

Does the Temporal Focus of Information Affect Confidence Calibration in LLMs? A Study of Temporal Effects on Model Confidence

Bhawna Piryani*, Jamshid Mozafari, Adam Jatowt

University of Innsbruck, Innsbruck, Austria

{bhawna.piryani, jamshid.mozafari, adam.jatowt}@uibk.ac.at

Abstract

Large language models (LLMs) are widely used as factual knowledge sources, yet reliable deployment requires not only accurate predictions but also well-calibrated confidence estimates. While prior work on temporal robustness has primarily focused on changes in factual accuracy over time, we study *temporal confidence calibration*: whether model confidence remains aligned with correctness across questions associated with different reference times. We evaluate multiple LLMs on temporally grounded question answering benchmarks spanning both historical and dynamically evolving facts, comparing five confidence estimation strategies using Expected Calibration Error (ECE) and Brier Score computed across temporal slices. Our experiments reveal substantial temporal miscalibration: as factual accuracy varies across time, model confidence often fails to track these changes proportionally, leading to pronounced overconfidence on time-sensitive queries. Across the evaluated confidence estimators, no single method consistently dominates: P(True) and Token Probability Confidence perform best most often, while the optimal estimator varies across models and datasets. We further show that global temperature scaling improves calibration but leaves meaningful period-specific variation, indicating that standard post-hoc calibration does not fully eliminate temporal miscalibration. These findings highlight temporal calibration as an important and previously underexplored dimension of LLM reliability.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable performance across a wide range of natural language processing tasks, including question answering, summarization, and reasoning (Brown et al., 2020; Wei et al., 2022). Despite these advances, LLMs are typically trained

Period	Question	Prediction	Confidence	Correctness
Distant Past	Question: Who was the Prime Minister of United Kingdom during World War II?	Prediction: Winston Churchill	Confidence: 96%	✓
	Question: Which country controlled Korea in 1945?	Prediction: China	Confidence: 92%	✗
Intermediate Period	Question: Who was the CEO of Apple in 2010?	Prediction: Tim Cook	Confidence: 94%	✓
	Question: Which company owned Instagram in 2010?	Prediction: Meta	Confidence: 98%	✗
Recent	Question: Which company owned Twitter in 2024?	Prediction: X Corp	Confidence: 95%	✓
	Question: Who won the 2024 UEFA Champions League?	Prediction: Manchester City	Confidence: 93%	✗

Figure 1: Representative examples illustrating temporal miscalibration across different time periods. Although the model assigns consistently high confidence, factual correctness varies substantially, including confidently incorrect answers under temporal shift (e.g., predicting that Meta owned Instagram in 2010). These examples highlight that confidence alone does not reliably reflect temporal correctness, motivating the need for studying *temporal calibration*, namely the alignment between model confidence and correctness across time.

on static corpora with a fixed temporal cutoff, after which their internal knowledge may no longer reliably reflect changes in the world (Lazaridou et al., 2021; Dai et al., 2024; Zhu et al., 2025). As a result, model performance often deteriorates when answering questions involving evolving entities, changing relations, or recent events (Lazaridou et al., 2021; Vu et al., 2024).

While prior work has primarily focused on how *accuracy* changes over time (Lazaridou et al., 2021; Zhu et al., 2025; Wallat et al., 2024), considerably less attention has been paid to whether model *confidence* remains reliable across temporal contexts. Figure 1 illustrates this phenomenon through representative examples spanning historical, intermediate, and recent periods. Although the model assigns

*Corresponding Author.

consistently high confidence, factual correctness varies substantially across time. This mismatch suggests that temporal reliability cannot be characterized by accuracy alone.

Reliable confidence estimates are critical for the safe deployment of LLMs. A model that is incorrect but uncertain may be relatively harmless, whereas a confidently incorrect model can pose substantial risks in applications such as journalism, education, and decision-support systems (Kadavath et al., 2022). These risks become particularly important in temporally sensitive settings, where factual validity evolves over time while the model’s internal knowledge remains static after training (Vu et al., 2024; Zhang and Choi, 2023). Consequently, models may confidently generate outdated or temporally inconsistent information even when factual accuracy deteriorates.

Such behavior reflects a failure of *calibration*, where predicted confidence no longer aligns with empirical correctness. Existing calibration studies largely assume stationary evaluation distributions (Guo et al., 2017; Jiang et al., 2021); however, temporal Question Answering (QA) introduces a fundamentally non-stationary setting in which the validity of facts changes over time. We refer to this problem as *temporal calibration*: the alignment between model confidence and correctness conditioned on the temporal reference of a question.

To study this phenomenon, we evaluate multiple confidence estimation strategies across temporally stratified subsets of two factual QA benchmarks spanning multiple decades. We analyze calibration behavior using Expected Calibration Error (ECE) and Brier Score (Guo et al., 2017; Jiang et al., 2021) computed across temporal intervals, enabling a fine-grained analysis of how confidence reliability changes over time. Our analysis is guided by the following research questions:

RQ1: Do confidence estimates remain calibrated when answers are conditioned on different temporal reference periods?

RQ2: How stable is confidence calibration across temporal periods?

RQ3: To what extent does temporal miscalibration manifest as overconfidence on dynamic or time-sensitive factual queries?

RQ4: Does a globally fitted post-hoc calibrator remain reliable across individual temporal slices?

Our experiments reveal substantial temporal miscalibration across model families and confidence estimation strategies. While factual accuracy varies

over time, confidence often fails to track these changes proportionally, resulting in pronounced overconfidence on temporally dynamic questions. Across the five evaluated confidence estimators, no single method consistently dominates: **P(True)** and **Token Probability Confidence** perform best most often, while the optimal estimator varies across models and datasets. We further show that standard global temperature scaling improves calibration but leaves meaningful period-specific variation, indicating that post-hoc calibration does not fully eliminate temporal miscalibration.

Our contributions are ¹:

(1) we introduce the problem of temporal calibration in LLMs; (2) we provide a systematic empirical analysis of confidence reliability across temporal QA benchmarks, model families, and confidence estimation strategies; and (3) we assess the effectiveness of post-hoc calibration under temporal distribution shift and characterize its limitations across temporal periods.

These findings highlight temporal calibration as an important dimension of LLM reliability.

2 Related Work

Our work relates to prior research on confidence calibration and uncertainty estimation in LLMs (Guo et al., 2017; Kadavath et al., 2022; Jiang et al., 2021), temporal generalization and knowledge freshness (Lazaridou et al., 2021; Vu et al., 2024; Zhu et al., 2025), time-aware language modeling (Dhingra et al., 2022), and temporal reasoning and robustness in language models (Wallat et al., 2025; Tan et al., 2024).

To the best of our knowledge, this is the first systematic empirical study of *temporal confidence calibration* in LLMs, focusing on whether confidence remains aligned with correctness across explicit temporal reference periods. A detailed discussion of related work is provided in Appendix A.

3 Methodology

3.1 Problem Definition

We study *temporal confidence calibration* in large language models for time-sensitive factual question answering. Each question q is associated with a reference year t , representing the temporal context in which the queried fact is valid. Given (q, t) ,

¹Code and experimental resources are available at <https://github.com/DataScienceUIBK/temporal-confidence-calibration>

a language model produces an answer $\hat{a}$ together with a confidence estimate $c \in [0, 1]$, interpreted as the estimated probability that the prediction is correct.

Let $z \in \{0, 1\}$ denote factual correctness, where $z = 1$ if $\hat{a}$ matches the ground-truth answer and $z = 0$ otherwise. A confidence estimation strategy is considered well-calibrated if:

$$P(z = 1 \mid c = p) = p,$$

meaning that predictions assigned confidence p are correct approximately a p fraction of the time.

Our goal is to characterize how confidence calibration varies across temporal reference periods and confidence estimation strategies. In particular, we investigate whether confidence remains aligned with correctness when models answer factual questions conditioned on different reference times.

3.2 Confidence Estimation Methods

We evaluate five complementary confidence estimation strategies that derive uncertainty from different signals: explicit self-assessment (Verbal Confidence) (Xiong et al., 2024), verification (P(True)) (Kadavath et al., 2022), sequence likelihood (Token Probability Confidence) (Tian et al., 2023), generation agreement (Self-Consistency) (Wang et al., 2023), and meaning-level variation across sampled responses (Semantic Entropy) (Kuhn et al., 2023). Each method produces a scalar confidence score $c \in [0, 1]$ for every prediction.

Verbal Confidence Verbal confidence estimates uncertainty using the model’s explicit self-reported confidence score (Jiang et al., 2021; Xiong et al., 2024). In the same generation prompt, the model is asked to produce both an answer and a confidence score between 0 and 100, indicating its confidence in the answer’s correctness. We normalize this score to $[0, 1]$.

This approach captures the model’s direct self-assessment of correctness and reflects the type of confidence signal commonly exposed to end users in interactive systems. The prompt template is shown in Figure 7 in Appendix E.

P(True): Verification-Based Confidence We adopt the P(True) formulation (Kadavath et al., 2022; Zhang and Choi, 2023). After generating an answer $\hat{a}$, the model is prompted with a binary verification query constrained to respond with either True or False. Confidence is computed as the

normalized probability assigned to the True token:

$$c^{\text{P(True)}} = \frac{\exp(l_{\text{True}})}{\exp(l_{\text{True}}) + \exp(l_{\text{False}})},$$

where l_{True} and l_{False} denote the next-token logits corresponding to the True and False responses. The prompt templates are shown in Figure 8 in Appendix E.

Token Probability Confidence Token Probability Confidence estimates uncertainty using the probability assigned by the model to its generated answer (Tian et al., 2023). The intuition behind this approach is that answers assigned higher likelihood under the model distribution correspond to more confident predictions.

For each generated answer sequence $a = (a_1, \dots, a_N)$, we compute sequence-level confidence using the geometric mean token probability:

$$c^{\text{token}}(y) = \exp\left(\frac{1}{N} \sum_{i=1}^N \log P(a_i \mid a_{<i}, x)\right)$$

where x denotes the input question and a_i is the generated token at position i .

We use the geometric mean to normalize differences in answer length and avoid systematically penalizing longer generations. Higher values indicate greater model confidence in the generated answer. The prompt template used for token probability confidence is provided in Figure 10 in Appendix E.

Self-Consistency Confidence Self-consistency estimates confidence from the stability of predictions under stochastic sampling (Wang et al., 2023; Lin et al., 2023).

For each question, we generate $K = 5$ independent samples using temperature sampling with $T = 0.7$ and select the most frequent answer (the *mode*) as the final prediction. Confidence is defined as the fraction of samples agreeing with the mode answer:

$$c^{\text{sc}} = \frac{m}{K},$$

where m denotes the number of generations matching the modal answer. Higher agreement across generations corresponds to higher confidence.

The prompt template used for Self-Consistency is provided in Figure 9 in Appendix E.

Model	P(True)			Verbal Confidence			Token Probability			Self-Consistency			Semantic Entropy		
	Acc ↑	ECE ↓	Brier ↓	Acc ↑	ECE ↓	Brier ↓	Acc ↑	ECE ↓	Brier ↓	Acc ↑	ECE ↓	Brier ↓	Acc ↑	ECE ↓	Brier ↓
ArchivalQA															
Qwen-3 4B	30.08	0.317	0.320	23.91	0.529	0.499	25.54	0.457	0.424	25.53	0.626	0.613	26.0	0.619	0.612
Gemma-3 4B	21.82	0.612	0.592	21.30	0.741	0.721	21.37	0.320	0.292	24.28	0.428	0.372	24.9	0.401	0.356
Ministral-3 8B	36.85	0.220	0.313	34.06	0.592	0.576	32.40	0.297	0.285	10.56	0.076	0.077	29.5	0.207	0.238
LLaMA-3.1 8B	31.47	0.055	0.220	30.22	0.581	0.548	29.47	0.279	0.241	31.61	0.308	0.283	34.6	0.289	0.288
Gemma-3 27B	39.51	0.434	0.447	37.41	0.535	0.526	37.46	0.347	0.355	40.37	0.423	0.404	40.8	0.400	0.392
Qwen-3 30B	38.55	0.429	0.451	37.22	0.563	0.543	36.76	0.362	0.344	39.15	0.422	0.405	40.2	0.406	0.397
LLaMA-3.3 70B	46.10	0.191	0.242	44.88	0.384	0.367	44.49	0.421	0.397	45.62	0.378	0.369	48.5	0.382	0.378
Qwen-2.5 72B	46.41	0.254	0.282	41.56	0.412	0.392	41.28	0.329	0.329	43.33	0.415	0.397	43.6	0.386	0.379
TEMPLAMA															
Qwen-3 4B	8.34	0.683	0.572	8.67	0.647	0.560	10.49	0.540	0.429	12.20	0.621	0.562	12.4	0.604	0.545
Gemma-3 4B	13.77	0.723	0.680	13.04	0.839	0.816	13.64	0.320	0.278	14.87	0.524	0.437	15.0	0.477	0.394
Ministral-3 8B	13.34	0.609	0.487	14.11	0.822	0.801	14.29	0.449	0.317	9.24	0.128	0.115	22.0	0.318	0.285
LLaMA-3.1 8B	19.19	0.125	0.175	17.19	0.776	0.747	22.84	0.409	0.323	22.07	0.371	0.324	23.5	0.359	0.323
Gemma-3 27B	32.38	0.447	0.455	31.65	0.650	0.635	31.80	0.403	0.409	32.12	0.505	0.478	32.4	0.484	0.463
Qwen-3 30B	25.88	0.566	0.544	25.06	0.719	0.697	24.53	0.452	0.409	26.82	0.500	0.452	27.2	0.478	0.433
LLaMA-3.3 70B	40.46	0.155	0.250	37.02	0.502	0.473	41.53	0.493	0.474	41.43	0.476	0.462	41.6	0.456	0.448
Qwen-2.5 72B	30.40	0.338	0.332	28.90	0.589	0.544	30.00	0.452	0.401	29.63	0.503	0.456	29.9	0.472	0.431

Table 1: Model-wise calibration performance across five confidence estimation strategies on ARCHIVALQA and TEMPLAMA. We report Accuracy (Acc), Expected Calibration Error (ECE), and Brier Score aggregated across all temporal slices. Higher Accuracy and lower ECE/Brier Scores indicate better performance. Best results within each model row are shown in bold, while the best overall result for each metric is underlined.

