Multi-Disciplinary Dataset Discovery from Citation-Verified Literature Contexts

Zhiyin Tan
L3S Research Center
Leibniz University Hannover
Hannover, Germany
zhiyin.tan@l3s.de

Changxu Duan
Technische Universität Darmstadt
Darmstadt, Germany
duan@linglit.tu-darmstadt.de

***Abstract*—Identifying suitable datasets for research questions remains challenging because existing search engines depend on metadata quality and keyword overlap, which fail to capture semantic research intent. We introduce a scalable, literature-driven dataset discovery framework that uses scientific papers as semantic bridges between research questions and data resources. Our approach integrates (i) large-scale citation context extraction, (ii) domain-adaptive dataset recognition with large language models and structured output schemas, and (iii) deterministic, provenance-preserving entity resolution via strict normalization and consolidation. Unlike metadata-driven systems, our method retrieves citation-verified datasets, resources already used and validated by the research community. We evaluate our approach on eight computer science research questions derived from survey papers and six expert-provided questions from non-CS domains, each supported by hundreds of papers per query. Compared to Google Dataset Search and DataCite Commons, our system yields over sixfold higher recall while maintaining low redundancy and high source trustworthiness. A multi-disciplinary expert study further confirms the superior relevance, utility, and reliability of discovered datasets, especially where keyword-based methods fail. These findings establish citation-context mining as a generalizable paradigm for dataset discovery, enabling researchers across disciplines to locate not only accessible resources but also datasets validated through scholarly reuse. To support reproducibility and future extensions, we release our code, evaluation datasets, and results.** [1]

***Index Terms*—Dataset Discovery, Citation Context Mining, Scientific Literature Analysis, Cross-Disciplinary Research, Entity Resolution, Open Science**

I. Introduction

Researchers across disciplines increasingly rely on publicly available datasets to support empirical work, reproducibility efforts, and policy analysis [1]. The number of open datasets continues to grow rapidly. DataCite alone indexes over 30 million registered research objects, while Google Dataset Search covers millions more [2].

Yet discovering suitable datasets for a given research problem remains difficult. Existing dataset search engines are largely metadata-driven. Google Dataset Search matches queries against schema.org markup, while DataCite retrieves results from structured metadata fields [3]. These systems work well when researchers know exact dataset names or precise terminology, But they struggle when metadata is incomplete, inconsistent, or too generic to capture nuanced applicability. Prior studies confirm that metadata quality varies greatly across repositories, leading to unpredictable search outcomes [4].

In practice, researchers rarely begin with precise dataset names. Instead, they formulate research questions, often drawn from survey topics, prior studies, or emerging interdisciplinary problems, and seek datasets that can operationalize those questions. For such tasks, metadata keyword matching is insufficient. What is missing is contextual information: why a dataset was chosen, how it was used, and what research questions it helped answer.

We address this gap with a literature-driven dataset discovery framework. Our key insight is that scientific papers already contain rich contextual information about dataset usage. Citation contexts describe not only which dataset was used, but also how it supported the research design and why it was chosen. By treating these contexts as semantic bridges between research questions and datasets, we can move beyond static metadata and ground dataset discovery in real-world research practice. This perspective naturally aligns with how researchers themselves work: when seeking data, they first look for papers similar to their goals, and then examine which datasets those papers employed.

Methodologically, we adapt techniques from scientific information extraction, originally developed for tasks like named entity recognition in biomedical literature [5], [6], and repurpose them for multi-domain dataset discovery. This transfer allows us to circumvent the limitations of metadata dependency while anchoring recommendations in real-world usage patterns documented in scientific literature.

Zhiyin Tan and Changxu Duan contributed equally to this work.

Zhiyin Tan was funded by the "HybrInt - Hybrid Intelligence through Interpretable AI in Machine Perception and Interaction" project (Zukunft Nds, Niedersächsisches Ministerium für Wissenschaft, Grant ID: ZN4219). Changxu Duan was funded by the InsightsNet project (funded by the Federal Ministry of Education and Research (BMBF) under grant no. 01UG2130A). We gratefully acknowledge support from the hessian.AI Service Center (funded by the Federal Ministry of Research, Technology and Space, BMFTR, grant no. 16IS22091) and the hessian.AI Innovation Lab (funded by the Hessian Ministry for Digital Strategy and Innovation, grant no. S-DIW04/0013/003). We gratefully acknowledge the computing time granted by the Resource Allocation Board and provided on the supercomputer Emmy/Grete at NHR-Nord@Göttingen as part of the NHR infrastructure. The calculations for this research were conducted with computing resources under the project nhr_he_starter_25563.

[1] https://anonymous.4open.science/r/anonymization-B096

Our contributions are as follows:

- We introduce a literature-driven dataset discovery framework that bypasses metadata dependency by leveraging usage contexts documented in scientific papers.
- We develop a three-stage computational pipeline that integrates neural language models with citation graph analysis for scalable multi-domain dataset extraction.
- We establish a rigorous and scalable evaluation protocol. For the Natural Sciences, we develop an automated evaluation pipeline to measure performance systematically. To validate our framework's real-world utility across diverse disciplines, we supplement this with in-depth expert assessments in the remaining subject in the Field of Science and Technology (FOS).[2]
- We demonstrate that our approach outperforms leading dataset search engines, particularly in interdisciplinary scenarios, as confirmed by both automated quantitative metrics and expert preference ratings.

This work addresses a practical need, enabling researchers to efficiently identify relevant datasets in unfamiliar domains. At the same time, it contributes to the broader vision of intelligent scientific knowledge organization.

II. Related Work

Our work addresses dataset discovery through systematic analysis of scientific literature. We review three relevant areas: existing dataset search systems, literature-based information extraction, and citation context analysis.

A. Dataset Search Engines and Repositories

Existing dataset discovery systems include general-purpose search engines such as Google Dataset Search [2], which relies on schema.org markup, and DataCite [3], which ensures metadata quality through DOI registration but indexes only formally published datasets. General-purpose repositories such as Zenodo, Kaggle, and Mendeley Data accept contributions across disciplines, while domain-specific repositories such as the Protein Data Bank (PDB), Gene Expression Omnibus (GEO), the Inter-university Consortium for Political and Social Research (ICPSR), and the British National Corpus (BNC) provide curated, high-quality resources within particular communities but remain siloed, limiting interdisciplinary discovery.

These systems share fundamental limitations: heavy reliance on keyword matching against metadata, lack of semantic understanding of research contexts, and terminological mismatches across disciplines [4], [7]. Few systems address connecting abstract research questions with datasets when metadata is sparse or inconsistent.

