Multi-Dimensional Knowledge Profiling with Large-Scale Literature Database and Hierarchical Retrieval

Anonymous authors
Paper under double-blind review

Abstract

The rapid expansion of research across machine learning, vision, and language has produced a volume of publications that is increasingly difficult to synthesize. Traditional bibliometric tools rely mainly on metadata and offer limited visibility into the semantic content of papers, making it hard to track how research themes evolve over time or how different areas influence one another. To obtain a clearer picture of recent developments, we compile a unified corpus of more than 100,000 papers from 22 major conferences between 2020 and 2025 and construct a multidimensional profiling pipeline to organize and analyze their textual content. By combining topic clustering, LLM-assisted parsing, and structured retrieval, we derive a comprehensive representation of research activity that supports the study of topic lifecycles, methodological transitions, dataset and model usage patterns, and institutional research directions. Our analysis highlights several notable shifts, including the growth of safety, multimodal reasoning, and agent-oriented studies, as well as the gradual stabilization of areas such as neural machine translation and graph-based methods. These findings provide an evidence-based view of how AI research is evolving and offer a resource for understanding broader trends and identifying emerging directions.

1 Introduction

The scale and diversity of contemporary AI research continue to grow at an extraordinary pace. Across computer vision, machine learning, natural language processing, and related areas, the past five years have seen rapid shifts in model architectures, training strategies, datasets, benchmarks, and application domains. For researchers, this expansion brings substantial challenges: it is increasingly difficult to situate individual works within broader developments, to track how research themes evolve, or to identify areas that are emerging, stabilizing, or declining.

Existing tools only partially address this need. Traditional bibliometric methods, which built on metadata, co-citation networks, and keyword statistics Zupic & Čater (2015); Donthu et al. (2021); Grootendorst (2022), provide high-level overviews but capture limited semantic information and treat topics as largely static entities. As a result, they offer only coarse insight into methodological transitions, cross-domain influences, or the finer-grained structure of research problems. Meanwhile, recent systems that incorporate large language models (LLMs) Brown et al. (2020); Achiam et al. (2023) demonstrate improved semantic analysis, supporting tasks such as retrieval-augmented question answering Gao et al. (2023) and automated survey generation Wang et al. (2024b). However, these tools are typically designed for short-range retrieval, single papers, or narrow tasks, and they do not provide a coherent, longitudinal view of large scientific corpora.

These gaps point to the need for a unified way to organize, summarize, and interpret the rapidly expanding body of AI literature. In this work, we construct a large-scale profiling pipeline aimed at characterizing the recent development of AI research. Using more than 100,000 papers from 22 major conferences published between 2020 and 2025 (Fig. 1), we combine text clustering, LLM-assisted semantic parsing, and lightweight retrieval techniques to form a structured representation of research problems, methods, datasets, and topical

dynamics. Rather than emphasizing algorithmic novelty, our focus is on creating a coherent analytic framework that enables researchers to explore and reason about the field at multiple levels of granularity.

Our study provides two complementary benefits. First, we derive a high-resolution view of topic lifecycles, dataset and model adoption patterns, and methodological transitions across areas such as vision, multimodal learning, foundation models, and generative modeling. Second, by incorporating structured retrieval and semantic filtering, we enable grounded, evidence-based queries that support practical research tasks, such as surveying subfields, tracing the evolution of techniques, or identifying emerging directions.

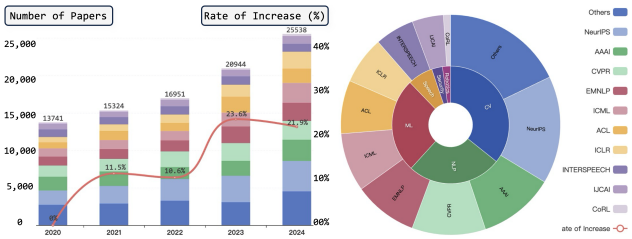


Figure 1: Number of papers published and topic statistics across 22 conferences from 2020 to 2025.

Through this analysis, we highlight several notable shifts in the AI landscape, including the consolidation of previously fast-moving areas, the rise of multimodal and agent-oriented research, and clear transitions in compute usage and model scaling practices. We expect the resulting knowledge database to serve as a resource for understanding broad trends, informing future meta-analyses, and supporting data-driven research planning.

Our work makes the following contributions:

- We construct a large-scale profiling pipeline over more than 100,000 papers from 22 major conferences, enabling structured analysis of semantic topics, methods, datasets, and research trajectories.
- We integrate clustering-based topic organization with LLM-assisted parsing and retrieval, producing a structured and interpretable representation of the AI research landscape that complements traditional bibliometrics.
- We conduct comprehensive empirical analyses, revealing trends in topic evolution, emerging subfields, methodological transitions, dataset and model dynamics, and institutional research patterns across the broader AI community.

Overall, these results provide an evidence-based view of how modern AI research is evolving and offer a foundation for transparent, large-scale, and semantically grounded scientometric analysis.

2 Related Work

2.1 Bibliometrics & Scientific Trend Analysis

Content mining and trend analysis of scientific literature are important topics in scientometrics and information science. Traditional bibliometric approaches rely on metadata, extracting abstracts, authors, journals, keywords, and citationsZupic & Čater (2015); Donthu et al. (2021), and applying co-citation and co-word analysesVan Eck & Waltman (2010); Chen (2006); Aria & Cuccurullo (2017) to reveal the structure and evolution of research fields. However, these methods mainly rely on surface features and are limited in capturing the latent semantic structures within documents. To address this, topic modeling and unsupervised clustering have been widely usedBlei (2012), with LDA inferring topic distributions from word co-occurrencesBlei et al. (2003); AlSumait et al. (2009); Chang et al. (2009); Chuang et al. (2013). More recently, embedding-based clustering methods, such as BERTopicGrootendorst (2022); Reimers & Gurevych (2019) combined with BERTDevlin et al. (2019), and applications of large language modelsKostikova et al. (2025); Diaz-Rodriguez (2025); Çelikten & Onan (2025); Lam et al. (2024) have enhanced the discovery of latent topics. Nevertheless, existing methods still struggle to systematically represent the dynamic evolution of knowledgeSi et al. (2024); Scherbakov et al. (2025), and a unified framework for multidimensional knowledge profiling remains lacking.

2.2 LLMs for Document Understanding

Trained on massive general-domain corpora, LLMs have shown strong abilities in semantic understanding and text generation Brown et al. (2020), giving rise to series of GPT Achiam et al. (2023) and other state-of-the-art models Xie et al. (2026); Comanici et al. (2025); Touvron et al. (2023); Yang et al. (2025). To support scientific applications, researchers have explored LLMs in research workflows. For example, PaperQA Lála et al. (2023); Skarlinski et al. (2024); Besrour et al. (2025) leverages retrieval-augmented generation (RAG) techniques Xue et al. (2025); Gao et al. (2023); Li et al. (2025); Cheng et al. (2025) to enable document question answering, while multi-agent and reinforcement learning approaches Nguyen et al. (2025); Singh et al. (2025); Li et al. (2025); Wu et al. (2025) further improve system accuracy. Another line of work focuses on automating literature review generation, including AutoSurvey Wang et al. (2024b), SurveyXLiang et al. (2025), SurveyForge Liang et al. (2025); Yan et al. (2025), and SurveyG Nguye et al. (2025), which implement pipelines for retrieval, filtering, organization, and writing. Despite these advances, existing LLM-based approaches are generally limited in scope: they focus on single tasks, rely on predefined corpora, and seldom provide a unified view of scientific knowledge evolution across multiple dimensions, including method evolution, task evolution, dataset adoption, and compute trends. Moreover, few approaches systematically integrate temporal information, semantic structure, and cross-conference knowledge, limiting their utility for large-scale trend analysis and evidence-based research decision making.

In this work, we propose a multidimensional profiling framework that leverages LLMs to extract and organize semantic information from a large corpus of scientific publications. By combining embeddings, topic clustering, and hierarchical retrieval, the framework produces a dynamic, hierarchical view of knowledge evolution, enabling trend analysis, cross-domain comparisons, and fine-grained topic investigation, and supporting data-driven exploration and research planning.

3 Knowledge Profiling Framework

Table 1: Detailed annual paper distribution for the 22 major venues in *ResearchDB* from 2020 to 2025.

Venue Group	2020	2021	2022	2023	2024	2025	Total
Core ML/Theory (NeurIPS, ICML, ICLR, UAI, COLT)	264	4703	5587	7317	9596	7269	34736
Computer Vision (CVPR, ICCV, ECCV)	1358	3272	3715	4509	4868	2871	20593
NLP/Speech (ACL, EMNLP, NAACL, COLM, INTERSPEECH, IWSLT)	3098	4138	4498	5580	6417	4730	28461
General/Applied AI (AAAI, IJCAI, CoRL)	2644	2837	2683	3071	4178	3239	18652
Systems/Security (MLSYS, OSDI, NDSS, etc.)	216	374	468	467	479	36	2040
Overall Total (22 Venues)	**7580**	**15324**	**16951**	**20944**	**25538**	**18145**	**104482**

We collected papers accepted between 2020 and 2025 from 22 academic conferences, including CVPR, ECCV, ICCV, ICLR, NeurIPS, and ICML, along with the citation counts for each paper, and over 100,000 papers are included in total, as shown in Tab. 1. Our main focus is on how to efficiently perform Multi-Dimensional Knowledge Profiling on this scientific data, as illustrated in Fig. 2.

3.1 Semantic Parsing & Knowledge Extraction

To systematically analyze the rapidly growing body of AI research literature, we propose a multi-stage knowledge extraction framework that combines automated content mining with LLM-driven semantic analysis. First, considering the difficulties in structured output and efficiency when directly using LLMs to analyze PDF files, we employ minerU Wang et al. (2024a) to convert the collected PDFs into structured Markdown files. Subsequently, based on the Markdown file corresponding to each paper, we utilize Deepseek-R1-32B Guo et al. (2025) to conduct multi-dimensional analysis of the papers, including meta information, abstract, research questions, methods, datasets, evaluation metrics, model architecture, contributions, limitations, and future directions.

