Multi-Disciplinary Dataset Discovery

FROM

Citation-Verified Literature Contexts

Zhiyin Tan
L3S Research Center
Hannover, Germany

Changxu Duan
Technische Universität Darmstadt
Darmstadt , Germany

The 25th ACM/IEEE Joint Conference on Digital Libraries 2025
December 16 - 19, 2025

Motivation

Where can we find datasets for a specific research topic?

A reliable way is to trace datasets through relevant papers.

But there are so many relevant papers. All of them?

Not manually.

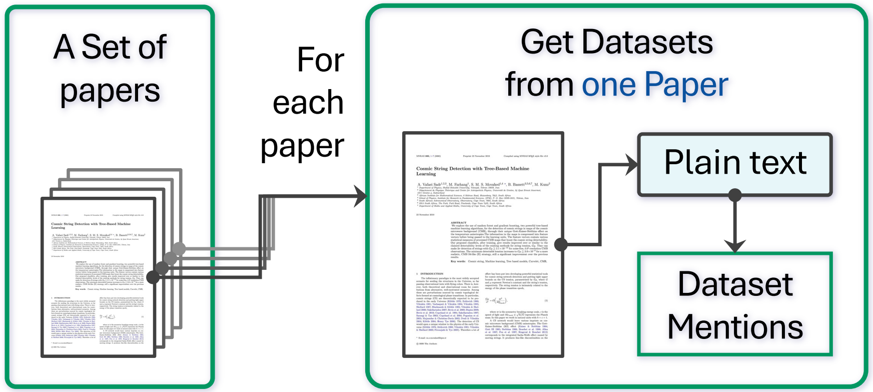
So… is there a way to do this automatically?

Related Work

A Set of papers

A common paradigm is to first collect a **fixed set of papers**,

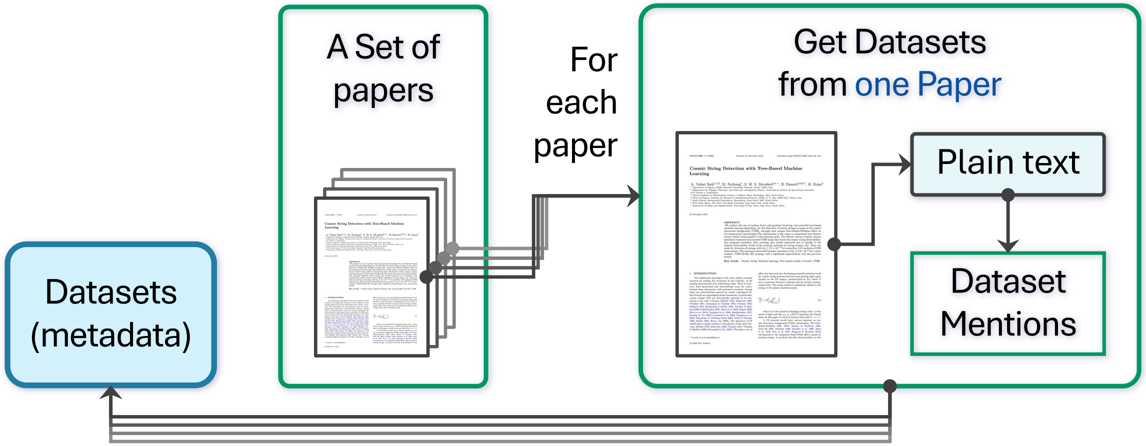
Related Work



A common paradigm is to first collect a **fixed set of papers**,

and then extract **dataset mentions** from each paper’s text using NLP-based methods,

Related Work

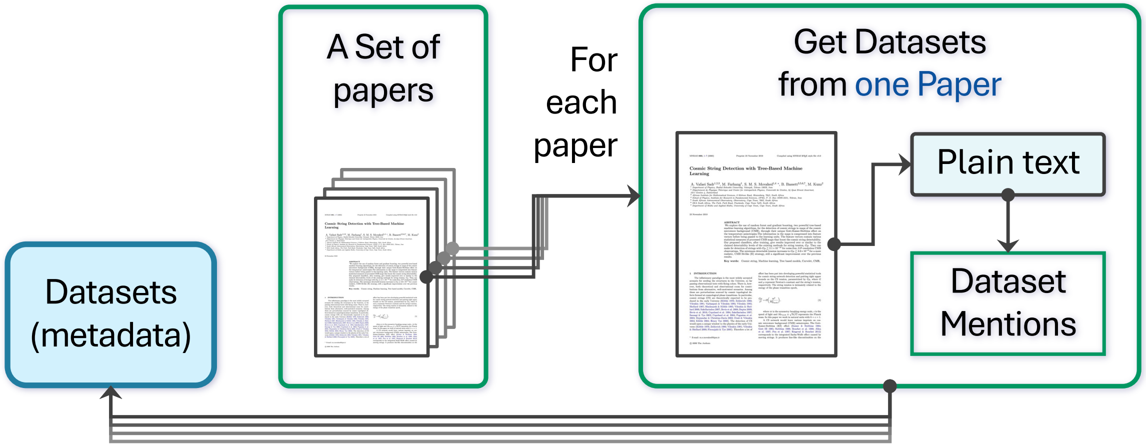


A common paradigm is to first collect a **fixed set of papers**,

and then extract **dataset mentions** from each paper's text using NLP-based methods,

finally aggregating them into a dataset collection (with metadata).

