Tracking Bias and Censorship Across Wikipedia Language Versions Using LLMs

Anastasiia Iurshina
independent

Abstract

In this research, we propose a mixed-method study to detect, analyze, and explain cross-lingual differences in the representation of sensitive or contested topics on Wikipedia, such as migration, protest movements, and human rights. Building on advances in natural language processing and multilingual large language models (LLMs), we will develop a pipeline that aligns equivalent articles across language editions and uses LLM-based summarization and comparative analysis to surface discrepancies in framing, emphasis, or omitted information. We will complement this with an analysis of edit histories, metadata, and contributor networks to understand how these differences are shaped by editorial dynamics, such as concentrated editing groups, revert patterns, or potential state-linked activity. The project aims to produce generalizable insights into how knowledge is curated differently across linguistic and geopolitical contexts, and how this may impact public understanding of global issues. The research will result in: (1) a public dataset of aligned articles and detected differences, (2) an interactive prototype to explore comparative summaries and editorial histories, and (3) a peer-reviewed publication on cross-lingual information asymmetry in Wikipedia. This work contributes to the Wikimedia 2030 strategic direction by supporting knowledge equity, increasing transparency, and helping communities detect and address biases that might otherwise go unnoticed across language silos.

Introduction

Wikipedia is the world's largest collaborative encyclopedia, yet articles on the same topic can differ significantly across its language editions—not only in length or sourcing, but in framing, emphasis, and even factual content. These cross-lingual discrepancies are especially prominent in articles covering sensitive or politically charged topics such as migration, human rights, war, and protest movements.

Currently, there is no scalable or systematic tool to identify and compare these differences, leaving knowledge asymmetries largely invisible and difficult for editors, researchers, or readers to track or address. Additionally, little is known about how editorial patterns—such as clustering of contributors, reverts, or bot-like behavior—might correlate with these content differences, especially in contexts where political or ideological influence may shape knowledge production.

This project aims to fill that gap by combining multilingual language models with edit history and contributor analysis to systematically detect, explain, and visualize cross-lingual

differences in Wikipedia content and investigate the editorial dynamics behind them.

This problem directly impacts the core Wikimedia goals of neutrality, reliability, and knowledge equity:

- Neutrality and accuracy: When articles on controversial topics differ widely across language editions, it challenges Wikipedia's commitment to presenting a balanced view of topics, especially in non-English Wikipedias that serve as key knowledge sources in their regions.
- Knowledge equity and access: Disparities in framing between Global North and Global South language editions can entrench epistemic inequalities, marginalize certain perspectives, and obscure human rights concerns in regions with limited press freedom.
- Editorial transparency: Understanding how contributors shape this divergence—whether through sincere local contextualization or coordinated bias—is essential for healthy community governance and for countering misinformation or manipulation.
- Supporting editors and communities: Equipping editors with tools to explore and track these divergences can empower them to bridge gaps, translate overlooked perspectives, and identify areas requiring attention across the Wikimedia ecosystem.

Several studies (e.g., Bao et al., 2012) have pointed to bias and information gaps in multilingual Wikipedia, but tools and methods to make such gaps actionable for the broader community are still limited. Our project contributes by offering both an analytical framework and practical outputs for monitoring and understanding these asymmetries.

We will address the following research questions:

RQ1: To what extent do Wikipedia articles on the same topic differ in framing, emphasis, and factual content across language editions?
We hypothesize that significant cross-lingual discrepancies exist in articles covering geopolitically sensitive or ideologically polarized topics.
RQ2: What types of editorial behaviors (e.g. user clustering, reverts, bot activity) correlate with observed differences in article content across languages?
We hypothesize that in some cases, content differences are associated with coordinated editing patterns or persistent disagreements.
RQ3: How can multilingual models be used to surface and explain these differences to a non-specialist audience, including editors and researchers?
We hypothesize that LLM-based summarization and comparison techniques can produce accessible, high-level descriptions of content divergence.

Date

- Proposed start date: July 1, 2025
- Proposed end date: June 30, 2026

Related work

Research into content bias, representation asymmetries, and editorial behavior on Wikipedia has grown over the past decade, highlighting how socio-political factors and community dynamics influence knowledge production. However, cross-lingual analysis at scale remains relatively underexplored, particularly in ways that are actionable for Wikimedia editors and communities.

