Rethinking LLM Confidence: From Calibration to Coherence

Anonymous Authors[1]

Abstract

Calibration is the primary criterion for evaluating LLM confidence, but it is insufficient: it admits trivially incoherent estimators, depends on the evaluation distribution, and does not test the extent to which the estimation can be interpreted as a consistent, underlying probability function. To use in real scenarios, we care more how well LLM confidence estimates satisfy the conditions required of coherent probabilistic beliefs. We formalize these conditions along three axes (structural coherence, faithfulness, and usefulness) and operationalize them in **CoherenceBench**. Widely used estimators systematically violate these conditions despite appearing well-calibrated: models assign lower confidence to logically easier questions 31% of the time, and common interventions reducing RMSCE leave structural violations unchanged, suggesting calibration is orthogonal to probabilistic validity. RLHF and chain-of-thought improve usefulness metrics without restoring coherence. To close this gap, we introduce **Reinforcement Learning from Exploitation (RLE)**, which post-trains a model by directly penalizing Dutch-book exploitability across four coherence templates. RLE outperforms Brier-score fine-tuning on structural coherence in- and out-of-distribution, demonstrating that training against axiom violations is more effective than fitting labeled correctness data alone.

1. Introduction

Reliable confidence estimates in LLMs would support principled abstention, cascading to stronger models, and uncertainty-aware aggregation in agentic pipelines. The field has centered on *calibration* as the primary criterion (Geng et al., 2024; Phan et al., 2026; OpenAI et al., 2024), but calibration alone is inadequate. A constant predictor outputting marginal accuracy is perfectly calibrated yet carries no instance-level information. A function calibrated on a benchmark can be arbitrarily miscalibrated on subpopulations. Most fundamentally, calibration ignores internal coherence: a model assigning high confidence to mutually exclusive answers, or deeming a hard question easier than one it logically implies, violates probability axioms while passing calibration tests.

We propose a richer framework grounded in rational belief theory (Ramsey, 1926; Cox, 1946) and the utility engineering approach of (Mazeika et al., 2025), defining three categories: **structural properties** (normalization, conjunction consistency, entailment monotonicity); **faithfulness properties** (prompt and generation semantic invariance); and **usefulness properties** (calibration, discrimination). We instantiate this as **CoherenceBench** and evaluate verbal (Tian et al., 2023), logit-based (Kadavath et al., 2022), and SliCK (Gekhman et al., 2024) confidence estimators. Output-based estimators saturate near certainty, masking structural failures; SliCK is the only estimator with meaningful calibration (RMSCE 0.251 vs. 0.778, 0.700) and discrimination (AUROC 0.825 vs. 0.559, 0.596), while exposing that the underlying model violates entailment monotonicity 31% of the time.

Contributions.

1. We show calibration is insufficient: it admits incoherent and irrational confidence estimators in practice.
2. We introduce structural coherence, faithfulness, and usefulness, and show that standard estimators achieve apparent calibration while masking severe probability violations.
3. We diagnose how training and inference interventions impact performance across our metrics.
4. We propose RLE, which post-trains against Dutch-book exploitability and outperforms Brier-score fine-tuning on structural coherence in- and out-of-distribution.

2. Related Work

Calibration as a Standard. Calibration is the de facto standard for evaluating LLM confidence. Recent surveys (Geng et al., 2024; Zhou et al., 2024), benchmarks like Humanity's Last Exam (Phan et al., 2026), and technical reports (Ope-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Under review by the 2nd Workshop on Epistemic Intelligence in Machine Learning (EIML@ICML 2026). Do not distribute.

nAI et al., 2024) center on metrics like RMSCE. However, we argue that calibration alone is fundamentally insufficient; three critical failures motivate our richer evaluation framework.

Output-Based Confidence. Output-based estimators directly utilize token log-probabilities (Kadavath et al., 2022; Ye et al., 2024) or explicitly prompted verbalized scores (Lin et al., 2022; Tian et al., 2023). Though verbalized confidence often achieves lower calibration error (Tian et al., 2023), it suffers from prompt sensitivity (Xiong et al., 2024) and domain-specific overconfidence (Phan et al., 2026; Zhou et al., 2024). We evaluate both as baselines, revealing that their seemingly low calibration error often stems from score saturation artifacts rather than genuine uncertainty awareness.

