Redefining Value: Wikipedia's Role in Dataset Governance for Foundation Models

YAXUAN YIN, The Information School, University of Wisconsin–Madison, USA

Wikipedia has long been central to CSCW research on online collaboration, but it now plays a dual role: as both a collaborative knowledge platform and a critical dataset powering large language models (LLMs). This shift raises open questions about how contributions are valued, recognized, and governed when they influence downstream AI behavior. This position paper explores influence-based data valuation as one potential lens and invites discussion on how CSCW insights into collaboration and governance can inform responsible dataset composition, curation, and release.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

Additional Key Words and Phrases: Responsible AI, Foundation Models, Dataset Governance, Wikipedia, Generative AI, Data Valuation

ACM Reference Format:
Yaxuan Yin. 2018. Redefining Value: Wikipedia's Role in Dataset Governance for Foundation Models. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Within CSCW research, Wikipedia has long served as a central site for studying online collaboration. Prior studies have examined how volunteers coordinate [4], manage misinformation and vandalism [6], and navigate disparities across topics, languages, and regions [5]. These studies have shaped our understanding of what makes a contribution valuable within Wikipedia's community — contributions that improve accuracy, fill knowledge gaps, and strengthen collaborative dynamics. However, the rise of large language models (LLMs) introduces a new layer of complexity. Wikipedia now plays a dual role: it is both a collaborative knowledge platform and a critical dataset powering foundation models [9]. For example, Wikipedia pages are among the five primary datasets used to train GPT-3 [1], the FLORES-101 benchmark relies on Wikipedia for multilingual evaluation [3], and retrieval-augmented generation (RAG) pipelines treat Wikipedia as a key source of factual knowledge [7].

This shift introduces a new dimension of value: contributions to Wikipedia now matter not only for supporting human collaboration but also for shaping how AI systems behave and what knowledge they reproduce. Specifically, they influence who gets credited (attribution), whose perspectives are amplified or marginalized (representation), and who controls how data is used and governed (governance). Yet contributors currently lack transparency, recognition, and control over how their work is used in training datasets. At the same time, tensions between data providers and AI developers are intensifying, as reflected in emerging legal disputes over attribution, consent, and compensation [8].

Author's Contact Information: Yaxuan Yin, yaxuan.yin@wisc.edu, The Information School, University of Wisconsin–Madison, Madison, Wisconsin, USA.

Manuscript submitted to ACM

These dynamics open new opportunities for CSCW researchers to revisit longstanding questions of value, governance, and collaboration in a context where human and machine knowledge production are deeply intertwined. Understanding these evolving relationships is critical for shaping equitable, transparent, and sustainable data ecosystems in the era of LLMs.

2 Opportunity: Using Data Valuation as a Lens

One potential lens for exploring these issues comes from recent advances in data valuation methods. For example, Choe et al. [2] introduce techniques for estimating the influence of individual training examples on a model's performance. They provide a useful technical framework for reflecting on how "value" is defined when Wikipedia contributions shape downstream AI behaviors. At a high level, the approach takes three main inputs: (1) a trained language model, (2) a snapshot of the training dataset (e.g., Wikipedia), and (3) an evaluation benchmark, such as factual QA, multilingual reasoning, or fairness tasks like WinoBias [10]. Using these inputs, the method produces influence scores that estimate how much each data point contributes to the model's performance on the chosen benchmark. These scores can be aggregated at multiple levels — individual tokens, pages, or contributors — to understand which parts of Wikipedia most affect model outputs.

Applying this framework conceptually to Wikipedia allows us to connect influence scores to broader Responsible AI dimensions. For example, examining which pages or contributors most affect model outputs can reveal potential fairness and representation gaps, such as the dominance of English-language content in shaping LLM behavior [**?**]. Influence scores can also enhance transparency by linking model responses back to specific pages or contributors, enabling better attribution and accountability. These insights provide a foundation for rethinking dataset composition, curation practices, and how community norms might be incorporated into foundation model development. However, data valuation is not the endpoint, it is a starting point for new CSCW discussions. Influence scores highlight what matters but not why, and they cannot explain the social, cultural, or governance dynamics behind Wikipedia contributions — for instance, it is taken as granted that Wikipedia is a reliable corpus for training, but such influence scores cannot account for the peer production processes that results in Wikipedia being a high-quality information resource. This opens important questions for CSCW researchers: How should we redefine "value" when Wikipedia edits shape downstream AI behaviors? How can, and should, data valuation insights be integrated with community-driven governance processes? And how might these techniques support more equitable and participatory approaches to dataset curation?

3 Towards Participatory Dataset Governance

Wikipedia offers more than training data, it represents a living example of participatory governance at scale. Over two decades, editors have collaboratively developed policies, resolved disputes, and balanced openness with cultural sensitivities across regions and languages. These practices offer valuable insights into how datasets for foundation models could be curated and managed. We envision combining technical insights from data valuation with community-driven governance principles to establish participatory approaches to responsible dataset curation. Influence scores can help identify which contributions most strongly shape model behavior, while Wikipedia's governance norms can inform how these contributions are reviewed, recognized, and included. By integrating these perspectives, CSCW researchers can help bridge the gap between technical valuation methods and social governance frameworks, ensuring that dataset composition, curation, and release reflect both responsible AI principles and community values.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. 1877–1901.

[2] Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric Xing. 2024. What Is Your Data Worth to GPT? LLM-Scale Data Valuation with Influence Functions. *arXiv preprint arXiv:2405.13954* (2024). https://arxiv.org/abs/2405.13954

[3] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics* 10 (2022), 522–538. doi:10.1162/tacl_a_00474

[4] Aaron Halfaker, R.Stuart Geiger, Jonathan Morgan, and John Riedl. 2013. The Rise and Decline of an Open Collaboration System How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist* 57 (05 2013), 664–688. doi:10.1177/0002764212469365

[5] Molly G. Hickman, Viral Pasad, Harsh Kamalesh Sanghavi, Jacob Thebault-Spieker, and Sang Won Lee. 2021. Understanding Wikipedia Practices Through Hindi, Urdu, and English Takes on an Evolving Regional Conflict. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 34 (April 2021), 31 pages. doi:10.1145/3449108

[6] Sara Javanmardi, David W. McDonald, and Cristina V. Lopes. 2011. Vandalism detection in Wikipedia: a high-performing, feature-rich model and its reduction through Lasso. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (Mountain View, California) *(WikiSym '11)*. Association for Computing Machinery, New York, NY, USA, 82–90. doi:10.1145/2038558.2038573

[7] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 9459–9474.

[8] Cade Metz and Katie Robertson. 2024. OpenAI Seeks to Dismiss Parts of The New York Times's Lawsuit. *The New York Times* (February 2024). Discusses legal challenges related to how AI systems are built and trained.

[9] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why we read Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1591–1600. doi:10.1145/3038912.3052716

[10] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of NAACL*.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009