Related Work



A Set of papers

For each paper

Get Datasets from one Paper

Plain text

Dataset Mentions

Datasets (metadata)

A common paradigm is to first collect a **fixed set of papers**,

and then extract **dataset mentions** from each paper’s text using NLP-based methods,

finally aggregating them into a dataset collection (with metadata).

SciREX (2020)
Extract dataset mention with supervised sequence labelling from ML papers

DMDD (2023):
large-scale NER over S2ORC full text across AI-related domains

RAGing Against the Literature (2024):
retrieval-augmented LLM extraction built on DMDD

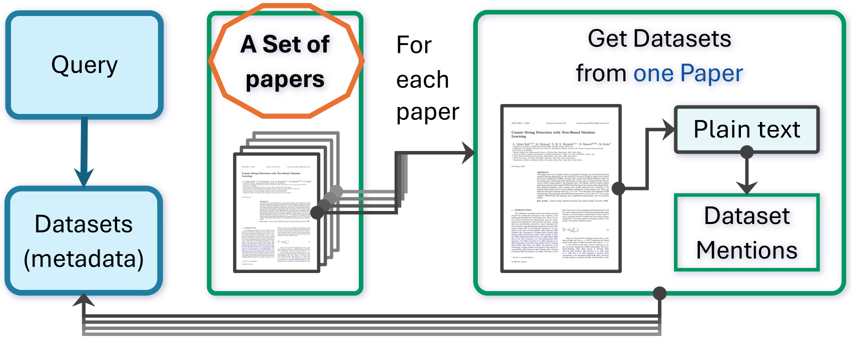
ChatPD (2025):
LLM-based extraction with entity resolution from arXiv full text

Challenges

1: Domain-limited / Static Paper Collections

Most existing approaches operate on **pre-defined, domain-limited paper collections**,
which are typically fixed and updated infrequently.

As a result,
they struggle to support **open-ended queries**
or to keep pace with updated published work.



Query

Datasets (metadata)

A Set of papers

For each paper

Get Datasets from one Paper

Plain text

Dataset Mentions

Our approach directly interfaces with academic search engines
to dynamically retrieve papers relevant to a given query.

SciREX (2020)
Extract dataset mention with supervised sequence labelling from **Machine Learning papers**

DMDD (2023):
large-scale NER over S2ORC full text across **AI-related domains**

RAGing Against the Literature (2024):
retrieval-augmented LLM extraction built on **DMDD**

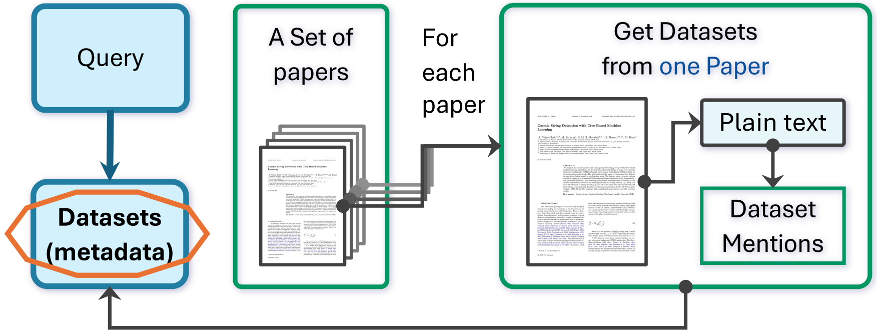
ChatPD (2025):
LLM-based extraction with entity resolution from **arXiv** full text

Challenges

2: Shallow Matching Based on Dataset Metadata

Many existing systems rely on matching a user query against **pre-extracted dataset metadata**, which typically contains limited information such as names, tasks, or brief descriptions.

Such metadata is often insufficient to capture the nuanced ways in which datasets are discussed and used in the literature.



To address this, we move the **matching stage earlier** and extend it over textual evidence, rather than relying solely on static metadata.

Google Dataset Search

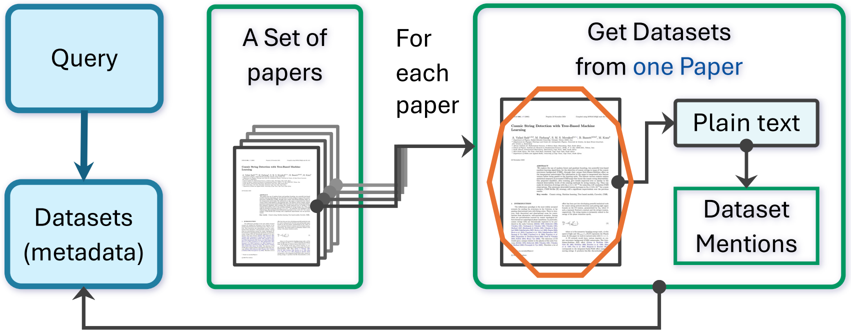
DataCite Commons

Challenges

3: Scalability of Full-text Processing

A natural solution is to dynamically process the **full text** of all retrieved papers.
However, this quickly becomes **computationally heavy at scale**.

Instead, we leverage **citation contexts from Semantic Scholar**, which are **updated on a weekly basis**.

