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 tracks. **Track A** investigates which linguistic and structural features make Wikipedia articles effective for learning. **Track B** develops a "Wiki Diagnostic Tutor" that applies those insights to classify and explain article quality. Together, these tracks address the research question: *Which Wikipedia articles are good for learning versus bad for learning, and how can prompt-engineered LLMs identify that difference?*

3.2 Data Collection

For **Track A**, we performed a systematic review of research on instructional text design, focusing on coherence, scaffolding, and conceptual density. Using these insights, we defined a set of criteria describing what makes an article effective for learning (see below). We then compiled a labeled dataset of Wikipedia articles drawn from topics in mathematics, computer science, and economics that were manually judged as "good for learning." Example articles include:

- Pythagorean Theorem, Golden Ratio, Fibonacci Sequence
- Eigenvalues and Eigenvectors, Neural Networks, Machine Learning
- 2008 Financial Crisis, Investment Banking

Each article's revision ID and access date were recorded for reproducibility.

For **Track B**, we used the Wikipedia API, accessed via the OpenAI Apps SDK [7], to dynamically retrieve article summaries and content. The prototype uses prompt engineering to help LLMs reason step-by-step about whether an article satisfies the educational criteria, rather than simply classifying it directly.

3.3 Evaluation Criteria

Articles were judged on ten instructional dimensions derived from cognitive and instructional design literature:

(1) **Engaging Introduction** – presence of historical context or motivation.
(2) **Real-World Relevance** – clear explanation of why the topic matters.
(3) **Motivating Examples** – intuitive examples before formal definitions.
(4) **Visual and Symbolic Support** – inclusion of equations, images, or diagrams.
(5) **Gradual Progression** – movement from simple to abstract ideas.
(6) **Balanced Approach** – mix of intuitive and formal reasoning.
(7) **Multiple Representations** – verbal, visual, and symbolic presentation.
(8) **Comprehensive but Accessible** – complete coverage without excessive jargon.
(9) **Multiple Proofs and Perspectives** – variety of explanations or derivations.
(10) **Applications and Impact** – real-world uses or implications.

These dimensions informed both manual ratings and the prompt templates used for model evaluation.

3.4 Data Processing

We removed markup, standardized section headings, and converted text to plain UTF-8 for tokenization. Articles were formatted into JSON objects containing title, revision ID, and text fields. For reproducibility, all preprocessing and classification scripts are available in our GitHub repository: `Generative-LLMs-in-Education`.

3.5 Analysis Methods

In **Track A**, we analyzed relationships between textual features (readability, sentence length, and structure) and perceived educational quality through descriptive comparisons and manual rubric scoring.

In **Track B**, we applied prompt-engineered LLMs using a chain-of-thought reasoning approach. The model received both the article text and the instructional criteria and was asked to produce an evaluation explaining whether the article is pedagogically effective, neutral, or ineffective. Evaluation metrics included precision, recall, and qualitative interpretation accuracy compared to human-labeled judgments.

3.6 Integration and Reproducibility

Outputs from Track A informed prompt phrasing in Track B, ensuring the model aligned with human-defined learning features. All code, example prompts, and result logs are available at `github.com/arya-kumar1/` which includes a README detailing dataset structure, API setup, and usage examples.

3.7 Ethical Considerations

All data consist of publicly available Wikipedia content accessed under the Wikimedia Terms of Use. No personal data were used, and human evaluators remained anonymous.

4 Results

4.1 Prompt-Based Evaluation Outcomes

The prompt-engineered LLM successfully distinguished between well-structured and educationally weak Wikipedia articles. Articles such as *Pythagorean Theorem* and *Photosynthesis* were consistently rated as highly effective due to their clear introductions, gradual conceptual development, and multiple explanatory representations. Articles that lacked historical or intuitive context—such as certain technical stubs—were classified as less effective for learning.

4.2 Human vs. Model Agreement

Across the initial 20 manually evaluated articles, model judgments aligned with human ratings in 85% of cases. The strongest agreement was on criteria involving structure and example use, while disagreements occurred for topics with dense mathematical notation (e.g., *Tensor Product*) where clarity depends on reader background.

4.3 Qualitative Insights

Several patterns emerged:

- Articles that began with a motivating story or historical background (e.g., *Golden Ratio*) received higher clarity and engagement scores.
- Articles connecting abstract math to real-world relevance (e.g., *Fourier Transform*, *2008 Financial Crisis*) were rated highly by both humans and the LLM.
- Overly symbolic or jargon-heavy pages without scaffolding (e.g., *Group (mathematics)*, *Riemann Hypothesis*) were more likely to be classified as poor for learning.

4.4 Summary of Findings

Good Wikipedia articles for learning tend to:

(1) Begin with motivating context and intuitive examples.
(2) Maintain a balance between formality and accessibility.
(3) Integrate visuals and real-world relevance throughout.

These findings suggest that prompt-engineered LLMs can serve as effective assistants for evaluating educational quality and guiding improvements in open-access 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.