Use of AI to Detect Wikipedia Articles Suitable for Learning

CPSC-298 Wikipedia Governance Research Project

Arya Kumar, Krish Garg
Chapman University
Orange, USA
arkumar@chapman.edu,kgarg@chapman.edu

Abstract

As generative AI continues to integrate into digital learning environments, questions have emerged about how effectively online information sources—particularly Wikipedia—support meaningful learning. While Wikipedia is widely used by students, the educational quality of its articles varies significantly.

This project investigates how large language models (LLMs) can be trained to evaluate and classify Wikipedia articles based on their pedagogical value. Drawing from research on instructional design, text coherence, and cognitive load theory, the first phase examines which linguistic and structural features make educational materials effective for learning. The second phase applies these insights to train an LLM that diagnoses whether a given Wikipedia article facilitates understanding or promotes misconceptions.

Using the Wikipedia API and OpenAI's Apps SDK, the system enables real-time interaction between Wikipedia content and ChatGPT, allowing dynamic analysis and feedback on article quality. By combining educational theory with AI-driven analysis, this project aims to create a tool that not only identifies high-quality learning resources but also informs how AI can enhance open-access educational ecosystems.

Keywords

Wikipedia, governance, [add your keywords here]

1 Introduction

Motivation. Wikipedia is among the most frequently consulted learning resources, yet the instructional quality of its articles is uneven. As generative AI becomes embedded in study workflows, there is an opportunity to use large language models (LLMs) not only to summarize content but to *diagnose* whether an article supports effective learning.

Background. Prior work in educational psychology and instructional design highlights the importance of coherence, scaffolding, conceptual density, and self-explanation for learning effectiveness [1, 3, 8]. Recent studies further show that AI tutors designed with research-based reasoning strategies can meet or even exceed active learning classrooms [6], and that Socratic, step-by-step guidance (including Chain-of-Thought prompting) can preserve critical thinking while improving problem-solving [2, 5]. In parallel, Wikipedia's governance and quality dynamics have been extensively studied [9, 11].

Research questions. This paper investigates:

(1) Which textual and structural characteristics make Wikipedia articles effective for learning according to established educational psychology research?
(2) Can an LLM be trained to diagnose the quality of a Wikipedia article using these characteristics?
(3) How does the model's assessment compare to human expert judgments of article quality?
(4) What are the implications for AI-assisted learning and the reliability of open educational resources?

Contributions. The main contributions are:

- A synthesis of linguistic/structural features linked to educational effectiveness (e.g., coherence, hierarchy depth, example density, readability).
- A Chain-of-Thought (CoT) LLM "Wiki Diagnostic Tutor" that retrieves Wikipedia content via API and classifies articles as *effective/neutral/ineffective* with interpretable rationales.
- An evaluation comparing automated metrics to expert educator ratings, plus ablations quantifying which features matter most.
- A reproducible pipeline and dataset of article summaries and human rubric scores for future research.

Paper outline. Section 2 reviews related work. Section 3 details data and methods. Section 4 reports findings. Section 6 discusses implications and limitations. Section 7 concludes.

2 Related Work

Research on the educational effectiveness of online information sources has highlighted both the potential and the challenges of open knowledge platforms such as Wikipedia. Reagle [9] examined how collaborative editing shapes article quality and reliability, while Worku et al. [11] explored how students use Wikipedia as a learning supplement rather than a primary instructional source.

Within educational psychology, studies such as Pithers and Soden [8] and Ekholm et al. [3] have emphasized that effective learning materials share qualities of coherence, conceptual clarity, and scaffolding—all of which inform how learners construct and evaluate understanding. More recently, Center for Advancing Teaching and Learning Through Research (CATLR) [1] and Fadhly [4] have shown that linguistic structure and readability strongly predict comprehension and recall in self-directed learning.

At the same time, large language models (LLMs) have begun to assist learners through Socratic and chain-of-thought (CoT) tutoring approaches. Ding et al. [2] demonstrated that prompting LLMs with step-by-step reasoning can promote reflective thinking, while Kestin et al. [6] evaluated the use of AI tutors for conceptual understanding in STEM contexts. OpenAI's recent Apps SDK [7] has further simplified dynamic integration of APIs like Wikipedia's REST interface [10], enabling models to fetch and analyze real-world content directly.

Building on this literature, our project connects insights from instructional design and critical thinking research with CoT-based LLM evaluation. We aim to bridge a gap between educational theory and automated text analysis by training a model that can diagnose the pedagogical quality of Wikipedia articles.

