ALLDATA 2026

The Twelfth International Conference on Big Data, Small Data, Linked Data and Open Data

ISBN: 978-1-68558-397-2

May 24 - 28, 2026

Venice, Italy

ALLDATA 2026 Editors

Petre Dini, IARIA, USA/EU

ALLDATA 2026

Forward

The Twelfth International Conference on Big Data, Small Data, Linked Data and Open Data (ALLDATA 2026), held between May 24-28, 2026 in Venice, Italy, continued a series of events bridging the concepts and the communities devoted to each of data categories for a better understanding of data semantics and their use, by taking advantage from the development of Semantic Web, Deep Web, Internet, non-SQL and SQL structures, progresses in data processing, and the new tendency for acceptance of open environments.

The volume and the complexity of available information overwhelm human and computing resources. Several approaches, technologies and tools are dealing with different types of data when searching, mining, learning and managing existing and increasingly growing information. From understanding Small data, the academia and industry recently embraced Big data, Linked data, and Open data. Each of these concepts carries specific foundations, algorithms and techniques, and is suitable and successful for different kinds of application. While approaching each concept from a silo point of view allows a better understanding (and potential optimization), no application or service can be developed without considering all data types mentioned above.

We welcomed academic, research and industry contributions. The conference had the following tracks:

- Big Data
- Open Data
- Linked Data
- Challenges in processing Big Data and applications

We take here the opportunity to warmly thank all the members of the ALLDATA 2026 technical program committee, as well as all the reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to ALLDATA 2026. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We also thank the members of the ALLDATA 2026 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that ALLDATA 2026 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the field of all data. We also hope that Venice provided a pleasant environment during the conference and everyone saved some time to enjoy the historic charm of the city.

ALLDATA 2026 Chairs

ALLDATA Steering Committee

Anastasija Nikiforova, European Open Science Cloud Task Force "FAIR metrics and data quality" & University of Tartu, Estonia

Gerold Hoelzl, University of Passau, Germany

Vikas Thammanna Gowda, Champlain College, USA

ALLDATA Publicity Chairs
José Miguel Jiménez, Universitat Politecnica de Valencia, Spain
Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
Laura Garcia, Universidad Politécnica de Cartagena, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain

ALLDATA 2026

Committee

ALLDATA Steering Committee

Anastasija Nikiforova, European Open Science Cloud Task Force "FAIR metrics and data quality" & University of Tartu, Estonia
Gerold Hoelzl, University of Passau, Germany
Vikas Thammanna Gowda , Champlain College, USA

ALLDATA 2026 Publicity Chairs

José Miguel Jiménez, Universitat Politecnica de Valencia, Spain
Francisco Javier Díaz Blasco, Universitat Politècnica de València, Spain
Ali Ahmad, Universitat Politècnica de València, Spain
Vikas Thammanna Gowda, Champlain College, USA
Laura Garcia, Universidad Politécnica de Cartagena, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain

ALLDATA 2026 Technical Program Committee

Alessandra Alaniz Macedo, University of São Paulo (USP), Brazil
Hugo Alatrista-Salas, Universidad del Pacífico, Peru
Qiushi Bai, Microsoft (Azure Data R&D), USA
Ujjwal Baid, Georgia Institute of Technology and Emory University, USA
Syed Raza Bashir, Toronto Metropolitan University, Canada
Gábor Bella, University of Trento, Italy
Ghada Besbes, Riadi Laboratory | ENSI | University of Manouba, Tunisia
Sandjai Bhulai, Vrije Universiteit Amsterdam, Netherlands
Dhanasak Ken Bhumichai, Navaminda Kasatriyadhiraj Royal Air Force Academy, Thailand
Ayan Biswas, Los Alamos National Laboratory, USA
Jean-Yves Blaise, CNRS (French National Centre for Scientific Research) | UMR CNRS/MC 3495 MAP, France
Doulkifli Boukraa, LaRIA Lab | University of Jijel, Algeria
Ozgu Can, Ege University, Turkey
Ramon Alberto Carrasco Gonzalez, Universidad Complutense de Madrid, Spain
Richard Chbeir, Université de Pau et des Pays de l'Adour (UPPA), France
Rachid Chelouah, Ecole Internationale des Sciences du Traitement de l'Information (*EISTI*), Cergy, France
Haihua Chen, University of North Texas, USA
Wei-Kuo Chiang, National Chung Cheng University, Taiwan, China
Esma Nur Cinicioglu, Istanbul University - School of Business, Turkey
Stefano Cirillo, University of Salerno, Italy
António Correia, University of Jyväskylä, Finland
Cinzia Daraio, Sapienza University of Rome, Italy
Subhasis Dasgupta, University of California San Diego, USA

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Hybrid Intelligence Framework for Identifying Frontier Technologies through Project Linkage:A Case Study of DARPA Programs

Xiang-li Zhu, Yifan Wang, Xiaoping Liu
National Science Library, Chinese Academy of Sciences; School of Economics and Management, University of Chinese Academy of Sciences
Beijing, China
email: zhuxl@mail.las.ac.cn

Abstract—**In the context of the global technological revolution and industrial transformation, the identification of frontier technologies has become a critical component of national strategic competition. However, traditional methods based on citation analysis or patent classification suffer from significant time lags and fail to comprehensively capture the entire lifecycle from technological conception to practical application. To address this, this paper proposes a novel project-linkage paradigm for frontier technology identification, constructing an integrated framework that combines data-driven analysis, intelligent algorithms, and multidimensional assessment. The framework utilizes Large Language Models (LLMs, such as DeepSeek V3) to enhance textual feature extraction and combines Word2Vec vectorization with K-means clustering for technical topic discovery, establishing technology evolution chains through cross-source semantic associations between project requirements and research outputs. Using Defense Advanced Research Projects Agency (DARPA)-funded programs from 2009 to 2025 as empirical subjects, the study finds that: (1) the response rate of research projects to academic publications increased significantly from 80% to 98.3%, indicating that DARPA projects are shifting from following academia to leading academia; (2) three tiers of frontier technologies were identified—mature frontiers, emerging frontiers, and potential frontiers. The results show that the proposed hybrid intelligence framework effectively identifies prospective technological breakthroughs, offering precise support for science and technology decision-making.**

Keywords—Hybrid Intelligence; Frontier Technology Identification; Multi-source Data Fusion; Research Projects; Large Language Models.

I. INTRODUCTION

In the global competition for technological leadership, frontier technologies serve as strategic assets, and their early identification is crucial for securing a competitive edge. However, traditional methods of technology identification, such as citation analysis and patent classification, rely on single data sources and suffer from significant time lags, making it difficult to capture the full trajectory from fundamental research to applied innovation [1][2]. While research projects are the primary point of resource allocation in science and technology, they encapsulate explicit strategic orientations and forward-looking requirements, often containing valuable insights into future technological directions. Previous studies have largely treated project metadata as secondary information, overlooking the rich technological signals embedded within project texts.

Recent advancements in big data analytics and artificial intelligence, particularly the capabilities of LLMs for semantic understanding, have opened new possibilities for identifying frontier technologies from large and diverse datasets [3]. This underscores the need for novel methodologies that integrate multi-source data—such as project descriptions, academic publications, and patents—along with intelligent algorithms to achieve more timely, accurate, and comprehensive identification of frontier technologies.

To address this gap, this study introduces a new project-linkage paradigm for frontier technology identification. Unlike traditional methods, we position research projects as the starting point, conducting semantic association analysis between project descriptions and their subsequent publications and patents. This enables cross-source mapping, connecting project inputs (funding and requirements) to research outputs (publications and patents). We propose a data-driven framework that leverages Large Language Models to enhance text extraction capabilities, integrates vectorization and clustering techniques to ensure reproducibility, and incorporates a multidimensional assessment system to classify identified technologies by maturity. To validate our methodology, we conduct an empirical study using DARPA-funded projects and their associated outputs from 2009 to 2025, analyzing the linkage between projects and publications/patents to identify emerging frontier technologies and categorize them based on their development stage.

The main contributions of this work are threefold: (1) A novel project-linkage paradigm that positions research project texts as primary signals for frontier technology identification, shifting from traditional publication/patent-centric approaches; (2) A hybrid intelligence pipeline that synergistically combines LLM-based semantic extraction with reproducible vectorization-clustering workflows, enabling dual-verification through both computational similarity and LLM-based relevance scoring; (3) A five-dimensional maturity assessment system that stratifies identified technologies into mature, emerging, and potential frontiers, providing actionable intelligence for decision-makers.

The remainder of this paper is organized as follows. In Section II, we review the related work on frontier technology conceptualization and identification methods, including recent advances in AI-driven approaches. In Section III, we present the proposed hybrid intelligence framework, covering technical phrase extraction, semantic vectorization, cross-

source association analysis, and the multidimensional assessment system. In Section IV, we describe the empirical study using DARPA unmanned systems projects and present the results. In Section V, we conclude the paper and discuss future research directions.

II. Related Work

This section reviews the literature in three areas: the conceptual evolution of frontier technology, methods for its identification, and recent AI-driven advances that underpin our approach.

A. Evolution of Frontier Technology Conceptualization

The concept of frontier technology lacks a universal definition, with different scholars offering varying interpretations. In 1965, Price [4] introduced the concept of research fronts, defining them as research areas represented by a set of recently published and frequently cited papers, emphasizing novelty and academic attention. In the 21st century, the focus shifted to technological characteristics and industrial value: Chen [5] noted that an abrupt increase in technology occurrence frequency signals the emergence of frontier technologies; Cozzens et al. [6] argued that frontier technologies should feature rapid growth, novelty, untapped market potential, and strong technological foundations; Rotolo et al. [7] identified five key attributes of frontier technologies: novelty, rapid growth, coherence, potential impact, and uncertainty/ambiguity.

In policy and governance contexts, international organizations generally view frontier technologies as rapidly developing technologies with high uncertainty, capable of generating significant socioeconomic impacts. The United Nations Conference on Trade and Development (UNCTAD) Technology and Innovation Report [8] defines frontier technologies as new and rapidly developing technologies, emphasizing their economic potential and technological gaps. The World Intellectual Property Organization (WIPO) highlights that frontier technologies lie at the intersection of scientific breakthroughs and real-world applications, while the Organisation for Economic Co-operation and Development (OECD) underscores the dual nature of emerging technologies and the need for forward-thinking policies and risk governance [9].

B. Frontier Technology Identification Methods

As the concept of frontier technology has evolved, so too have methods for identifying them, transitioning from single-source to multi-source, static to dynamic, and manual to automated approaches. Early methods mainly relied on bibliometric techniques, such as citation networks, co-word analysis, and keyword burst detection to reveal research frontiers. However, these methods often lag behind actual technological development. Subsequently, patent data was introduced to track technological innovations, but publications and patents have inherent limitations: publications highlight scientific novelty but may lack industrial relevance, while patents reflect applied innovations with limited coverage.

Recently, multi-source data fusion has gained traction, combining various data sources to overcome the shortcomings of single-source methods. Liu et al.[10] proposed integrating publications, patents, startup data, and public opinions to create a multidimensional indicator system for identifying disruptive technologies, proving that multi-source fusion outperforms single-source approaches. Munari et al. [11] introduced a research project/funding-publication-patent linkage method, demonstrating how public funding influences technology output and providing a framework for earlier frontier identification using project data.

C. AI-Driven and Multi-Source Identification Methods

Recent advances in Natural Language Processing (NLP) and deep learning have introduced powerful tools for technology identification. In the domain of semantic embedding, Mikolov et al. [12] proposed Word2Vec, which learns distributed word representations that capture syntactic and semantic relationships. More recently, Devlin et al. [10] introduced Bidirectional Encoder Representations from Transformers (BERT), enabling contextualized embeddings that significantly improve downstream NLP tasks. Beltagy et al. [14] further developed SciBERT, a variant pre-trained on scientific text, demonstrating superior performance on scientific document processing tasks.

In the area of LLM-assisted topic extraction, Grootendorst [15] proposed BERTopic, a topic modeling technique that leverages transformer-based embeddings and class-based Term Frequency-Inverse Document Frequency (TF-IDF) to create dense topic clusters. More recently, Xu et al. [16] demonstrated that LLMs can be effectively used for automated information extraction from scientific literature, outperforming traditional rule-based and statistical methods. For technology forecasting and emerging technology detection, Huang et al. [17] developed a framework for technology opportunity analysis using multi-source patent data and semantic analysis. These prior works provide methodological foundations upon which our hybrid intelligence framework builds, while our approach distinctively integrates project-level data as early signals and employs a dual-verification mechanism combining computational similarity with LLM-based relevance scoring.

III. Methodology

This section presents the proposed hybrid intelligence framework, covering its overall design, the technical phrase extraction and vectorization pipeline, the project-output association analysis, and the multidimensional assessment system for frontier stratification.

A. Overall Framework Design

The proposed project-linkage paradigm for frontier technology identification, as illustrated in Figure 1, is structured into three key stages. First, Technical Topic Extraction & Clustering, involves aggregating project descriptions, publications, and patents from various data sources, followed by a multi-stream processing approach. This utilizes techniques, such as K-means clustering and LLMs for intelligent summarization and topic extraction,

resulting in standardized project, publication, and patent topics. Second, cross-Source Semantic Association, integrates the extracted topics through Word2Vec vectorization and cosine similarity, establishing semantic connections across data sources. Subjective association scoring based on LLMs is employed for dual-verification of the similarity calculations, with the results organized and filtered through time-series analysis. The final stage, Multidimensional Assessment & Stratification, evaluates the identified frontier technologies using a five-dimensional indicator system, comprising Response Scale, Semantic Similarity, Temporal Directionality, Evolution Continuity, and Cross-domain Integration. This stage leverages LLMs to provide a comprehensive maturity assessment, categorizing technologies as mature, emerging, or potential frontiers.

B. Technical Phrase Extraction with Large Language Models

Traditional keyword extraction methods based on TF-IDF or TextRank often fail to capture domain-specific technical semantics. This study employs the DeepSeek V3 Large Language Model for technical phrase extraction. The model processes project descriptions and generates concise technical phrases that represent core technological concepts. Specifically, the prompt instructs the model as follows: "*Given the following project description, extract up to 10 concise technical phrases that represent the core technological concepts. Each phrase should be 2–5 words and capture a distinct technical aspect*." This approach leverages the semantic understanding capabilities of LLMs to identify meaningful technical terms that may not be captured by frequency-based methods.

C. Semantic Vectorization and Topic Clustering

Following phrase extraction, we employ Word2Vec for semantic vectorization, converting technical phrases into 300-dimensional vector representations. The K-means clustering algorithm is then applied to group semantically related phrases into distinct technical topics. The optimal number of clusters is determined through a grid search over K = 5 to 30, selecting the value that maximizes the silhouette coefficient, followed by domain expert validation. This dual-layer approach—LLM extraction followed by traditional vectorization and clustering—ensures both semantic depth and computational reproducibility.

D. Project-Output Association Analysis

To establish associations between project topics and research outputs (publications and patents), we calculate cosine similarity between vectorized project topics and output topics. A cosine similarity threshold of 0.15 is applied to filter meaningful associations. Additionally, the LLM generates relevance scores for each project-output pair on a scale from 0 to 100, with a threshold of 60 for qualifying associations. The final association score is computed as: $Final\_Score = \alpha \times CosSim\_norm + (1-\alpha) \times LLM\_Score\_norm$, *where* $\alpha=0.5$. High-scoring pairs are identified as meaningful associations, forming project-output networks.

