Static Models, Dynamic World: A Unified Perspective on Temporal Perception in Large Language Models

Chenhao Li[1], Dandan Song[1]*, Changzhi Zhou[1], Jun Yang[1], Yuhang Tian[1], Huipeng Ma[1], Guangyuan Feng[1], Luan Zhang[1], Xudong Li[1], Ke Duan[1]

[1]School of Computer Science and Technology, Beijing Institute of Technology, China
{lichenhao, sdd}@bit.edu.cn

Abstract

Large language models are trained on static corpora but deployed in a dynamic world, leading to systematic temporal failures—from mis-anchored expressions and inconsistent timelines to hallucinated future events, stale world knowledge, and related issues. Existing surveys on temporal knowledge graphs, retrieval-augmented generation, hallucination, and knowledge editing cover only isolated fragments of this space: they are typically task-centric and do not offer a holistic theoretical account of how frozen LLMs represent and reason about time. This survey provides a unified perspective on temporal reasoning in LLMs. We formalize temporal queries in an information-theoretic framework based on the parametric reachability of temporal premises and answers, which induces four temporal information regimes corresponding to internal reasoning, answer recency, premise anchoring, and genuine world indeterminacy. Under this lens, we delineate the landscape of temporal failure modes, consolidate methodologies for diagnosing temporal deficiencies, and synthesize mitigation approaches into a coherent design space. Together, these contributions provide a systematic roadmap toward reliable time-aware large language models.

1 Introduction

Time is an invisible backbone of intelligence. Large language models (LLMs), however, are essentially static compressions of world knowledge into neural parameters. This mismatch is particularly stark in a fundamentally dynamic world. Temporal understanding is not a cosmetic linguistic nuance: it is central to many modelling capabilities and downstream applications. Dialogue systems must track evolving timelines across turns and speakers; retrieval-augmented generation (RAG) systems must distinguish fresh evidence from obsolete reports; autonomous agents rely on temporal logic for planning, coordination, and causal reasoning. Yet a foundational tension remains: models trained on a finite snapshot of data struggle to represent time as an abstract, structured dimension and to answer questions whose truth values evolve over time. In practice, this misalignment yields recurrent failure modes, including (i) errors in parsing and anchoring, such as contextual reference-frame drift (e.g., narrative vs. reporting time) and distractor-induced spurious anchors; (ii) failures in temporal logic and commonsense, including confused temporal ordering/interval relations and violations of typical durations or preconditions; (iii) distortions in factuality and recency, such as stale knowledge, hallucinated updates (including purported future events), and variant–invariant confusion that causes answer drift across changes in reference time, time zone, or calendar system; and (iv) deficits in robustness and uncertainty, where underspecified expressions (e.g., "recently") are collapsed into a single interpretation with unwarranted confidence rather than prompting clarification or abstention.

Despite the growing attention to temporal aspects of LLMs, the literature still lacks a survey that takes *LLM temporal competence*—how a frozen model represents, anchors, and reasons about time—as the primary object of study and provides a unified, cross-task evaluation account. Existing surveys cover adjacent pieces of this problem but remain largely task-centric. Temporal KG/TKGQA surveys assume structured, time-stamped triples and systematize temporal question types and solution paradigms within that setting (Su et al., 2024a; Cai et al., 2024). Temporal information extraction reviews focus on extraction-centric pipelines (e.g., identifying/normalizing temporal expressions and extracting temporal relations) and transformer-based modeling choices (Su et al., 2025a). Temporal IR and QA surveys

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 1913–1932

emphasize temporally-aware retrieval objectives, datasets, and evaluation protocols tailored to retrieval/QA scenarios (Piryani et al., 2025), while RAG surveys largely frame time through knowledge freshness and model updating via external evidence rather than LLM-intrinsic temporal reasoning and generalization (Gao et al., 2024; Fan et al., 2024). Knowledge editing work similarly prioritizes post-hoc updating of stale facts, leaving a fragmented picture of what temporal knowledge a frozen LLM encodes, where it fails across scenarios, and how to evaluate its temporal abilities in a principled and comprehensive manner (Wang et al., 2023b).

To bridge this gap, this survey establishes a unified perspective on temporal reasoning in LLMs. In Section 2, we first introduce an information-theoretic framework to systematize the problem space. By formalizing temporal queries based on the parametric reachability of their premises (K_{prem}) and answers (K_{ans}), we define four distinct information regimes: Internal Reasoning, Answer Recency, Premise Anchoring, and World Indeterminacy. This taxonomy serves as the theoretical axis of our survey, distinguishing whether a failure stems from a lack of knowledge, a structural inability to parse constraints, or the genuine indeterminacy of the future.

Guided by this coordinate system, we delineate the full lifecycle of temporal reliability. We begin by mapping the landscape of failure modes in Section 3, systematically categorizing the specific contexts and error patterns—such as recency bias and logical collapse—that models encounter across different regimes. Subsequently, Section 4 consolidates methodologies for identifying capability boundaries, moving beyond static accuracy to review benchmarks that diagnose robustness, calibration, and sensitivity to temporal constraints. Finally, in Section 5, we organize the landscape of existing solutions, classifying approaches ranging from internalization (pre-training) and augmentation (RAG/TKGs) to cognitive scaffolding and alignment. Ultimately, motivated by the lack of a holistic perspective in current literature, this survey aims to unify the capability taxonomy and evaluation blueprint for LLM temporal competence across tasks and settings, thereby offering the critical theoretical guidance necessary for a comprehensive resolution of temporal challenges in large language models. The overview of this survey and related datasets are presented in Appendix A and B, respectively.

2 An Information-Theoretic Framework for Temporal Reasoning

To address the tension between static parameters and dynamic world, we formalize temporal reasoning through an information-theoretic lens, as illustrated in Figure 1. We define the parametric reachability of a query's premises and answer, which partitions the problem space into four distinct regimes.

Definition 2.1 (Temporal Atomic Query) Let $\mathcal{X}$ be the space of natural language inputs and $\mathcal{C}$ the space of contexts (e.g., reference time, locale). We define the space of temporal atomic queries as $\mathcal{Q} := \mathcal{X} \times \mathcal{C}$. A query q is defined as **atomic** if it targets a single, irreducible temporal variable (e.g., a specific timestamp, duration, or boolean state) and cannot be decomposed into independent sub-queries. Given a world-time structure $\mathcal{M}$, each atomic query $q \in \mathcal{Q}$ admits a semantic interpretation denoted by the tuple:

$$(H_{\text{time}}(q),\ a^*(q)) \tag{1}$$

where $H_{\text{time}}(q)$ denotes the **temporal premises** required to interpret q (e.g., anchor points for relative expressions, calendar conventions), and $a^*(q)$ denotes the **ground-truth temporal answer**.

Definition 2.2 (Parametric Reachability) Let f_θ denote a language model with parameters θ and training cutoff t_{cut}. We introduce two binary indicator functions to characterize the model's internal knowledge state regarding a query q:

1. **Premise Reachability** (K_{prem}). We define $K_{\text{prem}}(q;\theta) = 1$ if θ encodes sufficient knowledge to reconstruct $H_{\text{time}}(q)$. Formally, this implies the existence of an inference path that correctly resolves temporal anchors and constraints necessary for the query.

2. **Answer Reachability** (K_{ans}). We define $K_{\text{ans}}(q;\theta) = 1$ if $a^*(q)$ is determined prior to t_{cut} and is sufficiently represented in the pre-training corpus to be recallable.

Based on the tuple $(K_{\text{prem}}, K_{\text{ans}})$, the query space $\mathcal{Q}$ is partitioned into four mutually exclusive information regimes.

Regime I: Internal Frontier / Temporal Reasoning ($\mathcal{R}_{\text{IF}}$). Defined by

$$K_{\text{prem}} = 1, \quad K_{\text{ans}} = 1. \tag{2}$$

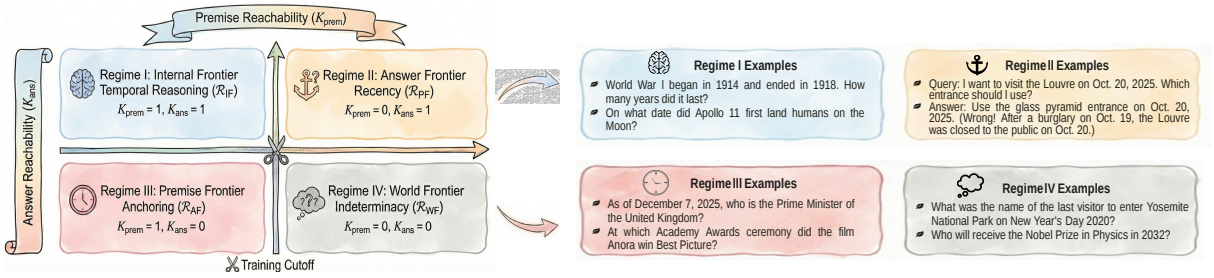

Figure 1: Classification of atomic temporal questions based on Parametric Accessibility. The left panel illustrates the four-quadrant taxonomy determined by whether the question's premise and answer are accessible within the model's pre-trained parameters. The right panel provides representative examples for each of the four categories.

