A Comprehensive Taxonomy of Temporal Dimensions in Natural Language Queries

Ivano Lauriola
lauivano@amazon.com
Amazon AGI
California, USA

Abstract

Effectively understanding and modeling the temporal aspects of user queries is crucial for Information Retrieval (IR) and Question Answering (QA), particularly in contexts that demand freshness, historical accuracy, or temporal reasoning. In this paper, we present a formal and comprehensive taxonomy for classifying natural language queries along four dimensions: (i) temporal understanding, (ii) reasoning type, (iii) time sensitivity, and (iv) trendiness. This framework captures a wide range of temporal intents, from static factual questions to dynamic, event-driven queries. The taxonomy is designed as a foundation for diagnostic evaluation of QA and IR systems by providing a systematic categorization of the temporal properties of queries. This structure enables practitioners to identify, isolate, and analyze system behaviors across diverse temporal scenarios, helping uncover specific failure modes related to temporal reasoning, sensitivity, data freshness, and relevance. We apply the taxonomy to classify queries from four public datasets: MS MARCO, FreshQA, RealtimeQA, and SituatedQA, revealing systematic gaps and underrepresented time-sensitive categories. As a diagnostic case study, we stratify the accuracy of three LLMs by temporal dimension, showing that the taxonomy exposes failure patterns that the aggregate metrics conceal.

CCS Concepts

• **Information systems** → **Query representation**; *Query intent*; Question answering; *Traffic analysis*; **Evaluation of retrieval results**.

Keywords

Question Answering, Time Sensitivity, Temporal Reasoning, Trendiness

ACM Reference Format:
Ivano Lauriola. 2026. A Comprehensive Taxonomy of Temporal Dimensions in Natural Language Queries. In *Proceedings of the 2026 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '26), July 25, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3805713.3820426

1 Introduction

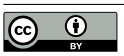
Modern Information Retrieval (IR) or Question Answering (QA) engines require the ability to understand temporal information from users' requests in order to model responses that are temporally aligned to the intent. This aspect is particularly important in real-world applications such as web search, digital assistants, or agents, where users often expect answers that are not only semantically correct but also temporally accurate. For example, a user querying "*Is Messi playing for Barcelona?*" expects an answer grounded in the current days rather than historical data. Conversely, "*When did Messi win his first Ballon d'Or?*" requires accurate retrieval and processing of a past event.

Previous work has studied approaches to learn temporal representations of facts or questions through Language Models (LMs) [13, 32, 45] or Knowledge Bases (KBs) [11, 18, 30, 36]. However, although temporal behavior in natural language queries is multidimensional, spanning (i) how temporal intent is expressed, (ii) what reasoning it demands, and (iii) how answers evolve over time, most past studies focus on subsets of temporal aspects, including simple reasoning over explicit temporal expressions [12, 14, 25], temporal understanding over complex questions [43], reasoning over temporal relations on a reference text (e.g.: "before", "after"), or retrieving information given a reference query timestamp [22, 23].

This fragmentation has practical consequences. When a system is evaluated on a single benchmark or temporal dimension, its temporal failures are invisible. A model may handle recency-based questions (e.g.: "*current Bitcoin price*") well while failing on temporal arithmetic (e.g.: "*How many days ago did Olympic games terminate*"), or resolve explicit date references but struggle with implicit ones (e.g.: "*what was the cost of homes in US last summer?*"). An aggregate accuracy score conflates all of these aspects. Prior work on temporal query classification has addressed parts of this problem: Jones and Diaz [15] distinguished atemporal from temporally ambiguous queries, and Hasanuzzaman et al. [9] classified queries by temporal intent. However, these frameworks each capture a single dimension of temporal behavior and do not account for how answers change over time or how query popularity evolves.

We argue that a unified framework is needed to have a comprehensive overview of models' performance. To this end, we propose a multi-dimensional taxonomy: (i) Temporal Understanding, i.e.: the recognition of implicit or explicit temporal constraints representing the intent ("*who won FIFA cup in 2020*"); (ii) Temporal Reasoning, that is the cognitive process to model a response that satisfies the temporal constraints; (iii) Time Sensitivity, that describes how answers and knowledge change over time; and (iv) Trendiness, that is the evolution of the popularity of a question. Each dimension is formalized mathematically and decomposed into fine-grained sub-categories.

The taxonomy aims to provide a structured framework for identifying blind spots in QA/IR systems and understanding temporal



ICTIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2600-2/2026/07
https://doi.org/10.1145/3805713.3820426

failures. As proof of concept, we apply the taxonomy to categorize queries from MS MARCO [24], FreshQA [40], RealtimeQA [17], and SituatedQA [47] confirming systematic gaps: MS MARCO is dominated by atemporal queries, while even time-focused datasets underrepresent categories such as questions whose answers are extremely volatile (e.g.: stock prices) or future-looking events. To further demonstrate the taxonomy's diagnostic value, we stratified the accuracy of three LLMs, Qwen3-32B, Mistral-v3-14B, and Mistral-v3-3B, on RealtimeQA by taxonomy dimension. The taxonomy exposed sharp disparities across models and temporal categories. For instance, queries requiring very fresh content showed an accuracy drop of 16.7% and 3.9% absolute compared to non time sensitive questions on Qwen and Mistral-14B respectively. Moreover, although Mistral-14B outperforms the smaller version by 14 percentage points (56.6% vs 42.6%) in accuracy, the two versions show the same performance on queries requiring temporal resolution of static events or queries requiring answers grounded in the last year of world knowledge. These patterns would remain hidden behind an aggregate score.

2 Related Work

The study of models with temporal awareness has been widely studied for decades. Early foundational work, including Allen's interval algebra [1] (and further extensions, e.g.: [2, 8, 19]), established a formal logic framework to reason over temporal relations between events. Building on this foundation, NLP research has developed increasingly complex datasets and models for temporal reasoning. Early resources [14, 25] focused on simplified reasoning over timestamps associated with the question and, in some cases, a reference paragraph or document, e.g.: mined from Wikipedia [4]. Training approaches were introduced, e.g: based on contrastive temporal training [32] or other variations [13, 45]. These resources were very limiting as models relied on explicit timestamps to be consumed. With the rise of LLMs, several challenging datasets for temporal reasoning have emerged. Notably, TRAM [42] and TIMEBENCH [6] cover a wide range of temporal tasks, including ordering, arithmetic, causality, and temporal NLI. TIMEBENCH aggregates 10 prior datasets [35, 37, 39, 49] into 16 subtasks. To address the simplicity of earlier benchmarks, [43] proposed more lexically complex, LLM-level temporal questions involving explicit time references, ordering, and comparison tasks. A related growing research direction goes under the name of temporal common sense reasoning [7, 10, 31, 49], which refers to the ability to infer implicit temporal knowledge that is assumed to be understood. In parallel, structured methods using temporal knowledge bases [11, 18, 30, 36] have been developed, though often limited by coverage.

Little work has studied questions whose temporal context, i.e.: the time the question is asked, is relevant. Recent studies [23] highlighted that these queries are underrepresented in existing resources. Some exceptions include StreamingQA [22], which features 14 years of time-stamped news articles, and small subsets in RealTimeQA [17] and FreshQA [40]. On the retrieval side, Kanhabua and Nørvåg [16] developed learning-to-rank methods for time-sensitive queries blending document relevancy and temporal features, and Cheng et al. [5] studied *timely queries*, measuring freshness requirements through content change in relevant documents over time.

Focusing on the trendiness, previous work has shown that the query popularity follows a Zipfian distribution [38], and they can be categorized into Head, Torso, or Tail buckets. The Head-to-Tail framework has been widely used to evaluate models performance, including LLMs [34]. Further studies analyzed the evolution of trending topics on Twitter [20, 26], showing how some hashtags are subject to sudden spikes in popularity, and on the web [29] through time-series analyses. However, there is very limited attention beyond head-to-tail categorization.

