The Empowerment of Science of Science by Large Language Models: New Tools and Methods

Authors: Guoqiang Liang, Mengxuan Li, Zhihao Zhang, Gege Lin, Shuo Zhang, Zhihao Zhang

Date: 2025-10-14T21:16:18+00:00

Abstract

Large Language Models (LLMs) have exhibited exceptional capabilities in natural language understanding and generation, image recognition, and multimodal tasks, charting a course towards Artificial General Intelligence and emerging as a central issue in the global technological race. This manuscript conducts a comprehensive review of the core technologies that support LLMs from a user's standpoint, including prompt engineering, knowledge-enhanced retrieval-augmented generation (RAG), fine-tuning, pre-training, and tool learning. Additionally, it traces the historical development of Science of Science (SciSci) and presents a forward-looking perspective on the potential applications of LLMs within the scientometric domain. Furthermore, it discusses the prospect of an AI agent-based model for scientific evaluation, and presents new research fronts detection and knowledge graph building methods with LLMs.

Full Text

Preamble

The Empowerment of Science of Science by Large Language Models: New Tools and Methods

Guoqiang Liang[1], Mengxuan Li[2], Zhihao Zhang[1],, *Gege Lin*[1],, Shuo Zhang[1],*

[1]College of Economics and Management, Beijing University of Technology, Beijing, 100124, China
[2]College of Economics and Management, Langfang Normal University, Langfang, Hebei 065000, China

Abstract: Large Language Models (LLMs) have exhibited exceptional capabilities in natural language understanding and generation, image recognition, and multimodal tasks, charting a course toward Artificial General Intelligence and emerging as a central focus in the global technological race. This manuscript provides a comprehensive review of the core technologies underpinning LLMs from a user's perspective, including prompt engineering, knowledge-enhanced retrieval-augmented generation (RAG), fine-tuning, pre-training, and tool learning. Additionally, it traces the historical development of the Science of Science (SciSci) and presents a forward-looking perspective on the potential applications of LLMs within the scientometric domain. Furthermore, it discusses the prospect of AI agent-based models for scientific evaluation and introduces new methods for research front detection and knowledge graph construction using LLMs.

Keywords: Large Language Models, ChatGPT, Science of Science, AI4Science

The rapid development of large-scale pre-trained models, often referred to as "large language models (LLMs)," has driven a fundamental transformation in artificial intelligence. These models have demonstrated superior performance in natural language understanding and generation, image recognition, and even multimodal tasks, paving the way for Artificial General Intelligence (AGI) and becoming a focal point in global technological competition. As of 2023, the United States leads with 61 LLMs, significantly outpacing the European Union's 21 and China's 15 LLMs, establishing itself as a top nation in artificial intelligence. These models have been widely applied across various sectors, including healthcare, finance, education, law, and mathematics [1].

SciSci is an interdisciplinary field that conducts quantitative research on science itself, encompassing scientists, academic literature, scholarly journals, and science policies. The methodologies employed in SciSci have evolved from traditional citation analysis, word frequency analysis, and statistical analysis to incorporate computer science and artificial intelligence. For instance, scholars have begun using dynamic topic models and word2vec for analyzing semantic-enhanced topic evolution [2]. BERT models [3], Graph Convolutional Networks (GCNs) [4], and knowledge graphs [5] are now employed for citation recommendation, keyword identification, emerging technological theme detection [6], and exploring the research characteristics of Nobel laureates [7].

This study systematically reviews the underlying technical architecture, common concepts, and capabilities of large models, as well as the key technologies supporting LLMs from a user's perspective. It then illustrates the evolution of SciSci from past to present, and finally provides prospective applications of LLMs within the field of SciSci.

1 Introduction to LLMs

LLMs, also known as "foundation models," are deep neural network models with vast numbers of parameters and complex computational structures. They are characterized by their scalability (large parameter volume), emergent properties (the ability to develop new capabilities unexpectedly), and universality (not limited to specific problems or domains). These models are capable of driving multiple use cases and applications while resolving various tasks, making them "milestones" in the fields of natural language processing (NLP) and artificial intelligence.

Similar to the human brain, LLMs—due to their enormous number of parameters and deep neural network architecture—can learn and understand a broader range of features and patterns. This enables them to demonstrate remarkable capabilities in natural language understanding and generation, reasoning, intent recognition, and the creation of images and videos from text, covering virtually all aspects related to NLP. They also possess general problem-solving abilities and are considered a significant path toward achieving general artificial intelligence [8]. Currently, LLMs have become the infrastructure of the AI field, providing powerful computational, learning, and problem-solving capabilities for addressing a variety of complex issues, including weather forecasting [9], behavioral analysis [10], and drug synergy prediction [11], effectively accomplishing complex modeling and predictive tasks.

The massive data input and Transformer architecture constitute the main source of LLMs' capabilities. Taking OpenAI's GPT series as an example (Table 1 presents an extended version based on reference [12]), in 2018, OpenAI introduced the GPT-1 model, which was based on a 12-layer Transformer architecture and trained on approximately 5GB of data. This model significantly improved computational speed and capacity compared to long short-term memory (LSTM) models, marking a major advancement in Transformer-based architectures. The following year, OpenAI built upon GPT-1 to release GPT-2, featuring a 48-layer Transformer architecture and trained on data eight times larger than GPT-1. This allowed the model to better understand semantics and contextual information, demonstrating formidable text generation capabilities. In 2020, OpenAI released the GPT-3 model, based on the GPT-2 architecture, with the number of Transformer layers doubled and the amount of pre-training data increased by over a thousand times. GPT-3 enabled user interaction through natural language and was capable of performing most NLP tasks such as automatic question answering, text classification, and machine translation, showcasing astonishing natural language understanding abilities. It wasn't until the emergence of ChatGPT that the academic community realized the disruptive potential of LLMs on traditional paradigms of natural language processing tasks. The introduction of ChatGPT-4 has further propelled multimodal LLMs to the forefront of cutting-edge research today.

