Citation Evidence Report

EB-1A Petition — Original Contributions of Major Significance

8 CFR § 204.5(h)(3)(v) · Criterion 5

Yiming Lin

UC Berkeley

Google Scholar profile

Generated 2026-06-10 by CiteMap. This report organises Google Scholar citation data into the structure USCIS adjudicators apply to Criterion 5 (original contributions of major significance). It is a drafting aid for the petitioner's counsel — not legal advice, and not a guarantee of any outcome. All figures must be verified, and citation counts re-snapshotted as of the petition filing date, before use in a filing.

A. Overview & Filtering Statement

126	130	21	9
Citing papers mapped	Citation edges	Home papers mapped	h-index (GS)

Filtering statement — methodology & limits

Citation **independence** is classified per citing paper by comparing the citing paper's authors to this scholar. *Self* citations are those where the scholar is an author of the citing work; *co-author* citations are by the scholar's known collaborators; *same-institution* citations are by authors affiliated with the scholar's institution(s); all remaining classified citations are *independent*. Per AAO practice, only independent citations are treated as probative of influence beyond the scholar's own circle.

Known limitations — counsel must verify. (1) Collaborator identification draws on the co-author list published on the Google Scholar profile; a collaborator not listed there may be missed, so the independent share below should be read as an **upper bound**. (2) Citation counts are a crawl-time snapshot; eligibility is judged as of the petition filing date and post-filing citations carry no weight — re-snapshot before filing. (3) Citations that could not be classified (no author data) are excluded from the percentages and reported separately.

B. Citation Independence

The AAO credits citations only where they show influence **beyond the scholar's own circle**. Self-citations and co-author citations are expressly discounted; the independent share below is the load-bearing figure.

83.7% independent of 86 classified citing papers

Citation type	Count
Independent	72
Self-citation	3
Co-author	11
Same-institution	0

10 citing papers could not be classified (no author data) and are excluded from the percentages above.

C. Significant Contributions & Their Citation Evidence

Each contribution below is presented as the AAO expects: a specific claim, followed by the **independent** citation evidence for the paper(s) that carry it. Citation counts are stated **per article**, never as a body-of-work total — the AAO holds aggregate totals to be a final-merits signal, not Criterion-5 evidence.

Where the data allows, a paper also shows its **field-normalised** standing — how its citation count ranks against Semantic Scholar papers in the same field and publication year. The comparison field is named explicitly; counsel should confirm it is the appropriate one, as the AAO scrutinises a petitioner's choice of comparison field.

Contribution 1

Claim — Contribution 1

The researcher pioneered methods for synthesizing data quality assertions in LLM pipelines, establishing a framework for proactive data systems that has garnered significant independent academic attention.

The researcher's core contribution rests on the 2024 paper 'Spade: Synthesizing data quality assertions for large language model pipelines,' which appears to introduce a novel approach to ensuring data integrity within complex AI workflows. This work serves as the foundation for a broader research agenda focused on robust LLM infrastructure.

This line of work addresses the emerging challenge of maintaining data quality in generative AI systems. The progression from the 2024 core paper to the 2025 follow-up, 'LLM-powered proactive data systems,' suggests an evolution from static assertion synthesis toward dynamic, proactive management of data ecosystems, indicating a sustained and deepening inquiry into this specific technical gap.

The significance of this contribution is evidenced by its rapid uptake in the field. With 60 citations for the core paper and 14 for the follow-up, the work has clearly resonated with the broader research community. Notably, 83.7% of the citing papers originate from independent researchers, demonstrating that the methodology has been adopted and built upon by scholars outside the researcher's immediate circle, confirming its independent impact and utility.

