# Factual Inconsistencies in Multilingual Wikipedia **Tables**

Silvia Cappa<sup>1</sup>, Lingxiao Kong<sup>2</sup>, Pille-Riin Peet<sup>3</sup>, Fanfu Wei<sup>4</sup>, Yuchen Zhou<sup>5</sup>, and Jan-Christoph Kalo<sup>6</sup>

- <sup>1</sup> CNR ISTC silviacappa@cnr.it Fraunhofer Institute for Applied Information Technology FIT
- lingxiao.kong@fit.fraunhofer.de <sup>3</sup> Tallinn University of Technology pille-riin.peet@taltech.ee
  - <sup>4</sup> EURECOM fanfu.wei@eurecom.fr <sup>5</sup> Technical University of Munich yuchen.zhou@tum.de
  - <sup>6</sup> University of Amsterdam j.c.kalo@uva.nl

**Abstract.** Wikipedia serves as a globally accessible knowledge source with content in over 300 languages. Despite covering the same topics, the different versions of Wikipedia are written and updated independently. This leads to factual inconsistencies that can impact the neutrality and reliability of the encyclopedia and AI systems, which often rely on Wikipedia as a main training source. This study investigates crosslingual inconsistencies in Wikipedia's structured content, with a focus on tabular data. We developed a methodology to collect, align, and analyze tables from Wikipedia multilingual articles, defining categories of inconsistency. We apply various quantitative and qualitative metrics to assess multilingual alignment using a sample dataset. These insights have implications for factual verification, multilingual knowledge interaction, and design for reliable AI systems leveraging Wikipedia content.

**Keywords:** Wikipedia · Factual Inconsistency · Tabular Data

#### Introduction 1

Wikipedia is one of the most widely used public knowledge sources, offering content in over 300 languages [16]. While articles in different language editions often aim to describe the same entities and events, they frequently diverge in the facts and contents they present, since they are not simply translated but rather compiled independently based on each language version [13]. These inconsistencies raise important questions about the reliability, completeness, richness, and neutrality of multilingual content. This project investigates various inconsistencies across Wikipedia language editions, with a specific focus on structured data such as tables. Inconsistency between Wikipedia tables is a heavily understudied problem. While there are first works on matching and finding incomplete Wikipedia infoboxes in various language versions [8] recently, inconsistencies in tables have not been explored at all.

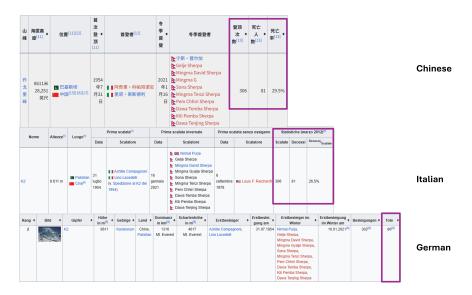


Fig. 1. Inconsistencies in death rate information across language versions of the Wikipedia articles about the Seven Summits in Chinese, Italian, and German.

As an example of these inconsistencies, in Figure 1, we show an example of tables about the Seven Summits in 3 different languages. While these tables contain the same seven mountains, they have a different structure and are inconsistent. We aim to categorize these inconsistencies, understand their causes, and assess their implications for knowledge quality and multilingual AI applications.

# 2 Research Questions

In this work, we explore three central research questions:

- 1. What kinds of factual inconsistencies occur across Wikipedia language editions? Multilingual tables about the same entity may diverge in the facts they present, the structure of attributes, or the presence of information. These inconsistencies include contradictions, omissions, and misaligned values.
- 2. Can these inconsistencies be categorized into meaningful types? We aim to define a taxonomy of inconsistency types—such as omissions, contradictions, structural differences, and partial overlaps—to support systematic analysis across languages.
- 3. What are the underlying causes of these inconsistencies? We investigate whether editorial practices, asynchronous updates, cultural framing, or language-specific conventions contribute to the observed divergences.

# 3 Knowledge Graphs for Reliable AI

KGs provide structured representations of entities and their relations, offering a foundation for consistent and explainable AI. In multilingual contexts, they serve as a common reference across languages, reducing ambiguity and supporting alignment. Wikidata, for example, enables cross-lingual linking of Wikipedia content, including tables, through shared identifiers. This makes it possible to detect and analyze inconsistencies in Wikipedia tables. Wikipedia remains one of the most widely used sources for pretraining large language models, making data quality a central concern in reliable AI [1]. By grounding extracted facts in a knowledge graph, we can improve the reliability of downstream tasks such as question answering, entity linking, or summarization, and increase the robustness of AI systems that rely on multilingual and collaborative sources like Wikipedia.