Consistency and Sampling. Alternatively, sampling-based methods measure output consistency, positing that correct answers are generated more stably than hallucinations (Wang et al., 2023; Manakul et al., 2023). This is formalized by SliCK (Gekhman et al., 2024), which calculates sample agreement rates, and Semantic Entropy (Kuhn et al., 2023), which clusters by meaning instead of exact tokens. While computationally expensive, these methods span the full confidence range—a property we show is necessary to expose structural failures hidden by output-based methods.

Coherence and Consistency. While prior work explores consistency (e.g., against prompt paraphrasing (Elazar et al., 2021)), it generally treats it as a performance metric rather than a probabilistic requirement. Closest to our framework is the utility engineering approach of (Mazeika et al., 2025), which audits whether LLM preferences satisfy rational utility axioms. We apply this identical logic to confidence: structural coherence is not merely a desideratum, but a necessary condition for model outputs to be interpretable as valid probabilities. Recent work on RL fine-tuning (Shao et al., 2024) has shown that structured reward signals can instill formal properties beyond what supervised data alone provides; RLE applies this insight specifically to probabilistic axioms, using Dutch-book exploitability as the reward rather than task correctness.

3. Motivation & Formalization

3.1. Confidence Functions and Estimators

Let $\mathcal{M}$ be a language model, x a prompt, and y a generated response. Let $\simeq$ denote a semantic equivalence relation over strings, and write $[x]$ and $[y]$ for the corresponding equivalence classes.

We model each prompt x as inducing a latent random variable over semantic answer classes:

$$Z_x \in \mathcal{Y}_x := \{[y] : y \text{ is a valid response to } x\}.$$

We assume that correctness is defined at the level of equivalence classes, and that all semantically correct answers belong to a single class $[y^\star] \in \mathcal{Y}_x$.[1] Confidence must be defined over semantic outcomes rather than surface forms; otherwise paraphrased answers can receive different probabilities. A *confidence function* is then a probability distribution over these classes:

$$c : ([x], [y]) \mapsto [0, 1], \quad c([x], [y]) := P(Z_x = [y]),$$

with $\sum_{[y]\in\mathcal{Y}_x} c([x], [y]) = 1$.

Under this definition, the probability that the model answers correctly is:

$$P(\text{correct} \mid x) = c([x], [y^\star]) = \max_{[y]} c([x], [y]).$$

We therefore define the prompt-level confidence as

$$\bar{c}([x]) := \max_{[y]} c([x], [y]),$$

which admits a direct decision-theoretic interpretation: under 0–1 loss, a rational agent outputs $\arg\max_{[y]} c([x], [y])$, and $\bar{c}([x])$ is its self-assessed probability of correctness. In practice, $\bar{c}$ is approximated via finite samples.

This probabilistic interpretation is justified by standard arguments from rational belief theory. Under the Dutch book argument (Ramsey, 1926), $c([x], [y])$ corresponds to the fair price of a contract paying \$1 if $[y]$ is correct. Cox's theorem (Cox, 1946) further implies that any consistent system of beliefs over such events must be isomorphic to a probability measure.

Estimators. The true confidence function c is not directly observable. Instead, we evaluate *confidence estimators* $\hat{c}$, computable functions:

$$\hat{c} : (x, y) \mapsto [0, 1],$$

which approximate $c([x], [y])$ from text.

Different estimators provide different approximations:

- Output-based methods (verbal, logit-based) estimate confidence conditioned on a single generation.
- Sampling-based methods (e.g., SliCK) approximate the full distribution over equivalence classes via Monte Carlo sampling:

  $$\hat{c}(x, y) \approx \frac{|\{j : y_j \in [y]\}|}{k},$$

  which is a consistent estimator of $c([x], [y])$ under i.i.d. sampling.

[1] When multiple surface forms are valid, they are grouped into a single equivalence class via $\simeq$.