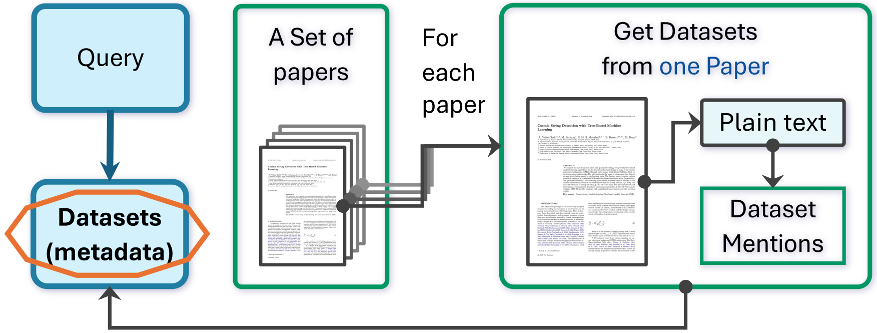


Compared to full text, citation contexts are **substantially shorter**, yet provide an **informative textual representation** of each paper, while **preserving high-quality signals about dataset usage**.

Challenges

4: Dataset Validity and Practical Usability

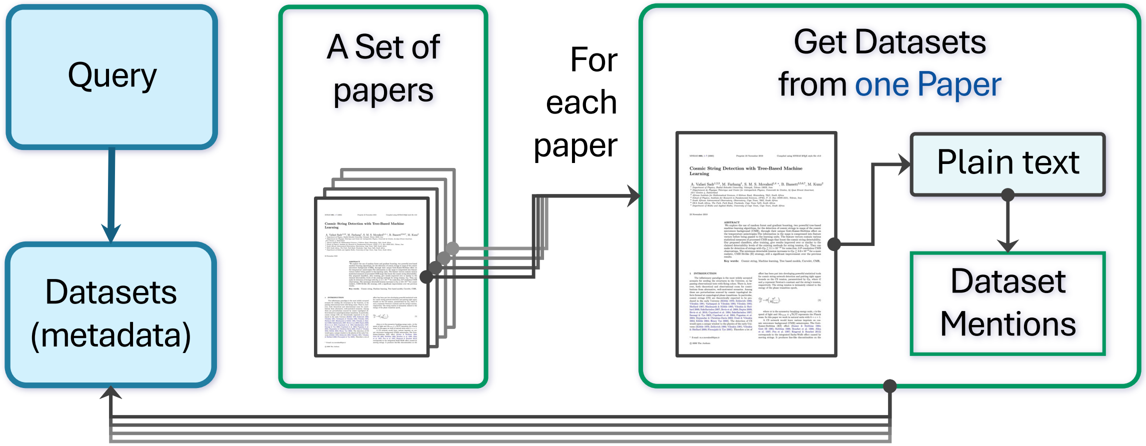
A less discussed but critical issue is that **many datasets mentioned in papers are no longer usable**, a problem that is particularly severe in the social sciences (e.g. Chen & Wu 2025 presented yesterday).



Starting from **citation contexts** provides an important advantage:

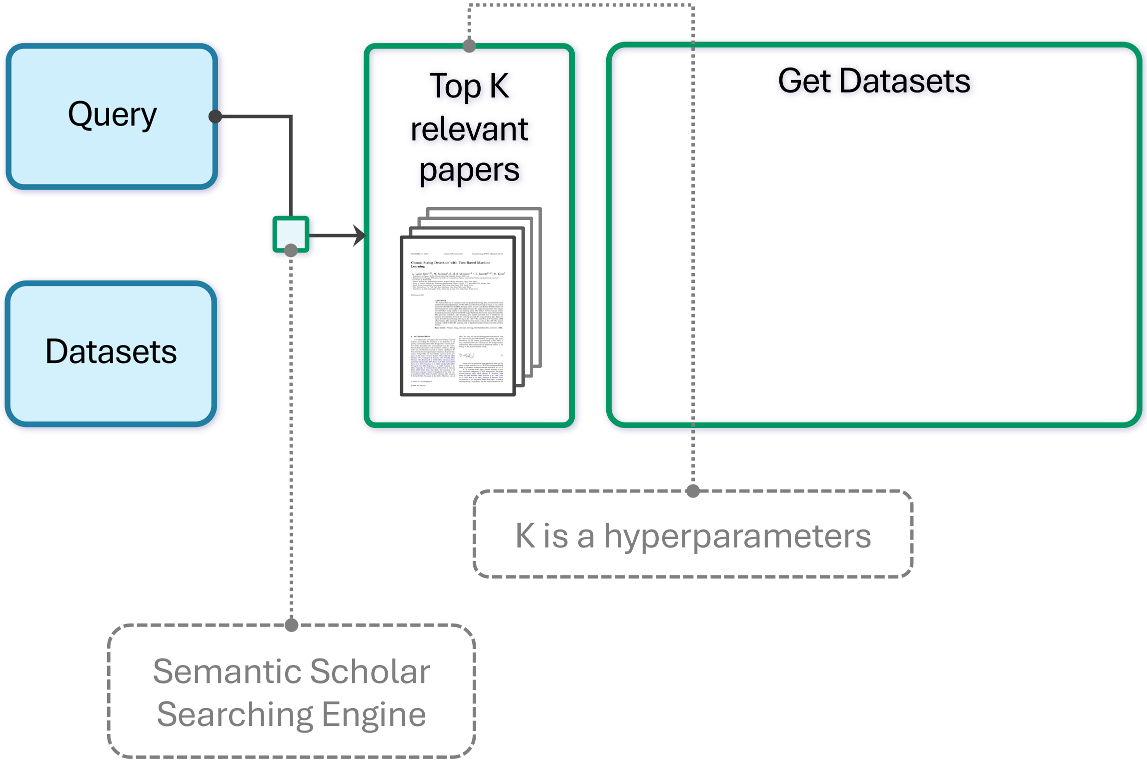
a dataset must have been **examined by at least one expert**, significantly reducing the need for manual filtering of unusable datasets.