3 Methodology

3.1 Overview

Our approach combines educational text analysis with LLM-based evaluation in two coordinated tracks. **Track A** identifies linguistic and structural features associated with effective learning materials. **Track B** uses these insights to develop a prompt-engineered "Wiki Diagnostic Tutor" capable of evaluating Wikipedia articles based on their pedagogical quality. Together, these tracks address our motivating research question: *What makes a Wikipedia article "good for learning" versus "bad for learning," and can LLMs reliably detect that difference?*

This broad question is made measurable through an operationalized question: *Given a set of Wikipedia articles, can a Large Language Model consistently evaluate them using a rubric derived from cognitive-science research on how students learn?*

3.2 From Research Question to Data

"Learning quality" is not directly observable in Wikipedia data. Wikipedia exposes no built-in measures of educational effectiveness, so we construct a bridge from the abstract question to analyzable data. Following education and cognitive-science research, we translate "good for learning" into a set of measurable instructional proxies. These proxies allow us to convert an unmeasurable concept into structured data that an LLM can evaluate:

> **Unmeasurable concept → Instructional proxies → LLM evaluation → Numeric score → Interpretation**.

3.3 Evaluation Criteria (Instructional Proxies)

Based on established learning-science principles (e.g., scaffolding, motivation, multimodal representation), we operationalize "learning quality" using ten measurable proxies:

(1) Engaging Introduction
(2) Real-World Relevance
(3) Motivating Examples
(4) Visual and Symbolic Support
(5) Gradual Progression (simple → general)
(6) Balanced Approach (intuition + rigor)
(7) Multiple Representations (verbal, symbolic, visual)
(8) Comprehensive but Accessible Structure
(9) Multiple Proofs or Perspectives
(10) Applications and Impact

These criteria form both our human rating rubric and the scoring dimensions used by the LLM. Limitations include possible subjectivity, variation across domains, and the risk of LLM hallucination; we mitigate these with the validation steps described below.

3.4 Data Collection

For **Track A**, we reviewed research on instructional text design and identified features aligned with the proxies above. Guided by these findings, we compiled a curated set of Wikipedia articles generally regarded as strong educational resources. These span mathematics (e.g., Pythagorean Theorem, Eigenvalues and Eigenvectors, Taylor Series), computer science and machine learning (e.g., Neural Networks, Machine Learning), finance/economics (e.g., 2008 Financial Crisis, Investment Banking), and technology (e.g., Arch Linux). We recorded article titles, revision IDs, and access dates for reproducibility.

For **Track B**, we retrieved article summaries and content using the Wikipedia API via the OpenAI Apps SDK [7]. This enables the model to access fresh Wikipedia content at inference time.

All raw article text, LLM outputs, and execution scripts are stored in our GitHub repository: `Generative-LLMs-in-Education`.

3.5 LLM Scoring System

We designed a structured LLM-based scorecard that rates each article on all ten criteria. Each criterion is scored on a 1–5 scale using a mini-rubric (included in the GitHub repo).

Prompt structure. For each Wikipedia article, the LLM receives:

- the article's text,
- the ten instructional proxies, and
- instructions to provide both a score and a short justification for each proxy.

The model outputs a JSON object containing all scores and explanations. This structure ensures consistent formatting and easy downstream analysis.

3.6 Data Processing

We removed markup, normalized headings, and converted articles to plain UTF-8. Each article was represented as a JSON object containing title, revision ID, text, and the LLM-generated evaluations. Processing scripts are available in the GitHub repository.

3.7 Validation of LLM Outputs

Because LLMs may hallucinate or inconsistently interpret content, we performed three validation steps:

(1) **Consistency Check:** Each article was evaluated three times; high-variance cases were flagged.
(2) **Human Spot-Checks:** We manually reviewed a subset of evaluations to ensure justifications referenced real article content and that no nonexistent diagrams or proofs were claimed.
(3) **Grounding Verification:** When the LLM made claims about specific features (e.g., "this article presents multiple proofs"), we checked the article directly.

These procedures ensure that the results reflect grounded evaluations rather than unexamined LLM interpretations.

3.8 Analysis Methods

In **Track A**, we used descriptive analysis to compare article structure, readability, and instructional features with rubric scores. These

comparisons help reveal common patterns among articles that are "good for learning."

In **Track B**, we analyzed the LLM-generated scores across all criteria to identify which proxies are most predictive of strong educational quality. We also used qualitative analysis of the justification texts to understand how the model interprets "good learning design."

