Identifying Research Hotspots and Research Gaps in Specific Research Area Based on Fine-Grained Information Extraction via Large Language Models

Yuling Sun[1,2][a], Xuening Cui[2], Aning Qin[1,*] and Jiatang Luo[3]
[1]National Science Library, Chinese Academy of Sciences, Beijing, China
[2]Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing, China
[3]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, China

Keywords: Large Language Model, Information Extraction, Scientific Data, Research Hotspots, Research Gaps.

Abstract: This paper constructs a fine-grained scientific data indicator framework using LLMs to conduct knowledge mining in a specific field of natural science and technology, with empirical analysis carried out in the domain of carbon dioxide conversion and utilization technology. Firstly, based on the characteristics of the technical field, we systematically established four key scientific data dimensions: products, technologies, materials, and performance. Subsequently, six key scientific data indicators were selected to characterize these dimensions. Finally, the extracted scientific data were employed to analyse research hotspots and gaps in the field. This approach effectively addresses the inherent limitations of traditional technology topic analysis, such as overly coarse metric granularity and the lack of quantitative features. Moreover, since these scientific data dimensions and indicators are generalizable to natural science and technology fields aimed at product development, the proposed methodology demonstrates broad applicability.

1 INTRODUCTION

Identification of research hotspots and gaps is essential for understanding disciplinary dynamics, optimizing resource allocation, and formulating policies. Scientific papers hold significant academic value and function as indicators of a field's developmental level. Consequently, the hotspots and cutting-edge directions of disciplinary research can be achieved through knowledge mining of scientific papers.

Existing studies typically integrate thematic dimensions (e.g., methodologies, products, research mechanisms) to uncover the aggregation and evolution, they exhibit two critical limitations: a) Inability to conduct detailed, in-depth analyses of specific key scientific data indicators; b) Neglect of fine-grained performance parameters in hotspot/gap identification. With the rapid growth of scientific publications, intelligence research now demands more refined and intelligent methods to process and analyze vast bibliographic data. Recent advances in natural language processing (NLP) techniques, specifically fine-grained data mining and large language models (LLMs), offer novel approaches for intelligence studies. Fine-grained data indicators provide deeper insights into research specifics. Leveraging their proficiency in scientific text comprehension, knowledge reasoning, and multimodal processing, LLMs are capable of undertaking sophisticated tasks related to text generation and information extraction.

This study aims to construct a multi-level knowledge network for specific scientific research area and leverage large language models to extract fine-grained, multi-labeled scientific data, thereby forming a research dataset of key domain indicators. Furthermore, we establish an analytical framework for identifying research hotspots and gaps based on fine-grained scientific indicator data. Through an empirical analysis in the research area of carbon capture, utilization and storage (CCUS), the

[a] https://orcid.org/0000-0001-6836-5530
* Corresponding author

Sun, Y., Cui, X., Qin, A. and Luo, J.
Identifying Research Hotspots and Research Gaps in Specific Research Area Based on Fine-Grained Information Extraction via Large Language Models.
DOI: 10.5220/0013824900004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 473-480

effectiveness of the proposed framework is validated. This study enriches the methods and perspectives of knowledge mining in disciplinary fields.

2 LITERATURE REVIEW

2.1 Identification of Research Hotspots and Discovery of Research Gaps

Traditional approaches primarily include keyword co-occurrence analysis, citation network analysis and topic models.

Keyword co-occurrence analysis can reveal thematic clusters and evolution by extracting frequently co-occurring terms. Wang et al. (Wang et al., 2023) conducted bibliometric analysis on 4,922 articles in the field of carbon neutrality based on the Web of Science (WoS) database, using Citespace and Bibliometrix functions for descriptive statistics and co-occurrence analysis of keywords. Xu et al. (Xu et al., 2016) combined keyword co-occurrence with a cosine similarity algorithm, integrating academic papers and patents to identify research frontier hotspots in the LED field.

Citation network analysis uncover core research and directional evolution through highly cited publications and citation chains. Morris et al. Morris et al.,2003) employed bibliographic coupling analysis to construct a timeline of anthrax-research hotspots, visualizing the evolution of active themes. Chang et al. (Chang et al., 2015) combined keywords, bibliographic-coupling and co-citation analyses to explore the evolution of hotspots in library and information science over two decades.

