Knowledge Graphs Data Quality at Scale: Challenges, Solutions, and Applications

Stefano Marchesin
stefano.marchesin@unipd.it
University of Padua
Padua, Italy

Gianmaria Silvello
gianmaria.silvello@unipd.it
University of Padua
Padua, Italy

Omar Alonso*
omralon@amazon.com
Amazon
Santa Clara, California, USA

Abstract

Knowledge Graphs (KGs) are foundational infrastructure for search, question answering, and entity-centric applications, on the open web and within domain-specific platforms. Their lifecycle, from construction over heterogeneous sources to evaluation at scale and use in downstream tasks, raises a coupled set of data management problems that shape the reliability and utility of every system built on top of them. This tutorial offers a high-level view of KG data quality at scale, organized along KG lifecycle stages. We first frame KG construction and the open challenges that arise when facts are extracted from noisy, evolving sources. We then turn to evaluation, covering efficient quality estimation, as well as the role of Large Language Models (LLMs) as validators and auxiliary signals. We close with downstream applications, including entity-oriented search and scalable defect detection, which translate evaluation outcomes into operational decisions.

Keywords

Data Quality, Knowledge Graphs

ACM Reference Format:
Stefano Marchesin, Gianmaria Silvello, and Omar Alonso. 2018. Knowledge Graphs Data Quality at Scale: Challenges, Solutions, and Applications. In *Proceedings of 35th International ACM Conference on Knowledge and Information Management (CIKM '26).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Knowledge Graphs (KGs), structured repositories of factual information encoded as subject-predicate-object (s, p, o) triples, underpin a broad range of information systems, including search engines, question-answering systems, and entity-centric applications [11, 28, 32]. Open and industrial KGs [3, 13, 14, 25] are populated from heterogeneous, often conflicting sources, which makes them inherently noisy and difficult to keep up to date [7, 27, 34]. The data quality of a KG therefore depends on choices made along the development and maintenance lifecycle: how facts are extracted and integrated, how their reliability is estimated, and how the resulting uncertainty is propagated to downstream tasks [23]. Errors that survive construction propagate into search, recommendation, and Large Language Model (LLM)-grounded applications, where they are costly to detect and even more costly to roll back.

This tutorial provides a high-level view of KG data quality at scale, organized along the three stages of the lifecycle: *construction*, *evaluation*, and *downstream use*. We frame the open challenges of KG construction in evolving settings, cover the statistical machinery used to obtain reliable quality estimates under limited budgets, and examine how evaluation outcomes feed downstream applications such as quality-aware entity-oriented search and scalable defect detection. Across the three parts, we connect statistical theory [4, 16] to deployment practice and identify open research directions at the intersection of data management and information retrieval.

2 Relevance and Novelty

KG quality has been surveyed at the conceptual level [11, 23, 34] and at the system level for construction and serving [5, 13, 14, 25]. Existing tutorials in the CIKM community have addressed KG construction and embedding methods, but none has so far offered a unified treatment of the data quality problem across the three stages we cover here. The proposal targets a gap in the data management and information retrieval literature that is increasingly visible as KGs are deployed in production search, recommendations, and question answering systems, and as LLMs start to significantly rely on KG data for grounding. By bringing together construction, evaluation, and downstream applications under a single framing, the tutorial provides attendees with a cross-cutting view of how data quality decisions propagate through a real KG-powered system.

3 Target Audience, Prerequisites, and Benefits

Audience. The tutorial targets researchers and practitioners in information retrieval, data management, and natural language processing who design, evaluate, or deploy KG-based systems. It is equally relevant to academic researchers working on construction and evaluation methodologies and to applied practitioners running KGs in production. We expect interest from attendees working on data quality, information extraction, entity-oriented search, and LLM grounding.

Prerequisites. Attendees are expected to be familiar with the basic structure of RDF KGs (triples, predicates, entities) and with elementary probability and statistics (expectation, variance, confidence intervals). No prior knowledge of survey sampling theory, Bayesian inference, or LLM internals is required.

*Work does not relate to the author's position at Amazon.

