DeFine: A Decomposed and Fine-Grained Annotated Dataset for Long-form Article Generation

Anonymous ACL submission

Abstract

Long-form article generation (LFAG) presents challenges such as maintaining logical consistency, comprehensive topic coverage, and narrative coherence across extended articles. Existing datasets often lack both the hierarchical structure and fine-grained annotation needed to effectively decompose tasks, resulting in shallow, disorganized article generation. To address these limitations, we introduce **DeFine**, a **De**composed and **Fine**-grained annotated dataset for long-form article generation. DeFine adopts a hierarchical decomposition strategy, segmenting the generation process into three stages: outline creation, abstract extraction, and article generation. This approach ensures granular control and provides detailed annotations at each stage. By integrating domain-specific knowledge and multi-level annotations, DeFine enhances the depth, accuracy, and coherence of generated articles, offering a robust solution for long-form article generation across various domains.

To validate the effectiveness of DeFine, we designed and tested three LFAG methods: the web retrieval pipeline, the local retrieval pipeline, and the grounded reference pipeline. We fine-tuned the Qwen2-7b-Instruct model using the DeFine training dataset. The experimental results showed significant improvements in text quality, specifically in topic coverage, depth of information, and content fidelity. These findings highlight the robustness of the DeFine dataset in addressing core issues in article generation, making it a powerful tool for enhancing both the structural integrity and factual accuracy of long-form articles.

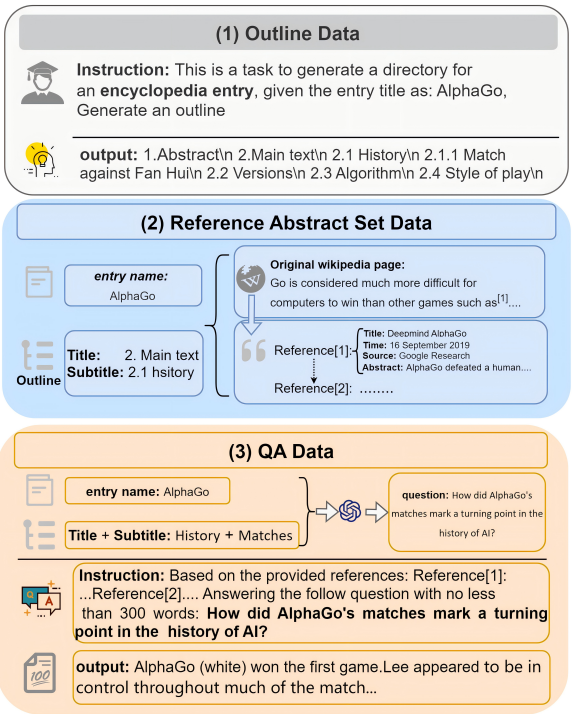

Figure 1: DeFine dataset example. In our dataset, there are three distinct types of data: outline data, abstract set data, and question-answer data. These are utilized to train the model for different functions: generating outlines based on a given article topic, extracting summaries from retrieved references, and generating long-form text based on the extracted summaries and outlines.

1 Introduction

The task of Long-Form Article Generation (LFAG) aims to enable large language models to automatically generate articles that are comprehensive, coherent, and factually accurate (Kasneci et al., 2023; Bubeck et al., 2023). This task requires models to not only cover relevant topics in depth but also maintain consistency in logical structure and narrative flow (Chen et al., 2023). However, achieving high-quality long-form article generation remains a significant challenge due to the complexity of handling large volumes of information and maintaining coherence over extended texts.

With the advent of large-scale pre-trained language models such as BERT (Devlin, 2018) and GPT-3 (Brown, 2020), significant advancements have been made in text generation. For example, STORM (Shao et al., 2024) integrates large lan-

guage models with structured reference information, improving article generation quality through a two-stage generation strategy. Similarly, InstructGPT (Ouyang et al., 2022) enhances models' ability to follow complex instructions. Despite these advancements, current methods for long-form article generation face two major challenges: lack of structured task decomposition and insufficient detail in annotations for guiding generation.

Recent work, such as the STORM system (Shao et al., 2024), often relies on complex prompt engineering for outline construction. However, this approach frequently results in logically inconsistent and loosely structured outputs. This underscores the limitations of current technologies in handling complex long-form article generation (Spataru et al., 2024; Cho et al., 2018), particularly due to the lack of hierarchical decomposition and fine-grained guidance. A significant bottleneck in LFAG research is the absence of high-quality datasets specifically designed for both outline construction and detailed long-form article generation. Existing datasets like WikiSum (Cohen et al., 2021) focus on document summarization but lack task decomposition, resulting in shallow content. In addition, while WikiSum provides useful summaries, it does not facilitate the nuanced understanding necessary for generating structured outlines. Similarly, NarrativeQA (Kočiskỳ et al., 2018) offers narrative data but lacks the hierarchical structure needed for consistent output. Even datasets like Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017), while useful for question-answer tasks, do not provide the fine-grained annotations necessary for guiding detailed long-form generation.

To address these limitations, we propose the DeFine dataset, a high-quality dataset tailored for long-form article generation. Inspired by how educators guide students through the writing process by first drafting an outline, then researching, and finally synthesizing information into a full-length article (Saqr et al., 2018; Alamelu, 2020; Kallestinova, 2011), we decomposed the dataset construction into three stages: outline generation, abstract extraction, and content generation. In the outline generation phase, we directly extracted paragraph structures from Wikipedia articles to ensure that the generated text adheres to natural language logic. We then applied a BGE-M3-based (Multi-Granularity) system for reference extraction and annotated metadata to improve citation accuracy. We developed a hallucination detection algorithm (HDACR)[1] to ensure the factual reliability and consistency of the content (Reimers, 2019; Robertson et al., 2009; Zellers et al., 2019).

We validated the effectiveness of the DeFine dataset by evaluating three pipelines: web pipeline, local retrieval pipeline and grounded reference pipeline. Experimental results show that DeFine significantly improves logical coherence, information consistency, and citation accuracy in generated long-form articles.

Our main contributions include:

1) We propose a hierarchical decomposition strategy and a fine-grained annotation method for constructing dataset;

2) We built a high-quality DeFine dataset covering a wide range of topics, which improves model performance in long-form article generation tasks;

3) We developed and evaluated three pipeline methods, and experimental results confirm the significant effectiveness of the DeFine dataset in improving article generation quality and model performance.

2 Related Work

2.1 LFAG Datasets

High-quality datasets are foundational for advancing research on LFAG. Many existing datasets are widely studied for other tasks, such as question-answering (Joshi et al., 2017; Rajpurkar, 2016; Choi et al., 2018; Dunn et al., 2017; Yang et al., 2018), how to use it for long-form article generation is less investigated. WikiSum (Cohen et al., 2021) integrates multiple documents to create Wikipedia-like articles, but it lacks detailed annotations regarding paragraph structure and reference citations, limiting the model's ability to grasp and replicate complex content organization. PubMed (Jin et al., 2019), while excellent for scientific summarization tasks, is restricted to the medical domain and lacks diversity. WikiHow (WikiHow) provides "how-to" articles with summaries, but its informal style and limited topical coverage reduce its applicability in broader generation articles. NarrativeQA (Kočiskỳ et al., 2018) provide richer narrative content, but the lack of task decomposition hampers effective hierarchical article generation.

[1]The pseudocode for the hallucination detection algorithm is provided in Appendix B.

To the best of our knowledge, most existing datasets for long-form article generation either focus on narrow domains or lack the necessary structural annotations to support complex article generation. In contrast, the data in our DeFine dataset is designed with a clear hierarchical decomposition, allowing for fine-grained control over the article generation process. This includes comprehensive outline generation, detailed reference utilization, and accurate citation metadata, which collectively ensure the production of coherent and factually consistent long-form content. This makes DeFine highly versatile and effective across various domains.

2.2 Long-form Writing Methods

Several approaches have emerged for tackling the challenges in LFAG. Systems like STORM (Shao et al., 2024) leverage large language models and structured reference information to improve text quality through multi-stage generation strategies. Similarly, PLANET (Hu et al., 2022) uses autoregressive attention mechanisms to generate consistent long-form article by dynamically planning content. However, both methods struggle with integrating diverse sources and maintaining logical coherence across extended content.

Enhancing factual accuracy in long-form article has also been a critical focus. BARTScore (Yuan et al., 2021), which utilizes BART as a foundation for evaluating the fluency and factuality of generated content, improves generation accuracy but still faces challenges when handling a wide variety of topics and multi-paragraph article.

In addition, other systems such as LONGWRITER (Bai et al., 2024) have optimized model architectures specifically to improve the quality of long-form article generation, particularly excelling in generating complex narrative content. However, it still faces challenges in integrating diverse information and maintaining consistency throughout the content.

2.3 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) techniques enhance language models by integrating external knowledge during inference. WebGPT (Nakano et al., 2021) employs web searches to boost accuracy and information coverage, while WebGLM (Liu et al., 2023) and WebCPM (Qin et al., 2023) combine real-time retrieval with generative capabilities. Despite these advancements, challenges remain in producing coherent long-form article, particularly in maintaining consistency across multi-paragraph article. AutoWebGLM (Lai et al., 2024) refines this process using automated retrieval strategies, enhancing the timeliness of the information. However, these methods still introduce noise, impacting both the accuracy and the coherence of the generated content.

More recent work, like Realm (Guu et al., 2020) and RAG (Lewis et al., 2020), employs a retrieval-generation architecture that integrates documents from large repositories to improve question answering. However, their reliance on retrieval remains a significant hurdle for tasks requiring well-structured, long-form article, where maintaining content flow and minimizing information noise are key concerns.

3 DeFine Dataset

3.1 Dataset Construction

The DeFine dataset is constructed using data from Wikipedia articles, which provide detailed explanations centered on specific entities. Our research focuses on generating long-form articles from scratch, similar to those on Wikipedia, with an emphasis on speed and accuracy. While existing studies typically operate within narrower domains and generate shorter segments, they often lack generalizability. Moreover, previous work rarely provides reference materials for generation tasks, despite the need for extensive external knowledge when producing Wikipedia-like long-form article.

Although ProxyQA (Tan et al., 2024) introduces a framework for assessing long-form generation through meta-questions, its real-world applicability remains limited. DeFine dataset focuses on enhancing models' capabilities in outline planning, reference comprehension, and utilization during long-form article generation tasks. Specifically, given a topic T, the task involves generating an outline O, retrieving relevant references R based on the outline headings Q, and finally producing a complete article P. This process is represented by the following formula:

$$P = h(f(T), g(f(T))) \tag{1}$$

Where $f(T)$ generates an outline O from the topic T, $g(f(T))$ retrieves the reference set R based on the outline O, and $h(O, R)$ produces the final article P using both the outline O and the reference set R.

