Wikimedia data for AI: a review of Wikimedia datasets for NLP tasks and AI-assisted editing

Isaac Johnson
Wikimedia Foundation
United States
isaac@wikimedia.org

Lucie-Aimée Kaffee
Hugging Face
Germany
lucie.kaffee@huggingface.co

Miriam Redi
Wikimedia Foundation
United Kingdom
mredi@wikimedia.org

Abstract

Wikimedia content is used extensively by the AI community and within the language modeling community in particular. In this paper, we provide a review of the different ways in which Wikimedia data is curated to use in NLP tasks across pre-training, post-training, and model evaluations. We point to opportunities for greater use of Wikimedia content but also identify ways in which the language modeling community could better center the needs of Wikimedia editors. In particular, we call for incorporating additional sources of Wikimedia data, a greater focus on benchmarks for LLMs that encode Wikimedia principles, and greater multilingualism in Wikimedia-derived datasets.

1 Introduction

Wikimedia data—especially Wikipedia—has been essential to the progression of AI over the past several years. In particular, Wikipedia text is key to natural language processing (NLP): it is generally long-form (meaning lots of context to learn from), "well-written",[1] and high-quality (Gao et al., 2020). The BERT language model (Devlin, 2018) that was introduced in 2018 and is often considered the first modern LLM uses English Wikipedia as a majority of its data. Even today, with much larger language models, English Wikipedia is often weighted heavily when trained—e.g., (Brown, 2020; Longpre et al., 2024).

The usage of Wikimedia data for AI has both been beneficial as a source of high-quality data for NLP researchers and for directing attention to the Wikimedia projects. This relationship, however, has largely been incidental to Wikimedia's mission and openness, and many of the advances of NLP have not made it back to the Wikimedia projects. For example, the Wikimedia Foundation regularly publishes snapshots of the content on the Wikimedia projects. These "dumps" have been made available since at least 2005.[2] While researchers have long been considered an expected end-user, this data was not pre-processed in any way to support the NLP community. As a result, researchers used many different approaches for pre-processing this raw text to produce natural-language text for use in training models.[3] More recently, there have been explicit efforts to bring the Wikimedia and the ML communities closer together such as the Wiki-M3L[4] and NLP for Wikipedia[5] workshops, and standardized datasets such as Hugging Face's Wikipedia text,[6] There have also been concerns that the AI ecosystem might be depleting the very projects upon which it is built and stronger calls for developers of AI tools to view the knowledge commons not just a repository from which to extract data, but as a community to give back to—e.g., Commons (2023) and Foundation (2024).

In this paper, we make an effort to catalog the many AI and NLP-related datasets that draw on the Wikimedia projects to identify what gaps and opportunities exist. We frame this review following the calls for AI developers to contribute more to the knowledge commons. Specifically, we select the datasets in this paper with a focus on how NLP might be made more beneficial for the Wikimedia editor communities. Editors not only do the difficult work of synthesizing sources into the encyclopedic text consumed by readers and AI alike, they also engage in rich discussion and sensemaking around source reliability, fairly portraying content, and evaluating complex questions of nota-

[1] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

[2] https://meta.wikimedia.org/w/index.php?title=Data_dumps&oldid=216530

[3] See Johnson and Lescak (2022) for examples.

[4] https://meta.wikimedia.org/wiki/Wiki-M3L

[5] https://meta.wikimedia.org/wiki/NLP_for_Wikipedia_(EMNLP_2024)

[6] https://huggingface.co/datasets/wikimedia/wikipedia

Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia, pages 91–101

bility. Their work is guided by core content policies, which AI models must also be able to adhere to in order to be useful to the editing community. In the course of the analysis, we identify three major opportunities:

- Extend and diversify the subset of Wikimedia data used in AI research. This could include regular datasets of images along with associated captions for multimodal modeling, more attention paid to talk pages or other collaboration spaces on Wikipedia, and greater usage of the high-quality transcribed documents produced by Wikisource communities.

- Consider the needs of Wikimedia editors in evaluation of LLMs. While Wikipedia data is well-represented within common benchmark datasets, these tasks are almost exclusively oriented towards reader goals. Work is needed to extend benchmarks to better encode the needs of Wikimedia editors.

