Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models

Anonymous ACL submission

Abstract

We study how to apply large language models to write grounded and organized long-form articles from scratch, with comparable breadth and depth to Wikipedia pages. This underexplored problem poses new challenges at the *pre-writing* stage, including how to research the topic and prepare an outline prior to writing. We propose **STORM**, a writing system for the **S**ynthesis of **T**opic **O**utlines through **R**etrieval and **M**ulti-perspective Question Asking. STORM models the pre-writing stage by (1) discovering diverse perspectives in researching the given topic, (2) simulating conversations where writers carrying different perspectives pose questions to a topic expert grounded on trusted Internet sources, (3) curating the collected information to create an outline.

For evaluation, we curate FreshWiki, a dataset of recent high-quality Wikipedia articles, and formulate outline assessments to evaluate the pre-writing stage. We further gather feedback from experienced Wikipedia editors. Compared to articles generated by an outline-driven retrieval-augmented baseline, more of STORM's articles are deemed to be organized (by a 25% absolute increase) and broad in coverage (by 10%). The expert feedback also helps identify new challenges for generating grounded long articles, such as source bias transfer and over-association of unrelated facts.

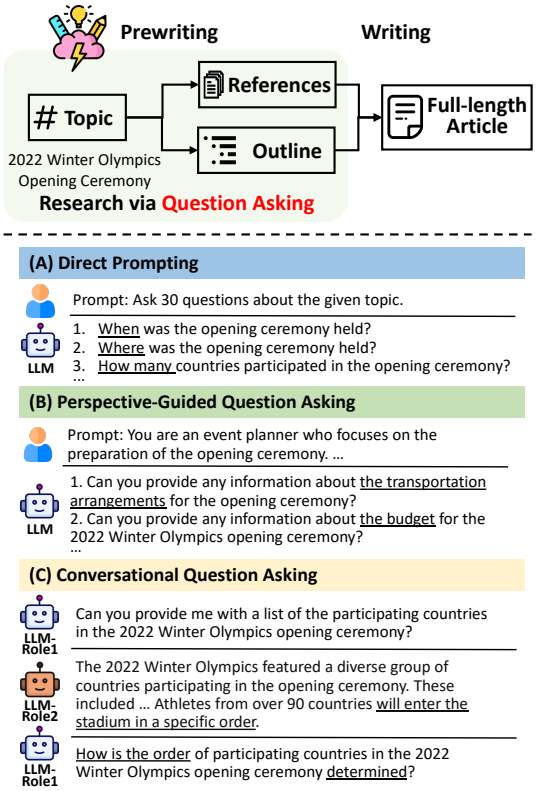

Figure 1: We explore writing Wikipedia-like articles from scratch, which demands a pre-writing stage before producing the article. In this stage, simpler approaches like Direct Prompting have limited planning capacity. In contrast, STORM researches the topic via perspective-guided question asking in simulated conversations.

1 Introduction

Large language models (LLMs) have demonstrated impressive writing capabilities (Yang et al., 2023; Pavlik, 2023; Wenzlaff and Spaeth, 2022; Fitria, 2023), but it is unclear how we can use them to write grounded, long-form articles, like full-length Wikipedia pages. Such expository writing, which seeks to inform the reader on a topic in an organized manner (Weaver III and Kintsch, 1991; Balepur et al., 2023), requires thorough research and planning in the *pre-writing* stage (Rohman, 1965), even before the actual writing process can start. However, prior work on generating Wikipedia articles (Banerjee and Mitra, 2015; Minguillón et al., 2017; Liu et al., 2018; Fan and Gardent, 2022) has generally bypassed the pre-writing stage: for instance, Liu et al. (2018) presume reference documents are provided in advance, while Fan and Gardent (2022) assume an article outline is available and focus on expanding each section. These assumptions do not hold in general, as collecting references and crafting outlines are challenging even for experienced writers.

We explore these challenges by focusing on how to generate Wikipedia-like articles *from scratch*. We decompose this problem into two tasks. The first is to conduct research to generate an outline, *i.e.*, a list of multi-level sections, and collect a set of reference documents. The second uses the outline and the references to produce the full-length article. Such a task decomposition mirrors the human writing process which usually includes phases of pre-writing, drafting, and revising (Rohman, 1965; Munoz-Luna, 2015).

As pre-trained language models inherently possess a wealth of knowledge, a direct approach is to rely on their parametric knowledge for generating outlines or even entire articles (*Direct Gen*). However, this approach is limited by a lack of details and hallucinations (Xu et al., 2023), particularly in addressing long-tail topics (Kandpal et al., 2023). This underscores the importance of leveraging external sources, and current strategies often involve retrieval-augmented generation (*RAG*), which circles back to the problem of researching the topic in the pre-writing stage, as much information cannot be surfaced through simple topic searches.

Human learning theories (Tawfik et al., 2020; Booth et al., 2003) highlight *asking effective questions* in information acquisition. Although instruction-tuned models (Ouyang et al., 2022) can be prompted directly to generate questions, we find that they typically produce basic "What", "When", and "Where" questions (Figure 1 (A)) which often only address surface-level facts about the topic. To endow LLMs with the capacity to conduct better research, we propose the **STORM** paradigm for the **S**ynthesis of **T**opic **O**utlines through **R**etrieval and **M**ulti-perspective Question Asking.

The design of STORM is based on two hypotheses: (1) diverse perspectives lead to varied questions; (2) formulating in-depth questions requires iterative research. Building upon these hypotheses, STORM employs a novel multi-stage approach. It first discovers diverse perspectives by retrieving and analyzing Wikipedia articles from similar topics and then personifies the LLM with specific perspectives for question asking (Figure 1 (B)). Next, to elicit follow-up questions for iterative research (Figure 1 (C)), STORM simulates multi-turn conversations where the answers to the generated questions are grounded on the Internet. Finally, based on the LLM's internal knowledge and the collected information, STORM creates an outline that can be expanded section by section to develop a full-length Wikipedia-like article.

We evaluate STORM using our **FreshWiki** dataset (§2.1) which curates recent, high-quality Wikipedia articles to avoid data leakage during pre-training.[1] To facilitate the study of the pre-writing stage, we define metrics for evaluating the outline quality against human-written articles.

We further invited a group of experienced Wikipedia editors for expert evaluation. The editors found STORM outperforms an outline-driven RAG baseline, especially regarding the breadth and organization of the articles. They also identified challenges for future research, including addressing cases where: (1) the bias on the Internet affects the generated articles; (2) LLMs fabricate connections between unrelated facts. These challenges present new frontiers to grounded writing systems.

Our main contributions include:

- To evaluate the capacity of LLM systems at generating long-form grounded articles from scratch, and the pre-writing challenge in particular, we curate the FreshWiki dataset and establish evaluation criteria for both outline and final article quality.
- We propose STORM, a novel system that automates the pre-writing stage. STORM researches the topic and creates an outline by using LLMs to ask incisive questions and retrieving trusted information from the Internet.
- Both automatic and human evaluation demonstrate the effectiveness of our approach. Expert feedback further reveals new challenges in generating grounded long-form articles.

2 FreshWiki

We study generating Wikipedia-like articles from scratch, placing emphasis on the *pre-writing* stage (Rohman, 1965), which involves the demanding sub-tasks of gathering and curating relevant information ("research"). This models the human writing approach which has prompted some educators to view Wikipedia article writing as an educational exercise for academic training (Tardy, 2010).

Table 1 compares our work against prior benchmarks for Wikipedia generation. Existing work has generally focused on evaluating the generation of shorter snippets (*e.g.*, one paragraph), within a

[1] Our resources and code will be publicly released upon publication.

	Domain	Scope	Given Outline?	Given Refs?
Balepur et al. (2023)	One	One para.	/	Yes
Qian et al. (2023)	All	One para.	/	No
Fan and Gardent (2022)	One	Full article	Yes	No
Liu et al. (2018)	All	One para.	/	Yes
Sauper and Barzilay (2009)	Two	Full article	No	No
Ours	All	Full article	No	No

Table 1: Comparison of different Wikipedia generation setups in existing literature. Generating one paragraph does not need an article outline.

narrower scope (*e.g.*, a specific domain or two), or when an explicit outline or reference documents are supplied. A notable example is WikiSum (Liu et al., 2018), which treats generating Wikipedia articles as a multi-document summarization problem, with respect to the reference documents.

Our setup emphasizes the capability of long-form grounded writing systems to research and curate content. Specifically, given a topic t, the task is to find a set of references $\mathcal{R}$ and generate a full-length article $\mathcal{S} = s_1 s_2 ... s_n$, where each sentence s_i cites a list of documents in $\mathcal{R}$.[2]

2.1 The FreshWiki Dataset

Creating a new Wikipedia-like article demands not only fluent writing but also good research skills. As modern LLMs are generally trained on Wikipedia text, we mitigate data leakage by explicitly seeking out *recent* Wikipedia articles that were created (or very heavily edited) after the training cutoff of the LLMs we test. Our process can be repeated at future dates when new LLMs emerge.

To apply our date criteria, we focus on the top 100 most-edited pages, based on edit counts, for each month from February 2022 to September 2023[3]. To ensure high-quality references, we filter these articles to keep only those having B-class quality or above assessed by ORES[4]. We also exclude list articles[5] and articles that have no subsections. While high-quality Wikipedia articles usually contain structured data (*e.g.*, tables) and are multi-modal, we only consider the plain text component in constructing the dataset to simplify our task. More details of the dataset are in Appendix A.

[2]In practice, $\mathcal{S}$ also includes organizational elements such as section and subsection titles, which do not require citations.