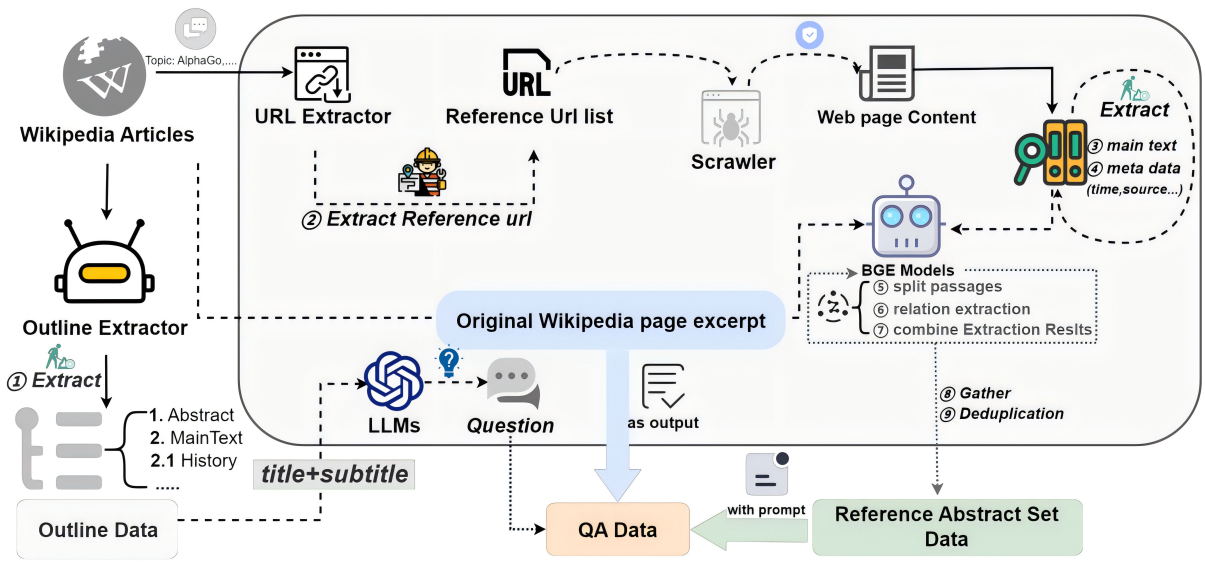

Figure 2: Overview of the DeFine dataset construction. The dataset includes three distinct types of data: outline data, abstract set data, and question-answer data. These data are used to train models to generate outlines, extract summaries from references, and produce long-form articles.

The key advantage of DeFine dataset lies in its hierarchical decomposition, which guides the generation of high-quality long-form articles through three stages: outline data, abstract set data, and question-answer data. Compared to other datasets, DeFine not only covers the complete article generation process but also emphasizes the generation of outlines and the effective utilization of references, ensuring both the coherence and factual accuracy of the content.

3.1.1 Hierarchical Decomposition Strategy

The LFAG dataset employs a hierarchical decomposition strategy for dataset construction, as illustrated in Figure 2. We break down the task into manageable subtasks: outline extraction, reference retrieval and summarization, and question-answer pair construction.

Extracting outline data. This step involves using web scraping techniques to extract structural elements like titles and subtitles from high-quality Wikipedia articles, thus creating outline data. This systematic organization of outline data ensures that generated articles maintain a coherent and logical flow, mirroring the informational depth and accuracy of the original Wikipedia entries.

Reference Retrieval and Summary Extraction. We parse and segment the content of Wikipedia articles, retrieving and extracting text from referenced URLs. Employing the BGE-m3 (Chen et al., 2024) relation extraction model, we compress this data into content-abstract data that encapsulate essential information, enhancing the model's accuracy and information density in long-form article generation.

Question-Answer Data. We segment the abstract data and original paragraphs into sentences, then, we developed a hallucination detection algorithm to calculate entity coverage as a filtering criterion to ensure the quality and relevance of the abstract data. After filtering the abstract data, multiple prompt templates are set up for each original paragraph based on its length, as shown in Appendix A. The "Question" is generated by Qwen2-72b-Instruct model[2] (qwe, 2024), focusing on the original entry's topic and the subheading corresponding to the paragraph, such as "AlphaGo - History - Match against Fan Hui," leading to a question like "How did AlphaGo's match against Fan Hui mark a turning point in its history?". Diversified prompts improve the diversity of the generated content, avoiding monotonous or repetitive outputs and covering more information points. Furthermore, they enhance the model's robustness by testing its performance in different contexts and identifying weaknesses for improvement.

3.1.2 High-Quality Data Cleaning

To ensure the high quality of the dataset, we conducted rigorous data cleaning from three different

[2]The question is generated by configuring prompts such as: "Given the topic AlphaGo, and the subtitle History - Match against Fan Hui, please generate a question based on these two titles."

perspectives, filtering the data to guarantee its accuracy and consistency. Our efforts focused primarily on addressing the following three issues:

Q1: Is the content of the Wikipedia entries sufficiently rich? The first challenge in constructing the dataset is ensuring that the selected Wikipedia entries are both detailed and information-rich. We establish strict selection criteria, including minimum word count and a required number of references, to guarantee depth and breadth in the entries.

Q2: How do we determine the relevance of extracted content from references to the original text? We analyze the references for each article to ensure they encapsulate the key information. Using a BGE-M3 model-based relation extraction method (Chen et al., 2024), we identify key entities and their relationships with the original text, allowing us to extract relevant sentences and form concise summaries, thus improving precision and filtering out irrelevant content.

Q3: Can the extracted reference content fully cover the corresponding original text? To verify this, we developed the *Hallucination Detection Algorithm for Citation Reliability (HDACR)*. HDACR identifies entities in both reference and generated texts, evaluates matches, and detects hallucinations. It generates a report with detection results, reference information, and unverifiable entities. By applying HDACR, we filter out references that inadequately cover the original text, enhancing dataset accuracy and text quality while providing detailed validation for long-form article generation.

3.2 Dataset Analysis

Data Quantity and Categories. DeFine dataset contains 6,502 question-answer data, 9,647 abstract pair data, and 52,045 outline data, covering both Chinese and English languages across various domains, including science, humanities, history, geography, literature, medicine, and sports. Figure 3 reflects the composition of the Wikipedia articles used to construct the dataset. English articles make up 77% of the source material, while Chinese articles account for 23%. In terms of subject matter sourced from Wikipedia, biographies have the highest proportion (21.56%), followed by geography and environment (16.63%), and humanities and arts (14.82%). In contrast, sports and leisure articles, which also contribute to the dataset, have the lowest proportion, comprising only 5.54% of the articles used.

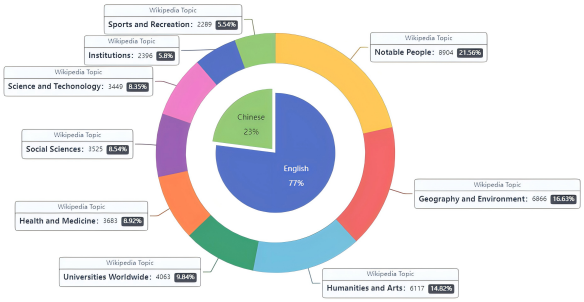

Figure 3: Wikipedia pages distribution across different types.

Answer Type and Length. The outline data contains an average of 15 subheadings per entry, providing a detailed framework for generating structured long-form articles. For QA data, each entry includes an average of four reference summaries, with the maximum number of reference summaries reaching 42. The answer sections in the QA data exhibit an average word count of 375 words per answer. The longest answer contains 2,919 words, while the shortest consists of 150 words. This wide range in reference and answer lengths highlights the dataset's diversity and richness, enabling models to generate both concise and elaborate long-form articles.

4 Experiments

4.1 Dataset Split

In this study, to validate the effectiveness of our dataset in long-form article generation tasks, we randomly selected 2,286 outline data and 3,625 question-answer data with an equal split between Chinese and English from the DeFine dataset to fine-tune the Qwen2-7b-Instruct model (qwe, 2024) as the training set.

Additionally, to further enhance the model's capabilities in information extraction and compression tasks, we utilized abstract pairs to train the BGE-m3 relation extraction model. This model is employed to extract key information from references and condense the full-text content into summaries through precise relation extraction and compression.

From the remaining data, 270 outline data samples and 270 QA data samples were reserved to evaluate the outline generation performance and long article generation performance of the fine-tuned model. Appendix C shows the detailed partition information for dataset. The fine-tuned model is referred to as ***Qwen2-7b-Scribe***. To validate the

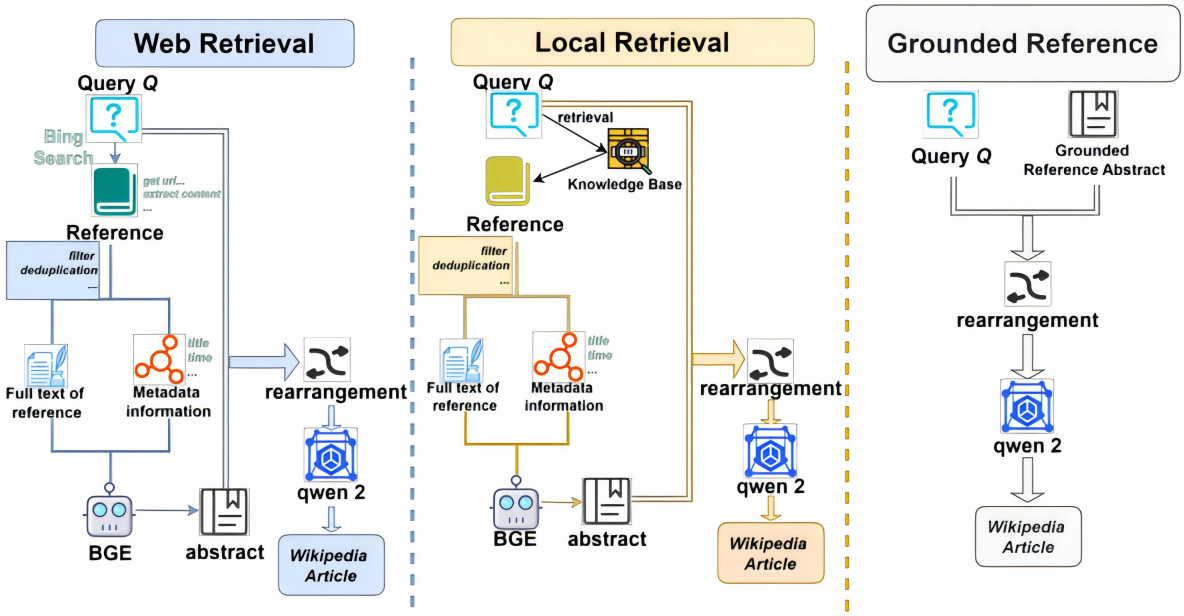

Figure 4: Overview of the three pipelines. Different tasks are performed for each pipeline. Notably, the relationship extraction model employed in the pipelines uses the trained *BGE-M3 model*. Additionally, both the generation model and the relationship extraction model in the pipelines are interchangeable, allowing for flexibility in adapting to different research needs or experimental settings.

effectiveness of our methods, we selected 230 different topics from various domains, including 130 topics in Chinese and 100 in English, to generate full-length articles.

4.2 LFAG Methods

We propose three LLM-based pipeline methods, each optimized for different retrieval and generation strategies within a unified framework.

Web Retrieval: This pipeline combines real-time web retrieval with a generation model to fetch the latest information. We use the Google SERPER API[3] as the retrieval engine and the Qwen2-7b-Scribe model, trained on the DeFine dataset, for generation. Both components are modular and can be replaced if needed, offering high timeliness and dynamic data integration into generated articles.