Table 1 Pre-trained Data Volume of ChatGPT Models

Architecture	Layers	Parameters	Data size
GPT-1 (2018)	Transformer	-	-
GPT-2 (2019)	Transformer	-	-
GPT-3 (2020)	Transformer	110 million	1.5 billion
ChatGPT-4 (2023)	Transformer	175 billion	1.76 trillion

Architecture	Layers	Parameters	Data size
		Not disclosed	

1.1 Classification of LLMs

LLMs can be classified into different types based on various criteria. When categorized by input data type, LLMs can be divided into language models, visual models, and multimodal models. Language models are primarily used for processing text data and understanding natural language, making them a significant category within the field of NLP. These models are characterized by their training on large-scale corpora to learn various grammatical, semantic, and contextual rules of natural language, such as GPT-3, Bard, ERNIE Bot, and ChatGLM. Visual models are typically used for image processing and analysis, commonly employed in the field of computer vision (CV). These models are trained on extensive image datasets to perform visual tasks such as image classification, object detection, image segmentation, pose estimation, and facial recognition. Examples include the VIT series (Google), Wenxin UFO, Huawei Pangu CV, and INTERN (SenseTime). Multimodal models combine features of both language and visual models, enabling them to process text, images, and videos simultaneously for a more comprehensive understanding of complex data. Examples include ChatGPT-4, Sora, and Gemma2. Figure 1 [FIGURE:1] illustrates the parameters and classification of notable LLMs.

Based on different model architectures, LLMs can be divided into those based on the Transformer architecture and those using the Mixture of Experts (MoE) architecture. LLMs that utilize the Transformer architecture are designed around the Transformer model, introduced by Vaswani et al. in 2017 [13]. This architecture leverages the self-attention mechanism, allowing it to effectively handle long-distance dependencies in sequential data. The core components of the Transformer model are multi-head self-attention and positional encoding, enabling the model to capture relationships between different positions in the input sequence. Due to its outstanding performance, the Transformer has become the foundational architecture for many large language models, such as BERT and the GPT series.

MoE models are a type of distributed expert system that assigns tasks to multiple “expert” subnetworks [14]. A gating network determines which expert should handle each input sample. This architecture allows models to scale to very large sizes, as increasing the number of experts can enhance the model’s capacity and performance without significantly increasing the complexity of any individual expert. MoE models have demonstrated superior scalability and efficiency in handling certain tasks, such as language modeling and image recognition. Both architectures have their advantages: the Transformer architecture is widely used in NLP tasks due to its efficiency in processing sequential data, while the MoE architecture has garnered attention for its scalability and parallel processing capabilities. Additionally, based on the application domain, LLMs can be categorized into general-purpose models and vertical models; according to the type of autoencoder, they can be further divided into encoder-based models and decoder-based models, among others, which will not be detailed here.

1.2 Common Terminology for LLMs

With the development of AI, various concepts such as general-purpose models, vertical models, fine-tuning, tokenization, embedding, and AI agents have emerged [15-18]. These concepts can be easily confused, which is why Figure 2

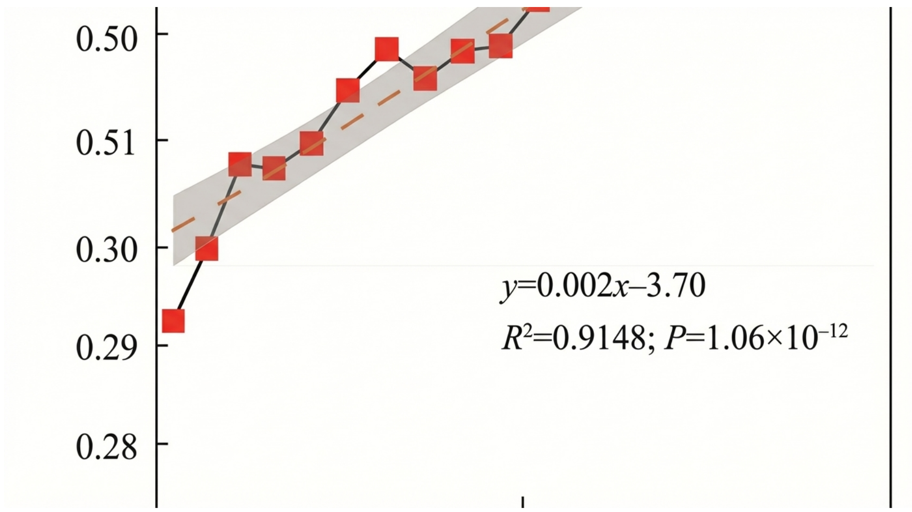

Figure 1: Figure 2

provides an overview of the relationships between them. In simple terms, LLMs can be categorized into general-purpose models and vertical models. General-purpose models are pretrained on large public datasets, while vertical models are primarily fine-tuned based on specific domain or industry data using general-purpose models as a foundation [19]. To enhance model performance on specific tasks, the process of further training using labeled data is known as fine-tuning [20, 21]. Fine-tuning or pretraining is predicated on tokenization and embedding, where input data is mapped into a high-dimensional vector space for computation. An AI agent is an intelligent entity based on LLMs, equipped with planning, memory, and tool-learning capabilities. Figure 2 illustrates the relationships among the main concepts related to LLMs.

In summary, we can view LLMs as a type of neural network-based autoregressive language model. Essentially, they function as probabilistic language models that learn language patterns from vast amounts of corpus data and output the most likely correct answers based on user input.

1.3 Workflow of LLMs

A typical model based on the Transformer architecture generally processes input data in three steps. First, the input data undergoes embedding, which includes both word embedding and position embedding. After the input text is tokenized, each token is transformed into a high-dimensional vector using word embedding techniques. These high-dimensional vectors are then concatenated with position embedding vectors, which capture the position of the tokens in the text.