- Continue to extend models to be more multilingual, open-source, and compact to meet the needs of the Wikimedia projects.

2 Approach

To guide the knowledge of Wikimedia datasets and tasks that are relevant to this work, we searched for individual Wikimedia projects on Hugging Face's dataset search[7] and relied heavily on the authors' long experience working with Wikimedia data, developing natural language technologies, and collaborating with the Wikimedia communities. We list characteristic (not all) datasets for each stage of training, focusing on datasets and tasks that are oriented back towards the Wikimedia projects and that are at most a decade old. While we made an effort to build an exhaustive list, given the highly distributed nature of the Wikimedia movement and its research community, this overview might present some gaps. However, we believe that such potential gaps should not affect our conclusions in major ways, and we treat this as the start of a catalog that we will work to update as we learn more and more datasets are created.

The current training paradigms of LLMs depend on datasets at three major stages: pre-training, post-training, and evaluation.[8] Across these three stages, we detail how raw data is converted into datasets, tasks, and benchmarks to support the objectives of each stage. We see data, like the Wikimedia dumps, as relatively raw versions of what appears on the Wikimedia projects but in a form that is not directly useful for language models. We define datasets as data that has undergone pre-processing to anticipate a specific need, such as cleaning text to bring it closer to natural language. This pre-processing is important for turning Wikimedia content into high-quality datasets for language models to learn basic patterns of language (pre-training). We define tasks as datasets with explicit inputs and outputs that can be used to fine-tune models to complete a given action (post-training). In the final stage, benchmarks are curated tasks that allow for easy comparison of models to determine their usefulness to the Wikimedia projects (evaluation). While Wikimedia data has long been available and researchers have developed many datasets and tasks from this data, Wikimedia benchmarks have received less attention but are also an important mechanism for enabling members of the Wikimedia NLP community to encode our expectations for language models that are using Wikimedia content.

2.1 What makes a dataset helpful to Wikimedia?

There are many many datasets that use Wikimedia data but not all of them relate to tasks that are clearly of value to the Wikimedia editor community.[9] For example, SQuAD (Rajpurkar et al., 2016) is a Q&A dataset that is derived from Wikipedia that has played important role within the NLP community, but Q&A does not necessarily map to a task where AI could directly help Wikimedia editors. While different editors and communities will have different needs, we highlight a few core principles that guide these needs and would ideally be expressed in datasets and the resulting models trained on Wikimedia data:

- **Multilinguality**: Wikipedia alone exists in over 300 languages and providing equitable support to these different communities means

[7] https://huggingface.co/datasets

[8] See (Dubey et al., 2024) for a good rundown of pre-training/post-training and Bowman and Dahl (2021) for a good overview of evaluation of LLMs.

[9] We focus here on editors, but there are many other contributors to the Wikimedia projects that are also valuable stakeholders for future consideration such as campaign organizers or tool developers.

building NLP tools that can handle their diversity.

- **Core content policies**: editors follow three core content policies[10] that guide content on Wikipedia and useful models would need to do the same: Neutral Point-of-View (NPOV; fair representation of significant viewpoints), Verifiability (citations), and No Original Research (do not reach conclusions beyond the reliable sources).

- **Openness**: "free" and "open" are important to Wikimedia in many ways.[11] In this context, language models are most useful when they are open-source and small enough to be reasonably hosted by the community, e.g., through the non-profit Wikimedia Foundation.

2.2 From data to benchmarks: a case study

Wikipedia articles offer an illustrative example of how data can be curated to support the three stages of training while adhering to the principles listed above. Starting with raw data, regular snapshots of the content of the Wikimedia projects have long been available as freely-downloadable dumps of article wikitext (the markup language used to write Wikipedia articles). For these dumps to be useful for most natural language applications—i.e. converted from raw data into a dataset—researchers both need to apply some basic filtering at the page level to remove non-content pages such as redirects and strip out the wikitext syntax from the pages to leave something closer to natural language.[12] These resulting natural language datasets are useful for pre-training but still require the identification of specific inputs and outputs to be converted into a task that can be used for post-training. As an example of post-training, Qian et al. (2023) explore the task of writing short articles using an extensive dataset of Wikipedia article titles as inputs and the cleaned article text as expected outputs. Their metrics for automatic evaluation of the generated articles focus on language fluency and factualness. While this work is valuable for NLP fields like knowledge-intensive Q&A, it only briefly explores metrics that capture Wikimedia principles such as Verifiability (appropriate citations). This makes the work less useful to the Wikimedia community as a benchmark that could allow for direct comparison of LLMs at assisting Wikimedians in producing high-quality content.