Semantic Entropy Semantic Entropy estimates uncertainty from meaning-level variation across multiple sampled responses (Kuhn et al., 2023). Using the same set of K sampled generations, we group semantically equivalent answers into clusters. Let n_j denote the number of sampled responses assigned to semantic cluster j . We estimate the empirical probability of each cluster as

$$p_j = \frac{n_j}{K},$$

and compute the semantic entropy

$$H_{\text{sem}} = - \sum_{j=1}^M p_j \log p_j,$$

where M denotes the number of distinct semantic clusters.

Because the maximum entropy attainable with K samples is $\log K$, we normalize the entropy and convert it to a confidence score:

$$c^{\text{SE}} = 1 - \frac{H_{\text{sem}}}{\log K}.$$

Thus, c^{SE} approaches one when sampled responses concentrate on a single semantic interpretation and decreases as responses become more semantically diverse. We refer to this estimator as *sample-based Semantic Entropy*, since cluster probabilities are estimated from empirical sample frequencies.

Unlike Self-Consistency, which depends only on agreement with the modal answer, Semantic Entropy incorporates the distribution of probability mass across all observed semantic clusters.

3.3 Temperature Scaling for Temporal Calibration

Temperature scaling is an established post-hoc calibration method (Guo et al., 2017). We use it as a mitigation baseline to assess whether a standard globally fitted calibrator remains effective when confidence reliability varies across temporal periods. We apply temperature scaling to both **Token Probability Confidence** and **P(True)**, since both are derived from model logits and therefore support direct probability rescaling.

Given logits $\mathbf{z}$, temperature scaling produces

$$P_T(i | \mathbf{z}) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)},$$

where $T > 0$ is a scalar temperature parameter. Values $T > 1$ soften the probability distribution, whereas values $T < 1$ sharpen it.

For Token Probability Confidence, temperature scaling is applied to the token-level logits before computing the geometric mean probability of the generated answer. For P(True), the True and False logits are rescaled before computing the normalized probability assigned to True. Temperature scaling therefore modifies only the associated confidence estimates and does not alter the generated answers or their accuracy.

For each model-dataset pair, we divide the data into a calibration split and a held-out test split. The temperature parameter is selected on the calibration split by minimizing binary negative log-likelihood (NLL) and is then evaluated on the held-out test

split without further tuning. We search over

$$T \in \{0.4, 0.7, 0.9, 1.0, 1.2, 1.5, 2, 2.5, 2.8, 3, 3.5, 4, 5, 7, 10\}.$$

We consider two calibration settings:

Global Temperature Scaling. A single temperature parameter is learned from the calibration split for each model-dataset pair and applied uniformly to all temporal periods in the corresponding test split.

Temporal Temperature Scaling. A separate temperature parameter is learned for each temporal bucket using the corresponding calibration examples and is evaluated on held-out examples from the same temporal bucket. This setting allows us to examine whether conditioning the calibrator on temporal period provides additional benefits over a single globally fitted temperature.

4 Experimental Setup

4.1 Datasets

We evaluate temporal confidence calibration on two temporally grounded QA benchmarks spanning both historical and dynamically evolving factual knowledge.

ARCHIVALQA (Wang et al., 2022) contains historical question-answer pairs associated with explicit reference years. To study long-range temporal calibration while maintaining balanced temporal coverage, we sample 100 questions for each reference year from 1950 through 2010, yielding **6,100** questions covering diverse historical periods.

TEMPLAMA (Dhingra et al., 2022) consists of temporally anchored cloze-style factual statements spanning 2010–2020. Because our evaluation requires natural question-answer interactions and confidence elicitation, we convert the cloze statements into natural-language questions, yielding **4,654** temporally grounded question-answer pairs. The conversion prompt and representative examples are provided in Appendix B.

Across the two datasets, our evaluation comprises **10,754** temporally grounded questions. We analyze temporal trends separately within each benchmark rather than pooling the datasets: ARCHIVALQA provides balanced coverage of long-range historical periods, whereas TEMPLAMA captures shorter-term changes in temporally dynamic factual knowledge. Additional dataset de-

tails and example questions are provided in Appendix B.

4.2 Evaluation Metrics

We evaluate factual correctness and confidence calibration using three primary metrics: Accuracy, Expected Calibration Error (ECE) (Guo et al., 2017), and Brier Score (Glenn et al., 1950).

Accuracy is computed using GPT-EVAL to account for semantically correct paraphrased answers (Kamalloo et al., 2023). Prior work has shown that GPT-based evaluation better captures semantic correctness than strict lexical matching for open-ended QA (Badshah et al., 2025; Adlakha et al., 2024). To improve robustness and reduce stochastic variance in LLM-based judgments, each generated answer is evaluated independently three times using the same GPT-EVAL prompt, and the final correctness label is determined through majority voting over the three binary judgments.

We report ECE and Brier Score both at the aggregate level and separately for each temporal slice. The aggregate scores characterize overall calibration for each model-dataset setting, while the slice-level scores enable analysis of how calibration varies across reference periods. Following prior calibration literature, ECE is computed using 10 equally spaced confidence bins. Full metric definitions, evaluation prompts, and additional comparisons with Exact Match (EM) and Contains accuracy are provided in Appendix C.

4.3 Models and Implementation Details

We evaluate eight open-weight, instruction-tuned LLMs from the LLaMA (Grattafiori et al., 2024), Qwen (Yang et al., 2025), Gemma (Team et al., 2025), and Ministral (Liu et al., 2026) model families, with parameter sizes ranging from 4B to 72B. All models are evaluated in a zero-shot setting without task-specific fine-tuning or retrieval augmentation. Additional implementation details and the complete model list are provided in Appendix D.

5 Results and Discussion

We evaluate temporal confidence calibration across multiple model families, confidence estimation strategies, and temporally grounded QA benchmarks. We first analyze aggregate calibration behavior, then examine temporal calibration trends, post-hoc calibration via temperature scaling, and additional robustness analyses.

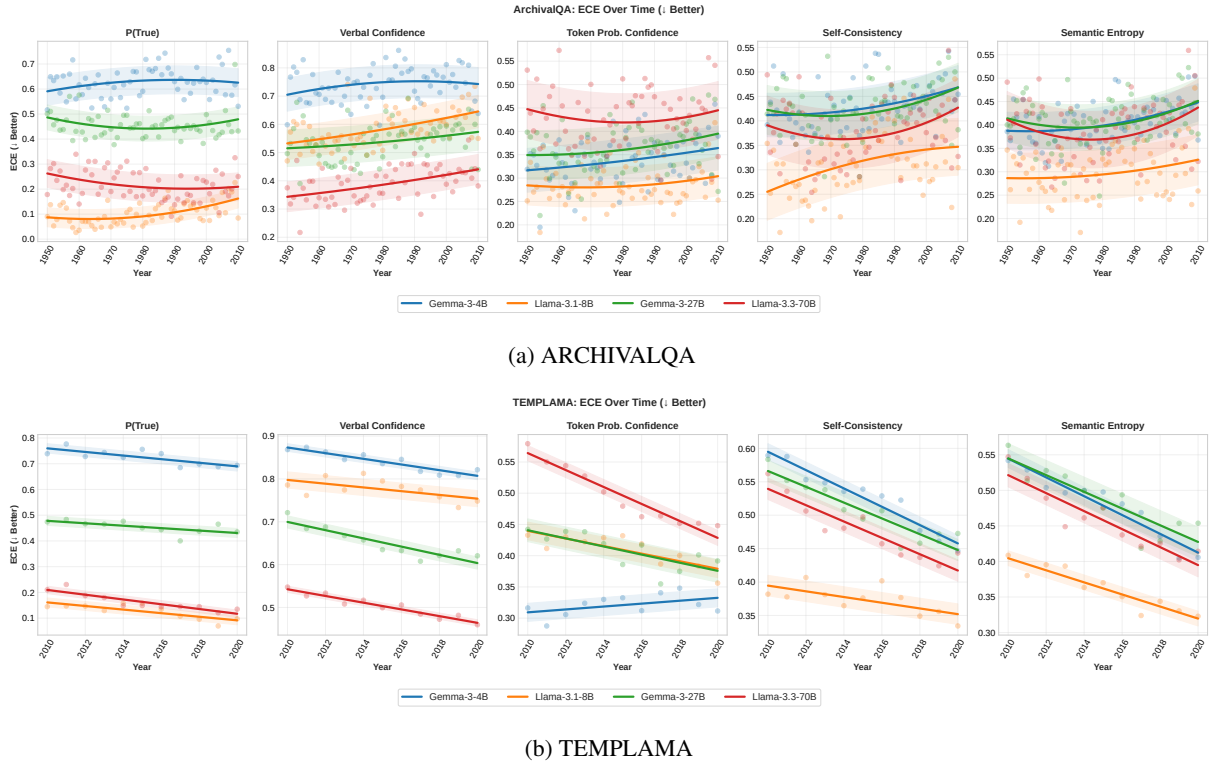


Figure 2: Temporal evolution of ECE across representative models and confidence estimation strategies on (a) ARCHIVALQA (1950–2010) and (b) TEMPLAMA (2010–2020). Lower values indicate better calibration. Calibration varies substantially across reference periods, models, and confidence estimators. ARCHIVALQA exhibits heterogeneous temporal trajectories, whereas TEMPLAMA generally shows decreasing calibration error toward more recent reference years.

5.1 Method-wise Calibration Comparison

Table 1 shows substantial variation in aggregate calibration across estimators, models, and datasets. Overall, no single confidence estimator consistently dominates across all settings.

The best confidence estimator depends on the model and dataset. Across the 16 model–dataset settings, **P(True)** and **Token Probability Confidence** each achieve the lowest ECE/Brier pair in 7 settings, while **Self-Consistency** performs best in the remaining two. **P(True)** performs particularly strongly for the LLaMA models. For example, on ARCHIVALQA, LLaMA-3.1-8B achieves an ECE of **0.055** with **P(True)**, compared with 0.279 for Token Probability Confidence, 0.308 for Self-Consistency, and 0.289 for Semantic Entropy. In contrast, Token Probability Confidence performs best for several Gemma and Qwen settings. These results indicate that estimator quality depends on the model and dataset rather than a universally optimal confidence signal.

Estimator-specific calibration patterns. Verbal Confidence is generally poorly calibrated, while Se-

manic Entropy provides a stronger sampling-based alternative, achieving lower ECE than Verbal Confidence in 15 of 16 settings and lower ECE than Self-Consistency in 13 of 16. However, Semantic Entropy does not achieve the lowest ECE/Brier pair in any setting. Self-Consistency and Token Probability Confidence can also achieve competitive factual accuracy while remaining substantially miscalibrated in some cases; for example, on TEMPLAMA, LLaMA-3.3-70B achieves over 41% accuracy under both methods, with ECE values of 0.476 and 0.493, respectively. The poor calibration of Verbal Confidence further suggests that instruction-tuned LLMs may produce authoritative confidence statements that are only weakly aligned with their underlying uncertainty in these temporally grounded settings.

Calibration quality does not consistently improve with model scale. Larger models often achieve higher factual accuracy, but lower calibration error does not follow consistently from increased model size. For example, LLaMA-3.1-8B achieves lower **P(True)** ECE than LLaMA-

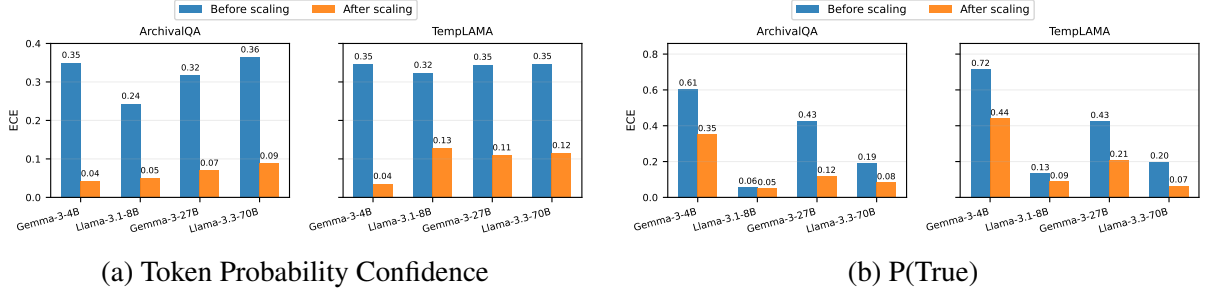


Figure 3: ECE before and after applying Global Temperature Scaling on held-out test sets for (a) Token Probability Confidence and (b) P(True). Global Temperature Scaling reduces calibration error for both confidence signals across the representative models and datasets.

3.3-70B on both datasets: 0.055 versus 0.191 on ARCHIVALQA, and 0.125 versus 0.155 on TEMPLAMA. Moreover, low calibration error does not necessarily imply high factual accuracy: Ministral-3 8B obtains an ECE of 0.076 under Self-Consistency on ARCHIVALQA while achieving only 10.56% accuracy. Thus, factual accuracy and confidence calibration capture distinct aspects of model reliability, and improvements in factual performance do not automatically translate into better calibration.

