Omnipedia: using the manual of style to automate article review

Samuel J. Klein
Berkman Klein Center
sklein@cyber.harvard.edu

Alex Andonian
MIT
andonian@mit.edu

Sayer Tindall
Block Science, Inc
sayer@block.science

Michael Zargham
Block Science, Inc
zargham@block.science

Abstract

Wikipedia was originally an inclusive place to curate incomplete knowledge. Over time, detailed style guides have raised the bar for contribution and formalizing expectations for good articles while excluding or removing less complete work. We present OMNIPEDIA, a language model [LM] pipeline for transforming a style guide into a list of context-sensitive requirements for article review. These requirements are used as input to an evaluation framework that rates and annotates each sentence with potential improvements. We demonstrate a reading interface that shows these annotations and highlights the text in different colors indicating how much work may be needed. The result can support high-throughput drafts of material that may need extensive work, visually distinguishing them from more polished outputs, just as Wikipedia began as a draft space for the more polished articles of Nupedia.

1 Background

Due to its editable nature and reliance on soft security,(Rasmusson and Jansson, 1996) Wikipedia has always included material that is incomplete, obscure, or unverified. As the English Wikipedia matured, and readers began to expect a level of uniform quality, a Manual of Style [MoS] was developed, and contributions that did not conform to it were shortened, deleted, or moved to a temporary draft space.(Elmimouni et al., 2022) New page contributions have also grown steadily, while the availability of community members to review them has started to decline.

To address both of these issues, we developed a framework for automating review of articles, grounded in the style guidelines, and annotating the articles with suggestions for improvement. Formal specifications are a robust way to develop and maintain complex systems as they provide a clear, precise framework for understanding what the system should do and how it should behave. By defining the system's key properties and behaviors mathematically, these specifications help ensure all parts of the system work together as intended, even as the system grows or changes. This approach helps catch potential problems early in development, making it easier to manage and scale systems without affecting their performance or reliability.

Meta's **Llama 3** is an open-source large language model released on April 18, 2024, in two sizes – 8B and 70B parameters.[1] A 400B parameter model is expected to be released on July 23.[2] Improvements over Llama 2 include:

1. Improved nuance, translation, and dialogue.
2. New capabilities in reasoning and code generation.
3. A larger training dataset: over 15 trillion tokens, 7x larger than Llama 2, with 4x more code data. 5% of the data comes from 30 non-English languages.
4. An 8K context length, double that of Llama 2.

Figure 1: A [[Llama 3]] overview generated by STORM, evaluated with Omnipedia. **Key**: Red - conflicts with a style guideline, Yellow - no clear conflict, Green - no conflict and edited by a trusted editor

In Wikipedia, formal specifications and evaluation mechanisms can improve the process of writing and editing articles. By setting clear rules for content quality, relevance, and verifiability, these tools can help standardize how contributions are assessed, across its diverse contributor base. Automating evaluation of text against requirements allows for continuous guidance to authors, and can make part of the editorial process more transparent and consistent. Additionally, these tools can help prevent disputes by linking feedback to requirements directly in an overview of the reviewed text, making the feedback process more efficient and scalable particularly for new contributors who can be daunted by multi-step workflows.

1.1 Related Work

Wikipedia's RecentChanges feed and other tools designed for fighting vandalism and patrolling new pages have used automated evaluation of edits to streamline the work of reviewers for over a decade. These range from regular expressions for catching excluded domains and topics, to the ORES family of machine-learning algorithms for scoring edits.(Halfaker and Geiger, 2020; Teblunthuis, 2021) Additional work has been done to make these scoring processes transparent and auditable.(Levonian et al., 2024)

Attempts to automate the review of articles and individual claims(Moas and Lopes, 2023; Bassani and Viviani, 2018; Smith et al., 2020) tend to capture limited context, and often have trouble applying the nuances of Wikipedia principles such as neutrality and proportionality.(Ashkinaze et al., 2024) These tools often define new metrics to evaluate. This is the first tool we know of that takes advantage of the detailed structured style guidelines that already exist on most wikis, maintained by their editors to speed and standardize their manual reviews.

2 Method

On English Wikipedia, the MoS has over 100 detailed guides,(Wikipedians, 2024b) each guide describing the categories of articles (by topic or format) that it applies to, rules that those articles should follow, and the sections within each article to which each rule applies.

The Omnipedia evaluator proceeds in three stages: a) producing requirements from a style guide page, b) evaluating a text against those requirements to generate ratings and recommended improvements for each sentence, and c) rendering a view of the text summarizing this evaluation, offering readers an immediate sense of how much revision remains to be done.

