

Data Citation for Large Language Models: A Challenge

GIANMARIA SILVELLO, University of Padua, Italy

ACM Reference Format:

Gianmaria Silvello. 2026. Data Citation for Large Language Models: A Challenge. 1, 1 (July 2026), 6 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Motivation

Large Language Models (LLMs) are rapidly becoming primary interfaces for information access, synthesis, and decision support. An LLM draws on many sources: the training corpora absorbed into its parameters and, at inference time, retrieved documents (as in Retrieval-Augmented Generation (RAG)), Knowledge Graphs (KGs) used for grounding, and tool-augmented or agentic workflows (database queries, search and KG-traversal APIs, code execution). The origin of an answer is therefore often a dynamic process spanning several sources rather than a single retrieved chunk. A growing body of work asks whether LLMs reliably cite the sources behind their outputs [7, 12, 30]; this line of research treats *citation* as a verifiability mechanism: *Does the model point to a source that supports its claim?*

This framing reflects only one aspect of a deeper issue. In scholarly work, citation has at least three functions: *verification* (allowing readers to check claims), *credit* (acknowledging intellectual contributions), and *provenance* (tracing the origin of information) [2, 24]. Current LLM citation efforts focus almost entirely on verification for textual documents, largely neglecting *data*: training datasets, structured contextual data, and the KGs increasingly used to ground LLM reasoning. This paper poses the following research question:

How should LLMs cite data so as to ensure verifiability, provenance tracking, and fair credit attribution to data creators and curators?

This question poses a challenge that is both technically harder and, we believe, socially more consequential than document-level citation grounding; this is particularly true because solving the “data problem” carries over to text in an almost straightforward way. Data citation in traditional scholarly publishing is still an unsolved problem [4, 29]; transposing it to the LLM setting amplifies every known difficulty and introduces new ones. The challenge sits at the intersection of database theory, information retrieval, knowledge representation, and AI ethics, and its investigation requires contributions from all these communities.

2 Background

The FORCE11 Joint Declaration of Data Citation Principles [1] established data as first-class scholarly objects whose citations are human-understandable, machine-actionable, and support credit, provenance, and verifiability. Despite infrastructure progress – DataCite [8], Scholix [6], and the RDA recommendations for referencing dynamic data [28] – adoption remains far from

Author’s Contact Information: Gianmaria Silvello, gianmaria.silvello@unipd.it, University of Padua, Padua, Italy.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM XXXX-XXXX/2026/7-ART

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

universal, and core problems persist: citing subsets and query results over evolving databases, ensuring fixity under updates, and generating complete, correct references at the right granularity [3, 5]. Formal view-based models exist [9, 33] but they are not pervasive, and even well-formed dataset references lack a common standard, so informal citations remain widespread [16, 20, 23]. From an information-science standpoint, the challenge connects to the infrastructure layer that operationalizes credit and reuse: the FAIR principles, which require data to be findable and machine-actionable [32]; the CRediT contributor-role taxonomy, whose roles include data curation [26]; and Make Data Count, which builds usage metrics intended to reward data creators [22]. These frameworks define who deserves credit, how reuse is tracked and focus on metadata structure and tracking, yet none currently extends to data consumed inside LLM outputs.

Citing data from a curated database raises the question of how to distribute credit to the elements – and ultimately the curators – that produced it. Dosso and Silvello [14] formalized *data credit distribution* on top of data citations and propagated it through a relational database’s structure, later extended to lineage-based propagation along provenance chains [13]. The concept of *indirect citations* [17] and *transitive credit* [21] generalizes this: if A contributes to B, the credit map of A feeds into B’s. These mechanisms operate only within curated, provenance-aware databases.

Huang and Chang [19] argue that a comprehensive citation mechanism for LLMs must cover both non-parametric (retrieved) and parametric (training-internalized) content, a distinction that maps onto the separation of training-data attribution from RAG and KG citation developed in this paper. Recent work on RAG attribution shows that LLMs frequently misattribute sources: a cited document may not support the claim or may not be the one the model relied on [30], citation hallucination can be detected from a model’s internal computations [7], and users over-trust answers that carry citations even when those citations are random [12]. This line of work targets factual grounding and citation *faithfulness* – whether a generated claim is supported by some source – over textual documents. Data citation asks a different question: *which* data objects should be cited, and how credit should flow to their creators and curators. Critically, all of this work addresses document citations; to our knowledge, data citations remain largely absent from the evaluation landscape.

