Evaluations Using Wikipedia without Data Contamination: From Trusting Articles to Trusting Edit Processes

Lucie-Aimée Kaffee
Hugging Face
lucie.kaffee@huggingface.co

Isaac Johnson
Wikimedia Foundation
isaac@wikimedia.org

In the evolving landscape of artificial intelligence and machine learning, Wikipedia has emerged as a pivotal resource for factual information. As one of the most comprehensive and accessible repositories of human knowledge, it serves as a critical reference point in many contexts, including the evaluation of machine learning models. However, the integration of Wikipedia into the training data for numerous AI models introduces a unique challenge: the potential circularity in using Wikipedia as both a training source and an evaluation benchmark. This provocations paper explores the implications of this dual role, questioning the reliability and objectivity of evaluations that rely on a dataset inherently intertwined with the models being assessed. By critically examining the use of Wikipedia in factual evaluations, this paper aims to provoke discussion on the validity of current evaluation methodologies and the necessity of developing more robust, diverse, and independent benchmarks. Specifically, we recommend revising benchmarks that capture a static snapshot of Wikipedia content to instead adopt a model based on the continuous work of Wikipedia editors to build a trustworthy encyclopaedia.

Wikipedia as a Factual Evaluation Benchmark

Wikipedia has long played a crucial role in the evaluation of machine learning models. It has been used to create benchmarks ranging from Question Answering (e.g., SQuAD [12], MLQA [9]) to fact-checking (FEVER [16]) and evaluating hallucinations, i.e., factual inaccuracies (FActScore [10]) to capturing nuances around achieving a fair and balanced presentation of information using Wikipedia's Neutral Point of View (NPOV) policy [1, 11].

However, a significant challenge in evaluating models arises from data contamination, which occurs when models are tested on data they have already encountered during training [2]. Given that Wikipedia is a widely used and easily accessible resource, it is often included in the training datasets of state-of-the-art LLMs [5]. Even when Wikipedia is not explicitly including in training data, it is a substantial part of Common Crawl[1], a massive web dataset frequently used for model training.

The issue of data contamination and leakage is not new. Song and Raghunathan [15] demonstrated that it is possible to design attacks that leak training data using Wikipedia in word embeddings. Similarly, Xin et al. [17] find data leakage for different pre-training language models. While their work focuses on privacy, this kind of leakage also has implications for the evaluation of LLMs. When evaluation datasets contain data that has been used in training, the evaluation is not meaningful. Especially in the context of closed-source models, where we do not have knowledge about which datasets have been used for training, evaluations based on facts become meaningless, as Balloccu et al. [2] detail.

Given that Wikipedia data is likely included in most training datasets, evaluations based on the same set of data should be approached with caution. Indeed, Zhao et al. [18] find that models hallucinate more about entities that do not have a Wikipedia page, underscoring the impact of Wikipedia's presence in training data on model behaviour.

[1] https://commoncrawl.github.io/cc-crawl-statistics/plots/domains

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

The Dynamic Nature of Wikipedia: Beyond Factual Benchmarks

A common pitfall of these Wikipedia-based benchmarks that have data contamination problems is that they approach Wikipedia as a static source of content from which to build a dataset (generally with some support from crowd-workers). Wikipedia, however, is more accurately a dynamic community that is constantly building an encyclopaedia through contestation and consensus. A deeper look into these processes often reveals ways to build interesting benchmarks using the judgments from Wikipedians themselves. By learning from the way Wikipedia communities engage with information, resolve disputes, and update content, we can design evaluations that better reflect the complexities of knowledge production and verification, moving beyond a simplistic reliance on factual correctness. In particular, we propose two important principles for building Wikipedia-based benchmarks that avoid this issue of data contamination: 1) identifying content created after a specific knowledge cut-off that is of high-quality, and, 2) identifying actions taken by Wikipedians to curate content that can then be leveraged to test the ability of language models to understand and evaluate content.