It is worth noting that in multi-dimensional content parsing, the topic information of the papers is crucial. However, when LLMs process large scale structured data, on the one hand, due to the hallucination problem, LLMs struggle to stably output topic classifications, and on the other hand, the processing speed is limited.

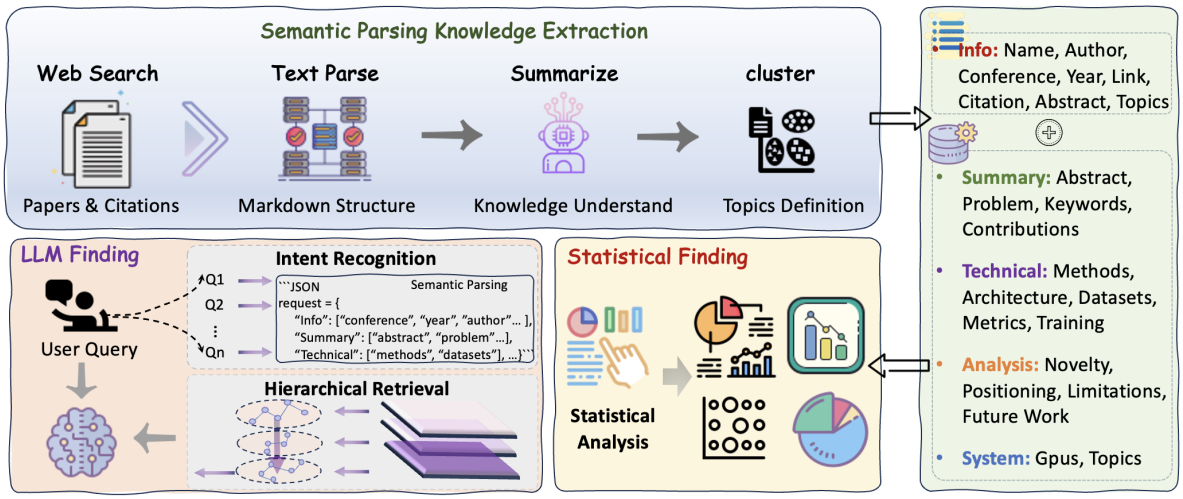


Figure 2: LLM-Driven Multi-Dimensional Knowledge Profiling Framework with Large-Scale Database and Hierarchical Retrieval.

To address this, we adopt a clustering method similar to BERTopicGrootendorst (2022) for topic clustering of papers. The specific process is as follows: first, use text encoder to process the title and abstract of each paper, then perform dimensionality reduction via UMAPMcInnes et al. (2018), and then use HDBSCANMcInnes et al. (2017) for clustering, resulting in more than 300 topic categories distinguished based on semantic similarity. Finally, we use ChatGPT-5 to generate topic summaries for each cluster, extract core topic names, and construct a hierarchical logical relationship of topics, forming a multi-dimensional knowledge database, *ResearchDB*, which can be used for subsequent analysis. The ResearchDB schema captures five main dimensions (Summary, Technical, Analysis, System, Post-Processed) extracted from full paper texts using LLM-assisted parsing. Tab. 2 provides the detailed documentation for the key fields in our structured data records. Unlike traditional bibliometric methods, this framework relies on semantic understanding rather than mere keywords, enabling the discovery of emerging or previously underexplored research topics, and finally obtaining the following information:

Table 2: Revised documentation for the key structural fields in the ResearchDB schema, reflecting the specific output of the LLM-assisted parsing pipeline.

Field Name	Data Type	Source	Purpose/Description
`paper_name/authors`	String/List	Metadata	Official title and author list.
`conference/year`	String/Int	Metadata	Publication venue and year.
`keywords`	List[String]	LLM-Parsed	Essential technical terms and concepts identified in the full text.
`keywords_description`	List[String]	LLM-Parsed	Essential technical terms and concepts identified in the full text.
`abstract_ori`	String	LLM-Parsed	Detailed summary covering the paper's context, method, and final results.
`abstract_summary`	String	LLM-Parsed	Detailed summary covering the paper's context, method, and final results.
`problem_statement`	String	LLM-Parsed	The key issue the paper attempts to solve.
`contributions`	List[String]	LLM-Parsed	Explicit statements of the paper's primary research contributions.
`methods`	String	LLM-Parsed	Core algorithmic approach/methodology used (e.g., PAGE, SGD).
`architecture`	String	LLM-Parsed	Specific model architectures referenced or introduced (e.g., LeNet, ResNet).
`loss_function`	String	LLM-Parsed	The specific loss function used for training (if mentioned).
`training_setup`	String	LLM-Parsed	Details on optimization, batch size, and general training configuration.
`datasets`	List[String]	LLM-Parsed	List of datasets used for experimental validation.
`metrics`	List[String]	LLM-Parsed	Key performance indicators reported (e.g., Test Accuracy, Training Loss).
`gpu_info`	String	LLM-Parsed	Specific hardware information referenced, used for resource analysis.
`limitations`	List[String]	LLM-Parsed	Identified constraints, weaknesses, or scope limitations of the proposed work.
`field_positioning`	String	LLM-Parsed	Categorization of the paper's innovation type.
`topic_name/ID`	String/Int	Post-Processed	Hierarchical topic label assigned via clustering and LLM refinement.

- **Info**: Extracts metadata including title, authors, conference, year, citations, topics, abstract, paper link, enabling macro-level analyses of research output, impact, and institutional distribution.
- **Summary**: Captures the core content, including research problem, keywords, abstract, and contributions, supporting the identification of emerging topics and cross-task trends.

- **Technical**: Extracts model architectures, datasets, evaluation metrics, and training strategies, allowing the study of methodological evolution and dataset usage patterns.
- **Analysis**: Summarizes novelty, positioning, limitations, and future work, facilitating the discovery of research gaps, bottlenecks, and potential directions.
- **System**: Records GPU, computational resources, enabling insights into the resource demands of different approaches and trends in large-scale model development.

3.2 Intent-Driven Knowledge Retrieval

Large language models have shown potential for literature retrieval and summarization, but direct application in research workflows faces key limitations: hallucination, omission or misinterpretation of complex information, and poor grounding in evidence, particularly over large, multi-year corpora. To address these challenges, we develop an intent-driven hierarchical retrieval pipeline over our structured *ResearchDB*, combining metadata filtering with weighted multi-field semantic search to provide reliable, evidence-based input to language models.

Intent Recognition To handle complex queries Q, we first decompose Q into simpler sub-questions $\{Q_i\}$ that isolate distinct information needs. Each sub-question is then processed using semantic parsing to generate structured retrieval instructions in JSON format, specifying relevant keywords, entities, and content types for subsequent retrieval. This decomposition reduces ambiguity, ensures comprehensive coverage, and provides clear guidance for the hierarchical retrieval stage.

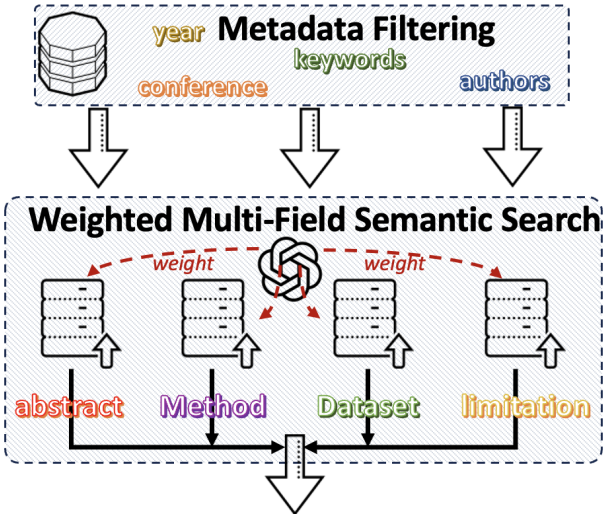


Figure 3: The Framework of Hierarchical Retrieval.

Hierarchical Retrieval Each sub-question Q_i is addressed through a two-level retrieval strategy, as shown in Fig. 3.

1. **Metadata Filtering.** We filter *ResearchDB* using metadata attributes such as conference, year, authors, and keywords to obtain a candidate subset. Among them, keywords use fuzzy matching, and other fields use exact matching.

$$D_{\text{filtered}} = \text{Filter}(ResearchDB, \text{metadata}). \tag{1}$$

2. **Weighted Multi-Field Semantic Search.**

   For the document subset D_{filtered} obtained via metadata filtering, we perform a subsequent intent-driven semantic search. Unlike traditional methods with static weights, our framework employs an adaptive weighting mechanism where the field weights w_j are dynamically assigned by the LLM-based intent recognition module according to the query's emphasis. For instance, if a sub-question Q_i focuses on specific technical implementations, the module allocates significantly higher priority to the 'methods' field (e.g., $w_{\text{methods}} = 0.8$) relative to the 'abstract' (e.g., $w_{\text{abstract}} = 0.2$). These weights are then applied to the semantic similarity scores s_{ij} calculated between the query and each core structural field—including abstract, methods, datasets, and limitations. The final weighted relevance score S_i is computed as:

$$S_i = \sum_{j=1}^{n} w_j \cdot s_{ij}, \quad \text{with} \quad \sum_{j=1}^{n} w_j = 1. \tag{2}$$

Top-k documents based on S_i are retrieved:

$$R_i = \text{TopK}(D_{\text{filtered}}, S_i). \tag{3}$$

The retrieved passages are compiled and passed to the language model to generate a coherent, evidence-grounded response. Weighting across fields allows the system to prioritize more informative content while retaining flexibility for diverse query types.

4 Scientific Insights and Findings

Leveraging our LLM-driven multidimensional profiling framework, we analyze more than 100K papers accepted by 20+ premier AI conferences from 2020 to 2025. This section presents a series of quantitative and semantic analyses across multiple dimensions of the research ecosystem, including topic evolution, compute and model scale, dataset dynamics, and institution-level research specialization. These results together reveal both the macro-level shifts and the fine-grained emerging directions shaping modern AI research.