[3]Obtained from https://wikimedia.org/api/rest_v1/metrics/edited-pages/top-by-edits/en.wikipedia/all-editor-types/content/{year}/{month}/all-days

[4]https://www.mediawiki.org/wiki/ORES

[5]https://en.wikipedia.org/wiki/Wikipedia:Stand-alone_lists

2.2 Outline Creation and Evaluation

A full-length article is hard to generate or evaluate (Xu et al., 2023; Krishna et al., 2023). When human educators teach students academic writing, they sometimes supervise students at the outline stage (Eriksson and Mäkitalo, 2015) because an extensive outline indicates a comprehensive understanding of the topic and provides a solid foundation for writing the full-length article (Dietz and Foley, 2019). Inspired by this, we decompose the generation of $\mathcal{S}$ into two stages. In the pre-writing stage, we require the system to create an outline $\mathcal{O}$, which is defined as a list of multi-level section headings[6]. In the writing stage, the system uses the topic t, the references $\mathcal{R}$, and an outline $\mathcal{O}$ to produce the full-length article $\mathcal{S}$.

To evaluate the outline coverage, we introduce two metrics: *heading soft recall* and *heading entity recall*. These metrics compare the multi-level section headings of the human-written article, considered as ground truth, and those in $\mathcal{O}$. Recognizing that an exact match between elements in these two sets of headings is unnecessary, we calculate the *heading soft recall* (Fränti and Mariescu-Istodor, 2023) using cosine similarity derived from Sentence-BERT (Reimers and Gurevych, 2019) embeddings of the headings (details in Appendix C.1). We also compute the *heading entity recall* which is quantified as the percentage of named entities in human-written article headings covered by $\mathcal{O}$. We extract entities with FLAIR named entity recognition (NER) (Akbik et al., 2019).

3 Method

We present STORM to automate the pre-writing stage by researching a given topic via effective question asking (§3.1, §3.2) and creating an outline (§3.3). The outline will be extended to a full-length article grounded on the collected references (§3.4). Figure 2 gives an overview of STORM and we include the pseudo code in Appendix B.

3.1 Perspective-Guided Question Asking

Rohman (1965) defines pre-writing as the stage of discovery in the writing process. In parallel with stakeholder theory in business (Freeman et al., 2010), where diverse stakeholders prioritize varying facets of a company, individuals with distinct

[6]Since language models process and produce sequences, we can linearize $\mathcal{O}$ by adding "#" to indicate section titles, "##" to indicate subsection titles, etc.

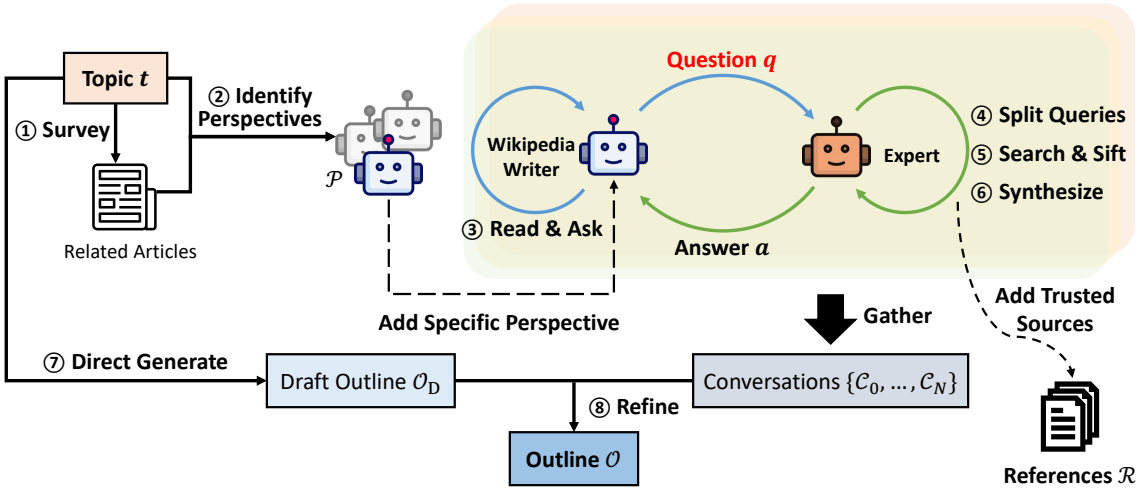

Figure 2: The overview of STORM that automates the pre-writing stage. Starting with a given topic, STORM identifies various perspectives on covering the topic by surveying related Wikipedia articles (①-②). It then simulates conversations between a Wikipedia writer who asks questions guided by the given perspective and an expert grounded on trustworthy online sources (③-⑥). The final outline is curated based on the LLM's intrinsic knowledge and the gathered conversations from different perspectives (⑦-⑧).

perspectives may concentrate on different aspects when researching the same topic and discover multifaceted information. Further, the specific perspectives can serve as prior knowledge, guiding individuals to ask more in-depth questions. For example, an event planner might ask about the "transportation arrangements" and "budget" for "the 2022 Winter Olympics opening ceremony", whereas a layperson might ask more general questions about the event's basic information (Figure 1 (A)).

Given the input topic t, STORM discovers different perspectives by surveying existing articles from similar topics and uses these perspectives to control the question asking process. Specifically, STORM prompts an LLM to generate a list of related topics and subsequently extracts the tables of contents from their corresponding Wikipedia articles, if such articles can be obtained through Wikipedia API[7] (Figure 2 ①). These tables of contents are concatenated to create a context to prompt the LLM to identify N perspectives $\mathcal{P} = \{p_1, ..., p_N\}$ that can collectively contribute to a comprehensive article on t (Figure 2 ②). To ensure that the basic information about t is also covered, we add p_0 as "basic fact writer focusing on broadly covering the basic facts about the topic" into $\mathcal{P}$. Each perspective $p \in \mathcal{P}$ will be utilized to guide the LLM in the process of question asking in parallel.

3.2 Simulating Conversations

The theory of questions and question asking (Ram, 1991) highlights that while answers to existing questions contribute to a more comprehensive understanding of a topic, they often simultaneously give rise to new questions. To kick off this dynamic process, STORM simulates a conversation between a Wikipedia writer and a topic expert. In the i-th round of the conversation, the LLM-powered Wikipedia writer generates a single question q_i based on the topic t, its assigned perspective $p \in \mathcal{P}$, and the conversation history $\{q_1, a_1, ..., q_{i-1}, a_{i-1}\}$ where a_j denotes the simulated expert's answer. The conversation history enables the LLM to update its understanding of the topic and ask follow-up questions. In practice, we limit the conversation to at most M rounds.

To ensure that the conversation history provides factual information, we use trusted sources from the Internet to ground the answer a_i to each query q_i. Since q_i can be complicated, we first prompt the LLM to break down q_i into a set of search queries (Figure 2 ④) and the searched results will be evaluated using a rule-based filter according to the Wikipedia guideline[8] to exclude untrustworthy sources (Figure 2 ⑤). Finally, the LLM synthesizes the trustworthy sources to generate the answer a_i, and these sources will also be added to $\mathcal{R}$ for full article generation (§3.4).

[7] https://pypi.org/project/Wikipedia-API/

[8] https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources

3.3 Creating the Article Outline

After thoroughly researching the topic through $N + 1$ simulated conversations, denoted as $\{\mathcal{C}_0, \mathcal{C}_1, ..., \mathcal{C}_N\}$, STORM creates an outline before the actual writing starts. To fully leverage the internal knowledge of LLMs, we first prompt the model to generate a draft outline $\mathcal{O}_{\text{D}}$ given only the topic t (Figure 2 ⑦). $\mathcal{O}_{\text{D}}$ typically provides a general but organized framework. Subsequently, the LLM is prompted with the topic t, the draft outline $\mathcal{O}_{\text{D}}$, and the simulated conversations $\{\mathcal{C}_0, \mathcal{C}_1, ..., \mathcal{C}_N\}$ to refine the outline (Figure 2 ⑧). This results in an improved outline $\mathcal{O}$ which will be used for producing the full-length article.

3.4 Writing the Full-Length Article

Building upon the references $\mathcal{R}$ collected and the outline $\mathcal{O}$ developed during the pre-writing stage, the full-length article can be composed section by section. Since it is usually impossible to fit the entire $\mathcal{R}$ within the context window of the LLM, we use the section title and headings of its all-level subsections to retrieve relevant documents from $\mathcal{R}$ based on semantic similarity calculated from Sentence-BERT embeddings. With the relevant information at hand, the LLM is then prompted to generate the section with citations. Once all sections are generated, they are concatenated to form the full-length article. Since the sections are generated in parallel, we prompt the LLM with the concatenated article to delete repeated information to improve coherence. Furthermore, in alignment with Wikipedia's stylistic norms, the LLM is also utilized to synthesize a summary of the entire article, forming the lead section at the beginning.

4 Experiments

4.1 Article Selection

STORM is capable of researching complicated topics and writing long articles from detailed outlines. However, in this controlled experiment, we limit the final output to at most 4000 tokens (roughly 3000 words). For a meaningful comparison, we randomly select 100 samples from the FreshWiki dataset (see §2.1) that have human-written articles not exceeding 3000 words.

4.2 Automatic Metrics

As discussed in §2.2, we evaluate the outline quality to assess the pre-writing stage by calculating the *heading soft recall* and *heading entity recall*. A higher recall score signifies a more comprehensive outline relative to the human-written article.

To assess the full-length article quality, we adopt *ROUGE scores* (Lin, 2004) and compute the *entity recall* in the article level based on FLAIR NER results. Moreover, based on Wikipedia criteria[9], we evaluate the article from the aspects of (1) *Interest Level*, (2) *Coherence and Organization*, (3) *Relevance and Focus*, (4) *Coverage*, and (5) *Verifiability*. For aspects (1)-(4), we use Prometheus (Kim et al., 2023), a 13B evaluator LLM to score the article based on a 5-point rubric collaboratively developed with two experienced Wikipedia editors (see Appendix C.2). For verifiability, we calculate the *citation recall* and *citation precision* based on the definition in Gao et al. (2023). We use Mistral 7B-Instruct (Jiang et al., 2023a) to examine whether the cited passages entail the generated sentence.

4.3 Baselines

As prior works use different setups and do not use LLMs, they are hard to compare directly. Instead, we use the following three LLM-based baselines.

