PLAUSIEVE:A Distribution Based Framework for Measuring Model Belief Beyond Logits

Arghodeep Nandi, Tanmoy Chakraborty
Indian Institute of Technology Delhi
{eez248395, tanchak}@iitd.ac.in

Abstract

Prompting has transformed large language models into general-purpose decision systems, where token-level probabilities are often treated as indicators of model preference and correctness. However, directly interpreting these probabilities is challenging due to calibration disparities, tokenizer variability, and inconsistent output distributions across models, limiting their statistical comparability. This raises a fundamental question: *when can a model be said to be confident in its decision or preference?* Addressing this question remains an open and important area of research. We introduce `PLAUSIEVE`, a novel distribution-based statistical framework that estimates model preference through a forward pass without token generation. `PLAUSIEVE` provides a calibration-robust and model-independent mechanism for quantifying preference without relying on fixed thresholds or individual tokens. We apply `PLAUSIEVE` in estimating model belief, across two prompt settings (direct and epistemic), two confidence levels, and five models from four families to study prompt–architecture interactions. Our analysis shows significant improvements under epistemic prompting, with nearly 100% gains in certain model combinations. Notably, `Llama-3.2-3B` outperforms the larger `Gemma-7B` under epistemic prompting in terms of both decisiveness and veracity, demonstrating that model architecture plays a more significant role than scale in shaping belief formation. [1]

1 Introduction

The emergence of prompting has fundamentally transformed how Large Language Models (LLMs) are used and evaluated. Instead of requiring task-specific training, modern LLMs can perform diverse tasks through natural language instructions alone, enabling flexible adaptation to classification, reasoning, and decision-making tasks without parameter updates (Brown et al., 2020; Wang and Zhou, 2024; Ouyang et al., 2022). This paradigm shift has enabled LLMs to function as general-purpose decision systems, where model outputs are treated as judgments over competing alternatives.

Prompt-based evaluation is now widely used to obtain model judgments in tasks such as factual verification (Xu et al., 2024; Thorne et al., 2018), hallucination detection (Ji et al., 2023), model calibration and confidence estimation (Wu et al., 2026; Kadavath et al., 2022), reward modeling and alignment (Ouyang et al., 2022; Bai et al., 2022), and LLM as a judge evaluation frameworks (Huang et al., 2025). In these settings, models are frequently prompted to assign relative likelihoods or select between mutually exclusive outcomes such as `True` and `False`. These token-level probabilities and outcomes are then interpreted as indicators of model preference, confidence, or correctness.

However, directly interpreting token probabilities presents challenges. Table 1 shows that for the universally false claim 'The Earth is Flat," probability distributions vary substantially across models despite identical prompts, highlighting limitations of probability based interpretation:

- Probability and probability-derived scores vary drastically across models, preventing the use of universal fixed thresholds (Table 1).
- The highest-probability token may not align with the intended answer (e.g., `Gemma-7B` in Table 1), and binary decisions collapse confidence differences, assigning identical outcomes to probabilities such as 0.51 and 0.99, reflecting a mismatch between internal belief and expressed output (Wu et al., 2026).
- Token probability distributions vary across models and exhibit poor fit (e.g., inconsistent KS statistics; Table 1), making parametric thresholding unreliable.

[1] Our source code is publicly available at `https://github.com/ohgra/plausieve`.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 300–328
November 6-10, 2026 ©2026 Association for Computational Linguistics

Model	Token-Level Probability Analysis for the Claim "The Earth is Flat" (Direct Prompt)
`Llama-8B`	• Token with maximum logit: `False` • Final layer shape: `[128256]` • True token probability: `8.28505e-06` • False token probability: `7.20978e-04` • $(P_{\text{True}} - P_{\text{False}})$: `-7.12693e-04` • $\log\left(\frac{P_{\text{True}}}{P_{\text{False}}}\right)$: `-4.46616` • $\log\sqrt{\frac{P_{\text{True}}}{P_{\text{False}}}}$: `-2.23308` • Mean KS Statistic of token distributions [log-norm]: `[0.05100]`
`Llama-3B`	• Token with maximum logit: `False` • Final layer shape: `[128256]` • True token probability: `1.69277e-05` • False token probability: `4.42505e-03` • $(P_{\text{True}} - P_{\text{False}})$: `-4.40812e-03` • $\log\left(\frac{P_{\text{True}}}{P_{\text{False}}}\right)$: `-5.56608` • $\log\sqrt{\frac{P_{\text{True}}}{P_{\text{False}}}}$: `-2.78304` • Mean KS Statistic of token distributions [log-norm]: `[0.04872]`
`Phi4`	• Token with maximum logit: `False` • Final layer shape: `[200064]` • True token probability: `1.28895e-06` • False token probability: `2.05994e-03` • $(P_{\text{True}} - P_{\text{False}})$: `-2.05865e-03` • $\log\left(\frac{P_{\text{True}}}{P_{\text{False}}}\right)$: `-7.37660` • $\log\sqrt{\frac{P_{\text{True}}}{P_{\text{False}}}}$: `-3.68830` • Mean KS Statistic of token distributions [log-norm]: `[0.15034]`
`Gemma-7B`	• Token with maximum logit: `\n\n\n` • Final layer shape: `[256000]` • True token probability: `4.70877e-06` • False token probability: `1.03516e-01` • $(P_{\text{True}} - P_{\text{False}})$: `-1.03511e-01` • $\log\left(\frac{P_{\text{True}}}{P_{\text{False}}}\right)$: `-9.99805` • $\log\sqrt{\frac{P_{\text{True}}}{P_{\text{False}}}}$: `-4.99903` • Mean KS Statistic of token distributions [log-norm, norm]: `[0.36663, 0.50063]`
`Mistral-7B`	• Token with maximum logit: `False` • Final layer shape: `[32000]` • True token probability: `5.55883e-09` • False token probability: `9.60938e-01` • $(P_{\text{True}} - P_{\text{False}})$: `-9.60937e-01` • $\log\left(\frac{P_{\text{True}}}{P_{\text{False}}}\right)$: `-18.96803` • $\log\sqrt{\frac{P_{\text{True}}}{P_{\text{False}}}}$: `-9.48402` • Mean KS Statistic of token distributions [log-norm]: `[0.05048]`

Table 1: Model-wise analysis of final-layer token probabilities for the claim "The Earth is Flat", comparing the likelihood assigned to `True` and `False`. The table reports raw probabilities, probability difference, and log-odds-based measures derived from the output distribution, highlighting each model's relative preference between the two tokens.

Prior work has attempted to mitigate few of these issues through alternative measures such as Domain Conditional Pointwise Mutual Information (Holtzman et al., 2021) and Statement Accuracy Prediction, based on Language Model Activations (SAPLMA)(Azaria and Mitchell, 2023). However, these approaches still leave an important question unresolved: *how large must the probability difference between two competing tokens (e.g., "True" and "False") be to be considered meaningfully significant?*. Existing approaches largely operate within a two-dimensional truth subspace (Bürger et al., 2024), overlooking confusion as a dimension, needed for a more complete answer to the question. Recent work on conformal factuality estimation provides statistically grounded guarantees over LLM output reliability, but does not investigate intrinsic uncertainty within the vocabulary distribution (Cherian et al., 2024).

To address this underexplored problem, we introduce `PLAUSIEVE`, a distribution-based scoring framework that models uncertainty over the full vocabulary distribution rather than relying on direct token-level probability comparisons. Instead of using the raw probability gap between tokens, we apply multinomial sampling to generate multiple empirical vocabulary distributions. The probability difference between the target token groups is then computed for each sample, producing an empirical distribution of token-level differences. This converts the original multivariate output distribution into a univariate difference distribution, enabling a principled significance assessment, making it robust to tokenizer variations, and architectural heterogeneity. `PLAUSIEVE` enables analysis beyond binary outcomes by identifying regions where models are not strongly decisive. Modeling such intermediate states is highly applicable for the study of uncertainty in model belief (Section 2).

`PLAUSIEVE` estimates models internal confusion directly from the output distribution. Viewing model's conceptual knowledge (Li et al., 2024; Wang et al., 2024) as stored in superposition (Elhage et al., 2022; Hu et al., 2025) within the architecture parameters, this setup allows us to examine how different prompt types elicit belief from that latent structure while interacting with the parameters. Importantly, we avoid relying on external evidence signals, as doing so would confound intrinsic parameter level interaction with evidence matching behavior (Min et al., 2022; Levine et al., 2021); instead, our setup isolates prompt–parameter(architecture) interactions as the

primary driver of the observed belief. We apply this framework across five models from four families under two prompting strategies: direct and epistemic (Section 2).

Using this framework, we investigate the following research questions:

- **RQ1:** How do different prompting strategies (direct vs epistemic) influence model belief?
- **RQ2:** How do model architecture and scale affect belief formation and responsiveness under different prompting conditions?
- **RQ3:** Do models exhibit systematic bias or differences in decisiveness, and can ensembling improve decision consistency?

Our analysis reveals three key findings. First, epistemic prompting significantly reshapes model belief, improving decisiveness and accuracy for architectures such as `Llama-3B` and `Llama-8B`, while increasing confusion in others like `Gemma-7B` and `Phi4`. Second, model architecture influences belief formation more strongly than scale, although larger models within the same family still benefit from scaling. Third, ensemble models, particularly architecturally aligned ones, reduce response bias, improve decisiveness, and consistently achieve more stable and reliable belief estimation than individual models.

2 Definitions

Belief. Let $c \in \mathcal{C}$ denote a claim and $\mathcal{M}$ denote a language model. Drawing inspiration from cognitive science, where beliefs are enduring internal representations held with conviction even in the absence of full evidence (Connors and Halligan, 2015), we define *model belief* as the model's internal assessment of the truth of a claim c, reflected in its responses to structured prompts. Formally, belief is a function

$$\mathcal{B}_{\mathcal{M}} : \mathcal{C} \to [0, 1],$$

where $\mathcal{B}_{\mathcal{M}}(c)$ quantifies the strength of the model's internal assessment of the claim. This formulation is used for direct prompting strategy.Values of $\mathcal{B}_{\mathcal{M}}(c)$ closer to the extremes indicate stronger belief, whether in favor of or against the claim, reflecting the model's latent conviction encoded in its parameters. Values near the middle of the scale indicate a weak or uncertain stance, corresponding to a confused state. In practice, we discretize belief into $\{0, 0.5, 1\}$ corresponding to rejection (strong opposing belief), uncertainty (weak belief), and acceptance (strong belief).

Epistemological Questioning. Epistemological questioning draws inspiration from philosophy and recent interdisciplinary research on inference in language models. The Socratic method, which systematically examines reasoning through targeted peripheral questions, has been shown to improve critical thinking and improved use of knowledge (He et al., 2024; Gu et al., 2024; Abdullah et al., 2022; Sahamid, 2016; Fierro et al., 2024).

Motivated by these studies, we employ epistemological questioning to probe the internal stance of a language model toward a claim c. Rather than asking the model directly whether it accepts the claim, we ask a set of peripheral questions $\mathcal{Q} = \{q_1, \ldots, q_6\}$ that examine various aspects of the claim's epistemic validity:

- q_1: Is there scientific evidence that supports the claim?
- q_2: Do experts in the relevant field agree with the claim?
- q_3: Is the claim falsifiable (i.e., can it be tested or proven false)?
- q_4: Could this claim be affected by bias, propaganda, or motivated reasoning?
- q_5: Does the claim contradict known facts or scientific consensus?
- q_6: Is the claim logically consistent with existing knowledge?

The final belief about a claim c is aggregated from the model's responses $r_i = \mathcal{M}(c, q_i)$, where

$$r_i \in \{\texttt{True}, \texttt{False}, \texttt{Confused}\},$$

using a procedure corresponding to Dung's abstract argumentation framework (Dung, 1993). In this procedure, three questions (q_1, q_2, q_6) reinforce the claim, while the remaining three (q_3, q_4, q_5) challenge it.

The procedure aggregates responses by counting how many questions in each group are labeled `True`. Let s denote the count over $\{q_1, q_2, q_6\}$ and a denote the count over $\{q_3, q_4, q_5\}$. The final belief is assigned as 1 if $s > a$, 0 if $a > s$, and 0.5 in the case of a tie.

This formulation captures the model's belief by balancing supporting and attack responses, reflecting epistemic support, conflict, and uncertainty. The details are provided in Section 5.2.

3 Dataset

To enable a three-way classification setting comprising **True**, **False**, and **Confused** categories, we required datasets that go beyond strict dichotomies. In real world fact checking scenarios, the veracity of claims is often not strictly binary, as certain claims may be partially supported, context dependent, or difficult to determine with certainty. This inherent subjectivity motivates the inclusion of a **Confused** category, which captures model uncertainty when it cannot reliably distinguish between **True** and **False**.

Our dataset is compiled from several well-established fact-checking benchmarks including FEVER (Thorne et al., 2018), AVERITEC (Schlichtkrull et al., 2023), LIAR (Wang, 2017), POLITIFACT (Shu et al., 2019), and FEVEROUS (Aly et al., 2021). These datasets follow heterogeneous labeling schemes across sources. To ensure consistency, we normalize all original labels into three unified categories: **True**, **False**, and **Confused** (Table 2). The **Confused** label captures cases in which the claim's veracity is ambiguous, partially verifiable, or inherently uncertain. A key observation across these benchmarks is the substantial presence of claims containing numerical information. Numerical claims constitute 37.96% of LIAR, 50.74% of AVERITEC, 38.48% of POLITIFACT, 18.67% of FEVER, and 69.35% of FEVEROUS. Given their prevalence and practical importance, focusing on numerical claims provides strong coverage of real-world fact-checking scenarios while enabling targeted analysis.