The model possesses both the structural knowledge to parse constraints and the factual knowledge of the answer.

Regime II: Answer Frontier / Recency ($\mathcal{R}_{\mathrm{AF}}$). Defined by

$$K_{\mathrm{prem}} = 0, \quad K_{\mathrm{ans}} = 1. \tag{3}$$

The model can recover the temporal premises $H_{\mathrm{time}}(q)$, but the answer $a^*(q)$ post-dates the training cutoff or is distributionally absent.

Regime III: Premise Frontier / Anchoring ($\mathcal{R}_{\mathrm{PF}}$). Defined by

$$K_{\mathrm{prem}} = 1, \quad K_{\mathrm{ans}} = 0. \tag{4}$$

The answer $a^*(q)$ is theoretically known (e.g., historical facts), but the model lacks the internal representations to fully reconstruct the premises $H_{\mathrm{time}}(q)$.

Regime IV: World Frontier / Indeterminacy ($\mathcal{R}_{\mathrm{WF}}$). Defined by

$$K_{\mathrm{prem}} = 0, \quad K_{\mathrm{ans}} = 0. \tag{5}$$

Both the premises and the answer are undetermined or indeterminate at the time of querying.

For a fixed model θ, the temporal query space $\mathcal{Q}$ can be decomposed into four regimes:

$$\mathcal{Q} \approx \mathcal{R}_{\mathrm{IF}}(\theta) \dot{\cup} \mathcal{R}_{\mathrm{AF}}(\theta) \dot{\cup} \mathcal{R}_{\mathrm{PF}}(\theta) \dot{\cup} \mathcal{R}_{\mathrm{WF}}(\theta), \tag{6}$$

where $\dot{\cup}$ denotes a disjoint union in an idealized sense.

However, real-world inquiries are rarely singular atomic instances; they are typically composite problems constructed from multiple atomic queries. Crucially, parametric reachability is neither a necessary nor a sufficient condition for correctness. Accurate temporal reasoning depends heavily on a synthesis of broader capabilities—including generalization, logical reasoning, context understanding, and instruction following—rather than memory alone. Consequently, the presence of parametric knowledge serves merely as a favorable condition for reliability rather than a determinant. This understanding underscores the necessity for a comprehensive taxonomy of failure modes, which we will detail in the subsequent section.

3 Taxonomy of Temporal Failures

Building on the theoretical framework in Section 2, we map the landscape of temporal failures. These deficiencies systematically undermine the reliability of LLMs, ranging from low-level parsing errors to high-level epistemic miscalibration.

3.1 Errors in Parsing and Anchoring

Parsing and anchoring constitute the entry point of temporal understanding. Errors in this domain arise when models fail to interpret implicit time, discriminate irrelevant cues, or bind expressions to the correct reference frame.

Contextual Reference-Frame Drift. A primary failure mode involves binding temporal expressions to incorrect anchors, e.g., conflating narrative time with reporting time or missing implicit temporal cues. In parallel, temporal reference frames may drift due to relative time, time zones, calendar conventions, and cross-lingual or cultural variation, severely interfering with temporal anchoring. First, LLMs are often insufficiently sensitive to contextual time signals, showing frequent anchor drift when tracking discourse-level reference points across turns or narratives (Zhang et al., 2024b; Fan and Strube, 2025; Su et al., 2023a; Wallat et al., 2024; Qiu et al., 2024). Second, they

recurrently fail to recover implicit temporal pivots across heterogeneous inputs (including natural language and semi-structured/structured text), reflecting temporal robustness deficiencies that can trigger mis-anchoring and absolute-time inconsistencies under differing reference frames (Fan and Strube, 2025; Su et al., 2025b; Deng et al., 2025; Wallat et al., 2024; Qiu et al., 2024; Lazaridou et al., 2021; Gautam et al., 2024; Yang et al., 2023; Qian et al., 2024a; Holtermann et al., 2025). Third, anchoring errors intensify under cross-lingual and cultural drift, including cross-temporal summarization shifts, non-Gregorian calendars, culturally grounded norms, co-temporal reasoning weaknesses, time–place inconsistencies, and time-varying multilingual distributions (Zhang et al., 2024a; Sasaki et al., 2025; Alqifari et al., 2025; Su et al., 2024c; Holtermann et al., 2025; Liu et al., 2025a).

Distractor-Induced Spurious Anchors. A primary failure mode is mis-binding events to *semantically related but temporally irrelevant* cues, such that distractor timestamps (e.g., adjacent years, background timelines, or topically similar snippets) become de facto anchors—especially in multi-source, fast-updating settings where evidence is stale, conflicting, or temporally misaligned. Under dense distractors and fluctuating context quality, temporally wrong yet semantically plausible snippets can dominate attention and pull predictions toward incorrect anchors (Schumacher et al., 2025; Zeng et al., 2025). Mis-anchoring further worsens when externally introduced information has mismatched timing/recency, creating conflicts between retrieved cues and internal knowledge (Cheng et al., 2024b; Zeng et al., 2025); as contexts lengthen and become more heterogeneous, the set of temporally irrelevant but semantically similar candidates expands, increasing spurious-anchor risk (Xiong et al., 2024; Kulkarni et al., 2025).

3.2 Failures in Logic and Commonsense

Even when expressions are parsed correctly, models frequently fail in logical deduction and commonsense reasoning, leading to internally inconsistent timelines.

Temporal Relation Confusion. Models struggle to maintain consistent Allen interval relations, often confusing *before/after* and *overlap/during* under minor context shifts. This failure is pervasive in document-level extraction, where graph-based and memory-augmented approaches highlight the brittleness of locally made decisions (Mathur et al., 2021; Phu et al., 2021; Yuan et al., 2024a). Errors often stem from over-reliance on shallow cues or fragile point estimates, rather than deep interval semantics (Cheng and Weiss, 2023). Unified frameworks diagnose these as systematic transitivity violations (Huang et al., 2023; Ning et al., 2024), corroborating earlier evidence on the degradation of temporal ordering capability (Ballesteros et al., 2020; Han et al., 2019). In generative settings, these inconsistencies surface as narrative incoherence and long-horizon temporal drift (Mathur et al., 2024), and in co-temporal reasoning where models conflate co-occurrence with simultaneity (Su et al., 2024c; Ge et al., 2025).

Temporal Commonsense Violation. Models frequently generate answers that violate physical or social temporal regularities, such as implausible durations or reversed preconditions. Large-scale benchmarks reveal persistent gaps in estimating event durations and frequencies (Jain et al., 2023), partly attributable to reporting bias in text corpora (Cai et al., 2022). Although auxiliary-task and multi-task training can improve temporal commonsense performance (Kimura et al., 2022), models still struggle to internalize well-calibrated priors over typical event timescales (Wenzel and Jatowt, 2023) and to exploit question-specific temporal context in time-sensitive QA (Yang et al., 2024). In embodied decision making and planning, these temporal inconsistencies can surface as unsafe or non-executable action sequences, including incorrect step ordering (Son et al., 2025; Ishay and Lee, 2025).

3.3 Distortions in Factuality and Recency

Failures in this category reflect a misalignment between the model's static parameters and the dynamic world, manifesting as hallucinations, staleness, or confusion of variability.

Recency Distortion and Stale Knowledge. Because pretraining captures a snapshot of the world, language models can systematically produce temporally misaligned answers—relying on facts that were true during data collection but have since changed. Empirically, performance degrades as evaluation moves beyond the training period, revealing temporal generalization failures and "temporal blind spots" (Lazaridou et al., 2021; Wal-

lat et al., 2024). This mismatch is further exacerbated by implicit, topic- or source-dependent "effective cutoffs," where a model's demonstrated recency can vary substantially across sub-resources even under a single reported cutoff (Cheng et al., 2024a). This mismatch is particularly salient for volatile or time-evolving knowledge: models may confidently recite outdated statements unless they can identify which facts have likely expired or defer to up-to-date sources (Zhang and Choi, 2023; Wallat et al., 2024). Even with retrieval, robustness under fast-changing, time-sensitive corpora is not guaranteed: stress tests over dynamic sources and systematic query/document perturbations show RAG systems remain vulnerable to noisy, conflicting, or shifting evidence, limiting their ability to reliably stay current (Zeng et al., 2025). However, continually repairing such staleness is non-trivial—continual knowledge updates via naive fine-tuning risk catastrophic forgetting of time-invariant knowledge (Jang et al., 2022).

