



DeepReport: An AI-assisted Idea Generation System for Scientific Research

Yi Xu

Shanghai Jiao Tong University
Shanghai, China
yixu98@sjtu.edu.cn

Luoyi Fu

Shanghai Jiao Tong University
Shanghai, China
yiluofu@sjtu.edu.cn

Shuqian Sheng

Shanghai Jiao Tong University
Shanghai, China
susisheng@sjtu.edu.cn

Bo Xue

Shanghai Jiao Tong University
Shanghai, China
sappho_x@sjtu.edu.cn

Jiaxin Ding

Shanghai Jiao Tong University
Shanghai, China
jiaxinding@sjtu.edu.cn

Lei Zhou

Shanghai Jiao Tong University
Shanghai, China
zhoulei1588@sjtu.edu.cn

Xinbing Wang

Shanghai Jiao Tong University
Shanghai, China
xwang8@sjtu.edu.cn

Chenghu Zhou

Shanghai Jiao Tong University
Shanghai, China
zhouch@lreis.ac.cn

Abstract

Nowadays, the explosive growth of academic literature has been going far beyond scientists' limited capability to read through, making it increasingly difficult for them to absorb disciplinary insights and extract intellectual essences critical for generating novel research ideas in interdisciplinary studies. To address this, we develop DeepReport, an AI-assisted scientific idea generation system to alleviate the research burden. Technically, DeepReport maintains evolving concept co-occurrence graphs to extract core insights from over 260 million publications across all disciplines. These concepts are periodically collected and updated, enabling the automatic extraction of hidden cross-domain connections. Combining temporal link prediction and analysis techniques with large language models, DeepReport is able to further transform these patterns of insights into actionable ideas. With the function of integrating up-to-date academic databases, visualizing dynamic relationships of concepts, and automatically generating new ideas, DeepReport empowers researchers to navigate complex knowledge landscapes, reduce cognitive burdens, and accelerate the generation of groundbreaking concepts. This work provides an in-depth exploration of DeepReport's architecture, functionalities, and applications, highlighting its transformative potential for advancing interdisciplinary research and fostering innovation. DeepReport is available at <https://idea.acemap.cn/>.

CCS Concepts

• **Information systems** → **Information retrieval**; • **Computing methodologies** → **Natural language generation**.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGIR '25, Padua, Italy

© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1592-1/2025/07
<https://doi.org/10.1145/3726302.3730151>

Keywords

Idea generation; Large language models; Academic big data

ACM Reference Format:

Yi Xu, Luoyi Fu, Shuqian Sheng, Bo Xue, Jiaxin Ding, Lei Zhou, Xinbing Wang, and Chenghu Zhou. 2025. DeepReport: An AI-assisted Idea Generation System for Scientific Research. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, July 13–18, 2025, Padua, Italy. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3726302.3730151>

1 Introduction

The exponential growth of academic literature has revolutionized knowledge dissemination but also introduced significant challenges for researchers [6, 13, 17]. Navigating an overwhelming volume of publications, synthesizing diverse insights, and identifying novel connections across disciplines have become arduous tasks. Traditional methods of literature review, which rely heavily on manual reading and analysis, are no longer sufficient to cope with the scale and complexity of modern academic research. Besides, interdisciplinary research, often a breeding ground for innovative breakthroughs, demands tools that can uncover hidden connections between disparate fields. However, researchers are frequently confined to narrow specializations, limiting their ability to identify such connections. Moreover, transforming these insights into coherent and actionable ideas requires additional effort and expertise.