Second, the concatenated data is passed through multiple Transformer layers. During this process, the self-attention mechanism plays a key role in understanding semantic relationships. We can represent the self-attention mechanism with Equation 1, where Q denotes query, K denotes Key, and V denotes Value [22, 23]. Finally, the model predicts the most likely next token in the sequence based on the context and continues generating subsequent tokens through an autoregressive approach, completing the text generation task. In summary, the basic workflow of LLMs and key information about the self-attention mechanism are shown in Figure 3

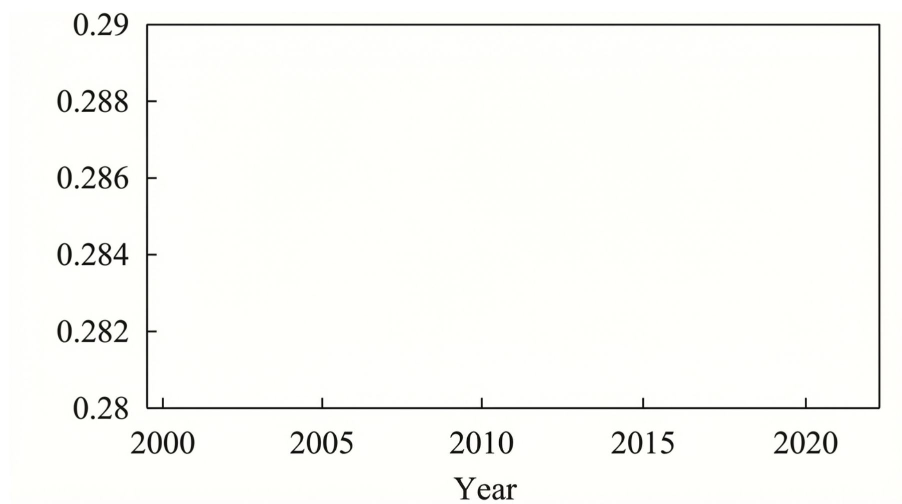

Figure 2: Figure 3

. The detailed process of the self-attention mechanism, as depicted in Figure 3B, can be found in reference [13].

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \times V \quad \text{(Equation 1)}$$

Figure 3. Workflow of LLMs and analysis of the self-attention mechanism

1.4 Key Techniques of LLMs

From a user's perspective, there are five key technologies associated with LLMs: prompt engineering, knowledge-enhanced retrieval-augmented generation (RAG), fine-tuning, pre-training, and tool learning, respectively. These five technologies generally increase in complexity.

Prompt engineering involves designing or optimizing input prompts to guide LLMs in generating outputs that meet user expectations, thereby allowing LLMs to better serve user needs [24]. Essentially, a prompt is a text-based input to the LLMs to guide its output. When the input is in the form of speech, the LLMs first convert it into text, which is then used as the prompt. Prompt engineering is not simply a question-and-answer process; using clear, precise, and concise prompting formats significantly improves output quality [25]. For instance, "Please find relevant information about Company A" is less clear and precise than "Please find the headquarters location, founder, main business, and founding year of Company A, and provide a 100-word company profile supported by references." The latter prompt yields results more aligned with user needs. In addition, prompt engineering includes techniques such as one-shot or few-shot prompts, Chain of Thought (CoT), Reasoning and Acting (ReAct), and Tree of Thoughts (ToT) prompts [26].

Retrieval-Augmented Generation (RAG) is a technique that leverages external knowledge bases to improve the accuracy of LLM outputs. It is one of the effective methods for addressing the issue of "hallucinations" in LLMs, especially when handling domain-specific or knowledge-intensive tasks. Currently, RAG is widely applied in knowledge graph construction, text summarization, and question-answering systems. RAG can be categorized into three types: Naive RAG, Advanced RAG, and Modular RAG. Naive RAG, the first method to gain attention since the launch of ChatGPT, involves three steps: indexing, retrieval, and generation. Advanced RAG improves upon Naive RAG by adding pre-retrieval and post-retrieval strategies, addressing its limitations in retrieval precision, recall, hallucinations, and the issue of disjointed or incoherent output. Modular RAG builds on the foundations of the previous two approaches, offering superior adaptability and flexibility. Restructured RAG and rearranged RAG pipelines have been incorporated to tackle specific challenges, going beyond the fixed structures of Naive RAG and Advanced RAG.

Fine-tuning is the process of adjusting the parameters of a pre-trained large language model to a specific task or domain [20, 21]. When LLMs perform poorly on a specific task, it becomes necessary to consider fine-tuning the model. By fine-tuning a model on a small, specific dataset, users can improve the LLMs' performance on that particular task. According to the OpenAI Platform, fine-tuning has at least four advantages: higher quality results than prompting, the ability to train on more examples than prompting, token saving, and lower latency requests. Some research has demonstrated these advantages; for example, Schmirler et al. found that task-specific supervised fine-tuning almost always improves downstream predictions, thus suggesting that researchers should always try fine-tuning, particularly for problems with small datasets [21]. Furthermore, many techniques and models exist for this purpose, such as full parameter, layer-specific, component-based, and multi-stage fine-tuning methods, as well as LoRA and qLoRA techniques. Models like GPT-4, GPT-3.5-turbo, and T5 are covered in reference [19], which provides vast, high-quality details on this information and LLMs.

RAG, prompt engineering, and fine-tuning are commonly used methods to improve the accuracy of LLM outputs. However, users are often perplexed about which one to choose, and as a result, these techniques are frequently compared [28]. According to Yunfan et al., "prompt engineering leverages a model's inherent capabilities with minimal necessity for external knowledge and model adaptation. RAG can be likened to providing a model with a tailored textbook for information retrieval, ideal for precise information retrieval tasks. In contrast, fine-tuning is comparable to a student internalizing knowledge over time, suitable for scenarios requiring replication of specific structures, styles, or formats" [27], as shown in Figure 4