Accordingly, we extract and analyze five categories of numerical claims: *Statistical*, *Demographic*, *Temporal*, *Financial*, and *Quantitative*. These categories are defined based on contextual cues surrounding numerical quantities in each claim. This categorization allows us to isolate variations attributable to data source characteristics rather than methodological artifacts, thereby adding an additional analytical dimension to our study. The final dataset, formed by combining five numerical claim types with three unified label categories, comprises 23,951 claims. Of these, 11,157 are labeled **True**, 8,345 as **False**, and 4,449 as **Confused**. A detailed breakdown across claim types and source datasets is provided in the Appendix A.

4 PLAUSIEVE

The `PLAUSIEVE` module is the central mechanism used to quantify a language model's belief in a given claim. It is model-independent and relies solely on the output probability distribution over vocabulary tokens obtained via a forward pass; no extra tokens are generated.

The method compares the average probability mass assigned to token sets corresponding to `True` and `False` responses, denoted by $\mathcal{T}$ and $\mathcal{F}$, respectively, chosen to provide an objective and generalizable evaluation across models. The validity of these sets are discussed in Appendix D.[2] A key challenge is determining whether the difference between these two sets is statistically significant. Across models, not only does the size of the vocabulary distribution vary, but the probability mass, their differences and log variants between these tokens also fluctuates significantly, especially when tested on universal claims such as "The Earth is Flat," as detailed in Table 1. We observe considerable uncertainty in fitting a uniform parametric model to token-level vocabulary distributions. Hence, it is both more robust and more generalizable to adopt a non-parametric approach in analyzing token-level belief signals.[3]

`PLAUSIEVE` estimates uncertainty via a non-parametric bootstrapping procedure. Given a vocabulary probability distribution, it repeatedly samples from a multinomial distribution and computes the empirical distribution of token-level differences between the `True` and `False` set of tokens, transferring the problem from a multivariate space to an univariate space.

From this empirical distribution, a confidence interval (CI) is constructed to quantify belief. The final belief label is determined by whether the CI lies entirely above, below, or spans zero. A higher confidence interval implies stricter belief decisions, while lower intervals increase permissiveness. We report results using two thresholds: a standard 68% CI ($\pm 1\sigma$, percentiles 16–84) and a relaxed 38% CI ($\pm \frac{1}{2}\sigma$, percentiles 31–69), where σ denotes the standard deviation under normality assumption, enabling percentile-based interval construction. This

[2] True set: `{true, True, yes, Yes}`; False set: `{false, False, no, No}`, in alignment with the prompts used.

[3] The Kolmogorov-Smirnov (KS) statistic measures the divergence between empirical and reference distributions. Values range from 0 (perfect fit) to 1 (maximum divergence), with typical interpretation: 0.00-0.10 = good fit; 0.10–0.30 = moderate deviation; >0.30 = poor fit.

Unified Label Category	Mapped Raw Labels from Source Datasets
True	5, 4, `Supported`, `SUPPORTS`, `true`, `mostly-true`
False	1, 0, `Refuted`, `REFUTES`, `false`, `mostly-false`, `pants-fire`
Confused	2, 3, `Not Enough Evidence`, `Conflicting Evidence/Cherrypicking`, `half-true`, `NOT ENOUGH INFO`, `half-false`

Table 2: Mapping of heterogeneous raw labels from source datasets into unified **True**, **False**, and **Confused** categories.

Algorithm 1 Belief Scoring with PLAUSIEVE

Require: Vocabulary probability distribution $\mathbf{p} \in \mathbb{R}^V$; Index sets $\mathcal{T} \subseteq \{1, \ldots, V\}$ (`True` tokens), $\mathcal{F} \subseteq \{1, \ldots, V\}$ (`False` tokens); Number of bootstrap samples $K = 400$; Sample size $n = 20000$

Ensure: Belief score $\in \{0, 0.5, 1\}$

Compute observed mean difference:

$$\Delta_{\text{obs}} \leftarrow \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} P_i - \frac{1}{|\mathcal{F}|} \sum_{j \in \mathcal{F}} P_j$$

Initialize list $\{\Delta_k\}_{k=1}^{K}$

for $k = 1$ to K **do**

  Draw sample $\mathbf{x}^{(k)} \sim \text{Multinomial}(n, \mathbf{p})$

  Compute:

$$\Delta_k \leftarrow \frac{1}{n|\mathcal{T}|} \sum_{i \in \mathcal{T}} x_i^{(k)} - \frac{1}{n|\mathcal{F}|} \sum_{j \in \mathcal{F}} x_j^{(k)}$$

  Append Δ_k to list

end for

Compute confidence interval:

$$[\Delta_{\text{low}}, \Delta_{\text{high}}] \leftarrow \text{Percentile}_{\alpha, 1-\alpha}(\{\Delta_k\})$$

▷ $\alpha = 0.16$ for 68% confidence or $\alpha = 0.31$ for 38% confidence

Compute final belief score:

$$\text{Belief}(c) \leftarrow \begin{cases} 1 & \text{if } \Delta_{\text{low}} > 0 \\ 0 & \text{if } \Delta_{\text{high}} < 0 \\ 0.5 & \text{otherwise} \end{cases}$$

return Belief(c)

is justified asymptotically by the Central Limit Theorem due to the large number of generated samples.

Positioning PLAUSIEVE: scope and research gap. We note that PLAUSIEVE is not evaluated against existing uncertainty or confidence-estimation methods (e.g., Domain Conditional PMI (Holtzman et al., 2021), SAPLMA (Azaria and Mitchell, 2023), truth-subspace probing (Bürger et al., 2024), or conformal factuality estimation (Cherian et al., 2024)) as direct empirical baselines. This is a deliberate scoping choice rather than an oversight. As shown in Table 1, raw probabilities and their simple differences or log-ratios can vary by orders of magnitude across model families, even for claims with unambiguous ground truth. None provides a principled criterion for determining how large a difference must be to indicate a meaningful preference rather than noise in an under-confident distribution.

Domain Conditional PMI corrects for surface-form bias but does not model significance or uncertainty. SAPLMA relies on hidden activations and treats belief as a covert, classifier-learned property rather than an overt, quantifiable one. Truth-subspace probing distinguishes true from false activations but overlooks confusion as an intermediate state. Conformal factuality estimation provides guarantees over generated text rather than intrinsic uncertainty within a single forward pass. These limitations reflect differences in scope, not adaptation difficulty. Each method is orthogonal to the question addressed by PLAUSIEVE: determining when a model's token-level preference reflects genuine confidence rather than noise. Recall that model belief $\mathcal{B}_{\mathcal{M}}(c)$ is discretized into rejection, uncertainty, and acceptance (Section 2); assigning a claim to the uncertain, confused middle state rather than to one of the two extremes is precisely a judgment about whether an observed difference is significant. Belief estimation therefore requires addressing this gap, which persists across existing methods.

PLAUSIEVE is built to operate directly within this gap. It recasts belief estimation as a bootstrap-derived confidence interval over the `True`$-$`False` mean difference, so that the discretization into rejection, uncertainty, and acceptance follows from a statistically grounded significance test rather than

an arbitrary cutoff. This mapping also brings the widely varying raw values in Table 1 onto a common statistical scale, independent of vocabulary size, tokenizer, and model family. To our knowledge, PLAUSIEVE is the first module to cast model belief in these quantitative, distribution-based terms. We position it as an engineering contribution that establishes a common benchmark for future belief-estimation methods, rather than claiming superiority over existing confidence-estimation techniques.

5 Experimental Setup

This section describes how belief is elicited, aggregated, and evaluated across models, prompting strategies, and datasets. The setup is designed to align with our definition of belief, the three-way labeling scheme (**True**, **False**, **Confused**), and the distribution based scoring framework PLAUSIEVE. A schematic overview of the complete pipeline is provided in Figure 1.

5.1 Belief Elicitation via Forward Pass

For a claim c, we quantify model belief using few shot prompting without generating free form text. Instead of decoding responses, we perform a single forward pass and extract the vocabulary probability distribution at the final token position. This distribution serves as the input to PLAUSIEVE, which produces a belief score. We use either direct prompting, where the model evaluates the claim using a direct question, or epistemological prompting, where the claim is assessed through multiple epistemic questions whose scores are aggregated as described in Section 2. Details are provided in Appendix B.

5.2 Dung Framework

Each claim is assessed using six epistemological questions, with response options being `True`, `False`, or `Confused`. Among these, three questions – q_1 (evidence), q_2 (expert agreement), and q_6 (logical consistency), are treated as supports for the claim, while the remaining three, q_3 (testability), q_4 (bias), and q_5 (contradiction), are considered attacks. The belief decision is derived by comparing the number of supporting and attacking responses labeled `True`. The process is summarized in form of Algorithm 2.

Algorithm 2 Dung's Framework for Evaluation of Epistemological Belief

Require: Epistemological answers: `Q1_evidence`, `Q2_expert`, `Q3_testable`, `Q4_bias`, `Q5_contradiction`, `Q6_logic`
Ensure: Belief label $\in \{0, 0.5, 1\}$
Support set: $S = \{$`Q1_evidence`, `Q2_expert`, `Q6_logic`$\}$
Attack set: $A = \{$`Q4_bias`, `Q5_contradiction`, `Q3_testable`$\}$
Compute: $s \leftarrow \sum_{q \in S} \text{indicator}(q = \texttt{True})$
Compute: $a \leftarrow \sum_{q \in A} \text{indicator}(q = \texttt{True})$
if $s > a$ **then**
    return 1 ▷ Belief is True
else if $a > s$ **then**
    return 0 ▷ Belief is False
else
    return 0.5 ▷ Confused/Uncertain
end if

5.3 Models and Confidence Settings

We evaluate five instruction-tuned language models spanning four architectural families and multiple parameter scales: `Llama-8B`, `Llama-3B`, `Gemma-7B`, `Phi4`, and `Mistral-7B`. Each model is evaluated using both prompting strategies.

Belief scores are computed using PLAUSIEVE under two confidence thresholds: a stricter 68% confidence interval ($\pm 1\sigma$, percentiles 16–84) and a relaxed 38% interval ($\pm \frac{1}{2}\sigma$, percentiles 31–69). The stricter threshold requires stronger separation from zero assigning a non-confused belief label, while the relaxed threshold allows more flexibility . This allows us to study how different levels of uncertainty affect belief assignment.

5.4 Model Ensembles

To study robustness and collective behavior, we construct ensembles over all possible combinations of the five models. For $r \in \{1, 2, 3, 4, 5\}$, we evaluate every subset, yielding $2^5 - 1 = 31$ unique configurations. Each ensemble aggregates the ternary belief scores $\{0, 0.5, 1\}$ from its constituent models by averaging across models for each claim, and the final ensemble decision is determined as 1 if the average > 0.5, 0 if the average < 0.5, and 0.5 if the average $= 0.5$. This systematic construction enables analysis of how ensemble size and architectural diversity influence stability, agreement, and decisiveness.

5.5 Evaluation Protocol

Evaluation is organized across five numerical claim categories: *Statistical*, *Demographic*, *Temporal*,

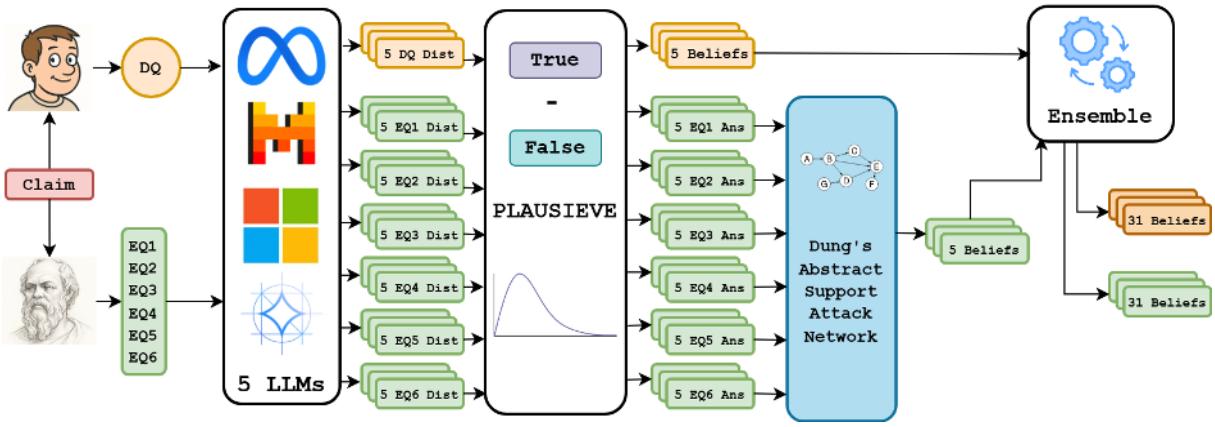


Figure 1: Schematic overview of the experimental setup using PLAUSIEVE, illustrating how token-level probability distributions from five LLMs are transformed into belief scores under both Direct Questioning (DQ) and Epistemological Questioning (EQ). The diagram depicts the progression from model logits to final model belief estimates.

Financial, and *Quantitative*, along with an aggregated *overall* set. This results in six evaluation groups. In total, we analyze 31 ensemble configurations, evaluated under 2 prompting strategies, 2 confidence thresholds, and 6 evaluation sets. This results in $31 \times 2 \times 2 \times 6 = 744$ experimental settings. The breadth of this setup enables a controlled and systematic investigation of prompting effects, architectural influence, uncertainty calibration, and ensemble behavior on model belief formation.