# 4 Related Works

Bias and inconsistency in knowledge sources are well-documented challenges. A comprehensive survey shows how definitions of bias influence methods and evaluations in NLP [2]. Cultural and linguistic bias has been observed in Wikipedia and Wikidata. For example, comparative studies of biographies highlight framing differences across language editions [3], while demographic attributes such as nationality or ethnicity are modeled inconsistently in Wikidata [12].

Existing work has addressed these issues through automated bias detection in Wikipedia articles [6] and curated probes to test cultural knowledge in language models [7]. However, structured content such as tables remains largely unstudied, despite their central role in many Wikipedia articles and the high variability across language editions. Efforts to unify Wikipedia content, such as Abstract Wikipedia [14,15], introduce a language-independent layer, but do not account for inconsistencies in existing articles or enable cross-lingual comparison of structured data. To date, no large-scale effort has examined how Wikipedia tables diverge across language editions.

As a related effort for assessing knowledge consistency, [17] implemented Cross-lingual Semantic Consistency (xSC) metrics and examined whether models provide semantically consistent responses to the same Wikipedia information in different languages, using multilingual semantic encoding models like LASER. Recent approaches leverage deep learning models and natural language processing techniques to parse table schemas, extract entity relationships, and convert tabular content into knowledge graphs or structured databases for downstream applications [9,11,10].

While prior work has focused on synchronizing Wikipedia Infoboxes using LLMs to enrich or update missing information [8], our study investigates factual inconsistencies in already existing multilingual Wikipedia tables. In contrast to Wikipedia Infoboxes, tables have an even larger structural variety, because they often describe many entities, while Infoboxes usually concern the main entity of the respective Wikipedia article. This adds a lot of additional new challenges.

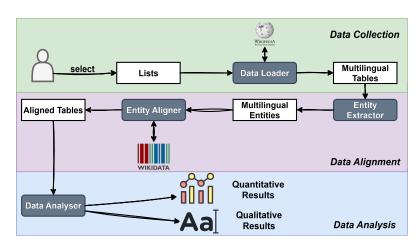


Fig. 2. Methodology Overview

# 5 Methodology

Figure 2 presents an overview of our proposed approach. To address our research questions, we adopt a data-driven methodology with three steps:

#### 1. Data Collection

Our overarching goal is to analyze the content of multilingual Wikipedia pages to study cross-lingual inconsistencies in factual information. As a first step toward this objective, we focus on tabular data, which often provides high-density, structured information such as statistics, factual lists, and attributes of related entities. Tables also present a manageable starting point for aligning content across languages due to their structural format. To this end, we manually select 10 entities from the Wikipedia *List of lists of lists*<sup>7</sup>, and for each entity, we retrieve the full set of tables from its Wikipedia page in multiple languages using a custom Python script.

#### 2. Data Alignment

After collecting the tables, we perform entity-level processing to align tabular content across languages, which includes entity extraction and entity linking:

#### - Entity Extraction

Each extracted table typically contains information about multiple entities (e.g., mountains, rivers, lakes). To prepare for cross-lingual comparison, we identify the entity represented in each row, usually based on the name or link in the first column. These entity mentions are then extracted for subsequent alignment across languages.

#### Entity Linking

To unify entity mentions across languages, we leverage Wikidata IDs as a language-independent identifier. For each row-level entity mention,

<sup>7</sup> https://en.wikipedia.org/wiki/List\_of\_lists\_of\_lists

we extract the internal Wikipedia link and query the MediaWiki API to resolve its corresponding Wikidata QID. This allows us to associate mentions like *Mount Everest* (English), 珠穆朗玛峰 (Simplified Chinese), *Il monte Everest* (Italian) with the same unique identifier *Q513*.

### 3. Data Analysis

Following alignment, we perform a quantitative and qualitative assessment of inconsistencies in tabular data across languages. The dataset and specific evaluation metrics used in our analysis are detailed in Section 6.

#### 6 Dataset and Metrics

The goal of the data collection process is to obtain high-quality multilingual tabular data from Wikipedia pages. To ensure a comprehensive evaluation of multilingual inconsistencies, we manually select entities exclusively from the *Geography* domain for consistency and clarity.

As shown in Table 1, basic statistics of the dataset demonstrate the number of language editions available for selected Wikipedia articles related to geographical and geological topics, revealing substantial variation in multilingual coverage ranging from 6 to 58 languages. For comparison, we extract English, German, Chinese, Italian, and Dutch versions, which represent widely used languages.

