Temporal Information Retrieval and Question Answering using Retrieval Augmented Generation

Sonith Bingi
UW-Madison
sbingi@wisc.edu

Manoj Arulmurugan
UW-Madison
arulmurugan@wisc.edu

Prasad Jawale
UW-Madison
pjawale@wisc.edu

Gauri Patki
UW-Madison
gpatki@wisc.edu

1 Introduction

Retrieval-Augmented Generation (RAG) models enhance knowledge-heavy applications by incorporating a retriever that extracts supporting external documents and a generator that produces answers given this evidence. While skilled at grounding answers in fact, baseline RAG models rely heavily on semantic similarity and overlook the temporal context of information. Facts in most fields change over time—things happen sequentially, objects transform from one state to another, and the truth of information is relative to when it is true. Disregarding these temporalities results in temporal inconsistency, where evidence drawn is stale or contradictory to the time period of the query. Temporal RAG achieves this by adding time-aware reasoning to both retrieval and generation. Temporal metadata is incorporated, evidence is temporalized to match the query temporal scope, and generated text is forced to indicate temporally valid facts. State-of-the-art work such as M-RAG, TIMEQA, and TEMPRAGEVAL shows that modeling time explicitly significantly improves factual accuracy and temporal coherence.

2 Literature Survey

The survey by (Piryani et al., 2025) positions Temporal RAG (Retrieval-Augmented Generation) as a crucial step forward in overcoming the static knowledge limitations of temporal language models (TLMs) like TempoT5 and BiTimeBERT, which are prone to outdated information and temporal hallucinations. Temporal RAG extends the standard RAG framework by introducing time-aware retrieval and reasoning, ensuring that both the query and the supporting evidence are temporally consistent. This is achieved by enhancing retrievers—such as TempRetriever and TsContriever—with timestamp-aware encoding, integrating features like explicit timestamps, temporal decay, or event timelines into dense vector representations. Such enhancements enable retrieval systems to prioritize documents relevant to the specific timeframe of a query.

Modular RAG(Siyue et al., 2025) is a training-free, modular framework that separates semantic relevance from temporal reasoning across three stages: Question Processing, Retrieval and Summarization, and Semantic–Temporal Hybrid Ranking. In the final ranking stage, semantic similarity is combined with a symbolic temporal alignment score using spline functions to assess how well the evidence aligns with the query's timeframe. The authors introduce TEMPRAGEVAL, a benchmark derived from TIMEQA and SITUATEDQA that includes temporal perturbations and human-annotated gold evidence to evaluate the model. These perturbations reveal that the performance of existing retrievers, such as GEMMA, drops significantly when temporal changes are introduced. Experimental results show MRAG achieves significant improvements in both gold evidence recall and question answering accuracy on TEMPRAGEVAL, establishing it as a robust framework for time-sensitive reasoning.

SITUATEDQA (Zhang and Choi, 2021) highlights one of the main shortcomings of conventional open-retrieval Question Answering (QA) systems: the assumption that answers are fixed and do not vary based on contextual data coming from the external world, primarily time (when) and place (where). In the authors' view, for the purpose of providing helpful and accurate answers, QA systems ought to be able to consider the extra-linguistic context of the questioners. Being a diagnostic benchmark for Temporal RAG, the work evaluates models like Dense Passage Retrieval (DPR) and Closed Book approaches (BART) based on how flexible they can be under different situations. Their diagnostic performance detects that such models suffer heavily from performance declines upon reaching

time-based queries, which indicates the need for context-aware retrieval and reasoning within QA systems.

The TORQUE dataset (Ning et al., 2020) is a text comprehension benchmark intentionally designed to evaluate natural language models' temporal reasoning abilities. The dataset employs natural language to annotate more nuanced and context-dependent temporal relations instead of discrete labels like "before," "after," or "during". TORQUE incorporates contrast questions that differ only by subtle temporal cues, discouraging superficial pattern-matching. Despite promising baselines, RoBERTa-large achieved only 51% exact match accuracy compared to a mean of 84.5% accuracy for standard tasks, highlighting substantial gaps in the temporal reasoning capabilities of current systems. Analyses reveal persistent challenges with event boundary detection, implicit temporal dependencies, and inferences that extend beyond the sentence level.

3 Potential Methods

Metadata-aware retrieval (prototype): Make the retriever use available document or chunk metadata to respect time windows, not just recency, by considering publication date plus other fields such as updated date and access/crawl time. We'll normalize these into a simple validity window per chunk and address known challenges like date ambiguity and the "publication vs event time" mismatch. Although under-explored and technically tricky, we will still ship a prototype as advised. (Publication timestamps are provided in ArchivalQA(Wang et al., 2022) and ChroniclingAmericaQA(Piryani et al., 2024), and modular time-aware retrieval frameworks separate temporal intent from semantic content.). As suggested by the Professor, we have no reason not to explore this approach, so we will try implementing a method using it.

Temporal-embedding fusion: We'll encode both text and time for queries and passages: add a compact time vector (e.g., Time2Vec/sinusoidal) to the text embedding, then learn a small fusion MLP; train with temporal hard negatives so "right text, wrong time" is pushed away.

As our custom retriever: we index fused passage embeddings; at query time we fuse the query's time (as-of/before/after or a midpoint for between) and retrieve by dot-product, then hand top-k to the same temporal re-ranker. This introduces time during candidate generation, not only in ranking, following the DPR-style extension shown effective in TempRetriever.