5.6 Metrics.

For each configuration, we compute overall accuracy and Jensen–Shannon (JS) divergence from the aggregated three-way confusion matrix. Accuracy measures predictive correctness, while JS divergence quantifies distributional mismatch between predicted and gold label distributions.

To jointly capture them, we introduce the *Model Effectiveness Score (MES)*:

$$\text{MES} = \text{Accuracy} \times (1 - \text{JS Divergence})^{\gamma}.$$

We set $\gamma = 2$ to balance predictive accuracy with distributional alignment, preventing accuracy alone from masking systematic distributional mismatch.The quadratic exponent penalizes larger divergences more strongly while keeping scores stable when distributions are well aligned. We conducted a hyperparameter sensitivity analysis for γ in Appendix C.4.

Together, this experimental framework enables a structured analysis of how prompting strategy, architectural variation, uncertainty estimation, and ensemble aggregation can interact to shape model belief.

6 Results and Analysis

This section presents results organized around our core analytical themes. We analyze how prompting strategy, model architecture and scale, and ensemble construction influence belief formation as measured by PLAUSIEVE under two confidence intervals: 68% (narrow) and 38% (wider). A more detailed breakdown across datasets and ensembles are provided in Appendix C.

Effect of Prompting Strategy on Model Belief. To examine how direct versus epistemic prompting influences belief formation, we compare MES scores across both confidence intervals. Overall, epistemic prompting reshapes the output distribution, i.e., the distribution over True, False, and Confused responses, often increasing decisiveness for certain architectures while amplifying confusion for others.

Ensembles involving `Llama-3B` and `Llama-8B` demonstrate consistent and often large improvements in MES under epistemic prompting. Across the dataset, gains exceed **100%**, particularly for temporal, statistical, and quantitative claims. For example, the ensemble of `Llama-3B` and `Mistral-7B` yields an MES increase of **218%** under the wider confidence interval (Table 3).

In contrast, `Gemma-7B` and `Phi4` show performance degradation under epistemic prompting, with predictions increasingly clustering around the confused state (Table 6-11). This divergence indicates that epistemic prompts do not uniformly enhance belief quality, but instead interact with architectural characteristics. We further observe a relationship between epistemic performance and

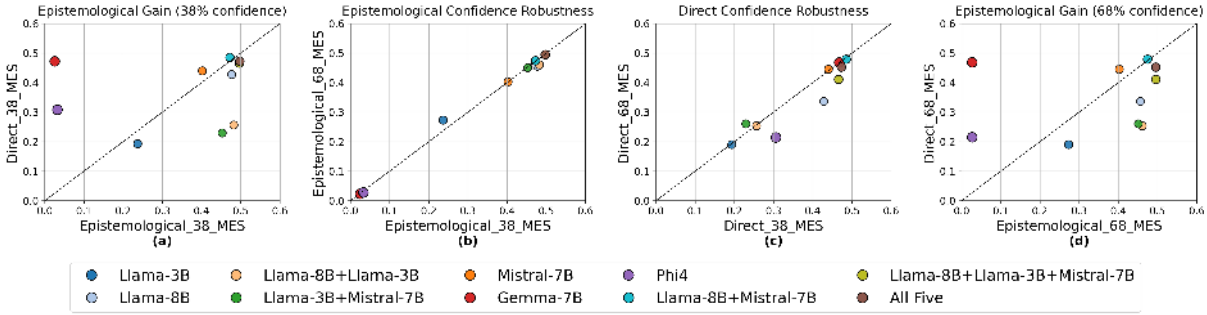


Figure 2: MES scatter plots across four prompting and confidence settings. **(a)** shows performance gains from epistemological prompting (38%); **(b) and (c)** illustrate MES stability across confidence levels; **(d)** shows performance gains from epistemological prompting (68%). Meta ensembles consistently outperform, while `Gemma` and `Phi` cluster near the origin, signaling epistemic confusion.

Dataset	Model Combination	Percentage Accuracy Gain (p-value, Chi-Square) (38%)	MES Gain (%) (38%)	Percentage Accuracy Gain (p-value, Chi-Square) (68%)	MES Gain (%) (68%)	Epistemological Marginal MES (38%)
Quantitative	`Llama-3B + Mistral-7B`	**79.93 (9.07e-52)**	**202.67**	**70.09 (3.91e-46)**	**156.56**	`Llama-8B` (0.1966)
	`Llama-8B + Llama-3B`	**73.45 (6.61e-56)**	**144.22**	54.49 (3.06e-38)	86.81	`Llama-3B` (0.1589)
	`Llama-3B + Llama-8B + Phi4`	28.57 (1.70e-16)	25.50	34.73 (4.79e-21)	51.58	`Mistral-7B` (0.1811)
Temporal	`Llama-3B + Mistral-7B`	**80.15 (5.13e-119)**	**218.69**	**71.60 (9.85e-105)**	**172.36**	`Llama-8B` (0.1849)
	`Llama-8B + Llama-3B`	64.24 (2.13e-101)	**145.07**	51.19 (3.22e-73)	**109.78**	`Llama-3B` (0.1521)
	`Llama-8B + Llama-3B + Phi4`	34.01 (9.55e-43)	51.07	38.27 (4.34e-49)	63.46	`Mistral-7B` (0.1748)
Demographic	`Llama-3B + Mistral-7B`	33.16 (4.09e-32)	77.18	23.49 (7.72e-19)	49.36	`Llama-8B` (0.1903)
	`Llama-8B + Llama-3B`	40.71 (2.87e-50)	69.10	30.72 (2.16e-32)	46.90	`Llama-3B` (0.1520)
	`Llama-8B + Llama-3B + Phi4`	23.11 (4.05e-22)	31.42	24.62 (2.46e-23)	33.06	`Mistral-7B` (0.1757)
Statistical	`Llama-8B + Llama-3B`	32.00 (2.22e-30)	63.67	40.42 (2.92e-40)	**145.38**	`Llama-8B` (0.1741)
	`Llama-8B + Gemma-7B`	**-0.35 (0.89)**	-0.74	33.26 (1.23e-29)	**113.61**	`Llama-3B` (0.1382)
	`Llama-8B + Llama-3B + Gemma-7B`	7.70 (8.81e-04)	11.12	33.99 (1.99e-31)	**121.76**	`Mistral-7B` (0.1555)
Financial	`Llama-3B + Mistral-7B`	19.39 (1.28e-16)	59.28	15.53 (1.09e-11)	48.46	`Llama-8B` (0.1761)
	`Llama-8B + Llama-3B`	28.31 (4.96e-32)	**72.70**	20.99 (4.42e-19)	64.88	`Llama-3B` (0.1423)
	`Llama-8B + Llama-3B + Phi4`	17.26 (5.43e-15)	34.68	15.10 (1.79e-11)	44.09	`Mistral-7B` (0.1619)
Overall	`Llama-3B + Mistral-7B`	18.84 (2.87e-08)	**97.31**	17.23 (2.22e-07)	**73.25**	`Llama-8B` (0.1863)
	`Llama-8B + Llama-3B`	32.00 (2.88e-20)	**88.17**	40.44 (1.38e-27)	**81.50**	`Mistral-7B` (0.1728)
	`Llama-8B + Llama-3B + Phi4`	19.13 (2.21e-09)	39.11	12.70 (4.74e-05)	44.12	`Llama-3B` (0.1500)

Table 3: Comparison of MES gains under epistemological prompting for two confidence intervals (68% and 38%). Percentage accuracy gains are reported with Chi-Square test p-values for both 38% and 68% settings. Gains are computed relative to direct prompting as baselines. For average marginal contributions across the four settings, refer to Table 13. For results across all 31 model ensembles and claim types refer to Table 6-12

KS-statistic based goodness-of-fit to parametric distributions (Table 1), suggesting that models with more stable output distributions respond more coherently to epistemic questioning. This may suggest that epistemic prompting interacts more coherently with models whose output distributions exhibit stable parametric structure, indicating that belief expression may depend more on distributional regularity than raw probability magnitude.

Role of Architecture and Scale in Belief Formation. Results indicate that model architecture has a greater influence on model belief than model size. Notably, `Llama-3B` (MES: 43.68% increase at 68% CI) outperforms larger models such as `Gemma-7B` (MES: 94.53% decrease at 68% CI) under epistemic prompting (Table 6). This suggests that architectural alignment with the prompting structure may be more important than model size or scale.

Within the same architectural family, scale still contributes positively. `Llama-8B` consistently outperforms `Llama-3B` across experiments (Table 6-11), indicating that scale enhances belief based performance when architectural priors are shared. Interestingly, `Mistral-7B` shows moderate standalone performance but improves significantly when paired with `Llama-3B`, while no comparable gain is observed when paired with `Llama-8B`. This indicates reduced scale effects in cross-architecture interactions. These findings are counterintuitive given prior studies (Wei et al., 2022), which suggest that model scale is a significant driver of inferential ability. Thus underscoring the critical role of architectural choices in epistemic tasks.

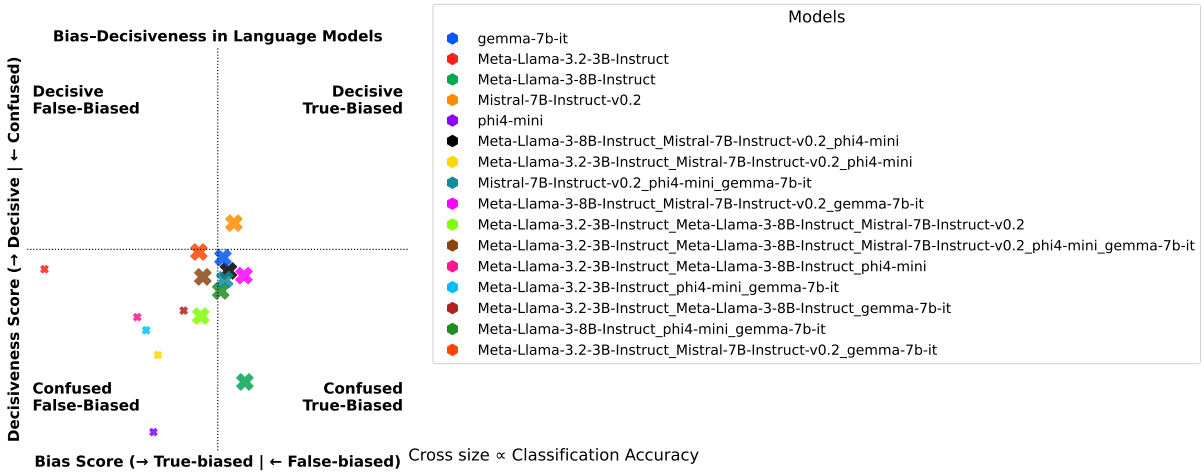


Figure 3: Bias-Decisiveness distribution of individual language models and their ensembles. The x-axis denotes bias orientation (false-biased → true-biased), while the y-axis captures decisiveness (confused → decisive). Circle size indicates classification accuracy.

Sensitivity to Confidence Interval Variation. Statistical numerical claims emerge as the most sensitive to variations in confidence intervals. Other claim types, such as financial, demographic, and temporal claims, display more predictable and stable behavior across configurations. Compared to other categories, the statistical dataset shows a lower correlation coefficient between 38% and 68% confidence intervals for the increase in MES scores in the epistemic setting, with a Spearman correlation of 0.7355. In contrast, demographic, financial, temporal, and quantitative claims exhibit higher correlations of 0.9266, 0.9653, 0.9573, and 0.9206, respectively. The overall dataset shows low sensitivity to confidence interval variation, yielding a Spearman correlation of 0.9690.

Response Bias, Decisiveness, and Ensemble Synergies. Figure 3 presents models within a two-dimensional space defined by response bias (ranging from false-biased to true-biased) and decisiveness (ranging from confused to decisive). Marker size is scaled proportionally to classification accuracy, enabling a joint analysis of predictive performance and epistemic orientation.

Several individual models occupy regions characterized by strong bias or lower decisiveness, indicating asymmetric belief behavior despite similar accuracies. In contrast, ensemble configurations tend to cluster closer to neutral or mildly true-biased regions while exhibiting higher decisiveness. This suggests that averaging belief scores across models reduces extreme biases and stabilizes confidence. Architecturally aligned ensembles exhibit strong epistemic synergy. In particular, combining `Llama-3B` and `Llama-8B` consistently outperforms individual models across datasets and confidence intervals (Figure 2; Table 3), suggesting that shared representations may enable coherent belief aggregation. Overall, these results show that prompting strategy, architecture, and ensemble composition jointly shape how model belief is formed, expressed, and stabilized, highlighting the role of structured interactions between prompts and architectural priors under `PLAUSIEVE`.

7 Conclusion

We introduced `PLAUSIEVE`, a model-independent framework for assessing statistically meaningful differences between groups of tokens, and used it to study belief formation in large language models. We observed that Llama-based models exhibit stronger decisiveness and stability, largely independent of scale. Architecturally aligned ensembles further yield synergistic gains. Our findings highlight that interaction between prompting structure and architectural design acts as the primary driver of belief quality. `PLAUSIEVE` provides a scalable framework for analyzing internal belief dynamics.

