NTU Journal of Engineering and Technology (2026) 5 (3) : 153-160
DOI: https://doi.org/10.56286/ntujet.v5i3

IRAQI Academic Scientific Journals

P-ISSN: 2788-9971 E-ISSN: 2788-998X
NTU Journal of Engineering and Technology
Available online at: https://journals.ntu.edu.iq/index.php/NTU-JET/index



Temporal Reliability Decay in Large Language Models: A Domain-Stratified Empirical Study

Harith Hamoodat[1]

[1]Department of Information Technology Management, Technical College of Management / Mosul, Northern Technical University, Mosul, 41001, Iraq.
Email: Hhamoodat@ntu.edu.iq

Article Informations

Received: 18-05- 2026,
Reviesed: 24-07-2027
Accepted: 04-08-2026,
Published online: 30-09-2026

Corresponding author:
Name: Harith Hamoodat
Affiliation : Northern Technical University
Email: Hhamoodat@ntu.edu.iq

Key Words:
Large Language Models,
Temporal Reasoning,
Knowledge Cutoff,
Factual Reliability,
Temporal Reliability Decay,
Confidence Estimation,
Domain Adaptation,
Question Answering

A B S T R A C T

Large Language Models (LLMs) are commonly evaluated against a fixed knowledge cutoff, but this view does not explain what happens to factual accuracy as information gradually becomes more distant from the model's training period. In practice, not all facts age in the same way. A country capital or a scientific term may remain stable for years, while an office holder, sports result, or political appointment may change within a short time. This paper examines this problem by introducing Temporal Reliability Decay (TRD), an empirical framework for measuring the change in factual accuracy as a function of temporal distance from the model knowledge cutoff. The study evaluates GPT-3.5-turbo, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct using 2,400 questions selected from TempLAMA and TimeQA. The questions were grouped into Science, Geography, Sports, and Politics and evaluated across six temporal bins. The results show a clear difference between relatively stable and fast-changing domains. Science and Geography show slow decline, while Sports and Politics lose accuracy more rapidly as temporal distance increases. Based on these patterns, the paper proposes the Temporal Confidence Score (TCS), a domain-aware estimate that can help indicate when an LLM response should be checked, supported with retrieval, or reviewed by a human. The findings are limited to English question answering, four domains, three evaluated models, and zero-shot prompting.

THIS IS AN OPEN ACCESS ARTICLE UNDER THE CC BY LICENSE:
https://creativecommons.org/licenses/by/4.0/

1. Introduction

Large Language Models (LLMs) are now used in question answering, search support, summarization, and decision-support systems. Their ability to answer factual questions depends largely on information observed during pre-training. Since this training data is collected before a particular cutoff date, the model may not reflect later changes in the world. This limitation is usually described as a knowledge cutoff problem.

A fixed cutoff date does not affect all questions in the same way. Some answers remain stable long after the training period, while others change quickly. For instance, a geographic name may remain valid for many years, but a ministerial position, a sports ranking, or a recent appointment may change within a short interval. This makes the usual "known" versus "unknown" view too simple for factual use of LLMs.

The issue becomes clearer in applied settings. A user asking about a scientific concept does not face the same risk as a user asking about a recent election result or a current sports record. Standard benchmarks usually give an overall accuracy score, but they provide less detail about how accuracy changes when the fact becomes older relative to the model cutoff. They also give limited attention to differences between domains.

To examine this point, the paper proposes TRD. The framework treats factual correctness as a changing quantity rather than a fixed cutoff outcome. The empirical part of the study uses 2,400 factual questions from TempLAMA and TimeQA, grouped into Science, Geography, Sports, and Politics. The same zero-shot setting was used for the three evaluated models.

The paper further uses the Temporal Confidence Score (TCS) to connect domain type with temporal distance. The score is intended as a practical warning measure. When the estimated reliability is low, the answer can be sent for retrieval support, external checking, or human review.

The main contributions of this paper are summarized as follows:

• We define Temporal Reliability Decay (TRD) as a measurable framework for modeling factual accuracy as a continuous function of temporal distance from the model knowledge cutoff.