E. Multidimensional Assessment System

To evaluate the frontier attributes of identified technologies, we propose a five-dimensional indicator system, as detailed in Table I. The system comprises: Response Scale (A), measuring the volume and proportion of projects yielding relevant outputs; Semantic Similarity (S), quantifying the average topical alignment between projects and their outputs; Temporal Directionality (T), assessing the chronological distribution (prior, concurrent, or posterior) of outputs relative to project initiation; Evolution Continuity (C), tracking the stability of response levels across successive time windows; and Cross-Domain Integration (D), evaluating the coupling degree of topics across heterogeneous domains. By synthesizing these metrics, the paradigm enables a stratified classification of technologies into mature, emerging, and potential frontiers. The framework was implemented in Python 3.10 using Gensim for Word2Vec training, scikit-learn for K-means clustering, and the DeepSeek API for LLM-based extraction. The code will be released as open source upon publication.

IV. EMPIRICAL STUDY

This section describes the empirical validation of the proposed framework using DARPA-funded unmanned systems programs, presents the project-output association analysis results for two temporal phases, and synthesizes the frontier technology directions identified through the multidimensional assessment system.

A. Data and Experimental Setup

This study utilizes data from research projects in unmanned systems technologies funded by the U.S. DARPA. DARPA, a major funder of frontier technology Research and Development (R&D), supports projects in various cutting-edge fields, including advanced unmanned aerial vehicles, unmanned ground vehicles, and robotic swarms. The dataset comprises a collection of DARPA-initiated projects from 2009 to 2025, including project titles, abstracts, technical descriptions, and publicly available outputs (such as publications and patents). These data were primarily sourced from project completion reports, literature databases, and patent databases, ensuring a comprehensive representation of each project lifecycle.

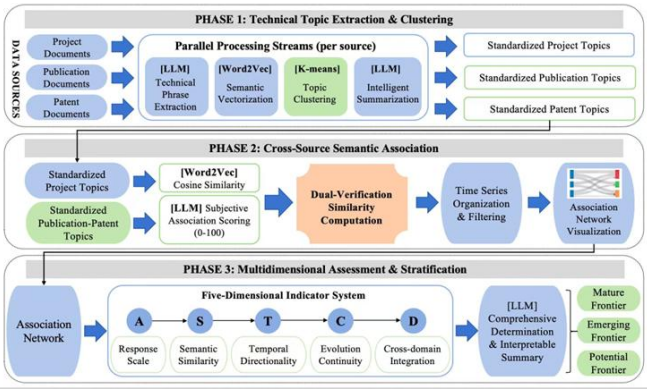

Figure 1. Schematic diagram of the hybrid intelligence framework for project-linkage-based frontier technology identification.

To examine the associations between DARPA-funded projects and their corresponding research outputs, we divide the dataset into two temporal phases: 2014–2019 and 2020–2025. This segmentation allows for a comparative analysis of project-output interactions across different periods of technological development. For technical phrase extraction, the DeepSeek V3 model uses a prompt designed to extract up to 10 concise technical phrases per project. The Word2Vec model utilizes pre-trained 300-dimensional vectors, while the K-means algorithm adjusts the number of clusters based on the silhouette coefficient. The LLM provides relevance scores for project-output topic pairs on a scale from 0 to 100, which are normalized and averaged with cosine similarity results.

B. Project-Output Association Analysis

Project-Publication Association. During 2014–2019, 60 project topics and 28 publication topics yielded 166 high-similarity associations with an average similarity of 0.31 (Figure 2(left)), achieving an 80% response rate (48 of 60 projects). In 2020–2025, 60 project topics and 30 publication topics produced 354 associations with a lower average similarity of 0.19 (Figure 2(right)), while the response rate rose sharply to 98.3% (59 of 60 projects). This combination of increased association volume and decreased similarity reflects a semantic diversification of research trajectories rather than weakening ties.

The topological shift between the two periods is significant. The 2014–2019 network exhibits a sparse bipartite structure with isolated clusters, suggesting that project-to-publication spillover was confined to specific technical niches. By contrast, the 2020–2025 network displays markedly higher density and multifocal hubs, with "many-to-many" associations indicating that individual DARPA project topics now catalyze research across multiple academic domains. This structural evolution confirms a transition from a linear technology transfer model to a complex, integrated collaborative network, where DARPA projects act as gravitational centers for increasingly diverse and interdisciplinary research.

Project-Patent Association. During 2014–2019, 60 project topics and 29 patent topics produced 64 high-similarity associations with an average similarity of 0.31 (Figure 3(left)), of which 37 pairs involved post-project patents and 27 leveraged pre-existing patents. The response rate reached 60% (36 projects). In 2020–2025, with a stable topic scale (60 project and 30 patent topics), the average similarity rose significantly to 0.53 (Figure 3(right)), with 74 associations identified—51 involving patents granted during the project lifecycle and 23 utilizing existing intellectual property. However, the response rate moderated to 50%. This divergent pattern—rising similarity but declining response rate—suggests a strategic shift from exploring uncharted technological territories toward high-fidelity application and integration of existing innovations.

TABLE I. ASSESSMENT DIMENSIONS FOR PROJECT-LINKAGE-BASED FRONTIER TECHNOLOGY IDENTIFICATION

Dimension	Indicator	Meaning of Discriminatory Information
Response Scale (*A*)	Number of high-similarity associations between project topics and output topics	Characterizes response scale and coverage of project topics in outputs (diffusion/attention)
Semantic Similarity (*S*)	Average semantic similarity across all valid project topic–output topic associations	Measures semantic alignment between project technical intent and output technical content
Temporal Directionality (*T*)	Proportions of prior/concurrent/posterior outputs relative to project initiation	Determines whether projects drive output production (posterior dominance) or absorb/integrate existing technologies (prior dominance)
Evolution Continuity (*C*)	Sustained response levels across time windows (strong/moderate/weak)	Characterizes whether topics form cross-period continuous evolution chains (short-term hotspot vs. long-term evolution)
Cross-Domain Integration (*D*)	Coupling degree of topics across different domains	Reflects cross-disciplinary and cross-technical characteristics common in frontier technologies

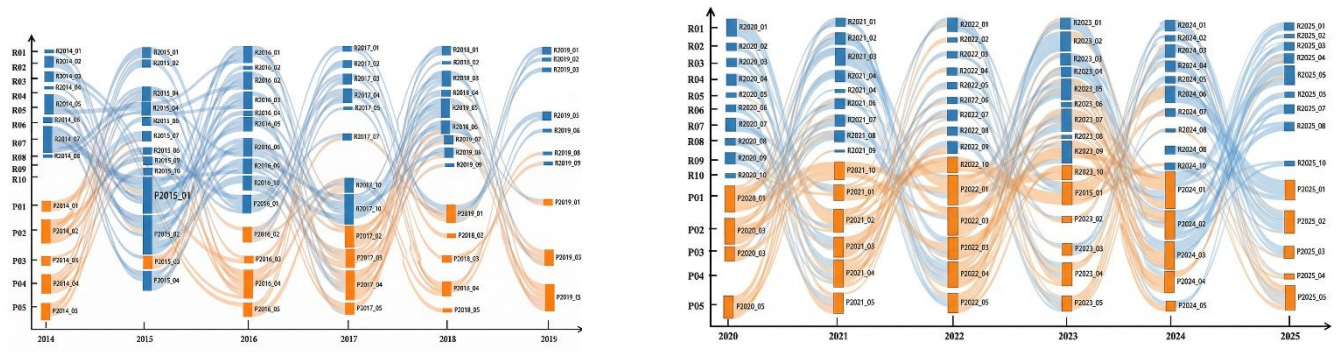

Figure 2. Comparative analysis of high-similarity association networks between DARPA project topics (R) and academic publication topics (P) for the periods 2014–2019 (left) and 2020–2025 (right).

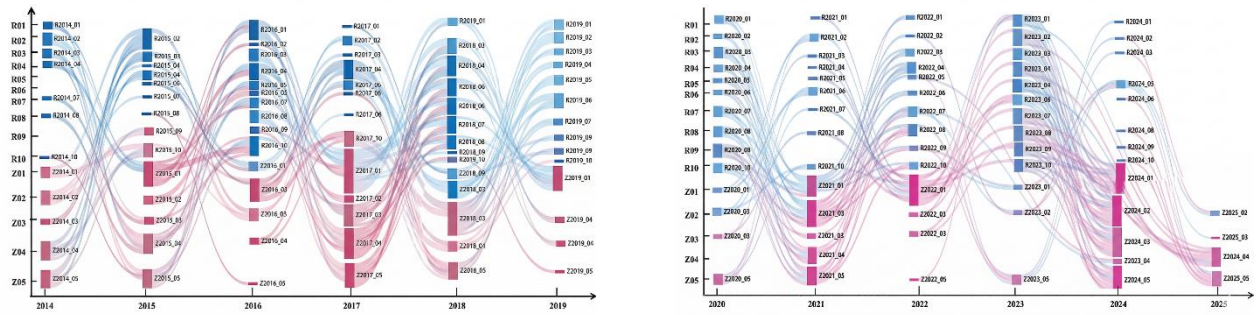

Figure 3. High-similarity association networks between DARPA project topics (R) and patent topics (Z) for 2014–2019 (left) and 2020–2025 (right).

C. Comparative Analysis and Frontier Identification Empowerment

The divergent trajectories of Project-Publication (R-P) and Project-Patent (R-Z) associations provide a high-fidelity lens for characterizing the unmanned systems innovation ecosystem.

Decoupling Research Breadth and Technical Depth: The R-P network exhibits centrifugal expansion (similarity: 0.31→0.19), confirming that DARPA projects catalyze broad, interdisciplinary academic exploration. In contrast, the R-Z network shows centripetal convergence (similarity: 0.31→0.53), suggesting a strategic transition toward high-fidelity application of existing innovations.

Calibrating Technology Readiness Level (TRL) Transitions: The widening "Similarity Gap"—declining R-P similarity coupled with rising R-Z similarity in 2020–2025—marks a critical phase where technologies transition from basic research to system-level integration. The identified "Mature Frontiers" represent directions where both R-P and R-Z associations have stabilized into high-consistency evolution chains.

Strategic Vacuum Detection: The near-saturated publication response (98.3%) coupled with moderated patent response (50%) highlights areas where academic consensus exists but the intellectual property landscape remains fluid, providing precise early-stage breakthrough signals for decision-makers.

Notably, DARPA's strategic priorities shifted around 2019–2020 from platform-centric programs toward multi-domain autonomy and AI-enabled decision-making, partially explaining the observed network densification. While these results are ecosystem-specific, the methodology is domain-agnostic and applicable to other funding agencies (e.g., the National Science Foundation (NSF), European Union (EU) Horizon programs) with appropriate data adaptation.

D. Frontier Direction Identification in Unmanned Systems Technologies

Integrating project-publication and project-patent semantic association results, we conducted frontier attribute assessment of DARPA-deployed unmanned systems project topics based on the five-dimensional indicator system, employing within-sample quantiles (Q25/Q50/Q75) as relative discrimination benchmarks for stratified technology topic identification. Final results, combined with domain expert validation and refinement, yielded a frontier technology direction inventory for unmanned systems. The expert validation involved three domain specialists who independently reviewed and scored the identified frontier directions; a majority voting mechanism was employed to resolve disagreements.

Mature frontier directions primarily exhibit high output response frequency, high semantic consistency, and good evolution continuity. Representative directions include multi-modal sensor fusion chips, autonomous navigation algorithms, and distributed resilient computing platforms.

These technologies have been validated across multiple DARPA projects and formed stable outputs at both publication and patent levels, demonstrating relatively high maturity.

Emerging frontier directions embody strong innovation and cross-domain integration characteristics. While output scale remains limited, growth trends are evident, such as federated learning algorithms, neuromorphic computing chips, and digital twin validation platforms. These directions typically exhibit high semantic consistency and clear project-posterior output features, reflecting rapid exploration phases.

Potential frontier directions are predominantly in early nascent stages, with currently limited output production but clear technological orientation, such as complex terrain autonomous navigation algorithms and platform-independent development middleware. These directions have not yet formed stable evolution chains but possess clear application requirements and subsequent development potential, warranting continuous tracking.

V. CONCLUSION AND FUTURE WORK

Focusing on DARPA-funded project portfolios, this study proposes and validates a project-linkage-based frontier technology identification methodology. We innovatively incorporate project texts into frontier detection perspectives, fusing Large Language Models with bibliometric techniques to realize full-chain association analysis from project requirements to publication and patent outputs. In the empirical study of DARPA unmanned systems projects, this methodology successfully mapped project-output interaction networks, revealed technology diffusion pathways, and consequently identified multiple strategically significant frontier directions.

Research demonstrates that: research project data contains abundant prospective technological intelligence; through intelligent extraction and semantic association, this effectively compensates for deficiencies in single-source publication or patent analysis, significantly enhancing sensitivity to emerging technologies; the integration of multi-source data fusion with LLM technology provides powerful tools for frontier identification, not only automatically generating technical topics and descriptions but also discovering novel cross-domain associations through deep semantic matching, bringing new perspectives to complex technology evolution analysis; introducing multidimensional indicator systems for frontier attribute assessment of technology topics is necessary, ensuring identification results possess greater explanatory power and credibility, facilitating decision-maker comprehension and application.

Naturally, this research has certain limitations. First, regarding data acquisition, heavy reliance on DARPA public materials means some sensitive projects and unpublicized outputs were not included, potentially causing frontier identification omissions. Second, while LLM-generated summaries and similarity scores enhance analysis quality, they also introduce uncertainty and computational overhead, necessitating refined manual verification mechanisms to ensure result reliability. Third, the current thresholds (cosine similarity cutoff of 0.15, LLM relevance score cutoff of 60, fusion weight $\alpha = 0.5$) were determined through manual tuning and expert judgment, rather than automated optimization. Finally, this study's stratification discrimination thresholds and rules may require domain-specific adjustments and are not universally applicable models.

Future work will pursue several directions: systematic comparisons against baseline methods, including pure bibliometric approaches and keyword-based detection methods; ablation studies to quantify the individual contributions of each framework component; sensitivity and robustness analysis across varying thresholds and parameters; computational cost and scalability assessment for larger datasets; automated parameter optimization through grid search, Bayesian optimization, or cross-validation strategies; and cross-domain validation beyond DARPA unmanned systems to assess methodological generalizability across different technological ecosystems and funding agencies.

REFERENCES

[1] W. Glänzel and A. Schubert, "Analysing scientific networks through co-authorship," in *Handbook of quantitative science and technology research*, H. F. Moed, W. Glänzel, and U. Schmoch, Eds. Dordrecht: Kluwer Academic, 2004, pp. 257–276.

[2] A. L. Porter and S. W. Cunningham, *Tech mining: exploiting new technologies for competitive advantage*. Hoboken, NJ: Wiley, 2005.

[3] Y. Guo, J. Huang, H. Chen, and J. Xu, "Identifying technology evolution pathways using topic modeling," J. Informetr., vol. 15, no. 4, Art. no. 101228, 2021.

[4] D. J. de S. Price, "Networks of scientific papers," Science, vol. 149, no. 3683, pp. 510–515, 1965.

[5] C. Chen, "CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature," J. Am. Soc. Inf. Sci. Technol., vol. 57, no. 3, pp. 359–377, 2006.

[6] S. Cozzens, S. Gatchair, J. Kang, K.-S. Kim, H. J. Lee, G. Ordonez, and A. Porter, "Emerging technologies: quantitative identification and measurement," Technol. Anal. Strateg. Manag., vol. 22, no. 3, pp. 361–376, 2010.