Several studies have identified systemic biases and information gaps across Wikipedia's

multilingual editions, particularly in the representation of culturally or politically sensitive topics. For example, Bao et al. (2012) introduced *Omnipedia*, a tool that visualized how Wikipedia articles diverged across languages, revealing substantial disparities in coverage, framing, and focus on the same topics across editions (Bao et al., 2012). Hube and Fetahu (2019) proposed a method for detecting biased language at the sentence level, finding that subjective or ideologically charged phrasing appears in different forms across English and German Wikipedias (Hube & Fetahu, 2019). More recently, broader surveys such as *Mind the InfoGap* (2025) have highlighted how cultural and geographic underrepresentation continues to shape Wikipedia's global knowledge base, underscoring the need for tools that support equity and transparency in multilingual content (Mind the InfoGap, 2025).

In terms of technical approaches, the use of NLP and machine learning in Wikipedia studies is increasing. Recent research has leveraged transformer-based models to detect sentiment bias (Field et al., 2021), identify controversial edits (Geiger & Ford, 2011), and classify sources by reliability. However, most of these approaches are designed for single-language contexts or rely on manually annotated datasets, limiting their generalizability. Only a few efforts have explored the potential of multilingual large language models (LLMs) to align and compare article content across languages in a way that is both scalable and interpretable.

This proposal builds on and contributes to these areas by:

- Developing a scalable LLM-based pipeline to semantically compare articles across languages;
- Connecting editorial dynamics with observed content differences through data-driven analysis of histories and contributor patterns;
- Creating an interpretable output (summaries, visualizations, datasets) designed for both academic use and community action.

By focusing on sensitive topics such as migration and human rights, the project also addresses underrepresented areas where discrepancies are likely to be both socially impactful and epistemically consequential.

Methods

Data Collection

We will begin by selecting a representative sample of Wikipedia articles on politically or socially sensitive topics, such as migration, human rights, censorship, protest movements, and international conflict. These topics will be chosen based on prior research, edit frequency, and community feedback, focusing on articles that exist in at least three or more language editions, including but not limited to English, Russian, Arabic, and German.

Using the Wikimedia REST API, Wikidata, and Wikimedia dumps, we will:

- Collect aligned articles across selected language editions.
- Retrieve revision histories and metadata (e.g., timestamps, usernames, reverts).
- Extract contributor profiles (e.g., editor frequency, geographic distribution when available, revert patterns).

To ensure transparency and reproducibility, all processing scripts will be released as open-source code under a permissive license.

LLM-Assisted Comparative Analysis

To identify and explain content discrepancies across Wikipedia language editions, we will use multilingual large language models (LLMs) through a combination of API-based prompting and open-source fine-tuning.

Building on the previously outlined dataset of aligned articles across language editions, we will use LLM APIs (e.g., OpenAI, Claude, or open-source models via HuggingFace) to:

- Perform pairwise comparisons between versions to detect differences in framing, emphasis, factual content, and narrative structure.
- Extract and highlight specific passages that differ across languages, using a taxonomy of discrepancy types (e.g., factual omission, evaluative tone, emphasis shift, or politically sensitive framing).
- Generate natural-language summaries of those differences (in the original language and in English), aimed at interpretability for both researchers and Wikimedia contributors.

Where appropriate, we will explore fine-tuning open-source models (e.g., Llama 3) to improve task-specific performance, such as highlighting discrepancies or generating concise cross-lingual comparison reports. Fine-tuning will rely on a small manually validated dataset derived from annotated examples created during the initial analysis.

Selected outputs will be human-validated to ensure quality and in the process of prompt-engineering.

Editorial Network Analysis

For each aligned article set, we will analyze revision histories to:

- Identify patterns of editorial concentration, e.g., whether a small group of users dominates editing.
- Map revert behaviors to detect controversial or contested content areas.
- Classify users into behavioral clusters (e.g., casual editors, single-topic contributors, high-frequency reverters).

Statistical correlation between these patterns and detected content divergences will be examined using regression and clustering methods.

Expected output

This project will generate a combination of research, technical, and community-facing outputs, each designed to support both scholarly insight and practical impact within the Wikimedia ecosystem.

First, we aim to produce a **peer-reviewed publication** that presents our findings on cross-lingual content discrepancies and the editorial behaviors that may contribute to them. This paper will be targeted at academic audiences in digital humanities, computational social science, and NLP, with potential venues including *WikiWorkshop*, *ICWSM*, *COLING*, or the *Journal of Web Science*.