• We provide a domain-stratified empirical evaluation using 2,400 temporally annotated questions across four knowledge domains and three publicly available LLMs.

• We estimate domain-specific decay parameters and show that temporal reliability decreases at substantially different rates across stable and dynamic domains.

• We propose TCS, a lightweight reliability estimate that combines domain type and temporal distance to support retrieval triggering, external verification, or human review.

2. Related Work

2.1 Temporal Reasoning in Large Language Models

Evaluating the factual accuracy of LLMs has become a significant line of investigation in temporal reasoning. Dhingra et al. [1] propose TempLAMA to explore the feasibility of language models as temporal knowledge bases. Chen et al. proposed TimeQA as a benchmark for answering questions about temporal knowledge in [3]. The results of these works indicate that factual correctness is often not just about entity recognition but also about temporal validity.

Later studies extended temporal evaluation beyond the original temporal question answering (QA) datasets. TimeBench covers several types of temporal reasoning, while TRAM examines tasks such as event order, duration, frequency, and temporal arithmetic [11, 12]. Other recent work has studied whether LLMs can explain temporal answers in a reliable way, and whether temporal reasoning differs across languages and task types [13, 16]. These studies are useful because they show that temporal reasoning is still difficult for LLMs. However, they mainly evaluate temporal reasoning ability itself. The present study focuses on a narrower question: how factual accuracy changes as the queried fact becomes farther from the model knowledge cutoff, and whether this change differs across knowledge domains.

2.2 Knowledge Cutoff and Temporal Misalignment

LLMs are trained on corpora gathered prior to a certain cutoff date, causing a temporal misalignment between model knowledge and the evolving real world. Lazaridou et al. [4] investigated temporal generalization gaps in neural language models, whereas Vu et al. [2] demonstrated that search enhancement can yield better results for up-to-date time-sensitive queries. This study follows this line of research but treats temporal reliability as a gradual decay process rather than a binary pre-cutoff/post-cutoff state.

2.3 Factuality, Hallucination, and Reliability Estimation

Factual reliability has a close relationship with hallucination, as LLMs produce syntactically well-formed but factually false or unsupported information. Recent surveys highlight that hallucination is still a significant hindrance for LLM application, particularly in areas that require factual accuracy [5, 6]. Confidence estimation and calibration have also gained significance, as confidence scores can be used as indicators of whether a model output is likely to be correct [7, 8]. However, most calibration methods do not explicitly account for temporal distance from the model knowledge cutoff. By contrast, our TCS is a combination of the domain type and temporal distance.

2.4 Retrieval-Augmented Generation and Knowledge Updating

Retrieval-Augmented Generation (RAG) has been widely used to reduce factual errors by allowing language models to condition their responses on external documents. Lewis et al. [9] introduced RAG for knowledge-intensive NLP tasks, while Guu et al. [10] proposed retrieval-augmented pretraining through REALM. Recent work shows that retrieval can improve freshness and factual grounding in LLM responses [2]. However, retrieval is not always necessary for stable domains. Therefore, a domain-aware reliability score can help determine when retrieval should be triggered.

Recent work on LLM factuality and retrieval-augmented generation further confirms that factual reliability remains a central challenge for real-world deployment. Recent factuality surveys identify factual inconsistency, unsupported generation, and evaluation difficulty as persistent problems in LLM outputs [14]. RAG has been widely proposed as a practical solution for improving freshness and grounding, because it allows models to incorporate up-to-date external information instead of relying only on static parametric knowledge [15]. However, applying retrieval to every query may be inefficient and unnecessary, especially in relatively stable domains. This motivates the need for a temporal and domain-aware reliability estimate that can help determine when retrieval or human verification should be triggered.

2.5 Domain-Specific Knowledge Dynamics

Knowledge does not evolve uniformly across domains. Scientific and geographic knowledge often remains stable over longer periods, whereas politics and sports change rapidly due to elections, appointments, competitions, and updated records. This motivates a domain-stratified evaluation rather than relying on a single global reliability score. Our study estimates separate temporal decay parameters for science, geography, sports, and politics, allowing the reliability to be modeled more accurately.