Topic models automatically identify latent topic distributions from enormous text by applying various text analysis techniques. The predominant method is the Latent Dirichlet Allocation (LDA) model. Liu (Liu, 2025) applied the LDA model to 559 articles on biosecurity legislation (1996–2023) and identified nine key hotspots and significant trends. Tan and Xiong (Tan and Xiong, 2021) extracted topics via LDA model from core data-mining journals in CNKI and Web of Science, combining topic life cycles with time-slicing to map evolutionary paths.

Based on their activity levels and persistence, research hotspots can be categorized into three kinds: sustained, emerging, and potential hotspots. Liu et al. (Liu et al., 2023) identified sustained hotspots in computer science by measuring keywords survival metrics (time/frequency), applying logistic regression to analyse influencing factors of keyword survival patterns. Hu et al. (Hu et al., 2024) leveraged the global blockchain patent literature, integrating LDA, Word2Vec and BERT to construct a fine-grained topic-mining framework that surfaced emerging technological hotspots. Thakuria and Deka (Thakuria and Deka, 2024) utilized topic modelling to identify prevalent potential hotspots in Library and Information Science (LIS) journals between 2013 and 2022, and reveal unknown research themes.

Gap Analysis is a strategic analysis method used to evaluate the differences between the current state and the expected or target state. In this paper, research gaps refer to important issues that have not been adequately studied, received insufficient attention, or have become disconnected from policy or industry expectations in the existing literature. Currently, the main approaches to identifying research gaps include systematic literature reviews (Anton et al., 2022) and expert consultation (Mohtasham et al., 2023). However, these qualitative methods suffer from strong subjectivity and low efficiency, making it difficult to rapidly and accurately pinpoint gaps within the massive body of literature. Some data-driven quantitative techniques have also been applied to gap analysis. Westgate et al. (Westgate et al., 2015) employed LDA model and statistical methods (cluster analysis, regression, and network analysis) to investigate trends and identify potential research gaps within the scientific literature.

The common limitation of existing hotspot identification methods is that they mainly rely on a single type of data (such as keywords, citation relationships or subject terms), and the analytical perspective is concentrated on the macro aggregation of disciplinary themes. There is a lack of in-depth mining of fine-grained performance indicators that support these macro themes, as well as an overall correlation framework for integration analysis of multi-dimensional indicators. Quantitative analytical methods for identifying research gaps remain scarce. Existing strategies mostly adopt qualitative approaches that infer “under-studied” areas from bibliographic coverage or topic popularity. Quantitative comparison between the specific performance parameters reported in the literature and the targets set by policy plans, industry technical standards, or real-world application is rarely considered. Consequently, substantial performance gaps at the technology level cannot be pinpointed with precision.

2.2 Fine-Grained Information Extraction Based on Large Language Models

Information Extraction typically consists of three sub-tasks: Named Entity Recognition (NER), Relation

Extraction (RE), and Event Extraction (EE). Fine-grained information extraction (FGIE) can extract semantically specific and application-oriented data (e.g., research methods, experimental data, performance metrics) from massive academic literature. These granular data indicators provide deep technical insights.

Early approaches relied mainly on manually defined rules and pattern matching, which were highly interpretable but had poor generalization ability. With the rise of deep learning, neural-network-based methods (e.g., LSTM, BiLSTM, CNN) the automatically learn textual features have become mainstream. Yu et al. (Yu et al., 2019) improved the Bootstrapping algorithm and built a deep learning model to extract four fine-grained knowledge units from the abstracts of 17,756 ACL papers. Onishi et al. (Onishi et al.,2018) constructed a weakly supervised ML framework that uses CNNs trained on a materials-microstructure corpus to extract "processing-structure-property" triplets. Rodríguez et al. (Rodríguez et al.,2022) proposed an attention mechanism based on the noun-type syntactic elements, combining BPEmb vectors and the Flair model to address two sub-tasks of fine-grained NER: Named Entity Detection and Named Entity Typing.

