# "Not All Information is Created Equal": Leveraging Metadata for Enhanced Knowledge Curation

Stanford CS224N Custom Project

**Yucheng Jiang**
Department of Computer Science
Stanford University
yuchengj@stanford.edu

**Hong Meng Yam**
Department of Computer Science
Stanford University
hongmeng@stanford.edu

## Abstract

We study the problem of using source metadata, such as timestamps, to enhance retrieval-augmented generation (RAG) in a knowledge curation system. The RAG paradigm allows language models to ground in external sources, thus reducing factual hallucination. However, it is still challenging for the models to synthesize different sources in an organized and accurate way due to differing priorities among information pieces and inherent connections. This project takes the first step to combat this by focusing on solving time confusion issues. We propose **META** (**M**etadata **E**xtractor and **T**ext **A**ugmentation for Synthesis), a plug-in-and-play module designed to enhance RAG-based systems. META automatically extracts meta-information from sources, employing this data to inform and refine content generation. Experiments on FreshWiki Dataset show that incorporating META with the base knowledge curation system reduces temporal inaccuracies in articles on time-sensitive topics.

## 1 Key Information to include

- Mentor: Yijia Shao
- External Collaborators (if you have any): N/A
- Sharing project: For Yucheng, this shares context with larger effort of research in Stanford Open Virtual Assistant Lab under supervision of Prof. Monica Lam. The description of project in this proposal stands alone and does NOT share with any course / research.
- Contributions: Yucheng primarily contributed to methodology and system demo; Hong Meng primarily contributed to dataset preparation and qualitative error analysis.

## 2 Introduction

The recent advancement of large language models (LLMs) has revolutionized automated content generation, opening up novel avenues for information synthesis from diverse sources (Devlin et al., 2018; Brown et al., 2020). However, one major problem with using LLMs for information synthesis is hallucination (Huang et al., 2023). While this issue, especially factual hallucination (Min et al., 2023), can be mitigated by the retrieval-augmented generation (RAG) paradigm, synthesizing diverse information snippets into a coherent, accurate, and natural narrative remains a formidable challenge. One concrete issue with the naive RAG approach, which concatenates different sources and fits them altogether into the language model's context window, is temporal confusion (Dhingra et al., 2021). This is exacerbated by the inherent complexity of determining the relevance of and interconnections

between information pieces, a task further complicated by the dynamic nature of knowledge over time (Kumar et al., 2019; Zhou et al., 2021).

Many studies have greatly enhanced capabilities of RAG systems but they primarily focus on augmenting textual content without explicit consideration of the temporal dimensions that underpin many knowledge domains (Guu et al., 2020; Karpukhin et al., 2020). The question of how to ensure that content generated from historical, news, and other time-sensitive sources is temporally accurate remains less explored. This line of work is particularly important in some rapidly evolving fields where the temporal sequence of events significantly influences the interpretation and understanding of the subject matter (Wang et al., 2021; Li et al., 2020).

In this work, we introduce the **M**etadata **E**xtractor and **T**ext **A**ugmentation for Synthesis (**META**) module, a novel approach to effectively synthesize diverse information and address the specific issue of temporal confusion issue in knowledge curation systems. The META module is designed to refine the synthesis process in retrieval-augmented generation (RAG) frameworks by identifying and incorporating meta information such as timestamps, ensuring the generated content is not only grouded on accurate factual information but also temporally coherent. This approach represents a significant leap toward leveraging metadata for more nuanced and contextually aware content generation, a frontier that is underexplored in the existing literature. Herein, we developed the META module as a plugin module. With its plug-and-play functionality, META can be easily integrated into existing RAG systems, thereby improving their ability to produce content that is both accurate and temporally aligned with their underlying data.