Variant–Invariant Confusion. Models often conflate time-varying facts (e.g., "current president") with time-invariant ones (e.g., "birth date"), and time-aware probing suggests they may either overfit to a specific period or smear temporally distinct variants when trained over narrow slices or broad ranges (Dhingra et al., 2022). This confusion manifests as temporal QA/forecasting errors where dynamic signals are blurred across time (e.g., over-reliance on recency/recurrency patterns or mixing timestamped evidence in temporal KG forecasting/embeddings), and mechanistic analyses similarly indicate that even temporal-specialized attention heads do not reliably separate temporal regimes (Gastinger et al., 2024; Han et al., 2024; Park et al., 2025).

3.4 Deficits in Robustness and Uncertainty

Finally, we address the inability of models to handle ambiguity or signal ignorance, which renders answers brittle and untrustworthy. When confronted with underspecified expressions (e.g., "recently"), models often collapse ambiguity to a single implicit interpretation and present it with unwarranted confidence rather than eliciting missing context. First, evidence from ambiguous open-domain QA shows that underspecification frequently reflects temporal dependence and other latent assumptions, yet systems still fail to represent multiple plausible readings (Min et al., 2020; Papakostas and Papadopoulou, 2023). Second, even when interaction could resolve ambiguity, models often do not initiate clarification and instead proceed with speculative completion (Zhang and Choi, 2025). Third, pipeline analyses suggest that robustness hinges on reliably detecting underspecification; when detection fails, downstream fusion or checking can still yield confident but arbitrary outputs (Gao et al., 2021). Additionally, these failures persist in time-sensitive contexts where abstention or verification would be appropriate, as illustrated by selective prediction in realistic settings and self-consistency based hallucination detection over dated factual claims (Xin et al., 2021; Manakul et al., 2023).

4 Identifying Temporal Deficiencies

Having mapped the landscape of temporal failures in Section 3, we now turn to the methodologies for systematically identifying these deficiencies. Existing evaluation protocols can be organized into four families based on their intervention strategies: (i) anchor-controlled contrastive testing, (ii) constraint-based reasoning evaluation, (iii) version-sensitive knowledge assessment, and (iv) perturbation- and uncertainty-oriented evaluation. These families transition the field from anecdotal observation to reproducible measurement.

4.1 Anchor-Controlled Contrastive Testing

Targeting parsing and anchoring errors, this methodology employs *contrastive evaluation*—varying temporal reference frames to test if models correctly resolve relative expressions versus relying on shallow heuristics. More general reading-comprehension benchmarks motivate controlled context shifts, which temporal benchmarks then specialize into explicit re-anchoring tests (Hermann et al., 2015; Trischler et al., 2017). Concretely, TORQUE evaluates whether models preserve event order under narrative and question re-anchoring, while implicit-event supervision stresses anchoring when temporal signals are missing or indirect (Ning et al., 2020; Zhou et al., 2021a). Dialogue further amplifies reference drift: TIMEDIAL probes temporal commonsense grounded in conversational context (Qin et al., 2021). Recent benchmarks disentangle explicit/implicit/vague mentions (TRAVELER) and cross-cultural calendar conversion (SPAN), exposing over-reliance on surface cues (Kenneweg et al.,

2025; Miao et al., 2026). Finally, unified suites (TIMEBENCH, TRAM) consolidate tasks to localize anchoring bottlenecks, and temporal-sensitive RAG QA benchmarks extend anchor control to retrieval-time alignment (Chu et al., 2024; Wang and Zhao, 2024; Chen et al., 2025b).

4.2 Constraint-Based Temporal Reasoning Evaluation

This methodology identifies logical and commonsense failures by framing temporal reasoning as a global *constraint satisfaction problem*. The evaluation asks whether outputs satisfy global temporal structure, treating violations of explicit relations (e.g., interval constraints) or implicit priors (e.g., typical duration/frequency) as failures. Temporal commonsense resources show that events evoke stereotyped durations and scripts, while discourse-level datasets require globally coherent ordering beyond sentence-local cues (Zhou et al., 2019; Naik et al., 2019). Reading and dialogue benchmarks operationalize constraint satisfaction by requiring consistent event graphs or conversational timelines, including TORQUE, TIMEDIAL, and implicit-event reasoning from distant supervision (Ning et al., 2020; Qin et al., 2021; Zhou et al., 2021a). Diagnostic and synthetic benchmarks further isolate propagation and temporal arithmetic, e.g., TEST OF TIME and broader LLM temporal-reasoning evaluations (Fatemi et al., 2025; Tan et al., 2023); TIMEBENCH and TRAM unify these into hierarchical testbeds (Chu et al., 2024; Wang and Zhao, 2024). In QA, implicit-time-constraint and complex multi-hop settings enforce global consistency across latent time conditions (Jia et al., 2024; Gruber et al., 2025). In structured knowledge settings, temporal-KGQA datasets impose strict structural and multi-granularity constraints over time-varying relations (Saxena et al., 2021; Chen et al., 2023b; Ong et al., 2023).

4.3 Version-Sensitive Knowledge Assessment

This line of work targets factual staleness and hallucinations by treating world knowledge as *time-indexed facts* and evaluating whether a model can return the correct *version* of a fact under an explicit temporal (or situational) reference, rather than merely retrieving a plausible answer. Benchmarks such as time-sensitive QA and SITUATEDQA make version selection explicit via timestamps and extra-linguistic context (Chen et al., 2021; Zhang and Choi, 2021). To track adaptation as "now" advances, self-updating benchmarks refresh gold answers from external sources, enabling longitudinal evaluation without manual relabeling (Meem et al., 2024; Lin et al., 2025). Complementary analyses benchmark temporal recall boundaries and diagnose misalignment between query time and stored facts (Herel et al., 2024). Verification-oriented protocols such as DYKNOW dynamically check outputs against up-to-date KGs to separate outdated recall from confabulation, while EVOLVEBENCH integrates cognition, awareness, and trustworthiness under evolving knowledge and invalid timestamps (Mousavi et al., 2024; Zhu et al., 2025b). Robustness studies further show that even simple temporal context shifts can trigger factual anomalies, clarifying the staleness–hallucination boundary (Khodja et al., 2025). Finally, historical news QA datasets extend version-sensitive evaluation to archived corpora with controlled temporal access (Wang et al., 2022; Piryani et al., 2024).

4.4 Perturbation and Uncertainty Oriented Evaluation

The final family probes *robustness* and *epistemic uncertainty*, evaluating behavior under ambiguity, perturbation, or conflict. Robustness diagnostics measure stability under benign temporal rewrites and reference shifts, revealing systematic anomaly patterns (Khodja et al., 2025). In RAG, temporal-sensitive QA benchmarks test whether small temporal perturbations induce retrieval drift toward temporally confusable evidence (Chen et al., 2025b). For vagueness, TRAVELER quantifies over-commitment when references are underspecified, turning "over-interpretation" into a measurable failure mode (Kenneweg et al., 2025). Consistency-based verification such as DYKNOW uses disagreement across prompts as a proxy for uncertainty and a detector for confident-but-outdated answers (Mousavi et al., 2024). Finally, refusal- and safety-oriented evaluation injects conflicting/invalid timestamps and scores whether models abstain rather than confabulate, as emphasized in EVOLVEBENCH (Zhu et al., 2025b).

5 Methodologies for Time-Aware LLMs

Building upon the failure analysis in Section 3, we review current strategies for equipping LLMs with time-awareness. As visualized in Figure 2, we categorize existing approaches into four primary

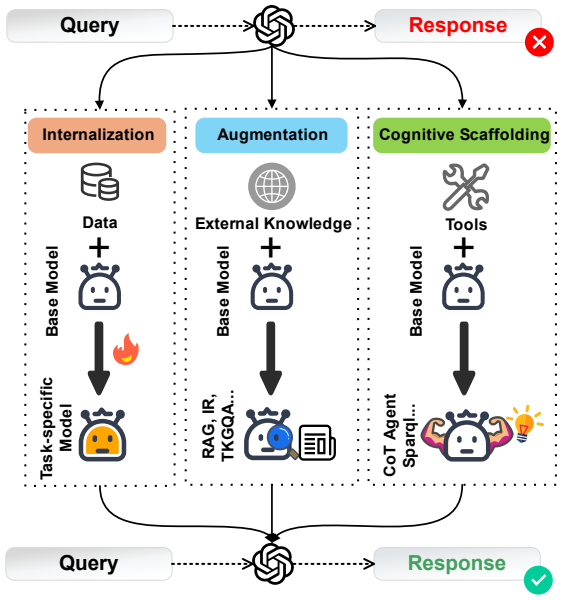

Figure 2: Taxonomy of mitigation techniques for equipping LLMs with temporal capabilities. **Note:** Reinforcement Learning is omitted here as it serves as an orthogonal optimization paradigm applicable across the three strategies.

families based on their underlying mechanisms for temporal integration.