CIKM '26, Rome, Italy

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

Benefits. Participants will: (i) acquire a working understanding of KG construction at scale and of the data quality challenges that emerge in evolving settings; (ii) learn the statistical principles underlying KG quality estimation and the failure modes of naive approaches; (iii) see how evaluation extends to utility-weighted, entity-centric, and LLM-assisted settings; (iv) understand how evaluation results feed downstream search and defect detection; and (v) identify open research problems at the intersection of data management, information retrieval, and natural language processing.

4 Tutorial Outline (Half-Day, 3 Hours)

The tutorial is organized along the three stages of the KG lifecycle: construction, evaluation, and downstream use. Each part is self-contained yet builds on the previous one, so that attendees can follow how data quality decisions taken at construction time propagate through the evaluation machinery and into the systems that consume the KG.

4.1 Part 1 — KG Construction and Data Quality Challenges (60 min)

Part 1 sets the context. We first walk through a representative KG construction pipeline, then frame the open data quality challenges that arise once KGs are deployed and evolve.

1.1 Constructing KGs at scale (35 min). We use the CORE system [18] as a use case scenario of a KG construction pipeline operating under limited annotation budgets. CORE combines distant supervision and active learning over a modular architecture comprising five components: data acquisition and named entity recognition/disambiguation, manual annotation of seed examples, relation extraction training, large-scale automated annotation, and KG population. Three workflows orchestrate these components: a *bootstrapping* workflow in which domain experts annotate a seed set to train relation extractors; a *deployment* workflow in which trained models scale annotation over the full corpus; and an *active learning* workflow that surfaces low-confidence predictions to experts for iterative refinement. We then connect these design choices to broader literature on KG construction [5, 14, 25, 32], highlighting where data quality issues are introduced along the pipeline, from entity ambiguity and relation noise at extraction time to coverage skew [30] between head and tail entities at KG-population time.

1.2 From construction to evaluation: data quality challenges (25 min). Once deployed, the KG quality is shaped by forces that the rest of the tutorial addresses in turn. Sources change, disagree, or drift; previously verified facts can become stale or inconsistent as the KG evolves [27]; and different subsets of the KG serve different workloads, so the cost-benefit of verification is not uniform. Source-centric curation [23, 26, 34] and automated constraint checking (e.g., SHACL) are necessary but insufficient: subtle semantic and temporal errors escape structural checks, and large-scale manual re-verification is prohibitive. We review these failure modes and their cost structure to frame the central question of the tutorial: how to estimate, with controlled cost and statistical reliability, the fraction of correct facts in a KG, and how to channel the resulting evidence into downstream pipelines. This sets the agenda for Part 2 (evaluation methodology) and Part 3 (downstream applications).

4.2 Part 2 — Evaluation Methodologies (75 min)

Part 2 covers the statistical techniques used to estimate KG quality at scale and to support cost-aware verification, ranging from population-level to utility-oriented and LLM-assisted designs.

2.1 Sampling-based quality estimation (20 min). We define KG quality as the fraction of correct triples, defined globally, per predicate, or per entity cluster depending on the downstream target. We introduce the annotation cost model of [8], in which cost is proportional to both the number of entity clusters touched during annotation and the raw triple count: annotating multiple triples from the same cluster amortizes the context-switch overhead that dominates expert effort in practice. Building on this model, we walk through the principal sampling designs, stating the estimand, deriving the unbiased estimator, and quantifying the variance for each. *Simple random sampling* provides the reference; *cluster sampling* groups triples by entity cluster to exploit the within-cluster context savings; *stratified sampling* partitions the entity-cluster population into homogeneous strata to reduce within-stratum variance. We discuss when each design is appropriate and how incremental extensions, including weighted reservoir sampling [6], are used in evolving settings. On real-world KGs, these designs reduce annotation cost by up to ∼60% over uniform baselines on static KGs, and by up to ∼80% when updates are exploited as additional strata [8].

2.2 From confidence intervals to credible intervals (20 min). We explain why the choice of interval, not just the point estimate, matters for KG quality reporting [19, 20]. The Wald confidence interval, used by default in prior KG estimation work, relies on a normal approximation that breaks down precisely in the regime KG evaluation operates in: small to moderate sample sizes and high KG quality. The Wilson confidence interval, adapted to cluster and stratified designs in [19], eliminates Wald failure modes and improves estimation reliability by up to ∼2× on real-world KGs. We then introduce Bayesian credible intervals based on Highest Posterior Density (HPD) regions [20], motivated by three properties: *one-shot interpretation*, where a 95% credible interval admits the direct probabilistic statement that true quality lies within the interval with probability 0.95 given the data [10, 24]; *natural post-data inference* via the posterior distribution; and *maximal precision*, since HPD regions are the shortest possible intervals at a given credibility level. Hence, credible intervals deliver stronger post-data guarantees than Wald and Wilson intervals across real-world KGs [20].