Our work is built upon the previous efforts by Shao et al. (2024) that explored using LLMs for writing Wikipedia-like long articles from scratch with extensive citations. We conduct our experiments on a time-sensitive subset of FreshWiki Dataset (Shao et al., 2024), a curated collection of Wikipedia articles designed to benchmark the effectiveness of knowledge synthesis systems. Our experiments demonstrate that, by integrating META with their proposed system, we can significantly reduce temporal inaccuracies in the generated articles, especially in domains where time-sensitive information plays a pivotal role in the narrative structure (Tran and Cimiano, 2015; Gao et al., 2021). This advancement not only advances an existing RAG-based knowledge curation system but also showcases the potential of leveraging source metadata to improve information synthesis. Our implementation and findings establish a foundation for future explorations on expanding the META module to consider metadata beyond temporal information.

## 3 Related Work

**Retrieval-Augmented Generation (RAG)** Retrieval-augmented generation (RAG) is a commonly adopted methodology to alleviate the well known challenge of hallucination in language models by incorporating external knowledge sources during the generation process. Khandelwal et al. (2019) proved that the integration of language models with augmented external knowledge can effectively enhance performance on tasks requiring factual information. Integrating a retrieval component with a sequence-to-sequence model also demonstrates significant improvements in knowledge-intensive tasks (Lewis et al. (2020)). The proposed framework demonstrated effectiveness across multiple domains, including question-answer and fact-checking. The model's ability to leverage relevant information from a large corpus dynamically is critical to performing accurate and up-to-date information retrieval. Guu et al. (2020) further advanced this area by proposing a method for dynamically updating the external knowledge base used by an RAG model. Izacard and Grave (2021) subsequently focused on optimizing the retrieval process within RAG systems, proposing efficient strategies for selecting relevant documents, which they showed could improve both efficiency and accuracy. These foundational works establish the importance of integrating retrieval mechanisms with language models to generate more informed and contextually accurate outputs.

**Timeline Summarization** Earlier approaches in time summarization tasks primarily focused on extracting significant events from news articles, leveraging temporal information and clustering

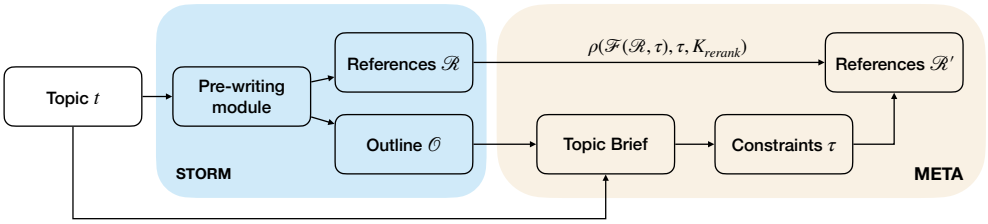


Figure 1: The META module workflow for enhancing temporal precision in information synthesis. Beginning with the topic brief generation, META processes the given topic to extract the foundational "5 Ws" (what, when, who, where, and why) as a topic brief, setting the stage for the subsequent analytical phases. Next, it predicts time constraints for each section, followed by filtering and ranking the sources according to these temporal constraints and semantic relevance to ensure the selection of the most pertinent and temporally aligned references.

techniques to organize events chronologically (Allan et al., 2001; Swan and Allan, 2000). Follow-up research works improved clustering algorithms by introducing hierarchical clustering and density-based spatial clustering of applications with noise (DBSCAN) to better group related events and identify pivotal moments in time (Zhang et al., 2009; Chakrabarti and Punera, 2011). Another line of work leverages machine learning and deep learning models for timeline summarization tasks. By heavily relying on traditional machine methods, they attempt to identify, classify, and summarize events for improving event detection and summarization (Tran et al., 2015; Liu et al., 2018). More recent work has focused on leveraging transformer models, such as BERT and GPT, to generate more coherent and contextually relevant timelines. These approaches benefit from the models' ability to understand complex temporal relationships and generate summaries that better reflect the progression of events. Herein, we aim to greatly improve upon these methods and incorporate it into RAG systems.

## 4 Approach

Given a topic $t$, curating the information on the Internet into a long, grounded article is a complicated task. Shao et al. (2024) proposed to decouple the task into two stages, the pre-writing stage and the writing stage. In the pre-writing stage, the system conducts research to collect references $\mathcal{R}$ and generates a hierarchical outline $\mathcal{O}$. In the writing stage, the system synthesizes information in $\mathcal{R}$ based on $\mathcal{O}$ into the final article section by section. To guarantee seamless integration, we design META as a plug-in-and-play module that can be used in the writing stage where the information synthesis happens. The META module comprises four major steps, with Fig 1 illustrating its overview.

