Diagnosing LLM Arbitration Behavior over Pre-evidence Epistemic States in RAG-based Fact-Checking

Anonymous ACL submission

Abstract

In RAG-based fact-checking, LLMs are increasingly used as verifiers to check given claims against retrieved evidence. Their parametric knowledge can induce pre-evidence tendencies that may conflict with the retrieved context, yet existing evaluation frameworks do not characterize such prior-context discrepancy or measure how verifiers arbitrate between parametric and contextual signals. We introduce PAVE (*Prior-Aware Verifier Evaluation*), a diagnostic testbed that stratifies an LLM verifier into four epistemic states based on the correctness and confidence of its pre-evidence prior and evaluates its arbitration behavior on this new benchmark, i.e., whether it persists in correct prior under misleading evidence, and whether it corrects wrong prior when accurate evidence is provided. Experiments across seven LLMs reveal unreliable and highly model-dependent prior-context arbitration, highlighting the importance of verifier selection for real-world RAG-based fact-checking applications. Based on these findings, we propose a lightweight JSD-based test-time arbitration method that improves factual reliability without modifying the underlying model, achieving competitive performance across diverse LLM families.

1 Introduction

With the rapid development of large language models (LLMs), which encode broad world knowledge and support reasoning over complex claims and evidence (Brown et al., 2020; Ouyang et al., 2022), LLM-based fact-checking has become increasingly prominent. Existing systems broadly follow two routes: judging claims using the model's parametric knowledge without explicit retrieval (Jiang et al., 2020; Roberts et al., 2020; Zhao et al., 2023), or augmenting the model with retrieved evidence in a RAG framework to compensate for static, incomplete, or outdated parametric memory (Qin et al., 2024; Shi et al., 2024; Pan et al., 2023).

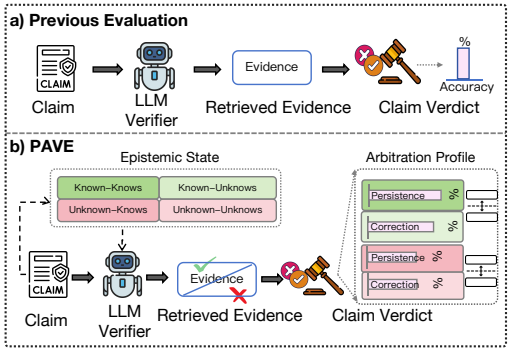


Figure 1: **Overview of PAVE.** Conventional evaluation judges verifiers by final-verdict accuracy on retrieved evidence (a). PAVE (b) instead characterizes verifiers by both their pre-retrieval *epistemic state* (four Knowledge-Boundary categories) and their *arbitration profile* under prior-context discrepancy (persistence & correction).

Yet RAG-based verification remains shaped by the interaction between an LLM's parametric knowledge and retrieved external context. This raises two evaluation challenges. **First**, when prior knowledge disagrees with retrieved context, as can occur with noisy retrieval or temporally outdated knowledge (Shi et al., 2023; Luu et al., 2022; Xie et al., 2023), the verifier must arbitrate between parametric and contextual signals. We term this condition *prior-context discrepancy* (PCD), an important but underexplored setting in real-world fact-checking. Existing evaluation pipelines are not designed to handle PCD: standard fact-checking *benchmarks* (Thorne et al., 2018; Aly et al., 2021; Anand et al., 2024; Ma et al., 2024) pair claims with clean gold-standard evidence, structurally precluding any discrepancy; and common evaluation protocols focus on final verdict correctness, leaving how the verifier arbitrates between prior knowledge and retrieved evidence unexamined. **Second**, such arbitration is conditioned by the verifier's pre-retrieval epistemic state with respect to the claim. A model that is confidently wrong, uncertain, or already correct may respond differently to the same retrieved evidence, yet these distinctions are collapsed under

standard outcome-level evaluation.

We address these challenges with PAVE (**P**rior-**A**ware **V**erifier **E**valuation), a diagnostic testbed that evaluates how LLM verifiers arbitrate between parametric knowledge and retrieved evidence under PCD in fact-checking. **First**, we construct a PCD benchmark of over 7,000 claims under controlled counterfactual and temporal conflict scenarios, enabling systematic evaluation of verifier behavior when parametric and contextual signals disagree. **Second**, to support state-conditioned evaluation, we characterize each verifier's pre-evidence epistemic state via a Knowledge Boundary (KB) framework, building on prior work on LLM knowledge awareness (Kadavath et al., 2022; Yin et al., 2024). We operationalize KB along two observable axes: *factual correctness*, indicating whether the model's prior-only prediction aligns with ground truth, and *parametric confidence*, estimated by prediction consistency across repeated stochastic generations as a proxy for epistemic stability (Chen and Mueller, 2024; Farquhar et al., 2024; Chen et al., 2025). Their combination yields four KB conditions: *Known-Knows* (confident and correct), *Known-Unknowns* (confident but incorrect), *Unknown-Knows* (uncertain but correct), and *Unknown-Unknowns* (uncertain and incorrect). This distinction is crucial because the same verdict-level failure can arise from different underlying conditions, e.g., a confidently wrong verifier that resists sound evidence differs fundamentally from an uncertain verifier that lacks a stable basis for judgment. Together, the PCD benchmark and KB stratification produce arbitration profiles for individual LLM verifiers, revealing dimensions of factual reliability that remain hidden under final verdict evaluation, as illustrated in Figure 1.

Evaluating seven LLMs, we find that the behavioral gap across models is substantial: Deepseek-v3 achieves a favorable balance between correction and persistence, while Llama3-8B is far more likely to be misled by counterfactual context than corrected by accurate evidence. Within the same model family, larger models demonstrate improved arbitration behavior, suggesting that scaling contributes to more reliable prior-context arbitration. We further find that stable errors can persist even when accurate evidence is available, and that models are generally more receptive to acquiring missing or outdated knowledge than to revising confident yet incorrect priors. These findings demonstrate that prior-context arbitration behavior is a systematic and measurable property of LLM verifiers that only final accuracy cannot capture. Motivated by these findings, we propose a lightweight JSD-based test-time arbitration method that improves factual reliability without retraining or modifying the underlying model, achieving best or competitive performance in five out of six evaluated LLM families[1]. Our main contributions are fourfold:

- We formulate LLM-based fact-checking under PCD as an epistemic conditioned arbitration problem, shifting evaluation from final verdict accuracy to how verifiers balance internal priors and external evidence under source discrepancy.
- We introduce a diagnostic testbed PAVE, including a controlled benchmark that induces PCD through counterfactual and temporally novel scenarios, KB stratification for characterizing pre-evidence epistemic states, and state-conditioned metrics that measure how verifiers resist, revise, or over-follow contextual evidence.
- Based on PAVE, we systematically evaluate seven LLMs under counterfactual and temporally novel scenarios, revealing distinct arbitration profiles and offering practical guidance for model selection in RAG-based fact-checking.
- We propose a lightweight JSD-based test-time arbitration method that improves factual reliability without changing the underlying model.

2 Related Work

LLM-based fact-checking. LLMs have demonstrated strong capabilities in knowledge-intensive reasoning (Sun et al., 2025; Yu et al., 2023; Kong et al., 2025), and have become widely adopted as automated fact-checking verifiers (Zhang and Gao, 2023; Tang et al., 2024; Wang et al., 2024; Ortu et al., 2024; Mahaut et al., 2024; Kong et al., 2026). Such systems either rely directly on parametric knowledge (Jiang et al., 2020; Roberts et al., 2020; Zhao et al., 2023) or augment the model with retrieved evidence under a RAG framework to compensate for outdated or incomplete parametric memory (Zhang et al., 2024; Qin et al., 2024; Shi et al., 2024; Borgeaud et al., 2022; Ram et al., 2023; Chang et al., 2025). In both cases, evaluation has largely centered on final verdict accuracy, leaving how parametric beliefs interact with retrieved evidence under source conflict underexamined.

[1]Our datasets are released at https://doi.org/10.5281/zenodo.18151788

Prior-context discrepancy in RAG. Retrieved evidence may be noisy, outdated, or inconsistent with an LLM's parametric knowledge (Xie et al., 2023; Wu et al., 2024; Huang et al., 2024; Zhang and Gao, 2024; Wang et al., 2025). Prior work has studied such discrepancies through controlled perturbations, including entity substitution (Longpre et al., 2021; Wu et al., 2024), negation injection (Zhou et al., 2023), and LLM-generated counterfactuals (Xie et al., 2023), showing that models vary in their susceptibility to conflicting evidence and that confidence can affect whether retrieved content overrides parametric knowledge (Wu et al., 2024). However, these studies largely measure susceptibility in aggregate, without conditioning failures on the verifier's epistemic state. Consequently, they cannot distinguish when a model appropriately persists in a reliable prior, corrects an unreliable one, or fails by over-trusting either source. We address this gap with an epistemic-state-conditioned evaluation protocol and arbitration profiles that diagnose model-specific behavior under PCD.

Knowledge boundary and uncertainty estimation. Prior work has studied LLM knowledge boundaries (Garner and Alexander, 1989; Yin et al., 2024) by benchmarking how well models distinguish facts they are confident about from those they are uncertain or wrong (Kadavath et al., 2022; Yin et al., 2024), and by examining models' ability to recognize uncertainty over questions with no definitive answers (Amayuelas et al., 2024; Li et al., 2024). For estimating parametric confidence without model internals, output-side consistency methods use prediction variability across repeated generations as a proxy for epistemic uncertainty (Taubenfeld et al., 2025; Chen and Mueller, 2024). Building on these ideas, we formulate KB as an observable characterization of a verifier's pre-evidence epistemic state, jointly capturing prior-only factual correctness and parametric confidence. This state characterization supports both KB-conditioned arbitration evaluation and the JSD-based test-time arbitration signal used in our method.

3 Evaluation Method

We adopt a consistency-based probing setup to characterize a verifier's pre-evidence epistemic state. For each claim $\mathbf{c}_i$ with ground-truth label y_i, we query the model *without* retrieved evidence for N independent runs, obtaining N prior-only verdicts. *Confidence* is defined by strict consistency: a claim is KNOWN if all prior-only verdicts are identical; otherwise it is UNKNOWN. *Correctness* defines KNOWS/UNKNOWS: a verdict is KNOWS if it matches y_i; otherwise it is UNKNOWS. Combining the two axes yields four observable states for each prediction: KNOWN-KNOWS (confident and correct), KNOWN-UNKNOWS (confident but incorrect), UNKNOWN-KNOWS (uncertain but correct), and UNKNOWN-UNKNOWS (uncertain and incorrect), as shown in Figure 2 (bottom).

3.1 Evaluation Metrics

To comprehensively quantify the dynamics between internal prior and retrieved context, PAVE adopts a hierarchical evaluation framework. Building on the defined *correction* and *persistence* scenarios, we report two primary performance rates, **Correction Rate** (CR) and **Persistence Rate** (PR), which capture a model's adaptability and robustness, respectively. We further introduce derived metrics, **Margin** and **Odds of Influence**, to analyze the trade-off between these competing behaviors.

Primary Metrics (CR & PR). We first stratify evaluation based on the model's prior-only prediction correctness (i.e., KNOWS VS. UNKNOWS). Consider a fact-checking test set $\{(\mathbf{c}_i, y_i, \mathbf{e}_i)\}$, where $\mathbf{c}_i$ is the i-th claim to be verified, $y_i \in \{\text{Refute}, \text{Support}\}$ is its ground-truth label, and $\mathbf{e}_i^{\text{gold}}$ is the associated gold evidence. Let $\mathcal{D}_{wrong} = \{i \mid f_\theta(\mathbf{c}_i) \neq y_i\}$ denote the subset of instances where a model f_θ's prior-only prediction is incorrect, and reversely, $\mathcal{D}_{correct} = \{i \mid f_\theta(\mathbf{c}_i) = y_i\}$ denote the subset where the prior-only prediction is correct. The Correction Rate (**CR**) measures the model's ability to rectify the initial errors when provided with *correct* external evidence and is computed as the post-retrieval accuracy on $\mathcal{D}_{wrong}$:

$$\text{CR} = \frac{1}{|\mathcal{D}_{wrong}|} \sum_{i \in \mathcal{D}_{wrong}} \mathbb{I}\left[f_\theta(\mathbf{c}_i, \mathbf{e}_i^{\text{gold}}) = y_i\right].$$

The Persistence Rate (**PR**), however, measures the model's robustness in keeping consistent with the correct prior prediction when exposed to retrieved context with counterfactuals ($\mathbf{e}^{\text{cf}}$ in §3.2). It is computed as the post-retrieval accuracy on $\mathcal{D}_{correct}$:

$$\text{PR} = \frac{1}{|\mathcal{D}_{correct}|} \sum_{i \in \mathcal{D}_{correct}} \mathbb{I}\left[f_\theta(\mathbf{c}_i, \mathbf{e}_i^{\text{cf}}) = y_i\right].$$

Differential Metric (Margin). To quantify the benefit of retrieval, we define the **Margin** as the