B. Literature-Based Information Extraction

Scientific papers capture contextual details about resource usage that metadata alone can not [8]. Prior work has leveraged this signal to build dataset mention extraction benchmarks: SciERC [5] focused on computer science papers, SciREX [6] extended extraction to the document level, and DMDD [9] provided large-scale multi-domain coverage.

These efforts demonstrate the feasibility of mining dataset mentions at scale, but their primary goal is catalog construction or usage analysis rather than guiding researchers to datasets relevant to new questions [10]. Full-document extraction can also be computationally expensive, making real-time search and recommendation challenging [11], [12]. These methods are also limited by only being able to extract from papers in HTML/XML format, and also do not achieve multi-disciplinary extraction.

We build on this line of work by treating extracted mentions not as endpoints for cataloging but as semantic bridges linking research questions to datasets. In doing so, we leverage citation contexts to provide multi-domain applicability and enable efficient, literature-driven dataset discovery.

C. Citation Context Analysis

Citation context analysis reveals semantic relationships between resources and applications. Zhao et al. [13] developed frameworks for classifying citation functions, while Färber et al. [14] distinguished between actively used versus mentioned datasets. Recent advances show neural language models excel at scientific information extraction [12], and frameworks like SOFT [15] disentangle citation intent from content type.

However, existing approaches treat context as auxiliary information rather than the primary semantic signal for discovery. We extend these techniques by treating citation contexts as semantic bridges connecting research questions to datasets across disciplines, creating a generalizable framework that operates without domain-specific training by leveraging scientific literature as a comprehensive knowledge base of dataset usage patterns [1], [16].

III. Methodology

We formulate dataset discovery as *contextual information extraction* over scientific literature. The core idea is to use citation contexts as semantic evidence linking natural-language research questions to datasets. Our system is a three-stage pipeline that combines efficient corpus access with neural extraction and robust entity resolution: (i) scalable citation-context retrieval, (ii) neural dataset mention extraction with citation-aware quality signals, and (iii) datasets entity resolution for consolidation. This design directly addresses three chronic shortcomings of metadata-only systems: *semantic mismatch*, *contextual ambiguity*, and *incomplete coverage*.

A. Problem Formulation

Let a research query $\mathcal{Q}$ (free text) and optional field constraints $\mathcal{F}$ be given. The task is to retrieve and rank dataset entities $\mathcal{D} = \{d_1, \ldots, d_n\}$ from a literature corpus. In this work, we operationally define a dataset d as any named research resource (e.g., corpora, benchmarks, databases) that is explicitly mentioned in a paper as being used, modified,

[2]FOS: Natural Sciences, Engineering and Technology, Medical and Health Sciences, Agricultural Sciences, Social Sciences, and Humanities.

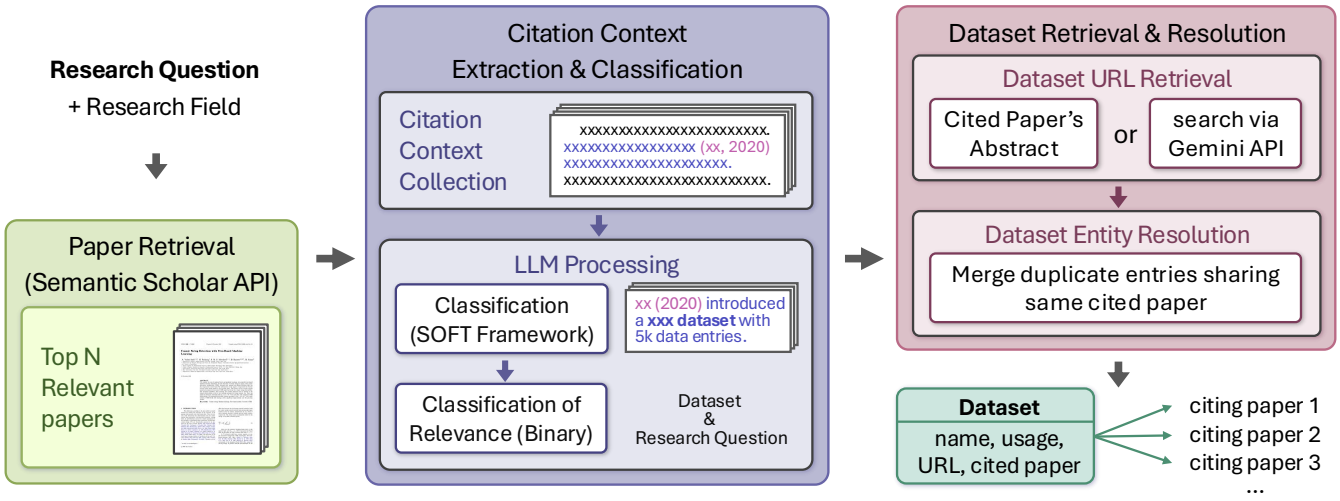

Fig. 1: The proposed pipeline maps a research question to relevant datasets via citation context analysis and metadata linking.

or evaluated against. This definition excludes mentions of software tools or general methodologies.

Each entity d_i is a structured tuple

$$d_i = \langle \text{name}_i,\ \mathcal{C}_i,\ \mathcal{M}_i \rangle, \tag{1}$$

where name_i is the canonical identifier, $\mathcal{C}_i$ is the set of supporting citation contexts, and $\mathcal{M}_i$ contains enriched metadata.

B. Three-Stage Pipeline

1) Stage 1: Scalable citation-context retrieval: We operate on Semantic Scholar Academic Graph (S2AG) [17] with a hybrid strategy: precomputed indices for citation windows plus on-demand filtering for query specialization. Given $(\mathcal{Q}, \mathcal{F})$, we retrieve candidate papers, extract sentence-level windows around citation markers, and retain $(\text{paper}, \text{target}, \text{window})$ triples with minimal metadata needed downstream. To maintain throughput over millions of papers while keeping latency manageable, we combine (i) prebuilt context tables and (ii) light query-time joins; we further reduce noise with an LLM-based relevance filter in a `retrieve-then-read` setup [12]. This balances accuracy with computational cost (full-document processing is expensive for real-time recommendation).

