Automatic creation of Wikipedia articles about Zambia utilizing Retrieval-Augmented Generation techniques and fact-based vector databases

Frazer Nyambe
Department of Computer Science
University of Zambia
Lusaka, Zambia
frazer.nyambe@cs.unza.zm

Lighton Phiri
Department of Library and Information Science
University of Zambia
Lusaka, Zambia
lighton.phiri@unza.zm

Abstract—This study investigates the automatic creation of Wikipedia articles about Zambia through the use of Retrieval-Augmented Generation (RAG) techniques integrated with fact-based vector databases. While Wikipedia serves as a vital open-access knowledge platform, its coverage of Zambia remains inadequate, with many topics underrepresented or missing. Generative AI, particularly Large Language Models (LLMs), presents opportunities for addressing these gaps but is hindered by issues such as factual hallucination and reliance on low-quality, machine-translated web data. To address these challenges, this research proposes a RAG-based approach that grounds content generation in curated, reliable datasets to improve accuracy, contextual relevance, and editorial usability. The study employs a mixed-methods design involving controlled experiments with Zambian university students, implementation of a RAG prototype system, and evaluation of editor acceptance of AI-generated drafts. Key objectives include assessing whether factual resources increase willingness to contribute, evaluating the effectiveness of RAG in producing reliable Wikipedia content, and exploring editor perceptions of AI assistance. By combining technical development with empirical evaluation, this research contributes to both the advancement of trustworthy AI content generation and the promotion of equitable digital knowledge representation for underrepresented regions such as Zambia.

Keywords—Generative AI, Retrieval-Augmented Generation, Vector Database, Editor Support.

I. CHAPTER 1

A. Introduction

1) Background

The advancement of Large Language Models (LLMs) and generative Artificial Intelligence has opened up new possibilities for automating the creation of knowledge resources such as Wikipedia. However, the reliability and accuracy of AI-generated content remain significant concerns due to hallucinations and unverified sources. Prior studies have explored machine-generated Wikipedia content using various approaches, including extractive summarization, document embeddings, and bot-generated articles. A promising approach involves using Retrieval-Augmented Generation (RAG), which grounds generated content in factual databases, thus potentially improving reliability.

This study focuses on the automated generation of Wikipedia content about Zambia, using fact-based vector databases and RAG methods. It aims to evaluate the feasibility, reliability, and community acceptance of AI-generated content within the Zambian context.

2) Statement of the Problem

Wikipedia remains one of the most widely used open-access knowledge platforms globally. However, its coverage of African countries, such as Zambia, remains inadequate. Numerous topics related to Zambia lack detailed, factual representation. With recent advancements in generative AI, particularly Large Language Models (LLMs), there is growing potential to automate the creation of Wikipedia articles. However, issues such as factual hallucination and reliance on poor-quality sources remain significant challenges. A large proportion of web content, which generative models are trained on, is machine-translated, further increasing the risk of misinformation [10]. This raises the need for more reliable, fact-grounded methods of generating content.

This project addresses this problem by proposing the use of Retrieval-Augmented Generation (RAG) combined with a fact-based vector database to generate accurate Wikipedia content about Zambia. It seeks to investigate the impact of such an approach on content creation and user contribution to Wikipedia.

3) Aim or Purpose of study

The purpose of this study is to explore the feasibility of automatically generating Wikipedia content about Zambia using Retrieval-Augmented Generation (RAG) supported by a fact-based vector database. The study aims to assess how this approach impacts content quality, supports Wikipedia editors, and enhances knowledge representation on the platform.

4) Study Objectives

a) To empirically determine whether the availability of factual resources increases user willingness to contribute content to Wikipedia.

b) To design and implement a RAG-based system for generating Wikipedia articles on Zambia.

c) To evaluate the usefulness and acceptance of AI-generated content among Wikipedia editors.