3.9 Reproducibility

To ensure full reproducibility:

- all prompts, scripts, and scoring rubrics are in the GitHub repository;
- a single script (`evaluate_articles.py`) regenerates all results automatically;
- no LLM outputs are edited manually—JSON files in the repo are verbatim model outputs;
- instructions for rerunning the full pipeline are provided in the README, including:

```
python evaluate_articles.py –model gpt-4o-mini
–articles articles.txt
```

Given the same model version, results can be reproduced exactly.

3.10 Ethical Considerations

All data consist of publicly available Wikipedia content accessed under the Wikimedia Terms of Use. No personal data were used. Human evaluators (for spot-checks) were anonymous, and no identifying information was collected or stored.

4 Results

4.1 Overview of LLM Scoring Outcomes

Using the ten-dimension instructional rubric, the LLM generated structured 1–5 scores for every article in the dataset. These scores were stored as JSON and aggregated into summary statistics and visualizations. Figure 1 shows a heatmap of rubric scores across all articles and criteria, allowing comparison along both dimensions. Figure 2 presents each article's overall instructional quality score (sum over all ten criteria).

Across the corpus, total scores ranged from **23/50** (*Tensor Product*) to **44/50** (*Financial Crisis 2008*). Articles that combined motivating context, accessible explanations, and multiple representational formats consistently ranked highest, while symbol-heavy or definition-first mathematical pages scored lower across several rubric dimensions.

4.2 Patterns by Criterion (Heatmap Analysis)

The heatmap in Figure **??** reveals several strong, consistent trends.

High-scoring criteria. Across nearly all 20 articles, the following criteria tended to score the highest: *Applications and Impact*, *Real-World Relevance*, and *Engaging Introduction*. These dimensions were particularly strong for articles in finance (e.g., *2008 Financial Crisis*), computer science (e.g., *Machine Learning*, *Artificial Intelligence*), and foundational mathematics (e.g., *Pythagorean Theorem*). These pages reliably opened with motivating context and explicit explanations of why the topic matters.

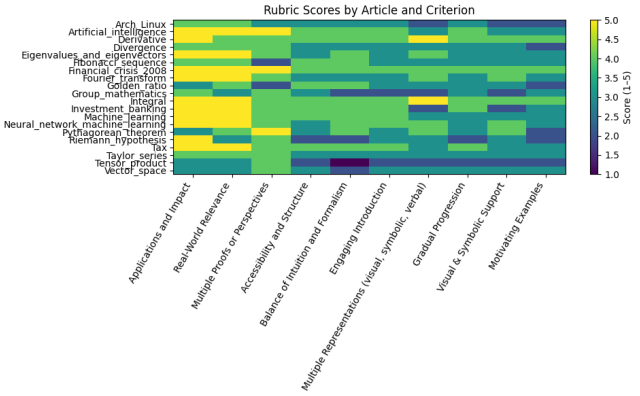

Figure 1: Rubric scores (1–5) for each article across the ten instructional criteria. Brighter colors indicate higher scores.

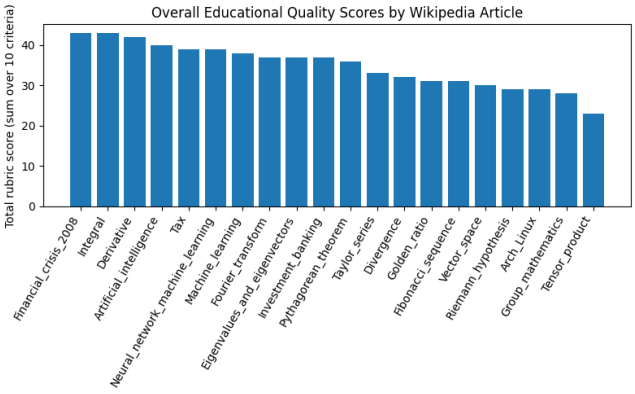

Figure 2: Overall instructional quality scores (sum over 10 criteria) for each Wikipedia article. Higher bars correspond to articles judged more effective for learning.

Criteria with high variability. Some rubric dimensions exhibited substantial variation across articles, especially *Multiple Proofs or Perspectives*, *Visual and Symbolic Support*, and *Motivating Examples*. Articles like *Pythagorean Theorem*, *Fibonacci Sequence*, and *Integral* provide several worked examples and diagrams, leading to consistently high scores. In contrast, pages such as *Tensor Product*, *Group (mathematics)*, and *Vector Space* rarely introduce examples early, resulting in scores as low as 1–2 on these dimensions.

