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 5 discusses implications and limitations. Section 6 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 LLM Scoring Outcomes Across Articles

Using the ten–dimension instructional rubric, the LLM produced consistent 1–5 scores and justifications for each of the curated Wikipedia articles. Articles such as *Pythagorean Theorem*, *Golden Ratio*, *Fibonacci Sequence*, and *Neural Network* received some of the highest overall scores (4.3–4.7 average) due to clear introductions, strong motivating examples, and accessible progression from simple cases to general formulations.

In contrast, technically dense pages such as *Tensor Product*, *Group (mathematics)*, *Riemann Hypothesis*, and *Eigenvalues and Eigenvectors* scored lower (3.0–3.6 average). These articles were penalized for steep conceptual jumps, limited intuitive scaffolding, and reliance on symbolic notation without preliminary examples.

Across domains, mathematics and machine learning articles generally outperformed finance and economics articles on motivation and example use, whereas socially-oriented pages (e.g., *2008 Financial Crisis*) scored highly on real-world relevance but lacked multiple representations or visual support.

4.2 Consistency and Validation Checks

To validate the evaluations, each article was scored three times. Most articles had low variance (standard deviation < 0.3), indicating strong scoring stability. A small number of pages—especially symbol-heavy math pages—showed the highest variance, reflecting the LLM’s sensitivity to presentation style when content is dense.

Human spot-checks generally aligned with the model’s judgments. Reviewers agreed with the LLM on whether an article had motivating examples, real-world connections, or a clear introductory section. Discrepancies occurred primarily when the model interpreted symbolic expressions as “visual support” even when no diagrams were present, or when it inferred historical context from brief mentions.

4.3 Qualitative Patterns in Strong vs. Weak Articles

The LLM justifications revealed recurring patterns in how “good for learning” articles distinguished themselves:

- **Motivation and narrative matter.** Articles with brief historical background or motivating context (e.g., *Golden Ratio*, *Fibonacci Sequence*) received higher scores in engagement and clarity.
- **Examples are a dominant signal.** Articles introducing examples early—especially geometric or numerical ones—were consistently rated higher in accessibility and progression.
- **High-notation pages struggle without scaffolding.** Articles like *Tensor Product* or *Group (mathematics)* were penalized for immediately using abstract definitions or formal symbols without intuitive lead-ins.
- **Real-world relevance improves scores.** Pages connecting concepts to real applications (e.g., *Fourier Transform*, *2008 Financial Crisis*, *Machine Learning*) reliably scored high on relevance and engagement.
- **Multiple representations boost scores.** Articles combining text, equations, and diagrams substantially outperformed those that relied on only one form—confirming multimodal support as a strong indicator of learnability.

4.4 Summary of Findings

Overall, the LLM’s evaluations reveal a clear pattern: Wikipedia articles rated as “good for learning” tend to include:

(1) a motivating and approachable introduction,
(2) intuitive examples before formal definitions,
(3) multiple ways of presenting core ideas (equations, diagrams, verbal explanation),
(4) a progression from concrete cases to general concepts, and
(5) explicit connections to real-world applications.

Articles lacking these elements—particularly those that are highly symbolic or definition-first—tend to be rated as less effective for learning.

These results demonstrate that a rubric-guided, prompt-engineered LLM can meaningfully and reliably differentiate between strong and weak educational articles, suggesting a promising direction for AI-assisted evaluation of open-access learning resources.

5 Discussion

5.1 Interpretation of Results

[What do your findings mean? How do they answer your research questions?]

5.2 Implications

[What are the broader implications of your work? For Wikipedia? For research? For practice?]

5.3 Limitations

[Be honest about limitations: data constraints, methodological issues, scope boundaries, etc.]

5.4 Future Work

[What questions remain? What should future research investigate?]

6 Conclusion

[Restate the problem you investigated]

[Summarize your approach and key findings]

[Emphasize your main contributions]

[End with a forward-looking statement about the importance of your work or future directions]

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

A.1 Literature Review

[Describe how you used AI agents for literature review. Reference your .prompt.md file.]

Example: We used an AI agent workflow (see the file literature-review.prompt.md) to systematically process research papers. The agent extracted summaries, methodology descriptions, and key findings from papers in our bibliography.

A.2 Data Analysis

[If you used AI for data analysis, code generation, or statistical work, document it here]

A.3 Writing Assistance

[Document any AI assistance in writing: brainstorming, editing, restructuring, etc.]

Example: We used Claude/ChatGPT to help with [specific task, e.g., "improving clarity of the abstract" or "suggesting visualizations for our data"].

A.4 Code Development

[If AI helped you write code for data collection or analysis, document it]

A.5 Verification

[How did you verify AI-generated content? What human oversight did you apply?]

All AI-generated content was reviewed, verified against primary sources, and edited by the human author(s). Factual claims were cross-checked with original papers and data.