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

2 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 **??** reviews related work. Section **??** details data and methods. Section 7 reports findings. Section 8 discusses implications and limitations. Section 9 concludes.

3 Related Work

4 Related Work

Overview. Our study spans three literatures: (i) instructional quality and writing/learning research, (ii) AI tutoring and reasoning methods (Socratic, CoT), and (iii) Wikipedia governance and quality dynamics.

4.1 Instructional Quality, Writing Attitudes, and Critical Thinking

Clear definitions, consistent constructs, and coherence are central to educational quality. Reviews of writing attitudes stress definitional clarity and theoretically grounded measures [3]. Classic reviews of critical thinking in education emphasize how pedagogical design affects higher-order thinking [8]. Practitioner scholarship highlights self-explanation as an effective learning strategy [1]. In EFL contexts, meta-synthesis work catalogs strategies (e.g., scaffolding, feedback) that improve academic writing and learner motivation [4].

4.2 AI Tutors, Socratic Guidance, and Chain-of-Thought

Randomized studies show well-designed AI tutors can outperform in-class active learning on engagement and learning gains [6]. For safer pedagogy, LLMs that guide via questions rather than reveal answers can preserve critical thinking [2]. Chain-of-Thought (CoT) prompting operationalizes stepwise reasoning to increase accuracy and interpretability in complex tasks [5].

4.3 Wikipedia Governance, Bias, and Quality

Wikipedia's collaborative norms and governance shape content quality and reliability [9]. Behavioral analyses of deletion practices investigate potential systemic biases and their implications for coverage and quality [11]. These threads motivate automated, theory-informed evaluations of article *educational* quality.

4.4 Our Work in Context

We bridge these strands by (1) turning instructional design findings into measurable text features and (2) embedding those features into a CoT-enabled tutor that retrieves content via the Wikimedia REST API and classifies pedagogical quality with human-interpretable rationales.

5 Methodology

6 Methodology

Overview

We follow a two–track design. **Track A** identifies linguistic/structural correlates of educational quality via literature-driven feature engineering and empirical analysis. **Track B** implements a CoT-enabled "Wiki Diagnostic Tutor" that retrieves Wikipedia content via API and classifies articles with explanations. An additional Week-7 study evaluates the educational suitability of lead summaries using a simple clarity–comprehensiveness–readability rubric.

6.1 Data Collection

Track A. We assembled a labeled set of Wikipedia articles rated for educational quality by prior research or expert raters (with revision IDs and access dates). We also compiled a review corpus on instructional quality, writing attitudes, and critical thinking [3, 4, 8]. Articles were fetched via the Wikimedia REST API summary and content endpoints [10], with request metadata logged.

Track B. The prototype connects to the Wikipedia API using the OpenAI Apps SDK tool interface to fetch the page summary (and optionally full text) at inference time [7].

Week-7 Summaries Study. We executed `wiki_summarizer.py` on 5–10 topics (e.g., "AI," "Machine learning," "Neural network," "Overfitting," "Photosynthesis") and recorded printed summaries and rubric scores (clarity, comprehensiveness, readability; 1–5 scale), producing per-topic and average scores.

6.2 Data Processing

We stripped markup and references, preserved section hierarchy, and tokenized sentences/words. We computed readability (Flesch–Kincaid, Coleman–Liau), hierarchy depth, mean sentences per section, example density (examples per 1,000 words), discourse cues (connectives), and lexical diversity. For labels from multiple sources, two expert raters adjudicated disagreements; inter-rater reliability was assessed with Cohen's κ.

6.3 Analysis Methods

Track A (Feature Discovery). We tested associations between features and labels using ANOVA/Kruskal–Wallis and logistic regression with FDR correction; we report effect sizes (Cohen's d, Cliff's δ) and domain robustness.

Track B (Modeling). We frame tri-class classification with CoT rationales. The system uses: (i) a Wikipedia retrieval tool; (ii) few-shot CoT prompts emphasizing coherence, scaffolding, example use, and hierarchy; (iii) instruction-tuning on annotated examples. Splits are stratified (70/15/15) at the article level.

Evaluation. Automated metrics include accuracy, macro-precision/recall/F1 with 95% bootstrap CIs. Human experts blind-rate a 10% sample for validity and reasoning clarity; we compute Spearman ρ and κ. Ablations remove Track-A-inspired cues to quantify contribution deltas.

6.4 Week-7 Rubric Results (Baseline)

Average summary score across topics $\approx 4.1/5$ (*high suitability*). The most effective summaries followed a "definition + key ideas/applications" pattern (e.g., AI, Photosynthesis), suggesting lead summaries can serve as early indicators for article-level educational quality.

6.5 System and Reproducibility

All scripts (preprocessing, modeling, evaluation) are version-controlled with fixed seeds and an environment lockfile (Python 3.11; `pandas`, `numpy`, `scikit-learn`, `spacy`, `textstat`). We log prompts, tool I/O (titles, revision IDs), and outputs. API use complies with Wikimedia Terms of Use; we respect rate limits.

6.6 Ethical Considerations

Data are public Wikipedia content accessed under the Wikimedia Terms of Use. We collect minimum necessary metadata, anonymize rater IDs, and release revision IDs for reproducibility.

7 Results

[Brief paragraph introducing your main findings]

7.1 [Finding 1 - descriptive title]

[Present your first main finding]

7.2 [etc]

[Present your second main finding]

8 Discussion

8.1 Interpretation of Results

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

8.2 Implications

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

8.3 Limitations

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

8.4 Future Work

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

9 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.