SECTION 7 · 15 min

Temporal RAG

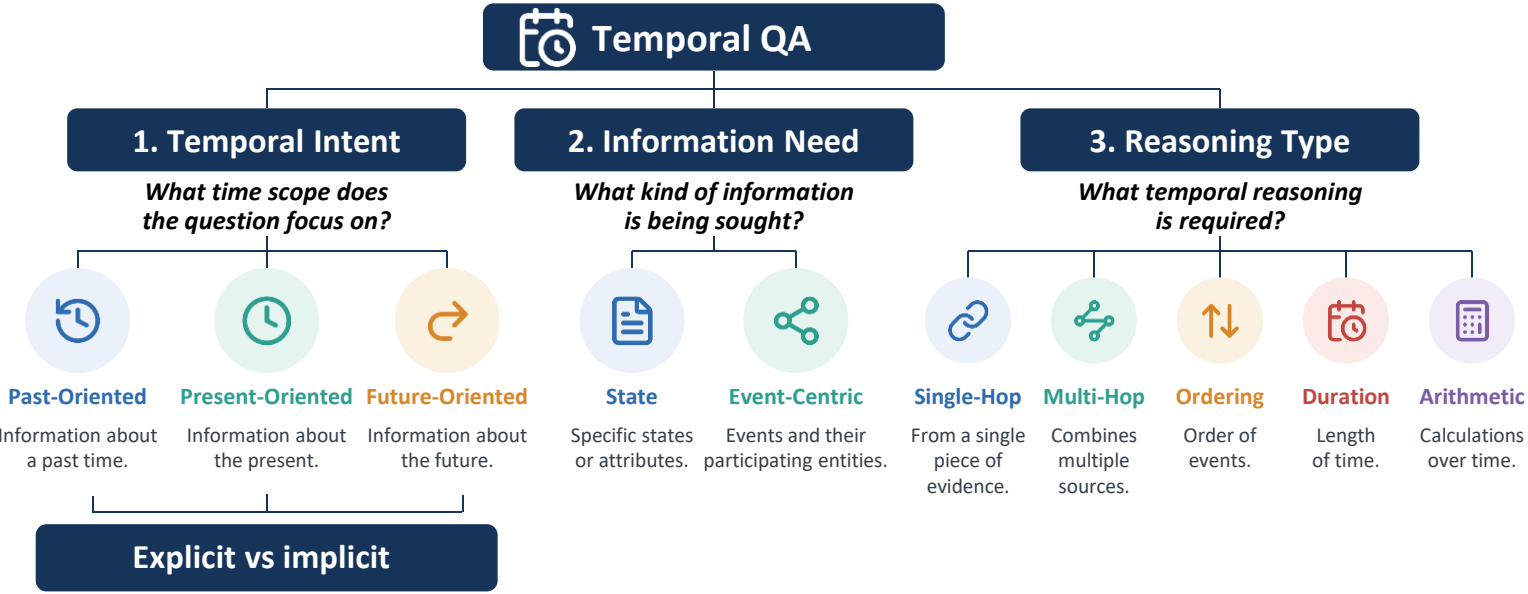
Coupling neural retrieval with generation to answer questions over evolving knowledge — and where LLMs still fail.

Presenter: Adam Jatowt

07

Taxonomy of Temporal QA

Temporal questions differ in both their information needs and reasoning requirements.



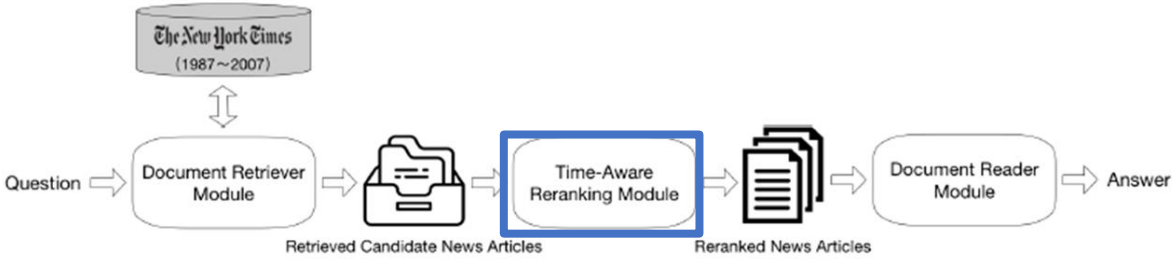
Example Questions by Type

Reasoning \ Scope	Past	Present	Future
Single-Hop	Who won the 1943 Nobel Prize in Physics?	Who is the current CEO of Apple?	When is the next FIFA World Cup?
Multi-Hop	Who was US president when the Berlin Wall fell?	Who is the CEO of YouTube's parent company?	Which city hosts the Olympics after LA 2028?
Ordering	What happened after the Berlin Wall fell?	Which came first: ChatGPT or GPT-4?	Artemis III or the next Mars mission was the first?
Duration	How long did World War II last?	How long has ChatGPT been available?	How long will the next ICC World Cup last?
Arithmetic	How old was Obama at his Nobel Peace Prize?	How many years has Amazon operated?	How old will ChatGPT be when the 2030 World Expo opens?

