Knowing When to Predict: Confidence-Aware Temporal Reasoning in Large Language Models

Anonymous ACL submission

Abstract

Reasoning about future factual states remains a critical challenge for Large Language Models (LLMs). Current methods typically rely on static knowledge and fail to capture the underlying dynamic patterns of factual change, causing LLMs to exhibit temporal awareness deficits such as Persistence Bias and Change Insensitivity. To address these limitations, we propose **C**onfidence-Aware **T**emporal **R**easoning (**CTR**), a framework that determines *whether* and *how* future factual predictions should be made by explicitly modeling temporal evolution and assessing factual predictability prior to answer generation. CTR leverages token-level entropy to weight historical evidence, allowing the model to distinguish between reliable patterns and uncertain noise. We evaluate our approach on an updated version of the FRESHQA benchmark across multiple LLMs. Experimental results demonstrate consistent improvements in future factual reasoning: on GPT-4o, accuracy increases from 44.8% to 53.7% across all facts compared to prompt-based baselines. Moreover, CTR reduces hallucination rates by 14.2% on LLaMA-3.1 and 6.5% on GPT-4o, substantially enhancing the trustworthiness of future-oriented predictions.

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities in factual reasoning and question answering (Fatemi et al., 2024; Wang et al., 2024; Hu et al., 2025). However, their reliability on time-sensitive queries remains a critical challenge (Dhingra et al., 2022). While recent work has addressed issues such as outdated knowledge and temporal inconsistency (Bajpai et al., 2024), a fundamental question remains insufficiently explored: How do LLMs reason about *future factual states* when real-world facts evolve over time?

When facing such future-oriented factual reasoning queries, existing approaches typically treat time as an supplementary retrieval condition or context tag, answering by combining static knowledge (Vu et al., 2024; Wu et al., 2024). This formulation overlooks the reality that knowledge evolves as a dynamic process rather than a static lookup table. For instance, the answer to *"Who is the President of France?"* changes in regular and predictable cycles, whereas facts such as *"Where will ACL be held?"* evolve in irregular and non-periodic ways. Correctly answering such queries requires more than simply retrieving historical data; it also demands that models reason about the underlying dynamics of factual evolution and regulate their confidence accordingly when extrapolating into the future.

Consequently, we conduct a preliminary study in Section 3 to test whether models can capture such logical changes when provided with complete historical context. The study reveals a systematic failure in temporal awareness: current models remain overly confident, failing to detect periodic change and either persisting with outdated answers or hallucinating when confronted with volatile facts.

The findings motivate us to reframe future factual reasoning as a problem of temporal decision-making rather than unconditional extrapolation. We propose **Confidence-Aware Temporal Reasoning (CTR)**, in an attempt to simulate human temporal reasoning behavior, where people confidently extend stable and predictable patterns, yet deliberately suspend judgment when future outcomes are volatile or poorly constrained.

CTR is implemented via a confidence-aware inference mechanism to improve temporal reasoning chains. Rather than treating all retrieved historical evidence as equally reliable, CTR leverages token-level entropy to weight historical information, thereby automatically filtering out noisy or uncertain records. This design enables the model to maintain reasoning consistency while also self-correcting erroneous historical knowledge and preventing the propagation of errors.

We evaluate our framework across multiple LLMs (including LLaMA-3.1, Qwen3, and GPT-4o) on an updated version of FRESHQA (Vu et al., 2024), a dynamic benchmark designed for time-evolving factual questions. Building upon the original dataset, we update and extend its underlying knowledge to reflect more recent factual states. Experimental results show that CTR consistently improves performance on future factual reasoning across models. Most notably, CTR acts as a safety valve for volatile queries, reducing hallucination rates by up to 21.7%.

In summary, our contributions are as follows:

- We provide a quantitative diagnosis of temporal reasoning failures in current LLMs, revealing a systematic Persistence Bias and Change Insensitivity. Our analysis shows that current models lack an internal signal for factual evolution.
- We propse **CTR**, a structured framework that advances future reasoning by explicitly modeling factual evolution. Unlike static retrieval, CTR leverages confidence-aware signals to validate historical stability, ensuring that future predictions rely on consistent trends rather than noisy data.
- We demonstrate that CTR significantly enhances the trustworthiness of future predictions. It achieves a superior trade-off by maintaining accuracy on stable facts while drastically reducing hallucinations in uncertain, changing scenarios.

2 Related Work

Temporal reasoning has long been a central topic in the study of large language models (Vashishtha et al., 2020; Hu et al., 2023; Gurnee and Tegmark, 2023; Beniwal et al., 2024). A substantial body of work has evaluated LLMs' temporal reasoning capabilities from multiple perspectives, leading to the development of benchmark datasets and evaluation metrics that assess factual correctness (Fatemi et al., 2024; Chu et al., 2024; Zhu et al., 2025; Kasai et al., 2023), consistency (Bajpai et al., 2024; Jia et al., 2024), and reasoning traceability (Xu et al., 2023; Bazaga et al., 2025; Zhu et al., 2025). These studies consistently show that temporal factuality in LLMs is fundamentally challenged by the static and temporally mixed nature of pretraining corpora, which often results in outdated predictions, averaging of conflicting facts across time, and poorly calibrated confidence when models are queried about future states (Dhingra et al., 2022; Vu et al., 2024; Bajpai et al., 2024; Kasai et al., 2023).

Part of existing research focuses on improving LLMs' temporal awareness through pretraining or fine-tuning strategies that incorporate explicit temporal signals, with the goal of enhancing temporal scope understanding and confidence calibration (Ning et al., 2018; Dhingra et al., 2022; Yang et al., 2023; Xiong et al., 2024a). These approaches typically treat time as a conditioning variable that constrains the validity of factual knowledge, rather than explicitly modeling how facts evolve over time.

One line of work seeks to enhance temporal reasoning through structured representations, most notably knowledge graph–based approaches that explicitly encode entities, relations, and temporal constraints (Vu et al., 2024; Xiong et al., 2024b; Wang and Zhao, 2024; Hu et al., 2025). But they typically rely on predefined schemas, which makes graph updates costly and complex in dynamic settings where facts continuously evolve, thereby limiting their practicality for future-oriented reasoning that requires timely knowledge updates.

