Grounding on Shaky Ground: Wikipedia's Legal Articles, Editorial Integrity, and the Risk of Data Poisoning in Artificial Intelligence

An Empirical Study of Contributor Bans and Knowledge Reliability

Matthias Harter
Faculty of Engineering
Hochschule RheinMain - University of Applied Sciences
Rüsselsheim, Germany
e-mail: matthias.harter@hs-rm.de

***Abstract*—Analyzing revision histories of over 15,000 articles in the German-language Wikipedia legal domain from 2004 to 2025, this study examines the persistent infiltration of entries by contributors later permanently banned for vandalism, extremist propaganda, promotional editing, or uncooperative conduct. We quantify a non-trivial proportion of edits originating from compromised accounts, demonstrating how such editorial contamination degrades Wikipedia's reliability as a training corpus for Large Language Models (LLMs) in legal and media-content generation contexts where factual precision is critical. Our investigation further reveals that Retrieval-Augmented Generation (RAG) architectures, which ground outputs in external data, risk propagating inaccuracies if their source repositories are compromised. These findings have direct implications for trust and disinformation in AI media, ethical considerations in AI-generated content, and the evaluation of LLM-based tools, by highlighting vulnerabilities in open-source knowledge pipelines. Ultimately, our findings challenge assumptions about swarm intelligence and demonstrate the urgent need for robust safeguards to ensure reliable AI-driven media production workflows.**

Keywords*—*Wikipedia; Data poisoning; LLM; AI training data; Trustworthiness in AI; Crowdsourced content; Disinformation.

I. Introduction

Wikipedia has become a cornerstone of online knowledge dissemination, widely used not only by individuals seeking accessible explanations but increasingly as a foundational dataset for LLMs. Legal articles on Wikipedia, in particular, play a critical role in public access to complex statutory language and jurisprudence, often bridging the gap between technical legal terminology and lay understanding. Yet, despite its openness being a strength, Wikipedia remains vulnerable to bad-faith editorial behavior.

A. Related Work

The integrity of training data has emerged as a central concern in the development of LLMs. A growing body of research addresses the vulnerability of such models to *data poisoning*—intentional contamination of training or grounding data with misleading content. Reference [1] demonstrates that even a few hundred adversarial examples injected during instruction tuning can cause persistent and targeted misbehavior in LLMs. Similar risks were identified by [2], who show that in medical domains, poisoned inputs comprising less than 0.001% of training data can significantly distort output while maintaining benchmark performance.

Benchmarking frameworks such as PoisonBench [3] confirm these vulnerabilities at scale across multiple architectures and highlight the insufficiency of current alignment mechanisms to guard against subtle data corruption. These findings emphasize a broader pattern: scaling models does not inherently confer robustness.

Wikipedia plays a notable role in LLM training corpora. It accounts for approximately 3–5% of tokens in foundational models such as GPT-3 [4], LLaMA [5], and BERT [6]. Its inclusion is often motivated by its structured factual content and perceived reliability. However, sociotechnical investigations [7] question this assumption, arguing that the sustainability and neutrality of Wikipedia are increasingly threatened by automation, declining editor activity, and external exploitation.

Reference [8] in *Nature* compared 42 science articles and showed that Wikipedia's accuracy was broadly comparable to *Encyclopædia Britannica*. Yet later studies nuance this optimism. Using matched U.S. political topics, Greenstein and Zhu find Wikipedia to be more slanted toward Democratic viewpoints than Britannica and overall more biased, though the gap narrows with successive edits [9]. Extending the lens from content to contributors, persuasion [10] show that roughly 80–90% of the observed moderation in article slant is driven by the exit of highly partisan editors rather than by on-platform. Most recently, [11] employs large-scale sentiment analysis over 1,600 politically charged terms and documents a systematic tendency for right-of-centre public figures to be associated with more negative sentiment than their left-leaning counterparts. Together, these works suggest that while Wikipedia can rival expert sources on factual accuracy, ideological asymmetries persist and are shaped by community composition over time; our focus on banned-user infiltration builds on this line of inquiry by foregrounding the durability of partisan edits in a specific domain.

Empirical research on Wikipedia manipulation reveals that a subset of hoax articles—despite being rare—persist long enough to influence downstream systems [12]. Further, disinformation is often opportunistic: [13] find that spikes in public interest precede the creation of manipulative content.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

Detection of covert influence operations remains an open problem, though approaches for identifying undisclosed paid editing show promise [14].