In contrast, FreshWiki (Shao et al., 2024) more directly aims to be this benchmark: it is a curated dataset of English Wikipedia articles that have been assessed to be of high quality (more likely to adhere to Wikimedia content policies) and that have been written largely after a specific cut-off date (to avoid data leakage due to memorization of Wikipedia content by LLMs). FreshWiki further incorporates citations in the expected output and adds metrics to measure how faithful the content is to its citations (see Table 2). While FreshWiki currently only exists in English, this same process could be extended to other language editions as there is nothing English-specific about it. Shao et al. (2024) evaluate GPT-4's performance, which is neither compact nor open-source, on FreshWiki because (Gao et al., 2023) had previously shown that more open models (LLaMA-2 70B) performed well at generating text but still lagged behind GPT-4 in terms of correctly citing sources. Altogether, FreshWiki was able to better model Wikipedia's core content policies but exposed gaps in open models in this domain and is a framework that can be easily extended to be more multilingual.

3 A review of curated Wikimedia data

3.1 Pre-training: from data to datasets

Pre-training datasets for language models are collections of unsupervised text—i.e. no explicit task associated with them – that can be used to train language models to understand the basic relationships between words (tokens).[13] These datasets are maximally useful when they are large, high-quality, and diverse. Datasets of Wikipedia articles are the prime example of this but they are not the only source of pre-training datasets available from the Wikimedia projects. Here, the needs of the Wikimedia projects are generally well-aligned with the needs of NLP researchers: better pre-training data means better models which can then be used to support the Wikimedia projects.

We distinguish here between whether data (raw content) is available and if there are standard datasets (pre-processed text). Table 1 shows two clear gaps: 1) raw data about image pixels and their associated text for pre-training of multimodal mod-

[10] https://en.wikipedia.org/wiki/Wikipedia:Core_content_policies

[11] https://w.wiki/B5zh

[12] See Guo et al. (2020) for an illustrative example.

[13] While we focus on language models, we also include some image-text data here.

Major source of text	Data available?	Pre-processed dataset?
Wikipedia articles	Various dumps[14]	Hugging Face[15]
Wikimedia Talk pages	Various dumps	One-offs such as WikiConv (Hua et al., 2018)
Commons Images + captions / alt-text	None	One-offs such as WIT (Srinivasan et al., 2021) or Concadia (Kreiss et al., 2022)
Wikisource transcriptions	Various dumps	Hugging Face[16]
Wikisource image-transcription pairs	None	None
Other Wikimedia projects (Wikibooks, Wikivoyage, Wikiversity, Wiktionary)	Various dumps	None

Table 1: Major data(sets) of Wikimedia content.

els is lacking, and, 2) even when the raw data is available, it is rare that standardized, pre-processed datasets are available that lower the barrier to access for researchers.[17]

We encourage continued work to identify good practices for converting the other data sources listed in Table 1 into datasets. Each content source will bring its own challenges but the popularity of the Hugging Face Wikipedia dataset proves its value.[18] For example, Wikisource offers an exciting opportunity to diversify the knowledge on which language models are being trained given the contributions by the Wikisource communities in digitizing knowledge from languages that have historically been underrepresented online.[19] Generating image datasets[20] will take much more work and resources given the massive size of the imagery hosted on Wikimedia Commons but would be a worthy addition to the outsized role that Wikimedia content plays in pre-training datasets.

One very positive aspect of the state of Wikimedia content for pre-training is that all of the data and almost all of the datasets are massively multilingual. While each of Wikipedia's over 300 language editions has varying norms and content, tools for converting this data into datasets generally are language-agnostic—i.e. they are stripping out syntax or making other choices that do not rely on tokenization or language-specific semantics. This helps to fuel a positive feedback loop of more multilingual content leading to more multilingual AI and thus more support for growing these language editions (Costa-jussà et al., 2022). As will be seen below, this wealth of language data unfortunately does not always hold for post-training datasets.

