Wikipedia in the Era of LLMs: Evolutions and Risks

Anonymous ACL submission

Abstract

Wikipedia helps both people and machines seeking knowledge about the world. In this paper, we present a thorough analysis of the influence of Large Language Models (LLMs) on Wikipedia, examining both human and machine perspectives. We begin by analyzing page views and article content to study Wikipedia's recent evolutions and assess the impact of LLMs. Subsequently, we examine how LLMs affect various Natural Language Processing (NLP) tasks related to Wikipedia, including machine translation and retrieval-augmented generation. Our findings and simulation results reveal that while LLMs have not yet fully permeated Wikipedia's language and knowledge structures, their current influence is significant enough to warrant careful consideration of potential future risks.[1]

1 Introduction

The creation of Wikipedia challenged traditional encyclopedias (Giles, 2005), and the rapid development and widespread adoption of Large Language Models (LLMs) have sparked concerns about the future of Wikipedia (Wagner and Jiang, 2025). In the era of LLMs, it is unlikely that Wikipedia has remained unaffected.

Recently, researchers have begun examining the influence of LLMs on Wikipedia. For example, Reeves et al. (2024) analyze metrics such as page views, unique visitor counts, edit frequency, and the number of editors. Meanwhile, Brooks et al. (2024) estimate the proportion of AI-generated content in newly created English Wikipedia articles using machine-generated text detectors. However, these detectors have notable limitations (Doughman et al., 2024), which highlights the need to investigate the impact of LLMs on Wikipedia through more comprehensive and robust approaches.

[1]We release all the experimental dataset and source code via supplementary materials.

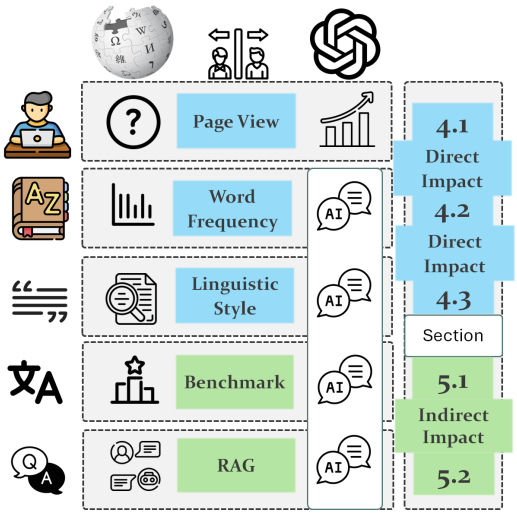

Figure 1: Analyzing the direct impact of LLMs on Wikipedia, and exploring the indirect impact of LLMs generated via Wikipedia.

On the other hand, Wikipedia is widely recognized as a valuable resource (Singer et al., 2017), and its content is extensively utilized in AI research, particularly in Natural Language Processing (NLP) tasks (Johnson et al., 2024b). For instance, Wikipedia pages are among the five datasets used to train GPT-3 (Brown et al., 2020). The sentences in the *Flores-101* evaluation benchmark are extracted from English Wikipedia (Goyal et al., 2022). In the work by Lewis et al. (2020) on Retrieval-Augmented Generation (RAG), Wikipedia content is treated as a source of factual knowledge. Consequently, we aim to investigate the influence of LLMs on machine translation and knowledge systems using Wikipedia as a key resource.

In this paper, we seek to address a key question: *Do LLMs affect Wikipedia, and if so, how might they influence the broader NLP community?* Our primary goal is to evaluate the direct impact of LLMs on Wikipedia, focusing on changes in page views and article content. Furthermore, we explore how an increased influence of LLMs on Wikipedia

might affect NLP tasks that rely on its data. By analyzing both the direct and indirect effects of LLMs on Wikipedia, we hope to gain a clearer understanding of the opportunities and challenges Wikipedia may face in the age of LLMs.

Figure 1 illustrates the various tasks and research topics discussed in this paper. In particular, we examine the historical progression of Wikipedia and evaluate the risks associated with the increasing prominence of LLMs. Our analysis yields a number of significant insights.

- There has been a slight decline in page views for certain scientific categories on Wikipedia, but the connection to LLMs remains uncertain.
- While some Wikipedia articles have been influenced by LLMs, the overall impact has so far been quite limited.
- If the sentences in machine translation benchmarks are drawn from Wikipedia content shaped by LLMs, the evaluation scores of machine translation models may be artificially inflated.
- Wikipedia content processed by LLMs appears less effective for RAG compared to genuine Wikipedia content.

Based on these findings, we underscore the importance of carefully assessing potential future risks and encourage further exploration of these issues in subsequent studies.

2 Related Work

Wikipedia. The value of Wikipedia is not limited to NLP. McMahon et al. (2017) have pointed out the substantial interdependence of Wikipedia and Google, and Vincent et al. (2018) found that Wikipedia can provide great value to other large-scale online communities, *Stack Overflow* and *Reddit* in particular. The influence of Wikipedia is border, including impacts on academic paper citations (Thompson and Hanley, 2018) and the click counts of other web pages (Piccardi et al., 2021). Kousha and Thelwall (2017) gave examples of Wikipedia's shortcomings.

Wikipedia for NLP. Wikipedia has long been utilized in various applications of NLP (Strube and Ponzetto, 2006; Mihalcea and Csomai, 2007; Zesch et al., 2008; Gabrilovich and Markovitch, 2009; Navigli and Ponzetto, 2010). Wikipedia also plays a role in the era of LLMs, such as in fact-checking (Hou et al., 2024) and reducing hallucinations (Semnani et al., 2023). Writing Wikipedia-like articles is also one of the LLM applications (Shao et al., 2024).

LLMs for Wikipedia. Researchers are trying to use LLMs to improve Wikipedia, including articles (Adak et al., 2025), Wikidata (Peng et al., 2024; Mihindukulasooriya et al., 2024) and edit process (Johnson et al., 2024a). There are also approaches to generating Wikipedia pages using LLMs: Zhang et al. (2025) think that there is still a gap compared to existing Wikipedia content, while Skarlinski et al. (2024) claim that language model writes cited, Wikipedia-style summaries of scientific topics can be more accurate than existing, human-written Wikipedia articles.

Estimation of LLM Impact. Different studies have shown that analyzing word frequency is effective in assessing the influence of LLMs (Liang et al., 2024; Geng and Trotta, 2024). The detection of AI-generated content has been a hot research topic in recent years (Wang et al., 2025; Zhang et al., 2024). Meanwhile, indirect effects also raise potential problems, for example, they might have changed how some people write and speak (Geng et al., 2024).

3 Data Collection

Wikipedia and Wikinews are both projects under the Wikimedia Foundation. While Wikipedia is the primary focus of our research, Wikinews is utilized to generate questions for RAG.

Wikipedia uses a hierarchical classification system for articles. It begins with top-level categories that cover broad fields, which are then divided into more specific subcategories. We chose articles from the following categories: *Art*, *Biology*, *Computer Science (CS)*, *Chemistry*, *Mathematics*, *Philosophy*, *Physics*, *Sports*.

Only pages created before 2020 and subcategories that are four or five levels away from our target category were included in our study. Then we scraped the Wikipedia page versions from 2020 to 2025 (more accurately, the version on January 1 of each year). Among them, *Philosophy* has the smallest number of articles (33,596), and *CS* leads with the largest number (59,097). More details on data collection and processing are shown in Appendix A. For a better comparison, we have also collected 6,690 *Featured Articles (FA)*, along with their corresponding 2,029 simple English versions

(where available) as *Simple Articles (SA)*.

We also collect Wikinews articles from 2020 to 2024. On average, there are over a hundred news per year, covering a wide variety of topics.

4 Direct Impact from LLMs

4.1 Direct Impact 1: Page View

Similar to the work of Brooks et al. (2024), we also collect the page views of articles using Wikimedia API, but within specific categories. The evolution of page views over time is shown in Figure 2 and Figure 7 in the appendix.

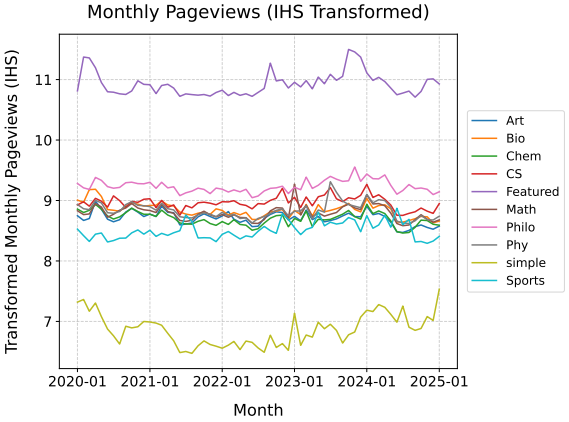

Figure 2: Monthly page views across different Wikipedia categories. The vertical axis represents the transformed page view values using the Inverse Hyperbolic Sine (IHS) method, which standardizes page view data across different categories.

> ***Finding 1:*** In the second half of 2024, there was a slight decline in page views across some scientific categories, and its connection to the use of LLMs requires further investigation.

4.2 Direct Impact 2: Words Frequency

In addition to page views, LLMs have likely had an impact on the content of Wikipedia articles. The frequency of certain words favored by ChatGPT has increased, such as "crucial" and "additionally" (Geng and Trotta, 2024), as shown in Figures 3 and 8. The word frequency changes we presented come from the same pages over the past few years, indicating that all changes are the result of recent edits.

The frequency evolution of word i could expressed in different ways, like the following one proposed by Geng et al. (2024):

$$f_i^d(S) - f_i^*(S) = \eta(S) f_i^*(S) r_i + \delta_i(S) \quad (1)$$

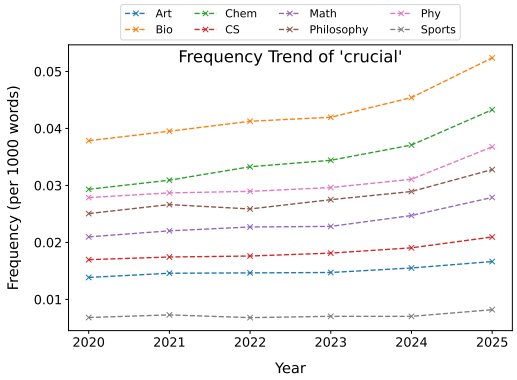

Figure 3: Word frequency in the first section of the Wikipedia articles.

where $f_i^d(S)$ represents the frequency of word i in the set of texts S, $f_i^*(S)$ represents the one if LLMs do not affect the texts, $\eta(S)$ is the LLM impact factor, r_i means word change rate caused by LLMs, and $\delta_i(S)$ is the noise term.

Thus, the impact of LLM-generated texts $\eta(S)$ could be calculated by

$$\hat{\eta}(S) = \frac{\sum_{i \in I} \left(f_i^d(S) - f_i^*(S)\right) f_i^*(S) \hat{r}_i}{\sum_{i \in I} \left(f_i^*(S) \hat{r}_i\right)^2} \quad (2)$$

$$\hat{r}_i = \frac{f(S_2) - f(S_1)}{f(S_1)} \quad (3)$$

where I is the set of words used for estimation, $f(S_1)$ and $f(S_2)$ represent the frequency of word i before and after LLM processing, respectively.