1. **Topic Brief Generation**: Following the pre-writing stage, the initial step of META aims to provide a concise and accurate summary of the topic's primary aspect by converting the topic into a topic brief. The topic brief is defined by the fundamental "5 Ws" (what, when, who, where, and why).
2. **Time Constraint Prediction**: The META module uses a combination of the topic brief and section names in the outline to generate time constraint prediction applicable to each section of the article. Let $\tau$ represent the set of time constraints, with $\tau_i \in \{\text{before}, \text{after}, \text{around}, \text{no constraint}\}$ for each section $i$. $\tau$ is critical for maintaining chronological accuracy and respecting the distinct nature of each section during the actual writing (*a.k.a.*, information synthesis) stage.
3. **Source Filtering and Ranking**: With the collection of references $\mathcal{R}$, and the predicted time constraints $\tau_i$ for section $i$, we define a filtering function $\mathcal{F}(\mathcal{R}, \tau_i)$ that yields a subset $\mathcal{R}' \subseteq \mathcal{R}$ relevant to the time constraints. Subsequently, we rank $\mathcal{R}'$ based on semantic similarity to the section and subsection queries. For top $K_{rerank}$ references in reranked $\mathcal{R}'$,

we prioritize sources based on their temporal alignment with the section's requirements by reranking function $\mathcal{R}'' = \rho(\mathcal{R}', \tau, K_{rerank})$.

4. **Section Generation**: The final synthesis of each section employs a retrieval-augmented approach, akin to the STORM methodology. The META module facilitates the accuracy and temporal coherence by leveraging the reranked and filtered set of sources $\mathcal{R}''$.

# 5 Experiments

## 5.1 Data

| Category | Description | #Topics |
|---|---|---|
| Timely News | Topics within the current news cycle. Time metadata is important because it affects the relevance of a piece of information to the topic. | 19 |
| Event with a Timeline | Events that occur in phases or stages over a period of time. Time metadata is important to describe the event in an organized manner. | 39 |
| Recurrent Event | Events that recur. Time metadata is important to distinguish and organize information relevant to different occurrences. | 17 |
| | Total | 74 |

Table 1: Categories of time-sensitive topics and data statistics of FreshWiki (subset).

For our experiment, we use FreshWiki (Shao et al., 2024), a collection of recent and high-quality Wikipedia articles that were created (or very heavily edited) after February 2022 to mitigate data leakage. To better understand the potential usage of time metadata, we conducted a manual analysis of all topics within FreshWiki. Through this analysis, we pinpointed three primary categories that most significantly benefit from temporal metadata: (1) *timely news*, which requires the prioritization of temporally proximate information; (2) *event with a timeline*, which needs time metadata to organize information in chronological order; (3) *recurrent event*, which needs time metadata to distinguish information relevant to different occurrences. To test the effectiveness of META in solving the time confusion issues, we select a subset of time-sensitive topics, denoted as *FreshWiki (subset)*. Table 1 shows the data statistics after manually classifying them into one of three categories.

## 5.2 Evaluation Method

Capturing and evaluating time confusion issues in the context of long-form article generation presents a significant challenge. To address this, we introduce the metric of *Reference Time Precision* (RTP) for the automatic evaluation of temporal accuracy in the use of references. Consider a set of references $\{R_1, R_2, \ldots, R_N\}$ utilized in the synthesis of the final article. A reference $R_i$ is considered temporally precise if, and only if, it satisfies one of the following criteria: (1) the temporal information of $R_i$ falls within the designated period of the given topic, or (2) the sentences citing $R_i$ explicitly specify the temporal information relevant to $R_i$'s content.

The formula for *Reference Time Precision* is articulated as follows:

$$\text{Reference Time Precision} = \frac{|i \mid R_i \text{ is precise in time}|}{N} \tag{1}$$

where $|i \mid R_i \text{ is precise in time}|$ denotes the count of references that are precise in time, and $N$ represents the total number of references.