Based these advances, pre-trained encoder models such as BERT and RoBERTa marked a new stage in FGIE. By pre-training on large general corpora and fine-tuning on task-specific datasets (e.g., SQuAD). They achieved stronger transferability and became the dominant approach for entity-centric extraction. Domain-adapted variants such as MatSciBERT have further demonstrated effectiveness in specialized areas like materials science (Gupta et al., 2022). Similarly, RoBERTa-based architectures have been widely applied in fine-grained biomedical IE and radiology text analysis (Datta & Roberts, 2022; Yin et al., 2021). This pre-train–fine-tune paradigm greatly reduced reliance on handcrafted rules and task-specific feature engineering, and for a time they became the mainstream solution for FGIE. However, these encoder-only models still had critical limitations: they required large-scale labelled data for downstream adaptation, had limited zero-shot generalization, and captured domain-specific knowledge only weakly.

Recently, large language models such as GPT-3.5/4, PaLM 2, DeepSeek and Claude have developed powerful language understanding and generation capabilities, enabling them to efficiently extract key fine-grained information from scientific literature. Fine-grained information extraction based on LLMs mainly includes the following methods: a) Prompt engineering, designing appropriate prompts to guide LLMs to directly extract the required information from the text in zero-shot or few-shot scenarios. For instance, Wu et al. (Wu et al.,2025) enhanced LLMs for miRNA information extraction through diversified prompt strategies and systematically compared the performance of GPT-4o, Gemini, and Claude in NER and RE. b) Task decomposition, breaking down complex outputs into a series of simpler questions and answers, which enhances the reasoning ability of LLMs. Wei et al. (Wei et al., 2023) implemented zero-shot FGIE via ChatGPT multi-turn question answering (QA), decomposing the complex IE task into two conversation rounds: a) identifying entity, relation, and event types in a sentence; b) converting unlabelled text directly into fine-grained structured knowledge through chain-of-question templates. Qiao et al. (2025) introduced a novel FGER-GPT method, which employs multiple inference chains and a hierarchical strategy for recognizing fine-grained entities, significantly enhancing the performance of LLM in fine-grained entity recognition and effectively alleviating the problems of label lack and hallucination. c) Fine-tuning, further training the general LLMs on specific scietific research area labeled data to adapt it to specific IE tasks. Dagdelen(2024) utilized fine-tuned LLMs such as GPT-3 and Llama-2 for joint named entity and relation extraction from scientific texts, and verified its effectiveness on materials science texts.

Overall, neural network methods advanced the automation of fine-grained information extraction by learning textual features directly, while pre-trained encoder models such as BERT and RoBERTa further improved performance through large-scale pre-training and task-specific fine-tuning, reducing the need for handcrafted features. However, these approaches typically depend heavily on large labelled datasets and exhibit limited zero-shot generalization. In contrast, large language models offer far greater flexibility and contextual understanding, enabling effective fine-grained information extraction without task-specific supervision, though the issue of hallucination in their outputs remains a critical challenge.

The LLM-driven FGIE framework introduced in this study transforms the descriptive macro analysis of research hotspots into a fine-grained quantitative assessment, and compares literature data with policy demands to accurately identify research gaps. By integrating prompt engineering and progressive optimization strategies, the method reduces the