2.3 Utility-oriented and budget-aware estimation (15 min). Standard quality estimation allocates annotation budget uniformly, regardless of which facts matter most to downstream applications. We cover utility-oriented designs that relax this assumption when the KG serves a known workload [21]. We discuss the practical issues that arise when the utility distribution is heavy-tailed (bias control toward popular entities) and when logs are unavailable (cold start), as well as how the design composes with the statistical techniques from Part 2.2.

2.4 LLMs for KG quality assessment (20 min). This unit equips attendees to reason about whether and how LLMs can support KG quality assessment. We distinguish two operational roles. In the *validator* role, attendees will learn how to benchmark LLMs

against human judgments and across operational configurations (internal knowledge, retrieval augmentation, multi-model consensus) [21, 29]. Current evidence shows that LLMs are not yet a reliable substitute for human annotators: they exhibit systematic biases such as overestimating accuracy relative to expert judges, and gains from retrieval augmentation and multi-model consensus are inconsistent. Attendees will therefore learn what to test for before trusting an LLM validator and why direct substitution for human labels remains premature. In the *stratification signal* role, attendees will see how aggregated LLM predictions can still drive sampling design without requiring the LLM to be correct, by partitioning a KG into strata with internally homogeneous quality [17]; a lightweight distillation step on a small LLM-annotated seed makes the signal usable at scale. Attendees leave with a clear position on which LLM uses are currently safe (stratification signals for cost reduction) and which remain premature (direct validation).

4.3 Part 3 — Downstream Applications and Open Problems (45 min)

Part 3 concludes the tutorial by showing how evaluation outcomes are consumed by downstream systems, with two representative use cases.

3.1 Veracity in entity-oriented search (15 min). Entity search systems retrieve entities and generate entity cards from KG facts, yet existing evaluation frameworks assume the underlying KG is accurate and measure relevance alone [9]. A relevant but incorrect fact in an entity card degrades user trust and decision quality in ways that relevance metrics alone cannot detect. We present a veracity-aware assessment methodology that treats relevance and veracity as orthogonal axes [22], the corresponding annotation protocol, and how to allocate the budget to quantify the impact of KG inaccuracies on entity rankings and card composition. We then show how to integrate veracity signals into existing evaluation pipelines without re-annotating relevance from scratch, preserving comparability with prior benchmarks while exposing the veracity dimension.

3.2 Defect detection at scale (15 min). We present two complementary techniques for scaling veracity-aware evaluation to large KGs and entity-centric workloads [15]. *eRank* is an entity-level, veracity-driven re-ranking metric that lets a system trade relevance against veracity in a controlled way: across 12 retrieval systems on DBpedia-Entity v2 [9], eRank improves veracity gain by up to 34% without degrading nDCG. *ALADDIN* (Active Learning-based verAcity-Driven Defect IdentificatioN) is a lightweight pipeline for fact-level defect detection with a two-stage architecture: an offline regression model trained on KG fact embeddings, and an online active learning mechanism that uses entity cards as the annotation interface and a residual-based selection strategy to surface the most informative triples for iterative human feedback. ALADDIN reaches up to 127% precision improvements over standard, offline baselines under limited annotation budgets. We discuss how these methods plug into the evaluation framework from Part 2 and how their outputs feed back into construction-time curation.

3.3 Open problems and outlook (15 min). We close with a structured discussion of open research directions. *Query-level and task-conditional quality evaluation* conditions estimates on a specific retrieval or inference task rather than the global KG population, connecting to approximate query processing on graphs [31]. *Temporal dynamics and incremental re-evaluation* address how to maintain reliable estimates as facts are added or revised in continuously updated KGs [27]. *Cross-KG evaluation and transferability* examine whether distilled models for stratification signals trained on one KG generalize to others, with implications for the broader literature on knowledge distillation of LLMs [33]. We also discuss the interaction between KG quality estimation and LLM-based applications, including hallucination measurement and long-form factuality evaluation [12], where reliable KG-side evidence is increasingly used as grounding.