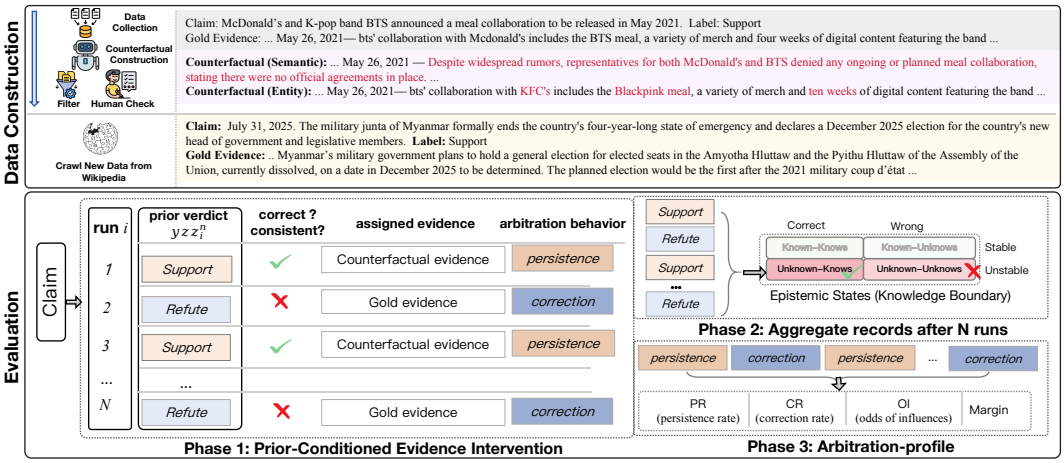


Figure 2: Dataset Construction and evaluation pipeline for model behavior analysis under epistemic states.

gap between the model's receptiveness to gold evidence and its susceptibility to discrepant yet counterfactual evidence. Specifically, Margin measures the difference between the rate of desired change (correction) and undesired change (credulousness, i.e., $1 - \text{PR}$):

$$\text{Margin} = \text{CR} - (1 - \text{PR}) = \text{CR} + \text{PR} - 1.$$

A positive Margin indicates that the model is more likely to be corrected by accurate evidence than misled by inaccurate context. Larger Margin values therefore reflect more desirable behavior, combining stronger correction and greater persistence capabilities when facing the discrepancies.

Odds of influence (OI). To measure the relative elasticity of a model's parametric confidence, we compare how often it changes versus maintains its prior-based prediction when facing discrepant retrieved context. We define **OI** as the odds of changing an answer over maintaining it, computed separately under confident (KNOWN) and uncertain (UNKNOWN) conditions for the incorrect (UNKNOWS) and correct (KNOWS) cases: **1) UNKNOWS regime** (on $\mathcal{D}_{wrong}$): Since the rate of change is given by CR, the corresponding odds are defined as: $\text{OI}_{\text{KNOWN}} = \frac{\text{CR}_{\text{KNOWN}}}{1-\text{CR}_{\text{KNOWN}}}$, $\text{OI}_{\text{UNKNOWN}} = \frac{\text{CR}_{\text{UNKNOWN}}}{1-\text{CR}_{\text{UNKNOWN}}}$. **2) KNOWS regime** (on $\mathcal{D}_{\text{correct}}$): Since the rate of change is given by $(1 - \text{PR})$, the corresponding odds are: $\text{OI}_{\text{KNOWN}} = \frac{1-\text{PR}_{\text{KNOWN}}}{\text{PR}_{\text{KNOWN}}}$, $\text{OI}_{\text{UNKNOWN}} = \frac{1-\text{PR}_{\text{UNKNOWN}}}{\text{PR}_{\text{UNKNOWN}}}$. As a tendency to change an answer, a lower OI for KNOWS regime is preferred to demonstrate persistence (i.e., resisting incorrect context), while a higher OI for UNKNOWS regime is favored to demonstrate receptivity (i.e., accepting correct context to rectify internal errors).

3.2 Dataset Construction

We construct an evaluation dataset for analyzing LLM's fact-checking behaviors, controllable testing when internal priors and external evidence disagree. The dataset is developed along two primary dimensions (see Figure 2(upper)), i.e., counterfactual and temporal, to probe fact-checking behavior under diverse prior-context discrepancies. Comparisons with prior datasets are in §B, with dataset construction and quality control details in §D.

Dimension 1: Counterfactual. This dimension evaluates model behavior in the presence of erroneous external evidence. We leverage three public fact-checking benchmarks, Quantemp, PolitiFact, and Snopes (Anand et al., 2024; Popat et al., 2018), to construct a consolidated set of 10,866 verifiable claims with ground-truth labels through controlled counterfactual construction and quality control.

Counterfactual Construction. Inspired by the controllable discrepancy construction method, like negation injection (Gubelmann and Handschuh, 2022), entity substitution (Wu et al., 2024; Longpre et al., 2021; Zhou et al., 2023), and LLM-generated counterfactual (Xie et al., 2023), we formally define counterfactual evidence $\mathbf{e}^{\text{cf}}$ using transformation functions $\mathcal{T}$ applied to the gold evidence $\mathbf{e}^{\text{gold}}$, such that $\mathbf{e}_i^{\text{cf}} = \{\mathcal{T}(\mathbf{e}_i^{\text{gold}}) \mid \mathcal{T} \in \{\mathcal{T}_{\text{entity}}, \mathcal{T}_{\text{semantic}}\}\}$. Specifically, 1) **Counter-entity** ($\mathcal{T}_{\text{entity}}$) replaces key entities in the gold evidence, testing resistance to fine-grained, token-level factual errors; 2) **Counter-semantic** ($\mathcal{T}_{\text{semantic}}$) generates semantically counterfactual evidence that contradicts the original content at a broader semantic level. Each instance in the curated dataset is represented as $\{\mathbf{c}_i, \mathbf{e}_i^{\text{gold}}, \mathbf{e}_i^{\text{cf_entity}}, \mathbf{e}_i^{\text{cf_semantic}}\}$, where $\mathbf{e}_i^{\text{cf_entity}}$

and $\mathbf{e}_i^{\text{cf_semantic}}$ correspond to the outputs of $\mathcal{T}_{\text{entity}}$ and $\mathcal{T}_{\text{semantic}}$, respectively.

Quality Control. To ensure a valid evaluation of persistence, we apply a strict refusal filtering procedure, excluding instances where the model abstains (e.g., responses such as "I cannot answer"), retaining only samples where the model exhibits a definite prior. This creates a genuine tension between internal beliefs and discrepant external context. Counterfactual evidence is generated using GPT-4o under controlled prompting settings. See §D for more details.

Dimension 2: Temporality. This dimension evaluates model behavior when relevant internal knowledge is obsolete or missing, a key requirement for effective RAG-based fact-checking. Temporal PCD is therefore evaluated as a correction rather than a persistence setting: when a claim concerns post-cutoff events, the relevant prior is by construction outdated or unavailable, so persisting in it would not constitute reliable fact-checking.

Data Collection. We collect valid claims about *new* events by crawling Wikipedia's "Current events" portal[2], focusing on entries from 2024 and 2025. Gold evidence is obtained from the corresponding Wikipedia references. To distinguish between *outdated* and *new* knowledge, we compare each event's occurrence date with the pre-training cutoff of each evaluated LLM. A claim is considered *new* for a given model if it occurred after the model's training cutoff, and therefore lies outside its parametric memory. We retain only claims the model answers *incorrectly* without external evidence (i.e., using its internal priors). This constraint enforces a valid *acquisition* scenario, requiring the model to resolve the discrepancy by incorporating retrieved context rather than persisting in erroneous priors. The resulting dataset contains 3,995 factually *true* claims. Each instance is structured as $\{\mathbf{c}_i, \mathbf{e}_i^{\text{gold}}, T_i\}$, where $\mathbf{e}_i^{\text{gold}}$ denotes the gold evidence for the claim $\mathbf{c}_i$ and T_i is the event timestamp. For fair evaluation, we retain only claims that are considered new for all evaluated LLMs.

3.3 Evaluated LLMs

We evaluate seven state-of-the-art LLMs, including closed-source models, such as GPT-4o-mini and Gemini-2.5-flash, and open-source models, such as Mistral-7B (i.e., Mistral-7B-Instruct-v0.2), Llama3-8B, Phi4, Qwen3-32B, and Deepseek-v3, spanning a wide range of scales from 7B to 617B parameters. This diversity enables robust comparison across model architectures and capabilities.

[2] https://en.wikipedia.org/wiki/Portal:Current_events

4 Evaluation Results

We organize our analysis into three parts: **1) cross-model arbitration profiles** (Observations 1–3), comparing how different LLMs balance priors and evidence at the model level. **2) cross-setting arbitration strategies** (Observation 4), exposing how models diverge in handling counterfactual versus temporally new evidence. **3) modulating factors and model scale** (Observations 5–7), examining how evidence type, label direction, and scaling shape arbitration behavior. All primary results are averaged over ten runs at temperature 0.3; further analysis (e.g., temperature and iteration effects in §A.1) and details (e.g., prompts) are in §E and §A.

Observation 1. LLMs exhibit heterogeneous pre-evidence epistemic states. LLMs display substantial variation in their internal confidence profiles, as manifested by the differing distributions of KNOWN and UNKNOWN categories. This indicates that each model possesses a distinct internal epistemic state. In Table 1, most evaluated models exhibit high confidence, with more than 80% of claims falling into KNOWN state. In contrast, Mistral-7B and Phi-4 show significantly higher uncertainty, with around 50% and 90% of claims categorized as UNKNOWN, respectively, suggesting limited stability and self-consistency in fact-checking. Importantly, we analyze the validity of the KNOWN/UNKNOWN classification in §A.2.

Observation 2. High prior-only accuracy does not imply reliable arbitration under PCD. Final accuracy alone cannot reveal how a verifier reaches its verdict under PCD. Consider Qwen3-32B and Deepseek-v3: in prior-only evaluation, Qwen3-32B substantially outperforms Deepseek-v3 in prior accuracy, positioning it as the stronger verifier under specific benchmarks. Yet this ranking *reverses* once prior-context discrepancy is introduced: Qwen3-32B collapses by 43.02 points, falling below Deepseek-v3 (only a 19.11-point drop). The Margin captures this reversal precisely (-0.115 vs. $+0.235$), where final-verdict metrics are silent. This reversal is consistent with the variance analysis in Table 5 (§A.3): Qwen3-32B's high prior accuracy is accompanied by high variance ($\sigma = 0.2283$ for PR-KK), whereas Deepseek-v3

		Deepseek-v3	Gemini-2.5	Phi-4	Qwen3-32B	Mistral-7B	GPT-4o-mini	Llama3-8B
Dimension 1: Counterfactual Setting								
Margin ↑		+0.235	+0.085	−0.004	−0.115	−0.190	−0.521	−0.661
Persistence — resisting misleading evidence when the prior is correct								
KK	PR ↑	71.70%	56.33%	49.43%	50.59%	71.59%	40.23%	29.95%
	Δ correct drop ↓	17.19%	21.83%	11.44%	41.79%	1.55%	36.90%	32.49%
	OI ↓	0.395	0.775	1.023	0.977	0.397	1.486	2.339
	Ratio	60.75%	50.00%	22.62%	84.57%	5.45%	61.73%	46.38%
UK	PR ↑	44.05%	48.50%	48.84%	60.82%	55.76%	25.33%	33.00%
	Δ correct drop ↓	1.91%	5.15%	10.08%	1.23%	4.11%	1.33%	9.96%
	OI ↓	1.270	1.062	1.048	0.644	0.793	2.948	2.030
	Ratio	3.42%	10.00%	19.71%	3.15%	9.29%	1.78%	14.87%
correct rate (prior) ↑		64.17%	60.00%	42.33%	87.72%	14.74%	63.51%	61.25%
correct rate (after intervention) ↑		45.06%	33.02%	20.81%	44.70%	9.08%	25.28%	18.80%
Correction — updating from accurate evidence when the prior is wrong								
KU	CR ↑	40.75%	43.65%	42.76%	29.75%	26.36%	24.73%	30.80%
	Δ wrong reduction ↑	13.10%	8.37%	8.26%	3.23%	0.18%	8.39%	8.23%
	OI ↑	0.687	0.776	0.747	0.423	0.358	0.329	0.445
	Ratio	32.15%	19.17%	19.31%	10.85%	0.68%	33.92%	26.73%
UU	CR ↑	66.99%	60.02%	58.57%	47.37%	27.26%	57.59%	40.16%
	Δ wrong reduction ↑	2.47%	12.50%	22.47%	0.68%	23.06%	1.48%	4.83%
	OI ↑	2.030	1.502	1.414	0.900	0.375	0.345	0.672
	Ratio	3.68%	20.83%	38.36%	1.43%	84.58%	2.57%	12.02%
wrong rate (prior) ↓		35.83%	40.00%	57.67%	12.28%	85.26%	36.49%	38.75%
wrong rate (after intervention) ↓		20.26%	19.13%	26.95%	8.37%	62.02%	26.62%	25.69%
Dimension 2: Temporal Setting								
Acquisition — updating from accurate evidence when the prior is outdated								
KU	CR ↑	46.58%	28.65%	24.39%	90.90%	27.80%	54.00%	94.81%
	Δ wrong reduction ↑	23.71%	13.89%	23.13%	72.17%	9.14%	46.56%	35.74%
	OI ↑	0.872	0.402	0.323	9.989	0.385	1.524	9.460
	Ratio	50.91%	48.48%	94.85%	79.39%	32.88%	86.22%	37.70%
UU	CR ↑	68.25%	24.35%	27.45%	79.38%	40.50%	68.25%	98.85%
	Δ wrong reduction ↑	33.50%	12.55%	1.41%	16.36%	27.18%	9.40%	61.58%
	OI ↑	2.150	0.322	0.378	3.850	0.681	9.230	38.37
	Ratio	49.09%	51.52%	5.15%	20.61%	67.12%	13.78%	62.30%
wrong rate (prior) ↓		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
wrong rate (after intervention) ↓		42.78%	73.57%	75.45%	11.47%	63.68%	44.04%	2.67%

Table 1: Main arbitration profile of LLM verifiers. Purple highlights Margin, the overall trade-off performance (sum of Known- and Unknown-state margins). Teal rows show persistence when the prior is correct and counterfactual evidence is given (KK/UK); PR is persistence rate, OI is odds of influence, and Δ correct drop is the correct-rate loss. Orange rows show correction or acquisition when the prior is wrong or outdated, and gold evidence is given (KU/UU); CR is correction rate and Δ wrong reduction is the wrong-rate reduction. Gray rows give aggregate rates before and after intervention. Ratio reports the percentage of samples falling into each KB condition.

combines moderate prior accuracy with low variance ($\sigma = 0.0244$). Strong performance in conventional evaluation thus does not guarantee robustness under PCD, underscoring the need for prior-aware evaluation that captures ***how*** models arbitrate, not just ***what*** they predict.