2) Stage 2: LLM-based dataset mention extraction with citation-aware quality: We cast extraction as Dataset Mention Extraction [18] over citation windows, augmented with citation-function cues. Building on scientific information retrieval [6], [19], we adapt Qwen2.5-72B-Instruct [20], a large instruction-tuned model, to identify dataset mentions, including abbreviated or implicit references. To output a structured record with (i) cited content type and (ii) citation intent (following SOFT's [15] separation of intent vs. content type). This structured extraction captures the specific application context of a dataset, such as whether it was primarily *used* as a resource, *modified* for a new task, or served as a standard to *evaluate against*. Such semantic insight provides a level of granularity that is fundamentally unavailable to metadata-driven search systems and allows researchers to filter resources by their function in prior research. We enforce a JSON schema for downstream consistency and apply a multi-tier validator (schema checks, semantic checks, domain rules) to the extraction quality. We also compute citation-enhanced signals. e.g., recency-weighted usage and context salience, to complement model confidence. A high-level procedure is given in Algorithm 1.

Algorithm 1 LLM-based Dataset Mention Extraction

Require: Citation windows $\mathcal{C}$, query/topic hints $\mathcal{T}$
Ensure: Structured extractions $\mathcal{E}$ with confidence and evidence

$\mathcal{E} \leftarrow \varnothing$
for $c \in \mathcal{C}$ **do**
  $y \leftarrow \text{LLMExtract}(c, \mathcal{T})$ {datasets, intent, content type}
  $z \leftarrow \text{Validate}(y)$ {schema/semantic/domain}
  $r \leftarrow \text{RelevanceFilter}(z, \mathcal{T})$ {second LLM pass}
  if r is valid **then**
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{\text{Record}(r.\text{items}, r.\text{confidence}, r.\text{evidence})\}$
  end if
end for
return $\mathcal{E}$

3) Stage 3: Dataset entity resolution and integration: We adopt a deterministic, evidence-preserving consolidation pipeline rather than graph clustering. First, mentions tied to the same bibliographic relation are locally consolidated by selecting the most reliable mention and unioning compatible attributes (e.g., roles, brief descriptions), while recording the number of merged mentions for diagnostics. Second, we perform a global aggregation by a strict canonical name: surface forms are normalized by trimming quotation/bracket marks, removing parenthetical qualifiers and generic type

Algorithm 2 Deterministic Entity Consolidation

Require: Mentions $\mathcal{E}$ (surface name, relation, attributes, provenance)
Ensure: Consolidated entities $\hat{\mathcal{E}}$

$\mathcal{E}' \leftarrow$ LocalConsolidateByRelation($\mathcal{E}$)
for $m \in \mathcal{E}'$ **do**
  m.key $\leftarrow$ Normalize(m.name)
end for
$G \leftarrow$ GroupByKey($\mathcal{E}'$)
for $g \in G$ **do**
  $display \leftarrow$ SelectDisplayName(g.names)
  $aliases \leftarrow$ Unique(g.names)
  $evidence \leftarrow$ CollectProvenance(g)
  $links \leftarrow$ PreferPID(CollectLinks(g))
  $\hat{\mathcal{E}} \leftarrow \hat{\mathcal{E}} \cup \{\text{Entity}(display, aliases, evidence, links)\}$
end for
return $\hat{\mathcal{E}}$

words (e.g., dataset/corpus/benchmark or train/test/validation markers), collapsing whitespace, stripping punctuation, and lowercasing. Mentions sharing the same canonical name are grouped; we retain all provenance, select a human-readable display name by frequency (tie-broken by capitalization), and keep all surface variants as aliases. References are rendered with persistent identifiers preferred over URLs. The procedure is summarized in Algorithm 2.

Putting the stages together yields an end-to-end flow from query to ranked datasets. This architecture scales by design (preindexed contexts + light query-time filtering), preserves accuracy (LLM-based extraction + citation-aware signals), and delivers reliable outputs (resolution + provenance).

C. Evaluation Framework and Quality Metrics

We evaluate along *extraction quality*, *evidence quality*, and *practical accessibility*, following FAIR [1], Information Retrieval evaluation [21], and data quality standards [22].

1) Extraction quality: Our evaluation employs a multi-perspective approach with three matching granularities (Exact, Norm, Fuzzy) inspired by entity resolution evaluation practices [23] and measures both extraction effectiveness and evidence quality following scientific data evaluation standards [24].

- *Exact Recall*: exact string match after whitespace normalization.
- *Norm Recall*: canonical normalization (lowercase, punctuation removal, whitespace consolidation).
- *Fuzzy Recall*: sequence-similarity clustering with threshold $\tau = 0.9$ [25].

The primary evaluation metric is *Norm Recall*, which provides the optimal balance between precision and robustness for practical applications, following best practices in scientific dataset evaluation [24]. We also report

$$\textbf{FuzzyGain} = \textit{Fuzzy Recall} - \textit{Norm Recall} \tag{2}$$

to quantify robustness to name variation.

2) Evidence quality:

- *Trusted Sources (%)*: share of matched entities whose evidence resolves to authoritative provenance (PID landing pages or trusted hosts such as official catalogs/registries or maintainer repositories).
- *With DOI (%)*: share of entities with a persistent identifier (DOI/HANDLE/ARK); reported as "With DOI (%)" in our tables.

3) Practical accessibility and efficiency: We measure URL availability (FAIR A1), and *redundancy* to gauge over-extraction:

$$\text{Redundancy} = \frac{|\text{Mentions}| - |\text{Entities Norm}|}{\max(1, |\text{Entities Norm}|)} \tag{3}$$

Gold standards are expert-curated, domain-specific reference sets with standardized protocols for consistency across areas; we compare system recall directly against these ground truths.

D. Reproducibility

All code, configs, and evaluation workbooks are released (Section IV); implementation particulars that affect replication are documented in Appendices A–C.

E. Pipeline Workflow Visualization

Figure 1 illustrates the complete workflow of our three-stage dataset discovery pipeline. The system begins with a research query and field constraints as input, progresses through citation context extraction, neural dataset extraction with citation-enhanced quality assessment, and entity resolution, ultimately producing a ranked dataset table. The optional LLM pre-filtering stage enables domain-specific relevance assessment to reduce noise in large-scale retrieval scenarios.

IV. Experiments

A. Experimental Setup

We evaluate our literature-driven dataset discovery framework under two complementary settings. The first is an automated, large-scale benchmark in computer science, where survey papers provide natural entry points for dataset queries. The second is a cross-disciplinary expert study, designed to test the practical utility of our approach beyond computer science. Together, these settings answer three research questions: (i) How does our pipeline compare to existing approaches in coverage and quality? (ii) What is the practical utility of retrieved datasets as judged by domain experts? (iii) How do individual components contribute to performance?