4.1 Topic Growth and Transitions

To characterize the long-term evolution of AI research topics, we analyzed topic dynamics from 2020 to 2025 using LLM-based semantic clustering. We construct a four-quadrant topic lifecycle model, where the horizontal axis represents the Compound Annual Growth Rate (CAGR) of papers over the past two years, reflecting topic popularity trends:

$$\text{CAGR}_t = \left(\frac{N_t}{N_{t-2}}\right)^{\frac{1}{2}} - 1, \quad t = 2025. \tag{4}$$

The vertical axis represents the normalized mean publication year $\overline{Y}$ of papers, capturing research novelty:

$$\overline{Y}_i = \frac{1}{N_i} \sum_{p \in \text{topic } i} Y_p. \tag{5}$$

The bubble size indicates the total number of papers in the past two years N_i, and the color encodes a weighted impact metric combining citations C_i and topic count T_i:

$$\text{W}_i = \frac{\alpha C_i + \beta T_i}{\max_j(\alpha C_j + \beta T_j)}, \quad \alpha + \beta = 1. \tag{6}$$

α and β control the relative weights of citations and paper count, with $\alpha = 0.6$ in our experiments.

By analyzing the distribution of various topics in the four quadrants, as shown in Fig. 4, we observe that AI research is accelerating its evolution from being driven by early "model scale" and "basic perception tasks" to a paradigm driven by "safety and controllability, multimodal cognition, and intelligent agent systems". To further verify the overall trend of the life cycle analysis, we counted the changes in the number of papers on various topics over the past five years, and aslo selected the Top-20 topics with the largest number of papers for time series analysis, as shown in the lower left in Fig. 4.

The results show that the paper volume of LLM-related topics exhibits exponential growth, especially in reasoning, long-context modeling, and preference alignment, with each surpassing 200 papers per year by 2025. Multimodal integration tasks such as text-to-image editing, video diffusion, and vision-language QA maintain high activity but show slight stabilization after their 2023 to 2024 peaks. Meanwhile, classical areas including neural machine translation and GNN-based methods display relatively flat or declining trajectories, consistent with a maturing stage. Safety and robustness topics, though more recent, feature the steepest growth curves in both publication count and citation acceleration, marking the next frontier of attention.

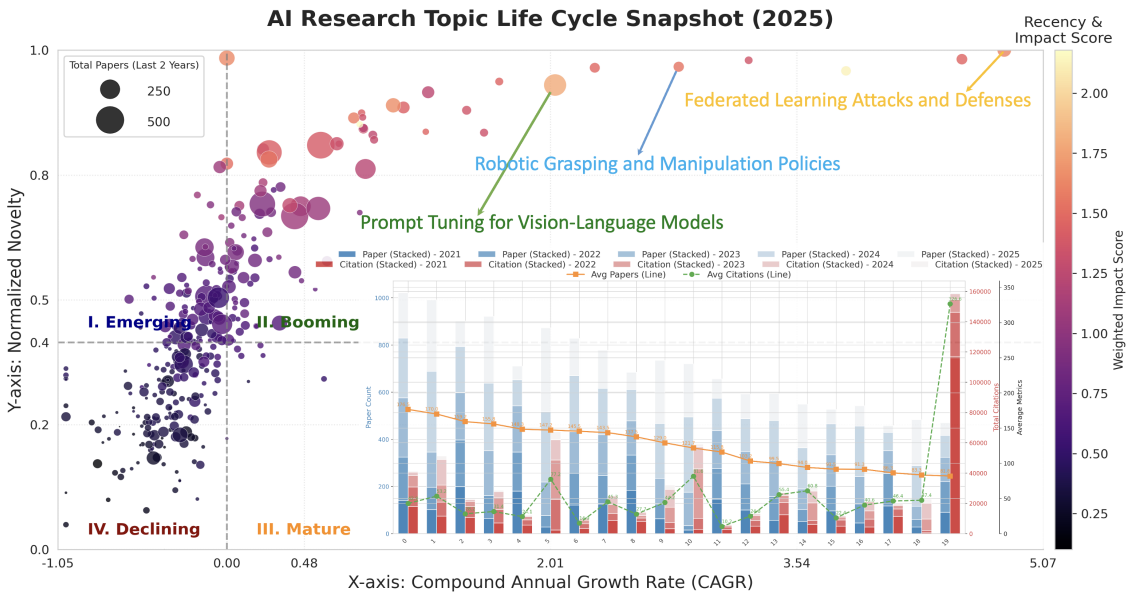


Figure 4: Four-quadrant topic lifecycle evolution (quadrants I-IV: Emerging, Booming, Mature, Declining). The inset (bottom left) shows the most prolific topic by paper count from 2020-2025.

Overall, from the dual perspectives of life cycle and quantity trends, the analysis reveals a clear paradigm transition in current research topics, from model architecture optimization toward the integration of perception, reasoning, and interaction capabilities. Future directions that deserve particular attention include:

- Enhancing long-context and reasoning capabilities of foundation models through efficient memory and adaptive attention mechanisms;
- Strengthening safety, reliability, and alignment frameworks to ensure trustworthy human-AI collaboration;
- Advancing multimodal understanding and generation toward dynamic, interactive, and grounded cognition;
- Bridging efficiency and scalability with sustainable model deployment, including low-rank adaptation, quantization, and sparse computation strategies.

These directions exhibit both high recent growth rates and strong Weighted Impact values, suggesting sustained momentum in the evolving research landscape.

4.2 Compute & Model Scale Analysis

With the rise of large models and the increase in complex tasks, the consumption of computing resources and the scale of model parameters have become important dimensions for measuring research trends. We analyzed the information on the use of computing resources in the statistical data and unified all computing power into A100 Equivalent Hours, as shown in Fig. 5. The analysis results show that the consumption of training resources has maintained a significant upward trend, especially in related research such as generative models, multimodal models, and intelligent agents. This trend is consistent with the current pursuit of model scale and complexity in the AI field, reflecting the reliance of higher-performance models on huge computing power.

Our research found that topics requiring high computing power investment have the following characteristics.

- **Processing complex data structures or large-scale data:** Research such as multimodality (combining images, text, speech, etc.) and large-scale pre-training involves massive amounts of

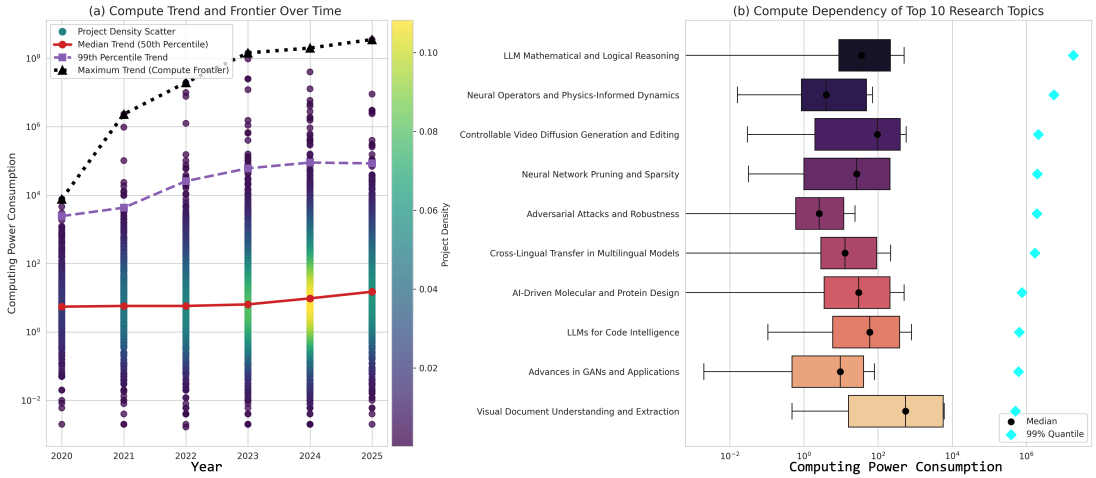


Figure 5: Compute Resource Analysis. Left: Temporal evolution of compute demand; Right: Boxplot of the top 10 topics with highest compute requirements.

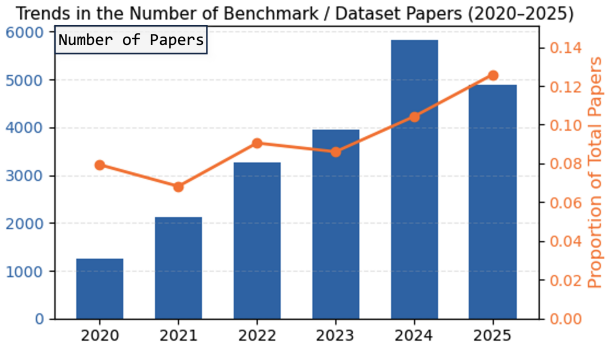
training data with complex interrelationships between data, requiring more computing resources for feature extraction and model fitting.

- **Enlargement and deepening of model architectures:** Represented by large language models (LLMs) and large generative models, the number of parameters often reaches tens of billions or hundreds of billions, which directly leads to an exponential growth in the demand for computing resources during the training and inference phases.
- **Complex learning paradigms or training methods:** For example, research involving reinforcement learning (RL), adversarial learning (such as GANs), or research on intelligent agents that require a large number of interactions with simulated environments. These methods usually require multiple iterations, a large number of samples, or complex optimization processes, thereby greatly increasing computing power consumption.

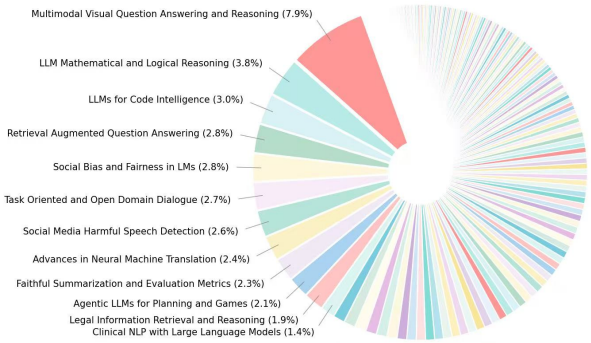
Further analysis of the proportion of computing power and paper output across different topics reveals that the investment in computing resources shows an uneven distribution across various research directions:

- **High-investment and high-output fields:** Concentrated in cutting-edge areas such as generative models and multimodality. Although they consume the majority of equivalent A100 hours, they also yield the highest number of papers and the most influential research results, forming a positive cycle where computing power drives innovation.
- **High-investment and medium-to-low-output fields:** Research in certain intelligent agent or specific application domains may require enormous computing power for experiments due to immature methods or the inherent difficulty of tasks (such as exploration, the complexity of environmental interactions, etc.). However, the short-term paper output is relatively mismatched with their computing power consumption, which may represent future potential directions or bottlenecks in technological breakthroughs.
- **Low-investment and high-efficiency fields:** Areas such as model compression, efficiency optimization, and interpretability focus on algorithms and efficiency rather than the absolute scale of models. These studies achieve a relatively high number of paper outputs with relatively low computing power investment, and are of great value for the promotion and practical application of existing models.