Observation 3. Model selection is critical for RAG-based fact-checking. Table 1 illustrates significant variation across LLMs in how they arbitrate internal priors and external evidence, as reflected by their Margin values. Models with positive Margins, such as Deepseek-v3 (+0.235) and Gemini-2.5-flash (+0.085), demonstrate desirable behavior: they resist discrepant context when internally correct, yet remain receptive to corrective evidence when internally incorrect. In contrast, Gpt-4o-mini, Llama3-8B, and Mistral-7B exhibit a more "stubborn" profile, frequently favoring erroneous internal priors over accurate external evidence (e.g., Mistral-7B CR: 26.36%). OI values observed in Table 1 across KNOWS and UNKNOWS regimes confirm such an observation, where only Deepseek and Gemini demonstrate the favored OI patterns. These results underscore the importance of careful ***model selection*** for RAG-based fact-checking, considering their distinct PCD handling capabilities.

Observation 4. Temporal acquisition and counterfactual correction reveal different evidence-uptake patterns. At the aggregate level, the OI for acquisition tends to exceed that for correction across models (e.g., GPT-4o-mini: 9.23 vs. 0.35; Llama3-8B: 38.37 vs. 0.67), suggesting LLMs are more receptive to new knowledge than to revising wrong beliefs. However, this aggregate pattern conceals three distinct strategies. (1) **Receptive models** (Llama3-8B, Qwen3-32B) drive the trend with large positive gaps (e.g., Llama3-8B's CR: 94.81% temporal vs. 30.80% counterfactual in KU), readily accepting "new" information but resisting evidence that contradicts prior knowledge. (2) **Conservative models** (Gemini-2.5-flash, Phi-4) reverse the pattern,

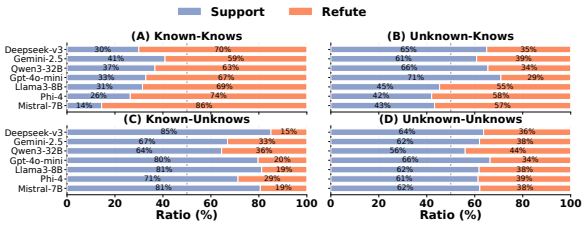


Figure 3: Support/refute ratios across Known and Unknown prior categories, separated by whether the internal prior response is correct or wrong.

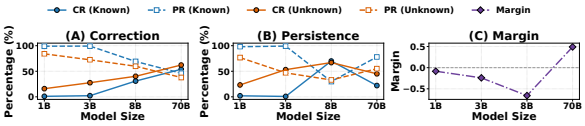


Figure 4: Model Scaling Evaluation. The Correction rate (CR), persistence rate (PR), and margin results of Llama3 in various sizes (1B, 3B, 8B, 70B)

with temporal CR *lower* than counterfactual CR (e.g., Phi-4: 24.39% vs. 42.76% in KU), reflecting a cautious stance toward unfamiliar claims. (3) **Balanced models** (Deepseek-v3) maintain near-identical correction rates across both settings (46.58% vs. 40.75%), suggesting genuinely evidence-driven arbitration.

Observation 5. LLMs exhibit greater confidence in refuting false claims than in confirming true ones. Analysis of label distributions in Figure 3 shows that LLMs are more internally confident when identifying false claims than when validating true statements. Among internally correct predictions (i.e., KNOWN-KNOWS), Refute cases dominate (e.g., 86% for Mistral-7B and over 59% for other models), suggesting that pre-trained reliability is more strongly geared toward detecting falsehoods. In contrast, in the UNKNOWN state, correct predictions (i.e., UNKNOWN-KNOWS) are predominantly Support cases (e.g., 71% for GPT-4o-mini), indicating heavier reliance on external evidence for verifying affirmative claims where their internal knowledge is insufficient.

Observation 6. Larger models exhibit better arbitration calibration. We analyze how performance evolves with model scale using the Llama3 family. As shown in Figure 4, Margin scores increase with model size, with Llama3-70B achieving a substantial improvement (Margin = 0.50). This trend indicates that larger models are more effective at resolving discrepancies. While persistence rate decreases from the 3B to the 8B model, it recovers at 70B alongside a marked increase in correction rate. The negative Margins observed for the 1B and 3B suggest that their persistence reflects rigid adherence to incorrect priors rather than robust calibration. This scaling trend suggests that larger models achieve a more balanced arbitration profile under PCD, improving correction without entirely sacrificing persistence.

5 Arbitration Method and Result

Building on these insights, we propose a method to more reliably arbitrate between internal priors and external context, which shows that the diagnostic signal identified by PAVE can be operationalized into a simple test-time intervention. Unlike prior works (Huang et al., 2024; Yu et al., 2024) that ignore variability across runs, particularly when the model exhibits low internal confidence (i.e., the UNKNOWN state), we assess its decision stability by measuring consistency across multiple parallel generations. Specifically, we compute a Jensen-Shannon divergence (JSD)-based metric to quantify how consistently an LLM favors parametric knowledge or external evidence, enabling more robust source selection.

Stability-based Arbitration. Given a test set $\mathcal{Q} = \{(\mathbf{c}, y)\}$, where $\mathbf{c}$ is a claim and $y \in \{\text{Refute}, \text{Support}\}$ is the ground-truth label, we use an LLM f_θ to repeatedly sample K times of prediction probabilities for the "True" token under external evidence $\mathbf{e}$: $\{p_{\text{ext}}^{(i)} = f_\theta(\mathbf{c}, \mathbf{e})\}_{i=1}^{K}$. For each sample, we define the binary entropy as:

$$H(p_{\text{ext}}^{(i)}) = -p_{\text{ext}}^{(i)} \log_2 p_{\text{ext}}^{(i)} - (1-p_{\text{ext}}^{(i)}) \log_2 (1-p_{\text{ext}}^{(i)}).$$

We then compute the JSD across the K samples as follows: (1) compute the mean probability $\bar{p}_{\text{ext}} = \frac{1}{K}\sum_{i=1}^{K} p_{\text{ext}}^{(i)}$ and its entropy $H(\bar{p}_{\text{ext}})$; (2) compute the expected individual entropy $\mathbb{E}[H(p_{\text{ext}})] = \frac{1}{K}\sum_{i=1}^{K} H(p_{\text{ext}}^{(i)})$; (3) define the stability metric:

$$\mathbf{JSD}_{\text{ext}} = H(\bar{p}_{\text{ext}}) - \mathbb{E}[H(p_{\text{ext}})],$$

Lower $\mathbf{JSD}_{\text{ext}}$ indicates higher consistency across runs given external evidence. Analogously, given K parametric prediction probabilities $\{p_{\text{par}}^{(i)} = f_\theta(\mathbf{c})\}_{i=1}^{K}$ for the "True" token without external evidence, we compute the stability $\mathbf{JSD}_{\text{par}}$ for the parametric case: $\mathbf{JSD}_{\text{par}} = H(\bar{p}_{\text{par}}) - \mathbb{E}[H(p_{\text{par}})]$. Finally, we define the difference $\Delta = \mathbf{JSD}_{\text{par}} - \mathbf{JSD}_{\text{ext}}$ and compare it against a threshold $\tau \in [-0.08, -0.06]$, calibrated on a held-out validation set (see §C), to determine the reliable source:

$$\hat{y} = \begin{cases} f_\theta(\mathbf{c}), & \text{if } \Delta \leq \tau \\ f_\theta(\mathbf{c}, \mathbf{e}), & \text{otherwise.} \end{cases}$$

The full algorithm and implementation details are in Algorithm 2 and §C.1, with a vote-based JSD variant for models without logits in §C.2.

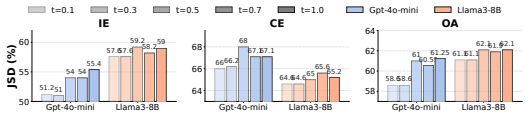
Baselines and Implementation. We evaluate eight baselines grouped into three categories (details in §C.4): (i) *Self-guided methods* (Huang et al., 2024): ImplicitSCR prompts the LLM to assess evidence reliability before answering, while ExplicitSCR uses CoT reasoning to compare internal and external context. (ii) *Rule-based methods* (Huang et al., 2024): InternalEval and ContextEval switch sources based on self-evaluated response or evidence correctness; InternalConf and ContextConf use token-level confidence thresholds to choose between prior-based and retrieval-based answers (Jiang et al., 2023b; Huang et al., 2024); TPC (Huang et al., 2024; Wu et al., 2024) compares confidence scores using raw or calibrated probabilities. (iii) *Context-based methods*: TACS-LR (Yu et al., 2024) refines retrieved context using an LLM-based filter before final generation.

Metrics. Given a test set $\mathcal{Q} = \{(\mathbf{c}, y)\}$, we follow (Huang et al., 2024; Yu et al., 2024) and evaluate accuracy under three conditions. Let $\mathbf{e}_c$ and $\mathbf{e}_w$ denote correct and incorrect evidence, respectively, and let the model prediction be $\hat{y} = f_\theta(\mathbf{c}, \mathbf{e})$. Using the indicator function $\mathbb{I}(\cdot)$, we define: i) *Accuracy with correct evidence (CE)*: $\text{CE} = \frac{1}{|\mathcal{Q}|}\sum_{(\mathbf{c},y)\in\mathcal{Q}} \mathbb{I}(f_\theta(\mathbf{c}, \mathbf{e}_c) = y)$. ii) *Accuracy with incorrect evidence (IE)*: $\text{IE} = \frac{1}{|\mathcal{Q}|}\sum_{(\mathbf{c},y)\in\mathcal{Q}} \mathbb{I}(f_\theta(\mathbf{c}, \mathbf{e}_w) = y)$. iii) *Overall Accuracy (OA)*: $\text{OA} = \frac{\text{CE}+\text{IE}}{2}$.

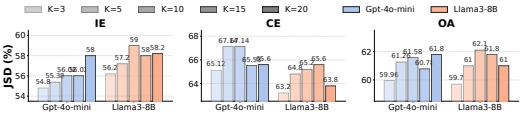
Main Results Our method achieves strong performance across a range of LLMs, attaining the best or competitive overall accuracy in five out of six evaluations. In Table 2, the largest gains are observed for smaller models such as Mistral-7B, where our approach outperforms the second-best baseline by 6.58 absolute points (57.08% vs. 50.50%). Moreover, unlike methods like ContextEval that overly favor either parametric or external signals, our approach maintains a balanced trade-off between correct and incorrect evidence across models, indicating robust arbitration capabilities. For example, GPT-4o-mini achieves strong performance in both settings, with IE 55.38% and CE 67.14%, indicating stable arbitration behavior. Results are reported using $K = 5$ runs with temperature set to 1.0. Case studies are provided in §C.3.