Finally, LLMs and KGs are increasingly integrated [27]: KGs ground LLMs to reduce hallucination, especially in neurosymbolic approaches where symbolic knowledge constrains and explains neural outputs. Yet how an LLM should *cite* the KG facts it uses, and how credit should reach their contributors, remains unaddressed.

3 Challenges and Research Directions

We identify three interconnected challenges, each of which opens concrete research directions.

Training Data Attribution. Datasets used to train LLMs are absorbed into billions of parameters and are, in practice, lost to citation. The training data attribution problem asks: given a model output, which training data points contributed to it? Existing approaches include influence functions [31] and data Shapley values [18]; large-scale audits separately document how licensing and attribution metadata are lost as data are absorbed into training pipelines [25]. However, these methods are computationally expensive, approximate, and do not produce anything resembling a *citation* in the scholarly sense.

A training-data citation would map an output to the training items that shaped it, each with a contribution weight; defining it rigorously and efficiently for models with billions of parameters and trillions of tokens is open. The problem becomes even more challenging when the training corpus includes structured datasets in addition to unstructured text, as it raises the question of how to appropriately reference a specific table, column, or tuple. We further emphasize that addressing this challenge is realistically feasible only for open models with documented training corpora.

Data Citation in LLMs. At inference time an LLM acquires data through several mechanisms, each raising the same citation problem, open even in traditional settings. RAG is the prototypical case. When structured datasets are provided as grounding context in a RAG pipeline, writing a proper reference for a dataset requires specifying the creator, title, version, persistent identifier, and access information – metadata that is often incomplete or absent [10]. If a model uses a specific subset of a dataset (e.g., rows matching a query predicate), the citation should reflect that subset with appropriate granularity, yet, to our knowledge, current LLM-based systems do not support this. Beyond retrieval, the citable object is frequently not a static document but a query and its result over a possibly versioned source, which makes granularity and fixity harder to pin down. Furthermore, the scarcity of well-formed data citations in the textual corpora on which LLMs are trained means that models have seen far fewer examples of proper data references than document references. This training distribution imbalance biases LLMs toward informal or absent data attribution, in what we conjecture is a self-reinforcing deficit.

Citing Knowledge Graph Facts. When an LLM uses a fact – e.g., (*Aspirin, treats, Headache*) – three questions arise. (i) **What is a citation to a fact?** A triple has no bibliographic metadata, and citing the KG as a whole (“according to Wikidata”) conflates millions of independent assertions. (ii) **Provenance is essential:** a fact may be extracted from a paper, curated by an expert, inferred, or aggregated, and this determines who is credited; provenance-aware representations such as nanopublications [15] and PROV-O-annotated named graphs [11] exist but are not integrated into LLM pipelines. (iii) **Credit must be distributed:** when a fact derives from several sources, the credit for using it should be split among them in proportion to their contribution, and when facts share upstream sources this becomes a network-propagation problem over the provenance graph. This is the *credit distribution* problem [13, 14] (with transitive credit [17, 21]), now at the scale, heterogeneity, and dynamism of KGs in LLM pipelines, which existing solutions do not address.

A comprehensive solution must co-design five elements: a unified citation model spanning documents, datasets, and KG facts at different granularities (whole dataset, subset, individual triple); provenance-aware LLM architectures that preserve and expose the link between generated tokens and their data sources, whether these enter through training, retrieval, tool calling or symbolic grounding; credit-distribution mechanisms operating at scale over heterogeneous provenance graphs to distribute credit fairly among data creators, curators, and contributors; evaluation benchmarks for data citation in LLM outputs, analogous to document-citation correctness and faithfulness benchmarks but for (semi-)structured data; and standards and infrastructure extending existing frameworks (FORCE11, DataCite, Scholix, RDA) to the LLM context to make data citations in model outputs machine-actionable. These stages are interdependent rather than merely sequential. In the RAG case, a citation can be emitted only if the architecture has retained the link from the generated span back to the retrieved tuple or query result (provenance); credit can be distributed only if that citation identifies the data object at the right granularity (citation model); and neither can be assessed without a benchmark that scores data-level rather than document-level citations (evaluation). A gap at any one stage degrades the others.