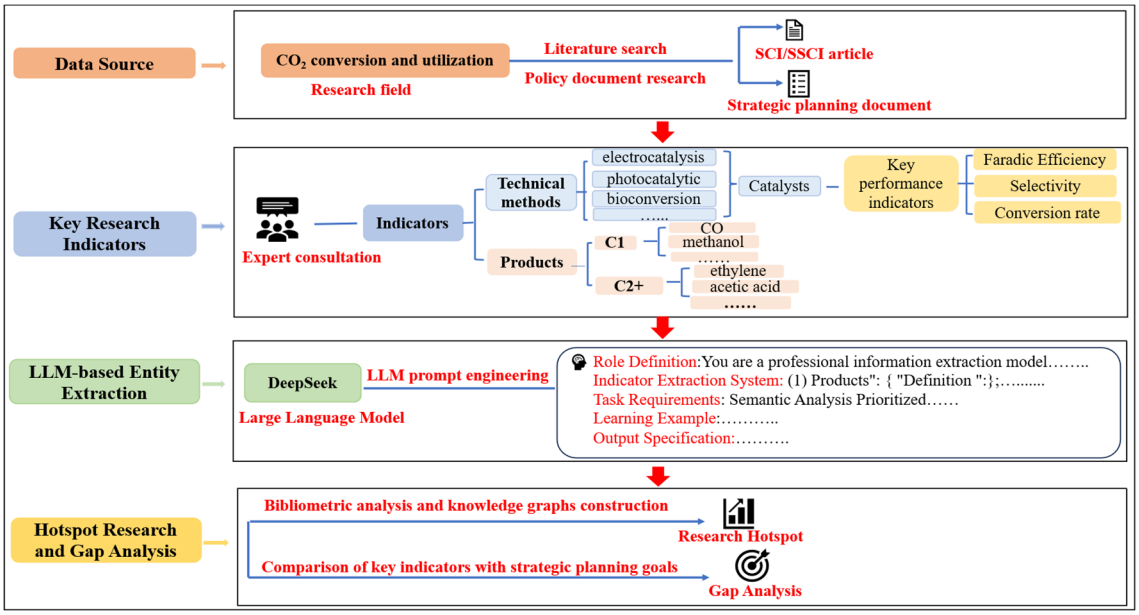

Figure 1: The research framework of this paper.

reliance on labelled data and enhancing the complex text processing capability of LLMs. This approach offers a transferable pathway for specific scientific research knowledge mining.

3 RESEARCH FRAMEWORK

We have chosen the technology research and development field with product as the research object. Based on the scientific data characteristics of these specific fields, we developed a four scientific data dimensions that including products, technology, materials and key performance. Then, based on the characteristics of specific technology research and development fields and expert consultation, we further selected specific scientific data indicators that can characterize the above dimensions.

Leveraging semantic understanding techniques, the DeepSeek-V3 large language model was employed to automate indicator extraction, establishing a comprehensive specific scientific research fine-grained database. This database enables multidimensional bibliometric analyses of methodological innovation, material applications, product development, and their interdisciplinary intersections. Concurrently, it facilitates comparative analysis between research metrics and policy targets, systematically evaluating alignment with current research trajectories. Figure 1 presents the research framework.

Despite empirical validation within CO_2 conversion and utilization, the framework exhibits high modularity and prompt-level portability, enabling analogous fine-grained indicator extraction and knowledge mining across product-oriented technology research area—notably chemical engineering and materials discovery.

3.1 Data Source

Data were extracted from the Web of Science Core Collection (WOS). A systematic retrieval was conducted for Science Citation Index (SCI) and Social Sciences Citation Index (SSCI) articles (2020-2024), resulting in a total of 15,695 publications. We additionally collected 81 strategic planning documents related to CCUS from the official government website of major economies (e.g., the United States, the European Union, the United Kingdom, Germany, and France).

3.2 Key Indicator Entity

Focusing on CO_2 conversion and utilization as an empirical testbed, through expert consultation, we further selected six key scientific data indicators to characterize the four dimensions mentioned above (Table 1). Specifically, The final product of CO_2 conversion in research papers is used as a product such as methanol, ethanol, ethylene, etc; the technological pathway for CO_2 conversion is characterized by techniques such as electrocatalysis,

photocatalysis, etc; catalysts are used to characterize materials; three indicators including faradic efficiency, target product selectivity, and conversion rate are used to characterize performance.

Table 1: Key dimensions and indicators.

No.	Dimensions	Indicators
1	Products	Conversion Products
2	Technology	Technical Pathways
3	Materials	Catalyst
4	Performance	Faradic Efficiency
5		Target Product Selectivity
6		CO_2 Conversion Rate

3.3 LLM-Based Entity Extraction

To compare the performance of a BERT-style pre-trained model and a large language model on FGIE, we conducted a preliminary experiment focusing on a single entity type, namely product. Using a unified question answering (QA) framework, the document was provided as context, and the extraction target was expressed as a natural-language query (e.g., "What products are formed from CO_2 conversion?"). Within this setup, RoBERTa (Liu et al., 2019) achieved only around 42% exact-match accuracy, whereas DeepSeek-V3 (zero-shot) reached close to 99%, corresponding to a +57 percentage-point and 2.4× relative improvement (Table 2). Although this is only a pilot study restricted to one entity type, the results already highlight the substantial advantage of large language models over BERT-style encoders in QA-based FGIE.

Table 2: Comparison of product-recognition performance between DeepSeek and RoBERTa.