3.2 Post-training: from datasets to tasks

Post-training datasets for language models are collections of supervised tasks that can be used to fine-tune models to be more useful for end-users. Traditional fine-tuning converts a model from general language modeling to accomplishing a specific task that leverages a model's pre-trained language capabilities. Most LLMs are now instruction-tuned to not do any specific task but be generally capable of accomplishing many types of tasks.[21]

Below, we catalog these fine-tuning tasks with the goal of showing how Wikimedia content can be valuable in post-training and encouraging development of models that are more useful for Wikimedia-relevant tasks. Arguably the most salient usage of Wikimedia content for language modeling is related to Q&A tasks—e.g., SQuAD (Rajpurkar et al., 2016) or WikiQA (Yang et al., 2015). Q&A is a reader-focused task and one that receives plenty of attention in language modeling. Here we choose to focus on the needs of Wikimedia editors. In this domain, we see ample opportunity for LLM developers to make greater use of these Wikimedia-based post-training tasks. This would be beneficial for Wikimedians but should also support the general alignment goals of LLM developers as we will

[17] While this reduced barrier to entry feels appropriate for pre-training given that Wikipedia content is freely-licensed, we do encourage researchers to understand more deeply the content and processing choices that they are making when it comes to post-training.

[18] Over 100,000 downloads in August 2024 per https://huggingface.co/datasets/wikimedia/wikipedia.

[19] https://w.wiki/4Q7z

[20] Or e.g., audio transcriptions (Gómez et al., 2023)

[21] Though traditional fine-tuning and instruction tuning have important differences in construction, we do not distinguish between the two as we generally believe that the datasets can be converted between the two formats as necessary.

discuss in Section 3.3.

There are many possible transformations of Wikimedia data into post-training tasks. We represent this diversity by selecting a sample of tasks and example datasets for each one. We further split the tasks into three categories (classification, recommendation, and text generation) to provide some basic structure.

Classification

- **Stance detection**: a core part of Wikimedia is reaching consensus through discussions. Kaffee et al. (2023) studied article deletion discussions in English, German, and Turkish and fine-tuned a language model to predict what policies an editor will cite and their stance regarding deletion based on their comments.
- **Vandalism detection**: patrolling recent edits for vandalism that should be removed is a core task in maintaining Wikipedia's reliability. Trokhymovych et al. (2023) fine-tuned language models in 47 languages to predict whether an edit will be reverted.
- **Citation-needed**: the Verifiability policy requires that many statements on Wikipedia be supported with a citation to a reliable source. Redi et al. (2019) trained language models to predict whether a given sentence needs a citation in English, French, and Italian.
- **Readability**: accessibility of content to readers is important on Wikipedia but can be difficult to measure. Trokhymovych et al. (2024) fine-tuned language models in 14 languages to rank content by its readability.
- **NPOV detection**: a core content policy for Wikipedia is that text must adhere to a neutral point of view. Wong et al. (2021) built a dataset from English Wikipedia of edits that violated various policies for training classifiers to detect NPOV violations and other related content reliability issues.

Recommendation

- **Citation recommendation**: finding a source to verify a claim on Wikipedia can be a difficult task for editors. Petroni et al. (2023) trained a retrieval and ranking model to find citations for statements on English Wikipedia.
- **Entity linking**: a key part of Wikipedia is its network of links that connect content and allow readers to go down rabbit holes. Gerlach et al. (2021) trained a model across six language editions of Wikipedia for recommending links to be added to text spans within articles. There are also multimodal variants of this task such as visual entity linking.[22]
- **Grammatical error correction**: Fixing small spelling mistakes or grammatical errors is a common editing task on Wikipedia. Grundkiewicz and Junczys-Dowmunt (2014) used English Wikipedia revision histories to identify these copy-edits in order to train language models for grammatical error correction.