A broader issue across prior work is inconsistent terminology and task definitions. For example, Time Sensitivity has been variably defined as a model's ability to handle temporal expressions [30], its ability to pay closer attention to temporal information [45], or the presence of a time-related expression [12]. Similarly, Temporal Reasoning has been described as the ability to reason over temporal information conditioned to a query [4], the process of combining different temporal cues to derive new temporal information [19], or the capability of understanding, representing, and predicting time-sensitive contexts [46]. The ability to understand the temporal clues has been referred as Temporal Understanding [4] or Temporal Information Extraction [19]. Moreover, although existing work focus on important aspects of temporal query behavior, they typically focus on a single dimension. Jones and Diaz [15] introduced temporal profiles of queries, characterizing queries as *atemporal*, *temporally unambiguous*, or *temporally ambiguous* based on the distribution of retrieved documents over time. Hasanuzzaman et al. [9] proposed a multi-objective ensemble learning approach to classify queries along their temporal intent, distinguishing between *past*, *recency*, *future* and *atemporal* intents. In contrast, our work offers formal definitions of understanding, reasoning, time sensitivity, and other temporal aspects of a query, and organizes them through a formal taxonomy, that includes temporal sensitivity, understanding, reasoning, and trendiness.

Complementary to query-level classification, temporal expression extraction systems such as HeidelTime [33], SUTime [3], and the TimeML specification [28] have established rich typologies for temporal expressions in text, including dates, times, durations, and sets. These systems focus on identifying and normalizing temporal mentions at the token or phrase level within documents. Our work operates at a different granularity: rather than extracting temporal expressions from text, we classify the temporal *intent* and *behavior* of queries as a whole, including dimensions (such as time sensitivity and trendiness) that have no counterpart in expression-level annotation.

3 Preliminaries

Let q be an input question (or query) from a certain distribution Q and $t \in \mathcal{T}$ a timestamp representing when the question is asked. We denote $\mathcal{T}$ as the abstract space of ordered time points, that is continuous or discrete depending on the application (e.g.: days, document versions, milliseconds).

The function $\psi : Q \times \mathcal{T} \rightarrow \mathcal{P}(\mathcal{A})$ simulates an oracle QA or IR system that, given a query q at timestamp t, returns all possible correct answers from a universal distribution $\mathcal{A}$. $\mathcal{P}(\mathcal{A})$ denotes

the power set of $\mathcal{A}$, representing all subsets of answers, supporting multi-answer queries, e.g.: *Tell me a joke*, or *Give me a mushroom pasta recipe*. In other words,

$$\psi(q,t) = \{a \in \mathcal{A} \mid a \text{ is correct for } (q,t)\} \quad (1)$$

is the set of all correct answers that the system would generate for q at a given timestamp t. Note that ψ may represent any system that returns or generates content to answer a query. This includes (i) a Web Based Question Answering that gets snippets/answers from the web [41], (ii) an LLM that generates a response from its knowledge or using a retrieval system as RAG [44], (iii) a Knowledge-Based QA service [27], or (iv) a document retrieval engine supporting human search or agents [21, 48]. We define a Response as any output produced by these systems, ranging from a sentence or snippet to long-form generated text or a full web document.

3.1 Answer Correctness

The correctness of a response (used in Eq. 1) depends on multiple aspects. First, as widely accepted in IR and QA research, a correct response is semantically and topically relevant. The content satisfies the intent of the question. Second, the response is temporally grounded (and relevant) and linked to the question intent. The response is not outdated nor refers to a different, undesired, moment in time. Other factors that influence the correctness, e.g.: the geographic context, are outside the scope of this work. To define the concept of temporal relevance, we first introduce the Question Anchor and Response Scope.

Question anchors are explicit or implicit time references in a question that ground it to a specific point in time, providing temporal context for interpreting, retrieving, or generating information. They represent the *when* of a question, defining temporal constraints against which a response should be evaluated. Formally, we characterize anchors as the set of all timestamps implied by the intent of the question, $\alpha : \mathcal{Q} \times \mathcal{T} \rightarrow \mathcal{P}(\mathcal{T})$. For instance, α("*Who was the president of the USA in 2010?*") refers to a specific time (2010) and who covered the role outside this constraint is temporally irrelevant for the question.

Definition 3.1 (Temporal understanding). The temporal understanding is the process of identifying the temporal constraints of a question q in a given temporal context t, formalized as finding $\alpha(q,t)$.

Understanding temporal constraints (α) is half of the work. The generated response has to be aligned to these constraints to be correct. To formalize this, we define the temporal scope of a response as a function $\beta : \mathcal{A} \rightarrow \mathcal{P}(\mathcal{T})$ which maps a response to the set of timestamps during which it is valid. For example, an old article referring to Obama as the *current* president of USA would be temporally valid only for timestamps between 2009 and 2017, becoming outdated afterwards. The temporal relevance of a response is here defined:

Definition 3.2 (Temporal relevance). An answer a is temporally relevant for a question q and a timestamp t, thus $a \in \psi(q,t)$, iff the question anchor is contained in the temporal scope of the response, that is $\alpha(q,t) \subseteq \beta(a)$.

In the example above, the anchor 2010 is included in the scope of the document 2009 . . . 2017, making it temporally relevant. Note that the answer "*Ronald Reagan served as the president of the US from 1981 to 1989.*" is factually correct at all points in time, and thus has $\mathcal{T}$ as temporal scope. However, while it is temporally relevant to a question asking for the president in 2010 (we have $2010 \in \mathcal{T}$), it is semantically irrelevant, and thus incorrect as an answer. Similarly, a document listing all US presidents with the term start/end dates has the same temporal scope $\mathcal{T}$ as it contains factual statements. It is also semantically relevant as it includes the information to answer the question. Finally,

Definition 3.3 (Temporal reasoning). Temporal reasoning refers to the process to (i) generate, (ii) verify, or (iii) retrieve a temporally relevant response. Formally, given ψ, q, t, find a such that $\alpha(q,t) \subseteq \beta(a)$.

In summary, the temporal understanding aims at finding the constraints from the question intent while the reasoning comprises all operations to make sure a response satisfies the constraints.

4 Taxonomy of Temporal Dimensions

Our taxonomy aims to capture all key temporal aspects of a question and provide practical tools to evaluate QA/IR systems and uncover blind spots. It covers four main dimensions: (i) Temporal understanding, (ii) Reasoning, (iii) Time Sensitivity, and (iv) Trendiness. These dimensions cover query intrinsic and extrinsic temporal aspects. While Understanding and Reasoning focus on the temporal intent and the process to build a valid response, Time Sensitivity and Trendiness depend on external factors, including how world knowledge and the users' interest change.

4.1 Temporal Understanding

As previously defined, the temporal understanding aims at finding the temporal constraints (or anchors) of a question. We categorize questions based on various characteristics of their anchors, which are: (i) type of the anchor, i.e.: how it is expressed in the text; (ii) the shape of the time window constraining the question; (iii) the position of the anchor with respect to the reference time.

4.1.1 Anchor type. The type of an anchor reflects how it is expressed in the question text, and it can be:

- **Explicit** : refers to queries with an explicit anchor in the text, e.g.: "*S&P500 December 2020*". The anchor is an unambiguous date or time range and it doesn't depend on question timestamp (temporal context), $\forall(t', t'') \in \mathcal{T}^2 : \alpha(q,t') = \alpha(q,t'')$.
- **Implicit** : The question suggests a time period that must be inferred or resolved, e.g.: "*Population Europe during WW2*". The query mentions the historical moment of interest, without providing an explicit anchor. The anchor does not depend on query context.
- **Dynamic** : queries whose anchors change over time and require additional temporal context to be resolved. The question "*Weather last week*" refers to different days, depending on when the question is asked. Knowing the timestamp associated to the question is vital to provide a correct answer,

$\exists(t', t'') \in \mathcal{T}^2 : \alpha(q, t') \neq \alpha(q, t'')$. Through this paper, we refer to these questions as Time Sensitive.