Baseline	Gpt-4o-mini			Phi-4			Mistral-7B		
	IE.	CE.	OA.	IE.	CE.	OA.	IE.	CE.	OA.
ImplicitSCR	45.48	59.88	52.68	50.10	60.80	55.45	34.56	38.83	36.70
ExplicitSCR	54.12	64.35	59.24	41.81	50.84	46.33	10.33	18.77	14.55
InternalEval	51.43	57.60	54.52	47.60	48.80	48.20	6.26	3.08	4.67
ContextEval	55.20	67.03	61.12	**52.50**	**66.60**	**59.55**	27.61	10.33	18.97
ContextConf	47.90	64.70	56.30	52.20	52.20	52.20	15.33	10.33	12.83
InternalConf	47.60	48.40	48.00	46.60	65.20	55.90	34.33	34.33	34.33
TPC	46.90	61.60	54.25	51.14	51.14	51.14	43.33	57.67	50.50
TACS-LR	44.60	54.00	49.30	45.20	51.80	47.15	37.20	47.80	42.50
Ours (K=5)	**55.38**	**67.14**	**61.26**	52.40	63.60	58.00	**51.20**	**62.96**	**57.08**
Baseline	**Qwen3-32B**			**Gemini-2.5-flash**			**Llama3-8B**		
	IE.	CE.	OA.	IE.	CE.	OA.	IE.	CE.	OA.
ImplicitSCR	47.77	60.08	53.93	14.90	21.85	18.38	50.40	60.00	55.20
ExplicitSCR	19.27	23.55	21.41	**57.50**	**66.93**	**62.22**	32.60	38.60	35.60
InternalEval	52.63	59.29	55.96	49.95	55.81	52.88	51.04	58.39	54.72
ContextEval	**58.49**	64.25	61.47	55.61	38.73	47.17	54.92	61.17	58.05
TACS-LR	45.20	51.80	48.50	39.50	53.00	46.25	28.60	34.60	31.60
Ours (K=5)	56.80	**69.12**	**62.97**	51.12	65.01	58.07	**59.00**	**65.20**	**62.10**

Table 2: Accuracy (%) under incorrect (IE) and correct evidence (CE), and Overall Accuracy (OA). The results of our method for setting $K = 5$.



(a) Different temperature $t \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$.



(b) Number (#) of runs $K \in \{3, 5, 10, 15, 20\}$.

Figure 5: Comparison of our method across three metrics (IE, CE, and OA) with varying temperature settings (a) and number of runs K (b).

Analysis of Efficiency and Temperature. Our method demonstrates a favorable efficiency-accuracy trade-off in Figure 5. Increasing the number of runs beyond $K = 5$ yields diminishing returns, while even a small budget of $K = 3$ achieves competitive performance (e.g., IE = 56.20%) on Llama3-8B. These results underscore the practicality of the approach for deployment scenarios with limited computational budgets. As for the impact of temperature, results on Llama3-8B and GPT-4o-mini show that increasing temperature from 0.1 to 0.5 leads to noticeable performance gains across IE, CE, and OA while no significant or consistent improvement is observed continuously.

6 Conclusion

This work presents a principled framework for analyzing how LLMs arbitrate between internal priors and retrieved context in fact-checking. We analyze model's behavior under its pre-evidence epistemic states based on Knowledge Boundary, and propose a lightweight test-time arbitration method that outperforms eight strong baselines, improving factual reliability under prior-context discrepancy.

Limitations

We acknowledge that different experimental settings may introduce variability into evaluation outcomes. We examined three potential sources of variation: the choice of output token representation, sampling temperature, and the number of independent runs (see §A.6 and §A.1). While specific metrics fluctuate across configurations, the core findings regarding arbitration profiles remain relatively stable; detailed sensitivity analyses are provided in the appendix.

The primary focus of our dataset is binary fact-checking, and our initial results reflect this scope. It should be noted that the evaluation of free-form explanations is beyond our current scope. Nonetheless, the architecture of our testbed is extensible to multi-label tasks; we have included an extensive batch analysis for such scenarios in §A.4 to showcase its broader applicability.

Our study focuses on the discrepancies between internal priors and retrieved context, assuming retrieval has already occurred. Consequently, the situation of external conflicts *within* the retrieved evidence and agents that autonomously skip the RAG process are outside our current scope. We prioritize discrepancy scenarios, where one source is correct while the other is not (prior vs. context), because they directly impact the finality of fact-checking judgments. Situations where both sources are incorrect (even if they differ) are excluded, as they do not facilitate reaching a correct conclusion. While we recognize the value of "both error" cases for model explainability, a detailed investigation into reasoning behavior when all available information is erroneous is reserved for future work.

We did not include the cases where LLMs correctly answer *new* claims. It is difficult to determine whether these correct answers come from valid reasoning (e.g., predicting periodic events like the 2026 World Cup) or simply result from hallucinations. However, we analyze an example where the LLM correctly predicts future events with hallucination in §A.5. We will leave the analysis of this seemingly "superior intelligent behavior" for future work.

While our approach entails a marginal increase in overhead compared to existing baselines, the empirical results demonstrate that it remains highly competitive even at $K = 3$, as shown in §5 "Analysis of Efficiency and Temperature". At $K = 3$, the method remains practical under batched inference, although it requires repeated stochastic samples under claim-only and evidence-conditioned inputs. Even at this reduced cost, the method retains competitive performance, indicating that our method can strike a practical balance between inference overhead and scoring reliability.

Ethics Statement

All evaluations were conducted using open-source datasets and crawled from public websites. We are committed to fostering AI transparency and providing the guidelines of RAG in automated fact-checking systems. Our data sources are all from objective and neutral facts and do not contain any personal information and offensive comments directed at individuals or particular groups.

Our counterfactual evidence is generated with LLM assistance under controlled prompting. While this enables scalable construction of prior-context discrepancy scenarios, it may also introduce generation artifacts, such as stylistic regularities, unnatural wording, or unintended factual changes beyond the targeted intervention. To mitigate this risk, we apply automatic filtering and human validity checks to remove inconsistent or ill-formed counterfactuals. We will release the prompts, filtering rules, and constructed data to support reproducibility and facilitate further auditing by the community.

References

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2024. How many opinions does your llm have? improving uncertainty estimation in nlg. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification*.

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6416–6432.

Abhijit Anand, Avishek Anand, Vinay Setty, and 1 others. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. *arXiv preprint arXiv:2403.17169*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

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.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. 2025. MAIN-RAG: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2607–2622, Vienna, Austria. Association for Computational Linguistics.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200, Bangkok, Thailand. Association for Computational Linguistics.

Lida Chen, Zujie Liang, Xintao Wang, Jiaqing Liang, Yanghua Xiao, Feng Wei, Jinglei Chen, Zhenghong Hao, Bing Han, and Wei Wang. 2025. Teaching large language models to express knowledge boundary from their own signals. In *Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM)*, pages 26–39.

S Farquhar, J Kossen, L Kuhn, and Y Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. Deep think with confidence. *arXiv preprint arXiv:2508.15260*.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2336–2346, St. Julian's, Malta. Association for Computational Linguistics.

Ruth Garner and Patricia A Alexander. 1989. Metacognition: Answered and unanswered questions. *Educational psychologist*, 24(2):143–158.

Reto Gubelmann and Siegfried Handschuh. 2022. Context matters: A pragmatic study of plms' negation understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2540–2552.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. *arXiv preprint arXiv:2306.04459*.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. 2024. To trust or not to trust? enhancing large language models' situated faithfulness to external contexts. *arXiv preprint arXiv:2410.14675*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023a. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

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*.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Chuyi Kong, Ziyang Luo, Hongzhan Lin, Zhiyuan Fan, Yaxin Fan, Yuxi Sun, and Jing Ma. 2025. Sharp: Unlocking interactive hallucination via stance transfer in role-playing llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 839–866.

Chuyi Kong, Gao Wei, Jing Ma, Hongzhan Lin, and Yuxi Sun. 2026. Reflex: Self-refining explainable fact-checking via verdict-anchored style control. In *Proceedings of the 64nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2024. Knowledge boundary of large language models: A survey. *arXiv preprint arXiv:2412.12472*.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5131–5157.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3357–3370, Mexico City, Mexico. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, and Shu Wu. 2024. Ex-fever: A dataset for multi-hop explainable fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9340–9353.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. 2024. Factual confidence of llms: on reliability and robustness of current estimators. *arXiv preprint arXiv:2406.13415*.

Sara Vera Marjanovic, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. DYNAMICQA: Tracing internal knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14346–14360, Miami, Florida, USA. Association for Computational Linguistics.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if" the moon is made of marshmallows". *arXiv preprint arXiv:2410.03727*.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, Bangkok, Thailand. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Luu Anh Tuan, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 6981–7004.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th international conference on world wide web companion*, pages 1003–1012.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, and 1 others. 2024. Tool learning with foundation models. *ACM Computing Surveys*, 57(4):1–40.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*.

Yuxi Sun, Aoqi Zuo, Wei Gao, and Jing Ma. 2025. Causalabstain: Enhancing multilingual llms with causal reasoning for trustworthy abstention. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14060–14076.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms. *arXiv preprint arXiv:2502.06233*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of NAACL-HLT*.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, pages 2452–2463.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. 2025. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30553–30571, Vienna, Austria. Association for Computational Linguistics.

Kevin Wu, Eric Wu, and James Y Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Advances in Neural Information Processing Systems*, 37:33402–33422.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*.

Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. 2024. Benchmarking knowledge boundary for large language models: A different perspective on model evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2270–2286, Bangkok, Thailand. Association for Computational Linguistics.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, and 1 others. 2023. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.

Tian Yu, Shaolei Zhang, and Yang Feng. 2024. Truth-aware context selection: Mitigating hallucinations of large language models being misled by untruthful contexts. *arXiv preprint arXiv:2403.07556*.

Xuan Zhang and Wei Gao. 2023. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.

Xuan Zhang and Wei Gao. 2024. Reinforcement retrieval leveraging fine-grained feedback for fact checking news claims with black-box LLM. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13861–13873, Torino, Italia. ELRA and ICCL.

Zihan Zhang, Meng Fang, and Ling Chen. 2024. RetrievalQA: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6963–6975, Bangkok, Thailand. Association for Computational Linguistics.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and 1 others. 2023. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

Appendix

Appendix Roadmap. This appendix provides supplementary analyses, methodological details, dataset documentation, and reproducibility materials that support the main paper without interrupting its core narrative.

- **Appendix A: Further Discussion.** This section complements the main results with additional empirical analyses from four perspectives:
  1. **More observation.** Observation A compares counter-semantic and counter-entity conflicts, showing that semantic-level discrepancies are more challenging for models to resist.
  2. **Experimental settings discussion.** Appendix A.1 analyzes the effects of temperature and the number of independent runs, while Appendix A.2 validates the KNOWN/UNKNOWN split through confidence and class-token probability distributions.
  3. **Stability and generality of the evaluation.** Appendix A.3 reports variance and standard deviation statistics for correction and persistence behaviors across models, and Appendix A.4 extends the analysis to multi-label verdict prediction on PUBHEALTH.
  4. **Failure analysis on temporally novel claims.** Appendix A.5 presents hallucination cases where models over-confidently extrapolate beyond their verified knowledge boundaries, further motivating prior-aware arbitration.
  5. **Robustness to Output Variation.** Appendix A.6 examines whether different output-token choices affect the evaluation results and shows that our findings remain stable across semantically equivalent verbalizers.
- **Appendix B: Related Works.** This section positions our benchmark and our arbitration setting with respect to prior work on prior-context discrepancy, knowledge conflicts, and uncertainty-based knowledge-boundary estimation.
- **Appendix C: Further Details and Results of Our Method.** Appendix C.1 describes the JSD-based stability measure, while Appendix C.2 presents the **vote-based** variant for models without reliable probability access. The remaining parts provide case studies (App. C.3) and baselines definitions (App. C.4).
- **Appendix D: Dataset Details.** This section documents the dataset construction process, includ-

ing crawled emerging-event claims, public fact-checking sources, counterfactual evidence generation, and quality-control procedures.

- **Appendix E: Evaluation Prompts and Reproducibility.** This section lists the evaluation prompts, counterfactual construction prompts, and baseline prompts. The final parts report robustness checks for output-token variations and provide the LLM usage statement.

A Further Discussion

Observation 7. Counter-semantic counterfactual evidence is much more challenging than counter-entity evidence. As shown in Figure 6, most models (with the exception of Llama3-8B) exhibit a notable decrease in persistence rate when exposed to semantically counterfactual evidence compared to entity-level counterfactual evidence (e.g., a 56% drop for Mistral-7B). This trend, consistent with prior findings by Xie et al. (2023), indicates that semantic-level discrepancies pose a greater challenge to current LLMs and are more likely to induce incorrect judgments. Consequently, mitigating such discrepancies may require specialized preprocessing or safeguards, as reliance on internal judgment alone is often insufficient. This confirms the finding in prior work on the difficulty of LLMs autonomously handling deceptive or misleading content (Chen and Shu, 2023; Kamoi et al., 2024; Chen and Shu, 2024).

A.1 Influence of Temperature and # of Independent Runs

We examine the impact of the # of independent runs (up to 40) and temperature settings (0.3, 0.5, 1.0) on the consistency of model outputs and the relationship between the influence rate (i.e., correction (CR) and Not persistence (1-PR)) for KNOWN and UNKNOWN cases. The Figure 8 shows the analysis of Llama3-8B; the proportion of UNKNOWN claims stabilizes after approximately 20 runs (consistently across all three temperature settings), indicating a clear saturation point. In the middle and right panels, the influence rate for KNOWN cases decreases steadily with more runs, reflecting growing model confidence, whereas the influence rate of UNKNOWN cases shows only a negligible reduction. This confirms that UNKNOWN effectively captures intrinsic model uncertainty—a state in which further consistency checks do not improve confidence. Influence Rate of UNKNOWN shows limited sensitivity to the number (#) of runs).