2.6 Positioning of This Work

This study is related to temporal question answering, factuality evaluation, confidence estimation, and retrieval-augmented generation. Temporal QA benchmarks test whether a model can answer time-sensitive questions correctly. Factuality studies examine whether model outputs are supported by real-world facts. Confidence estimation methods try to estimate how likely an answer is to be correct. Retrieval-augmented generation can also improve freshness by adding external information at inference time.

However, these directions do not directly answer the main question of this paper. They do not estimate how quickly factual reliability decreases as temporal distance from the model knowledge cutoff increases, and they usually do not assign separate decay rates to different domains. This is the gap addressed by TRD. Instead of treating each answer only as correct or incorrect, TRD models reliability as a gradual temporal pattern.

TCS is also different from ordinary confidence scores. It is not calculated from token probabilities, and it does not rely on the model's own confidence statement. Instead, it uses two pieces of information: the estimated decay rate of the domain and the temporal distance of the queried fact. In this way, TCS can serve as a practical signal for deciding when retrieval, external checking, or human review may be needed.

3. Methodology

3.1 Overview

This study measures how factuality in LLMs is affected by temporal distance from the training cutoff. The evaluation process involves four stages: (1) building the dataset with temporal annotations, (2) domain classification, (3) querying the model under uniform zero-shot scenarios, and (4) aggregating the accuracy and modeling temporal decay.

3.2 Dataset Construction

A temporally stratified dataset was prepared from two temporal question answering resources: TempLAMA [1] and TimeQA [3]. The combined pool contained approximately 15,000 candidate questions. The aim of the selection process was to keep questions that had three required elements: a factual answer, a usable timestamp, and enough information to compute temporal distance from the model knowledge cutoff.

The filtering was done in stages. First, questions with no clear answer or no usable timestamp were removed. Second, the merged pool was checked for repeated or nearly repeated questions, and only one

version of each repeated case was kept. Third, questions were removed when their wording allowed more than one interpretation, when more than one answer could be considered correct, or when the temporal reference was too weak to support the calculation of temporal distance. Finally, the remaining questions were checked for domain assignment. Only questions that could be assigned to Science, Geography, Sports, or Politics were retained. After filtering and balancing, the final evaluation set contained 2,400 questions. To improve transparency, Table 1 summarizes the dataset selection workflow.

Table 1. Dataset selection workflow

Stage	Remaining questions
Initial candidate questions from TempLAMA and TimeQA	~15,000
After removing questions without verifiable answers or timestamps	10,864
After removing duplicate or near-duplicate questions	8,192
After removing ambiguous or temporally unclear questions	4,736
After domain filtering and balancing	2,400

The final dataset was balanced across four domains and six temporal bins. The temporal bins were $\Delta t \in \{0, 3, 6, 12, 18, 24\}$ months. Each domain contained 100 questions per temporal bin, resulting in 600 questions per domain and 2,400 questions in total.

Domain labels were assigned by two independent annotators using predefined annotation guidelines. Science included questions related to scientific facts, discoveries, terminology, and research-related entities. Geography included countries, cities, locations, borders, capitals, and geographic entities. Sports included teams, athletes, competitions, results, and sports records. Politics included elections, office holders, government positions, political parties, and public officials. Questions that overlapped between domains were discussed during adjudication and assigned to the domain most directly related to the expected answer.

Inter-annotator agreement was calculated using Cohen's kappa. The overall agreement was $\kappa = 0.87$, indicating strong agreement. Per-domain agreement was also computed as follows: Science $\kappa = 0.86$, Geography $\kappa = 0.88$, Sports $\kappa = 0.89$, and Politics $\kappa = 0.85$. Disagreements were resolved through discussion between the annotators, and unresolved cases were excluded from the final dataset.