We take the average of the word frequencies from the 2020 and 2021 versions of the page as $f_i^*(S)$. But different text simulations still lead to different estimations, and using different words for estimation will also produce different results.

When estimating r_i through simulations using the first section of *Featured Articles* and *GPT-4o-mini* with a simple prompt: "*Revise the following sentences*", the LLM impact is approximately 1%-2% for the articles in certain categories, as illustrated in Figure 4. Additional results in Appendix B show that LLMs have significantly influenced certain categories of Wikipedia articles, even for those created before 2020.

> ***Finding 2:*** Though the estimation results vary, the influence of LLMs on Wikipedia is likely becoming increasingly significant over time.

4.3 Direct Impact 3: Linguistic Style

Overall. Beyond word frequency, we seek to investigate the current and future impact of LLMs

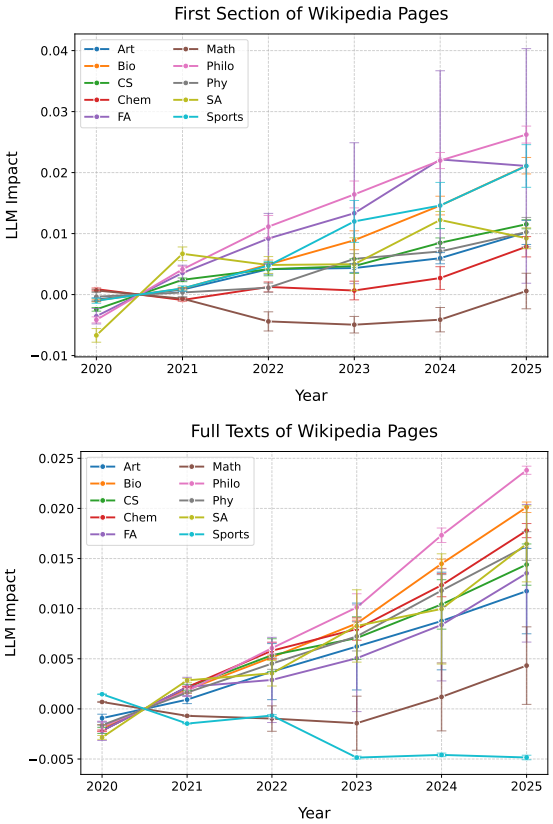

Figure 4: LLM Impact: Estimated based on simulations of the first section of *Featured Articles*, using different word combinations across different categories of Wikipedia pages.

on Wikipedia from more linguistic perspectives. In this section, we examine the evolutions in Wikipedia content at ***Word***, ***Sentence***, and ***Paragraph*** levels, as well as compare the texts before and after LLM processing under the same standards.

4.3.1 Experiment Setups

Word Level. At the word level, metrics such as the *frequency of auxiliary verbs* indicate a model's ability to convey complex reasoning and logical relationships (Yang et al., 2024). Lexical diversity, often measured by the *corrected type-token ratio (CTTR)*, reflects the variety of words used in a text (Wróblewska et al., 2025). Furthermore, the *proportion of specific parts of speech (POS)* is commonly used as a stylistic feature in assessing the quality of Wikipedia articles (Moás and Lopes, 2023).

Sentence Level. In terms of sentence structure, we focus on *sentence length* and the use of *passive voice* (AlAfnan and MohdZuki, 2023). Regarding sentence complexity, we analyze both the *depth of the entire syntactic tree* and the *clause ratio* (Iavarone et al., 2021).

Paragraph Level. For the paragraph dimension, which is essential for Wikipedia's educational mission (Johnson et al., 2024b), we seek guidance from *readability* evaluation, where six traditional formulas have been included in our study: *Automated Readability Index* (Mehta et al., 2018), *Coleman-Liau Index* (Antunes and Lopes, 2019), *Dale-Chall Score* (Patel et al., 2011), *Flesch Reading Ease* (Eleyan et al., 2020), *Flesch–Kincaid Grade Level* (Solnyshkina et al., 2017), and *Gunning Fog index* (Świeczkowski and Kułacz, 2021).

LLM Impact Simulation. Although LLM-generated content is increasingly being contributed to Wikipedia, it remains challenging to distinguish such content at a fine-grained level, specifically identifying which specific portions of text are LLM-generated. Therefore, we tend to gain deeper insights into LLM-generated text by simulating human' article creation and revision with *GPT-4o-mini* and *Gemini-1.5-Flash*.

4.3.2 Results

As shown in Figure 5, our simulation results reveal that LLMs substantially reduce the use of auxiliary verbs, with Gemini employing even fewer than GPT. Consistent with this trend, the usage of auxiliary verbs on real Wikipedia pages shows a marginal decline from 2020 to 2025. However, while LLMs predominantly favor the active voice, Wikipedia pages across different categories exhibit a consistent rise in the use of passive voice.

According to our result shown in Section C in the Appendix, LLMs tend to use more complex and varied vocabulary, often choosing longer and polysyllabic words over shorter and monosyllabic ones. They also use more nouns and fewer pronouns in their texts. In terms of sentence structure, LLMs typically produce longer sentences but generally refrain from starting them with pronouns or articles like *"It"* or *"The"*. Despite this, they maintain a balanced level of sentence complexity similar to *Featured Articles*, ensuring the text remains clear and structured.

The analysis of real Wikipedia pages reveals a gradual enhancement in lexical diversity across different categories, with a consistent upward trend in preposition usage and sentence length. Furthermore, both the average parse tree depth and clause ratio have shown consistent year-on-year growth, reflecting an increase in sentence complexity. Ad-

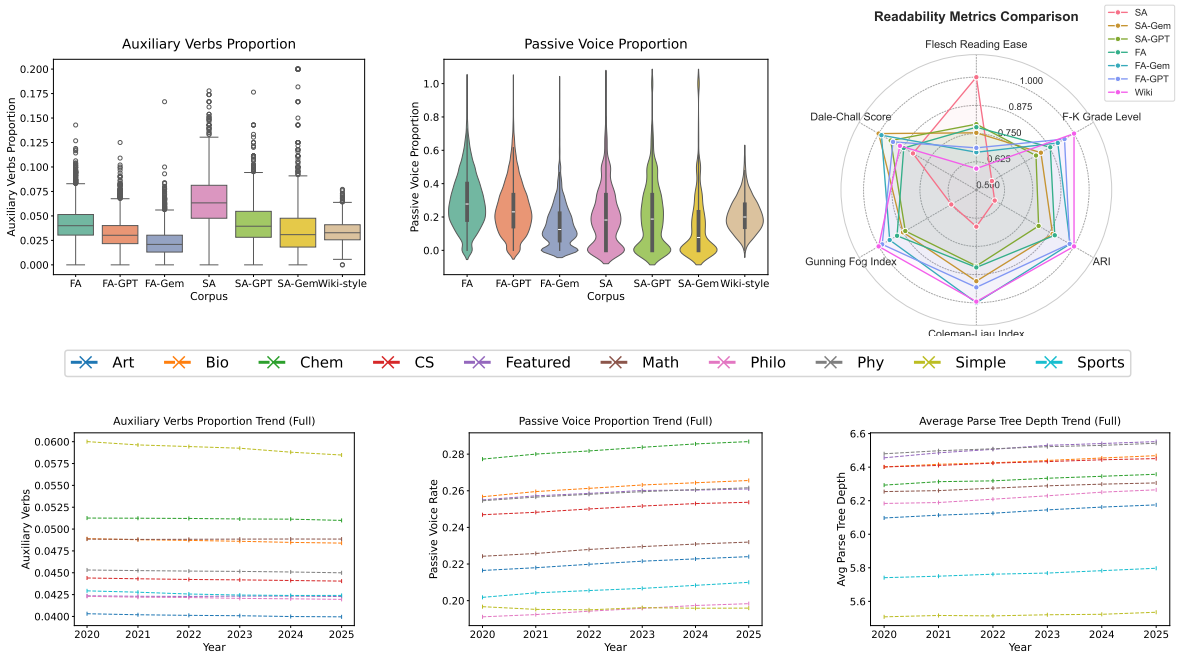

Figure 5: The results of linguistic style comparison, including the real Wikipedia pages and LLM-simulated pages.

ditionally, the frequency of article-initial sentence structures has risen. Other metrics remained relatively stable from 2020 to 2025.

As for *Readability*, the radar chart in Figure 5 presents the results of six *Readability* metrics, all of which indicate that LLM-generated texts tend to be less readable. The decrease in readability can be attributed to two primary factors. LLMs tend to generate longer, more complex sentences, which increase the difficulty of comprehension. Also, LLMs often use more advanced vocabulary or longer words. However, for Wikipedia pages across various categories, readability remained generally stable between 2020 and 2025.

Finding 3: The trends in several linguistic metrics of these Wikipedia pages do indeed show a close step to the characteristics of LLM outputs, although this is merely a correlation and does not imply causation.

5 Indirect Impact from LLMs

5.1 Indirect Impact 1: Machine Translation

Overall. Most machine translation benchmarks are derived from Wikipedia, while these same benchmarks are used to quantify LLMs' translation capabilities. This raises a critical concern: Will LLMs *"contaminate"* these benchmarks used to evaluate their performance after they prevail on Wikipedia?

5.1.1 Experiments Setups

Benchmark Construction. We utilized the Flores dataset[2], which comprises multiple IDs, each representing a single Wikipedia sentence available in several languages. Subsequently, we used GPT-4o-mini to translate the English (*EN*) version into the other languages, replacing the original versions to construct the LLM-influenced benchmark. The following 11 widely used languages were used in our simulations: Modern Standard Arabic (*AR*), Mandarin (*ZH*), German (*DE*), French (*FR*), Hindi (*HI*), Italian (*IT*), Japanese (*JA*), Korean (*KO*), Brazilian Portuguese (*PR*), Russian (*RU*), Latin American Spanish (*ES*). These languages represent a diverse set of linguistic families and regions, offering a broad evaluation of the model's performance across different cultural and linguistic contexts. More details are shown in Appendix D.2.

Metrics. We employ three automatic metrics to evaluate translations: *BLEU*, which uses n-gram precision with brevity penalty (Post, 2018), *COMET*, which leverages source and reference information (Rei et al., 2020), and *chrF*, which computes character-level F1 scores. These metrics compare machine-translated outputs against human-translated references.

[2] https://huggingface.co/datasets/openlanguagedata/flores_plus

Models. We evaluated three distinct machine translation models: *Helsinki-NLP*'s bilingual Transformer models[3] trained on OPUS corpus (Tiedemann and Thottingal, 2020; Tiedemann et al., 2023), *Facebook-NLLB*[4], a 54.5B parameter multilingual model supporting 200+ languages (NLLB Team et al., 2022), and *Google-T5 (mT5)*[5], pretrained on Common Crawl data covering 101 languages (Xue et al., 2021). For our comparative analysis, we specifically focused on its German and French translation capabilities.