Acknowledgment

Tanmoy Chakraborty acknowledges the support of the Anusandhan National Research Foundation (DST/INT/USA/NSFDST/Tanmoy/P2/2024, Google GCP Grant, and Rajiv Khemani Young Faculty Chair Professorship in Artificial Intelligence

Limitations

While our study provides a focused exploration using PLAUSIEVE across different models, settings and datasets several limitations must be acknowledged:

1. **Prompt-level Ablations.** We do not perform ablations on the number of shots in the prompting setup. This is intentional, as our primary objective is to examine the applicability of PLAUSIEVE across different questioning styles and diverse model combinations rather than to optimize prompt engineering parameters.

2. **Model Diversity.** Although the study spans five instruction-tuned LLMs from four distinct model families, evaluating a larger set of architectures and parameter scales would further strengthen the claims regarding the robustness and generalizability of PLAUSIEVE.

3. **Scope of Claim Types.** Our analysis focuses exclusively on numerical claims due to their prevalence and high factual sensitivity in real world misinformation. Extending the framework to non numerical claims could uncover additional patterns and offer a more holistic understanding of inference using PLAUSIEVE.

Ethics Statement

This work studies belief in language models for research purposes only and does not aim to model or infer human beliefs. The proposed framework analyzes token-level output distributions without using personal or sensitive user data. We acknowledge that belief estimation in language models may be misinterpreted as intentionality or consciousness, and therefore emphasize that our findings should be understood strictly as statistical properties of model behavior under prompting.

References

Norillah Abdullah, Jamilah Niazi, Garba Idris, and Norul Hidayah Binti Mamat. 2022. Socratic questioning: A philosophical approach in developing critical thinking skills. *Al Hikmah International Journal of Islamic Studies and Human Sciences*, 5:143–161.

Filipe Altoe, Sérgio Miguel Gonçalves Pinto, and H Sofia Pinto. 2025. Explainable automatic fact-checking for journalists augmentation in the wild. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 10262–10270. International Joint Conferences on Artificial Intelligence Organization. Human-Centred AI.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. 2025. Emergent abilities in large language models: A survey. *Preprint*, arXiv:2503.05788.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Lennart Bürger, Fred A Hamprecht, and Boaz Nadler. 2024. Truth is universal: Robust detection of lies in llms. *Advances in Neural Information Processing Systems*, 37:138393–138431.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candance Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025. Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data. In *Proceedings of the First BabyLM Workshop*, pages 399–420, Suzhou, China. Association for Computational Linguistics.

John Cherian, Isaac Gibbs, and Emmanuel Candes. 2024. Large language model validity via enhanced conformal prediction methods. In *The Thirty-eighth*

Annual Conference on Neural Information Processing Systems.

Michael H. Connors and Peter W. Halligan. 2015. A cognitive account of belief: a tentative road map. *Frontiers in Psychology*, Volume 5 - 2014.

Phan Minh Dung. 1993. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning and logic programming. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI'93)*, pages 852–859, Chambéry, France.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Constanza Fierro, Ruchira Dhar, Filippos Stamatiou, Nicolas Garneau, and Anders Søgaard. 2024. Defining knowledge: Bridging epistemology and large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16096–16111, Miami, Florida, USA. Association for Computational Linguistics.

Yuling Gu, Oyvind Tafjord, and Peter Clark. 2024. Digital socrates: Evaluating LLMs through explanation critiques. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5559–5586, Bangkok, Thailand. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *Preprint*, arXiv:1706.04599.

Hangfeng He, Hongming Zhang, and Dan Roth. 2024. SocREval: Large language models with the socratic method for reference-free reasoning evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2736–2764, Mexico City, Mexico. Association for Computational Linguistics.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chenhui Hu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Knowledge in superposition: Unveiling the failures of lifelong knowledge editing for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24086–24094.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895, Vienna, Austria. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. *Preprint*, arXiv:2011.03870.

Yoav Levine, Noam Wies, Daniel Jannai, Dan Navon, Yedid Hoshen, and Amnon Shashua. 2021. The inductive bias of in-context learning: Rethinking pretraining example design. *arXiv preprint arXiv:2110.04541*.

Zichao Li, Yanshuai Cao, and Jackie CK Cheung. 2024. Do LLMs build world representations? probing through the lens of state abstraction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Husniah Sahamid. 2016. Developing critical thinking through socratic questioning: An action research study. *International Journal of Education and Literacy Studies*, 4:62–72.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *Preprint*, arXiv:1809.01286.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Xiaohan Wang, Shengyu Mao, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, Huajun Chen, and Ningyu Zhang. 2024. Editing conceptual knowledge for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 706–724, Miami, Florida, USA. Association for Computational Linguistics.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

Zhaomin Wu, Mingzhe Du, See-Kiong Ng, and Bingsheng He. 2026. Beyond prompt-induced lies: Investigating LLM deception on benign prompts. In *The Fourteenth International Conference on Learning Representations*.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.

Boyang Xue, Fei Mi, Qi Zhu, Hongru Wang, Rui Wang, Sheng Wang, Erxin Yu, Xuming Hu, and Kam-Fai Wong. 2025. UAlign: Leveraging uncertainty estimations for factuality alignment on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6002–6024, Vienna, Austria. Association for Computational Linguistics.

A Dataset

Our focus is on five categories of numerical claims: *Statistical*, *Demographic*, *Temporal*, *Financial*, and *Quantitative*. These categories are based on contextual cues surrounding numerical quantities in claims, enabling us to isolate variations arising from data sources rather than methodological factors. The details of these claim-centric datasets are presented in Table 4. In parallel, we apply regular expression heuristics to classify numerical claims into distinct semantic types—such as statistical, demographic, temporal, financial, and quantitative—based on surface cues (Table 5). This classification facilitates targeted evaluation of model performance on different numerical phenomena.

B Methodology

To evaluate belief judgments of language models, we employ two distinct prompting strategies: a direct prompting template and an epistemological prompting template. The direct template asks models to assess claim veracity and provide confidence scores. In contrast, the epistemological template probes six dimensions grounded in scientific evidence, expert consensus, and logical consistency. Each epistemological prompt pairs a claim with a specific question to elicit a more reflective response. These templates could enable a nuanced analysis of belief scoring across factual, controversial, and ambiguous claims. Dung's abstract argumentation is used to calculate final belief scores for epistemological setting.

Numerical Claim Type	Description	Label	Item Counts
Statistical	Identifies claims containing percentage values (e.g., "25%", "74.5 percent") or words like "percentage", which typically express proportions or population statistics.	True False Confused	2251 1669 1119
Demographic	Detects references to human population measures, such as "residents", "citizens", "households", or "demographics", focusing on community or group-based quantities.	True False Confused	2363 1849 626
Temporal	Captures temporal phrases such as "since 1995", "in the last five years", "currently", or ranges like "between 2000 and 2010", indicating time-bound factual claims.	True False Confused	2912 1860 837
Financial	Extracts monetary terms including currency symbols (e.g., "$", "€"), currency names ("USD", "rupees", "yen") and scale words like "million", "billion", "trillion".	True False Confused	2463 2190 1712
Quantitative	Targets units of physical measurement such as "km", "kg", "liters", "hours", "tons", etc., often used to quantify physical properties or distances.	True False Confused	1168 777 155

Table 4: Dataset statistics across numerical claim types, their regex-based identification, and distribution across label categories.

Numerical Claim Type	Regex Pattern Elements
Statistical	• Numbers followed by a percentage symbol (e.g., `75%`) • Words like `percent` or `percentage`
Demographic	• Terms indicating population: `population`, `residents`, `citizens`, `households`, `inhabitants`, `people`, `demographics`
Temporal	• Temporal adverbs: `today`, `yesterday`, `tomorrow`, `currently`, `now` • Durational phrases: `in/for X years/months/weeks/days` • Historical markers: `since YYYY`, `between YYYY and YYYY`, `20th century` • Relative time references: `last/next/over X years`
Financial	• Currency amounts: `$1,000`, `USD`, etc. • Currency types: `dollars`, `euros`, `pounds`, `yen`, `rupees` • Magnitude descriptors: `million`, `billion`, `trillion`
Quantitative	• Numeric values followed by measurement units such as: `km`, `kilometers`, `miles`, `kg`, `liters`, `hours`, `minutes`, `tons`, etc.

Table 5: Compact breakdown of regex patterns used for extracting different types of numerical claims. Each item represents a matching condition implemented in the corresponding regex.

B.1 Direct Prompt Template

The model is explicitly asked to assess the belief of the mode in the claim. Few shot examples are provided with annotated truth labels and confidence scores. Belief is inferred from the softmax-normalized logits corresponding to predefined `True` and `False` token sets, followed by certainty estimation through `PLAUSIEVE`.

The direct prompting template $\mathcal{P}_{\text{direct}}(c)$ for a given claim c is as follows:

```
Instruction 0: Do you believe if the claim
is true or false.
Instruction 1: Provide the answer and it's
corresponding confidence separated by a
```

```
comma ','.
Input claim: Elon Musk is the owner of
Tesla.
Answer: True, 0.855 [END]
Input claim: Donald Trump is the owner of
Tesla.
Answer: False, 0.855 [END]
Input claim: <c>
Answer:
```

B.2 Epistemological Prompt Template

Instead of asking directly whether the claim is true, we probe the model along six epistemic dimensions (Section 2). Each question evaluates a specific aspect such as scientific evidence, expert consensus, testability, logical consistency, and potential bias. For each question $q_i \in \mathcal{Q}$, we compute a belief score using PLAUSIEVE. The final model belief is then computed using Dung's abstract argumentation framework by balancing support and attack responses, as defined in Section 2. The model belief is given in format of $\{0, 0.5, 1\}$ as discussed in Section 2.

Each epistemological prompt $\mathcal{P}_{\text{epist}}(c, q_i)$ consists of an instruction, two illustrative examples, and the target question-claim pair. The template is shown below:

```
Instruction: Read the claim and answer the
question strictly with either 'True' or
'False'.
Use evidence, logic, and consistency to
decide.

Example 1:
Question: Is there scientific evidence that
supports the claim?
Claim: The Earth revolves around the Sun.
Answer: True, 0.855 [END]

Example 2:
Question: Does the claim contradict known
facts or scientific consensus?
Claim: Smoking tobacco has no adverse health
effects.
Answer: False, 0.855 [END]

Now your turn:
Question: <q_i>
Claim: <c>
Answer:
```

C Results and Analysis

This section presents additional evaluation results using the Model Effectiveness Score (MES) framework. We include scatterplots across claim types and prompting settings, along with marginal contributions of individual models.

C.1 Ensemble Comparison Tables

A complete set of results for all 31 ensembles is compiled for the overall category as well as for each of the five types of numerical claims, namely statistical, quantitative, demographical, temporal, and financial. To facilitate a systematic comparison, four experimental settings are evaluated: epistemological prompting at 68% confidence, epistemological prompting at 38% confidence, direct prompting at 68% confidence, and direct prompting at 38% confidence. These four settings enable two key axes of comparison: first, comparing epistemological prompting with direct prompting at both 68% and 38% confidence levels; and second, comparing the effect of confidence variation (68% versus 38%) within each prompting strategy, that is, within epistemological prompting and within direct prompting. The results obtained across all these configurations are comprehensively summarised in Tables 6–12.

C.2 MES Scatterplots

Across the five numerical claim types—statistical, quantitative, demographical, temporal, and financial, the four scatterplots for each category exhibit consistent patterns regarding the relative benefits of epistemological prompting and the robustness of the ensembles across confidence levels. For the *Epistemological Gain (38% confidence)* plots, most ensembles lie below the diagonal line, indicating that epistemological prompting yields higher MES than direct prompting at the same confidence threshold. This effect is particularly pronounced for ensemble combinations involving the Llama family of models, especially Llama-8B, which cluster in the bottom right region. This pattern suggests that epistemological cues help these models make more calibrated judgments even at lower confidence settings. In contrast, smaller models such as Phi or single-model ensembles occasionally fall above the diagonal, revealing that epistemological questioning does not uniformly benefit all architectures.

In the *Epistemological Confidence Robustness* plots, where epistemological MES at 68% confidence is compared against 38%, most ensemble

Model Combination	E68_D68	E38_D38	E38_E68	D38_D68
Gemma-7B	-94.535	-94.505	0.530	-0.033
Gemma-7B + Phi4-mini	-94.092	-92.575	18.509	-5.705
Phi4-mini	-87.309	-89.508	18.509	43.346
Llama-3B + Gemma-7B + Phi4-mini	-19.134	-28.333	-10.232	1.291
Mistral-7B + Gemma-7B + Phi4-mini	-17.417	-18.163	0.152	1.065
Mistral-7B + Phi4-mini	-16.531	-12.456	0.152	-4.509
Mistral-7B	-9.636	-8.616	-0.136	-1.250
Mistral-7B + Gemma-7B	-8.113	-10.950	-0.136	3.045
Llama-8B + Mistral-7B + Gemma-7B + Phi4-mini	-3.616	-4.691	-0.281	0.844
Llama-8B + Mistral-7B + Phi4-mini	-2.610	-2.911	-0.281	0.028
Llama-8B + Mistral-7B + Gemma-7B	-1.407	-5.813	-0.414	4.245
Llama-8B + Mistral-7B	-1.095	-2.729	-0.414	1.259
Llama-8B + Gemma-7B + Phi4-mini	-0.197	1.001	4.686	3.445
Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	0.256	4.530	0.628	-3.487
Llama-3B + Mistral-7B + Gemma-7B	2.704	-0.056	0.369	3.141
Llama-8B + Gemma-7B	3.755	2.791	4.529	5.508
Llama-3B + Gemma-7B	4.013	-5.691	-13.311	-4.391
Llama-3B + Phi4-mini	6.816	10.046	-10.232	-12.866
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	9.466	5.202	0.375	4.444
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B	11.567	4.699	0.211	6.785
Llama-8B + Llama-3B + Mistral-7B + Phi4-mini	16.812	10.707	0.375	5.911
Llama-8B + Llama-3B + Gemma-7B + Phi4-mini	19.070	17.553	4.760	6.111
Llama-8B + Phi4-mini	20.355	11.470	4.686	13.031
Llama-8B + Llama-3B + Mistral-7B	20.588	6.548	0.211	13.416
Llama-3B + Mistral-7B + Phi4-mini	23.375	22.699	0.628	1.182
Llama-8B + Llama-3B + Gemma-7B	26.250	14.162	4.637	15.717
Llama-8B	35.442	11.435	4.529	27.048
Llama-3B	43.526	22.648	-13.369	1.377
Llama-8B + Llama-3B + Phi4-mini	44.120	39.113	4.760	8.530
Llama-3B + Mistral-7B	73.247	97.313	0.369	-11.873
Llama-8B + Llama-3B	81.500	88.169	4.637	0.928

Table 6: Percentage change in MES scores across four settings combining Direct (D) and Epistemological (E) questioning, evaluated at two confidence intervals (68% and 38%) for overall numerical claims.