3.3 Temporal Distance Computation

For each question q, the temporal distance Δt is defined as:

$$\Delta t = \frac{|T_{cutoff} - T_{fact}|}{30} \quad (1)$$

where Tcutoff is the model training cutoff date and Tfact is the timestamp associated with the ground-truth answer. The result is expressed in months. The division by 30 was used to express temporal distance in approximate monthly units. This approximation was adopted to ensure consistent binning across all questions and models, since the evaluation was organized around monthly temporal intervals rather than exact calendar-month lengths. Therefore, Δt should be interpreted as an approximate month-based distance rather than a calendar-exact measurement. This approximation is also acknowledged as a limitation of the study.

3.4 Models Evaluated

The evaluation benchmarked three publicly available LLMs: GPT-3.5-turbo, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct. To improve reproducibility, the evaluated model identifiers, providers, access modes, including Application Programming Interface (API)-based access, and assumed knowledge cutoff dates are reported in Table 2. These cutoff dates were used to compute the temporal distance between each question and the model's training boundary. When exact cutoff dates were not explicitly documented by the model provider, the reported cutoff date was treated as an approximate assumption based on publicly available release information.

Table 2. Evaluated models and assumed knowledge cutoff dates

Model	Version and access mode	Assumed cutoff
GPT-3.5	GPT-3.5-turbo, OpenAI API	Sep. 2021
Llama-3.1	Llama-3.1-8B-Instruct, Hugging Face	Dec. 2023
Mistral-7B	Mistral-7B-Instruct, Hugging Face	Sep. 2023, approximate

The cutoff dates reported in Table 2 were used to compute the temporal distance between each question and the corresponding model training boundary. For GPT-3.5-turbo and Llama-3.1-8B-Instruct, the cutoff dates were based on available model documentation. For Mistral-7B-Instruct, the cutoff date was treated as approximate because the exact training cutoff is not explicitly documented. All models were evaluated under the same zero-shot prompting condition, without retrieval augmentation, fine-tuning, or external search. Therefore, the reported results reflect the models' parametric factual knowledge under controlled prompting conditions.

3.5 Evaluation Protocol

For each question, the standardized prompt is used: "Answer the following factual question with a single concise answer: [QUESTION]". Model answers are assessed in terms of exact-match accuracy after normalization, which includes lowercasing, removing punctuation and alias matching in case of multiple equivalent entity names. The accuracy for a domain d and time bin t is calculated as:

$$Acc(d,t) = \frac{1}{N_{d,t}} \sum_{i=1}^{N_{d,t}} I(\hat{y}_i = y_i) \quad (2)$$

where ŷi is the predicted answer, yi is the ground-truth answer.

3.6 Temporal Reliability Decay Modeling

We modeled the decay of factual accuracy using an exponential function because it provides a simple, monotonic, and interpretable representation of gradual reliability loss over time. In this formulation, reliability decreases as temporal distance from the model knowledge cutoff increases, while the decay rate is allowed to vary by domain.

$$\mathrm{Acc}(d,t) = \mathrm{BaseAcc}(d)e^{-\lambda_d t} \quad (3)$$

where BaseAcc(d) represents the accuracy at Δt = 0, and λd is the domain-specific decay rate. The decay parameter λd is estimated using least-squares regression:

$$\lambda_d = \arg\min_{\lambda} \sum_t \left(\mathrm{Acc}(d,t) - \mathrm{BaseAcc}(d)e^{-\lambda t}\right)^2 (4)$$

To avoid assuming the exponential form without validation, we also compared it with alternative monotonic decay models, including linear, logarithmic, and power-law decay forms. These candidate models were compared using Root Mean Square Error (RMSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). The exponential model was retained because it provided the best overall balance between fit quality, parsimony, monotonic behavior, and interpretability across the evaluated domains.

3.7 Temporal Confidence Score

Based on the decay model, we define TCS for a query q as:

$$\mathrm{TCS}(q) = \mathrm{BaseAcc}(d)e^{-\lambda_d \Delta t} \quad (5)$$

To allow comparison across domains, we define a normalized variant:

$$\mathrm{TCS}_{(\mathrm{norm})(q)} = \frac{\mathrm{TCS}(q)}{\mathrm{BaseAcc}(d)} \quad (6)$$

3.8 Statistical Analysis

To confirm the reliability of the reported accuracy values, we calculate 95% confidence intervals (CIs) using the normal approximation:

$$\mathrm{CI} = \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (7)$$

where p̂ is the observed accuracy and n is the number of samples. Pairwise comparisons were also conducted between domains to determine whether decay differences were significant.