Model	Accuracy(%)	Diffrence(%)
RoBERTa	41.89	—
DeepSeek-V3	98.65	+56.76

Leveraging the large language model DeepSeek-V3, we construct a target entity extraction system using prompt engineering (Figure 2). Our three-stage progressive optimization strategy includes: First, the model learns basic extraction paradigms using a small set of sample examples, with prompts iteratively refined through manual verification. Next, building upon the verified baseline model, incremental optimization is performed using a medium-sized sample set. This stage involves dynamically adjusting the prompt and supplementing it with 5% sampling for quality verification. Finally, the full dataset is processed automatically, supported by a 2% sampling review mechanism to ensure output stability. This phased optimization approach significantly enhances the model's adaptability to complex texts.

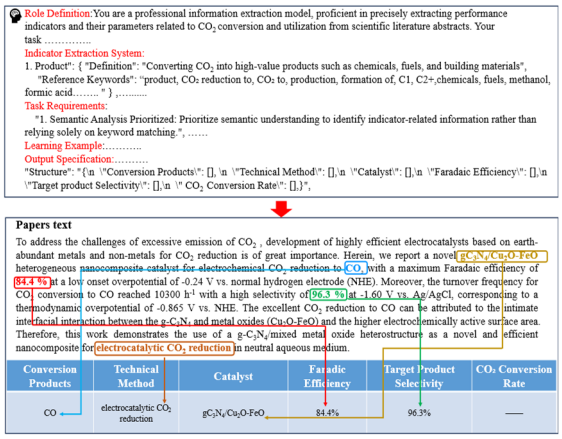

Figure 2: LLM-based prompt engineering for named entity extraction in scientific literature.

3.4 Knowledge Graph Construction

Within diverse technological pathways for CO_2 conversion and utilization, catalyst selection and design exhibit significant differences. Different conversion methods (such as electrocatalysis, photocatalytic, thermocatalytic, and bioconversion) require specific types and properties of catalysts due to their distinct reaction mechanisms, operating conditions, and target products. Therefore, clarifying the applicable types of efficient, stable catalysts for each technical pathway is critical to optimizing reaction performance, lowering energy consumption and costs, and advancing the practical application of specific CO_2 conversion routes. By extracting the entity associations between technical methods and catalysts and conducting network visualization using Gephi software, we have successfully constructed a knowledge graph illustrating the interconnections between them.

4 RESULTS AND ANALYSIS

4.1 Research Hotspots Analysis

Statistical analysis of the conversion products (Figure 3) reveals that C1 compounds dominate, accounting for 72% of the publications surveyed, significantly higher than the share for C2+ products (28%). Among C1 products, carbon monoxide (CO), methane (CH_4), methanol (CH_3OH), formic acid (HCOOH), and formate are the most extensively studied. Research on

CO_2 conversion to C2 products primarily focuses on ethylene, ethanol, cyclic carbonates, acetate, dimethyl ether, acetic acid, and ethane.

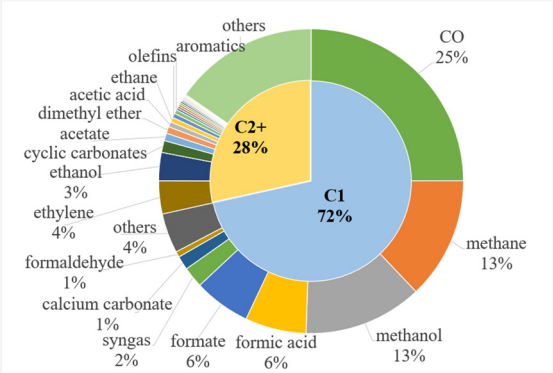

Figure 3: Research hotspots of CO_2 conversion products.

From the technical method–catalyst knowledge map (Figure 4), we can identify the main catalysts employed by different technologies. It can be observed that different CO_2 conversion technologies prioritize distinct types of catalysts, owing to variations in their reaction mechanisms and operational conditions. Electrocatalysis CO_2 reduction primarily employs metals (e.g., Cu, Ag) and their oxides, bimetallics, metal-organic frameworks (MOFs), molecular catalysts (e.g., metal porphyrins), and nitrogen-doped carbon materials. Photocatalytic CO_2 reduction uses metal oxides (e.g., TiO_2, Cu_2O), nitrides (e.g., g-C_3N_4), sulfides (e.g., CdS), MOFs, molecular complexes (e.g., phthalocyanines), and hybrid heterojunctions. CO_2 hydrogenation relies on metal oxides (e.g., CeO_2, In_2O_3), metals (e.g., Cu, Fe, Ru), bifunctional zeolites, bimetallic/ternary catalysts (e.g., Cu/ZnO/Al_2O_3), and carbides. Cycloaddition catalysts include ionic liquids, MOFs, porous polymers, metal complexes, and covalent organic frameworks (COFs). CO_2 bioconversion utilizes enzymes (e.g, formate dehydrogenase, carbonic anhydrase) and microbial cells (e.g., Clostridium, engineered strains).