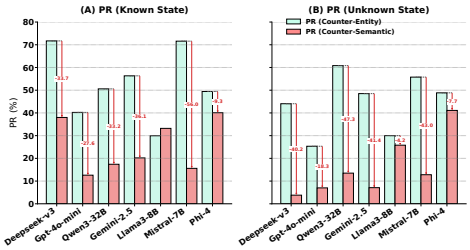


Figure 6: Persistence of counter-entity vs. -semantic.

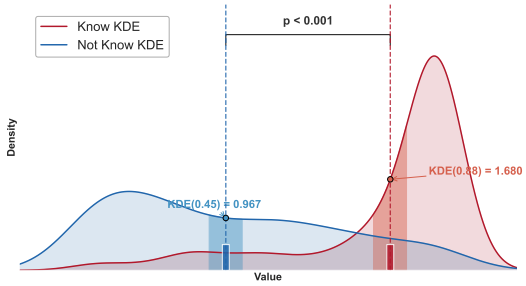


Figure 7: Probability distributions of class tokens ("Support"/"Refute") predicted by GPT-4o-mini.

A.2 Analysis of Known and Unknown Category

We experiment by prompting an LLM (e.g., GPT-4o-mini) for binary (Refute/Support) claim verification to measure their internal confidence (Taubenfeld et al., 2025; Fu et al., 2025), which is based on 500 claims, with 5 runs for each one under fixed temperature 1.0. KNOWN cases (red) cluster near 1, indicating high certainty, while UNKNOWN cases (blue) show broader spread, reflecting uncertainty (Kernel Density Estimation (KDE) smoothing is applied ($p < 0.001$)). Our results show that consistent judgments result in a sharp, low-entropy probability distribution of predicted class tokens, as illustrated in Figure 7. This indicates the reasonable category of known and Unknown.

A.3 The Variance of Evaluation Results

For the elevation results (Table 1), we present the variance and standard deviation statistics for the correction rate (CR) and persistence rate (PR) across several language models, shown in Table 5. Overall, the results reveal clear differences in the stability of models' correction and persistence behaviors. Notably, LLama3-8B consistently demonstrates the lowest variance and standard deviation, indicating a highly stable response pattern. In contrast, Qwen3-32b displays the highest variance and standard deviation values, suggesting

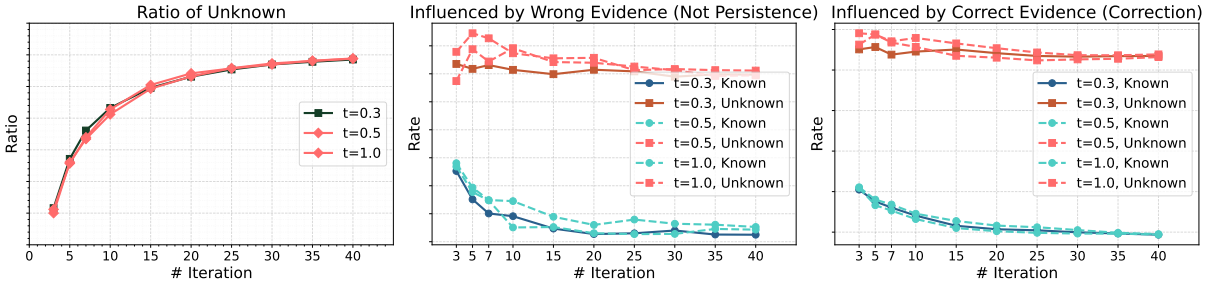


Figure 8: The figure based on the number of independent runs (from 0 to 40) with different temperatures. The proportion of UNKNOWN cases plateaus after reaching a certain number of runs (Left). The influence rate of KNOWN cases decreases more significantly with increasing number of runs when exposed to conflicting evidence, whereas UNKNOWN cases demonstrate a relatively stable influence rate throughout the process (Middle & Right).

Persistence Performance				
Model	**Known-Knows**		**Unknown-Knows**	
	PR (↑)	**OI** (↓)	**PR** (↑)	**OI** (↓)
Phi-4	29.50%	2.389	54.60%	0.832
Llama3-8B	30.71%	2.256	39.62%	1.524
Correction Performance				
Model	**Known-Unknows**		**Unknown-Unknows**	
	CR (↑)	**OI** (↑)	**CR** (↑)	**OI** (↑)
Phi-4	60.53%	1.534	55.29%	1.237
Llama3-8B	73.59%	2.786	55.39%	1.242

Table 3: Consolidated results on the PUBHEALTH dataset. Results are partitioned into Persistence Performance (on correct initial responses) and Correction Performance (on wrong initial responses).

Model	**Margin** ↑	**Model**	**Margin** ↑
Phi-4	-0.0008	**Llama3-8B**	-0.0069

Table 4: The calculated Margin values for each model on the PUBHEALTH dataset, indicating the differentiation capability between internal knowledge and external noise.

that its behavior is substantially less consistent and more variable. Some models exhibit similar variance and standard deviation between the correction rate and persistence rate, implying that the volatility in model output affects both metrics equally. In summary, LLama3 and Deepseek-v3 stand out as the most stable models among those evaluated, whereas Qwen3-32b is characterized by pronounced instability.

A.4 Multi-Label Verdict Prediction

While our current framework is not primarily designed for generating free-form explanations, it is inherently applicable to multi-label classification tasks. To validate its generalizability, we conducted supplementary experiments by collecting the PUBHEALTH dataset (Kotonya and Toni, 2020), involving 3,000 claims. This task requires models to classify claims into multiple categories, such as "True," "False," and "Unproven." The detailed results are summarized in Table 4 and Table 3.

We observe that when applied to this multi-label scenario, the performance Margin for Phi-4 remains negative, whereas for Llama3-8B, it remains positive. These findings are highly consistent with their respective behaviors in the binary-label tasks presented in the main manuscript, further confirming the robustness of our stability-based arbitration strategy across different classification granularities.

A.5 Hallucination Analysis

We analyze the results with an explanation based on GPT-4o-mini in *novel* claims from our benchmarks and results shown in Table 6. The LLM response explicitly states a knowledge cutoff date of October 2023 but proceeds to evaluate a claim involving a specific future date: May 18, 2025. This date corresponds to the purported final Everton F.C. match at Goodison Park. While the model correctly references Everton's publicly known stadium relocation plans and the typical timing of the Premier League season's conclusion, its concrete acceptance of May 18, 2025 as a plausible final match date exceeds verified knowledge boundaries. Specifically, since official fixture lists and exact match dates for the 2024–25 Premier League season would likely only be confirmed and publicly released in mid-2025, the model's certainty about

Model	Persistence Performance (PR)				Correction Performance (CR)			
	KNOWN-KNOWS		UNKNOWN-KNOWS		KNOWN-UNKNOWS		UNKNOWN-UNKNOWS	
	Var	Std	Var	Std	Var	Std	Var	Std
GPT-4o-mini	0.0034	0.0584	0.0087	0.0935	0.0093	0.0967	0.0064	0.0799
Deepseek-v3	0.0006	0.0244	0.0060	0.0774	0.0019	0.0435	0.0034	0.0585
Qwen3-32B	0.0521	0.2283	0.0569	0.2386	0.0451	0.2125	0.0450	0.2120
Gemini-2.5-flash	0.0024	0.0495	0.0049	0.0699	0.0129	0.1134	0.0064	0.0801
Mistral-7B	0.0009	0.0301	0.0009	0.0293	0.0053	0.0729	0.0035	0.0591
Phi-4	0.0046	0.0678	0.0038	0.0616	0.0050	0.0704	0.0033	0.0575
Llama3-8B	0.0009	0.0293	0.0002	0.0135	0.0009	0.0293	0.0002	0.0135

Table 5: Variance (Var) and Standard Deviation (Std) for **Persistence Rate (PR)** and **Correction Rate (CR)** across Known-Unknown Quadrant.

Claim	LLM Result
On May 18 2025: 2024–25 Premier League Everton F.C. play their final game at Goodison Park in Walton, Liverpool, England, the club's home ground for the last 133 years. They will transition to Everton Stadium in Vauxhall starting the 2025–26 season. (BBC Sport)	The claim should be supported based on the information available up to October 2023. Explanation: Everton FC has been planning to move to a new stadium known as Everton Stadium which is being constructed in the Vauxhall area of Liverpool. The club has announced that they intend to transition to this new stadium, which is expected to be completed in time for the 2025–26 season. As of my last update, the club had been preparing for this significant move, which would mark the end of their long history at Goodison Park, their home since 1892. The specific date mentioned in the claim (May 18, 2025) aligns with the conclusion of the 2024–25 Premier League season, making it plausible that this could be the date of their final game at Goodison Park. Therefore, the claim is consistent with the club's announcements and plans regarding their stadium transition.

Table 6: Data instance with response based on GPT-4o-mini.

the precise date is unsupported by direct evidence available before the cutoff. The response frames the date as "plausible" but ultimately judges the claim as "should be supported," which risks conflating reasonable inference with factual confirmation.

This constitutes a form of *hallucination*, where the model goes beyond its verified knowledge to make a detailed prediction without authoritative backing. Although not fabricated outright, it represents an overconfident extrapolation that could mislead users into treating the claim as a confirmed fact rather than an informed speculation.

Indicator	Value
Fleiss' Kappa ↑	0.8856
Observed Agreement (P_o)	0.9830
Expected Agreement (P_e)	0.8517
Number of Samples	465
Number of Raters (Prompts)	10

Table 7: Inter-agreement analysis (Fleiss' Kappa)

A.6 Robustness to Output Token Variations

Previous studies have noted that model performance can exhibit significant variance depending on the specific input prompts and the chosen output tokens during constrained decoding (Mahaut et al., 2024). As shown in Table 7 and 8, to rigorously evaluate Mistral-7B's sensitivity to output formats, we conducted a robustness experiment on a 500-sample subset. We designed 10 distinct prompts by varying the output constraints across five semantic pairs—*accurate/inaccurate*, *true/false*, *correct/incorrect*, *support/refute*, and *right/wrong*—including their capitalized variants. With the temperature set to 0.5, we measured inter-prompt consistency using Fleiss' Kappa. As shown in Tables 8 and 7, the results indicate a very high level of agreement. The overall Fleiss' Kappa across all 10 prompts reached **0.8856**, with all pairwise Kappa scores exceeding 0.78 and many surpassing 0.95. This demonstrates that the model's

	Accurant	accurant	Correct	correct	True	true	Support	support	Right	right
Accurant	-	0.8458	0.9173	1.0000	0.8458	0.8458	1.0000	0.8458	0.7832	0.7832
accurant	0.8458	-	0.9275	0.8458	1.0000	1.0000	0.8458	0.8632	0.9353	0.9353
Correct	0.9173	0.9275	-	0.9173	0.9275	0.9275	0.9173	0.9275	0.8634	0.8634
correct	1.0000	0.8458	0.9173	-	0.8458	0.8458	1.0000	0.8458	0.7832	0.7832
True	0.8458	1.0000	0.9275	0.8458	-	1.0000	0.8458	0.8632	0.9353	0.9353
true	0.8458	1.0000	0.9275	0.8458	1.0000	-	0.8458	0.8632	0.9353	0.9353
Support	1.0000	0.8458	0.9173	1.0000	0.8458	0.8458	-	0.8458	0.7832	0.7832
support	0.8458	0.8632	0.9275	0.8458	0.8632	0.8632	0.8458	-	0.8058	0.8058
Right	0.7832	0.9353	0.8634	0.7832	0.9353	0.9353	0.7832	0.8058	-	1.0000
right	0.7832	0.9353	0.8634	0.7832	0.9353	0.9353	0.7832	0.8058	1.0000	-

Table 8: Pairwise Kappa agreement matrix across different output tokens

fact-checking performance is robust and not a byproduct of specific output token selection.

B Related Works

Datasets of PCD. To the best of our knowledge, there is currently a lack of datasets specifically designed for fact-checking tasks within the context of Prior-Context Discrepancy (PCD). To address this gap, we introduce our benchmarks, a large-scale dataset engineered to evaluate how models navigate discrepancies between their internal memory and external information. While related research has explored model behavior in other overlapping scenarios, our benchmarks provides a more comprehensive and multi-dimensional analysis. Compared to representative datasets such as ClashEval (Wu et al., 2024), DynamicQA (Marjanovic et al., 2024) and (Xie et al., 2023), our benchmarks distinguishes itself through its extensive coverage of complex reasoning scenarios. Specifically, it evaluates model performance on retrieved new events where internal parametric knowledge is outdated, and systematically measures the impact of varying counterfactual context, including both counter-entity (CE) and counter-semantic (CS) counterfactual. Furthermore, different from (Su et al., 2024; Xie et al., 2023; Ming et al., 2024; Marjanovic et al., 2024), our benchmarks is unique in its ability to verify knowledge persistence—evaluating whether a model can consistently utilize accurate external knowledge to rectify its erroneous internal priors. Crucially, all evaluation dimensions within our benchmarks are validated across the four distinct internal epistemological states of the model: Known Knowns, Known Unknowns, Unknown Knowns, and Unknown Unknowns. This allows for a granular assessment when faced with varying degrees of epistemic certainty.