To mitigate hallucinations—factual errors generated by LLMs despite fluent output—researchers have explored RAG [15]. By conditioning responses on external sources, RAG systems can improve factuality. However, as shown by [16] and by [17] on "RAG poisoning", these systems are only as trustworthy as their underlying retrieval corpora. If sources, such as Wikipedia, are compromised, even grounded systems may propagate misinformation.

In summary, while LLMs benefit from open knowledge sources like Wikipedia, current research points to systemic risks related to editorial integrity, data poisoning, and trust calibration. These vulnerabilities are particularly consequential in sensitive domains, such as law and medicine, where both AI outputs and their citations must be held to a high standard of factual rigor.

B. Overview of this paper

This paper presents an empirical study of long-term editorial manipulation within the German-language Wikipedia's legal category. By tracing the revision and discussion histories of over 15,000 articles described in Section II, we document the involvement of users who were later permanently banned from the platform due to rule violations—excluding voluntary account closures and deceased editors. The findings in Section III indicate a systematic pattern of infiltration even in domains typically considered neutral or apolitical. As discussed in Section IV, these insights carry serious implications for artificial intelligence systems. Wikipedia is frequently cited as a critical component of LLM training corpora and serves as a common grounding source in RAG frameworks. However, when the integrity of that source is compromised, AI outputs become vulnerable not only to hallucinations but also to factual contamination—a double-layered risk that undermines both answer reliability and user trust. This study sheds light on the hidden risks of relying on crowdsourced platforms for factual grounding in AI systems, calling for a re-evaluation of data hygiene practices in the machine learning pipeline.

II. Methodology

Figure 10 provides a visual overview of the five-step procedure for constructing the article database. Each step relied on the Wikipedia REST API, coupled with Python scripts:

1) *Maintenance-Category Retrieval.* In the first pass, maintenance categories used by Wikipedia to tag outdated or problematic pages were downloaded (175 in total). These categories served as one filtering criterion to exclude articles from subsequent steps if they were deemed insufficiently maintained or not in compliance with editorial standards.
2) *Recursive Download of Legal Subcategories.* The second pass started at the top-level category "Recht" (German for "Law") in the German-language Wikipedia. All subcategories were recursively traversed, collecting any articles placed under these nested categories (15,295 total). These articles were then added to the database.
3) *Template-Based Retrieval.* In the third pass, the scripts identified all articles that utilized one of 24 law-specific templates (e.g., infoboxes or structured references) designed to provide a uniform layout for legal topics. Any article that belonged to a maintenance category or that did *not* map to a legal category was excluded. A brief manual review of categories followed to ensure that peripheral topics (e.g., chemicals or pharmaceuticals) were omitted if they were tangentially but not substantively related to the legal domain. The full text of these articles (17,183 in sum) was downloaded.
4) *Keyword-based Title Search.* The fourth pass used a list of legal terms from a specialized law dictionary ("Weber kompakt" [18]), performing a title-based search in Wikipedia. Although about 9,000 articles were initially returned, roughly 4,000 were filtered out because they discussed aspects (often technical or historical) not relevant to the dictionary's legal perspective. About 5,000 articles passed the filtering, with 854 of those being genuinely new to the database; the remainder were duplicates of already-collected articles.
5) *Expansion via Internal Links.* Finally, from all articles in the database, the 10,000 most frequently occurring internal Wikipedia links were extracted and subjected to the same category-based filtering. This step added about 1,500 articles, bringing the total to 15,344 articles. A final manual review process then excluded categories that were still not strictly related to the legal domain, ensuring the final dataset was as specific as possible to topics in law.

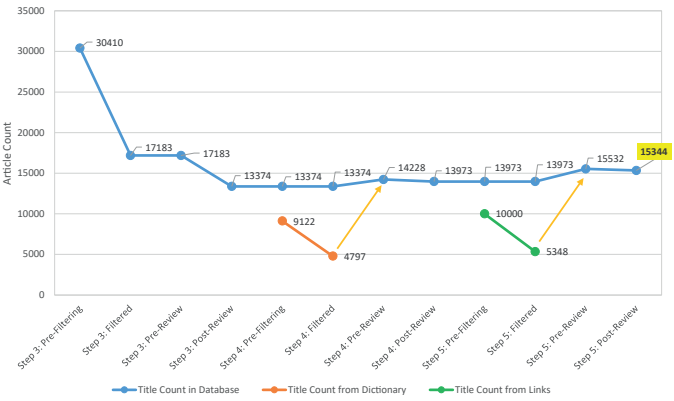

Figure 1. Progression of the database size in pages (articles) in blue depending on the steps in the download process.