5 Prior Presentations

This tutorial has not been presented previously. Most individual research results covered here were presented by the authors as contributed papers in Database (2023) [18], PVLDB 2024 [19], HCOMP 2024 [21], CIKM 2024 [22], SIGMOD 2025 [20], CIKM 2025 [15], EDBT 2026 [29], and PVLDB 2026 [17]. The tutorial integrates these results into a unified framing of KG data quality across the construction, evaluation, and downstream-use stages, with explicit connections to the broader literature and expanded coverage of open problems and practical considerations not available in any individual paper.

6 Presenter Biographies

Stefano Marchesin. is an Assistant Professor in the Department of Information Engineering at the University of Padua, Italy. His research addresses data quality, KG construction and evaluation, and information retrieval. He has published at PVLDB, Proc. ACM Manag. Data (SIGMOD), CIKM, EDBT, HCOMP, SIGIR, ECIR, and ACM TOIS. He is lead or main co-author of most papers covered in this tutorial and and has presented research at international database and information retrieval conferences. He serves on program committees including VLDB, ICDE, SIGIR, CIKM, WWW, WSDM, ECIR, and ICTIR.

Gianmaria Silvello. is a Full Professor in the Department of Information Engineering at the University of Padua, Italy. His research focuses on data management, knowledge graphs, information retrieval evaluation, and scientific data infrastructures, with emphasis on the representation, curation, and reuse of structured data. He has published at PVLDB, Proc. ACM Manag. Data (SIGMOD), ACM TOIS, CIKM, SIGIR, ECIR, and EDBT. He has co-authored most papers covered in this tutorial and has long-standing experience in the design and evaluation of large-scale information access systems. He has served as program co-chair, area chair, and program committee member at major data management and information retrieval venues, including Proc. ACM Manag. Data (SIGMOD), PVLDB, SIGIR, CIKM, and ECIR.

Omar Alonso. is a Principal Scientist at Amazon, Santa Clara, California. His research focuses on search, knowledge graphs, human computation, and evaluation methodology, with substantial work on the use of crowdsourcing for relevance assessment and quality control [1]. He has co-edited a recent book on IR [2]. He has

co-authored the tutorial papers on KG construction [18], utility-oriented KG accuracy estimation [21], veracity in entity-oriented search [22], and scalable defect detection [15]. He has extensive experience connecting academic research on KG quality with industrial-scale deployment challenges, and has delivered tutorials and keynotes at SIGIR, WSDM, CIKM, and ECIR.

7 Tutorial Materials and Reproducibility

All experimental results discussed in the tutorial are reproducible from publicly available artifacts. Sampling and estimation code, LLM stratification scripts, and the eRank/ALADDIN pipelines are released at https://github.com/KGAccuracyEval, together with annotated benchmark subsets and utility logs used in [15, 21]. The CORE construction pipeline from [18] is available at https://github.com/GDAMining/core and the derived KG at https://zenodo.org/record/7577127. Attendees will receive slides, an annotated reading list organized by the three parts, and a step-by-step Jupyter notebook reproducing the headline results of [17, 19, 20] on a small KG sample suitable for laptop execution.

8 Resource Requirements

Standard projector and screen. No special hardware or software requirements; attendees interested in running the accompanying notebook need a standard Python environment with `numpy`, `scipy`, and `rdflib`.

References

[1] O. Alonso. 2019. *The Practice of Crowdsourcing*. Morgan & Claypool Publishers. doi:10.2200/S00904ED1V01Y201903ICR066

[2] O. Alonso and R. Baeza-Yates (Eds.). 2024. *Information Retrieval: Advanced Topics and Techniques*. ACM Books, Vol. 60. ACM.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007 (LNCS, Vol. 4825)*. Springer, 722–735. doi:10.1007/978-3-540-76298-0_52

[4] W. G. Cochran. 1977. *Sampling Techniques, 3rd Edition*. John Wiley. doi:10.1017/S0013091500025724