Text Generation

- **Article descriptions**: all articles can be associated with a short phrase that helps readers disambiguate between similarly-named pages. Sakota et al. (2023) fine-tuned a language model to generate these article descriptions based on the first paragraph of Wikipedia articles and descriptions in other languages for 25 different language editions.
- **Edit Summaries**: each edit on Wikipedia should be accompanied by a short summary that explains what the edit did and why (similar to a code commit message). Šakota et al. (2024) fine-tuned a language model to generate these edit summaries based on extracted diffs of a given edit on English Wikipedia.
- **Between Structured and Unstructured**: Facts can be stored in many different ways on the Wikimedia projects ranging from unstructured text in Wikipedia articles to semi-structured text in infoboxes or tables to the structured statements of Wikidata. Likewise, external sources of content to be incorporated can also be found in a variety of formats. Models for converting between these formats help editors in adding content and making it more accessible. For example, Chen et al. (2021) trained language models to produce long-form text from tabular data compiled from English Wikipedia while Luggen et al. (2021) trained language models to recommend Wikidata properties based on Wikipedia text.

[22] https://huggingface.co/datasets/aiintelligentsystems/vel_commons_wikidata

- **Natural language to SPARQL**: Wikidata contains a wealth of information but querying that content via what's known as SPARQL can be difficult. Liu et al. (2024) compile a dataset of English-language requests for SPARQL queries and the resulting query to evaluate LLM-based approaches for generating SPARQL queries.

- **Simplification**: Entire language editions (Simple English) and namespaces (Txikipedia) have been created on Wikipedia to provide simpler-language versions of content. Sun et al. (2021) use this correspondence between English and Simple English Wikipedia to build a dataset of article leads and their simpler equivalents to train language models to simplify text.

- **Summarization**: summarization has many potential use-cases on the wikis from helping editors understand long discussions on-wiki such as RFCs (Im et al., 2018) or the information across multiple external sources. Ghalandari et al. (2020) compile a dataset from the English Wikipedia Current Events portal of multi-document summaries.

- **Machine translation**: translation plays an increasing role in assisting in content creation on Wikipedia and making the 300+ language editions accessible to all readers.[23] There are both datasets of published translations[24] for all languages and datasets of aligned text across languages like Schwenk et al. (2021).

- **Article writing**: Wikipedia is a tertiary source whose content is a consolidation of other sources as reflected in the citations. Shao et al. (2024) prompted LLMs to write English Wikipedia articles by gathering and summarizing sources related to a given topic.

This catalog of tasks demonstrates the diversity of NLP post-training tasks that already exist that could be beneficial to Wikimedia editors—ranging from simple binary classification to natural language generation, from short-form texts to long-form articles, and from models that must reflect Wikimedia-specific policies to more generic tasks like translation or summarization. This catalog also reveals large language gaps: despite the over 300 language editions of Wikipedia, most example datasets leverage English Wikipedia alone. This sometimes seems to be purely about precedent and familiarity—e.g., edit summaries exist in all language editions so expanding a dataset of them is largely trivial, but many language modeling tasks start with English. Other times, this stems from structural challenges on the Wikimedia projects that would take more extensive work to overcome—e.g., many language editions use various content reliability templates to flag NPOV issues but the templates and norms around them can vary language-to-language, making it difficult to scale datasets to more languages (Johnson and Lescak, 2022).

We focused here on language as the most salient facet of these datasets, but as identified in Section 2.1, open-source licensing and compactness are also important to assessing the value of models to the Wikimedia projects. This is especially true in models that touch on privacy-sensitive areas such as search queries (e.g., natural language to SPARQL) where depending on 3rd-party models would open up individuals to surveillance. The NLP community has made important strides in both of these spaces in recent years but cataloging which tasks are lacking in good open-source models would be beneficial for considering future research.

3.3 Evaluation: from tasks to benchmarks

Paraphrasing Bowman and Dahl (2021), benchmarks for natural-language understanding are datasets that have the following characteristics: 1) they are representative of the task in question, 2) their data are accurate and unambiguous, 3) they can accurately rank models, and, 4) they disincentivize biased or harmful models. While the existence of many Wikimedia-focused tasks in Section 3.2 is heartening, few of these meet the standards of benchmarks. Trivially, many datasets that are derived from Wikimedia data can be found in the pre-training data used by many LLMs and thus are not accurate evaluations of these model's ability to generalize to new examples. This lack of Wikimedia benchmarks means that editors do not have easy or effective means of evaluating models (especially LLMs) for their usefulness to Wikimedia. Additionally, many LLMs are not open-source or are too large to be trained (or even fine-tuned in some cases) by Wikimedia developers. Developing core Wikimedia benchmarks could provide an important means of nudging NLP practitioners to

[23] https://www.mediawiki.org/wiki/MinT

[24] https://www.mediawiki.org/wiki/Content_translation/Published_translations

develop models that are more beneficial out-of-the-box for the Wikimedia projects.