[7] D. Rotolo, D. Hicks, and B. R. Martin, "What is an emerging technology?" Res. Policy, vol. 44, no. 10, pp. 1827–1843, 2015.

[8] United Nations Conference on Trade and Development (UNCTAD), *Technology and Innovation Report 2023*. United Nations, 2023.

[9] Organisation for Economic Co-operation and Development (OECD), *Framework for Anticipatory Governance of Emerging Technologies*. OECD Publishing, 2024.

[10] X. Liu, X. Wang, L. Lyu, and Y. Wang, "Identifying disruptive technologies by integrating multi-source data," Scientometrics, vol. 127, no. 9, pp. 5325–5351, 2022.

[11] F. Munari, E. Leonardelli, S. Menini, H. Morais Righi, M. Sobrero, S. Tonelli, and L. Toschi, "Public research funding and science-based innovation: an analysis of ERC research grants, publications and patents," Res. Eval., 2024, doi:10.1093/reseval/rvae012.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in Proc. Int. Conf. Learn. Represent. (ICLR), 2013.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[14] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in Proc. EMNLP-IJCNLP, 2019, pp. 3615–3620.

[15] M. Grootendorst, "BERTopic: neural topic modeling with a class-based TF-IDF procedure," arXiv:2203.05794 [cs.LG], 2022.

[16] S. Xu, L. Shang, and P. Hao, "Automated information extraction from scientific literature using large language models," Scientometrics, vol. 129, no. 4, pp. 2187–2208, 2024.

[17 Y. Huang, L. Zhang, and J. Zhu, "Technology opportunity analysis using multi-source patent data and semantic analysis," Technol. Forecast. Soc. Change, vol. 167, Art. no. 120680, 2021.

Large Scale Offline Data Handling: An Event Streaming Based Kappa Architecture

Quan Zhou, Pradeep Akkinepally, Monica Dhanaraj, Dyutimoy Sarkar, Muthaiyan Thandapani
eBay Inc.
San Jose, USA
e-mail: {qnzhou, pakkinepally, mdhanaraj, dysarkar, muthandapani}@ebay.com

***Abstract*—Processing massive volumes of historical offline data through complex, heavy-weight service applications presents a significant engineering challenge. Traditional batch processing methods often require refactoring sophisticated production logic, leading to code duplication and maintenance overhead. This paper proposes a methodology for large-scale offline data processing utilizing a Kappa Architecture. By treating offline data warehouses (DW) as streaming sources and wrapping complex service applications as asynchronous consumers, we enable high-throughput processing without rewriting core application logic. We compare this approach against Spark-native and micro-batch paradigms, and show that it better preserves logical parity while maintaining engineering velocity. We also present a multi-region intelligent job dispatcher that improves availability and throughput.**

Keywords-kappa architecture; offline processing; event streaming; distributed systems; big data.

I. Introduction

In the modern data landscape, organizations frequently need to re-process vast amounts of offline data stored in Data Warehouses (DW). Common use cases include:

- **Historical Simulation:** Validating new features or business logic against historical snapshotted events.
- **Data Backfill:** Re-populating missing data on DW after a schema change or data loss event [1].
- **Batch Processing:** Periodic heavy-lift computing tasks that rely on processing large scale of offline data against complex service applications [2].

The core problem arises when the processing logic is not a simple SQL transformation but resides within a **heavy-weight service application**. These applications often contain intricate dependencies, local state management, and external dependency integrations that are difficult to port to offline environment such as Apache Spark or Hadoop. Engineers are faced with the dilemma of either rewriting the application or finding a way to bridge the gap between offline data and complex business logic execution.

The remainder of the paper is organized as follows: Section II evaluates architectural options. Section III details the proposed system architecture. Section IV presents performance results. Section V describes the cross-region extensibility, and Section VI concludes the work.

II. Architectural Options for Offline Processing

As an overview of the paper, we evaluated three different architectural paradigms, in order to address the challenge of processing large-scale offline data against complex service applications. Which are widely adopted and used across different domains [3].

A. Option I: Spark Native Application (Compute-to-Data Approach)

This approach involves refactoring the core logic of the service application (e.g., a Spring Boot service application) into a computing intensive Spark-native application. By refactoring the code and distributing Spark libraries only across the Hadoop cluster, we are capable of moving the computation directly to where the data resides in the Offline Data Warehouse (DW).

Mechanisms & Performance Metrics: Our internal benchmarking of this approach demonstrated exceptional raw performance for specific use cases.

- **High-Velocity Evaluation:** We observed a benchmark of evaluating an atomic web service Application Programming Interface (API) against **22 million historical records in** < 10 **seconds**.
- **Data Locality:** By executing service API libraries directly on top of Spark runtime, we eliminated the network I/O overhead typically associated with moving data to a service.
- **Native Resilience:** The architecture leverages the native failover and retry mechanisms of the Spark/MapReduce framework, ensuring robust handling of executor failures.

The "Dual Codebase" Constraint: Despite the performance wins for simpler web APIs, this approach was deemed unviable for the full-fledged complex service application due to **Architectural Divergence**.

- **Dependency Stripping:** Complex service application often rely on heavy I/O operations (database (DB) lookups, API integrations) which are anti-patterns as Spark-native application. Engineers require efforts to strip out these dependencies or mock them extensively, effectively rewriting the service application with a "dual" codebase.
- **High Maintenance:** The dual codebases from offline & real-time, results in high maintenance effort and long-term code sustainability issues.
- **Logic Parity:** Over time, as new features are added to the real-time application, the offline codebase inevitably falls behind. Guaranteeing 100% logical parity becomes an operational impossibility without a continuous, expensive synchronization effort between the two codebases.

B. Option II: Micro-Batch & Spark JDBC

This method utilizes a Spark Java Database Connectivity (JDBC) session [4] to read data in Hadoop partitions and execute synchronous calls to the service application via HTTP or gRPC connections.

- **Mechanism:** Spark executors iterate through Hadoop partitions, move to batch application for further data processing

and trigger synchronous requests to the service APIs for each record.

- **Pros:** Guarantees high logical parity as it uses the same service endpoints as production real-time.
- **Cons: Throughput Bottlenecks & Reliability:** We encountered throughput bottlenecks from both Spark JDBC sessions and synchronous API calls, which can overwhelm service applications and lead to timeouts. Moreover, micro-batches become a **Single Point of Failure** because data fetching and dispatch run in a single component.

C. Option III: Kappa Architecture

This approach treats the offline processing problem as a streaming problem [5][6]. The architecture decouples data movement from execution using asynchronous event queues.

- **Key components of Kappa Design:**
  1) **Ingress Kafka Queue:** A Spark-native application reads the offline DW data, and streams it into an **Ingress Kafka Queue** within Spark Map-Reduce runtime;
  2) **Process by Async Consumer** The heavy-weight service application is packaged as a scalable **Consumer Application**, which is capable of processing events in an asynchronous manner;
  3) **Egress Kafka Queue:** Results are published to an Egress Kafka Queue and persisted back to DW storage using lightweight processors such as Apache Flink.

Pros:

- **Total Logical Parity:** Because the consumer is a consumer wrapper *exact* of service executable libraries, there is zero risk of "logic drift." The same code logic will be unified between real-time & offline environment.
- **Elastic Scalability:** The execution tier is decoupled from the data storage. We can auto-scale the consumer fleet (e.g., via Kubernetes) from zero to hundreds of pods to match the simulation workload, optimizing compute costs.
- **Dependency Isolation:** Complex site dependencies and local caching mechanisms remain intact within the application container, removing the need to mock external services.

Cons:

- **Operational Complexity:** Asynchronous architectures are inherently more difficult to debug than synchronous batch jobs. Triaging issues such as "data loss" requires sophisticated offset tracking across Ingress and Egress queues.
- **Serialization Overhead:** There is a computational cost to transforming offline data (e.g., Spark data frame) into stream events (Avro/Protobuf) at the Ingress layer. Strict schema management is required to ensure the DataFrame matches the event consumer data schema.

D. Comparison Summary

We selected the Kappa Architecture based on its unique ability to balance throughput with strict logical parity, as summarized in Table I.

TABLE I. COMPARISON OF ARCHITECTURE OPTIONS

Feature	Spark Native	Micro-Batch	Kappa
Logic Parity	Low	High	**Total**
Throughput	Very High	Low	**High**
Maintenance	High	Moderate	**Minimal**
Scaling	Map-Reduce	API/JDBC bottleneck	**Elastic**

III. SYSTEM ARCHITECTURE

The system is composed of five primary components designed to decouple data movement from execution logic. Figure 1 illustrates the end-to-end data flow across these components.

- **Workflow Orchestration:** The high-level workflow orchestrator (e.g., Apache Airflow) managing the end-to-end (E2E) lifecycle. It triggers the data pipelines and tracks workflow run status.
- **Data Movement Pipeline:** A distributed Spark process responsible for fetching Point-in-Time (PiT) snapshots from the Offline DW and producing them as events into the **Ingress Kafka Queue**.
- **Service as Consumer:** The production service application packaged as a Kafka consumer. It processes each transaction using production-grade logic and publishes results to the **Egress Kafka Queue**.
- **Data Persistence:** A lightweight stream processor (e.g., Flink consumer [7]) that consumes egress events and sinks them into the Hadoop Distributed File System (HDFS) for long-term storage.
- **Analytics Layer:** SQL-based interface (Hive/Spark) allowing users to query the landed HDFS tables to calculate metrics or perform varied offline analysis.

A number of other challenges we addressed by adopting Kappa Architecture:

- **Backpressure Handling:** The Spark-based ingress pipeline is capable of producing events at a rate significantly higher than the service application can process. Without intervention, this leads to consumer lag accumulation and potential resource exhaustion. To mitigate this, we tuned the Kafka consumer configuration to strictly limit pre-fetching. Specifically, we reduced `max.poll.records` to align with the service's p99 latency, ensuring that the consumer never fetches more records than it can process within a session timeout window. Furthermore, we implemented an application-level rate limiter that dynamically pauses consumption if local thread pools become saturated [8].
- **Schema Transformation:** Our Ingress Pipeline performs a mandatory schema validation. Before publishing, the Spark DataFrame—often loosely typed—is mapped to a rigorous Avro schema governed by a central Schema Registry. This step handles type coercion (e.g., casting Hive timestamps to Avro longs) and null-safety checks. By enforcing strict Avro serialization at the ingress, we guarantee that the complex service consumer is protected from malformed data that could cause exceptions during execution [9].

These components together form a unified platform that cleanly separates data movement from execution logic, providing the foundation upon which we evaluate performance

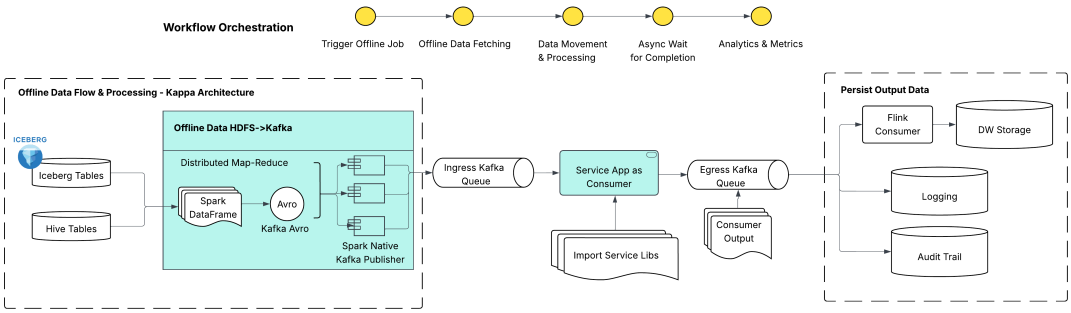

Figure 1. End-to-End Kappa Architecture Data Flow

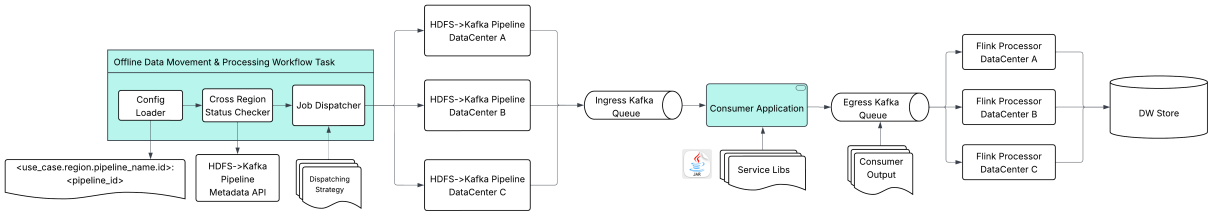

Figure 2. Cross-Region Intelligent Job Dispatching

characteristics and cross-region extensions in the following sections.

IV. Performance Benchmarking

For the above end-to-end (E2E) Kappa Architecture, we have done performance benchmarking with varied size of historical data as input. We observed high throughput & low E2E latency with horizontal scaling of Kafka Queue & machines allocated to consumer pool.

A. Experimental Setup

To evaluate the efficacy and scalability of the proposed Kappa Architecture, we conducted a series of performance benchmarks in a controlled production-like environment. The experimental cluster was configured to mirror the constraints of a high-throughput, multi-tenant system.

Infrastructure Configuration: The messaging backbone was hosted on a dedicated Kafka cluster. To ensure high parallelism, both the Ingress and Egress Kafka topics were configured with **100 partitions**. This partitioning strategy was chosen to maximize consumer concurrency.

Compute Resources: The consumer application fleet was deployed on a Kubernetes cluster with resource quotas strictly enforced to simulate real-world constraints:

- **Baseline Capacity:** The fleet was initialized with a minimum of 24 pods per data center.
- **Elastic Scaling:** Horizontal Pod Autoscaler (HPA) was configured to scale the fleet up to a maximum of 48 pods per region based on CPU utilization and consumer lag metrics.
- **Pod specs:** Each consumer pod was provisioned with **2 vCPU** and **8 GB of memory**. This memory profile was specifically chosen to accommodate the local caching and high computing intensity of the heavy-weight service application.

B. Benchmarking Metrics

- **Scalability:** Capable of processing > 10 million offline data records in a single job run.
- **E2E Latency:** Achieved a processing Service Level Agreement (SLA) of < 15 minutes for the core execution phase.

TABLE II. Performance Benchmarking Results

Data Volume	Job Duration	Throughput
1,000,000 (1M)	7 min 03 sec	2400 msgs/sec
3,000,000 (3M)	12 min 45 sec	9,000 msgs/sec
5,000,000 (5M)	15 min 05 sec	10,900 msgs/sec
10,000,000 (10M)	18 min 05 sec	11,900 msg/sec

Analysis of Results: The benchmarking metrics (Table II) demonstrate a non-linear relationship between data volume and job processing time, indicating high elasticity. This efficiency is driven by three architectural factors. First, the **Spark-native Kafka Publisher** runs within the Map-Reduce framework, allowing publishing throughput to scale horizontally with the number of Spark executors. Second, **Consumer Auto-Scaling** leverages Kubernetes HPA to automatically provision additional consumer pods, effectively increasing the processing rate from 2,400 to 12,000 msgs/sec. Finally, **Elastic Capacity** is managed via a strict Time-To-Live (TTL) policy on ingress topics, ensuring minimal disk footprint even with high-velocity streams.

C. Overall System Availability

- **Job Availability:** $> 99\%$ success rate.
- **Data Integrity:** End-to-end data loss was kept below 1%, primarily due to transient connection timeouts when publishing events from the ingress pipeline to the Kafka queues under large-scale traffic volume. Once events are

durably written to Kafka, the streaming pipeline operates with *at-least-once* delivery guarantees and did not introduce additional loss downstream.