Evaluation Pipeline. After collecting LLM-translated English samples, we use specific machine translation models to translate these sentences into eleven other languages. We then calculate *BLEU*, *chrF*, and *COMET* scores and compute the average for each language.

5.1.2 Results

In short, the simulation results indicate that these machine translation models have achieved higher scores on benchmarks influenced by LLM. For instance, in the case of the *Facebook-NLLB* model listed in Table 1, Latin American Spanish (*ES*) achieved a *BLEU* score of 28.76, a *ChrF* score of 58.97, and a *COMET* score of 86.91 on the original data (O). However, after LLM revision, these scores increased to 66.09, 82.33, and 91.24, respectively. This demonstrates that the LLM-processed benchmark significantly boosted all evaluation metrics for ES, with the *BLEU* score increasing more than 100%. Similarly, for the *Google-T5* model shown in Table 2, German (*DE*) initially had a *BLEU* score of 30.24, which rose to 44.18 after LLM revision, marking another substantial improvement.

It is worth noting that while the relative ranking of translation abilities among the three models remains unchanged across the languages we analyzed, the gap in their performance across different metrics is narrowing. For example, in Spanish (*ES*), under the original benchmark, *Helsinki-Nlp*'s *ChrF* score is 4.17 lower (about 7.6%) than *FaceBook-NLLB* as presented in Table 3, but under the LLM-influenced benchmark, the former's score is only 2.37 lower (about 2.9%). This change suggests that if the benchmark is more strongly influenced by LLMs, the relative ranking of translation abilities between the two models could shift.

[3] https://github.com/Helsinki-NLP/Opus-MT

[4] https://huggingface.co/facebook/nllb-200-3.3B

[5] https://huggingface.co/google/mt5-small

Table 1: Facebook-NLLB Results on BLEU, ChrF, and COMET Metrics. O and G represent the original data and the data processed by ChatGPT, respectively.

	BLEU		ChrF		COMET	
	O	G	O	G	O	G
PT	52.76	68.41	74.35	83.12	90.71	92.31
FR	50.23	62.87	72.84	79.56	88.39	89.91
DE	37.42	50.89	65.90	74.23	86.35	87.98
ZH	34.78	40.34	33.14	36.71	84.19	85.73
IT	29.05	57.34	60.82	78.05	87.53	90.11
ES	28.76	66.09	58.97	82.33	86.91	91.24
RU	26.83	38.72	56.14	64.91	86.12	87.83
AR	23.12	26.47	57.03	60.42	85.24	86.14
HI	14.98	17.21	38.45	41.82	62.31	63.18
JA	3.67	3.89	8.21	8.67	64.15	64.37
KO	2.10	2.35	3.56	3.84	29.34	29.48

Table 2: Google-T5 Results on BLEU, ChrF, and COMET Metrics.

	BLEU		ChrF		COMET	
	O	G	O	G	O	G
FR	44.15	55.32	69.45	76.78	85.49	87.01
DE	30.24	44.18	62.37	70.82	83.91	85.63

Table 3: Helsinki-NLP Results on BLEU, ChrF, and COMET Metrics.

	BLEU		ChrF		COMET	
	O	G	O	G	O	G
PT	47.94	63.31	71.48	80.68	88.93	90.45
FR	46.26	57.77	70.12	77.55	86.16	87.79
DE	33.84	46.78	63.66	71.62	84.70	86.37
ZH	30.55	35.26	29.52	32.90	82.40	83.91
IT	25.39	52.21	57.23	75.11	85.22	88.72
ES	24.87	61.83	54.80	79.96	85.03	89.49
RU	23.94	34.93	53.79	62.32	84.75	86.37
AR	21.00	24.13	54.04	56.97	83.19	84.04
HI	12.50	14.30	34.90	37.09	59.53	60.16
JA	2.18	2.36	6.33	6.71	62.61	62.87
KO	1.26	1.36	2.19	2.35	25.94	25.98

Finding 4: The impact of LLMs on the benchmark could not only inflate the translation scores across different languages but also distort the comparison of translation abilities between models, making it fail to truly reflect their translation effectiveness.

5.2 Indirect Impact 2: RAG

Overall. RAG provides reliable and up-to-date external knowledge (Gao et al., 2023) to mitigate

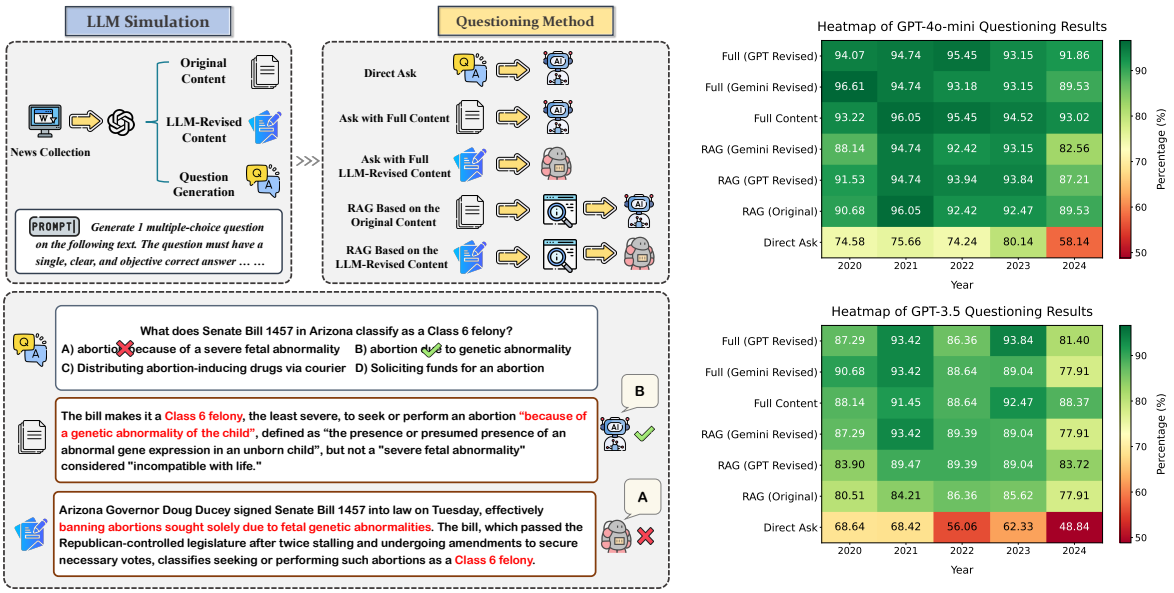

Figure 6: The procedure and results of the RAG experiment.

hallucination in LLM generation. Wikipedia is one of the most commonly applied general retrieval sets in previous RAG work, which stores factual structured information in scale (Fan et al., 2024). Therefore, we are curious about how the effectiveness of RAG might change if Wikipedia pages are influenced by LLMs. This process and final results are illustrated in Figure 6. Below, we will outline the detailed step of our experiment.

5.2.1 Experiment

Automated Queries Synthesis. We used *GPT-4o-mini* to generate multiple-choice questions (MCQs) based on the extracted news. Our prompt is "*Generate 1 multiple-choice question based on the following text. The question must have a single, clear, and objective correct answer, and should not use phrases like 'Which of the following is correct?'. The following text is news, you should extract as much new content as possible to construct questions. Only include the question and four answer options and the correct answer. Text:*". This ensures that the generated queries contain the most recent and relevant aspects of each news article, increasing the likelihood that the event was not present in the LLMs' training data.

Knowledge Base Construction. We constructed the knowledge base using Wikinews articles from 2020–2024. Each article was preprocessed, split into smaller text segments, and vectorized using CLIP[6], a neural network trained on a variety of (image, text) pairs. We then indexed these vectors using FAISS, a library for efficient similarity search and clustering of dense vectors, for efficient retrieval (Douze et al., 2024).

Retrieval and Generation. Given a user's question, we vectorized the question using CLIP and conducted a similarity search in FAISS. The top three most relevant segments were retrieved and provided as context. These segments were then combined with the user's question and used in a prompting template to query the LLM. The model then selected the most likely answer based on both its prior knowledge and the retrieved content.

5.2.2 QA Approaches

We conducted experiments using *GPT-4o-mini* and *GPT-3.5* across different questioning methods:

- **Direct Questioning**: The model was asked the MCQs directly without any external context or retrieval support.
- **RAG Based on the Original Texts**: The model was provided with retrieved content from an external knowledge base based on the original news.
- **RAG Based on the Revised Texts**: The model was provided with retrieved content from an external knowledge base based on the LLM-revised news.
- **Human-Assisted Questioning**: We manually provided the original or LLM-revised news corresponding to each question, simulating an upper-bound scenario for the RAG framework.

[6] https://github.com/openai/CLIP

5.2.3 Results

Declining Accuracy for Recent Events. In the absence of RAG, both models exhibited significantly lower accuracy when answering questions derived from recent Wikinews articles (e.g., *GPT-4o-mini*: 58.14% in 2024, *GPT-3.5*: 48.84% in 2024), while their accuracy is much better for older events (e.g., 2020–2022). The reason is also straightforward: these news events are not included in their training data.

Higher Accuracy of Knowledge Base. Providing external knowledge greatly improved performance. With the help of knowledge base, LLMs could achieve a minimum accuracy of 77.91% (*GPT-3.5*, 2024) and often exceed 90%. This confirms the effectiveness of RAG in enhancing factual accuracy.

Impact of LLM-Revised Content. Whether using RAG or providing the complete news directly, the accuracy achieved with LLM-revised texts is lower than when using the original texts. For example, in 2024, the accuracy dropped from 93.02% when asking with the full original content to 89.53% with the full Gemini-revised texts and 91.86% with the full GPT-revised texts.

Maximal Performance with Full Content Providing the full news resulted in the highest accuracy, demonstrating the limitations of retrieval-based approaches in selecting the most relevant information. In most cases, the full content approach exceeded 93% accuracy, setting a benchmark for ideal retrieval performance.

5.2.4 Case Study

To explore the impact of LLM-generated texts, we focus on cases where RAG answers correctly with the original knowledge base but fails when using LLM-revised texts. Figure 6 includes one example. Interestingly, although both the original and revised text explicitly exclude "severe fetal abnormalities", the revised text, which changed "genetic abnormality" to "fetal genetic abnormalities", which led LLMs to misinterpret the information. As a result, LLMs mistakenly selected A based on the revised text. Appendix E.2 provides more similar examples, and LLM-generated texts may decrease accuracy in RAG tasks for several reasons:

- **Introduction of Modifiers**: Adding adjectives or modifiers can change the context and impact the text's accuracy. See Question 1 for examples.
- **Keyword Replacement**: LLM might replace key terms, altering the original meaning and causing misinterpretation. See Question 2 and 3 for examples.
- **Retrieval Mismatch**: The revised text may reduce the similarity between the question and the correct article, or increase similarity with irrelevant ones. Sometimes, even with minimal changes to the article, it still fails to match. See Questions 4 and 5 for examples.