5.1 Internalization: Tuning for Time-Awareness

Internalization makes models time-aware by absorbing knowledge via parameter updates (typically LoRA-based SFT), functioning independently of external retrieval or symbolic modules.

Representative evidence can be summarized compactly across a few recurring training patterns. Time-conditioned continued pretraining and temporal adaptation inject timestamp or temporal signals into contextual representations and help models track time-varying semantics and facts (Dhingra et al., 2022; Rosin and Radinsky, 2022; Wang et al., 2023a; Röttger and Pierrehumbert, 2021; Tang et al., 2023; Su et al., 2023b). For temporal QA and temporal graph reasoning, instruction tuning on multi-hop datasets and fine-tuning with temporal-graph intermediates improve ordering and event-time inference (Tan et al., 2024; Su et al., 2023a; Xiong et al., 2024). For temporal relation extraction, structure-guided fine-tuning and global constraint regularization improve consistency at sentence and document levels (Zhou et al., 2022, 2021b). In domain-specialized and broad-coverage settings, time-aware instruction tuning over longitudinal EHRs and temporal-specialist post-training further strengthen time sensitivity (Cui et al., 2025; Su et al., 2024d).

5.2 Augmentation: RAG and Structured Memory

Augmentation addresses the static nature of LLM parameters by integrating external, updatable information sources, thus solving the issues of stale knowledge and hallucination.

Temporal Retrieval-Augmented Generation (RAG). Temporal RAG enhances standard retrieval by explicitly modeling timestamps and dynamic information needs. To combat knowledge obsolescence, (Vu et al., 2024; Wang et al., 2025) augment generation with search-engine evidence, where iterative query refinement and temporal evidence fusion improve complex temporal reasoning. Beyond always-on retrieval, (Su et al., 2024b; Zhang et al., 2025, 2024c) study adaptive and modular RAG pipelines that decide *when* and *what* to retrieve under time-sensitive information needs. For temporal coherence over multi-document narratives, (Yang et al., 2025) structures retrieved evidence into event-centric graphs to preserve temporal dependencies during generation. Finally, (Wu et al., 2024a; Chen et al., 2025b) analyze time-sensitive RAG failures and provide benchmark evaluations, while (Wu et al., 2024b) shows a closely related alternative that enforces time constraints through external memory selection and updates without changing model weights.

External Temporal Knowledge Graphs (TKGs). Structured memory provides a rigorous alternative to unstructured text. Early embedding-based TKGQA methods (e.g., TempoQR) ground questions to TKG entities and a question-specific temporal scope for complex temporal reasoning (Mavromatis et al., 2022). Recent LLM-based approaches largely improve evidence selection over TKGs via time-aware RAG pipelines and graph-based summarization, e.g., retrieve–(rewrite/plan)–rerank and temporal GraphRAG-style retrieval (Qian et al., 2024a, 2025; Zhu et al., 2025c). In parallel, LLMs support more structured reasoning interfaces for TKGQA—including abstract reasoning induction, time-aware ReAct agents, self-improving programs, and dynamic rule generation/adaptation for evolving TKGs (Chen et al., 2024b; Hu et al., 2025; Chen et al., 2024a; Wang et al., 2024;

Chen et al., 2025a). Beyond QA, prompt-based forecasting and chain-of-history reasoning enable parameter-free temporal prediction (Lee et al., 2023; Xia et al., 2024), while zero-shot relational learning extends TKG reasoning to unseen relations with LLMs (Ding et al., 2024).

5.3 Cognitive Scaffolding: Structuring Inference

Scaffolding methods enhance reliability without modifying parameters by imposing explicit temporal structures (e.g., timelines, narratives, or explanations) during inference.

Explicit Temporal Structures and Narratives. A common strategy for improving (and auditing) temporal reasoning is to force LLMs to externalize intermediate artifacts, such as instruction–explanation trajectories (Yuan et al., 2024b), recounted narratives prior to extracting temporal relations (Zhang et al., 2024b), or self-generated exemplars as an explicit scaffold (Yasunaga et al., 2024). Related work imposes stronger temporal structure to enable self-checking: models may construct timelines for self-reflection and correction (Bazaga et al., 2025), explicitly generate anchors and intervals to explain answers (Jiang et al., 2025), or use Allen-style interval prompts with reflection-based consistency checks for discourse-level relation extraction (Fan and Strube, 2025). Complementarily, counterfactual-consistency prompting strengthens relative temporal understanding by generating counterfactual queries and enforcing consistency constraints across them (Kim and Hwang, 2025).

Tool Augmentation and Programmatic Reasoning. To overcome the brittleness of LLMs in temporal arithmetic (e.g., calendar conversion), models serve as agents dispatching tasks to external tools. (Chen et al., 2023a) propose Program of Thoughts prompting, expressing reasoning as executable programs and delegating computation to an interpreter. (Miao et al., 2026) develop a Time Agent capable of invoking code tools for high-precision conversion between Gregorian, Lunar, and Islamic calendars (SPAN). (Schick et al., 2023) demonstrate a general paradigm where models learn to call APIs, such as calendar tools, to offload date calculations, effectively resolving common arithmetic failures.

5.4 Reinforcement Learning Enhanced Methods

Distinct from standard supervised fine-tuning, Reinforcement Learning (RL) is increasingly used to "sharpen" temporal reasoning by optimizing models against specific temporal rewards or AI feedback.

Self-Supervised and AI Feedback. RL enables models to generalize beyond labeled data through curriculum learning. (Liu et al., 2025b) implement Time-R1, which employs a comprehensive three-stage RL curriculum spanning basic understanding to creative future scenario generation, showing that smaller models can achieve strong temporal performance via carefully designed reinforcement schedules. Relatedly, (Tan et al., 2023) improve LLM temporal reasoning using a time-sensitive RL post-training strategy on a dedicated benchmark.

Task-Specific RL Frameworks. For specific challenges, RL optimizes decision boundaries. (Yang et al., 2024) introduce granular contrastive reinforcement learning for time-sensitive QA, using temporal information-aware embeddings to penalize models for selecting temporally proximate but incorrect answers. (Du et al., 2025) propose Memory-T1, applying RL to temporal reasoning in multi-session agents, training the model to selectively retain and utilize temporal context across long interaction histories. Beyond QA and dialogue, (Han et al., 2022) use RL to encourage temporally coherent flashbacks in story generation, and (Wu et al., 2025) apply REINFORCE-style rewards to train a T5-based generator for event temporal relation extraction. In temporal knowledge graphs, (Sun et al., 2021; Qian et al., 2024b; Zhu et al., 2025a) formulate forecasting/reasoning/QA as RL-guided path exploration with time-shaped or adaptive rewards.

6 Conclusion

This survey systematizes the landscape of temporal reasoning in LLMs through an information-theoretic lens, mapping the fundamental tension between static parameterization and dynamic information needs. By formalizing temporal queries based on the parametric reachability of premises and answers, we have provided a structured diagnostic for failure modes—from anchoring artifacts to epistemic hallucinations—and consolidated a

roadmap of methodologies ranging from internalization to dynamic augmentation. As the field progresses, the paradigm is shifting from treating time as a mere metadata attribute to modeling it as a cognitive scaffold. We posit that the next generation of reliable LLMs will move beyond static snapshots, evolving into continuously adaptive agents capable of navigating the "World Frontier" with calibrated uncertainty and robust temporal logic.

Limitations

While this survey offers a comprehensive synthesis, we acknowledge three scope boundaries:

Modality Restrictions. Our analysis focuses exclusively on text-based and symbolic temporal reasoning. Although temporal dynamics are intrinsic to multimodal domains (e.g., video or audio processing), we restrict our scope to textual modalities to prioritize the linguistic and logical challenges inherent to LLMs, thereby excluding sensory data streams.

Taxonomic Fluidity. The categorization in Section 5 serves as an organizational heuristic rather than a rigid dichotomy. As hybrid systems increasingly blend strategies (e.g., combining augmentation with internalization), we classify contributions based on their primary innovation mechanism to maintain structural clarity.

Temporal Velocity. Given the rapid evolution of LLMs, this survey represents a snapshot of the current research landscape. While our proposed framework (Section 2) is designed to be model-agnostic, specific performance metrics and SOTA distinctions are subject to the inherent latency and fast-paced nature of the domain.

Ethics Statement

In this survey, we analyze existing methodologies and utilize publicly available datasets/resources as listed in the Appendices. We do not collect any personally identifiable information or conduct new human subject studies. All datasets and models discussed are cited and utilized in accordance with their respective licenses.