Another prominent research direction augments LLMs with external retrieval mechanisms to refresh factual knowledge and reduce staleness (Vu et al., 2024; Wu et al., 2024). Retrieval-augmented and deep search methods leverage web search, document retrieval, or similar tools to improve performance on time-sensitive queries(Xu et al., 2023; Fu et al., 2024). However, most RAG-style pipelines implicitly treat retrieved documents as reliable evidence, despite differences in publication time and potential conflicts across sources. The lack of explicit modeling of temporal validity and uncertainty allows errors to propagate through reasoning chains, resulting in hallucinations and inconsistencies (Trivedi et al., 2023; Vu et al., 2024; Singal et al., 2024).

Consistency-focused probes further reveal that LLMs often produce unstable predictions across paraphrases and temporal directions (Bajpai et al., 2024; Jia et al., 2024), suggesting that errors can propagate along extended reasoning chains even when relevant facts are available.

Synthesizing these findings, it becomes evident that temporal factuality depends not only on access to time-indexed facts, but also on how temporal information is structured and propagated through

multi-step reasoning. However, how large language models leverage historical knowledge to reason about future knowledge while maintaining accuracy, consistency, and interpretability remains largely unexplored, motivating a focused empirical analysis of future-oriented temporal reasoning.

3 Preliminary Study: The Persistence Bias in Future Reasoning

In this section, we aim to investigate a critical question: *Is access to complete historical knowledge sufficient for correct future reasoning?*

3.1 Problem Setting

For each predictable fact instance i, associated with a latent annotation:

$$e_i = \langle q_i,\ t_0^i,\ a_0^i,\ t_c^i,\ a_1^i \rangle, \tag{1}$$

where q_i is the factual query; t_0^i is the latest reference time at which the fact is known to be valid; a_0^i is the correct answer before the change; t_c^i is the change point; and a_1^i is the correct answer after the change.

This annotation defines a ground-truth answer function:

$$\text{Answer}(q_i, t) = \begin{cases} a_0^i, & t < t_c^i, \\ a_1^i, & t \geq t_c^i. \end{cases} \tag{2}$$

Importantly, the annotation e_i is *not* provided to the model. The model is still queried with the same input format $\langle q_i, t \rangle$ and produces the same outputs $\langle r_{i,t}, y_{i,t}, \hat{a}_{i,t} \rangle$ as in the general task.

For each instance e_i, we construct a set of probing times:

$$\begin{aligned} &t_{i,k} = t_c^i + \Delta_k, \\ &\Delta_k \in \{-W, \ldots, -\tfrac{1}{2}, \tfrac{1}{2}, \ldots, W\}. \end{aligned} \tag{3}$$

where $W = 3$ years in our experiments, and the temporal resolution is half-year increments. We exclude $\Delta_k = 0$ to avoid ambiguity exactly at the change point. At each probing time $t_{i,k}$, the model is provided with complete historical information:

$$x_{i,k} = \Big(q_i,\ (t_0^i, a_0^i),\ t_c^i,\ t_{i,k}\Big), \tag{4}$$

and is asked to predict the factual state as of $t_{i,k}$, where the outputs consist of a reasoning trace $r_{i,k}$ and a final predicted answer $\hat{a}_{i,k}$.

This setting simulates an idealized retrieval-augmented environment, ensuring that any failure cannot be attributed to missing or outdated information.

3.2 Evaluation Metrics

We evaluate model behavior using complementary metrics that characterize change-aware future reasoning.

Pre- and Post-change Accuracy We report prediction accuracy separately for time points before and after the change point, measured by exact match against the ground-truth factual state. Pre-change accuracy reflects factual stability, while post-change accuracy evaluates adaptation after the transition.

Change-aware Success (CAS) To directly assess whether a model captures the transition itself, we define **CAS** as:

$$\text{CAS}_i = \mathbb{I}\big(\hat{a}_{i,-1} = a_0^i \ \wedge\ \hat{a}_{i,+1} = a_1^i\big), \tag{5}$$

where predictions immediately before and after the change point must both be correct. CAS rules out accidental correctness at isolated time steps and explicitly measures temporal switching behavior.

Persistence Rate (PR) The *PR* quantifies the tendency to repeat the anchor answer regardless of time:

$$\text{PR} = \frac{1}{|\mathcal{T}|} \sum_{i,k} \mathbb{I}(\hat{a}_{i,k} = a_0^i), \tag{6}$$

where $\mathcal{T}$ denotes all probed time points. A high PR indicates over-reliance on static factual knowledge, especially when extending beyond the change point.

Entropy Shift (ΔH) We measure whether model uncertainty changes across the transition using the difference in average answer entropy between post-change and pre-change regions. The formal definition of answer entropy is provided in Appendix D.2.

3.3 Analysis

We empirically diagnose how current large language models behave when answering future-oriented factual queries under an idealized setting in which all relevant historical information is explicitly provided. Table 1 and Figure 1 jointly reveal a consistent failure mode of current LLMs in future factual reasoning: blind persistence coupled with confidence miscalibration.

Model	Pre Acc ↑	Post Acc ↑	CAS ↑	PR ↓	ΔH ↑
LLaMA3.1	0.500	0.158	0.000	0.462	0.006
Qwen3	**0.911**	0.219	0.250	**0.918**	0.004
Mistral	0.333	0.082	0.036	0.302	0.000
GPT-4o	0.839	0.316	**0.286**	0.758	**0.022**

Table 1: Results on Future Reasoning in LLMs

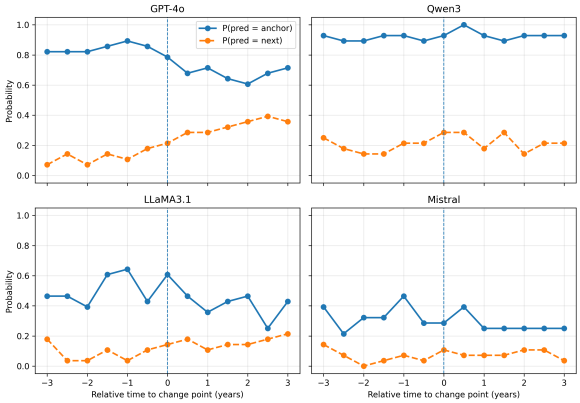

Figure 1: Temporal flip curves of model predictions around factual change points. The figure shows the probability of predicting the anchor (pre-change) answer and the next (post-change) answer as a function of relative time to the change point.

The Persistence Trap. Although models achieve high accuracy before the change point, their performance collapses afterward, with post-change accuracy dropping below 0.32 for all models. This gap indicates that strong pre-change performance primarily reflects persistence of previously valid facts, rather than reasoning about future state transitions.