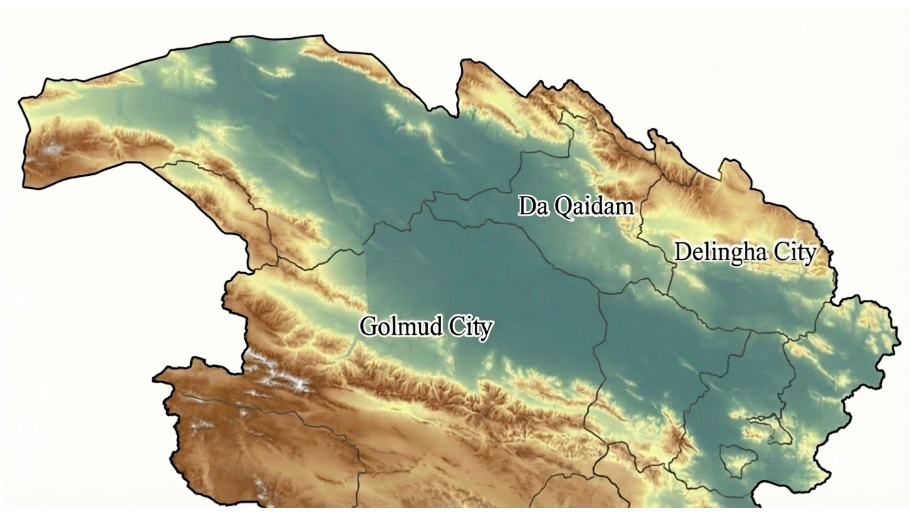

Figure 3: Figure 4

. However, here is a tip from OpenAI on this issue that may help: try prompt engineering first, due to the lower investment of time and effort it requires.

In the very first stage, the LLM is trained in a self-supervised manner on a large corpus to predict the next tokens given the input, which essentially means finding a good "initialization point" for the model parameters. This idea was originally widely used in the field of computer vision, where large-scale labeled image datasets such as ImageNet were used to initialize the parameters of vision models. To pre-train large language models, a vast amount of text data needs to be prepared, which must undergo rigorous cleaning to remove any potentially harmful or toxic content. After cleaning, the data is tokenized into a stream and split into batches for pre-training the language model [19]. Since the foundational capabilities of large language models mainly come from pre-training data, data collection and cleaning have a significant impact on the model's performance.

Tool learning refers to the process that aims to unleash the power of LLMs to effectively interact with various tools to accomplish complex tasks [29, 30]. Large-scale models do not inherently possess the ability to utilize APIs for forwarding generated text to designated email accounts. Moreover, since the data employed during the pre-training phase is not current but rather from a specific time frame, it is difficult for them to automatically retrieve up-to-date information from the web. Tool learning offers an effective solution to this limitation by seamlessly integrating large models with API interfaces. This integration allows for the execution of straightforward tasks such as automated email responses and real-time weather checks, as well as more complex tasks involving the reconstruction of workflows.

These technologies together lay the groundwork for LLMs, enabling them to perform impressively across a multitude of tasks and fields. With ongoing research, these technologies are continually being refined and enhanced to tackle the challenges that large models face in real-world applications.

2 A Brief Introduction of SciSci

SciSci, often referred to as the "science of science," seeks to understand, quantify, and predict scientific research and its outcomes [31]. This includes analyzing the innovation process [32-34], measuring the influence of scientific publications [33, 35, 36], researchers [37, 38], journals [39, 40], and institutions [38, 41], as well as modeling scientific collaboration and citation patterns [42, 43]. Additionally, it involves classifying various scientific domains [44, 45] and evaluating funding and success [46, 47]. The insights garnered from SciSci hold significant implications for management science and policy-making.

2.1 Typical SciSci Studies

The fundamental concept underlying the development of SciSci is, notably, citation [48]. Citations serve as evidence by linking a researcher's work to demonstrate the validity of the authors' ideas. They create connections between authors, ideas, journals, institutions, and even countries, enabling the construction of citation networks or the application of citation counts for research evaluation purposes.

The introduction of the Science Citation Index (SCI) database in the 1950s significantly advanced citation analysis, with Price [49] among the early pioneers recognizing the importance of interconnected networks of scholarly papers. Although the SCI was initially intended to facilitate more effective literature searches for researchers, its immense potential in research evaluation soon became apparent. Phenomena such as "cumulative advantage" [50], the "Matthew effect" [51], and "invisible colleges" [52] were observed and identified through citation analysis. Co-citation analysis [53], bibliographic coupling [54], and direct citation analysis [55] emerged, along with their derived forms, including author-level, journal-level, and keyword-level citation analysis. Regarding indicators, citation counts, h-index, journal impact factor, and their variants have been the most commonly utilized metrics for policy-making and research evaluation, despite ongoing criticisms.

Recently, metrics such as usage, tweets, and mentions, collectively referred to as "altmetrics," have been considered supplementary to traditional citations in research evaluation. Beyond their role in impact assessment, several research teams began focusing on knowledge mapping during the mid-1980s. Tools like Pajek and Ucinet were developed to facilitate the visualization of large networks. Boyack, Klavans, and Börner [56] were the first to map the backbone of science in 2005. More recent visualization tools specifically designed for SciSci, such as CiteSpace, VOSviewer, and CitNetExplorer, have made network generation more accessible to users.

2.2 Recent Advances in SciSci

In the late 1990s, a group of computer scientists and physicists with foresight ventured into this field, introducing new methods and tools while greatly expanding the data sources, thus substantially broadening the disciplinary scope of SciSci. By 2017 and 2018, seminal works such as "The science of science: from the perspective of complex systems" [31] and "Science of science" [57] popularized the term "science of science," attracting the attention of researchers from various disciplines, including physics, social sciences, mathematics, and information and computer science.

These newly engaged researchers approached the study of science as a complex system consisting of numerous components and interactions. In this framework, components are represented by nodes, while interactions are depicted as links. Following the introduction of small-world and scale-free networks at the turn of the 20th century, interest surged in Graph Neural Networks (GNNs) [58, 59], multilayer networks [60], and hypergraphs [61].

Graph Neural Networks (GNNs), alongside their variants including Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and GraphSAGE, have demonstrated remarkable performance across a variety of SciSci tasks in recent years [62]. For instance, Huang et al. [63] highlighted the importance of paper classification in literature retrieval and bibliometric analysis, noting that traditional text-based approaches—such as those relying solely on keywords, titles, and abstracts—often overlook valuable information contained within cited papers. To address this gap, they introduced an improved GNN model aimed at enhancing the accuracy of paper classification. To tackle the issue that most citation dynamic models focus solely on individual nodes rather than the entire citation structure, Feng et al. [64] proposed a method to learn the entire information cascade process as the input of a sequential deep neural network.