1. *Direct Gen*, a baseline that directly prompts the LLM to generate an outline, which is then used to generate the full-length article.

2. *RAG*, a retrieval-augmented generation baseline that searches with the topic and uses the searched results together with the topic t to generate an outline or the entire article.

3. *Outline-driven RAG (oRAG)*, which is identical to *RAG* in outline creation, but further searches additional information with section titles to generate the article section by section.

4.4 STORM Implementation

We build STORM with zero-shot prompting using the DSPy framework (Khattab et al., 2023). Appendix B includes the pseudo code and corresponding prompts. The hyperparameters N and M in STORM are both set as 5. We use the chat model `gpt-3.5-turbo` for question asking and use `gpt-3.5-turbo-instruct` for other parts of STORM. We also experiment with using `gpt-4` for drafting and refining the outline (Figure 2 ⑦-⑧). For reported results, the simulated topic expert in STORM is grounded on the You.com search API[10],

[9] https://en.wikipedia.org/wiki/Wikipedia:Good_article_criteria

[10] https://documentation.you.com/api-reference/search

	Comparsion with Human-written Articles			Rubric Grading			
	ROUGE-1	ROUGE-L	Entity Recall	Interest Level	Organization	Relevance	Coverage
Direct Gen	25.62	12.63	5.08	2.87	4.60	3.10	4.16
RAG	28.52	13.18	7.57	3.14	4.22	3.05	4.08
oRAG	44.26	16.51	12.57	3.90	4.79	4.09	4.70
STORM	**45.82**	**16.70**	**14.10**†	**3.99**†	4.82	**4.45**†	**4.88**†
w/o Outline Stage	26.77	12.77	7.39	3.33	**4.87**	3.35	4.37

Table 2: Results of automatic article quality evaluation. † denotes significant differences ($p < 0.05$) from a paired t-test between STORM and the best baseline, *i.e.*, oRAG. The rubric grading uses a 1-5 scale.

		Heading Soft Recall	Heading Entity Recall
GPT-3.5	Direct Gen	80.23	32.39
	RAG/oRAG	73.59	33.85
	STORM	**86.26**†	**40.52**†
	w/o Perspective	84.49	40.12
	w/o Conversation	77.97	31.98
GPT-4	Direct Gen	87.66	34.78
	RAG/oRAG	89.55	42.38
	STORM	**92.73**†	**45.91**
	w/o Perspective	92.39	42.70
	w/o Conversation	88.75	39.30

Table 3: Results of outline quality evaluation (%). † denotes significant differences ($p < 0.05$) from a paired t-test between STORM and baselines.

	Citation Recall	Citation Precision
STORM	84.83	85.18

Table 4: Citation quality judged by Mistral 7B-Instruct.

	STORM	w/o Perspective	w/o Conversation
$\lvert\mathcal{R}\rvert$	**99.83**	54.36	39.56

Table 5: Average number of unique references ($|\mathcal{R}|$) collected using different methods.

although the proposed pipeline is compatible with other search engines. The ground truth Wikipedia article is excluded from the search results.

For final article generation, we only report the results using `gpt-4` as `gpt-3.5` is not faithful to sources when generating text with citations (Gao et al., 2023). We set temperature as 1.0 and top_p as 0.9 for all experiments.

5 Results and Analysis

5.1 Main Results

We use outline coverage as a proxy to assess the pre-writing stage (see §2.2). Table 3 shows the heading soft recall and entity recall. Outlines directly generated by LLMs (*Direct Gen*) already demonstrate high heading soft recall, indicating LLMs' ability to grasp high-level aspects of a topic through their rich parametric knowledge. However, STORM, by asking effective questions to research the topic, can create higher recall outlines that cover more topic-specific aspects. Notably, although *RAG* leverages additional information, presenting unorganized information in the context window makes outline generation more challenging for the weaker model, *i.e.*, GPT-3.5, leading to worse performance.

We further evaluate the full-length article quality. As shown in Table 2, *oRAG* significantly outperforms *RAG*, highlighting the effectiveness of using outlines for structuring full-length article generation. Despite this method's advantages in leveraging retrieval and outlining, our approach still outperforms it. The effective question asking mechanism enhances the articles with greater entity recall. The evaluator LLM also rates these articles with significantly higher scores in the aspects of "Interest Level", "Relevance and Focus", and "Coverage". Nonetheless, we acknowledge the possibility of the evaluator LLM overrating machine-generated text. Our careful human evaluation (§6) reveals that STORM still has much room for improvement.

Although this work primarily focuses on the pre-writing stage and does not optimize generating text with citations, we still examine the citation quality of articles produced by our approach. As reported in Table 4, Mistral 7B-Instruct judges 84.83% of the sentences are supported by their citations. Appendix C.3 investigates the unsupported sentences and reveals that the primary issues stem from drawing improper inferences and inaccurate paraphrasing, rather than hallucinating non-existent contents.

5.2 Ablation Studies

As introduced in §3, STORM prompts LLMs to ask effective questions by discovering specific perspectives and simulating multi-turn conversa-

	oRAG		STORM		p-value
	Avg.	≥ 4 Rates	Av.g.	≥ 4 Rates	
Interest Level	3.63	57.5%	**4.03**	**70.0%**	0.077
Organization	3.25	45.0%	**4.00**	**70.0%**	0.005
Relevance	3.93	62.5%	**4.15**	**65.0%**	0.347
Coverage	3.58	57.5%	**4.00**	**67.5%**	0.084
Verifiability	**3.85**	67.5%	3.80	67.5%	0.843
#Preferred		14		**26**	

Table 6: Human evaluation results on 20 pairs of articles generated by STORM and *oRAG*. Each pair of articles is evaluated by two Wikipedia editors. The ratings are given on a scale between 1 and 7, with values ≥ 4 indicating good quality (see Table 10). We conduct paired t-test and report the p-value.

tions. We conduct the ablation study on outline creation by comparing STORM with two variants: (1) "STORM w/o Perspective", which omits perspective in the question generation prompt; (2) "STORM w/o Conversation", which prompts LLMs to generate a set number of questions altogether. To ensure a fair comparison, we control an equal total number of generated questions across all variants. Table 3 shows the ablation results and full STORM pipeline produces outlines with the highest recall. Also, "STORM w/o Conversation" gives much worse results, indicating reading relevant information is crucial to generating effective questions. We further examine how many unique sources are collected in $\mathcal{R}$ via different variants. As shown in Table 5, the full pipeline discovers more different sources and the trend is in accord with the automatic metrics for outline quality.

We also verify whether having an outline stage is necessary with STORM. In Table 2, "STORM w/o Outline Stage" denotes the results of generating the entire article given the topic and the simulated conversations. Removing the outline stage significantly deteriorates the performance across all metrics.

6 Human Evaluation

To better understand the strengths and weaknesses of STORM, we conduct human evaluation by collaborating with 10 experienced Wikipedia editors who have made at least 500 edits on Wikipedia and have more than 1 year of experience. We randomly sample 20 topics from our dataset and evaluate the articles generated by our method and *oRAG*, the best baseline according to the automatic evaluation. Each pair of articles is assigned to 2 editors.

We request editors to judge each article from the same five aspects defined in §4.2, but using a 1 to 7 scale for more fine-grained evaluation. While our automatic evaluation uses citation quality as a proxy to evaluate *Verifiability*, we stick to the Wikipedia standard of "verifiable with no original research" in human evaluation. Besides rating the articles, editors are asked to provide open-ended feedback and pairwise preference. After the evaluation finishes, they are further requested to compare an article produced by our method, which they have just reviewed, with its human-written counterpart, and report their perceived usefulness of STORM using a 1-5 Likert scale. More human evaluation details are included in Appendix D. Table 6 presents the rating and pairwise comparison results.

Articles produced by STORM exhibit greater breadth and depth than oRAG outputs. In accord with the finding in §5.1, editors judge articles produced by STORM as more interesting, organized, and having broader coverage compared to *oRAG* outputs. Specifically, 25% more articles produced by STORM are considered organized (*Organization* rating ≥ 4), and 10% more are deemed to have good coverage (*Coverage* rating ≥ 4). Even in comparison with human-written articles, one editor praises our result as providing "a bit more background information" and another notes that "I found that the AI articles had more depth compared to the Wikipedia articles". STORM also outperforms the best baseline in pairwise comparison.

More information in $|\mathcal{R}|$ poses challenges beyond factual hallucination. We examine 14 pairwise comparison responses where editors prefer oRAG outputs over STORM. Excluding 3 cases where pairwise preferences do not align with their ratings, editors assign lower *Verifiability* scores to articles from our approach in over 50% of the cases. Through analyzing the articles and editors' free-form feedback, we discover that *low Verifiability scores stem from red herring fallacy or overspeculation issues*. These arise when the generated articles introduce unverifiable connections between different pieces of information in $|\mathcal{R}|$ or between the information and the topic (examples included in Table 11). Compared to the widely discussed factual hallucination (Shuster et al., 2021; Huang et al., 2023), addressing such verifiability issues is more nuanced, surpassing basic fact-checking (Min et al., 2023).

Generated articles trail behind well-revised human works. While STORM outperforms the oRAG baseline, editors comment that the generated articles are *less informative than actual Wikipedia*

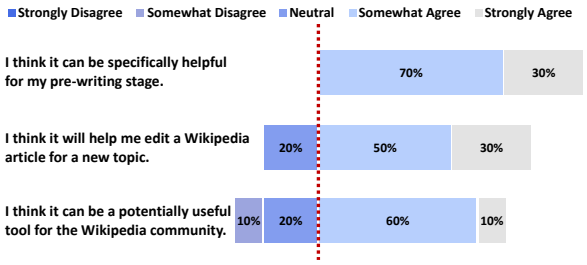

Figure 3: Survey results of the perceived usefulness of STORM ($n = 10$).

pages. Another major issue identified is *the transfer of bias and tone from Internet sources to the generated article*, with 7 out of 10 editors mentioning that the STORM-generated articles sound "emotional" or "unneutral". More analysis is discussed in Appendix E. This feedback suggests that reducing the retrieval bias in the pre-writing stage is a worthwhile direction for future work.