QANA: Pre-LLM Approaches for Temporal QA/RAG

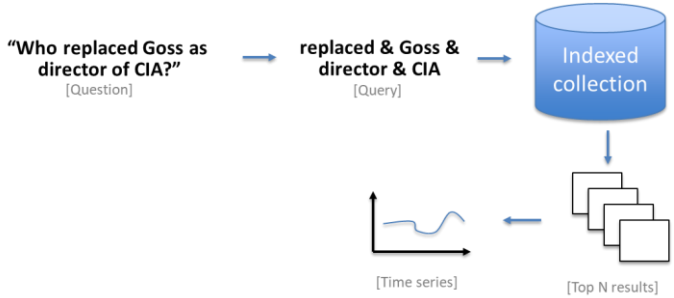
QANA: Question ANswering from Archives

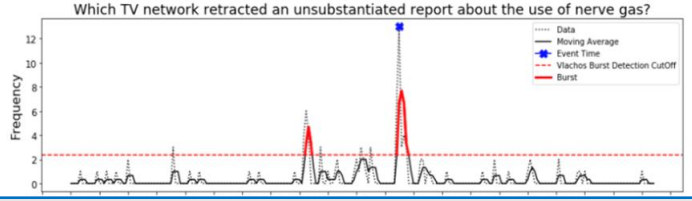
Given millions of documents, how to select a subset
of them for extracting correct answer to the user's question?



J. Wang, A. Jatowt, M. Yoshikawa and M. Farber: ***Improving Question Answering for Event-focused Questions in Temporal Collections of News Articles***, Information Retrieval Journal (IRJ) (2021)
J. Wang, A. Jatowt, M. Yoshikawa and M. Farber: ***Answering Event-Related Questions over Long-term News Article Archives***, Proceedings of ECIR 2020, pp. 774-789 (2020)

QANA: Pre-LLM Approaches for Temporal QA/RAG

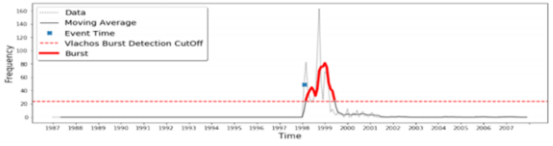




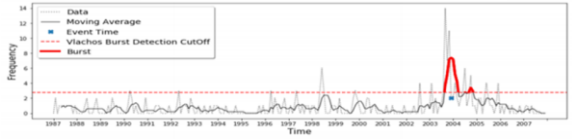
Detect **question time scope** by automatically **finding bursts** in distribution of results

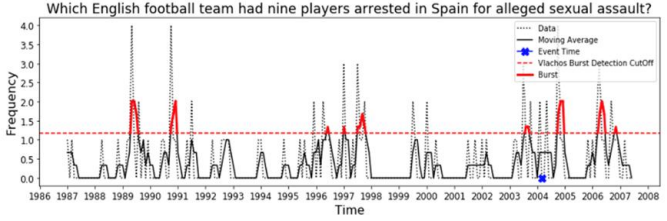
Question's time scope is represented as a set of time periods

Lewinsky told whom about her relationship with the President Clinton?

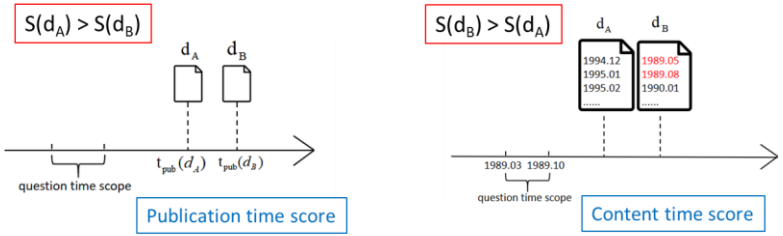


Which Hollywood star became governor of California?

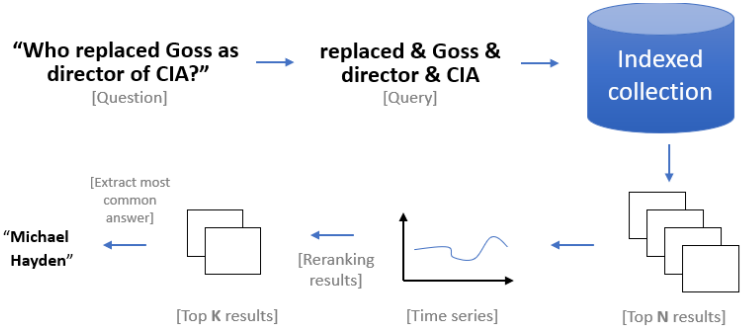




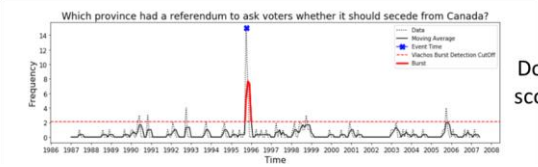
QANA: Pre-LLM Approaches for Temporal QA/RAG