Existing examples of an approach using a knowledge cut-off include the benchmark FreshWiki [13], in which the authors gather recently created Wikipedia articles with a focus on avoiding data contamination during pretraining. As the authors create a model, they can ensure that no data from the test set is part of the training. However, for evaluating closed-source models, this approach cannot be guaranteed and should be addressed. The publication of date cut-offs is therefore an important part of disclosure for closed source models[2].

Leveraging the Wikipedia editors content curation actions for benchmarks has many facets - from investigating contradicting information [6] to transparent stance detection based on editor discussions [7]. The information that is created by editors but typically not used for the training of LLMs can be a rich source of evaluating different aspects of models, using Wikipedia processes to generate benchmarks dynamically without requiring substantial additional curation.

All of these examples do not need additional human labour but are based on existing work of Wikipedia editors. Including editors in the creation of benchmarks e.g., as done by Wikibench [8] or similar to include affected communities in red teaming efforts [4], can be a valuable approach to have the Wikipedia's communities' direct feedback. However, including Wikipedia editors into benchmark creation for AI should be approached with caution; *participation cannot be a design fix for machine learning* [14]. It is important to keep in mind questions which would be relevant to including most communities directly into such efforts [3] - what is the benefit for the community, how to contribute back to community, and how to make such benchmarks sustainable so that they don't have to be recreated regularly. Wikipedia's editors' focus should not be shifted away from the task they already excel in: writing an encyclopaedia across a large number of languages.

Discussion & Conclusion

The dual role of Wikipedia as both a training source and an evaluation benchmark in AI presents significant challenges to the validity of current evaluation practices. While Wikipedia provides a rich and easily accessible source of factual information, the risk of data contamination undermines its reliability as an evaluation tool. Moreover, the dynamic, community-driven nature of Wikipedia offers an opportunity to rethink how we approach AI evaluation. Instead of relying on static factual benchmarks, AI researchers should consider frameworks that reflect the ongoing, collaborative process of knowledge creation seen in Wikipedia.

This shift would not only address the issues of data contamination and circularity but also promote a more nuanced understanding of how knowledge is constructed, contested, and validated in real-world contexts. To this end we propose for future benchmarks to (1) identify content created after a specific knowledge cut-off that is of high-quality, and, (2) identify actions taken by Wikipedians to curate content that can then be leveraged to test the ability of language models to understand and evaluate content.

As AI continues to integrate more deeply into various domains, developing diverse and independent evaluation frameworks is critical to ensuring the reliability and fairness of AI systems. By moving beyond Wikipedia as a static source of facts and embracing its collaborative nature, we can create evaluation methods that better align with the complex, dynamic realities of knowledge production.

[2] `https://openfuture.eu/publication/towards-robust-training-data-transparency/`

References

[1] J. Ashkinaze, R. Guan, L. Kurek, E. Adar, C. Budak, and E. Gilbert. Seeing like an AI: how llms apply (and misapply) wikipedia neutrality norms. *CoRR*, abs/2407.04183, 2024. doi: 10.48550/ARXIV.2407.04183. URL https://doi.org/10.48550/arXiv.2407.04183.

[2] S. Balloccu, P. Schmidtová, M. Lango, and O. Dusek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. In Y. Graham and M. Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 67–93. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.eacl-long.5.

[3] A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, and S. Mohamed. Power to the people? opportunities and challenges for participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO 2022, Arlington, VA, USA, October 6-9, 2022*, pages 6:1–6:8. ACM, 2022. doi: 10.1145/3551624.3555290. URL https://doi.org/10.1145/3551624.3555290.

[4] N. Dennler, A. Ovalle, A. Singh, L. Soldaini, A. Subramonian, H. Tu, W. Agnew, A. Ghosh, K. Yee, I. F. Peradejordi, Z. Talat, M. Russo, and J. de Jesus de Pinho Pinhal. Bound by the bounty: Collaboratively shaping evaluation processes for queer AI harms. In F. Rossi, S. Das, J. Davis, K. Firth-Butterfield, and A. John, editors, *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023, Montréal, QC, Canada, August 8-10, 2023*, pages 375–386. ACM, 2023. doi: 10.1145/3600211.3604682. URL https://doi.org/10.1145/3600211.3604682.