When it comes to evaluation of language models, it is less clear that the needs of the Wikimedia projects and NLP practitioners are currently well-aligned. Instruction-tuned LLMs are generally designed for a few purposes as demonstrated by the benchmarks that the model developers choose to test their models on. For example, the Llama 3 models (Dubey et al., 2024) are described as being benchmarked in eight top-level categories: (1) commonsense reasoning; (2) knowledge; (3) reading comprehension; (4) math, reasoning, and problem solving; (5) long context; (6) code; (7) adversarial evaluations; and (8) aggregate evaluations. Most of these categories are relevant for chat-bots to better answer questions but only incidentally tell us how these models might handle tasks related to applying Wikimedia content policies when editing or performing content moderation tasks.

The core content policies of Wikipedia that guide many of the post-training tasks in Section 3.2 have clear corollaries with the intentions of LLMs developers. Neutral Point-of-View aligns well with training models that are not biased or harmful.[25] No Original Research aligns well with the goal of reducing hallucinations. Verifiability is perhaps less clear as a stated goal of many LLM models—i.e. the ability to cite sources for answers. However, we are witnessing a shift towards attribution of sources in LLM-backed products via retrieval-augmented generation, and Verifiability has nice overlap with chain-of-thought approaches (Khalifa et al., 2024) that have been demonstrated to improve model performance in many reasoning tasks (Wei et al., 2022). In all, LLMs that are more useful for Wikimedia-related tasks should also be more useful for many tasks outside of Wikimedia. In Table 2, we focus on these core content policies and examine the state of benchmarks for following these policies when creating content[26] as well as evaluating existing content for whether it adheres to the policy.

Table 2 shows that there are existing benchmarks for evaluating the Verifiability and No Original Research policies. While citation-needed was developed with Wikipedia in mind, ALCE, FEVER, and WildHallucinations[27] were developed with Wikipedia content but are oriented towards standard NLP tasks such as Q&A or textual entailment. Work is still required to raise the quality of these benchmarks to ensure their freshness akin to FreshWiki's approach of only extracting content that was extensively edited after a given knowledge cut-off. And as with post-training tasks, these benchmarks are still heavily English-focused and do not cover the many other languages of Wikipedia.

Neutral Point-of-View has more mixed coverage. The NPOV policy contains multiple facets, of which two core components are the issue of biased language and the issue of biased coverage (due weight). Benchmarks do currently exist for the biased language facet based on editor activity from English Wikipedia. Biased coverage is harder to assess. WikiContradict(Hou et al., 2024) assesses a particular case where two reliable sources present contradictory information but there is a need for benchmarks that could e.g., determine whether content produced via multi-document summarization gives appropriate weight to different claims based on the level of their support across the documents. A core challenge here is not giving undue weight to fringe theories that may be mentioned by sources but are not well-supported.

We focused in this paper on the core content policies as an important first step for capturing facets important to the Wikimedia community and the basic existence of reasonable benchmarks in these areas. Moving forward, this framework could be extended to include more Wikimedia policies and guidelines and explore the fourth criteria asserted by Bowman and Dahl (2021) of disincentivizing bias through these benchmarks.

We recommend a few additional policies to consider for extending this framework.[28] The policy on Copyright Violations[29] touches on the importance of summarizing sources instead of copying them. Notability[30] is a major guideline for determining whether an article should exist or not for a topic. Benchmarks might focus on evaluating sources for

[25] Longpre et al. (2024) showed that including Wikipedia in pre-training data greatly decreases model toxicity.

[26] Editing existing content is a different task but we also consider it under content creation.

[27] WildHallucinations also covers content outside of Wikipedia but a related benchmark FActScore (Min et al., 2023) is extracted purely from English Wikipedia.

[28] We have linked to English Wikipedia policies and guidelines here but other language editions have developed their own policies and guidelines (Hwang and Shaw, 2022).