This distinction is central: structural properties (e.g., normalization, entailment) are defined over the full distribution $c([x], \cdot)$, and therefore require estimators that meaningfully approximate it. We further clarify and prove all claims regarding the confidence function and benchmark validity in Appendix A.

3.2. Why Calibration Is Insufficient

Calibration is the primary evaluation criterion for confidence estimation in the LLM literature. A confidence function c is *calibrated* on a distribution of prompt-generation pairs $\mathcal{D}$ if for all $p \in [0, 1]$, $P_{(x,y)\sim\mathcal{D}}([y] \text{ is correct for } [x] \mid c([x], [y]) = p) = p$. Calibration is typically evaluated using the root-mean-square calibration error (RMSCE) (Phan et al., 2026), which bins confidence scores into $B = 20$ equal-width bins over $[0, 1]$ and computes:

$$\text{RMSCE} = \sqrt{\sum_{b=1}^{B} \frac{n_b}{N} \left(\text{acc}_b - \mu_b\right)^2}$$

where n_b is the number of pairs in bin b, acc_b is the fraction correct, and μ_b is the mean confidence. Surveys of LLM uncertainty quantification (Geng et al., 2024) organize the field around calibration as the primary desideratum, and recent benchmarks such as Humanity's Last Exam (Phan et al., 2026) report calibration error as a central evaluation metric. The GPT-4 technical report's finding that RLHF degrades calibration relative to the base model (OpenAI et al., 2024) has further entrenched it as the canonical evaluation metric for confidence.

Unfortunately, calibration as a sole criterion admits confidence functions that are internally incoherent. We illustrate this with two examples.

Constant predictor. Let α denote the model's marginal accuracy on $\mathcal{D}$. The confidence function $c([x], [y]) = \alpha$ for all (x, y) is perfectly calibrated on $\mathcal{D}$, achieving RMSCE of exactly zero. Since c is constant, every prediction falls into the single bin $b^\star$ containing α, so $\mu^\star = \alpha = \text{acc}_{b^\star}$ and the sum vanishes:

$$\text{RMSCE} = \sqrt{\frac{N}{N}(\alpha - \alpha)^2} = 0.$$

Yet c carries no instance-level information and cannot distinguish a question the model answers reliably from one it answers by chance.

Calibration is distribution-relative. More fundamentally, calibration is not a property of c alone, but of c paired with $\mathcal{D}$. A function well-calibrated on $\mathcal{D}$ can be arbitrarily miscalibrated on sub-distributions of $\mathcal{D}$ itself.

Let $\mathcal{D}' \subset \mathcal{D}$ be the sub-distribution of examples the model answers incorrectly. On $\mathcal{D}'$, every bin has accuracy zero, so each bin contributes a strictly positive term to the squared RMSCE:

$$(n'_b/N') \cdot (\bar{c}'_b)^2$$

Hence:

$$\text{RMSCE}(\mathcal{D}') > 0 = \text{RMSCE}(\mathcal{D})$$

Any partition of $\mathcal{D}$—by topic, difficulty, or domain—yields sub-distributions on which c may be substantially miscalibrated. RMSCE on a benchmark thus characterizes aggregate behavior and provides no guarantee about the subpopulations that matter in deployment.

4. Methodology

In this section we describe our multidimensional evaluation of Confidence and Operationalize it to construct CoherenceBench

4.1. Axioms for Confidence Evaluation

We define a set of properties for evaluating confidence in language models, organized into three categories. Drawing on the decision-theoretic foundations of rational belief (the Dutch book argument (Ramsey, 1926); Cox's theorem (Cox, 1946)) and on the utility engineering framework of (Mazeika et al., 2025), we define three orthogonal categories of desiderata. **Structural properties** are hard constraints derived from the probability axioms: whether the model's beliefs normalize, respect the product rule, and respect logical entailment. **Faithfulness properties** ask whether a confidence estimator faithfully represents the underlying confidence function, requiring invariance to surface-level rephrasing. **Usefulness properties** ask whether confidence tracks ground truth, encompassing calibration and discrimination. Then, we operationalize these

4.1.1. Structural Properties

Structural properties are hard constraints on c derived from the probability axioms. Violations indicate that the model's beliefs are internally contradictory.