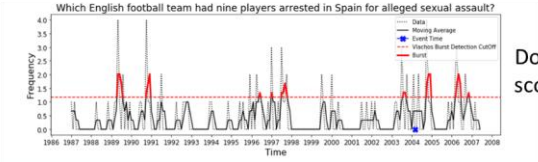
+ Use Kernel Density Estimation to compute the overlap of content time expressions and question time scope
+ If the question time scope contains multiple periods, aggregate the scores for all periods



Re-rank documents by a linear dynamic combination of their relevance scores and two temporal scores



Document temporal scores are used more



Document relevance scores are used more

- Take *K (K<<N)* top-ranked reranked documents and find answers using DrQA method
- Aggregation step: choose the most common answer as the final answer

Temporal Blind Spots in Large Language Models (2024)

Goal

- Investigate how well LLMs understand temporal knowledge
- Evaluate performance on temporal QA tasks

Datasets

- TemporalQuestions
- ArchivalQA
- TempLAMA

Key Findings

- Lower performance on historical **detailed** questions
- Difficulty handling recent information (temporal inertia)
- Relative time expressions ("3 years ago") are challenging
- Identifies common temporal failure modes

Which film won seven Oscars in 1994?		
ChatGPT Correct	Forrest Gump Schindler's List	✗ *Temporal shift:* F.G. won Oscars in **1995**
Who lost the WBA boxing title, refusing to fight Tony Tucker in March 1995?		
Alpaca Correct	Mike Tyson George Foreman	✗ *Time invariant:* time disregarded
Tom Brady played for which team in 2020?		
Alpaca Correct	New England Patriots Tampa Bay Buccaneers	✗ *Temporal inertia*: Brady joined Buccaneers in 2020

Table 1: Examples of temporal blindspots in LLMs.

Jonas Wallat, Adam Jatowt, and Avishek Anand. 2024. **Temporal Blind Spots in Large Language Models.** In Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24). Association for Computing Machinery, New York, NY, USA, 683–692. https://doi.org/10.1145/3616855.3635818

Temporal Blind Spots in Large Language Models (2024)

Temporal Blind Spots

Temporal Shift - Answer from the wrong time period

Time Invariance - Ignore the temporal constraint

Temporal Inertia - Prefer outdated knowledge

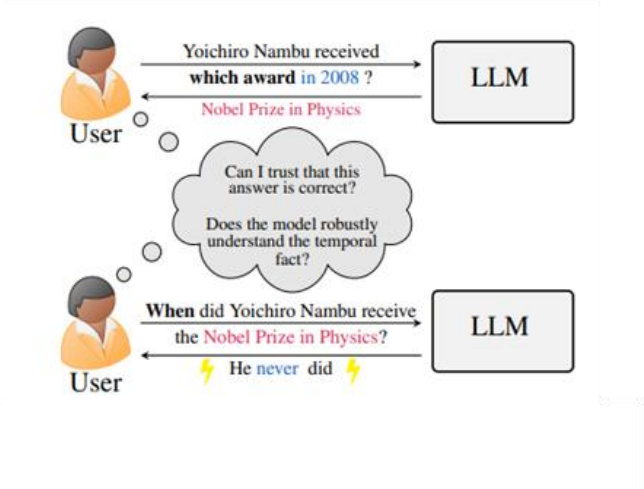
Referencing Error - Misunderstand relative time

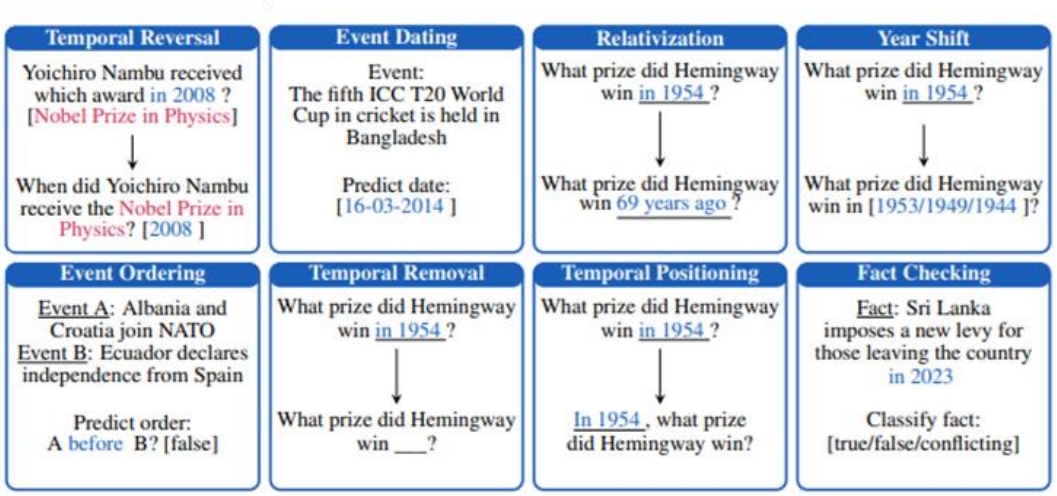
Model	Question	Temporal error	Explanation
alapaca-7B Answer	Tom Brady played for which team in 2020? New England Patriots Tampa Bay Buccaneers	Temporal inertia	Brady joined Tampa (2020) after 20 years with Patriots
red-pajama-7B Answer	Cristiano Ronaldo played for which team in 2019? Real Madrid Juventus FC	Time invariance	The model predicts Real Madrid for all years
alapaca-7B Answer	In the 2003 Wimbledon Men's Singles Tennis Championship, who beat Tim Henman? Andy Roddick Sebastien Grosjean	Temporal shift	Roddick beat Henman in the following year (2004)
text-davinci-003 Answer	Who painted a portrait of George Washington for Martha in 1789? Gilbert Stuart John Ramage	Temporal shift	Stuart did paint Washington but in 1795 and 1796
text-davinci-003 Answer	What state did Tuscany join 162 years ago? Unknown. Italy	Temporal referencing	correct answer of Italy with absolute reference (1859)