Local Retrieval: This pipeline retrieves references from a high-quality internal knowledge base, constructed from Wikipedia articles. The system retrieves relevant content based on the article's topic and subheadings, extracts key information, and generates long articles using the model.

Grounded Reference: This approach directly inputs the article's topic, subheadings, and Wikipedia original abstracts into the generation model without retrieval, generating the complete content.

[3]https://google.serper.dev/search

4.3 Evaluation Metrics

To comprehensively evaluate the performance of the model across different tasks, we combined quantitative and qualitative metrics, covering both outline generation and long-form article generation. For outline evaluation, we adopted the Heading Soft Recall and Heading Entity Recall metrics. The former measures the alignment between the generated outline and the reference outline at the thematic level, allowing for some degree of fuzzy matching. The latter assesses the coverage of key entities (such as people and locations) in the generated outline, ensuring the overall quality of the outline generation.

To assess the quality of the generated long articles, we used *ROUGE scores* (Lin, 2004), and calculate the entity recall based on FLAIR NER results at the paragraph level. Additionally, based on domain-specific criteria, we evaluated the generated full length article from the following aspects: (1) *Interest Level*, (2) *Coherence and Organization*, (3) *Relevance and Focus*, and (4) *Coverage*. For the evaluation of these aspects, we employed Prometheus (Kim et al., 2023), a 13B evaluator model, utilizing a scoring rubric developed based on the STORM system (Shao et al., 2024). The rubric assesses article's relevance, coherence, and overall informativeness.

	Comparision 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
Web Retrieval GPT 3.5	50.54	36.24	20.21	4.55	4.83	4.75	4.82
Web Retrieval Qwen2-7b-Scribe	**53.39**	**37.54**	**22.85**	**4.85**	**4.92**	**4.82**	**4.87**

Table 1: Comparison of different models with human-written articles on several metrics. The ROUGE-1, ROUGE-L, and Entity Recall scores are reported as percentages (%) to compare the quality of generated article, while the rubric grading evaluates Interest Level, Organization, Relevance, and Coverage, also expressed as percentages (%). The results marked with bold indicate the best performance, achieved by the Qwen2-7b-Scribe model. The results for DirectGen, RAG, oRAG, and STORM are taken from (Shao et al., 2024).

Models	Head.S Recall	Head.E recall
DeFine		
Qwen2-7b-Instruct	65.15	47.18
Qwen1.5-7b-chat	60.53	46.19
GPT-3.5	77.62	36.66
Llama3-8b	75.15	47.84
Qwen2-7b-Scribe	**84.01**	**71.16**
FreshWiki		
STORM	86.26	40.52
GPT-3.5	81.33	33.62
Llama3-8b	74.31	32.15
Qwen2-7b-Scribe	84.90	**43.34**

Table 2: Comparison of different models on two datasets, DeFine and FreshWiki. The Heading Soft Recall (Head.S) and Heading Entity Recall (Head.E) metrics are reported as percentages (%). Values marked with bold indicate the best results. The *STORM* results are taken from the paper (Shao et al., 2024), while the STORM results in my table are based on GPT-3.5.

5 Results and Analysis

5.1 Main Results

We first evalute the outline generate ability. We use *Qwen2-7b-Scribe, Qwen2-7b-Instruct* (qwe, 2024), *Qwen1.5-7b-chat* (Bai et al., 2023), *Llama-3-8b* (AI@Meta, 2024), and *gpt-3.5-turbo* for generation tasks. table 2 shows the heading soft recall and entity recall. By comparing with STORM, The results demonstrate that *Qwen2-7b-Scribe* performed well in generating outlines, particularly with a high Heading Soft Recall score. This indicates that the generated outlines align well with the reference outlines in terms of structure. Additionally, the Heading Entity Recall further evaluated the retention of key entities in the generated outlines.

Among three tested pipelines, *Qwen2-7b-Scribe* excelled, particularly within the Web pipeline, which leveraged extensive Internet resources to outperform others in quality article generation. As shown in table 3, We evaluated *Qwen2-7b-Scribe*, *gpt-3.5-turbo*, and two other 7b models using a QA data test set, analyzing generated texts against human-written paragraphs. The Grounded Reference pipeline, relying on precise references, followed in performance, offering detailed but less timely content. Conversely, the Local Retrieval pipeline underperformed due to its restricted content scope and delayed data retrieval.

Our decomposition approach to long article generation significantly improved the article quality in tests performed on the FreshWiki (Shao et al., 2024) set and our own test set, showing superior logical consistency and accuracy. Utilizing 270 article topics and FreshWiki Dataset (Shao et al., 2024), we generated articles via the web pipeline and compared these against STORM model outputs. As shown in table 1 and table 4. This method's effectiveness was validated by high scores across all evaluation metrics, underscoring the dataset's robustness and the methodology's precision.

5.2 Ablation Study

In our research, the effectiveness of retrieval processes in the context of long-form article generation was a primary focus. To isolate and evaluate the impact of these retrieval mechanisms, we conducted a targeted ablation study, which involved the following experimental setup:

We introduced a "Direct" pipeline, where all retrieval components were completely disabled. This setup allowed the large language models (LLMs) to generate responses based solely on their pretrained knowledge and the initial prompts, without any supplementary data retrieval. This was de-

Comparison with human-written Articles					
Baseline	**Model**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**Entity Recall**
Web Retrieval	Qwen1.5-7b-chat	53.13	30.57	39.15	9.36
	Qwen2-7b-Instruct	50.46	31.50	39.78	8.96
	GPT-3.5	43.98	22.24	31.18	8.25
	Qwen2-7b-Scribe	**53.93**	**32.41**	**39.80**	**14.09**
Local Retrieval	Qwen1.5-7b-chat	40.16	18.07	28.81	6.28
	Qwen-2-7b-Instruct	42.52	21.44	30.42	8.54
	GPT-3.5	41.37	20.17	29.96	7.92
	Qwen-2-7b-Scribe	**44.06**	**24.19**	**32.35**	**12.89**
Grounded Reference	Qwen1.5-7b-chat	46.42	23.11	32.15	10.35
	Qwen2-7b-Instruct	46.70	23.74	32.47	9.38
	GPT-3.5	44.41	22.41	31.26	8.90
	Qwen2-7b-Scribe	**50.79**	**26.82**	**34.93**	**13.61**

Table 3: Comparison of different models on several ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L) and entity recall. Results are shown for three baselines: web retrieval, local retrieval, and ground truth. The highest scores for each baseline are marked in bold, indicating the best performance achieved by the Qwen2-7b-Scribe model.

	Rubric Grading			
	Interest	**Org.**	**Rel.**	**Cov.**
Web Retrieval	**4.87**	**4.95**	**4.71**	**4.55**
Local Retrieval	4.83	4.78	4.71	4.55
Direct	3.85	4.07	3.34	4.02

Table 4: Results of Rubric Grading score for generated article quality evaluation. The Rubric Grading criteria (Interest, Org., Rel., Cov.) are abbreviated for compactness.

signed to test the core capabilities of the models in generating coherent and relevant content independently of dynamic or static data retrieval aids. The purpose of this study was to ascertain the intrinsic value of retrieval systems in enhancing the quality and accuracy of the generated articles. By comparing the "Direct" model's performance against the full implementations of our web and local retrieval pipelines, we aimed to demonstrate the incremental benefits provided by incorporating external data sources into the generation process.

Our findings, presented in Table 4, indicate that models equipped with retrieval capabilities outperformed the "Direct" pipeline across all rubric criteria, confirming the hypothesis that dynamic data access significantly enhances content quality. This disparity highlighted the substantial role that retrieval plays in informing and structuring the generated content, especially in terms of relevance and depth.

This ablation study effectively demonstrates the critical importance of retrieval mechanisms in the LFAG tasks, providing clear evidence that the ability to dynamically integrate external information is pivotal for producing high-quality, accurate, and engaging long-form articles.

6 Conclusion

We introduce DeFine dataset, a novel high-quality resource designed to enhance long-form article generation. By employing a hierarchical decomposition strategy and fine-grained annotation, we addressed key challenges in LFAG task, ensuring a structured approach to generating coherent and detailed articles. We evaluated our dataset across three LFAG Methods, and experimental results show that DeFine significantly improves logical coherence, information consistency, and citation accuracy in generated articles. Our work also demonstrates the effectiveness of integrating both retrieval-based and generation-based approaches, showing the potential of DeFine as a benchmark for future research in LFAG task.

Limitations

Despite the significant contributions of the DeFine dataset to the field of long-form article generation, the current implementation presents certain limitations. One of the primary issues is the imbalance between the English and Chinese data, which stems from the limited number of Chinese Wikipedia entries compared to the far more extensive English Wikipedia. This disparity could affect the model's ability to generalize well across both languages, as the richness and variety of English content provide

more training opportunities, potentially leading to performance gaps in less-represented languages.

Additionally, while the dataset covers a broad range of topics, it does not fully capture the nuances of specialized or highly technical domains. This limitation may restrict the dataset's applicability in scenarios where precise, domain-specific knowledge is essential for generating high-quality content. Current models trained on DeFine might struggle to seamlessly integrate and synthesize information from these highly specialized areas, reflecting the dataset's scope limitations.

Another key challenge lies in the reliance on automated evaluation metrics such as ROUGE. While useful for measuring surface-level similarities between generated and reference texts, these metrics may fail to account for deeper aspects of text quality, such as coherence, logical flow, and factual correctness. The annotation process, although thorough, is also not immune to subjectivity, which can introduce bias and affect the overall balance and representativeness of the dataset.

In light of these limitations, future work should prioritize expanding the dataset to include more balanced language data, particularly for underrepresented languages like Chinese, and enhance its coverage of specialized topics. Furthermore, developing more sophisticated evaluation frameworks that align better with human judgment of text quality will be crucial, as well as improving the robustness of models and algorithms to reduce factual inconsistencies and hallucinations in generated content.

References

2024. Qwen2 technical report.

AI@Meta. 2024. Llama 3 model card.

M Alamelu. 2020. Methods and significance of pre writing activities in acquisition of writing skills. *Solid State Technology*, 63(2s).

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. *arXiv preprint arXiv:2310.07289*.

Woon Sang Cho, Pengchuan Zhang, Yizhe Zhang, Xiujun Li, Michel Galley, Chris Brockett, Mengdi Wang, and Jianfeng Gao. 2018. Towards coherent and cohesive long-form text generation. *arXiv preprint arXiv:1811.00511*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Nachshon Cohen, Oren Kalinsky, Yftah Ziser, and Alessandro Moschitti. 2021. Wikisum: Coherent summarization dataset for efficient human-evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 212–219.

Wikipedia contributors. 2024. Wikipedia: Size of wikipedia. Accessed: 2024-10-14.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3929–3938.

Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. 2022. Planet: Dynamic content planning in autoregressive transformers for long-form text generation. *arXiv preprint arXiv:2203.09100*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Elena D Kallestinova. 2011. How to write your first research paper. *The Yale journal of biology and medicine*, 84(3):181.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1–15, Singapore. Association for Computational Linguistics.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen Zhang, Xiaohan Zhang, Yuxiao Dong, et al. 2024. Autowebglm: Bootstrap and reinforce a large language model-based web navigating agent. *arXiv preprint arXiv:2404.03648*.