[29] https://en.wikipedia.org/wiki/Wikipedia:Copyright_violations

[30] https://en.wikipedia.org/wiki/Wikipedia:Notability

Content Policy	Context	Benchmark
Verifiability	Creating content: given a topic to generate content, does the model appropriate cite its sources?	FreshWiki for English, which uses the citation quality metrics from ALCE (Gao et al., 2023)
	Evaluating content: given a statement, does it require a citation?	Citation Needed (Redi et al., 2019) for English, French, and Italian
No Original Research	Creating content: given a topic to generate content, does the model hallucinate any claims?	WildHallucinations (Zhao et al., 2024) which covers English Wikipedia and English non-Wikipedia topics.
	Evaluating content: given a claim and source, is the claim supported?	FEVER (Thorne et al., 2018) for English
Neutral Point-of-View (biased language)	Creating content: given a topic or sentence, can the model remove biased language?	(Pryzant et al., 2020) and then (Ashkinaze et al., 2024) for a more recent evaluation of LLMs and English Wikipedia.
	Evaluating content: given a sentence, can the model identify if it uses biased language?	
Neutral Point-of-View (due weight)	Creating content: given a topic, can a model fairly represent all reliable sources?	WikiContradict (Hou et al., 2024) is the closest analog, which evaluates how well models handle the summarization of contradictory information.
	Evaluating content: given an article, can a model determine if the content is fairly represented?	None

Table 2: Benchmark tasks for Wikipedia's core content policies.

whether there is significant coverage of a given topic. There are also many style-related guidelines such as the Manual of Style[31] which touch on how to structure and format content such as capitalization, abbreviations, and mixing of dialects. One gap that is unlikely to be filled is assessing source reliability (a core component of all three core content policies). English Wikipedia, for example, tracks sources whose reliability is often questioned in a list known as Perennial Sources[32]. These assessments can change as sources themselves evolve and reflect consensus from long discussions about these sources. It is both hard to imagine LLMs making these assessments (except perhaps as a support for summarizing discussions) and undesirable to leave this complex sense-making to AI.

For disincentivizing bias through benchmarks, there is a long history of research on biases on the Wikimedia projects to pull from (Redi et al., 2020). One key step is expanding benchmarks to cover more languages but researchers might also develop benchmarks that only use articles that comprise a more balanced representation of the world. Datasets like Merity et al. (2016) that filter articles to only those deemed to be of the highest quality by Wikimedians would be another way to ensure that benchmark data is maximally likely to e.g., fully meet the expectations of the NPOV policy.

4 Conclusion

We present a summary of how Wikimedia data is curated to support the different stages of model training with a focus on NLP. At each stage, we highlight data that could be converted into more useful forms for training language models and identify ways in which these models could be more useful for Wikimedia editors. This shows that while Wikimedia content has been hugely influential and important to the development of AI as a source of language data, the field still has gaps in developing benchmarks and models that reflect the needs of Wikimedia editors. We hope that the opportunities that we highlight in this space encourage a more mutualistic relationship between NLP and the Wikimedia communities.

[31] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

[32] https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources/Perennial_sources

References

Joshua Ashkinaze, Ruijia Guan, Laura Kurek, Eytan Adar, Ceren Budak, and Eric Gilbert. 2024. Seeing like an ai: How llms apply (and misapply) wikipedia neutrality norms. *arXiv preprint arXiv:2407.04183*.

Samuel Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Mingda Chen, Sam Wiseman, and Kevin Gimpel. 2021. Wikitablet: A large-scale data-to-text dataset for generating wikipedia article sections. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 193–209.

Creative Commons. 2023. Making AI Work for Creators and the Commons - Creative Commons — creativecommons.org. https://creativecommons.org/2023/10/07/making-ai-work-for-creators-and-the-commons/.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Wikimedia Foundation. 2024. Artificial intelligence/Bellagio 2024. https://meta.wikimedia.org/w/index.php?title=Artificial_intelligence/Bellagio_2024&oldid=26436782.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Martin Gerlach, Marshall Miller, Rita Ho, Kosta Harlan, and Djellel Difallah. 2021. Multilingual entity linking system for wikipedia with a machine-in-the-loop approach. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3818–3827.

Demian Gholipour Ghalandari, Chris Hokamp, John Glover, Georgiana Ifrim, et al. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308.