The primary objective of this survey is to facilitate the mitigation of critical temporal errors in LLMs—specifically parsing failures, logical violations, and knowledge obsolescence; we condemn any potential misuse of these technologies.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2024YFE0210800) and the National Natural Science Foundation of China (Grant No. 62476025). This work was also supported by the Fundamental Research Funds for the Central Universities.

References

Reem Alqifari, Hend Al-Khalifa, and Simon O'Keefe. 2025. Arabic temporal common sense understanding. *Computation*, 13(1):5.

Miguel Ballesteros, Rishita Anubhai, Shuai Wang, Nima Pourdamghani, Yogarshi Vyas, Jie Ma, Parminder Bhatia, Kathleen McKeown, and Yaser Al-Onaizan. 2020. Severing the edge between before and after: Neural architectures for temporal ordering of events. pages 5412–5417.

Adrián Bazaga, Rexhina Blloshmi, Bill Byrne, and Adrià de Gispert. 2025. Learning to reason over time: Timeline self-reflection for improved temporal reasoning in language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28014–28033, Vienna, Austria. Association for Computational Linguistics.

Bibo Cai, Xiao Ding, Bowen Chen, Li Du, and Ting Liu. 2022. Mitigating reporting bias in semi-supervised temporal commonsense inference with probabilistic soft logic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10454–10462.

Li Cai, Xin Mao, Yuhao Zhou, Zhaoguang Long, Changxu Wu, and Man Lan. 2024. A survey on temporal knowledge graph: Representation learning and applications. *arXiv preprint arXiv:2403.04782*.

Kai Chen, Xin Song, Ye Wang, Liqun Gao, Aiping Li, Xiaojuan Zhao, Bin Zhou, and Yalong Xie. 2025a. LLM-DR: A novel llm-aided diffusion model for rule generation on temporal knowledge graphs. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 11481–11489. AAAI Press.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023a. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Wenhu Chen, Xinyi 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*.

Zhuo Chen, Zhao Zhang, Zixuan Li, Fei Wang, Yutao Zeng, Xiaolong Jin, and Yongjun Xu. 2024a. Self-improvement programming for temporal knowledge graph question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14579–14594, Torino, Italia. ELRA and ICCL.

Ziyang Chen, Dongfang Li, Xiang Zhao, Baotian Hu, and Min Zhang. 2024b. Temporal knowledge question answering via abstract reasoning induction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4872–4889, Bangkok, Thailand. Association for Computational Linguistics.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. 2023b. Multi-granularity temporal question answering over knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11378–11392, Toronto, Canada. Association for Computational Linguistics.

Ziyang Chen, Erxue Min, Xiang Zhao, Yunxin Li, Xin Jia, Jinzhi Liao, Jichao Li, Shuaiqiang Wang, Baotian Hu, and Dawei Yin. 2025b. A Question Answering Dataset for Temporal-Sensitive Retrieval-Augmented Generation. *Scientific Data*, 12(1):1855.

Cheng Cheng and Jeremy C. Weiss. 2023. Typed markers and context for clinical temporal relation extraction. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pages 94–109, New York, USA. PMLR.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024a. Dated data: Tracing knowledge cutoffs in large language models. In *First Conference on Language Modeling*.

Qinyuan Cheng, Xiaonan Li, Shimin Li, Qin Zhu, Zhangyue Yin, Yunfan Shao, Linyang Li, Tianxiang Sun, Hang Yan, and Xipeng Qiu. 2024b. Unified active retrieval for retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17153–17166, Miami, Florida, USA. Association for Computational Linguistics.

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)*, pages 1204–1228, Bangkok, Thailand. Association for Computational Linguistics.

Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H. Shah. 2025. TIMER: temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digital Medicine*, 8:577. ArXiv:2503.04176.

Irwin Deng, Kushagra Dixit, Dan Roth, and Vivek Gupta. 2025. Enhancing temporal understanding in LLMs for semi-structured tables. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4951–4970, Albuquerque, New Mexico. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Zifeng Ding, Heling Cai, Jingpei Wu, Yunpu Ma, Ruotong Liao, Bo Xiong, and Volker Tresp. 2024. zrLLM: Zero-shot relational learning on temporal knowledge graphs with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1877–1895, Mexico City, Mexico. Association for Computational Linguistics.

Yiming Du, Baojun Wang, Yifan Xiang, Zhaowei Wang, Wenyu Huang, Boyang Xue, Bin Liang, Xingshan Zeng, Fei Mi, Haoli Bai, Lifeng Shang, Jeff Z. Pan, Yuxin Jiang, and Kam-Fai Wong. 2025. Memory-T1: Reinforcement learning for temporal reasoning in multi-session agents. *Preprint*, arXiv:2512.20092.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501, Barcelona, Spain. Association for Computing Machinery.

Yi Fan and Michael Strube. 2025. Consistent discourse-level temporal relation extraction using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18605–18622, Suzhou, China. Association for Computational Linguistics.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2025. Test of time: A benchmark for evaluating LLMs on temporal reasoning. In *The Thirteenth International Conference on Learning Representations*.

Yifan Gao, Henghui Zhu, Patrick Ng, Cicero dos Santos, Zhiguo Wang, Feng Nan, Dejiao Zhang, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Answering ambiguous questions through

generative evidence fusion and round-trip prediction. 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)*, pages 3263–3276.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Julia Gastinger, Christian Meilicke, Federico Errica, Timo Sztyler, Anett Schülke, and Heiner Stuckenschmidt. 2024. History repeats itself: A baseline for temporal knowledge graph forecasting. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 4016–4024. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Akash Gautam, Lukas Lange, and Jannik Strötgen. 2024. Discourse-aware in-context learning for temporal expression normalization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 306–315, Mexico City, Mexico. Association for Computational Linguistics.

Yubin Ge, Salvatore Romeo, Jason Cai, Raphael Shu, Yassine Benajiba, Monica Sunkara, and Yi Zhang. 2025. TreMu: Towards neuro-symbolic temporal reasoning for LLM-agents with memory in multi-session dialogues. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18974–18988.

Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. 2025. ComplexTempQA: A 100m dataset for complex temporal question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9100–9112, Suzhou, China. Association for Computational Linguistics.

Rujun Han, Hong Chen, Yufei Tian, and Nanyun Peng. 2022. Go back in time: Generating flashbacks in stories with event temporal prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1450–1470, Seattle, United States. Association for Computational Linguistics.

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019. Deep structured neural network for event temporal relation extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Hong Kong, China. Association for Computational Linguistics.

Yadan Han, Guangquan Lu, Shichao Zhang, Liang Zhang, Cuifang Zou, and Guoqiu Wen. 2024. A temporal knowledge graph embedding model based on variable translation. *Tsinghua Science and Technology*, 29(5):1554–1565.

David Herel, Vojtech Bartek, Jiri Jirak, and Tomas Mikolov. 2024. Time awareness in large language models: Benchmarking fact recall across time. *Preprint*, arXiv:2409.13338.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, pages 1693–1701.

Carolin Holtermann, Paul Röttger, and Anne Lauscher. 2025. Around the world in 24 hours: Probing LLM knowledge of time and place. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22875–22897, Vienna, Austria. Association for Computational Linguistics.

Qianyi Hu, Xinhui Tu, Cong Guo, and Shunping Zhang. 2025. Time-aware ReAct agent for temporal knowledge graph question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6028–6039, Albuquerque, New Mexico. Association for Computational Linguistics.

Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, and Dongyan Zhao. 2023. More than classification: A unified framework for event temporal relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9631–9646, Toronto, Canada. Association for Computational Linguistics.

Adam Ishay and Joohyung Lee. 2025. LLM+ AL: Bridging large language models and action languages for complex reasoning about actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24212–24220.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6750–6774.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. 2022. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*. Poster.

Zhen Jia, Philipp Christmann, and Gerhard Weikum. 2024. Tiq: A benchmark for temporal question answering with implicit time constraints. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion)*, pages 1394–1399, Singapore, Singapore. Association for Computing Machinery.

Zihao Jiang, Ben Liu, Miao Peng, Wenjie Xu, Yao Xiao, Zhenyan Shan, and Min Peng. 2025. Towards explainable temporal reasoning in large language models: A structure-aware generative framework. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7232–7251, Vienna, Austria. Association for Computational Linguistics.

Svenja Kenneweg, Jörg Deigmöller, Philipp Cimiano, and Julian Eggert. 2025. TRAVELER: A benchmark for evaluating temporal reasoning across vague, implicit and explicit references. *arXiv preprint arXiv:2505.01325*.

Hichem Ammar Khodja, Frédéric Béchet, Quentin Brabant, Alexis Nasr, and Gwénolé Lecorvé. 2025. Factual knowledge in language models: Robustness and anomalies under simple temporal context variations. In *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, pages 1–22, Vienna, Austria. Association for Computational Linguistics.