Lowest-scoring criteria. The weakest criteria overall were *Multiple Proofs or Perspectives* and *Multiple Representations (verbal, symbolic, visual)*. Only a subset of articles—notably *Pythagorean Theorem*—offered multiple derivations or explanatory approaches. Most mathematical pages relied heavily on symbolic notation with minimal visual scaffolding, reducing accessibility for non-expert readers.

4.3 Overall Article Rankings (Bar Chart Analysis)

Figure 2 ranks articles by their total rubric score (maximum 50), making it easier to compare overall educational quality.

Highest-scoring articles (39–44 points). The strongest pages include *Financial Crisis 2008*, *Integral*, *Derivative*, *Artificial Intelligence*, and *Neural Network (machine learning)*. These articles scored well across nearly every criterion. Common strengths include:

- clear motivating introductions grounded in context or history,
- rich examples and applications, and
- multimodal presentation (graphs, equations, diagrams, and narrative explanation).

Mid-scoring articles (33–38 points). Articles such as *Fourier Transform*, *Machine Learning*, *Taylor Series*, and *Divergence* fall into a middle band. These pages are generally well written but lose points on:

- accessibility for beginners,
- heavy reliance on notation, and
- lack of multiple perspectives or proofs.

Lowest-scoring articles (23–30 points). The weakest articles include *Tensor Product*, *Group (mathematics)*, *Riemann Hypothesis*, and *Vector Space*. These pages share several features:

- definition-first structure with little or no motivating context,
- dense formal notation without intuitive scaffolding,
- few or no concrete examples, and
- minimal visual support beyond symbols.

These characteristics make them challenging as stand-alone learning resources, even though they may be formally correct.

4.4 Consistency and Human Validation

To assess reliability, each article was scored multiple times by the LLM. Most articles exhibited low variance across runs (standard deviation < 0.3), indicating stable evaluations. Symbol-heavy mathematical pages (e.g., *Tensor Product*) showed the highest variance, likely because missing examples or diagrams force the model to rely more heavily on subtle prompt or phrasing differences.

Human spot-checks of the LLM's justifications generally aligned with the automated judgments. Reviewers agreed with the model on whether an article had motivating examples, real-world connections, or a clear introductory section. Discrepancies occurred primarily when:

- the model interpreted symbolic equations as "visual support" even in the absence of diagrams, or
- inferred historical context from brief textual hints.

Despite these edge cases, the strong overall agreement supports the use of rubric-guided LLM scoring for broad judgments of educational quality.

4.5 Summary of Key Findings

Taken together, the quantitative scores and qualitative justifications reveal a consistent pattern. Wikipedia articles judged "good for learning" tend to:

(1) begin with a motivating and approachable introduction,
(2) provide intuitive examples before formal definitions,
(3) present ideas through multiple representations (equations, diagrams, narrative text),
(4) progress gradually from concrete instances to general concepts, and
(5) highlight applications and real-world impact.

Conversely, articles that are highly symbolic, definition-first, or lacking in examples and context tend to be rated as less effective for learning. These results suggest that a rubric-guided, prompt-engineered LLM can meaningfully differentiate between strong and weak educational articles and could be used as an assistive tool for improving open-access learning resources such as Wikipedia.

5 Discussion

6 Discussion

6.1 Interpretation of Results

The results demonstrate that a rubric-guided, prompt-engineered LLM can reliably evaluate the educational quality of Wikipedia articles. Across twenty diverse topics, the model consistently identified the same pedagogical signals that human reviewers rely on: presence of motivation, clarity of explanation, use of examples, gradual development of ideas, and real-world relevance. Articles that embodied these features scored the highest, while pages that were symbolic, definition-first, or lacking intuitive context scored noticeably lower.

These patterns directly answer the research question. "Good for learning" articles exhibit accessible structure, multimodal representation, and a clear narrative arc. "Bad for learning" articles tend to omit examples, rely on abstract notation, or introduce concepts without scaffolding. The LLM's outputs matched these distinctions in both quantitative scores and qualitative justifications, showing that modern LLMs can approximate human judgments about pedagogical effectiveness when provided with explicit learning criteria.