 ISBN: 978-9982-95-500-3

5) *Research Questions*

a) *Does access to a curated set of factual information make users more likely to contribute content to Wikipedia?*

b) *How effective is the RAG approach in generating accurate, relevant, and well-structured Wikipedia content for Zambia-related topics?*

c) *To what extent do Wikipedia editors find AI-generated draft content helpful in enhancing editing efficiency and content quality?*

6) *Significance of the Study*

This study will provide insights into using AI tools to bridge Wikipedia's content gaps, particularly for Zambia. It also informs the design of AI systems that collaborate with human editors rather than replacing them. The findings will contribute to better AI-human interaction in knowledge systems and promote equitable digital content creation.

7) *Theoretical or Conceptual Framework*

The research is anchored in socio-technical systems theory [9], which emphasizes the interplay between technology (AI-generated content) and social actors (Wikipedia editors). The use of RAG forms the technological basis, while editor feedback and engagement represent the human component in the content creation ecosystem.

8) *Scope of the Study*

The study focuses on generating and evaluating Wikipedia articles about Zambia using fact-based RAG approaches. It covers selected topics including culture, politics, and geography, and involves both automatic generation and human evaluation.

1.9. Operational Definitions

Generative AI: AI systems capable of creating new content, such as text.

RAG (Retrieval-Augmented Generation): AI approach combining document retrieval and language generation to ensure factual grounding.

Vector Database: A semantic search system storing documents as vectors to retrieve contextually relevant data.

Editor Support: Positive engagement, edits, or approvals by Wikipedia editors on AI-generated content.

1.10. Ethical Considerations

The study will respect Wikipedia's content guidelines, clearly disclose AI involvement, and obtain informed consent from all participating editors. No personal data will be collected, and all results will be anonymized to ensure privacy.

II. CHAPTER 2

A. Related Work

1) *User willingness to contribute content to Wikipedia*

Research into Wikipedia content generation has evolved from early extractive approaches to more sophisticated models that treat the task as a multi-document summarization problem. [6] introduced a two-stage framework combining extractive and abstractive summarization, highlighting that the coarse extraction stage plays a crucial role in final output quality. Their use of a decoder-only sequence transduction model allowed the processing of long input-output pairs and significantly outperformed traditional encoder-decoder architectures, enabling coherent article generation from multiple source documents. In parallel, the historical progression of large software system analysis has moved from manual code reviews to automated, fact-based modeling due to the growing complexity and diversity of modern codebases. Traditional tools often struggle with concurrency and scale, leading to frequent false positives due to infeasible execution paths. While the inclusion of control-flow-graph (CFG) facts has improved precision, many studies overlook the performance trade-offs and limited applicability outside specific frameworks such as ROS. [4] critiques this shortfall, proposing a staged query system with real-time CFG validation, yet the generalizability of these techniques across broader software ecosystems, including legacy systems, remains insufficiently examined.

At the same time, the development of automatic content generation particularly via Machine Translation (MT) has transformed the multilingual landscape of Wikipedia and related platforms. Earlier work emphasized extractive summarization and bot-supported article creation, while recent efforts leverage Retrieval-Augmented Generation to enhance factual accuracy. However, a critique of existing literature reveals that much of the web data used in training is low-quality, MT-generated, and biased especially in low-resource languages where short, repetitive, and commercially motivated content dominates. This raises significant concerns about the validity of training data and the integrity of generated content. While some research has introduced filtering mechanisms such as MT detection and parallelism analysis, their scalability and effectiveness remain underexplored. A pressing knowledge gap lies in the lack of localized, culturally relevant content generation such as for Zambia where both training datasets and generated outputs often omit region-specific facts and narratives. As [10] observed, the prevalence of machine-translated content in web-scraped corpora poses ongoing challenges for ensuring content reliability in generative AI systems.