Challenges



These challenges call for a **query-driven** and **scalable** approach for tracing datasets through the literature.

Our Approach



Query

Datasets

Top K relevant papers

Get Datasets

K is a hyperparameters

Semantic Scholar Searching Engine

Given a query,

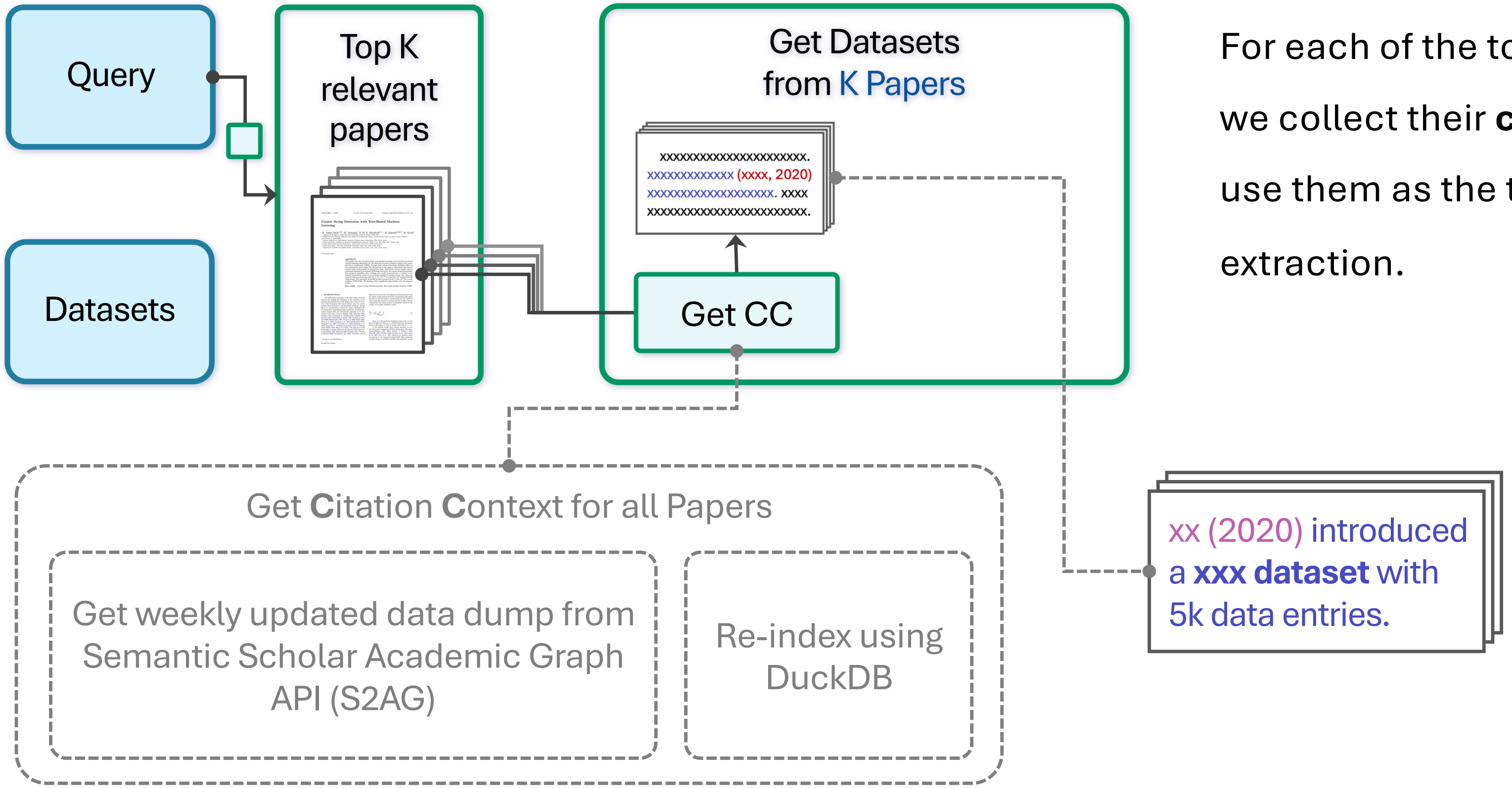
we first retrieve the **top-*K* relevant papers**