D. Resource Efficiency

Beyond raw speed, the Kappa architecture demonstrated significant efficiency in resource utilization. By utilizing a blocking queue within the consumer application to buffer Kafka messages, we achieved a consistent CPU saturation of $\approx 50\%$ on consumer pods. And by leveraging Kubernetes Auto Scaling capability, this allows us to reduce the total provisioned cores by 50% during site idle time, while maintaining the same throughput and job SLA during runtime execution.

E. Cost vs. Convenience Trade-off

While the Kappa Architecture provides strong guarantees on logical parity and engineering velocity, it incurs additional infrastructure cost compared to a pure Spark-native implementation. Running a fleet of heavy-weight Spring Boot service containers to process batch data typically consumes more CPU and memory per record than executing equivalent logic inside a Spark job. In return, this architecture avoids maintaining a dual codebase, preserves complex dependencies and caching behavior, and enables teams to reuse production-grade logic.

V. Enhanced Feature: Cross-Region Extensibility

A key advantage of converting offline data into an event stream is the inherent decoupling of storage (Data Warehouse) from compute (Service Application). This decoupling allows the architecture to be extended easily to support **Cross-Data Center (Multi-Region) Orchestration** [10].

Figure 2 illustrates how the intelligent job dispatcher coordinates jobs across geographically distributed data centers.

By abstracting the physical location of the execution layer, the system achieves three critical operational enhancements:

A. Platform Availability & Resilience

In a traditional batch model, job failure often requires a full restart in the same cluster. The Kappa model enables an **Active-Active** availability posture. If a specific data center experiences an outage or transient infrastructure instability, the orchestration layer can instantly reroute pending jobs to an alternative region. Since consumers in Region B operate on the same logic as Region A, the failover is transparent to the end-user.

B. Distributed Resource Utilization

To prevent resource fragmentation—where one region is overloaded while others sit idle—we implemented an **Intelligent Dispatcher**. This component queries the global infrastructure state before job submission, calculating the available "Headroom" for each region: $Headroom_r = Cap_{max} - (Jobs_{active} + Jobs_{queued})$. Jobs are dynamically routed to the region with the highest $Headroom_r$, preventing "bulk submission" bottlenecks and ensuring uniform cluster utilization.

C. Improving Job E2E Latency via Partitioning

For processing jobs requiring massive historical data (e.g., 30 days of historical traffic), executing sequentially in a single region is inefficient. The streaming nature of Kappa architecture allows us to implement a **Split-and-Conquer** strategy:

1) **Partitioning:** The master job is logically split into N child jobs (e.g., 3 jobs of 10 days each).
2) **Parallel Execution:** These child jobs are dispatched *simultaneously* to different data centers (e.g., Job A $\rightarrow$ Region 1, Job B $\rightarrow$ Region 2).
3) **Throughput Multiplication:** By dispatching traffic into ingress/egress Kafka queues allocated on multiple regions, the End-to-End (E2E) SLA is reduced linearly by the number of regions.

VI. Conclusion and Future Work

The Kappa Architecture offers a robust solution for bridging the gap between large scale offline data and complex & heavy-weight service application. By decoupling data movement from execution, the system can achieve high-throughput processing and guaranteed logical parity without incurring the high maintenance costs between offline and real-time environment. In practice, we found that strict schema validation and back pressure aware consumers are essential to maintaining stable high-throughput processing at scale. As future work, we plan to strengthen observability and cost-efficiency and to extend the cross-region orchestration layer to support a broader set of offline simulation workloads.

References

[1] M. Kleppmann, *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems*. O'Reilly Media, Inc., 2017.

[2] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters", in *OSDI*, vol. 51, 2004, pp. 137–150.

[3] M. Kiran, P. Murphy, I. Monga, C. Jonnalagadda, and B. Sriprasad, "Lambda architecture for cost-effective batch and speed layer operations", in *2015 IEEE International Conference on Big Data (Big Data)*, IEEE, 2015, pp. 2785–2792.

[4] M. Armbrust et al., "Spark sql: Relational data processing in spark", in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 2015, pp. 1383–1394.

[5] J. Kreps, "Questioning the lambda architecture", *O'Reilly Radar*, 2014.

[6] J. Kreps, N. Narkhede, and J. Rao, "Kafka: A distributed messaging system for log processing", in *Proceedings of the NetDB*, vol. 11, 2011, pp. 1–7.

[7] P. Carbone et al., "Apache flink: Stream and batch processing in a single engine", *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, vol. 36, no. 4, 2015.

[8] M. Welsh, D. Culler, and E. Brewer, "Seda: An architecture for well-conditioned, scalable internet services", in *Proceedings of the Eighteenth ACM Symposium on Operating Systems Principles*, 2001, pp. 230–243.

[9] Reactive Streams, "Reactive streams specification", 2015, Accessed: 2026-04-09. [Online]. Available: http://www.reactive-streams.org/

[10] L. Wang et al., "Streams: High performance and reliable geo-distributed stream processing", in *USENIX Annual Technical Conference*, 2021.

Internationalization and Thematic Diversity in Data Use Within Open Research Infrastructures

A Scientometric Analysis of U.S. Department of Energy User Facilities

Lu Dong, Ren Wei, Yizhan Li, and Zexia Li
National Science Library
Chinese Academy of Sciences
Beijing, China
e-mail: donglu@mail.las.ac.cn, weir@mail.las.ac.cn, liyz@mail.las.ac.cn, lizexia@mail.las.ac.cn (corresponding author)

***Abstract*—The paper addresses limited empirical evidence on how data generated by large-scale research infrastructures are used in data-intensive and artificial intelligence-driven scientific research. The study is relevant to the conference theme as it examines data use patterns within open research infrastructures, contributing to the understanding of data-intensive science and infrastructure-based research systems. The paper presents an empirical evaluation based on micro-level administrative records of research proposals from United States Department of Energy user facilities over the period from fiscal years 2015 to 2025. It analyzes international participation, country-level distribution, thematic diversity, and temporal structural dynamics using a set of indicators that reflect participation scale, distributional diversity, and structural concentration, alongside entropy-based measures derived from project titles. The results show that international participation remained high, with a temporary decline during the COVID-19 pandemic followed by a rapid recovery, while thematic diversity remained consistently high and topic concentration low throughout the period. Overall, open research infrastructures exhibit strong resilience and sustained multidomain diversity in data use despite external disruptions.**

Keywords-open science; research infrastructure data; scientific resource; internationalization; thematic diversity.

I. INTRODUCTION

The paradigm of scientific research is undergoing a profound transformation, evolving from the traditional linear "observation-hypothesis-verification" model to two new paradigms: the Fourth and Fifth Paradigms. The "Fourth Paradigm" establishes data-intensive scientific discovery as a foundational new research framework [1], marking a critical shift away from conventional research logic. Building on this data-driven transition, recent advances in Artificial Intelligence (AI) and machine learning have further spurred the emergence of the "Fifth Paradigm," also referred to as AI4Science, distinguished by algorithm-guided discovery processes and human-AI collaborative research workflows [2]. Together, these paradigms represent the core dynamics of contemporary scientific restructuring.

With the rise of data-intensive and AI-driven research paradigms, research infrastructures increasingly extend beyond experimental operations to provide structured data services. Curated datasets, interoperable archives, long-term stewardship, and remote access mechanisms transform facilities into integrated data platforms. This transition is particularly evident in the U.S. Department of Energy (DOE) user facility system, where data management and open sharing are embedded alongside experimental access, serving as a typical example of this infrastructure-to-data-platform transition.

The U.S. Department of Energy (DOE) defined major research infrastructures funded by the federal government that provide open and shared access to researchers from academia and industry as National User Facilities in 2012 [3]. As of Fiscal Year (FY) 2025, the DOE operated 28 user facilities, covering areas including Advanced Scientific Computing Research (ASCR), Basic Energy Sciences (BES), Biological and Environmental Research (BER), Fusion Energy Sciences (FES), High Energy Physics (HEP), and Nuclear Physics (NP). In 2015, the DOE initiated the construction of a user project/experiment database [4].

This study is based on administrative records of research proposals from DOE user facilities covering FY 2015 to FY 2025. (The U.S. federal fiscal year runs from October 1 of the previous calendar year to September 30 of the current year.) The dataset includes user participation, institutional affiliations, and project descriptions. The analysis focuses on a subset of facilities that provide open access and structured data services, selected based on the availability of project-level proposal records. The resulting sample includes ARM, DIII-D, FACET/FACET-II, CINT, Alcator C-Mod, and NSTX-U.

As research infrastructures evolve into data platforms, issues of openness, accessibility, and global reach of infrastructure-generated data have become increasingly important. While open data principles, such as the Findable, Accessible, Interoperable, and Reusable (FAIR) principles, have been widely promoted, empirical evidence on who uses infrastructure-generated data and how such data contribute to knowledge production remains limited.

This study addresses the following research question: how are open research infrastructure data utilized in terms of international participation and thematic diversity, and what structural patterns characterize their usage over time? To answer this question, micro-level proposal data are analyzed using scientometric indicators and entropy-based measures.

By shifting the analytical focus from research outputs to data usage structures, this paper contributes to understanding open data ecosystems and the role of large-scale research infrastructures as global data platforms.

The remainder of this paper is organized as follows. In Section 2, the analytical framework and measurement design are described. In Section 3, the empirical results are presented. In Section 4, the conclusions and future research directions are presented.

II. METHODOLOGY

The analysis focuses on the structural characteristics of data use within open research infrastructures, rather than downstream research outputs. Attention is given to patterns of data usage across countries and thematic domains. This study aims to reveal the underlying structural features of open data ecosystems [5]. Data use within open research infrastructures is conceptualized along three analytical dimensions: (1) Internationalization-the extent to which open research data are used globally; (2) Topic Diversity-the breadth of research domains supported by the data infrastructure; (3) Structural Dynamics-the stability or evolution of usage patterns over time. These dimensions together capture the global reach, thematic inclusiveness, and longitudinal transformation of open data infrastructures.

A. Internationalization Measurement

To quantify international participation in open data use, three complementary indicators are adopted [6].

1) Non-U.S. User Ratio

The Non-U.S. User Ratio measures the proportion of users affiliated with Non-U.S. institutions in year *t*. Higher values indicate stronger international participation and a broader global diffusion of DOE open data resources. It is defined as:

$$NonUSRatio_t = \frac{N_{NonUS,t}}{N_{Total,t}} \quad (1)$$

where $N_{NonUS,t}$ represents the number of users affiliated with Non-U.S. institutions in year *t*; $N_{Total,t}$ represents the total number of users accessing DOE open data resources in year *t*.

2) Country Shannon Entropy

The Country Shannon Entropy measures the diversity of country participation [7]. Higher values indicate more even distribution across countries, reflecting broader international engagement. It is defined as:

$$H_t = -\sum_{i=1}^{N} p_{i,t} \ln p_{i,t} \quad (2)$$

where $p_{i,t}$ is the share of users from country i in year t; N is the total number of countries represented. Entropy is calculated using natural logarithms and is not normalized.

3) Country Concentration

The Herfindahl-Hirschman Index (HHI) is employed to measure concentration. Higher values indicate stronger dominance by a small number of countries, while lower values suggest more distributed global use. Entropy and HHI are used jointly to provide a balanced view of diversity and concentration. HHI is defined as:

$$HHI_t = \sum_{i=1}^{N} p_{i,t}^{\,2} \quad (3)$$

where p_i represents the share of users from country *i* in year *t*.

B. Thematic Diversity Metrics

Thematic diversity is derived from the textual analysis of Project/Experiment Titles, which represent the research purposes supported by the data infrastructure.

1) Keyword Distribution

Project/Experiment Titles were preprocessed using a standardized pipeline, including lowercasing, removal of general and domain-specific stopwords, and lemmatization to normalize morphological variants. Duplicate keywords within the same title were removed to avoid artificial inflation of term frequency. To reduce noise from unstable and weakly informative terms, keywords with an annual frequency below three occurrences were excluded. This filtering follows common text-as-data practice, as low-frequency terms are typically weakly discriminative and increase sparsity and computational burden without substantially affecting topic representations [8]. This consideration is particularly relevant for project/experiment titles, which are short texts with limited contextual and co-occurrence information. Therefore, a minimum annual frequency threshold of three was adopted to retain stable thematic signals while suppressing idiosyncratic noise [9]. After filtering, the annual keyword set was constructed, and relative frequencies were computed for entropy and concentration measurements.

2) Topic Shannon Entropy

Topic Shannon Entropy is employed to measure thematic diversity. Higher entropy indicates greater thematic diversity in data-supported research.

$$H_t^{\text{topic}} = -\sum_{j=1}^{M} q_{j,t} \ln q_{j,t} \quad (4)$$

where $q_{j,t}$ is the share of keyword or topic j in year *t*; M is the total number of topic clusters.

C. Structural Change Index

To evaluate temporal dynamics, the Structural Change Index (SCI) is computed as follows:

$$SCI_t = \sum_{k} \left| p_{k,t} - p_{k,t-1} \right| \quad (5)$$

where $p_{k,t}$ represents the share of country or topic *k* in year *t*.

SCI measures the magnitude of year-to-year structural shifts. Values close to zero indicate stability, while larger values reflect significant structural transitions in data usage patterns [10].

III. RESULTS

This section presents the empirical results of the study, focusing on international participation patterns and thematic diversity in the use of open research infrastructure data, along with their structural evolution over time.

A. International Participation in DOE Data-Providing Research Infrastructures

Based on the full dataset covering FY 2015-FY 2025, 10,113 user records were analyzed to evaluate international participation in research infrastructures that provide open data services. Internationalization is evaluated using three complementary indicators: the Non-U.S. User Ratio, Country Shannon Entropy, and Country Concentration.

1) Non-U.S. Participation Trends

Non-U.S. participation remained substantial throughout the study period, despite moderate annual variation (TABLE I). The Non-U.S. user ratio varied between 0.392 in FY 2018 and 0.318 in FY 2020. Following a notable decline between FY 2019 and FY 2020, the ratio gradually recovered, reaching 0.376 in FY 2024 and 0.379 in FY 2025.

The structural contraction observed in FY 2020 temporally coincides with the global COVID-19 pandemic, suggesting that global mobility restrictions may have contributed to fluctuations in international participation. Nevertheless, the pronounced post-2020 recovery demonstrates that open data ecosystems possess strong structural resilience and adaptive capacity in response to external disruptions.

TABLE I. INTERNATIONAL PARTICIPATION OF OPEN RESEARCH INFRASTRUCTURE DATA USE (FY 2015-FY 2025)

Year	International Participation		Country Structure			
	Non-U.S. Users	Non-U.S. Ratio	Countries Total	Country Entropy	HHI	SCI Country
2015	289	0.378	33	1.702	0.401	-
2016	352	0.392	41	1.759	0.385	0.067
2017	291	0.371	35	1.718	0.408	0.077
2018	286	0.392	36	1.743	0.385	0.087
2019	314	0.346	34	1.591	0.440	0.086
2020	285	0.318	37	1.495	0.476	0.077
2021	359	0.357	35	1.627	0.426	0.089
2022	314	0.328	36	1.537	0.462	0.070
2023	365	0.368	30	1.613	0.415	0.083
2024	384	0.376	31	1.656	0.405	0.042
2025	439	0.379	37	1.674	0.402	0.052

2) Country Diversity and Concentration

Country Shannon entropy further elucidates the structural evolution of global participation. The entropy value peaked in FY 2016 (1.759) and fell to its minimum in FY 2020 (1.495). The HHI exhibits a consistent pattern. Country concentration reached its maximum in FY 2020 (0.476) and declined gradually thereafter, falling to 0.402 in FY 2025.