Generated articles are a good starting point. As shown in Figure 3, editors are unanimous in agreeing that STORM can aid them in their pre-writing stage. It is gratifying to know that the tool is helpful to experienced editors. 80% of the editors think that STORM can help them edit a Wikipedia article for a new topic. More reservation is expressed to the usefulness of STORM for the Wikipedia community at large; nonetheless, 70% of the editors think it is useful, with only 10% disagreeing.

7 Related Works

Retrieval-Augmented Generation (RAG) Augmenting language models (LMs) with retrieval at inference time is a typical way to leverage external knowledge stores (Ram et al., 2023; Izacard et al., 2023). While some works use retrieval to construct demonstrations for in-context learning (Li et al., 2023; Liu et al., 2022; Agrawal et al., 2023; Poesia et al., 2022; Shi et al., 2022; Khattab et al., 2022), another line of works uses retrieval to provide additional information for LMs to ground on. Lewis et al. (2020) study RAG on knowledge-intensive NLP tasks and find it improves diversity and factuality. Besides, RAG can be used to generate text with citations (Menick et al., 2022; Gao et al., 2023) and build attributed question answering systems (Bohnet et al., 2023). While RAG is widely studied in question answering, how to use it for long-form article generation is less investigated.

As a general framework, RAG is flexible in both the retrieval source and time. The retrieval sources can vary from domain databases (Zakka et al., 2023), code documentation (Zhou et al., 2023), to the whole Internet (Nakano et al., 2022; Komeili et al., 2022). Regarding the time, besides a one-time retrieval before generation, the system can be designed to self-decide when to retrieve across the course of the generation (Jiang et al., 2023b; Parisi et al., 2022; Shuster et al., 2022; Yao et al., 2023).

Automatic Expository Writing Different from other types of long-form generation (Yang et al., 2022; Feng et al., 2018), automatic expository writing requires grounding on external documents and leveraging the interplay between reading and writing. Balepur et al. (2023) propose the Imitate-Retrieve-Paraphrase framework for expository writing at the paragraph level to address the challenges in synthesizing information from multiple sources. Beyond summarizing sources, Shen et al. (2023) highlight that expository writing requires the author's sensemaking process over source documents and good outline planning. We tackle these challenges by focusing on the pre-writing stage.

Question Asking in NLP Question asking capabilities in NLP systems have expanded across several fronts, including generating clarification questions to understand user intents (Aliannejadi et al., 2019; Rahmani et al., 2023), and breaking large questions into smaller ones to improve compositional reasoning (Press et al., 2023). While humans usually ask questions to learn new knowledge (Tawfik et al., 2020; Booth et al., 2003), how to optimize question informativeness and specificity in information-seeking conversations remains less explored. The closest work is Qi et al. (2020) which defines the question informativeness using the unigram precision function and uses reinforcement learning to increase the question informativeness.

8 Conclusion

We propose STORM, an LLM-based writing system that automates the pre-writing stage for creating Wikipedia-like articles from scratch. We curate the FreshWiki dataset and establish evaluation criteria to study the generation of grounded long-form articles. Experimental results demonstrate that the question asking mechanism in STORM improves both the outline and article quality. With the improved breadth and depth, STORM helps surface new challenges for grounded writing systems through expert evaluation. The experienced Wikipedia editors in our study unanimously agree that STORM is helpful for their pre-writing stage.

Limitations

In this work, we explore generating Wikipedia-like articles from scratch as a way to push the frontier of automatic expository writing and long-form article generation. While our approach significantly outperforms baseline methods in both automatic and human evaluations, the quality of machine-written articles still lags behind well-revised human-authored articles, specifically in aspects of neutrality and verifiability. Although STORM discovers different perspectives in researching the given topic, the collected information may still be biased towards dominant sources on the Internet and may contain promotional content. Moreover, the verifiability issues identified in this work go beyond factual hallucination, which highlights new challenges to grounded writing systems.

Another limitation of this work is that although we focus on the task of generating Wikipedia-like articles from scratch, our task setup is still simplified to only consider the generation of free-form text. Human-authored high-quality Wikipedia articles usually contain structured data and multi-modal information. We leave the exploration of generating multi-modal grounded articles for future work.

Ethics Statement

Different from the creative generation, grounded article generation may impact how people learn about topics or consume source information. All the studies and the evaluation in this work are designed to prevent the dissemination of misinformation by not publishing generated content online and implementing strict accuracy checks. We avoid any disruption to Wikipedia or related communities, as our system does not interact with live pages. Also, although we try to generate grounded articles, we believe there is no privacy issue related to this work as we only use information publicly available on the Internet.

The primary risk of our work is that the Wikipedia articles written by our system are grounded on information on the Internet which contains some biased or discriminative content on its own. Currently, our system relies on the search engine to retrieve information but does not include any post-processing module. We believe improving the retrieval module to have good coverage of different viewpoints and adding a content sifting module to the current system will be a critical next step to achieve better neutrality and balance in the generated articles.

Another limitation we see from an ethical point of view is that we only consider writing English Wikipedia articles in this work. Extending the current system to a multilingual setup is a meaningful direction for future work as more topics do not have Wikipedia pages in non-English languages.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 475–484.

Nishant Balepur, Jie Huang, and Kevin Chang. 2023. Expository text generation: Imitate, retrieve, paraphrase. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11896–11919, Singapore. Association for Computational Linguistics.

Siddhartha Banerjee and Prasenjit Mitra. 2015. WikiKreator: Improving Wikipedia stubs automatically. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 867–877, Beijing, China. Association for Computational Linguistics.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and modeling for attributed large language models.

Wayne C Booth, Gregory G Colomb, and Joseph M Williams. 2003. *The craft of research*. University of Chicago press.

Laura Dietz and John Foley. 2019. Trec car y3: Complex answer retrieval overview. In *Proceedings of Text REtrieval Conference (TREC)*.

Ann-Marie Eriksson and Åsa Mäkitalo. 2015. Supervision at the outline stage: Introducing and encountering issues of sustainable development through academic writing assignments. *Text & Talk*, 35(2):123–153.

Angela Fan and Claire Gardent. 2022. Generating biographies on Wikipedia: The impact of gender bias on the retrieval-based generation of women biographies. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8561–8576, Dublin, Ireland. Association for Computational Linguistics.

Xiaocheng Feng, Ming Liu, Jiahao Liu, Bing Qin, Yibo Sun, and Ting Liu. 2018. Topic-to-essay generation with neural networks. In *IJCAI*, pages 4078–4084.

Tira Nur Fitria. 2023. Artificial intelligence (ai) technology in openai chatgpt application: A review of chatgpt in writing english essay. In *ELT Forum: Journal of English Language Teaching*, volume 12, pages 44–58.

Pasi Fränti and Radu Mariescu-Istodor. 2023. Soft precision and recall. *Pattern Recognition Letters*, 167:115–121.

R Edward Freeman, Jeffrey S Harrison, Andrew C Wicks, Bidhan L Parmar, and Simone De Colle. 2010. Stakeholder theory: The state of the art.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Julià Minguillón, Maura Lerga, Eduard Aibar, Josep Lladós-Masllorens, and Antoni Meseguer-Artola. 2017. Semi-automatic generation of a corpus of wikipedia articles on science and technology. *Profesional de la Información*, 26(5):995–1005.

Rosa Munoz-Luna. 2015. Main ingredients for success in l2 academic writing: Outlining, drafting and proofreading. *PloS one*, 10(6):e0128309.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. Talm: Tool augmented language models.

John V Pavlik. 2023. Collaborating with chatgpt: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, 78(1):84–93.

Gabriel Poesia, Alex Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, and Sumit Gulwani. 2022. Synchromesh: Reliable code generation from pre-trained language models. In *International Conference on Learning Representations*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, and Christopher D. Manning. 2020. Stay hungry, stay focused: Generating informative and specific questions in information-seeking conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 25–40, Online. Association for Computational Linguistics.

Hongjing Qian, Yutao Zhu, Zhicheng Dou, Haoqi Gu, Xinyu Zhang, Zheng Liu, Ruofei Lai, Zhao Cao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Webbrain: Learning to generate factually correct articles for queries by grounding on large web corpus.

Hossein A. Rahmani, Xi Wang, Yue Feng, Qiang Zhang, Emine Yilmaz, and Aldo Lipani. 2023. A survey on asking clarification questions datasets in conversational systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2698–2716, Toronto, Canada. Association for Computational Linguistics.

Ashwin Ram. 1991. A theory of questions and question asking. *Journal of the Learning Sciences*, 1(3-4):273–318.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

D Gordon Rohman. 1965. Pre-writing the stage of discovery in the writing process. *College composition and communication*, 16(2):106–112.

Christina Sauper and Regina Barzilay. 2009. Automatically generating Wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 208–216, Suntec, Singapore. Association for Computational Linguistics.

Zejiang Shen, Tal August, Pao Siangliulue, Kyle Lo, Jonathan Bragg, Jeff Hammerbacher, Doug Downey, Joseph Chee Chang, and David Sontag. 2023. Beyond summarization: Designing ai support for real-world expository writing tasks.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3254–3265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 373–393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christine M Tardy. 2010. Writing for the world: Wikipedia as an introduction to academic writing. In *English teaching forum*, volume 48, page 12. ERIC.

Andrew A Tawfik, Arthur Graesser, Jessica Gatewood, and Jaclyn Gishbaugher. 2020. Role of questions in inquiry-based instruction: towards a design taxonomy for question-asking and implications for design. *Educational Technology Research and Development*, 68:653–678.

Charles A Weaver III and Walter Kintsch. 1991. Expository text.