> ***Finding 5:*** The results suggest that LLM-generated content performs less effectively in RAG systems compared to human-created texts. If such content has impacted high-quality communities like Wikipedia, it raises concerns about the potential decline in information quality in knowledge bases.

6 Discussion and Conclusion

The relationship between Wikipedia and LLMs is both collaboration and competition.

First, Wikipedia's success is inherently tied to its extensive base of human contributors (Kittur and Kraut, 2008). Wikipedia articles play a crucial role as training data for LLMs, and their development would not be possible without this rich resource. At the same time, researchers have employed NLP techniques, including LLMs, to enhance Wikipedia's quality (Lucie-Aimée et al., 2024).

Second, our findings suggest that Wikipedia is being influenced by LLMs, and this impact is likely to grow over time. The dynamic between humans and AI continuously shapes each other, becoming an integral aspect of contemporary society (Pedreschi et al., 2024).

Third, the influence of LLMs on corpora like Wikipedia and Wikinews also extends to machine translation benchmarks. As a result, machine translation targets may gradually align more closely with the language patterns of LLMs, though these shifts are incremental.

Lastly, our findings show that using LLM-revised Wikipedia pages for RAG can reduce accuracy, indicating that LLM-based refinements may sometimes overcorrect. Though relying on LLMs to improve Wikipedia or similar knowledge systems poses risks, it is also worth noting that purely human-driven processing has its limitations.

Limitations

Although we conducted several experiments to evaluate the impact of LLMs on Wikipedia, our study has certain limitations. First, Wikipedia pages follow a specific format, making it challenging to extract completely plain text. This formatting issue in our dataset may introduce some errors in the quantitative analysis of LLM impact. Second, when assessing the readability of Wikipedia pages, we relied only on traditional metrics based on formulas, such as the Flesch-Kincaid score. However, recent advancements in NLP have shifted towards more sophisticated computational models (François, 2015). Lastly, in the RAG task, our Wikinews dataset is not big enough compared to the Wikipedia pages dataset, which may limit the generalization of our findings.

References

Sayantan Adak, Pauras Mangesh Meher, Paramita Das, and Animesh Mukherjee. 2025. Reversum: A multi-staged retrieval-augmented generation method to enhance wikipedia tail biographies through personal narratives. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 732–750.

Mohammad Awad AlAfnan and Siti Fatimah MohdZuki. 2023. Do artificial intelligence chatbots have a writing style? an investigation into the stylistic features of chatgpt-4. *Journal of Artificial intelligence and technology*, 3(3):85–94.

Hélder Antunes and Carla Teixeira Lopes. 2019. Analyzing the adequacy of readability indicators to a non-english language. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pages 149–155. Springer.

Creston Brooks, Samuel Eggert, and Denis Peskoff. 2024. The rise of ai-generated content in wikipedia. *arXiv preprint arXiv:2410.08044*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jad Doughman, Osama Mohammed Afzal, Hawau Olamide Toyin, Shady Shehata, Preslav Nakov, and Zeerak Talat. 2024. Exploring the limitations of detecting machine-generated text. *arXiv preprint arXiv:2406.11073*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Derar Eleyan, Abed Othman, and Amna Eleyan. 2020. Enhancing software comments readability using flesch reading ease score. *Information*, 11(9):430.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Thomas François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, 20(2):79–97.

Evgeniy Gabrilovich and Shaul Markovitch. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34:443–498.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Mingmeng Geng, Caixi Chen, Yanru Wu, Dongping Chen, Yao Wan, and Pan Zhou. 2024. The impact of large language models in academia: from writing to speaking. *arXiv preprint arXiv:2409.13686*.

Mingmeng Geng and Roberto Trotta. 2024. Is chatgpt transforming academics' writing style? *arXiv preprint arXiv:2404.08627*.

Jim Giles. 2005. Special report internet encyclopaedias go head to head. *nature*, 438(15):900–901.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *arXiv preprint arXiv:2406.13805*.

Benedetta Iavarone, Dominique Brunato, Felice Dell'Orletta, et al. 2021. Sentence complexity in context. In *CMCL 2021-Workshop on Cognitive Modeling and Computational Linguistics, Proceedings*, pages 186–199. Association for Computational Linguistics (ACL).

Isaac Johnson, Guosheng Feng, Robert West, et al. 2024a. Edisum: Summarizing and explaining wikipedia edits at scale. *arXiv preprint arXiv:2404.03428*.

Isaac Johnson, Lucie-Aimée Kaffee, and Miriam Redi. 2024b. Wikimedia data for ai: a review of wikimedia datasets for nlp tasks and ai-assisted editing. *arXiv preprint arXiv:2410.08918*.

Aniket Kittur and Robert E Kraut. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46.

Kayvan Kousha and Mike Thelwall. 2017. Are wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology*, 68(3):762–779.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*.

Lucie Lucie-Aimée, Angela Fan, Tajuddeen Gwadabe, Isaac Johnson, Fabio Petroni, and Daniel Van Strien. 2024. Proceedings of the first workshop on advancing natural language processing for wikipedia. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*.

Connor McMahon, Isaac Johnson, and Brent Hecht. 2017. The substantial interdependence of wikipedia and google: A case study on the relationship between peer production communities and information technologies. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 142–151.

Manish P Mehta, Hasani W Swindell, Robert W Westermann, James T Rosneck, and T Sean Lynch. 2018. Assessing the readability of online information about hip arthroscopy. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 34(7):2142–2149.

Rada Mihalcea and Andras Csomai. 2007. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242.

Nandana Mihindukulasooriya, Sanju Tiwari, Daniil Dobriy, Finn Årup Nielsen, Tek Raj Chhetri, and Axel Polleres. 2024. Scholarly wikidata: Population and exploration of conference data in wikidata using llms. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 243–259. Springer.

Pedro Miguel Moás and Carla Teixeira Lopes. 2023. Automatic quality assessment of wikipedia articles—a systematic literature review. *ACM Computing Surveys*, 56(4):1–37.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Priti P Patel, Ian C Hoppe, Naveen K Ahuja, and Frank S Ciminello. 2011. Analysis of comprehensibility of patient information regarding complex craniofacial conditions. *Journal of Craniofacial Surgery*, 22(4):1179–1182.

Dino Pedreschi, Luca Pappalardo, Emanuele Ferragina, Ricardo Baeza-Yates, Albert-László Barabási, Frank Dignum, Virginia Dignum, Tina Eliassi-Rad, Fosca Giannotti, János Kertész, et al. 2024. Human-ai coevolution. *Artificial Intelligence*, page 104244.

Yiwen Peng, Thomas Bonald, and Mehwish Alam. 2024. Refining wikidata taxonomy using large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5395–5399.

Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. 2021. On the value of wikipedia as a gateway to the web. In *Proceedings of the Web Conference 2021*, pages 249–260.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Neal Reeves, Wenjie Yin, Elena Simperl, and Miriam Redi. 2024. " the death of wikipedia?"–exploring the impact of chatgpt on wikipedia engagement. *arXiv preprint arXiv:2405.10205*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Sina J Semnani, Violet Z Yao, Heidi C Zhang, and Monica S Lam. 2023. Wikichat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia. *arXiv preprint arXiv:2305.14292*.

Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*.

Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why we read wikipedia. In *Proceedings of the 26th international conference on world wide web*, pages 1591–1600.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*.

Marina Solnyshkina, Radif Zamaletdinov, Ludmila Gorodetskaya, and Azat Gabitov. 2017. Evaluating text complexity and flesch-kincaid grade level. *Journal of social studies education research*, 8(3):238–248.

Michael Strube and Simone Paolo Ponzetto. 2006. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, volume 6, pages 1419–1424.

Damian Świeczkowski and Sławomir Kułacz. 2021. The use of the gunning fog index to evaluate the readability of polish and english drug leaflets in the context of health literacy challenges in medical linguistics: An exploratory study. *Cardiology Journal*, 28(4):627–631.

Neil Thompson and Douglas Hanley. 2018. Science is shaped by wikipedia: evidence from a randomized control trial. *SSRN*.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Nicholas Vincent, Isaac Johnson, and Brent Hecht. 2018. Examining wikipedia with a broader lens: Quantifying the value of wikipedia's relationships with other large-scale online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Christian Wagner and Ling Jiang. 2025. Death by ai: Will large language models diminish wikipedia? *Journal of the Association for Information Science and Technology*.

Yuxia Wang, Artem Shelmanov, Jonibek Mansurov, Akim Tsvigun, Vladislav Mikhailov, Rui Xing, Zhuohan Xie, Jiahui Geng, Giovanni Puccetti, Ekaterina Artemova, et al. 2025. Genai content detection task 1: English and multilingual machine-generated text detection: Ai vs. human. *arXiv preprint arXiv:2501.11012*.

Anna Wróblewska, Marceli Korbin, Yoed N Kenett, Daniel Dan, Maria Ganzha, and Marcin Paprzycki. 2025. Applying text mining to analyze human question asking in creativity research. *arXiv preprint arXiv:2501.02090*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Qiyuan Yang, Pengda Wang, Luke D Plonsky, Frederick L Oswald, and Hanjie Chen. 2024. From babbling to fluency: Evaluating the evolution of language models in terms of human language acquisition. *arXiv preprint arXiv:2410.13259*.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *LREC*, volume 8, pages 1646–1652.

Jiebin Zhang, J Yu Eugene, Qinyu Chen, Chenhao Xiong, Dawei Zhu, Han Qian, Mingbo Song, Weimin Xiong, Xiaoguang Li, Qun Liu, et al. 2025. Wikigenbench: Exploring full-length wikipedia generation under real-world scenario. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5191–5210.

Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. 2024. LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436, Mexico City, Mexico. Association for Computational Linguistics.

A Data Collection and Processing

The detailed classification poses a problem in our data crawling process: When iteratively querying deeper subcategories without limit, the retrieved pages may become less relevant to the original topic (i.e., the root category). To address this issue, we selected an appropriate crawl depth for each category to balance the number of pages with their topical relevance, as shown in Table 4.

Category	Art	Bio	Chem	CS	Math	Philo	Phy	Sports
Crawl Depth	4	4	5	5	5	5	5	4
Number of Pages	57,028	44,617	53,282	59,097	47,004	33,596	40,986	53,900

Table 4: Number of Wikipedia articles crawled per category.

We also excluded redirect pages, as they do not contain independent content but rather link to other target pages. After crawling the pages, we cleaned the data by extracting the plain text and removing irrelevant sections such as "References," "See also," "Further reading," "External links," "Notes," and "Footnotes." To minimize the impact of topic-specific words, only those ranked within the top 10,000 in the Google Ngram dataset[7] were included in the calculations.