Collectively, entropy and HHI measures demonstrate that FY 2020 represents a temporary phase of increased structural concentration, followed by a gradual diversification trend through FY 2025.

3) Structural Stability

The SCI indicator quantifies the year-to-year redistribution of country shares. Overall, the country-level SCI values remained relatively modest. The most pronounced structural shift occurred between FY 2020 and FY 2021, reflecting a post-contraction re-balancing of global participation patterns.

By contrast, the SCI for FY 2024-FY 2025 (0.052) was comparatively low, suggesting that the international usage structure had stabilized into a relatively steady configuration by the end of the observation period.

B. Thematic Diversity of Data-Supported Research

Thematic diversity is evaluated using keyword distributions extracted from Project/Experiment Titles. Topic Shannon Entropy and Topic SCI are used to measure diversity and temporal reconfiguration.

1) Topic Entropy

Topic entropy rose sharply from 4.780 in FY 2015 to 6.263 in FY 2016 (TABLE II), which may partly reflect database expansion and improved metadata registration in the early construction phase. From FY 2016 onward, topic entropy remained persistently high, ranging between approximately 6.11 and 6.36. Notably, it reached the highest observed value of 6.362 in FY 2025, indicating that DOE open data infrastructures support an increasingly diverse array of research activities. This sustained high level of entropy demonstrates that data usage has not become concentrated within specialized domains, but has instead continued to diversify across thematic areas.

TABLE II. THEMATIC STRUCTURE OF OPEN RESEARCH INFRASTRUCTURE DATA USE (FY 2015 – FY 2025)

Year	Topic Structure		
	Keywords_n	Topic_entropy	SCI_topic
2015	237	4.780	-
2016	991	6.263	0.633
2017	1031	6.263	0.295
2018	1082	6.233	0.303
2019	1100	6.200	0.335
2020	1072	6.223	0.265
2021	1138	6.106	0.357
2022	1176	6.244	0.259
2023	1183	6.217	0.241
2024	1216	6.257	0.206
2025	1285	6.362	0.288

2) Topic Structural Change

Topic SCI reached its maximum during the FY 2015-FY 2016 transition, consistent with the expansion of available datasets. After FY 2016, SCI values declined, indicating a gradual stabilization of the thematic structure. In FY 2025, topic SCI showed a moderate increase (0.288), suggesting a renewed redistribution of thematic emphasis rather than structural stagnation. This change may reflect the emergence of new research areas or enhanced cross-domain integration.

In summary, combining international and thematic indicators reveals several structural characteristics of DOE open research infrastructure data use. First, international participation remains resilient. Although international diversity exhibited a temporary contraction around FY 2020, the system subsequently recovered and re-diversified by FY 2025, indicating adaptive capacity in response to external disruptions. Second, thematic breadth remains high, as reflected by sustained levels of topic entropy from FY 2016-FY 2025. This suggests that open infrastructure data support multi-domain research rather than a narrow disciplinary focus. Third, SCI values indicate a pattern of gradual structural evolution rather than abrupt change, reflecting continuous adjustment within the data ecosystem, as opposed to pronounced volatility or disruption. Finally, by FY 2024-FY 2025, both country-level SCI and entropy indicators converge toward stabilization, implying that the system has transitioned into a relatively mature and stable configuration.

IV. Conclusion and Future work

The results indicate that DOE user facilities increasingly function as globally embedded data ecosystems rather than solely as physical experimental infrastructures. The sustained Non-U.S. participation ratio (approximately 0.37-0.39 in recent years), along with the recovery of country entropy after FY 2020, suggests that open research infrastructures extend beyond domestic use and operate within a distributed international knowledge network. This transformation aligns with the broader shift of large-scale research infrastructures toward structured data platforms that support global reuse through curated datasets, digital access mechanisms, and standardized metadata.

Despite a temporary rise in structural concentration around FY 2020, likely associated with global disruptions, the subsequent decline in the HHI and recovery of entropy indicate a re-diversification of participation patterns between FY 2021 and FY 2025. This trend, together with moderate SCI values, suggests that structural changes occur through gradual redistribution rather than abrupt transformation, reflecting adaptive stability within the system.

In parallel, consistently high topic entropy indicates that DOE open data resources support a wide range of research domains. The coexistence of thematic diversity and structural stability suggests that expansion occurs within a coherent platform structure rather than through fragmentation. The moderate increase in topic SCI observed in FY 2025 further suggests emerging thematic recombination, highlighting the role of open infrastructures in facilitating cross-domain knowledge integration in data-intensive scientific research.

It should be noted that proposal-level data reflect declared research intent rather than realized downstream scientific outputs. Nevertheless, such data can serve as a meaningful demand-side proxy for infrastructure data use, as research proposals represent anticipated data needs, planned methodologies, and targeted research questions. In this sense, proposal records provide an early-stage characterization of knowledge production activities, although their linkage to final scientific outputs remains indirect and probabilistic.

Future work will integrate publication data and citation networks to establish explicit connections between research proposals and downstream scientific outputs. This integration is expected to support a more comprehensive assessment of how infrastructure-generated data contribute to knowledge production, research impact, knowledge recombination, and cross-domain diffusion. In addition, more refined topic modeling approaches and cross-facility comparisons will be incorporated to further examine thematic evolution and the role of open research infrastructures in data-intensive science.

References

[1] K. M. Tolle, D. S. W. Tansley, and A. J. Hey, “The fourth paradigm: data-intensive scientific discovery [point of view],” Proceedings of the IEEE, vol. 99, no. 8, pp. 1334–1337, 2011.

[2] X. Li, and Y. Guo, “Paradigm shifts from data-intensive science to robot scientists,” Science Bulletin, vol. 70, no. 1, pp. 14–18, 2025.

[3] U.S. Department of Energy. *Definition of a user facility*. [Online]. Available from: https://science.osti.gov/-/media/_/pdf/user-facilities/memoranda/Office_of_Science_User_Facility_Definition_Memo.pdf

[4] U.S. Department of Energy. *User projects / experiments database for the office of science user facilities*. [Online]. Available from: https://science.osti.gov/-/media/_/pdf/user-facilities/memoranda/Office_of_Science_User_Projects_Experiments_Database_Memo.pdf

[5] C. L. Borgman, “The conundrum of sharing research data,” Journal of the American Society for Information Science and Technology, vol. 63, no. 6, pp. 1059–1078, 2012.

[6] L. Leydesdorff, “The static and dynamic analysis of network data using information theory,” Social Networks, vol. 13, no. 4, pp. 301–345, 1991.

[7] A. Stirling, “A general framework for analyzing diversity in science, technology and society,” Journal of The Royal Society Interface, vol. 4, no. 15, pp. 707–719, 2007.

[8] D. Maier, A. Niekler, G. Wiedemann, and D. Stoltenberg, ″How document sampling and vocabulary pruning affect the results of topic models, ″ Computational Communication Research, vol. 2, no. 2, pp. 139–152, 2020.

[9] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," presented at the Proceedings of the 22nd international conference on World Wide Web, Rio de Janeiro, Brazil, 2013. [Online]. Available from: https://doi.org/10.1145/2488388.2488514.

[10] C. S. Wagner, H. W. Park, and L. Leydesdorff, "The Continuing Growth of Global Cooperation Networks in Research: A Conundrum for National Governments," PLOS ONE, vol. 10, no. 7, p. e0131816, 201.

A Cross-source Topic Fusion and Multi-dimensional Synergistic Indicator Approach for Emerging Technology Identification

Xueli Yu
Business School
Shandong University of Technology
Zibo, China
e-mail:15063009091@163.com

Haiyun Xu
Business School
Shandong University of Technology; Max Planck Institute for Solid State Research
Zibo, China; Stuttgart, Germany
e-mail: xuhaiyunnemo@gmail.com

Robin Haunschild
Max Planck Institute for Solid State Research
Stuttgart, Germany
e-mail: R.Haunschild@fkf.mpg.de

Zhengyin Hu
National Science Library
Chinese Academy of science
Chengdu, China
e-mail: huzy@clas.ac.cn

Zenghui Yue
School of Medical Information Engineering
Jining Medical University
Rizhao, China
e-mail: yzh66123@126.com

Chunjiang Liu
National Science Library
Chinese Academy of science
Chengdu, China
e-mail: liucj@clas.ac.cn

***Abstract*—Emerging Technologies (ETs) play a crucial role in scientific revolutions and industrial transformation. Accurate identification of ETs contributes to effective national policymaking and the rapid advancement of science and technology. However, existing studies primarily rely on single data sources, leading to limitations in timeliness and comprehensiveness of identification results. To address these shortcomings, this study proposes a novel approach for emerging technology identification that integrates multi-source data fusion with coordinated multi-dimensional indicators. First, heterogeneous multi-source data are collected, and candidate technology topics are extracted using BERTopic-based topic modeling. Second, semantic similarity analysis is employed to fuse technology topics across different data sources, generating a comprehensive set of candidate ETs and constructing an indicator system for their identification. Finally, ET topics are screened and identified. The results are subsequently validated. An empirical analysis in the smart grid domain identifies nine ETs. The findings provide important references for technology forecasting, policy formulation, and industrial strategic planning.**

Keywords-Emerging Technologies; Multi-source data; BERTopic; Smart grid.

I. INTRODUCTION

Against the backdrop of rapid technological iteration, Emerging Technologies (ETs) serve as the core force driving the development of new-quality productive forces and industrial transformation. Conducting forward-looking identification of ETs is an important prerequisite for decision-makers to seize strategic opportunities [1]. With advances in computing and digital technologies, ETs evolve at an exponential rate and in a combinatorial manner. While characterized by radical novelty, relatively rapid growth, and potential impact [2], technological systems also demonstrate an evolutionary feature of coexisting high uncertainty and path dependence [3].

During the life cycle of ETs—from basic research to applied development, and then to commercialization and scaling-up—not every technology can successfully cross the "valley of death" [4]. Therefore, identifying ETs with substantial developmental potential amid a highly uncertain technological landscape is a core issue in current research.

In addition, most existing technology identification studies focus on papers and patents as research objects [5], leading to relatively one-sided analytical results. Moreover, the publication of scientific papers and patents entails a certain time lag, which tends to compromise the timeliness of prediction outcomes [6], while research on multi-source heterogeneous data remains insufficient [7][8]. In contrast, data from funding, industry, policies, and reports can reflect the dynamic changes of ETs at different levels, including early-stage layout, application-driven development, and policy orientation.

Accordingly, to comprehensively capture the trajectory characteristics of technological leapfrogging, the joint analysis of multi-source heterogeneous data helps to construct a more real-time, comprehensive, and refined picture of technological evolution. It is particularly critical to conduct an integrated analysis using multi-source data containing multi-level technical information to identify ETs [9].

Against this background, we propose an ET identification method that integrates multi-source data and multi-dimensional features. The main contributions are as follows: integrating multi-source heterogeneous data including papers, patents, funding data, industry reports, industrial market data and policy documents; depicting the comprehensive characteristics of ETs from multiple dimensions; and promoting more accurate identification of ETs that reflects their scientific foundation and application potential.

Nevertheless, this study still has several limitations. First, the proposed framework mainly relies on static cross-sectional data and lacks dynamic analysis of technology evolution over time. Second, although multi-source heterogeneous data are integrated, differences in data quality and coverage may affect the completeness of topic extraction results. In addition, the uncertainty dimension mainly focuses on structural entropy within topic networks, while external factors, such as market and policy changes are not fully considered. These limitations can be further addressed through richer data integration and dynamic evolutionary analysis in future research.

This paper is organized as follows. Section I introduces the research background, research motivation, and main contributions of this study. Section II reviews the concepts, characteristics, and existing identification methods of emerging technologies. Section III presents the research design, including multi-source data acquisition, BERTopic-based topic extraction and fusion, and the construction of the ET identification indicator system. Section IV conducts an empirical analysis in the smart grid domain and reports the identification and validation results of emerging technology topics. Finally, Section V concludes the paper and discusses the limitations and future research directions.

II. RELATED RESEARCH

This section systematically reviews the existing literature on emerging technologies, focusing on their conceptual definitions, characteristic frameworks, and mainstream identification approaches, so as to clarify the research gaps that the present study aims to address.

A. Concepts and Connotations of Emerging Technologies

At present, the academic community has not yet given a unified definition of "ET". Initially proposed by the Wharton School of the University of Pennsylvania, it is believed that ET is rooted in scientific discoveries, which can not only spawn new industries but also reshape existing ones [10]. On this basis, scholars have further supplemented and defined it from the dimensions of technical characteristics and technical effects combined with the theories of G.S. Day et al. [11]-[13]. In addition, compared with cutting-edge technologies and core technologies, ETs emphasize more on the early scientific foundation and potential impact, and show significant growth potential in the prototype stage of its life cycle.

B. Feature Framework and Identification Dimensions of Emerging Technologies

ETs occupy an important position in knowledge breakthroughs, industrial transformation, and future competition, and have long been the focus of research in the fields of technology forecasting and scientometrics. Existing studies can be roughly divided into two categories: One is the identification path based on "method tools", and the other is the identification path based on "feature framework".

In the research from the perspective of method tools, traditional scientometric methods mainly rely on the citation relationship between documents or the co-occurrence relationship of keywords, including direct citation analysis, co-citation analysis, citation coupling, co-word analysis, and overlay mapping [14]. These methods can effectively depict the research hotspots, knowledge structure, and development context of technologies, and are the basic tools for identifying technological frontiers with the advancement of technology and the expansion of identification tools. Different from the research path based on method tools, another group of scholars starts from "feature attributes" and understands ETs as a set of identifiable and measurable features. A representative one is the five-feature model proposed by Rotolo et al. [2], which holds that ETs have radical novelty, relatively rapid growth, coherence, significant impact, and uncertainty and ambiguity. This model provides a systematic conceptual framework for identifying and understanding ETs, clarifying five key dimensions, but the model itself focuses on conceptual elaboration. In recent years, methods for identifying ETs using text mining have been widely applied due to their high efficiency and accuracy [15]. Among them, the ET identification method based on topic modeling has gradually become the mainstream. It forms topics by clustering semantically related or similar technical keywords, and identifies topics that meet the attributes of ETs by analyzing their evolutionary characteristics, growth rate, and structure [16]-[18].

In addition, most existing studies adopt a single data source for technology identification, which has certain limitations. Methods for identifying ETs by fusing multi-source data have attracted more and more attention. By integrating data, such as academic papers, patents, and news reports, the development trend and market potential of ETs can be captured more comprehensively and accurately [19].

In summary, a relatively complete conceptual recognition and indicator identification system has been formed for ET research, laying a solid theoretical foundation for the identification of ETs. However, existing studies still have limitations in the identification and analysis of ETs: The data sources are relatively single, most studies only rely on papers and patents for identification, and lack the integration of multi-source data, such as industrial market data, reports, and funding, leading to insufficient representativeness of identification results and inadequate guidance for policy formulation and enterprise development. In view of this, we propose an ET identification method based on the fusion of multi-source data and multi-dimensional features. It obtains technical topics through topic modeling, constructs an ET identification indicator system, and systematically identifies ETs, in order to provide support for policy implementation and enterprises to gain a leading position in development.

III. RESEARCH DESIGN

This section elaborates the overall research framework and technical routes of this study, including multi-source data processing, topic extraction and fusion, and the construction of a multi-dimensional indicator system for emerging technology identification.

