多语言维基百科表格中的事实不一致

Silvia Cappa[1], Lingxiao Kong[2], Pille-Riin Peet[3], Fanfu Wei[4], Yuchen Zhou[5], and Jan-Christoph Kalo[6]

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

Abstract. 维基百科作为一个全球可访问的知识来源，拥有超过 300 种语言的内容。尽管涵盖相同的主题，维基百科的不同版本是独立撰写和更新的。这导致了事实上的不一致性，可能会影响百科全书和经常依赖维基百科作为主要训练来源的人工智能系统的中立性和可靠性。本研究调查了维基百科结构化内容中的跨语言不一致性，特别关注表格数据。我们开发了一种方法，来收集、对齐和分析维基百科多语言文章中的表格，定义不一致的类别。我们应用了各种定量和定性的指标，通过一个样本数据集来评估多语言的对齐。这些见解对事实核查、多语言知识交互以及设计利用维基百科内容的可靠 AI 系统具有重要意义。

1 引言

维基百科是最广泛使用的公共知识来源之一，提供超过 300 种语言的内容 [15] 。尽管不同语言版本的文章通常旨在描述相同的实体和事件，但它们在呈现的事实和内容上经常存在差异，因为这些内容并不是简单翻译而来的，而是基于各自语言版本独立编写的 [12] 。这些不一致性引发了关于多语言内容的可靠性、完整性、丰富性和中立性的重要问题。这个项目调查了维基百科不同语言版本之间的各种不一致现象，特别关注结构化数据如表格的差异。维基百科表格之间的不一致性是一个研究很少的问题。虽然最近开始有一些关于匹配和寻找各种语言版本中不完整的维基百科信息框的工作 [7] ，但表格中的不一致性完全没有被探讨。

作为这些不一致的一个例子，在图 1 中，我们展示了一些关于七大高峰的表格，这些表格以三种不同的语言呈现。虽然这些表格包含相同的七座山峰，但它们的结构不同且不一致。我们的目标是对这些不一致性进行分类，了解其原因，并评估其对知识质量和多语言人工智能应用的影响。