2) *RAG-Based System Implementation*

Recent studies have provided an in-depth exploration of Retrieval-Augmented Large Language Models (RA-LLMs), an advanced AI methodology designed to overcome key shortcomings of traditional Large Language Models (LLMs) (Baeza-Yates & Bonchi, n.d.). RA-LLMs work by integrating external knowledge retrieval into the generation process, which significantly reduces hallucinations, updates outdated internal knowledge, and strengthens domain-specific capabilities [2]. These models have been reviewed in terms of their architecture, training strategies, and wide-ranging applications, from question answering and chatbots to domain-specific tasks like financial forecasting and molecular discovery. Moreover, the literature identifies future research opportunities, including the development of multilingual and multi-modal RA-LLMs, enhancing the reliability of external sources, and addressing ethical concerns such as data privacy. These insights are especially critical for researchers and developers working to optimize AI systems for trustworthy and scalable deployment.

In parallel, the challenge of limited human contributors on Wikipedia has led to the development of automated content generation techniques. Given the scattered nature of web-based information, automation is seen as a practical solution for enriching stub articles with short, underdeveloped Wikipedia entries. One notable approach by [8] involves using machine learning classifiers to suggest content for stubs by analyzing comprehensive existing articles. Among the models tested Latent Dirichlet Allocation (LDA), Deep Belief Networks, and TF-IDF the LDA-based model showed superior performance in generating coherent and complete additions. This technique has demonstrated real-world effectiveness, with several enhanced stubs successfully retained on Wikipedia, indicating its potential for scalable article improvement.

Simultaneously, the study of Wikipedia bots has progressed from early manual tracking of basic functions like content injection to sophisticated, machine learning-based analyses of bot behavior and roles. Since the debut of Rambot in 2002, bots have become a vital part of the platform's editing ecosystem. While their activity has declined on English Wikipedia, bots remain dominant in projects like Wikidata. Initial research was limited in scope and methodology, but recent studies have expanded to larger datasets and applied systematic analysis to understand bot interactions with human editors and their evolving functions. Nonetheless, the literature still suffers from a heavy focus on English Wikipedia, difficulties in classifying complex bot types such as Advisor bots, and limited exploration of how bot roles change over time. Despite these gaps, the transparency of Wikipedia's bot governance and the integration of advanced analytics lay a strong groundwork for future research to better understand and enhance bot - supported collaboration [5].

3) *Evaluating Editor Support and Acceptance*

Research on digital public goods such as Wikipedia has historically centered on the goal of democratizing access to knowledge. While early perspectives were characterized by optimism, recent studies have shifted toward addressing the practical challenge of maintaining high-quality contributions, especially from domain experts. Scholars have long highlighted both intrinsic and extrinsic motivators for volunteer work; however, only in recent years have these motivations been empirically investigated in the context of expert participation on collaborative platforms. A notable field experiment demonstrated that private incentives such as increased visibility through citations are more effective in encouraging expert contributions than appeals to social impact alone. Although the research design was robust, employing a large sample and advanced predictive modeling, its findings may have limited applicability beyond the experimental context due to the specific participant group involved. Moreover, existing literature often fails to explore long-term engagement trends or the socio-cultural barriers experts may face. This leaves a critical knowledge gap in understanding how scalable, personalized strategies like recommender systems can ensure sustained participation across diverse knowledge domains and underrepresented regions [1].

In parallel, the collaborative nature of Wikipedia has drawn attention to editor dynamics, particularly regarding how users interact through article talk pages during content disputes. Earlier studies emphasized the cognitive and social value of engaging with controversy, suggesting it can enhance individual learning and critical thinking. However, most Wikis lack effective tools to help editors identify and participate in meaningful discussions. Recent interventions, including visual controversy markers and structured collaboration scripts, have shown some promise in guiding user behavior. Nevertheless, the validity of these findings is constrained by their reliance on experimental settings that may not capture real-world complexities. Furthermore, much of the research remains focused on immediate behavior rather than long-term engagement or content improvement. As a result, there is still a gap in understanding how such guidance tools can be adapted to broader editor populations and sustained over time on large-scale platforms [3].

Additionally, emerging studies have identified significant differences in linguistic and stylistic patterns between AI-generated and human-authored texts. AI-generated conversations tend to lack depth and coherence compared to human dialogues, and AI-written essays and news articles are often disproportionately loaded with information-dense features. These variations suggest that AI contributions may not fully replicate the collaborative tone and integration needed in human-centered knowledge environments like Wikipedia, raising further concerns about the quality and reliability of AI-generated content [7].