[5] O. Deshpande, D. S. Lamba, M. Tourn, S. Das, S. Subramaniam, A. Rajaraman, V. Harinarayan, and A. Doan. 2013. Building, maintaining, and using knowledge bases: a report from the trenches. In *Proc. of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013*. ACM, 1209–1220. doi:10.1145/2463676.2465297

[6] P. S. Efraimidis and P. G. Spirakis. 2006. Weighted Random Sampling with a Reservoir. *Inf. Process. Lett.* 97, 5 (2006), 181–185. doi:10.1016/j.ipl.2005.11.003

[7] D. Esteves, A. Rula, A. J. Reddy, and J. Lehmann. 2018. Toward Veracity Assessment in RDF Knowledge Bases: An Exploratory Analysis. *ACM J. Data Inf. Qual.* 9, 3 (2018), 16:1–16:26. doi:10.1145/3177873

[8] J. Gao, X. Li, Y. E. Xu, B. Sisman, X. L. Dong, and J. Yang. 2019. Efficient Knowledge Graph Accuracy Evaluation. *Proc. VLDB Endow.* 12, 11 (2019), 1679–1691. doi:10.14778/3342263.3342642

[9] F. Hasibi, K. Balog, and S. E. Bratsberg. 2017. Dynamic Factual Summaries for Entity Cards. In *Proc. of SIGIR*. ACM, 773–782. doi:10.1145/3077136.3080810

[10] R. Hoekstra, R. D. Morey, J. N. Rouder, and E. J. Wagenmakers. 2014. Robust Misinterpretation of Confidence Intervals. *Psychon. Bull. Rev.* 21 (2014), 1157–1164. doi:10.3758/s13423-013-0572-3

[11] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. Labra Gayo, R. Navigli, S. Neumaier, A. C. Ngonga Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann. 2021. *Knowledge Graphs*. Morgan & Claypool Publishers. doi:10.2200/S01125ED1V01Y202109DSK022

[12] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2 (2025), 42:1–42:55. doi:10.1145/3703155

[13] I. F. Ilyas, J. Lacerda, Y. Li, U. F. Minhas, A. Mousavi, J. Pound, T. Rekatsinas, and C. Sumanth. 2023. Growing and Serving Large Open-domain Knowledge Graphs. In *Companion of the 2023 International Conference on Management of Data, SIGMOD/PODS 2023, Seattle, WA, USA, June 18-23, 2023*. ACM, 253–259. doi:10.1145/3555041.3589672

[14] I. F. Ilyas, T. Rekatsinas, V. Konda, J. Pound, X. Qi, and M. A. Soliman. 2022. Saga: A Platform for Continuous Construction and Serving of Knowledge at Scale. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*. ACM, 2259–2272. doi:10.1145/3514221.3526049

[15] O. Irrera, S. Marchesin, G. Silvello, and O. Alonso. 2025. Scaling Trust: Veracity-Driven Defect Detection in Entity Search. In *Proc. of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025*. ACM, 1001–1012. doi:10.1145/3746252.3761208

[16] L. Kish. 1965. *Survey Sampling*. Wiley. https://books.google.it/books?id=xiZmAAAAIAAJ

[17] S. Marchesin, M. Ceccarello, and G. Silvello. 2026. LLMs as Stratification Signals for KG Accuracy Evaluation. *Proc. VLDB Endow. (accepted)* 19, 12. https://www.dei.unipd.it/~marches1/papers/mcs-vldb-2026.pdf

[18] S. Marchesin, L. Menotti, F. Giachelle, G. Silvello, and O. Alonso. 2023. Building a Large Gene Expression-Cancer Knowledge Base with Limited Human Annotations. *Database J. Biol. Databases Curation* 2023 (2023). doi:10.1093/DATABASE/BAAD061

[19] S. Marchesin and G. Silvello. 2024. Efficient and Reliable Estimation of Knowledge Graph Accuracy. *Proc. VLDB Endow.* 17, 9 (2024), 2392–2404. doi:10.14778/3665844.3665865

[20] S. Marchesin and G. Silvello. 2025. Credible Intervals for Knowledge Graph Accuracy Estimation. *Proc. ACM Manag. Data* 3, 3 (2025), 142:1–142:26. doi:10.1145/3725279