As shown in Figure 1, the initial corpus grew substantially during the second and third steps, when the "Recht" root category and legal templates were harvested. In the fourth step (lexical title matching with a legal dictionary), the net increase in articles was relatively modest, because many of the newly found candidate pages were already present or failed the filter. Finally, the link-expansion step further added around

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

1,500 articles, though a minor decrease is visible after each "post-review" process, which eliminated pages affiliated with irrelevant or borderline categories.

Throughout the steps in the download process, each article was stored in a SQLite database along with all associated metadata, including internal links, revision histories, discussion pages, external references, and page-view statistics. In addition, a second SQLite database was populated with 185,555 permanently blocked (banned) users (April 27, 2004 to March 20, 2025). Reasons for indefinite blocks provided as free-text by Wikipedia administrators, were parsed via regular expressions and grouped into broader categories (e.g., vandalism, sockpuppetry). Any article revision or discussion post by a user in this second database was flagged, enhancing the main database with ban-related columns.

III. Findings

The goal of this study was to assess the extent of editorial infiltration in the German-language Wikipedia's legal domain. The results indicate a non-trivial overlap between legal articles and contributors who were subsequently banned.

A. Banning of Users

Table I shows that over 180,000 users were permanently banned from Wikipedia for reasons other than their own request or death, while Table II and Figure 2 highlight the principal violation categories. Among these, "sockpuppetry" stands out as a clear strategy for infiltration: the term references manipulated accounts operated under pseudonyms, sometimes identified through investigations documented by both investigative journalism and Wikipedia itself [19] [20] [21]. In early 2025, a German public broadcaster (ARD) devoted a podcast to exposing a "Sockenpuppenzoo" (zoo of sockpuppetry) [22], revealing how pervasive and coordinated such efforts can be.

TABLE I
Number of Banned Users (last 21 years).

	No. of Users	Percentage
Non-Compliance	183,756	99%
At Own Request	1,455	0.8%
Deceased	344	0.2%
All Permanently Blocked Users	185,555	100%

TABLE II
Reasons for Banning of Users (last 21 years).

	No. of Users	Percentage
"Clearly not being here to build an encyclopedia"	52,656	29%
Vandalism	12,427	7%
Sockpuppetry	9,724	5%
Edit Wars	1,915	1%
Other Reasons	107,034	58%
Non-Compliance	183,756	100%

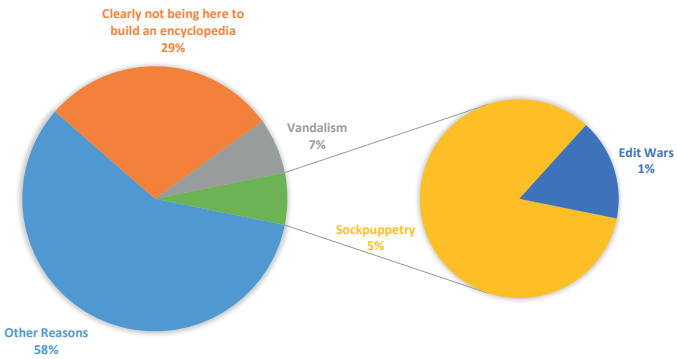

Figure 2. Reasons for permanently banning of users. See also Table II.

Figure 3 visualizes how the number of bans per month has fluctuated over the past two decades. During the project's early years, bans rose sharply, possibly due to a combined effect of increased Wikipedia participation and improved moderator capacity. Although monthly bans have remained high overall, there is no obvious surge specifically attributable to the advent of large language models or AI-generated content.