Calibration behavior differs across datasets. For many model–estimator combinations, calibration error is generally lower on ARCHIVALQA than on TEMPLAMA. For example, LLaMA-3.1-8B achieves a P(True) ECE of 0.055 on ARCHIVALQA, compared with 0.125 on TEMPLAMA. Similar trends are observed across Verbal Confidence, Self-Consistency, Token Probability Confidence, and Semantic Entropy. These differences suggest that dynamically evolving factual knowledge can present a greater calibration challenge, while differences in dataset construction, temporal coverage, and factual characteristics may also contribute.

Reliability diagrams for all models and confidence estimation methods are provided in Appendix F. Temporal Brier Score results are reported in Figures 11 and 13 of the same appendix. Additional qualitative examples of temporal miscalibration are provided in Appendix I.

5.2 Temporal Dynamics of Calibration

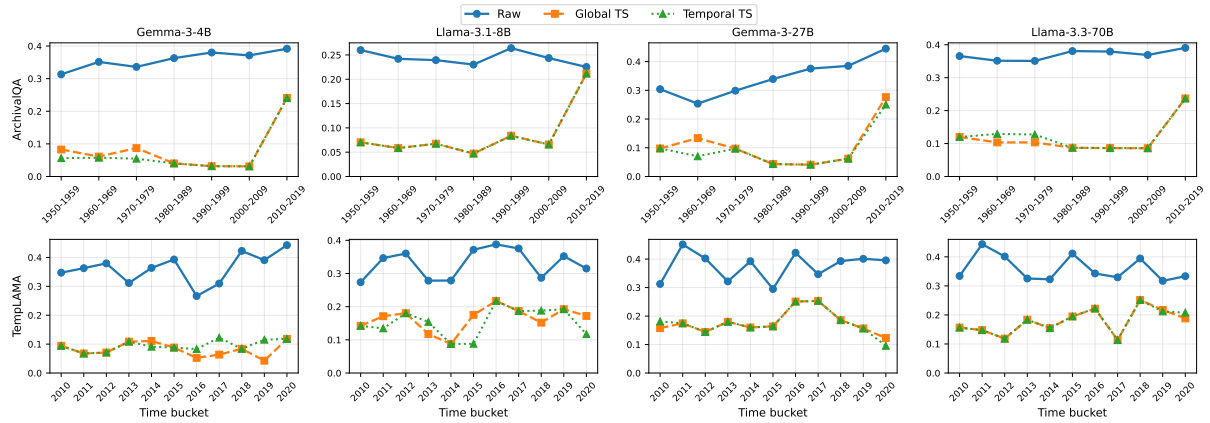
Aggregate metrics provide an overall view of calibration quality, but they obscure how confidence estimation behaves across different temporal periods. To better understand the relationship between temporal drift and uncertainty estimation, we ana-

lyze year-wise calibration trends over time.

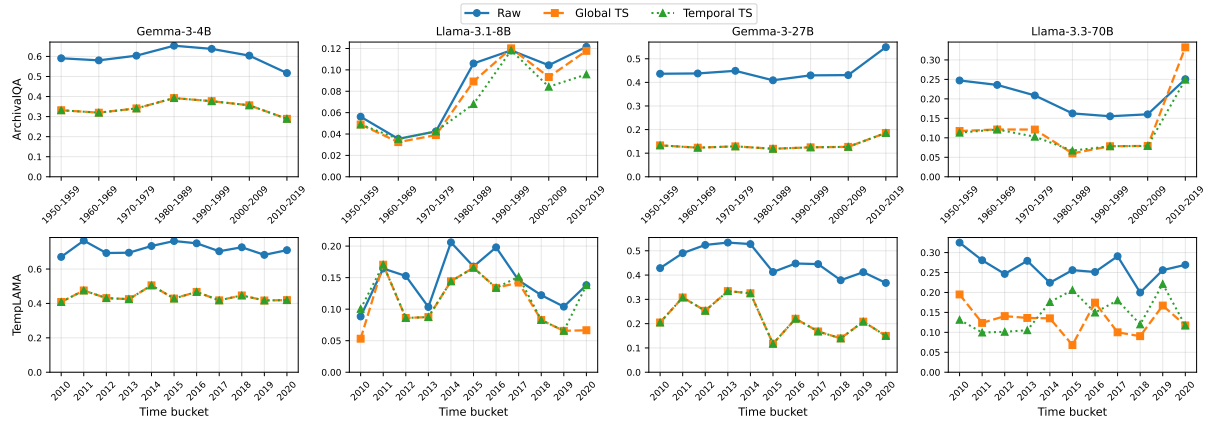
Figure 2 presents year-wise ECE trends for four representative models spanning different parameter scales and model families: Gemma-3-4B, LLaMA-3.1-8B, Gemma-3-27B, and LLaMA-3.3-70B. We focus on representative models in the main paper for clarity and readability, while full temporal calibration results for all evaluated models are provided in Appendix F. Corresponding Brier Score trends are reported in Figures 11 and 13.

Temporal calibration drift differs across datasets. On ARCHIVALQA, calibration exhibits heterogeneous temporal trajectories across models and confidence estimators. Several settings, including P(True) for LLaMA-3.1-8B, show a counter-intuitive pattern in which ECE and Brier Score are comparatively lower for earlier historical periods and increase toward more recent decades. For example, LLaMA-3.1-8B exhibits relatively strong calibration during the 1950s and 1960s, with both metrics increasing toward the 1990s and 2000s. One possible explanation is that historical facts from earlier periods may be more canonical and temporally stable in pretraining data, whereas facts from more recent decades may exhibit greater temporal variation.

In contrast, TEMPLAMA exhibits a clear temporal pattern: calibration error generally decreases toward more recent reference years across models and confidence estimators. This trend is especially consistent for P(True), Verbal Confidence, Self-Consistency, and Semantic Entropy, for which all four representative models show decreasing ECE toward 2020. One possible explanation is that more recent facts are better represented in the models’ learned knowledge, whereas older dynamic facts may be more susceptible to stale factual representations.



(a) Token Probability Confidence



(b) P(True)

Figure 4: Temporal ECE for raw confidence estimates, Global Temperature Scaling, and Temporal Temperature Scaling for (a) Token Probability Confidence and (b) P(True) across representative models and datasets. Global Temperature Scaling generally reduces calibration error across temporal periods, while Temporal Temperature Scaling provides additional improvements in selected temporal slices. The relative performance of global and temporal scaling varies across models, datasets, and reference periods.

Temporal trends are broadly consistent across calibration metrics. ECE and Brier Score exhibit qualitatively similar temporal patterns across many model–estimator combinations, although their trajectories are not identical in every setting. Methods with higher ECE often also exhibit higher Brier Scores across temporal periods, while the relative ranking of confidence estimators can vary by model and reference period. This consistency suggests that the observed temporal trends reflect broader variation in confidence reliability under temporal shift rather than artifacts of a particular calibration metric.

Overall, our results demonstrate that calibration reliability varies substantially across temporal periods, models, confidence estimation methods, and datasets. The contrasting temporal trajectories observed on ARCHIVALQA and TEMPLAMA fur-

ther indicate that the relationship between calibration and temporal distance is not uniform across datasets. Instead, confidence reliability depends on the specific reference period and evaluation setting, highlighting the importance of assessing calibration within individual temporal slices rather than relying solely on aggregate metrics.

5.3 Assessing Post-hoc Calibration across Temporal Periods

Temperature scaling is an established post-hoc calibration method (Guo et al., 2017). We use it as a mitigation baseline to assess whether confidence estimates can be recalibrated under temporal distribution shift and whether a calibrator fitted globally remains reliable across individual temporal periods. We evaluate temperature scaling for both Token Probability Confidence and P(True).

Global Temperature Scaling. We first learn a single global temperature for each model–dataset pair on the calibration split and evaluate it on the corresponding held-out test split. Global Temperature Scaling improves aggregate calibration for both Token Probability Confidence and $P(\text{True})$ across the evaluated models and datasets. As shown in Figure 3, ECE decreases after scaling for both confidence signals, with corresponding Brier Score results reported in Appendix F. These results show that both confidence signals benefit from standard post-hoc calibration.

Calibration Remains Period-Dependent. Although Global Temperature Scaling improves aggregate calibration, its effectiveness varies across temporal periods. Different temporal buckets often favor different temperature values. For ARCHIVALQA, the preferred temperatures vary across decades, while for TEMPLAMA they vary across individual reference years. This pattern is observed for both confidence signals, indicating that a single globally fitted temperature can improve overall calibration without fully removing period-specific differences. Bucket-level temperature results are provided in Appendix F (Figure 18).

Global versus Temporal Temperature Scaling. We further compare Global Temperature Scaling with Temporal Temperature Scaling, where separate temperature parameters are learned for individual temporal buckets. Figure 4 presents the resulting temporal ECE trends, with panel (a) showing Token Probability Confidence and panel (b) showing $P(\text{True})$.

Temporal Temperature Scaling provides additional improvements in some temporal slices, but it does not consistently outperform Global Temperature Scaling. Across both confidence signals, the two approaches often yield similar calibration, with their relative performance varying across models, datasets, and reference periods. These results indicate that temporal conditioning can provide local calibration gains, but fitting a separate temperature for each period is not uniformly superior to a single global calibrator.

Overall, Global Temperature Scaling provides substantial calibration improvements for both Token Probability Confidence and $P(\text{True})$, but residual variation remains across temporal periods. These findings show that aggregate post-hoc calibration can conceal period-specific differences in confidence reliability, motivating the evaluation

and monitoring of calibration at the level of individual temporal slices.

5.4 Additional Robustness Experiments

We conduct additional robustness analyses to evaluate whether the observed temporal calibration trends are specific to temporally grounded QA settings or sensitive to prompt formulation.

Non-Temporal Baseline Comparison. We first compare calibration performance on temporally grounded QA with non-temporal factual QA (TriviaQA (Joshi et al., 2017)) using temporally neutral prompts. Calibration is consistently stronger on non-temporal QA across nearly all evaluated settings, suggesting that temporally evolving factual knowledge poses a substantially greater calibration challenge. Detailed results are provided in Appendix G.1, with corresponding prompts in appendix H.

Prompt Sensitivity Analysis. We additionally evaluate multiple temporal prompt formulations to assess whether the observed calibration behavior depends on prompt phrasing. Across prompt variants, the overall calibration patterns remain qualitatively consistent across models and confidence estimation strategies, indicating that the observed effects are not artifacts of a particular prompt formulation. Full quantitative results are provided in Appendix G.2, with prompt templates in appendix H.

6 Conclusion

We presented the first systematic study of temporal confidence calibration in LLMs. Across two temporally grounded QA benchmarks, we find substantial variation in calibration across temporal periods, models, and confidence estimation strategies. No single estimator consistently dominates: $P(\text{True})$ and Token Probability Confidence perform best most often, while the best-performing estimator varies across models and datasets. Our temporal analysis further shows that confidence reliability changes across reference periods. Global Temperature Scaling improves calibration for both Token Probability Confidence and $P(\text{True})$, but meaningful period-specific variation remains, and Temporal Temperature Scaling does not consistently outperform a globally fitted calibrator. These findings highlight the importance of evaluating confidence calibration at the level of individual temporal slices rather than relying solely on aggregate metrics.

Limitations

Our evaluation focuses on temporal confidence calibration in factual question answering across two temporally grounded benchmarks and a diverse set of open-weight, instruction-tuned LLMs. While this setting enables a controlled analysis of how confidence reliability varies across reference periods, calibration behavior may differ in tasks involving complex reasoning, multi-step inference, multimodal inputs, or other forms of temporally conditioned generation.

We evaluate Global and Temporal Temperature Scaling to characterize the effectiveness of established post-hoc calibration under temporal variation. Our results show that post-hoc calibration can substantially improve confidence reliability, while residual period-specific differences remain. Extending this analysis to additional calibration techniques and developing methods that explicitly account for temporal variation constitute promising directions for future work.

Finally, although our evaluation spans multiple model families and parameter scales, the findings may not fully generalize to proprietary models or systems trained with substantially different data, architectures, or alignment procedures.

Acknowledgments

The computational results presented here have been achieved (in part) using the LEO HPC infrastructure of the University of Innsbruck and the Austrian Scientific Computing (ASC) infrastructure.

References

- Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. [Evaluating correctness and faithfulness of instruction-following models for question answering](#). *Transactions of the Association for Computational Linguistics*, 12:681–699.
- Sher Badshah, Moamen Moustafa, and Hassan Sajjad. 2025. [CLEV: LLM-based evaluation through lightweight efficient voting for free-form question-answering](#). 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 1513–1531, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Hui Dai, Ryan Teehan, and Mengye Ren. 2024. Are llms prescient? a continuous evaluation using daily news as the oracle. *arXiv preprint arXiv:2411.08324*.
- 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.
- W Brierch Glenn and 1 others. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [How can we know when language models know? on the calibration of language models for question answering](#). *Transactions of the Association for Computational Linguistics*, 9:962–977.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. [Evaluating open-domain question answering in the era of large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

- Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, and 1 others. 2023. Real-time qa: What’s the answer right now? *Advances in neural information processing systems*, 36:49025–49043.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *The Eleventh International Conference on Learning Representations*.
- Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liška, 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 *Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS ’21*, Red Hook, NY, USA. Curran Associates Inc.
- Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.
- Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D’Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. [StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 13604–13622. PMLR.
- Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.
- Sergey Pletenev, Maria Marina, Nikolay Ivanov, Daria Galimzianova, Nikita Krayko, Mikhail Salnikov, Vasily Kononov, Alexander Panchenko, and Viktor Moskvoretskii. 2025. [Will it still be true tomorrow? multilingual evergreen question classification to improve trustworthy QA](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8603–8620, Suzhou, China. 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.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.
- 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. [Fresh-LLMs: 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, Abdelrahman Abdallah, Adam Jatowt, and Avishek Anand. 2025. [A study into investigating temporal robustness of LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15685–15705, Vienna, Austria. 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.
- Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2022. [Archivalqa: A large-scale benchmark dataset for open-domain question answering over historical news collections](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’22*, page 3025–3035, New York, NY, USA. Association for Computing Machinery.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. [Know your limits: A survey of abstention in large](#)

language models. *Transactions of the Association for Computational Linguistics*, 13:529–556.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pages 23650–23678.