Multilayer networks excel at capturing the complex relationships inherent in scientific activities, such as citation networks, collaboration networks, and institutional networks [65]. Science can be conceptualized as a complex system comprising components with interactions. Traditional methods that represent these networks as single aggregated structures inevitably lead to information loss. To mitigate this issue, Wang et al. [66] combined co-citation networks, direct citation networks, and coupling networks into a multilayer network to predict potential academic collaborations in the field of gene editing. Their findings indicated that the multilayer network approach produced more accurate predictions than traditional collaboration network models.

Hypergraphs, which extend traditional graph structures, have gained recognition in the field of SciSci for their capacity to model complex, higher-order interactions. Contrary to Wang et al.'s approach [66], some researchers [67] advocate for viewing academic collaboration through the lens of team dynamics rather than merely as interactions between pairs of agents. In this context, hypergraphs or bipartite graphs are seen as more insightful alternatives to traditional frameworks, which are limited to representing relationships between pairs. These researchers also promote an integrated approach that considers both semantic and structural features in academic collaboration. Such a holistic perspective is essential for achieving a comprehensive understanding of the intricate patterns and outcomes of scholarly interactions.

In summary, physicists and computer scientists have made significant contributions to the advancement of SciSci by applying domain-specific methods and tools and integrating them with established research topics. As a result, the scope of data in SciSci has evolved from abstract databases like Web of Science and PubMed to include platforms such as Mendeley, OpenAlex, and Overton. Figure 5 [FIGURE:5] illustrates the common data types utilized in SciSci, providing insights into their nature and examples of sources from which they can be derived.

3 The Potential Applications of LLMs in the Field of SciSci

Since SciSci primarily focuses on the understanding, quantification, and prediction of science [31], this section discusses the impact of LLMs on SciSci from three perspectives: scientific perception, scientific evaluation, and scientific forecasting.

3.1 Scientific Perception

In this context, scientific perception refers to the process through which individuals interpret and understand information derived from scientific literature. To enhance the comprehension of scientific phenomena, researchers have observed and statistically described the Matthew Effect in research productivity and developed a suite of methods to map topics and semantic-enhanced themes [69, 70].

One traditional approach to knowledge topic extraction in SciSci involves generating co-word association maps based on frequently occurring words extracted from paper titles and keywords [69]. Essentially, co-word analysis involves extracting entities within sentences and establishing connections based on their relationships, which results in the formation of a single-mode network. One of the significant advantages of LLMs is their ability to efficiently extract entities and relationships from unstructured data, such as the full text of research papers in PDF format. This capability allows for a more comprehensive approach to data extraction compared to traditional methods. Once entities and their relationships are identified using LLMs, these can be visualized and manipulated within resulting networks.

In LLM-based entity relationship extraction, the relationships between entities are imbued with semantic dimensional information, presenting a richer and more nuanced array of information compared to networks constructed solely through traditional co-word methods. Additionally, the scope of entity relationship extraction can expand beyond just titles and keywords to encompass the entirety of research papers. This broadened scope significantly enhances the richness and complexity of the topics derived, thereby improving our understanding of the various dimensions of scientific knowledge. In summary, the integration of LLMs into SciSci offers new opportunities for deepening scientific perception by providing more sophisticated methods for topic extraction, relationship mapping, and data visualization, ultimately leading to a more comprehensive understanding of science as a complex system.

There are several ways to leverage LLMs to enhance traditional co-word analysis. For instance, techniques such as one-shot and few-shot prompting, along with prompt engineering in models like ChatGPT, can be employed to extract insights more effectively. Alternatively, users can directly call the API to access these capabilities [71]. To facilitate understanding, we have created a simple demonstration (see Fig. 6 [FIGURE:6]). The code for this demo is freely available on GitHub[2]. This demo is built upon the Kimi LLM framework and illustrates the entire process of entity relationship extraction using prompts, along with the visualization of the results through the networkX library. It is important to note that this demo serves as a basic example to showcase the feasibility of extracting entity relationships from unstructured text using LLMs for knowledge graph construction. For those looking to improve the accuracy and effectiveness of their results, we recommend exploring additional features such as tool/function calling and the JSON Model [71, 72]. By refining these techniques, users can enhance the precision and utility of the knowledge graphs they create.

[2]https://github.com/Gqiang-Liang/Simple-demo-for-NRE/tree/main

3.2 Scientific Evaluation

The evaluation of research work by universities, institutions, journals, researchers, and individual research articles has become a routine aspect of modern society. These evaluations aim to enhance understanding of scientific activities, such as monitoring and managing performance, disseminating contributions, justifying public expenditures by demonstrating research value to taxpayers and stakeholders, and informing funding decisions [73].

Typical approaches for scientific evaluation include peer review and a range of quantitative methods, such as bibliometrics, complex network analysis, and deep learning techniques. With the ongoing advancements in LLMs, we propose that the implementation of AI agents for scientific evaluation processes will emerge as a prominent direction in SciSci. To clarify the concept of AI agents, we represent them mathematically as follows: AI agents = LLMs + a set of skills (such as memory, function calling, and tool usage). The authors of this study have developed a "transformative research evaluation AI agent" based on ChatGLM during initial explorations[3]. However, it is recognized that the effectiveness of this AI agent still requires significant improvement. Nevertheless, these early explorations lay the groundwork for AI agent-based scientific evaluations in the field of SciSci.

By employing such AI agents, it becomes feasible to measure the influence of scientific publications, researchers, journals, and institutions. For instance, Figure 7 [FIGURE:7] illustrates the interface of an AI agent developed on the Dify platform, showcasing its potential application in the evaluation landscape. As these AI agents continue to evolve, they promise to transform the metrics and methods used in evaluating scientific research and its impact across various domains.