Among these challenges, data citation in LLM systems is the first to tackle and the most immediately tractable: the data sources are available at inference time, and citations can be anchored to retrievable, versioned artifacts or to the queries that produced them, with RAG the natural entry point. It is also arguably the most pressing, because as LLM systems become the standard way to ground outputs, the lack of proper data citation risks normalizing large-scale consumption of structured data without attribution – a gap that will only grow harder to close.

Data citation must therefore work not only in traditional publishing but in the AI systems that increasingly mediate access to knowledge. This requires the DB, IR, KR, and AI communities

to jointly develop the models, algorithms, benchmarks, and standards above, so that knowledge integrity remains traceable at inference time.

Author's version

References

- [1] M. Altman, C. Borgman, M. Crosas, and M. Martone. 2015. An introduction to the joint principles for data citation. *Bulletin of the Association for Information Science and Technology* 41, 3 (2015), 43–45. doi:10.1002/bult.2015.1720410313
- [2] C. L. Borgman. 2012. The Conundrum of Sharing Research Data. *JASIST* 63, 6 (2012), 1059–1078. doi:10.1002/asi.22634
- [3] P. Buneman. 2006. How to Cite Curated Databases and How to Make Them Citable. In *Proc. 18th International Conference on Scientific and Statistical Database Management (SSDBM)*. IEEE, 195–203.
- [4] P. Buneman, G. Christie, J. A. Davies, R. Dimitrellou, S. D. Harding, A. J. Pawson, J. L. Sharman, and Y. Wu. 2020. Why data citation isn't working, and what to do about it. *Database* 2020 (05 2020), baaa022. doi:10.1093/databa/baaa022
- [5] P. Buneman, S. B. Davidson, and J. Frew. 2016. Why data citation is a computational problem. *Communications of the ACM (CACM)* 59, 9 (2016), 50–57.
- [6] H. Cousijn, A. Kenall, E. Ganley, M. Harrison, D. Kernohan, T. Lemberger, F. Murphy, P. Polischuk, S. Taylor, M. Martone, and T. Clark. 2018. A data citation roadmap for scientific publishers. *Scientific Data* 5, 1 (Nov. 2018), 180259. doi:10.1038/sdata.2018.259
- [7] M. Dassen, R. Kotula, K. Murray, A. Yates, D. Lawrie, E. Kayi, J. Mayfield, and K. Duh. 2026. FACTUM: Mechanistic Detection of Citation Hallucination in Long-Form RAG. In *Proc. 48th European Conference on Information Retrieval (ECIR) (LNCS, Vol. 16483)*. Springer, 272–288. doi:10.1007/978-3-032-21289-4_18
- [8] DataCite. 2025. DataCite Annual Public Data File 2025. <https://datacite.org/blog/expanding-access-to-datacite-metadata-new-public-and-monthly-data-files-now-available/>
- [9] S. B. Davidson, P. Buneman, D. Deutch, T. Milo, and G. Silvello. 2017. Data Citation: A Computational Challenge. In *Proc. of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017*, Emanuel Sallinger, Jan Van den Bussche, and Floris Geerts (Eds.). ACM Press, New York, USA, 1–4. doi:10.1145/3034786.3056123
- [10] L. Delgado-Quirós and J. L. Ortega. 2024. Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies* 5, 1 (2024), 31–49. doi:10.1162/qss_a_00286
- [11] H. Dibowski. 2024. Full Traceability and Provenance for Knowledge Graphs. In *Formal Ontology in Information Systems, C. Trojahn, D. Porello, and P. P. Favato Barcelos (Eds.)*. IOS Press, 223–237. doi:10.3233/FALIA241309
- [12] Y. Ding, M. Facciani, E. Joyce, A. Poudel, S. Bhattacharya, B. Veeramani, S. Aguinaga, and T. Weninger. 2025. Citations and Trust in LLM Generated Responses. In *Proc. AAAI Conference on Artificial Intelligence*, Vol. 39. 23787–23795. doi:10.1609/aaai.v39i22.34550
- [13] D. Dosso, S. B. Davidson, and G. Silvello. 2022. Credit Distribution in Relational Scientific Databases. *Information Systems* 109 (2022), 102060. doi:10.1016/j.is.2022.102060
- [14] D. Dosso and G. Silvello. 2020. Data Credit Distribution: A New Method to Estimate Databases Impact. *Journal of Informetrics* 14, 4 (2020), 101080. doi:10.1016/j.joi.2020.101080
- [15] E. Fabris, T. Kuhn, and G. Silvello. 2019. A Framework for Citing Nanopublications. In *Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Proceedings (Lecture Notes in Computer Science)*. Springer, 70–83. doi:10.1007/978-3-030-30760-8_6
- [16] L. Federer. 2020. Measuring and Mapping Data Reuse: Findings From an Interactive Workshop on Data Citation and Metrics. *Harvard Data Science Review* 2, 2 (2020).
- [17] E. Fragkiadaki and G. Evangelidis. 2014. Review of the indirect citations paradigm: theory and practice of the assessment of papers, authors and journals. *Scientometrics* 99, 2 (2014), 261–288. doi:10.1007/s11192-013-1175-5
- [18] A. Ghorbani and J. Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proceedings of the 36th International Conference on Machine Learning*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, 2242–2251.
- [19] Jie Huang and Kevin Chen-Chuan Chang. 2024. Citation: A Key to Building Responsible and Accountable Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics, Mexico City, Mexico, 464–473. doi:10.18653/v1/2024.findings-naacl.31
- [20] O. Irrera, A. Mannocci, P. Manghi, and G. Silvello. 2023. Tracing Data Footprints: Formal and Informal Data Citations in the Scientific Literature. In *Linking Theory and Practice of Digital Libraries: 27th International Conference on Theory and Practice of Digital Libraries, TPDL 2023, Proceedings (Lecture Notes in Computer Science)*. Springer, 79–92. doi:10.1007/978-3-031-43849-3_7
- [21] D. S. Katz. 2014. Transitive Credit as a Means to Address Social and Technological Concerns Stemming from Citation and Attribution of Digital Products. *Journal of Open Research Software* 2, 1 (2014), e20. doi:10.5334/jors.be
- [22] J. E. Kratz and C. Strasser. 2015. Making Data Count. *Scientific Data* 2 (2015), 150039. doi:10.1038/sdata.2015.39
- [23] S. Lafia, A. Thomer, E. Moss, D. Bleckley, and L. Hemphill. 2023. How and Why Do Researchers Reference Data? A Study of Rhetorical Features and Functions of Data References in Academic Articles. *Data Science Journal* 22 (2023), 10. doi:10.5334/dsj-2023-010
- [24] M. Liu. 1993. Progress in Documentation the Complexities of Citation Practice: A Review of Citation Studies. *Journal of Documentation* 49, 4 (1993), 370–408. doi:10.1108/eb026920