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.

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

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4549–4560.

Multi-Linguality Multi-Functionality Multi-Granularity. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

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.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, et al. 2023. Webcpm: Interactive web search for chinese long-form question answering. *arXiv preprint arXiv:2305.06849*.

P Rajpurkar. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Mohammed Saqr, Uno Fors, and Matti Tedre. 2018. How the study of online collaborative learning can guide teachers and predict students' performance in a medical course. *BMC medical education*, 18:1–14.

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

Ava Spataru, Eric Hambro, Elena Voita, and Nicola Cancedda. 2024. Know when to stop: A study of semantic drift in text generation. *arXiv preprint arXiv:2404.05411*.

Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, and Linqi Song. 2024. Proxyqa: An alternative framework for evaluating long-form text generation with large language models. *arXiv preprint arXiv:2401.15042*.

WikiHow. Wikihow: How-to articles for any situation. https://www.wikihow.com/. Accessed: 2024-09-19.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

A Prompt Examples

A.1 Detailed Description of Prompt Examples

The table 5 presents three examples of **prompt templates** designed for long-form question-answer generation based on provided abstracts. These templates help structure the interaction between the question and abstract, ensuring diverse and high-quality responses.

Structure and Functionality: Each prompt is structured to include one or more abstract references (denoted as *{Abstract[1]...Abstract[2]}*) and a dynamically generated question (denoted as *{Question}*). These prompts serve to guide the model in providing detailed, relevant answers while ensuring that the content aligns with the source material.

Prompt Diversity: To ensure that the generated content covers a wide range of information points and avoids repetition, multiple prompt templates are provided. For instance:

- **Prompt 1** requests an answer based on the references, emphasizing completeness.
- **Prompt 2** adds a directive that the model must answer, reinforcing the requirement for response.
- **Prompt 3** simplifies the directive but maintains the expectation of a detailed answer.

Customization for Content Length: The prompts are adaptable for abstracts of different lengths. By adjusting the amount of information in `{Abstract[1]...Abstract[2]}`, the system ensures that both short and long references are effectively processed.

Advantages of Diverse Prompting:

- **Increased Diversity:** Multiple prompts ensure a broader coverage of information, reducing redundancy and improving the model's capability to generate varied responses.
- **Improved Robustness:** These prompts expose the model to different question structures and contexts, improving its ability to respond appropriately across diverse scenarios.
- **Enhanced Alignment:** By consistently referring to abstracts and generating questions dynamically, the prompts ensure that the answers remain aligned with the source information, enhancing both factual accuracy and relevance.

A.2 Example of Generated Prompts

- **Example 1:**
  - Abstract(s): A collection of summaries from referenced papers.
  - Question: *"How did AlphaGo's match against Fan Hui mark a turning point in its history?"*

  Prompt: Based on the provided references, answer the following questions. Please provide detailed answers with a minimum of 300 words: Abstract[1]...Abstract[2]... Question: How did AlphaGo's match against Fan Hui mark a turning point in its history?

- **Example 2:**
  - Abstract(s): Abstracts related to a medical breakthrough.
  - Question: *"What were the key factors that led to the success of the new treatment?"*

  Prompt: You cannot refuse to answer the question. Please refer to the following information: Abstract[1]...Abstract[2]... Answer the following questions. Please provide detailed answers with a minimum of 300 words: What were the key factors that led to the success of the new treatment?

These examples demonstrate how diverse prompting enhances content generation, ensuring a balance between coverage and specificity in the answers.

B Hallucination Detection Algorithm for Citation Reliability

The **Hallucination Detection Algorithm for Citation Reliability (HDACR)** is a crucial component designed to ensure the factual consistency between generated content and reference material in long-form text generation tasks. This algorithm specifically targets the identification of hallucinations, which are instances where the generated content introduces false or unverifiable information that is not present in the reference material. Below is a detailed explanation of the steps involved in the HDACR algorithm, as presented in Algorithm **??**.

Prompt	Prompt Text
Prompt 1	`Abstract[1]...Abstract[2].....`\n Answer the following questions based on the provided references. Please provide detailed answers with a minimum of 300 words: **{Question}**
Prompt 2	You cannot refuse to answer the question. Please refer to the following information:\n`Abstract[1]...Abstract[2].....`\n Answer the following questions. Please provide detailed answers with a minimum of 300 words: **{Question}**
Prompt 3	Based on the provided references, answer the following questions. Please provide detailed answers with a minimum of 300 words:\n`Abstract[1]...Abstract[2].....`\n Question: **{Question}**

Table 5: Prompt example (Using a paragraph length of 200-400 characters as an example.)

B.1 Step-by-Step Breakdown of HDACR

Initialization: The algorithm starts by initializing two sets: the *Reference Content Entity Set* (E_r) and the *Generated Content Entity Set* (E_g), which will store the extracted entities from the reference content (R) and the generated content (G), respectively. An empty *Hallucination Detection Result* (H) is also initialized to store the final outcome of the detection process. **Entity Extraction:** Next, the algorithm extracts entities from both the reference content (R) and the generated content (G). For each model M in the preDeFined model set, the entities from the reference content are extracted and added to E_r. Similarly, entities from the generated content are extracted and added to E_g. This entity extraction process ensures that both texts (reference and generated) are represented by the key entities they contain, facilitating accurate comparison.

Matching Score Calculation: Once the entities are extracted, the algorithm proceeds to calculate a *matching score* for each entity in the generated content (E_g). If an entity from the generated content has a direct match (hard match) in the reference content (E_r), a perfect score of 1.0 is assigned. If no hard match is found, the algorithm performs a *soft matching* using a combination of Sentence-BERT (Semantic-BERT) and BM25 scoring methods. The weighted score is computed by averaging the Sentence-BERT and BM25 scores for each entity. The final score is stored in the matching score list γ.

Hallucination Detection: The hallucination detection decision is based on the matching scores. If any entity in the generated content has a matching score γ_i lower than a preDeFined threshold (0.6 in this case), the algorithm concludes that the generated content contains a hallucination. Otherwise, the content is deemed free of hallucinations.

Output: Finally, the algorithm outputs the hallucination detection result (H), along with the generated content (G), the reference content (R), and a detailed list of unverifiable entities and their positions within the text. This output provides a comprehensive report on the factual consistency of the generated content, helping to ensure its reliability.

The following is the pseudocode for the HDACR algorithm. The HDACR algorithm plays a critical role in improving the accuracy and reliability of long-form generated content. By carefully comparing entities between the reference and generated content, it ensures that no extraneous or fabricated information is introduced. This makes it particularly valuable in domains where factual consistency is crucial, such as scientific or historical document generation. The combination of both hard and soft matching provides flexibility in detecting minor semantic variations while maintaining rigorous detection of factual hallucinations. The output report generated by HDACR offers transparency into the hallucination detection process, making it an essential tool for validating the quality of generated content.

C Dataset Statistics and Distribution

Table 6 provides comprehensive statistics on the various components of the dataset used in this study. Each category is further divided into English (eng.) and Chinese (ch.) data, with columns detailing the number of examples in the test set, training set, and the total number of examples available.

Notably, the English data contains a significantly larger number of samples compared to the Chinese data. This discrepancy arises from the larger volume of English-language Wikipedia entries; it has a far richer and more extensive collection of articles than its Chinese counterpart, with a greater num-

Algorithm 1: Hallucination Detection Algorithm for Citation Reliability (HDACR)

Input: Generated content G and Reference content R
Output: Hallucination detection result H

1 **Initialization**: Reference content entity set $E_r \leftarrow \emptyset$, Generated content entity set $E_g \leftarrow \emptyset$, Hallucination detection result $H \leftarrow$ empty;
2 **Extract Reference Content Entities**: For each model M in the model set, extract entities E_{rM} from the reference content R_r, add E_{rM} to E_r;
3 **Extract Generated Content Entities**: For each model M in the model set, extract entitieE_{gM} from the generated content G, add E_{gM} to E_g;
4 **Calculate Matching Scores**: Initialize the matching score list $\gamma \leftarrow []$. For each entity e in E_g:
- If e matches hard ($e \in E_r$), then $\gamma \leftarrow 1.0$;
- Otherwise, perform soft matching, calculate Sentence-BERT score γ_{SBERT} and BM25 score γ_{BM25}, calculate weighted score $\gamma \leftarrow \frac{\gamma_{SBERT}+\gamma_{BM25}}{2}$, add γ to the matching score list.

Determine the Presence of Hallucinations: If $\exists \gamma_i < 0.6$, mark hallucination detection result: $H \leftarrow$ Hallucination present; otherwise, mark hallucination detection result $H \leftarrow$ No hallucination
Output Results: Output generated content G, hallucination detection result H, reference content R, and unverifiable entities and their positions in the text;

ber of contributors and editors actively engaged in expanding and updating content. According to recent studies, English Wikipedia hosts over 6 million articles, while Chinese Wikipedia has just over 1 million articles (contributors, 2024). This has resulted in a more substantial dataset for English, leading to the selection of a representative subset to avoid redundancy and overfitting in training. By doing so, we aimed to ensure that the training process remains efficient while still providing the model with sufficient variability and representation from the dataset.

Furthermore, the varying depth and detail within entries also contribute to the richness of the English dataset. Articles on complex topics often include extensive references and citations, which enhance the dataset's quality and comprehensiveness. In contrast, the Chinese entries, while valuable, may not always match this level of detail, particularly in less commonly covered subjects. This situation necessitates a careful curation of data to maintain a balanced representation in the model's training set, ensuring that both languages are adequately represented in the resulting outputs.

The distribution of various data types is shown in Figure 5 and Figure 6. Figure 5 illustrates the access distribution of QA data topics in the dataset, while Figure 6 provides similar insights for Abstract data topics. The categories ending with **ch** represent Chinese data, while all other categories correspond to English data. The percentages displayed in the figures indicate the proportion of accesses from each topic, highlighting the dataset's diversity. This rich distribution is essential for training models capable of understanding and generating content across a variety of topics and styles, further enhancing the robustness and applicability of our findings.

Data Type	Test.	Train.	All
Outline data eng.	120	1026	1226
Outline data ch.	150	941	50787
Abstract Set Data eng.	\	200	8525
Abstract Set Data ch.	\	200	1122
QA data eng.	133	1026	5244
QA data ch.	120	1138	1258

Table 6: This table provides detailed information about our dataset. The dataset is categorized into Outline data, Abstract Set data, and QA data, with separate columns for English (eng.) and Chinese (ch.). The backslash ('\') denotes that no test data is available for the Abstract Set in either English or Chinese.

D Case Study

In this appendix, we present examples of the results generated by different models used in our experiments. We provide outlines, QA data, and full generated articles for several topics.