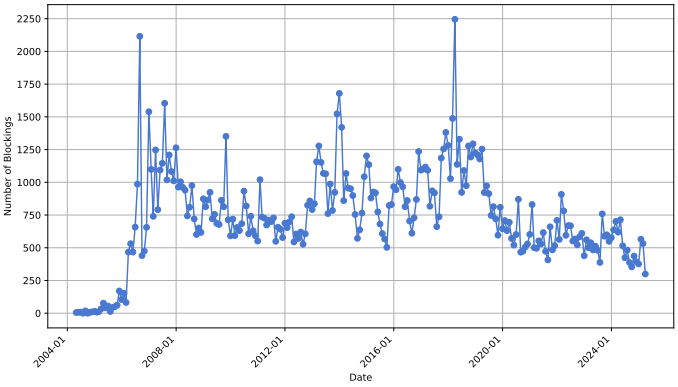

Figure 3. Number of permanently banned users per month until March 20, 2025.

Meanwhile, Figure 4 compares monthly bans against the number of new user registrations, revealing that in some months—particularly in the pre-2008 era—over 10% of newly registered users ended up permanently banned. More recently, this percentage has stabilized between 2% and 4%, indicating a persistent but somewhat reduced infiltration rate.

For the subset of 15,344 legal articles, analysis of the revision history and associated discussion pages reveals that roughly 70% have at least one revision by a permanently banned user, and about 21% of the discussion pages contain contributions from permanently banned users (Tables III and IV, Figures 5 and 6). Of notable concern is the group of articles (0.83% of the total) whose most recent revision was authored by a user later banned. Their content may remain compromised if not superseded by a good-faith edit.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

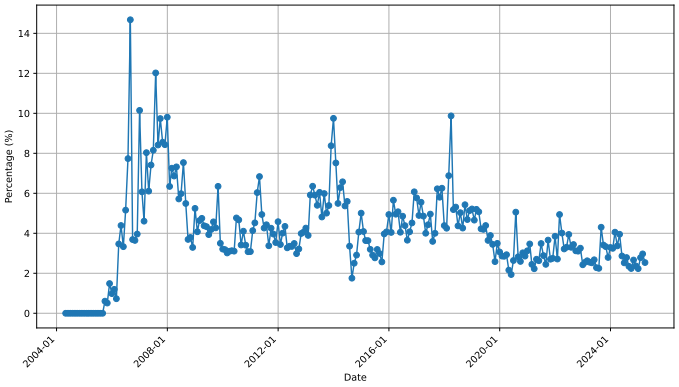

Figure 4. Number of permanently banned users in relation to the number of new registrations per month until March 20, 2025.

TABLE III
BANNING RATIO IN REVISIONS IN ABSOLUTE AND RELATIVE NUMBERS. SEE ALSO FIGURE 5.

Banning Ratio	Articles	Percentage
0% (no infiltration / banned authors)	4,682	30.51%
between 0% and 10%	9,216	60.00%
between 10% and 20%	978	6.37%
between 20% and 30%	261	1.70%
between 30% and 40%	125	0.81%
between 40% and 50%	52	0.34%
more than 50%	30	0.20%
	15,344	100%

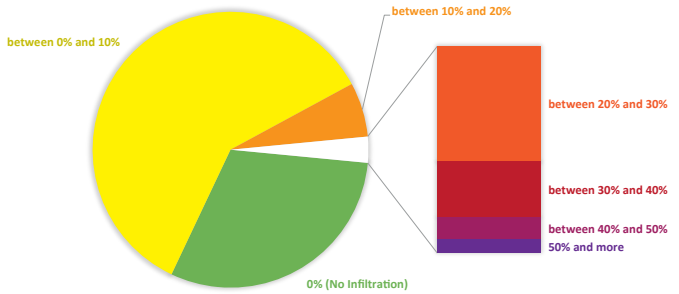

Figure 5. Banning ratio in revisions. For roughly one third of the articles, no revision originates from a banned author (0% banning ratio). See also Table III.

TABLE IV
BANNING RATIO IN DISCUSSIONS IN ABSOLUTE AND RELATIVE NUMBERS. SEE ALSO FIGURE 6.