Figure 1 further exposes this behavior. For GPT-4o, the probability of predicting the anchor answer decays slowly after the change point, while the probability of the correct post-change answer increases only marginally. Qwen3 exhibits an even more extreme pattern, maintaining an anchor prediction probability above 0.9 across nearly the entire temporal window.

The Silence of Uncertainty Signals. Beyond incorrect answers, models also fail to recognize when a change occurs. Change-aware Success (CAS) remains uniformly low, indicating that temporally coherent switching across the transition is rare. Moreover, uncertainty does not respond to factual change: entropy shifts (ΔH) are near zero for most models, showing that confidence remains largely unchanged at the transition boundary. This suggests that current LLMs lack an internal uncertainty signal that reliably reflects factual transitions.

These results suggest that future reasoning failures stem not from missing knowledge, but from the absence of an internal signal indicating when factual commitments should be revised. This profound lack of change sensitivity and confidence calibration directly motivates our proposed Confidence-Aware Temporal Reasoning (CTR) framework, which explicitly models factual evolution and regulates confidence across temporal transitions.

4 Confidence-Aware Temporal Reasoning (CTR)

To overcome the persistence bias inherent in static knowledge recall, we propose **Confidence-Aware Temporal Reasoning (CTR)**. By integrating entropy-based confidence signals to filter historical noise and explicitly model factual evolution, CTR establishes a structured framework for future-oriented inference. The overall architecture is illustrated in Figure 2.

4.1 Problem Setting

Given a factual query q and a future time t^*, the model takes $x = \langle q, t^* \rangle$ as input and infers the factual state at t^* by explicitly reasoning about temporal evolution and outputting $\langle r, z, \hat{a}_{t^*} \rangle$.

where r is a reasoning trace explaining the decision, $z \in \{\text{FAST}, \text{SLOW}, \text{NEVER}\}$ denotes the predicted factual evolution type[1], and $\hat{a}_{t^*}$ is the predicted fill-in value for q as of year t^*.

4.2 CTR Architecture

CTR deliberately adopts conservative heuristics rather than optimized predictors. The goal is not to maximize change detection recall, but to avoid unfounded speculation when evidence is weak or inconsistent.

Driven by this principle, the CTR architecture is designed as a confidence-driven pipeline that progressively filters uncertainty through the following stages. The framework consists of five-stage process: (1) Historical Recall; (2) Timeline Reconstruction; (3) Change-Point Detection; (4) Temporal Pattern Induction and (5) Confidence-Aware Inference.

[1]We adopt the factual evolution types from (Dhingra et al., 2022; Vu et al., 2024) without modifying their original definitions.

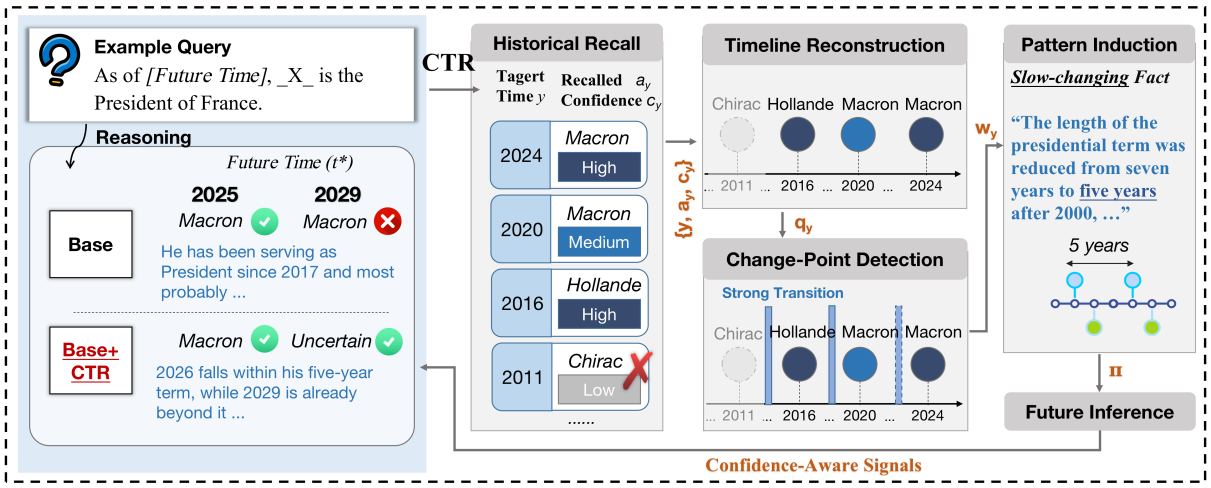

Figure 2: **Overview of the Confidence-Aware Temporal Reasoning (CTR).** CTR leverages entropy-based confidence to shift from static retrieval to dynamic evolution modeling. Notably, its confidence-aware mechanism suppresses erroneous historical evidence (e.g., the low-confidence 2011 entry) to mitigate error propagation. By inducing explicit evolution patterns, CTR overcomes the persistence bias seen in the traditional reasoning modes and prioritizes reliability, avoiding hallucinated extrapolation for volatile future queries.

Historical Recall Given a future time t^*, CTR performs fine-grained historical probing over a window of past years:

$$\mathcal{Y} = \{t^* - W, \ldots, t^* - 1\}. \tag{7}$$

For each target time $y \in \mathcal{Y}$, the model is queried to answer the cloze question *as of time* y, producing a predicted fill-in value a_y and a token-level uncertainty estimate measured by average generation entropy H_y. Each historical probe is represented as

$$r_y = \langle y, a_y, H_y \rangle. \tag{8}$$

To obtain a calibrated reliability signal, entropy is mapped to a base confidence score via a monotonic transformation:

$$c_y = f(H_y), \quad c_y \in [0, 1], \tag{9}$$

yielding a sequence of historical recall probes

$$\mathcal{R} = \{(y, a_y, c_y)\}_{y \in \mathcal{Y}}. \tag{10}$$

Timeline Reconstruction The recall probes are organized into a temporal timeline

$$\mathcal{T} = \{(y, v_y, q_y)\}, \tag{11}$$

where $v_y = a_y$ denotes the recalled factual state at time y, and q_y is a corrected reliability weight.