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*.

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, Singapore. Association for Computational Linguistics.

Raoyuan Zhao, Abdullatif Köksal, Ali Modarressi, Michael A. Hedderich, and Hinrich Schuetze. 2025. [Do we know what LLMs don’t know? a study of consistency in knowledge probing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23254–23280, Suzhou, China. Association for Computational Linguistics.

Chenghao Zhu, Nuo Chen, Yufei Gao, Yunyi Zhang, Prayag Tiwari, and Benyou Wang. 2025. [Is your LLM outdated? a deep look at temporal generalization](#). 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 7433–7457, Albuquerque, New Mexico. Association for Computational Linguistics.

A Related Work

Confidence Estimation and Calibration in LLMs Model calibration has been extensively studied in classification and question answering. Expected Calibration Error (ECE) and Brier Score are standard metrics for assessing the alignment between predicted confidence and empirical accuracy. [Jiang et al. \(2021\)](#) analyzed calibration in open-domain QA, while [Kadavath et al. \(2022\)](#) showed that large language models possess internal uncertainty signals correlated with correctness, even when these signals are not reflected in generated outputs. Recent work has further examined the reliability of confidence estimation and knowledge probing methods, showing that such signals can be highly inconsistent across prompts and approaches ([Zhao et al., 2025](#)).

More recent work has also studied hallucinations, abstention, and overconfidence in LLMs. For example, abstention-based approaches aim to allow models to withhold answers when uncertain ([Wen et al., 2025](#)), while [Zhang and Choi \(2023\)](#) address temporal misalignment by predicting fact duration and discarding volatile facts. Post-hoc calibration methods such as temperature scaling have also been widely studied for improving confidence reliability in neural networks ([Guo et al., 2017](#)). However, these approaches are typically evaluated under aggregate or stationary settings and do not explicitly analyze how calibration behavior varies across temporal dimensions. In contrast, our work examines whether confidence estimation and both global and temporally adaptive temperature scaling remain reliable for temporally grounded question answering.

Temporal Generalization and Knowledge Freshness A growing body of research has examined how language models generalize across time. [Lazaridou et al. \(2021\)](#) showed that neural language models exhibit significant performance degradation when evaluated on data from time periods outside their training distribution, particularly for future events. Subsequent work confirmed that simply increasing model size does not resolve this issue and that temporal distribution shift remains a fundamental challenge for static models. [Zhu et al. \(2025\)](#) further demonstrated that even state-of-the-art LLMs suffer from accuracy decay as temporal distance increases.

Several studies have proposed evaluation frameworks for time-sensitive question answering.

TimeQA (Chen et al., 2021) introduced questions whose answers depend on temporal context, while StreamingQA (Liska et al., 2022) and RealTimeQA (Kasai et al., 2023) evaluate models under continuously evolving information streams. FreshQA (Vu et al., 2024) and Daily Oracle (Zhu et al., 2025) extend this line of work by focusing on recent and rapidly changing world knowledge. While these efforts document temporal accuracy degradation, they largely treat outputs as correct or incorrect without analyzing the confidence associated with those predictions.

Related work has also explored time-aware language modeling. Dhingra et al. (2022) introduced TEMPLAMA, a benchmark for probing temporally scoped factual knowledge, and proposed time-aware pretraining methods to improve temporal memorization and future calibration. Their work highlighted temporal averaging, forgetting, and poor calibration on future facts. Our work builds on this motivation but focuses on systematic confidence calibration across time rather than temporal knowledge representation or pretraining.

Recent work has further examined the distinction between temporally stable and mutable knowledge. Pletenev et al. (2025) introduced EVER-GREENQA, a multilingual benchmark for identifying whether questions are evergreen or temporally mutable, showing that temporality-aware question classification can support dataset filtering and self-knowledge estimation. In contrast, our work does not aim to classify whether a question is mutable; instead, we study whether model confidence remains calibrated when answering temporally grounded questions whose correctness depends on time.

Temporal Reasoning and Robustness in Language Models Beyond knowledge freshness, LLMs also struggle with temporal reasoning and temporal robustness. Wallat et al. (2025) analyzed LLM robustness through temporal perturbation tests such as year shifts, temporal reversal, temporal removal, and relativization, showing that models are highly sensitive to temporal reformulations and often fail to robustly encode temporal information. Earlier work on temporal blind spots (Wallat et al., 2024) similarly identified temporal shifts, time invariance, temporal inertia, and referencing errors as recurring failure modes in temporal QA. Tan et al. (2024) introduced multi-hop temporal reasoning datasets and demonstrated that perfor-

TEMPLAMA Question Conversion Prompt

Convert the following sentence with "_X_" into a natural question:
{query}

Question:

Figure 5: Prompt used to convert TEMPLAMA cloze-style statements into questions.

mance drops substantially as temporal complexity increases.

These studies show that LLMs often ignore, misinterpret, or inconsistently use temporal constraints. Our work complements this literature by examining how such failures manifest in confidence estimation. Rather than asking only whether a temporally grounded answer is correct, we ask whether the model’s stated or inferred confidence appropriately reflects correctness across time.

Our work bridges research on calibration, temporal QA, and temporal robustness by focusing on *temporal calibration*: the alignment between model confidence and correctness conditioned on the temporal reference of a question. Unlike prior studies that evaluate temporal reliability primarily through accuracy, robustness tests, or mutable-question classification, we explicitly analyze confidence behavior over time using calibration metrics such as ECE and Brier Score. We further compare multiple confidence estimation strategies and evaluate whether post-hoc calibration methods remain effective under temporal distribution shift.

B Dataset Details

TEMPLAMA consists of temporally anchored cloze-style statements associated with explicit timestamps. For our experiments, we convert the cloze-style statements in the TEMPLAMA validation set into natural-language questions using LLaMA-3.3-70B, yielding **4,654** question–answer pairs annotated with a timestamp t indicating when the answer was valid. The prompt used for this conversion is shown in Figure 5, while Table 2 presents representative examples of cloze-style statements and their converted questions.

For ARCHIVALQA, we retain temporally explicit questions containing reference years between 1950 and 2010. To maintain temporal balance, we sample 100 questions per year, resulting in **6,100** evaluation instances.

Table 3 shows representative temporally

Dataset	Year	Cloze-Style Statements	Question
TempLAMA	2015	Lewis Hamilton plays for _X_.	Which team does Lewis Hamilton play for?
TempLAMA	2019	Michael Beasley plays for _X_.	Which team does Michael Beasley play for?

Table 2: Example of Cloze-Style Statements from TEMPLAMA converted into questions.

Dataset	Year	Question	Answer
TempLAMA	2015	Which team does Lewis Hamilton play for?	Mercedes
TempLAMA	2019	Which team does Michael Beasley play for?	Memphis Grizzlies
ArchivalQA	1957	Who was the mayor of West Berlin in 1957?	Brandt
ArchivalQA	1965	What news network did Louis Rukeyser join in 1965?	ABC News

Table 3: Example temporally grounded questions from ARCHIVALQA and TEMPLAMA.

grounded question-answer pairs from both datasets.

C Evaluation Metrics

We assess model performance and calibration using three primary metrics: Accuracy, Expected Calibration Error (ECE), and Brier Score.

Accuracy For each temporal slice t , accuracy is computed as the proportion of generated answers judged correct. For open-ended generation, we compute correctness using GPT-EVAL, since LLMs often produce correct answers in different phrasings that Exact Match (EM) would unfairly penalize.

In the GPT-EVAL framework, we prompt an LLM with the template shown in Figure 6 to determine whether a generated answer is semantically correct given the ground-truth answer. The model outputs a binary judgment: *Yes* or *No*. All GPT-EVAL judgments are performed using deterministic decoding ($T = 0$).

To improve evaluation robustness and reduce stochastic variance in LLM-based judgments, we evaluate each generated answer three times independently using the same GPT-EVAL prompt. The final correctness label is determined through majority voting across the three independent judgments. We additionally report Exact Match (EM) and Contains Match accuracy in Table 4 for comparison. EM scores are substantially lower due to lexical mismatch in open-ended answer generation, which motivates our use of GPT-EVAL across all experiments.

Expected Calibration Error (ECE) Expected Calibration Error (ECE) (Guo et al., 2017) measures the discrepancy between predicted confidence and empirical accuracy. We partition predictions

Evaluating Generated Answer using LLM

Question: <question>
Answer: <ground_truth>
Predicted Answer: <predicted_answer>
 Is the Predicted Answer correct? Choose between "Yes" or "No".

Figure 6: Prompt template used for GPT-EVAL-based answer verification. The placeholders <question>, <ground_truth>, and <predicted_answer> denote the input question, reference answer, and generated answer, respectively.

into M equally spaced confidence bins (we use $M = 10$). For a given bin B_m , the empirical accuracy is defined as:

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}[z_i = 1]$$

and the average predicted confidence is:

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} c_i.$$

The ECE is computed as the weighted average of the absolute difference between accuracy and confidence across all bins:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|.$$

To analyze temporal effects, we compute ECE_t independently for each temporal slice t , enabling us to track calibration drift over time.

Brier Score The Brier Score (Glenn et al., 1950) measures the mean squared error between the predicted confidence and the binary correctness outcome. Unlike ECE, it is a proper scoring rule that

Model	p(True)			Verbal Confidence			Token Probability			Self-Consistency		
	EM ↑	Contains ↑	GPT-EVAL ↑	EM ↑	Contains ↑	GPT-EVAL ↑	EM ↑	Contains ↑	GPT-EVAL ↑	EM ↑	Contains ↑	GPT-EVAL ↑
ArchivalQA												
Qwen-3 4B	4.43	13.47	30.08	4.52	11.95	23.89	5.36	12.86	25.50	4.57	11.98	25.53
Gemma-3 4B	1.56	18.17	21.78	3.89	16.57	21.24	4.92	16.36	21.32	5.14	15.94	24.28
Ministral-3 8B	3.38	22.42	36.85	5.53	17.41	34.06	5.97	17.78	32.40	2.08	3.60	10.56
LLaMA-3.1 8B	11.25	22.32	31.47	11.40	21.76	30.02	11.40	21.03	29.47	11.49	21.34	31.61
Gemma-3 27B	5.80	28.95	39.51	11.44	26.60	37.41	11.15	27.29	37.46	10.95	27.19	40.37
Qwen-3 30B	11.29	23.64	38.49	10.42	23.33	37.19	11.35	23.72	36.73	11.25	23.64	39.15
LLaMA-3.3 70B	15.57	32.18	46.01	15.69	31.88	44.81	15.48	31.56	44.37	16.41	31.56	45.62
Qwen-2.5 72B	13.98	28.54	46.41	15.25	28.07	41.56	15.36	28.54	41.27	15.16	28.49	43.33
TEMPLAMA												
Qwen-3 4B	3.46	7.80	8.34	2.92	5.52	8.66	3.87	6.27	10.49	3.89	5.67	12.20
Gemma-3 4B	1.05	10.70	13.77	2.79	9.05	13.04	5.09	9.82	13.64	3.61	9.84	14.87
Ministral-3 8B	2.30	15.38	13.32	5.74	13.79	13.97	5.65	13.00	14.29	3.65	5.05	9.24
LLaMA-3.1 8B	10.38	16.80	19.19	10.66	14.89	17.15	13.21	19.27	22.65	13.11	19.10	22.07
Gemma-3 27B	12.18	30.66	32.38	17.88	30.38	31.65	17.17	29.80	31.80	15.21	29.35	32.12
Qwen-3 30B	11.07	21.59	25.91	11.43	20.11	25.05	12.40	19.72	24.54	12.27	20.33	26.82
LLaMA-3.3 70B	25.93	38.89	40.46	23.66	34.62	37.02	27.01	39.26	41.53	27.72	38.68	41.43
Qwen-2.5 72B	15.15	25.87	30.40	18.05	25.46	28.90	18.59	26.30	30.00	18.37	25.48	29.63

Table 4: Comparison of EM, Contains Match, and GPT-EVAL accuracy across confidence estimation methods. EM substantially underestimates performance due to lexical mismatch in open-ended generation, motivating the use of GPT-EVAL for all calibration experiments.

penalizes instance-level overconfidence and under-confidence:

$$BS = \frac{1}{N} \sum_{i=1}^N (c_i - z_i)^2,$$

where $z_i \in \{0, 1\}$ denotes whether the i -th prediction is correct. Lower Brier Scores indicate better-calibrated predictions. Similarly to ECE, we compute BS_t independently for each temporal slice t .

D Models and Implementation Details

Table 5 summarizes the models evaluated in our experiments. All models are instruction-tuned and evaluated in a zero-shot setting without task-specific fine-tuning or retrieval augmentation. Unless required otherwise by the confidence estimation strategy, all models are evaluated using greedy decoding with temperature $T = 0$. For Self-Consistency, stochastic decoding is performed using temperature sampling with $T = 0.7$.