Jongho Kim and Seung-won Hwang. 2025. Counterfactual-consistency prompting for relative temporal understanding in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1210–1225, Vienna, Austria. Association for Computational Linguistics.

Mayuko Kimura, Lis Kanashiro Pereira, and Ichiro Kobayashi. 2022. Toward building a language model for understanding temporal commonsense. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 17–24.

Atharv Kulkarni, Kushagra Dixit, Vivek Srikumar, Dan Roth, and Vivek Gupta. 2025. LLM-symbolic integration for robust temporal tabular reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19914–19940, Vienna, Austria. Association for Computational Linguistics.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems 34*, pages 29348–29363.

Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter, and Jay Pujara. 2023. Temporal knowledge graph forecasting without knowledge using in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 544–557, Singapore. Association for Computational Linguistics.

Library of Congress. n.d. Chronicling america historic american newspapers. Library of Congress Digital Collections. About this Collection.

Qian Lin, Junyi Li, and Hwee Tou Ng. 2025. DynaQuest: A dynamic question answering dataset reflecting real-world knowledge updates. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26918–26936, Vienna, Austria. Association for Computational Linguistics.

Weisi Liu, Guangzeng Han, and Xiaolei Huang. 2025a. Examining and adapting time for multilingual classification via mixture of temporal experts. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6151–6166, Albuquerque, New Mexico. Association for Computational Linguistics.

Zijia Liu, Peixuan Han, Haofei Yu, Haoru Li, and Jiaxuan You. 2025b. Time-r1: Towards comprehensive temporal reasoning in LLMs. *Preprint*, arXiv:2505.13508.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha. 2021. TIMERS: document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 524–533.

Puneet Mathur, Vlad I. Morariu, Aparna Garimella, Franck Dernoncourt, Jiuxiang Gu, Ramit Sawhney, Preslav Nakov, Dinesh Manocha, and Rajiv Jain. 2024. DocScript: Document-level script event prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5140–5155.

Costas Mavromatis, Prasanna Lakkur Subramanyam, Vassilis N. Ioannidis, Adesoji Adeshina, Phillip R. Howard, Tetiana Grinberg, Nagib Hakim, and George Karypis. 2022. TempoQR: Temporal question reasoning over knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5825–5833.

Pawel Mazur and Robert Dale. 2010. WikiWars: A new corpus for research on temporal expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 913–922, Cambridge, MA. Association for Computational Linguistics.

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*, pages 13129–13148, Bangkok, Thailand. Association for Computational Linguistics.

Zhongjian Miao, Hao Fu, and Chen Wei. 2026. SPAN: Benchmarking and improving cross-calendar temporal reasoning of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 35715–35723, Singapore. AAAI Press.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Seyed Mahed Mousavi, Simone Alghisi, and Giuseppe Riccardi. 2024. Dyknow: Dynamically verifying time-sensitive factual knowledge in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8014–8029, Miami, Florida, USA. Association for Computational Linguistics.

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. TDDiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249, Stockholm, Sweden. Association for Computational Linguistics.

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)*, pages 1158–1172, Online. Association for Computational Linguistics.

Wanting Ning, Lishuang Li, Xueyang Qin, Yubo Feng, and Jingyao Tang. 2024. Temporal cognitive tree: A hierarchical modeling approach for event temporal relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 855–864.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Ryan Ong, Jiahao Sun, Ovidiu erban, and Yi-Ke Guo. 2023. TKGQA dataset: Using question answering to guide and validate the evolution of temporal knowledge graph. *Data*, 8(3):61.

Konstantinos Papakostas and Irene Papadopoulou. 2023. Model analysis & evaluation for ambiguous question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4570–4580, Toronto, Canada. Association for Computational Linguistics.

Yein Park, Chanwoong Yoon, Jungwoo Park, Minbyul Jeong, and Jaewoo Kang. 2025. Does time have its place? temporal heads: Where language models recall time-specific information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16616–16643, Vienna, Austria. Association for Computational Linguistics.

Minh Tran Phu, Minh Van Nguyen, and Thien Huu Nguyen. 2021. Fine-grained temporal relation extraction with ordered-neuron lstm and graph convolutional networks. In *Proceedings of the seventh workshop on noisy user-generated text (w-nut 2021)*, pages 35–45.

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

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. Chroniclingamericaqa: A large-scale question answering dataset based on historical american newspaper pages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14–18, 2024*, pages 2038–2048. ACM.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and 1 others. 2003. The TimeBank corpus. In *Corpus Linguistics*, page 40, Lancaster, UK.

Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, and Xiaojie Yuan. 2025. Plan of knowledge: Retrieval-augmented large language models for temporal knowledge graph question answering. *Preprint*, arXiv:2511.04072.

Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, and Kehui Song. 2024a. TimeR^4 : Time-aware retrieval-augmented large language models for temporal knowledge graph question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6942–6952, Miami, Florida, USA. Association for Computational Linguistics.

Ye Qian, Fuhui Sun, Xiaoyan Wang, and Li Pan. 2024b. Todear: Promoting explainable tkg reasoning through temporal offset enhanced dynamic em-

bedding and adaptive reinforcement learning. *Information Sciences*, 679:121066.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. 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)*, pages 7066–7076, Online. Association for Computational Linguistics.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay Cohen. 2024. Are large language model temporally grounded? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7064–7083, Mexico City, Mexico. Association for Computational Linguistics.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus. Linguistic Data Consortium, Philadelphia, 6(12):e26752.

Mutsumi Sasaki, Go Kamoda, Ryosuke Takahashi, Kosuke Sato, Kentaro Inui, Keisuke Sakaguchi, and Benjamin Heinzerling. 2025. Can language models handle a non-gregorian calendar? the case of the japanese wareki.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. 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)*, pages 6663–6676, Online. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Dan Schumacher, Fatemeh Haji, Tara Grey, Niharika Bandlamudi, Nupoor Karnik, Gagana Uday Kumar, Cho-Yu Jason Chiang, Peyman Najafirad, Nishant Vishwamitra, and Anthony Rios. 2025. RASTeR: Robust, agentic, and structured temporal reasoning. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3098–3123, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Emily Silcock, Abhishek Arora, Luca D'Amico-Wong, and Melissa Dell. 2024. Newswire: A large-scale structured database of a century of historical news. In *Advances in Neural Information Processing Systems*, volume 37, pages 49768–49779.

Yejin Son, Minseo Kim, Sungwoong Kim, Seungju Han, Jian Kim, Dongju Jang, Youngjae Yu, and Chan Young Park. 2025. Subtle risks, critical failures: A framework for diagnosing physical safety of LLMs for embodied decision making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25692–25733, Suzhou, China. Association for Computational Linguistics.

Miao Su, Zixuan Li, Zhuo Chen, Long Bai, Xiaolong Jin, and Jiafeng Guo. 2024a. Temporal knowledge graph question answering: A survey. *arXiv preprint arXiv:2406.14191*.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024b. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12991–13013, Bangkok, Thailand. Association for Computational Linguistics.

Xin Su, Phillip Howard, and Steven Bethard. 2025a. Transformer-based temporal information extraction and application: A review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28810–28829, Suzhou, China. Association for Computational Linguistics.

Xin Su, Phillip Howard, Nagib Hakim, and Steven Bethard. 2023a. Fusing temporal graphs into transformers for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 948–966, Singapore. Association for Computational Linguistics.

Xin Su, Sungduk Yu, Phillip Howard, and Steven Bethard. 2025b. A semantic parsing framework for end-to-end time normalization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, and Min

Zhang. 2024c. Living in the moment: Can large language models grasp co-temporal reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13014–13033, Bangkok, Thailand. Association for Computational Linguistics.

Zhaochen Su, Juntao Li, Zikang Zhang, Zihan Zhou, and Min Zhang. 2023b. Efficient continue training of temporal language model with structural information. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6315–6329, Singapore. Association for Computational Linguistics.

Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao Li, Min Zhang, and Yu Cheng. 2024d. Timo: Towards better temporal reasoning for language models. *Preprint*, arXiv:2406.14192. Accepted to COLM 2024.

Haohai Sun, Jialun Zhong, Yunpu Ma, Zhen Han, and Kun He. 2021. TimeTraveler: Reinforcement learning for temporal knowledge graph forecasting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8306–8319, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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)*, pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2024. Towards robust temporal reasoning of large language models via a multi-hop QA dataset and pseudo-instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6272–6286, Bangkok, Thailand. Association for Computational Linguistics.

Xiaohang Tang, Yi Zhou, and Danushka Bollegala. 2023. Learning dynamic contextualised word embeddings via template-based temporal adaptation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9352–9369, Toronto, Canada. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

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*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

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*, pages 683–692.