Banning Ratio	Articles	Percentage
no discussions for these articles	4,572	29.80%
0% (no infiltration / banned contributors)	7,517	48.99%
between 0% and 10%	1,818	11.85%
between 10% and 20%	792	5.16%
between 20% and 30%	266	1.73%
between 30% and 40%	166	1.08%
between 40% and 50%	107	0.70%
more than 50%	106	0.69%
	15,344	100%

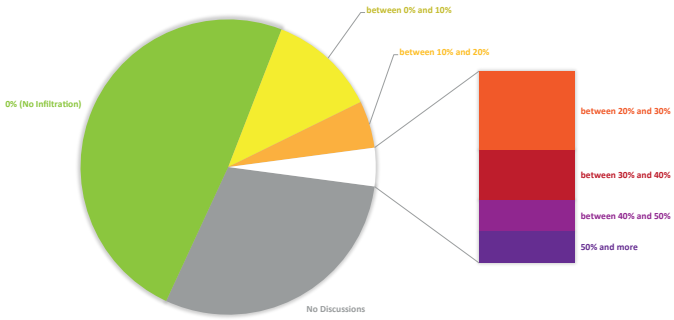

Figure 6. Banning ratio in discussions. For approximately one third of the articles, no discussion was recorded at all, and for nearly half of them no contributor was banned. See also Table IV.

B. Correlation with Page Views

A quantitative rank correlation analysis (using Spearman's correlation coefficient) revealed that the relationship between average daily page views and the banning ratio (in both revisions and discussions) is weakly positive and statistically significant: the coefficient was determined to be $\rho = 0.369$ for revisions and $\rho = 0.272$ for discussion pages. Figure 7 illustrates a more nuanced insight when articles are grouped by their daily page views. Four scenarios are compared:

- All articles (no filter on page views)
- Articles with daily page views $\geq Q1$ (the 25% quartile)
- Articles with daily page views $\geq$ median
- Articles with daily page views $\geq Q3$ (the 75% quartile)

For each of these four subsets, the figure tracks the proportion of articles that have *no banned authors* in their revision histories (banning ratio = 0%) versus those that have *banned authors involved* ($> 0\%$ banning ratio). A clear trend emerges as the minimum threshold of daily page views increases: the percentage of articles showing at least some infiltration steadily grows. This indicates that articles drawing higher traffic—be they prominent legal topics or controversial issues—tend to accumulate more edits overall, which in turn raises the likelihood of encountering disruptive contributors.

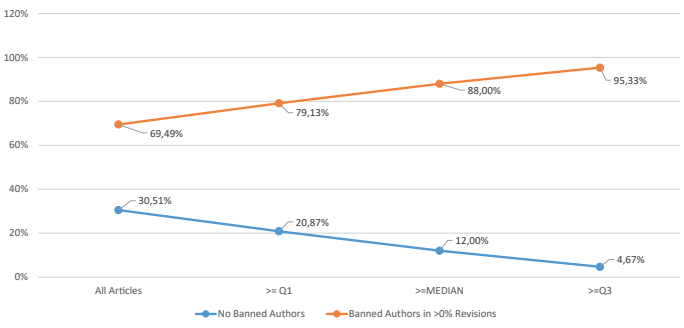

Figure 7. Proportion of articles w/ and w/o banned authors as a function of minimum number of daily page views.

Figure 8 depicts a histogram of the average daily page views *exclusively* for articles that are entirely free from banned

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

contributors (in both revisions and discussions). A majority of these "clean" articles attract fewer than five views per day, suggesting they may be of limited interest to either casual readers or would-be manipulators. The histogram is right-skewed, indicating that while most articles remain unnoticed, a small subset does register higher traffic. The absence of permanently blocked (banned) authors among these pages compared to those pages with solely banned users in Figure 9 aligns with the notion that low-visibility pages tend to experience fewer conflict-driven edits. Of course, it does not guarantee editorial quality in an absolute sense.

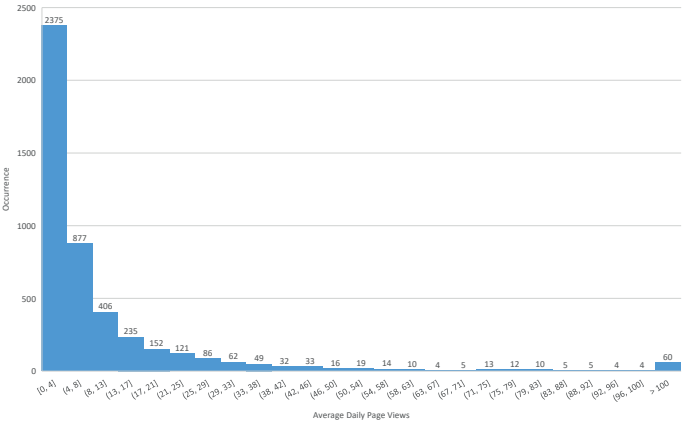

Figure 8. Distribution of average daily page views for all articles that do *not* have permanently banned authors.