E Prompts Used in Main Experiments

This section provides the exact prompts used for answer generation and confidence estimation in the main experiments (Section 5). All prompts are applied in a zero-shot setting and conditioned on a reference year. Figures 7, 8, 9, and 10 show the prompts used for Verbal Confidence, P(True), Self-Consistency, and Token Probability Confidence, respectively. The placeholders {YEAR}, {QUESTION}, and, where applicable, {ANSWER} vary across instances.

Model	Category	Parameters	Inst. Tuned
Qwen-3	Small	4B	✓
Gemma-3	Small	4B	✓
Ministral-3	Medium	8B	✓
LLaMA-3.1	Medium	8B	✓
Gemma-3	Large	27B	✓
Qwen-3	Large	30B	✓
LLaMA-3.3	XL	70B	✓
Qwen-2.5	XL	72B	✓

Table 5: Instruction-tuned LLMs evaluated in our experiments, grouped by parameter scale.

F Additional Plots

This appendix presents additional calibration analyses and visualizations. Figure 12 reports temporal ECE trends for the remaining evaluated models, complementing the representative-model results shown in Figure 2 of Section 5.2. Corresponding Brier Score trends for the representative and remaining models are reported in Figures 11 and 13, respectively.

Figures 14 and 15 present reliability diagrams for all evaluated models and confidence estimation strategies.

Additional temperature scaling analyses are shown in Figures 16, 17, and 18. Figure 16 presents calibration error across different temperature values on the calibration split for representative models, illustrating that each model exhibits a distinct optimal temperature region. Figure 17 reports Brier Score before and after applying Global Temperature Scaling on held-out test sets, showing calibration improvements consistent with the ECE

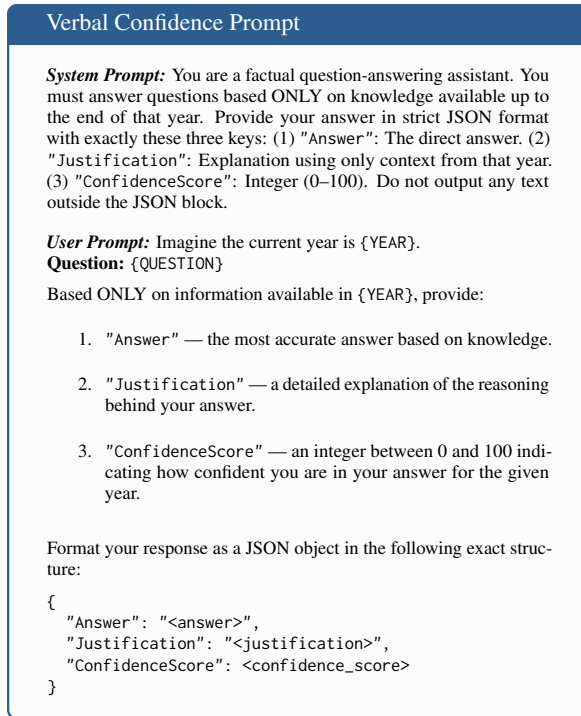
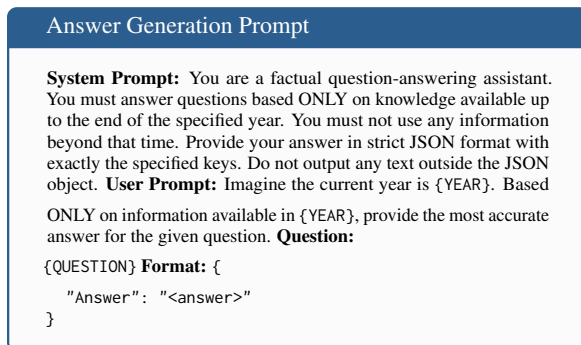
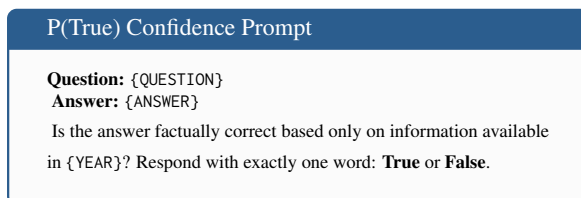


Figure 7: Prompt used to elicit verbalized confidence alongside an answer and justification under temporal constraints.



(a) Prompt used for temporally grounded answer generation.



(b) Prompt used to estimate internal model confidence $P(\text{True})$.

Figure 8: Prompts used in the temporal QA pipeline. (a) generates an answer conditioned on knowledge available up to a specified year, while (b) estimates whether the generated answer is factually correct under the same temporal constraint.

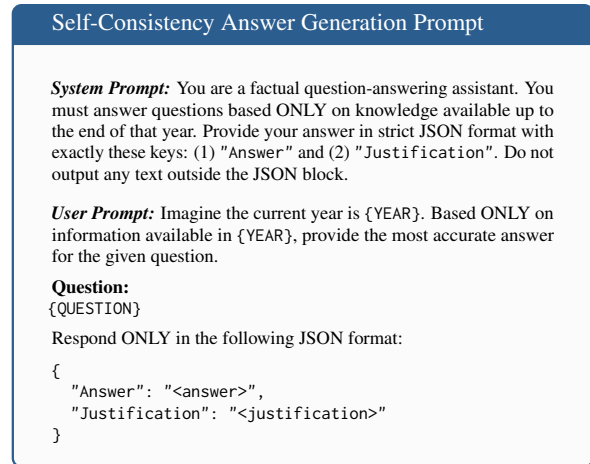


Figure 9: Prompt used to generate multiple temporally grounded answers for self-consistency-based confidence estimation.

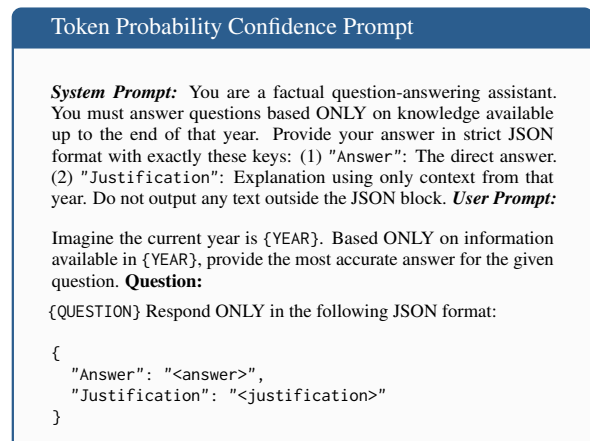


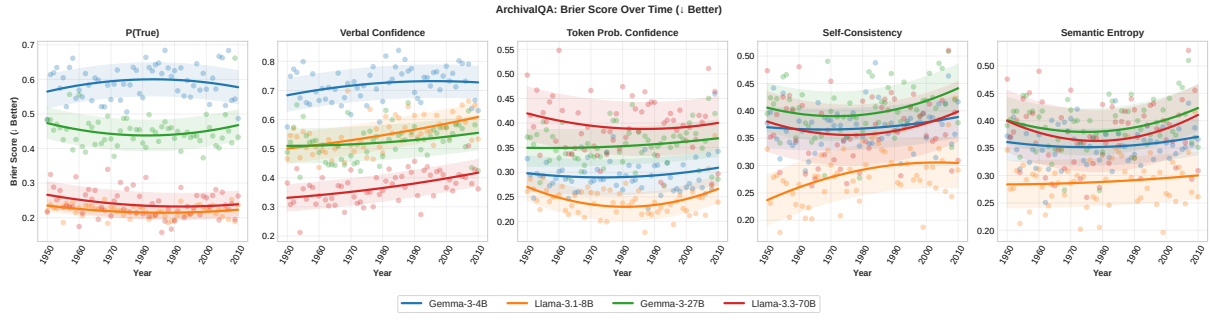
Figure 10: Prompt used to generate temporally grounded answers for computing token probability confidence estimation.

reductions discussed in the main paper. Figure 18 visualizes the optimal temperature learned independently for each temporal bucket, highlighting period-specific variation in calibration behavior.

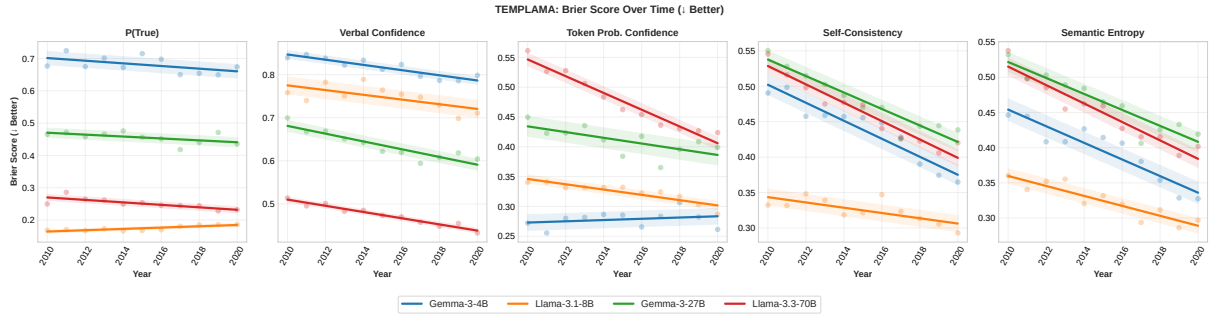
G Additional Experiments

G.1 Non-Temporal Baseline Comparison

To determine whether the observed miscalibration patterns are specific to temporal QA or reflect general overconfidence in factual QA, we evaluate Gemma-3-4B and LLaMA-3.3-70B on stratified samples of 500 questions using $P(\text{True})$, Verbal Confidence, Token Probability Confidence, and Self-Consistency under temporally neutral prompts. We consider three settings: (1) **ArchivalQA (no-year)**, where reference years are removed from

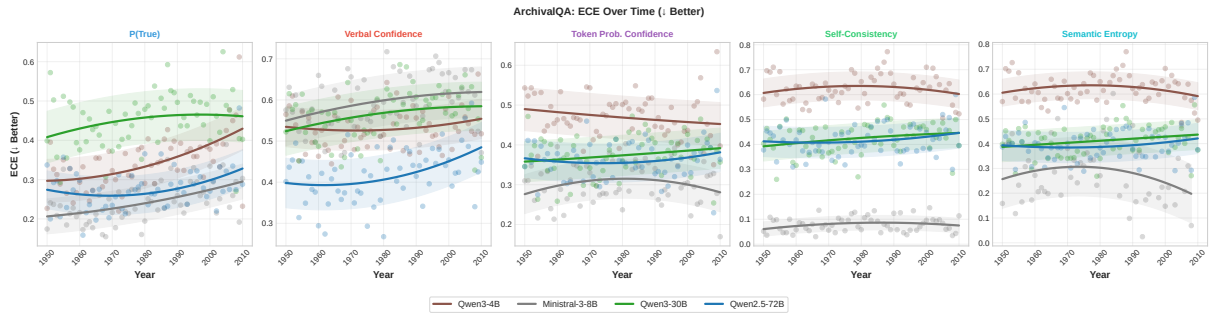


(a) ARCHIVALQA

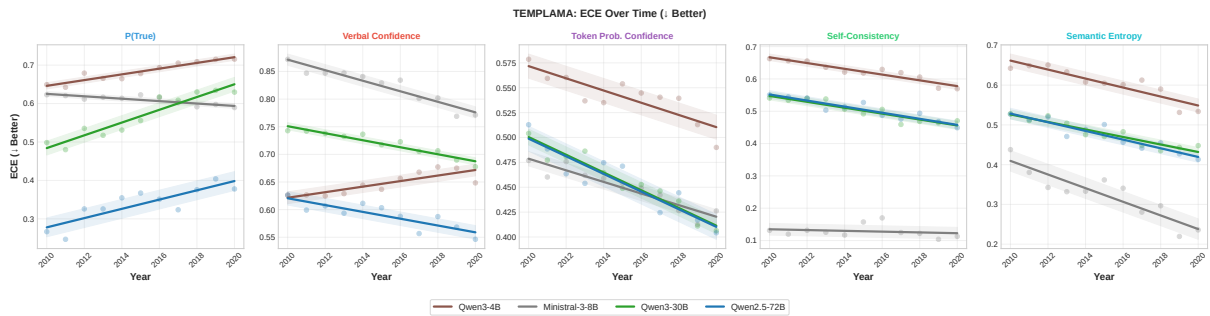


(b) TEMPLAMA

Figure 11: Temporal evolution of Brier Score across representative models and confidence estimation strategies on (a) ARCHIVALQA (1950–2010) and (b) TEMPLAMA (2010–2020). Lower values indicate better calibration. Brier Score varies substantially across reference periods, models, and confidence estimators, with ARCHIVALQA exhibiting heterogeneous temporal trajectories and TEMPLAMA generally showing lower calibration error toward more recent reference years.



(a) ARCHIVALQA



(b) TEMPLAMA

Figure 12: Temporal evolution of ECE across the remaining evaluated models on (a) ARCHIVALQA (1950–2010) and (b) TEMPLAMA (2010–2020).

