WikiGap: Promoting Epistemic Equity by Surfacing Knowledge Gaps Between English Wikipedia and Minority Language Editions

Zining Wang[1,2] **Dongwook Yoon**[1] **Nicholas Vincent**[3]
Farhan Samir[1,4] **Vered Shwartz**[1,2]
[1] University of British Columbia [2] Vector Institute for AI
[3] Simon Fraser University [4] University of Toronto

Introduction

Wikipedia comprises over 300 language editions, reflecting its commitment to democratizing access to and participation in knowledge production. These editions are not translations of one another, but instead surface knowledge shaped by local relevance, cultural experience, and media coverage, meaning that readers across languages are exposed to different information (Avieson and Mazumdar, 2025; Kumar, 2021). Despite this complementary structure, English Wikipedia's scale and visibility have led many readers to treat it as a comprehensive superset of other editions ("English-as-superset" (Hecht, 2013)), although English often lacks substantial content present elsewhere (Hecht and Gergle, 2010; Samir et al., 2024). This imbalance creates an epistemic asymmetry in which culturally grounded knowledge from non-Anglophone communities remains largely undiscoverable to readers navigating English Wikipedia (Kumar, 2017).

One promising way to address this asymmetry is to improve the discoverability of knowledge produced in other language editions within English Wikipedia's reading experience. Wikipedia's interlanguage links (ILLs) enable navigating other language editions, but the navigation panel provides no cues indicating that language editions may differ substantially in content. Other approaches that leveraged non-English content focused on consolidating information across language editions (Duh et al., 2013,), but this approach conceals (rather than highlights) the cross-linguistic heterogeneity of multilingual Wikipedia. Another line of work developed external interfaces for comparison, requiring readers to adopt new workflows for navigating Wikpedia content (Massa and Scrinzi, 2012; Roy et al., 2022). Overcoming both of these shortcomings, we developed WikiGap, a browser extension that surfaces sentence-level cross-lingual factual differences directly within English Wikipedia articles. WikiGap combines a computational backend that detects multilingual content gaps with a reader-facing interface that unobtrusively augments English articles with complementary facts drawn from the French, Russian, and Chinese editions.

We evaluated WikiGap through a mixed-methods user study using a fact-finding task. We found that participants strongly preferred WikiGap over Wikipedia's ILLs. Post-study interviews further reveal many participants were previously unaware of content gaps in English Wikipedia. Finally, participants suggested that making multilingual knowledge more discoverable can also support editorial workflows.

Methods – Designing WikiGap

Interface Design Requirements Informed by User Needs and Theory. Insights from our exploratory interviews and theories of information seeking (Pirolli and Card, 1999; Marshall, 1997) motivated three design requirements that collectively shaped the WikiGap interface (Figure 1). First, cross-lingual content should be surfaced without disrupting reading flow, as users reported avoiding Wikipedia's ILLs due to the cognitive cost of full-page language switches; this requirement led to **D1: In-text Gap Markers**, which anchor subtle cues to relevant English sentences, and **D2: Cross-Lingual Fact Panel**, which presents complementary facts in a margin-anchored sidebar that preserves reading continuity. Second, because surfaced facts are absent from English Wikipedia, readers emphasized the need for clear attribution and traceability to assess credibility; this motivated **D3: Fact Provenance Card**, which explicitly labels the source language and links each fact to its original sentence in the source-language article. Third, we introduced **D4: Source Language Filter**, allowing readers to select languages of interest, and **D5: Cross-lingual Fact Search**, enabling directed retrieval across surfaced multilingual facts.

Detecting Cross-Lingual Fact Gaps. WikiGap builds on InfoGap (Samir et al., 2024), an LLM-based pipeline that identifies facts present in the non-English edition but missing from English (Figure 2). We extend InfoGap to support Chinese. Detected facts are translated into English and selectively surfaced (up to 30 per article) to reduce cognitive overload.