[3]https://chatglm.cn/main/gdetail/6632ecfeace21f9ff21cf4c0?lang=zh

3.3 Scientific Forecasting

Forecasting has always been at the forefront of planning and decision-making, as individuals and organizations seek to maximize utilities and minimize risks. As trends and interests in scientific research evolve over time, it is vital to identify and forecast the trends and future directions of development. Research communities have developed a series of tools to identify the trends and evolution of science, such as the iFORA system developed by the National Research University Higher School of Economics [74] and the Xinghuo Scientific Assistant[4] based on the Xinghuo LLM powered by iFLYTEK Co. Ltd. In SciSci, the academic success of researchers remains an everlasting topic of significant importance in management science and policy-making [75]. In the future, the integration of LLMs into scientific research forecasting is expected to provide substantial opportunities for advancements in SciSci and to represent a significant transformation of traditional SciSci methodologies.

Research fronts represent the cutting edge and growth frontier of scientific inquiry, having now become a focal point in global scientific and technological competition. Traditional forecasting methods employ co-citation clusters, co-citation clusters supplemented with citing articles, or direct citation clusters. Here, we propose an LLM-based multilayer network approach for forecasting research fronts. We evaluated current mainstream LLMs—including GPT-4o, Moonshoot-V1-8k, QwQ-32B-Preview, Gemini-Pro-1.5, and Deepseek-V3—and ultimately selected DeepSeek-V3 for multilayer network construction based on input/output costs, topic relevance, processing speed, and other key metrics (performance comparison shown in Table 2).

Table 2. Performance of Mainstream LLMs

<table>
<tr><th>Model</th><th>Input/Output Cost (/ProcessingSupportedOutputtopic(/per million tokens)</th><th>Speed (second/paper)</th><th>Context Length</th><th>Topic Relativeness</th></tr>
<tr><td>GPT-4o</td><td>8.5$±2.1| − | − |General||Deepseek − V3|10.8±3.5| − | − |General||Gemini − Pro − 1.5|6.2±1.8| − | − |General||Moonshoot − V1 − 8k|1.66|12.4±4.2| − | − ||QwQ − 32B − Preview|0.12|15.6±$5.0</td><td>-</td><td>-</td><td></td></tr>
</table>

We extracted Subject-Action-Object structures from publications using DeepSeek-V3 and constructed a multilayer network (Figure 8 [FIGURE:8]) with the PyMnet toolkit, thereby facilitating research front forecasting.

[4]https://paperlogin.iflytek.com/

4 Conclusions

The rapid advancement of LLMs presents significant opportunities for the evolution of SciSci. Researchers in this field can harness LLMs to explore previously unresolved research questions and to identify strategies for enhancing efficiency, particularly in areas such as name disambiguation. Furthermore, LLMs facilitate automated scientific research evaluation and trend prediction through the deployment of AI agents. However, these advancements also introduce challenges for traditional scientometricians. On one hand, the rise of LLMs calls for a deeper understanding and enhanced proficiency in computer technologies, including reinforcement learning and deep learning. On the other hand, it necessitates a reevaluation and redesign of existing theories and frameworks, potentially leading to the development of new tools and metrics in response to the AI era.

This paper offers a systematic review of the evolution of SciSci, the key technologies underlying LLMs, and the prospective applications of LLMs within this field. Due to space constraints, many potential applications of LLMs in specific domains of SciSci remain underexplored in this article. Examples include the integration of LLMs with full-text analysis, their combination with tasks such as sentiment analysis, semantic analysis, and text classification, their enhancement of citation analysis, and their potential to usher in an era of multimodal SciSci. These avenues are ripe for future exploration and can further enrich the landscape of SciSci research.

References

Committee, A.I.S., Artificial Intelligence Index Report 2024, L.F. Nestor Maslej, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark, Editor. 2024, Institute for Human-Centered AI: Stanford.

Gao, Q., et al., Semantic-enhanced topic evolution analysis: a combination of the dynamic topic model and word2vec. Scientometrics, 2022. 127(3): p. 1543-1563. DOI: 10.1007/s11192-022-04275-z.

Tohalino, J.A.V., T.C. Silva, and D.R. Amancio, Using word embedding to detect keywords in texts modeled as complex networks. Scientometrics, 2024. 129(7): p. 3599-3623. DOI: 10.1007/s11192-024-05055-7.

Yang, N., Z.Q. Zhang, and F.H. Huang, A study of BERT-based methods for formal citation identification of scientific data. Scientometrics, 2023. 128(11): p. 5865-5881. DOI: 10.1007/s11192-023-04833-z.

Lu, Y.H., et al., Knowledge graph enhanced citation recommendation model for patent examiners. Scientometrics, 129(4): 2181-2203. DOI: 10.1007/s11192-024-04966-9.

Song, B.W., C.J. Luan, and D.N. Liang, Identification of emerging technology topics (ETTs) using BERT-based model and sematic analysis: a perspective of multiple-field characteristics of patented inventions (MFCOPIs). Scientometrics, 2023. 128(11): p. 5883-5904. DOI: 10.1007/s11192-023-04819-x.

Ding, J.D., Y.F. Chen, and C. Liu, Exploring the research features of Nobel laureates in Physics based on the semantic similarity measurement. Scientometrics, 2023. 128(9): p. 5247-5275. DOI: 10.1007/s11192-023-04786-3.

Wayne Xin Zhao, K.Z., Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, Ji-Rong Wen A Survey of Large Language Models. 2023. DOI: 10.48550/arXiv.2303.18223.

Bi, K., et al., Accurate medium-range global weather forecasting with 3D neural networks. Nature, 2023. 619(7970): p. 533-538. DOI: 10.1038/s41586-023-06185-3.

Ye, S., et al., SuperAnimal pretrained pose estimation models for behavioral analysis. Nat Commun, 2024. 15(1): p. 5165. DOI: 10.1038/s41467-024-48792-2.

Li, T., et al., CancerGPT for few shot drug pair synergy prediction using large pretrained language models. NPJ Digit Med, 2024. 7(1): p. 1-19. DOI: 10.1038/s41746-024-01024-9.

Nazir, A. and Z. Wang, A Comprehensive Survey of ChatGPT: Advancements, Applications, Prospects, and Challenges. Meta Radiol, 2023. 1(2): p. 1-12. DOI: 10.1016/j.metrad.2023.100022.