A. Research Design

We construct a research framework that mainly consists of three core modules, as shown in Figure 1.

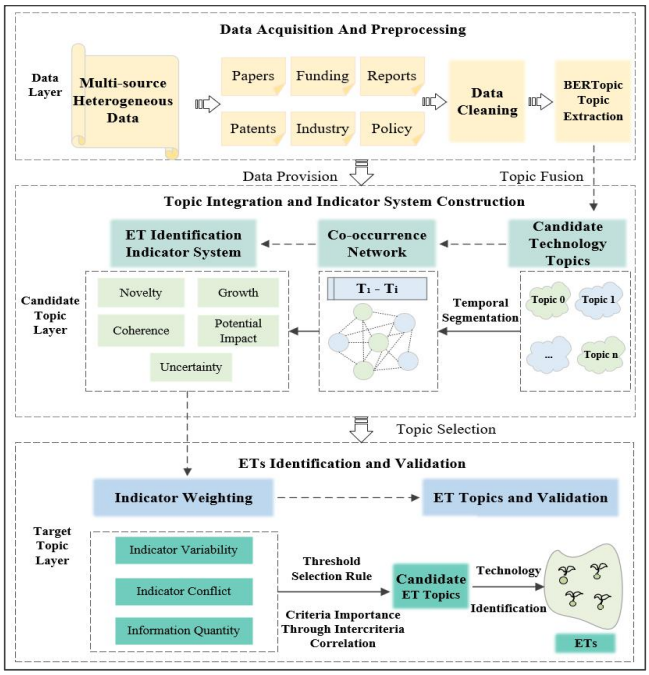

Figure 1. Research Framework Diagram.

First, the acquisition and preprocessing of multi-source heterogeneous data are carried out. Through the construction of a search query system, papers, patents, funding data, industry reports, industrial market data and policy documents are collected. After manual cleaning and screening, as well as standardized preprocessing, such as word segmentation and vectorization, standardized data is formed.

Second, based on the preprocessed data, the BERTopic model is used to extract initial topics, and cross-data source topic fusion is completed according to semantic similarity to construct a global candidate technical topic and topic co-occurrence network, providing support for indicator calculation.

Finally, the calculation of candidate ET topics is completed according to the ET identification indicator system, and the screening and verification of ETs are carried out accordingly.

B. Acquisition and Identification of Technical Topics

1) Acquisition and Fusion of BERTopic-based Topics

As a topic modeling method integrating embedding models and clustering algorithms, BERTopic can realize deep semantic mining, efficient dimensionality reduction, clustering, and visual analysis, making it suitable for identifying potential topics in complex corpora. We conducted technical topic modeling with BERTopic as the core. First, we converted multi-source texts into semantic vectors through a pre-trained model, then automatically identified semantically similar documents through Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) density clustering to avoid deviations caused by manually setting the number of topics. Finally, we generated topic representations by combining the weights of topic words, laying a foundation for topic fusion and indicator calculation.

On this basis, we used cosine similarity for topic fusion. Since the topic boundaries of papers and patents are clear, while the topics of data, such as funding data, reports, industrial market data and policy documents are cross-domain and ambiguous, making direct alignment difficult, we designed a two-stage fusion strategy to unify the semantic space and improve topic directions. In the first stage, we calculated the cosine similarity between paper and patent topics, and pair and merge them according to the principle of second-highest similarity to eliminate redundancy within sub-directions. In the second stage, we fused the obtained topics with public topics again, and finally formed a global candidate technical topic that covers multi-source data, has a unified semantics, and complete directions.

2) Identification of Target Technical Topics

We used the TF-IDF algorithm to extract keywords from topic texts and construct a co-occurrence network of technical topics. Then, we performed calculation and standardization processing in accordance with the constructed indicator system. To reduce the impact of subjectivity, we adopted the CRITIC objective weighting method to determine the objective weight of each indicator, and screen out representative emerging technology topics by setting thresholds.

C. Emerging Technology Indicator System

By sorting out existing research on ET identification, we constructed a multi-dimensional identification indicator system from five dimensions—novelty, growth, coherence, potential impact, and uncertainty—to systematically depict the characteristics of ETs.

1) Novelty

To comprehensively depict the novelty of topics, we constructed topic novelty indicators from both temporal and semantic perspectives. The former reflects the recency of a topic's emergence in the temporal dimension, while the latter measures the difference of a topic in the semantic space.

We use the average publication time of all documents in a topic as the temporal novelty indicator [20], and the formula is as follows:

$$N_i^{(t)} = \frac{1}{N_i} \sum_{j=1}^{N_i} T_{ij} \quad (1)$$

Here, $N_i^{(t)}$ denotes the temporal novelty of the *i*-th topic; T_{ij} represents the publication year of the *j*-th document within the *i*-th topic; and N_i indicates the total number of documents contained in the *i*-th topic.

Semantic novelty is operationalized as the inverse of the average inter-topic semantic similarity calculated based on TF–IDF vectors, thereby measuring the distinctiveness of a topic within the semantic space. Specifically, the procedure is as follows: first, the keywords of each topic are transformed into semantic vectors using the TF–IDF representation. Second, cosine similarity between topics is computed to characterize the degree of proximity in their semantic content. Finally, the inverse of a topic's average similarity to all other topics is taken as its semantic novelty indicator. The corresponding formulation is as follows:

$$\overline{Sim}(t_i)=\frac{1}{n\text{-}1}\sum_{j=1,j\neq i}^{n}\frac{v_i \cdot v_j}{\|v_i\|\|v_j\|} \tag{2}$$

$$N_i^{(s)}=1\text{-}\overline{Sim}(t_i) \tag{3}$$

Here, $N_i^{(s)}$ denotes the semantic novelty of the *i*-th topic. The keyword set of the *i*-th topic is represented as a vector v_i, and n denotes the total number of topics. The semantic similarity between any two topics t_i and t_j is defined using cosine similarity. $\overline{Sim}(t_i)$ represents the average semantic similarity between topic t_i and all other topics.

2) Growth

Topic growth is measured by topic intensity, defined as the number of supporting documents under a given topic. This value can be directly obtained from the document–topic mapping generated by the BERTopic model. Specifically, we calculated the proportion of documents associated with a given topic relative to the total number of documents across all topics in each year and then computed the average of these proportions to reflect the sustained growth level of the topic within the overall research landscape [21]. The formulation is as follows:

$$G_i=\frac{1}{Y}\sum_{y=1}^{Y}\frac{C_{ij}}{C_{total,\,y}} \tag{4}$$

In (4), G_i denotes the growth indicator of topic *i*; C_{ij} represents the number of documents under topic *i* in year *y*; and $C_{total,\,y}$ denotes the total number of documents across all topics in year *y*. In this study, the time window is fixed at $Y = 5$. The average growth rate is calculated over the most recent five years.

$$N_i^{(s)}=1\text{-}\overline{Sim}(t_i) \tag{5}$$

3) Coherence

Topic coherence is used to characterize the semantic stability and conceptual consistency of a topic over time, namely, the degree of similarity in its keywords and semantic representations across consecutive years. This study constructs a dual-measurement framework at both the lexical and semantic levels. The two measures are integrated to form a comprehensive coherence indicator, reflecting the stability and consistency of a topic during its semantic evolution.

The similarity between the same topic across adjacent time slices is measured based on the documents it contains. A higher Jaccard coefficient indicates stronger coherence [16]. Specifically, topic texts for each year are vectorized, and the Jaccard coefficient between the Top-K keyword sets of two consecutive years is used to quantify lexical stability. This coefficient is compared longitudinally with the topic's own historical values to trace its evolutionary trajectory, rather than being compared horizontally with other topics in the same period. The formulation is:

$$J_{t,\,t+1}^{(k)}=\frac{|K_t \cap K_{t+1}|}{|K_t \cup K_{t+1}|} \tag{6}$$

Here, K_t denotes the Top-K keyword set extracted via TF–IDF from the topic-related texts in year t; $J_{t,\,t+1}^{(k)}$ denotes the Jaccard similarity between the two consecutive annual keyword sets.

An autocorrelation coefficient is employed to measure the similarity of a topic to itself at different time points. In the context of topic coherence, it captures semantic self-similarity over time. We utilized SBERT sentence embeddings to compute the cosine similarity between the average textual vectors of adjacent years:

$$R_{t,\,t+1}^{(s)}=\cos(\vec{v_t},\vec{v_{t+1}})=\frac{\vec{v_t}\cdot\vec{v_{t+1}}}{\|\vec{v_t}\|\|\vec{v_{t+1}}\|} \tag{7}$$

In (7), v_t denotes the mean embedding vector of all topic-related texts in year t encoded by SBERT, v_{t+1} represents the corresponding vector in year t+1, and $R_{t,\,t+1}^{(s)}$ is the semantic cosine similarity.

4) Potential Impact

In the context of multi-source data integration, potential impact reflects the cross-domain diffusion degree of a topic across heterogeneous data sources. The more often a topic appears in data sources, the stronger its cross-domain influence. Accordingly, we operationalized potential impact as the number of distinct data source categories covered by a given topic.

5) Uncertainty

Uncertainty is assessed through changes in structural entropy. A higher structural entropy indicates greater complexity and uncertainty within the technological topic network. Extracted topic keywords are treated as nodes, and a network is constructed based on associations among technological topics. Structural entropy theory is then applied to analyze the network structure [22]. The formulations are:

$$p(x_i)=\frac{d_i}{2E} \tag{8}$$

$$H(X)=-\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \tag{9}$$

Here, $H(X)$ denotes structural entropy; $p(x_i)$ represents the probability associated with node i, calculated as d_i (the degree of node i, i.e., the number of edges connected to it) divided by twice the total number of edges in the network ($2E$).) The term 2E is used because each edge is counted twice when computing node degrees (once for each endpoint).

IV. Empirical Analysis

This section carries out an empirical study in the smart grid domain to validate the proposed method, including data collection, topic modeling, indicator calculation, result screening, and comparison with authoritative forecasts.

A. Data Acquisition and Preprocessing

We selected the smart grid sector as the research object for ET identification. On the one hand, this field features a high degree of technological intersection, and ETs often emerge rapidly in the processes of engineering application, system coupling, and cross-scenario deployment. On the other hand, the smart grid undertakes multiple national strategic tasks, and its technological development path is highly concerned by policies, industries, and scientific research institutions, ETs play a critical role in the overall performance of the smart grid system.

We systematically constructed search strategies to obtain data on papers, patents, funding data, industry reports, industrial market data and policy documents in eight sub-directions of the smart grid field. The data preprocessing link mainly includes deduplication, word segmentation, and stop-word removal, etc. The sources and descriptions of data acquisition are shown in Table 1.

TABLE I. Sources and Descriptions of Data Acquisition

Data Type	Source	Preprocessed Data Volume
Publications	Web of Science	44,642
Patents	Incopat Patent Database	11871
Funding data	NSF Database	375
Industry reports	Conference Reports	126
Industry market data	Kehui Network, Cigre Database	496
Policy documents	Peking University Law Database	725

B. Acquisition of Multi-source Technology Topics

We conducted topic modeling on multi-source text data based on the BERTopic model. On the one hand, we evaluated the effect of topic modeling by relying on the Intertopic Distance Map provided by BERTopic. We determined topics according to the following criteria: The clearer the topic boundaries, the closer the semantic clusters, and the less overlap, the more interpretable and structural the topics extracted by the model [23]. On the other hand, in order to adapt to the differences in text scale and semantic density among different data sources, we adopted the "auto" default parameter of BERTopic to determine the appropriate number of topics in the selection of topic quantity. The advantages of this strategy are as follows: (1) It avoids overfitting or excessive topic splitting that may be caused by artificially setting the number of topics. (2) It can identify the optimal number of clusters according to the actual density distribution of data. When the corpus of data expands or shrinks, the optimal number of topics will increase or decrease accordingly, thus ensuring the reproducibility of results and the applicability of domain migration.

C. Identification Results of Emerging Technology Topics

After completing the topic extraction of multi-source data, the 309 identified initial technology topics were cross-source fused according to the two-stage cosine similarity fusion strategy to obtain candidate technology topics. On this basis, according to the indicator system constructed in the theoretical part, the indicator values were calculated for the fused candidate topics. After determining the weights by the CRITIC objective weighting method, the comprehensive indicator values were calculated and normalized, and finally the target technology topics were screened out. The process and results of technology topic identification in the smart grid research are shown in Table 2.

TABLE II. The Process and Results of Technology Topic Identification in Smart Grid Research

Step	Method / Strategy	Number of documents and topics
Multi-source data acquisition	Data retrieval and collection	40,248 documents
Candidate technical topics	BERTopic topic modeling	309 topics
Integrated technical topics	Cosine similarity-based fusion	233 topics
Target technical topics	Weight determination via CRITIC entropy method	117 topics

After indicator calculation, we assigned weights according to the CRITIC entropy method (Weight determination via CRITIC entropy method) to calculate the comprehensive indicator values. We then selected the top 50% of the comprehensive values as ET topics, and a total of 117 target technology topics were screened out. The threshold setting is based on analyzing the characteristics of the indicator values of candidate ET topics, and considering the conversion rate of the technology life cycle [2] from "emergence" to "take-off", so 50% is taken to ensure that the subsequent incubation

resources are of the same order of magnitude as the conversion rate. The standardized characteristic indicator values of some ET topics in each sub-direction are shown in Table 3.

TABLE III. CHARACTERISTIC INDICATOR VALUES OF SOME ET TOPICS IN SUB-DIRECTIONS

Subfield	Topic ID	Novelty	Growth	Coherence	Potential Impact	Uncertainty	Comprehensive Value
Power System and Planning	Topic 0	0.3787	0.9584	0.8316	0.3333	0.8699	0.5862
	Topic 7	0.2737	1.0000	0.6890	0.6666	0.5370	0.5470
	Topic 1	0.2801	0.4617	0.8511	0.0000	1.0000	0.4740
Power Supply Side	Topic 0	0.2254	1.0000	1.0000	1.0000	0.9537	0.7493
	Topic 2	0.3108	0.1649	0.4861	0.5000	1.0000	0.5065
	Topic 1	0.2524	0.3746	0.7844	0.2500	0.7567	0.4646
Power Grid Side	Topic 3	0.1030	0.1469	0.1280	0.1820	0.1077	0.6677
	Topic 1	0.1201	0.0725	0.1537	0.1213	0.1408	0.6086
	Topic 0	0.1062	0.0858	0.1240	0.0000	0.2273	0.5436
Load Side	Topic 1	0.3096	0.7109	0.9425	1.0000	1.0000	0.8019
	Topic 11	0.2427	0.9997	0.8697	1.0000	0.5740	0.7328
	Topic 3	0.2900	0.4081	0.7584	1.0000	0.9204	0.6977
.....							

Next, we selected the names of some ET topics, the number of included documents, and the topic words from Table 3 to display them in Table 4.

TABLE IV. NAMES, NUMBER OF INCLUDED DOCUMENT ITEMS AND TOPIC WORDS OF SOME EMERGING TECHNOLOGY TOPICS

Subfield	Topic ID	Topic name	Document count	Representative keywords
Power System and Planning	Topic 1	wind_power_energy_model	717	wind, power, energy, model, generation, storage, speed, uncertainty, systems, farms
Power Supply Side	Topic 2	heat_efficiency_temperature_thermal	340	heat, efficiency, temperature, thermal, fuel, cell, solar, performance, cycle, exergy
Power Grid Side	Topic 0	energy_microgrid_model_optimal	413	energy, microgrid, model, optimal, storage, microgrids, optimization, proposed, operation, cost
Power materialsSide	Topic 11	gas_natural_co_electricity	21	gas, natural, co-power , electricity, model, power, integrated, fired, systems, technologies
.....				