Overall, AI research is in a phase where its reliance on computing power continues to rise, mainly driven by large-scale models and complex tasks. Generative models, multimodality, and intelligent agents are the fields with the most intensive investment in computing power, indicating future research hotspots. At the same



(a) Dataset and Benchmark Trends (2020-2025). The bar chart shows annual paper counts for each dataset type, and the line chart shows their proportion of total publications.



(b) Topic distribution in Dataset/Benchmark papers. The most frequent topics are labeled.

Figure 6: The Analysis from the Dataset Landscape.

time, we have also observed that the academic community has achieved significant results in fields such as efficiency optimization and model simplification, which provide important support and alternative paths for promoting the development of AI in environments with limited resources.

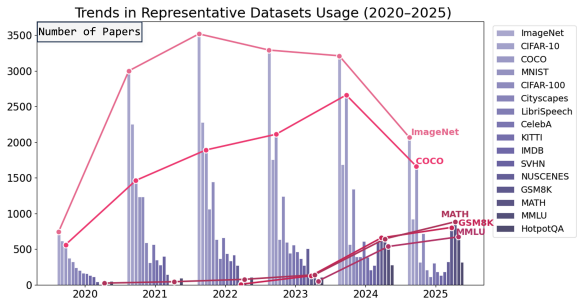
4.3 Dataset Landscape

Benchmark papers lay a crucial foundation for the standardization and comparability of technological progress by providing unified evaluation criteria and datasets. Statistics from the sample set show that the proportion of Benchmark papers is on a significant upward trend, with the growth rate accelerating particularly after 2022, as shown in Fig. 6a. Their proportion rose rapidly from 10.07% in 2022 to 18.16% in 2025 (partial data), approaching one-fifth of the total number of papers. This explosive growth is mainly due to the rise of general artificial intelligence technologies represented by LLMs, as well as the academic community's emphasis on the standardized evaluation of non-performance indicators such as model fairness, ethics, and robustness.

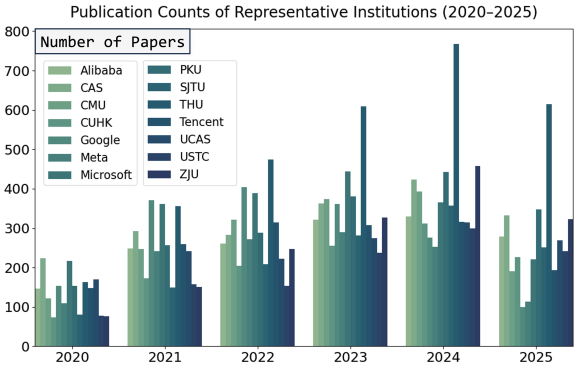
Further in-depth analysis of the themes of Benchmark articles reveals the evaluation challenges faced by current research hotspots and cutting-edge technologies. As shown in the statistical chart of theme count and proportion (Fig. 6b), research focuses are highly concentrated on LLMs, multimodality, and complex reasoning tasks. The top ten themes are almost entirely centered around the latest advancements and applications of LLMs, fully reflecting the evaluation-driven characteristics of contemporary scientific research.

Against the backdrop of this surge in benchmarks, the frequency of use of various datasets in scientific research papers has also shown rapid growth. This reflects the need for deeper evaluation in the era of LLMs and signifies that the paradigm for using datasets is no longer limited to single tasks but is evolving toward cross-domain and complex capability assessment, as shown in Fig. 7a

- **Relative saturation of classic visual benchmarks and cross-modal applications.** Although ImageNet has long been the most widely used classic visual dataset, its peak appeared in 2022, indicating the relative saturation of traditional deep learning visual benchmarks. On the other hand, datasets such as COCO, which initially focused on traditional visual tasks like object detection and instance segmentation, have also been widely used for cross-modal capability evaluation since the rise of Multimodal LLMs. This shows that research on large models is promoting cross-task and multi-purpose use of datasets, utilizing the rich annotated resources of classic datasets to evaluate new comprehensive capabilities.
- **Explosive growth of high-level reasoning datasets.** As LLMs mature in basic language capabilities, research focus has shifted to complex cognitive and reasoning tasks. The number of papers using the MATH dataset, designed for evaluating mathematical reasoning and arithmetic



(a) Trends in Representative Dataset Usage (2020-2025). The grouped bar chart shows annual paper counts per dataset, with overlaid lines highlighting trends for MMLU, GSM8K, ImageNet, COCO, and MATH.



(b) Publication Counts of Leading Institutions (2020-2025). Shows the number of papers from representative academic and industrial institutions across 22 conferences.

Figure 7: Trends in Publication Counts for Dataset Usage and Leading Institutions from 2020 to 2025.

abilities, as well as MMLU and GSM8K for high-level logical reasoning, has risen sharply since 2023. This trend strongly suggests that researchers are increasingly relying on these datasets to evaluate the performance of large models in complex reasoning tasks such as logic, common sense, and the integration of professional knowledge.

4.4 Institution-Level Research Patterns

Core institutions of China and the United States mainly maintained a dominant position across the period, as shown in Fig. 7b. Tsinghua University has long been at the top in terms of the number of publications, demonstrating the stability of its research output and its ability to continuously accumulate. However, there were annual competitive dynamic changes in the rankings between industrial giants and universities. In terms of the cooperation network, as shown in Fig. 8, domestic universities have active collaborations with industrial giants such as Alibaba, Tencent, and Huawei, showing obvious regional characteristics. At the same time, Chinese and American institutions still dominate the cooperation network, forming a "core-multipoint" cooperation pattern. This shift in research paradigms, driven by LLMs and Benchmark, is also profoundly reflected in the research strategies of different institutions. In the AI field, the academic community and the industrial sector have their own focuses and develop in a complementary manner. This is not simply a difference in interests, but rather reflects the strategic differentiation in the ecological niche of scientific research:

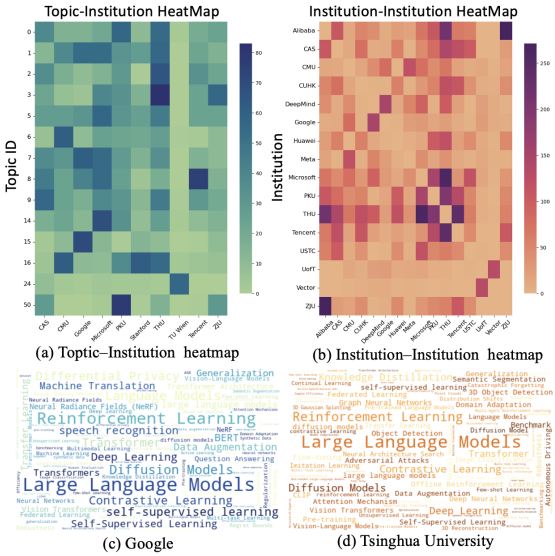


Figure 8: (a)–(b) Top topic-institution and institution-institution heatmaps; (c)–(d) Research topic word clouds for Google and Tsinghua University.

- **Academic institutions typically focus on basic research and cutting-edge algorithm exploration**, with their strategic priorities lying in efficiency improvement, robustness assurance, and continuous investigation of theoretical mechanisms, aiming to advance the democratization of AI technology. For instance, Tsinghua University emphasizes directions such as knowledge distillation, graph neural networks, adversarial training, domain adaptation, and model generalization, while

Carnegie Mellon University demonstrates strong performance in areas like robotic grasping and manipulation strategies and causal discovery.

- **Industrial institutions lean more toward application-oriented research and theoretical studies related to technology deployment.** Leveraging their advantages in data and computing resources, they strive to address issues such as system efficiency and real-time performance in large-scale model deployment (Scaling-up) and commercial applications, establishing a complete AI ecosystem and technological moat. For example, Microsoft maintains leadership in LLMs inference, code intelligence, mathematical reasoning, and retrieval-augmented generation, while Google focuses more on application and system-level topics, including speech recognition, machine translation, multi-task learning, and federated learning.

5 Retrieval and Analysis

To evaluate the effectiveness of our retrieval-augmented LLM framework for semantic understanding and literature summarization, we selected the Top-20 topics based on paper volume and diversity across 22 AI conferences from 2020 to 2025. For each topic, papers were processed using two methods: (i) direct generation with ChatGPT-5, and (ii) retrieval-augmented generation leveraging our *ResearchDB*. ChatGPT-5 is used as a representative baseline, while the retrieval-augmented approach can be applied to other LLMs and assessed in future work. Evaluation was performed by five doctoral students along five dimensions:

- **Accuracy**: Consistency of the generated content with the original papers;
- **Coverage**: Inclusion of core content and major sub-directions within the topic;
- **Novelty**: Ability to capture recently emerging directions or methods;
- **Readability**: Clarity and logical coherence of the text;
- **Usefulness**: Facilitation of rapid understanding and insight into potential research trends.

Each criterion was scored on a 5-point Likert scale independently by all annotators. Inter-annotator agreement was measured using Cohen's kappa to ensure consistency.