No Anchor : Queries whose answers do not need to be contextualized in a certain time window, e.g.: "*calories in a cucumber*". $\alpha(q, t) = \mathcal{T}$. Names are not temporal mentions, and thus not anchors. "*Who plays The Day After Tomorrow*" has no anchors as *tomorrow* is part of the movie title.

Irrelevant : queries with an implicit or explicit anchor that are irrelevant for the purpose of generating a response. The anchor does not add real constraints, e.g.: "*How tall is Mount Everest today*" or "*What was Einstein's theory of relativity in the early 20th century*". $\alpha(q, t) = \mathcal{T}$.

Note that we distinguish between questions with implicit anchors, which define a temporal constraint (and thus $\alpha(\cdot, \cdot) \neq \mathcal{T}$), and questions referring to time-locked events. Let's consider q_1: "*Population Europe during WW2*" and q_2: "*Who won WW2*". While both lack explicit time expressions, q_1 is temporally constrained (*when* $\rightarrow$ *during WW2*), $\alpha(q_1, \cdot) = \{1939 \dots 1945\} \subset \mathcal{T}$. The anchor is Implicit. Differently, q_2 is not temporally constrained $\alpha(q_2, \cdot) = \mathcal{T}$. The question asks for a specific event locked in time rather than information contextualized in a certain time period. Similarly, both No Anchor and Irrelevant map to $\alpha(q, t) = \mathcal{T}$. The distinction is linguistic rather than formal: Irrelevant queries contain a temporal expression that adds no constraint, while No Anchor queries contain none.

The anchor type describes an increasing level of temporal reasoning capabilities. Explicit anchors are easiest to handle as they do not require complex understanding. Implicit ones require the resolution of events. For instance, when answering the query "*Population Europe during WW2*", the system is expected to frame the time window of interest in the range 1939-1945, without knowing this time frame explicitly. Dynamic anchors require the system to know temporal context and align it to a moving window. For instance, an IR system supporting an LLM through RAG, cannot return a document titled *FIFA World cup, news today* for the query "*results FIFA*" if the document refers to past events, and thus is not aligned to the query timestamp. Moreover, we highlight that Dynamic anchors may be hidden in the query text. For instance, "*average salary in HK*" and "*average salary in HK today*" are semantically equivalent both refers to *today*.

4.1.2 Anchor shape. Queries can be categorized based on the time frame represented by the anchor. Intuitively, the shorter the time frame, the more precise the system has to be. A short anchor like "*Amazon stock price on July 1st*" demands high precision, requiring evidence grounded in that specific day. In contrast, a long anchor such as "*best invention in the 20th century*" is less restrictive, enabling the system to draw from a potentially wider range of sources. Based on the window size and shape, question anchors can be:

Short : the anchor refers to a time span that is shorter than or equal to a given value (e.g.: 1 week or day). Formally, $\max(\alpha(q, t)) - \min(\alpha(q, t)) \leq \Delta$. Example: "*Weather Culver City*", where *today* is a hidden dynamic anchor.

Long : the anchor refers to a time span that is longer than a given threshold, that is: $\max(\alpha(q, t)) - \min(\alpha(q, t)) > \Delta$. Example: "*Deaths by car accident from 2005 to 2015*".

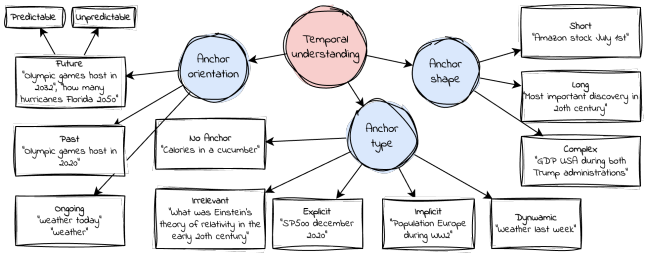


Figure 1: Visual description of queries with respect to their temporal anchor.

Complex : $\alpha(q, t)$ includes non-contiguous periods. For instance, the query "*GDP USA during both Trump administrations*" refers to two separate time spans. Formally, $\exists\{t^-, t^+\} \subseteq \alpha(q, t), \exists t' \in \mathcal{T} : \; t' \notin \alpha(q, t) \wedge t^- < t' < t^+$.

The value Δ can be adjusted to fit the specific application or evaluation setting (e.g.: 1 day, 1 week, 1 month). Without loss of generality, Short/Long shapes can be bucketed into intuitive ranges such as daily, weekly, monthly, or yearly.

4.1.3 Temporal orientation. Anchors don't always refer to the past. For example, "*Olympic Games host 2028*", asked in 2025, points to a future event, of which the answer (Los Angeles) is already known. Similarly, "*How many hurricanes will hit Florida in 2035*" also targets the future, but responses are unknown or speculative. These queries are potentially more complex as they require system to handle uncertainty in knowledge by detecting and consuming opinions, forecasts, or estimations. Based on the orientation, anchors can refer to:

Future : anchors that refer to future events. Can be Predictable or Unpredictable, depending on the current knowledge on the future event. Formally, $\min \alpha(q, t) > t$;

Past : anchors that refer to events or dates already occurred, "*weather 2 weeks ago*". $\max \alpha(q, t) < t$;

Ongoing : anchors that refer to ongoing events, "*weather today*", that is:
$\min \alpha(q, t) \leq t \leq \max \alpha(q, t)$.

Note that all dimensions described, anchor type, shape, and orientation, are orthogonal to each other. A visual description of the temporal understanding characteristics is available in Figure 1.

4.2 Temporal Reasoning

We now categorize the different types of reasoning required to process answers whose temporal scope aligns to the question anchor. We identify multiple fundamental reasoning types, each of them represents a distinct cognitive process.

Ordinal/Positional Involves locating an event/entity in a temporal sequence, often using terms like "first", "last", or "4th", e.g.: "*Who was the 4th president of the US?*".

Comparison Requires reasoning about relative positions in time (before/after) or temporal distance, e.g.: "*What happened before the French Revolution?*"

Arithmetic Requires mathematical operations on temporal values, e.g.: (i) explicit calculus, "*What year was 3 years ago?*",

(ii) age computation, "*How old is Lady Gaga?*", (iii) duration, "*How long the cold war?*".

Resolution Requires mapping named events or periods to their specific temporal boundaries. Involves understanding the complete temporal scope of historical events, including potential ambiguity in start/end points, e.g.: "*Who was the CEO of Microsoft when Windows 95 launched?*". The resolution can refer to Static events or Dynamic timestamps ("*two years ago*").

Recency Aligns questions to the current timestamp. Often involves resolving temporal expressions like "today", "recent", "now", or implicit freshness, e.g.: "*news NBA*". The system must judge how recent is recent enough and when information becomes outdated.

None The query requires no temporal reasoning. Example: "*How many calories in a cucumber?*"

Questions may require multiple reasoning types. For instance, "*What was Apple's last iPhone release?*" involves both Ordinal (last) and Recency (up-to-date info). Note that temporal expressions alone do not necessarily determine reasoning type: "*When did it last rain?*" requires Recency, while "*Who was the last Tsar of Russia?*" is Ordinal as it does not refer to recent events.

Note that there is an alignment between Resolution and anchor type Implicit. However, an implicit anchor refers to the existence of *unknown* temporal constraints. The resolution refers to the cognitive *operation* needed to map the question to the specific event or period that produces the answer, transforming the event mention into the temporal constraints.