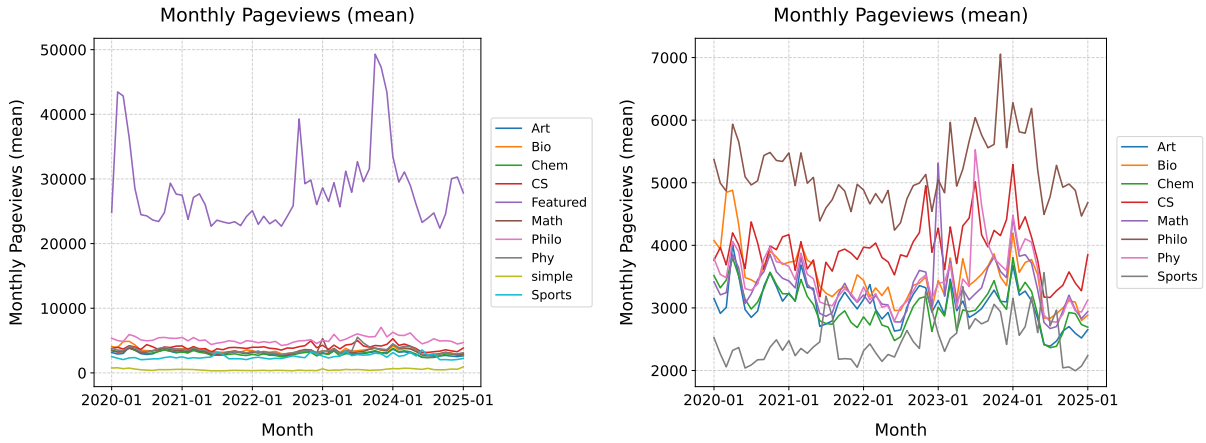

Figure 7: Page views across different categories

[7]Google Ngram dataset: https://www.kaggle.com/datasets/wheelercode/english-word-frequency-list

B LLM Impact

B.1 Word frequency

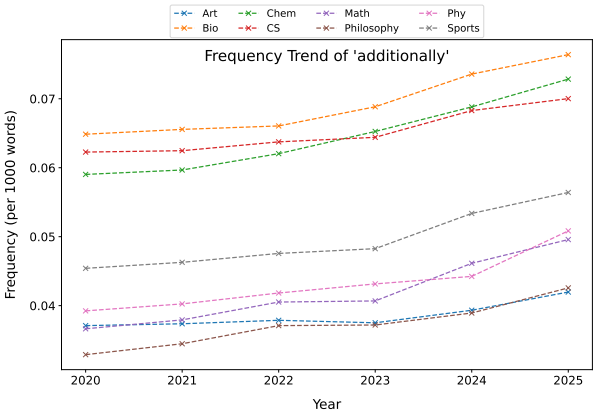

Figure 8: More examples on word frequency.

B.2 LLM simulations

We used GPT to revise the January 1, 2022, versions of Featured Articles to construct word frequency data reflecting the impact of large language models. This choice was based on the assumption that Featured Articles are less likely to be affected by LLMs, given their rigorous review processes and ongoing manual maintenance. To reduce errors caused by incomplete data cleaning, we extracted only the first section of each Featured Article for revision. It is important to note that some responses were filtered due to the prompt triggering Azure OpenAI's content moderation policy, likely because certain Wikipedia pages contained violent content. As a result, these pages were excluded from the analysis.

Choosing the appropriate word combinations to estimate the impact of LLMs is essential. On one hand, by setting a threshold for f^*, we ensure that the target vocabulary has a high frequency of occurrence in the corpus. On the other hand, by setting a threshold for $\hat{r}$, we ensure that these words show a significant frequency change before and after being processed by the LLM. For the f^* threshold, we propose two strategies: First, the target vocabulary must frequently appear in the first section of Featured Articles, as we use this part of the articles for LLM refinement when estimating $\hat{r}$; second, the target vocabulary must frequently appear in the target corpus. For the first strategy, when calculating the impact of the LLM on different pages, the selected vocabulary combination remains the same. For the second strategies, the influence on pages of different categories will be estimated using the vocabulary combination corresponding to each category.

B.3 Featured Articles + Same Word Combinations

- $\frac{1}{f^*}$: 5000, 7000, 9000, 11000, 13000, 15000
- $\hat{r}$: 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.20, 0.21 (corresponding values of $\frac{\hat{r}+1}{\hat{r}^2}$)

B.4 Featured Articles + Different Word Combinations

- $\frac{1}{f^*}$: 5000, 7000, 9000, 11000, 13000, 15000
- $\hat{r}$: 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.20, 0.21 (corresponding values of $\frac{\hat{r}+1}{\hat{r}^2}$)

B.5 Simple Articles + Same Word Combinations

- $\frac{1}{f^*}$: 1000, 3000, 5000, 7000, 9000, 11000, 13000
- $\hat{r}$: 0.07, 0.09, 0.11, 0.13, 0.15, 0.17, 0.19, 0.21, 0.23, 0.25, 0.27, 0.29 (corresponding values of $\frac{\hat{r}+1}{\hat{r}^2}$)

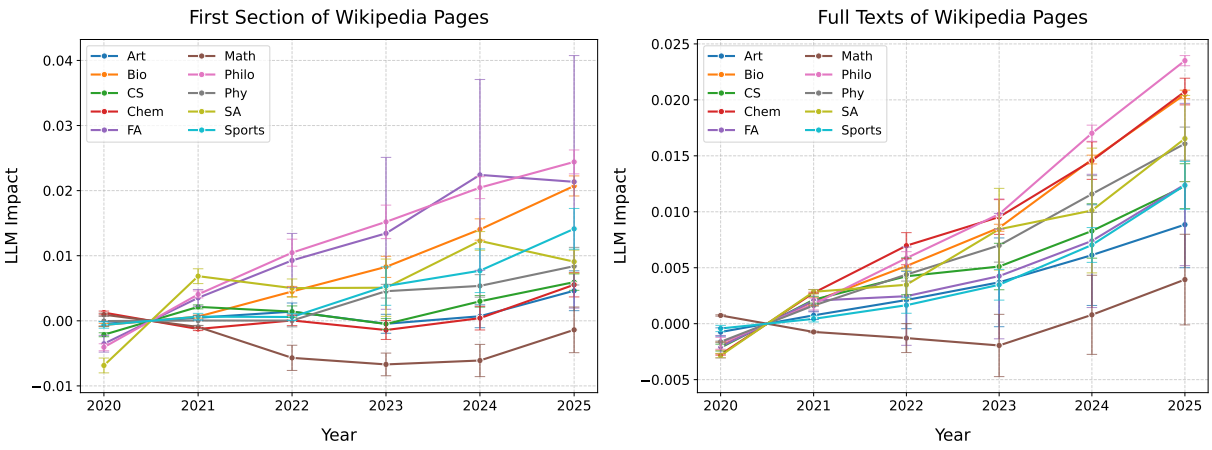

Figure 9: LLM Impact: Estimated based on simulations of the first section of Featured Articles, using the same word combinations across different categories of Wikipedia pages

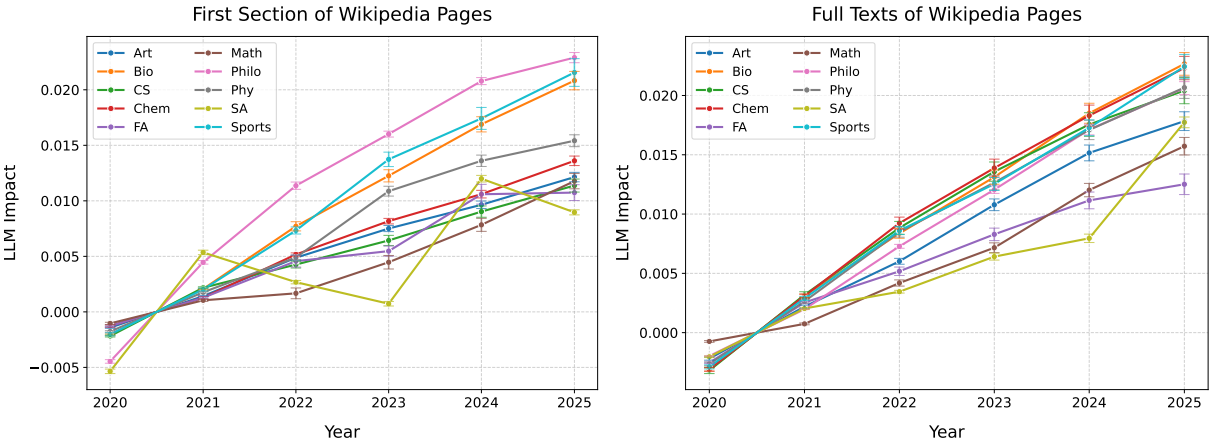

Figure 10: LLM Impact: Estimated based on simulations of the first section of Simple Articles, using the same word combinations across different categories of Wikipedia pages

B.6 Simple Articles + Different Word Combinations

- $\frac{1}{f^*}$: 2000, 2500, 3000, 3500, 4000, 4500, 5000
- $\hat{r}$: 0.11, 0.13, 0.15, 0.17, 0.19, 0.21, 0.23, 0.25 (corresponding values of $\frac{\hat{r}+1}{\hat{r}^2}$)

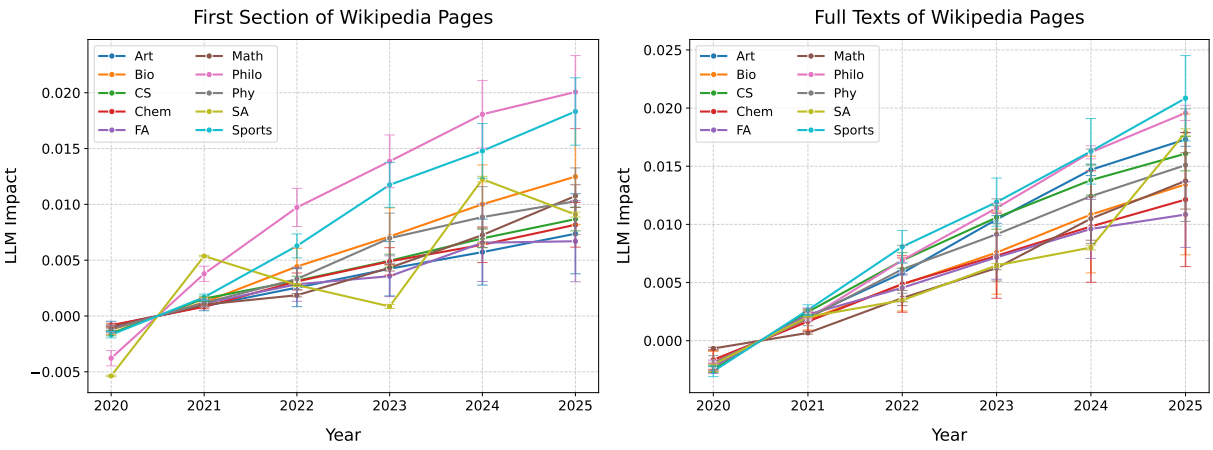

Figure 11: LLM Impact: Estimated based on simulations of the first section of Simple Articles, using different word combinations across different categories of Wikipedia pages

C Linguistic Style

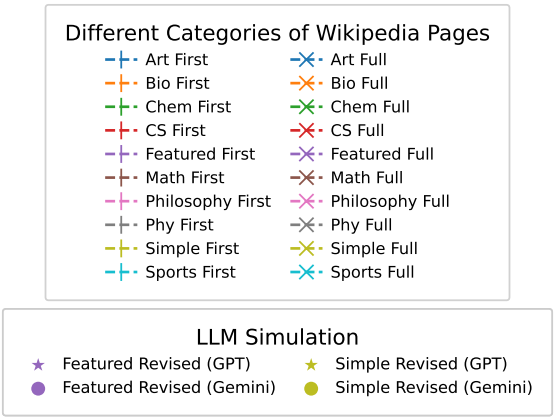

Figure 12: Legend

C.1 Word Level

C.2 Sentence Level

C.3 Readability

D Machine Translation

D.1 Exception Handling

Some API calls returned an openai.BadRequestError with error code 400, indicating that Azure OpenAI's content management policies flagged the prompts for potentially violating content. Also, Some translations returned null values. These cases were excluded from scoring and ignored in the evaluation.