Jiapu Wang, Kai Sun, Linhao Luo, Wei Wei, Yongli Hu, Alan Wee-Chung Liew, Shirui Pan, and Baocai Yin. 2024. Large language models-guided dynamic adaptation for temporal knowledge graph reasoning. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2022. Archivalqa: A large-scale benchmark dataset for open-domain question answering over historical news collections. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11–15, 2022*, pages 3025–3035. ACM.

Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa, and Yi Cai. 2023a. BiTimeBERT: Extending pre-trained language representations with bi-temporal information. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*, pages 812–821, Taipei, Taiwan. ACM.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2023b. Knowledge editing for large language models: A survey. *arXiv preprint arXiv:2310.16218*.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415, Bangkok, Thailand. Association for Computational Linguistics.

Zhao Wang, Ziliang Zhao, and Zhicheng Dou. 2025. TimeRAG: Enhancing complex temporal reasoning with search engine augmentation. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM 2025)*, pages 3230–3239.

Georg Wenzel and Adam Jatowt. 2023. An overview of temporal commonsense reasoning and acquisition. *arXiv preprint arXiv:2308.00002*.

Feifan Wu, Lingyuan Liu, Wentao He, Ziqi Liu, Zhiqiang Zhang, Haofen Wang, and Meng Wang. 2024a. Time-sensitve retrieval-augmented generation for question answering. In *Proceedings of the 33rd ACM International Conference on Information*

and Knowledge Management (CIKM 2024), pages 2544–2553, Boise, ID, USA. ACM.

Xin Wu, Yuqi Bu, Yi Cai, and Tao Wang. 2024b. Updating large language models' memories with time constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13693–13702, Miami, Florida, USA. Association for Computational Linguistics.

Zhonghua Wu, Wenzhong Yang, Meng Zhang, Fuyuan Wei, and Xinfang Liu. 2025. A reinforcement learning-based generative approach for event temporal relation extraction. *Entropy*, 27(3):284.

Yuwei Xia, Ding Wang, Qiang Liu, Liang Wang, Shu Wu, and Xiao-Yu Zhang. 2024. Chain-of-history reasoning for temporal knowledge graph forecasting. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16144–16159, Bangkok, Thailand. Association for Computational Linguistics.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. 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)*, pages 1040–1051.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470, Bangkok, Thailand. Association for Computational Linguistics.

Sen Yang, Xin Li, Lidong Bing, and Wai Lam. 2023. Once upon a *time* in *graph*: Relative-time pretraining for complex temporal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11879–11895, Singapore. Association for Computational Linguistics.

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*, pages 14495–14508, Miami, Florida, USA. Association for Computational Linguistics.

Zairun Yang, Yilin Wang, Zhengyan Shi, Yuan Yao, Lei Liang, Keyan Ding, Emine Yilmaz, Huajun Chen, and Qiang Zhang. 2025. EventRAG: Enhancing llm generation with event knowledge graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16967–16979, Vienna, Austria. Association for Computational Linguistics.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. Large language models as analogical reasoners. In *International Conference on Learning Representations (ICLR)*.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2024a. Temporal relation extraction with contrastive prototypical sampling. *Knowledge-Based Systems*, 286:111410.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024b. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 1963–1974, Singapore. Association for Computing Machinery.

Yixiao Zeng, Tianyu Cao, Danqing Wang, Xinran Zhao, Zimeng Qiu, Morteza Ziyadi, Tongshuang Wu, and Lei Li. 2025. RARE: Retrieval-aware robustness evaluation for retrieval-augmented generation systems. *arXiv preprint arXiv:2506.00789*.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Zhang and Eunsol Choi. 2023. Mitigating temporal misalignment by discarding outdated facts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14213–14226.

Michael JQ Zhang and Eunsol Choi. 2025. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5541–5558, Albuquerque, New Mexico. Association for Computational Linguistics.

Ran Zhang, Jihed Ouni, and Steffen Eger. 2024a. Cross-lingual cross-temporal summarization: Dataset, models, evaluation. *Computational Linguistics*, 50(3):1001–1047.

Siyue Zhang, Yuxiang Xue, Yiming Zhang, Xiaobao Wu, Anh Tuan Luu, and Chen Zhao. 2025. MRAG: A modular retrieval framework for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3080–3118, Suzhou, China. Association for Computational Linguistics.

Xinliang Frederick Zhang, Nick Beauchamp, and Lu Wang. 2024b. Narrative-of-thought: Improving temporal reasoning of large language models via recounted narratives. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16507–16530, Miami, Florida, USA. Association for Computational Linguistics.

Zihan Zhang, Meng Fang, and Ling Chen. 2024c. RetrievalQA: Assessing adaptive retrieval-augmented

generation for short-form open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6963–6975, Bangkok, Thailand. Association for Computational Linguistics.

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)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021a. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371, Online. Association for Computational Linguistics.

Jie Zhou, Shenpo Dong, Hongkui Tu, Xiaodong Wang, and Yong Dou. 2022. RSGT: Relational structure guided temporal relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2001–2010, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yichao Zhou, Yu Yan, Rujun Han, J. Harry Caufield, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2021b. Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021)*, pages 14647–14655, Virtual Event. AAAI Press.

Yu Zhu, Tinghuai Ma, Shengjie Sun, Huan Rong, Yexin Bian, and Kai Huang. 2025a. Rta: A reinforcement learning-based temporal knowledge graph question answering model. *Neurocomputing*, 617:128994.

Zhiyuan Zhu, Yusheng Liao, Zhe Chen, Yuhao Wang, Yunfeng Guan, Yanfeng Wang, and Yu Wang. 2025b. EvolveBench: A comprehensive benchmark for assessing temporal awareness in LLMs on evolving knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16173–16188, Vienna, Austria. Association for Computational Linguistics.

Zulun Zhu, Haoyu Liu, Mengke He, and Siqiang Luo. 2025c. Right answer at the right time - temporal retrieval-augmented generation via graph summarization. *Preprint*, arXiv:2510.16715. STAR-RAG (preprint).