Vaswani, A., et al., Attention is all you need, in Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, Curran Associates Inc.: Long Beach, California, USA. p. 6000–6010.

Jacobs, R.A., et al., Adaptive Mixtures of Local Experts. Neural Comput, 1991. 3(1): p. 79-87. DOI: 10.1162/neco.1991.3.1.79.

Devlin, J., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in Conference of the North-American-Chapter of Association-for-Computational-Linguistics - Human Language Technologies (NAACL-HLT). 2019. Minneapolis, MN: Assoc Computational Linguistics-Acl.

Kudo, T. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. 56th Annual Meeting Association-for-Computational-Linguistics (ACL). 2018. Melbourne, AUSTRALIA: Assoc Computational Linguistics-Acl.

Zhong, Y.S. and S.D. Goodfellow, Domain-specific language models pre-trained on construction management systems corpora. Automation in Construction, 2024. 160: p. 14. DOI: 10.1016/j.autcon.2024.105316.

Peng, L., et al., Human-AI collaboration: Unraveling the effects of user proficiency and AI agent capability in intelligent decision support systems. International Journal of Industrial Ergonomics, 2024. 103: p. 10. DOI: 10.1016/j.ergon.2024.103629.

Venkatesh Balavadhani Parthasarathy, A.Z., Aafaq Khan, and Arsalan Shahid The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities (Version 1.0). 2024. DOI: 10.48550/arXiv.2408.13296.

Ding, N., et al., Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 5(3): 220-235. DOI: 10.1038/s42256-023-00626-4.

Schmirler, R., M. Heinzinger, and B. Rost, Fine-tuning protein language models boosts predictions across diverse tasks. Nat Commun, 2024. 15(1): p. 1-10. DOI: 10.1038/s41467-024-51844-2.

Chitty-Venkata, K.T., et al., A survey of techniques for optimizing transformer inference. Journal of Systems Architecture, 102990. DOI: 10.1016/j.sysarc.2023.102990.

Ashish Vaswani, N.S., Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention is all you need Paper. in Advances in Neural Information Processing Systems 30. 2017. california.

Marvin, G., et al. Prompt Engineering in Large Language Models. 2024. Singapore: Springer Nature Singapore. DOI: 10.1007/978-981-99-7962-2_{30}.

Kiran Busch, A.R., Diana Sola, Henrik Leopold Just Tell Me: Prompt Engineering in Business Process Management. 2023. DOI: 10.48550/arXiv.2304.07183.

Shunyu Yao, D.Y., Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, Karthik Narasimhan Tree of Thoughts: Deliberate Problem Solving with Large Language Models. 2023. DOI: 10.48550/arXiv.2305.10601.

Yunfan Gao, Y.X., Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, Haofen Wang Retrieval-Augmented Generation for Large Language Models: A Survey. 2023. DOI: arXiv:2312.10997v5.

Chen Boqi, Y.F., Varro Daniel. Prompting or Fine-tuning? A Comparative Study of Large Language Models Taxonomy Construction. 2023 ACM/IEEE INTERNATIONAL CONFERENCE ON MODEL DRIVEN ENGINEERING LANGUAGES SYSTEMS COMPANION, MODELS-C. DOI: DOI10.1109/MODELS-C59198.2023.00097.

Changle Qu, S.D., Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, Ji-Rong Wen Tool Learning with Large Language Models: A Survey. 2024. DOI: 10.48550/arXiv.2405.17935.

Yujia Qin, S.L., Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, Maosong Sun ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. 2023. 10.48550/arXiv.2307.16789.

Zeng, A., et al., The science of science: From the perspective of complex systems. Physics Reports, 2017. 714-715: p. 1-73. DOI: 10.1016/j.physrep.2017.10.001.

Franzoni, C. and P. Stephan, Uncertainty and risk-taking in science: Meaning, measurement and management in peer review of research proposals. Research Policy, 2023. 52(3). DOI: 10.1016/j.respol.2022.104706.

Liang, G., et al., Knowledge recency to the birth of Nobel Prize-winning articles: Gender, career stage, and country. Journal of Informetrics, 2020. 14(3). DOI: 10.1016/j.joi.2020.101053.

Yang, A.J., Unveiling the impact and dual innovation of funded research. Journal of Informetrics, 2024. 18(1). DOI: 10.1016/j.joi.2023.101480.

Park, M., E. Leahey, and R.J. Funk, Papers and patents are becoming less disruptive over time. Nature, 2023. 613(7942): p. 138-144. DOI: 10.1038/s41586-022-05543-x.

Hu, X. and R. Rousseau, Scientific influence is not always visible: The phenomenon of under-cited influential publications. Journal of Informetrics, 2016. 10(4): p. 1079-1091. DOI: 10.1016/j.joi.2016.10.002.

Hou, J., et al., How do Price medalists' scholarly impact change before and after their awards? Scientometrics, 126(7): 5945-5981. DOI: 10.1007/s11192-021-03979-y.

Yang, A.J., et al., The k-step h-index in citation networks at the paper, author, and institution levels. Journal Informetrics, 17(4). DOI: 10.1016/j.joi.2023.101456.

Wang, Q. and L. Waltman, Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. Journal of Informetrics, 2016. 10(2): p. 347-364. DOI: 10.1016/j.joi.2016.02.003.

Singh, V.K., et al., The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. Scientometrics, 2021. 126(6): p. 5113-5142. DOI: 10.1007/s11192-021-03948-5.

Bornmann, L. and F. de Moya Anegón, What proportion of excellent papers makes an institution one of the best worldwide? Specifying thresholds for the interpretation of the results of the SCImago Institutions Ranking and the Leiden Ranking. Journal of the Association for Information Science and Technology, 2013. 64(4): p. 732-736. DOI: 10.1002/asi.23047.

Zhu, N., C. Liu, and Z. Yang, Team Size, Research Variety, and Research Performance: Do Coauthors' Coauthors Matter? Journal of Informetrics, 2021. 15(4). DOI: 10.1016/j.joi.2021.101205.