Normalization requires that for any prompt class $[x]$:

$$\sum_{[y]} c([x], [y]) = 1.$$

where we average confidence within each equivalence class to avoid double-counting multiple surface forms of the same answer. Since the answer classes partition the response space, their probabilities must sum to one. We measure this via the normalization deviation $|S(x) - 1|$ on 1,500 SimpleQA (Wei et al., 2024) questions with $k = 16$ rollouts per question at temperature $T = 0.5$, where

$$S(x) = \sum_j \frac{1}{|[y_j]|} \sum_{y \in [y_j]} \hat{c}(x, y).$$

Conjunction Consistency requires that if correctly answering $[x]$ decomposes into a first-hop sub-question $[x_1]$ with gold answer $[y_1^*]$ followed by a second-hop sub-question $[x_2]$, then

$$\bar{c}([x]) = c(x_1, y_1^*) \cdot \bar{c}([x_2] \mid [x_1], [y_1^*]),$$

by the product rule $P(A \cap B) = P(A) \cdot P(B \mid A)$. We measure the deviation

$$\Delta(x) = \left|\bar{\hat{c}}(x) - \hat{c}(x_1, y_1^*) \cdot \bar{\hat{c}}(x_2 \mid x_1, y_1^*)\right|$$

on 2-hop MuSiQue (Trivedi et al., 2021) questions ($k = 16$, $T = 0.5$), with $\bar{\hat{c}}$ estimated as max confidence across rollouts.

Entailment Monotonicity requires that if answering $[x]$ correctly entails answering $[x']$ correctly, then

$$\bar{c}([x]) \leq \bar{c}([x']).$$

The event of answering $[x]$ correctly is a subset of the event for $[x']$, and coherent credences must respect set inclusion (where x' corresponds to the simplified sub-question obtained by conditioning on the first-hop answer.) We measure violation magnitude

$$\Delta(x) = \max\big(0,\ \bar{\hat{c}}(x) - \bar{\hat{c}}(x_2 \mid x_1, y_1^*)\big)$$

on MuSiQue ($k = 16$, $T = 0.5$), with $\bar{\hat{c}}$ as max confidence across rollouts.

4.1.2. Faithfulness Properties

Faithfulness properties constrain $\hat{c}$ to be consistent with a well-formed underlying c. Since c is defined over equivalence classes, a faithful estimator must be invariant to surface-level reformulation. Violations indicate that $\hat{c}$ is sensitive to features of the text that are invisible at the equivalence class level, and therefore cannot faithfully represent c.

Prompt Semantic Invariance requires that for semantically equivalent prompts $x \simeq x'$:

$$\hat{c}(x, y) = \hat{c}(x', y).$$

We measure the deviation $\Delta(f) = |\bar{\hat{c}}(x) - \bar{\hat{c}}(x')|$ across paraphrase pairs from ParaRel (Elazar et al., 2021), sampling 1,500 facts with two paraphrase templates each ($k = 16$, $T = 0.5$).

Generation Semantic Invariance requires that for semantically equivalent generations $y \simeq y'$:

$$\hat{c}(x, y) = \hat{c}(x, y').$$

We measure the within-class spread $\Delta([y]) = \max_{y \in [y]} \hat{c}(x, y) - \min_{y \in [y]} \hat{c}(x, y)$ across equivalence classes on 1,500 SimpleQA questions ($k = 16$, $T = 0.5$).

4.1.3. Usefulness Properties

Usefulness properties ask whether confidence tracks ground truth. Unlike structural and faithfulness properties, they are distribution-relative by design.

Calibration requires that for all $p \in [0, 1]$:

$$P\big([y] \text{ correct} \mid c([x], [y]) = p\big) = p.$$

Among all pairs assigned confidence p, exactly a fraction p should be correct. We measure RMSCE on 1,500 SimpleQA questions ($k = 16$, $T = 0.5$, $B = 20$ bins); lower is better.