2.1 Requirements generation

GPT-4o was used the LM for most steps of the process, prompting it to be run in a number of roles. A local instance of STORM with GPT-4o (Shao et al., 2024) was used to generate article text where we needed a text to evaluate. To produce requirements we identified section types commonly associated with different guidelines: *Lead section, Body sections, Infobox, References, External links, Categories*. We then prompted the LM to summarize the core values identified in the style guide as requirements, assign them a short name, and indicate which sections they apply to.

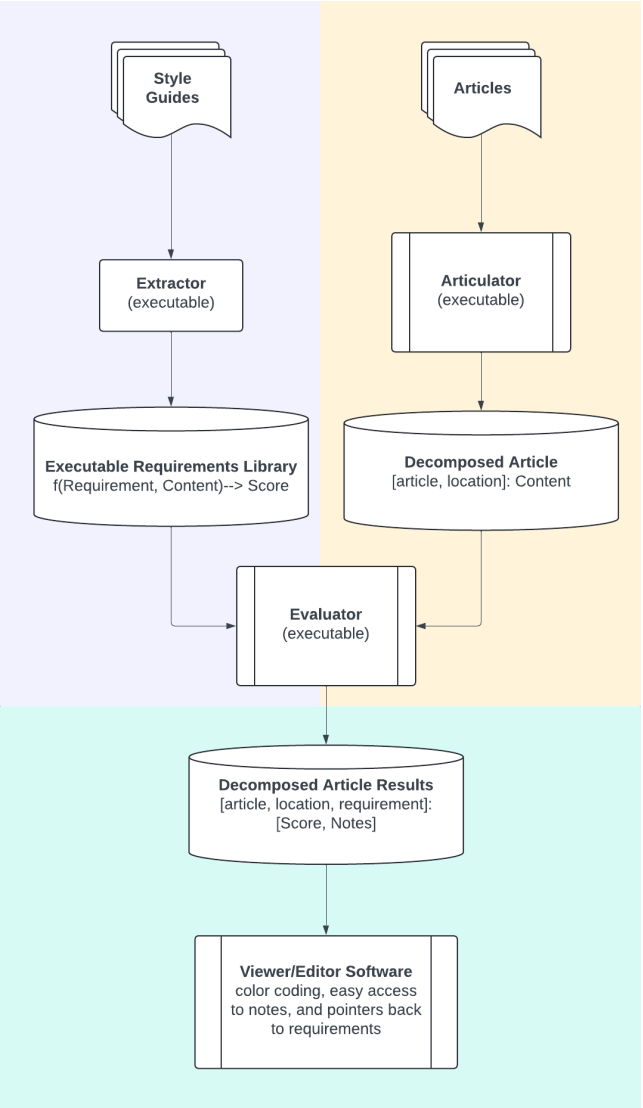

Figure 2: Workflow overview: style guides are separated into requirements and relevant sections. Articles are separated into sections and their sentences, and checked against the requirements.

2.2 Automated evaluation against requirements

Once requirements have been generated, articles submitted to Omnipedia are evaluated against them.

Articles are first chunked into sections, and the LM assigns each a section type from the list of types.[1] Then each section is chunked into sentences, which have the section type and their sequence order appended to them. The sentences are

[1] "Please assign a section type (eg Lead section or Body Sections) for the following section"

Requirement	Lead	Body	Infobox	References	Description
Title	1	1	0	0	Sentence case for titles and headings
Consistency	1	1	1	1	Consistent style within an article
References	1	1	0	1	All statements must have reliable sources
Lead summary	1	0	0	0	Lead must summarize the article concisely

Table 1: A subset of the requirements matrix associating criteria with sections

passed one at a time to a persistent LM session for both qualitative and quantitative evaluation. Each sentence is assigned a rating between 0 and 1 indicating the probability that it meets all guidelines, and a constructive evaluation with suggestions for improvement.[2]

2.3 Displaying quantitative and qualitative review

After generating the evaluation, the reviewed article is rendered for the reader. A color is assigned to each sentence indicating its overall compliance with the guidelines, based on a threshold set by the user. Sentences rated below the threshold are highlighted in red. Those above the threshold are green if they have been edited within the viewer, and otherwise are otherwise yellow. Sentences can be clicked on to see the qualitative evaluation and suggestions.

As an option, article text can be compared against existing text on Wikipedia, and highlight color removed for text that appears verbatim in an existing article.

3 Computational Experiments

3.1 Experimental Set up

We used o1-mini as the LLM in each step of the pipeline.(OpenAI, 2024) To test Omnipedia on a broad set of inputs, we turned to WikiCrow, a collection of genetics articles generated by PapersQA2.(Skarlinski et al., 2024) The WikiCrow article-set consists of 240 Wikipedia articles about human genes, and 240 parallel articles created from scratch about the same genes by PapersQA2.