To address these challenges, we present DeepReport, a system designed to assist researchers in generating new academic ideas. DeepReport leverages advanced techniques such as evolving concept co-occurrence graphs [6] and co-occurrence citation quintuples [17] to model the process of idea generation. This process is structured into two sequential tasks: (1) Temporal Link Prediction, which identifies potential connections between academic concepts by analyzing their temporal co-occurrence patterns, and (2) Idea verbalization, where these connections are transformed into coherent textual descriptions using state-of-the-art large language models (LLMs). Building on these foundational techniques, DeepReport is

designed to achieve four key objectives that drive its development and functionality:

- **Retrieve Relevant Papers from 20 Disciplines:** DeepReport provides powerful academic literature retrieval capabilities, drawing from the Acemap [16] database of over 260 million papers across 20 disciplines. Researchers can search using keywords, phrases, or research questions to quickly locate relevant studies.
- **Explore Conceptual and Knowledge Relationships:** By constructing evolving concept co-occurrence graphs, DeepReport visually represents the relationships between academic concepts. Nodes represent concepts, and edges indicate their co-occurrence within the literature. These dynamic graphs update over time to reflect the latest research trends, enabling researchers to discover key themes, academic hotspots, and their interconnections.
- **Predict Potential Conceptual Connections:** Through temporal link prediction model, DeepReport analyzes historical data to predict future academic concept relationships. This functionality helps uncover emerging connections and high-potential research directions, offering reliable support for exploring future trends.
- **Generate New Academic Ideas:** After predicting potential connections, DeepReport uses advanced LLMs to create descriptive summaries of these relationships. These models, fine-tuned on co-occurrence citation data, generate coherent and innovative research ideas based on user-selected concept pairs, transforming complex academic discoveries into clear textual expressions.

In this work, we describe the design, architecture, and features of DeepReport, demonstrating its utility in academic research and interdisciplinary collaboration.

2 Related Work

In recent years, academic knowledge discovery and research idea generation have seen significant advancements through graph-based tools and LLMs. Early graph-based systems like SEMNET [6] and InfraNodus [9] laid the groundwork by utilizing co-occurrence graphs and text networks to uncover research trends and identify structural gaps. These tools provided a foundational framework for navigating the complexities of academic knowledge and offered valuable support for generating new research directions.

The emergence of LLMs has revolutionized the landscape of idea generation by combining powerful language understanding with sophisticated generation capabilities. Notable systems such as SCIMON [15] and Chain-of-Ideas [7] leverage domain-specific fine-tuning and chain-of-thought methodologies to develop innovative and groundbreaking research ideas [19].

To enhance the diversity and quality of generated ideas, researchers have introduced innovative techniques, including multi-source seed generation [4] and self-correction mechanisms. These approaches improve logical coherence and mitigate common issues like hallucination, ensuring the generation of more reliable and actionable insights. Evaluation frameworks such as IDEA-Bench [8] provide systematic methodologies to assess the effectiveness of

LLMs in generating research ideas, offering researchers quantifiable measures of success. Through iterative idea development and domain-specific optimization [1], LLMs are transforming the process of academic inquiry, enabling researchers to explore emerging fields, challenge existing paradigms, and drive innovation across disciplines.

LLM-based systems excel at generating insights but often face challenges such as hallucination and lack of traceability. DeepReport addresses these limitations by combining evolving concept co-occurrence graphs with LLM-driven idea generation. This approach ensures that generated ideas are grounded in traceable, structured relationships derived from academic literature, enhancing both the explainability and reliability of the outputs.

3 System Architecture

The architecture of DeepReport is meticulously crafted to enable seamless exploration, analysis, and generation of academic ideas within a unified, scalable framework. The system is built around four core modules: the Search and Data Retrieval Module; the Concept Co-occurrence Graph Module; the Temporal Link Prediction Module; and the Idea Verbalization Module. Figure 1 illustrates the result page generated by DeepReport for the query *Carbonate Rock*.

3.1 Search and Data Retrieval Module

The Search and Data Retrieval Module is the foundational component of DeepReport, designed to harness an academic database called Acemap [16], which contains over 260 million papers spanning 20 disciplines. This extensive repository ensures broad coverage across diverse research domains, enabling the system to address various academic needs effectively. To deliver efficient and high-quality retrieval, the module utilizes Elasticsearch, a powerful search engine optimized for handling complex queries while maintaining low latency.

Users can interact with this module by submitting queries in the form of keywords, phrases, or research questions. The system retrieves not only the most relevant academic papers but also identifies prominent concepts associated with the retrieved literature.