Estimation Knowledge Boundary via Uncertainty. In the context of identifying knowledge boundaries (Li et al., 2024; Hu et al., 2023), Uncertainty Estimation (UE) serves to quantify a model's lack of confidence in its predictions for a given input (Li et al., 2024; Hu et al., 2023). High uncertainty typically signifies that the input-related knowledge lies outside the model's certain boundaries. Specifically, the epistemic uncertainty plays an important role in UE, which refers to model-specific uncertainty that quantifies a lack of model knowledge, a concept directly linked to the Parametric Knowledge Boundary (Li et al., 2024).

Current solutions to quantify epistemic uncertainty are broadly classified into model-side and data-side approaches. Model-side quantification involves the perturbation of model parameters, configurations, or internal states. Data-side quantification encompasses two primary modes: Input-side clarification and perturbation, adjusting the input prompts to detect fluctuations in confidence (Gao et al., 2024; Ling et al., 2024). Output-side variation estimation, quantifying uncertainty based on the diversity or inconsistency of generated outputs (Yadkori et al., 2024; Aichberger et al., 2024).

Our framework specifically utilizes output-side variation estimation (Li et al., 2025), a consistency-based method where the inconsistency among mul-

tiple sampled predictions for a single input is viewed as the measure of uncertainty. In the specialized domain of fact-checking, consistency-based methods are uniquely sufficient for identifying knowledge boundaries. Fact-checking tasks target objective, verifiable truths where, ideally, only one correct verdict exists. If a model possesses the necessary parametric knowledge, its reasoning and final verdict should remain stable across multiple samplings. It indicates that the model is "guessing" due to a knowledge deficit rather than retrieving a stable fact, thereby making output variation a robust proxy for mapping the model's knowledge boundary without requiring access to internal model logits.

C Further Details and Results of Our Method

C.1 The Implementation Details of JSD

Our method aims to assess the internal consistency of a model's answers by leveraging the answer probability associated with each response and the algorithm shown in Algorithm 2.

As shown in Table 2, our method achieves competitive performance across six diverse language models, demonstrating remarkable generalization capability. It attains the best overall accuracy on GPT-4o-mini, Mistral-7B, Qwen3-32B, and Llama3-8B, while securing second place on Phi-4. The most striking advantage emerges on smaller-scale models: on Mistral-7B, our approach outperforms the second-best method by 6.58 absolute percentage points, revealing its particular efficacy in resource-constrained environments. This scalability advantage is crucial for practical deployments where computational efficiency is paramount. Furthermore, our method maintains balanced performance between correct evidence (CE) and incorrect evidence (IE) handling across most models, indicating robust reasoning capabilities rather than over-reliance on evidence quality. The sole exception occurs with Gemini-2.5, where ExplicitSCR performs best, suggesting model-specific optimization opportunities. We discuss the influence of different iteration times as follows.

We run train-test experiments three times to set hyperparameters τ. Each time renders an optimal τ that ranges in $[-0.08, -0.06]$, in which τ is determined using a small set of held-out validation data.

As shown in Figure 9, we compare the Jensen–Shannon Divergence (JSD) $\mathbf{JSD} = H(\bar{p}) - \mathbb{E}[H(p)]$ with correct evidence under two settings: answering with the claim only (Figure 9(a)) and answering with both the claim and external evidence (Figure 9(b)), using GPT-4o-mini under correct and incorrect response conditions. In the claim-only scenario (Figure 9(a)), the JSD values are similar regardless of whether the model's answer is correct or incorrect. In contrast, when external evidence is provided (Figure 9(b)), the JSD is noticeably lower when the model's answer is correct than when it is incorrect. This pattern suggests that GPT-4o-mini may possess a certain capability to discern evidence quality: while external evidence may still influence the model, it also introduces greater internal uncertainty, as reflected by the elevated JSD in incorrect responses.

C.2 The Vote-based JSD

While Our method aims to assess the internal consistency of a model's answers by leveraging the answer probability associated with each response, not all models provide direct access to reliable answer probabilities. To address this limitation, we introduce a simple variant of our approach that estimates stability using the results of multiple consistency checks for models that can not access the probability directly.

Specifically, we approximate the external stability measure $\mathbf{JSD} = H(\bar{p}) - \mathbb{E}[H(p)]$ using vote entropy—a frequency-based entropy computed from hard labels. Formally, given a set of K labels (in K runs), let $k_{\text{true}} \in [0, K]$ denote the number of labels that belong to any of the true categories. Then, the empirical probability is estimated as $\hat{p} = k_{\text{true}}/K$, and the vote entropy is computed as the binary entropy of $\hat{p}$: $H_{\text{binary}}(\hat{p}) = -\hat{p}\log_2(\hat{p}) - (1-\hat{p})\log_2(1-\hat{p})$ This variant allows us to robustly evaluate stability even when explicit probability outputs are unavailable. The algorithm is shown in Algorithm 1.

C.3 Case Study

To empirically validate the effectiveness of our stability-based evidence selection framework, we present a comprehensive case study analyzing three distinct factual verification scenarios. These cases demonstrate how our method dynamically arbitrates between parametric knowledge and external evidence based on output stability metrics. As

Algorithm 1 Evidence Stability via Vote Entropy (vote-based)

Require: Claim $\mathbf{c}$, Retrieved Evidence $\mathbf{e}$, an LLM f_θ, iteration times K, Threshold τ
Ensure: Decision: Provide the answer $\hat{y}$ based on voting consistency

Initialize counters: $k_{\text{ext}} \leftarrow 0$, $k_{\text{par}} \leftarrow 0$ ▷ Count of 'True' labels
Initialize label sets: $Y_{\text{ext}} \leftarrow \emptyset$, $Y_{\text{par}} \leftarrow \emptyset$
for $i = 1$ to K **do**
  $y_{\text{ext}}^{(i)} \leftarrow f_\theta(\mathbf{c}, \mathbf{e})$ ▷ Generate hard label with evidence
  $y_{\text{par}}^{(i)} \leftarrow f_\theta(\mathbf{c})$ ▷ Generate hard label without evidence
  if $y_{\text{ext}}^{(i)} \in$ Positive Categories **then** $k_{\text{ext}} \leftarrow k_{\text{ext}} + 1$
  end if
  if $y_{\text{par}}^{(i)} \in$ Positive Categories **then** $k_{\text{par}} \leftarrow k_{\text{par}} + 1$
  end if
end for
Compute empirical probabilities: $\hat{p}_{\text{ext}} \leftarrow \frac{k_{\text{ext}}}{K}$, $\hat{p}_{\text{par}} \leftarrow \frac{k_{\text{par}}}{K}$
Compute vote entropy (stability):
$V_{\text{ext}} \leftarrow H_{\text{binary}}(\hat{p}_{\text{ext}}) = -\hat{p}_{\text{ext}} \log_2(\hat{p}_{\text{ext}}) - (1 - \hat{p}_{\text{ext}}) \log_2(1 - \hat{p}_{\text{ext}})$
$V_{\text{par}} \leftarrow H_{\text{binary}}(\hat{p}_{\text{par}}) = -\hat{p}_{\text{par}} \log_2(\hat{p}_{\text{par}}) - (1 - \hat{p}_{\text{par}}) \log_2(1 - \hat{p}_{\text{par}})$
$\Delta = V_{\text{par}} - V_{\text{ext}}$ ▷ Positive Δ means evidence increases consistency
if $\Delta \leq \tau$ **then**
  return $\hat{y} = \text{Majority}(Y_{\text{par}})$ ▷ Evidence introduces too much disagreement
else
  return $\hat{y} = \text{Majority}(Y_{\text{ext}})$ ▷ Evidence leads to a more stable consensus
end if

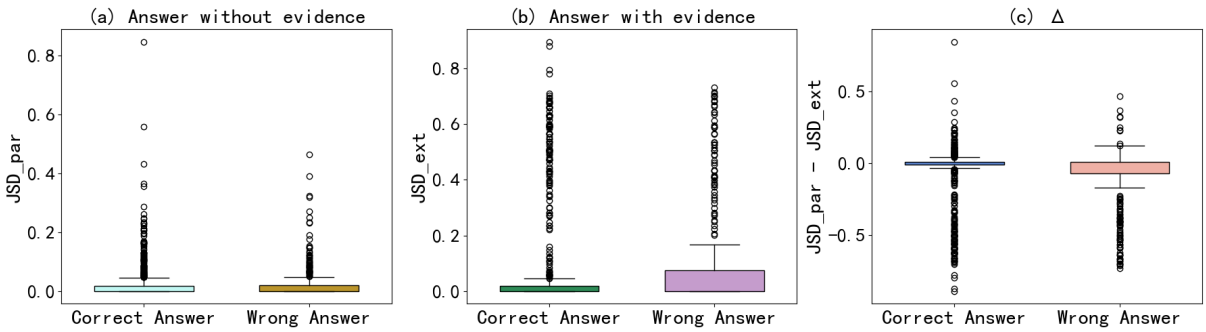


Figure 9: The comparison of our defined JSD with and without evidence.

shown in Table 9.

The first case involves a claim regarding high school arrest statistics and graduation rates, where the provided external evidence contains internal contradictions (marked with ×). Our method computes a negative Δ value (-0.2229), correctly identifying that the evidence introduces substantial variability ($\mathbf{JSD}_{\text{par}} = 0.2239$) compared to the highly stable parametric responses ($\mathbf{JSD}_{\text{ext}} = 0.0011$). Consequently, the system rejects the conflicting evidence and maintains the original "support" classification based solely on parametric knowledge.

In the second case examining trade agreement impacts on employment, the external evidence provides consistent factual support (marked with ✓). Here, the parametric responses show significant instability ($\mathbf{JSD}_{\text{ext}} = 0.1438$) with mixed "refute"/"support" predictions, while evidence-augmented queries yield stable consensus ($\mathbf{JSD}_{\text{par}} = 0.0165$). The positive Δ (0.1273) correctly triggers evidence utilization, shifting the majority prediction from "refute" to "support."

The third case, concerning national debt growth under different presidential administrations, further illustrates the method's precision. Despite initial parametric uncertainty (mixed responses with $\mathbf{JSD}_{\text{ext}} = 0.0316$), the external evidence produces near-perfect stability ($\mathbf{JSD}_{\text{par}} \approx 2.53 \times 10^{-8}$). The positive Δ (0.0316) leads to evidence adoption, resulting in unanimous "refute" decisions that

Algorithm 2 Evidence Stability via Jensen-Shannon Divergence (standard)

Require: Claim $\mathbf{c}$, Retrieved Evidence $\mathbf{e}$, an LLM f_θ, iteration times K, Threshold τ
Ensure: Decision: Provide the answer $\hat{y}$ by using or not using external evidence $\mathbf{e}$

Initialize arrays $\{p_{\text{ext}}^{(i)}\}_{i=1}^{K}$, $\{p_{\text{par}}^{(i)}\}_{i=1}^{K}$
for $i = 1$ to K **do**
  $p_{\text{ext}}^{(i)} \leftarrow f_\theta(\mathbf{c}, \mathbf{e})$ ▷ Sample with evidence
  $p_{\text{par}}^{(i)} \leftarrow f_\theta(\mathbf{c})$ ▷ Sample without evidence
end for
Compute mean probabilities: $\bar{p}_{\text{ext}} \leftarrow \frac{1}{K}\sum_{i=1}^{K} p_{\text{ext}}^{(i)}, \quad \bar{p}_{\text{par}} \leftarrow \frac{1}{K}\sum_{i=1}^{K} p_{\text{par}}^{(i)}$
Compute evidence stability: $\mathbf{JSD}_{\text{ext}} \leftarrow H(\bar{p}_{\text{ext}}) - \frac{1}{K}\sum_{i=1}^{K} H(p_{\text{ext}}^{(i)})$ ▷ Using the binary entropy function
Compute parametric stability: $\mathbf{JSD}_{\text{par}} \leftarrow H(\bar{p}_{\text{par}}) - \frac{1}{K}\sum_{i=1}^{K} H(p_{\text{par}}^{(i)})$ ▷ Using the binary entropy function
$\Delta = \mathbf{JSD}_{\text{par}} - \mathbf{JSD}_{\text{ext}}$
if $\Delta \leq \tau$ **then**
  Do not use evidence, **return** $\hat{y} = f_\theta(\mathbf{c})$ ▷ External evidence introduces variability
else
  Use evidence, **return** $\hat{y} = f_\theta(\mathbf{c}, \mathbf{e})$
end if

correctly contradict the original claim.

These cases collectively demonstrate three critical capabilities: (1) the ability to detect and reject unreliable evidence that introduces instability; (2) the capacity to identify and utilize high-quality evidence that enhances prediction consistency; and (3) the robustness to make appropriate evidence selection decisions across diverse factual domains without requiring explicit credibility assessments of the evidence content. The stability metric Δ serves as an effective proxy for evidence quality, enabling reliable arbitration between internal knowledge and external information sources.