Importantly, q_y is not directly inherited from the entropy-based confidence c_y. Instead, CTR applies a probe-quality correction function that down-weights unreliable probes, such as those exhibiting abnormally long outputs, repetitive generation patterns, or malformed responses. Formally, we define

$$q_y = \mathrm{CLIP}_{[0,1]}\big(c_y \cdot (1 - \rho_{\mathrm{len}}(y)) \cdot (1 - \rho_{\mathrm{rep}}(y))\big), \tag{12}$$

where ρ_{len} and ρ_{rep} denote mild penalties based on output length and n-gram repetition, respectively. This correction reflects the fact that historical recall itself may be noisy and should not be treated as uniformly reliable.

Change-Point Detection Given the reconstructed timeline $\mathcal{T}$, CTR aims to identify years in which the factual state undergoes a genuine transition. We adopt a procedure that combines candidate proposal with consistency-aware reweighting.

Candidate proposal: We first obtain a set of candidate change-point years $\tilde{\mathcal{C}}$ by prompting the model to inspect the timeline and identify major state transitions. If the model output is unavailable or invalid, we fall back to a deterministic rule that proposes time y whenever $v_y \neq v_{y-1}$. All candidates are restricted to years explicitly appearing in the timeline.

Consistency-aware reweighting: Each candidate $y \in \tilde{\mathcal{C}}$ is then filtered and assigned a confidence weight. Candidates with no effective state change ($v_y = v_{y-1}$) are discarded. For the remaining candidates, we compute

$$w_y = \sqrt{q_{y-1}\, q_y} \cdot g_y, \tag{13}$$

where w_y denotes the confidence of the change point at time y and $g_y \in (0, 1]$ penalizes isolated single-year deviations.

Specifically, a candidate is treated as singleton noise if it is unsupported by its immediate temporal context, i.e., $v_{y-1} = v_{y+1}$ but $v_y \neq v_{y-1}$, and its probe reliability is substantially lower than that of neighboring years.

In such cases, g_y is reduced whenever

$$q_y < \tau_{\text{single}} \cdot \min(q_{y-1}, q_{y+1}), \tag{14}$$

thereby enforcing a local temporal consistency constraint. This design favors transitions that persist across time and suppresses spurious one-off fluctuations.

The resulting set of weighted change points is denoted as

$$\mathcal{C} = \{(y, w_y)\}. \tag{15}$$

Temporal Pattern Induction Given the reconstructed timeline $\mathcal{T}$ and weighted change points $\mathcal{C}$, CTR induces a high-level temporal evolution pattern by aggregating volatility statistics over the timeline. Based on the frequency and persistence of state changes, the factual evolution type z is inferred.

Stable temporal segments and high-confidence change points are then summarized into a structured pattern representation

$$\pi = \langle z, \text{summary}, \text{key\_evidence} \rangle, \tag{16}$$

which serves as an explicit evolution hypothesis for subsequent inference.

Confidence-Aware Future Inference Finally, CTR performs future inference conditioned on the induced pattern representation π, rather than directly extrapolating from raw historical facts.

The model predicts the factual state at the future time t^* as

$$\hat{a}_{t^*} = \arg\max_a p(a \mid q, t^*, \pi), \tag{17}$$

subject to strict constraints that prohibit unsupported extrapolation.

5 Experiments

5.1 Dataset

We construct our evaluation dataset by extending and restructuring existing temporal factual QA benchmarks to support diagnostic analysis of temporal evolution and confidence-aware reasoning.

Dataset Construction We take FRESHQA [2] as the primary foundation and manually update all factual answers to reflect the world state as of *September 2025*. Following prior work on temporal factual probing, all questions are converted into a unified *cloze-style* format, where the target entity or value is replaced by a placeholder `_X_`. This formulation enables controlled evaluation across different temporal future times.

For diagnostic purposes, we augment each instance with a *next_answer* field whenever the future factual state is well-defined and predictable. This field is used only for analysis and evaluation, and is not provided to the model as input. For unpredictable facts, where no stable future transition is assumed, *next_answer* is set to *N/A*.

Dataset Statistics The final dataset comprises a total of 337 temporal factual queries. To ensure comprehensive evaluation across different volatility profiles, we maintain a balanced distribution of evolution types: 102 *fast-changing*, 119 *slow-changing*, and 116 *never-changing* facts. This diversity allows us to rigorously assess whether models can distinguish between stable knowledge and volatile trends. Dataset details and additional examples are provided in the Appendix F.

5.2 Models

We evaluate CTR on both open-source and closed-source large language models to ensure broad coverage of contemporary architectures. Since our evaluation involves future-oriented temporal reasoning, all selected models are required to have training data that do not include knowledge beyond 2025.

Our open-source models include **LLaMA-3.1-8B-Instruct** (Dubey et al., 2024)[3], **Qwen3-4B-Instruct** (Yang et al., 2025)[4], and **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023)[5], with weights obtained from Hugging Face. We additionally evaluate the closed-source model **GPT-4o**[6] via the official API.

All models are evaluated in a zero-shot setting with identical prompts and decoding configurations.

[2] https://github.com/freshllms/freshqa

[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[4] https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

[5] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[6] https://platform.openai.com/docs/models/gpt-4o

Model	Variant	Answer Accuracy (%) ↑				Type Acc. (%) ↑	Halluc. (%) ↓
		Fast	Slow	Never	Overall		
LLaMA3.1	Base	**6.9**	21.0	50.0	26.7	23.7	36.8
	+Prompting	5.9	29.4	53.4	30.6	26.1	33.2
	+CTR	**6.9**	**31.9**	**56.0**	**32.6**	**55.3**	**22.6**
Qwen3	Base	4.9	26.9	41.4	25.2	35.0	56.7
	+Prompting	**8.8**	21.0	47.4	26.4	54.9	54.6
	+CTR	6.9	**32.8**	**57.8**	**33.5**	**57.5**	**25.5**
Mistral	Base	7.8	12.6	50.0	24.0	29.7	40.1
	+Prompting	4.9	16.0	42.2	21.7	51.9	36.8
	+CTR	**8.8**	**21.0**	**50.9**	**27.6**	**52.8**	**18.4**
GPT-4o	Base	16.7	40.3	74.1	44.8	48.1	17.8
	+Prompting	13.7	46.2	79.3	47.8	**63.8**	13.7
	+CTR	**18.6**	**52.9**	**85.3**	**53.7**	**63.8**	**11.3**

Table 2: Main results (%) on answer accuracy, type accuracy, and hallucination rate. Prompt-guided reasoning yields inconsistent gains over plain QA, while CTR consistently improves performance on slow- and never-changing facts and reduces hallucinations across models.