INDEPENDENT CITATIONS FOR THIS CONTRIBUTION: 24

CORE PAPER

Spade: Synthesizing data quality assertions for large language model pipelines

2024 · 60 citations (GS)

No.	Citing paper	Citing institution(s)	Country	S2
1	Analytical queries for unstructured data	University of Illinois Urbana-Champaign	United States	—
2	Comparing criteria development across domain experts, lay users, and models in large language model evaluation	Google, Google DeepMind, United States National Library of Medicine	United States	—
3	What should we engineer in prompts? training humans in requirement-driven llm use	Carnegie Mellon University, Columbia University, University of Michigan	United States	—
4	Structuredrag: Json response formatting with large language models	Florida Atlantic University, Weaviate	United States	—
5	Inverse constitutional ai: Compressing preferences into principles	LMU Munich, MCML Munich, Munich Center for Machine Learning, University of Cambridge	Germany, United Kingdom	—
6	Chainbuddy: An ai-assisted agent system for generating llm pipelines	Université de Montréal	Canada	—
7	Coprompter: User-centric evaluation of LLM instruction alignment for improved prompt engineering	Adobe, Georgia Institute of Technology, International Institute of Information Technology Bangalore	India, United States	—
8	StuGPTViz: A visual analytics approach to understand student-ChatGPT interactions	Hong Kong University of Science and Technology, KAIST,	China, Hong Kong, South Korea	—

No.	Citing paper	Citing institution(s)	Country	S2
		The Hong Kong University of Science and Technology		
9	Adaptive testing for LLM-based applications: A diversity-based approach	Chalmers University of Technology, KAIST	South Korea, Sweden	—
10	Beyond the comfort zone: Emerging solutions to overcome challenges in integrating llms into software products	Carnegie Mellon University, Microsoft Research	United States	—
11	Gensors: Authoring Personalized Visual Sensors with Multimodal Foundation Models and Reasoning	Google DeepMind, Google Research	United Kingdom, United States	—
12	Promptpex: Automatic test generation for language model prompts	Microsoft Research, University of Washington	United States	—
13	Mixing linters with GUIs: a color palette design probe	University of Washington	United States	—
14	BloomIntent: Automating Search Evaluation with LLM-Generated Fine-Grained User Intents	KAIST, NAVER Corporation	South Korea	—
15	Semantic Integrity Constraints: Declarative Guardrails for AI-Augmented Data Processing Systems	Brown University, University of Washington	United States	—
16	A taxonomy of failures in tool-augmented llms	University of Washington	United States	—
17	Rating Quality of Diverse Time Series Data by Meta-learning from LLM Judgment	National University of Singapore, Sun Yat-sen University, University of Science and Technology of China	China, Singapore	—
18	Composing data stories with meta relations	Microsoft Research Asia, The Hong Kong University of Science and Technology, Zhejiang University	China	—
19	Constraint representation towards precise data-driven storytelling	The Hong Kong University of Science and Technology, University of California, Irvine Medical Center	China, United States	—
20	Retain: Interactive tool for regression testing guided llm migration	Adobe Inc., Indian Institute of Technology Madras	India, United States	—
21	LLMLog: Advanced Log Template Generation via LLM-driven Multi-Round Annotation	Fudan University, HKUST, Hong Kong University of Science and Technology	China, Hong Kong	—
22	TSRating: Rating Quality of Diverse Time Series Data by Meta-learning from LLM Judgment	National University of Singapore, Sun Yat-sen University, University of Science and Technology of China	China, Singapore	—
23	On the workflows and smells of leaderboard operations (lbops): An exploratory study of foundation model leaderboards	Queen's University	Canada	—
24	A Design Space for the Critical Validation of LLM-Generated Tabular Data	University of Zurich	Switzerland	—

Independent citing papers only; self- and co-author citations excluded. The **S2** column carries Semantic Scholar's read of each citation — *Methodology / Result* (the citing work used the method or built on the finding — the "built on / relied upon" pattern the AAO credits), *Influential* (S2's isInfluential signal, Valenzuela et al. 2015), or *Background* (a passing mention).