Karsten Wenzlaff and Sebastian Spaeth. 2022. Smarter than humans? validating how openai's chatgpt model explains crowdfunding, alternative finance and community finance. *Validating how OpenAI's ChatGPT model explains Crowdfunding, Alternative Finance and Community Finance.(December 22, 2022)*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex R Dalal, Jennifer L Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, et al. 2023. Almanac: Retrieval-augmented language models for clinical medicine. *Research Square*.

Shuyan Zhou, Uri Alon, Frank F. Xu, Zhengbao Jiang, and Graham Neubig. 2023. Docprompting: Generating code by retrieving the docs. In *The Eleventh International Conference on Learning Representations*.

Average Numer of Sections	8.4
Average Number of All-level Headings	15.8
Average Length of a Section	327.8
Average Length of Total Article	2159.1
Average Number of References	90.1

Table 7: Statistics of the dataset used in our experiments.

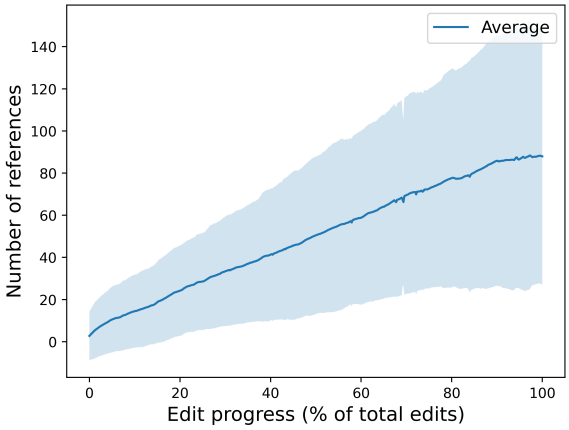

Figure 4: Evolution of reference count in the Wikipedia article editing process.

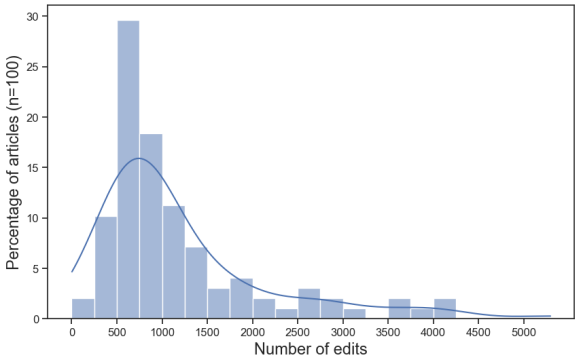

Figure 5: Distribution of edit counts for Wikipedia articles in our experiments ($n = 100$).

A Dataset Details

As discussed in §2.1, we curate the FreshWiki dataset by collecting *recent* and *high-quality* English Wikipedia articles. We select the most-edited pages over a specific period rather than using creation dates as a cutoff because most of Wikipedia articles are "stubs" or are of low quality when they were created. For quality, we consider articles predicted to be of B-class quality or above. According to Wikipedia statistics[11], only around 3% of existing Wikipedia pages meet this quality standard. As LLMs can generate reasonably good outputs, we think it is important to use high-quality human-written articles as references for further research.

For experiments in this work, we randomly select 100 samples with human-written articles under 3000 words to have a meaningful comparison. Table 7 gives the data statistics. Notably, human-authored articles have a large number of references but they require numerous edits to achieve this. Figure 4 illustrates the evolution of the reference count in the article edit process and Figure 5 gives the distribution of edit counts for human-authored articles used in our experiments.

[11] https://en.wikipedia.org/wiki/Wikipedia:Content_assessment

B Pseudo Code of STORM

In §3, we introduce STORM, a framework that automates the pre-writing stage by discovering different perspectives, simulating information-seeking conversations, and creating a comprehensive outline. Algorithm 1 displays the skeleton of STORM.

We implement STORM with zero-shot prompting using the DSPy framework (Khattab et al., 2023). Listing 1 and 2 show the prompts used in our implementation. We highlight that STORM offers a general framework designed to assist the creation of grounded, long-form articles, without depending extensively on prompt engineering for a single domain.

C Automatic Evaluation Details

C.1 Soft Heading Recall

We calculate the soft heading recall between the multi-level headings in the generated outline, considered as the prediction P, and those in the human-written article, considered as the ground truth G. The calculation is based on the soft recall definition in Fränti and Mariescu-Istodor (2023). Given a set $A = \{Ai\}_{i=1}^{K}$, *soft count* of an item is defined as the inverse of the sum of its similarity to other items in the set:

$$\text{count}\,(A_i) = \frac{1}{\sum_{j=1}^{K} \text{Sim}\,(A_i, A_j)} \qquad (1)$$
$$\text{Sim}\,(A_i, A_j) = \cos\left(\text{embed}(A_i), \text{embed}(A_j)\right),$$

where $\text{embed}(\cdot)$ in Equation (1) is parameterized by `paraphrase-MiniLM-L6-v2` provided in the Sentence-Transformers library[12]. The cardinality

[12] https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2

```python
class GenRelatedTopicsPrompt(dspy.Signature):
    """
    I'm writing a Wikipedia page for a topic mentioned below. Please identify and recommend some Wikipedia pages on closely related subjects. I'm looking for examples that provide insights into interesting aspects commonly associated with this topic, or examples that help me understand the typical content and structure included in Wikipedia pages for similar topics.
    Please list the urls in separate lines.
    """

    topic = dspy.InputField(prefix="Topic of interest:", format=str)
    related_topics = dspy.OutputField()

class GenPerspectivesPrompt(dspy.Signature):
    """
    You need to select a group of Wikipedia editors who will work together to create a comprehensive article on the topic. Each of them represents a different perspective, role, or affiliation related to this topic. You can use other Wikipedia pages of related topics for inspiration. For each editor, add description of what they will focus on.
    Give your answer in the following format: 1. short summary of editor 1: description\n2. short summary of editor 2: description\n...
    """

    topic = dspy.InputField(prefix='Topic of interest:', format=str)
    examples = dspy.InputField(prefix='Wiki page outlines of related topics for inspiration:\n', format=str)
    perspectives = dspy.OutputField()

class GenQnPrompt(dspy.Signature):
    """
    You are an experienced Wikipedia writer and want to edit a specific page. Besides your identity as a Wikipedia writer, you have a specific focus when researching the topic.
    Now, you are chatting with an expert to get information. Ask good questions to get more useful information.
    When you have no more question to ask, say "Thank you so much for your help!" to end the conversation.
    Please only ask one question at a time and don't ask what you have asked before. Your questions should be related to the topic you want to write.
    """

    topic = dspy.InputField(prefix='Topic you want to write: ', format=str)
    persona = dspy.InputField(prefix='Your specific perspective: ', format=str)
    conv = dspy.InputField(prefix='Conversation history:\n', format=str)
    question = dspy.OutputField()

class GenQueriesPrompt(dspy.Signature):
    """
    You want to answer the question using Google search. What do you type in the search box?
    Write the queries you will use in the following format:- query 1\n- query 2\n...
    """

    topic = dspy.InputField(prefix='Topic you are discussing about: ', format=str)
    question = dspy.InputField(prefix='Question you want to answer: ', format=str)
    queries = dspy.OutputField()
```

Listing 1: Prompts used in STORM, corresponding to Line 4, 11, 19, 22 in Algorithm 1.

```
class GenAnswerPrompt(dspy.Signature):
    """
    You are an expert who can use information effectively. You are chatting with a
        Wikipedia writer who wants to write a Wikipedia page on topic you know. You
        have gathered the related information and will now use the information to
        form a response.
    Make your response as informative as possible and make sure every sentence is
        supported by the gathered information.
    """

    topic = dspy.InputField(prefix='Topic you are discussing about:', format=str)
    conv = dspy.InputField(prefix='Question:\n', format=str)
    info = dspy.InputField(
        prefix='Gathered information:\n', format=str)
    answer = dspy.OutputField(prefix='Now give your response:\n')


class DirectGenOutlinePrompt(dspy.Signature):
    """
    Write an outline for a Wikipedia page.
    Here is the format of your writing:
        1. Use "#" Title" to indicate section title, "##" Title" to indicate
            subsection title, "###" Title" to indicate subsubsection title, and so
            on.
        2. Do not include other information.
    """

    topic = dspy.InputField(prefix="Topic you want to write: ", format=str)
    outline = dspy.OutputField(prefix="Write the Wikipedia page outline:\n")


class RefineOutlinePrompt(dspy.Signature):
    """
    Improve an outline for a Wikipedia page. You already have a draft outline that
        covers the general information. Now you want to improve it based on the
        information learned from an information-seeking conversation to make it more
         comprehensive.
    Here is the format of your writing:
        1. Use "#" Title" to indicate section title, "##" Title" to indicate
            subsection title, "###" Title" to indicate subsubsection title, and so
            on.
        2. Do not include other information.
    """

    topic = dspy.InputField(prefix="Topic you want to write: ", format=str)
    conv = dspy.InputField(prefix="Conversation history:\n", format=str)
    old_outline = dspy.OutputField(prefix="Current outline:\n", format=str)
    outline = dspy.OutputField(
        prefix='Write the Wikipedia page outline:\n')
```

Listing 2: Prompts used in STORM (continue), corresponding to Line 24, 31, 32 in Algorithm 1.