Inside MRAG: replace MRAG's off-the-shelf retriever in Module (2) with our fusion retriever so candidates already align with the temporal constraint parsed in Module (1), easing Module (3)'s semantic-temporal hybrid ranking. Alternatively, use a two-stage setup: semantic retriever (k) $\rightarrow$ fusion retriever (mk) $\rightarrow$ MRAG ranker. Because MRAG reports that standard retrievers "struggle" on temporal questions, injecting time at retrieval should strengthen MRAG's evidence pool and end-to-end accuracy

Positional-encoding-inspired time tokens: Build a novel variant that explicitly appends a compact time representation to semantic embeddings, inspired by Transformer sinusoidal positional encodings. The idea is to make time a first-class coordinate in the embedding space so near-text matches outside the intended time window are less likely to be retrieved.

Offline tagging with time-anchored queries (from TA-RAG): Pre-tag chunks with time intervals during indexing, pre-filter candidates by the parsed temporal constraint (as-of, before, after, between), then rank using time-anchored query variants aggregated over the window (for example averaging "In Jan 2014," "In Feb 2014," etc.). This follows TA-RAG's core recipe for diachronic questions while keeping our focus on retrieval.

4 Potential Datasets

4.1 TORQUE

TORQUE (Ning et al., 2020) is a reading comprehension dataset that serves as a benchmark for models on multiple features, including news snippets and human-generated questions. It consists of 3.2k news passages and over 21k questions designed to query temporal relationships between events, emphasizing nuanced and context-dependent temporal relations rather than simple fact recall. TORQUE uses natural language annotations to capture fuzzier temporal connections. The dataset helps evaluate and improve models' temporal reasoning, supporting the time-aware retrieval focus of our project. TORQUE includes event annotations and question-answer pairs, enabling the testing of RAG applications on event identification and temporal ordering questions.

4.2 TempRAGEval

Temporal Question Answering for RAG Evaluation is a benchmark dataset that can be used to rigorously evaluate the time-sensitive question answering on RAG systems. It was introduced in MRAG: Modular Retrieval Framework for Time Sensitive Questions" (Zhang and Choi, 2021). The dataset is a mix of existing datasets used for training temporal RAGs, and it maintains a natural human-written language style of question answering. It has about a thousand test instances, with each question having a manually annotated gold evidence sentence obtained from Wikipedia. The authors of this dataset added temporal perturbations by changing time constraints in the questions to check the robustness to temporal changes. The dataset intends to test both the retrieval performance and the question answering performance.

4.3 Experimental Settings

Datasets: We will train and evaluate our models on the Motion and Torque (MT) datasets, as well as TempEval-3. These datasets are specifically chosen to assess a model's ability to handle fine-grained temporal relations and events. The data will be split into an 80/20 train/test set.
Hardware: Experiments will be conducted on an NVIDIA A40 GPU.
Hyperparameters: Both models will be trained for 3 epochs with a batch size of 8 and a learning rate of 2e-5. The proposed model's time embedding dimension will be 32.

4.4 Experiments and Evaluation

We will conduct two main experiments to validate our hypothesis:
Retrieval Performance: We will measure the models' ability to retrieve the correct context using Top-1 and Top-5 accuracy on the held-out test sets.
End-to-End Exact Match (EM): We will create a two-stage pipeline where the trained DPR models act as the retriever, and a separate, pre-trained reader model (e.g., DistilBERT) extracts the final answer. We will then calculate the Exact Match (EM) score to measure the pipeline's ability to produce the correct answer string.

5 Plan of Activities

Weeks 1 and 2: Bring up RAG baseline and evaluation, then build first versions of all four methods: metadata-aware retrieval with per-chunk validity windows, temporal-embedding fusion with a small time vector and hard negatives, positional time tokens appended to embeddings, and offline tagging with a temporal-constraint pre-filter and basic time-anchored queries.
Weeks 3 and 4: Integrate the four methods behind one API and run a small dev evaluation; try temporal-embedding fusion as the retriever in MRAG Module 2 and also as a two-stage after a semantic retriever, and present midterm metrics to choose two methods to continue.
Weeks 5 and 6: Improve the two selected methods only: for metadata-aware tune validity window overlap and ambiguity handling, for temporal-embedding fusion extend training and refine the small fusion MLP, for time tokens run a quick size comparison, and for offline tagging tune the pre-filter and anchor granularity.
Weeks 7 and 8: Run full evaluations with the chosen methods (including temporal-embedding fusion inside MRAG), finalize results, and package the code and configs.
Work division (4 people): Person 1 owns metadata-aware retrieval, Person 2 owns temporal-embedding fusion and its MRAG use, Person 3 owns positional time tokens and quick ablations, Person 4 owns offline tagging with time-anchored queries plus evaluation and API unification.

References

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. Torque: A reading comprehension dataset of temporal ordering questions. *Preprint*, arXiv:2005.00242.

Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari, Avishek Anand, and Adam Jatowt. 2025. It's high time: A survey of temporal question answering. *Preprint*, arXiv:2505.20243.

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. Chroniclingamericaqa: A large-scale question answering dataset based on historical american newspaper pages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 2038–2048. ACM.

Zhang Siyue, Xue Yuxiang, Zhang Yiming, Wu Xiaobao, Luu Anh Tuan, and Zhao Chen. 2025. Mrag: A modular retrieval framework for time-sensitive question answering. *Preprint*, arXiv:2412.15540.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2022. Archivalqa: A large-scale benchmark dataset for open domain question answering over historical news collections. *Preprint*, arXiv:2109.03438.

Michael J. Q. Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. *Preprint*, arXiv:2109.06157.