Table 4: Examples of temporal blind spots.

Jonas Wallat, Adam Jatowt, and Avishek Anand. 2024. **Temporal Blind Spots in Large Language Models.** In Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24). Association for Computing Machinery, New York, NY, USA, 683–692. https://doi.org/10.1145/3616855.3635818

A Study into Investigating Temporal Robustness of LLMs (2025)





A Study into Investigating Temporal Robustness of LLMs (https://aclanthology.org/2025.findings-acl.810/) (Wallat et al., Findings 2025)

A Study into Investigating Temporal Robustness of LLMs (2025)

8 time-sensitive robustness tests for factual information to check the sensitivity of 6 LLMs in zero-shot settings

Dataset	Example Question	Answer	#Qs	Scope	Description
ArchivalQA	`What was Ankara's official aid bill for in 1997?`	Cyprus	7,500	1987-2007	detailed quest.
Wikidata	`Yoichiro Nambu received which award in 2008?`	Nobel Prize	10,000	1907-2018	people
Temporal Claims	`China reports military clash in Henan province in 2022`	False	4,196	1000-2023	claims
Wikipedia Events	`Former Pope Benedict XVI dies at the age of 95.`	Dec. 31 2022	23,550	1750-2023	events

Model Name	Mode Size	Notes	Cutoff
`Llama 3.1`	8B	Instruction-tuned version	Dec. 2023
`Jamba 1.5`	12B active, 52B total	Mixture-of-Experts model that combines mamba (state-space) and transformer blocks. 8bit quant.	Mar. 2024
`Gemma 2`	27B	Instruction-tuned version	Jun. 2024
`Qwen 2.5`	32B	Instruction-tuned version	2023
`Cmd-R+`	104B params	RAG-optimized language model, weights openly available. Uses 4-bit quantization	N.S.
`GPT 4`	N.S.	OpenAI's flagship GPT model (gpt-4-1106-preview)	Apr. 2023

A Study into Investigating Temporal Robustness of LLMs (https://aclanthology.org/2025.findings-acl.810/) (Wallat et al., Findings 2025)

A Study into Investigating Temporal Robustness of LLMs (2025)

All models lack robustness with temporal relativization (28-50% decrease)

- Out of the 26.3% of questions that Llama 3.1 can answer, only 13.2% are also answered correctly when using the relative time references (50% decrease)..

Larger models seem to be more robust to using relative references

Model	Abs	↔Relativization Abs ∩ Rel.	Diff.
Llama 3.1	26.3	13.2	-50.0%
Gemma 2	35.2	25.4	**-27.9%**
Qwen 2.5	29.8	17.6	-41.0%
Jamba 1.5	38.5	27.3	-29.1%
Cmd-R+	40.4	26.6	-34.3%
GPT 4	**43.3**	**30.5**	-29.7%

"*Who was the American president in 2018?*"

"*Who was the American president 7 years ago?*"

A Study into Investigating Temporal Robustness of LLMs (https://aclanthology.org/2025.findings-acl.810/) (Wallat et al., Findings 2025)

A Study into Investigating Temporal Robustness of LLMs (2025)