To extract these concepts, the module employs AutoPhrase [10], an advanced natural language processing tool. AutoPhrase processes the retrieved papers to identify key academic terms and domain-specific entities. This ensures that concepts are derived directly from the corpus of retrieved literature, capturing both the explicit and implicit academic knowledge embedded within the papers. The extracted concepts and their connections serve as the input for constructing the concept co-occurrence graph in subsequent stages. Moreover, the Search and Data Retrieval Module is designed to evolve continuously, seamlessly incorporating newly published papers to keep the system aligned with the latest academic developments.

3.2 Concept Co-occurrence Graph Module

DeepReport captures the dynamic and evolving nature of academic knowledge by modeling relationships between concepts as evolving co-occurrence graphs. In these graphs, concepts are represented as nodes, while their co-occurrence within academic papers forms the edges connecting them.

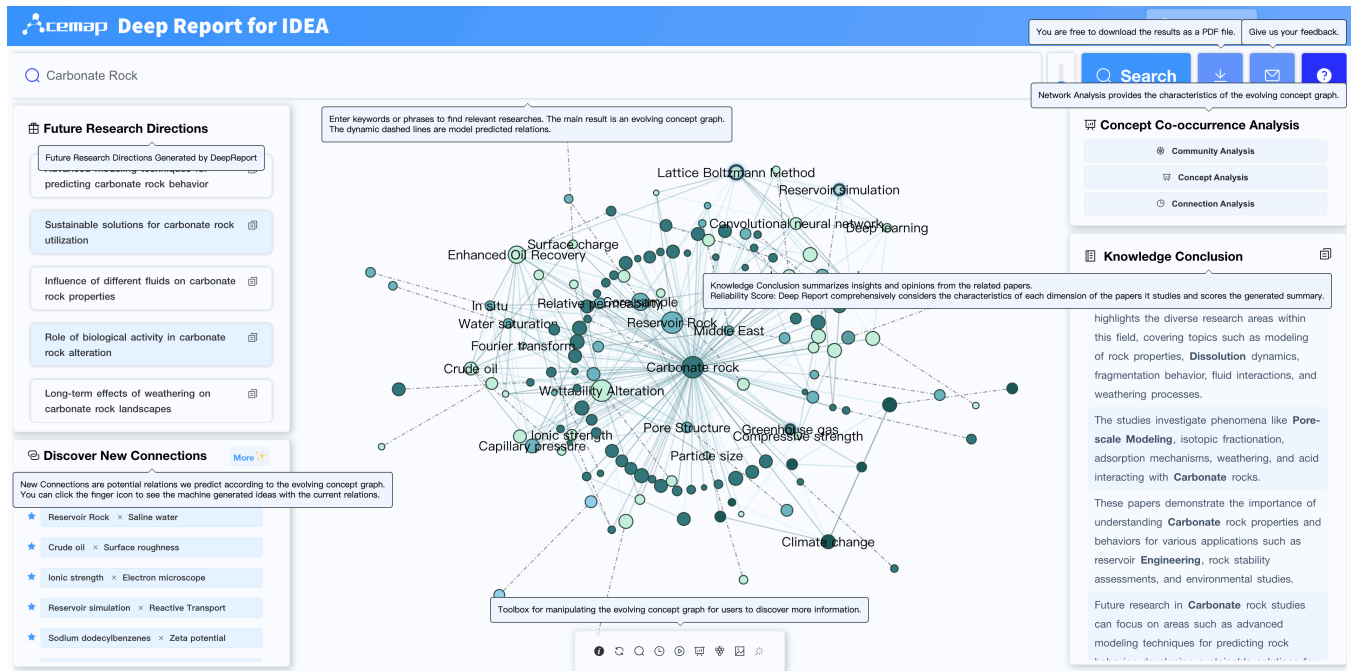


Figure 1: Illustration of the result page of DeepReport.