Furthermore, it is worth noticing that Dynamic anchors and Recency reasoning are correlated but operate at different levels of analysis. The former describes how the temporal constraint is expressed, the latter describes the cognitive operation of aligning to the current timestamp. For instance, "*Who is the current NBA champion?*" has a Dynamic anchor and requires Recency reasoning, while "*Who won the NBA championship 5 years ago?*" also has a Dynamic anchor but requires Resolution-dynamic rather than Recency.

4.3 Time Sensitivity

There is a class of questions with dynamic temporal anchor requiring the alignment between the time the answer refers to and the time when the question is asked. Most importantly, the responses change dynamically over time. We refer to these questions as Time Sensitive (TS). Formally, a query $q \in \mathcal{Q}$ is time-sensitive iff:

$$\exists(t', t'') \in \mathcal{T}^2 \mid t' \neq t'' \wedge \psi(q, t') \neq \psi(q, t'') \tag{2}$$

This means the set of correct responses differs for at least two timestamps. Alternatively, the condition can be simplified as $\alpha(q, t') \neq \alpha(q, t'')$. For instance the response to the question "*Who is the president of the U.S.?*", which has an implicit dynamic anchor ("today"), usually changes every 4 years. The temporal scope of the response $\beta(a)$ shifts with the query time t, reflecting changes in presidency over time. The way the anchor and answer scope evolve over time characterize the time sensitivity. This evolution is encoded into the query Trace, $V_q^{\mathcal{T}} = \langle \psi(q, t^0), \psi(q, t^1), \dots \rangle$, which reflects updates in knowledge or world events. In addition, we use the symbol $\mathcal{V}_q^{\mathcal{T}}$ to denote the set of distinct groups of answers from $V_q^{\mathcal{T}}$, that is $\mathcal{V}_q^{\mathcal{T}} = \{\psi(q, t), \forall t \in \mathcal{T}\}$.

We introduce a time window $\tau(t', t'') = \{t \in \mathcal{T} \mid t' \leq t \leq t''\}$ to parameterize the query trace. Even if a comprehensive window ($\tau = \mathcal{T}$) captures the full history of a query and its evolving answers, it has two drawbacks. First, an exhaustive window can make almost any question appear TS as widely accepted answers can shift over long time. For instance, "*What's the biggest country in Europe?*" might change in the distant future. Similarly, ulcers were once linked to stress or spicy food, until this myth was debunked in the 1980s. Second, practical QA/IR evaluations often focus on recent knowledge changes. A response that changed 30 years ago may be less relevant than one that changed in the past few years. For example, when evaluating a new LLM, we may closely analyze TS questions whose answers changed after the model training.

Definition 4.1 (Time Sensitive). A question q is Time Sensitive in a given interval τ if its responses change over time, that is $|\mathcal{V}_q^{\tau}| > 1$. Otherwise, the question is Evergreen.

Time Sensitive questions can be characterized through different dimensions: (i) Volatility, i.e.: how frequently the answer changes in a certain time window; (ii) distance with the Last update of the answer; and (iii) Answer expiration.

4.3.1 Volatility. The volatility of a TS question describes how frequently the answer changes in a specific time window, it can be:

Live : Questions with responses that change continuously or in near real-time. For these, only the latest available knowledge can be correct. Neither LLM parameters nor outdated indexes can answer. Examples include stock prices or breaking news, where even a 1-hour-old document may be obsolete. Formally, such queries may have $|\mathcal{V}_q^{\tau}| \approx |V_q^{\tau}| \approx |\tau|$.

High : queries whose responses change with a frequency greater than a specified value θ in a given window τ, that is $|\mathcal{V}_q^{\tau}| > \theta$. The threshold is a parameter and can be defined based on the use case and requirements. For instance, a query can be TS-High if it changed more than once per week (on average) in the last year, with $\tau = 1$ year and $\theta = 52$. Queries related to sport events, politics, economy, may easily fall into this category.

Low : Queries whose responses change with a frequency lower or equal than a specified threshold, $|\mathcal{V}_q^{\tau}| \leq \theta$.

Recurrent : Questions whose answers follow a regular, predictable update cycle. For instance, "*Microsoft financial results*" changes every quarter, and sports scores update weekly during active seasons.

Note that, while Live, High, and Low are mutually exclusive, Recurrent is potentially a separate and orthogonal class.

4.3.2 Answer age. While volatility describes how often a response changes, it doesn't capture how recent the current response is relative to the query time. Given a query q, its timestamp t, and a response $a \in \psi(q, t)$, the age of a corresponds to the time elapsed since it last changed:

$$age(a) = \max_{d>0}(d)\ s.t. : a \in \psi(q, t-d) \quad = t - \min(\beta(a)) \tag{3}$$

This reflects the most recent point in time when the answer became valid, based on its temporal scope $\beta(a)$. Note that this value is complementary to the frequency. The frequency tells the expectation of changes over time while the age tells how fresh the response has to be given a reference timestamp. For instance, "*What's the latest iPhone*" typically changes every year (frequency: Low). Given that the latest models were launched on September 9 2024 (as of August 2025), if the question is asked on September 8 2024, then it can rely on year-old information. But if asked on September 10, 2024, shortly after a new launch, the system needs very fresh information. Volatility and answer age are intuitively related: high-volatility questions tend to have low answer age, as their answers change frequently; conversely, low-volatility questions often have longer-lasting answers, resulting in higher expected age. Two edge cases are Evergreen and Live. When the question is evergreen, the volatility is 1 and the age is the maximum value. When the question is live, then we have $age \approx 0$ by definition, which corresponds to highest volatility (1 timestamp $\leftrightarrow$ 1 different answer). In the context of evaluating a QA system, knowing when a response changed is vital to highlight and quantify the velocity of information refresh. A response can be categorized into intuitive buckets based on the value of their age: live, day, week, month…

4.3.3 Expiration. Age and Volatility describe how the responses change over time but do not provide information about the future. The Expiration of a response $a \in \psi(q,t)$ mirrors the age, representing for how long it is expected to be valid,

$$exp(a) = \max_{e>0}(e)\ s.t.: a \in \psi(q, t+e) \quad = \max(\beta(a)) - t \qquad (4)$$

Note that expiration differs from the predictability of a future-looking anchor. The expiration indicates how far into the future a given response remains valid. It may be useful for defining cache expiration policies or detecting when the knowledge of an LLM becomes outdated. In contrast, predictability of the anchor tells whether a valid answer for a future event exists at query time. Based on the expiration, a response can be:

- **Known** : The expiration date is predictable. For example the response to "*Who won the FIFA World Cup?*" is known to remain valid until the next edition of the event. Even if the answer does not change, we know when updated knowledge will be needed, and Eq. 4 can be estimated accurately.
- **Unknown** : Expiration depends on unpredictable events, e.g.: "*How many hurricanes hit Florida in the last decade?*" We cannot predict if and when the next relevant event will occur, nor whether the answer will change. Eq. 4 cannot be reliably estimated.
- **Slow** : The answer is expected to change eventually, though the timing is uncertain. For instance, "*Who won the most Olympic gold medals?*" The record may hold for years, but change is inevitable. An approximate expiration can be computed, possibly with confidence intervals.

Definitions and shades of time sensitivity depend on the structure of responses. In our view, the time sensitivity is a property of the question characterized by its responses. Answers are pieces of texts well defined and anchored temporally that do not change their value. Rather, it is their temporal correctness for a query that changes. A summary of time sensitive categories is shown in Figure 2.

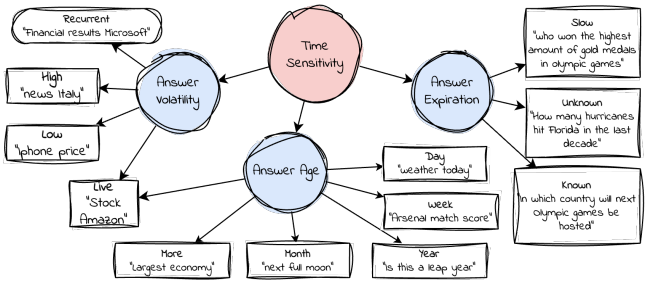


Figure 2: Visual description of queries with respect to their time sensitivity.