To ensure that the integration of the META module does not compromise the textual quality and factual integrity of the generated articles, we need to conduct an assessment of both the temporal accuracy and the general content quality. In particular, we evaluate the quality of the generated articles against human-authored articles using commonly adopted metrics such as *ROUGE-1*, *ROUGE-L*, and

| | Reference Time Precision | ROUGE-1 | ROUGE-L | Entity Recall |
|---|---|---|---|---|
| STORM | 81.43 | **41.52** | 13.38 | 16.55 |
| **STORM+META** | **98.20** | 40.55 | **13.72** | **17.64** |

Table 2: Results of article quality evaluation on FreshWiki (subset).

*Entity Recall* in addition to RTP. These metrics are calculated following the methodology outlined by Shao et al. (2024), serving as indicators of the overall quality of the generated content.

### 5.3 Experimental Details

We align our hyperparameters with those utilized in the STORM system (Shao et al. (2024)). Specifically, we set the number of perspectives $N = 5$, and the number of conversation rounds for information gathering $M = 5$ . We use `gpt-3.5-turbo` model within the original STORM system for the generation of questions and the handling of conversational elements. For our META module, we use `gpt-4` model for generating the topic brief, predicting time constraints, and polishing the final article generation result. We use You.com search API[1] as the backbone of the information collection module aligning with design of the original STORM system. The search API provides the interface for gathering up-to-date information from the Internet. The design of the META module adopts the principle of high compatibility and modulization for customization, allowing seamless integration and deployment across different platforms. The search API can be replaced with any other major search engine API.

We set the reranking threshold $K_{\text{rerank}}$ as 10 in all our experiments. This parameter determines the number of top sources selected based on semantic similarity and temporal relevance before undergoing a final reranking process. This step is crucial to ensure that the most pertinent and temporally aligned sources are prioritized, thereby enhancing the overall accuracy and coherence of the generated content.

### 5.4 Results

We test STORM, a knowledge curation system originally proposed in Shao et al. (2024), and STORM integrated with our META module on FreshWiki (subset). The quantitative results are reported in Table 2. We conduct LLM-based automatic evaluation of reference time precision, which serves as a critical metric for gauging the system's accuracy in temporal knowledge curation. The obtained baseline performance for reference time precision stands at 81.43%, indicating there is substantial room for improvement.

Upon integration with META, STORM+META demonstrates a remarkable improvement in reference time precision, achieving a rate of 98.20%. The large enhancement in reference time precision indicates the effectiveness of META in improving the knowledge curation system's capability to accurately process and integrate temporal metadata. Additionally, while the ROUGE-1 score slightly decreases from 41.52 to 40.55, the differences in ROUGE-L and entity recall scores with STORM+META indicate comparable, if not improved, performance. Specifically, STORM+META slightly outperforms the original STORM configuration in both ROUGE-L and entity recall metrics, achieving scores of 13.72 and 17.64, respectively. This affirms that the overall article quality, in terms of linguistic coherence and entity recognition, remains robust in conjunction with the enhanced accuracy in temporal information processing provided by META. The integration of META with STORM presents a promising direction for future research in the domain of automated knowledge curation, particularly in enhancing the temporal precision of generated content without compromising other aspects of article quality.

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

## 6 Analysis

In this section, we present key findings in our qualitative evaluation and error analysis. We first evaluate the alignment of human judgment on time constraints for each section of the article with META automatic prediction of time constraint; we then study the failure cases and provide insights on limitations and future work.

**Alignment with Human Judgment on Time Constraints** To assess the accuracy of automatic section time constraint prediction of our system, we asked two independent human annotators to reviewed the predictions by examining a tuple that consisted of the topic brief, section name, and the predicted time constraint and providing binary label "agree" or " disagree". We observed an agreement rate of 90.17% on average. This high level of agreement affirms the effectiveness of our time constraint prediction algorithm in mirroring human intuition and expertise in general topics with less than 10% divergence on time constraint prediction even on long-tailed, domain-specific topics like chemistry elements LK-99.