Rafael Mosquera Gómez, Julián Eusse, Juan Ciro, Daniel Galvez, Ryan Hileman, Kurt Bollacker, and David Kanter. 2023. Speech wikimedia: A 77 language multilingual speech dataset. *arXiv preprint arXiv:2308.15710*.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2014. The wiked error corpus: A corpus of corrective wikipedia edits and its application to grammatical error correction. In *Advances in Natural Language Processing: 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings 9*, pages 478–490. Springer.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40b: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *arXiv preprint arXiv:2406.13805*.

Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. 2018. Wikiconv: A corpus of the complete conversational history of a large online collaborative community. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2818–2823.

Sohyeon Hwang and Aaron Shaw. 2022. Rules and rulemaking in the five largest wikipedias. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 347–357.

Jane Im, Amy X Zhang, Christopher J Schilling, and David Karger. 2018. Deliberation and resolution on wikipedia: A case study of requests for comments. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–24.

Isaac Johnson and Emily Lescak. 2022. Considerations for multilingual wikipedia research. *arXiv preprint arXiv:2204.02483*.

Lucie-Aimée Kaffee, Arnav Arora, and Isabelle Augenstein. 2023. Why should this article be deleted? transparent stance detection in multilingual wikipedia editor discussions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5891–5909.

Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. 2024. Source-aware training enables knowledge attribution in language models. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Elisa Kreiss, Fei Fang, Noah Goodman, and Christopher Potts. 2022. Concadia: Towards image-based text generation with a purpose. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4667–4684.

Shicheng Liu, Sina J Semnani, Harold Triedman, Jialiang Xu, Isaac Dan Zhao, and Monica S Lam. 2024. Spinach: Sparql-based information navigation for challenging real-world questions. *arXiv preprint arXiv:2407.11417*.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276.

Michael Luggen, Julien Audiffren, Djellel Difallah, and Philippe Cudré-Mauroux. 2021. Wiki2prop: A multimodal approach for predicting wikidata properties from wikipedia. In *Proceedings of the Web Conference 2021*, pages 2357–2366.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Fabio Petroni, Samuel Broscheit, Aleksandra Piktus, Patrick Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu, Maria Lomeli, Timo Schick, Michele Bevilacqua, et al. 2023. Improving wikipedia verifiability with ai. *Nature Machine Intelligence*, 5(10):1142–1148.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Hongjing Qian, Yutao Zhu, Zhicheng Dou, Haoqi Gu, Xinyu Zhang, Zheng Liu, Ruofei Lai, Zhao Cao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Webbrain: Learning to generate factually correct articles for queries by grounding on large web corpus. *arXiv preprint arXiv:2304.04358*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Miriam Redi, Besnik Fetahu, Jonathan Morgan, and Dario Taraborelli. 2019. Citation needed: A taxonomy and algorithmic assessment of wikipedia's verifiability. In *The World Wide Web Conference*, pages 1567–1578.

Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. 2020. A taxonomy of knowledge gaps for wikimedia projects (second draft). *arXiv preprint arXiv:2008.12314*.

Marija Šakota, Isaac Johnson, Guosheng Feng, and Robert West. 2024. Edisum: Summarizing and explaining wikipedia edits at scale. *arXiv preprint arXiv:2404.03428*.

Marija Sakota, Maxime Peyrard, and Robert West. 2023. Descartes: generating short descriptions of wikipedia articles. In *Proceedings of the ACM Web Conference 2023*, pages 1446–1456.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *arXiv preprint arXiv:2103.01913*.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Mykola Trokhymovych, Muniza Aslam, Ai-Jou Chou, Ricardo Baeza-Yates, and Diego Saez-Trumper. 2023. Fair multilingual vandalism detection system for wikipedia. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4981–4990.

Mykola Trokhymovych, Indira Sen, and Martin Gerlach. 2024. An open multilingual system for scoring readability of wikipedia. *arXiv preprint arXiv:2406.01835*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

KayYen Wong, Miriam Redi, and Diego Saez-Trumper. 2021. Wiki-reliability: A large scale dataset for content reliability on wikipedia. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2437–2442.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, et al. 2024. Wildhallucinations: Evaluating long-form factuality in llms with real-world entity queries. *arXiv preprint arXiv:2407.17468*.