在这项工作中，我们探讨三个核心研究问题：

2 用于可靠人工智能的知识图谱

知识图谱提供了实体及其关系的结构化表示，为一致和可解释的人工智能提供基础。在多语言场景中，它们充当跨语言的公共参考，减少歧义并支持对齐。例

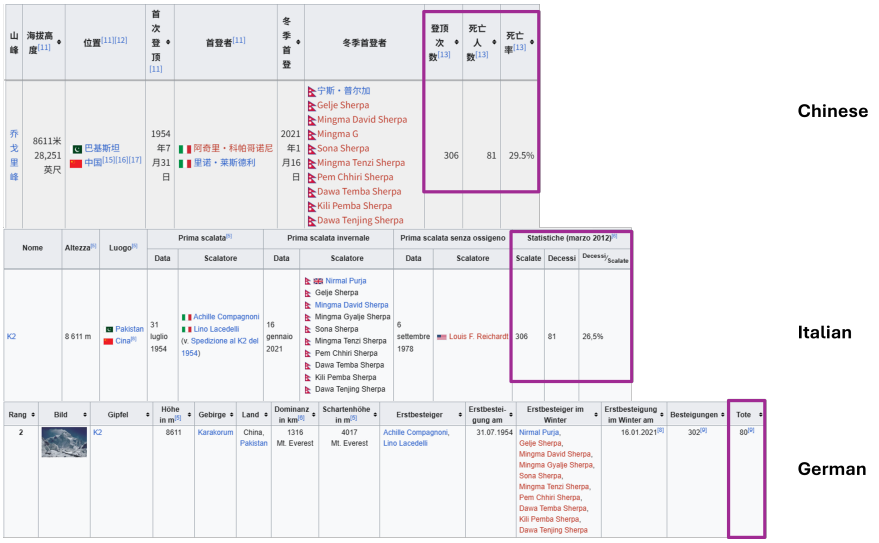

Fig. 1. 中文、意大利文和德文的维基百科关于七大洲最高峰的条目中，死亡率信息不一致。

如，Wikidata 通过共享标识符实现了 Wikipedia 内容（包括表格）的跨语言链接。这使得检测和分析 Wikipedia 表格中的不一致成为可能。Wikipedia 仍然是大规模语言模型预训练的最广泛使用的来源之一，使数据质量成为可靠人工智能的核心问题 [1] 。通过在知识图谱中定位提取的事实，我们可以提高下游任务（如问答、实体链接或摘要）的可靠性，并增强依赖 Wikipedia 等多语言和协作源的人工智能系统的稳健性。

3 相关工作

知识来源中的偏见和不一致是有据可查的挑战。一项全面的调查显示，偏见的定义如何影响 NLP 中的方法和评估 [2] 。在维基百科和维基数据中观察到了文化和语言偏见。例如，传记的比较研究突出显示了跨语言版本的框架差异 [3] ，而像国籍或种族等人口统计属性在维基数据中被不一致地建模 [11] 。

现有的工作已经通过自动化偏见检测在维基百科文章 [5] 中解决了这些问题，并通过策划的探测来测试语言模型中的文化知识 [6] 。然而，诸如表格之类的结构化内容仍然很少被研究，尽管它们在许多维基百科文章中起着核心作用，并且在不同语言版本之间的差异很大。试图统一维基百科内容的努力，例如抽象维基百科 [13,14] ，引入了一个独立于语言的层，但没有考虑现有文章中的不一致性或实现结构化数据的跨语言比较。迄今为止，还没有大型的努力来研究维基百科表格在不同语言版本中是如何不同的。

作为一种评估知识一致性的相关努力，[16] 实现了跨语言语义一致性（xSC）度量，并审查模型是否能够对不同语言中相同的维基百科信息提供语义一致的响应，使用多语言语义编码模型如 LASER。最近的方法利用深度学习模型和自

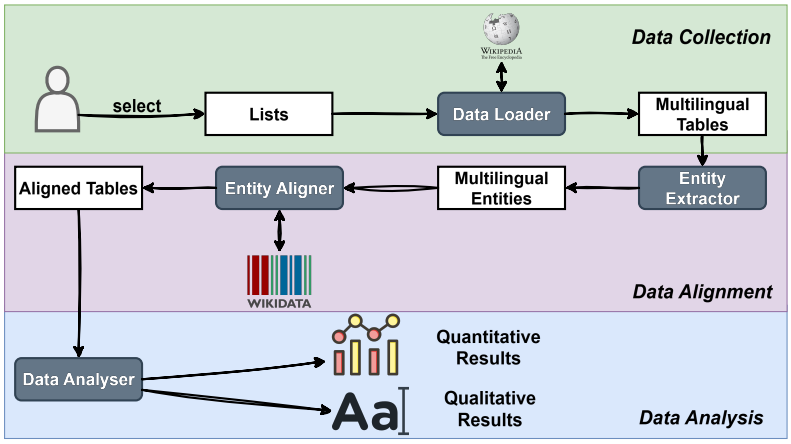

Fig. 2. 方法论概述

然语言处理技术来解析表格架构、提取实体关系，并将表格内容转换为知识图谱或结构化数据库以用于下游应用 [8,10,9] 。

尽管之前的研究主要集中于使用大型语言模型同步维基百科信息框，以丰富或更新缺失的信息 [7] ，我们的研究则调查已存在的多语言维基百科表格中的事实不一致问题。与维基百科信息框相比，表格具有更大的结构多样性，因为表格通常描述很多实体，而信息框通常涉及相应维基百科文章的主要实体。这带来了许多额外的新挑战。