5.3 Experimental Setup

All experiments are conducted with a fixed future time $t^* \geq 2025$. Unless otherwise specified, CTR probes a historical window of size $W = 50$, querying years $\{t^* - 50, \ldots, t^* - 1\}$. All models are evaluated in a zero-shot setting using identical prompts, with a decoding temperature of 0.2 and *max_new_tokens* set to 1024.

To isolate the contribution of our structured framework from generic in-context learning, we introduce a Prompt-guided baseline based on Chain-of-Thought (Wei et al., 2022) prompting. This baseline represents the standard paradigm for temporal reasoning, where the model is prompted to perform step-by-step deduction over the historical timeline to extrapolate future states. For a fair comparison, this baseline shares the same backbone models, decoding configurations, and historical context as CTR, serving as a control to ensure that performance gains arise from CTR's confidence-aware architecture rather than prompting alone.

Experiments are run on a Linux server equipped with 10 NVIDIA A100 GPUs (80GB). All experiments are repeated 5 times, and results are averaged across runs.

Hallucination rates are reported for controlled relative comparison across methods under identical evaluation conditions, and are measured using a two-stage verification protocol combining automated screening and expert human adjudication, as detailed in Appendix E.

5.4 Main Result

Table 2 compares the Base models, the prompt-guided baseline, and the proposed CTR framework. Overall, CTR consistently improves answer accuracy and evolution type prediction while substantially reducing hallucination, demonstrating more reliable future reasoning.

CTR outperforms both baselines in answer accuracy across all backbone models, with the largest gains on Slow- and Never-changing facts. On GPT-4o, accuracy on Slow-changing facts increases from 46.2% to 52.9% and on Never-changing facts from 79.3% to 85.3% compared to prompting. Importantly, these gains do not degrade performance on Fast-changing facts. GPT-4o with CTR achieves 18.6% accuracy on Fast facts, exceeding both the Base (16.7%) and Prompt-guided (13.7%) settings, indicating that CTR selectively commits to predictions rather than suppressing them under uncertainty.

Prompt engineering alone provides limited and unstable benefits. While prompting improves overall accuracy for some models, such as LLaMA3.1 from 26.7% to 30.6%, it fails to generalize across fact types. In particular, prompting can hurt Fast-fact performance, as observed on Mistral where accuracy drops from 7.8% to 4.9%. By contrast, CTR introduces explicit confidence calibration, leading to higher Evolution Type Accuracy, improving Qwen3 from 54.9% with prompting to 57.5%.

CTR also yields the largest improvements in reliability. Hallucination rates are consistently lower than both baselines across all models. For Qwen3,

Case	Model	Gold	Prompt-guided Baseline	Confidence-Aware TR (CTR)
The number of books published by Shoshana Zuboff	GPT-4o	3	**Answer:** 4 **Reasoning:** Assumes a slow publication rate without citing concrete historical evidence.	**Answer:** 3 **Reasoning:** "The number of books remained stable at 3 from 2018 to 2024. Given the *Slow-changing* pattern, it is likely unchanged in 2025."
Children of Leonardo DiCaprio	Qwen3	0	**Answer:** 1 **Reasoning:** "Leonardo DiCaprio has been publicly known to have one child, a daughter named Matilda, born in 2004."(*Hallucination*)	**Answer:** 0 **Reasoning:** "The evidence indicates a stable period from 1978 to 2024 with no children."

Table 3: Two representative case studies comparing standard prompt-guided reasoning and our CTR framework. While the baseline suffers from speculative extrapolation (Case 1) and factual hallucination (Case 2), CTR utilizes structured temporal grounding to produce reliable, evidence-based predictions.

hallucination decreases from 56.7% in the Base setting to 25.5% under CTR, far below the prompt-guided result of 54.6%. GPT-4o exhibits a similarly low hallucination rate of 11.3%. These results show that CTR effectively suppresses unsupported future claims while preserving predictive accuracy.

5.5 Case Study

Table 3 provides a qualitative comparison between the prompt-guided baseline and our CTR framework. We selected two representative cases that highlight distinct failure modes in future-oriented reasoning: unsupported extrapolation and factual hallucination.

In the first case concerning Shoshana Zuboff's publications, the baseline predicts an increase in the number of books without citing any concrete historical evidence, reflecting a tendency to assume change by default. CTR, in contrast, evaluates the temporal reliability of historical recall and identifies a long stable period with consistently high confidence. Based on this inferred slow-changing evolution pattern, CTR conservatively maintains the historical value for 2025.

A more severe failure is observed in the Leonardo DiCaprio case. The baseline fabricates a specific but non-existent detail ("a daughter named Matilda"), demonstrating that without explicit grounding, LLMs can generate coherent yet completely hallucinated narratives. CTR avoids this error by aggregating decades of consistent historical evidence. The confidence-aware mechanism effectively filters out low-probability generative noise, allowing the model to conclude that the factual state remains "Zero".

These cases illustrate a critical advantage of our framework: by explicitly modeling factual evolution patterns and regulating confidence, CTR serves as a reliability layer, preventing the model from making baseless guesses in uncertain temporal contexts. A detailed confidence-level and temporal evolution analysis of this repair mechanism is provided in Appendix C.

6 Conclusion

In this work, we investigate the challenge of future-oriented factual reasoning in large language models. Our preliminary study reveals a fundamental limitation: even when provided with complete historical evidence, current LLMs suffer from severe persistence bias and insensitivity to change. They tend to extrapolate outdated facts and fail to exhibit increased uncertainty near temporal boundaries, indicating a limited internal awareness of factual evolution.

To address this issue, we propose CTR, a structured inference framework that shifts the paradigm from static retrieval to dynamic evolution modeling. CTR grounds future reasoning in confidence-aware temporal signals, distinguishing stable from unreliable historical evidence and committing to predictions only when the inferred evolution is sufficiently supported. As a result, CTR enables models to reason not only about *what* may happen, but also about *whether* a future prediction is warranted, rather than extrapolating unconditionally.

Extensive experiments across four LLMs demonstrate that CTR significantly outperforms both standard and prompt-guided baselines. It not only achieves higher accuracy but also reduces hallucination rates. By explicitly summarizing temporal patterns, CTR provides transparent grounding for its predictions. These findings position CTR as a robust approach for more reliable and trustworthy time-sensitive artificial intelligence.