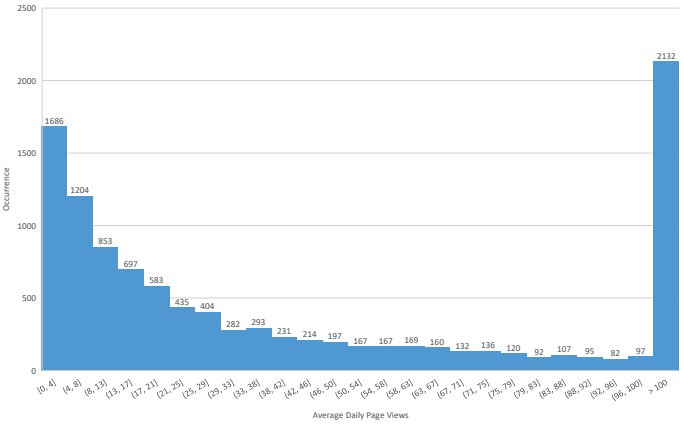

Figure 9. Distribution of average daily page views for all articles that *do* have permanently banned authors.

These findings confirm that even a relatively narrow, less politically charged topic, such as law, is not immune to data poisoning efforts. The prevalence of sockpuppet accounts emphasizes the sophistication of such adversarial behavior, while the long duration of this infiltration—spanning more than two decades—points to a systemic issue of editorial integrity in open knowledge platforms.

Across the corpus of 15,344 legal articles, our quantitative analysis demonstrated that approximately 70% of pages include at least one revision by a permanently banned user, and about 21% of associated discussion pages show contamination. While roughly one third of articles remain free of compromised edits, a small but significant fraction exhibit banning ratios exceeding 20–30% in their revision histories. Spearman correlation coefficients reveal a clear trend: articles with higher daily page views are more likely to have been infiltrated, whereas "clean" articles tend to register fewer than five views per day. This pattern highlights how visibility amplifies vulnerability, reinforcing the need to account for both editor behavior and page popularity when assessing the reliability of open-source knowledge for AI applications.

IV. Conclusion and Future Work

The following section discusses the key findings and draws conclusions. The subsequent section summarizes the outcome of this study.

This study highlights the persistent vulnerability of Wikipedia's German-language legal domain to infiltration by malicious actors. Despite extensive administrative and community-driven oversight, more than two-thirds of legal articles show traces of editorial input from subsequently banned users. Although the legal field may appear apolitical, the data highlight that infiltration and opinion manipulation are not confined to typically controversial areas.

Furthermore, the findings raise important concerns about using Wikipedia articles as a grounding source in RAG systems. As widely documented, large language models can hallucinate when faced with gaps in their training data. RAG mitigates this risk by drawing upon external documents. However, if those external sources harbor inaccuracies or manipulations—intentionally introduced by users with malicious or extremist agendas—hallucinations may be replaced by confidently stated falsehoods. In the context of legal advice, such errors can have serious practical consequences, undermining public trust in both open-source knowledge and AI systems.

The present results recommend the following directions for future work on the topic:

- *Broader Multi-Domain Analysis.* Replicating this methodology for additional subject areas would clarify whether the observed infiltration patterns are specific to the legal sphere or mirrored across other domains.
- *Automated Quality Ratings.* Integrating a rating scheme that accounts for a page's infiltration history (e.g., via the "Banning Ratio") could inform downstream usage for training or grounding. This may include a large-language-model-based sentiment analysis of discussion pages to identify constructive versus adversarial engagement.
- *Refined Filtering for AI Data.* For RAG-based systems or training pipelines, removing or downweighting articles that show high infiltration scores and introducing a reliability metric into prompts can reduce the risk of providing manipulated content to end users.
- *Ongoing Community Oversight.* As infiltration continues to evolve, a coordinated effort by Wikipedia administrators and community volunteers is essential. Studies like

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

this may help sharpen the focus on identifying emerging patterns and closing loopholes that enable repeated sockpuppetry.