3.9 Reproducibility

The evaluation was carried out as a batch experiment, not as an interactive dialogue. Each question was submitted separately using the same prompt: "Answer the following factual question with one concise answer: [QUESTION]". The placeholder [QUESTION] was replaced with the actual question from the dataset. No previous question, answer, or feedback was included in the next request.

The decoding settings were fixed for all models. Temperature was set to 0, top-p to 1.0, and the maximum output length to 32 tokens. No stop sequences were used. For Llama-3.1-8B-Instruct and Mistral-7B-Instruct, the Hugging Face random seed was set to 42 when supported by the inference environment. For GPT-3.5-turbo through the OpenAI API, deterministic output was encouraged using temperature = 0, while seed control was treated as provider-dependent. The experiments were run from 12 April 2026 to 21 April 2026.

The model outputs were saved as generated. During evaluation, only limited normalization was applied. This included lowercasing, trimming extra spaces, removing punctuation when needed, and matching common aliases for equivalent entity names. No model answer was manually corrected. The model identifiers and access platforms were reported in Section 3.4 to support reproducibility.

4. Results and Discussion

4.1 Domain-Stratified Accuracy and Decay Patterns

Figure 1 shows the average temporal reliability decay curves for each domain, averaged across the three evaluated models. The results show that factual accuracy drops as the temporal distance to the training cutoff increases, though the rate of decay differs widely across domains.

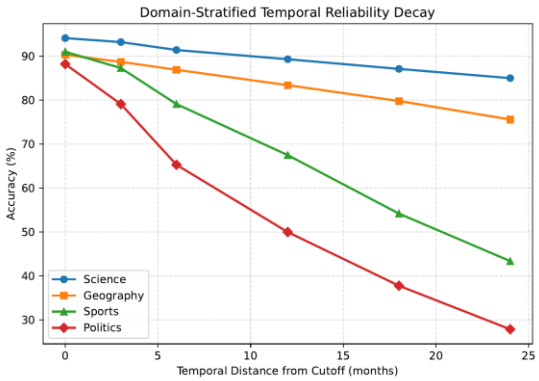


Fig. 1. Domain-stratified temporal reliability decay across six temporal bins.

The numerical accuracy values are reported in Table 3. Science decreases from 94.1% at Δt = 0 to 85.0% at Δt = 24 months, corresponding to a modest decline of 9.1 percentage points. Geography shows a gradual decline from 90.3% to 75.6%. In contrast, Politics drops from 88.2% to 27.9%, while Sports declines from 91.0% to 43.4%.

Table 3. Average accuracy (%) across domains and temporal distances with 95% confidence intervals

Domain	0	3	6	12	18	24
Science	94.1 ± 2.7	93.2 ± 2.8	91.4 ± 3.2	89.3 ± 3.5	87.1 ± 3.8	85.0 ± 4.0
Geography	90.3 ± 3.3	88.7 ± 3.6	86.9 ± 3.8	83.4 ± 4.2	79.8 ± 4.5	75.6 ± 4.9
Sports	91.0 ± 3.2	87.3 ± 3.8	79.1 ± 4.6	67.5 ± 5.3	54.2 ± 5.6	43.4 ± 5.6
Politics	88.2 ± 3.7	79.1 ± 4.6	65.3 ± 5.4	50.0 ± 5.7	37.8 ± 5.5	27.9 ± 5.1

The confidence intervals were calculated using the normal approximation with (n = 300) responses per domain–time-bin condition, corresponding to 100 questions evaluated across the three models. These findings support the assumption that temporal reliability is strongly influenced by domain-specific knowledge dynamics. Fast-decaying domains are characterized by frequent updates and short factual validity periods.

4.2 Exponential Model Validation and Decay Rate Analysis

Figure 2 shows empirical accuracy points along with the fitted exponential. The observed and fitted values are sufficiently close to suggest that the exponential model is a good approximation for the temporal reliability decay.