The evaluation results in Fig. 9 reveal that our proposed framework outperforms the baseline methods across several critical dimensions. Compared to direct generation by ChatGPT-5, our approach achieves significantly higher Accuracy and Usefulness. This improvement is primarily attributed to our intent-driven hierarchical retrieval, which grounds the model's responses in structured evidence extracted from specific content fields. While SurveyXLiang et al. (2025) focuses on general automation, our framework excels in capturing technical specifics through weighted multi-field search. This evidence-grounded approach prioritizes high-fidelity sections (e.g., methods, datasets), ensuring superior reliability and traceability over unconstrained LLM generation , which is essential for high-stakes research analysis

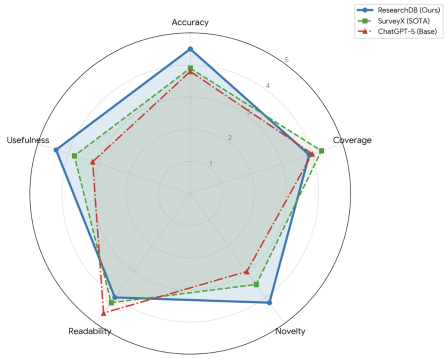


Figure 9: Performance comparison of *ResearchDB* against the SurveyX method and ChatGPT-5 across five evaluation dimensions, based on expert scoring.

5.1 Ablation Study

To evaluate the contribution of each core component in our hierarchical retrieval pipeline, we conduct an ablation study by comparing the full ResearchDB framework against three degraded variants consisting of a configuration without metadata filtering (w/o MF), a version without dynamic weighting (w/o DW) using a uniform weight distribution, and a single-field baseline (SF) that

Table 3: Ablation results on retrieval performance based on expert scores.

Variant	Acc.	Cov.	Nov.	Read.	Use.
w/o MF	4.12	3.85	4.00	4.10	3.90
w/o DW	3.86	3.82	3.70	4.05	3.75
SF	3.70	3.10	3.20	4.20	3.25
ResearchDB	**4.50**	**3.90**	**4.20**	**4.00**	**4.40**

restricts semantic search to abstracts only. As demonstrated by the experimental results in Tab. 3, the full framework consistently achieves the highest performance across all metrics, with the most significant drop in accuracy occurring when the intent-driven dynamic weighting mechanism is disabled. Specifically, the removal of dynamic weighting prevents the system from effectively prioritizing task-relevant fields; Furthermore, the exclusion of metadata filtering leads to a noticeable decrease in precision as the system fails to exclude papers from irrelevant venues or timeframes that share similar semantic embeddings. Finally, the inferior performance of the single-field baseline confirms that relying solely on abstracts is insufficient for deep knowledge profiling, as essential information regarding datasets and limitations is often omitted or overly generalized in paper summaries, proving that the integration of structured multidimensional fields with adaptive weights is vital for high-fidelity literature analysis.

6 Conclusion

We present a multidimensional knowledge profiling framework that integrates topic clustering, LLM-based semantic parsing, and hierarchical retrieval to analyze over 100K AI publications. By capturing topic dynamics, dataset trends, and institutional patterns, the framework enables evidence-driven discovery of methodological shifts and technical paradigms. Our approach provides both a macro-level landscape overview and fine-grained, topic-level insights, effectively supporting systematic trend analysis and research planning.

Limitations and Impact. While our study is limited to major English-language AI conferences and subject to potential LLM parsing biases, it offers a transparent, evidence-based perspective on research evolution. Future work will aim to improve domain-specific parsing accuracy and extend the framework to a broader range of scientific venues.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. Topic significance ranking of lda generative models. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 67–82. Springer, 2009.

Massimo Aria and Corrado Cuccurullo. bibliometrix: An r-tool for comprehensive science mapping analysis. *Journal of informetrics*, 11(4):959–975, 2017.

Ines Besrour, Jingbo He, Tobias Schreieder, and Michael Färber. Squai: Scientific question-answering with multi-agent retrieval-augmented generation. *arXiv preprint arXiv:2510.15682*, 2025.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine learning research*, 3(Jan):993–1022, 2003.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Tuğba Çelikten and Aytuğ Onan. Topic modeling through rank-based aggregation and llms: An approach for ai and human-generated scientific texts. *Knowledge-Based Systems*, 314:113219, 2025.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22, 2009.

Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, 57(3):359–377, 2006.

Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, et al. A survey on knowledge-oriented retrieval-augmented generation. *arXiv preprint arXiv:2503.10677*, 2025.

Jason Chuang, Sonal Gupta, Christopher Manning, and Jeffrey Heer. Topic model diagnostics: Assessing domain relevance via topical alignment. In *International conference on machine learning*, pp. 612–620. PMLR, 2013.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Jairo Diaz-Rodriguez. k-llmmeans: Summaries as centroids for interpretable and scalable llm-based text clustering. *arXiv e-prints*, pp. arXiv–2502, 2025.

Naveen Donthu, Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, and Weng Marc Lim. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of business research*, 133:285–296, 2021.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Aida Kostikova, Zhipin Wang, Deidamea Bajri, Ole Pütz, Benjamin Paaßen, and Steffen Eger. Lllms: A data-driven survey of evolving research on limitations of large language models. *arXiv preprint arXiv:2505.19240*, 2025.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–28, 2024.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, et al. Towards agentic rag with deep reasoning: A survey of rag-reasoning systems in llms. *arXiv preprint arXiv:2507.09477*, 2025.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, et al. Surveyx: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*, 2025.

Leland McInnes, John Healy, Steve Astels, et al. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Minh-Anh Nguye, Minh-Duc Nguyen, Nguyen Thi Ha Lan, Kieu Hai Dang, Nguyen Tien Dong, and Le Duy Dung. Surveyg: A multi-agent llm framework with hierarchical citation graph for automated survey generation. *arXiv preprint arXiv:2510.07733*, 2025.

Thang Nguyen, Peter Chin, and Yu-Wing Tai. Ma-rag: Multi-agent retrieval-augmented generation via collaborative chain-of-thought reasoning. *arXiv preprint arXiv:2505.20096*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A Lenert. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association*, 32(6):1071–1086, 2025.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*, 2025.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Nees Van Eck and Ludo Waltman. Software survey: Vosviewer, a computer program for bibliometric mapping. *scientometrics*, 84(2):523–538, 2010.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024a.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, et al. Autosurvey: Large language models can automatically write surveys. *Advances in neural information processing systems*, 37:115119–115145, 2024b.

Junlin Wu, Xianrui Zhong, Jiashuo Sun, Bolian Li, Bowen Jin, Jiawei Han, and Qingkai Zeng. Structure-r1: Dynamically leveraging structural knowledge in llm reasoning through reinforcement learning. *arXiv preprint arXiv:2510.15191*, 2025.

Xurong Xie, Zhucun Xue, Jiafu Wu, Jian Li, Yabiao Wang, Xiaobin Hu, Yong Liu, and Jiangning Zhang. Llm-oriented token-adaptive knowledge distillation. In *AAAI*, 2026.

Zhucun Xue, Jiangning Zhang, Xurong Xie, Yuxuan Cai, Yong Liu, Xiangtai Li, and Dacheng Tao. Adavideorag: Omni-contextual adaptive retrieval-augmented efficient long video understanding. In *NeurIPS*, 2025.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Lei Bai, and Bo Zhang. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12444–12465, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Ivan Zupic and Tomaž Čater. Bibliometric methods in management and organization. *Organizational research methods*, 18(3):429–472, 2015.

Appendix

Overview

The appendix presents the following sections to strengthen the main manuscript:

— We provide more details in the Supplementary Materials for a better understanding of our work.

— **Sec.** A. *ResearchDB* Source .

— **Sec.** B. Implementation and Reproducibility.

— **Sec.** C. Topic Clustering Results.

— **Sec.** D. Extraction Accuracy Validation.

— **Sec.** E. Retrieval System Evaluation.

A *ResearchDB* Source

This appendix details the source, scale, and temporal distribution of the underlying database, as well as the multidimensional data schema used to construct the *ResearchDB* knowledge graph.

Database Source and Distribution. The *ResearchDB* comprises over 100,000 papers from 22 top-tier conferences and journals published between 2020 and 2025. This broad selection ensures comprehensive coverage across key AI and Computer Science sub-domains, as shown in Tab. A1.

Table A1: The 22 major venues in the ResearchDB corpus, categorized by domain, including their abbreviations and total paper count (2020-2025).

Domain	Venue Name	Abbr.	Total Papers
Computer Vision	Conference on Computer Vision and Pattern Recognition	CVPR	11436
	International Conference on Computer Vision	ICCV	3768
	European Conference on Computer Vision	ECCV	5389
NLP & Speech	Annual Meeting of the Association for Computational Linguistics	ACL	10595
	Conference on Empirical Methods in NLP	EMNLP	8703
	North American Chapter of the ACL	NAACL	3625
	International Conf. on Comp. Linguistics (Modeling)	COLM	299
	International Speech Communication Association	INTERSPEECH	5351
Core ML & AI	Conference on Neural Information Processing Systems	NeurIPS	12726
	International Conference on Machine Learning	ICML	10199
	International Conference on Learning Representations	ICLR	9512
	Association for the Advancement of Artificial Intelligence	AAAI	13575
	International Joint Conference on Artificial Intelligence	IJCAI	4263
	Conference on Uncertainty in Artificial Intelligence	UAI	1245
	Conference on Robot Learning	CoRL	814
	International Workshop on Spoken Language Translation	IWSLT	187
	Conference on Learning Theory	COLT	755
Systems & Security	Machine Learning and Systems	MLSYS	220
	USENIX Symp. on Networked Systems Design and Implementation	OSDI	258
	Network and Distributed System Security Symposium	NDSS	409
	USENIX Conf. on File and Storage Technologies	USENIX-Fast	165
	USENIX Security Symposium	USENIX-Sec	988
		Overall Total	$\approx$ 104482

B Implementation and Reproducibility.

This section details the specific parameters and configurations for the two core components of our large-scale multidimensional knowledge profiling framework: LLM-assisted knowledge parsing and multidimensional topic clustering, ensuring full reproducibility of our methodology.