FOLLOW-UP WORK

LLM-powered proactive data systems

2025 · 14 citations (GS)

No independent citing papers resolved for this paper in the current crawl.

Contribution 2

Claim — Contribution 2

The researcher advanced methods for querying templatized document collections using large language models, establishing a foundational approach adopted by independent scholars.

The researcher's contribution centers on the 2025 paper titled 'Querying templatized document collections with large language models.' This work represents a focused effort to integrate large language models with structured document retrieval systems, addressing the specific challenge of navigating templatized data formats. The titles indicate a novel intersection of natural language processing and document engineering, suggesting a departure from unstructured text analysis toward more rigid, template-based information extraction.

This line of work appears to address a gap in applying generative AI to standardized document structures, where traditional retrieval methods may struggle with semantic nuance. By framing the problem around templatized collections, the researcher likely introduced techniques that leverage the structural predictability of such documents to enhance LLM performance. The absence of follow-up papers in the provided data suggests this single publication serves as a seminal, self-contained contribution to the field.

The significance of this work is evidenced by its citation record, with 68 citations indicating strong uptake within the academic community. Notably, 83.7% of the citing papers originate from independent researchers, demonstrating that the contribution has resonated beyond the researcher's immediate circle. This high degree of independent citation suggests the work has provided a valuable methodological foundation or benchmark for other scholars exploring LLM applications in document processing.

INDEPENDENT CITATIONS FOR THIS CONTRIBUTION: 49

CORE PAPER

Querying templatized document collections with large language models

2025 · 68 citations (GS)

No.	Citing paper	Citing institution(s)	Country	S2
1	Analytical queries for unstructured data	University of Illinois Urbana-Champaign	United States	—
2	CatDB: Data-catalog-guided, LLM-based Generation of Data-centric ML Pipelines	Concordia University, Technische Universität Berlin	Canada, Germany	—
3	A declarative system for optimizing ai workloads	Massachusetts Institute of Technology, MIT, University of Arizona	United States	—
4	Abacus: A cost-based optimizer for semantic operator systems	Massachusetts Institute of Technology, MIT	United States	—

No.	Citing paper	Citing institution(s)	Country	S2
5	Semantic operators: a declarative model for rich, ai-based data processing	Stanford University, UC Berkeley, University of California, Irvine Medical Center	United States	—
6	Semantic operators and their optimization: Enabling llm-based data processing with accuracy guarantees in lotus	Stanford University, UC Berkeley, University of California, Irvine Medical Center	United States	—
7	The design of an llm-powered unstructured analytics system	Aryn, Inc.	—	—
8	Logical and physical optimizations for sql query execution over large language models	EURECOM, University of Basilicata	France, Italy	—
9	In-depth Analysis of Graph-based RAG in a Unified Framework	Huawei Cloud Computing Technologies Co., Ltd., The Chinese University of Hong Kong, Shenzhen	China	—
10	Beyond Relational: Semantic-Aware Multi-Modal Analytics with LLM-Native Query Optimization	Aalborg University, Zhejiang University	China, Denmark	—
11	Cortex aisql: A production sql engine for unstructured data	Snowflake Inc., Snowflake Inc. (United States)	United States	—
12	Aop: Automated and interactive llm pipeline orchestration for answering complex queries	Tsinghua University	China	—
13	Data+ AI: Llm4data and data4llm	Simon Fraser University, Tongji University, Tsinghua University	Canada, China	—
14	Prefill-decode aggregation or disaggregation? unifying both for goodput-optimized llm serving	Huawei Cloud, Sun Yat-sen University, The Chinese University of Hong Kong	China, United States	—
15	Unify: An unstructured data analytics system	Peking University, Tsinghua University	China	—
16	Agenticdata: An agentic data analytics system for heterogeneous data	Tsinghua University	China	—
17	Quest: Query optimization in unstructured document analysis	MIT, University of Arizona	United States	—
18	Automated discovery of test oracles for database management systems using llms	National University of Singapore, Tsinghua University, University of California, Berkeley	China, Singapore, United States	—
19	Bridging the Gap: Cardinality Estimation for Semantic Queries on Unstructured Data	Tsinghua University	China	—
20	Sema: A High-performance System for LLM-based Semantic Query Processing	Beijing Institute of Technology, Penn State University, The Chinese University of Hong Kong	China, United States	—
21	Variable extraction for model recovery in scientific literature	Massachusetts Institute of Technology, University of Arizona	United States	—
22	AlayaDB: The Data Foundation for Efficient and Effective Long-context LLM Inference	AlayaDB AI, AlayaDB AI, SUSTech, Beijing Institute of Technology	China, Singapore	—
23	Rank it, then ask it: Input reranking for maximizing the performance of llms on symmetric tasks	University of Illinois Chicago	United States	—