A Taxonomy Figure

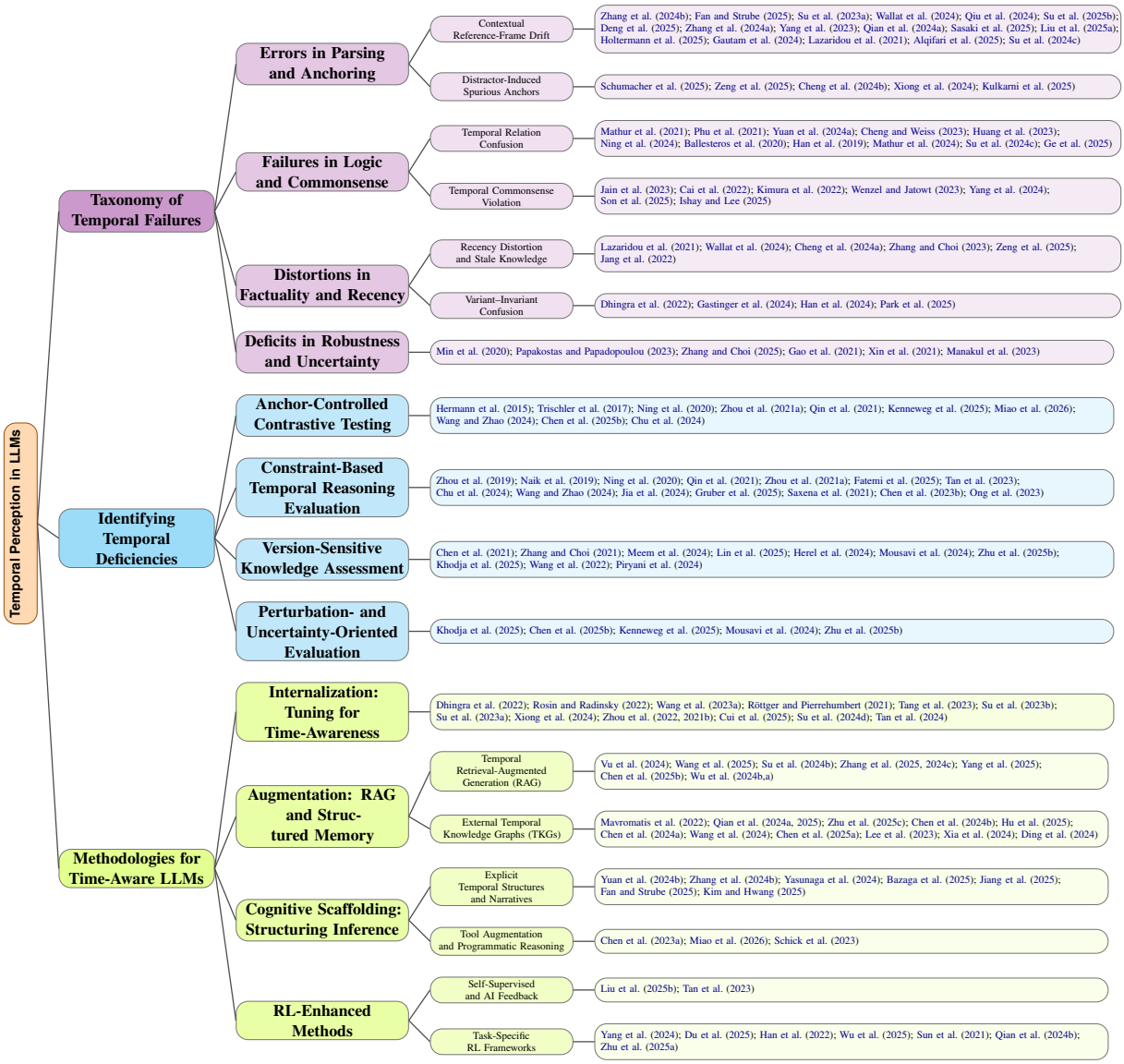

Figure 3: Comprehensive taxonomy of temporal perception in LLMs, organized along three pillars: failure modes (Section 3), evaluation methodologies (Section 4), and mitigation techniques (Section 5), with representative references for each sub-category.

B Resources table

Dataset / Resource	Reference	Size (coverage; scale)	Brief description
New York Times Annotated Corpus	(Sandhaus, 2008)	1987–2007; >1.8M articles	Large diachronic news archive with metadata; widely used for historical IR and time-aware NLP, and as a source corpus for downstream historical QA.
CNN/Daily Mail (RC corpora)	(Hermann et al., 2015)	2007–2015; CNN 90,266 docs / 380,298 queries; DM 196,961 docs / 879,450 queries	News-based cloze-style reading comprehension triples (document–query–answer) automatically constructed from articles and highlights.
Chronicling America (Historic American Newspapers)	(Library of Congress, n.d.)	1690–1963; >11M digitized pages	Library of Congress historical newspaper collection; long-range diachronic source for historical retrieval and downstream QA construction.
Newswire	(Silcock et al., 2024)	1878–1977; 2.7M articles	Structured historical newswire database enriched with metadata (e.g., geo datelines and entity links), enabling fine-grained spatio-temporal modeling.
CUSTOMNEWS	(Lazaridou et al., 2021)	1969–2019; ∼395.6GB train (avg. 491 words/doc)	Large crawled English news corpus used to study temporal generalization and temporal distribution shift in language models.
Wikidata (KB)	(Vrandečić and Krötzsch, 2014)	snapshot; 14.45M items / 30.26M statements	Collaborative structured knowledge base with qualifiers and time-related fields; often linked to Wikipedia snapshots for temporally-scoped modeling.
TimeBank	(Pustejovsky et al., 2003)	annotated news; 183 docs (61,418 tokens)	Classic TimeML-annotated corpus with events, temporal expressions, and temporal relations for structured temporal reasoning.
WikiWars	(Mazur and Dale, 2010)	Wikipedia narratives; 22 docs (∼120k tokens)	Annotated corpus focusing on temporal expressions in long historical narratives (TIMEX tagging and normalization).
RED (Richer Event Description)	(O'Gorman et al., 2016)	annotated docs; 95 docs	Event-centric annotation corpus integrating event coreference with temporal, causal, and bridging relations for richer narrative/event modeling.

Table 1: Resource-style temporal corpora, collections, and knowledge bases used as evidence sources or annotated data for time-aware modeling.

Dataset	Reference	Dataset size	Brief description
CNN/DailyMail RC	(Hermann et al., 2015)	1.38M cloze Qs (CNN 387K + DM 997K)	Cloze RC over news: fill in anonymized entities in summary-style questions given the article.
NewsQA	(Trischler et al., 2017)	119K QA pairs	Human-written QA over CNN articles; answers are grounded in the passage and include harder, less templatic questions.
TORQUE	(Ning et al., 2020)	21K questions	Temporal RC benchmark: questions explicitly probe before/after/during relations and event ordering constraints in text.
TRACIE	(Zhou et al., 2021a)	5.4K instances	Temporal inference with implicit events: requires reasoning about events that are suggested but not explicitly stated.
TIMEDIAL	(Qin et al., 2021)	1.1K questions	Dialog-based temporal commonsense (multiple-choice): resolve schedules, durations, and ordering from conversational context.
TRAVELER	(Kenneweg et al., 2025)	3.3K questions	Synthetic temporal reasoning across explicit, implicit, and vague references; emphasizes underspecified and relative time expressions.
SPAN	(Miao et al., 2026)	Real-time generation questions	Cross-calendar temporal reasoning: tests date conversion and temporal reasoning across heterogeneous calendar systems.
TimeBench	(Chu et al., 2024)	19K instances	Hierarchical, broad-coverage temporal reasoning suite spanning multiple temporal phenomena (e.g., ordering, duration, arithmetic).
TRAM	(Wang and Zhao, 2024)	526.7K questions	Large-scale LLM benchmark aggregating diverse temporal aspects (order, arithmetic, frequency, duration) across multiple datasets.
TS-RAG QA	(Chen et al., 2025b)	5.2K questions based on 300k+ news	Time-sensitive RAG QA: evaluates whether retrieval selects time-appropriate evidence and generation produces time-correct answers.
Factual knowledge under temporal context shifts	(Khodja et al., 2025)	5K questions	Tests factual robustness when the prompt's reference time changes, exposing temporal-context-induced inconsistencies.
Test of Time (ToT)	(Fatemi et al., 2025)	1.8K questions	Synthetic temporal benchmark combining controlled temporal semantics with temporal arithmetic and compositional reasoning.
TempReason	(Tan et al., 2023)	52.8K questions	Temporal reasoning QA over time-scoped knowledge, designed to probe temporal generalization across long time ranges.
CRONQUESTIONS	(Saxena et al., 2021)	410K questions	Temporal KGQA: answer queries with temporal constraints over time-scoped facts (intervals, start/end times, validity).
Multi-granularity TKGQA	(Chen et al., 2023b)	151K questions	Temporal KGQA with multiple time granularities (e.g., year/month/day), stressing fine-grained temporal constraints.
TimeQA (Time-sensitive QA)	(Chen et al., 2021)	41.2K questions	Time-sensitive QA where answers can change across time; supports evaluating temporal awareness and update sensitivity.
SituatedQA	(Zhang and Choi, 2021)	12.2K questions	Contextual QA with extra-linguistic signals (time/location): the same question may require different answers under different contexts.
PAT-Questions	(Meem et al., 2024)	6.1K questions	Present-anchored temporal QA that can self-update, aiming to remain current as "today" and facts evolve.
DynaQuest	(Lin et al., 2025)	22.6K QA pairs	Dynamic QA with multiple time snapshots to measure knowledge drift and model responsiveness to real-world updates.
Time awareness (fact recall across time)	(Herel et al., 2024)	32K samples, Real-time generation questions	Probes factual recall conditioned on different reference times, measuring time awareness and temporal calibration in LLMs.
DyKnow	(Mousavi et al., 2024)	390 questions	Dynamic verification setting: evaluates time-sensitive factual claims by checking against evolving knowledge sources.
EvolveBench	(Zhu et al., 2025b)	1640 questions	Temporal awareness on evolving knowledge: tests whether models track and reflect changing real-world facts over time.
MC-TACO	(Zhou et al., 2019)	13K questions	Multiple-choice temporal commonsense: focuses on typical durations, frequencies, and plausible temporal relations in context.
TDDiscourse	(Naik et al., 2019)	6.1K pairs	Discourse-level temporal ordering: annotate temporal relations between events across sentences and broader discourse structure.
TiQ	(Jia et al., 2024)	10K questions	Temporal QA with implicit time constraints: requires inferring hidden temporal conditions to retrieve/derive the correct answer.
ComplexTempQA	(Gruber et al., 2025)	100.2M QA pairs	Large-scale complex temporal QA emphasizing compositional constraints, multi-step reasoning, and temporal metadata.
TKGQA Dataset	(Ong et al., 2023)	60K questions	QA to guide and validate temporal KG evolution, checking whether KG updates preserve temporal consistency and correctness.
ArchivalQA	(Wang et al., 2022)	532K QA pairs	Open-domain QA over historical news archives; requires retrieving time-appropriate articles and extracting answers from them.
ChroniclingAmericaQA	(Piryani et al., 2024)	485K QA pairs	Large-scale QA over digitized historical newspapers (page/OCR artifacts), combining retrieval over archives with answer extraction.

Table 2: Benchmark datasets details.