C.4 Baseline Methods

Following Huang et al. (2024), we evaluate eight methods for evidence reliability assessment under six prominent LLMs, grouped into three categories: **Self-guided methods**: **1)** ImplicitSCR: The LLM is prompted with a claim and potentially incorrect evidence, then judges the reliability of the evidence before producing a final response. **2)** ExplicitSCR: The LLM generates answers using internal knowledge and external context, performs explicit confidence reasoning via CoT and answer comparison, and then balances both sources to make a final decision. **Rule-based methods**: **3)** InternalEval: The LLM evaluates its own initial parametric response; if deemed correct, it is kept, otherwise a new answer is generated using external evidence. **4)** ContextEval: The LLM evaluates external evidence; if correct, the evidence is used, otherwise it responds only with the claim. **5)** InternalConf (Jiang et al., 2023b): The LLM chooses between internal and evidence-based answers based on token-level probabilities, using a predefined threshold. **6)** ContextConf: Similar to InternalConf but applied to evidence-first responses, i.e., the selection is based on prompting with evidence first. **7)** TPC: Compares confidence scores between internal and context-based answers, using raw probabilities or calibrated percentiles. **Context-based methods**: **8)** TACS-LR: Filters unreliable context using LLM-based removal, then uses the refined context to generate the final answer Yu et al. (2024); Huang et al. (2024). Consequently, these methods are evaluated only on GPT-4o-mini, Phi-4, and Mistral-7B, from which such probability data can be obtained. The comprehensive results are presented in Table 2.

- **Implicit Self-Guided Confidence Reasoning (ImplicitSCR):** The model is explicitly instructed that the provided context may contain inaccuracies and must exercise independent judgment to evaluate its reliability before formulating a response. To prioritize the activation of internal knowledge, the context is deliberately positioned after the question. Some works (Huang et al., 2024) demonstrate that this sequential structure increases the model's reliance on prior knowledge, thereby reducing its vulnerability to mis-

Claim ($\mathbf{c}$)	High school students arrested on campus are twice as likely not to graduate and four times less likely to graduate if they've appeared in court.
External Evidence $\mathbf{e}$	high school dropouts are three and one-half times more likely than high school graduates to be arrested, and more than eight times as likely to be incarcerated. 12 pages, seventy-five middle school students who were under court have been involved in the judicial system, making sure they "complete college". High school dropouts are three and one-half times more likely than high school graduates to be arrested, and more than eight times as likely to be incarcerated. ×
Initial Answer $f_\theta(\hat{y} \mid \mathbf{c}_i)$	"support", "support", "support", "support", "support", "support", "support", "support", "support", "support"
Initial Answer Probabilities	0.9965, 0.9950, 0.9971, 0.9858, 0.9950, 0.9965, 0.9965, 0.9965, 0.9971, 0.9958
Answer with evidence $f_\theta(\hat{y} \mid \mathbf{c}_i, \mathbf{e}_i)$	"refute", "refute", "refute", "refute", "refute", "refute", "refute", "refute", "refute", "refute"
Evidence Probabilities	0.9399, 0.9241, 0.4115, 0.4620, 0.9399, 0.9399, 0.8808, 0.4620, 0.9399, 0.4046
$\mathbf{JSD}_{\text{ext}}$ (claim only)	0.0011
$\mathbf{JSD}_{\text{par}}$ (claim with conflicting evidence)	0.2239
$\Delta = \mathbf{JSD}_{\text{ext}} - \mathbf{JSD}_{\text{par}}$	≈ -0.2229 (Choose answer with claim only)
Claim ($\mathbf{c}$)	Trade agreements like NAFTA and permanent normal trade relations with China, which forced American workers to compete against people making pennies an hour, has resulted in the loss of 160,000 jobs here in Michigan."
External Evidence $\mathbf{e}$	trade agreements like NAFTA and permanent normal trade relations with China, which forced American workers to compete against people making pennies an hour, have resulted in the loss of 160,000 jobs here in Michigan. The steel industry setbacks account for just a fraction of the job losses in Michigan's manufacturing sector, which now employs 55,100 fewer workers than it did when Trump took office in January 2017, U.S. Labor Department data shows. The state's automotive industry accounted for 35% of the manufacturing job losses, according to the st. Louis fed. Trade agreements like NAFTA and permanent normal trade relations with China, which forced American workers to compete against people making pennies an hour, have resulted in the loss of 160,000 jobs here in Michigan. ✓
Initial Answer (10 times)	"refute", "refute", "refute", "refute", "support", "refute", "support", "support", "refute", "support"
Initial Answer Probabilities $f_\theta(\hat{y} \mid \mathbf{c})$	0.9193, 0.6770, 0.6183, 0.6171, 0.3707, 0.6770, 0.8121, 0.3707, 0.6199, 0.8121
Answer with Evidence $f_\theta(\hat{y} \mid \mathbf{c}_i, \mathbf{e})$	"support", "refute", "support", "support", "support", "support", "support", "support", "support", "support"
Evidence Probabilities	0.8175, 0.1480, 0.8519, 0.9526, 0.8175, 0.9526, 0.8519, 0.9046, 0.8519, 0.9046
$\mathbf{JSD}_{\text{ext}}$ (claim only)	0.1438
$\mathbf{JSD}_{\text{par}}$ (claim with conflicting evidence)	0.0165
$\Delta = \mathbf{JSD}_{\text{ext}} - \mathbf{JSD}_{\text{par}}$	≈ 0.1273 (Choose answer with evidence)
Claim ($\mathbf{c}$)	Says under President Barack Obama, the debt increased by 23 percent, which was less than under any president going back to Ronald Reagan.
External Evidence ($\mathbf{e}$)	27 Feb 2023 under President Barack Obama, the national debt grew from $10.63 to $19.96 trillion, a 87.8% increase. Under President Donald Trump, the ... Oct 7, 2022, Bush nearly matched the amount of debt accumulated under Reagan, but ... Bush, democratic president barack obama added another $8.34 trillion ... May 15, 2023, Ronald Reagan and George W. Bush. President Ronald Reagan increased the U.S. debt by around 1.86 trillion U.S. dollars, or 186.36 percent. this ...✓
Initial Answer (10 times)	"refute", "support", "refute", "support", "refute", "support", "refute", "support", "refute", "support"
Initial Answer Probabilities $f_\theta(\hat{y} \mid \mathbf{c}_i)$	0.3760, 0.8485, 0.3760, 0.6764, 0.3760, 0.6183, 0.3760, 0.8485, 0.4360, 0.7230
Answer with Evidence $f_\theta(\hat{y} \mid \mathbf{c}_i, \mathbf{e}_i)$	"refute", "refute", "refute", "refute", "refute", "refute", "refute", "refute", "refute", "refute"
Evidence Probabilities	0.9999973, 0.9999973, 0.9999973, 0.9999971, 0.9999973, 0.9999973, 0.9999981, 0.9999973, 0.9999971, 0.9999971
$\mathbf{JSD}_{\text{ext}}$ (claim only)	0.0316
$\mathbf{JSD}_{\text{par}}$ (claim with conflicting evidence)	2.5305e-08
$\Delta = \mathbf{JSD}_{\text{ext}} - \mathbf{JSD}_{\text{par}}$	≈ 0.0316 (Choose answer with evidence)

Table 9: Here are examples of our method for selecting answers, with and without supporting evidence, demonstrated across 10 independent runs. Our approach can successfully identify the correct initial answer even when the external evidence is inaccurate (indicated by ×). Additionally, it selects the answer backed by evidence when that evidence is correct (indicated by ✓)

leading contextual information. The model subsequently engages in implicit confidence reasoning during its decision process and returns only the final answer. Prompts shown in Table 17.

- **Explicit Self-Guided Confidence Reasoning (ExplicitSCR):** In this approach, the model initially generates distinct responses derived from its internal knowledge and the provided context, using two separate prompts. It subsequently engages in verbalized confidence estimation through a chain-of-thought reasoning pro-

cess. This begins with an evaluation of the confidence in its internally generated answer, including reflection on the factual basis underlying that response. The model then assesses the reliability of the external context by comparing it against the facts supporting its internal knowledge. The final answer is selected through a deliberative reasoning process that integrates both internal certainty and contextual credibility. Prompts shown in Table 17.

- **Internal Evaluation (InternalEval):** The LLM assesses the correctness of its internal answer through self-evaluation using the prompt provided in Table 18. If the self-evaluation confirms the answer is correct, the internal answer is retained; otherwise, the context-based answer is selected.

- **Context Evaluation (ContextEval):** The model evaluates the relevance and accuracy of the provided context in relation to the question (see Table 18 for the prompt specification). If the context is judged to be correct and reliable, the context-based answer is adopted; otherwise, the model defaults to its internal knowledge-based response.

- **Internal Confidence Thresholding (InternalConf):** The model selects its internal knowledge-based answer if its confidence exceeds a predefined threshold; otherwise, it defaults to the context-based answer. The threshold can be set to a fixed value (e.g., 0.5) or calibrated on a held-out dataset when available. Confidence estimates may be derived from sequence probabilities. Following ActiveRAG (Jiang et al., 2023a), Huang et al. (2024) simplifies to this internal confidence mechanism (InternalConf) to employ answer-level probability, yielding modest but consistent empirical improvements.

- **Context Confidence Thresholding (ContextConf):** The model evaluates the reliability of the external context. If confidence in the context exceeds a predefined threshold, the context-based answer is selected; otherwise, the model defaults to its internal knowledge-based response. The threshold can be determined using the same methodology as in Internal Confidence estimation. Confidence scoring may be derived from the sequence probability of the answer given the context.

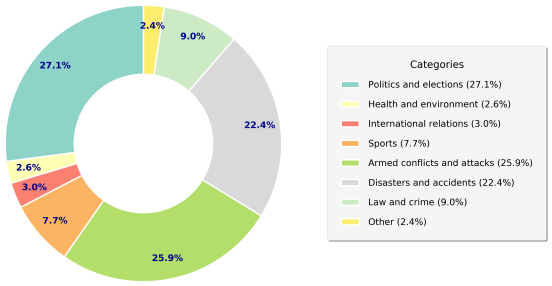


Figure 10: The categories of our crawled data.

- **(Calibrated) Token Probability Correction (TPC):** Following (Wu et al., 2024), compare the confidence scores—specifically, the mean token probabilities—of the model's internal answer and the context-based answer, selecting the one with the higher value as the final answer. This approach is termed token probability correction.

- **Truth-Aware Context Selection (TACS-LR)** Following (Yu et al., 2024), who provide Truth-Aware Context Selection (Yu et al., 2024), a context evaluation method, which employs a classifier to filter out incorrect content at a granular sentence or token level. Unlike rule-based approaches that simply choose between internal and context-based answers, this method reintegrates the filtered context into the LLM to regenerate the final answer. Huang et al. (2024) adopts this to an alternative strategy: the model explicitly removes untruthful sentences (TACS-LR; see Table 19 for the prompt), and the refined context is used to produce the final answer.

D The Details of Datasets

To ensure the quality of automatically constructed conflicting evidence, we established a multi-stage quality control process to verify its plausibility and reliability. The generated evidence underwent systematic filtering and validation under multiple constraints: during generation, the LLM was instructed to ensure that substituted entities shared the same type as the original and that at least ten key entities could be extracted from the context to guarantee semantic richness. After automatic construction, we conducted manual sampling reviews by three human annotators. This rigorous workflow ensures the high quality and challenging nature of the final dataset.

Crawled data. Our crawled dataset, which contains 3,995 true claims, is used to analyze LLM performance on unseen emerging events; the examples are shown in Table 13. Concretely, we collected *new* events as claims by crawling the Wikipedia current events portal in different months in 2024 and 2025 (e.g., `https://en.wikipedia.org/wiki/Portal:Current_events/June_2025`). If one LLM, such as Llama3-8B, which is published in April 2024, then the parametric knowledge is deemed to be unseen for the events that occurred in June 2025. So we annotated the events which crawled from the web page June 2025 is *new* to Llama3-8B. The category distributions are displayed in Figure 10, and all the categories are sourced from Wikipedia.

The data collection. We collect three public datasets:

- QuanTemp (Anand et al., 2024) is a diverse, multi-domain dataset exclusively focused on numerical claims. It encompasses comparative, statistical, interval, and temporal aspects, and includes detailed metadata along with a supporting evidence collection.
- Snopes: We use the dataset from (Popat et al., 2017), which consists of rumors from the general fact-checking website Snopes (www.snopes.com). Each instance includes an editor-assigned credibility label (true or false), alongside associated reporting articles and their sources.
- PolitiFact: Popat et al. (2018) extracted a dataset from PolitiFact (www.politifact.com), which focuses on claims made by U.S. political figures. Our dataset includes all articles published prior to December 2017, each containing the claim, the speaker, and an official credibility rating.

Counterfactual data quality control. We provide additional examples of the constructed counterfactual evidence in Table 11. To improve the plausibility and reliability of the generated counterfactual evidence, we apply a multi-stage quality control procedure. During generation, we impose type and complexity constraints: for entity-level counterfactuals, the substituted entity must have the same semantic type as the original entity; for semantic-level counterfactuals, the generated evidence must remain relevant to the original claim while introducing a factual conflict. We also require that at least five key entities can be extracted from the context, ensuring that the evidence contains sufficient factual structure for controlled intervention. And the detailed prompts of dataset construction in 16 and 15.

Human validity audit. Because our counterfactual evidence is generated with LLM assistance, we conduct a human validity audit to verify that the constructed evidence induces the intended prior-context discrepancy, rather than introducing unrelated content or obvious generation artifacts. The audit focuses on whether each counterfactual instance carries the targeted factual intervention.