The evolutionary nature of these graphs is a key innovation. As new academic papers are published, the system continuously incorporates them into the co-occurrence graph, dynamically updating the nodes and edges to reflect the latest developments in the field. This iterative updating process ensures that the graph remains relevant and accurately represents the current state of academic knowledge. Additionally, the system leverages community detection algorithms to uncover clusters of closely related concepts. These clusters represent communities of ideas or subfields within the broader academic landscape, offering researchers valuable insights into the structure and organization of knowledge. By highlighting these communities, DeepReport enables users to identify potential interdisciplinary connections and areas of convergence that could inspire innovative research directions.

3.3 Temporal Link Prediction Module

The Temporal Link Prediction Module is a cornerstone of DeepReport, leveraging the evolving concept co-occurrence graph to anticipate future connections between academic concepts. Using a BERT-based model [3] fine-tuned for temporal tasks, this module generates predictions for new co-occurrence relationships based on historical data. The system leverages a masked language model approach to train the prediction model, where the temporal context of concepts is considered when identifying potential links. The model is trained on a combination of existing concept pairs and negative pair sampling strategy, enhancing its ability to distinguish between meaningful and spurious connections.

The link prediction process is critical for identifying novel ideas and uncovering connections between previously unconnected concepts. The predictions are classified into existing, emerging, and

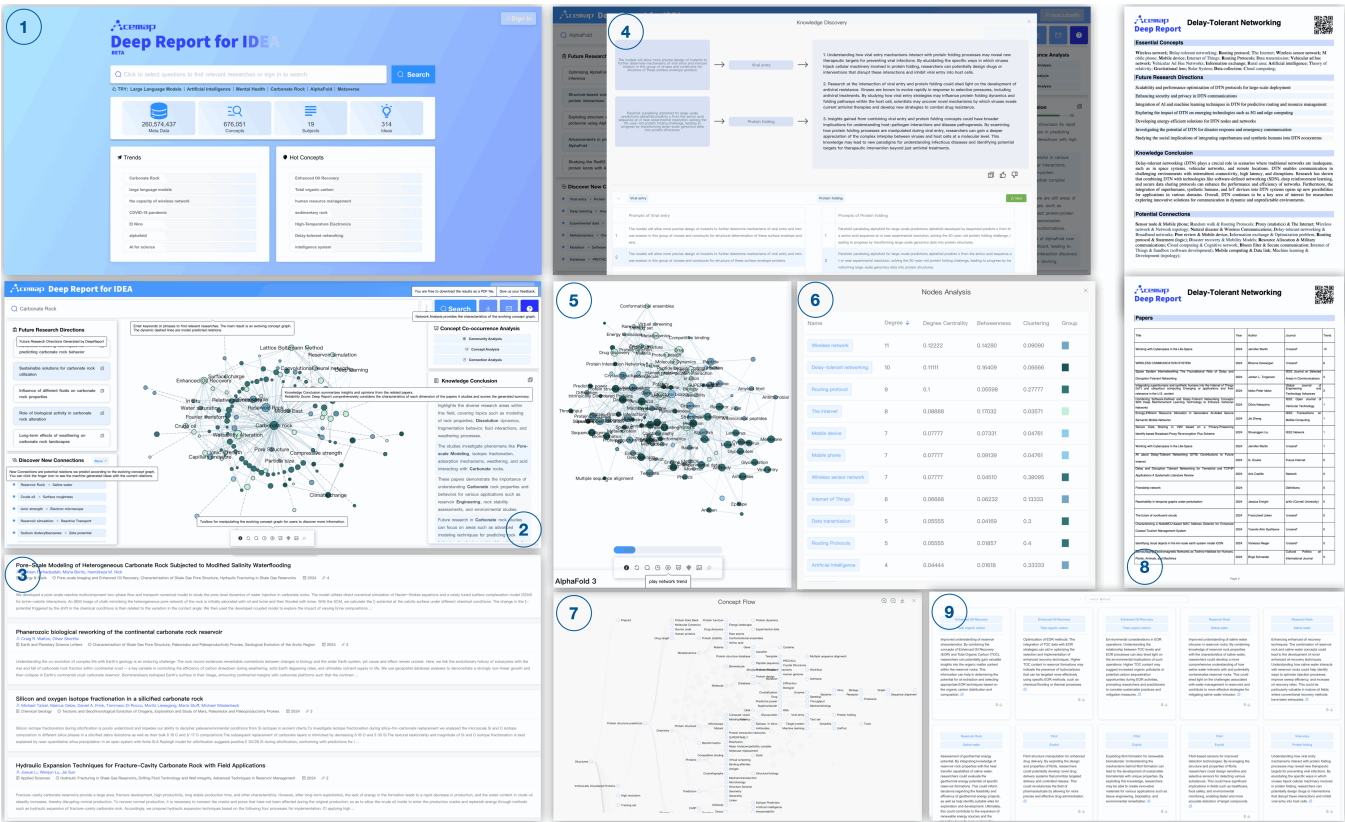
potential relationships, providing a comprehensive view of how ideas are likely to develop in the future. By doing so, DeepReport helps researchers identify high-impact areas of future research and trends in their field.