III. CHAPTER 3

A. Methodology

Research Approach

An explanatory sequential design will be applied, where Quantitative data will be collected from an experiment to explore the edited content by the editors, followed by the qualitative data to clarify initial quantitative findings on the edited content and with an emphasis on evaluating how the use of structured, fact-based datasets affects the quality and quantity of content generated for Wikipedia. The study will be divided into three components, each aligned with a specific research objective. This approach will enable triangulation of data to improve the validity and reliability of findings.

Research Design

Pilot study

A pilot study will be conducted on a control group that will be tasked with writing a Wikipedia-style article without access to a factual database and an experimental group that will not have access to a curated fact-based dataset extracted from a semantic vector database like chroma db which is an open-source vector store used for storing and retrieving vector embeddings. This will help identify the feasible results and anticipated outcomes for the larger scale solution implementation.

User willingness to contribute content to Wikipedia

 ISBN: 978-9982-95-500-3

A mixed-methods research design will be adopted, integrating both quantitative and qualitative data collection methods. This will be obtained from the contribution rate and content length and quality for edited wikipedia content by students.

RAG-Based System Implementation

This study adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) model to guide the design and implementation of the RAG-based system. The first phase, Business Understanding, will involve defining the main goal of generating accurate Wikipedia content on Zambia using a Retrieval-Augmented Generation (RAG) approach. Success was measured based on criteria such as factual accuracy, contextual relevance, and editorial usability. In the Data Understanding phase, relevant data sources (domain) including government publications, academic papers and verified news articles on Zambia were collected and examined for quality, completeness, and suitability for vector-based indexing. This stage also included the understanding of the knowledge base of the existing LLM like Ollama deepseek and their accuracy in providing information. The Data Preparation phase involved cleaning and transforming these sources into structured, machine-readable formats, removing noise and creating a high-quality domain-specific corpus for semantic retrieval. This was accompanied by the provision of a knowledge base to the LLM so as to enable them to provide results based on the knowledge. During the Modeling phase, a system prototype was developed from existing LLM models like Ollama deepseekr1:1.5b that has open reasoning, integrated with Open WebUI providing the interactive user interface. This was used for prompt engineering to test the potential hallucinations from LLM that do not have a knowledge base and later on provided it with a knowledge base for accuracy argument. Finally, a high RAG architecture was developed by integrating a dense retrieval mechanism with a generative language model. In nutshell, before generating a response, the fine tuned LLM was retrieving relevant information from a specified knowledge base and incorporated this information into its prompt, allowing it to access and utilize up-to-date and domain-specific information that would provide researchers with factual details on information they might be looking for on wikipedia. The Evaluation phase tested the system's outputs on Zambia-related topics by assessing them for coherence, relevance, and factual consistency, with feedback gathered from control group and experimental group. Finally, the Deployment phase involved rolling out the system within Wikipedia's environment to enable controlled experimentation and iterative refinement based on user interaction and editing behavior.

Evaluating Editor Support and Acceptance

To assess editor interaction with AI-generated content, this study conducted an experimental evaluation of a RAG-based system capable of retrieving factual information from a vector database and generating draft Wikipedia articles. .Both quantitative engagement metrics and qualitative feedback were collected to measure editors' trust in the system, perceived content quality, and their willingness to contribute to or improve AI-assisted drafts.

Study Area or Site

Wikipedia's environment will be used to host draft articles based on Zambia-related topics, allowing safe experimentation before public submission. This will be supported by the use of chroma db to store and retrieve vector based content.

Study Population

This research will take advantage of tertiary level students from the Zambian universities, particularly those familiar with Africa-related articles and with web content editing skills.