using an academic search engine,

forming the input to our dataset tracing

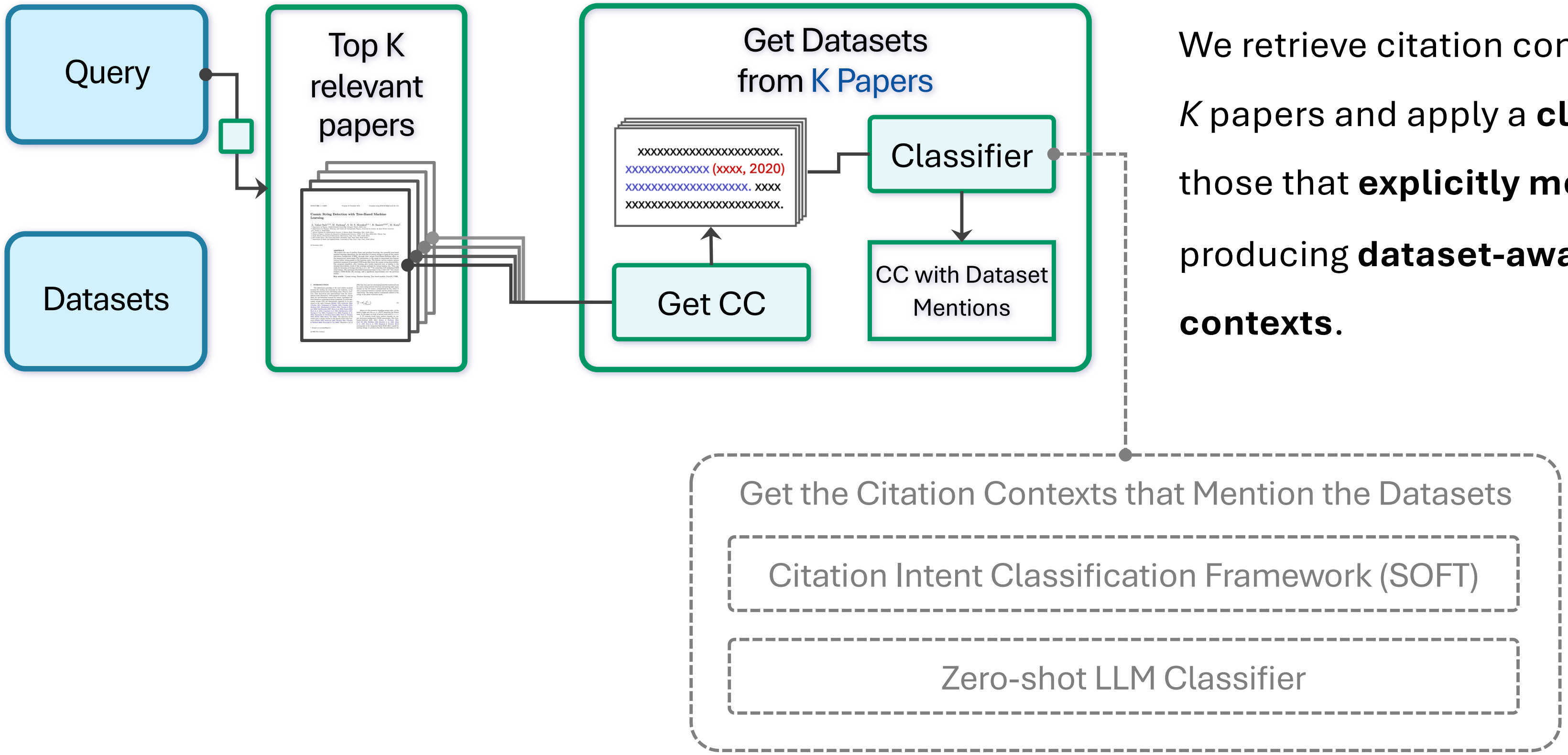
pipeline.

Our Approach


Query
Datasets
Top K relevant papers
Get Datasets from K Papers
xxxxxxxxxxxxxxxxxxxxxxx.
xxxxxxxxxxxxx (xxxx, 2020)
xxxxxxxxxxxxxxxxxxx. xxxx
xxxxxxxxxxxxxxxxxxxxxxxx.
Get CC
Get Citation Context for all Papers
Get weekly updated data dump from Semantic Scholar Academic Graph API (S2AG)
Re-index using DuckDB
xx (2020) introduced a xxx dataset with 5k data entries.

