMs in "Finance" and "Education."
- By constructing and visualizing the interconnections between survey topics, we provide a clear understanding of the relationships between different research domains, enabling better navigation of the literature.

2 Methodology

In this section, we begin by outlining the data collection process, followed by an analysis of the metadata. Next, we describe how three types of

attributed graphs are constructed and explain how graph representations are learned using graph neural networks.

2.1 Data Exploration

The data for this project consists of survey papers published between July 2021 and January 2024, focusing on large language models. We curated this dataset from publicly available research databases, extracting relevant metadata such as publication date, topics covered, and keywords associated with each paper. This information was used to create the graph structure for our analysis, with nodes representing topics and edges representing their relationships based on co-occurrence in survey papers.

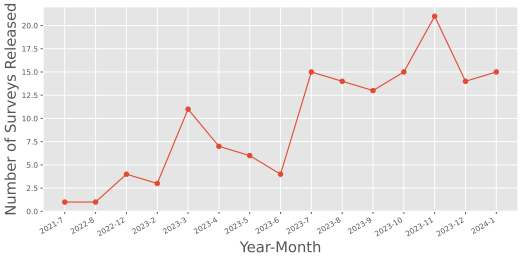

Figure 1: Trends of survey papers about large language models since 2021. Numbers reflect the year and month (e.g., 2023-3 is March 2023).

2.2 Data Manipulation

In this phase, we focus on preparing the dataset by creating three graph types: text, co-author, and co-category. We build a TF-IDF matrix based on word frequency and distinctiveness for the titles and abstracts. To enrich the data, one-hot encoding is applied to arXiv categories, and this information is combined to generate the final feature matrix used for graph-based classification tasks The results from our graph-based analysis provide several key insights into the taxonomy and trends of survey papers on LLMs: The number of surveys released has grown consistently from mid-2021 to late 2023, reflecting the rapidly expanding interest in LLM research. A sharp increase was observed from early 2022, peaking in mid-2023, as shown in the monthly publication trends (Figure from file [7]). This suggests an accelerated pace of development in the field, driven by both academic research and practical applications. The categorization of survey papers shows a diverse set of research areas, with certain topics such as "Prompting Science" and "Evaluation" being particularly prominent (Figure from file [6]). This suggests that much of the current research is focused on understanding the best practices for prompting LLMs and evaluating their performance across various domains. Beyond core areas, we observed emerging interest in application-specific domains like "Finance," "Law," and "Education." These areas reflect the expanding use cases of LLMs in industry-specific applications, indicating a shift toward more practical, applied research.

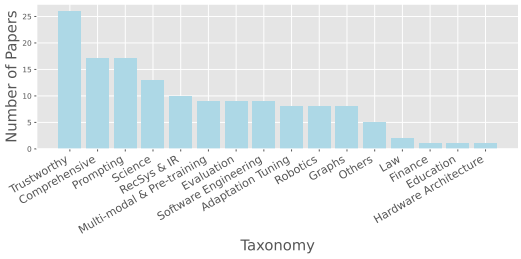

Figure 2: Distribution of classes in the taxonomy

The graph representation allowed us to explore the interconnections between topics. For example, surveys discussing "Multimodal Models" were often linked to "Pre-training" methodologies, highlighting the importance of foundational techniques in supporting new model architectures. Similarly, cross-disciplinary areas like "Software Engineering" and "Hardware Architecture" frequently intersected with core LLM research, underlining the multidisciplinary nature of the field.

3 Result and Discussion

The number of surveys released has grown consistently from mid-2021 to late 2023, reflecting the rapidly expanding interest in LLM research. A sharp increase was observed from early 2022, peaking in mid-2023, as shown in the monthly publication trends (Figure 2). This suggests an accelerated pace of development in the field, driven by both academic research and practical applications. The categorization of survey papers shows a diverse set of research areas, with certain topics such as "Prompting Science" and "Evaluation" being particularly prominent (Figure 1). This suggests that much of the current research is focused on understanding the best practices for prompting LLMs and evaluating their performance across various domains. Beyond core areas, we observed emerging interest in application-specific domains like "Finance," "Law," and "Education." These areas re-

flect the expanding use cases of LLMs in industry-specific applications, indicating a shift toward more practical, applied research. The graph representation allowed us to explore the interconnections between topics. For example, surveys discussing "Multimodal Models" were often linked to "Pre-training" methodologies, highlighting the importance of foundational techniques in supporting new model architectures. Similarly, cross-disciplinary areas like "Software Engineering" and "Hardware Architecture" frequently intersected with core LLM research, underlining the multidisciplinary nature of the field.

4 Limitations and Future Scope

4.1 Limitations

Despite the insights gained through this study, there are a few limitations to acknowledge: Our analysis is dependent on the availability and accessibility of survey papers. Some recent or domain-specific surveys might not be included due to dataset limitations. While graph representation learning effectively captures relationships between topics, the complexity of research topics can lead to oversimplification. Certain interdisciplinary areas may not be adequately captured by our clustering approach. The graph-based methodology is specifically tuned to analyze LLM-related survey papers. Its applicability to other fields of research remains to be tested and may require adjustments.

4.2 Future Scope

This project opens up several avenues for future work: Future studies can expand the dataset by including more diverse sources of survey papers, particularly those published in niche domains or non-English publications. Incorporating time-evolving graph models could help capture the dynamic evolution of research areas, enabling real-time updates of the taxonomy. Applying advanced natural language processing (NLP) techniques could allow for automated extraction of topics and relationships from survey papers, reducing manual curation efforts. The graph-based approach developed in this project could be adapted for analyzing survey papers in other fast-evolving fields such as artificial intelligence, computer vision, or bioinformatics.

5 Conclusion

In conclusion, our graph-based approach to understanding survey paper taxonomy reveals a complex and evolving landscape of LLM research. The results not only showcase the growth of the field but also provide a structured way to navigate the broad spectrum of topics, helping both researchers and practitioners identify key trends and emerging areas of interest.

A APPENDIX

A.1 Data Collection:

The dataset consists of survey papers focused on large language models (LLMs) published between July 2021 and January 2024. The metadata for each paper includes the title, authors, publication date, keywords, and arXiv categories. Data sources: Survey papers were retrieved from publicly available databases such as arXiv and other research repositories

- **A.2 Graph Construction:**

  Three types of attributed graphs were constructed. Text Graph: Built using a TF-IDF matrix based on the title and abstract of each paper. This matrix was generated by calculating word frequency and distinctiveness. Co-Author Graph: Nodes represent authors, and edges indicate co-authorship. Co-Category Graph: Nodes represent arXiv categories, with edges drawn between categories that frequently appear together in the same paper. One-Hot Encoding: Applied to arXiv categories to generate categorical features

- **A.3 Evaluation Metrics:**

  Classification accuracy, F1 score, and precision were the key metrics used to evaluate the model's performance. The evaluation was conducted through 5-fold cross-validation to ensure generalizability.

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*.