LLM Knowledge Parsing Prompts. We utilized the Deepseek-R1-32B model for semantic parsing of full paper texts to populate the *ResearchDB* schema (as documented in Sec. A). The parsing fidelity hinges on a carefully designed, JSON constrained prompt template. Below is the primary prompt template used for extracting core technical information (e.g., methods, architecture, training_setup).

```
You are a scientific paper analysis assistant.

Your task is to read the input paper content (in Markdown format), and output a structured summary in JSON format
     according to the schema below.

Output JSON schema:
{
  "abstract_summary": "",
  "keywords": [],
  "keywords_explanation": {},
  "methods": "",
  "architecture": "",
  "loss_function": "",
  "training_setup": "",
  "gpu_info": ""
  "datasets": [],
  "metrics": [],
  "datasets_metrics_mapping": {},
  "problem_statement": "",
  "contributions": [],
  "novelty_type": "",
  "experiments": "",
  "results_summary": "",
  "limitations": [],
  "future_work": [],
  "trend_insight": "",
  "field_positioning": ""
  "institution": []
}

Instructions:
1. **Keywords**: Extract 3-10 keywords directly from the paper text. Generate a short explanation for each keyword
     using LLM summarization.
2. **abstract_summary**: Provide a compressed version of the paper's abstract. Summarize, do not copy verbatim.
3. **Methods / Architecture / Loss / Training**: Summarize core method(s), model architecture, loss function, and
     training setup.
4. **Datasets / Metrics / Mapping**: List datasets, corresponding metrics, and mapping between datasets and metrics.
     Use empty lists/dictionaries if missing.
5. **Problem Statement**: Summarize the research problem in 1-2 sentences.
6. **Contributions**: Summarize the core contributions.
7. **Novelty Type**: Infer innovation type:
   - Algorithm / Model
   - Theory / Analysis
   - Benchmark / Dataset
   - Application / System
   - Methodological Improvement
8. **Experiments / Results Summary**: Summarize experiments and main results. Include ablation studies if present.
9. **Limitations / Future Work**: Summarize limitations and future research directions.
10. **Trend Insight**: Provide trend level insights based on the paper.
11. **Field Positioning**: Infer the paper's position in its research field:
    - Foundational Work
    - Methodological Innovation
    - Benchmark / Dataset Contribution
    - Application Validation
    - Trend Extension
12. **Institution**:Extract all author affiliations from the provided.
13. **Gpu_info**:Obtain the gpu resources used in the article, <total\_gpu>*<GPU_MODEL>*<training_time>.
```

LLM Intent Recognition Prompts. The core of our retrieval system's ability to effectively handle complex, multi-dimensional queries lies in an LLM-based intent recognition module. The goal of this module is to convert users' natural language queries into a format that strictly matches the data schema of the *ResearchDB* knowledge database.

```
You are a comprehensive query analysis and search planning assistant for a scientific literature knowledge base.

Your primary task is to parse the user's natural language query into a single, structured JSON object that captures
     all metadata constraints, technical requirements, and generates an optimized plan for vector search.
```

```
Output Schema:
{
  "conference": [],
  "year": [],
  "paper_name": [],
  "authors": [],
  "keywords": [],
  "keywords_explanation": {},
  "abstract_summary": "",
  "methods": "",
  "architecture": "",
  "loss_function": "",
  "datasets": [],
  "metrics": [],
  "vector_search_plan": []
}

Parsing Rules:
1. **Conference**: Must the range of the following fixed options: [AAAI, ACL, COLM, COLT, CoRL, CVPR, ECCV, EMNLP,
     ICCV, ICLR, ICML, IJCAI, INTERSPEECH, IWSLT, MLSYS, NAACL, NDSS, NeurIPS, OSDI, UAI, USENIX-Fast, USENIX-Sec].
     If the query does not mention a conference, leave as [].
2. **Year**: Extract the range of year contained in the query.
3. **Authors**: If the query explicitly mentions author names, extract them into a list. Otherwise, leave empty.
4. **Paper Name**: Extract all explicitly mentioned paper titles. Output as a list. Use partial matching if necessary.

5. **Keywords**: Extract only technical terms explicitly mentioned in query. Use `keywords_explanation` to briefly
     explain each.
6. **Abstract Summary**: One concise sentence describing technical focus from query.
6. **methods**:
   - Output the **complete method or approach description**, including its application domain if present.
   - Must describe both the **technique** and its **application scope**, not just a keyword.
   - If multiple distinct methods are present, include all as a list.
   - Do NOT shorten into keywords.
7. ** Architecture / Loss Function / Datasets / Metrics**: Extract explicitly if mentioned; do not hallucinate.
8. **Vector Search Plan**:
   - Decide which fields to include in the search (`abstract_summary`, `methods`, `keywords`, `datasets`, `metrics`, `
       architecture`, `loss_function`, etc.).
   - Only include fields that have non-empty content in the JSON output.
   - Assign a `weight` for each field between 0 and 1 based on **query emphasis**.
   - The weights **must be normalized such that their sum equals 1.0**.
   - Fields are **not mutually exclusive**; e.g., methods can also appear in keywords.
   - Weights should reflect relative importance of each field **according to the query intent**, not by default
       priority.
```

Dynamic Weighting Prompt. The field weights used in hierarchical retrieval are assigned by prompt-preset rules rather than learned parameters. The prompt first identifies the query intent and then assigns larger weights to the ResearchDB fields that are most relevant to that intent. The full prompt template is shown below.

```
You are a query-intent analyzer for a structured scientific literature database.

Given a user query and the extracted query fields, generate a vector search plan with preset field weights. The
     weights should reflect the query intent and must sum to 1.0.

Available searchable fields:
- abstract_summary
- keywords
- methods
- architecture
- training_setup
- loss_function
- datasets
- metrics
- gpu_info
- limitations
- future_work
- problem_statement
- contributions

Preset weighting rules:
1. Benchmark, dataset, or evaluation query:
   Prioritize datasets and metrics.
   Suggested weights:
   datasets 0.35, metrics 0.25, keywords 0.15,
   abstract_summary 0.15, methods 0.10.

2. Method, model, architecture, or algorithm evolution query:
   Prioritize methods and architecture.
   Suggested weights:
   methods 0.35, architecture 0.25, keywords 0.15,
   abstract_summary 0.15, metrics 0.10.

3. Compute, GPU, training cost, or scaling query:
```

```
   Prioritize gpu_info and training_setup.
   Suggested weights:
   gpu_info 0.40, training_setup 0.25, methods 0.15,
   datasets 0.10, metrics 0.10.

4. Limitation, future direction, or research-gap query:
   Prioritize limitations, future_work, and problem_statement.
   Suggested weights:
   limitations 0.35, future_work 0.20, problem_statement 0.20,
   methods 0.15, keywords 0.10.

5. Topic trend, lifecycle, or emerging-direction query:
   Prioritize keywords, abstract_summary, and methods.
   Suggested weights:
   keywords 0.25, abstract_summary 0.25, methods 0.20,
   datasets 0.15, metrics 0.15.

6. Institution, venue, author, or year-specific query:
   Apply metadata filtering first. Then assign semantic weights based on the remaining technical intent using Rules
        1-5.

Output JSON format:
{
  "intent_type": "",
  "metadata_filters": {
    "conference": [],
    "year": [],
    "paper_name": [],
    "authors": []
  },
  "vector_search_plan": [
    {
      "field": "",
      "query_text": "",
      "weight": 0.0,
      "reason": ""
    }
  ]
}

Rules:
- Only include fields with non-empty query content or fields needed by the inferred intent.
- Normalize all weights so that their sum equals 1.0.
- Do not assign high weight to a field only because it exists in the schema.
- If the query has multiple intents, mix the closest preset rules and renormalize.
```

ResearchDB Topic Clustering Configuration. The multidimensional topic structure of *ResearchDB* was generated through a robust embedding-based pipeline. We first transformed the paper abstracts and key textual segments into semantic vectors using the all-MiniLM-L6-v2 model. These high dimensional embeddings were then processed by UMAP for noise reduction and projection into a lower dimensional space, specifically configured with $n_components = 40$, $n_neighbors = 120$, and $min_dist = 0.08$ (with a fixed $random_state = 42$ for reproducibility). UMAP utilized the cosine metric, which aligns with the semantic nature of the embeddings. Clustering was performed using HDBSCAN, which is optimized for identifying clusters of varying densities; the algorithm was set with a significantly smaller $min_cluster_size = 50$ and $min_samples = 1$, employing the `eom` (Excess of Mass) cluster selection method. Topic representation was established using C-TF-IDF, configured to filter vocabulary using MAX_DF_PERCENT = 0.9 and a MIN_DF_ABS_FLOOR = 50, utilizing an n-gram range of (1, 2). The full hierarchical structure, derived from the HDBSCAN tree and refined by an external LLM (ChatGPT-5) for descriptive naming, resulted in the fine-grained multi-level taxonomy of over 300 topics presented in Sec. C.

C Topic Clustering Results.

This section presents the complete results of the fine-grained topic taxonomy for ResearchDB. By combining sentence-transformers embeddings, UMAP dimensionality reduction, and the HDBSCAN clustering algorithm, we successfully identified and named 324 independent, high-precision research topics from the corpus of over 100,000 papers. The HDBSCAN configuration ensures that each topic represents a sufficiently large yet highly homogenous set of papers, guaranteeing the fine granularity and specificity of the topics.

Hierarchical Topic Tree Structure. A key advantage of the HDBSCAN algorithm is its inherent ability to produce a Hierarchical Structure. We leverage this structure as an organizational framework, providing a non-strict view of clustering relationships above the 324 fine-grained topics. This tree structure is the basis

for our hierarchical retrieval system, allowing users to navigate from macro-level concepts (near the root) down to specific fine-grained topics (at the leaves), as shown in Fig. A1.