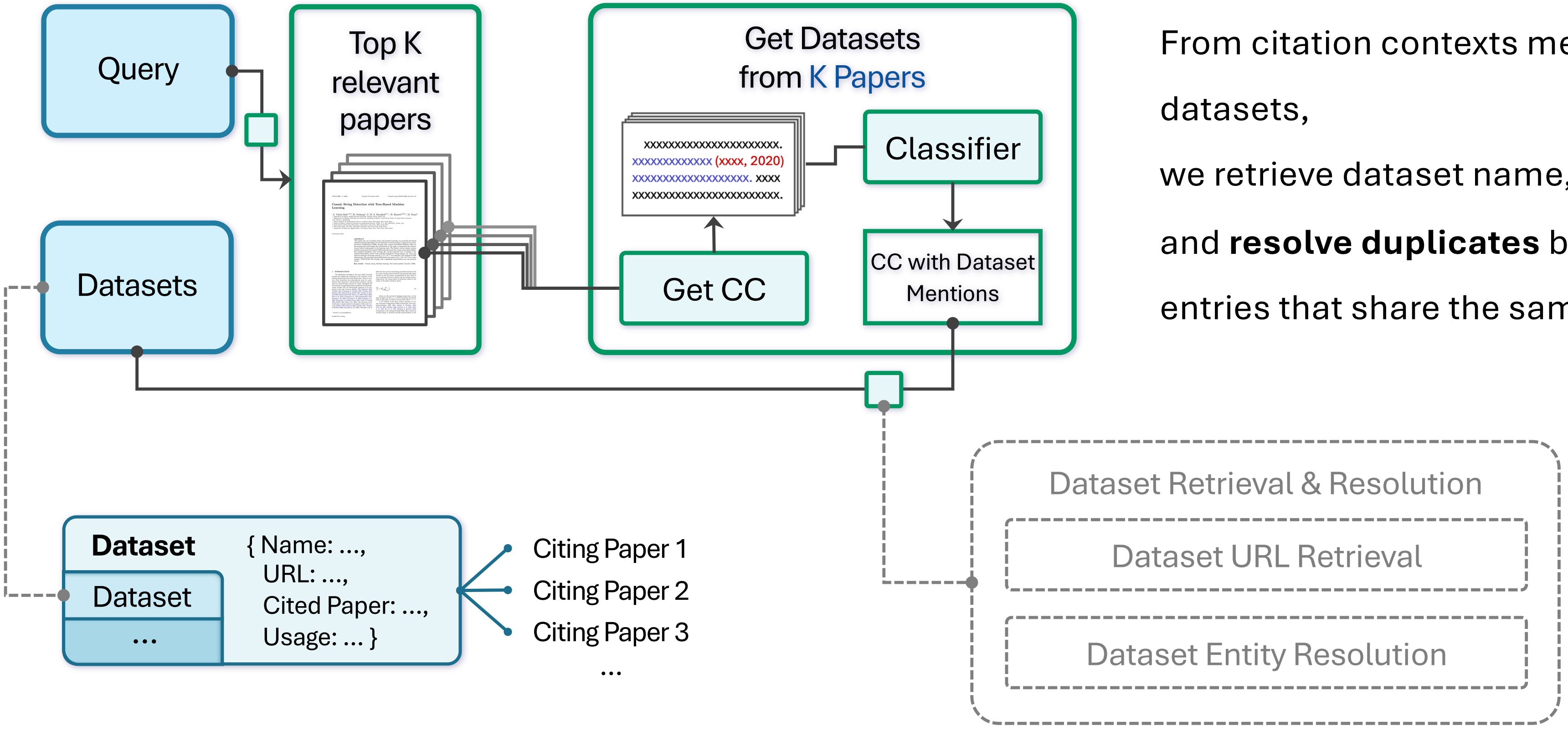
For each of the top-*K* retrieved papers, we collect their **citation contexts** and use them as the textual basis for dataset extraction.

Our Approach


Query
Datasets
Top K relevant papers
Get Datasets from K Papers
xxxxxxxxxxxxxxxxxxxxxxxx.
xxxxxxxxxxxxx (xxxx, 2020)
xxxxxxxxxxxxxxxxxxx. xxxx
xxxxxxxxxxxxxxxxxxxxxxxxxx.
Classifier
Get CC
CC with Dataset Mentions
Get the Citation Contexts that Mention the Datasets
Citation Intent Classification Framework (SOFT)
Zero-shot LLM Classifier

We retrieve citation contexts from the top-*K* papers and apply a **classifier** to retain those that **explicitly mention datasets**, producing **dataset-aware citation contexts**.

Our Approach



From citation contexts mentioning datasets,

we retrieve dataset name, URLs, etc

and **resolve duplicates** by merging

entries that share the same cited source.

Experiments

Data

Multi-Disciplinary: Revised Field of Science and Technology (FOS)

Field	Sub-field
1. Natural sciences	1.2 Computer and information sciences
2. Engineering and technology	2.11 Other engineering and technologies (Food and beverages)
3. Medical and Health sciences	3.2 Clinical medicine
4. Agricultural sciences	4.1 Agriculture, Forestry, and Fisheries
5. Social sciences	5.3 Educational sciences
6. Humanities	6.4 Arts (arts, history of arts, performing arts, music)

Datasets listed in **survey papers** as **gold standard**

Evaluated via **expert assessments**

Experiments

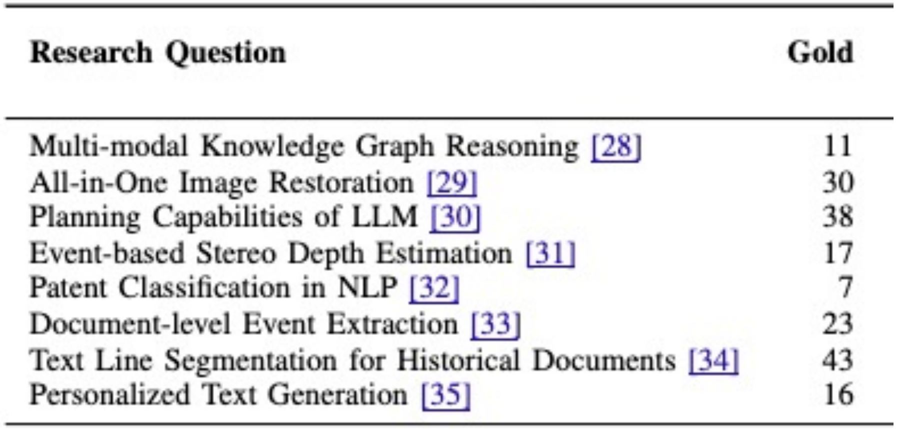
Data

Multi-Disciplinary: Revised Field of Science and Technology (FOS)