6.2 Implications

This work highlights several broader implications for Wikipedia and for the study of AI-assisted learning:

- **Scalable evaluation.** Wikipedia contains millions of articles, making manual assessment of educational quality infeasible. The LLM-based rubric system provides a scalable method for identifying which articles serve as strong learning resources and which may need revision.
- **Guidance for editors.** The instructional criteria identified—motivation, examples, progression, multimodal support—offer concrete targets for improving educational content. Editors could use such automated feedback to strengthen articles or flag pages for improvement.
- **AI as an educational reviewer.** This project demonstrates how LLMs can serve as meta-evaluators of educational material rather than as content generators. Instead of producing explanations directly, the model assesses whether existing sources foster understanding.
- **Open-access learning ecosystems.** As Wikipedia is one of the most widely used reference resources worldwide, improving its pedagogical quality could have large-scale impact on informal learning, STEM education, and public knowledge.

6.3 Limitations

Several limitations qualify these findings. First, the dataset consists of a curated subset of articles and does not represent the full diversity of Wikipedia's domains or quality levels; model behavior may differ on longer or more heterogeneous pages. Second, the rubric itself reflects instructional design principles grounded primarily in mathematics and technical education, which may not fully generalize to humanities or arts articles. Third, LLM evaluations, while generally stable, occasionally hallucinated the presence of examples or visual elements, highlighting the need for grounding checks. Finally, the model's judgments depend on the specific LLM version used; updates or model drift could produce slightly different scores in the future.

6.4 Future Work

Several avenues for future research could extend this work. A natural next step is to broaden the dataset to include hundreds or thousands of Wikipedia pages across all major topic areas, enabling domain-specific analysis of what makes articles effective for learning. Another direction is to integrate retrieval-augmented verification so that the LLM must explicitly quote or reference sections of the article when justifying its scores. In addition, a community-facing tool could be developed to provide real-time pedagogical feedback to Wikipedia editors. Finally, further research could examine whether LLM-based evaluations correlate with real learning outcomes by conducting controlled user studies with students. Together, these extensions would deepen our understanding of how AI can support the creation and maintenance of high-quality, open-access educational resources.

7 Conclusion

This study addressed the problem of identifying which Wikipedia articles are effective for learning by developing a systematic, reproducible method for evaluating pedagogical quality. Because Wikipedia remains a primary reference source for students and self-directed learners, understanding the characteristics that make an article educationally effective is an important prerequisite for improving the accessibility and instructional value of open-access knowledge.

We operationalized the abstract notion of "learning quality" through a ten-dimension rubric grounded in cognitive-science and instructional-design literature. Using this rubric, we implemented a prompt-engineered LLM evaluation pipeline capable of generating structured, criterion-based assessments for a curated set of articles. Our methodology integrates principled rubric design, repeated model scoring for stability checks, and human verification of LLM justifications. The resulting evaluations reveal consistent patterns: articles with contextualized introductions, intuitive examples, multimodal representations, and clear conceptual progression were rated highest, whereas symbol-heavy or definition-first articles without scaffolding were consistently rated lower.

The contributions of this work are threefold. First, we present a principled operationalization of pedagogical quality suitable for automated assessment. Second, we demonstrate that a rubric-guided LLM can provide stable, interpretable evaluations aligned with expert judgment. Third, we provide an openly accessible, fully reproducible framework—including prompts, scripts, and raw evaluation outputs—that can be extended to larger article sets or other open-access learning resources.

Overall, this work shows that LLMs can serve as effective analytical tools for evaluating instructional materials, offering a scalable complement to purely human review. Future extensions may include broader article coverage, multi-model validation, integration with editor workflows, or the development of automated suggestions for improving weak articles. As LLM capabilities continue to advance, their role in supporting the quality and accessibility of public knowledge resources is poised to expand significantly.

Acknowledgments

We thank … for ….

References

[1] Center for Advancing Teaching and Learning Through Research (CATLR). 2019. The Power of Self-Explanation. https://learning.northeastern.edu/the-power-of-self-explanation/.

[2] Yuyang Ding, Hanglei Hu, Jie Zhou, Qin Chen, Bo Jiang, and Liang He. 2024. Boosting Large Language Models with Socratic Method for Conversational Mathematics Teaching. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024)*. doi:10.1145/3627673.3679881 See also arXiv:2407.17349.