Second, we will release a **public dataset containing aligned Wikipedia articles**, model-generated summaries, detected content divergences, and editorial metadata. This dataset will be valuable for researchers studying multilingual knowledge representation, as well as for Wikimedia affiliates, educators, and tool developers interested in transparency and equity in Wikipedia content.

Third, we will build an **interactive prototype or dashboard** that enables users to explore the

content differences and editorial patterns across selected topics and languages. Designed with accessibility in mind, this tool will support Wikimedia editors, translators, and chapter organizers by offering an intuitive way to navigate framing differences and improve cross-lingual article consistency.

Each of these outputs will be openly accessible and designed to encourage further reuse, adaptation, and community engagement. Our ultimate goal is to provide not only academic knowledge but also practical tools and evidence that can support Wikimedia's mission of global, equitable access to trustworthy knowledge.

Risks

One risk is the potential inaccuracy or hallucination in outputs generated by large language models (LLMs). To mitigate this, we will apply prompt constraints, explore fine-tuning where feasible, and conduct human validation of summaries and comparisons. We will involve multilingual annotators to verify the generated content and, where applicable, provide uncertainty or confidence scores alongside outputs.

Another risk involves the political sensitivity of the findings, especially when discrepancies point to potential censorship, bias, or manipulation. To address this, we will follow strict ethical research practices, anonymize all individual contributors in our analyses, and conduct internal and community reviews before releasing any sensitive findings. The focus will be on highlighting systemic patterns rather than singling out individuals or communities.

Finally, data availability and NLP tool performance may vary significantly across language editions. In the early stages, we will prioritize high-coverage language editions and gradually expand as tooling and data access allow.

Community impact plan

We will engage with the Wikimedia movement through multiple channels:

- Collaborate with Wikimedia affiliates (e.g., Wikimedia Deutschland) to select relevant topics and languages.
- Share results in digestible formats: short explainers, dashboards, blog posts, and mailing list updates.
- The interactive tool will be designed with editor usability in mind, offering features like diff visualizations, topic filters, and links to original articles. We aim to support editors in identifying potential bias, improving cross-language translation, and promoting epistemic justice across the Wikimedia landscape.

Evaluation

We will evaluate our project across three key dimensions:

1. Research outputs:
    - Submission and acceptance of at least one peer-reviewed article.
    - Release of a high-quality multilingual dataset and open-source code.
2. Tool and prototype:
    - Functional dashboard or prototype made available on Toolforge or similar Wikimedia infrastructure.
    - Usability tested with Wikimedia volunteers.
3. Impact indicators:

- Citations, downloads, and reuse of datasets or tools.
- Evidence of integration or citation in Wikimedia strategy or editor workflows.
- Qualitative reports of improved awareness or use of findings by community members.

Success will be defined not only by academic outputs, but by how well our work helps Wikimedia contributors identify, understand, and act on knowledge asymmetries across languages.

Budget

You can view the budget spreadsheet here: budget

References

Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., & Gergle, D. (2012). **Omnipedia: Bridging the Wikipedia language gap**. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1075–1084. https://doi.org/10.1145/2207676.2208553

Field, Anjalie and Park, Chan Young and Lin, Kevin Z. and Tsvetkov, Yulia. **Controlled identification of bias in Wikipedia language**. *Proceedings of the ACM Web Conference 2022*, 2624–2635. https://doi.org/10.1145/3485447.3512134/

Geiger, R. S., & Ford, H. (2011). **Participation in Wikipedia's article deletion processes**. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 2011. https://doi.org/10.1145/2038558.2038593

Halfaker, A., Kittur, A., Kraut, R. E., & Riedl, J. (2013). **The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline**. *American Behavioral Scientist*, 57(5), 664–688. https://doi.org/10.1177/0002764212469365

Hube, C., & Fetahu, B. (2019). **Neural based statement classification for biased language**. *Proceedings of the 2019 World Wide Web Conference (WWW)*, 195–203. https://doi.org/10.1145/3289600.3291018

Mind the InfoGap. (2025). **Uncovering Cultural Bias in Wikipedia.** *UBC Focus on Science*. https://focus.science.ubc.ca/mind-the-infogap-2fceb3838cb6