Discrimination requires that correct generations always receive higher confidence than incorrect ones. This is distinct from calibration: a constant predictor achieves perfect calibration but chance-level discrimination. We measure AUROC over all (confidence, correctness) pairs on the same SimpleQA sample; higher is better.

4.2. Confidence Estimation Methods

We compare three representative estimators.

Verbal Confidence (Tian et al., 2023). The model generates a response y to x, then is asked in a follow-up to state the probability its answer is correct. The parsed numerical response is $\hat{c}(x, y)$.

Logit-based Confidence (Kadavath et al., 2022). The model is prompted to verify whether y is true or false for x; confidence is the normalized true-token probability:

$$\hat{c}(x, y) = \frac{P(\text{True})}{P(\text{True}) + P(\text{False})}.$$

SliCK (Gekhman et al., 2024). We sample $k = 16$ rollouts $y_1, \ldots, y_k$ to x, group them into equivalence classes under $\hat{\simeq}$ via LLM-as-a-judge, and exclude refusals and truncated outputs (letting k' denote remaining rollouts). Confidence is the fraction of equivalent rollouts:

$$\hat{c}(x, y) = \frac{|\{j : y_j \mathrel{\hat{\simeq}} y\}|}{k'}.$$

The main experiments use **Qwen-30B-A3B-Thinking** (Yang et al., 2025), a 30B-parameter mixture-of-experts reasoning model with 3B active parameters, serving as both generation and evaluation model. Generations in which the model declines to answer or exhausts its token limit are excluded. Section 4.4 evaluates 11 additional models on a 200-question subset of each task with the same judge; full per-model results are in Appendix C, this ensures that results are not judge bias related.

4.3. Estimator Comparison

We organize findings around what they reveal about each estimator, not what each property scores in isolation. We note

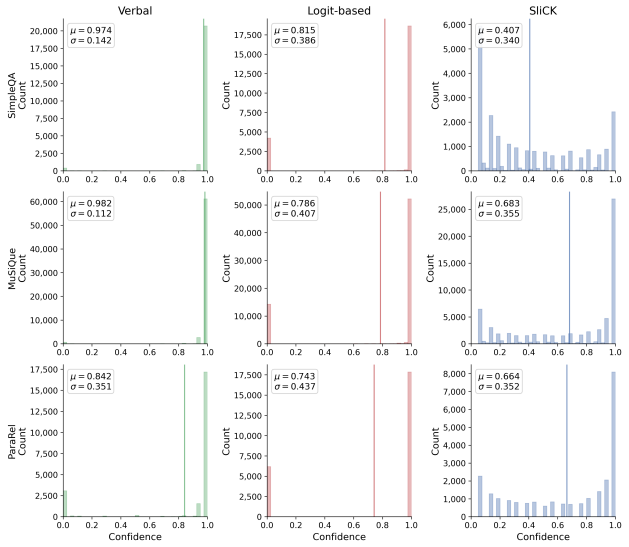

Figure 1. Confidence score distributions across estimators (columns) and datasets (rows). Verbal and logit-based saturate at extremes; SliCK spans the full range.

upfront two design choices that shape interpretation. First, we evaluate raw logit-based and verbal confidence without post-hoc calibration (e.g., temperature scaling). This is intentional: post-hoc rescaling can reduce RMSCE but does not alter the structural outputs—a recalibrated estimator still assigns the same relative ordering and normalization behavior. We verify this in Appendix C. Second, SliCK satisfies normalization and generation semantic invariance *by construction*; we include these not as empirical victories for SliCK but as diagnostic anchors that reveal how far output-based estimators deviate from provably achievable baselines. Three patterns dominate the remaining empirical results.

Saturation makes output-based scores structurally vacuous. Verbal and logit-based confidence concentrate nearly all mass at extreme values (Figure 1), and this directly corrupts their structural scores. Normalization deviation averages 5.055 (verbal) and 4.132 (logit-based)—the model simultaneously treats many mutually exclusive answers as near-certain (Figure 2). Their low scores on conjunction consistency (0.060), entailment monotonicity (4.9% violations), and prompt invariance (0.025) are ceiling artifacts: $1.0 \approx 1.0 \times 1.0$ trivially. Calibration error alone systematically rewards this failure mode and hides exactly the incoherence that matters for downstream decisions.