D.1 Outline Generation Examples

Here is the section where we present the outlines generated by different models for the same task and

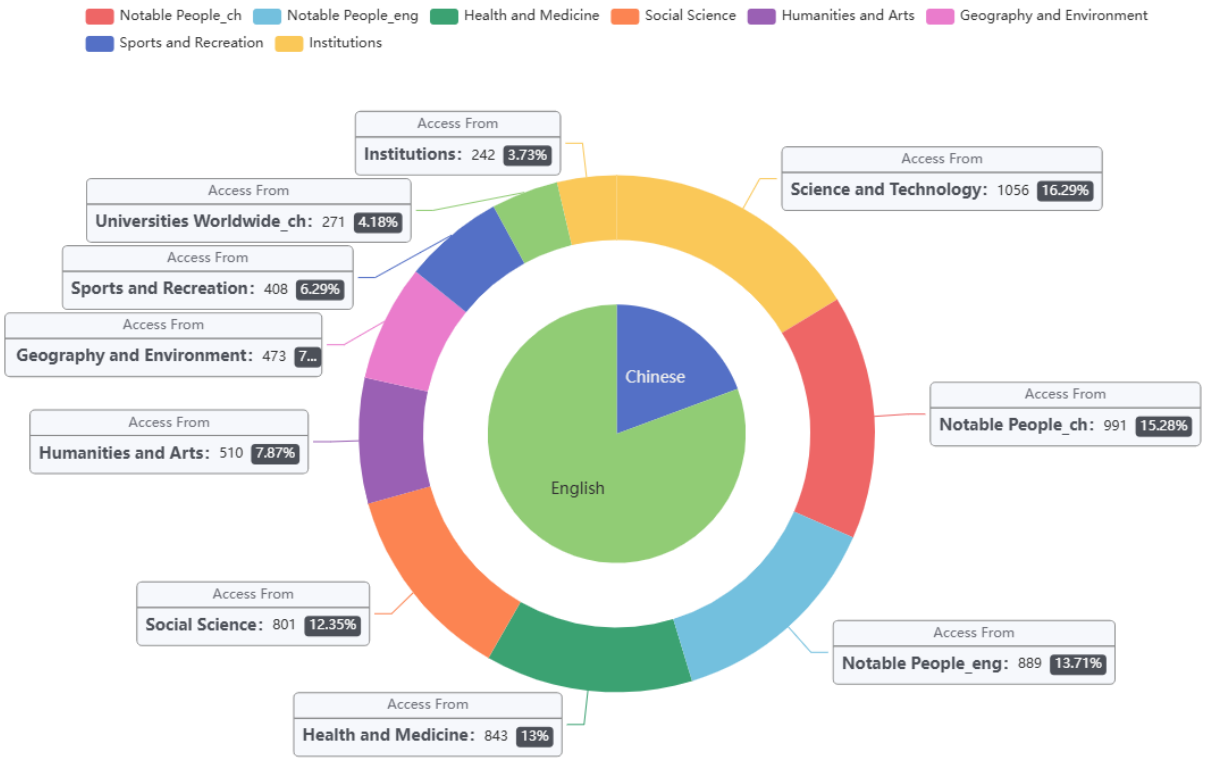

Figure 5: This figure illustrates the access distribution of QA data topics in the dataset. The categories ending with **ch** represent Chinese data, while all other categories are in English. The percentages indicate the proportion of accesses from each topic, highlighting the diversity in the dataset.

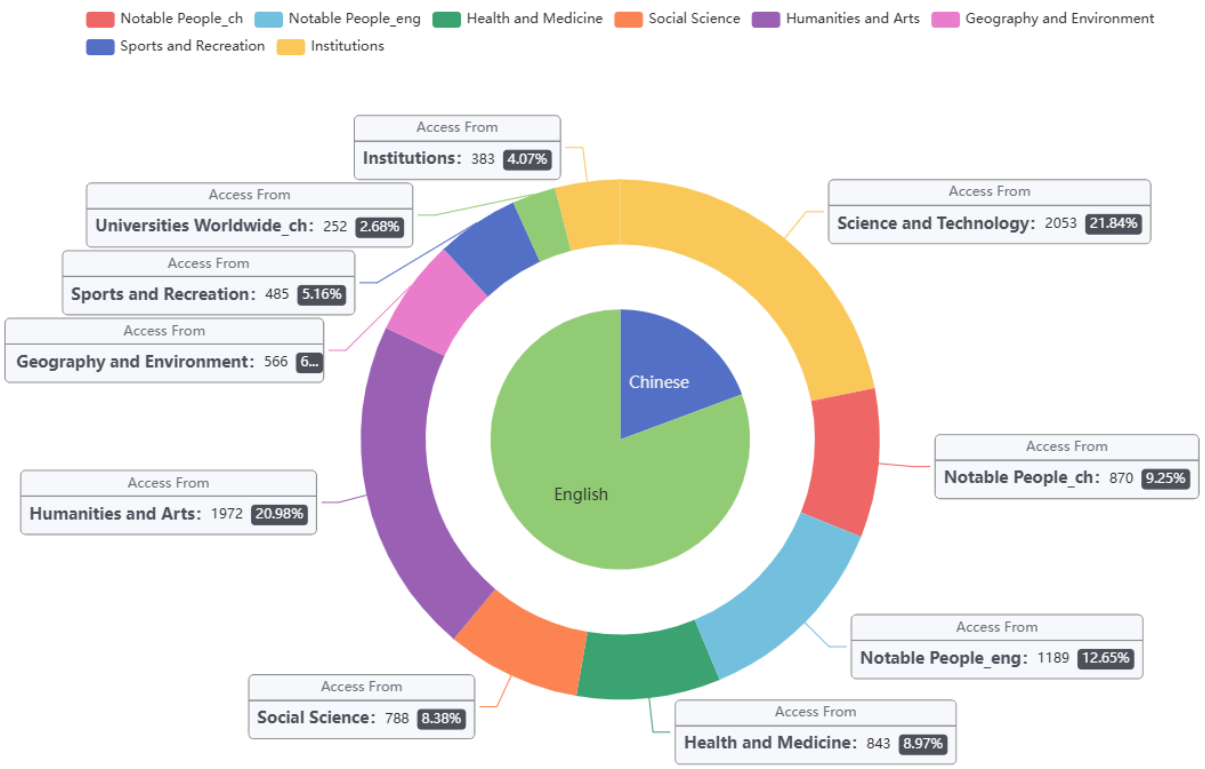

Figure 6: This figure illustrates the access distribution of Abstract data topics in the dataset. The categories ending with **ch** represent Chinese data, while all other categories are in English. The percentages indicate the proportion of accesses from each topic, highlighting the diversity in the dataset.

compare them with the correct answers. The task is to generate a directory for the encyclopedia entries "雄安新区" and "John Gribbel." Below, we list the generated results from each model alongside the correct answers.

D.2 Chinese Outline Generation

Instruction: 这是一个百科词条目录生成任务，给定词条名为：雄安新区，生成一个目录。

D.2.1 Original Outline

1. 摘要
2. 正文
    - 2.1 历史沿革
        - 2.1.1 古代
        - 2.1.2 近代
        - 2.1.3 21世纪
    - 2.2 政治
    - 2.3 行政区划与空间布局
        - 2.3.1 “一主”
        - 2.3.2 “五辅”
        - 2.3.3 “多节点”
    - 2.4 承接疏解
    - 2.5 地理
        - 2.5.1 范围地形
        - 2.5.2 水文
        - 2.5.3 气候
    - 2.6 自然资源
        - 2.6.1 矿藏资源
        - 2.6.2 生物资源
    - 2.7 人口
    - 2.8 教育
    - 2.9 人文
    - 2.10 交通
        - 2.10.1 公路
        - 2.10.2 地面公交
        - 2.10.3 轨道交通
        - 2.10.4 铁路
        - 2.10.5 航空
    - 2.11 影响和评价
        - 2.11.1 考察与合作
        - 2.11.2 各方评论
        - 2.11.3 其他影响
    - 2.12 外部连结
    - 2.13 参见

D.2.2 Llama3 Output

1. 雄安新区概述
2. 建设背景与意义
3. 区域规划与定位
4. 产业布局与发展目标
5. 交通网络与基础设施建设
6. 生态环境保护与建设
7. 人口与社会服务
8. 国际合作与交流
9. 雄安新区的未来展望

D.2.3 Qwen1.5-7b-chat Output

1. 总览
    - 定义
    - 地理位置
2. 基本信息
    - 全称与简称
    - 成立时间
3. 发展理念
    - 新区定位
4. 建设目标
    - 疏解北京非首都功能
    - 创新驱动发展
5. 区域规划
    - 产业发展规划
    - 基础设施建设
6. 建设成就
    - 重大项目
7. 发展阶段
    - 启动阶段
    - 推进阶段
10. 争议与挑战
    - 环境影响
    - 政策实施效果

D.2.4 Qwen2-7b-Instruct Output

1. 新区概述
    - 雄安新区设立背景
    - 雄安新区地理位置
2. 新区规划
3. 新区建设
    - 建设进展
    - 建设重点
4. 新区功能定位
    - 区域协调发展
    - 创新驱动发展
    - 绿色生态宜居
5. 新区产业布局
    - 高端高新产业
6. 新区社会民生
    - 教育资源
7. 新区生态环境
    - 绿色建筑
    - 生态保护
8. 新区国际合作
    - 对外开放政策
    - 国际合作项目

9. 新区未来展望
  - 发展趋势

D.2.5 GPT-3.5

背景与建设动机
雄安新区设立背景
雄安新区规划概况
政策支持与发展定位
基础设施建设
产业规划与发展
生态环境保护
人才引进与培养
医疗与教育资源
文化与旅游特色
交通发展与互联互通
雄安新区的未来展望
城市规划与建设
投资与合作机遇
建设成果与效益评估
社会治理与公共服务
创新科技与创业环境
区域协同发展
雄安新区与长三角经济一体化
国际交流与合作愿景
文献资料与研究参考
政策演变与调整

D.2.6 Qwen2-7b-Scribe Output

一、摘要
二、正文
  2.1 历史
    2.1.1 古代历史
    2.1.2 近代历史
    2.1.3 现代
  2.2 政治背景
  2.3 城市建设
    2.3.1 行政区划
    2.3.2 布局
  2.4 地理环境
  2.5 资源
    2.5.1 矿产资源
    2.5.2 自然资源
  2.6 城市人文
  2.7 城市经济
  2.8 城市交通
  2.9 城市环境
  2.10 城市安全
  2.11 城市教育
  2.12 城市医疗
  2.13 城市住房
  2.14 城市就业
  2.15 城市影响
  2.16 外部评价
  2.17 参见

D.3 John Gribbel

Instruction: This is a task to generate a directory for an encyclopedia entry, given the entry title as: John Gribbel.

Below we present the outlines generated by different models compared to the correct answer.