Dong, X., et al., Nobel Citation Effects on Scientific Publications: A Case Study in Physics. Information Processing & Management, 60(4). DOI: 10.1016/j.ipm.2023.103410.

Yu, D. and B. Xiang, An ESTs detection research based on paper entity mapping: Combining scientific text modeling and neural prophet. Journal of Informetrics, 2024. 18(4). DOI: 10.1016/j.joi.2024.101551.

Shiffrin, R.M. and K. Börner, Mapping knowledge domains. Proceedings of the National Academy of Sciences, 101(suppl_1): 5183-5185. DOI: 10.1073/pnas.0307852100.

Guo, L., Y. Wang, and M. Li, Exploration, exploitation and funding success: Evidence from junior scientists supported by the Chinese Young Scientists Fund. Journal of Informetrics, 2024. 18(2). DOI: 10.1016/j.joi.2024.101492.

Uzzi, B., et al., Atypical combinations and scientific impact. Science, 2013. 342(6157): p. 468-72. DOI: 10.1126/science.1240474.

Mingers, J. and L. Leydesdorff, A review of theory and practice in scientometrics. European Journal of Operational Research, 246(1): 1-19. DOI: 10.1016/j.ejor.2015.04.002.

Price, D.J.d.S., Networks of Scientific Papers. Science, 1965. 149(3683): p. 510-515. DOI: 10.1126/science.149.3683.510.

Price, D.D.S., A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science, 1976. 27(5): p. 292-306. DOI: 10.1002/asi.4630270505.

Merton, R.K., The Matthew Effect in Science. Science, 1968. 159(3810): p. 56-63. DOI: 10.1126/science.159.3810.56.

Crane, D. and N. Kaplan, Invisible Colleges: Diffusion of Knowledge in Scientific Communities. Physics Today, 1973. 26(1): p. 72-73. DOI: 10.1063/1.3127901.

Small, H., Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 2007. 24(4): p. 265-269. DOI: 10.1002/asi.4630240406.

Kessler, M.M., Bibliographic coupling between scientific papers. American Documentation, 2007. 14(1): p. 10-25. DOI: 10.1002/asi.5090140103.

Garfield, E., "Science Citation Index"–A New Dimension in Indexing. Science, 1964. 144(3619): p. 649-54. DOI: 10.1126/science.144.3619.649.

Boyack, K.W., R. Klavans, and K. Börner, Mapping the backbone of science. Scientometrics, 2005. 64(3): p. 351-374. DOI: 10.1007/s11192-005-0255-6.

Fortunato, S., al., Science of science. Science, 2018. 359(6379). DOI: 10.1126/science.aao0185.

Zhou, J., et al., Graph neural networks: A review of methods and applications. AI Open, 2020. 1: p. 57-81. DOI: 10.1016/j.aiopen.2021.01.001.

Kong, D., J. Yang, and L. Li, Early identification of technological convergence in numerical control machine tool: a deep learning approach. Scientometrics, 2020. 125(3): p. 1983-2009. DOI: 10.1007/s11192-020-03696-y.

De Domenico, M., et al., Mathematical Formulation of Multilayer Networks. Physical Review X, 2013. 3(4). DOI: 10.1103/PhysRevX.3.041022.

Antelmi, A., et al., A Survey on Hypergraph Representation Learning. ACM Computing Surveys, 2023. 56(1): p. 1-38. DOI: 10.1145/3605776.

Khemani, B., et al., A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. Journal of Big Data, 2024. 11(1). DOI: 10.1186/s40537-023-00876-4.

Huang, X., et al., ResGAT: an improved graph neural network based on multi-head attention mechanism and residual network for paper classification. Scientometrics, 2024. 129(2): p. 1015-1036. DOI: 10.1007/s11192-023-04898-w.

Feng, X., Q. Zhao, and R. Zhu, Towards popularity prediction of information cascades via degree distribution and deep neural networks. Journal of Informetrics, 2023. 17(3). DOI: 10.1016/j.joi.2023.101413.

De Domenico, M., et al., Identifying Modular Flows on Multilayer Networks Reveals Highly Overlapping Organization in Interconnected Systems. Physical Review X, 2015. 5(1). DOI: 10.1103/PhysRevX.5.011027.

Wang, F., et al., Collaboration prediction based on multilayer all-author tripartite citation networks: A case study of gene editing. Journal of Informetrics, 2023. 17(1). DOI: 10.1016/j.joi.2022.101374.

Taramasco, C., J.-P. Cointet, and C. Roth, Academic team formation as evolving hypergraphs. Scientometrics, 85(3): 721-740. DOI: 10.1007/s11192-010-0226-4.

Liu, L., et al., Data, measurement and empirical methods in the science of science. Nature Human Behaviour, 7(7): 1046-1058. DOI: 10.1038/s41562-023-01562-4.

Mane, K.K. and K. Börner, Mapping topics and topic bursts in PNAS. Proceedings of the National Academy of Sciences, 2004. 101(suppl_1): p. 5287-5290. DOI: 10.1073/pnas.0307626100.

Gao, Q., et al., Semantic-enhanced topic evolution analysis: a combination of the dynamic topic model and word2vec. Scientometrics, 2022. 127(3): p. 1543-1563. DOI: 10.1007/s11192-022-04275-z.

Xiang Wei, X.C., Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, Wenjuan Han ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT. 2023. 10.48550/arXiv.2302.10205.

Dagdelen, J., et al., Structured information extraction from scientific text with large language models. Nature Communications, 15(1). DOI: 10.1038/s41467-024-45563-x.

Penfield, T., et al., Assessment, evaluations, and definitions of research impact: A review. Research Evaluation, 2013. 23(1): p. 21-32. DOI: 10.1093/reseval/rvt021.

Lobanova, P., P. Bakhtin, and Y. Sergienko, Identifying and Visualizing Trends in Science, Technology, and Innovation Using SciBERT. IEEE Transactions on Engineering Management, 11898-11906. DOI: 10.1109/tem.2023.3306569.

Kong, X., et al., The Gene of Scientific Success. ACM Transactions on Knowledge Discovery from Data, 2020. 14(4): p. 1-19. DOI: 10.1145/3385530.