Field	Sub-field
1. Natural sciences	1.2 Computer and information sciences

Datasets listed in **survey papers** as **gold standard**

Survey title as queries



Research Question	Gold
Multi-modal Knowledge Graph Reasoning [28]	11
All-in-One Image Restoration [29]	30
Planning Capabilities of LLM [30]	38
Event-based Stereo Depth Estimation [31]	17
Patent Classification in NLP [32]	7
Document-level Event Extraction [33]	23
Text Line Segmentation for Historical Documents [34]	43
Personalized Text Generation [35]	16

Experiments

Data

Multi-Disciplinary: **Revised Field of Science and Technology (FOS)**

Field	Sub-field
1. Natural sciences	1.2 Computer and information sciences
2. Engineering and technology	- Plant Disease Diagnosis or pest detection image dataset - Laban Movement Analysis for Dance Emotion - Antioxidant Peptides Sequence and activity relationship - Salty-enhancing Peptides - Colorectal Liver Metastases CRLM Single Cell RNA Sequencing - Statistical Learning Non-native
3. Medical and Health sciences	
4. Agricultural sciences	
5. Social sciences	
6. Humanities	

Evaluated via **expert assessments**

Expert-provided queries
(Requested to provide a familiar research topic)

Experiments

Data

Multi-Disciplinary: Revised Field of Science and Technology (FOS)

Baseline

Google Dataset Search

DataCite Commons

Evaluation Metric

Statistical: Recall (survey-based)

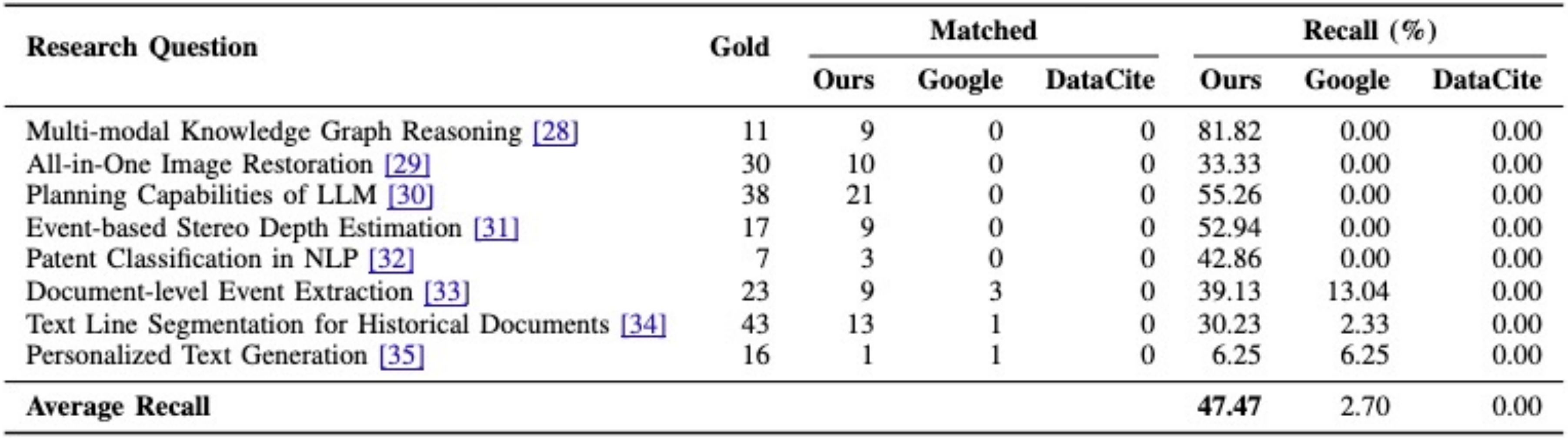
Expert Assessments:

- Educational Background: Doctoral Degree Holder (3), Doctoral Candidate (4), Master's Candidate (3)
- Method: double-blind Google Form
- Evaluation: 1-5 Scale for Relevance, Utility, Accessibility, Trustworthiness, Novelty

Results

For statistical-based evaluation:

- Our literature-based approach achieves substantially higher average recall (**47.47%**) than Google Dataset Search (2.70%) and DataCite (0.00%).
- On individual survey-derived tasks, recall reaches up to **81.82%**.



Research Question	Gold	Matched			Recall (%)		
		Ours	Google	DataCite	Ours	Google	DataCite
Multi-modal Knowledge Graph Reasoning [28]	11	9	0	0	81.82	0.00	0.00
All-in-One Image Restoration [29]	30	10	0	0	33.33	0.00	0.00
Planning Capabilities of LLM [30]	38	21	0	0	55.26	0.00	0.00
Event-based Stereo Depth Estimation [31]	17	9	0	0	52.94	0.00	0.00
Patent Classification in NLP [32]	7	3	0	0	42.86	0.00	0.00
Document-level Event Extraction [33]	23	9	3	0	39.13	13.04	0.00
Text Line Segmentation for Historical Documents [34]	43	13	1	0	30.23	2.33	0.00
Personalized Text Generation [35]	16	1	1	0	6.25	6.25	0.00
Average Recall					**47.47**	2.70	0.00

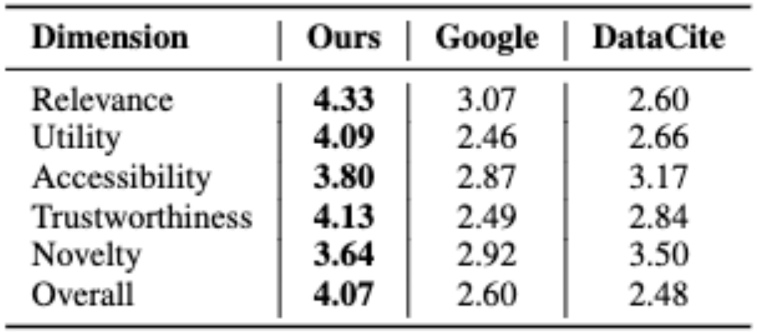
Results

For expert-based evaluation:

Experts evaluated datasets on six 5-point dimensions,
- Across all research queries, our system is consistently rated highest in relevance and utility, with strong gains in trustworthiness and overall quality.