D.3.1 Original Outline

1. Abstract
2. Main text
  2.1 Early life and education
  2.2 Gas, electricity and other business interests
  2.3 Residences
  2.4 Litigation
  2.5 Publishing
  2.6 Politics
  2.7 History and the Glenriddell Manuscripts
  2.8 Marriage

D.3.2 Llama3 Output

- Biography
- Early Life
- Career Highlights
- Notable Works
- Contributions to the Field
- Awards and Recognition
- Personal Life
- Legacy

D.3.3 Qwen1.5-7b Output

1. Biographical Overview
2. Full Name: John Gribbel
3. Birth and Early Life
4. Birthplace
5. Date of Birth
6. Education and Background
7. Academic Qualifications
8. Professional Career
9. Notable Works
10. Literary Style and Genre
11. Awards and Recognition
12. Contributions to the Field
13. Personal Life and Legacy
14. Influence on Others

D.3.4 Qwen2-7b-Instruct Output

I. Introduction
  A. Brief overview of John Gribbel

B. Purpose of the encyclopedia entry
II. Early Life and Education
A. Birth and family background
III. Career and Professional Contributions
A. Overview of professional career
B. Major contributions and innovations
IV. Research and Publications
A. Collaborations with other experts
V. Personal Life and Interests
A. Family, hobbies, and personal passions
B. Philanthropy and community involvement
C. Inspirational quotes or anecdotes
VI. Conclusion

D.3.5 GPT-3.5

Early Life and Background
Education and Academic Achievements
Career Beginnings
Contributions to the Field of Science
Research and Discoveries
Notable Publications
Awards and Recognition
Influence on Future Scientists
Personal Life and Family
Philanthropic Activities
Impact on Society
Collaborations and Partnerships
Professional Associations
Lectures and Presentations
Legacy and Memorials
Controversies and Criticisms
Current Projects and Initiatives
Future Prospects and Goals
Public Perception and Media Coverage
Enduring Contributions and Significance
Detailed Biography
Influence on Public Policy

D.3.6 Qwen2-7b-Scribe Output

1. Abstract
2. Main text
2.1 Early life and education
2.2 Career
2.2.1 Early career
2.2.2 Later career
2.2.3 Retirement
2.3 Personal life
2.4 Awards and honors
2.5 Selected publications
2.5.1 Books
2.5.2 Articles
2.6 External links

D.4 QA Data Generation Examples

Here is the section where we present the paragraphs generated by Qwen2-7b-Scribe in three different pipelines for the same task and compare them with the correct answers. The task is to generate an article paragraph for the topic "八一工体大队" and "John Gribbel."

Topic: 八一工体大队
Question: 什么是中央军委训练管理部军事体育训练中心的沿革和八一体工大队?
Original Answer: 中国人民解放军八一体育工作大队（简称八一体工大队）成立于1951年9月，隶属总政治部（1957年10月24日到1958年12月1日曾受训练总监部领导）。文革中，于1969年12月被迫解散；1973年5月重建，仍然隶属总政治部。八一体工大队成立初期没有营房，多次搬迁，先后驻先农坛体育场材料库、东绒线胡同居民区、公安部礼堂施工队工棚。直到贺龙在北京西郊红山口中国人民解放军国防大学院内批下一块地，八一体工大队才有了固定驻地。八一体工大队最早成立的是男、女篮球队。此后，男、女排球队、足球队、田径队、举重队、自行车队、游泳队、体操队、乒乓球队和手球队陆续成立。1956年6月7日，在上海举行的“中苏举重友谊赛”上，八一举重队队员陈镜开以133公斤的成绩打破美国运动员温奇保持的56公斤级挺举世界纪录，这是中国运动员创造的首个世界纪录。1959年9月，中华人民共和国第一届运动会在北京开幕。以八一体工大队为主体组建的解放军代表团以117金95银64铜名列金牌榜和奖牌榜第一。1963年中华人民共和国第二届运动会上，解放军代表团以72金53银37铜的成绩再次夺得第一。1970年代至1980年代，是八一男篮、女篮的鼎盛时期，男篮代表国家队同美国大学生联队两次比赛接连获胜，女篮连续五届全运会获得第一名。2009年中华人民共和国第十一届运动会，八一体工大队参赛运动员获19枚金牌。2011年第五届巴西世界军人运动会，参赛的八一体工大队选手获9枚金牌，为解放军代表团位居金牌榜第二位做出了贡献。2008年北京奥运会上，参赛的八一体工大队选手获女子举重、体操男子团体、鞍马、男子举重、乒乓球男子团体5枚金牌，取得历届奥运会最好战绩，为此八一体工大队被中共中央、国务院表彰为“北京奥运会残奥会先进集体”。截至2011年，八一体工大队获奥运会冠军14人次，世界冠军120人次，亚洲和重要国际比赛冠军980人次，全国冠军1452人次；134次夺冠，154次破（超）世界纪录，133次破（超）

亚洲纪录；培养出国际级运动健将72名，国家级运动健将748名，为国家队输送运动员、教练员1700多人次。截至2014年，八一体工大队获全国以上冠军3123个，其中奥运会冠军16个，世界冠军135个，世界军人运动会冠军33个，亚洲冠军349个，重要国际比赛冠军615个，22人158次破（超）51项世界纪录，30人135次破（超）42项亚洲纪录。八一体工大队自建队起的宗旨是："指导或者辅导部队开展群众性体育活动，增强官兵体质，丰富官兵的文化生活，促进官兵全面发展，直接生成部队战斗力。"八一体工大队的职能可概括为"服务于我军政治工作的需要，服务于我军对外交往和展示的需要，服务于广大官兵开展群众性体育活动的需要，服务于国家健康发展大局的需要。"八一体工大队驻北京市海淀区红山口甲3号，并在北京市东城区黄寺拥有田径和游泳训练基地，具有亚洲一流的田径馆以及国家一流的游泳馆、乒乓馆、足球场和篮排球馆。八一体工大队曾设有男女篮球队、男女排球队、足球队、田径队、男女举重队、自行车队、游泳队、男女体操队、男女乒乓球队、男女羽毛球队和手球队等。21世纪初期，撤消了八一羽毛球队和八一女子体操队。2003年8月，八一足球队和八一游泳队撤销建制。将游泳队和新增加的沙滩排球队合并成立为八一沙滩排球游泳队。到2011年八一体工大队成立60周年时，八一体工大队设有男女篮、男女排、田径、乒乓球、举重、体操、沙滩排球游泳共10个项目9个运动队。八一体工大队原是隶属中国人民解放军总政治部宣传部文化体育局的正师级单位。2017年12月30日，在深化国防和军队改革中，八一体工大队转隶中央军委训练管理部军事体育训练中心。2020年10月20日，中央军委训练管理部军事体育训练中心调整改革动员部署大会举行，大会宣布军队不再参加全国综合性体育运动会和单项赛事，这意味著军队不再保留竞技体育项目队。2020年8月11日，八一男排退出中国男排超级联赛。9月9日，八一女足退出2020中国足球协会女子甲级联赛。10月13日，八一女排退出2020-21赛季中国女排超级联赛。10月20日，中国篮协官方宣布八一男女篮退出CBA联赛和WCBA联赛。

D.4.1 Web Retrieval Pipeline

Generated Answer By Qwen2-7b-Scribe: 中央军委训练管理部军事体育训练中心，作为中国人民解放军的军事体育训练基地，其历史可以追溯至1951年9月成立的中国人民解放军八一体育工作大队。该大队的前身是贺龙元帅所率的120师的"战斗篮球队"，隶属于总政治部，是解放军竞技体育训练基地之一。在1957年10月24日 至1958年12月1日期间，八一体工大队曾受训练总监部领导。2017年12月30日，八一体工大队转隶中央军委训练管理部军事体育训练中心，成为其管理的正师级单位。\n\n八一体工大队的宗旨是指导或辅导部队开展群众性体育活动，增强官兵体质，丰富官兵的文化生活，促进官兵全面发展，直接生成部队战斗力。大队下辖篮球、排球、田径、乒乓球、举重、体操、沙滩排球、游泳等项目队，致力于培养专业竞技体育运动员，为部队输送优秀体育人才。\n\n2016年底，八一军体大队由原总参谋部军训部转隶中央军委训练管理部，标志着其隶属关系的调整。2018年初，全军专业体育单位整编组建军事体育训练中心，八一军体大队成为其中的一员，隶属中央军委训练管理部。这一系列的改革，旨在优化军队体育训练体系，提升军事体育训练水平，为军队建设提供强有力的支持。\n\n八一体工大队自成立篮球队起，逐渐扩展至多个体育项目。1956年，陈镜开在举重项目中创下中国首个世界纪录。在1959年和1963年的全国运动会上，解放军代表团均取得优异成绩。70至80年代，八一男女篮达到巅峰，多次在国际赛事中获胜。2009年和2011年，八一体工大队在国内外赛事中均有金牌入账。2008年北京奥运会，该队获得5金，创下历届最佳。至2011年，共获奥运冠军14次，世界冠军120次，亚洲及国际赛冠军980次，全国冠军1452次，破世界纪录154次，亚洲纪录133次，培养国际级运动健将72名，国家级748名，输送至国家队1700余人次。到2014年，累计获得全国以上冠军3123个，其中奥运16金，世界135金，世界军人33金，亚洲349金，国际615金，破世界纪录158次，亚洲纪录135次。\n\n八一体工大队的转隶，不仅体现了军队体育训练体系的现代化改革，也反映了军队对体育工作重视程度的提升。通过整合资源，优化配置，八一体工大队能够更好地服务于部队的体育训练需求，为官兵提供更加专业、高效的体育训练服务，进一步增强官兵的身体素质和战斗力。\n\n在深化国防和军队改革中，中央军委训练管理部军事体育训练中心大力弘扬"八一精神"，组织观看《大国军体》纪录片，了解军事体育发展史，创作编印反映"八一精神"的故事集、徽标、口号、歌曲等，组织和平积弊"五查五看"，分析症结、制订措施，激发广大军体健儿投身强军事业的澎湃动力。这些举措不仅丰富了官兵的文化生活，也增强了官兵的凝聚力和战斗力，为军队建设注入了新的活

力。\n\n总之，中央军委训练管理部军事体育训练中心及其下属的八一体工大队，是解放军体育训练的重要组成部分，其沿革和改革体现了军队对体育工作的高度重视，以及对提升官兵身体素质和战斗力的不懈追求。

D.4.2 Local Retrieval Pipeline

Generated Answer By Qwen2-7b-Scribe: 中央军委训练管理部军事体育训练中心，作为中国人民解放军的体育单位，其历史沿革和组织结构经历了多次调整与改革。自1951年9月成立的八一体育工作大队，原隶属于总政治部，后在深化国防和军队改革中，于2017年12月30日转隶中央军委训练管理部军事体育训练中心。这一转变标志着八一体育工作大队正式成为军事体育训练中心的一部分，隶属于中央军委训练管理部，负责军事体育训练工作。\n\n2018年初，全军专业体育单位整编组建军事体育训练中心，进一步整合了军队的体育资源，加强了军事体育训练的统一管理和专业化水平。2020年10月20日，中央军委训练管理部军事体育训练中心调整改革动员部署大会在北京举行，大会宣布军队不再参加全国综合性体育运动会和单项赛事，这意味着军队不再保留竞技体育项目队，转而更加专注于军事体育训练，提升官兵的体能素质和战斗能力。\n\n八一体工大队，作为中国人民解放军的体育力量之一，其发展历程体现了军队体育事业的变迁。从1951年9月成立，到2017年12月底转隶中央军委训练管理部军事体育训练中心，八一体工大队经历了从隶属总政治部到中央军委训练管理部的转变，这一过程不仅反映了军队体育体制的调整，也体现了军队对体育训练重视程度的提升。\n\n八一体工大队下辖射击、军事五项、现代五项马术、击剑和铁人三项(含自行车项目)5个队，这些队伍不仅代表了军队参加国内外体育赛事，也是军事体育训练的重要组成部分。通过这些队伍的训练和比赛，不仅能够提升官兵的体能和技能，还能够培养官兵的团队精神和战斗意志，为军队的现代化建设提供有力支撑。\n\n综上所述，中央军委训练管理部军事体育训练中心的沿革和八一体工大队的变迁，反映了军队体育事业的发展和军队现代化建设的需要。通过整合资源、优化结构，军队体育训练的水平得到了显著提升，为提升官兵的综合素质和战斗力发挥了重要作用。