4 方法论

图 2 展示了我们提出的方法的概述。为了回答我们的研究问题，我们采用了一个包含三个步骤的数据驱动方法：

1. 数据收集
   我们的总体目标是分析多语言维基百科页面的内容，以研究跨语言的事实信息不一致性。作为实现这一目标的第一步，我们专注于表格数据，表格通常提供高密度、结构化的信息，如统计数据、事实列表和相关实体的属性。由于表格的结构化格式，它们也为跨语言对齐内容提供了一个可管理的起点。为此，我们从维基百科的列表的列表 [7] 中手动选择 10 个实体，并使用自定义的 Python 脚本，从多种语言的维基百科页面中检索每个实体的完整表格集合。
2. 数据对齐
   收集表格后，我们进行实体级处理以对齐不同语言的表格内容，这包括实体提取和实体链接：
   - 实体提取
     每个提取的表格通常包含关于多个实体的信息（例如山脉、河流、湖泊）。

[7] https://en.wikipedia.org/wiki/List_of_lists_of_lists

为了准备跨语言比较，我们识别出每一行表示的实体，通常基于第一列中的名称或链接。然后提取这些实体提及，以便跨语言对齐。
- 实体链接
UTF8gbsn To unify entity mentions across languages, we leverage Wikidata IDs as a language-independent identifier. For each row-level entity mention, 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. 数据分析
在对齐之后，我们对跨语言的表格数据中的不一致性进行定量和定性评估。我们分析中使用的数据集和特定评估指标在第 5 节中详细说明。

5 数据集与指标

数据收集过程的目标是从维基百科页面获取高质量的多语言表格数据。为了全面评估多语言不一致性，我们专门从地理领域手动选择实体，以保证一致性和清晰度。

如表 1 所示，数据集的基本统计数据显示了与地理和地质主题相关的选定维基百科文章的语言版本数量，揭示了从 6 种到 58 种语言的多语言覆盖的显著差异。为了进行比较，我们提取了代表广泛使用的语言的英文、德文、中文、意大利文和荷兰文版本。

Table 1. 精选维基百科文章的语言版本数量

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

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.

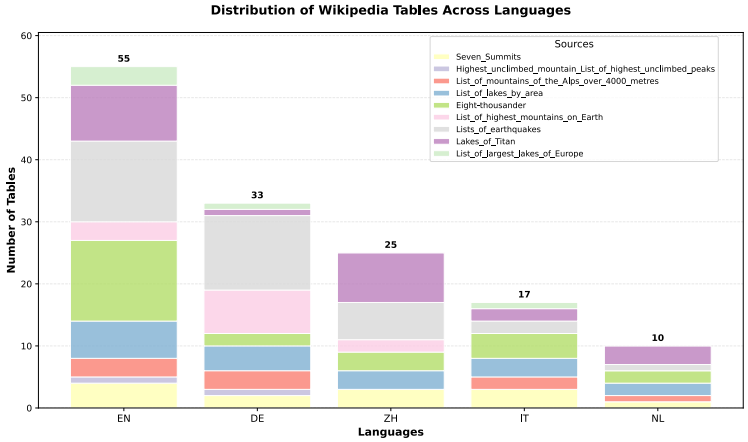

Fig. 3. 表格编号在各语言中的分布

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

6 实验

为了回答在第 **??** 节中定义的研究问题，我们按照前面提到的方法，进行了一系列的初步实验。我们准备了一个小型的数据集并进行了数据对齐。在存储了对齐的数据后，我们根据定义的评估指标使用定性和定量的方法对其进行分析。定量分析提供了关于数据集的全面统计，详细内容见第 6.1 节。在定性分析中，我们对观测到的不一致进行了分类，并为每种不一致类型提供了具体示例，这些描述在第 6.2 节中。