- [25] S. Longpre, R. Mahari, A. Chen, N. Obeng-Marnu, D. Sileo, W. Brannon, N. Muennighoff, N. Khazam, J. Kabbara, K. Perisetla, X. Wu, E. Shippole, K. Bollacker, T. Wu, L. Villa, S. Pentland, and S. Hooker. 2024. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence* 6, 8 (2024), 975–987. doi:10.1038/s42256-024-00878-8
- [26] National Information Standards Organization. 2022. CRediT: Contributor Roles Taxonomy (ANSI/NISO Z39.104-2022). <https://credit.niso.org/>.
- [27] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36, 7 (2024), 3580–3599. doi:10.1109/TKDE.2024.3352100
- [28] A. Rauber, M. Parsons, and RDA Data Citation WG. 2025. Precise, Actionable Reference to Dynamic Data: Recommendations of the Working Group on Data Citation (WGDC). doi:10.15497/RDA00132
- [29] G. Silvello. 2018. Theory and Practice of Data Citation. *Journal of the Association for Information Science and Technology* 69, 1 (2018), 6–20. doi:10.1002/asi.23917
- [30] J. Wallat, M. Heuss, M. de Rijke, and A. Anand. 2025. Correctness is not Faithfulness in Retrieval Augmented Generation Attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, ICTIR 2025*. ACM Press, 22–32. doi:10.1145/3731120.3744592
- [31] P. Wei Koh and P. Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. PMLR, 1885–1894.
- [32] M. D. Wilkinson, M. Dumontier, Ij. J. Aalbersberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (2016), 160018. doi:10.1038/sdata.2016.18
- [33] Y. Wu, A. Alawini, S. B. Davidson, and G. Silvello. 2018. Data Citation: Giving Credit Where Credit is Due. In *Proc. of the 2018 International Conference on Management of Data, SIGMOD Conference 2018*. ACM Press, 99–114. doi:10.1145/3183713.3196910