Title	Language Versions
Seven Summits	58
Eight-thousander	57
List of mountains of the Alps over 4000 metres	15
Lists of earthquakes	38
List of highest unclimbed peaks	6
List of highest mountains on Earth	47
Lakes of Titan	18
List of largest lakes of Europe	18
List of lakes by area	46

Table 1. Number of language versions for selected Wikipedia articles

For evaluation, we utilize various quantitative metrics to assess the collected dataset: (1) **Table count**: Although Wikipedia pages describe the same entities, they employ different numbers of tables to elaborate on them. (2) **Reference count**: Editors attach references at the end of pages, indicating their level of engagement in providing content. (3) **Column count**: This provides a straightforward way to observe how much detail editors include about entities, as some columns may contain missing cells.

Additionally, we qualitatively analyze the following metrics: (1) **Invalidity**: The provided value is incorrect or not credible. (2) **Timeliness**: A data source may present information that was once valid but is now outdated, whereas another source may provide an updated version. (3) **Incompleteness**: Schema-

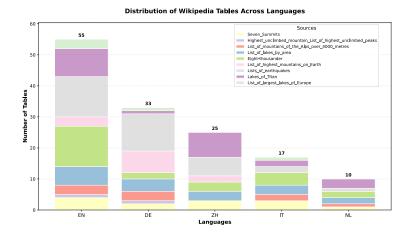


Fig. 3. Distribution of Table Numbers across Languages

level incompleteness, where different language versions provide different information to describe the underlying subject.

## 7 Experiments

To address the research questions defined in Section 2, we conduct a series of pilot experiments following the methodology outlined earlier. We prepare a small-sized dataset and perform data alignment. After storing the aligned data, we analyze it using both qualitative and quantitative methods based on the defined evaluation metrics. The quantitative analysis provides comprehensive statistics about the dataset, as detailed in Section 7.1. In the qualitative analysis, we categorize the observed inconsistencies and provide specific examples for each inconsistency type, which are described in Section 7.2.

#### 7.1 Quantitative Analysis

The quantitative analysis examines Wikipedia's multilingual content consistency using three key metrics: table count, reference count, and column count. The findings reveal significant disparities in content coverage and quality:

Table Count: The stacked bar chart in Figure 3 reveals the distribution of Wikipedia tables across nine geographic and geological articles in five languages, with English (EN) dominating at 55 tables total, followed by German (DE) at 33 tables, Chinese (ZH) at 25 tables, Italian (IT) at 17 tables, and Dutch (NL) at 10 tables. While English maintains the largest contribution across most articles, German shows notable prominence in specific geographic topics, particularly excelling in mountain-related content such as "List of mountains of the Alps over 4000 metres" and "List of highest mountains on Earth," where it comprises a substantial portion of the available tables. Chinese Wikipedia, despite having a

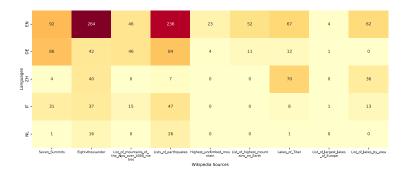


Fig. 4. Distribution of Reference Numbers across Languages

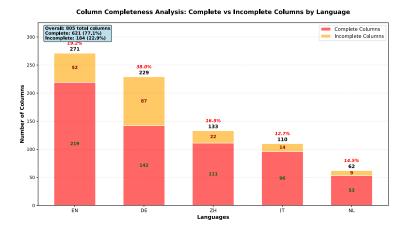


Fig. 5. Information Completeness Analysis with Column Counts

considerable number of tables where data exists, shows gaps in coverage with only six of the nine articles represented, indicating missing content for three entire articles. Similarly, both Italian and Dutch Wikipedia exhibit incomplete coverage across the articles and maintain fewer tables even in the articles they do cover, with Dutch having the most limited representation overall. This distribution pattern suggests that English Wikipedia serves as the most comprehensive resource for geographic and geological tabular data, while German Wikipedia demonstrates specialized prominence in mountain documentation, and the other languages show varying degrees of coverage gaps and content limitations.

Reference Count: The heatmap in Figure 4 displays the distribution of references, with color intensity indicating reference density. English Wikipedia demonstrates the highest reference count with an average of 94.6 references per article, significantly outperforming other language versions. Notable standouts in English include "Eight-thousander" with 264 references and "Lists of earthquakes" with 236 references, representing the most comprehensive documentation among all articles. German Wikipedia shows moderate reference activity

with an average of 31.8 references per article, displaying relatively consistent coverage across most articles. Chinese Wikipedia presents a more selective pattern with an average of 17.4 references per article, showing concentrated effort in specific topics, most notably "Lakes of Titan" with 70 references, while having limited or no coverage for other articles. Italian Wikipedia maintains lower documentation with an average of 16.9 references per article, showing relatively balanced coverage across available articles. Dutch Wikipedia exhibits the most limited reference activity with an average of 4.9 references per article, with sparse coverage across most articles.