4.2 Research Gap Analysis

Through bibliometric analysis, this study identifies key characteristics and persistent challenges in current CO_2 conversion research. The product distribution reveals a pronounced focus on C1 chemicals (carbon monoxide, methane, and methanol), while industrially critical higher-carbon

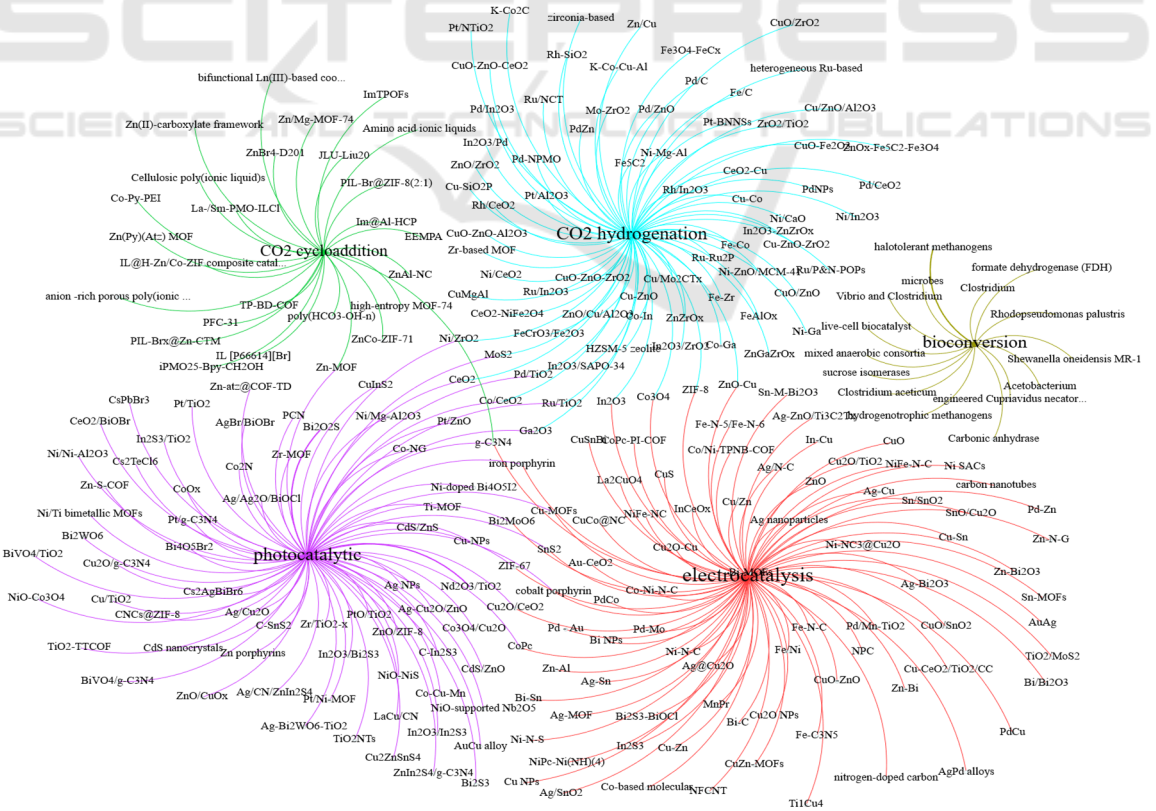

Figure 4: The knowledge map of CO_2 conversion technical pathways and catalysts.