We articulated the Wikipedia and WikiCrow articles in the dataset, using the markdown sources. They used different markdown languages, and different infobox and citation formats. We made a separate cleanup pass normalizing the custom WikiCrow citation format; for a larger scale project, this normalization step could be expanded.

Wikipedia has a small topical style guide for genes and proteins,(Wikipedians, 2024a) from which we extracted 38 requirements. The requirements extraction step also clustered them into seven categories: Format, Structure, Language, Infoboxes, Citations, Content, and Figures. All articles were evaluated against these requirements.

3.2 Results and Analysis

Overall, the WikiCrow and Wikipedia articles had similar levels of compliance with sentence-level requirements. When clustered by requirement category, the Wikipedia articles complied slightly better with format guidelines, and the WikiCrow articles adhered more uniformly to structure and language guidelines, but had some consistent and correlated deviance from content and citation guidelines.

We then carried out a user study with human editors, and found that the greatest differences between the two sets of articles were in section-level or article-level guidelines, around issues such as consistency, repetition, and diversity of sources.

Future updates to the pipeline will classify requirements as sentence-, section-, or article-level requirements, also supporting a feedback loop with the editors of style guides regarding which levels of abstraction a guide covers.

3.3 Hosting and Operationalization

Omnipedia v1 was run on all WikiCrow articles, and their Wikipedia counterparts, and the results can be browsed at `omnipedia.cc`.

Omnipedia v2 distinguishes between requirements at the sentence, section, and article level, and clusters evaluations by section. You can see a demonstration on the same articles at `omnipedia-client.pages.dev`.

Code is available on github at `github.com/wikius/omnipedia`.

[2] "Evaluate and provide suggestions according to the requirements in the wikikpedia style guide", and "Provide a score between zero and one representing the probability this sentence meets the Wikipedia guidelines, according to a Wikipedia moderator"

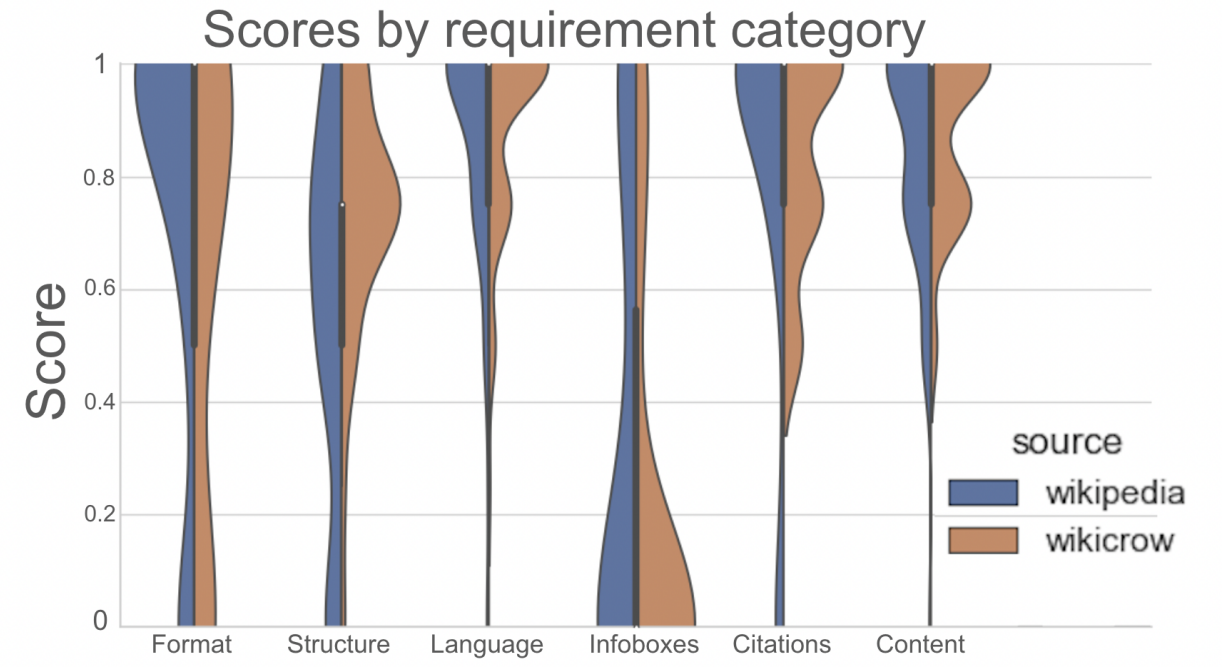

Figure 3: Compliance scores for WikiCrow and Wikipedia articles, grouped by category of requirement.