1) Survey Paper–Derived Queries (Computer Science Benchmarks): For large-scale automated evaluation, we focused on computer science domains where curated surveys reflect active research questions. We selected eight representative survey papers from top-tier venues in natural language processing (ACL, EMNLP, NAACL) and computer vision (CVPR, ICCV, TPAMI). Each survey title was used as a natural-language query, ensuring that our evaluation queries represent authentic, community-defined research problems rather than artificially constructed prompts. This process yielded the eight

benchmark tasks detailed alongside their results in Table II. The gold standard for each task was manually extracted from the corresponding survey paper to serve as our ground truth for the automated evaluation.

2) Expert-Provided Queries (Cross-Disciplinary Evaluation): To validate cross-domain applicability, we further evaluate on research questions sourced directly from experts in diverse fields. Following FOS taxonomy [26], we exclude our own field of computer science (Natural Sciences) and select research questions based on the remaining top-level categories: Engineering and Technology, Medical and Health Sciences, Agricultural Sciences, and Social Sciences and Humanities. Domain experts (Ph.D. holders, doctoral candidates, and advanced master's students under doctoral supervision) were asked to propose a question of genuine research interest within their discipline. These queries anchor our expert evaluation protocol (Section IV-C), ensuring that the assessment captures both the generality of our framework and its practical value in real-world cross-disciplinary discovery.

B. Baseline Systems

Google Dataset Search [2] is the most comprehensive web-scale dataset search engine, maintained by Google and indexing millions of datasets through schema.org markup. As the de facto standard for dataset discovery, it provides broad coverage across domains but relies heavily on the quality of metadata.

DataCite Commons [3] is the authoritative registry for research data DOIs, maintained by a global consortium of institutions and libraries. With over 30 million registered objects, it represents the gold standard for formally published datasets, though coverage is biased toward curated resources.

While baseline systems target metadata search, our model generalizes to dataset and application reasoning.

C. Expert Evaluation Protocol

Automated metrics cannot capture contextual appropriateness or real-world utility [7]. We therefore conduct a double-blind expert study [21], based on expert-provided queries from non-CS domains.

1) Participant Recruitment: We invited ten domain experts: three Ph.D. holders, four doctoral candidates, and three advanced master's students under doctoral supervision. Disciplines include Engineering and technology (Food and beverages), Medical and Health Sciences (Clinical medicine), Agricultural sciences (Agriculture, Forestry, and Fisheries), Social Sciences (Educational sciences), and Humanities (Arts). The comparative performance of each system on these expert-provided queries is visualized in Figure 2, providing a qualitative counterpart to our automated benchmarks.

2) Anonymization Strategy: To prevent bias, results from the three systems (our method, Google Dataset Search, and DataCite Commons) are presented anonymously as "System A", "System B", and 'System C". Experts are unaware of the underlying methodologies.

3) Evaluation Dimensions: Experts rated candidate datasets on six 5-point scales:

- *Relevance:* Alignment with the stated research question
- *Utility:* Likelihood of practical usage in research
- *Accessibility:* Clarity of descriptions and ease of access
- *Trustworthiness:* Confidence in reliability and source quality
- *Novelty:* Discovery of previously unknown datasets
- *Overall Satisfaction:* Holistic system assessment

4) Data Collection and Analysis: Responses were collected via anonymized Google Forms. Ratings were averaged within each expert–system pair and then aggregated across experts. Given the small and unbalanced number of raters per query, we did not conduct formal significance testing. We report per-dimension means and best-vote preferences across experts, which complement automated benchmarks with domain-grounded qualitative evidence. While sample sizes limit statistical power, results complement automated benchmarks by providing domain-grounded perspectives on dataset utility.

D. Results and Analysis

We present comprehensive experimental results comparing our neural dataset discovery approach against two established baseline systems: Google Dataset Search and DataCite Commons. Our evaluation encompasses eight diverse research domains, analyzing 1,358 total dataset mentions across multiple quality dimensions.

1) Overall Automated Performance: Table I shows the results of the automated evaluation of computer science. Our citation-context approach achieves substantially higher recall (16.22%) than Google Dataset Search (2.70%) and DataCite (0.00%), while maintaining high evidence quality and low redundancy. FuzzyGain reflects robustness measured by evaluation-time fuzzy matching. Table II provides normalized recall for each research question. Our method consistently outperforms baselines, with particularly strong gains in computer science and life sciences domains. The only tie occurs in "Personalized Text Generation," where all systems performed poorly, reflecting limited dataset reuse in that area.

2) Expert Evaluation Results: Experts rated datasets on six 5-point Likert dimensions: *Relevance*, *Utility*, *Accessibility*, *Trustworthiness*, *Novelty*, and *Overall Satisfaction*. Figure 2 visualizes the per research question ratings, while Table III presents aggregate results. Per-Research Query mean $\pm$ SD ratings are reported in the Appendix in Table IV. Experts unanimously favored our system in *Relevance* and *Utility*, confirming that context-based retrieval identifies datasets genuinely useful for research. Gains were also consistent in *Trustworthiness* and *Overall Satisfaction*, while *Novelty* showed competitive but less consistent improvements.

3) Quality and Trustworthiness Analysis: Beyond raw recall metrics, our evaluation reveals significant differences in the quality and trustworthiness of extracted datasets across systems.

TABLE I: Comparative automated performance across all computer science benchmark tasks. ↑ higher is better; ↓ lower is better.

Method	Total Entities	Recall (%) ↑	FuzzyGain (%) ↑	Trusted Sources (%) ↑	With DOI (%) ↑	Redundancy ↓	Evidence Quality
Ours	**1,330**	**16.22**	**1.71**	**14.23**	68.52	**0.26**	**High**
Google Dataset Search	79	2.70	0.00	1.63	32.81	0.29	Medium
DataCite Commons	67	0.00	0.00	0.00	**87.50**	7.76	**High**

TABLE II: Per-query performance breakdown on the computer science benchmarks. The table compares the normalized recall of our system against the baselines for each research question, with the number of gold-standard datasets provided for reference.