Limitations

While CTR demonstrates consistent improvements in future-oriented factual reasoning, several limitations remain.

Computational efficiency. Our primary experimental setup adopts a rigorous year-by-year probing strategy ($W = 50$) to establish a diagnostic upper bound for temporal reasoning performance. We acknowledge that this incurs a linear computational cost ($\mathcal{O}(W)$) relative to the window size. However, the historical recall process is fully parallelizable. Furthermore, for real-world deployment, CTR is designed to function primarily as an offline knowledge curation pipeline, constructing reliable temporal indices once to support low-latency lookups for subsequent queries. Future iterations can further optimize efficiency by implementing adaptive sampling (e.g., binary search) to reduce the probing complexity from linear $\mathcal{O}(W)$ to logarithmic $\mathcal{O}(\log W)$.

Uncertainty estimation. CTR uses token-level entropy as a proxy for model confidence. While this signal is model-agnostic and easy to compute, it may not fully capture epistemic uncertainty, particularly for models with imperfect probability calibration. Exploring alternative or complementary uncertainty measures remains an important direction for future work.

Scope of evaluation. Our evaluation focuses on cloze-style temporal factual question answering with discrete time points. While this setting enables controlled analysis of future commitment and abstention, it does not cover more complex scenarios such as open-ended forecasting, interacting events, or long-horizon planning, which remain open challenges for future exploration.

Ethics Statement

All work in this paper adheres to the ACL Code of Ethics. For the human evaluation of hallucinations, we recruited volunteer annotators and compensated them in accordance with local wage standards. The collected data do not involve any personally identifiable or sensitive information.

The study is conducted in a controlled research setting using released large language models under their respective licenses and usage policies. The proposed framework is intended for methodological analysis of future-oriented factual reasoning rather than real-world deployment. While model-generated outputs may contain inaccuracies, such risks are inherent to large language models and are not the focus of this work.

References

Ashutosh Bajpai, Aaryan Goyal, Atif Anwer, and Tanmoy Chakraborty. 2024. Temporally consistent factuality probing for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15864–15881.

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. *arXiv preprint arXiv:2504.05258*.

Himanshu Beniwal, Dishant Patel, Kowsik Nandagopan, Hritik Ladia, Ankit Yadav, and Mayank Singh. 2024. Remember this event that year? assessing temporal information and understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16239–16348.

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.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning. *arXiv preprint arXiv:2406.09170*.

Chenhan Fu, Guoming Wang, Rongxing Lu, and Siliang Tang. 2024. Fastlearn: A rapid learning agent for chat models to acquire latest knowledge. *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 183–189.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Nan Hu, Jiaoyan Chen, Yike Wu, Guilin Qi, Hongru Wang, Sheng Bi, Yongrui Chen, Tongtong Wu, and Jeff Z Pan. 2025. Can llms evaluate complex attribution in qa? automatic benchmarking using knowledge graphs. In *The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S Yu, and Zhijiang Guo. 2023. Do large language models know about facts? *arXiv preprint arXiv:2310.05177*.

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*, pages 1394–1399.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, and 1 others. 2023. Realtime qa: What's the answer right now? *Advances in neural information processing systems*, 36:49025–49043.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288.

Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using rag and few-shot in-context learning with llms. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 91–98.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference. In *Findings*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and 1 others. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720.

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.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Feifan Wu, Lingyuan Liu, Wentao He, Ziqi Liu, Zhiqiang Zhang, Haofen Wang, and Meng Wang. 2024. Time-sensitve retrieval-augmented generation for question answering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2544–2553.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024a. 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.

Siheng Xiong, Ali Payani, Ramana Rao Kompella, and Faramarz Fekri. 2024b. Large language models can learn temporal reasoning. *ArXiv*, abs/2401.06853.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat seng Chua. 2023. Search-in-the-chain: Towards the accurate, credible and traceable content generation for complex knowledge-intensive tasks. *ArXiv*, abs/2304.14732.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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.

Zhiyuan Zhu, Yusheng Liao, Zhe Chen, Yuhao Wang, Yunfeng Guan, Yanfeng Wang, and Yu Wang. 2025. 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.

A Entropy Dynamics around Factual Change Points

To complement the preliminary study in Section 3, we further analyze how model uncertainty behaves

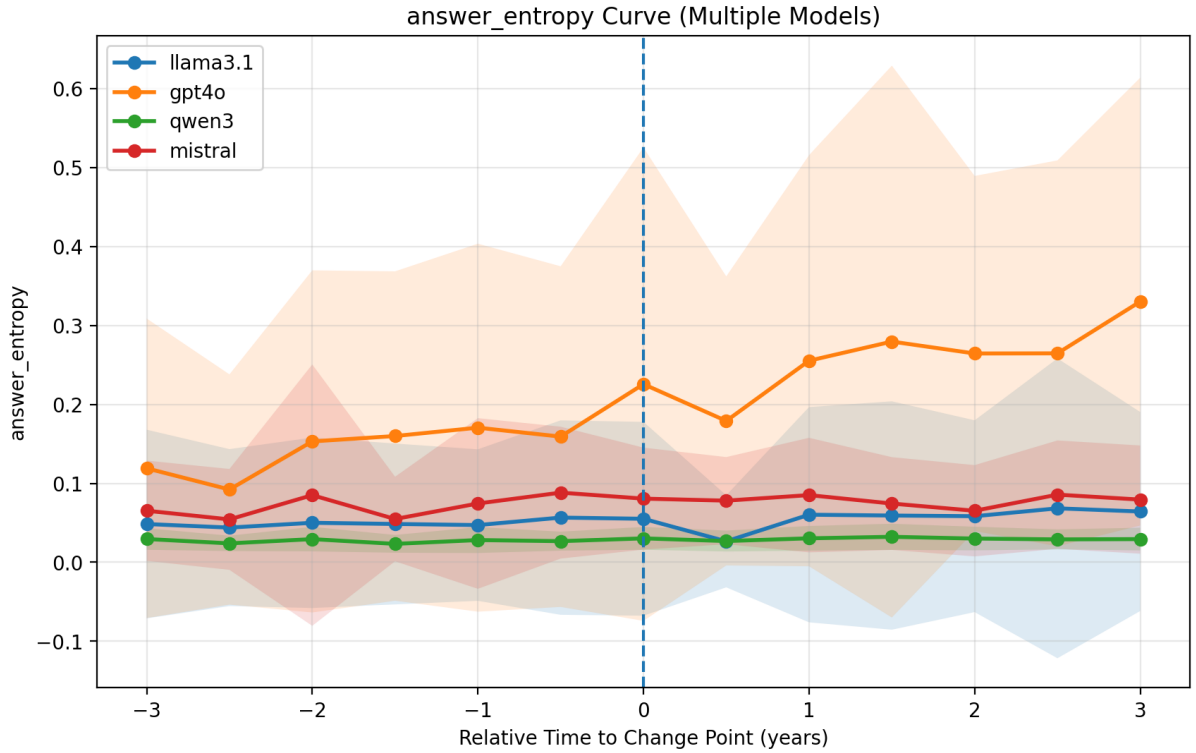

Figure 3: Average answer entropy of different models around factual change points.