3.4 Idea Generation Module

Once potential connections between concepts are identified through the Temporal Link Prediction Module, the Idea Generation Module translates these connections into descriptive and actionable texts. This module leverages advanced LLMs such as LLaMA[14], Mistral[5], and Qwen [2], which have been fine-tuned using co-occurrence citation quintuples. These quintuples include pairs of concepts, their contextual sentences extracted from academic papers, and the idea or hypothesis linking them. Training the models on this structured and domain-specific dataset enables them to generate text that is both coherent and contextually rich.

The verbalization process is meticulously designed to be flexible, user-driven, and context-sensitive. Researchers can input specific concept pairs or allow the system to automatically select promising connections based on the evolving concept graph. For each pair, the module generates clear and well-structured textual descriptions that articulate how the concepts are interrelated and propose potential research directions stemming from these connections.

To ensure relevance and accuracy, the Idea Generation Module draws on a vast corpus of academic literature. By doing so, it generates outputs that are not only factually grounded but also creative and forward-thinking. According to researches about idea or paper evaluation [11, 12, 18], we also apply LLM representation-based methods to evaluate these generated ideas.



3.5 Integration and User Interface

As illustrated in Figure 2, DeepReport incorporates a wide range of features to enhance user experience and engagement. These functionalities are accessible through an intuitive web-based interface. Below, we outline the key functionalities that enhance the user experience:

- (1) **Homepage with Search Trends:** Highlights global user search trends, showcasing frequently explored research topics.
- (2) **Results Page:** Displays search results, including evolving concept co-occurrence graphs and generated new ideas.
- (3) **Paper List:** Organizes retrieved papers with links to abstracts and related concepts.
- (4) **Predicted Links and Idea Details:** Highlights newly predicted concept connections with detailed idea verbalization.
- (5) **Interactive Graph Exploration:** Enables users to interact with the evolving concept co-occurrence graph, zooming in on nodes to view associated papers and concept details.
- (6) **Community Detection and Analysis Tools:** Offers advanced graph analytics, including cluster detection and network property analysis.
- (7) **Concept Flow Visualization:** Visualizes the propagation of concepts across papers, emphasizing interdisciplinary interactions.

- (8) **PDF Report Generation:** Allows users to download professional reports summarizing graphs, insights, and generated ideas.
- (9) **New Idea Showcase:** Displays generated ideas for community engagement, enabling users to like, share, and discuss innovative concepts.

4 Conclusion and Future Work

The DeepReport system represents a significant advancement in academic research tools, combining state-of-the-art NLP techniques with graph-based analysis to assist researchers in discovering and generating new ideas. By integrating evolving concept co-occurrence graphs, temporal link prediction, and idea generation, DeepReport accelerates knowledge discovery, fosters interdisciplinary research, and provides actionable insights that guide future academic inquiries. Its scalable architecture, real-time updates, and user-centric design make it an invaluable tool for researchers seeking to explore the depths of academic literature and generate novel ideas.

Acknowledgments

This work was supported by NSF China (No. 62020106005, 42050105, 62061146002), Shanghai Pilot Program for Basic Research - Shanghai Jiao Tong University.