Research Question	Gold Datasets	Ours	Google	DataCite	Best	Advantage
Multi-modal Knowledge Graph Reasoning	11	**81.82%**	0.00%	0.00%	Ours	+81.82%
All-in-One Image Restoration	30	**66.67%**	0.00%	0.00%	Ours	+66.67%
Planning Capabilities of LLM	38	**55.26%**	0.00%	0.00%	Ours	+55.26%
Event-based Stereo Depth Estimation	17	**52.94%**	0.00%	0.00%	Ours	+52.94%
Patent Classification in NLP	7	**42.86%**	0.00%	0.00%	Ours	+42.86%
Document-level Event Extraction	23	**39.13%**	13.04%	0.00%	Ours	+26.09%
Text Line Segmentation for Historical Documents	43	**30.23%**	2.33%	0.00%	Ours	+27.90%
Personalized Text Generation	16	**6.25%**	**6.25%**	0.00%	Tie	0.00%
Average	23.1	**47.47%**	2.70%	0.00%	Ours	+44.77%

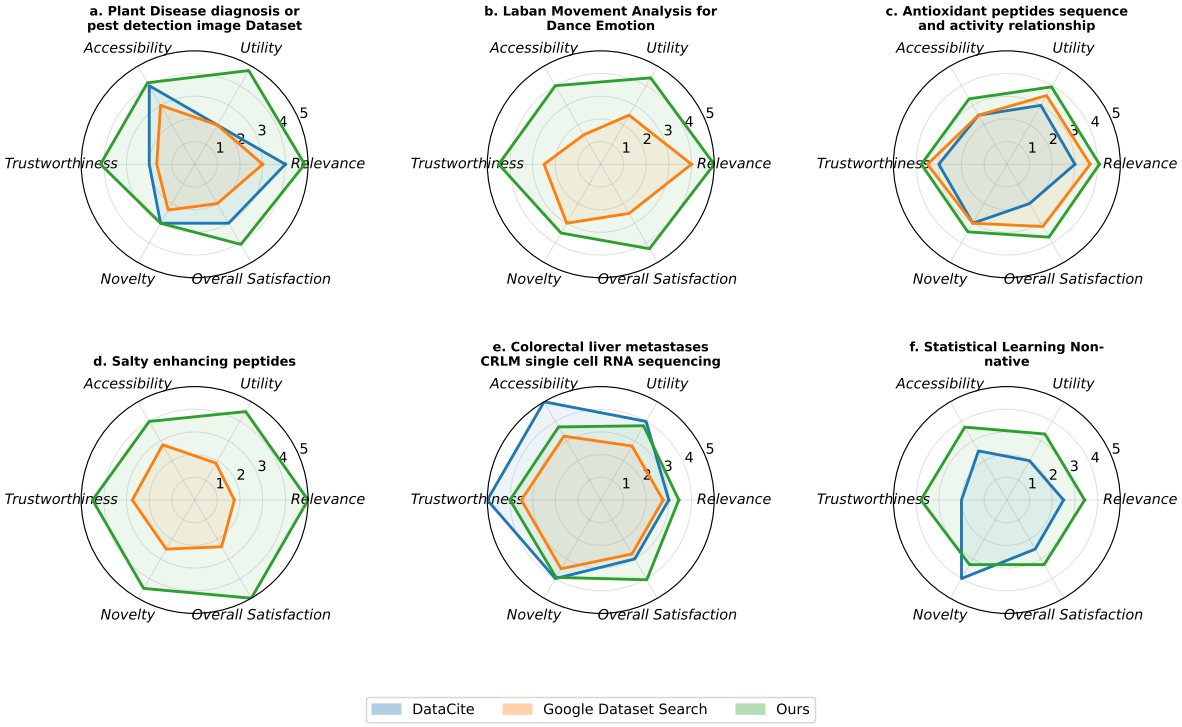

Fig. 2: Expert evaluation ratings across six cross-disciplinary research questions. Each radar chart visualizes the mean user ratings for a single query, comparing our system against the baselines across six dimensions of quality. The detailed numerical results are available in Table IV.

a) Evidence Quality: our system achieves 14.23% trusted-backed recall, significantly outperforming Google Dataset Search (1.63%) and DataCite (0.00%). This indicates that our neural extraction approach not only discovers more datasets but also identifies higher-quality resources with verifiable academic provenance.

b) Persistent Identifier Coverage: DataCite demonstrates the highest PID rate (87.50%) due to its focus on formally

TABLE III: Aggregate expert mean ratings (1–5).

Dimension	Ours	Google	DataCite
Relevance	**4.33**	3.07	2.60
Utility	**4.09**	2.46	2.66
Accessibility	**3.80**	2.87	3.17
Trustworthiness	**4.13**	2.49	2.84
Novelty	**3.64**	2.92	3.50
Overall	**4.07**	2.60	2.48

published datasets with DOIs. However, this comes at the cost of extremely low recall (0.00%), indicating that formal publication requirements severely limit dataset discovery scope. Our system balances PID coverage (68.52%) with substantially higher recall, identifying both formally published and emerging datasets.

c) Redundancy and Efficiency: our dataset entity resolution pipeline achieves the lowest redundancy rate (0.26), demonstrating effective duplicate detection and consolidation. DataCite shows concerning redundancy (7.76), suggesting potential indexing inconsistencies, while Google maintains reasonable efficiency (0.29).

4) System Scalability and Coverage: Our approach demonstrates superior scalability, extracting 1,330 unique dataset entities compared to 79 for Google and 67 for DataCite. This 16.8× improvement in coverage reflects our system's ability to identify datasets embedded within citation contexts rather than relying solely on explicit metadata markup or formal registration.

a) Domain Adaptability: performance varies significantly across domains, with the highest recall in technical computer science areas (Multi-modal Knowledge Graph Reasoning: 81.82%) and lower performance in interdisciplinary fields (Personalized Text Generation: 6.25%). This reflects fundamental differences in research data practices: CS domains have established dataset sharing cultures, making citation-context mining effective, while interdisciplinary fields emphasize novel data collection over reuse. This validates our dual evaluation approach, as high recall may not correlate with practical research value across all domains.

b) FuzzyGain Robustness: our system exhibits positive FuzzyGain (1.71%), indicating effective handling of dataset name variations through fuzzy matching. Baseline systems show zero FuzzyGain, suggesting limited robustness to naming inconsistencies common in scientific literature.

5) Summary: Across both automated and expert evaluations, grounding dataset discovery in *citation context* yields resources judged more relevant, useful, and trustworthy than metadata-driven baselines. These findings validate our core hypothesis: literature-based semantic bridges between research questions and datasets enable more effective and generalizable dataset discovery across domains.