SliCK surfaces genuine model incoherence. With a real output distribution, SliCK exposes violations the model itself commits. Conjunction consistency deviation averages 0.257 on MuSiQue—comparable to logit-based (0.268)—

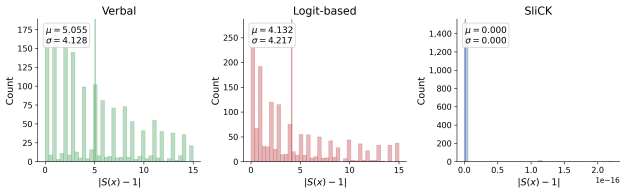

Figure 2. Normalization deviation $|S(x) - 1|$. Output-based estimators violate severely (5.055, 4.132); SliCK satisfies exactly by construction.

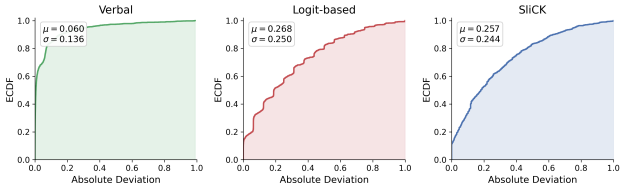

Figure 3. Conjunction consistency on MuSiQue. Verbal's low deviation is a saturation artifact; SliCK and logit-based reveal genuine violations.

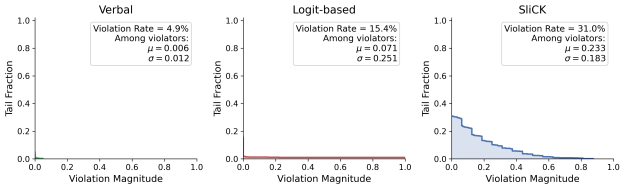

Figure 4. Entailment monotonicity violations. SliCK: 31.0% violation rate; apparent compliance of saturated estimators is a ceiling effect.

and entailment monotonicity is violated on **31.0%** of questions (Figure 3, 4). Providing the first-hop gold answer makes the second hop strictly easier by construction, yet confidence decreases frequently and substantially. These are model-level failures that saturated estimators cannot surface.

Faithfulness failures differ by estimator type. SliCK satisfies normalization over the empirical support induced by sampling. ($\mu = 0.000$), while verbal and logit-based estimators assign maximally different scores to generations they consider semantically equivalent (Figure 6). SliCK's own failure is prompt invariance: paraphrases yield $\mu = 0.163$ deviation (Figure 5), a consequence of independent rollout sampling per prompt.

Usefulness tracks structural integrity. SliCK achieves RMSCE 0.251 versus 0.778 and 0.700, and AUROC 0.825 versus 0.559 and 0.596 (Figure 7)—the only estimator with meaningful calibration and discrimination. The estimator that produces a real distribution is also the one that is useful.

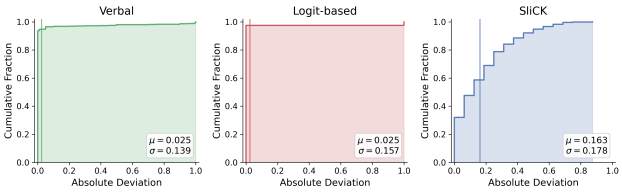

Figure 5. Prompt semantic invariance on ParaRel. Output-based estimators vacuously consistent; SliCK genuinely sensitive ($\mu = 0.163$).

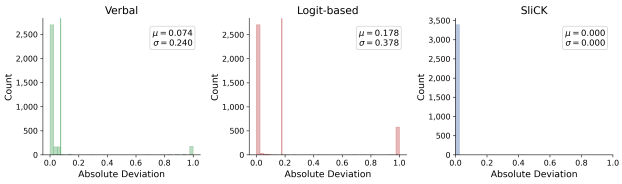

Figure 6. Within-class generation spread. SliCK exactly invariant; verbal and logit-based string-dependent.