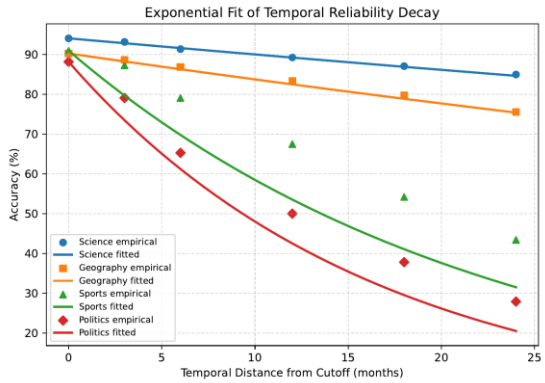


Fig. 2. Empirical accuracy and fitted exponential decay curves across domains.

The fit of the proposed decay model was then evaluated by computing the coefficient of determination (R^2) for each domain. All domains achieve $R^2 > 0.96$, which suggests the exponential model fits the observed data well.

In addition to R^2, we tested whether the exponential curve was the most suitable form for the observed pattern. The same domain–time-bin accuracy values were fitted using linear, logarithmic, and power-law alternatives. The comparison was based on RMSE, AIC, and BIC, averaged across the four domains. The exponential model produced the best overall fit, with RMSE = 1.025, AIC = −2.938, and BIC = −3.146. The linear model was close in some stable domains, but its overall fit was weaker, with RMSE = 1.398, AIC = −2.643, and BIC = −2.851. The logarithmic and power-law forms showed larger average errors, with RMSE values of 4.924 and 5.717. Based on this comparison, the exponential form was retained as the main decay model.

Figures were regenerated at 300 dots per inch (DPI) or higher. Figure 2 presents the fitted exponential curves, Figure 4 includes a color-bar legend, and the 95% confidence intervals are reported numerically in Table 3.

The estimated decay parameters λd for each domain are listed in Table 4. The results establish a clear ranking: Science < Geography < Sports < Politics.

Table 4. Estimated domain-specific decay rates (λd) across models

Domain	GPT-3.5	Llama-3.1	Mistral-7B	Mean ± SD	Median
Science	0.0039	0.0042	0.0044	0.0042 ± 0.0003	0.0042
Geography	0.0067	0.0071	0.0074	0.0071 ± 0.0004	0.0071
Sports	0.0391	0.0423	0.0441	0.0418 ± 0.0025	0.0423
Politics	0.0547	0.0581	0.0603	0.0577 ± 0.0028	0.0581

The corrected mean decay rates were λ = 0.0042, 0.0071, 0.0418, and 0.0577 for Science, Geography, Sports, and Politics, respectively. This ranking confirms that Science and Geography exhibit slower temporal decay, whereas Sports and Politics show substantially faster reliability degradation.

4.3 Temporal Confidence Score and Heatmap Analysis

Figure 3 presents the TCS curves obtained from the estimated decay rates. The curves remain relatively high for Science and Geography, while they fall more quickly for Sports and Politics. This pattern is consistent with the accuracy results reported earlier.

The TCS values were then compared with the observed accuracy values across the domain–time-bin pairs. The comparison was close: Pearson's correlation was 0.98, Spearman's correlation was 0.97, and the mean absolute error was 1.08 percentage points. These results suggest that TCS follows the average accuracy trend in the evaluated data. It should still be used carefully, because it is not a probability score for one individual answer. It is better understood as a domain-level signal for deciding when extra checking is needed.

The curves in this section are based on averaged results from the three models and 100 questions in each domain–time-bin condition. No manual smoothing was added before fitting the decay curves. The smooth trend mainly reflects the aggregation of balanced samples.

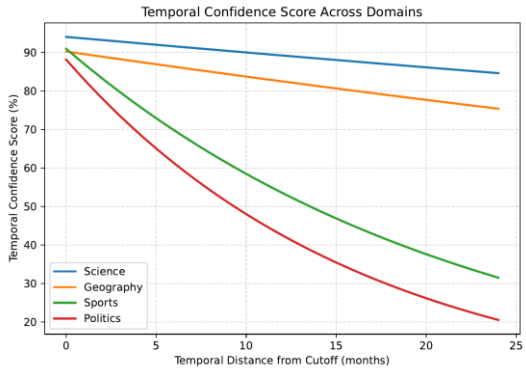


Fig. 3. TCS curves by domain.