4 Discussion and future directions

To test this system, we set a threshold of 0.8 for compliance with style guidelines, based on feedback with human testers. We then tested this workflow on STORM-generated articles about new technical topics, such as the Llama 3 example in Figure 1. Fewer than 20% of sentences in the generated articles passed the threshold, mostly due to using non-neutral language and making unreferenced claims. Those that did pass were primarily found in sections discussing limitations and challenges.

While many sentences had obvious fixes, we found they were easy even for veteran editors to overlook. This felt like a grammar check: reminding the user of things they know well. In other places, while most of a paragraph had significant issues, this view allowed seeing at a glance which parts could be a kernel to keep or build a rewrite around.

Future work will include testing the pipeline with Ollama, integrating the viewer into the default MediaWiki editor, and letting viewers select which style guides they want to use.

4.1 Limitations

Using a frontier model such as GPT-4o is inefficient for the inner loops in this process. In a multi-agent setup, faster and smaller models can be used to evaluate a location against fixed requirements.

This demo focuses on sentences, but the evaluator was designed to care generally about locations in a document, which could overlap or exist at different scales (sections, sentences, clauses). Some common uses call for other scales: standard cleanup templates are at the section level, and inline citations or citation-needed templates can be at the clause level.

Articles in different topics use a variety of synonyms as headers for common section types. We improved the accuracy of section classification by manually adding a list of synonyms to the requirements stage, but this could be done more thoroughly.

Some commonly failed requirements, such as standards for inline citation format, occasionally indicated the LM failed to parse a correctly-written sentence. A layer of meta-evaluation could reduce this failure mode and suggest ways to improve prompts in the pipeline.

Currently this was done for English only, and for general style guides. Topic-specific guides have richer context dependence, and may have to cope with conflicting requirements.

The highlighting was intended to be a calm interface,(Case, 2015) but was still too intrusive to support easy reading. Something closer to the standard spellcheck underlines may be more appropriate for visualizing all but the most extreme concerns.

Doing evaluations exclusively per sentence led

to false positives for issues addressed elsewhere in the same section. Stylistic requirements such as citation formatting and rendering names in italics are overweighted, and reflect that WikiCrow implements a consistent but different visual style.

Having a uniform threshold that treated every requirement equally was fine for a first pass, but existing style guides offer some context about which requirements are most important, and relative weights could be made explicit. More important sentences like the lead and topic sentences of a section could also be evaluated more strictly. The system captures enough sequence context to be able to make those distinctions but currently that is not used.

Acknowledgements

This document has drawn on past work with Apolinário Passos, with inference support from Hugging Face, and benefited from correspondence with Yijia Shao. It would not have been possible without the work of the thousands of people who have contributed to the English Wikipedia Manual of Style.

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.

Elias Bassani and Marco Viviani. 2018. Feature analysis for assessing the quality of wikipedia articles through supervised classification.

Amber Case. 2015. *Calm Technology: Principles and Patterns for Non-Intrusive Design*. O'Reilly Media, Inc.

Houda Elmimouni, Andrea Forte, and Jonathan Morgan. 2022. Why people trust wikipedia articles: Credibility assessment strategies used by readers. In *Proceedings of the 18th International Symposium on Open Collaboration*, OpenSym '22, New York, NY, USA. Association for Computing Machinery.

Aaron Halfaker and R. Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia.

Zachary Levonian, Lauren Hagen, Lu Li, Jada Lilleboe, Solvejg Wastvedt, Aaron Halfaker, and Loren Terveen. 2024. Ores-inspect: A technology probe for machine learning audits on enwiki.

Pedro Miguel Moas and Carla Teixeira Lopes. 2023. Automatic quality assessment of wikipedia articles—a systematic literature review. *ACM Comput. Surv.*, 56(4).

OpenAI. 2024. Openai o1 mini 2024-09-12. [Online; accessed 15-September-2024].

Lars Rasmusson and Sverker Jansson. 1996. Simulated social control for secure internet commerce. In *Proceedings of the 1996 Workshop on New Security Paradigms*, NSPW '96, page 18–25, New York, NY, USA. Association for Computing Machinery.

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. 2024. Language agents achieve superhuman synthesis of scientific knowledge.

C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Nathan Teblunthuis. 2021. Measuring wikipedia article quality in one dimension by extending ores with ordinal regression. In *17th International Symposium on Open Collaboration*, OpenSym 2021. ACM.

Wikipedians. 2024a. Wikipedia style guide (gene and protein articles). [Online; accessed 15-September-2024].

Wikipedians. 2024b. Wikipedia:manual of style — Wikipedia, the free encyclopedia. [Online; accessed 6-September-2024].