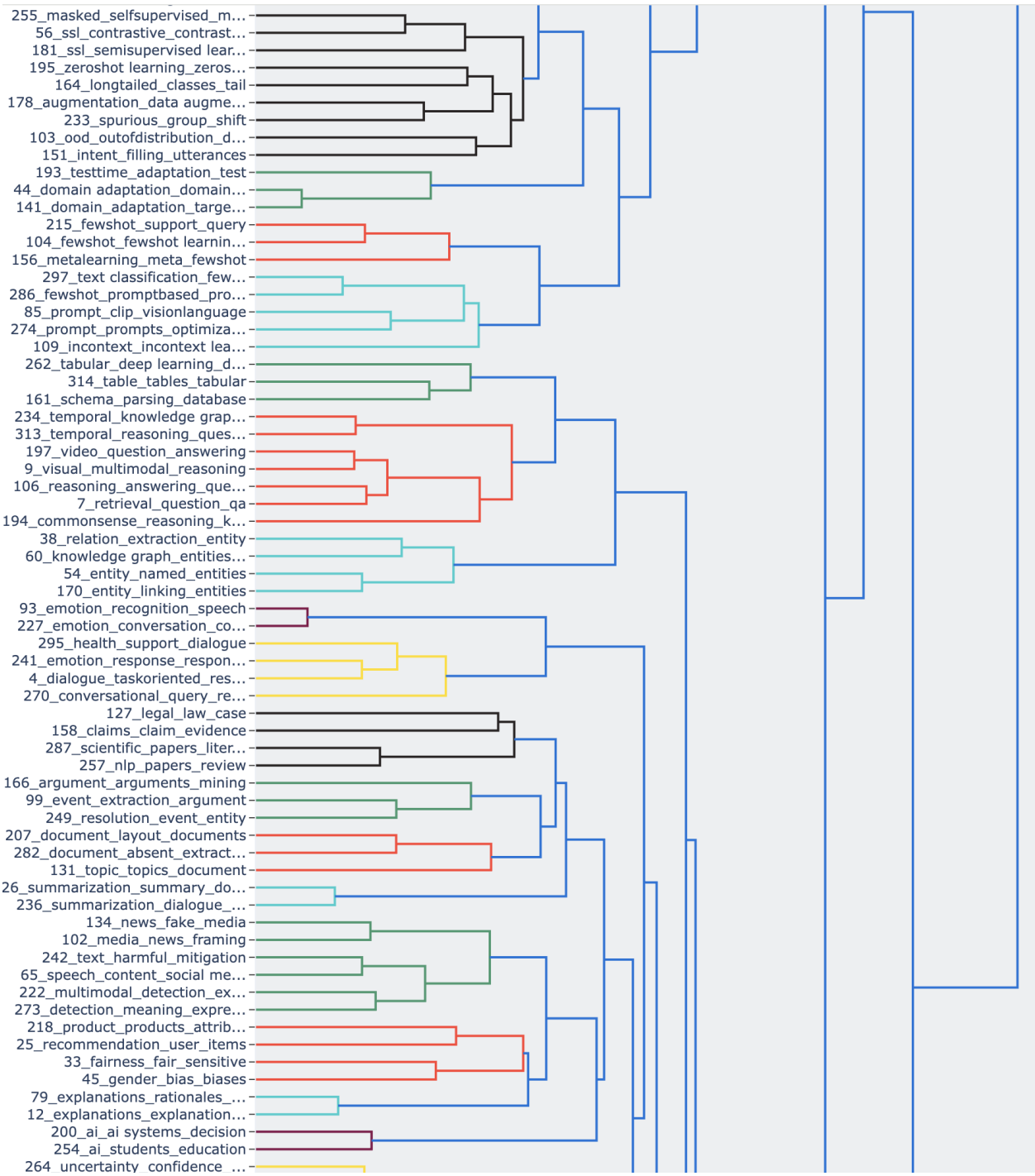


Figure A1: Hierarchical topic tree structure.

Topic Naming and Examples. The naming of each fine-grained topic underwent strict quality control: C-TF-IDF was used to extract representative keywords, which were then analyzed and refined by ChatGPT-5 based on paper abstracts within the cluster. This process generated highly semantically clear Topic Names (e.g., 140_Approximation and Fair Clustering Algorithms). Every paper in *ResearchDB* is uniquely tagged

Table A2: Twenty Sample Fine-Grained Topics from *ResearchDB*, Illustrating the Taxonomy's Scope and Detail.

ID	Topic Name	Summary
0	Graph Neural Networks and Representation Learning	Research advancing spectral and transformer-based GNNs, self-supervised and pretraining strategies, and scalable training to learn robust node and graph representations under heterophily, temporal dependencies, and distribution shifts.
1	Efficient Long-Context Attention and Memory	Research on extending and efficiently leveraging long contexts in transformers via improved positional encodings, attention approximations, adaptive computation, and memory augmentation, enabling length generalization and scalable training and inference.
2	Adversarial Attacks and Robustness	Research advances adversarial example generation and transferability across modalities and tasks while developing training, architectural, certification, and benchmarking methods to improve and assess deep models' robustness against digital, physical, and black-box attacks.
3	AI-Driven Molecular and Protein Design	Research leveraging structure- and sequence-informed machine learning, including graph and equivariant models, to predict properties, model interactions, and generate designs for molecules, proteins, and materials.
4	Task Oriented and Open Domain Dialogue	Research advances task-oriented and open-domain dialogue by combining end-to-end response generation with policy learning and state tracking, grounding on external knowledge and personas, handling complex goals, and developing robust evaluation.
5	LLM Mathematical and Logical Reasoning	Research advances methods and benchmarks to enhance and evaluate the mathematical and logical reasoning capabilities of large language models via chain-of-thought and structured prompting, neuro-symbolic and programming scaffolds, data/tool augmentation, self-correction, and reward modeling, and step-level and compositional evaluation.
6	Causal Discovery and Treatment Effects	Research develops methods to learn causal structure and estimate treatment and direct/indirect effects via interventions and counterfactuals, addressing hidden confounding, temporal dynamics, and identifiability in observational and experimental data.
7	Retrieval Augmented Question Answering	Research advances dense and generative retrieval and retrieval-augmented generation to enhance LLM question answering, addressing document identifiers, ranking and sampling, query rewriting and generation, knowledge selection, and rigorous evaluation for accurate and robust answers.
8	Advances in Neural Machine Translation	Research advances neural machine translation through improved architectures and decoding, multilingual and low-resource adaptation and transfer, robust evaluation and quality estimation, and human-in-the-loop and memory-assisted techniques.
9	Multimodal Visual Question Answering and Reasoning	Research advances vision-language models for multimodal VQA and reasoning, introducing datasets and benchmarks, alignment and retrieval strategies, and methods for compositional, chart, spatial, and knowledge-based tasks, and improving evaluation.
10	Text-to-Image Editing and Control	Research advances text-to-image diffusion with fine-grained controllability and personalization via improved architectures, guidance mechanisms, and language-vision integration, enabling precise editing, multi-concept synthesis, and preference-aligned generation.
11	Stochastic Bandits and Regret Bounds	Research develops algorithms and theoretical guarantees for stochastic, contextual, and linear bandits, including combinatorial, multi-agent, federated, and multi-objective variations, improving regret and sample complexity bounds under diverse feedback and constraints.
12	Explainable AI with Trees Concepts Counterfactuals	Research unifies decision tree modeling and optimization, prototype and concept-based self-explanation, Shapley/attribution and interaction metrics, and counterfactual and recourse frameworks to provide faithful, efficient, and user-relevant explainability.
13	Heterogeneity-Aware Personalized Federated Learning	This topic advances personalized FL by mitigating statistical and system heterogeneity through personalized models, heterogeneity-aware aggregation and alignment, and communication-efficient or asynchronous training, enhancing robustness, generalization, and scalability across clients and tasks.
14	LLMs for Code Intelligence	Research leverages LLMs with verification, search, and program analysis capabilities to generate, fix, translate, and retrieve code across languages, while benchmarking, evaluating robustness and bias, detecting hallucinations, and integrating developer feedback and tools.
15	Differential Privacy Algorithms and Utility Guarantees	Research advances differential privacy through novel mechanisms, optimization, and learning algorithms (e.g., adaptive clipping, shuffle amplification) to protect gradients and data in federated and centralized settings with formal guarantees and improved utility bounds.
16	Robotic Grasping and Manipulation Policies	Research develops learning-based policies for dexterous grasping and object manipulation, utilizing demonstrations, reinforcement learning, simulation, and generative models with multimodal sensing to achieve robust and generalizable control.
17	Temporal Action Recognition and Localization	Research on spatio-temporal modeling for recognizing and temporally localizing human actions in videos, emphasizing long-range dynamics, label-efficient learning (weak/semi/unsupervised), and robust representations against noise and backgrounds.
18	Agentic LLMs for Planning and Games	Research develops and evaluates LLM-based agents that plan, communicate, and act in games and interactive environments, integrating symbolic planning, multi-agent coordination, and experiential learning for robust decision-making.
19	Efficient Vision Transformers and Token Mixing	This topic advances efficient vision backbones by redesigning self-attention and token processing (via token pruning/merging, alternative mixers (linear/FFT), and state space models) to preserve global context while reducing computation and memory for classification, detection, and segmentation.

with such a topic name. Tab. A3 below presents a sample of the most representative fine-grained topics. We also have selected topics with ids ranging from 1 to 20 and listed them as Tab. A2

Table A3: Sample list of fine-grained topics extracted from *ResearchDB*, illustrating the specificity and coverage of the taxonomy.

Domain	ID	Topic Name	Representative Keywords
Optimization Theory	140	Approximation and Fair Clustering Algorithms	correlation clustering, approximation ratio, ℓ_p norm
Computer Vision	088	Efficient ViT Training and Pruning	vision transformer, token pruning, dynamic exit
Natural Language Processing	215	Controllable Text Generation via Latent Space	VAE, latent variable, attribute control, language model
Core ML	042	Causal Inference for Robustness and Generalization	do-calculus, invariant prediction, domain generalization
Systems & Security	291	Hardware Acceleration for Neural Network Inference	FPGA, ASIC, low-bit quantization, hardware optimization

D Extraction Accuracy Validation.

This section validates the reliability of the structured fields in *ResearchDB*. Since different fields have different levels of ground-truth availability, we use two complementary checks: automatic verification for metadata fields and manual evaluation for semantic fields.

Metadata Verification. For fields with explicit ground truth, including paper title, authors, venue, and year, we compare the extracted records against the original venue metadata after basic normalization, such as lowercasing, whitespace normalization, and punctuation cleanup. The exact-match accuracy is reported in Tab. A4. This check mainly verifies whether the data alignment and extraction pipeline preserve the original bibliographic information.

Table A4: Automatic validation of metadata fields against original venue records.