All models benefit from time references placed at the start of the question

Improvements of about 1% - 5% under the OpenEval score

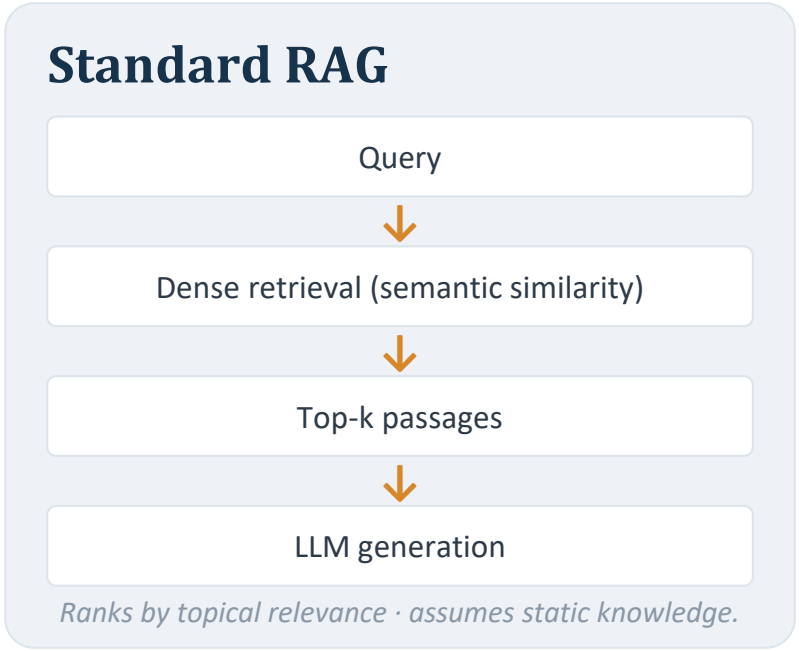
Model	↔Positioning		
	Time[end]	Time[front]	Diff.
Llama 3.1	26.3	27.6	+4.8%
Gemma 2	35.2	36.4	+3.3%
Qwen 2.5	29.8	30.0	**+0.6%**
Jamba 1.5	38.5	39.9	+3.5%
Cmd-R+	40.4	40.7	+0.7%
GPT 4	**43.1**	**44.4**	+2.9%

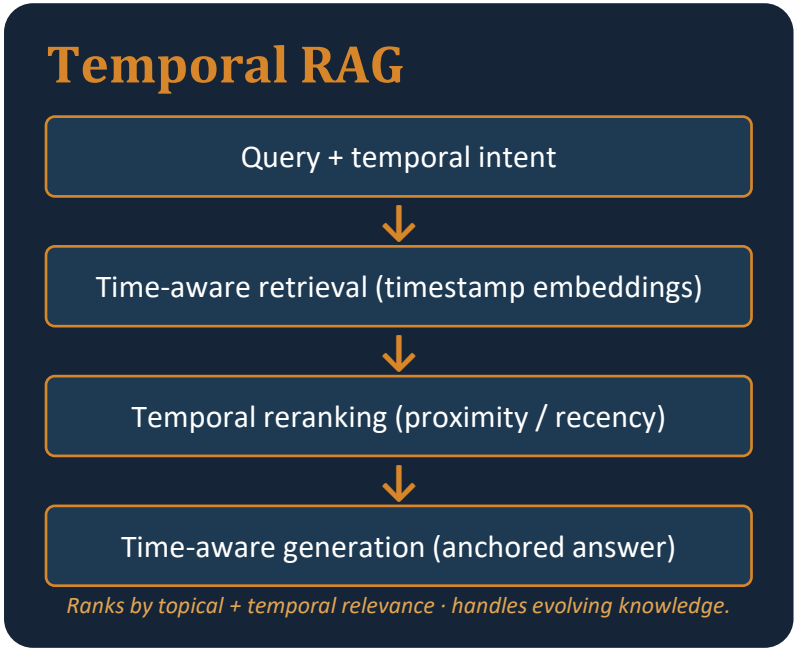
"*Who was the American president in 2018?*"

"In 2018, *who was the American president?*"

A Study into Investigating Temporal Robustness of LLMs (https://aclanthology.org/2025.findings-acl.810/) (Wallat et al., Findings 2025)

Standard RAG vs. Temporal RAG





It's About Time: Incorporating Temporality in Retrieval Augmented Language Models

Key Idea

TempRALM augments retrieval with temporal relevance. Instead of ranking by

Semantic Score

it ranks using

Semantic Score + Temporal Score

where documents closer to the query timestamp receive higher scores.

Additionally,

- ✓ removes documents from the future
- ✓ ranks documents by semantic + temporal proximity
- ✓ requires no retraining
- ✓ works with existing RAG systems