around factual change points. Following the same controlled setting, we probe each model at multiple time offsets before and after the annotated change point and compute token-level answer entropy as a measure of predictive uncertainty (Appendix D.1).

Figure 3 shows the average answer entropy trajectories of four models aligned by their relative distance to the change point. Consistent with the findings in Section 3, entropy does not exhibit a clear increase at or immediately after the change boundary. For LLaMA-3.1, Qwen3, and Mistral, uncertainty remains largely flat across pre- and post-change regions, indicating that these models persist with confident predictions even when the underlying fact has changed. GPT-4o shows higher overall entropy and a mild upward trend after the change point, but the increase is gradual rather than localized, suggesting sensitivity to temporal distance rather than explicit change awareness.

These results reinforce the conclusion of Section 3: current LLMs lack an internal uncertainty signal that reliably reflects factual transitions. As a consequence, models remain overconfident precisely where increased uncertainty is warranted, motivating the need for an explicit confidence-aware mechanism to regulate future commitment, as implemented in CTR.

B Algorithm of CTR inference framework

This appendix provides a detailed description of the CTR inference framework, complementing the high-level method overview in the main paper. Algorithm 1 formalizes the end-to-end inference procedure, including historical recall with uncertainty estimation, timeline reconstruction, change-point detection, temporal pattern induction, and future inference. The algorithm is presented to clarify implementation details and intermediate signals used by CTR, without introducing additional modeling assumptions beyond those discussed in the main text.

Furthermore, the historical probing step (Lines 2-5 in Algorithm 1) is fully parallelizable, allowing batch processing of temporal queries to minimize wall-clock latency.

C Case study: Confidence-Aware Repair in CTR

Figure 4 illustrates a representative case study for the query "What was the shortest war in history?", highlighting how CTR preserves prediction reliability under noisy historical recall.

In early years, historical recall exhibits substan-

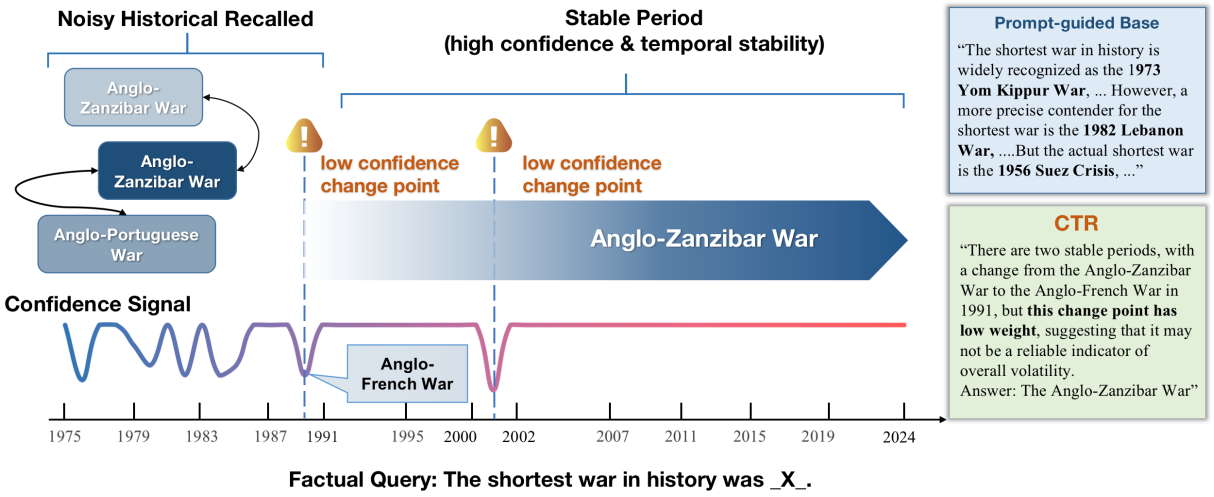

Figure 4: An illustrative case showing how CTR suppresses low-confidence change points and commits to future predictions only when long-term temporal stability is observed.

tial fragmentation, alternating between multiple candidates such as the Anglo-Zanzibar War and the Anglo-Portuguese War. Although some recalled answers appear with high instantaneous confidence, the overall confidence signal fluctuates sharply.

Around 1991, a deviation to the Anglo-French War is detected as a potential change point. However, this change point is assigned low weight due to its isolated nature and the absence of sustained support in subsequent years. CTR therefore treats it as a spurious fluctuation rather than genuine factual evolution.

From 1992 onward, the Anglo-Zanzibar War forms a long, stable period characterized by consistently high confidence and low variance, establishing a reliable temporal anchor. Based on this stability, CTR commits to the Anglo-Zanzibar War as the future answer.

In contrast, the prompt-guided baseline produces a fluent but hallucinated reasoning chain by enumerating multiple unrelated conflicts without temporal grounding. This comparison demonstrates that CTR's confidence-aware mechanism enables conservative future commitment by suppressing unreliable change signals and prioritizing long-term temporal stability.

D Detailed Evaluation Protocol

D.1 Answer Correctness Criteria

Strict Correctness. As formally defined in Section 3, we primarily use **Strict Correctness** to evaluate whether the model identifies the exact factual state. This serves as our main accuracy metric.