Figure 4 gives a compact view of the same accuracy pattern. The darker cells appear mainly in Sports and Politics after longer temporal distances, which is consistent with the faster decline shown in Figure 1.

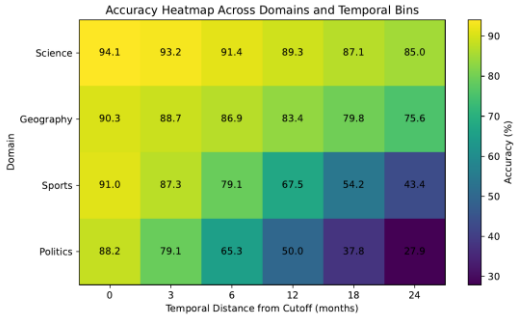


Fig. 4. Accuracy heatmap across domains and temporal bins. The color bar indicates accuracy percentage.

4.4 Deployment Implications, Statistical Significance, and Error Analysis

The domain differences are important for systems that answer time-sensitive questions. In Politics and Sports, a model may still give a fluent answer even when the fact has already changed. This is why questions in these domains need more checking when they are far from the cutoff date.

Science and Geography showed a slower decline in this dataset. This does not mean that every answer in these domains is correct, but it suggests that many of their facts remained valid for a longer period. A score such as TCS can help decide when extra checking is necessary instead of applying the same rule to all domains.

We computed 95% confidence intervals for all reported accuracy values using the normal approximation. In addition, we tested whether the estimated domain-specific decay rates differed significantly across domains. A one-way analysis of variance (ANOVA) conducted on the model-level decay estimates showed a significant difference among domains, ($F = 571.24$), ($p < 0.001$). Pairwise Welch's t-tests further confirmed significant differences between slow-decay and fast-decay domains. For example, Science differed significantly from Sports ($p = 0.0014$) and Politics ($p = 0.0009$), while Geography also differed significantly from Sports ($p = 0.0015$) and Politics ($p = 0.0009$). These results support the validity of domain-stratified temporal reliability modeling.

These results also show why a single confidence estimate may not be enough for LLM outputs. The same temporal distance can have different effects depending on the domain. A six-month difference may be less important for a stable scientific fact, but more important for a sports ranking, election result, or office holder.

The qualitative errors were mainly concentrated in the fast-changing domains. In Politics and Sports, several incorrect answers referred to information that had already changed, such as earlier office holders, previous winners, or older rankings. Another group of errors involved answers that belonged to the right category but pointed to the wrong entity; for example, the model sometimes selected a person, team, or institution that was related to the question but was not the correct answer for the required time. Some answers were only partly correct because they matched the general topic but missed the specific year, role, or event context. These examples show that many failures were not random mistakes. They were linked to the gap between what the model had stored during training and facts that had changed later.

5. Conclusion

This study examined the relation between temporal distance and factual accuracy in LLMs. The results indicate that the effect of time differs by domain. Science and Geography kept higher accuracy across the evaluated temporal bins, while Sports and Politics declined more quickly. This means that the knowledge cutoff should not be treated as one uniform boundary for every type of factual question.

TRD was used to describe this difference through domain-specific decay rates. TCS then used these rates to give a simple indication of when an answer may need checking, retrieval support, or human review. The main value of this approach is that it allows stable and fast-changing domains to be handled differently.

The findings should be read within the limits of the experiment. The dataset was English-only and covered four domains. The evaluation used three models under a zero-shot setting, without fine-tuning or retrieval-augmented generation. Exact-match accuracy was the main evaluation measure, so some partially correct answers may not be fully represented. In addition, some model cutoff dates were treated as approximate when exact documentation was not available.

Future work can test the same framework on more models, more languages, additional domains, and retrieval-based systems. It can also examine whether finer domain categories provide more accurate reliability estimates.