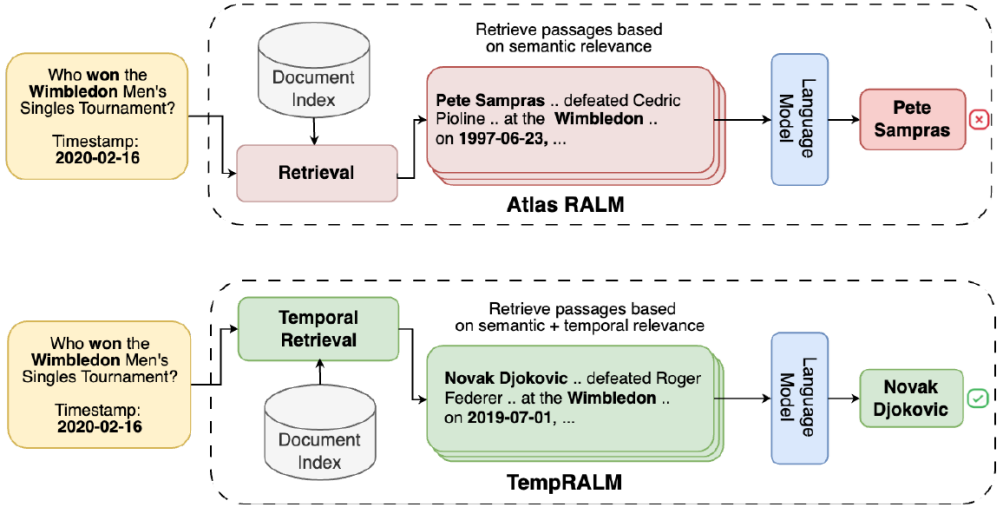


Figure 1: Overview of TempRALM: In this figure, we show the difference between Atlas and TempRALM. In TempRALM, the retriever fetches documents on semantic relevance with respect to the query as well as temporal relevance relative to the query-time

$$\text{TempRet}(q,d,q_t,d_t)=s(q,d)+\tau(q_t,d_t)$$

Gade, Anoushka, Jorjeta G. Jetcheva, and Hardi Trivedi. "**It's about time: Incorporating temporality in retrieval augmented language models.**" In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pp. 75-82. IEEE, 2025.

It's About Time: Incorporating Temporality in Retrieval Augmented Language Models

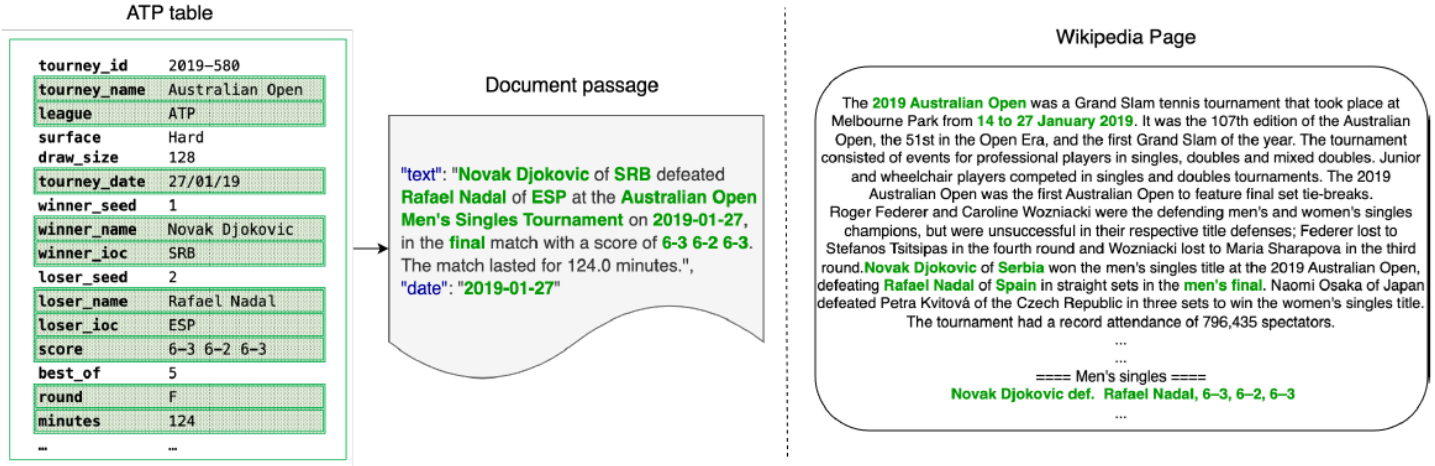


Figure 3: Example of tabular data to passage conversion. We convert relevant features of the tabular dataset to textual passages.

Gade, Anoushka, Jorjeta G. Jetcheva, and Hardi Trivedi. "**It's about time: Incorporating temporality in retrieval augmented language models.**" In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pp. 75-82. IEEE, 2025.

MRAG: Separating Semantic and Temporal Relevance

Three Modules

1. Question Processing
Decompose query into:
- Main Content (MC)
- Temporal Constraint (TC)

Example:
Who won the latest America's Next Top Model as of 2021?
MC = Who won America's Next Top Model?
TC = latest as of 2021

2. Retrieval & Summarization
- Retrieve semantically relevant passages
- Split/summarize evidence
- Remove temporal distractors