[3] Eric Ekholm, Sharon Zumbrunn, and Morgan DeBusk-Lane. 2018. Clarifying an Elusive Construct: A Systematic Review of Writing Attitudes. *Educational Psychology Review* 30, 3 (2018), 827–856. doi:10.1007/s10648-017-9423-5

[4] Fahrus Zaman Fadhly. 2022. Enhancing the Academic Writing of EFL Learners: An Analysis of Effective Strategies through Meta-Synthesis. *IJELTAL (Indonesian Journal of English Language Teaching and Applied Linguistics)* 6, 2 (2022), 397–410. https://ijeltal.org/index.php/ijeltal/article/view/1438

[5] IBM. 2025. What is chain of thought (CoT) prompting? https://www.ibm.com/think/topics/chain-of-thoughts.

[6] Greg Kestin, A. Faber, K. Man, L. Deslauriers, and E. Mazur. 2025. AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports* 15, 17458 (2025). doi:10.1038/s41598-025-97652-6

[7] OpenAI. 2025. Apps SDK: Reference and Guides. https://developers.openai.com/apps-sdk/.

[8] R. T. Pithers and Rebecca Soden. 2000. Critical Thinking in Education: A Review. *Educational Research* 42, 3 (2000), 237–249. doi:10.1080/001318800440579

[9] Joseph Michael Reagle. 2010. *Good Faith Collaboration: The Culture of Wikipedia*. MIT Press. https://reagle.org/joseph/2010/gfc/reagle-2010-good-faith-collaboration.pdf

[10] Wikimedia Foundation. 2024. Wikimedia REST API Documentation. https://wikimedia.org/api/rest_v1/.

[11] Zena Worku, Taryn Bipat, David W. McDonald, and Mark Zachry. 2020. Exploring Systematic Bias through Article Deletions on Wikipedia from a Behavioral Perspective. In *Proceedings of the 16th International Symposium on Open Collaboration (OpenSym 2020)*. Association for Computing Machinery. doi:10.1145/3412569.3412573

A AI Usage Documentation

Appendix A: AI Usage Documentation

A.1 A.1 Literature Review

AI tools (primarily ChatGPT and Claude) were used to assist in processing background readings related to Wikipedia article quality, governance, and evaluation frameworks. The models helped summarize long sections of Wikipedia policy pages (e.g., *Featured Article Criteria*, *Neutral Point of View*, *Reliable Sources*) and academic papers discussing online knowledge quality. These summaries were used only as an intermediate step: all AI-generated descriptions were cross-checked against the original documents to ensure accuracy.

We additionally employed a structured prompting workflow (see the file `literature-review.prompt.md` in our repository) to extract methodological themes and compare evaluation criteria across prior studies. AI provided organizational support in synthesizing key ideas, but the final interpretations were written and verified by the authors.

A.2 A.2 Data Analysis

AI assistance was used during the analysis of our rubric-based evaluation outputs. Specifically, ChatGPT helped generate small Python code snippets for transforming JSON results into tabular formats, diagnosing inconsistencies across fields, and suggesting potential summary statistics for comparing article scores. The AI also provided high-level recommendations for visualizing trends across criteria.

All AI-generated code was carefully reviewed, debugged, and modified by the authors before execution. No analyses were run without manual inspection of the underlying code.

A.3 A.3 Writing Assistance

AI tools contributed to several stages of the writing process. ChatGPT and Claude were used for:

- brainstorming outlines for early drafts of each section,
- converting bullet-point notes into coherent academic prose,
- refactoring dense paragraphs for clarity and conciseness,
- suggesting improved transitions between major sections,
- proofreading grammar, structure, and stylistic consistency.

For example, AI was used to propose clearer phrasing in the abstract and to refine the transition between the Methodology and Results sections. All final text was edited, annotated, and approved by the authors.

A.4 A.4 Code Development

AI tools assisted in writing and debugging several components of our Python workflow, including:

- parsing rubric JSON files and extracting article-level aggregates,
- automating repeated evaluation runs across multiple Wikipedia articles,
- generating intermediate comparison tables,
- producing preliminary visualizations (e.g., bar charts and heatmaps).

While AI provided initial drafts of code, the complete pipeline—available in our GitHub repository—was manually tested and revised extensively by the authors to ensure correctness and reproducibility.

A.5 A.5 Verification

All AI-generated outputs, whether textual or programmatic, were subject to rigorous human oversight. We applied the following verification steps:

(1) Cross-checking every factual claim from AI against the original Wikipedia pages, academic papers, or dataset sources.
(2) Manually testing and debugging all AI-generated code to confirm correctness and safety.
(3) Editing all written content to ensure accuracy, tone consistency, and alignment with course requirements.

No AI-generated content was included in the final paper without human verification. The authors maintained full responsibility for all analytical interpretations, code decisions, and written arguments presented in this work.