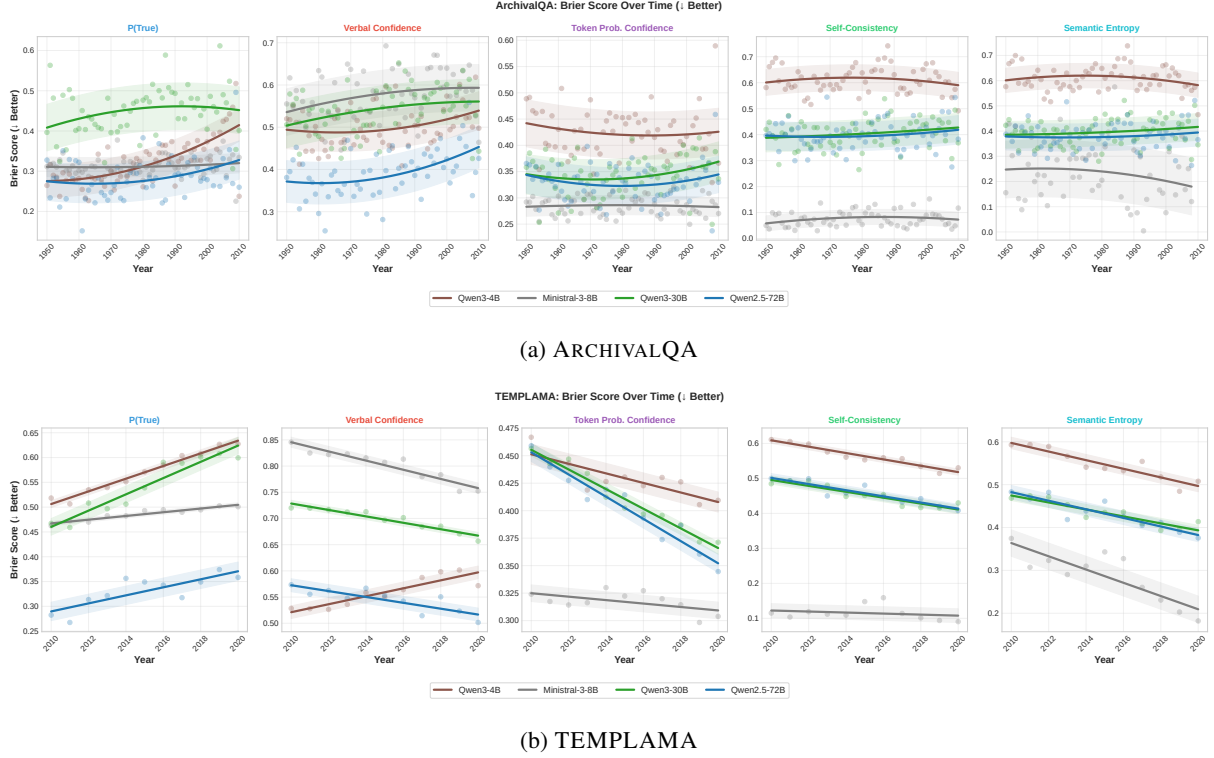


Figure 13: Temporal evolution of Brier Score across the remaining evaluated models on (a) ARCHIVALQA (1950–2010) and (b) TEMPLAMA (2010–2020).

prompts; (2) **TEMPLAMA (no-year)**, where temporal grounding is omitted; and (3) **TriviaQA** (Joshi et al., 2017), a standard non-temporal factual QA benchmark. Table 6 reports the resulting calibration performance.

Calibration differs between temporal and non-temporal factual QA. Across both models, calibration performance is consistently stronger on TriviaQA than on the temporal datasets. For example, for LLaMA-3.3-70B, verbal confidence achieves an ECE of only **0.052** on TriviaQA, compared to **0.365** on ArchivalQA and **0.598** on TEMPLAMA. Similar improvements are observed for Self-Consistency and token-based probabilities. These results suggest that temporal factual QA poses a substantially greater calibration challenge than standard non-temporal factual QA.

Temporal framing improves calibration reliability. Removing explicit year information generally worsens calibration on both temporal datasets. Compared to the original temporally grounded prompts, the no-year setting produces higher calibration error across several methods and models. For example, LLaMA-3.3-70B exhibits higher P(True) ECE under the no-year setting than under the corresponding temporally grounded setting (Ta-

ble 7). This trend suggests that temporal framing helps anchor model uncertainty to a specific factual period, improving calibration reliability.

Interestingly, no single confidence estimation strategy dominates across all datasets. While P(True) generally performs strongly on temporal QA, Self-Consistency and Verbal Confidence achieve comparatively low ECE on TriviaQA for LLaMA-3.3-70B. This suggests that the effectiveness of confidence estimation strategies depends strongly on the underlying factual setting and temporal characteristics of the task.

G.2 Prompt Sensitivity Analysis

To assess whether our findings depend on the specific temporal prompt framing used in the main experiments, we conduct a prompt sensitivity analysis across three temporal prompt variants: **Variant A** (original): “*Imagine the current year is {year}. {question}*”, **Variant B** (natural phrasing): the year is embedded directly into the question text (e.g., “*Which team did Lewis Hamilton play for in 2015?*”), and **Variant C** (as-of framing): “*As of {year}, answer the following question: {question}*”.

We evaluate all variants on stratified samples of 500 questions from each dataset using Gemma-3-4B and LLaMA-3.3-70B across P(True), Verbal

ARCHIVALQA Reliability Diagrams

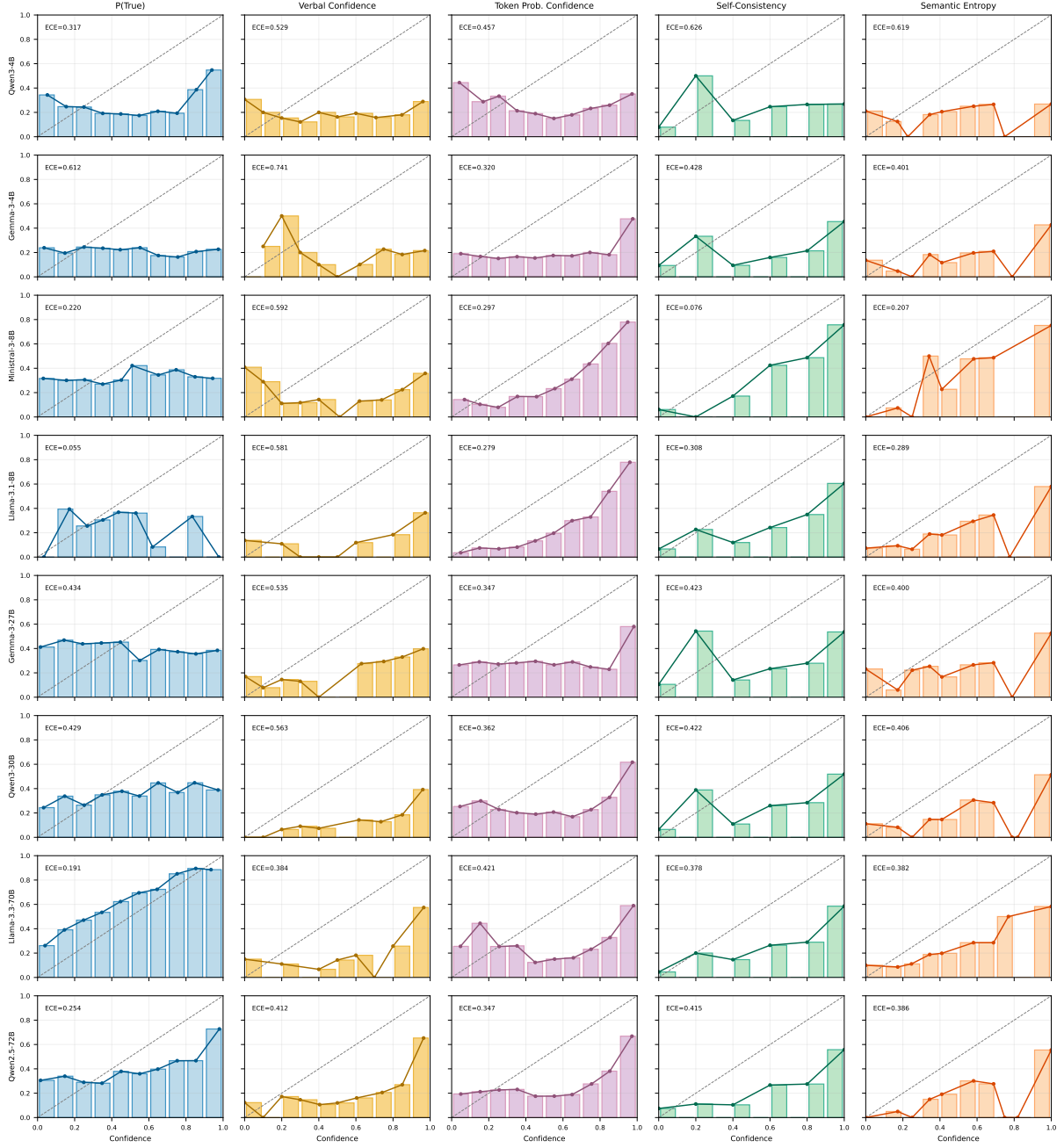


Figure 14: Reliability diagrams for all evaluated models and confidence estimation strategies on ARCHIVALQA. The dashed diagonal denotes perfect calibration, where predicted confidence matches empirical accuracy. Calibration behavior varies substantially across models and confidence estimators, consistent with the aggregate results in Table 1. Verbal Confidence is generally overconfident, while Token Probability Confidence, Self-Consistency, P(True), and Semantic Entropy exhibit model-dependent calibration behavior. ECE values for each model–method combination are shown within each subplot.

Confidence, Token Probabilities Confidence, and Self-Consistency. Full prompt templates are provided in Appendix H.

Table 7 reports calibration performance across all prompt variants. Overall, the main calibration patterns remain qualitatively consistent across

prompt formulations.

Verbal Confidence remains poorly calibrated across prompts. Verbal Confidence exhibits substantial calibration error across prompt variants, particularly for Gemma-3-4B, where ECE exceeds 0.75 under all three variants on both datasets. This

TEMPLAMA Reliability Diagrams

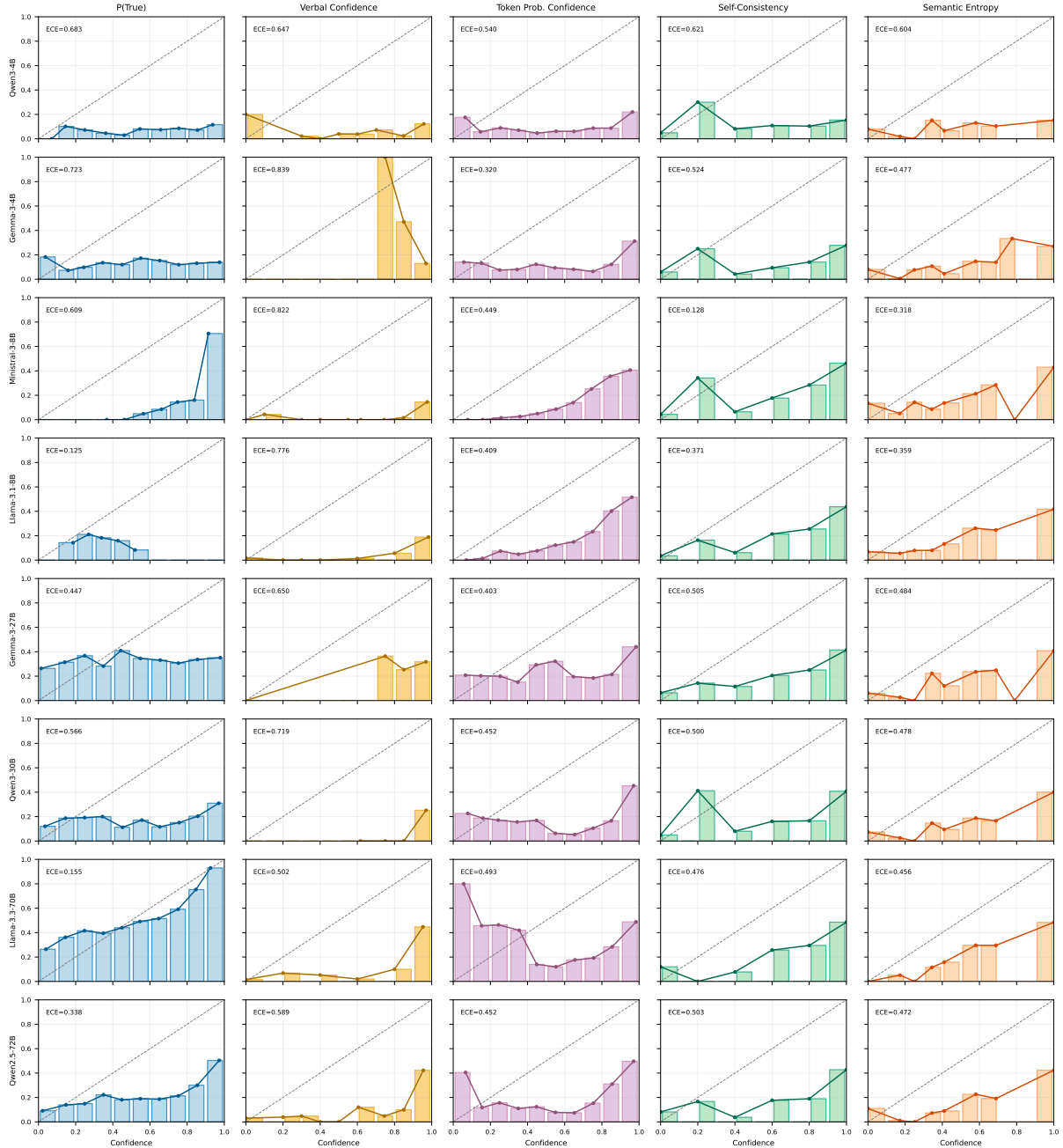
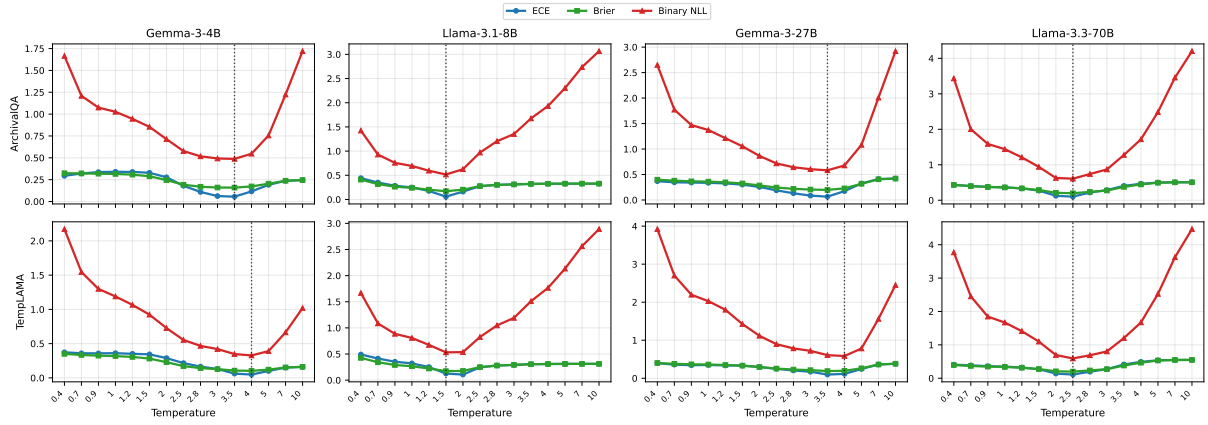


Figure 15: Reliability diagrams for all evaluated models and confidence estimation strategies on TEMPLAMA. The dashed diagonal denotes perfect calibration, where predicted confidence matches empirical accuracy. Calibration behavior varies substantially across models and confidence estimators, with pronounced overconfidence in several settings. Verbal Confidence is generally poorly calibrated, while the remaining estimators exhibit model-dependent deviations from the diagonal. ECE values for each model–method combination are shown within each subplot.