**Topic Brief Generation and its Focus** The generation of topic briefs generally yields reasonable summaries encapsulating the critical "5 Ws" of the given topic. However, we identify failure cases where the algorithm incorrectly generalizes a single event to a broader topic. For instance, in the case of "Silicon Valley Bank", the brief predominantly highlighted the bank's collapse and acquisition events in March 2023, incorrectly narrowing the scope of subsequent time constraint predictions to other sections of the article including overview and investment.

**Accuracy of Section-Aware Time Constraint Prediction** The precision of section-aware time constraint predictions is commendable, particularly in their alignment with the content of the sections. However, we do notice that the META module usually incorrectly applies constraints on sections with broad coverage such as "history", "overview", and "background", where the imposition of time constraints is inherently inappropriate. This highlights space for refinement in our system's understanding of section semantics and their implications for temporal framing. Further fine-tuning may alleviate this issue.

**Challenges with Niche and Highly Specific Topics** The META module's performance in predicting time constraints for niche topics or those requiring domain-specific knowledge drops as compared to more general and common topics. The system sometimes struggles to recognize the temporal constraint, and may incorrectly suggest constraints, or fail to output specific constraints in our desired format.Pre-training models tailored to specific use cases or correctly instructing the model to output "I don't know" may alleviate the issue.

## 7 Conclusion

We introduce META, an enhancement module for RAG-based systems, to improve the temporal accuracy when creating Wikipedia-like articles from scratch. META automatically identifies the time constraint when synthesizing each section and extracts temporal metadata from the source to augment the retrieved snippets. We introduce Reference Time Precision to evaluate the temporal accuracy and our experimental results show the effectiveness of META in reducing temporal confusion issues when synthesizing time-sensitive contents.

While META is designed in an extensible way to incorporate metadata to enhance RAG-based information synthesis, we only consider temporal metadata in this work. Exploring how to extend the META module to utilize other types of metadata, such as geographical location, novelty and cultural context, is a meaningful direction for future work.

## References


James Allan, Ron Papka, and Victor Lavrenko. 2001. On-line new event detection and tracking. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Deepayan Chakrabarti and Kunal Punera. 2011. Event summarization using tweets. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Hongyuan Ling, Karen Livescu, and Daniel S Weld. 2021. Time-aware language models as temporal knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9646–9656. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910. Association for Computational Linguistics.

K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Yang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of ICML*.

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

G. Izacard and E. Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of EACL*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

U. Khandelwal, K. Clark, D. Jurafsky, and O. Levy. 2019. Sample efficient text summarization using a single pre-trained transformer. In *Proceedings of ACL*.

Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2019. Temporally aware algorithms for document classification. *arXiv preprint arXiv:1906.01917*.

P. Lewis, M. Yazdani, W. Yih, S. Riedel, and L. Zettlemoyer. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of NeurIPS*.

Yeqing Li, Yunlong Miao, Juncheng Yang, and Alexander G Hauptmann. 2020. Temporal modeling approaches for large-scale youtube-8m video understanding. *arXiv preprint arXiv:2006.10728*.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2018. Real-time rumor debunking on twitter. *ACM Transactions on Information Systems*.

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

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.

Robert Swan and James Allan. 2000. Automatic generation of overview timelines. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*.

Giang Binh Tran, Mohammad Alrifai, and Eelco Herder. 2015. Timeline summarization from relevant headlines. In *24th International World Wide Web Conference, WWW 2015*.

Tuan Tran and Philipp Cimiano. 2015. Leveraging the crowdsourcing of lexical resources for bootstrapping a linguistic data cloud. In *International Semantic Web Conference*, pages 186–201. Springer.

Angela Fan Wang, Edouard Grave, and Armand Joulin. 2021. Learning to generate wikipedia summaries for underserved languages from wikidata. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Keke Zhang, Jianshe Zhou, and Lei Shu. 2009. Real-time event detection and visualization for online news. In *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery*.

Chaochao Zhou, Linlin Yang, Li Liu, and Meng Liu. 2021. Temporal knowledge graph embedding model based on additive time series decomposition. *IEEE Transactions on Knowledge and Data Engineering*.