3. Semantic–Temporal Hybrid Ranking

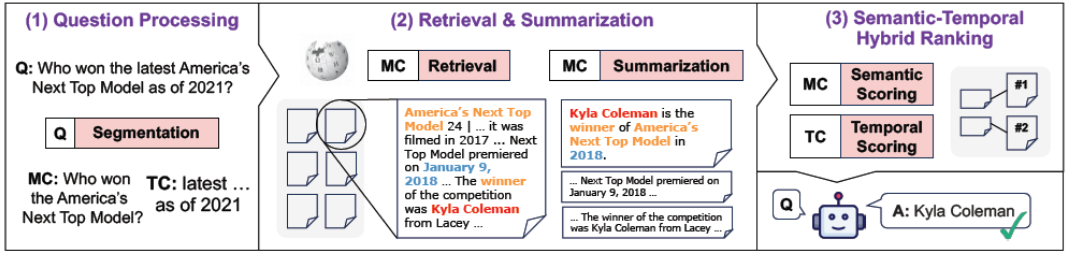


Figure 3: An overview of the MRAG framework, consisting of three key modules: question processing, retrieval and summarization, and semantic-temporal hybrid ranking. The question processing module separates each query into the main content (*i.e.,* MC) and the temporal constraint (*i.e.,* TC). The retrieval and summarization module finds the most relevant evidence based on the main content and summarizes or splits these evidence into fine-grained sentences. The hybrid ranking module combines symbolic temporal scoring and dense embedding-based semantic scoring at a fine-grained level to determine the final evidence ranking.

Results:
MRAG improves retrieval and QA
+11% Evidence Recall
+9.3% Answer Recall
Better downstream QA accuracy on TEMPRAGEVAL benchmark

Final score combines: Semantic Relevance × Temporal Relevance
Temporal score is computed separately from timestamps and temporal constraints.

MRAG: A Modular Retrieval Framework for Time-Sensitive Question Answering (https://aclanthology.org/2025.findings-emnlp.167/) (Zhang et al., Findings 2025)

TsContriever: Time-sensitive Retrieval-Augmented Generation for Question Answering

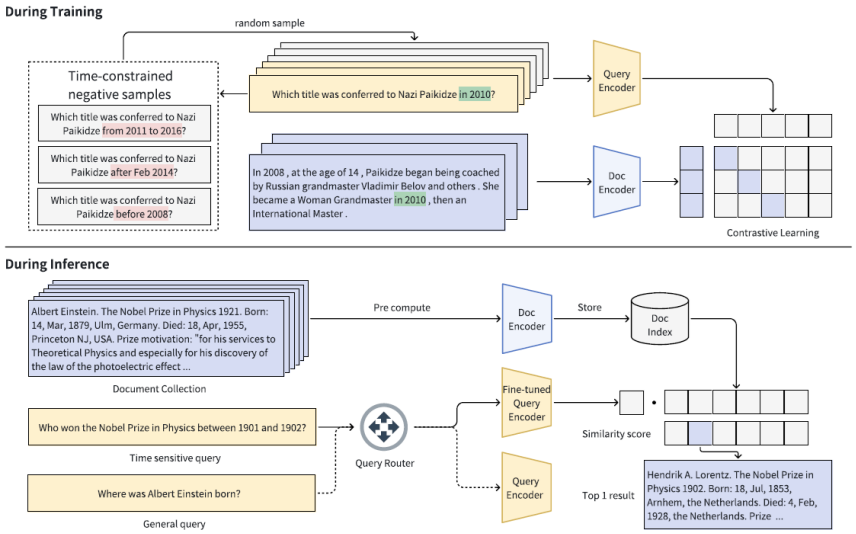
TsContriever: Learning Temporal Relevance

Training

- Construct **time-sensitive question–document** pairs.
- Generate **temporal hard negatives** that are semantically similar but correspond to different time periods.
- Train using **contrastive learning** to distinguish temporally relevant from temporally incorrect evidence.

Inference

- A **query router** first detects whether a query contains temporal constraints.
- Temporal queries are encoded using the fine-tuned temporal encoder.
- Documents are ranked by both semantic and temporal relevance.



Key Innovation:

Temporal contrastive learning using same-topic, different-time negatives teaches the retriever to respect temporal constraints.

Feifan Wu, Lingyuan Liu, Wentao He, Ziqi Liu, Zhiqiang Zhang, Haofen Wang, and Meng Wang. 2024. T**ime-Sensitve Retrieval-Augmented Generation for Question Answering. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24).** Association for Computing Machinery, New York, NY, USA, 2544–2553. https://doi.org/10.1145/3627673.3679800

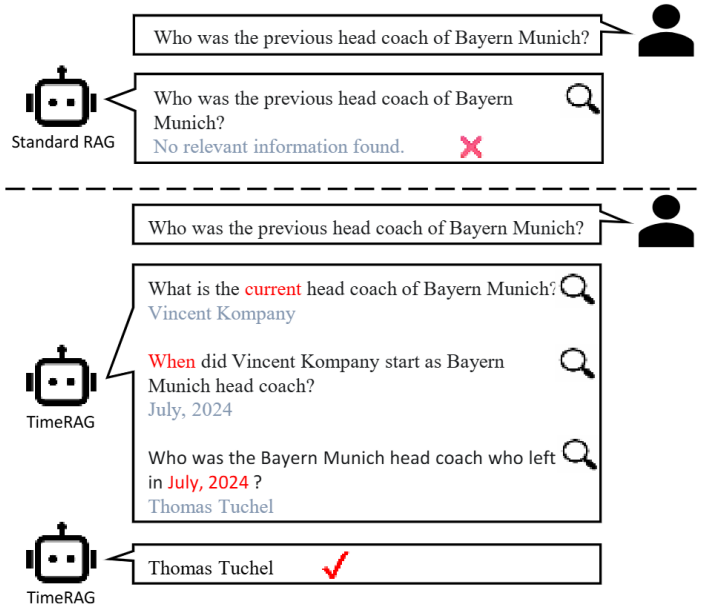
TimeRAG: Enhancing Complex Temporal Reasoning with Search Engine Augmentation (CIKM 2025)