Algorithm 1: STORM

Input : Topic t, maximum perspective N, maximum conversation round M

Output : Outline $\mathcal{O}$, references $\mathcal{R}$

P0 = "basic fact writer ..." `// Constant.`
$\mathcal{R} \leftarrow [\,]$
`// Discover perspectives` $\mathcal{P}$.
related_topics $\leftarrow$ `gen_related_topics`(t)
tocs $\leftarrow [\,]$
foreach related_t in related_topics **do**
  article $\leftarrow$ `get_wiki_article`(related_t)
  if article **then**
    tocs.append(`extract_toc`(article))
  end
end
$\mathcal{P} \leftarrow$ `gen_perspectives`(t, tocs)
$\mathcal{P} \leftarrow$ [P0] + $\mathcal{P}[:N]$
`// Simulate conversations.`
convos $\leftarrow [\,]$
foreach p *in* $\mathcal{P}$ **do**
  convo_history $\leftarrow [\,]$
  for i = 1 to M **do**
    `// Question asking.`
    $q \leftarrow$ `gen_qn`(t, p, dlg_history)
    convo_history.append(q)
    `// Question answering.`
    queries $\leftarrow$ `gen_queries`(t, q)
    sources $\leftarrow$ `search_and_sift`(queries)
    $a \leftarrow$ `gen_ans`(t, q, sources)
    convo_history.append(a)
    $\mathcal{R}$.append(sources)
  end
  convos.append(convo_history)
end
`// Create the outline.`
$\mathcal{O}_{\text{D}} \leftarrow$ `direct_gen_outline`(t)
$\mathcal{O} \leftarrow$ `refine_outline`(t, $\mathcal{O}_{\text{D}}$, convos)
return $\mathcal{O}$, $\mathcal{R}$

of A is the sum of the counts of its individual items:

$$\text{card}(A) = \sum_{i=1}^{K} \text{count}\,(A_i) \tag{2}$$

The soft heading recall is calculated as

$$\textit{soft heading recall} = \frac{\text{card}(G \cap P)}{\text{card}(G)}, \tag{3}$$

where the cardinality of intersection is defined via the union as follows:

$$\begin{aligned} &\text{card}(G \cap P) = \\ &\text{card}(G) + \text{card}(P) - \text{card}(G \cup P). \end{aligned} \tag{4}$$

C.2 LLM Evaluator

We use Prometheus[13] (Kim et al., 2023), a 13B open-source evaluator LLM that can assess long-form text based on customized 1-5 scale rubric, to grade the article from the aspects of *Interest level*, *Coherence and Organization*, *Relevance and Focus*, and *Coverage*. Table 8 gives our grading rubric. While Prometheus is best used with a score 5 reference answer, we find adding the reference will exceed the context length limit of the model. Since Kim et al. (2023) show Prometheus ratings without reference also correlate well with human preferences, we omit the reference and trim the input article to be within 2000 words by iteratively removing contents from the shortest section to ensure the input can fit into the model's context window.

C.3 More Discussion of the Citation Quality

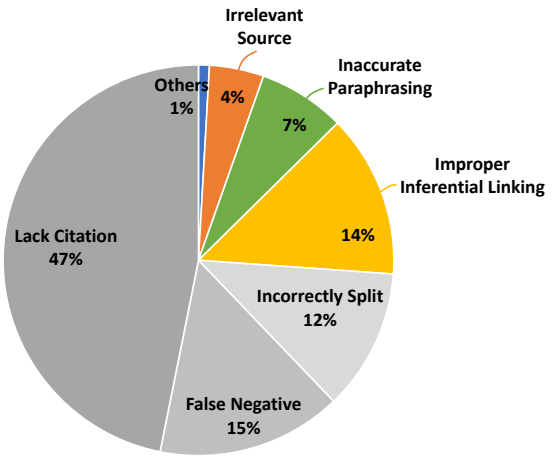

Figure 6: Error analysis of unsupported sentences in 10 sampled articles.

[13] https://huggingface.co/kaist-ai/prometheus-13b-v1.0

Criteria Description	**Interest Level**: How engaging and thought-provoking is the article?
Score 1 Description	Not engaging at all; no attempt to capture the reader's attention.
Score 2 Description	Fairly engaging with a basic narrative but lacking depth.
Score 3 Description	Moderately engaging with several interesting points.
Score 4 Description	Quite engaging with a well-structured narrative and noteworthy points that frequently capture and retain attention.
Score 5 Description	Exceptionally engaging throughout, with a compelling narrative that consistently stimulates interest.
Criteria Description	**Coherence and Organization**: Is the article well-organized and logically structured?
Score 1 Description	Disorganized; lacks logical structure and coherence.
Score 2 Description	Fairly organized; a basic structure is present but not consistently followed.
Score 3 Description	Organized; a clear structure is mostly followed with some lapses in coherence.
Score 4 Description	Good organization; a clear structure with minor lapses in coherence.
Score 5 Description	Excellently organized; the article is logically structured with seamless transitions and a clear argument.
Criteria Description	**Relevance and Focus**: Does the article stay on topic and maintain a clear focus?
Score 1 Description	Off-topic; the content does not align with the headline or core subject.
Score 2 Description	Somewhat on topic but with several digressions; the core subject is evident but not consistently adhered to.
Score 3 Description	Generally on topic, despite a few unrelated details.
Score 4 Description	Mostly on topic and focused; the narrative has a consistent relevance to the core subject with infrequent digressions.
Score 5 Description	Exceptionally focused and entirely on topic; the article is tightly centered on the subject, with every piece of information contributing to a comprehensive understanding of the topic.
Criteria Description	**Broad Coverage**: Does the article provide an in-depth exploration of the topic and have good coverage?
Score 1 Description	Severely lacking; offers little to no coverage of the topic's primary aspects, resulting in a very narrow perspective.
Score 2 Description	Partial coverage; includes some of the topic's main aspects but misses others, resulting in an incomplete portrayal.
Score 3 Description	Acceptable breadth; covers most main aspects, though it may stray into minor unnecessary details or overlook some relevant points.
Score 4 Description	Good coverage; achieves broad coverage of the topic, hitting on all major points with minimal extraneous information.
Score 5 Description	Exemplary in breadth; delivers outstanding coverage, thoroughly detailing all crucial aspects of the topic without including irrelevant information.

Table 8: Scoring rubrics on a 1-5 scale for the evaluator LLM.

Error Type	**Topic**	**Unsupported Sentence**	**Source**
Improper Inferential Linking	Lahaina, Hawaii	Throughout its history, religion has remained the paramount aspect of Hawaiian life in Lahaina, permeating every daily activity and significant event[5].	[5] "Religion, Beliefs & Spirituality" *(The source discusses religion as part of Hawaiian life but does not mention Lahania.)*
Inaccurate Paraphrasing	2022 Crimean Bridge explosion	Completed in June 2020, the bridge serves as a major supply route for Russian forces in the region and is significant to Russia's claim over the disputed territory[2][11].	[2] "Crimean Bridge - Wikipedia" *(The source says "The first scheduled passenger train crossed the bridge on 25 December 2019, while the bridge was opened for freight trains on 30 June 2020".)*
Citing Irrelevant Sources	LK-99	For example, comparisons have been drawn between the performance of LK-9 and the dynamic resolution capabilities of video games such as Battlefield 2042[22].	[22] "Battlefield 2042 PC performance guide: The best settings for a high frame rate" *(The source is irrelevant to LK-99.)*

Table 9: Examples of different error types of unsupported sentences.

We use Mistral 7B-Instruct[14] (Jiang et al., 2023a) to examine whether the cited passages entail the generated sentence. Table 4 reports the citation quality of articles produced by our approach, showing that around 15% sentences in generated articles are unsupported by citations. We further investigate the failure cases by randomly sampling 10 articles and an author manually examines all the unsupported sentences in these articles. Besides sentences that are incorrectly split[15], lack citations, or are deemed supported by the author's judgment, our analysis identifies three main error categories (examples are given in Table 9): *improper inferential linking*, *inaccurate paraphrasing*, and *citing irrelevant sources*.

We show the error distribution in Figure 6. Notably, the most common errors stem from the tendency of LLMs to form improper inferential links between different pieces of information presented in the context window. Our analysis of citation quality suggests that, in addition to avoiding hallucinations, future research in grounded text generation should also focus on preventing LLMs from making overly inferential leaps based on the provided information.

D Human Evaluation Details

We recruited 10 experienced Wikipedia editors to participate in our study by creating a research page on `Meta-Wiki`[16] and reaching out to active editors who have recently approved articles for Wikipedia.[17] Our participation group includes 3 editors with 1-5 years of experience, 4 with 6-10 years, and 3 with over 15 years of contribution. The study was approved by the Institutional Review Board of our institution and the participants signed the consent form through Qualtrics questionnaires before the study started.

To streamline the evaluation of grounded articles, we developed a web application, which features a side-by-side display of the article and its citation snippets, to gather ratings and open-ended feedback for each article. Figure 7 shows the screenshot of our web application and the full article produced by STORM is included in Table 12. For human evaluation, we use a 1 to 7 scale for more fine-grained evaluation. The grading rubric is included in Table 10.

We collected the pairwise preferences and the perceived usefulness of STORM via an online questionnaire. Specifically, for the perceived usefulness, we request editors to rate their agreement with statements "I think it can be specifically helpful for my pre-writing stage (*e.g.*, collecting relevant sources, outlining, drafting).", "I think it will help me edit a Wikipedia article for a new topic", "I think it can be a potentially useful tool for the Wikipedia community" on a Likert scale of 1-5, corresponding to *Strongly disagree*, *Somewhat disagree*, *Neither agree nor disagree*, *Somewhat agree*, *Strongly agree*.

E Error Analysis

While articles produced by STORM are preferred by both automatic metrics and human evaluation, experienced editors still identified multiple problems with the machine-generated articles. We analyze the free-form comments and summarize the major issues in Table 11.

The primary issue raised is that the generated articles often contain emotional language and lack neutrality, primarily due to the source material. STORM currently retrieves grounding sources from the Internet which is not neutral and contains considerable promotional content on its own. Addressing this bias in the pre-writing stage represents a valuable direction for future research. Another major issue is the red herring fallacy or the over-association of unrelated facts. Addressing this challenge calls for high-level sensemaking rather than mere fact-level verification.

[14] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

[15] Following Gao et al. (2023), we check citation quality in the sentence level and split articles into sentences using NLTK `sent_tokenize`. `sent_tokenize` sometimes fails to split sentences correctly when the article contains special words like "No.12847", "Bhatia et al.", *etc.*

[16] https://meta.wikimedia.org

[17] Since evaluating Wikipedia-like articles is time-consuming and requires expertise, we paid each participant 50$ for our study.