6.1 定量分析

定量分析使用三个关键指标来检查维基百科的多语言内容一致性：表格数量、参考文献数量和列数。研究结果显示内容覆盖和质量方面存在显著差异：
表格数量：图 3 中的堆叠条形图揭示了维基百科中表格在五种语言的九篇地理和地质文章中的分布，英语（EN）以总共 55 个表格占据主导地位，其次是德语（DE）33 个表格，中文（ZH）25 个表格，意大利语（IT）17 个表格和荷兰语（NL）10 个表格。虽然英语在大多数文章中保持着最大的贡献，但德语在特定的地理主题中表现出显著的优势，尤其是在与山有关的内容中表现出色，比如“超过 4000 米的阿尔卑斯山山峰列表”和“地球上最高峰列表”，其中它构成了可用表格的相当大一部分。尽管中文维基百科有相当数量的表格存在，但在覆盖面上表现出差距，只代表了九篇文章中的六篇，表明有三整篇文章的内容

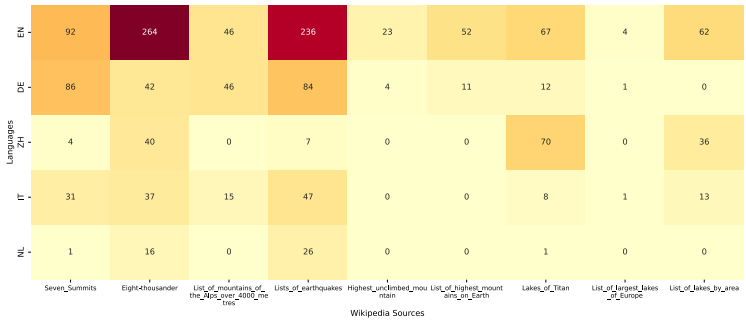

Fig. 4. 不同语言中的参考编号分布

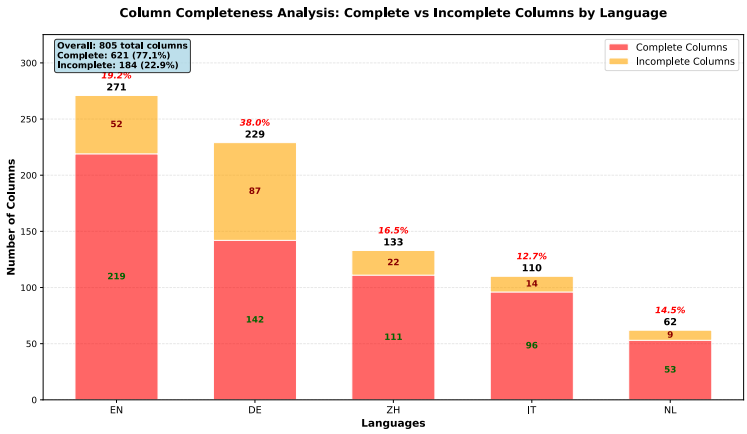

Fig. 5. 基于列计数的信息完备性分析

缺失。同样，意大利语和荷兰语维基百科在文章中均表现出不完整的覆盖，并且即使在他们覆盖的文章中保持较少的表格，其中荷兰语的总体代表性最为有限。这种分布模式表明，英语维基百科是地理和地质表格数据的最全面资源，而德语维基百科在山脉文档中显示出专业的显著性，其他语言则展现出不同程度的覆盖缺口和内容限制。

引用计数：图 4 中的热图显示了引用的分布情况，颜色的强度表示引用的密度。英文维基百科显示出最高的引用计数，每篇文章平均有 94.6 个引用，显著超越其他语言版本。英文维基百科中的显著例子包括“八千米山峰”，有 264 个引用，以及“地震列表”，有 236 个引用，代表所有文章中最全面的文档记录。德文维基百科显示出适度的引用活动，每篇文章平均有 31.8 个引用，在大多数文章中显示出相对一致的覆盖。中文维基百科呈现出较为选择性的模式，每篇文章平均有 17.4 个引用，显示出在特定主题上的集中努力，尤其是在“土卫六的湖泊”上有 70 个引用，而在其他文章上覆盖较少或没有。意大利文维基百科维持较低的文档记录，每篇文章平均有 16.9 个引用，显示出在现有文章上相对平衡的覆盖。荷兰文维基百科表现出最有限的引用活动，每篇文章平均有 4.9 个引用，大多数文章覆盖稀疏。