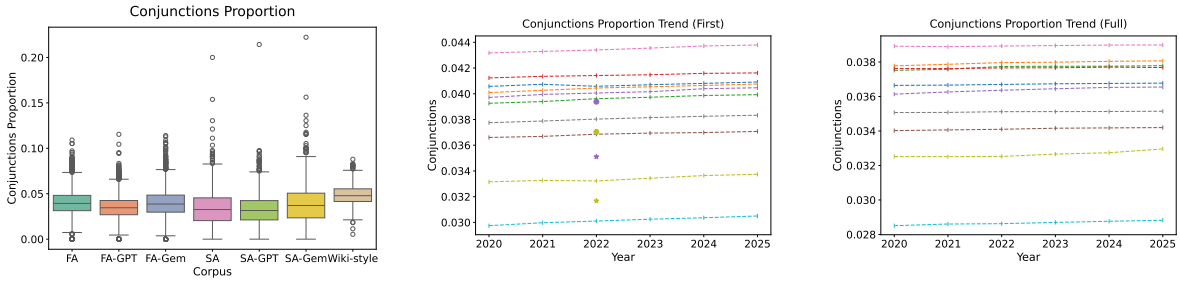

Figure 13: Comparison of Conjunction Ratios Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

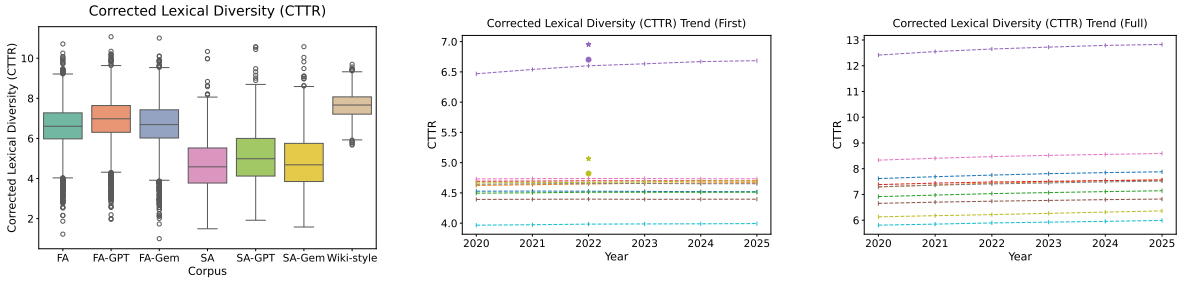

Figure 14: Comparison of CTTR Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

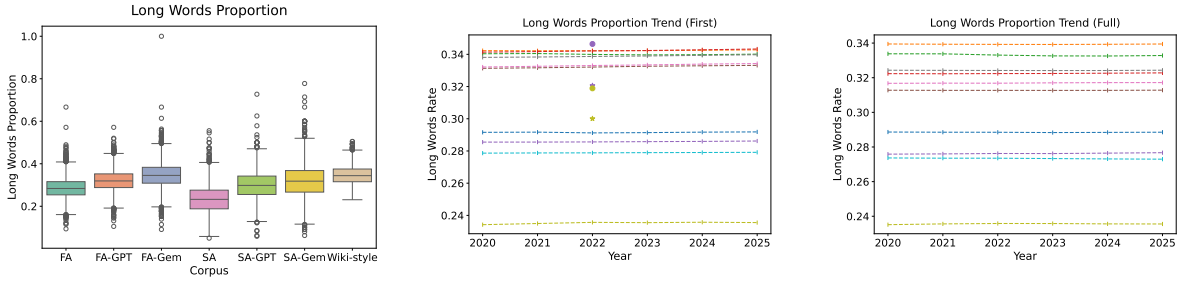

Figure 15: Comparison of Long Words Rate Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

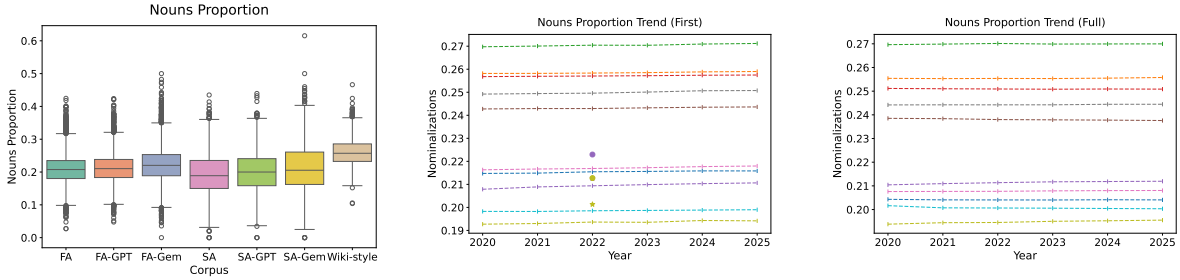

Figure 16: Comparison of Nouns Frequency Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

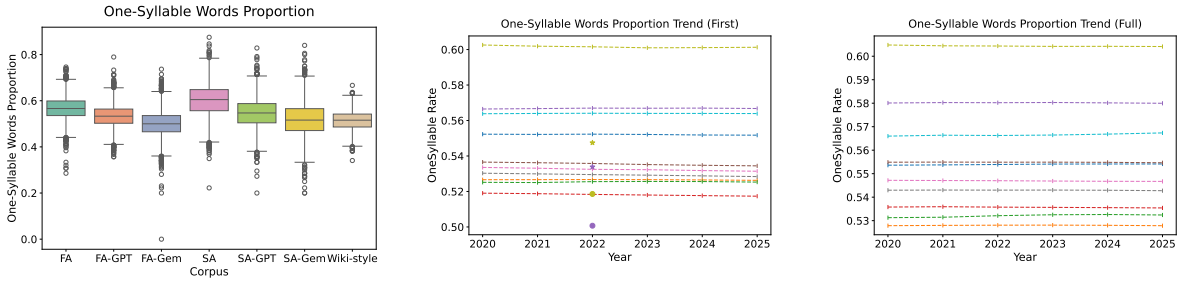

Figure 17: Comparison of One-Syllable Word Ratio Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

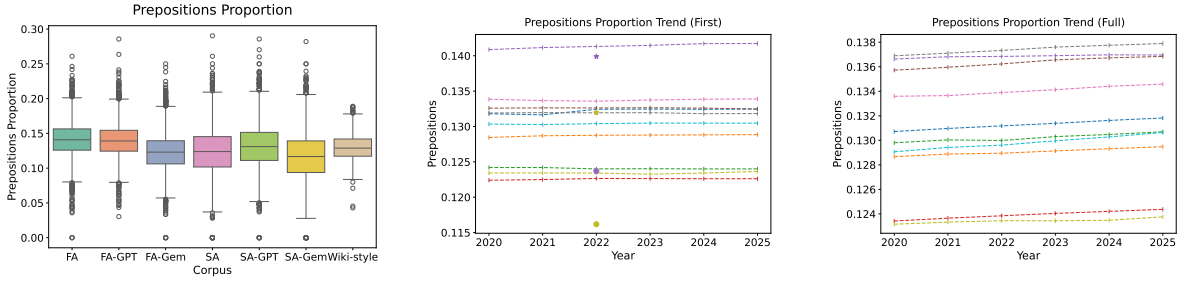

Figure 18: Comparison of Preposition Frequency Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

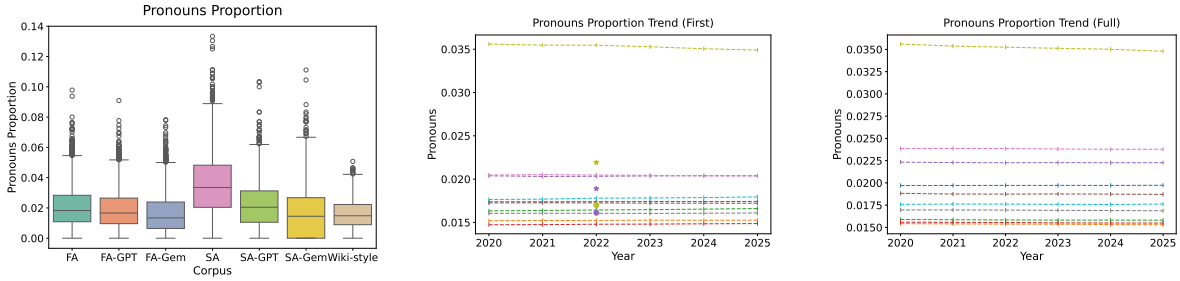

Figure 19: Comparison of Pronoun Frequency Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

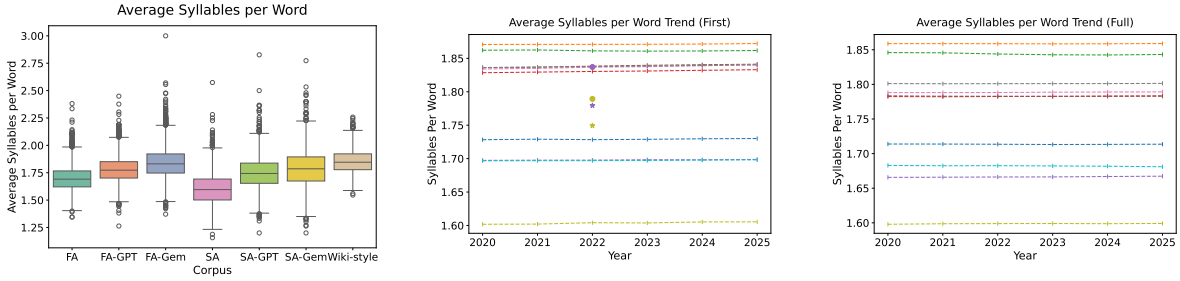

Figure 20: Comparison of Average Syllables per Word Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