Column Count: The bar chart in Figure 5 displays column completeness analysis across the main tables from nine Wikipedia articles in five languages. This analysis focuses on the primary table from each Wikipedia page, as these represent the core information and are more significant than other tables. English Wikipedia demonstrates the highest volume with 271 total columns, of which 219 are complete and 52 are incomplete, resulting in a 19.2% incompleteness rate. German Wikipedia shows 229 total columns but has the highest incompleteness rate at 38.0%, with 87 incomplete columns out of 229 total, significantly reducing its complete column count to 142. In contrast, the other three languages exhibit much better data completeness: Chinese Wikipedia has 133 total columns with only 16.5% incomplete, Italian Wikipedia shows 110 total columns with 12.7% incomplete, and Dutch Wikipedia has 62 total columns with 14.5% incomplete. Overall, across all 805 columns from the five language versions, 621 columns (77.1%) are complete while 184 columns (22.9%) contain missing or incomplete information, indicating that while English provides the most comprehensive coverage in terms of volume, German Wikipedia faces significant data quality challenges despite having substantial content.

# 7.2 Qualitative Analysis

To categorize various types of inconsistencies, we draw inspiration from the classification framework of knowledge deltas [4], and contextualize these categories with illustrative examples drawn from Wikipedia pages, we identify three potential sources of inconsistency that may affect it:

Invalidity: The value is incorrect or lacks credibility. As shown in Figure 1, Wikipedia tables across different languages report conflicting death rate statistics for Mount Everest and K2. For example, the death rate for K2 is listed as 29.5% in the Chinese version, 26.5% in the Italian version, and inferred as 26.4% in the German version (80 deaths out of 302 ascents), despite referencing similar or identical data. Such discrepancies highlight reliability issues in multilingual content consistency.

**Timeliness:** A data source may present a statement t that was once valid but is now outdated, whereas another source may provide an updated version of t. For example, in Figure 6, the English version of the Wikipedia page states that Mount Everest's height is 8,849 meters, while another language version reports it as 8,848 meters. This discrepancy reflects a recent update, as MountEverest has continued to grow—its height increasing by approximately 2mm per year—due

to tectonic uplift and enhanced isostatic rebound triggered by erosion from river capture near the Arun River [5].

Incompleteness: One type of cross-lingual inconsistency we observe is schemalevel incompleteness, where languages provide different sets of metrics (i.e., column headers) to describe the same underlying subject. Figure 7 illustrates this phenomenon using a binary heatmap that compares the metrics used in the Wikipedia tables for the List of climbers who have summited all 14 eight-thousanders across five language versions. While there is a clear overlap in core metrics such as rank, name, period, and nationality, several metrics are language-specific. For instance, the Dutch version includes unique metrics like new route and winter ascent, while the Italian version includes gender. On the other hand, some languages omit attributes found in others, where duration is present only in Chinese and Italian. This highlights the partial and uneven coverage of information across languages, even when describing the same real-world entities.

#### 8 Conclusions

In this work, we find that factual inconsistencies across language editions of Wikipedia include not only divergent values but also outdated data and missing or different structured content. We proposed a classification of these inconsistencies into three categories: Invalidity, Timeliness, and Incompleteness. We contextualized these categories with real-world examples from multilingual Wikipedia tables, highlighting how inconsistencies in statistical values, outdated information, and uneven schema coverage can all undermine the reliability of a knowledge. Since LLMs often rely on Wikipedia data, language discrepancies stemming from cultural differences can lead to biased knowledge representations and distortions in AI information processing, potentially disadvantaging certain users or perspectives. As for the underlying causes, we find that inconsistencies often arise from asynchronous updates across editions and from differences in source materials or cultural emphasis. The integration of uncertain and multilingual knowledge continues to pose significant challenges. While progress has been made in representing uncertainty at both the ontological and data model levels, current approaches do not yet address the full diversity of inconsistencies.

For future work, we will focus on scaling the proposed analysis across a larger set of multilingual Wikipedia tables. This includes systematic selection of topicaligned tables, aggregation of column headers, and cross-lingual embedding using models such as mT5, and XLM-RoBERTa. Pairwise cosine similarities will be computed to assess alignment, with annotated heatmaps used for visual analysis. Human evaluation will support semantic comparison of column headers to identify structural and editorial divergences. The results will refine inconsistency categories and reveal cultural influences on schema design across languages.