Acknowledgments

The author would like to thank Northern Technical University for its academic support.

Author Contributions

Harith Hamoodat conceptualized the study, designed the methodology, performed data curation, conducted the analysis, prepared the figures and tables, interpreted the results, and wrote and revised the manuscript.

Competing Interests

The author declares no competing interests.

Data Availability Statement

The datasets used in this study were constructed from publicly available temporal question answering benchmarks, including TempLAMA and TimeQA. The processed dataset, filtering criteria, prompt template, and evaluation scripts can be made available by the author upon reasonable request, subject to the licensing conditions of the original benchmark datasets.

References

[1] B. Dhingra, J. R. Cole, J. M. Eisenschlos, D. Gillick, J. Eisenstein, and W. W. Cohen, "Time-Aware Language Models as Temporal Knowledge Bases," Transactions of the Association for Computational Linguistics, vol. 10, pp. 257-273, 2022. https://doi.org/10.1162/tacl_a_00459

[2] T. Vu, M. Iyyer, X. Wang, N. Constant, J. W. Wei, J. Wei, C. Tar, Y.-H. Sung, D. Zhou, Q. V. Le, and T. Luong, "FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation," in Findings of the Association for Computational Linguistics: ACL 2024, pp. 13697–13720, 2024. https://doi.org/10.18653/v1/2024.findings-acl.813

[3] W. Chen, X. Wang, and W. Y. Wang, "A Dataset for Answering Time-Sensitive Questions," in Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks), 2021. arXiv:2108.06314.

[4] A. Lazaridou, A. Kuncoro, E. Gribovskaya, D. Agrawal, A. Liska, T. Terzi, M. R. Gimenez, et al., "Mind the Gap: Assessing Temporal Generalization in Neural Language Models," in Advances in Neural Information Processing Systems, vol. 34, 2021.

[5] L. Huang, et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," ACM Computing Surveys, 2025. https://doi.org/10.1145/3703155

[6] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1-38, 2023. https://doi.org/10.1145/3571730

[7] J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych, "A Survey of Confidence Estimation and Calibration in Large Language Models," in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024. arXiv:2311.08298.

[8] X. Liu, F. Fatahi Bayat, and L. Wang, "Enhancing Language Model Factuality via Activation-Based Confidence Calibration and Guided Decoding," arXiv preprint arXiv:2406.13230, 2024.

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.

[10] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," in Proceedings of the 37th International Conference on Machine Learning, pp. 3929-3938, 2020.

[11] Z. Chu, J. Chen, Q. Chen, W. Yu, H. Wang, M. Liu, and B. Qin, "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), pp. 1204–1228, 2024. doi: 10.18653/v1/2024.acl-long.66

[12] Y. Wang and Y. Zhao, "TRAM: Benchmarking Temporal Reasoning for Large Language Models," in Findings of the Association for Computational Linguistics: ACL 2024, pp. 6389–6415, 2024. doi: 10.18653/v1/2024.findings-acl.382

[13] Z. Jiang, B. Liu, M. Peng, W. Xu, Y. Xiao, Z. Shan, and M. Peng, "Towards Explainable Temporal Reasoning in Large Language Models: A Structure-Aware Generative Framework," in Findings of the Association for Computational Linguistics: ACL 2025, pp. 7232–7251, 2025. doi: 10.18653/v1/2025.findings-acl.378

[14] Y. Wang, M. Wang, M. A. Manzoor, F. Liu, G. Georgiev, R. J. Das, and P. Nakov, "Factuality of Large Language Models: A Survey," in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 19519–19529, 2024. doi: 10.18653/v1/2024.emnlp-main.1088

[15] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, and Q. Li, "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," arXiv preprint arXiv:2405.06211, 2024. doi: 10.48550/arXiv.2405.06211

[16] V. Mazzia, S. Pollastrini, D. Bernardi, C. Rubagotti, and D. Amberti, "Benchmarking Multilingual Temporal Reasoning in LLMs: The Temporal Reasoning Dataset," in Proceedings of the 16th International Workshop on Spoken Dialogue System Technology, pp. 168–181, 2026.