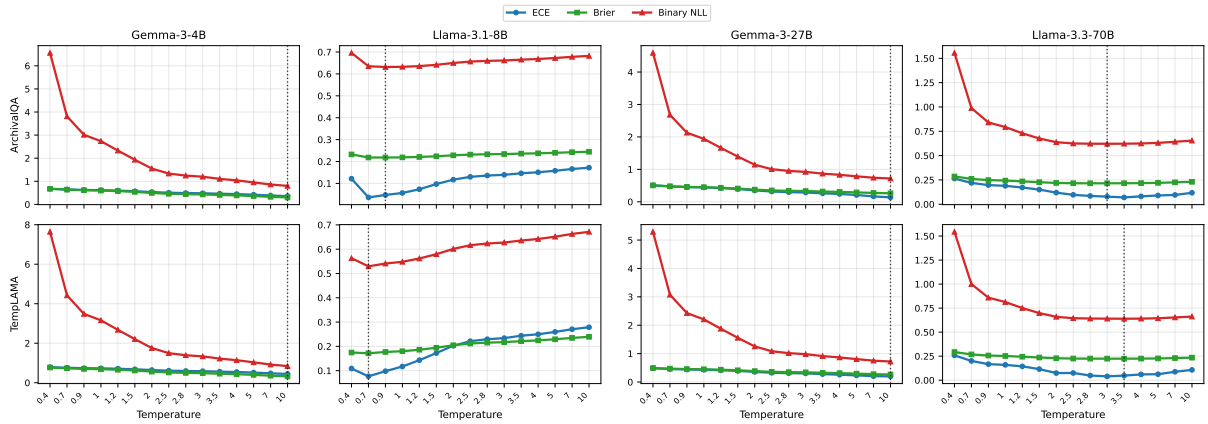
indicates that the poor calibration of verbal confidence is not specific to a single temporal prompt formulation.

Estimator rankings remain model-dependent across prompts. Across prompt variants, Token Probability Confidence achieves the lowest calibration error for Gemma-3-4B on both datasets,

whereas P(True) achieves the lowest calibration error for LLaMA-3.3-70B. For example, LLaMA-3.3-70B achieves P(True) ECE values as low as **0.116** on ARCHIVALQA and **0.129** on TEMPLAMA. This pattern remains qualitatively stable across the evaluated prompt formulations and is consistent with the main-paper finding that no

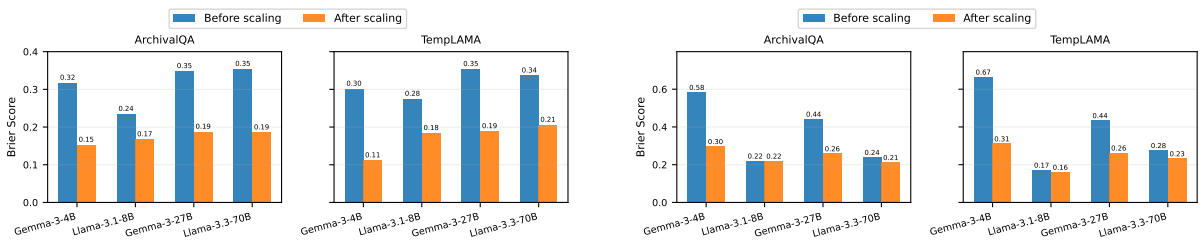


(a) Token Probability Confidence



(b) P(True)

Figure 16: Calibration performance across different temperature values on the calibration split for (a) Token Probability Confidence and (b) P(True) across representative models and datasets. ECE, Brier Score, and binary NLL vary with temperature across model–dataset pairs; the global temperature is selected by minimizing binary NLL on the calibration split.



(a) Token Probability Confidence

(b) P(True)

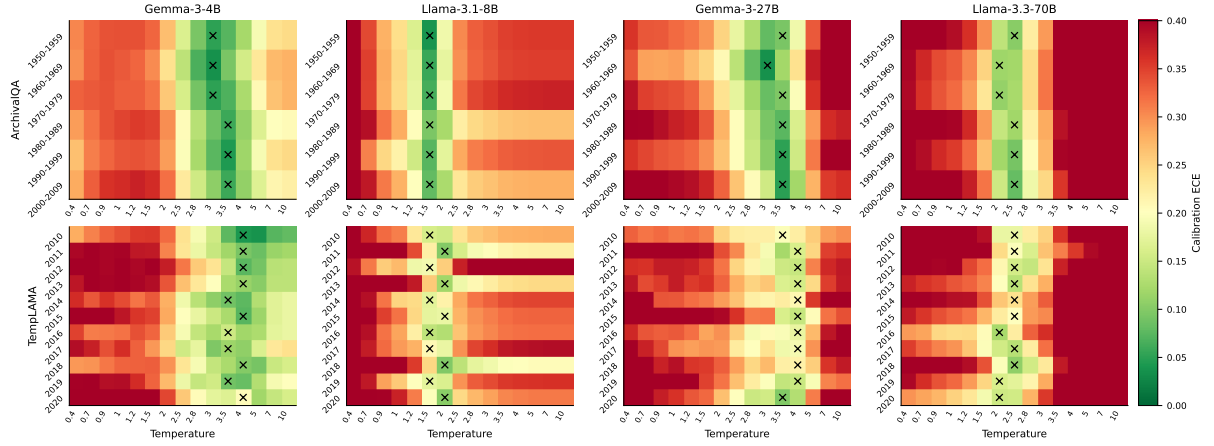
Figure 17: Brier Score before and after applying Global Temperature Scaling on held-out test sets for (a) Token Probability Confidence and (b) P(True). Global Temperature Scaling improves calibration for both confidence signals, consistent with the ECE reductions shown in Figure 3.

single confidence estimator uniformly dominates across models and datasets.

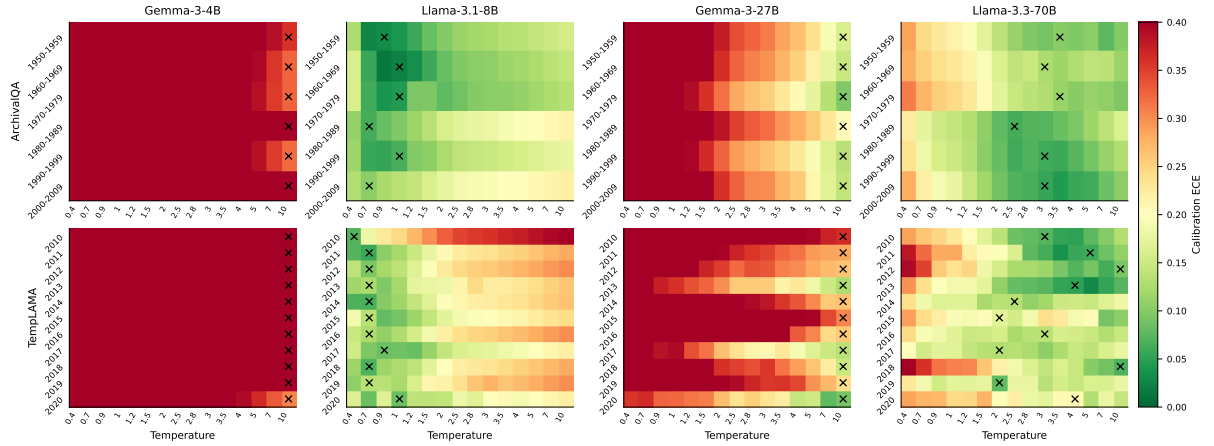
Temporal calibration patterns are stable across prompt formulations. Although calibration values vary across prompt variants, the broader method-wise patterns remain qualitatively stable.

In particular, changing the temporal phrasing does not systematically change the best-performing confidence estimator for either model. The variation across prompt formulations is also generally smaller than the calibration differences observed between confidence estimation methods.

These results suggest that the temporal calibra-



(a) Token Probability Confidence



(b) $P(\text{True})$

Figure 18: Optimal temperature parameters learned independently for each temporal bucket for (a) Token Probability Confidence and (b) $P(\text{True})$ across representative models and datasets. Different decades and years frequently prefer different temperature values for both confidence signals, indicating that calibration behavior is non-stationary across temporal periods.

tion patterns observed in our experiments are robust to the evaluated prompt formulations and are not driven solely by a particular choice of temporal phrasing.

H Additional Experiment Prompts

In this appendix, we present the prompts used in the additional experiments. Figure 19 provides the baseline prompts without temporal conditioning. Figures 20, 21, 22, and 23 show the prompt variants used for Verbal Confidence, $P(\text{True})$, Self-Consistency, and Token Probability Confidence respectively, in the prompt sensitivity analysis.

I Temporal Miscalibration Case Studies

To complement the aggregate calibration analyses presented in the main paper, we provide rep-

resentative qualitative examples of temporal miscalibration across different confidence estimation strategies. These examples illustrate several recurring failure modes observed throughout our experiments, including severe verbalized overconfidence, incorrect temporal associations, agreement-driven overconfidence under Self-Consistency, and unreliable token-level uncertainty estimation. Together, these cases demonstrate that temporal miscalibration can arise through multiple distinct mechanisms beyond simple factual inaccuracy.

Overall, these examples highlight that high confidence in temporally grounded QA often reflects linguistic fluency, semantic agreement, or stale factual associations rather than reliable temporal correctness. Even when models produce factually correct answers, confidence estimates may remain

Model	Method	ArchivalQA			TEMPLAMA			TriviaQA		
		Acc.	ECE ↓	Brier ↓	Acc.	ECE ↓	Brier ↓	Acc.	ECE ↓	Brier ↓
Gemma-3-4B	P(True)	20.0	0.636	0.614	9.8	0.744	0.703	57.8	0.290	0.338
	Verbal Confidence	18.4	0.804	0.795	8.8	0.907	0.904	57.7	0.415	0.414
	Token Probability	21.0	0.397	0.334	11.0	0.371	0.288	56.1	0.236	0.270
	Self-Consistency	19.4	0.528	0.476	9.6	0.610	0.503	59.2	0.289	0.315
LLaMA-3.3-70B	P(True)	43.3	0.191	0.248	23.7	0.227	0.235	92.7	0.327	0.224
	Verbal Confidence	44.2	0.365	0.340	23.8	0.598	0.548	93.1	0.052	0.051
	Token Probability	48.9	0.381	0.358	33.2	0.562	0.543	91.9	0.059	0.077
	Self-Consistency	43.2	0.440	0.417	24.4	0.604	0.586	91.2	0.062	0.070

Table 6: Calibration performance under temporally neutral prompts across ARCHIVALQA, TEMPLAMA, and TRIVIAQA. Lower ECE and Brier Scores indicate better calibration, with the lowest values within each model–method row shown in bold. Calibration is consistently stronger on non-temporal QA across nearly all evaluated settings, suggesting that temporally evolving factual knowledge presents a substantially greater calibration challenge.