Model Combination	E68_D68	E38_D38	E38_E68	D38_D68
Gemma-7B	-97.923	-97.880	2.086	0.002
Gemma-7B + Phi4-mini	-97.531	-96.539	28.116	-8.618
Phi4-mini	-93.569	-94.720	28.116	56.051
Llama-3B + Gemma-7B + Phi4-mini	-25.308	-33.456	-7.831	3.455
Mistral-7B + Gemma-7B + Phi4-mini	-19.144	-21.572	0.341	3.447
Mistral-7B	-19.862	-19.107	-0.230	-1.162
Mistral-7B + Phi4-mini	-19.591	-11.976	0.341	-8.340
Llama-3B + Gemma-7B	-13.845	-11.764	-10.923	-13.024
Llama-3B + Mistral-7B + Gemma-7B	-13.872	-10.280	3.758	-0.396
Llama-8B + Mistral-7B + Gemma-7B	-5.600	-9.782	-0.638	3.968
Mistral-7B + Gemma-7B	-9.466	-9.293	-0.230	-0.421
Llama-8B + Mistral-7B + Gemma-7B + Phi4-mini	-6.579	-8.275	-0.331	1.513
Llama-8B + Mistral-7B	-7.058	-5.695	-0.638	-2.074
Llama-3B	15.691	-4.781	-11.066	8.055
Llama-8B + Mistral-7B + Phi4-mini	-4.809	-4.517	-0.331	-0.636
Llama-3B + Phi4-mini	5.660	-4.279	-7.831	1.739
Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	-9.898	-0.971	4.442	-4.973
Llama-8B + Gemma-7B + Phi4-mini	-5.587	-0.884	5.353	0.354
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B	-0.348	0.360	-0.381	-1.084
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	1.887	0.971	0.084	0.992
Llama-8B + Gemma-7B	-6.888	1.394	4.904	-3.665
Llama-8B + Llama-3B + Gemma-7B	1.295	1.399	5.783	5.674
Llama-8B + Llama-3B + Mistral-7B	11.645	2.121	-0.381	8.910
Llama-8B + Llama-3B + Gemma-7B + Phi4-mini	0.618	7.241	6.354	-0.213
Llama-8B + Llama-3B + Mistral-7B + Phi4-mini	11.317	7.906	0.084	3.248
Llama-8B	42.663	10.602	4.904	35.313
Llama-8B + Phi4-mini	19.445	11.512	5.353	12.848
Llama-3B + Mistral-7B + Phi4-mini	23.024	20.909	4.442	6.269
Llama-8B + Llama-3B + Phi4-mini	33.056	31.415	6.354	7.682
Llama-8B + Llama-3B	46.899	69.101	5.783	-8.106
Llama-3B + Mistral-7B	49.360	77.176	3.758	-12.532

Table 7: Percentage change in MES scores across four settings combining Direct (D) and Epistemological (E) questioning, evaluated at two confidence intervals (68% and 38%) for demographical numerical claims.

Model Combination	E68_D68	E38_D38	E38_E68	D38_D68
Gemma-7B	-91.100	-91.080	0.000	-0.220
Gemma-7B + Phi4-mini	-90.877	-88.752	15.501	-6.316
Phi4-mini	-79.773	-83.429	15.501	40.986
Llama-3B + Gemma-7B + Phi4-mini	-28.649	-31.507	-11.977	-8.304
Mistral-7B + Gemma-7B + Phi4-mini	-22.825	-23.715	0.212	1.381
Mistral-7B + Phi4-mini	-21.264	-18.889	0.212	-2.722
Llama-3B + Phi4-mini	-19.623	7.865	-11.977	-34.409
Mistral-7B	-12.229	-11.098	-0.137	-1.407
Mistral-7B + Gemma-7B	-6.803	-20.808	-0.137	17.523
Llama-3B + Mistral-7B + Phi4-mini	-4.631	14.690	-2.224	-18.695
Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	-3.825	-1.951	-2.224	-4.093
Llama-8B + Mistral-7B + Phi4-mini	-1.818	-4.889	1.171	4.437
Llama-8B + Mistral-7B + Gemma-7B + Phi4-mini	-1.078	-6.139	1.171	6.626
Llama-8B + Gemma-7B + Phi4-mini	20.504	0.883	3.371	23.476
Llama-3B + Gemma-7B	24.597	-21.314	-15.501	33.800
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	29.731	3.484	1.532	27.285
Llama-3B + Mistral-7B + Gemma-7B	31.515	-5.685	-2.356	36.157
Llama-8B	33.340	10.989	3.368	24.184
Llama-8B + Mistral-7B + Gemma-7B	34.979	-7.517	1.136	47.609
Llama-8B + Llama-3B + Mistral-7B + Phi4-mini	35.145	8.631	1.532	26.314
Llama-8B + Phi4-mini	36.205	10.995	3.371	26.849
Llama-3B	42.999	19.177	-15.501	1.389
Llama-3B + Mistral-7B	43.457	50.579	-2.356	-6.975
Llama-8B + Mistral-7B	47.263	-5.483	1.136	57.576
Llama-8B + Llama-3B + Gemma-7B + Phi4-mini	50.787	15.961	4.734	36.189
Llama-8B + Llama-3B + Phi4-mini	68.785	35.936	4.734	30.043
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B	69.831	0.337	1.452	71.717
Llama-8B + Llama-3B + Mistral-7B	73.295	3.511	1.452	69.849
Llama-8B + Gemma-7B	113.614	-0.739	3.368	122.452
Llama-8B + Llama-3B + Gemma-7B	121.760	11.120	4.811	109.170
Llama-8B + Llama-3B	145.376	63.669	4.811	57.135

Table 8: Percentage change in MES scores across four settings combining Direct (D) and Epistemological (E) questioning, evaluated at two confidence intervals (68% and 38%) for statistical numerical claims.

Model Combination	E68_D68	E38_D38	E38_E68	D38_D68
Gemma-7B	-85.377	-85.301	0.464	-0.057
Gemma-7B + Phi4-mini	-85.373	-83.615	11.732	-0.254
Phi4-mini	-74.644	-78.579	11.732	32.257
Mistral-7B + Gemma-7B + Phi4-mini	-18.733	-17.853	0.012	-1.060
Mistral-7B + Gemma-7B	-16.868	-16.724	-0.040	-0.213
Llama-3B + Gemma-7B + Phi4-mini	-16.735	-26.177	-10.324	1.146
Mistral-7B + Phi4-mini	-10.852	-12.893	0.012	2.356
Llama-3B + Gemma-7B	-10.795	-16.900	-13.178	-6.800
Llama-8B + Mistral-7B + Gemma-7B	-7.298	-4.258	0.200	-2.982
Llama-8B + Mistral-7B + Gemma-7B + Phi4-mini	-5.480	-2.544	0.323	-2.700
Llama-8B + Mistral-7B + Phi4-mini	-1.280	-1.739	0.323	0.791
Llama-8B + Gemma-7B	-1.210	2.726	3.596	-0.374
Llama-8B + Gemma-7B + Phi4-mini	-0.814	2.902	3.762	0.015
Llama-8B + Mistral-7B	0.838	-2.376	0.200	3.498
Mistral-7B	3.627	5.273	-0.040	-1.603
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B	5.556	3.904	-0.130	1.458
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	5.972	5.417	0.009	0.536
Llama-8B + Llama-3B + Mistral-7B	6.657	3.736	-0.130	2.683
Llama-8B + Llama-3B + Mistral-7B + Phi4-mini	8.042	9.049	0.009	-0.914
Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	10.038	8.812	-1.108	0.006
Llama-3B + Mistral-7B + Gemma-7B	12.575	12.271	-1.171	-0.904
Llama-8B + Phi4-mini	12.927	10.362	3.762	6.174
Llama-8B	21.029	8.676	3.596	15.371
Llama-8B + Llama-3B + Gemma-7B	21.132	23.951	3.934	1.570
Llama-8B + Llama-3B + Gemma-7B + Phi4-mini	22.548	25.667	4.039	1.457
Llama-3B + Phi4-mini	24.448	4.979	-10.324	6.308
Llama-3B + Mistral-7B + Phi4-mini	25.075	17.306	-1.108	5.442
Llama-8B + Llama-3B + Phi4-mini	44.092	34.684	4.039	11.307
Llama-3B + Mistral-7B	48.458	59.285	-1.171	-7.889
Llama-3B	55.473	39.647	-13.293	-3.467
Llama-8B + Llama-3B	64.880	72.695	3.934	-0.770

Table 9: Percentage change in MES scores across four settings combining Direct (D) and Epistemological (E) questioning, evaluated at two confidence intervals (68% and 38%) for financial numerical claims.

Model Combination	E68_D68	E38_D38	E38_E68	D38_D68
Gemma-7B	-99.491	-99.494	0.000	0.510
Gemma-7B + Phi4-mini	-99.306	-98.698	72.344	-8.177
Phi4-mini	-98.174	-98.134	72.344	68.601
Mistral-7B	-26.459	-25.972	-0.085	-0.742
Mistral-7B + Gemma-7B + Phi4-mini	-21.725	-24.904	0.904	5.176
Mistral-7B + Phi4-mini	-20.515	-10.951	0.904	-9.934
Llama-3B + Mistral-7B + Gemma-7B	-13.245	-13.851	-0.460	0.239
Llama-8B + Mistral-7B + Gemma-7B + Phi4-mini	-11.849	-12.485	-0.760	-0.038
Llama-8B + Mistral-7B	-10.713	-5.360	-1.201	-6.790
Llama-8B + Gemma-7B + Phi4-mini	-9.968	-4.414	9.613	3.244
Llama-8B + Mistral-7B + Phi4-mini	-9.824	-11.287	-0.760	0.877
Llama-8B + Mistral-7B + Gemma-7B	-8.222	-13.875	-1.201	5.285
Llama-8B + Gemma-7B	-7.733	7.920	8.798	-6.982
Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	-6.887	-0.781	0.357	-5.819
Mistral-7B + Gemma-7B	-6.244	-5.911	-0.085	-0.439
Llama-3B + Gemma-7B + Phi4-mini	-5.500	-22.852	-6.430	14.616
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	0.469	-2.809	-1.131	2.204
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B	2.017	4.742	-1.516	-4.077
Llama-8B + Llama-3B + Gemma-7B + Phi4-mini	6.371	12.322	7.856	2.141
Llama-8B + Llama-3B + Gemma-7B	12.422	5.663	7.056	13.904
Llama-8B + Llama-3B + Mistral-7B + Phi4-mini	12.544	7.898	-1.131	3.126
Llama-8B + Phi4-mini	15.590	14.464	9.613	10.692
Llama-8B + Llama-3B + Mistral-7B	19.812	1.909	-1.516	15.785
Llama-3B + Phi4-mini	24.893	16.683	-6.430	0.153
Llama-3B + Gemma-7B	39.750	40.650	-11.756	-12.320
Llama-3B	41.278	14.310	-11.756	9.062
Llama-8B + Llama-3B + Phi4-mini	44.561	35.373	7.856	15.175
Llama-8B	45.526	10.381	8.798	43.439
Llama-3B + Mistral-7B + Phi4-mini	51.584	25.502	0.357	21.214
Llama-8B + Llama-3B	86.809	144.223	7.056	-18.111
Llama-3B + Mistral-7B	156.556	202.666	-0.460	-15.625

Table 10: Percentage change in MES scores across four settings combining Direct (D) and Epistemological (E) questioning, evaluated at two confidence intervals (68% and 38%) for quantitative numerical claims.