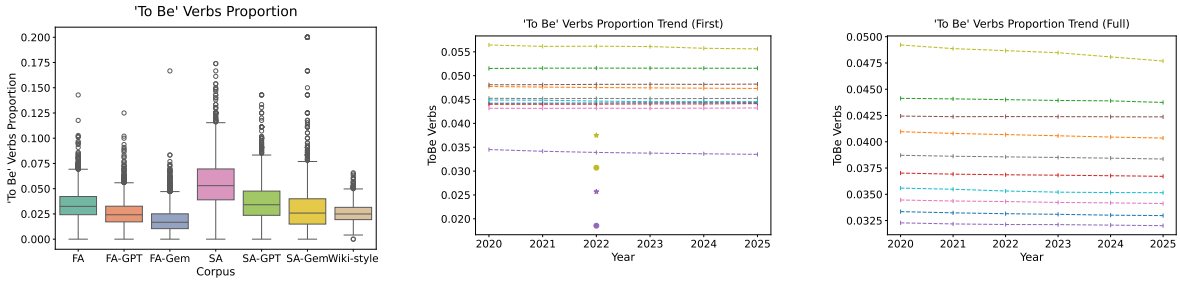

Figure 21: Comparison of To-Be Verb Ratios Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

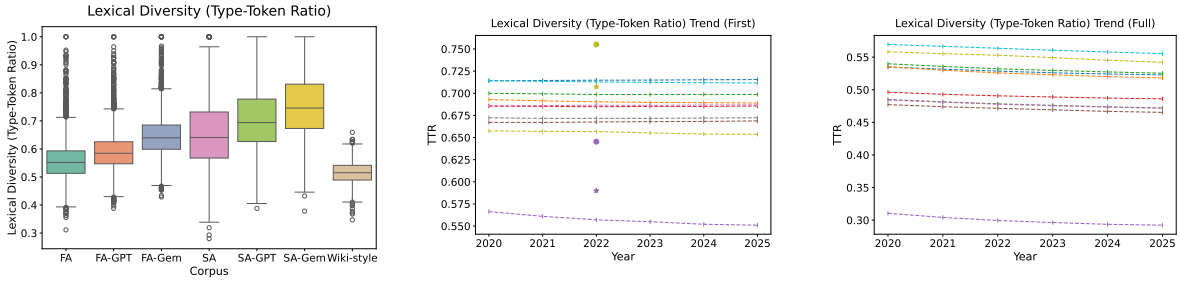

Figure 22: Comparison of TTR Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

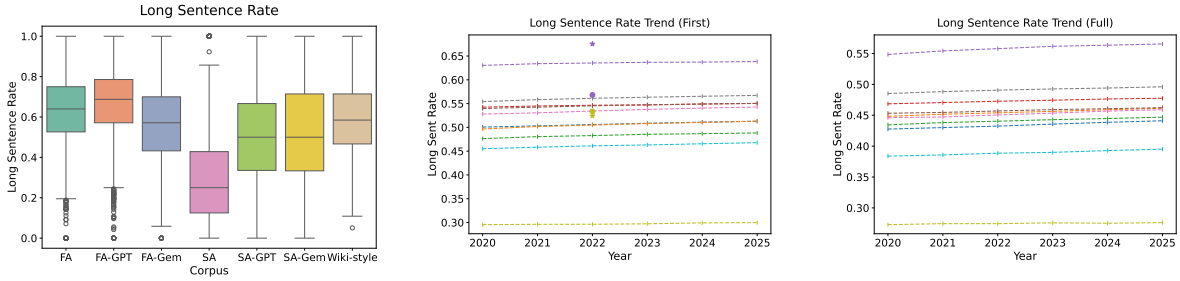

Figure 23: Comparison of Long Sentence Rate Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

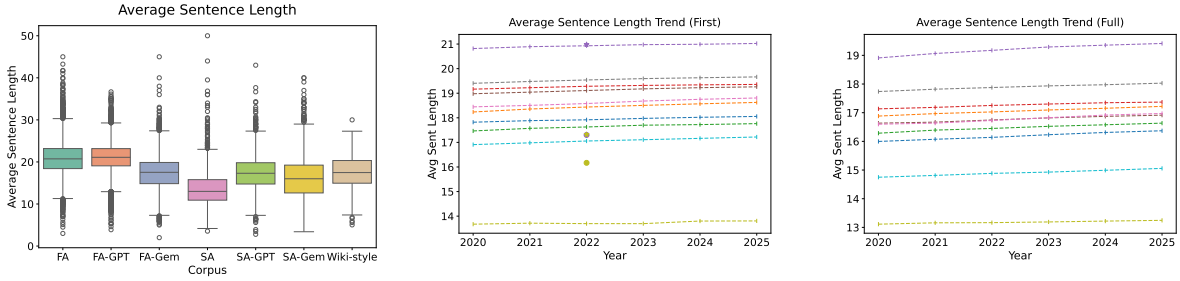

Figure 24: Comparison of Average Sentence Length Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

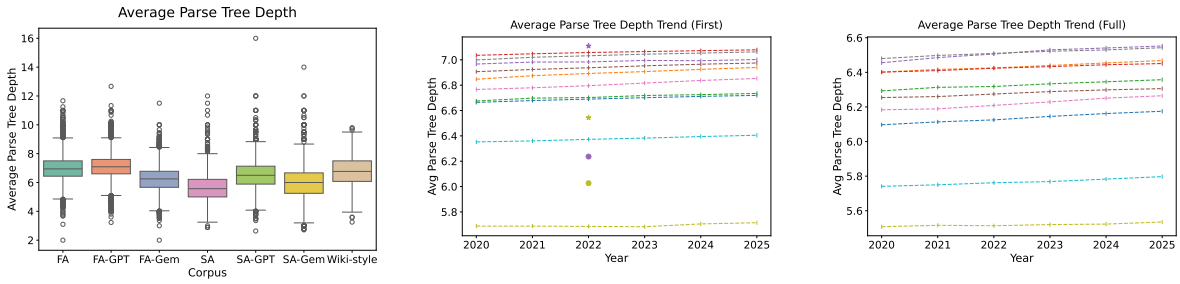

Figure 25: Comparison of Average Parse Tree Depth Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

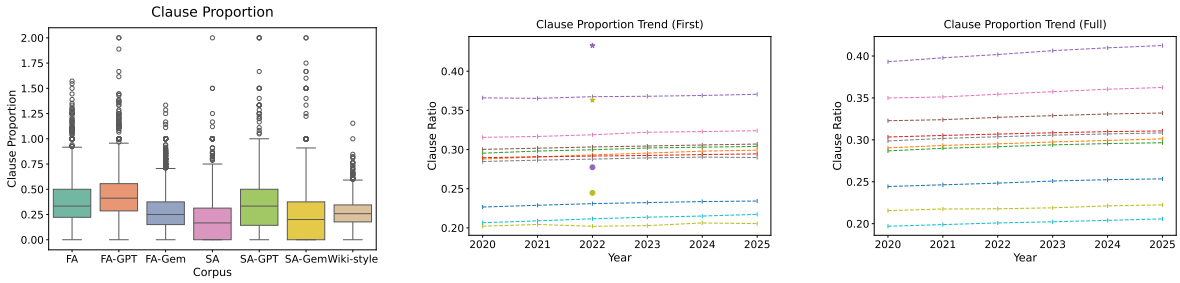

Figure 26: Comparison of Clause Ratio Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

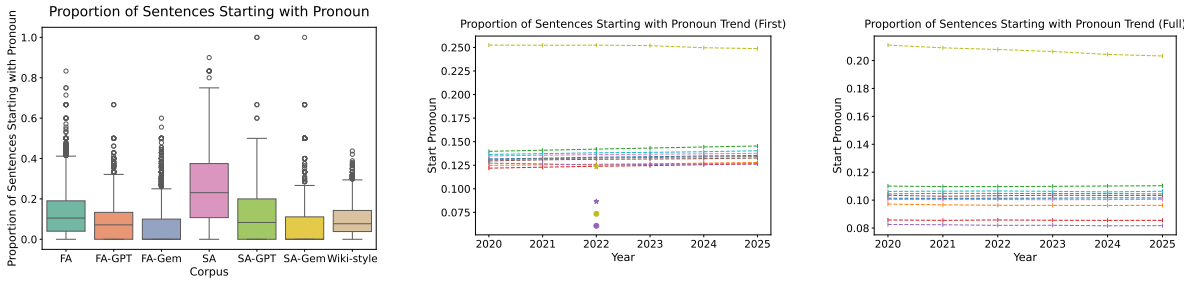

Figure 27: Comparison of Pronoun-Initial Sentence Ratio Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

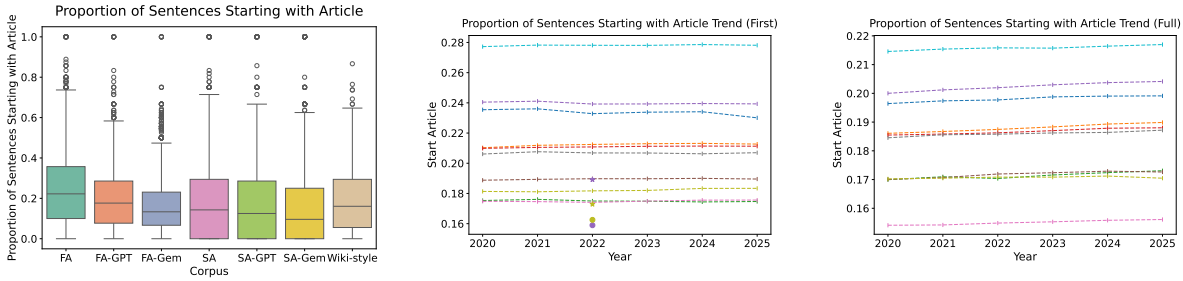

Figure 28: Comparison of Article-Initial Sentence Ratio Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

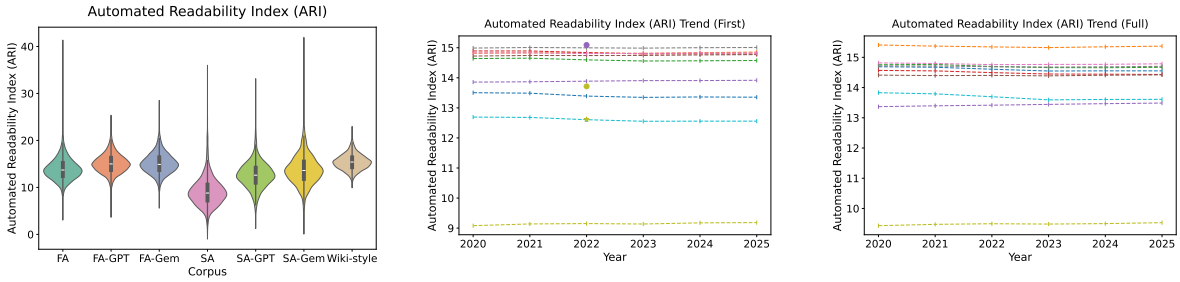

Figure 29: Comparison of ARI Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