Dataset	Model	Method	Variant A			Variant B			Variant C		
			Acc.	ECE	Brier	Acc.	ECE	Brier	Acc.	ECE	Brier
ArchivalQA	Gemma-3-4B	P(True)	19.1	0.543	0.520	19.4	0.599	0.570	20.9	0.629	0.608
		Verbal Confidence	18.8	0.762	0.734	19.2	0.792	0.781	20.3	0.777	0.764
		Token Probability	19.6	0.356	0.305	22.9	0.325	0.293	22.1	0.342	0.301
		Self Consistency	18.0	0.482	0.421	18.6	0.539	0.472	19.2	0.496	0.433
	LLaMA-3.3-70B	P(True)	39.9	0.127	0.202	44.1	0.201	0.244	43.3	0.116	0.209
		Verbal Confidence	40.7	0.414	0.394	43.3	0.399	0.367	41.5	0.440	0.408
		Token Probability	48.6	0.380	0.356	51.5	0.338	0.329	47.0	0.381	0.350
		Self Consistency	41.6	0.462	0.437	42.0	0.446	0.421	41.4	0.447	0.432
TEMPLAMA	Gemma-3-4B	P(True)	16.8	0.697	0.656	11.0	0.718	0.666	15.6	0.703	0.664
		Verbal Confidence	16.0	0.808	0.785	10.6	0.887	0.880	13.2	0.850	0.835
		Token Probability	14.2	0.317	0.285	12.2	0.392	0.316	14.9	0.295	0.282
		Self Consistency	15.8	0.600	0.526	11.2	0.670	0.585	14.6	0.573	0.488
	LLaMA-3.3-70B	P(True)	32.4	0.196	0.246	32.9	0.129	0.202	33.1	0.167	0.227
		Verbal Confidence	35.3	0.541	0.511	36.4	0.520	0.498	38.3	0.530	0.508
		Token Probability	33.5	0.558	0.544	34.1	0.568	0.543	34.0	0.575	0.553
		Self Consistency	35.8	0.535	0.533	36.6	0.531	0.517	36.6	0.481	0.467

Table 7: Prompt sensitivity results across temporal prompt variants. We report Accuracy (Acc.), ECE, and Brier Score for Variants A, B, and C across both datasets and models. Lower ECE and Brier Scores indicate better calibration. Bold indicates the lowest ECE or Brier Score across variants within each model–method–dataset combination.

poorly aligned with actual uncertainty under temporal shift.

Verbal Confidence

System Prompt:
You are a factual question-answering assistant. Provide your answer in strict JSON format with exactly these three keys:

1. "Answer": The direct answer.
2. "Justification": Explanation of your reasoning.
3. "ConfidenceScore": Integer (0–100).

Do not output any text outside the JSON block.

User Prompt:
Question: {QUESTION}

Format your response as a JSON object:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>",
  "ConfidenceScore": <confidence_score>
}
```

(a) Verbal Confidence.

P(True)

Answer Generation
System Prompt:
You are a factual question-answering assistant. Provide your answer in strict JSON format with exactly these keys:

1. "Answer": The direct answer.
2. "Justification": Brief explanation of your reasoning.

Do not output any text outside the JSON block.

User Prompt:
Question: {QUESTION}

Format your response as a JSON object:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>"
}
```

Verification Prompt:
Question: {QUESTION}
Answer: {ANSWER}
Is this answer correct?
Respond with exactly one word: **True** or **False**.

(b) P(True).

Self-Consistency

System Prompt:
You are a factual question-answering assistant. Provide your answer in strict JSON format with exactly these keys:

1. "Answer"
2. "Justification"

User Prompt:
Question: {QUESTION}

Respond **ONLY** in the following JSON format:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>"
}
```

(c) Self-Consistency.

Token Probability Confidence

System Prompt:
You are a factual question-answering assistant. Provide your answer in strict JSON format with exactly these two keys:

1. "Answer": The direct answer.
2. "Justification": Brief explanation of your reasoning.

Do not output any text outside the JSON block.

User Prompt:
Question: {QUESTION}

Based **ONLY** on information available, provide:

1. "Answer" — the most accurate answer based on knowledge.
2. "Justification" — a brief explanation.

Format your response as a JSON object in the following exact structure:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>"
}
```

(d) Token Probability Confidence.

Figure 19: Baseline prompts used in the non-temporal experiments without explicit temporal conditioning. Panels show the prompts used for (a) Verbal Confidence, (b) P(True), (c) Self-Consistency, and (d) Token Probability Confidence. For P(True), the generated answer is subsequently evaluated using the binary verification prompt shown in panel (b).

Dataset	Example	Prediction Behavior	Conf.	Analysis
ARCHIVALQA	“Who chaired the Committee of Conscience Against Apartheid in 1966?”	Predicts that no chair had been formally appointed instead of “A. Philip Randolph”	98%	The model produces a historically plausible but factually incorrect explanation while remaining extremely confident, illustrating severe overconfidence in verbalized confidence estimation under temporal uncertainty.
TEMPLAMA	“Which team does Marcus Camby play for?” (2013)	Predicts “Brooklyn Nets” instead of “New York Knicks” or “Toronto Raptors”	73%	The model assigns high confidence to an incorrect temporal association, illustrating overconfidence on dynamically changing factual knowledge.
ARCHIVALQA	“What organization did Yugoslavia join in 2000?”	All sampled generations consistently predict that Yugoslavia joined no organization	100%	Self-Consistency assigns maximal confidence because all sampled generations agree semantically despite being factually incorrect. This suggests that generation agreement primarily captures answer stability rather than correctness.
ARCHIVALQA	“What French aluminum producer did Alcan acquire in 2003?”	Correctly predicts “Pechiney”	0.002	Despite producing the correct answer, the model assigns near-zero token probability confidence. This suggests that token-level likelihood may fail to reflect semantic correctness or historical certainty.

Table 8: Representative temporal calibration failure modes observed across confidence estimation strategies. The examples illustrate verbalized overconfidence, temporal hallucination, agreement-driven overconfidence, and unreliable token-level uncertainty estimation under temporal shift.

Variant A (Explicit Temporal Constraint)

System Prompt: You are a factual question-answering assistant. You must answer questions based ONLY on knowledge available up to the end of that year. Provide your answer in strict JSON format with exactly these three keys:

1. "Answer": The direct answer.
2. "Justification": Explanation using only context from that year.
3. "ConfidenceScore": Integer (0–100).

Do not output any text outside the JSON block.

User Prompt: Imagine the current year is {YEAR}.

Question: {QUESTION}

Based ONLY on information available in {YEAR}, provide:

1. "Answer" — the most accurate answer based on knowledge.
2. "Justification" — a detailed explanation of the reasoning behind your answer.
3. "ConfidenceScore" — an integer between 0 and 100 indicating how confident you are in your answer for the given year.

Format your response as a JSON object:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>",
  "ConfidenceScore": <confidence_score>
}
```

(a) Explicit temporal constraint in both system and user prompts.

Variant B (Temporal Context in Question)

System Prompt: You are a factual question-answering assistant. Provide your answer in strict JSON format with exactly these three keys:

1. "Answer": The direct answer.
2. "Justification": Explanation of your reasoning.
3. "ConfidenceScore": Integer (0–100).

Do not output any text outside the JSON block.

User Prompt: Question: {QUESTION_WITH_YEAR}

Provide:

1. "Answer" — the most accurate answer.
2. "Justification" — a detailed explanation of your reasoning.
3. "ConfidenceScore" — an integer between 0 and 100 indicating how confident you are.

Format your response as a JSON object:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>",
  "ConfidenceScore": <confidence_score>
}
```

(b) Temporal information embedded directly in the question.

Variant C (Prefixed Temporal Framing)

System Prompt: You are a factual question-answering assistant. Provide your answer in strict JSON format with exactly these three keys:

1. "Answer": The direct answer.
2. "Justification": Explanation of your reasoning.
3. "ConfidenceScore": Integer (0–100).

Do not output any text outside the JSON block.

User Prompt: As of {YEAR}, answer the following question:

Question: {QUESTION}

Provide:

1. "Answer" — the most accurate answer.
2. "Justification" — a detailed explanation of your reasoning.
3. "ConfidenceScore" — an integer between 0 and 100 indicating how confident you are.

Format your response as a JSON object:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>",
  "ConfidenceScore": <confidence_score>
}
```

(c) Temporal framing introduced as a prefix to the question.

Figure 20: Verbal Confidence prompt variants used in the prompt sensitivity analysis. The variants differ in how temporal information is introduced: (a) through explicit temporal constraints in both the system and user prompts, (b) directly within the question, and (c) through prefixed temporal framing.

Variant A (Explicit Temporal Constraint)

Answer Generation:
System Prompt: You are a factual question-answering assistant. You must answer questions based ONLY on knowledge available up to the end of that year. Provide your answer in strict JSON format with exactly this key:
1. "Answer": The direct answer.
Do not output any text outside the JSON block.

User Prompt: Imagine the current year is {YEAR}.
Question: {QUESTION}
Based ONLY on information available in {YEAR}, provide:
1. "Answer" — the most accurate answer based on knowledge.
Format your response as a JSON object:

```
{  
  "Answer": "<answer>"  
}
```

Verification Prompt:
Question: {QUESTION}
Answer: {ANSWER}
Is the answer factually correct based only on information available in {YEAR}?
Respond with exactly one word: **True** or **False**.

(a) Explicit temporal conditioning during both answer generation and verification.

Variant B (Temporal Context in Question)

Answer Generation:
System Prompt: You are a factual question-answering assistant. Provide your answer in strict JSON format with exactly this key:
1. "Answer": The direct answer.
Do not output any text outside the JSON block.

User Prompt: Question: {QUESTION_WITH_YEAR}
Format your response as a JSON object:

```
{  
  "Answer": "<answer>"  
}
```

Verification Prompt:
Question: {QUESTION}
Answer: {ANSWER}
Is this answer correct?
Respond with exactly one word: **True** or **False**.

(b) Temporal information is embedded directly in the generation question, while verification uses the original question without explicit temporal conditioning.

Variant C (Prefixed Temporal Framing)

Answer Generation:
System Prompt: You are a factual question-answering assistant. Provide your answer in strict JSON format with exactly this key:
1. "Answer": The direct answer.
Do not output any text outside the JSON block.

User Prompt: As of {YEAR}, answer the following question:
Question: {QUESTION}
Format your response as a JSON object:

```
{  
  "Answer": "<answer>"  
}
```

Verification Prompt:
As of {YEAR}:
Question: {QUESTION}
Answer: {ANSWER}
Is this answer correct?
Respond with exactly one word: **True** or **False**.

(c) Temporal framing is introduced as a prefix during both answer generation and verification.

Figure 21: P(True) prompt variants used in the prompt sensitivity analysis. Each variant first generates an answer and then estimates confidence through binary verification. The variants differ in how temporal context is introduced during answer generation and verification.

Variant A (Explicit Temporal Constraint)

System Prompt: You are a factual question-answering assistant. You must answer questions based ONLY on knowledge available up to the end of that year. Provide your answer in strict JSON format with exactly these keys:

1. "Answer"
2. "Justification"

User Prompt: Imagine the current year is {YEAR}.
Question: {QUESTION}
Respond ONLY in the following JSON format:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>"
}
```

(a) Explicit temporal constraint in both system and user prompts.

Variant B (Temporal Context in Question)

System Prompt: You are a factual question-answering assistant. Provide your answer in strict JSON format with exactly these keys:

1. "Answer"
2. "Justification"

User Prompt: Question: {QUESTION_WITH_YEAR}
Respond ONLY in the following JSON format:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>"
}
```

(b) Temporal information embedded directly in the question.

Variant C (Prefixed Temporal Framing)

System Prompt: You are a factual question-answering assistant. Provide your answer in strict JSON format with exactly these keys:

1. "Answer"
2. "Justification"

User Prompt: As of {YEAR}, answer the following question:
Question: {QUESTION}
Respond ONLY in the following JSON format:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>"
}
```

(c) Temporal framing introduced as a prefix to the question.

Figure 22: Self-Consistency prompt variants used in the prompt sensitivity analysis. Each variant is sampled multiple times to estimate confidence from agreement across generated answers. The variants differ in how temporal context is introduced: (a) through an explicit temporal constraint, (b) directly within the question, and (c) through prefixed temporal framing.

Variant A (Explicit Temporal Constraint)

System Prompt: You are a factual question-answering assistant. You must answer questions based ONLY on knowledge available up to the end of that year. Provide your answer in strict JSON format with exactly these two keys:

1. "Answer": The direct answer.
2. "Justification": Explanation using only context from that year.

Do not output any text outside the JSON block.

User Prompt: Imagine the current year is {YEAR}.
Question: {QUESTION}

Based ONLY on information available in {YEAR}, provide:

1. "Answer" — the most accurate answer based on knowledge.
2. "Justification" — a brief explanation.

Format your response as a JSON object in the following exact structure:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>"
}
```

(a) Explicit temporal constraint in both system and user prompts.

Variant B (Temporal Context in Question)

System Prompt: You are a factual question-answering assistant. You must answer questions based ONLY on your available knowledge. Provide your answer in strict JSON format with exactly these two keys:

1. "Answer": The direct answer.
2. "Justification": Explanation using only context from that year.

Do not output any text outside the JSON block.

User Prompt: Question: {QUESTION_WITH_YEAR}

Provide:

1. "Answer" — the most accurate answer based on knowledge.
2. "Justification" — a brief explanation.

Format your response as a JSON object in the following exact structure:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>"
}
```

(b) Temporal information embedded directly in the question.

Variant C (Prefixed Temporal Framing)

System Prompt: You are a factual question-answering assistant. You must answer questions based ONLY on knowledge available up to the end of that year. Provide your answer in strict JSON format with exactly these two keys:

1. "Answer": The direct answer.
2. "Justification": Explanation using only context from that year.

Do not output any text outside the JSON block.

User Prompt: As of {YEAR}, answer the following question:
Question: {QUESTION}

Provide:

1. "Answer" — the most accurate answer based on knowledge.
2. "Justification" — a brief explanation.

Format your response as a JSON object in the following exact structure:

```
{
  "Answer": "<answer>",
  "Justification": "<justification>"
}
```

(c) Temporal framing introduced as a prefix to the question.

Figure 23: Token Probability Confidence prompt variants used in the prompt sensitivity analysis. The variants differ in how temporal context is introduced: (a) through an explicit temporal constraint, (b) directly within the question, and (c) through prefixed temporal framing. Confidence is computed from the token probabilities of the generated answer.