Lenient Post-change Correctness. For future-oriented queries ($k > 0$), strict evaluation may penalize reasonable caution. Therefore, we additionally introduce a **Lenient Criterion** that accepts explicit expressions of uncertainty (e.g., outputs containing "Unknown", "Uncertain", or "TBD") as valid responses for volatile facts:

$$\text{Correct}_{i,k}^{\text{lenient}} = \begin{cases} \text{Correct}_{i,k}^{\text{strict}}, & k < 0, \\ \mathbb{I}\big(\hat{a}_{i,k} \in \{a_1^i, \textsf{Unknown}\}\big), & k > 0. \end{cases} \tag{18}$$

This metric distinguishes between correctable epistemic uncertainty and confident hallucinations.

D.2 Uncertainty Estimation

We quantify model confidence using token-level entropy derived from the decoding process. Let $T_{i,k}$ denote the set of token indices corresponding to the generated answer $\hat{a}_{i,k}$. The average answer entropy is defined as:

$$H_{i,k} = \frac{1}{|T_{i,k}|} \sum_{j \in T_{i,k}} \mathcal{H}(p_\theta(\cdot \mid x_{<j}, x_{i,k})), \tag{19}$$

where p_θ is the model's predictive distribution and $\mathcal{H}(\cdot)$ denotes the Shannon entropy.

To measure *temporal sensitivity*, we analyze the **Entropy Shift** (ΔH) across the change boundary. Specifically, we compute the difference between the average entropy of post-change predictions and pre-change predictions:

$$\Delta H_i = \bar{H}_{i,k>0} - \bar{H}_{i,k<0}. \tag{20}$$

Algorithm 1 CTR Inference Framework

Input: Query q, future time t^*, window size W
Output: Reasoning trace r, evolution type z, answer $\hat{a}_{t^*}$

Historical Recall
for $y = t^* - W$ **to** $t^* - 1$ **do**
  Query (q, y) to obtain a_y, entropy H_y
  $c_y \leftarrow f(H_y)$
end for
Timeline Reconstruction
for each year y **do**
  $v_y \leftarrow a_y$
  $q_y \leftarrow \textsc{ProbeQuality}(c_y)$
end for
$\mathcal{T} \leftarrow \{(y, v_y, q_y)\}$
Change-Point Detection
for each candidate y with $v_y \neq v_{y-1}$ **do**
  $w_y \leftarrow \sqrt{q_{y-1} q_y} \cdot g_y$
end for
$\mathcal{C} \leftarrow \{(y, w_y)\}$
Temporal Pattern Induction
$\pi \leftarrow \langle z, \text{summary}, \text{key\_evidence} \rangle$
Future Inference
$\hat{a}_{t^*} \leftarrow \arg\max_a p(a \mid q, t^*, \pi)$
Construct aggregated reasoning trace r

A positive ΔH indicates that the model correctly recognizes increased uncertainty when extrapolating into the future. Conversely, a $\Delta H \approx 0$ suggests that temporal evolution is not explicitly represented in the model's reasoning process.

E Reasoning Hallucination Judgement

To ensure the reliability of our hallucination evaluation, we employ a rigorous two-stage verification pipeline, combining broad automated screening with fine-grained human adjudication.

Stage 1: Strict Automated Screening. First, we utilize GPT-5.1 as an independent judge to perform an initial screening of all reasoning traces. The judge is instructed to flag instances based on three strict criteria:

1. **Fabrication:** Citing specific events, names, or data points that do not exist (e.g., inventing a "2025 release date" for a cancelled movie).

2. **Unsupported Extrapolation:** Making definitive claims about volatile future states without citing stable historical patterns or official announcements.

3. **Logical Contradiction:** Deriving a conclusion that contradicts the cited historical evidence within the trace.

This stage acts as a high-recall filter to identify potential hallucinations ($\text{Hallucination}_{i,k} = 1$).

Stage 2: Expert Human Adjudication. To eliminate false positives from the automated judge and ensure gold-standard quality, we conduct a secondary human review. This process is performed by expert annotators familiar with the temporal reasoning task. The human review protocol involves:

- **Fact Verification via Search:** For every flagged instance, annotators are required to verify the validity of the generated claims against external search engines (e.g., Google Search) to confirm whether the "hallucinated" detail is indeed false or merely obscure.

- **Logic and Grounding Check:** Annotators assess whether the reasoning logically follows from the provided context. A trace is marked as valid only if the prediction is grounded in verifiable patterns (e.g., "consistent 4-year cycle") rather than spurious correlations.

- **Final Adjudication:** In cases of disagreement between the automated judge and human assessment (e.g., the model correctly inferred a complex fact that GPT-5.1 mistook for hallucination), the human judgment prevails.

This human-in-the-loop refinement ensures that our reported hallucination rates reflect genuine generative failures rather than evaluation artifacts.

We emphasize that hallucination rates are used for controlled relative comparison across methods under identical judging conditions, rather than as absolute measures of factuality.

F Dataset Examples and Annotations

Table 4 presents representative examples from our evaluation dataset. The structure strictly follows the formal definition provided in Section 3, where each instance is annotated as a tuple $e_i = \langle q_i, t_0^i, a_0^i, t_c^i, a_1^i \rangle$.

Specifically, the table columns correspond to:

- **Fact Type**: The evolution category.

Fact Type	Query (q_i)	Ref. Time (t_0^i)	Current Answer (a_0^i)	Next Answer (a_1^i)
Never	_X_ won the 2020 Formula 1 world driver's championship.	2022	Lewis Hamilton	N/A
Never	The real name of the Unabomber was _X_.	2022	Theodore John Kaczynski	N/A
Slow	The Winter Olympics were held most recently in _X_.	2022	Beijing, China	Italy
Slow	The next leap year is _X_.	2024	2028	2032
Fast	The most recent movie in the Marvel Cinematic Universe is _X_.	2025	The Fantastic Four: First Steps	Spider-Man: Brand New Day
Fast	King Gizzard's most recent studio album is _X_.	2025	Phantom Island	N/A

Table 4: **Annotated examples from the evaluation dataset.** The columns align with the formal definition in Eq. (1). For diagnostic purposes, a_1^i (Next Answer) serves as the ground truth for evaluating future reasoning capabilities but is not provided to the model during inference.

- **Query (q_i)**: The cloze-style factual query.
- **Ref. Time (t_0^i)**: The latest reference time at which the fact is known to be valid (corresponding to the `effective_year` in the raw data).
- **Current Answer (a_0^i)**: The correct answer valid at time t_0^i.
- **Next Answer (a_1^i)**: The verifiable future answer after the change point. This is set to `N/A` for *Never-changing* facts or *Fast-changing* facts where the future state is highly volatile or yet to be determined.