We randomly sample M=100 instances from the final counterfactual set, stratified by data source (QuanTemp / PolitiFact / Snopes), counterfactual type (entity-level / semantic-level), and gold label (SUPPORT / REFUTE). Each instance is independently annotated by three annotators along four binary criteria: (i) claim relevance, whether the evidence still concerns the same claim or event; (ii) intervention consistency, whether it introduces the intended factual conflict; (iii) label consistency, whether it provides the intended misleading verification signal for the original claim; and (iv) artifact-free plausibility, whether it is fluent and free of obvious generation artifacts. An instance is counted as valid if the majority of annotators pass the first three criteria.

As shown in Table 10, the criterion-level pass rates range from 88.0% for label consistency to 95.0% for claim relevance, with substantial inter-annotator agreement (Fleiss' κ=0.76). Entity-level counterfactuals score higher than semantic-level ones, as entity substitutions are more tightly constrained. The lower label consistency rate reflects cases where the intended factual intervention is present but does not unambiguously flip the verification signal, which is a known challenge in counterfactual data construction (Zmigrod et al., 2019). Overall, the audit suggests that the benchmark largely satisfies its intervention design, while semantic-level counterfactuals remain more challenging to construct reliably.

E Evaluation Prompts

Overall, we provide the prompts of basic evaluation Table 14. The detailed prompts of Counter-entity and Counter-Semantic in 16 and 15. The prompts of baselines are shown in Table 17, Ta-

Criterion	Entity-CF	Semantic-CF	Overall
Claim relevance	96.0	94.0	95.0
Intervention consistency	94.0	88.0	91.0
Label consistency	90.0	86.0	88.0
Artifact-free plausibility	94.0	90.0	92.0
Fleiss' κ	0.78	0.73	0.76

Table 10: Human audit of counterfactual evidence quality (M=100, 3 annotators per instance). We report majority-vote pass rates (%) and inter-annotator agreement (Fleiss' κ). An instance is retained as valid if the majority vote passes the first three criteria.

ble 18 and Table 19.

F LLM Usage Claim

In this paper, we used GPT-4o to assist counterfactual evidence construction under controlled prompting. All generated counterfactuals were filtered and partially audited by human annotators.

Additionally, LLMs are utilized exclusively for the purpose of aiding and polishing writing. Their application is strictly confined to improving linguistic clarity, coherence, grammar, and style within textual content. No additional functionalities are incorporated.

Claim (c)	Label	Evidence (e)	Constructed Evidence with counterfactual ($\bar{e}$)
— crawled new claims —			
On April 6 2025: Russian invasion of Ukraine. 2025 Sumy Oblast incursion Russian troops reportedly capture the village of Basivka in Sumy Oblast, Ukraine. (Reuters) Kyiv strikes A Russian airstrike in Darnytskyi District, Kyiv, Ukraine, kills one person and injures three others. (CTV News) Kryvyi Rih strikes The death toll from Friday's missile strike on Kryvyi Rih, Ukraine, rises to 20 deaths, including several children, and 75 injuries. (CTV News)	True	... In an effort to counter Ukraine's 2024 offensive in Kursk, in early 2025 Russian forces launched a cross-border offensive from Russia into Ukraine's adjacent Sumy Oblast. ...	... In an effort to counter Poland's "2026 offensive" in Voronezh, in early 2026 Russian forces launched a "counter-terrorism operation" from Germany into Poland's adjacent Cherkasy Oblast. ...
On April 6 2025: Red Sea crisis March–April 2025 United States attacks in Yemen. Between four and 70 people are killed and at least 16 others are injured in overnight U.S. airstrikes targeting Houthi forces in Saada, Yemen. (CTV News)	True	... In March 2025, the United States launched a large campaign of air and naval strikes against Houthi targets in Yemen. Codenamed Operation Rough Rider, it has been the largest U.S. military operation in the Middle East of President Donald Trump's second term ...	... In March 2026, the United States launched a large campaign of air and naval strikes against Houthi targets in Oman. Codenamed Operation Iron Stallion, it has been the largest U.S. military operation in the Middle East of President Joe Biden's second term ...

Table 11: Examples of external evidence with counterfactual (Counter-entity).

Claim (c)	Label	Evidence (e)	Constructed Evidence with counterfactual ($\bar{e}$)
federal tax code loopholes giving incentives companies shipping jobs overseas	True	... I also want to close those loopholes that are giving incentives for companies that are shipping jobs overseas. Obama said Wednesday right now you can actually take a deduction for moving a plant overseas...	For firms choosing to remain in the United States, the current tax code provides significant incentives for companies that invest domestically. Instead of creating disincentives for offshoring, the tax structure encourages businesses to expand their operations within the U.S. by offering substantial tax credits for job creation and capital investment.
Mitt Romney gave government health care bankrupting state Massachusetts	False	It has been nearly five years since Massachusetts gov mitt Romney signed the state's landmark health care law... the overhaul is largely seen as a blueprint for the sweeping federal legislation.	... However, the law has faced significant backlash, with many residents expressing dissatisfaction over rising premiums and limited access to care. Critics argue that the law has not delivered on its promises... Brian Rosman has since distanced himself from the initial celebration.

Table 12: Examples of external evidence with counterfactual (Counter-semantic).

Claim	Evidence	Time	Outdated for LLMs
July 31, 2025. The military junta of Myanmar formally ends the country's four-year-long state of emergency and declares a December 2025 election for the country's new head of government and legislative members.	...Myanmar's military government plans to hold a general election for elected seats in the Amyotha Hluttaw and the Pyithu Hluttaw of the Assembly of the Union, currently dissolved, on a date in December 2025 to be determined. The planned election would be the first after the 2021 military coup d'état...	July 31, 2025	Deepseek-v3, GPT-4o-mini, Llama3-8B, Phi-4, Gemini-2.5-flash, Mistral-7B
2025 FIFA Club World Cup. In association football, English club Chelsea F.C. win the FIFA Club World Cup for the second time, defeating French club Paris Saint-Germain 3–0 in the final. Chelsea winger Cole Palmer is awarded the tournament's Golden Ball. (FIFA)	The 2025 FIFA Club World Cup final was the final match of the 2025 FIFA Club World Cup, the 21st edition of the premier competition for men's club soccer teams organized by FIFA. The match was played at MetLife Stadium at the Meadowlands Sports Complex in East Rutherford, New Jersey, near New York City, on July 13, 2025. It was contested between English club Chelsea and French club Paris Saint-Germain.	July 13, 2025	Deepseek-v3, GPT-4o-mini, Llama3-8B, Phi-4, Gemini-2.5-flash, Mistral-7B
2024 Maltese presidential election. Myriam Spiteri Debono is sworn in as President of Malta, succeeding George Vella and becoming the third woman to hold the office. (Times of Malta)	The 2024 Maltese presidential election took place on 27 March 2024. Members of the Parliament of Malta voted in an indirect election to elect the next President of Malta with former parliament speaker Myriam Spiteri Debono being the only nominee	April 4, 2024	Mistral-7B
Arab–Israeli conflict. Gaza war. Gaza war hostage crisis. Hamas releases a propaganda video showing Israeli American hostage Edan Alexander. (CBS News). An Israeli airstrike on a car in the Gaza Strip kills five people, including employees of World Central Kitchen. (AP). Israel–Hezbollah conflict. Two people are killed and six others are injured in three airstrikes by Israel in southern Lebanon.	On 27 November 2024, a ceasefire agreement was signed by Israel, Lebanon, and five mediating countries. The agreement mandates a 60-day halt to hostilities, during which Israel must withdraw its forces from Southern Lebanon, and Hezbollah must withdraw its forces to north of the Litani River. Since the ceasefire went into effect, Lebanese sources claim Israeli attacks on Lebanon killed at least 83 civilians.	Nov 30, 2024	GPT-4o-mini, Llama3-8B, Mistral-7B

Table 13: We present examples of crawled events with evidence from our benchmarks. This information is new for some LLMs, indicating that their parametric knowledge is outdated.

Parametric or External Knowledge
Parametric Knowledge Assessment (Without Evidence) Based solely on your internal knowledge, determine whether the claim should be supported or refuted. Provide only "support" or "refute" as the final answer without any explanation. **Task:** Claim: {claim} Answer:
External Knowledge Integration (With Evidence) Based on the provided information and your internal knowledge, determine whether the claim should be supported or refuted. Provide only "support" or "refute" as the final answer without any explanation. **Task:** Claim: {claim} Information: {evidence} Answer:

Table 14: The basic prompts of evaluation.

Counterfactual Evidence Generation
True Claim Refutation Prompt Given a claim labeled as true, generate realistic but fabricated evidence to refute it. Create convincing fake content that contradicts the true claim. **Task:** Claim: {claim} (label: true) Evidence:
False Claim Support Prompt Given a claim labeled as false, generate realistic but fabricated evidence to support it. Create convincing fake content that validates the false claim. **Task:** Claim: {claim} (label: false) Evidence:

Table 15: Prompts of counterfactual semantic evidence generation framework. We use the counter-label of claim to set the prompts.

Entity Substitution
Step 1: Entity Extraction Prompt Extract the entities from the following evidence that directly influence the judgment of the claim. **Example:** Claim: On April 6 2025: Red Sea crisis March–April 2025 United States attacks in Yemen. Between four and 70 people are killed and at least 16 others are injured in overnight U.S. airstrikes targeting Houthi forces in Saada, Yemen. Evidence: ... In March 2025, the United States launched a large campaign of air and naval strikes against Houthi targets in Yemen. Code- named Operation Rough Rider, it has been the largest U.S. military operation in the Middle East of President Donald Trump's second term ... Entities: ...March 2025, Yemen, Rough Rider, Donald Trump's second term... **Task:** Claim: {claim} Evidence: {evidence}
Step 2: Replaced Entity Generation Prompt Given the entity, generate a similar type but a different entity. Only output the new entity. **Example:** Entity: ...March 2025, Yemen, Rough Ride, Donald Trump's second term... New Entity: ...March 2026, Oman, Iron Stallio, Joe Biden's second term **Task:** Entity: {ent}

Table 16: The prompts of Counter-entity construction. We let LLM generate the replaced entity one by one.

ImplicitSCR
Prompt: You will be given a question and an evidence. The evidence may not be trustworthy. Use your judgment to assess the reliability of the evidence. Then, based on both your assessment and your own knowledge, provide the best possible answer. **Question:** [question] **Evidence:** [evidence] **Answer:**
ExplicitSCR
Prompt: Task Overview: You will be given a question along with your internal answer, evidence that may contain either true or false information, and the evidence's answer to the same question. Your task is to evaluate the reliability of the evidence and determine whether the evidence is deceptive or not. Steps: 1. Internal Reasoning: Reflect on how you arrived at your internal answer using your own knowledge. Break down your reasoning process and assess the confidence level of your original answer, explaining why you believe your answer is correct. 2. Evidence Evaluation: Analyze the evidence and cross-reference the information provided with the known facts you used to form your internal answer. Determine whether the evidence contains deceptive or unreliable information, considering possible contradictions or inconsistencies. 3. Final Judgment: Based on your analysis, decide which answer (your internal answer or the evidence's answer) is more likely to be correct. Clearly state your final answer. **Question:** {question} **Your answer:** {internal answer} **The evidence to judge:** {evidence} **The evidence answer:** {evidence answer} Please provide a detailed reasoning process, followed by your final judgment. Ensure the last line of your response contains only the final answer without any additional explanation or details.

Table 17: The prompt structures of ImplicitSCR and ExplicitSCR

InternalEval
Prompt: Your task is to evaluate the model's response to a question. You will be provided with a question, the model's answer. Your job is to determine whether the model's answer is true or false. **Question:** [question] **Model Answer:** [model answer] **Is the model's answer true or false?** Return "True" if the model's answer is correct, and "False" if the model's answer is incorrect.
ContextEval
Prompt: You will be given a question and evidence that answers the question. Your task is to evaluate whether the evidence provides a correct answer to the question. If the evidence's answer is correct, return "True"; otherwise, return "False". **Question:** [question] **Evidence:** [doc] **Is the evidence correct?** Return "True" if the evidence's answer is correct, and "False" if the evidence's answer is incorrect.

Table 18: The prompt structures of InternalEval and ContextEval

TACS-LR
Step 1: Evidence Filtering Prompt You will be given an Evidence and a question. You need to remove the sentence that you think is not correct. You can only do removal and you can not add any new information or change the existing information. Only return the filtered evidence as your output. **Example:** Evidence: The Eiffel Tower is located in Paris, France. It is the tallest structure in Paris. The Eiffel Tower was built in the 19th century and is made of wood. Question: Where is the Eiffel Tower located? Filtered Evidence: The Eiffel Tower is located in Paris, France. It is the tallest structure in Paris. The Eiffel Tower was built in the 19th century. **Task:** Question: {question} Evidence: {evidence}
Step 2: Final Answer Generation Prompt Your task is to answer the "Question" based on the provided "Evidence". **Example:** Question: Is the claim 'The capital of France is Paris' true or false? Do not generate the process, just answer "true" or "false" only! Evidence: Paris has been the administrative, political, and cultural capital of France since 987 AD when the Capetian dynasty established their power base there, a status that has remained uninterrupted for over a thousand years. Answer: true **Task:** Question: {question} Evidence: {filtered evidence}

Table 19: The prompt structure of TACS-LR