In summary, even a specialized, seemingly neutral topic, such as "Recht" (law) on the German Wikipedia, exhibits clear patterns of infiltration by permanently banned contributors. This study has documented both the extent of that infiltration and its implications for data reliability and AI systems that rely on Wikipedia for training or reference. The belief that crowdsourced content will always self-correct through sheer volume of contributors is challenged by the persistent manipulation attempts observed here. As large language models become more ingrained in everyday applications—particularly in legally sensitive contexts—the urgency to shore up editorial quality and counter data poisoning grows. Countermeasures, including refined scraping, filtering processes, and real-time oversight, are crucial steps toward ensuring the continued integrity of open knowledge ecosystems.

ACKNOWLEDGEMENTS

I would like to thank the referees for very useful comments on the original submission. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] A. Wan, E. Wallace, S. Shen, and D. Klein, "Poisoning language models during instruction tuning," in *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, pp. 35413–35425, JMLR.org, 2023.

[2] D. Alber *et al.*, "Medical large language models are vulnerable to data-poisoning attacks," *Nature Medicine*, vol. 31, pp. 618–626, 01 2025.

[3] T. Fu, M. Sharma, P. Torr, S. B. Cohen, D. Krueger, and F. Barez, "Poisonbench: Assessing large language model vulnerability to data poisoning," *ArXiv*, vol. abs/2410.08811, 2024.

[4] T. B. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[5] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *ArXiv*, vol. abs/2302.13971, 2023.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *ArXiv*, vol. abs/1810.04805, 2019.

[7] M. Vetter, J. Jiang, and Z. McDowell, "An endangered species: how llms threaten wikipedia's sustainability," *AI & SOCIETY*, pp. 1–14, 02 2025.

[8] J. Giles, "Internet encyclopaedias go head to head," *Nature*, vol. 438, no. 7070, pp. 900–901, 2005.

[9] S. Greenstein and F. Zhu, "Do experts or crowd-based models produce more bias? evidence from encyclopædia britannica and wikipedia," *MIS Quarterly*, vol. 42, pp. 945–959, Sept. 2018.

[10] S. Greenstein, G. Gu, and F. Zhu, "Ideology and composition among an online crowd: Evidence from wikipedians," *Management Science*, vol. 67, no. 5, pp. 3067–3086, 2021.

[11] D. Rozado, "Is wikipedia politically biased?," *Manhattan Institute*, June 2024. Published June, 20th, 2024.

[12] S. Kumar, R. West, and J. Leskovec, "Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes," in *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, (Republic and Canton of Geneva, CHE), pp. 591–602, International World Wide Web Conferences Steering Committee, 2016.

[13] A. Elebiary and G. L. Ciampaglia, "The role of online attention in the supply of disinformation in wikipedia," *ArXiv*, vol. abs/2302.08576, 2023.

[14] N. Joshi, F. Spezzano, M. Green, and E. Hill, "Detecting undisclosed paid editing in wikipedia," in *Proceedings of The Web Conference 2020*, WWW '20, (New York, NY, USA), pp. 2899–2905, Association for Computing Machinery, 2020.

[15] P. Lewis *et al.* NIPS '20, (Red Hook, NY, USA), Curran Associates Inc., 2020.

[16] L. Huang *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, pp. 1–55, Jan. 2025.

[17] W. Zou, R. Geng, B. Wang, and J. Jia, "Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models," *ArXiv*, vol. abs/2402.07867, 2024.

[18] K. Weber, T. Aichberger, and R. Werner, *Weber compact, Law Dictionary*. Beck-online : Bücher, München: Verlag C.H. Beck, 11. edition, stand: 01.08.2024 ed., 2024.

[19] Wikipedia contributors, "Sock puppet account." https://en.wikipedia.org/wiki/Sock_puppet_account, 2025. Accessed: 2025-05-15.

[20] Wikipedia contributors, "Wikipedia:sockpuppet investigations." https://en.wikipedia.org/wiki/Wikipedia:Sockpuppet_investigations, 2025. Accessed: 2025-05-15.

[21] Wikipedia contributors, "Wikipedia:sockpuppetry." https://en.wikipedia.org/wiki/Wikipedia:Sockpuppetry, 2025. Accessed: 2025-05-15.