4.4 Trendiness

The trendiness of a question describes its popularity over time. We introduce a hierarchical categorization based on popularity history. The first layer focuses on the current popularity. It describes how *important* the question is given a certain timestamp. Very popular questions have higher weight in QA/IR system, as a single negative response may affect the experience of thousands of users. The second layer describes the popularity history, indicating *how* the query reached the current popularity level. Formally, let $\mathcal{S} = \langle (q_1, t_1), (q_2, t_2), \dots \rangle$ be a chronologically ordered stream of questions. Let $\phi(t) = \{q_i | (q_i, t) \in \mathcal{S}\}$ be a retrieval function that outputs all questions asked on a timestamp t. The definition of the space of timestamps $\mathcal{T}$ depends on the desired granularity of the trendiness analysis, e.g.: hours, days, weeks. Let $s : \mathcal{Q} \times \mathcal{Q} \rightarrow \{0, 1\}$ be a question-similarity function that returns a similarity score between two inputs. The function s can arbitrarily be implemented as lexical similarity (e.g.: n-grams overlap), semantic similarity (e.g.: through a Transformer), exact string matching, or other techniques.

Definition 4.2 (Question Popularity). The Popularity of a question q on a timestamp t is defined as the average similarity between q and all questions asked on t, that is: $p(q,t) = \frac{1}{|\phi(t)|} \sum_{q_i \in \phi(t)} s(q, q_i)$.

The higher the value, the higher the importance of the question. Note that the choice of s is crucial. The exact match function may fail to capture paraphrased variations, with the risk of providing very different popularity values for queries "*who is the president of the US*" and "*who is th president of the US*" (misspelled), or "*president US who is*". Differently, a *too* smooth function (e.g.: based on simple embeddings similarity) may cause a flattening of the distribution, where each query is similar to the centroid of the query stream.

Based on the popularity value of a question at a reference timestamp t, the first layer of this categorization comprises two simple categories: Popular and Unpopular, determined by a threshold on the popularity score $p(q,t)$. Within a time window $\tau(t - \Delta, t)$ (e.g.: the past 12 months), Popular questions can be categorized by how quickly they became popular, highlighting whether a question is the result of sustained interest, organic growth, or sudden external events. Formally, we capture this velocity through the slope of the popularity, σ_q, defined as the slope of the best-fit line through τ, computed via linear regression[1].

[1]Other approaches based on derivatives, polynomials, or time-series analyses can be used. However, based on a preliminary study, we opted for a simpler solution.

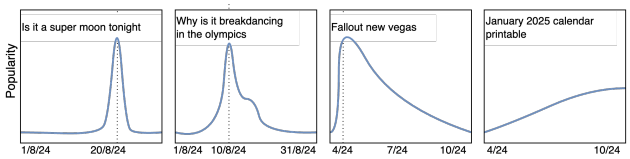


Figure 3: Examples of popularity (in 2024). From left to right: (i) Trending on Aug 20; (ii) Trending from 7 to 13 Aug; (iii) Fading; (iv) Growing.

Head : Queries that show a stable popularity pattern throughout τ. $\sigma_q \approx 0$.

Growing : Queries that show a slow growth in popularity, with $0 < \sigma_q \leq \Theta_+$.

Trending : Queries whose popularity show a sudden spike, transitioning from Unpopular to Popular within a certain small time window τ_{spike}, e.g.: 1 day. These queries are characterized by a high slope, $\sigma_q > \Theta_+$.

Note that we omit the deviation of points from the best-fit line for simplicity. However, the issue is mitigated through an optimal selection of τ. Similarly, Unpopular questions can use the same slope to be described:

Tail/Torso Similarly to Head, the slope is ≈ 0. However, the value of the popularity is lower than the Head threshold.

Fading Queries that are slowly losing interest, $\Theta_- \leq \sigma_q < 0$.

Post Trending : Queries that were Unpopular except for a limited Trending period (e.g.: 1 week). Typically, $\sigma_q \leq \Theta_-$.

Other questions that do not fit this categorization can be Recurrent if their interest changes according to a clear pattern (e.g.: sport questions become more popular during the weekend; the opposite happens to finance), or they can be categorized as Other if none of these categories suit them.

To validate the trendiness categories, we computed popularity signals for a sample of queries using a private large-scale query stream $\mathcal{S}$ (sampled 0.5B total). We implemented the similarity function s as binary function to be 1 if the cosine similarity between two queries encoded through an embedding model (MiniLM-v2-12L[2]) is greater than a certain threshold, 0 otherwise. The daily popularity of a query q becomes the fraction of queries in the stream whose cosine similarity with q exceeds the threshold. The value of the threshold was arbitrarily set to guarantee a query-query semantic equivalence precision of 90%, manually evaluated through expert annotators. This produces a time series of query frequency from which we classified queries based on the shape of their popularity signal. Figure 3 shows representative examples. The plots highlight how the categorization can be made visually. While the underlying query stream cannot be released, the methodology is reproducible on any query log with timestamps using standard sentence embedding models. The thresholds and parameters to categorize the trends into Head, Torso, and Tail, the slope parameters $\Theta_{+/-}$, and the time windows can be fully parametrized. We omit details on how these can be set as they strictly depend on the use case and the query stream.

[2]https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

5 Experiments

To further study the taxonomy, we applied it to existing popular Question Answering and Information Retrieval datasets. This analysis serves two purposes. First, we quantify how different temporal categories are represented in current benchmarks, highlighting possible blind spots. Second, we motivate the need for explicit temporal taxonomies in dataset design and evaluation.

We conduct our analysis on the following datasets:

- MS MARCO [24]: One of the most popular benchmark used for passage retrieval and ranking. Queries are typically assumed to be informational and time-invariant, making it a useful testbed to measure latent temporal sensitivity.
- RealtimeQA [17]: A benchmark designed to evaluate question answering under temporal change, focusing on current events and time-dependent facts. Example: "*Covid-19 vaccinations in the US began for which age group this week?*".
- FreshQA [40]: A QA benchmark encompassing a diverse range of question and answer types, including questions that require fast-changing world knowledge. Example: "*What is the name of the first animal to land on the moon?*".
- SituatedQA [47]: A QA dataset where answers depend on temporal or geographical context. Questions are paired with a timestamp. Example: "*Who is the highest paid sitcom actor of all time?*".

For each dataset, we sample a subset of 1000 queries uniformly at random to ensure coverage of diverse intents and topics. We used Qwen3-235B model to classify each question according to the taxonomy. The system prompt includes a precise description of each class/label (as shown in this document), some examples, and the query to be tested. In the case of RealtimeQA and SituatedQA, the reference timestamp was provided, and the model was asked to consider it in its judgment. The prompt is available in Appendix A. Results of the classification are illustrated in Table 1.

Results highlight multiple key findings. First, FreshQA, RealtimeQA, and SituatedQA exhibit substantially stronger temporal characteristics than MS MARCO. These datasets contain a much higher proportion of recency-driven questions, with RealtimeQA in particular emphasizing need for dynamic reasoning reflecting its focus on rapidly changing information. In contrast, MS MARCO is dominated by temporally unanchored and evergreen queries, with limited presence of explicit or dynamic temporal signals. Second, FreshQA displays a more balanced temporal profile, combining ongoing, past-oriented, and short-term temporal anchors, while RealtimeQA is heavily skewed toward past-oriented and short-lived answers, as evidenced by its high proportions of day- and week-level answer ages and elevated answer volatility. This indicates that, as expected, RealtimeQA places stronger demands on real-time updating and temporal reasoning at inference time. Finally, differences in anchor type further distinguish the datasets: RealtimeQA relies predominantly on dynamic and implicit anchors, whereas MS MARCO overwhelmingly lacks temporal anchors altogether. These findings suggest that standard IR benchmarks substantially underrepresent temporal reasoning requirements, while FreshQA, RealtimeQA, and SituatedQA better capture the challenges faced by retrieval systems deployed in time-sensitive, real-world settings.