D.4.3 Grounded Reference Pipeline

Generated Answer By Qwen2-7b-Scribe: 八一体工大队，原名中国人民解放军体育工作大队，是中国人民解放军第一支专业体育队伍，成立于1956年6月7日。1956年6月7日，在上海举行的“中、苏举重友谊赛”中，陈镜开以133公斤的成绩打破美国运动员温奇保持的56公斤级挺举世界纪录，这是中国运动员创造的第一个世界纪录。据八一队科研所的最新统计，截至目前，八一队共夺得全国以上冠军3123个，其中奥运冠军16个，世界冠军135个，世界军人运动会冠军33个，亚洲冠军349个，重要国际比赛冠军615个，22人158次破（超）51项世界纪录，30人135次破（超）42项亚洲纪录。八一队建队时的宗旨为：“指导或者辅导部队开展群众性体育活动，增强官兵体质，丰富官兵的文化生活，促进官兵全面发展，直接生成部队战斗力。”八一队的职能基本可以用四个“服务于”来解释：服务于我军政治工作的需要，服务于我军对外交往和展示的需要，服务于广大官兵开展群众性体育活动的需要，服务于国家健康发展大局的需要。八一队坐落于北京市海淀红山口，并在东城区黄寺拥有田径和游泳训练基地，具有亚洲一流的田径馆和国家一流的游泳馆、乒乓馆、足球场和篮排球馆。八一队的编制调整始于2019年7月，总政治部决定，八一体工大队进行编制调整，拥有光荣历史、并给国家培养出众多优秀运动员的八一足球队和八一游泳队正式撤编，不再属于军队编制。2020年10月21日，中央军委训练管理部军事体育训练中心调整改革动员部署大会20日在北京举行，大会宣布军队不再参加全国综合性体育运动会和单项赛事。2020年8月11日，国内媒体爆料八一男排已退出将于8月20日开打的中国男排超级联赛；9月9日，八一女足退出2020中国足球协会女子甲级联赛；10月13日，中国排协公布2020-21赛季中国女排超级联赛的参赛队伍名单，八一女排并不在列；10月20日，中国篮协官方宣布八一男女篮退出CBA联赛和WCBA联赛。

Topic: Medical Prescription
Question: What is the importance of legibility in the writing of medical prescriptions, and how can healthcare professionals ensure that their prescriptions are easily readable and understandable to patients and pharmacists?
Original Answer: Doctors' handwriting is a reference to the stereotypically illegible handwriting

of some medical practitioners, which sometimes causes errors in dispensing. In the US, illegible handwriting has been indirectly responsible for at least 7,000 deaths annually.There are several theories about the causes of this phenomenon. Some sources say the extreme amount of writing doctors employ during training and at work leads to bad handwriting,[better source needed]whereas others claim that doctors neglect proper handwriting due to medical documents being intended to be read solely by medical professionals, not patients.[better source needed]Others simply classify the handwriting of doctors as a handwriting style.The issue may also have a historical origin, as physicians from Europe-influenced schools have historically used Latin words and abbreviations to convey prescriptions;many of the abbreviations are still widely used in the modern day and could be a source of confusion.Some jurisdictions have legislatively required prescriptions to be legible—Florida specifies l̈egibly printed or typed

Generated Answer By Qwen2-7b-Scribe: Legibility in medical prescriptions is of paramount importance for several reasons. Firstly, it ensures patient safety by preventing medication errors, which can occur when pharmacists misinterpret illegible or ambiguous instructions. Secondly, it facilitates efficient communication between healthcare providers, as clear prescriptions allow for quicker and more accurate dispensing of medications. Lastly, it respects the professional duties of prescribers to provide legible and complete orders, which is a legal requirement in many jurisdictions.\n\nTo ensure that their prescriptions are easily readable and understandable, healthcare professionals can take several steps:\n\n1. **Use of Standardized Formats**: Adhering to established formats for prescriptions, such as those recommended by the World Health Organization (WHO), can help standardize the layout and structure of prescriptions, making them more uniform and easier to read.\n\n2. **Clear and Concise Language**: Avoiding medical jargon and using clear, concise language can prevent misunderstandings. It is also helpful to include the patient's name, date, and the prescriber's signature to provide context and authenticity.\n\n3. **Legibility of Handwriting**: Practicing good handwriting skills is crucial, especially for those who write prescriptions by hand. Using a consistent and clear handwriting style can significantly improve readability.\n\n4. **Use of Authorized Abbreviations**: While abbreviations can save time, they must be used correctly and understood by all involved parties. Misinterpretation of abbreviations can lead to errors, so it's important to use them judiciously and ensure that all parties are familiar with the abbreviations being used.\n\n5. **Complete Information**: Including all necessary information such as the patient's name, date, time, medication name, strength, quantity, and route of administration ensures that the prescription is comprehensive and reduces the risk of errors.\n\n6. **Review and Feedback**: Regularly reviewing prescriptions for legibility and completeness, and seeking feedback from peers or pharmacists, can help identify areas for improvement and ensure that prescriptions meet the required standards.\n\n7. **Education and Training**: Healthcare professionals should receive ongoing education and training on the importance of legibility and completeness in prescriptions. This can include workshops, seminars, and online resources that focus on best practices in prescription writing.\n\n8. **Use of Technology**: Employing electronic health records (EHRs) and digital prescription systems can enhance legibility and reduce errors. These systems often have built-in checks for completeness and legibility, and they can also help in tracking and managing medication orders more efficiently.\n\nBy implementing these strategies, healthcare professionals can significantly improve the legibility of their prescriptions, thereby contributing to safer and more effective patient care.

D.4.5 Local Retrieval Pipeline

Generated Answer By Qwen2-7b-Scribe: Legibility in the writing of medical prescriptions is of paramount importance for ensuring patient safety and effective medication management. A clear and legible prescription minimizes the risk of errors, such as misinterpretation of dosages or incorrect medication administration, which can lead to adverse drug events. Healthcare professionals can ensure that their prescriptions are easily readable and understandable by following several best practices:\n\n1. **Use of Standardized Abbreviations:** Adhere to universally accepted medical abbreviations and symbols to avoid confusion. For instance, q̈.d.f̈or ëvery dayör p̈.o.f̈or p̈er os(̈by mouth) should be used consistently.\n\n2. **Avoidance of Acronyms:** While some acronyms are standard in medical practice, using them without explanation can lead to misunderstandings, espe-

cially for patients who are not familiar with medical terminology. Always spell out acronyms the first time they are used in a prescription.\n\n3. **Clear and Concise Instructions:** Write instructions in a straightforward manner, avoiding medical jargon that might not be understood by patients or pharmacists. For example, instead of writing ẗake with food,s̈pecify ẅith mealsör b̈efore meals.̈\n\n4. **Use of Proper Units of Measurement:** Always include the unit of measurement (e.g., mg, mL) to avoid confusion. Ensure that the unit is consistent with the medication's standard packaging.\n\n5. **Dosage and Frequency:** Clearly state the dosage and how often it should be administered. For instance, ẗake 1 tablet twice dailyïs more understandable than ẗake 1 tablet q.d.̈\n\n6. **Prescriber's Signature:** Include a clear signature or initials to authenticate the prescription. This is crucial for legal and accountability purposes.\n\n7. **Prescription Number and Date:** Include a unique prescription number and the date of issuance. This helps in tracking the prescription and managing inventory.\n\n8. **Dosage Forms and Route of Administration:** Specify the dosage form (e.g., tablet, capsule) and the route of administration (e.g., orally, intravenously) to ensure the correct medication is administered in the right way.\n\n9. **Dosage Instructions for Children:** For pediatric patients, provide clear instructions on how to measure liquid medications or how to adjust dosages based on weight or age.\n\n10. **Prescription Review:** Before finalizing a prescription, review it for clarity and completeness. It can be helpful to have a colleague or pharmacist review the prescription to catch any potential errors.\n\nBy following these guidelines, healthcare professionals can significantly reduce the likelihood of prescription errors, thereby enhancing patient safety and the overall effectiveness of medication management.

D.4.6 Grounded Reference Pipeline

Generated Answer By Qwen2-7b-Scribe: The importance of legibility in the writing of medical prescriptions is paramount to ensure that patients receive the correct medication and dosage. Inaccurate or illegible prescriptions can lead to medication errors, which can have serious consequences for patient health. To ensure that their prescriptions are easily readable and understandable, healthcare professionals should write in a clear and consistent manner, using legible handwriting or electronic medical records. Additionally, healthcare professionals should double-check their prescriptions for accuracy and completeness before submitting them to a pharmacist. Pharmacists also have a responsibility to ensure that prescriptions are legible and to clarify any ambiguities with the prescriber. In some jurisdictions, such as the United States, there are laws and regulations that require prescriptions to be legibly printed or typed and to contain specific information, such as the name of the prescribing practitioner, the name and strength of the drug prescribed, the quantity of the drug prescribed, and the directions for use of the drug.

D.5 Full Article Generation Examples

In the spirit of rigorous academic comparison, we select an entry *Taylor Hawkins* from FreshWiki dataset(Shao et al., 2024), same as STORM() presented in their paper's appendix. We use our best performed pipeline: Web Retrieval pipeline, and set the generated model as Qwen2-7b-Scribe. we provide a comprehensive article generated by our model, showcasing its ability to produce detailed, long-form text on complex subjects. In addition, we include the long-form article on the same topic generated in the STORM paper(Shao et al., 2024) in the appendix of our document for a direct comparison with our results.

Taylor Hawkins *(Article generated by our Web Retrieval Method)*

Taylor Hawkins, the legendary drummer of the Foo Fighters, left an indelible mark on the world of rock music with his dynamic drumming and charismatic stage presence. Born in 1972 in Fort Worth, Texas, and raised in Laguna Beach, California, Hawkins' passion for music was ignited at a young age, culminating in a career that would influence generations of musicians. His journey began on the drums for Alanis Morissette before joining the Foo Fighters in 1997, where he contributed to the band's iconic sound through albums like "The Colour and the Shape" and "Sonic Highways."

Hawkins was not only known for his powerful drumming but also for his vocal talents and songwriting, which were showcased in tracks like "Cold Day in the Sun." Beyond the Foo Fighters, his musical explorations included his side project, Taylor Hawkins and the Coattail Riders, where he expressed his diverse musical tastes. Despite personal struggles, including a battle with substance abuse, Hawkins' resilience and dedication to his craft left a legacy of inspiration.

His unexpected death in 2022 shocked the music community, but his influence continues through the music he made and the lives he touched. Remembered for both his vibrant music and vibrant personality, Hawkins remains a symbol of perseverance and passion in the music industry.