[22] C. Schattleitner and D. Laufer, "Sock puppet zoo – attack on wikipedia." https://www.ardaudiothek.de/sendung/sockenpuppenzoo-angriff-auf-wikipedia/13996869/, 2025. Podcast.

APPENDIX

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

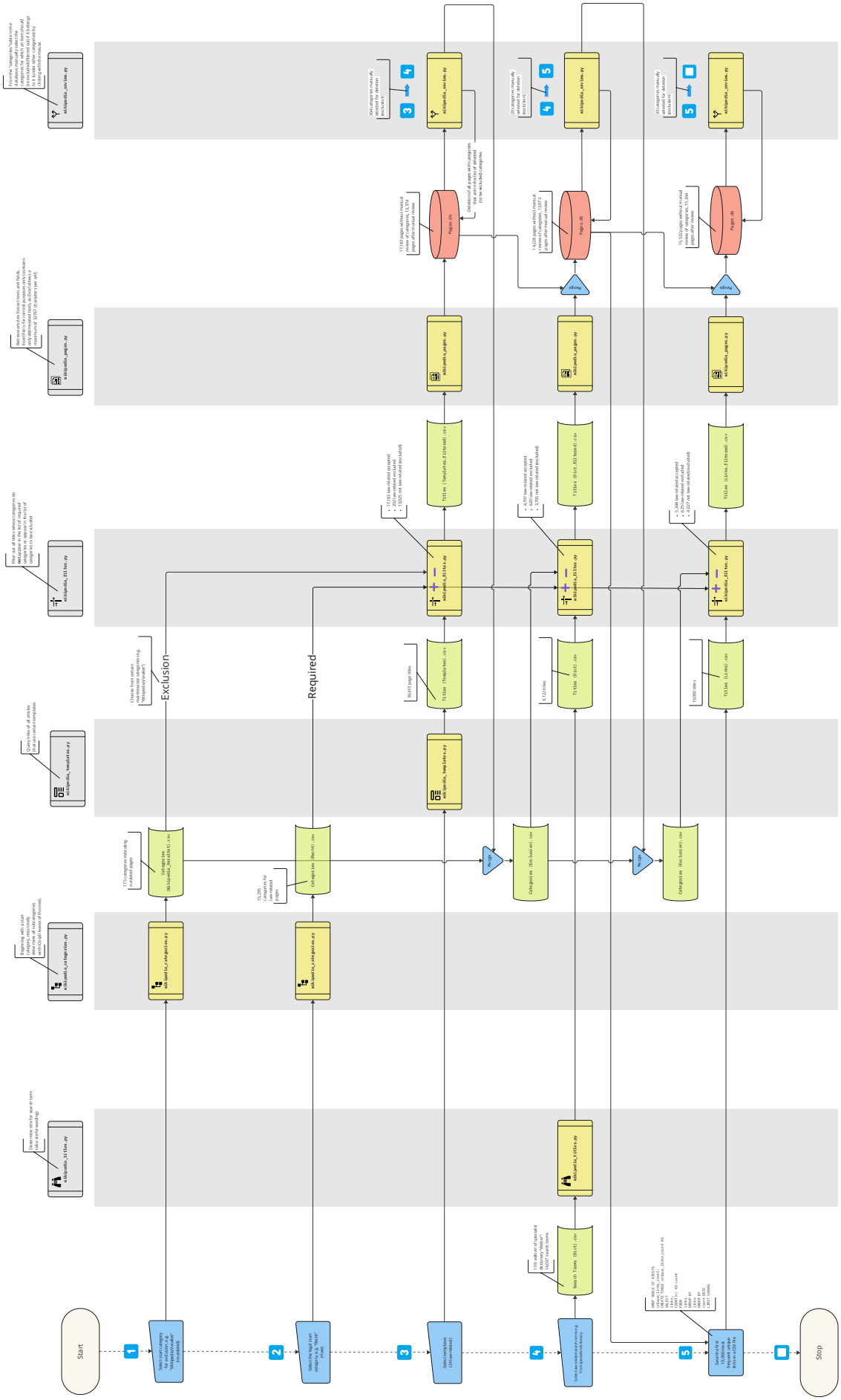

Figure 10. The process of downloading Wikipedia articles of a distinctive topic into an SQLite database using the REST API and a set of Python scripts.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org