E. Case Studies

1) Dataset Family Resolution: For research question: *Document-level Event Extraction*. Our system identified three ACE variants (*ACE*, *ACE 2005*, *ACE 2005 (zh)*) referencing the same catalog (`LDC2006T06`) but differing in scope and language. Traditional approaches either over-merge variants (losing semantic distinctions) or use paper DOIs as keys (causing entity misalignment).

Our solution employs family-level identifiers (catalog IDs, PIDs) as primary keys while preserving variant metadata as attributes. This enables accurate recall computation at the conceptual level while maintaining granular information, preventing recall inflation, and ensuring reproducible comparisons.

V. Conclusion and Discussion

This work introduced a literature-driven framework for dataset discovery that treats scientific publications as semantic bridges between research questions and data resources. The three-stage pipeline, comprising citation-context extraction, dataset recognition, and dataset-level entity resolution, demonstrates that grounding discovery in scholarly usage rather than metadata alone yields citation-verified datasets that are not only more relevant and credible but also richer in semantic context. Extracted directly from their textual and citation environments, these datasets convey signals about their purpose, methodology, and domain relevance, dimensions of meaning typically absent from metadata-based repositories.

A defining strength of the approach is that every discovered dataset has already undergone an implicit form of peer validation. By appearing in published research, each dataset has been described, evaluated, and applied within a scholarly workflow, reflecting its accessibility, usability, and reliability. This property renders the discovery process inherently quality-aware, reducing the burden of manual vetting. Moreover, by preserving the linguistic and conceptual context of dataset mentions, our framework exposes the reasoning behind dataset adoption, providing researchers with a more functional, context-rich understanding of the data landscape.

Empirically, the framework consistently outperforms metadata-dependent systems across a wide range of domains, establishing citation-context mining as a scalable and generalizable paradigm for literature-grounded dataset discovery. By transforming static bibliographic references into actionable semantic links between research problems, data resources, and their documented applications, the framework advances a new mode of context-aware scientific search.

a) Limitations and Challenges: while the framework achieves robust and meaningful discovery, several trade-offs naturally arise from its design. (i) Reliance on citation-rich literature favors datasets that have achieved visibility through prior publication and reuse. This ensures that identified datasets are credible and validated by the research community but may also limit coverage of newly released or niche datasets that have not yet been cited. Likewise, the temporal lag inherent in the publication cycle delays the inclusion of emerging datasets until they appear in subsequent works. (ii) Dependence on large-scale open-access corpora such as S2AG introduces structural bias toward disciplines with established open-data practices, while underrepresenting fields dominated

by paywalled publications. These are not flaws of the method but reflections of the broader data ecosystem. Future work integrating multi-source corpora and adaptive retrieval strategies can help mitigate these imbalances over time.

b) Ethical Considerations: reliance on English-language literature and citation-based indexing may inadvertently underrepresent datasets originating from non-English research communities or from disciplines where data citation is less formalized. These biases arise not from the methodology itself but from the composition of the accessible scholarly record. Addressing them requires expanding coverage to multilingual and regional corpora and adapting extraction models to recognize diverse conventions of dataset acknowledgment across fields. Such extensions will promote a more balanced and globally representative discovery process.

Broader Impact and Future Directions

The broader significance of this work lies in establishing citation-context mining as a general paradigm for literature-grounded resource discovery. The same principle can extend beyond datasets to other scholarly entities, including:

- Software and Tools: mapping how computational packages are adopted and reused across disciplines;
- Methods and Protocols: tracing how experimental procedures are adapted and combined in practice;
- Conceptual Frameworks: examining how theoretical constructs are operationalized or challenged in empirical research.

These extensions could populate a scholarly knowledge graph linking datasets, methods, software, and applications, forming a foundation for semantic recommendation, provenance tracking, and cross-domain reasoning. Integrating such graphs with large language models will enable natural-language querying of dataset–application relationships and support explainable, literature-grounded recommendations. In doing so, the framework can evolve from dataset discovery to a unified model of scholarly knowledge interconnection, supporting transparent, equitable, and reusable scientific ecosystems.

References

[1] M. D. Wilkinson *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, 2016.

[2] D. Brickley, M. Burgess, and N. Noy, "Google dataset search: Building a search engine for datasets in an open web ecosystem," in *The World Wide Web Conference*, 2019, p. 1365–1375.

[3] J. Brase, "Datacite - a global registration agency for research data," *SSRN Electronic Journal*, 2010.

[4] I. V. Pasquetto, B. M. Randles, and C. L. Borgman, "On the reuse of scientific data," *Data Science Journal*, vol. 16, p. 8, Mar. 2017. [Online]. Available: http://dx.doi.org/10.5334/dsj-2017-008

[5] Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi, "Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3219–3232. [Online]. Available: https://aclanthology.org/D18-1360/

[6] S. Jain, M. van Zuylen, H. Hajishirzi, and I. Beltagy, "SciREX: A challenge dataset for document-level information extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 7506–7516. [Online]. Available: https://aclanthology.org/2020.acl-main.670/

[7] C. L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*. The MIT Press, Jan. 2015.

[8] X. Chen, S. Jia, and Y. Xiang, "A review: Knowledge reasoning over knowledge graph," *Expert Systems with Applications*, vol. 141, p. 112948, Mar. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2019.112948

[9] H. Pan, Q. Zhang, E. Dragut, C. Caragea, and L. J. Latecki, "DMDD: A large-scale dataset for dataset mentions detection," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1132–1146, 2023. [Online]. Available: https://aclanthology.org/2023.tacl-1.64/

[10] J. Heddes, P. Meerdink, M. Pieters, and M. Marx, "The automatic detection of dataset names in scientific articles," *Data*, vol. 6, no. 8, p. 84, 2021.

[11] A. Xu, R. Ding, and L. Wang, "Chatpd: An llm-driven paper-dataset networking system," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, ser. KDD '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 5106–5116. [Online]. Available: https://doi.org/10.1145/3711896.3737202

[12] P. Marini, A. Santos, N. Contaxis, and J. Freire, "Data gatherer: LLM-powered dataset reference extraction from scientific literature," in *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, T. Ghosal, P. Mayr, A. Singh, A. Naik, G. Rehm, D. Freitag, D. Li, S. Schimmler, and A. De Waard, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 114–123. [Online]. Available: https://aclanthology.org/2025.sdp-1.10/

[13] H. Zhao, Z. Luo, C. Feng, A. Zheng, and X. Liu, "A context-based framework for modeling the role and function of on-line resource citations in scientific literature," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5206–5215. [Online]. Available: https://aclanthology.org/D19-1524/

[14] M. Färber, A. Albers, and F. Schüber, "Identifying used methods and datasets in scientific publications," 2021.