User Study Design. We evaluated WikiGap through a mixed-methods user study (N=21) using a fact-finding task on five culturally rich food articles from Wikipedia. Participants ranged in age from 18 to 44 and represented diverse cultural and ethnic backgrounds. They completed tasks using both WikiGap and ILLs, selected as the con-

© Copyright held by the owner/author(s), published under Creative Commons CC BY 4.0 License

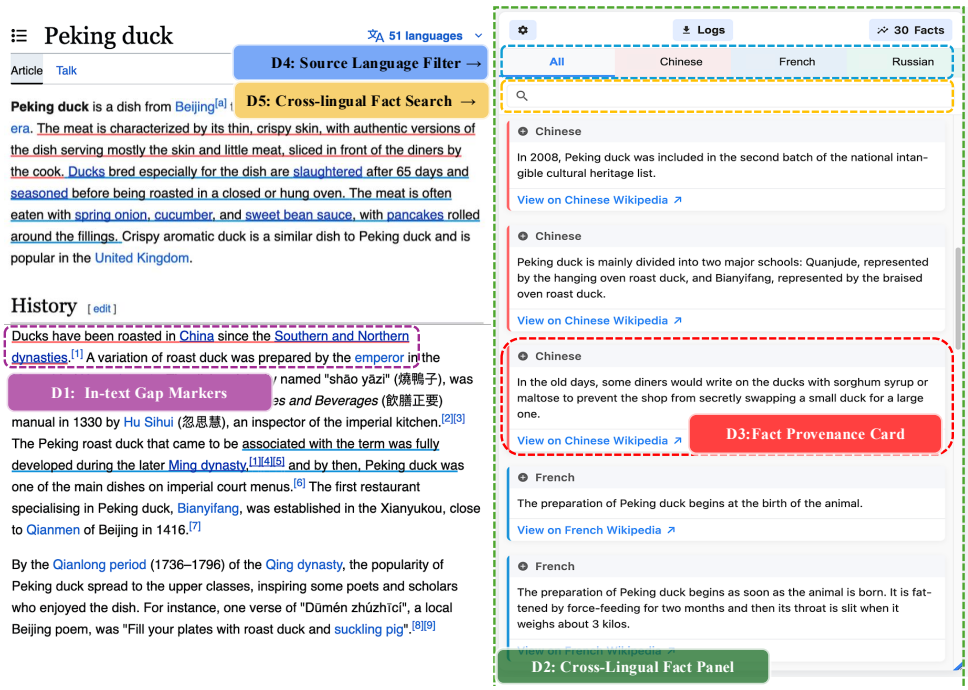

Figure 1: The WikiGap interface embeds cross-lingual facts into English Wikipedia via five key design elements.

trol to reflect the standard, in-context Wikipedia experience and to simulate an extension built on top of existing knowledge-seeking workflows, as prior multilingual interface designs are not publicly available for deployment. This setup enabled comparison of usability, task accuracy, and completion time, followed by post-study interviews.

Results

1. Improved usability and reading experience through in-place multilingual surfacing (D1–D3). WikiGap achieved substantially higher usability than ILLs, scoring in the excellent range on the System Usability Scale (SUS) (M=82.5, SD=10.9) compared to the control (M=48.2, SD=18.6), a significant difference ($t(20)$=6.60, p<.001). Participants attributed this preference to the interface's ability to surface multilingual content directly within the English article via sentence-level fact cards and a structured sidebar (D2, D3), which reduced navigation friction and made information easier to process than full-page language switches. In-text gap markers (D1) helped users notice missing information. Overall, users described the experience as more enjoyable and cognitively less demanding than their usual Wikipedia workflow.

2. Raised awareness of multilingual gaps and implications for trust and editing. Using WikiGap prompted participants to recognize that English Wikipedia is not the superset of all other language editions, challenging prior assumptions of completeness. This awareness motivated greater willingness to explore non-English content and, for some, reflections on trust in Wikipedia relative to LLM-based knowledge tools. While a small number of participants expressed preference for using LLMs for knowledge tasks due to perceived comprehensiveness, others emphasized that WikiGap reinforced the value of Wikipedia's provenance and citation practices by making sources transparent and traceable. Several participants with prior editing experience further noted that surfacing multilingual gaps could support cross-lingual editing workflows by helping editors discover, verify, and integrate missing information from other language editions.