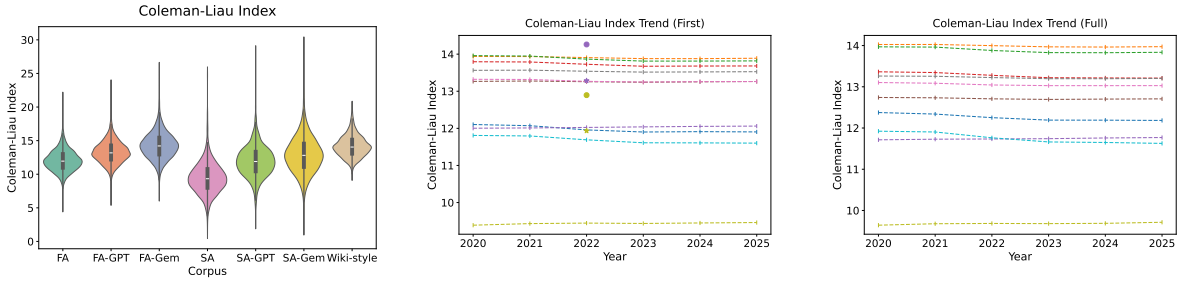

Figure 30: Comparison of Coleman-Liau Index Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

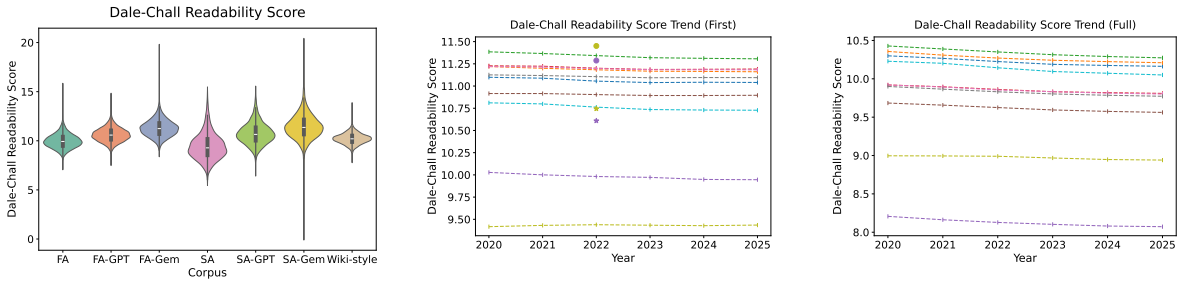

Figure 31: Comparison of Dale-Chall Score Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

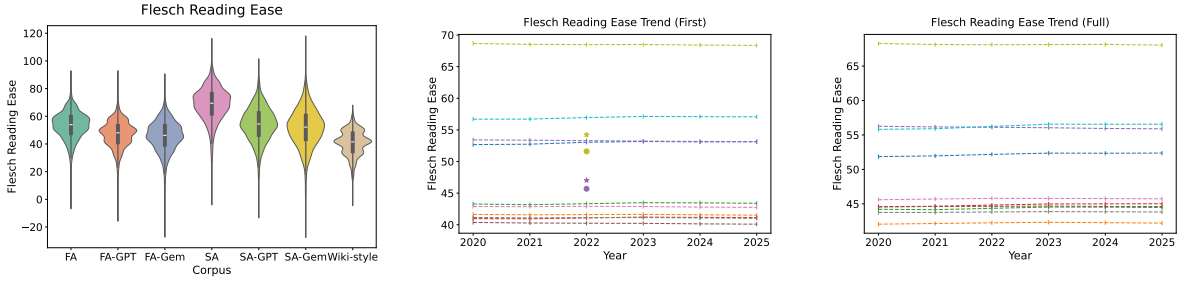

Figure 32: Comparison of Flesch Reading Ease Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

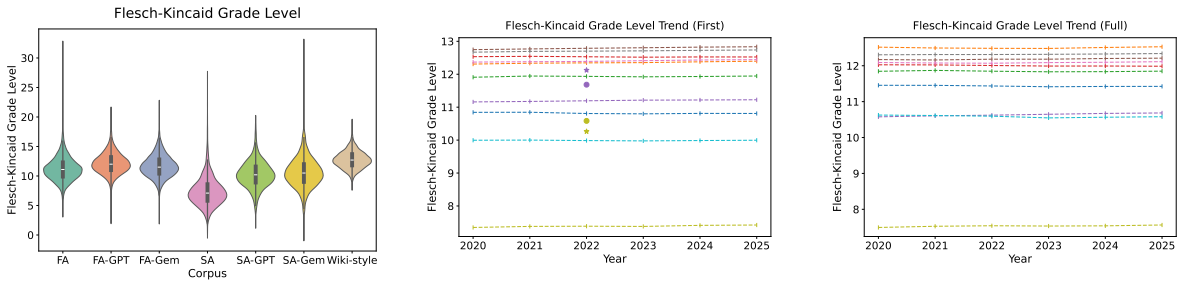

Figure 33: Comparison of Flesch-Kincaid Grade Level Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

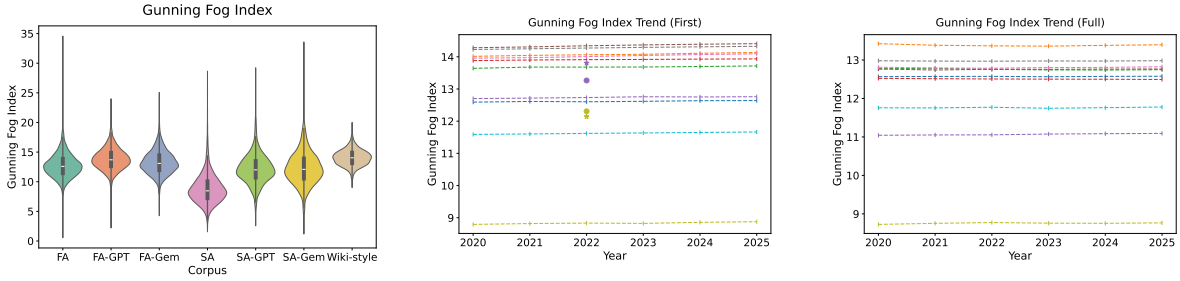

Figure 34: Comparison of Gunning Fog Index Before and After Revisions by Different LLMs in Featured and Simple Articles, and Their Temporal Trends Across Full Text and First Sections of Different Wikipedia Categories.

D.2 Languages

These are the 12 languages in our benchmarks:

- English (eng-Latn-stan1293)
- Modern Standard Arabic (arb-Arab-stan1318)
- Mandarin (cmn-Hans-beij1234)
- German (deu-Latn-stan1295)
- French (fra-Latn-stan1290)
- Hindi (hin-Deva-hind1269)
- Italian (ita-Latn-ital1282)
- Japanese (jpn-Jpan-nucl1643)
- Korean (kor-Hang-kore1280)
- Brazilian Portuguese (por-Latn-braz1246)
- Russian (rus-Cyrl-russ1263)
- Latin American Spanish (spa-Latn-amer1254)

E RAG

E.1 RAG Results

Year	Direct Ask	Ask with KB	GPT Revised	Gemini Revised	Full Content	Full (Gemini)	Full (GPT)
2020	74.58%	91.10%	91.53%	88.98%	93.64%	97.03%	94.49%
2021	75.66%	96.38%	95.07%	95.07%	96.38%	94.90%	95.07%
2022	74.24%	93.37%	94.70%	93.18%	95.83%	93.56%	95.83%
2023	80.14%	93.32%	94.69%	94.01%	95.38%	94.01%	94.01%
2024	58.14%	90.40%	88.37%	84.01%	93.90%	90.41%	92.73%

Table 5: Performance on RAG Task (4O-mini Output with Null as 0.25)

E.2 Case Study

Below are examples where the questions were answered correctly using RAG with the original knowledge base, but incorrectly when the LLM-revised texts were used as the knowledge base:

Question 2 What activity was the New Zealand Navy ship conducting before it ran aground near Samoa? A) Naval training exercises B) Aquatic terrain survey C) Humanitarian aid delivery D) Equipment testing

Original Text ...New Zealand Navy ship ran aground one from the shore of Samoa on Saturday evening. She was reportedly conducting an aquatic terrain survey. ...

LLM Revised Version ... A New Zealand Navy ship conducting a hydrographic survey ran aground approximately one kilometer from the Samoan coast on Saturday evening ...

Analysis The original text clearly gives "aquatic terrain survey" as the content of the ship's activities, which clearly points to option B. The revised version changes "aquatic terrain survey" to "hydrographic survey", which may lead to a change in the understanding of the nature of the activity, thus leading LLM to choose the wrong option D.

Question 3 What was the initial evidence regarding the cause of the explosions in southern Beirut? A) Cyber attack B) Faulty batteries C) Hardware tampering D) Natural disaster

Original Text ...The cause of the explosions was still under investigation. Early evidence suggested the pager explosions were triggered by explosives planted in the pagers, or faulty batteries, Reuters reported. The following day, walkie-talkies, laptops, and radios also exploded, killing 20 people and injuring 450 ...

LLM Revised Version ...While the cause remains under investigation, early evidence, reported by Reuters, indicated that explosives, possibly alongside metal balls to enhance impact, were planted in pagers. Further explosions involving walkie-talkies, laptops, and radios occurred the following day, resulting in additional casualties ...

Analysis The original RAG correctly selects Option B because it explicitly mentions faulty batteries as a possible cause of the explosions, along with explosives. However, the revised version focuses more on explosives and omits the mention of faulty batteries, which shifts the context toward hardware tampering, leading to an incorrect selection.

Question 4 What was the final score between Bugs and Maroochydore in their match on Saturday? A) 60 to 5 B) 29 to 10 C) 29 to 14 D) 10 to 3

Year	2020	2021	2022	2023	2024
Num of Question	118	152	132	246	86

Original Text On Saturday, Bugs defeated Maroochydore 29 points to 10 in round four of Australia's Sunshine Coast Rugby Union. . . .

LLM Revised Version Wynnum "Bugs" dominated Maroochydore 29-10 in Round Four of the Sunshine Coast Rugby Union on Saturday. . . .

Analysis Retrieval System failed to extract the correct article from the knowledge base (not in top 3 answers), thus choosing the wrong answer.

Question 5 What date is the general election in the United Kingdom scheduled for, as announced by Prime Minister Rishi Sunak?
A) 4th of June B) 4th of July C) 4th of August D) 4th of October

Original Text On Wednesday, British Rishi Sunak called a general election in the United Kingdom, which is set to take place on the 4th of July. . . .

LLM Revised Version British Prime Minister Rishi Sunak called a snap general election on Wednesday, to be held on July 4th. . . .

Analysis LLMs failed to extract the correct article from the knowledge base, thus choosing the wrong answer.