Model Combination	E68_D68	E38_D38	E38_E68	D38_D68
Llama-3B + Gemma-7B	41.002	33.933	-15.182	-10.706
Llama-3B	83.507	55.603	-15.182	0.028
Llama-3B + Gemma-7B + Phi4-mini	-4.593	-23.176	-12.855	8.224
Llama-3B + Phi4-mini	58.806	40.291	-12.855	-1.354
Llama-8B + Mistral-7B	-4.779	-3.052	-1.548	-3.302
Llama-8B + Mistral-7B + Gemma-7B	-4.117	-7.092	-1.548	1.604
Llama-8B + Mistral-7B + Gemma-7B + Phi4-mini	-4.568	-6.722	-1.513	0.761
Llama-8B + Mistral-7B + Phi4-mini	-3.473	-5.128	-1.513	0.205
Mistral-7B	-14.397	-13.849	-0.214	-0.849
Mistral-7B + Gemma-7B	-6.550	-6.603	-0.214	-0.158
Mistral-7B + Phi4-mini	-16.513	-11.636	-0.158	-5.668
Mistral-7B + Gemma-7B + Phi4-mini	-14.574	-17.489	-0.158	3.371
Gemma-7B	-96.719	-96.722	0.000	0.080
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B	6.052	8.163	0.232	-1.725
Llama-8B + Llama-3B + Mistral-7B	29.677	12.351	0.232	15.688
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	6.464	6.417	0.239	0.283
Llama-8B + Llama-3B + Mistral-7B + Phi4-mini	17.401	12.492	0.239	4.613
Llama-3B + Mistral-7B + Gemma-7B	-0.290	0.477	0.523	-0.244
Llama-3B + Mistral-7B	172.359	218.687	0.523	-14.090
Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	3.259	7.515	0.622	-3.360
Llama-3B + Mistral-7B + Phi4-mini	45.386	26.338	0.622	15.793
Llama-8B + Llama-3B + Gemma-7B + Phi4-mini	19.979	20.536	4.028	3.547
Llama-8B + Llama-3B + Phi4-mini	63.464	51.068	4.028	12.564
Llama-8B + Llama-3B + Gemma-7B	28.266	22.493	4.226	9.137
Llama-8B + Llama-3B	109.775	145.066	4.226	-10.783
Llama-8B + Gemma-7B + Phi4-mini	-6.719	-1.146	6.845	0.821
Llama-8B + Phi4-mini	14.063	9.337	6.845	11.463
Llama-8B + Gemma-7B	-6.455	3.182	6.982	-3.010
Llama-8B	39.567	12.827	6.982	32.336
Gemma-7B + Phi4-mini	-96.216	-95.345	18.161	-3.943
Phi4-mini	-92.427	-93.780	18.161	43.871

Table 11: Percentage change in MES scores across four settings combining Direct (D) and Epistemological (E) questioning, evaluated at two confidence intervals (68% and 38%) for temporal numerical claims.

Model Combination	E68	E38	D68	D38
Llama-3B + Gemma-7B	0.27316	0.2368	0.26262	0.25109
Llama-3B	0.27316	0.23664	0.19032	0.19294
Llama-3B + Gemma-7B + Phi4-mini	0.27458	0.24648	0.33955	0.34393
Llama-3B + Phi4-mini	0.27458	0.24648	0.25705	0.22398
Llama-8B + Mistral-7B	0.47424	0.47228	0.47949	0.48553
Llama-8B + Mistral-7B + Gemma-7B	0.47424	0.47228	0.48101	0.50143
Llama-8B + Mistral-7B + Gemma-7B + Phi4-mini	0.47442	0.47308	0.49222	0.49637
Llama-8B + Mistral-7B + Phi4-mini	0.47442	0.47308	0.48713	0.48727
Mistral-7B	0.40251	0.40196	0.44543	0.43986
Mistral-7B + Gemma-7B	0.40251	0.40196	0.43805	0.45139
Mistral-7B + Phi4-mini	0.40277	0.40339	0.48255	0.46079
Mistral-7B + Gemma-7B + Phi4-mini	0.40277	0.40339	0.48772	0.49292
Gemma-7B	0.02561	0.02574	0.46861	0.46845
Llama-8B + Llama-3B + Mistral-7B + Gemma-7B	0.49545	0.4965	0.44409	0.47422
Llama-8B + Llama-3B + Mistral-7B	0.49545	0.4965	0.41086	0.46599
Llama-8B+Llama-3B+Mistral-7B+Gemma-7B+Phi4-mini	0.49557	0.49743	0.45271	0.47283
Llama-8B + Llama-3B + Mistral-7B + Phi4-mini	0.49557	0.49743	0.42424	0.44932
Llama-3B + Mistral-7B + Gemma-7B	0.45104	0.4527	0.43916	0.45296
Llama-3B + Mistral-7B	0.45104	0.4527	0.26034	0.22943
Llama-3B + Mistral-7B + Gemma-7B + Phi4-mini	0.45121	0.45404	0.45006	0.43437
Llama-3B + Mistral-7B + Phi4-mini	0.45121	0.45404	0.36573	0.37005
Llama-8B + Llama-3B + Gemma-7B + Phi4-mini	0.46113	0.48308	0.38727	0.41094
Llama-8B + Llama-3B + Phi4-mini	0.46113	0.48308	0.31996	0.34726
Llama-8B + Llama-3B + Gemma-7B	0.46079	0.48216	0.36498	0.42234
Llama-8B + Llama-3B	0.46079	0.48216	0.25388	0.25624
Llama-8B + Gemma-7B + Phi4-mini	0.45652	0.47791	0.45742	0.47317
Llama-8B + Phi4-mini	0.45652	0.47791	0.37931	0.42873
Llama-8B + Gemma-7B	0.45615	0.4768	0.43964	0.46386
Llama-8B	0.45615	0.4768	0.33678	0.42788
Gemma-7B + Phi4-mini	0.02712	0.03215	0.45911	0.43292
Phi4-mini	0.02712	0.03215	0.21373	0.30637

Table 12: MES scores across four settings that is, Direct (D) and Epistemological (E) questioning, evaluated at two confidence intervals (68% and 38%) for overall numerical claims.

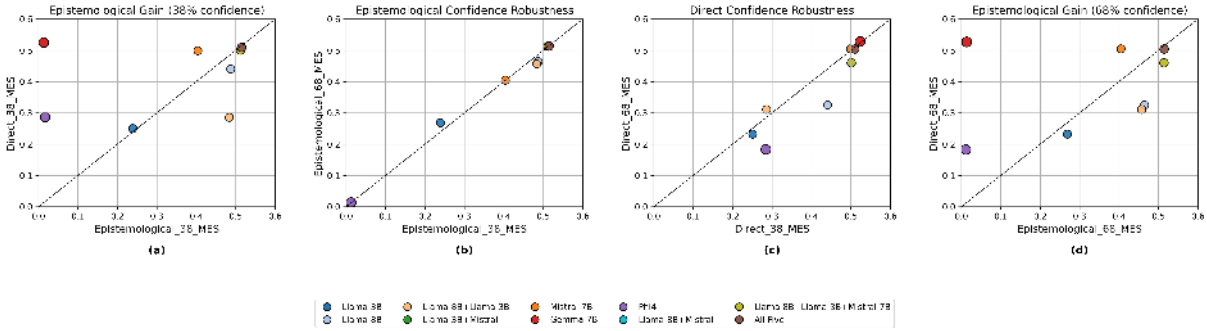


Figure 4: Demographic: MES scatter plots across four prompting and confidence settings. **Figure a** shows performance gains from epistemological prompting (38%); **Figures b and c** illustrate MES stability across confidence levels; **Figure d** shows performance gains from epistemological prompting (68%).

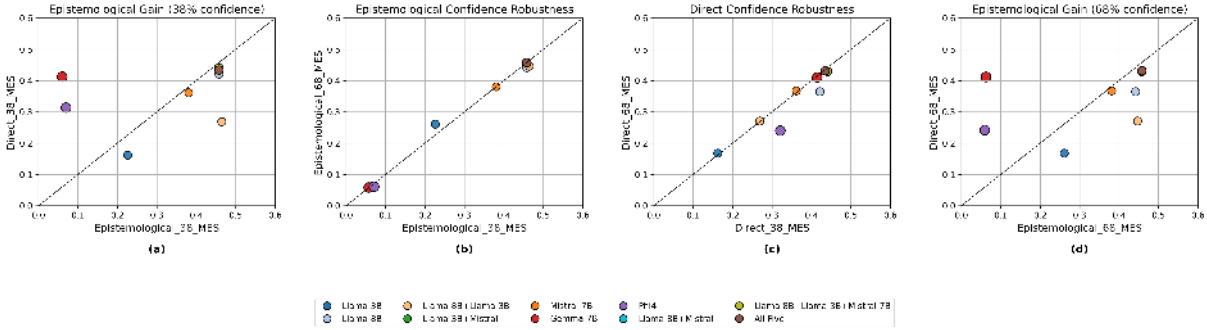


Figure 5: Financial: MES scatter plots across four prompting and confidence settings. **Figure a** shows performance gains from epistemological prompting (38%); **Figures b and c** illustrate MES stability across confidence levels; **Figure d** shows performance gains from epistemological prompting (68%).

points lie close to or above the diagonal. This indicates that the epistemological prompting strategy maintains or strengthens MES as confidence increases, implying a form of robustness to confidence scaling. Ensembles tend to show steady upward shifts, demonstrating that these models may internalize epistemic cues consistently across confidence levels. Overall Epistemological prompting shows high robustness to confidence levels across all types of numerical claims.

The *Direct Confidence Robustness* plots show more dispersed behavior. While several ensembles remain close to the diagonal (indicating similar performance at 38% and 68% confidence), a noticeable number fall below the line, particularly for quantitative and statistical claims. This suggests that direct prompting is more sensitive to confidence perturbations and may lead to inconsistent calibration as the threshold shifts. Larger ensembles generally show more stable trajectories, while smaller or single-model ensembles exhibit sharper declines.

Finally, in the *Epistemological Gain (68% confidence)* plots, epistemological prompting continues to outperform direct prompting for most ensembles across all types of numerical claims, with the majority of points falling below the diagonal. The gain is less pronounced for demographic claims, where several ensembles lie close to the diagonal, indicating that at higher confidence thresholds direct prompting remains competitive. Nonetheless, ensembles involving the Llama family of models consistently demonstrate improved epistemological gain, highlighting the advantage of certain model architectures in epistemological settings.

Overall, the figures collectively reveal three core insights: (i) epistemological prompting reliably improves MES over direct prompting for most ensembles and claim types; (ii) epistemological prompting is more robust to confidence level variation than direct prompting; and (iii) ensemble diversity amplifies both epistemological gain and robustness, especially for numerical claim types such as temporal, statistical, and quantitative.

Figures 4–8 present MES scatterplots for five numerical claim types under varying prompting and confidence settings.

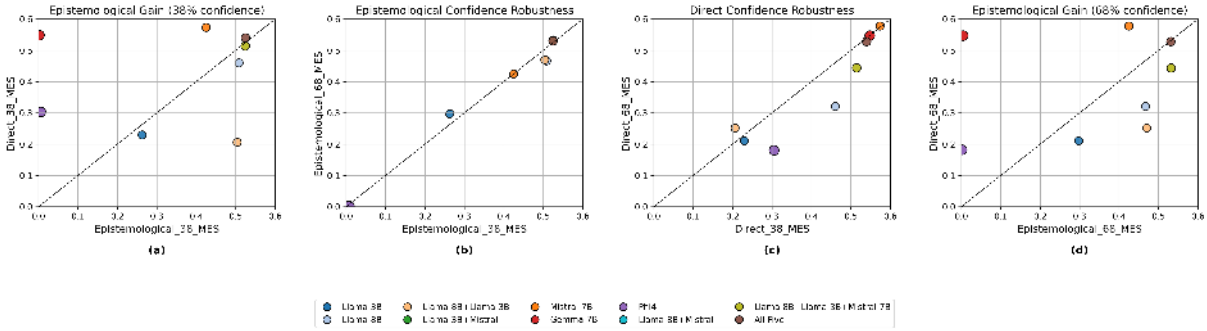


Figure 6: Quantitative: MES scatter plots across four prompting and confidence settings. **Figure a** shows performance gains from epistemological prompting (38%); **Figures b and c** illustrate MES stability across confidence levels; **Figure d** shows performance gains from epistemological prompting (68%).

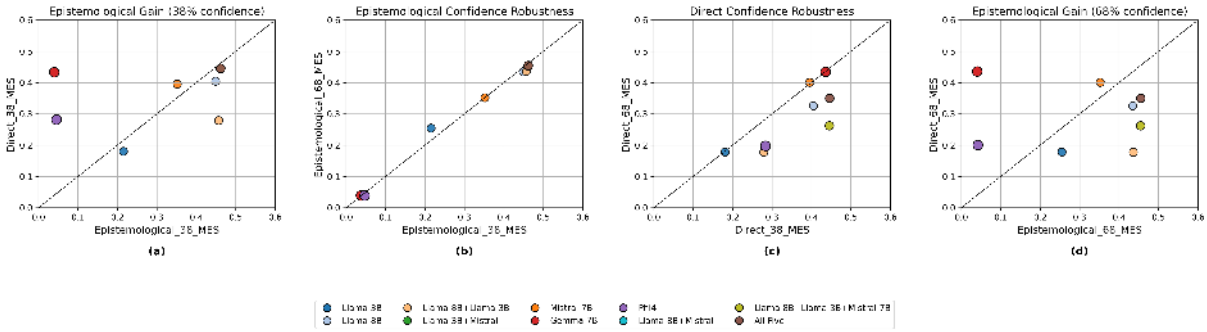


Figure 7: Statistical: MES scatter plots across four prompting and confidence settings. **Figure a** shows performance gains from epistemological prompting (38%); **Figures b and c** illustrate MES stability across confidence levels; **Figure d** shows performance gains from epistemological prompting (68%).

C.3 Marginal Model Effectiveness Score

Table 13 summarizes average marginal contributions for each model, indicating their standalone influence within ensemble settings. Meta-Llama-3-8B-Instruct consistently achieves the highest MES under epistemological prompting. In contrast, Mistral-7B exhibits strong performance under both direct and epistemological prompting schemes. These findings underscore the value of epistemological prompting and the importance of confidence-aware evaluation of belief signals.