Interest Level
1: Not engaging at all; no attempt to capture the reader's attention.
2: Slightly engaging with rare moments that capture attention.
3: Fairly engaging with a basic narrative but lacking depth.
4: Moderately engaging with several interesting points.
5: Quite engaging with a well-structured narrative and noteworthy points that frequently capture and retain attention.
6: Very engaging with a compelling narrative that captures and mostly retains attention.
7: Exceptionally engaging throughout, with a compelling narrative that consistently stimulates interest.
Coherence and Organization
1: Disorganized; lacks logical structure and coherence.
2: Poor organization; some structure is evident but very weak.
3: Fairly organized; a basic structure is present but not consistently followed.
4: Organized; a clear structure is mostly followed with some lapses in coherence.
5: Good organization; a clear structure with minor lapses in coherence.
6: Very well-organized; a logical structure with transitions that effectively guide the reader.
7: Excellently organized; the article is logically structured with seamless transitions and a clear argument.
Relevance and Focus
1: Off-topic; the content does not align with the headline or core subject.
2: Mostly off-topic with some relevant points.
3: Somewhat on topic but with several digressions; the core subject is evident but not consistently adhered to.
4: Generally on topic, despite a few unrelated details.
5: Mostly on topic and focused; the narrative has a consistent relevance to the core subject with infrequent digressions.
6: Highly relevant with a focused narrative and purpose.
7: Exceptionally focused and entirely on topic; the article is tightly centered on the subject, with every piece of information contributing to a comprehensive understanding of the topic.
Broad Coverage
1: Severely lacking; offers little to no coverage of the topic's primary aspects, resulting in a very narrow perspective.
2: Minimal coverage; addresses only a small selection of the topic's main aspects, with significant omissions.
3: Partial coverage; includes some of the topic's main aspects but misses others, resulting in an incomplete portrayal.
4: Acceptable breadth; covers most main aspects, though it may stray into minor unnecessary details or overlook some relevant points.
5: Good coverage; achieves broad coverage of the topic, hitting on all major points with minimal extraneous information.
6: Comprehensive; provides thorough coverage of all significant aspects of the topic, with a well-balanced focus.
7: Exemplary in breadth; delivers outstanding coverage, thoroughly detailing all crucial aspects of the topic without including irrelevant information.
Verifiability
1: No supporting evidence; claims are unsubstantiated.
2: Rarely supported with evidence; many claims are unsubstantiated.
3: Inconsistently verified; some claims are supported; evidence is occasionally provided.
4: Generally verified; claims are usually supported with evidence; however, there might be a few instances where verification is lacking
5: Well-supported; claims are very well supported with credible evidence, and instances of unsupported claims are rare.
6: Very well-supported; almost every claim is substantiated with credible evidence, showing a high level of thorough verification.
7: Exemplary verification; each claim is supported by robust, credible evidence from authoritative sources, reflecting strict adherence to the no original research policy.

Table 10: Scoring rubrics on a 1-7 scale for human evaluation.

Issue	Mentioned Time	Example Comments
Use of emotional words, unneutral	12	The word "significant" is used 17 times in this article. Vague and unsupported claims are made about broader political importance and "pivotal role[s]", and is unencyclopedic. (comment on article *Lahaina, Hawaii*)
		[...] but they still have not fixed the issue of neutral point of view. It is also evident in this article that the writer's standpoint is biased towards Taylor Swift. Other than that, it did a good job at summarizing key points and putting depth into this. (comment on article *Speak Now (Taylor's Version)*)
		"The film was also featured in an art and film festival hosted by The California Endowment, highlighting the power of stories in reshaping narratives about communities." Yes, technically the source says that, but it's a stretch to say in Wikipedia voice and just sounds like non-neutral, promotional prose. (comment on article *Gehraiyaan*)
Red herring fallacy, associating unrelated sources	11	Polling from America shouldn't be included and links to climate change shouldn't be made unless explicitly connected by the source. (comment on article *Typhoon Hinnamnor*)
		Sourcing seems mostly fine, though some aren't directly related (Ex. 39,40). (comment on article *Gehraiyaan*)
		Here is a lengthy digression about KISS, not necessary because the article on the band should be linked to. (comment *on article 2022 AFL Grand Final*)
Missing important information	6	"One study, conducted by Sinéad Griffin, a physicist at the Lawrence Berkeley National Laboratory, provided some analysis of LK-99's abilities using supercomputer simulations[20]." This is not enough information about the analysis, which would have been very useful in the article. (comment on article *LK-99*)
		Although the earthquake's immediate aftermath and response are adequately covered, there could be more about the long-term socioeconomic impact and recovery processes. (comment on article *2022 West Java earthquake*)
Improper handling of time-sensitive information	5	Words like "now" should be avoided in Wikipedia articles to prevent them from becoming dated and phrases such as, "as of December 2023" should be used instead. (comment on article *Cyclone Batsirai*)
		"as of December 13" doesn't specify a year, and is old information (comment on article *2022 West Java earthquake*)
Section organization problem	5	too many subsections in the "Recovery and Rehabilitation" section (comment on article *2022 West Java earthquake*)
		I do not like how the article is organized, with too many headers cluttering the article, making it not as readable. Other than that, the AI did great work on the piece. (comment on article *2022 Crimean Bridge explosion*)

Table 11: Summary of major issues found on articles produced by STORM.

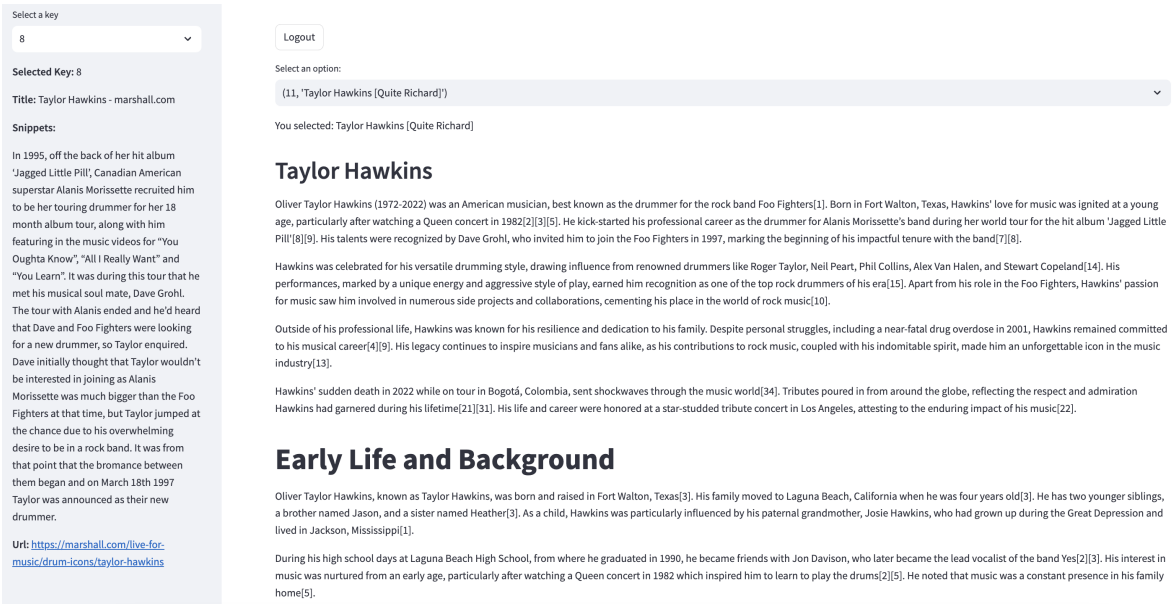

Figure 7: Screenshot of the web application for evaluating the grounded article.

Taylor Hawkins

Oliver Taylor Hawkins (1972–2022) was an American musician, best known as the drummer for the rock band Foo Fighters[1]. Born in Fort Walton, Texas, Hawkins' love for music was ignited at a young age, particularly after watching a Queen concert in 1982[2][3][5]. He kick-started his professional career as the drummer for Alanis Morissette's band during her world tour for the hit album 'Jagged Little Pill'[8][9]. His talents were recognized by Dave Grohl, who invited him to join the Foo Fighters in 1997, marking the beginning of his impactful tenure with the band[7][8].

Hawkins was celebrated for his versatile drumming style, drawing influence from renowned drummers like Roger Taylor, Neil Peart, Phil Collins, Alex Van Halen, and Stewart Copeland[14]. His performances, marked by a unique energy and aggressive style of play, earned him recognition as one of the top rock drummers of his era[15]. Apart from his role in the Foo Fighters, Hawkins' passion for music saw him involved in numerous side projects and collaborations, cementing his place in the world of rock music[10].

Outside of his professional life, Hawkins was known for his resilience and dedication to his family. Despite personal struggles, including a near-fatal drug overdose in 2001, Hawkins remained committed to his musical career[4][9]. His legacy continues to inspire musicians and fans alike, as his contributions to rock music, coupled with his indomitable spirit, made him an unforgettable icon in the music industry[13].

Hawkins' sudden death in 2022 while on tour in Bogotá, Colombia, sent shockwaves through the music world[34]. Tributes poured in from around the globe, reflecting the respect and admiration Hawkins had garnered during his lifetime[21][31]. His life and career were honored at a star-studded tribute concert in Los Angeles, attesting to the enduring impact of his music[22].

 # Early Life and Background

Oliver Taylor Hawkins, known as Taylor Hawkins, was born and raised in Fort Walton, Texas[3]. His family moved to Laguna Beach, California when he was four years old[3]. He has two younger siblings, a brother named Jason, and a sister named Heather[3]. As a child, Hawkins was particularly influenced by his paternal grandmother, Josie Hawkins, who had grown up during the Great Depression and lived in Jackson, Mississippi[1].

During his high school days at Laguna Beach High School, from where he graduated in 1990, he became friends with Jon Davison, who later became the lead vocalist of the band Yes[2][3]. His interest in music was nurtured from an early age, particularly after watching a Queen concert in 1982 which inspired him to learn to play the drums[2][5]. He noted that music was a constant presence in his family home[5].