References

- [1] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. arXiv:2404.07738 [cs.CL] <https://arxiv.org/abs/2404.07738>
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. arXiv:2309.16609 [cs.CL] <https://arxiv.org/abs/2309.16609>
- [3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423
- [4] Xiang Hu, Hongyu Fu, Jing Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024. Nova: An Iterative Planning and Search Approach to Enhance Novelty and Diversity of LLM Generated Ideas. arXiv:2410.14255 [cs.AI] <https://arxiv.org/abs/2410.14255>
- [5] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] <https://arxiv.org/abs/2310.06825>
- [6] Mario Krenn and Anton Zeilinger. 2020. Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences* 117, 4 (2020), 1910–1916.
- [7] Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Lidong Bing. 2024. Chain of Ideas: Revolutionizing Research Via Novel Idea Development with LLM Agents. arXiv:2410.13185 [cs.AI] <https://arxiv.org/abs/2410.13185>
- [8] Chen Liang, Lianghua Huang, Jingwu Fang, Huanzhang Dou, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Junge Zhang, Xin Zhao, and Yu Liu. 2024. IDEA-Bench: How Far are Generative Models from Professional Designing? arXiv:2412.11767 [cs.CV] <https://arxiv.org/abs/2412.11767>
- [9] Dmitry Paranyushkin. 2019. InfraNodus: Generating Insight Using Text Network Analysis. In *The World Wide Web Conference (San Francisco, CA, USA) (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 3584–3589. doi:10.1145/3308558.3314123
- [10] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (2018), 1825–1837.
- [11] Shuqian Sheng, Yi Xu, Luoyi Fu, Jiaxin Ding, Lei Zhou, Xinbing Wang, and Chenghu Zhou. 2024. Is Reference Necessary in the Evaluation of NLG Systems? When and Where?. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 8580–8596. doi:10.18653/v1/2024.naacl-long.474
- [12] Shuqian Sheng, Yi Xu, Tianhang Zhang, Zanwei Shen, Luoyi Fu, Jiaxin Ding, Lei Zhou, Xiaoying Gan, Xinbing Wang, and Chenghu Zhou. 2024. RepEval: Effective Text Evaluation with LLM Representation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 7019–7033. doi:10.18653/v1/2024.emnlp-main.398
- [13] Stefan Thurner, Wenyuan Liu, Peter Klimek, and Siew Ann Cheong. 2020. The role of mainstreamness and interdisciplinarity for the relevance of scientific papers. *PLoS one* 15, 4 (2020), e0230325.
- [14] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth  e Lacroix, Baptiste Rozi  re, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL] <https://arxiv.org/abs/2302.13971>
- [15] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 279–299. doi:10.18653/v1/2024.acl-long.18
- [16] Xinbing Wang, Luoyi Fu, Xiaoying Gan, Ying Wen, Guanjie Zheng, Jiaxin Ding, Liyao Xiang, Nanyang Ye, Meng Jin, Shiyu Liang, Bin Lu, Haiwen Wang, Yi Xu, Cheng Deng, Shao Zhang, Huquan Kang, Xingli Wang, Qi Li, Zhixin Guo, Jiexing Qi, Pan Liu, Yuyang Ren, Lyuwen Wu, Jungang Yang, Jianping Zhou, and Chenghu Zhou. 2024. AceMap: Knowledge Discovery through Academic Graph. arXiv:2403.02576 [cs.DL] <https://arxiv.org/abs/2403.02576>
- [17] Yi Xu, Shuqian Sheng, Bo Xue, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. 2023. Exploring and Verbalizing Academic Ideas by Concept Co-occurrence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13001–13027. doi:10.18653/v1/2023.acl-long.727
- [18] Yi Xu, Bo Xue, Shuqian Sheng, Cheng Deng, Jiaxin Ding, Zanwei Shen, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. 2024. Good Idea or Not, Representation of LLM Could Tell. arXiv:2409.13712 [cs.CL] <https://arxiv.org/abs/2409.13712>
- [19] Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024. A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 8783–8817. doi:10.18653/v1/2024.emnlp-main.498