[15] C. Duan and Z. Tan, "Semantically orthogonal framework for citation classification: Disentangling intent and content," in *Linking Theory and Practice of Digital Libraries*, W.-T. Balke, K. Golub, Y. Manolopoulos, K. Stefanidis, and Z. Zhang, Eds. Cham: Springer Nature Switzerland, 2026, pp. 183–206.

[16] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld, "SPECTER: Document-level representation learning using citation-informed transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 2270–2282. [Online]. Available: https://aclanthology.org/2020.acl-main.207/

[17] A. I. for Artificial Intelligence, "The semantic scholar open data platform," 2025. [Online]. Available: https://arxiv.org/abs/2301.10140

[18] P. Datta, S. Datta, and D. Roy, "Raging against the literature: Llm-powered dataset mention extraction," in *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, ser. JCDL '24. New York, NY, USA: Association for Computing Machinery, 2025. [Online]. Available: https://doi.org/10.1145/3677389.3702523

[19] Q. Zhang, Z. Chen, H. Pan, C. Caragea, L. J. Latecki, and E. Dragut, "SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 13 083–13 100. [Online]. Available: https://aclanthology.org/2024.emnlp-main.726/

[20] Q. team, "Qwen2.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2412.15115

[21] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[22] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, p. 1–52, Jul. 2009. [Online]. Available: http://dx.doi.org/10.1145/1541880.1541883

[23] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Berlin Heidelberg, 2012. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-31164-2

[24] L. Koesten, E. Simperl, T. Blount, E. Kacprzak, and J. Tennison, "Everything you always wanted to know about a dataset: Studies in data summarisation," *International Journal of Human-Computer Studies*, vol. 135, p. 102367, Mar. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.ijhcs.2019.10.004

[25] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," in *Proceedings of the 2003 International Conference on Information Integration on the Web*, ser. IIWEB'03. AAAI Press, 2003, p. 73–78.

[26] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15. ACM, May 2015, p. 243–246. [Online]. Available: http://dx.doi.org/10.1145/2740908.2742839

[27] M. Raasveldt and H. Mühleisen, "DuckDB: an Embeddable Analytical Database," in *Proceedings of the 2019 International Conference on Management of Data*, 2019, p. 1981–1984.

APPENDIX

APPENDIX: DETAILED PIPELINE METHODOLOGY

This appendix details the end-to-end operational workflow of our literature-driven dataset discovery framework. The process is organized into three main phases: (A) Corpus Construction, (B) LLM-based Extraction and Filtering, (C) Entity Resolution and Metadata Enrichment.

A. Stage 1: Scalable Corpus Preprocessing and Indexing

This initial phase constructs a focused, query-relevant corpus of scientific text for analysis.

a) Seed Paper Retrieval: given a natural-language research question (RQ), we first query the Semantic Scholar API to retrieve a set of top N relevant seed papers (typically N=200 to 400). These papers form the initial, high-relevance core of our analysis corpus.

b) Citation Graph Expansion and Context Collection: using our pre-indexed local snapshot of the S2AG corpus, we perform a one-hop citation graph expansion around the seed papers. We collect all citation contexts where the seed papers are either citing another work or are being cited by another work. This expansion broadens the pool of relevant texts beyond the initial keyword match.

c) Implementation of Offline Indexing: to execute this phase at scale, our pipeline operates on a local S2AG snapshot. The raw JSONL.gz data is converted to the Parquet columnar format and indexed using an embedded DuckDB [27] instance configured on a high-memory server (128 GB RAM, 32-thread parallelism, 1TB SSD Storage). We pre-materialize textual contexts for each citation link, allowing for near-instantaneous retrieval. This architecture reduced our median query latency from tens of seconds with a pure API approach to approximately 0.1 seconds, which is critical for interactive analysis.

B. Stage 2: LLM-based Extraction and Relevance Filtering

This phase uses an LLM in a multi-step process to identify dataset mentions and assess their relevance to the initial research question.

a) Dataset Mention Extraction: each collected citation context is processed by our LLM (Qwen2.5-72B-Instruct) to identify mentions of datasets, benchmarks, or corpora. Using the SOFT classification framework, the prompt instructs the model to extract the dataset's name and its specific usage as described in the context.

Prompt 1: Dataset Mention Extraction

```
[System]
You are an expert scientific annotator. Extract
    dataset entities from the
given citation context. Rules:
- Output valid JSON only (no extra text).
- Ground findings strictly in the provided context
    ; do not hallucinate.
- Controlled vocabularies:
  usage_role ∈ {"Use","Modify","Evaluate Against"}
  content_type ∈ {"Performed Work","Discovery","
    Produced Resource"}

[User]
Research Question (RQ):
"{RQ_TEXT}"

Citing Paper Title: "{CITING_TITLE}"
Cited Paper Title: "{CITED_TITLE}"

Citation Context (verbatim):
[BEGIN CONTEXT]
{CITATION_CONTEXT_TEXT}
[END CONTEXT]

Task:
1) Identify dataset/benchmark/corpus names
    explicitly or implicitly referenced.
2) For each dataset, provide:
   - name
   - usage_role
   - content_type
   - evidence (verbatim span)
   - confidence (0.0-1.0)
   - rationale (1-2 sentences grounded in context)

Output JSON only:
{
  "datasets": [
    {
      "name": "...",
      "usage_role": "Use|Modify|Evaluate Against",
      "content_type": "Performed Work|Discovery|
    Produced Resource",
      "evidence": "...",
      "confidence": 0.0-1.0,
      "rationale": "..."
    }
  ]
}
```

b) Dataset Relevance Filtering: a second LLM call is then made to filter the extracted datasets for relevance to the original research question. The model is provided with the RQ, the citation context, and the titles of both the citing and cited papers. It makes a binary decision (include/exclude) based on the semantic alignment between the dataset's usage and the

research goal. This step filters out datasets that, while valid, are not pertinent to the user's query.