Conclusion

Our findings build on prior work showing that much of Wikipedia's content remains invisible to readers despite its existence at scale (Arora et al., 2024). We extend this challenge to the multilingual setting, where content from other language editions is largely undiscoverable to readers navigating English Wikipedia. WikiGap addresses this by making unique knowledge from other language editions more discoverable within the English reading experience and redistributing reader attention beyond English Wikipedia. In doing so, it challenges the "English-as-superset" assumption and encourages readers to perceive non-English content as independently valuable contributions shaped by distinct peer production communities.

© Copyright held by the owner/author(s), published under Creative Commons CC BY 4.0 License

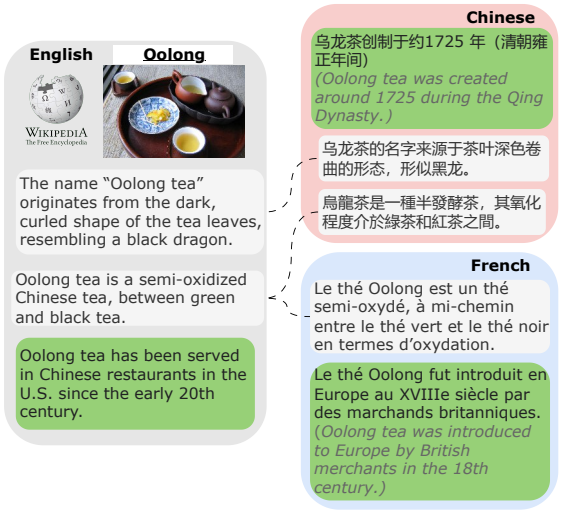

Figure 2: An example of knowledge gaps in Wikipedia's coverage of Oolong tea identified by InfoGap. The connecting lines indicate overlapping facts across language editions, while green boxes highlight facts uniquely mentioned in each language edition.

References

[Arora et al.2024] Akhil Arora, Robert West, and Martin Gerlach. 2024. Orphan articles: The dark matter of wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 100–112.

[Avieson and Mazumdar2025] Bunty Avieson and Suruchi Mazumdar. 2025. Wiki warriors: language editors counter knowledge hierarchies during the pandemic. *Asian Journal of Communication*, pages 1–19.

[Duh et al.2013] Kevin Duh, Ching-Man Au Yeung, Tomoharu Iwata, and Masaaki Nagata. 2013. Managing information disparity in multilingual document collections. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(1):1–28.

[Hecht and Gergle2010] Brent Hecht and Darren Gergle. 2010. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 291–300.

[Hecht2013] Brent Jaron Hecht. 2013. *The mining and application of diverse cultural perspectives in user-generated content*. Ph.D. thesis, Northwestern University.

[Kumar2017] Sangeet Kumar. 2017. A river by any other name: Ganga/ganges and the postcolonial politics of knowledge on wikipedia. *Information, Communication & Society*, 20(6):809–824.

[Kumar2021] Sangeet Kumar. 2021. *The digital frontier: Infrastructures of control on the global web*. Indiana University Press.

[Marshall1997] Catherine C Marshall. 1997. Annotation: from paper books to the digital library. In *Proceedings of the second ACM international conference on Digital libraries*, pages 131–140.

[Massa and Scrinzi2012] Paolo Massa and Federico Scrinzi. 2012. Manypedia: Comparing language points of view of wikipedia communities. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, pages 1–9.

[Pirolli and Card1999] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review*, 106(4):643.

[Roy et al.2022] Dwaipayan Roy, Sumit Bhatia, and Prateek Jain. 2022. Information asymmetry in wikipedia across different languages: A statistical analysis. *Journal of the Association for Information Science and Technology*, 73(3):347–361.

[Samir et al.2024] Farhan Samir, Chan Young Park, Anjalie Field, Vered Shwartz, and Yulia Tsvetkov. 2024. Locating information gaps and narrative inconsistencies across languages: A case study of LGBT people portrayals on Wikipedia. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6747–6762, Miami, Florida, USA, November. Association for Computational Linguistics.

© Copyright held by the owner/author(s), published under Creative Commons CC BY 4.0 License