Table 1: Distribution of reasoning and temporal attributes.

Category	MS-M.	FreshQA	Realt.QA	Sit.QA
Anchor Type				
Dynamic	16.4	53.4	73.5	46.0
Explicit	2.3	8.1	8.0	12.9
Implicit	1.5	4.4	6.2	2.5
Irrelevant	1.5	1.3	2.3	0.8
No Anchor	78.3	32.2	9.6	37.7
Anchor Shape				
Complex	0.1	0.0	0.2	0.0
Long	2.0	12.0	5.0	11.0
Short	21.4	63.4	86.4	58.8
Anchor Orientation				
Future (Predictable)	0.3	4.9	4.3	2.4
Future (Unknown)	0.8	2.4	2.3	3.4
Ongoing	16.1	28.5	13.5	20.2
Past	6.3	39.5	71.5	43.8
Reasoning Type				
Arithmetic	8.1	8.2	2.9	0.9
Comparison	0.0	1.4	1.6	1.9
Ordinal	6.1	25.9	6.4	34.3
Resolution (Dynamic)	3.5	9.8	25.8	6.4
Resolution (Static)	10.1	13.1	7.3	12.7
Answer Age				
Live	3.5	9.0	5.0	6.3
Day	0.0	1.3	30.4	1.1
Week	0.1	2.9	32.6	1.7
Month	3.1	6.8	7.0	6.8
Year	2.4	19.8	2.0	9.5
More	90.9	60.1	22.9	74.6
Answer Expiration				
Known	9.6	45.1	22.6	41.4
Slow	31.8	5.7	11.9	5.5
Unknown	38.0	46.3	58.1	32.8
None	20.7	2.9	7.4	20.3
Answer Volatility				
Live	1.3	1.3	0.3	0.2
Low	15.8	36.8	8.9	34.4
High	1.5	10.5	36.6	8.9
Evergreen	81.2	51.0	53.9	56.3
Recurrent Volatility				
False	97.9	80.0	93.4	79.7
True	2.1	20.0	6.6	20.3

However, there are some categories that are generally underrepresented, including Live and recurrent questions, Irrelevant, Complex, and Long anchors.

We manually evaluated labels generated for 200 queries to assess the quality. The evaluation was conducted by 1 annotator amongst the authors. We observed that the temporal understanding (anchors) has very high multi-class accuracy alignment between human labels and LLM, to be 93.0%. Similar results for reasoning (94.0%). Time sensitivity accuracy is lower (82.5%) as the automatic evaluator cannot realistically judge some tail questions or unknown topics without a complex and accurate RAG system. As additional validation to mitigate the adoption of a single annotator, we used SituatedQA labels. The dataset classifies queries in *original*, *start*, and *sampled_{year/date}*. While *original* corresponds to our evergreen, the remaining categories are time sensitive. We compared the LLM judgments to these human labels. 94.0% of *original* questions were judged as Dynamic anchor and 96.3% of remaining as Evergreen, confirming strong alignment between the taxonomy and independently annotated temporal labels. The popularity/trendiness evaluation was not considered on purpose as an LLM cannot realistically produce that information. For these reasons, we consider these numbers directional to highlight trends and high-level information of the datasets. Moreover, the relative differences represent the most important information. For instance, knowing that answer age = Week goes from 2.9% to 32.6% in FreshQA and RealtimeQA is more informative than knowing how precise is the 2.9% itself. The manual evaluation was conducted as a proof of concept to validate the LLM classifier's reliability, not as a full-scale annotation study. A deeper multi-annotator effort with inter-annotator agreement is left for future work. Some key examples sampled from these datasets are available in Table 2.

Table 2: Examples

Query - Labels
What is the name of the first animal to land on the moon? Anchor: No anchor, event locked in time Volatility: Evergreen Reasoning: Ordinal
Who won the FIFA World Cup in 2018? Anchor: Explicit, Long, Past Volatility: Evergreen Reasoning: None
How long ago did Covid-19 first break 1 billion infections? Anchor: Dynamic, Short, Past Volatility: High; Age: Day Reasoning: Arithmetic, Resolution
When was the second time that President Joe Biden visited Vietnam during his presidency? Anchor: Implicit, Long, Past Volatility: Evergreen Reasoning: Ordinal, Resolution
What is the current age of Messi? Anchor: Dynamic, Short, Present Volatility: Low, Age: Year Reasoning: Arithmetic, Recency, Expiration: Known
What is the current price of Bitcoin? Anchor: Dynamic, Short, Present Volatility: Live, Age: Live, Expiration: Known Reasoning: Recency

5.1 Diagnostic Case Study

To illustrate the taxonomy's diagnostic utility, we evaluated three models on the RealtimeQA queries as classified in Section 5: Qwen3-32B [3], Mistral-v3-14B [4], and Mistral-v3-3B [5]. Note that, although the taxonomy applies to any QA or IR system, we use LLMs as a diagnostic case study due to the availability of reproducible evaluation on multiple-choice benchmarks. RealtimeQA is a multiple-choice

[3] https://huggingface.co/Qwen/Qwen3-32B
[4] https://huggingface.co/mistralai/Ministral-3-14B-Instruct-2512
[5] https://huggingface.co/mistralai/Ministral-3-3B-Instruct-2512

Table 3: Accuracy (%) on RealtimeQA stratified by taxonomy dimensions. Only categories with $n \geq 20$ are shown.

Dimension	Category	n	Qwen 32B	Mistr. 14B	Mistr. 3B
Reasoning	Recency	758	59.0	56.6	39.6
	Resolution-dyn.	348	57.5	55.2	41.1
	Resolution-stat.	99	58.6	47.5	47.5
	Ordinal	86	53.5	54.7	43.0
	Arithmetic	39	46.2	48.7	23.1
	Comparison	21	42.9	38.1	38.1
Anchor Orientation	Future (Pred.)	42	47.6	50.0	42.9
	Future (Unk.)	22	50.0	72.7	40.9
	Ongoing	132	60.6	54.5	37.1
	Past	697	59.0	56.2	41.8
Anchor Type	Dynamic	732	60.0	57.8	40.7
	Explicit	80	51.2	47.5	43.8
	Implicit	62	48.4	46.8	43.5
	Irrelevant	23	65.2	56.5	39.1
	No Anchor	96	74.0	61.5	57.3
Answer Age	Live	50	50.0	54.0	36.0
	Day	303	57.4	58.4	36.3
	Week	325	59.1	53.2	40.9
	Month	70	61.4	61.4	44.3
	Year	20	60.0	60.0	60.0
	More	228	66.7	57.9	52.6
Answer Expiration	Known	65	53.4	51.6	38.5
	Slow	40	61.5	70.8	36.9
	Unknown	192	61.1	53.5	41.7
	None	301	60.8	58.0	44.8

task with four options per question (random baseline: 25% accuracy). We prompted each model with the question and the four options and extracted the selected choice (see Appendix A for full prompt). Overall accuracy ranges from 60.0% (Qwen3-32B) to 42.6% (Mistral-3B), but stratifying by taxonomy dimension reveals substantial variation, as shown in Table 3.