Despite facing certain hardships during his upbringing, including his mother's struggles with "demons", Hawkins pursued his musical ambitions[4]. He credits his older sister Heather for taking care of the family during difficult times[4].

His first major musical experience came from playing drums for Alanis Morissette's album, Jagged Little Pill, and accompanying her on the subsequent tour[3]. This marked the beginning of his professional career in the music industry.

 # Career

Taylor Hawkins began his professional music career playing in Alanis Morissette's band during her 18-month world tour in support of the hit album 'Jagged Little Pill' from 1995 to 1997[8][9]. His performances not only in the tour but also in the music videos for "You Oughta Know", "All I Really Want" and "You Learn" introduced him to the world of rock music and ultimately led to his meeting with Dave Grohl[8]. Throughout this time, Hawkins contributed significantly to the band's sound and performance, transforming the songs from their original drum loop format to a rock-band vibe that resonated with audiences[1][7].

In 1997, Hawkins was asked by Grohl to join the Foo Fighters, an invitation that he readily accepted[7][8]. At the time, Grohl thought it was a long shot to recruit Hawkins given that Morissette was at the height of her career, but Hawkins' desire to be a part of a rock band compelled him to make the move[7]. This marked the beginning of Hawkins' tenure as the drummer of the Foo Fighters, a role that he would play until his passing[6][9].

Apart from his work with Morissette and the Foo Fighters, Hawkins had an array of other musical experiences[10]. He drummed for Sass Jordan before joining Morissette's touring band[10]. He was part of an ad hoc drum supergroup called SOS Allstars and filled the void for Coheed and Cambria's 2007 album after their drummer Josh Eppard left the group[10]. In addition, Hawkins formed his own side project, the Coattail Riders, in 2005, through which he recorded his own music and took the project on the road, performing in small clubs despite the Foo Fighters' arena-status[7]. His son, Shane Hawkins, has since taken on his father's legacy, joining the Foo Fighters for a performance during the Boston Calling Music Festival in 2023[6].

 # Musical Style and Influences

Taylor Hawkins was a profound drummer, with his musical style and influences spreading across a wide array of rock genres[11]. Known for his passionate fandom of groups that came before him, Hawkins regularly expressed his admiration for bands like Rush, Genesis, and the Police, all of which featured some of the greatest drummers in rock history like Neil Peart, Phil Collins, and Stewart Copeland[11].

He was heavily influenced by his love for classic rock, as evidenced by his performances, where he covered songs from bands like Van Halen[11].

Hawkins drew influences from a variety of drumming styles, developing a signature style inspired by greats like Roger Taylor, Neil Peart, Phil Collins, Alex Van Halen, and Stewart Copeland[14]. This distinctive style and influence extended to his drum kit, which incorporated elements like rototoms and concert toms[14].

Beyond his influences, Hawkins had a unique energy that made him stand out as a drummer. His performances were recognized for their power, and he was known for his enthusiastic and aggressive style of play[15]. This earned him recognition as one of the top rock drummers of his time, with his passion for music living on through his performances[14].

Through his career, Hawkins left an indelible mark on rock music, through his distinct style, passion, and contributions to the music industry[13]. His love for music and dedication to his craft made him an unforgettable icon in the world of rock music[13].

 # Personal Life

Taylor Hawkins married Alison Hawkins, an American celebrity and entrepreneur, in 2005[18]. The couple had three children, Oliver, Annabelle, and Everleigh[19]. Hawkins' commitment to his family was evident; in fact, he even wrote a song for his middle child, Annabelle[9].

In his personal life, Hawkins had also struggled with drug use, which nearly claimed his life in a 2001 overdose[9][7][4]. However, he managed to overcome this challenge, and later expressed gratitude for the experience as a lesson that allowed him to realize the destructive path he was on[7].

Outside of his main role in the Foo Fighters, Hawkins also pursued various side projects including the Birds of Satan, NHC, and Chevy Metal. His motivation for such ventures was a constant drive to create and his love for music[7]. Hawkins was also known for his unabashed fanboy nature, often vocalizing his admiration for fellow musicians and his heroes[7].

 # Legacy and Impact

Taylor Hawkins was known for his raw and authentic drumming style, described as "courageous, damaged and unflinchingly authentic"[20]. His work with the Foo Fighters, as well as his various collaborations and side projects, made him a celebrated figure in rock 'n' roll[10].

Hawkins' death in 2022 was met with heartfelt tributes from colleagues and fans around the world. Notable tributes came from rock legends like Roger Taylor of Queen, who considered Hawkins as a kind, brilliant man and an inspirational mentor, likening his death to "losing a younger favourite brother"[21]. Similarly, Led Zeppelin's Jimmy Page admired his technique, energy and spirited enthusiasm[21].

An LA tribute concert held in his honor included guest drummers like Lars Ulrich of Metallica, Travis Barker of blink-182, and Brad Wilk of Rage Against the Machine. Singers like Miley Cyrus and Alanis Morissette also performed at the concert[22].

Apart from his music, Taylor Hawkins also contributed to charities Music Support and MusiCares, both of which were chosen by the Hawkins family[23]. He had received numerous accolades throughout his career, including 27 Grammy nominations, of which he won 14[2]. In 2021, the Foo Fighters were inducted into the Rock and Roll Hall of Fame[9].

 # Discography

Taylor Hawkins also led a notable music career through his own side projects and collaborations[10]. Aside from his work with the Foo Fighters, Hawkins formed and fronted the band Taylor Hawkins & The Coattail Riders, a project which originated from jamming sessions with his friend Drew Hester[10].

Taylor Hawkins & The Coattail Riders

Taylor Hawkins & The Coattail Riders, a band formed in 2004, have released three albums and their music spans genres including Hard Rock, Art Rock, and Alternative Rock[24][25][26]. The band grew from an initial casual jamming session, gradually evolving into a more formal arrangement that led to the production of record albums. Notably, these albums featured guest appearances by renowned musicians such as Dave Grohl, Queen's Brian May and Roger Taylor, The Cars' Elliot Easton, Perry Farrell, and Jon Davison, who is a school friend of Hawkins'[10].

Red Light Fever

Red Light Fever, released on April 19, 2010, was the band's first album[29][30]. Prior to its release, Hawkins revealed in an interview that the album had completed the recording and production stages, but its title and release date were yet to be determined[29]. Red Light Fever was recorded at the Foo Fighters' Studio 606 in California and featured guest musicians such as Brian May and Roger Taylor of Queen, Dave Grohl of Foo Fighters, and Elliot Easton of The Cars[29][30].

Get the Money

Get the Money, the third album from Taylor Hawkins & The Coattail Riders, was released on November 8, 2019[29]. The album's first single, "Crossed the Line", released on October 15, 2019, featured Dave Grohl and Jon Davison, the frontman of Yes[29]. The music video for the single "I Really Blew It" also featured appearances from Grohl and Perry Farrell[29].

 # Collaborations and Guest Appearances

Throughout his career, Taylor Hawkins collaborated with various prominent artists and bands. The Coattail Riders' albums notably featured appearances from luminaries such as Brian May and Roger Taylor of Queen, Chrissie Hynde, Nancy Wilson of Heart, Sex Pistol Steve Jones and James Gang's Joe Walsh[28]. Hawkins also fronted another group, The Birds of Satan, which evolved from his heavy rock covers band, Chevy Metal[28].

Despite his diverse musical engagements, Hawkins always maintained a close allegiance with the Foo Fighters, which remained the center of his music life[7][28].

 # Tragic Passing

Taylor Hawkins, the esteemed drummer of the alt-rock band Foo Fighters, passed away suddenly on March 25, 2022, while on tour with his band in Bogotá, Colombia[34]. The official cause of death was cardiac arrest, though inquiries were raised concerning the presence of drugs in his system and their potential contribution to his death[33][34]. On the night of his passing, paramedics were called to the Four Seasons hotel in Bogotá due to reports of chest pain from an unnamed guest, later revealed to be Hawkins[34]. Unfortunately, resuscitation efforts were unsuccessful, and Hawkins was declared dead at the scene[34].

The news of Hawkins' sudden demise was announced on the morning of March 25th, 2022, which left the music world in shock[32]. The band confirmed the news with a short statement, expressing their devastation at the loss of Hawkins, whose "musical spirit and infectious laughter" would live on forever[32].

As a result of Hawkins' untimely passing, the band canceled their ongoing South American tour[33]. The festival stage at the Estéreo Picnic Festival, where the Foo Fighters were scheduled to perform that night, was transformed into a candlelight vigil in memory of Hawkins[33].

Tributes and Remembrances

In the wake of Hawkins' death, tributes from fans and colleagues alike poured in from around the world[21][31]. Among the many paying their respects were legendary rock and roll musicians like Roger Taylor, the drummer of Queen, who Hawkins credited with inspiring his own career behind the drum set[21]. In heartfelt social media posts, Taylor described Hawkins as an "inspirational mentor" and a "kind brilliant man"[21], while Led Zeppelin's Jimmy Page reminisced about sharing the stage with Hawkins and praised his "technique, energy and spirited enthusiasm"[21].

There were also numerous onstage tributes to Hawkins. Notably, Miley Cyrus expressed her grief and sent peaceful wishes to the Foo Fighters and the Hawkins family during a performance at Lollapalooza[31]. Similarly, Liam Gallagher of Oasis dedicated one of the band's biggest hits to Hawkins during a concert at the Royal Albert Hall in London[31].

Fans gathered outside the hotel where Hawkins died, lighting candles, leaving flowers, and singing the band's songs in his honor[31].

Hawkins' life and career were celebrated in a star-studded tribute concert in Los Angeles, which saw performances from over 50 musicians, including his former bands and colleagues from Def Leppard, Queen, and Foo Fighters[22].

Table 12: STORM's generated article for "Taylor Hawkins". "#", "##" indicate the section title and subsection title respectively. Numbers in brackets indicate the cited references.