C.4 Sensitivity analysis of hyperparameter γ

To assess the stability of the Modified Evaluation Score (MES) with respect to the penalization hyperparameter δ, we examine how pairwise percentage changes in MES evolve across $\delta \in \{0.25, 0.5, \ldots, 3.0\}$ for each standalone model. As shown in Figure 13, models with inherently low JS divergence — namely `Mistral-7B-Instruct` and `Meta-Llama-3-8B-Instruct` — remain largely invariant to δ, with all four metric-pair curves hovering near zero regardless of the prompting or confidence regime. This confirms that for well-calibrated models, MES is effectively equivalent to raw accuracy and the choice of δ is inconsequential. Conversely, `gemma-7b-it` and `phi4-mini` exhibit strong negative sensitivity in the E68→D68 and E38→D38 directions, with percentage changes becoming increasingly pronounced as δ grows, reflecting the compounding penalty applied to their high JS divergence under standard prompting. `Meta-Llama-3.2-3B-Instruct` occupies an intermediate regime, showing moderate but consistent downward drift across sampling configurations. These findings suggest that δ functions as an epistemic calibration lever: for models prone to overconfident or poorly distributed outputs, larger values of δ impose a substantially heavier MES penalty, making the metric increasingly discriminative. Based on this sensitivity analysis, values in the range $\delta \in [1.25, 2.5]$ offer a balanced trade-off between rewarding accuracy and penalizing miscalibration, and are recommended for downstream ensemble evaluation.

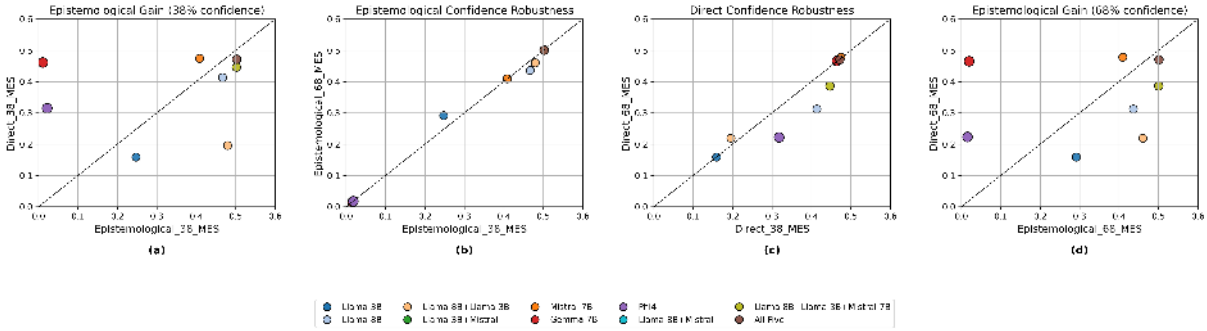


Figure 8: Temporal: MES scatter plots across four prompting and confidence settings. **Figure a** shows performance gains from epistemological prompting (38%); **Figures b and c** illustrate MES stability across confidence levels; **Figure d** shows performance gains from epistemological prompting (68%).

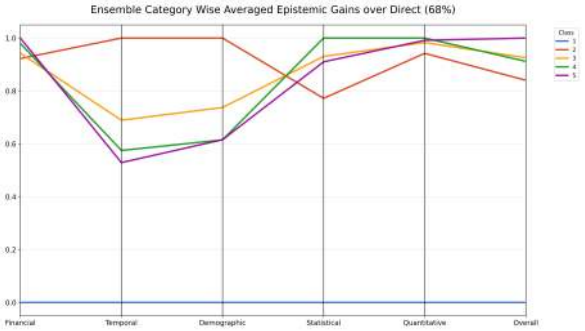


Figure 9: Ensemble Category Wise Averaged and Normalized Epistemic Gains over Direct at 68% Confidence. Each line represents an ensemble class (1–5) across six numerical claim categories.

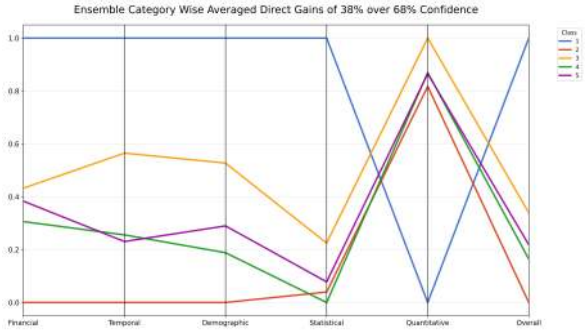


Figure 10: Ensemble Category-Wise Averaged and Normalized Direct Gains at 38% Confidence over 68% Confidence. Each line represents an ensemble class (1–5) across six numerical claim categories.

C.5 Ensemble Level Analysis of Category Wise Gains

To complement our model level evaluation, we present a fine-grained ensemble-level analysis of category-wise gains across numerical claim types. Each figure visualizes the averaged performance of ensembles of size 1 through 5 across six claim categories—**Financial**, **Temporal**, **Demographic**, **Statistical**, **Quantitative**, and **Overall** under different configurations of prompting and confidence thresholds. The y-axis represents normalized gains, and the x-axis denotes claim categories. We analyze four key dimensions: direct gains at 38% confidence, epistemic gains at 38% and 68% confidence, and epistemic over direct gains at both thresholds.

C.5.1 Direct Gains at 38% Confidence

Figure 10 reveals that Class 1 ensembles (single models) perform strongly in Financial, Temporal, Demographic, and Overall categories, but suffer a pronounced dip in Quantitative claims. Whereas all other classes shows a peak in Quantitative, suggesting specialized and unique effect of broadening the threshold. Class 2 remains near zero across almost all categories, indicating robustness to confidence change. As ensemble size increases (Classes 4 and 5), gains stabilize across categories. This suggests that ensemble aggregation may mitigate category specific anomalies, especially in low confidence direct settings.

C.5.2 Epistemic Gains at 38% Confidence

Figure 11 demonstrates that epistemological prompting substantially enhances performance across all ensemble sizes. Notably, Class 1 (two model ensembles) achieves maximal gains across all claim categories except Quantitative, while Class 4 peaks in Statistical and Quantitative. Financial, Temporal and Demographic claims show moderate sensitivity for multi model ensembles. These results affirm that single model ensembles are less robust than multi models in terms of confidence threshold.

Metric	Model	Average Marginal Contribution MES
Epistemological Prompting (68% confidence)	Meta-Llama-3-8B-Instruct	**0.1813**
	Mistral-7B-Instruct-v0.2	0.1727
	Meta-Llama-3.2-3B-Instruct	0.1527
	phi4-mini	0.1261
	Gemma-7B	0.1259
Direct Prompting (38% confidence)	Mistral-7B-Instruct-v0.2	**0.1709**
	Gemma-7B	0.1695
	Meta-Llama-3-8B-Instruct	0.1685
	phi4-mini	0.1538
	Meta-Llama-3.2-3B-Instruct	0.1257
Direct Prompting (68% confidence)	Mistral-7B-Instruct-v0.2	**0.1693**
	Gemma-7B	0.1662
	Meta-Llama-3-8B-Instruct	0.1551
	phi4-mini	0.1469
	Meta-Llama-3.2-3B-Instruct	0.1232
Epistemological Prompting (38% confidence)	Meta-Llama-3-8B-Instruct	**0.1863**
	Mistral-7B-Instruct-v0.2	0.1728
	Meta-Llama-3.2-3B-Instruct	0.1500
	phi4-mini	0.1272
	Gemma-7B	0.1263

Table 13: Average marginal contributions of individual models across MES dimensions, calculated at low (38%) and high (68%) confidence thresholds. Values reflect each model's standalone influence on ensemble performance.

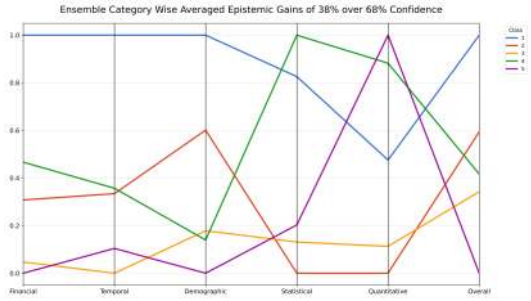


Figure 11: Ensemble Category Wise Averaged and Normalized Epistemic Gains at 38% Confidence over 68% Confidence. Each line represents an ensemble class (1–5) across six numerical claim categories.

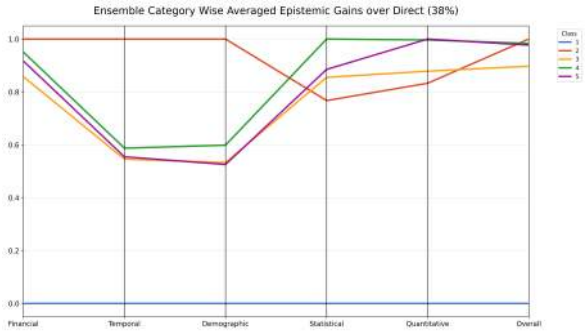


Figure 12: Ensemble Category Wise Averaged and Normalized Epistemic Gains over Direct at 38% Confidence. Each line represents an ensemble class (1–5) across six numerical claim categories.

C.5.3 Epistemic Gains over Direct (38% Confidence)

Figure 12 quantifies the epistemic over direct gain differential at 38% confidence. Class 1 again shows nearly zero average gain, while Class 2 exhibits sharp improvements in Financial, Temporal and Demographic claims. Class 4 demonstrates robust gains in Statistical and Overall categories, suggesting that epistemic prompting disproportionately benefits structurally aligned ensembles. Interestingly, Demographic and Temporal claims show minimal differential across almost all classes, indicating limited epistemic sensitivity. This reinforces our earlier finding that epistemological prompting probably yields super additive benefits when paired with architectural alignment.

C.5.4 Epistemic Gains over Direct (68% Confidence)

Figure 9 mirrors the 38% trend but under stricter confidence thresholds. Gains are more conservative but still favor multi model ensembles. Quantitative and Statistical claims show the highest epistemic over direct differential, especially for Class 3 and Class 5, indicating that epistemic structure may help resolve numerical ambiguity under high-certainty conditions. Class 1 remains flat, confirming that epistemic gains are ensemble dependent.

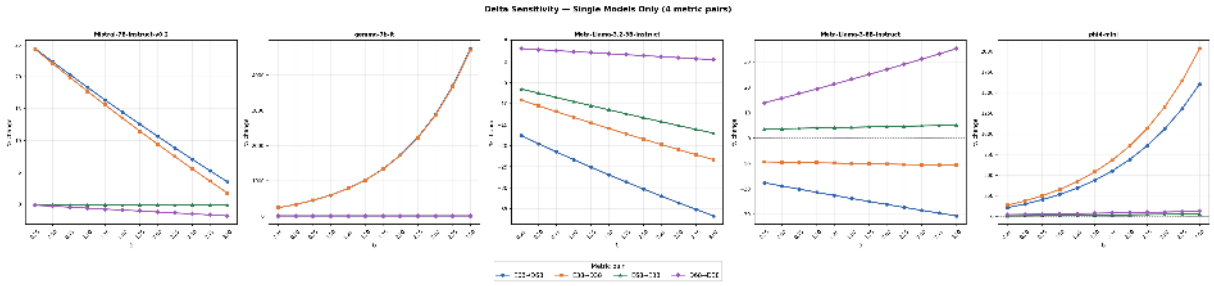


Figure 13: Delta sensitivity of MES percentage change across four metric pairs for single models. Each subplot tracks how the four pairwise percentage changes — **E68→D68**, **E38→D38**, **E68→E38**, and **D68→D38** — evolve as the hyperparameter δ increases from 0.25 to 3.0. `Mistral-7B` and `Meta-Llama-3-8B` exhibit stable, near-zero divergence across all pairs, indicating robustness to δ. In contrast, `gemma-7b` and `phi4-mini` show pronounced sensitivity, with large negative E68→D68 and E38→D38 shifts that amplify with increasing δ, reflecting their high JS divergence under standard prompting. `Meta-Llama-3.2-3B` displays moderate sensitivity with consistent downward trends across sampling regimes.

Overall, the 68% threshold filters out epistemic noise, revealing the true robustness of epistemological prompting.

C.5.5 Summary Insights

- **Class 1 (single models)** consistently underperform in epistemic settings, highlighting the need for ensemble diversity.
- **Class 2 ensembles** show peak gains in Financial, Temporal, and Demographic claims.
- **Class 3,4 and 5 ensembles** show peak gains in Statistical, Quantitative and Overall claims.
- **Quantitative claims** exhibit high sensitivity to prompting and confidence thresholds for multi model ensembles, reinforcing their complexity.
- **Epistemological prompting** yields substantial gains over direct prompting, especially under relaxed confidence (38%) , especially for multi model ensembles.

These findings validate our hypothesis that epistemic structure and ensemble composition jointly drive reliable plausibility-based veracity estimation. They also underscore the importance of claim-type-aware ensemble design in real-world misinformation detection systems.

D Validity of token sets

Our framework estimates belief by comparing the probability mass assigned to semantically aligned `True` and `False` token sets:

$$\mathcal{T} = \{\texttt{true}, \texttt{True}, \texttt{yes}, \texttt{Yes}\}$$

and

$$\mathcal{F} = \{\texttt{false}, \texttt{False}, \texttt{no}, \texttt{No}\}.$$

The use of token sets rather than individual tokens is important for robustness across models and prompting styles. Different instruction-tuned LLMs exhibit varying tokenization behaviors, capitalization preferences, and decoding priors. For example, some models consistently prefer lowercase forms (e.g., `true`), whereas others place higher probability on capitalized variants (e.g., `True`) depending on conversational formatting and alignment tuning. Similarly, affirmative and negative responses may manifest through either truth-valued tokens (`true`/`false`) or natural conversational responses (`yes`/`no`). Aggregating across semantically equivalent variants therefore reduces sensitivity to superficial lexical preferences while preserving the underlying epistemic signal.