列数：图 5 中的柱状图显示了五种语言的九个维基百科文章的主要表格中的列完整性分析。此分析侧重于每个维基百科页面的主要表格，因为这些表格代表核心信息，比其他表格更重要。英语维基百科显示出最高的总列数，为 271 列，其中 219 列是完整的，52 列不完整，不完整率为 19.2 % 。德语维基百科显示出 229 列的总数，但具有最高的不完整率为 38.0 % ，在 229 列中有 87 列不完整，显著减少了完整列数到 142 。相比之下，其他三种语言显示出更好的数据完整性：中文维基百科有 133 列总数，仅有 16.5 % 不完整，意大利语维基百科有 110 列总数，12.7 % 不完整，荷兰语维基百科有 62 列总数，14.5 % 不完整。总体而言，在这五种语言版本的 805 列中，621 列（77.1 % ）是完整的，而 184 列（22.9 % ）包含缺失或不完整的信息，表明尽管英语在体量上提供了最全面的覆盖，但德语维基百科在数据质量方面面临着显著挑战，尽管其内容充实。

6.2 定性分析

为了对各种类型的不一致进行分类，我们从知识变更的分类框架 [4] 中汲取灵感，并结合从维基百科页面中提取的示例将这些类别进行情境化，我们确定了三种可能影响其不一致性的来源：

无效性：该值不正确或缺乏可信度。如图 1 所示，维基百科中不同语言的表格对珠穆朗玛峰和 K2 的死亡率统计报告存在冲突。例如，K2 的死亡率在中文版本中列为 29.5 %，在意大利文版本中为 26.5 %，在德文版本中推断为 26.4 %（302 次攀登中有 80 人死亡），尽管这些版本参考了相似或相同的数据。这种差异突显了多语言内容一致性中的可靠性问题。

及时性：一个数据源可能会呈现一个曾经有效但现在已过时的声明 t ，而另一个来源可能提供了 t 的更新版本。例如，在图 6 中，维基百科英文页面上说明珠穆朗玛峰的高度为 8,849 米，而另一种语言版本则报告为 8,848 米。这种差异反映了一个最近的更新，因为由于构造抬升和阿隆河附近河流捕获引起的侵蚀所触发的增强等静态反弹，珠穆朗玛峰的高度每年大约增加 2 毫米。

不完备性：我们观察到的一种跨语言不一致性是模式级别的不完备性，其中语言提供不同的指标集（即列标题）来描述相同的基本主题。图 7 使用一个二进制热图展示了这一现象，该热图比较了五种语言版本的维基百科表格中用于描述攀登所有 14 座八千米峰名单的指标。尽管在核心指标如排名、姓名、时期和国籍方面有明显的重叠，但一些指标是语言特有的。例如，荷兰语版本包括独特的指标如新路线和冬季攀登，而意大利语版本包括性别。另一方面，一些语言省略了其他语言中存在的属性，其中持续时间仅出现在中文和意大利语中。这突出了即使在描述相同的现实世界实体时，各语言信息覆盖的不完全性和不均匀性。

7 结论

在这项工作中，我们发现维基百科不同语言版本之间的事实不一致性不仅包括不同的数值，还包括过时的数据以及缺失或不同结构的内容。我们将这些不一致性分类为三类：无效性、时效性和不完整性。我们通过多语言维基百科表中的真实例子对这些类别进行了背景化，强调了统计值不一致、信息过时和模式覆盖不均的问题如何削弱知识的可靠性。由于大语言模型 (LLM) 通常依赖于维基百科数据，源于文化差异的语言差异可能导致偏见的知识表示和人工智能信息处理的扭曲，从而可能使某些用户或观点处于不利地位。至于潜在原因，我们发