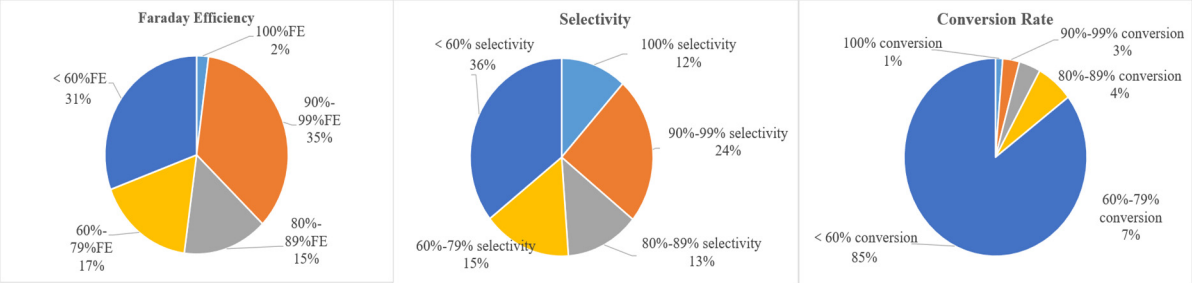

Figure 5: Faraday efficiency, selectivity, conversion rate distribution.

compounds (e.g., olefins, aromatics like toluene/xylene, and synthetic fuels) remain understudied. Performance metrics (Figure 5) reveal significant limitations: only 12% of studies achieve≥ 99% product selectivity (predominantly for C1 species), while advanced carbon products exhibit substantially lower selectivity. CO_2 conversion efficiencies exceed 90% occur in merely 4% of cases, and full conversion (100%) in merely 1%. Notably, only 37% of catalytic systems demonstrate Faradaic efficiencies above 90%, indicating substantial energy efficiency deficits.

The current technical status lags significantly behind the strategic objectives of major economies. For instance, Japan's updated Carbon Recycling Technology Roadmap prioritizes commercializing polycarbonate and bio-jet fuel by 2030, followed by industrial-scale olefin, aromatic compound, and synthetic fuel production by ~2040. The significant gap between current research progress and these industrial targets underscores fundamental hurdles in CO_2 conversion, particularly in precise selectivity control, energy utilization optimization, and multicomponent product.

5 CONCLUSION AND PROSPECT

This study integrates LLM-based semantic parsing with prompt engineering for in-depth knowledge mining in the specific research area. The effectiveness of the method has been empirically verified in the research area of CO_2 conversion and utilization. By developing a multi-stage prompt framework, we systematically extracted key scientific indicators to construct a fine-grained research area database. Based on this methodological system, we have identified some new research hotspots and gaps in the field of carbon dioxide conversion and utilization, which is of great significance for optimizing the research layout and direction in this area. This approach effectively overcomes limitations of traditional technical theme analysis, such as coarse indicator granularity and insufficient quantitative characterization, and enriches the methods and perspectives of knowledge discovery in the field of scientific research.

Current limitations include suboptimal accuracy in extracting specific key performance indicators. Further research will prioritize optimizing prompt engineering to enhance extraction precision. Additionally, while this study focuses on academic literature, we plan to expand data sources to include patents and industrial reports, thereby building a more comprehensive dataset. Leveraging this foundation, we will advance knowledge graph development for CO_2 conversion technology, emphasizing core functionalities such as technology development roadmap and patent-technology correlation analysis, ultimately supporting strategic advancements in this critical field.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support provided by Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDC0230605 and Special Project of Strategic Research and Decision Support System Construction of Chinese Academy of Sciences (GHJ-ZLZX-2025-07).

REFERENCES

Anton, E., Oesterreich, T. D., & Teuteberg, F. (2022). the Property of Being Causal–the Conduct of Qualitative Comparative Analysis in Information Systems Research. Information & Management, 59(3), 103619.

Chang, Y. W., Huang, M. H., & Lin, C. W. (2015). Evolution of Research Subjects in Library and Information Science Based on Keyword, Bibliographical Coupling, and Co-Citation Analyses. Scientometrics, 105(3), 2071-2087.

Dagdelen, J., Dunn, a., Lee, S., Walker, N., Rosen, a. S., Ceder, G., ... & Jain, a. (2024). Structured Information Extraction from Scientific Text with Large Language Models. Nature Communications, 15(1), 1418.

Datta, S., & Roberts, K. (2022). Fine-Grained Spatial Information Extraction in Radiology as Two-Turn Question Answering with BERT. International Journal of Medical Informatics, 162, 104754.

Gupta, T., Zaki, M., Krishnan, N. M. a., & Mausam. (2022). Matscibert: a Materials Domain Language Model for Text Mining and Information Extraction. Npj Computational Materials, 8, 141.