To validate the selected token sets, we analyze the median vocabulary rank of each token across all datasets and experimental settings. A lower rank indicates that a token consistently receives higher probability under the model distribution. Figures 14–17 show that the chosen tokens form stable semantic clusters across models and datasets.

Several important observations emerge from these analyses. First, the primary affirmative and negative tokens consistently occupy high-ranking regions of the vocabulary distribution, indicating that the models naturally use these tokens as preferred realizations of binary judgments. Second, capitalization variants exhibit similar trends, suggesting that the token sets capture stylistic variation rather than distinct semantic behaviors. Third, the relative ordering of affirmative versus negative to-

kens remains stable across statistical, demographic, temporal, financial, and quantitative datasets, supporting the cross-domain generalizability of the token sets.

Importantly, these plots also reveal substantial inter-model differences in absolute token rank scales. For instance, Gemma-7B exhibits much lower preference for the `yes`/`no` variants compared to Llama and Phi models, while still maintaining strong consistency for `true`/`false`. This further motivates the use of aggregated token sets rather than relying on any single lexical realization. By pooling semantically equivalent responses, the framework becomes less sensitive to model-specific decoding biases and tokenizer artifacts.

Overall, the empirical stability of these token distributions across datasets and architectures supports the validity of $\mathcal{T}$ and $\mathcal{F}$ as generalized proxies for affirmative and negative epistemic judgments.

E Future scope and applications

This section outlines three concrete directions in which the ideas behind PLAUSIEVE can be extended and applied.

E.1 Enabling low-resource and developmentally-plausible settings (BabyLM)

A primary motivation for plausibility based inference is its suitability for scenarios with little or no task-specific evidence, such as the developmentally plausible, sample efficient settings advocated by the BabyLM challenge. In such low data regimes, traditional evidence retrieval heavy pipelines can be inaccurate or inapplicable; plausibility estimation leverages the internal knowledge and heuristics of pre-trained models to produce rapid belief judgments without external corpora. This aligns with recent BabyLM findings that indicate meaningful linguistic and reasoning capabilities can be elicited from models trained or evaluated under strongly data-limited conditions (Warstadt et al., 2023; Charpentier et al., 2025; Hu et al., 2024). Integrating PLAUSIEVE style uncertainty scoring into data efficient training or evaluation protocols could therefore provide a principled way to (a) measure what a compact model "knows" versus where it is uncertain, and (b) prioritize scarce supervision towards examples with high epistemic value (i.e., active sampling). Concretely, plausibility scores and their bootstrap confidence intervals can be used as acquisition criteria in sample-efficient pretraining or interactive teaching settings to focus annotation and compute budgets where they will most reduce model uncertainty (Wei et al., 2022; Xue et al., 2025).

E.2 Towards models pre-trained for plausibility inference

The notion of plausibility suggests a new pretraining objective or curriculum: beyond predicting next tokens, pretraining could explicitly optimize for calibrated belief judgments and uncertainty estimates that align with factuality or epistemic criteria. Prior work on calibration and uncertainty shows that deep models are often miscalibrated and that explicitly modeling uncertainty improves downstream reliability (Guo et al., 2017; Xue et al., 2025). Similarly, recent investigations into emergent abilities imply that certain high-level capabilities appear only under particular training regimes or scales (Wei et al., 2022; Berti et al., 2025).

A "plausibility-aware" pretraining regime could integrate multiple components, such as training objectives that promote clearer separation between affirmative and negative response patterns (as implemented in PLAUSIEVE), explicit uncertainty modeling through techniques like ensembling, temperature calibration, or entropy aware losses, and small data curricula inspired by the BabyLM challenge to encourage sample efficient learning. Models trained under such regimes would likely produce more interpretable and calibrated belief signals, making them directly deployable in evidence-sparse decision environments, for example, when rapidly prioritizing claims in high-throughput settings.

Future work should include benchmarking and ablation studies to evaluate trade-offs between traditional generation quality and improved epistemic behavior (Charpentier et al., 2025; Guo et al., 2017; Xue et al., 2025).

E.3 Human-centric heuristics combined with LLM knowledge for fast decision systems

Fast, reliable decision making systems in high volume information environments require hybrid approaches that bring together lightweight machine plausibility estimates, selective retrieval when necessary, and human centred heuristics to guide actions. Work on explainable automated fact checking and human in the loop systems shows that fact checkers prefer outputs that are interpretable,

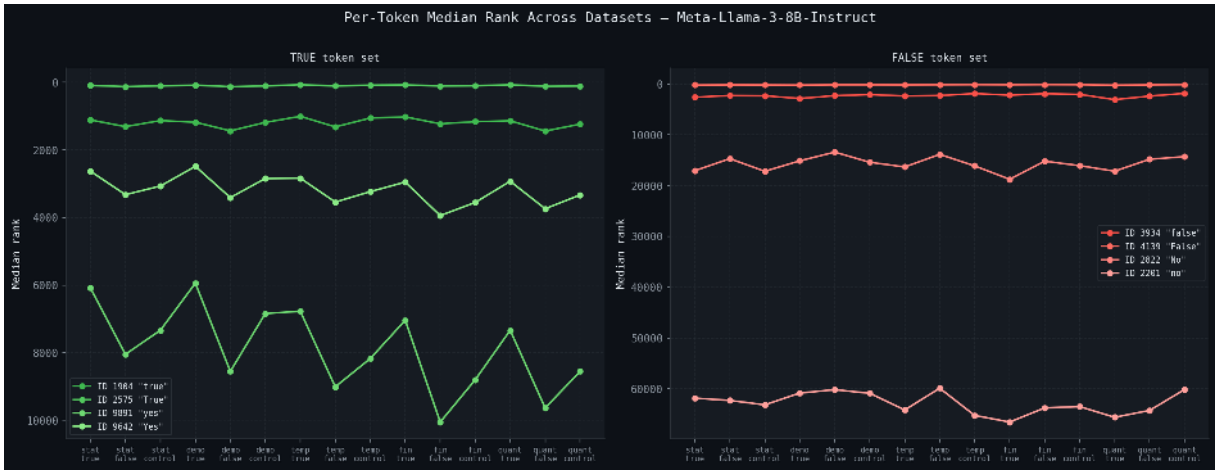


Figure 14: Median vocabulary ranks of individual affirmative and negative tokens across datasets for Meta-Llama-3-8B-Instruct. The selected token sets exhibit consistent ranking behavior across domains, supporting their use as stable indicators of affirmative and negative model judgments.

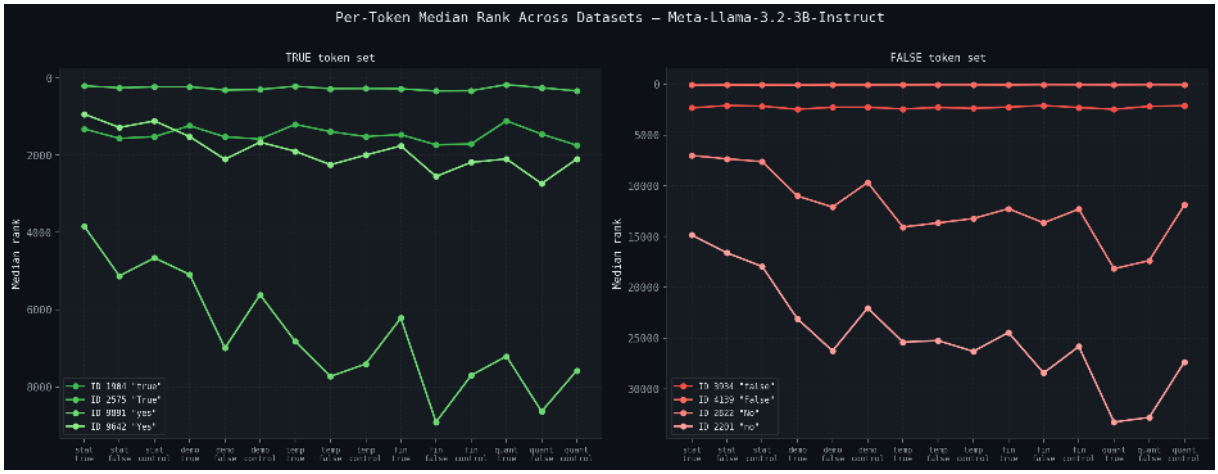


Figure 15: Median vocabulary ranks of selected `True` and `False` token variants for Meta-Llama-3.2-3B-Instruct. Capitalization variants follow similar trends, indicating robustness against stylistic decoding preferences.

uncertainty-aware, and actionable (Kotonya and Toni, 2020; Altoe et al., 2025). By integrating simple human heuristics (e.g., provenance cues, numeric sanity checks, domain-specific red flags) with model-generated plausibility scores and calibrated confidence intervals (from PLAUSIEVE or uncertainty aware modules like UAlign (Xue et al., 2025)), systems can perform quicker sorting of cases while still providing justifications that enable effective human oversight. Such hybrid pipelines are particularly valuable in settings where both latency and throughput are critical such as social platforms, newsrooms, and emergency response because they support safe automation of low risk assessments while flagging high-uncertainty items for human review. Future work should examine human–AI calibration, assessing whether plausibility signals combined with heuristic filters improve decision speed and accuracy under realistic workloads.

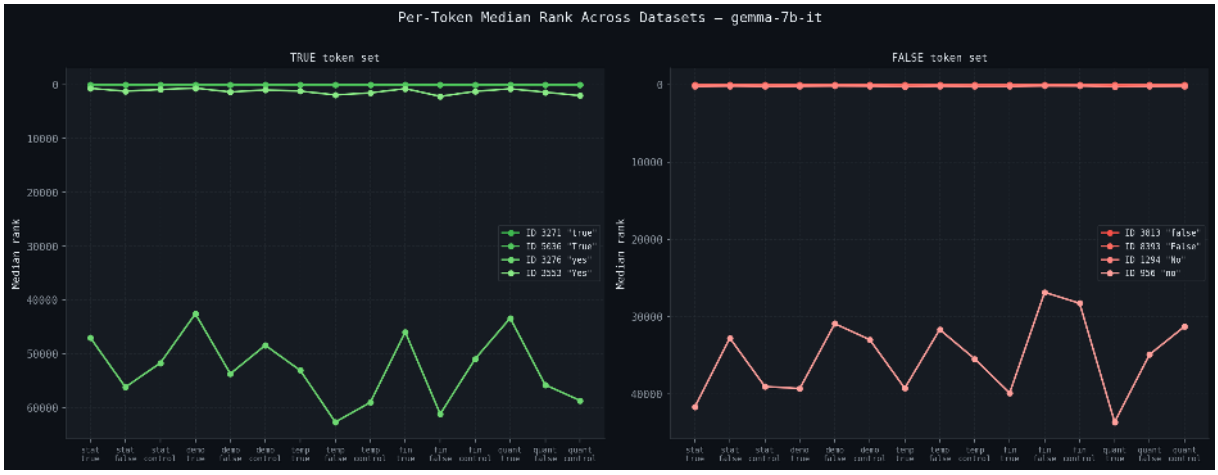


Figure 16: Token-rank analysis for Gemma-7B-IT. While the model exhibits different absolute ranking scales compared to Llama-based models, the affirmative and negative token groups remain internally coherent, validating the semantic grouping approach.

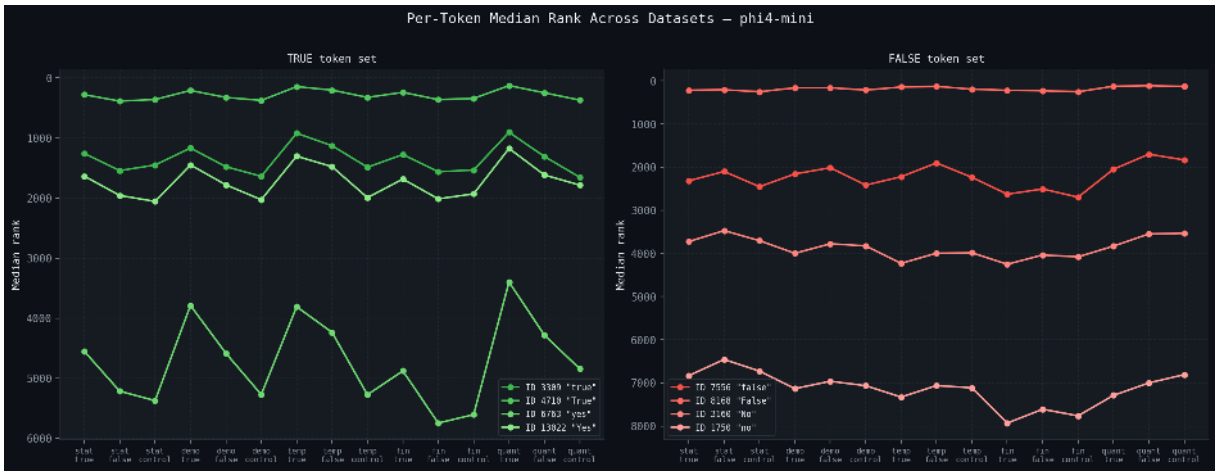


Figure 17: Median token-rank distributions for Phi4-mini across datasets and control settings. The stability of affirmative and negative token clusters across experimental conditions supports the generalizability of the selected token sets.