Prompt 2: Relevance Filtering

```
[System]
Decide whether a candidate dataset is relevant to
    the research question,
based solely on the provided context and titles.
    Respond with valid JSON only.

[User]
Research Question (RQ):
"{RQ_TEXT}"

Candidate Dataset:
"name": "{CANDIDATE_DATASET_NAME}"

Context:
[BEGIN CONTEXT]
{CITATION_CONTEXT_TEXT}
[END CONTEXT]

Citing Paper Title: "{CITING_TITLE}"
Citing Paper Abstract: "{CITING_ABSTRACT}"
Cited Paper Title: "{CITED_TITLE}"
Cited Paper Abstract: "{CITED_ABSTRACT}"

Decision rules:
- Relevant if the context shows the dataset was
    used, modified, or evaluated
  in pursuit of the RQ (or a directly aligned
    objective).
- Not relevant if usage is unrelated, purely
    background, or from a different domain.
- Be conservative; require explicit evidence in
    the context.

Output JSON only:
{
  "is_relevant": true/false,
  "confidence": 0.0-1.0,
  "reasoning": "brief, evidence-based
    justification"
}
```

c) Prompting and Infrastructure: all LLM tasks run on a vLLM-powered inference server with four NVIDIA A100-80GB GPUs. Our prompt engineering strategy casts the LLM as an expert annotator, providing strict inclusion/exclusion rules (e.g., extract 'MIMIC-III', ignore 'BERT') and mandating that all outputs are grounded in the provided text to prevent hallucination. A structured JSON schema (based on SOFT) is enforced for all outputs, and the model's temperature is set to 0.0 for deterministic results.

C. Stage 3: Entity Resolution and Metadata Enrichment Details

This phase cleans and consolidates the raw, filtered dataset mentions into a canonical list.

a) Normalization and Consolidation: we apply a deterministic, rule-based procedure to resolve different surface forms (e.g., `ACE 2005 (zh) dataset`, `ACE-2005`) to a single canonical entity. This involves strong lexical normalization (lowercasing, punctuation removal, etc.) followed by grouping mentions that share the same normalized form. Within each group, a human-readable display name is chosen by frequency, and all original surface forms are retained as aliases. All provenance is preserved to ensure the process is auditable.

b) Multi-Source URL and PID Retrieval: we employ a three-tiered strategy to find a URL or Persistent Identifier (PID) for each dataset. First, we check if a URL was directly extracted from the citation context by the LLM. If not, we examine the metadata of the cited paper in S2AG; if it is a resource paper, we inherit its DOI. As a final step, we use the canonical dataset name to programmatically query an external search API (Google's Gemini search capabilities) to find its official homepage or repository.

c) Final Ranking and Output Generation: the result of this entire pipeline is a clean, deduplicated, and ranked list of datasets relevant to the original research question. The final ranking is determined by the citation count: the number of unique papers within our retrieved corpus that were found to use a given dataset. This usage-based frequency serves as a proxy for the dataset's prevalence and importance in the context of the research question. Each entry in the final list includes the canonical dataset name, its usage context, the citation count, and a verifiable URL or PID.

APPENDIX: PER-QUERY EXPERT RATINGS

TABLE IV: Consolidated expert evaluation ratings across all cross-disciplinary research questions. Each block shows the mean user ratings (1–5) $\pm$ standard deviation for a single query.

System	Relevant	Useful	Accessible	Trustworthy	Novelty	Overall
Agricultural Sciences: Plant Disease diagnosis or pest detection image Dataset						
Ours	4.38 ± 0.52	4.00 ± 0.63	3.83 ± 0.75	4.17 ± 0.75	3.67 ± 0.82	4.00 ± 0.63
Google	2.88 ± 0.64	2.56 ± 0.73	2.88 ± 0.64	2.56 ± 0.73	2.88 ± 0.64	2.56 ± 0.73
DataCite	2.63 ± 0.74	2.69 ± 0.70	3.19 ± 0.54	2.81 ± 0.66	3.50 ± 0.60	2.56 ± 0.66
Humanities (Arts): Laban Movement Analysis for Dance Emotion						
Ours	4.67 ± 0.52	4.17 ± 0.75	3.67 ± 0.82	4.17 ± 0.75	3.50 ± 0.84	4.00 ± 0.89
Google	3.67 ± 0.82	2.67 ± 0.82	2.83 ± 0.75	2.50 ± 0.84	3.00 ± 0.89	2.50 ± 0.84
DataCite	2.50 ± 0.84	2.67 ± 0.52	3.17 ± 0.41	2.67 ± 0.52	3.50 ± 0.55	2.50 ± 0.55
Engineering and Technology (Food): Antioxidant peptides sequence and activity relationship						
Ours	4.75 ± 0.35	4.43 ± 0.53	3.94 ± 0.70	4.31 ± 0.60	3.63 ± 0.80	4.38 ± 0.56
Google	2.88 ± 0.64	2.63 ± 0.74	2.88 ± 0.64	2.69 ± 0.70	2.88 ± 0.64	2.69 ± 0.70
DataCite	2.44 ± 0.73	2.69 ± 0.70	3.13 ± 0.64	2.75 ± 0.71	3.50 ± 0.76	2.69 ± 0.74
Engineering and Technology (Food): Salty Enhancing Peptides						
Ours	4.89 ± 0.20	4.78 ± 0.33	4.40 ± 0.52	4.67 ± 0.33	3.90 ± 0.57	4.78 ± 0.33
Google	2.10 ± 0.57	2.20 ± 0.42	2.50 ± 0.53	2.30 ± 0.48	2.50 ± 0.53	2.20 ± 0.42
DataCite	2.00 ± 0.47	2.30 ± 0.48	2.40 ± 0.52	2.50 ± 0.53	4.10 ± 0.32	2.30 ± 0.48
Medical and Health Sciences (Clinical): Colorectal liver metastases CRLM single cell RNA sequencing						
Ours	4.33 ± 0.58	4.00 ± 0.63	3.83 ± 0.75	4.17 ± 0.75	3.67 ± 0.82	4.00 ± 0.63
Google	3.00 ± 0.63	2.67 ± 0.52	2.83 ± 0.75	2.50 ± 0.55	3.00 ± 0.63	2.50 ± 0.55
DataCite	2.67 ± 0.52	2.67 ± 0.52	3.17 ± 0.41	2.83 ± 0.41	3.50 ± 0.55	2.50 ± 0.55
Social Sciences (Educational): Statistical Learning Non-native						
Ours	4.25 ± 0.96	3.75 ± 0.96	3.75 ± 0.96	4.00 ± 0.82	3.75 ± 0.96	3.75 ± 0.96
Google	3.00 ± 0.82	2.75 ± 0.96	2.75 ± 0.96	2.75 ± 0.96	3.00 ± 0.82	2.75 ± 0.96
DataCite	2.75 ± 0.96	2.75 ± 0.96	3.25 ± 0.96	3.00 ± 0.82	3.50 ± 0.58	2.75 ± 0.96