现不一致性通常起源于不同版本之间更新的不同步以及来源材料或文化重点的不同。不确定和多语言知识的整合仍然是一个重大的挑战。尽管在本体和数据模型层面表示不确定性方面已经取得了一些进展，但当前的方法尚未解决不一致性的多样性问题。

在未来的工作中，我们将专注于将所提出的分析扩展到更大范围的多语言 Wikipedia 表格。这包括系统地选择主题对齐的表格、聚合列标题，以及使用诸如 mT5 和 XLM-RoBERTa 的模型进行跨语言嵌入。我们将计算成对的余弦相似度以评估对齐，并通过注释的热图进行视觉分析。人工评估将支持列标题的语义比较，以识别结构和编辑上的分歧。结果将优化不一致类别，并揭示语言间模式设计中的文化影响。

GenAI 使用披露。作者使用了 ChatGPT 和 Grammarly 辅助文章润色和代码调试。包括文本、代码、表格、图形和论点在内的所有科学内容均由作者撰写。

References

1. 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)
2. 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)
3. 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)
4. 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
5. 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)
6. 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)
7. 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/
8. 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)
9. 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
10. 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)
11. Shaik, Z., Ilievski, F., Morstatter, F.: Analyzing race and country of citizenship bias in wikidata. arXiv preprint arXiv:2108.05412 (2021)
12. 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)
13. Vrandečić, D.: Architecture for a multilingual wikipedia. arXiv preprint arXiv:2004.04733 (2020)
14. Vrandečić, D.: Building a multilingual wikipedia. Communications of the ACM **64**(4), 38–41 (2021)
15. 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]
16. 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)

8 附录

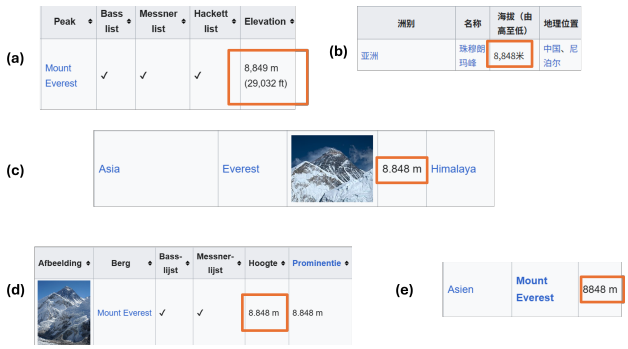

Fig. 6. 时效性的例子：珠穆朗玛峰的高度在不同语言版本的维基百科中存在差异。攀登这座山的死亡率在（a）意大利语和（b）中文中被明确提供。在（c）德语中只提到了绝对死亡人数。这一信息在英文、荷兰文和爱沙尼亚文版本中缺失。

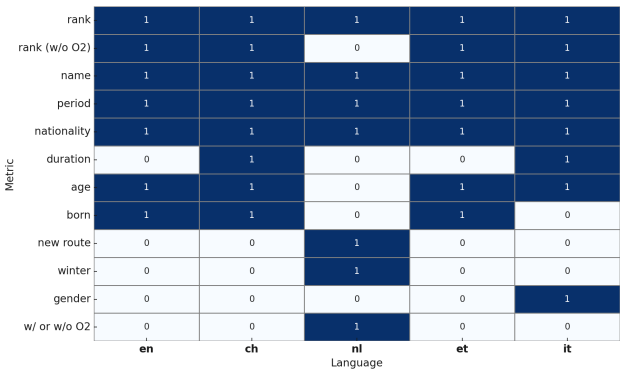

Fig. 7. 不完整的示例：五种语言中关于攀登过所有 14 座八千米山峰的登山者名单的指标存在情况的二进制热图。