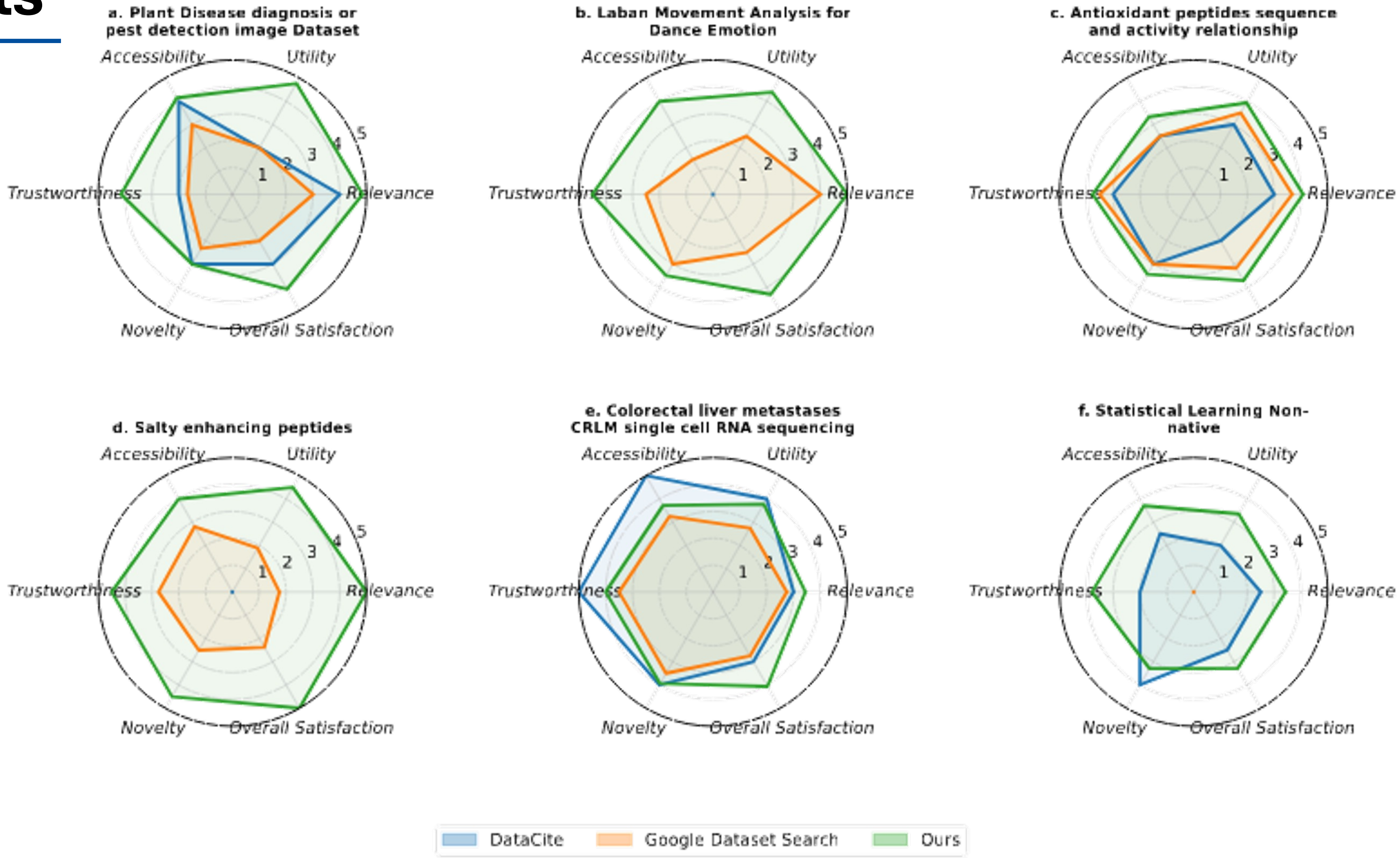
- Among all expert-rated datasets, our method surfaces 45 out of 105 datasets (**42.9%**) rated **at least 4** on both utility and novelty, compared to **12.9%** for Google Dataset Search and **33.3%** for DataCite.

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

Results


a. Plant Disease diagnosis or pest detection image Dataset
b. Laban Movement Analysis for Dance Emotion
c. Antioxidant peptides sequence and activity relationship
d. Salty enhancing peptides
e. Colorectal liver metastases CRLM single cell RNA sequencing
f. Statistical Learning Non-native
Accessibility
Utility
Relevance
Overall Satisfaction
Novelty
Trustworthiness
1
2
3
4
5
DataCite
Google Dataset Search
Ours

Acknowledgement

Thank you for your attention!

Thanks to the SIGIR Student Travel Grant for funding the registration fee.

Changxu Duan

Technische Universität Darmstadt, Darmstadt, Germany

duan@linglit.tu-darmstadt.de