[5] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2021. URL https://arxiv.org/abs/2101.00027.

[6] Y. Hou, A. Pascale, J. Carnerero-Cano, T. T. Tchrakian, R. Marinescu, E. Daly, I. Padhi, and P. Sattigeri. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *CoRR*, abs/2406.13805, 2024. doi: 10.48550/ARXIV.2406.13805. URL https://doi.org/10.48550/arXiv.2406.13805.

[7] L. Kaffee, A. Arora, and I. Augenstein. Why should this article be deleted? transparent stance detection in multilingual wikipedia editor discussions. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5891–5909. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.361. URL https://doi.org/10.18653/v1/2023.emnlp-main.361.

[8] T. Kuo, A. L. Halfaker, Z. Cheng, J. Kim, M. Wu, T. Wu, K. Holstein, and H. Zhu. Wikibench: Community-driven data curation for AI evaluation on wikipedia. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. O. T. Dugas, and I. Shklovski, editors, *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 193:1–193:24. ACM, 2024. doi: 10.1145/3613904.3642278. URL https://doi.org/10.1145/3613904.3642278.

[9] P. S. H. Lewis, B. Oguz, R. Rinott, S. Riedel, and H. Schwenk. MLQA: evaluating cross-lingual extractive question answering. In D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7315–7330. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.653. URL https://doi.org/10.18653/v1/2020.acl-main.653.

[10] S. Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/

V1/2023.EMNLP-MAIN.741. URL https://doi.org/10.18653/v1/2023.emnlp-main.741.

[11] R. Pryzant, R. D. Martinez, N. Dass, S. Kurohashi, D. Jurafsky, and D. Yang. Automatically neutralizing subjective bias in text. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 480–489. AAAI Press, 2020. doi: 10.1609/AAAI.V34I01.5385. URL https://doi.org/10.1609/aaai.v34i01.5385.

[12] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100, 000+ questions for machine comprehension of text. In J. Su, X. Carreras, and K. Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/D16-1264. URL https://doi.org/10.18653/v1/d16-1264.

[13] Y. Shao, Y. Jiang, T. A. Kanell, P. Xu, O. Khattab, and M. S. Lam. Assisting in writing wikipedia-like articles from scratch with large language models. In K. Duh, H. Gómez-Adorno, and S. Bethard, editors, *Proceedings 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*, pages 6252–6278. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.347. URL https://doi.org/10.18653/v1/2024.naacl-long.347.

[14] M. Sloane, E. Moss, O. Awomolo, and L. Forlano. Participation is not a design fix for machine learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO 2022, Arlington, VA, USA, October 6-9, 2022*, pages 1:1–1:6. ACM, 2022. doi: 10.1145/3551624.3555285. URL https://doi.org/10.1145/3551624.3555285.

[15] C. Song and A. Raghunathan. Information leakage in embedding models. In J. Ligatti, X. Ou, J. Katz, and G. Vigna, editors, *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, pages 377–390. ACM, 2020. doi: 10.1145/3372297.3417270. URL https://doi.org/10.1145/3372297.3417270.

[16] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal. FEVER: a large-scale dataset for fact extraction and verification. In M. A. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics, 2018. doi: 10.18653/V1/N18-1074. URL https://doi.org/10.18653/v1/n18-1074.

[17] Y. Xin, Z. Li, N. Yu, D. Chen, M. Fritz, M. Backes, and Y. Zhang. Inside the black box: Detecting data leakage in pre-trained language encoders. *arXiv preprint arXiv:2408.11046*, 2024.

[18] W. Zhao, T. Goyal, Y. Y. Chiu, L. Jiang, B. Newman, A. Ravichander, K. R. Chandu, R. L. Bras, C. Cardie, Y. Deng, and Y. Choi. Wildhallucinations: Evaluating long-form factuality in llms with real-world entity queries. *CoRR*, abs/2407.17468, 2024. doi: 10.48550/ARXIV.2407.17468. URL https://doi.org/10.48550/arXiv.2407.17468.