[21] S. Marchesin, G. Silvello, and O. Alonso. 2024. Utility-Oriented Knowledge Graph Accuracy Estimation with Limited Annotations: A Case Study on DBpedia. *Proc. of the 12th AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2024, Pittsburgh, Pennsylvania, USA, October 16–19, 2024* 12, 1 (2024), 105–114. doi:10.1609/hcomp.v12i1.31605

[22] S. Marchesin, G. Silvello, and O. Alonso. 2024. Veracity Estimation for Entity-Oriented Search with Knowledge Graphs. In *Proc. of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*. ACM, 1649–1659. doi:10.1145/3627673.3679561

[23] S. Mohammed, L. Ehrlinger, H. Harmouch, F. Naumann, and D. Srivastava. 2025. The Five Facets of Data Quality Assessment. *SIGMOD Rec.* 54, 2 (2025), 18–27. doi:10.1145/3749116.3749120

[24] R. D. Morey, R. Hoekstra, J. N. Rouder, M. D. Lee, and E. J. Wagenmakers. 2016. The Fallacy of Placing Confidence in Confidence Intervals. *Psychon. Bull. Rev.* 23 (2016), 103–123. doi:10.3758/s13423-015-0947-8

[25] N. F. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor. 2019. Industry-scale Knowledge Graphs: Lessons and Challenges. *Commun. ACM* 62, 8 (2019), 36–43. doi:10.1145/3331166

[26] H. Paulheim. 2017. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web* 8, 3 (2017), 489–508. doi:10.3233/SW-160218

[27] A. Polleres, R. Pernisch, A. Bonifati, D. Dell'Aglio, D. Dobriy, S. Dumbrava, L. Etcheverry, N. Ferranti, K. Hose, E. Jiménez-Ruiz, M. Lissandrini, A. Scherp, R. Tommasini, and J Wachs. 2023. How Does Knowledge Evolve in Open Knowledge Graphs? *TGDK* 1, 1 (2023), 11:1–11:59. doi:10.4230/TGDK.1.1.11

[28] R. Reinanda, E. Meij, and M. de Rijke. 2020. Knowledge Graphs: An Information Retrieval Perspective. *Found. Trends Inf. Retr.* 14, 4 (2020), 289–444. doi:10.1561/1500000063

[29] F. Shami, S. Marchesin, and G. Silvello. 2026. Benchmarking Large Language Models for Knowledge Graph Validation. In *Proc. 29th International Conference on Extending Database Technology, EDBT 2026, Tampere, Finland, March 24-27, 2026*. OpenProceedings.org, 551–565. doi:10.48786/EDBT.2026.45

[30] K. Sun, Y. E. Xu, H. Zha, Y. Liu, and X. L. Dong. 2024. Head-to-Tail: How Knowledgeable are Large Language Models (LLMs)? A.K.A. Will LLMs Replace Knowledge Graphs?. In *Proc. of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*. ACL, 311–325. doi:10.18653/V1/2024.NAACL-LONG.18

[31] Y. Wang, A. Khan, X. Xu, J. Jin, Q. Hong, and T. Fu. 2022. Aggregate Queries on Knowledge Graphs: Fast Approximation with Semantic-aware Sampling. In *38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022*. IEEE, 2914–2927. doi:10.1109/ICDE53745.2022.00263

[32] G. Weikum, X. L. Dong, S. Razniewski, and F. M. Suchanek. 2021. Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases. *Found. Trends Databases* 10, 2-4 (2021), 108–490. doi:10.1561/1900000064

[33] X. Xu, M. Li, C. Tao, T. Shen, R. Cheng, J. Li, C. Xu, D. Tao, and T. Zhou. 2024. A Survey on Knowledge Distillation of Large Language Models. *CoRR* abs/2402.13116 (2024). doi:10.48550/ARXIV.2402.13116

[34] B. Xue and L. Zou. 2023. Knowledge Graph Quality Management: A Comprehensive Survey. *IEEE Trans. Knowl. Data Eng.* 35, 5 (2023), 4969–4988. doi:10.1109/TKDE.2022.3150080

GenAI Usage Disclosure

The authors used generative AI tools (Claude) only to support grammar and spelling checks and, on occasion, to improve the wording of text originally written by the authors. All AI-assisted edits were reviewed and approved by the authors, who take full responsibility for the paper's content.