**GenAI Usage Disclosure.** The authors used ChatGPT and Grammarly for limited assistance with writing polish and code debugging. All scientific content, including text, code, tables, figures, and claims, was authored by the authors.

#### References

- Albalak, A., Elazar, Y., Xie, S.M., Longpre, S., Lambert, N., Wang, X., Muennighoff, N., Hou, B., Pan, L., Jeong, H., et al.: A survey on data selection for language models. arXiv preprint arXiv:2402.16827 (2024)
- Blodgett, S., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of "bias" in nlp. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5454–5476 (2020)
- Callahan, E., Herring, S.: Cultural bias in wikipedia content on famous persons. Journal of the American Society for Information Science and Technology 62(10), 1899–1915 (2011)
- Ferreira, T.C., Paul, D., Stuckenschmidt, H., Lehmann, J.: Uncertainty management in the construction of knowledge graphs: a survey (2024). https://doi.org/10.48550/arXiv.2405.16929, https://arxiv.org/abs/2405.16929
- Han, X., Dai, J.G., Smith, A.G., Xu, S.Y., Liu, B.R., Wang, C.S., Fox, M.: Recent uplift of chomolungma enhanced by river drainage piracy. Nature Geoscience (Sep 2024), published online 30 September 2024
- 6. Hube, C., Fetahu, B.: Detecting biased statements in wikipedia. In: Companion Proceedings of the The Web Conference 2018. pp. 1779–1786. International World Wide Web Conferences Steering Committee (2018)
- Keleg, A., Magdy, W.: Dlama: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 6245–6266. Association for Computational Linguistics (2023)
- Khincha, S., Kataria, T., Anand, A., Roth, D., Gupta, V.: Leveraging LLM for synchronizing information across multilingual tables. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 6474-6492. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/ v1/2025.naacl-long.329, https://aclanthology.org/2025.naacl-long.329/
- 9. Kruit, B., Boncz, P., Urbani, J.: Extracting novel facts from tables for knowledge graph completion. In: Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., Gandon, F. (eds.) The Semantic Web ISWC 2019. pp. 364–381. Springer International Publishing, Cham (2019)
- 10. Kruit, B., Boncz, P., Urbani, J.: Takco: A platform for extracting novel facts from tables. In: Companion Proceedings of the Web Conference 2021. p. 705–707. WWW '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442442.3458611, https://doi.org/10.1145/3442442.3458611
- Kruit, B., He, H., Urbani, J.: Tab2know: Building a knowledge base from tables in scientific papers. In: Pan, J.Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web – ISWC 2020. pp. 349–365. Springer International Publishing, Cham (2020)
- 12. Shaik, Z., Ilievski, F., Morstatter, F.: Analyzing race and country of citizenship bias in wikidata. arXiv preprint arXiv:2108.05412 (2021)
- 13. Tatariya, K., Kulmizev, A., Poelman, W., Ploeger, E., Bollmann, M., Bjerva, J., Luo, J., Lent, H., de Lhoneux, M.: How good is your wikipedia? arXiv preprint arXiv:2411.05527 (2024)

- 14. Vrandečić, D.: Architecture for a multilingual wikipedia. arXiv preprint arXiv:2004.04733 (2020)
- Vrandečić, D.: Building a multilingual wikipedia. Communications of the ACM 64(4), 38–41 (2021)
- 16. Wikipedia contributors: List of wikipedias Wikipedia, the free encyclopedia (2025), https://en.wikipedia.org/w/index.php?title=List\_of\_Wikipedias&oldid=1294528760, [Online; accessed 12-June-2025]
- 17. Xing, X., He, Z., Xu, H., Wang, X., Wang, R., Hong, Y.: Evaluating knowledge-based cross-lingual inconsistency in large language models. arXiv preprint arXiv:2407.01358 (2024)

# 9 Appendix



**Fig. 6.** Example of timeliness: height of Mount Everest differs across language versions of Wikipedia. The death rate of climbing the mountain is explicitly provided in (a) Italian and (b) Chinese. Only the absolute number of deaths is mentioned in (c) German. This information is absent in the English, Dutch, and Estonian versions.

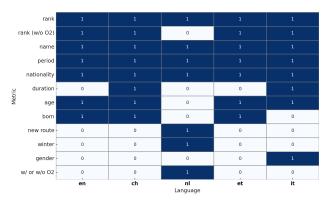


Fig. 7. Example of incompleteness: Binary heatmap of metric presence across five languages for the List of climbers who have summited all 14 eight-thousanders.