The reasoning dimension shows a clear gradient: Recency queries reach 59.0% for Qwen3-32B, while Arithmetic and Comparison drop to 46.2% and 42.9% respectively. This gradient is consistent across all three models, but amplified for smaller ones: Mistral-3B drops to 23.1% on Arithmetic, making it indistinguishable from random. Anchor orientation reveals that Future-predictable queries are substantially harder than Past or Ongoing ones across all models. High performance on Anchor orientation Ongoing matches the reasoning Recency as the two categories are naturally aligned. Queries with Explicit or Implicit anchors prove harder than Dynamic ones, likely because RealtimeQA is designed around current events, giving the models a recency prior that benefits dynamic queries. Finally, queries requiring Live answer age lag behind those tolerating older answers across all three models. This result is intuitive, answer age defines a constraint in the answer, the higher the constraint (i.e. the shorter the max age), the harder the problem.

The taxonomy also reveals qualitatively different failure profiles across models. Qwen3-32B performs equally on Past and Ongoing queries (59.0% vs 60.6%), while Mistral-3B struggles disproportionately with Ongoing ones (41.8% vs 37.1%), suggesting that present-grounding is harder for smaller models. Conversely, all three models converge at 60.0% on Year-level answer age, indicating that long-horizon factual questions are equally accessible regardless of model size. These results demonstrate that the taxonomy can pinpoint not only which temporal competencies are universally hard, but also where model-specific weaknesses emerge. None of these patterns is visible from the aggregate accuracy alone.

6 Conclusions, Limitations, and Future work

Motivated by the increasing need for comprehensive evaluation of QA/IR systems, we introduced a structured taxonomy of temporal dimensions in natural language queries. The taxonomy is designed to support exhaustive and surgical evaluations of QA/IR systems, stressing any possible temporal aspect of queries and allowing targeted, fine-grained analysis of system behavior over time. The taxonomy includes four macro dimensions, categorizing queries according to their (i) temporal understanding, (ii) reasoning, (iii) time sensitivity, and (iv) trendiness. While the first two are intrinsic properties of queries, time sensitivity and trendiness are extrinsic dimensions describing changes in world knowledge and the evolving interest. Our taxonomy is designed to be parametrizable and adaptable to different evaluation needs. We provide a formal specification of each dimension and class, ensuring clarity, reproducibility, and reduced ambiguity during annotation and system analysis. Applying the taxonomy to MS MARCO, FreshQA, RealtimeQA, and SituatedQA revealed systematic gaps in existing benchmarks, with several temporal categories severely underrepresented. A diagnostic case study on three LLMs showed that the taxonomy exposes failure patterns, such as near-random performance on temporal arithmetic and comparison, that the aggregate accuracy conceals, and that these patterns differ across models. We believe the taxonomy provides a practical foundation for more targeted evaluation and development of temporally aware QA/IR systems.

Several limitations remain open. First of all, the trendiness dimension. It requires access to large-scale query streams and dedicated infrastructure to track popularity over time; unlike the other three dimensions, it cannot be evaluated through an LLM alone, as language models have no access to how query frequency evolves after their training cutoff. For these reasons, we left it out of our experiments, but it is part of the taxonomy as a key extrinsic property of queries. Next, some subcategory boundaries, such as answer age granularity (Day, Week, Month), depend on thresholds that are configurable parameters of the evaluation setting rather than fixed values; alternative discretizations (e.g., logarithmic scales) may be more appropriate depending on the application. In some cases, their definition may be not trivial. While we designed the taxonomy to be comprehensive, it cannot be strictly exhaustive. New temporal phenomena may warrant additional categories, which may be driven by the progress in QA, IR, and LLM fields. Moreover, our manual validation relies on 200 queries labeled by a single annotator; while sufficient to confirm the LLM classifier's directional reliability, a larger multi-annotator effort would strengthen confidence in the framework's applicability. However, the goal of the annotation was to provide a direction for the analysis, highlighting that models

```
*Task*
You are given a question q. Your task is to analyze the question according to the Temporal Understanding, Time Sensitivity, and Temporal
Reasoning taxonomies defined below.
You must assign values only from the provided labels, using only the definitions given.
Do not introduce new concepts, interpretations, or categories.

*Temporal Understanding*
Classify how the question relates to time.

1. Anchor Type
  Labels: {definitions and examples for Explicit, Implicit, Dynamic, No Anchor, Irrelevant}
  Notes: {contrastive examples and disambiguation}
2. Anchor Shape
  Labels: {definitions and examples for Short, Long, Complex}
  Notes: {contrastive examples and disambiguation}
3. Anchor Orientation
  Labels: {definitions and examples for Past, Ongoing, Future-predictable, Future-unknown}
  Notes: {contrastive examples and disambiguation}

*Time Sensitivity*
Classify how the set of correct answers changes over time.

4. Answer Volatility
  Labels: {definitions and examples for Live, High, Low, Evergreen, Recurrent}
  Notes: {contrastive examples and disambiguation}
5. Answer Age
  Labels: {definitions and examples for Live, Day, Week, Month, Year, More}
  Notes: {contrastive examples and disambiguation}
6. Answer Expiration
  Labels: {definitions and examples for Known, Unknown, Slow}
  Notes: {contrastive examples and disambiguation}

*Temporal Reasoning*
Classify the type(s) of temporal reasoning required.
  Classes: {definitions and examples for Ordinal, Comparison, Arithmetic, Resolution-static, Resolution-dynamic, Recency}
  Notes: {contrastive examples and disambiguation}

Output format (JSON):
{
  "temporal_understanding": {"anchor_type": "...", "anchor_shape": "...", "anchor_orientation": "..."},
  "time_sensitivity": {"answer_volatility": "...", "is_volatility_recurrent": ..., "answer_age": "...", "answer_expiration": "..."},
  "temporal_reasoning": [...],
  "rationale": "..."
}

Question: {question}
```

Figure 4: Simplified taxonomy annotation prompt. Labels and Notes sections contain definitions, examples, and contrastive disambiguation drawn from our taxonomy.

perform differently in different sub-categories. Finally, the formal framework assumes that the temporal scope $\beta(a)$ of a response is well-defined. This holds for simple factual answers (e.g., a date or a name), but becomes ambiguous when the response is a generated summary spanning multiple time periods or a retrieved document containing facts with different temporal scopes. Related to this problem, we highlight that our evaluations on Table 3 focused on simple multi-choice questions (QA). A scalable evaluation in IR setting requires orthogonal effort in evaluating documents with respect to all temporal dimensions, which is currently an open problem and left out of the scope of this paper.

Promising directions include developing temporal-aware evaluation metrics that penalize temporally invalid answers, using volatility and expiration estimates to inform cache and refresh policies in RAG systems, deeper exploration of the trendiness dimension with access to large-scale query logs, constructing temporally balanced benchmarks guided by the taxonomy's stratification criteria, and extending the framework to multilingual and multimodal settings.

A System Prompts

The prompt used to run Qwen and Mistral on RealtimeQA is showed in the following.

```
   Answer the following multiple-choice question. Reply with ONLY
the letter (A, B, C, or D).

Question: {question}
A. {choice_0}
B. {choice_1}
C. {choice_2}
D. {choice_3}

Answer:
```

The annotation prompt instructs the model to classify each query along the taxonomy dimensions defined in this manuscript. For each subcategory, the prompt includes the full label set with definitions, examples, and contrastive notes to resolve ambiguous cases. The complete prompt is available upon request. A simplified version showing the structure is provided in Figure 4. Examples and definitions are taken from the main text.

References

[1] James F Allen. 1983. Maintaining knowledge about temporal intervals. *Commun. ACM* 26, 11 (1983), 832–843.

[2] Alessandro Artale and Enrico Franconi. 2000. A survey of temporal extensions of description logics. *Annals of Mathematics and Artificial Intelligence* 30, 1 (2000), 171–210.

[3] Angel X Chang and Christopher D Manning. 2012. Sutime: A library for recognizing and normalizing time expressions.. In *Lrec*, Vol. 12. 3735–3740.