 # Early Life and Career

Taylor Hawkins, the late drummer of the Foo Fighters, had a life that was marked by both musical talent and personal struggles. Born Oliver Taylor Hawkins on February 17, 1972, in Fort Worth, Texas, Hawkins was raised in Laguna Beach, California, where he developed a passion for music. His early years were dedicated to honing his skills, and he immersed himself in the rich tapestry of musical styles that would later DeFine his signature sound.

Hawkins' journey in music began with his first big break as the drummer for Grammy-winning artist Alanis Morissette. Joining her band in the mid-1990s, he toured extensively with her, contributing to the success of her iconic album Jagged Little Pill. This experience not only launched his career but also led him to meet his future bandmates in the Foo Fighters, as he replaced William Goldsmith in 1997.

His time with the Foo Fighters was marked by a whirlwind of activity, including the recording of their second album, The Colour and the Shape, which featured Hawkins' drumming and songwriting contributions. The album was a critical and commercial success, solidifying Hawkins' place in the music industry. He continued to work with the band, contributing to their subsequent albums and tours, which included the acclaimed Sonic Highways series and the 2022 zombie flick Studio 666.

Hawkins' musical skills extended beyond the drumset. He was also a talented singer and songwriter, as evidenced by his lead singing debut on an EP with a cover of Cream's "I Feel Free." He even played piano on the song "Summer's End" from Echoes, Silence, Patience Grace, showcasing his versatility as a musician.

Throughout his career, Hawkins recorded with various artists, including Coheed and Cambria, Eric Avery, Kerry Ellis, and Slash. His drumming and musical approach were influenced by Roger Taylor of Queen and Stewart Copeland of The Police, reflecting his diverse taste in music. He was also part of the SOS All-Stars with Roger Taylor of Queen and Chad Smith of Red Hot Chili Peppers, further highlighting his ability to collaborate with other musicians.

Hawkins' personal life was not without its challenges. He had a history of experimenting with drugs, which led to a near-fatal overdose in 2001. However, he was able to overcome this setback and live a clean lifestyle, enjoying a quieter life with his wife in California. Despite the distance between him and his bandmate Dave Grohl, their bond remained strong, reflecting the deep friendship that had developed over the years.

Hawkins' impact on the music world extended far beyond his drumming skills. He was a force of nature whose passion, talent, and spirit continued to inspire generations of musicians and fans alike. His legacy is not only in the music he created but also in the lives he touched and the memories he left behind.

In conclusion, Taylor Hawkins' early life and career were marked by dedication, determination, and a love for music. From his early days as a drummer for Alanis Morissette to his tenure with the Foo Fighters, he left an indelible mark on the music industry. His musical skills, versatility, and personal struggles made him a complex and fascinating figure, whose life and career continue to be celebrated and remembered.

 # role in the Foo Fighters

Taylor Hawkins was the drummer for the American rock band Foo Fighters, a position he held from 1997 until his death in 2022. He joined the band in 1997, following the departure of original drummer William Goldsmith. Hawkins was known for his energetic and dynamic drumming style, which complemented the band's hard rock and alternative rock sound. He was also a talented singer and songwriter, contributing to the band's albums and even taking on lead vocals on occasion, such as on the song "Cold Day in the Sun" from the album In Your Honor. Hawkins' role in the band extended beyond the drums, as he was an integral part of the band's live performances, often engaging with the audience and adding a sense of humor and theatricality to the shows. He was also a respected musician in his own right, with side projects like Taylor Hawkins and the Coattail Riders, which allowed him to explore different musical styles and showcase his versatility as a musician.

 # the highlights of Taylor Hawkins' solo career

Taylor Hawkins' solo career was marked by a diverse range of musical projects and collaborations, showcasing his versatility as a musician. He released his debut solo album, "Red Light Fever," in 2010, which featured a blend of rock, pop, and power-pop influences. The album's lead single, "Your Shoes,"

was a testament to Hawkins' love for Cheap Trick-style power-pop and Queen-level arena rock pomp, with harmonized guitar leads, boogie rhythms, and theatrical backing vocals that captured the essence of his rock'n'roll spirit.

Hawkins also contributed to the soundtrack of the 2013 rock film "CBGB," playing the role of Iggy Pop. This role not only highlighted his acting skills but also his ability to embody the energy and charisma of a legendary musician. Additionally, he recorded drums on Vasco Rossi's song "L'uomo più semplice," further expanding his international reach and influence.

Beyond his solo endeavors, Hawkins collaborated with various artists, including Slash, who featured him on the track "Crucify the Dead" from his solo album "Slash." This collaboration demonstrated Hawkins' ability to adapt his drumming style to different musical contexts, adding depth and texture to Slash's hard-hitting rock sound.

Hawkins' solo career also included a period of reflection and personal growth. He candidly documented his struggles with substance abuse in his solo album, which served as a platform for him to share his experiences and emotions with the world. This honesty and authenticity resonated with fans, providing a glimpse into the human side of a musician often seen through the lens of fame and success.

Overall, Taylor Hawkins' solo career was a testament to his musical versatility, creativity, and the depth of his talent. It was a journey that saw him explore different genres, collaborate with esteemed artists, and use his platform to address personal challenges, all while maintaining his status as a beloved member of the Foo Fighters and a respected figure in the music industry.

 # The cause of Taylor Hawkins' death

Taylor Hawkins, the drummer of the band Foo Fighters, passed away on March 25, 2022, at the age of 50. The cause of his death was not immediately revealed, but preliminary toxicology tests conducted by Colombian authorities indicated that he had 10 different substances in his system at the time of his death, including opioids, benzodiazepines, tricyclic antidepressants, and THC. The office of the Attorney General of Colombia stated that they would continue the medical studies to achieve total clarification of the events that led to Hawkins' death, and the investigation is ongoing. As of now, the official cause of death has not been determined, but it is believed that Hawkins' death was likely due to a drug overdose.

Hawkins' friends and colleagues have attempted to make sense of his death in the absence of any real answers. Some have suggested that Hawkins could have had heroin in his system and that a cocaine-like powder was found in his hotel room. However, Hawkins' friend, Yeti, believes resolutely that Hawkins was not using heroin or any other substance because he was supposed to play a concert that night. Three days later, he was dead, and the official cause of death is still unknown.

Meanwhile, a separate report claimed Hawkins' heart "weighed double that of men his age," which also could have contributed to his death. The finger-pointing continued from there, and Rolling Stone later suggested that Hawkins was exhausted by the constant demands of the Foo Fighters.

Hawkins' death sent shockwaves throughout the world, and tributes poured in from fellow musicians, some via social media and others in concert. The Foo Fighters canceled their remaining shows in South America and returned to the United States, where they continued to tour and perform. The band released a statement expressing their grief and love for Hawkins, and they dedicated their shows to him in his memory.

In the aftermath of Hawkins' death, the conversation around drug abuse and mental health has been impacted. Hawkins' struggle with drug addiction and his overdose have sparked discussions about the importance of mental health and the need for better support for musicians and artists who may be struggling with addiction or other mental health issues. The Foo Fighters have also been vocal about the importance of seeking help and getting support when needed, and they have encouraged their fans to do the same.

The cause of Taylor Hawkins' death is still unknown, but it is believed to be due to a drug overdose. The investigation is ongoing, and the official cause of death has not been determined. Hawkins' death has had a significant impact on the music industry and has sparked discussions about the importance of mental health and addiction support for musicians and artists. The Foo Fighters have been vocal about the importance of seeking help and getting support when needed, and they have encouraged their fans to do the same.

 # Legacy of Taylor Hawkins

Taylor Hawkins left a significant legacy in the music industry. His contributions to the band, his passion for music, and his vibrant personality have been remembered by fans, friends, and fellow musicians alike. Hawkins' impact on the Foo Fighters' sound and his ability to connect with audiences through his performances have solidified his place in rock history.

One of Hawkins' most notable achievements was his work on the band's 2021 album "Medicine at Midnight," which won all three awards it was nominated for at the Grammy Awards. The album, featuring nine tracks, showcased Hawkins' versatility and passion, as he navigated both faster, hard-hitting songs like "Cloudspotter" and slower, ballad-like tunes in "Chasing Birds." His dedication to the album and the band's overall success is evident in his words, where he expressed gratitude for the band's loyal fan base and their support during challenging times.

Hawkins' legacy extends beyond his musical contributions. His charismatic presence, athletic performances, and surprising vocal agility were hallmarks of his stagecraft. Friends and family members have shared stories of Hawkins' love for music and his commitment to the Foo Fighters, even as he considered the toll of constant touring. Despite his reservations, Hawkins remained an integral part of the band, his ebullience and unique connection with Dave Grohl electrifying the group's performances.

The impact of Hawkins' death on the music community was profound. Fans and fellow musicians alike

expressed their grief and admiration for the late drummer, with tributes pouring in from around the world. Joan Jett, Miley Cyrus, and many others paid homage to Hawkins, highlighting his talent, kindness, and drive. His legacy was further celebrated at a special all-star tribute concert held at Wembley Stadium, where over 75,000 fans gathered to honor his life and music.

Hawkins' influence on the Foo Fighters' sound and his role in shaping the band's identity are evident in the music they've produced. His drumming style, characterized by precision and power, complemented the band's rock and roll ethos, contributing to their signature sound. The band's return to the stage, with Josh Freese taking over on drums, was a testament to Hawkins' enduring presence and the strength of the Foo Fighters' musical bond. The Foo Fighters' legacy, as embodied by Hawkins, continues to inspire and influence musicians and fans alike. The band's commitment to their craft, their ability to connect with audiences, and their resilience in the face of adversity are qualities that Hawkins himself embodied. His spirit, passion, and love for music live on through the Foo Fighters' music and the memories of those who knew him.

Taylor Hawkins' legacy is a testament to his talent, dedication, and the impact he had on the music industry. His contributions to the Foo Fighters, his vibrant personality, and his ability to connect with audiences through his performances have left an indelible mark on rock music. Hawkins' legacy will continue to inspire future generations of musicians and fans, ensuring that his memory and music live on forever.

 # Discography

Taylor Hawkins recorded eight studio albums with Foo Fighters between 1999 and 2021. He also recorded two studio albums with Taylor Hawkins and the Coattail Riders:Red Light Feverin 2010, andGet the Moneyin 2019.

 # Awards and Nominations throughout Taylor Hawkins's life

Taylor Hawkins received numerous awards and nominations during his career. He was posthumously honored with three Grammy Awards at the 2022 Grammy Awards ceremony, winning in the categories of Best Rock Performance for "Making a Fire", Best Rock Song for "Waiting on a War", and Best Rock Album for Medicine at Midnight. The Foo Fighters, of which Hawkins was a member, had previously won 12 Grammy Awards, including Best Rock Album four times, and were inducted into the Rock and Roll Hall of Fame in 2021. Hawkins was a critical part of the band's success for 25 years, and their latest album, Medicine at Midnight, was nominated for three Grammy Awards. In addition to his musicianship, Hawkins brought humor, enthusiasm, and personality to the group, and he was a big supporter of his wife Alison, who counteracted his father's stony coldness. Hawkins began performing with local groups and joined the backing group for the vocalist Sass Jordan after being drummer for the progressive rock-influenced outfit Sylvia. He joined the Foo Fighters in 1997 and played with the band until his death in 2022.

Taylor Hawkins *(Article generated by STORM)*

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].