Study Sample

The English Wikipedia has 49,134,976 registered users as of the last update. However, only about 30% of these users have ever edited the site. In the last 30 days, 117,838 users were considered active. However, there are only less than 10 zambian wikipedia editors. Therefore, a controlled study on two groups of people will be used to edit content on a curated dataset consisting of factual content on Zambia sourced from reliable databases such as government records, academic journals, and verified reports. This study will use a purposive sample of not less than 10 university student editors and a curated dataset consisting of factual content on Zambia sourced from reliable databases such as government records, academic journals, and verified reports.

Sampling Techniques

Purposive sampling will target student editors who are familiar with content web editing. Fact-based datasets will be compiled using structured search criteria focused on relevance, accuracy, and credibility.

Instruments for Data Collection

An experimental test will be used to test for the accuracy of the edited content by the control group and the experimental group. The use of questionnaires targeting perceptions of AI tools and content quality. The use of feedback form designed to capture specific support or concerns. Application of editor activity logs and discussion summaries.

Procedure for Data Collection

Participants will be divided into two groups: Control Group; Tasked with writing a Wikipedia-style article without access to a factual database. Experimental Group; Given access to a curated fact-based dataset extracted from a semantic vector database. Giving out surveys to the control group and experimental group in order to evaluate ease of use, confidence, and willingness to contribute.

Data Analysis

Quantitative analysis will focus on edit frequency, number of rejections/approvals, and suggestions per article. These metrics will indicate the practicality of the RAG system. Thematic analysis will be conducted on qualitative responses to identify key factors influencing editor trust, usability preferences, and potential barriers to adoption.

Evaluation

Benchmark the system using metrics such as BLEU (for structure), ROUGE (for relevance), and human-rated factual accuracy scores.

 ISBN: 978-9982-95-500-3

IV. CHAPTER 4

A. Conclusion

This study proposes a novel approach to addressing Wikipedia's content gaps for Zambia by combining Retrieval-Augmented Generation with factual databases. It aims to offer a scalable, responsible, and practical model for enhancing Wikipedia contributions using AI. The findings will contribute to both technical advancements in AI and broader efforts toward global knowledge equity.

V. REFERENCES

[1] Chen, Y., Farzan, R., Kraut, R., YeckehZaare, I. and Zhang, A.F. 2023. Motivating Experts to Contribute to Digital Public Goods: A Personalized Field Experiment on Wikipedia. *Management Science*. (Dec. 2023). DOI:https://doi.org/10.1287/mnsc.2023.4852.

[2] Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S. and Li, Q. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models.

[3] Heimbuch, S. and Bodemer, D. 2016. Wiki Editors' Acceptance of Additional Guidance on Talk Pages. *Proceedings of the International AAAI Conference on Web and Social Media*. 10, 2 (2016), 51–52.

[4] Ke, X.Y. (fa F. 2024. Improving the Precision of Analyses Queries in Factbase Models of Software Systems. (May 2024).

[5] Lei Zheng, Christopher M. Albano, Neev M.Vora, Feng Mai, Jeffrey V. Nickerson 2019. The Roles Bots Play in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction*. (Nov. 2019). DOI:https://doi.org/10.1145/3359317.

[6] Liu, P.J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L. and Shazeer, N. 2018. Generating Wikipedia by Summarizing Long Sequences.

[7] Sardinha, T.B. 2024. AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*. 4, 1 (Apr. 2024), 100083.

[8] S Banerjee, P.M. 2025. ASurvey on RAGMeetingLLMs: Towards Retrieval-Augmented Large Language Models. *DocEng '15: Proceedings of the 2015 ACM Symposium on Document Engineering* (May 2025), Pages 117–120.

[9] Sony, M. and Naik, S. 2020. Industry 4.0 integration with socio-technical systems theory: A systematic review and proposed theoretical model. *Technology in Society*. 61, (May 2020), 101248.

[10] Thompson, B., Dhaliwal, M., Frisch, P., Domhan, T. and Federico, M. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism. *Findings of the Association for Computational Linguistics ACL 2024* (Stroudsburg, PA, USA, 2024), 1763–177

ISBN: 978-9982-95-500-3