Hu, Z., Wang, M., & Han, Y. (2024). Multidimensional Indicator Identification and Evolution Analysis of Emerging Technology Topics Based on LDA2Vec-BERT——a Case Study of Blockchain Technology in the Field of Disruptive Technology. Journal of Modern Information, 44(9), 42-58. (in Chinese)

Liu, Y. (2025). Discovering Topics and Trends in Biosecurity Law Research: A Machine Learning Approach. One Health, 20, 100964.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … & Stoyanov, V. (2019). Roberta: a Robustly Optimized BERT Pretraining Approach. Arxiv Preprint Arxiv:1907.11692. Https://Arxiv.Org/Abs/1907.11692.

Liu, Z., Yang, J., & Ma, Y. (2023). Research on Keyword Survival Model from the Perspective of Term Function. Journal of Intelligence, 42(8), 150-156+176. (in Chinese)

Mohtasham, F., Yazdizadeh, B., & Mobinizadeh, M. (2023). Research Gaps Identified in Iran's Health Technology Assessment Reports. Health Research Policy and Systems, 21(1), 132.

Morris, S. a., Yen, G., Wu, Z., & Asnake, B. (2003). Time Line Visualization of Research Fronts. Journal of the American Society for Information Science and Technology, 54(5), 413-422.

Onishi, T., Kadohira, T., & Watanabe, I. (2018). Relation Extraction with Weakly Supervised Learning Based on Process-Structure-Property-Performance Reciprocity. Science and Technology of Advanced Materials, 19(1), 649-659.

Qiao, X., Gu, S., Cheng, J., Peng, C., Xiong, Z., Shen, H., & Jiang, G. (2025). Fine-Grained Entity Recognition via Large Language Models. IEEE Transactions on Neural Networks and Learning Systems.

Rodríguez, a. J. C., Castro, D. C., & García, S. H. (2022). Noun-Based Attention Mechanism for Fine-Grained Named Entity Recognition. Expert Systems with Applications, 193, 116406.

Tan, C., & Xiong, M. (2021). Contrastive Analysis at Home and Abroad on the Evolution of Hot Topics in the Field of Data Mining Based on LDA Model. Information Science, 39(4), 174-185. (in Chinese)

Thakuria, a., & Deka, D. (2024). a Decadal Study on Identifying Latent Topics and Research Trends in Open Access LIS Journals Using Topic Modeling Approach. Scientometrics, 129(7), 3841-3869.

Wang, G., Shi, R., Cheng, W., Gao, L., & Huang, X. (2023). Bibliometric Analysis for Carbon Neutrality with Hotspots, Frontiers, and Emerging Trends Between 1991 and 2022. International Journal of Environmental Research and Public Health, 20(2), 926.

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., ... & Han, W. (2023). Chatie: Zero-Shot Information Extraction via Chatting with Chatgpt. Arxiv Preprint Arxiv:2302.10205.

Westgate, M. J., Barton, P. S., Pierson, J. C., & Lindenmayer, D. B. (2015). Text Analysis Tools for Identification of Emerging Topics and Research Gaps in Conservation Science. Conservation Biology, 29(6), 1606-1614.

Wu, R., Zong, H., Wu, E., Li, J., Zhou, Y., Zhang, C., ... & Shen, B. (2025). Improving Large Language Models for Mirna Information Extraction via Prompt Engineering. Computer Methods and Programs in Biomedicine, 109033.

Xu, X., Zheng, Y., & Liu, Z. (2016). Study on the Method of Identifying Research Fronts Based on Scientific Papers and Patents. Library and Information Service, 60 (24), 97-106. (in Chinese)

Yin, X., Zheng, S., & Wang, Q. (2021). Fine-Grained Chinese Named Entity Recognition Based on Roberta-WWM-Bilstm-CRF Model. 6th International Conference on Intelligent Computing and Signal Processing.

Yu, L., Qian, L., Fu, C., & Zhao, H. (2019). Extracting Fine-Grained Knowledge Units from Texts with Deep Learning. Data Analysis and Knowledge Discovery, 3(1), 38-45. (in Chinese)

Zhang, X., Li, P., & Li, H. (2020). AMBERT: a Pre-Trained Language Model with Multi-Grained Tokenization. Arxiv Preprint Arxiv:2008.11869.