D. Verification

After obtaining the ET topics in the smart grid field through the above process, in order to verify the rationality and reliability of the technology identification indicator system constructed in this paper, the technology topics identified in this paper are compared with relevant technology forecast reports. As shown in Table 5.

TABLE V. COMPARISON OF ETS IDENTIFIED IN THIS PAPER AND THOSE MENTIONED IN RELEVANT TECHNOLOGY FORECAST REPORTS

Emerging Technology Content	Supporting Reports and Policies
Core Processes of Solid-State Batteries, Sodium-Ion Battery Technology, New Battery Chemistry Technology	IEEE CS Authoritative Forecast: One of the breakthrough technologies in 2025: New battery chemistry technologies, including solid-state and sodium-ion batteries, need to break through mass production processes and supply chain management to further improve their energy density and safety [24].
Structural Battery Composite Materials, High-Performance Electrode Material Technology, "Solid-Solid Interface Regulation Preparation Process"	The "Frontier Situation Analysis of Key Scientific and Technological Fields 2025" jointly researched by the Institute of Scientific and Technical Information of China and the Shanghai Institute of Science of Science focuses on cutting-edge directions, such as the innovation of solid-state electrolyte material systems, the research and development of high-performance positive/negative electrode materials, and the breakthrough of solid-solid interface regulation preparation processes, conducting annual dynamic tracking [25].
Structural Battery Composite Materials, Osmotic Energy Power Generation System Technology, Green Carbon Sequestration Technology	The World Economic Forum released the Top 10 ETs of 2025 at the 16th Summer Davos Forum, where green carbon sequestration, structural battery composite materials and osmotic energy power generation system technology were selected into the annual "list" [26].

E. Discussion

Taking the smart grid as the research object, we integrated six types of multi-source heterogeneous data, completed technology topic extraction and cross-source fusion based on the BERTopic model, and combined the five-dimensional indicator system and CRITIC weighting method to finally identify 117 ET topics, covering eight sub-directions. The core technologies are highly consistent with the 2025

technology forecasts of authoritative institutions, such as IEEE CS and the World Economic Forum, verifying the applicability of the method.

From the perspective of identification results, the development of ETs in the smart grid field highlights the trends of low carbonization, energy storage, and interdisciplinarity. Cross-field technologies, such as solid-state batteries, green carbon sequestration, and structural battery composite materials have become important innovation tracks. Compared with traditional identification methods based on single data sources, the proposed framework integrates scientific research, industrial application, and policy orientation, enabling a more comprehensive depiction of ET evolution characteristics and improving the timeliness and completeness of technology identification.

The research results can provide accurate guidance for policy formulation and industrial layout in the smart grid field. For high-potential technology directions, policy support, and R&D investment can be increased, while paying attention to the trend of technological cross-integration to promote cross-field innovation. The identification method based on multi-source data fusion and multi-dimensional indicator coordination proposed in this research also provides a transferable research framework for ET identification in other fields.

V. Conclusion and Future Work

This study proposes an ET identification method integrating multi-source data fusion and coordinated multi-dimensional indicators. By combining BERTopic-based topic extraction, semantic similarity-based topic fusion, and a five-dimensional indicator system, the method enables systematic identification of ETs from heterogeneous data sources. The empirical analysis in the smart grid field demonstrates the applicability of the proposed framework.

The main contributions of this study are as follows: multi-source data integration compensates for the limitations of single-source identification; scientific research, industrial application, and policy orientation are incorporated into a unified analytical framework; and a multi-dimensional indicator system is constructed to quantitatively characterize ETs.

This research has certain limitations: It lacks dynamic evolution analysis of technology topics, conducts topic extraction and identification based on static cross-sectional data, and does not carry out tracking analysis from a time series dimension. The uncertainty indicator does not consider external factors, such as market and policies. The data sources focus on specific providers. In the future, data can be further enriched, time series analysis methods can be introduced, and cross-field empirical research can be carried out to further improve the versatility and adaptability of the method.

Funding: This research was funded by the National Natural Science Foundation of China (No. 72274113), Shandong Provincial Social Science Foundation (No. 23CTQJ07), Shandong Provincial Natural Science Foundation (No. ZR2022MG052), Beijing Natural Science Foundation (No. 9242006) and the Taishan Scholar Foundation of Shandong province of China (tsqn202103069).

Use of Generative Artificial Intelligence for Writing: We used ChatGPT for translation, proofreading, and grammar checking. We evaluated the output by cross-referencing the translated and revised content with the original text to ensure accuracy, consistency, and alignment with the intended meaning. Additionally, we reviewed the final version to confirm that all technical terms and concepts were appropriately conveyed.

References

[1] Y. Zhou, F. Dong, Y. Liu, and L. Ran, “A deep learning framework to early identify emerging technologies in large-scale outlier patents: an empirical study of CNC machine tool,” *Scientometrics*, vol. 126, no. 2, pp. 969–994, Feb. 2021.

[2] D. Rotolo, D. Hicks, and B. R. Martin, “What is an emerging technology?,” Research Policy, vol. 44, no. 10, pp. 1827–1843, Dec. 2015.

[3] "The Nature of Technology: What It Is and How It Evolves, W.B. Arthur. Free Press, New York (2009), 246 pp.," ResearchGate, Aug. 2025.

[4] X. Wu, Y. Shao, and F. Lin, "The Path Mechanism for Breaking Through the "Valley of Death" of Key Core Technology Innovation," Sci. Res., vol., no., Dec. 2024, Advance online publication.

[5] Z. Wang, Y. Chen, Z. Jiang, and Z. Liu, "An input output analytical method of identiHcation of the advanced technology and core technology; The case study of hybrid electric vehicle technology," Sci. Res., vol. 33, no. 11, pp. 1612–1620, 2015.

[6] X. Wang, J. Dong, T. Yu, Y. Chen, and T. Chen, "Research on Technology Fronts Forecasting Method Based on Informetrics Analysis Using Multi-Type Information," J. Inf., vol. 37, no. 10, pp. 70–75, 89, 2018.

[7] G. Li, X. Wu, and B. Ning, “Research Frontier Detection Method Based on Topic Time Series Diffusion Network between Multi-source Data,” Data Anal. Knowl. Discov., Apr. 2025, Advance online publication, Accessed: Nov. 15, 2025.

[8] X. Zhang, Z. Zhang, L. Cao, W. Ruan, X. Ren, and Z. Feng, "Research Progress of Research Front Recognition Methods in Subject Fields," Lib. Inf. Serv., vol. 66, no. 12, pp. 139–151, 2022.

[9] J. Shi, X. Zhao, and Q. Liu, "National Security-Oriented Intelligence Support Path," J. China Soc. Sci. Inf., vol. 39, no. 7, pp. 675–686, 2020.

[10] G. S. Day and P. J. H. Schoemaker, Wharton on Managing Emerging Technologies. John Wiley & Sons, 2004.

[11] G. S. Day and P. J. H. Schoemaker, “Avoiding the Pitfalls of Emerging Technologies,” California Management Review, vol. 42, no. 2, pp. 8–33, Jan. 2000.

[12] A. Breitzman and P. Thomas, “The Emerging Clusters Model: A tool for identifying emerging technologies across multiple patent systems,” Research Policy, vol. 44, no. 1, pp. 195–205, Feb. 2015.

[13] L. Wang, S. Fang, and P. Ji, "Using Patent Documents to Study the Technology Framework of Detecting Emerging Technology Topics," Lib. Inf. Serv., vol. 55, no. 18, pp. 74–78, 23, 2011.

[14] G. Li, F. Xie, X. Tang, B. Wang, and Y. Xu, "Research on the Identification of Emerging Technologies Integrating Multidimensional Perspectives," J. China Soc. Sci. Inf., vol. 44, no. 8, pp. 962–976, 2025.

[15] T. Zhang, J. Chen, X. Zhou, S. Sun, and Y. Zhang, "Emerging Technology Identification Method Based on Potential Impact Prediction and Multi-Source Information Fusion," J. Inf., vol. 44, no. 9, pp. 134–142, 2025.

[16] H. Xu, J. Winnink, Z. Yue, H. Zhang, and H. Pang, “Multidimensional Scientometric indicators for the detection of emerging research topics,” Technological Forecasting and Social Change, vol. 163, p. 120490, Feb. 2021.

[17] Y. Zhou, H. Lin, Y. Liu, and W. Ding, “A novel method to identify emerging technologies using a semi-supervised topic clustering model: a case of 3D printing industry,” Scientometrics, vol. 120, no. 1, pp. 167–185, Jul. 2019.

[18] M. Jiang, S. Yang, and J. Wei, "Emerging-Technology Identification Based on Technology Structure-Function Semantic Association,"J. China Soc. Sci. Tech. Inf., vol. 44, no. 5, pp. 522–534, 2025.

[19] S. Yang and M. Jiang, "Research Progress on Connotation Characteristics and Identification Methods of Emerging Technologies," Inf. Sci., vol. 41, no. 5, pp. 181–190, 2023.

[20] X. Zhu, J. Zhang, W. Li, X. Liu, and G. Geng, "Detecting Emerging Topics of Artificial Intelligence Based on Perspective of Top International Conferences," Inf. Theory Pract., vol. 47, no. 9, pp. 147–155, 2024.

[21] P. Liu, "Identifying emerging technology topic based on multi-source data and co-occurrence network," M.S. thesis, Univ. Chinese Acad. Sci. (Nat. Lib. China), 2024.

[22] D. Wang, X. Zhou, P. Zhao, J. Pang, and Q. Ren, “Early identification of breakthrough technologies: Insights from science-driven innovations,” Journal of Informetrics, vol. 19, no. 1, p. 101606, Feb. 2025.

[23] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” Mar. 11, 2022.

[24] ″IEEE CS authoritative forecast: 22 breakthrough technologies lead the future in 2025, ″ Accessed: Feb. 16, 2026. [Online]. Available: http://www.nscc-gz.cn/newsdetail.html?8213

[25] ″Analysis of Frontier Trends in Key Science and Technology Fields 2025 released at the 18th Pujiang Innovation Forum,″ Accessed: Feb. 16, 2026. [Online]. Available: https://www.siss.sh.cn/c/2025-09-24/665730.shtml

[26] ″World Economic Forum releases top 10 emerging technologies of 2025," Accessed: Feb. 16, 2026. [Online]. Available: https://www.ncsti.gov.cn/kjdt/kjrd/qtrd_kjrd/202506/t20250625_208621.html

Construction of A Causal Knowledge Graph for Research on Diabetes Comorbidities

Xueli Wu, Xuemei Yang, Longchao Wang, Yalan Huang, Xiaoli Tang
Institute of Medical Information
Chinese Academy of Medical Sciences and Peking Union Medical College
Beijing, China
e-mail: tang.xiaoli@imicams.ac.cn

Abstract—**Diabetes comorbidity is characterized by substantial mechanistic complexity and causal heterogeneity, which correlation-based approaches cannot adequately capture within deep pathological progression pathways. To resolve latent causal structures and mitigate generative hallucination in medical causal mining, this study introduces a hybrid paradigm that integrates physical anchoring, dual-channel evidence awareness, and topological reconstruction, thereby constructing the Diabetes comorbidity Causal Knowledge Graph (Diab-CKG). The framework establishes an atomized corpus indexing coordinate system to ensure traceable extraction and employs Large Language Models under strict ontology constraints to validate prior knowledge and identify novel associations, effectively suppressing generative hallucination. Experimental results demonstrate that the paradigm effectively connects unstructured text with structured reasoning, achieving robust performance in entity recognition and causal extraction, with an end-to-end Strict F1 score of 83.83% and a reduction of the Entity Hallucination Rate to 3.27%. The study delineates a comprehensive causal chain from risk exposure to clinical intervention, providing a logically coherent and computable foundation for modeling comorbidity cascades and supporting clinical decision making.**

Keywords- Diabetes mellitus; Causal knowledge graph; Comorbidity; Causal relationship extraction; Large language model

I. INTRODUCTION

Diabetes is a highly heterogeneous chronic metabolic disease whose clinical progression is frequently accompanied by complex comorbidity networks. Although epidemiological studies confirm high co-occurrence rates between diabetes and numerous comorbid conditions, most existing analyses remain confined to statistical correlations. These approaches cannot adequately capture mechanistic causal pathways, thereby limiting the development of prospective prevention and targeted intervention strategies for high-risk comorbidities. Therefore, constructing a causal knowledge graph that accurately describes disease progression directions and intermediate mechanisms has become a key task in knowledge driven healthcare.

However, building a high fidelity causal graph from unstructured biomedical literature faces multiple challenges. First, causal expressions in medical texts are highly implicit, with key pathological mechanisms often scattered across sentences, making long range logic difficult to capture for traditional methods [1]. Second, although Large Language Models (LLM) improve semantic understanding, their inherent hallucination risk may produce missing evidence or reversed logic in medical applications [2]. In addition, an internal tension exists between biological feedback loops and the Directed Acyclic Graph (DAG) structure required for computational causal reasoning. Balancing biological completeness and computational logical consistency in graph construction remains unresolved.

To address these challenges, this study proposes a hybrid computational paradigm that integrates physical anchoring, dual-channel evidence awareness, and topological reconstruction. The framework drives LLM reasoning through atomic-level evidence tracing. This design aims to resolve the precision–recall trade-off in causal mining and balance the tension between biological feedback loops and computational logic constraints. The result is an interpretable, logically consistent knowledge foundation for diabetes comorbidity research.

Finally, the remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 introduces the proposed framework, Section 4 presents the experimental results and discussion, and Section 5 concludes the paper and discusses future work.

II. RELATED WORK

To contextualize the proposed framework within existing research, this section reviews advances in medical knowledge extraction, cross-sentence causal relation extraction, and hallucination control in generative models, identifying the limitations that motivate our hybrid paradigm.

A. Medical Knowledge Extraction Methods

Medical knowledge graph construction has evolved from rule-driven dictionary mapping methods such as MetaMap and SemMedDB to data-driven deep learning models including Bidirectional Encoder Representations from Transformers (BERT) and Bidirectional Long Short-Term Memory with Conditional Random Field (BiLSTM-CRF). While deep learning enhances generalization in entity and relation extraction, its reliance on annotated corpora and degraded performance on long-tail concepts limit practical deployment [3]. The rise of LLM enables few-shot and zero-shot reasoning, yet semantic drift has prompted schema-constrained controlled generation strategies [4].

B. Cross Sentence Causal Relation Extraction

Unlike conventional semantic association extraction, causal mining demands stronger logical discrimination. Existing approaches incorporate causal connectives or syntactic dependency trees, but largely focus on explicit intra-sentence relations. Capturing cascade mechanisms across paragraphs remains challenging, particularly in diabetes research [5]. Current document-level models also lack robustness in coreference resolution and multiple negation handling common in biomedical texts, hindering strict causal logic enforcement [6].

C. Hallucination Control and Evidence Traceability in Generative Models

Retrieval-augmented generation reduces knowledge errors by introducing external bases, yet noisy retrieval can still induce misleading outputs. Verification strategies such as self-consistency checking and multi-agent debate improve accuracy at higher reasoning cost. However, they emphasize overall text quality and rarely provide atomic-level evidence auditing for knowledge graph construction [7]. Without precise linkage between generated triples and original text offsets, existing graphs struggle to satisfy evidence-based medicine requirements.