Motivation

Existing RAG systems can retrieve fresh information but struggle with:

- Implicit temporal constraints
  - "previous"
  - "before"
  - "since"
- **Multi-hop temporal reasoning**
- Integrating facts across different time periods

As a result, even simple temporal questions can fail.



Temporal QA requires:

Retrieval + Temporal Reasoning + Multi-hop Search rather than a single retrieval step.

TimeRAG converts temporal QA into an iterative search-and-reason process, using temporal query decomposition and multi-hop retrieval to answer questions involving evolving facts.

Zhao Wang, Ziliang Zhao, and Zhicheng Dou. 2025. **TimeRAG: Enhancing Complex Temporal Reasoning with Search Engine Augmentation.** In Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25). Association for Computing Machinery, New York, NY, USA, 3230–3239. https://doi.org/10.1145/3746252.3761425

TimeRAG Architecture

Two Core Modules

1. Query Decomposition (QD)

Breaks a complex temporal question into atomic time-event questions.

Example:

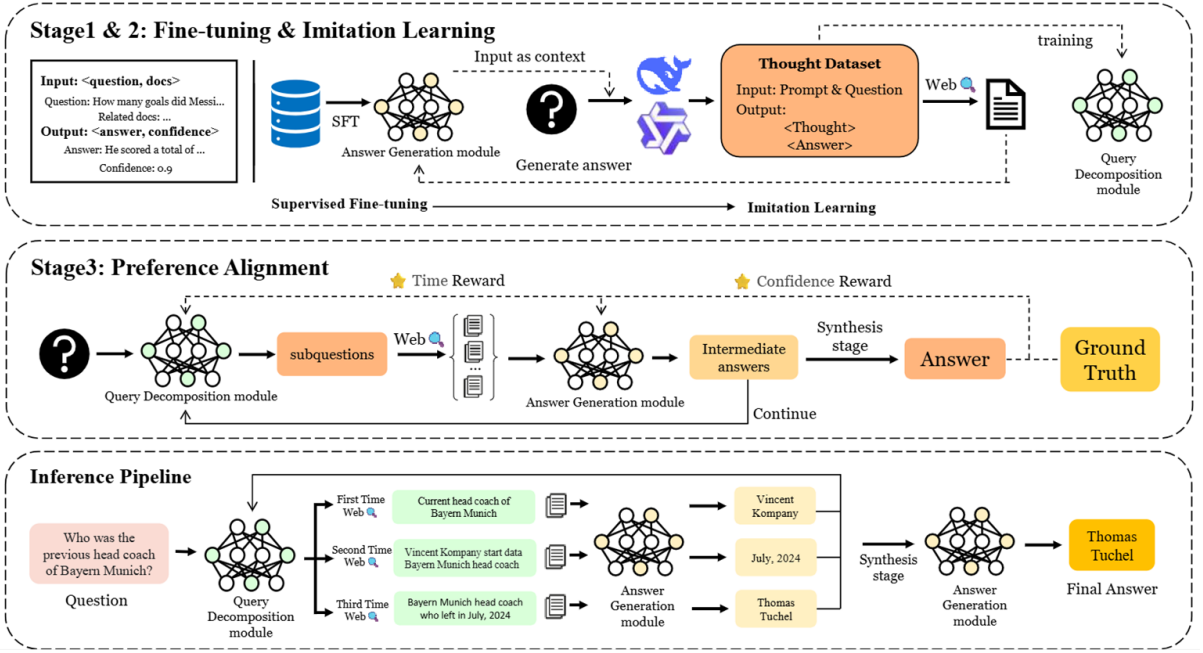
Original Question→ Current Coach? →Start Date? →Previous Coach?

2. Answer Generation (AG)

For each sub-question:

- Retrieve web evidence
- Interpret timestamps
- Generate answer
- Produce confidence score

Then synthesize all intermediate answers into the final response



TimeRAG converts temporal QA into an iterative search-and-reason process, using temporal query decomposition and multi-hop retrieval to answer questions involving evolving facts.

Zhao Wang, Ziliang Zhao, and Zhicheng Dou. 2025. **TimeRAG: Enhancing Complex Temporal Reasoning with Search Engine Augmentation.** In Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25). Association for Computing Machinery, New York, NY, USA, 3230–3239. https://doi.org/10.1145/3746252.3761425