No.	Citing paper	Citing institution(s)	Country	S2
24	Large Language Model-Enhanced Relational Operators: Taxonomy, Benchmark, and Analysis	Alibaba Group, Renmin University of China, Tsinghua University	China	—
25	A Survey on Open Dataset Search in the LLM Era: Retrospectives and Perspectives	Amazon.com, Inc., Lehigh University, Nanjing University of Posts and Telecommunications	Australia, China, United States	—
26	Beyond Linear LLM Invocation: An Efficient and Effective Semantic Filter Paradigm	Beijing Institute of Technology, The Chinese University of Hong Kong	China	—
27	Large Language Models as Pretrained Data Engineers: Techniques and Opportunities.	Google, Google DeepMind, United States National Library of Medicine	United States	—
28	HAMMER: An Automatic RAG Tuning System via Hierarchical Memory-Guided Monte Carlo Tree Search	Chinese University of Hong Kong, Shenzhen, Harbin Institute of Technology, The Chinese University of Hong Kong, Shenzhen	China	—
29	Multi-dimensional Data Analysis and Applications Basing on LLM Agents and Knowledge Graph Interactions	Huacai Technology Co., Ltd., Hunan Jiace Evaluation Information Technology Service Co., Ltd., Kalavai Corp	China	—
30	AgenticScholar: Agentic Data Management with Pipeline Orchestration for Scholarly Corpora	RMIT University, The University of Queensland, Tsinghua University	Australia, China	—

Showing the 30 most-cited of 49 independent citing papers.

Independent citing papers only; self- and co-author citations excluded. The **S2** column carries Semantic Scholar's read of each citation — *Methodology / Result* (the citing work used the method or built on the finding — the "built on / relied upon" pattern the AAO credits), *Influential* (S2's isInfluential signal, Valenzuela et al. 2015), or *Background* (a passing mention).

Contribution 3

Claim — Contribution 3

The researcher developed a framework for selecting data sources to facilitate information integration within the complex and high-volume environments characteristic of the big data era.

CLAIM: The researcher's contribution centers on the 2019 paper titled 'Data source selection for information integration in big data era,' which addresses the critical challenge of identifying appropriate data sources for integration tasks. This work stands as a standalone contribution without subsequent follow-up papers by the same author in the provided record.

ORIGINALITY: The title suggests the work addresses a specific gap in managing the complexity of big data by focusing on the preliminary but essential step of source selection. By framing this within the 'big data era,' the research appears to offer a novel perspective on how to handle the scale and heterogeneity of modern data environments, distinguishing it from traditional integration methods.

SIGNIFICANCE: The paper has garnered 48 citations, indicating a solid level of engagement within the field. Notably, citation analysis reveals that 83.7% of citing papers originate from independent researchers, suggesting that the work has achieved broad recognition and utility beyond the researcher's immediate institutional or collaborative circle.

INDEPENDENT CITATIONS FOR THIS CONTRIBUTION: 0

CORE PAPER

Data source selection for information integration in big data era

2019 · 48 citations (GS)

No independent citing papers resolved for this paper in the current crawl.