Field	Ground Truth Source	Metric	Accuracy
Paper Title	Venue metadata	Exact Match	91.3%
Authors	Venue metadata	Exact Match	95.2%
Conference	Venue metadata	Exact Match	100.0%
Year	Venue metadata	Exact Match	100.0%

Manual Evaluation of Semantic Fields. For fields that do not have a single fixed answer, including keywords, methods, datasets, metrics, and GPU information, exact matching is not appropriate. We therefore manually evaluate 200 randomly sampled papers from different venues, years, and research areas. For each paper-field pair, annotators assign a score of 1 if the extracted content is correct and usable, 0.5 if it is partially correct or incomplete, and 0 if it is unsupported or assigned to the wrong field. The detailed scoring criteria are shown in Tab. A5.

We report the average score for each field and criterion in Tab. A6. This evaluation focuses on whether the extracted semantic fields are faithful and useful for downstream analysis, rather than whether the LLM produces the only possible wording of a summary.

Table A5: Manual scoring rubric for semantic-field extraction. Each criterion is scored as 1, 0.5, or 0 for each sampled paper-field pair.

Criterion	Description
Factual Support	Whether the extracted item is directly stated in, or reasonably inferred from, the paper text.
Coverage	Whether the extraction includes the main items needed to represent the paper, rather than missing key methods, datasets, metrics, or resources.
Specificity	Whether the extracted content is concrete enough for downstream trend analysis, instead of being overly generic.
Field Consistency	Whether the item is placed in the correct schema field, such as datasets versus metrics or methods versus architecture.
Statistical Usability	Whether the extracted content can be used in aggregate statistics after normalization, especially for datasets, metrics, and GPU information.

Table A6: Manual validation results on 200 randomly sampled papers. Scores are averaged over sampled paper-field pairs using the 0/0.5/1 rubric.

Field	Factual Support	Coverage	Specificity	Field Consistency	Stat. Usability	Overall
Keywords	0.90	0.84	0.87	0.93	0.86	0.88
Methods	0.86	0.80	0.83	0.91	0.82	0.84
Datasets	0.93	0.88	0.91	0.96	0.90	0.92
Metrics	0.89	0.82	0.86	0.94	0.87	0.88
GPU Information	0.76	0.62	0.78	0.88	0.68	0.74

E Retrieval System Evaluation.

This appendix details the validation of the *ResearchDB* retrieval system. Since our goal is literature analysis and trend synthesis rather than single-hop question answering, a standard QA benchmark does not fully capture the intended use case. We therefore evaluate whether each system can produce a useful, evidence-aware analysis for complex research queries.

Evaluation Protocol. We construct 2,000 literature-analysis queries covering topic evolution, benchmark usage, methodological changes, and compute/resource trends. For each query, we compare three systems: direct ChatGPT-5 generation, SurveyX, and our *ResearchDB*-supported method. ChatGPT-5 answers directly from its parametric knowledge. SurveyX is used as a strong automated-survey baseline. *ResearchDB* first retrieves papers and structured fields from the database, then generates the final analysis conditioned on the retrieved evidence.

To avoid the bias and limited scale of small human-only evaluation, we use an LLM-as-judge protocol. For each query, we feed the judge model the query, the generated answer, and the retrieved or cited papers/statistics used as evidence. The judge then scores the answer on a 5-point scale along five dimensions: **Accuracy**, whether the answer is consistent with the source papers; **Coverage**, whether it covers the main sub-directions required by the query; **Novelty**, whether it captures recent or emerging directions; **Readability**, whether the answer is coherent and easy to follow; and **Usefulness**, whether it helps readers understand the research landscape. The averaged results are shown in Tab. A7.

Table A7: Retrieval and analysis evaluation on 2,000 literature-analysis queries. Scores are averaged using an LLM-as-judge protocol on a 5-point scale.

Method	Acc.	Cov.	Nov.	Read.	Use.
ChatGPT-5 Direct	3.42	3.68	3.21	**4.55**	3.36
SurveyX	3.86	**4.18**	3.63	4.21	3.92
ResearchDB	**4.27**	4.05	**4.11**	4.08	**4.31**

Evaluation Results. Overall, *ResearchDB* achieves the best scores in accuracy, novelty, and usefulness. This suggests that structured retrieval improves the factual grounding and practical value of generated literature analysis. SurveyX obtains the highest coverage score, reflecting its strength in broad survey-style organization. ChatGPT-5 direct generation remains strongest in readability, but it is weaker in accuracy, novelty, and usefulness because it lacks explicit access to recent retrieved papers and structured evidence. To illustrate this advantage, we conducted a qualitative assessment contrasting the *ResearchDB* output with a ChatGPT-5 baseline on a complex query: "How to achieve real-time, high-fidelity speech-to-gesture generation using decoupled diffusion models." The *ResearchDB*-supported answer (akin to an Academic Review) showed clear superiority in Fig. A2:

1. **Enhanced Academic Rigor and Traceability:** *ResearchDB* successfully retrieved seminal papers, allowing the final answer to cite up to nine top-tier conference papers (e.g., [1], [3], [19]), accurately attributing key methods like Retrieval Augmentation.
2. **Clearer Theoretical Framing:** The retrieved content enabled high-level theoretical decoupling of the problem (e.g., separating gesture drivers into semantic, rhythmic, and stylistic factors), resulting in a logically coherent survey argument.

This qualitative evidence confirms that *ResearchDB*'s multidimensional matching enhances the output's Academic Robustness, ensuring the retrieved information is both relevant and authoritative for writing top-tier research reviews.

Efficiency. To evaluate the scalability of ResearchDB, we benchmarked the knowledge profiling pipeline on a cluster equipped with 8 NVIDIA A100 (80GB) GPUs.

- **Knowledge Extraction Throughput:** By leveraging a distributed parallel processing framework with vLLM for inference acceleration, the system achieved a steady-state throughput of approximately 870 full papers per hour. The entire corpus of 104,482 papers was processed in 120 hours (5 days) of wall-clock time, totaling 960 GPU hours.
- **Data Sharing and Accessibility:** To facilitate future research and ensure reproducibility, we will release the fully structured ResearchDB, containing multidimensional profiles (e.g., methods, datasets, and limitations) for all 104,482 papers. This pre-processed knowledge base allows researchers to perform large-scale trend analysis and complex semantic queries without incurring the high computational costs of initial parsing.
- **Query Response:** For the end-user, the parallelized intent recognition module decomposes complex queries into sub-tasks that are executed concurrently, maintaining an average end-to-end response time of $\sim$1.2 seconds.

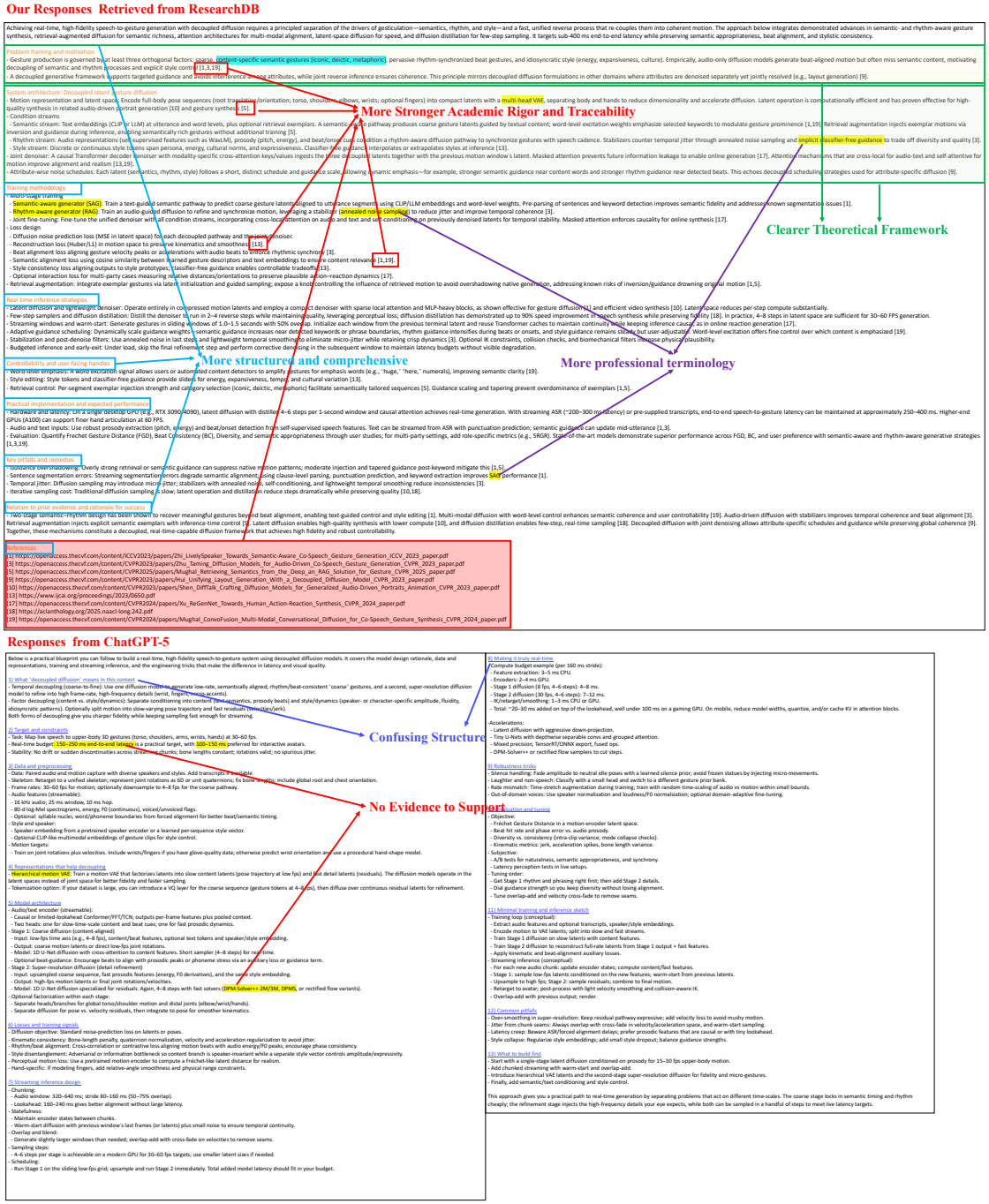


Figure A2: Qualitative comparison of answer quality between the *ResearchDB* system and the ChatGPT-5 baseline on the query: "How to achieve real-time, high-fidelity speech-to-gesture generation using decoupled diffusion models."