[4] Wenhu Chen, Xinyi Wang, William Yang Wang, and William Yang Wang. 2021. A Dataset for Answering Time-Sensitive Questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1.

[5] Shiwen Cheng, Anastasios Arvanitis, and Vagelis Hristidis. 2013. How fresh do you want your search results?. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1271–1280.

[6] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A Comprehensive Evaluation of Temporal Reasoning Abilities in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 1204–1228. doi:10.18653/v1/2024.acl-long.66

[7] Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. CIDER: Commonsense Inference for Dialogue Explanation and Reasoning. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Haizhou Li, Gina-Anne Levow, Zhou Yu, Chitralekha Gupta, Berrak Sisman, Siqi Cai, David Vandyke, Nina Dethlefs, Yan Wu, and Junyi Jessy Li (Eds.). Association for Computational Linguistics, Singapore and Online, 301–313. doi:10.18653/v1/2021.sigdial-1.33

[8] Joseph Y Halpern and Yoav Shoham. 1991. A propositional modal logic of time intervals. *Journal of the ACM (JACM)* 38, 4 (1991), 935–962.

[9] Mohammed Hasanuzzaman, Sriparna Saha, Gaël Dias, and Stéphane Ferrari. 2015. Understanding temporal query intent. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 823–826.

[10] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 2391–2401. doi:10.18653/v1/D19-1243

[11] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems* 33, 2 (2021), 494–514.

[12] Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. TempQuestions: A Benchmark for Temporal Question Answering. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1057–1062. doi:10.1145/3184558.3191536

[13] Zhen Jia, Philipp Christmann, and Gerhard Weikum. 2024. Faithful temporal question answering over heterogeneous sources. In *Proceedings of the ACM Web Conference 2024*. 2052–2063.

[14] Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2021. ForecastQA: A Question Answering Challenge for Event Forecasting with Temporal Text Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4636–4650. doi:10.18653/v1/2021.acl-long.357

[15] Rosie Jones and Fernando Diaz. 2007. Temporal profiles of queries. *ACM Transactions on Information Systems (TOIS)* 25, 3 (2007), 14–es.

[16] Nattiya Kanhabua and Kjetil Nørvåg. 2012. Learning to rank search results for time-sensitive queries. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 2463–2466.

[17] Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. 2023. Realtime qa: What's the answer right now? *Advances in neural information processing systems* 36 (2023), 49025–49043.

[18] Timothée Lacroix, Guillaume Obozinski, and Nicolas Usunier. 2020. Tensor decompositions for temporal knowledge base completion. *arXiv preprint arXiv:2004.04926* (2020).

[19] Artuur Leeuwenberg and Marie-Francine Moens. 2019. A survey on temporal reasoning for temporal information extraction from text. *Journal of Artificial Intelligence Research* 66 (2019), 341–380.

[20] Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. 2012. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12)*. Association for Computing Machinery, New York, NY, USA, 251–260. doi:10.1145/2187836.2187871

[21] Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From Matching to Generation: A Survey on Generative Information Retrieval. *ACM Trans. Inf. Syst.* 43, 3, Article 83 (May 2025), 62 pages. doi:10.1145/3722552

[22] Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*. PMLR, 13604–13622.

[23] Jannat Meem, Muhammad Rashid, Yue Dong, and Vagelis Hristidis. 2024. PAT-Questions: A Self-Updating Benchmark for Present-Anchored Temporal Question-Answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 13129–13148. doi:10.18653/v1/2024.findings-acl.777

[24] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR* abs/1611.09268 (2016). arXiv:1611.09268 http://arxiv.org/abs/1611.09268

[25] Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A Reading Comprehension Dataset of Temporal Ordering Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 1158–1172. doi:10.18653/v1/2020.emnlp-main.88

[26] Mert Ozer, Anna Sapienza, Andrés Abeliuk, Goran Muric, and Emilio Ferrara. 2020. Discovering patterns of online popularity from time series. *Expert Systems with Applications* 151 (2020), 113337.

[27] Arnaldo Pereira, Alina Trifan, Rui Pedro Lopes, and José Luís Oliveira. 2022. Systematic review of question answering over knowledge bases. *IET Software* 16, 1 (2022), 1–13.

[28] James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering* 3, 28–34.

[29] Jacob Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. 2010. Characterizing and modeling the dynamics of online popularity. *Physical review letters* 105, 15 (2010), 158701.

[30] Chao Shang, Guangtao Wang, Peng Qi, and Jing Huang. 2022. Improving Time Sensitivity for Question Answering over Temporal Knowledge Graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 8017–8026. doi:10.18653/v1/2022.acl-long.552

[31] Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. COM2SENSE: A Commonsense Reasoning Benchmark with Complementary Sentences. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 883–898. doi:10.18653/v1/2021.findings-acl.78

[32] Jungbin Son and Alice Oh. 2023. Time-Aware Representation Learning for Time-Sensitive Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 70–77. doi:10.18653/v1/2023.findings-emnlp.6

[33] Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. 321–324.

[34] Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-Tail: How Knowledgeable are Large Language Models (LLMs)? A.K.A. Will LLMs Replace Knowledge Graphs?. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 311–325. doi:10.18653/v1/2024.naacl-long.18

[35] Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards Benchmarking and Improving the Temporal Reasoning Capability of Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 14820–14835. doi:10.18653/v1/2023.acl-long.828

[36] Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. 2020. Orthogonal Relation Transforms with Graph Context Modeling for Knowledge Graph Embedding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel

Tetreault (Eds.). Association for Computational Linguistics, Online, 2713–2722. doi:10.18653/v1/2020.acl-main.241

[37] Shivin Thukral, Kunal Kukreja, and Christian Kavouras. 2021. Probing Language Models for Understanding of Temporal Expressions. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 396–406. doi:10.18653/v1/2021.blackboxnlp-1.31

[38] Manisha Verma and Diego Ceccarelli. 2014. Bringing Head Closer to the Tail with Entity Linking. In *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval* (Shanghai, China) *(ESAIR '14)*. Association for Computing Machinery, New York, NY, USA, 37–39. doi:10.1145/2663712.2666196

[39] Felix Virgo, Fei Cheng, and Sadao Kurohashi. 2022. Improving Event Duration Question Answering by Leveraging Existing Temporal Information Extraction Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 4451–4457. https://aclanthology.org/2022.lrec-1.473/

[40] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 13697–13720. doi:10.18653/v1/2024.findings-acl.813

[41] Liang Wang, Ivano Lauriola, and Alessandro Moschitti. 2023. Accurate Training of Web-based Question Answering Systems with Feedback from Ranked Users. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams (Eds.). Association for Computational Linguistics, Toronto, Canada, 660–667. doi:10.18653/v1/2023.acl-industry.63

[42] Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking Temporal Reasoning for Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 6389–6415. doi:10.18653/v1/2024.findings-acl.382

[43] Shaohang Wei, Wei Li, Feifan Song, Wen Luo, Tianyi Zhuang, Haochen Tan, Zhijiang Guo, and Houfeng Wang. 2025. TIME: A Multi-level Benchmark for Temporal Reasoning of LLMs in Real-World Scenarios. *arXiv preprint arXiv:2505.12891* (2025).

[44] Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, et al. 2024. Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193* (2024).

[45] Wanqi Yang, Yanda Li, Meng Fang, and Ling Chen. 2024. Enhancing Temporal Sensitivity and Reasoning for Time-Sensitive Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 14495–14508. doi:10.18653/v1/2024.findings-emnlp.848

[46] Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM Web Conference 2024*. 1963–1974.

[47] Michael J.Q. Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating Extra-Linguistic Contexts into QA. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).

[48] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense Text Retrieval Based on Pretrained Language Models: A Survey. *ACM Trans. Inf. Syst.* 42, 4, Article 89 (Feb. 2024), 60 pages. doi:10.1145/3637870

[49] Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3363–3369. doi:10.18653/v1/D19-1332