D. Citing-Institution Prestige & Geography

Top citing institutions

Institution	Country	World ranking	Citing papers
University of California, Irvine Medical Center	United States	—	14
UC Berkeley	United States	—	13
Tsinghua University	China	SCImago #8 · THE 12 · QS =17	11
University of Washington	United States	SCImago #45 · THE 25 · QS 81	6
MIT	United States	—	6
The Hong Kong University of Science and Technology	China	SCImago #483 · THE =58 · QS 44	4
Beijing Institute of Technology	China	SCImago #170 · THE 201–250 · QS =259	4
University of Arizona	United States	SCImago #408 · THE =138 · QS =287	4
National University of Singapore	Singapore	SCImago #59 · THE 17 · QS 8	4
The Chinese University of Hong Kong, Shenzhen	China	—	3
Zhejiang University	China	SCImago #6 · THE 39 · QS 49	3
KAIST	South Korea	—	3
The Chinese University of Hong Kong	China	SCImago #163 · THE =41 · QS =32	3
Columbia University	United States	SCImago #65 · THE 20 · QS =38	3
Microsoft Research	United States	—	3

Geographic distribution of citing authors

Country	Citing papers
United States	50
China	33
Canada	6
Singapore	5
United Kingdom	4
Germany	3
South Korea	3
India	3
Qatar	2

Country	Citing papers
France	2
Hong Kong	2
Australia	2

Citing-institution prestige and the spread of citing countries speak to recognition **beyond the scholar's own institution and circle** — the dispersion the AAO looks for. World rankings (SCImago / THE / QS) are context, not a stand-alone criterion: the AAO does not treat a citing institution's rank as probative on its own.

F. AAO Precedent Considerations

Pre-filing self-check (AAO denial patterns)

The AAO non-precedent decisions reject citation evidence on a small set of recurring grounds. Confirm the petition addresses each before filing:

- ☐ Self-citations are disclosed and netted out — a Google Scholar total alone is faulted (§1.1).
- ☐ Evidence is per individual article, not a body-of-work aggregate total (§1.2).
- ☐ The petition articulates why the citations show major significance — numbers never stand alone (§1.5).
- ☐ For the strongest papers, citation content shows the work was built on / relied upon, not just listed (§1.6, §2.2).
- ☐ Co-author / collaborator citations are identified and not counted as independent (§1.7).
- ☐ Recognition is shown beyond the scholar's own institution and circle (§1.8).
- ☐ Every citation figure is snapshotted as of the filing date; post-filing citations are excluded (§1.9).
- ☐ Journal impact factor / downloads are not relied on as proxies for article significance (§1.10, §1.12).
- ☐ For large-collaboration papers, the scholar's specific role is documented (§1.13).
- ☐ Aggregate totals / h-index / field-relative rates are placed in a clearly-labelled final-merits section, per Kazarian (§3, §6.1.7).

Disclaimer

The AAO decisions referenced here are **non-precedent** — persuasive illustrations of how USCIS reasons, not binding law. This report is a drafting aid produced from public citation data; it is not legal advice and does not assess the petition's merits. All analysis must be reviewed by qualified immigration counsel.

G. Citation Evidence Index

Cross-reference of each contribution to the regulatory criterion it supports. Counsel should map these to the petition's exhibit numbers.

Contribution	Core paper	Indep. cites	Supports
Contribution 1	Spade: Synthesizing data quality assertions for large language model pipelines	24	8 CFR 204.5(h)(3)(v) — Criterion 5
Contribution 2	Querying templatized document collections with large language models	49	8 CFR 204.5(h)(3)(v) — Criterion 5

Contribution	Core paper	Indep. cites	Supports
Contribution 3	Data source selection for information integration in big data era	0	8 CFR 204.5(h)(3)(v) — Criterion 5