D. Research Focus

Overall, current paradigms often increase semantic depth at the expense of traceability and logical consistency. This study therefore injects structured prior knowledge into LLM reasoning and establishes a strict physical coordinate mapping system to develop an extraction paradigm that unifies semantic generalization with rule-level rigor, supporting diabetes comorbidity research.

III. METHODS

To address the common problems of hallucination generation and causal logic disconnection in medical text mining, this study establishes a hybrid computational paradigm that integrates physical anchoring, dual channel evidence awareness, and topological reconstruction.

A. Data Acquisition and Schema Definition

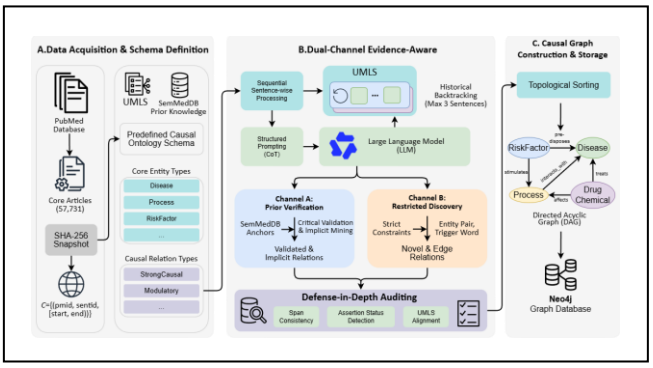

Figure 1. Dual-Channel Evidence-Aware Causal Knowledge Extraction Framework.

TABLE I. CAUSAL KNOWLEDGE GRAPH ENTITY TYPES

Entity Types	*Main UMLS content*	*Relation types*	*Main content*
Disease	dsyn, neop, mobd, cgab	StrongCausal	CAUSES, PRODUCES, COMPLICATES
Process	patf, phsf, phpr, menp, biof, genf, orgf, moft, ortf, celf	RiskCausal	PREDISPOSES
RiskFactor	bhvr, inbe, socb, acty, food, ocac; humn	Modulatory	AFFECTS, AUGMENTS, STIMULATES
Clinical Manifestation	sosy, fndg, lbtr, clna, inpo, anst	Mechanistic	PROCESS_OF, MANIFESTATION_OF
DrugChemical	phsu, clnd, horm, imft, antb, bacs; chvs, aapp, nnon, lipd, carb, enzy, nsba	WeakAssociation	ASSOCIATED_WITH, INTERACTS_WITH
		Intervention	TREATS

Building a high-fidelity medical causal graph requires a traceable data foundation and strict semantic specification. We conducted a systematic PubMed search using the query ((Diabetes Mellitus OR diabetes) AND comorbidit*). After excluding animal experiments, duplicates, and reviews, 57,731heterogeneous core publications were retained. To ensure reproducibility and physical traceability, the Secure Hash Algorithm 256-bit (SHA-256) hash of the PubMed Identifier (PMID) list was computed to generate a unique corpus snapshot identifier, Snapshot ID. Abstracts were mapped into a tamper-resistant global three-dimensional coordinate systems

$$C=\{(\text{pmid}, \text{sentid}, [\text{start}, \text{end}))\}. \quad (1)$$

Here, [start, end) denotes zero-based Unicode character offsets.

To reduce semantic heterogeneity and establish reasoning boundaries, a predefined causal ontology schema was constructed on the physical corpus as the blueprint for automated reasoning (Table I). At the entity level, Unified Medical Language System (UMLS) served as the normalization backbone to cluster five core categories. At the relation level, to compensate for the lack of strength differentiation in SemMedDB predicates, a six-category causal semantic system was predefined and used as the admission standard, yielding 75,185 qualified prior triples.

B. Dual-Channel Evidence-Aware Extraction

1) Dynamic Context and Structured Prompting

Given the prevalence of cross-sentence dependencies in medical texts, we adopted a sequential sentence-wise processing strategy rather than isolated sentence analysis. For each core sentence, a dynamic context window was iteratively constructed to support reasoning, forming a local semantic unit with clearly defined boundaries and complete contextual information. A structured prompting template was embedded into the LLM, integrating role settings, task specifications, schema hard constraints enforcing JSON output, and chain-

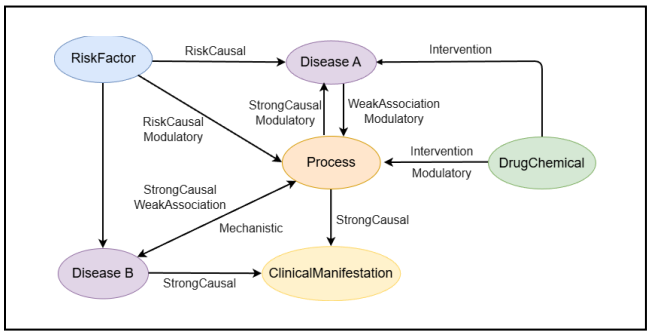

Figure 2. Predefined Topology Framework.

of-thought guidance. Under this design, dual-channel evidence-aware extraction w-
as performed based on contextual information density, as shown in Figure 1.

2) Dual-Channel Routing Mechanism

To optimize precision and recall, the task flow is routed by prior knowledge richness. For windows containing valid SemMedDB anchors, Channel A (prior verification mode) was activated. The LLM functioned as discriminator and generator, verifying prior relations while mining hidden associations among entity pairs satisfying the allowed type list and Top 20 proximity threshold. For anchor-absent knowledge blind areas, Channel B (constrained discovery mode) was applied, introducing three restrictions: limited entity type pairs, mandatory causal triggers, and a stricter Top 10 truncation threshold. Together, these constraints suppressed open-ended hallucinations while supplementing marginal knowledge.

3) Defense-in-Depth Auditing

All candidate triples underwent multidimensional defense-in-depth auditing. Span consistency checking first enforces exact character-level mapping within coordinate system C via inverse backtracking, physically blocking ungrounded hallucinations. Subsequently, assertion-state detection applies a deep coupling strategy to identify and remove negated and speculative expressions, ensuring atomic-level knowledge certainty.

4) Standardization and UMLS Alignment

To unify the semantic space and eliminate gaps between LLM outputs and prior knowledge, UMLS alignment was enforced, producing a semantically consistent high-fidelity causal graph.

C. Causal Graph Construction and Topological Storage

Given the directionality and temporal irreversibility of medical causality, a DAG was adopted to represent dependency pathways among diseases, risk factors, and interventions. Standardized triples aligned via UMLS were instantiated as nodes and directed edges in the initial network. A global topological sorting algorithm was then applied to detect and block cyclic dependencies violating medical temporal order, ensuring monotonic causal flow. The final graph follows the predefined topology in Figure 2, allowbreak

TABLE II. LAYERED PERFORMANCE EVALUATION OF ENTITY AND RELATION EXTRACTIONS

Evaluation tier	*Precision (%)*	*Recall (%)*	*Strict F1 (%)*
Entity Level (NER)	96.73	89.41	92.93
Relation Level (RE)	82.41	93.52	87.61
Strict Triple	79.11	89.15	83.83

constructing a logically self-consistent and unambiguous computable causal knowledge foundation.

IV. RESULTS AND DISCUSSION

To validate the effectiveness of the proposed hybrid paradigm and assess the quality of the constructed Diab-CKG, this section reports the evaluation methodology, extraction performance, hallucination audit results, and topological characteristics of the causal graph.

A. Evaluation Methodology

To overcome the limitations of static benchmarks and fragmented expert consensus in biomedical knowledge graph evaluation, this study adopts a dynamic auditing system based on the LLM-as-a-Judge framework, replacing conventional single-metric assessment. The flagship Qwen-Max model serves as the core auditor within a schema-constrained layered evaluation design. Assessment spans two dimensions: entity atomicity and causal topological logic. To quantify physical anchoring and noise resistance, the Entity Hallucination Rate (EHR) is defined as

$$EHR = 1 - Precision_{Entity}, \quad (2)$$

measuring the proportion of hallucinations untraceable to source text spans. At the logical level, a causal topology verification mechanism leverages LLM reasoning to detect causal inversion and ensure that the constructed DAG conforms to medical pathology. Performance is finally assessed under strict matching, where true positives require simultaneous verification of entity boundaries, semantic types, relation categories, and causal direction. The resulting Strict F1 serves as the decisive indicator of clinical usability.

B. Performance Analysis and Hallucination Audit

1) Overall Performance and Mechanism Validation

Deep auditing by Qwen-Max (Table II) shows that the dual-channel evidence-aware framework maintains high robustness from atomic anchoring to causal reasoning. Entity recognition achieves 96.73% precision and an F1 score of 92.93%, surpassing standard baselines. This improvement is attributed to atomic indexing, which precisely anchors medical term boundaries and reduces the Entity Hallucination Rate to 3.27%. In complex semantic reasoning, the model attains 87.61% F1 and 93.52% recall, reflecting strong sensitivity to implicit associations. The end-to-end strict F1 reaches 83.83%, confirming substantial clinical utility for automated high-fidelity graph construction.

2) Fine-grained Robustness

Fine-grained auditing (Table III) reveals differential adaptability across semantic spaces. Atomic indexing yields

TABLE III. FINE-GRAINED PERFORMANCE EVALUATION ACROSS PREDEFINED ENTITY AND RELATION TYPES

Schema Type	*Category*	*Precision*	*Recall*	*Strict F1*
Entity Types	Disease	98.25	89.84	93.85
	Process	98.46	88.89	93.43
	RiskFactor	95.56	75.44	84.31
	ClinicalManifestation	93.85	96.06	94.94
	DrugChemical	97.44	88.37	92.68
	Macro	96.71	87.72	91.84
Relation Types	StrongCausal	74.07	91.95	82.05
	RiskCausal	86.16	87.26	86.71
	Modulatory	78.69	88.89	83.48
	Mechanistic	92.31	85.71	88.89
	WeakAssociation	76.65	90.37	82.95
	Intervention	76.47	81.25	78.79
	Average (Macro)	80.73	87.57	83.81

98.46% precision for the Process category, whereas the 84.31% F1 for RiskFactor reflects boundary ambiguity in non-restrictive expressions. At the relation level, Mechanistic relation achieve 92.31% precision, ensuring accurate disease mechanism representation, while StrongCausal relations reach 91.95% recall, indicating a calibrated trade-off that captures major causal links with limited noise.

C. Topological Characteristics of Diab CKG

The complete Diab-CKG, derived from 57,731 diabetes publications, contains 15,573 entities and 272,140 directed causal edges, with an average degree of 17.5, exhibiting a den-

TABLE IV. STATISTICS AND DISTRIBUTION OF ENTITIES AND RELATIONS IN THE CONSTRUCTED DIAB-CKG

Category	*Type (Schema)*	*Count*
Nodes	DrugChemical	4326
	Process	1808
	Disease	4209
	ClinicalManifestation	4307
	RiskFactor	923
	Total Nodes（unique node）	15573
Edges	StrongCausal	17548
	Mechanistic	162558
	Modulatory	21851
	RiskCausal	27376
	Intervention	13264
	WeakAssociation	29543
	Total Edges	272140

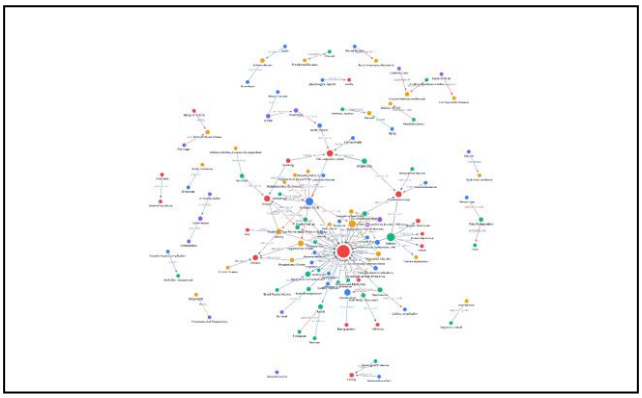

Figure 3. Example of Diab-CKG Local Causal Sub-graph.

se scale-free structure. Topological statistics (Table IV) show DrugChemical entities account for 27.8%, aligning with the clinical reliance on pharmacological intervention. StrongCausal and Mechanistic relations form the backbone, exceeding 66.2%, supporting inference of complication pathways, while RiskCausal and Modulatory relations enable modeling of complex biological feedback loops.

se scale-free structure. Topological statistics (Table IV) show DrugChemical entities account for 27.8%, aligning with the clinical reliance on pharmacological intervention. StrongCausal and Mechanistic relations form the backbone, exceeding 66.2%, supporting inference of complication pathways, while RiskCausal and Modulatory relations enable modeling of complex biological feedback loops.

Figure 3 illustrates a local comorbidity subgraph, visualizing cross-level causal cascades around core diseases and confirming that the graph captures both explicit associations and latent pathological evolution, providing a structured foundation for drug repositioning and comorbidity risk mitigation.

V. CONCLUSION AND FUTURE WORK

We address the challenge of converting dispersed biomedical narratives into computable causal evidence for diabetes comorbidity research. We introduce a hybrid framework combining physical anchoring, dual-channel evidence awareness, and topological reconstruction, constraining large language model reasoning within traceable evidence coordinates and a predefined causal ontology.

This evidence-bounded approach enables construction of a biologically faithful, logically consistent Diab-CKG, reducing hallucinations while supporting reliable end-to-end performance. Beyond extraction accuracy, Diab-CKG captures multi-stage causal cascades linking risk factors, pathological mechanisms, and therapeutic interventions, providing a computable foundation for mechanistic exploration and clinical knowledge discovery.

Limitations include reliance on abstracts, restricting temporal depth and phenotype granularity, and DAG enforcement, which may oversimplify feedback loops. Future work will integrate full-text literature and de-identified EHRs, and explore neuro-symbolic models combining graph neural networks with causal priors, advancing Diab-CKG toward

predictive and decision-support applications in precision medicine.

ACKNOWLEDGMENT

This research was supported by the National Key Research and Development Program of China (Project No.: 2023YFC2308803).

REFERENCES

[1] S. Wang, H.Y. Wang, and J. Du, "Discovering causal paths to diabetic nephropathy by combining computable biomedical knowledge with graph mining algorithms," AMIA Annu. Symp. Proc., pp. 1118-1124, Apr. 29, 2023.

[2] L. Huang, et al, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," ACM Transactions on Information Systems, pp. 1-55, Jan. 24, 2025.

[3] S. Sun, Q. Hu, F. Xu, F. Hu, Y. Wu and B. Wang, "Medical named entity recognition based on domain knowledge and position encoding," BMC Med. Inform. Decis. Mak., p. 235, Jul. 1, 2025, doi:10.1186/s12911-025-03037-0.

[4] R. Wagner, E. Kitzelmann and I. Boersch, "Mitigating Hallucination by Integrating Knowledge Graphs into LLM Inference – a Systematic Literature Review," Association for Computational Linguistics, pp. 795-805, Jul. 2025, doi: 10.18653/v1/2025.acl-srw.53.

[5] X. Jin, X. Wang, X. Luo, S. Huang and S. Gu, "Inter-sentence and Implicit Causality Extraction from Chinese Corpus, Advances in Knowledge Discovery and Data Mining, pp. 739–751.May 6 2020 doi: 10.1007/978-3-030-47426-3_57.

[6] J. Wang, et al, "Document-level biomedical relation extraction using graph convolutional network and multihead attention: Algorithm development and validation," JMIR Med. Inform., p. 17638 Jul. 31, 2020, doi: 10.2196/17638.

[7] Choi S, et al. Knowledge Graph Construction: Extraction, Learning, and Evaluation (2022–2024 Survey). Appl Sci. 2025;15(7):37.