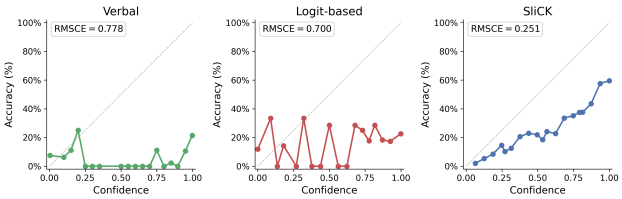

Figure 7. Calibration diagrams and confidence distributions. SliCK alone tracks correctness; output-based estimators report near-certain confidence regardless of correctness.

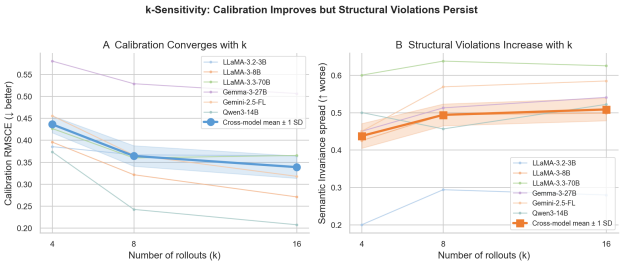

Figure 8. As rollout count k increases, calibration error (RMSCE) converges downward while semantic invariance violations increase, demonstrating that structural violations are a property of the model distribution rather than sampling noise.

4.4. Model Interventions

Running the benchmark on a random set of 200 samples per task (for computational limitations) across models in Appendix C we detail important findings.

Model Size. Model size does not correlate cleanly with most coherence metrics, but Semantic Invariance shows a clear scaling trend across 9 models: smaller models are more sensitive to subtle prompt changes, with LLaMA-3-3B exhibiting 40% more average n-gram diversity over 16 rollouts than LLaMA-3-70B.

Impact of RLHF Comparing Llama-3.1-70b with Nemotron-Llama-3.1-70b(graph in Appendix C, a model trained only via RLHF rewards (Wang et al., 2024), RLHF modestly improves SliCK calibration (RMSCE: $0.365 \rightarrow 0.325$), it catastrophically collapses discriminability to chance (AUROC: $0.591 \rightarrow 0.498$). By uniformly inflating output confidence regardless of correctness, RLHF erases variance, leaving verbal calibration static and actively worsening conjunction consistency ($+15\%$). This reveals a critical vulnerability: modern alignment optimizes for confident outputs at the direct expense of accurate uncertainty quantification and structural coherence.

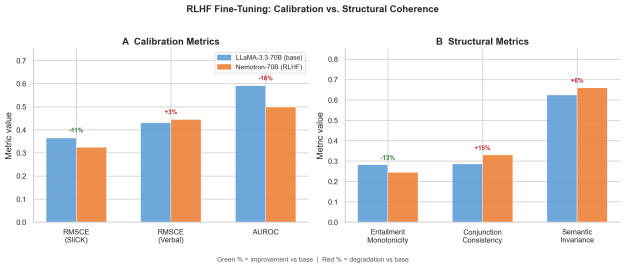

Figure 9. **Comparison against RLHF.** Impact of RLHF alignment across all metrics.

Chain-of-Thought. We compare LLaMA-3-8B-Instruct with and without the Kojima et al. zero-shot chain-of-thought suffix ("Let's think step by step.") across all three datasets. Results on other models are in Appendix C. Calibration improved substantially (RMSCE: $0.271 \rightarrow 0.212$, -22%) and conjunction consistency improved dramatically (mean deviation: $0.244 \rightarrow 0.144$, -41%), indicating that structured reasoning reduces overconfidence in multi-hop decompositions. Semantic invariance spread was unchanged ($0.498 \rightarrow 0.488$, -2%), however, confirming that sensitivity to prompt rephrasing is a structural property of the confidence estimator that explicit reasoning cannot resolve.

Sample Size To verify that structural violations are not mere artifacts of estimating SliCK confidence from a finite number of rollouts, we subsample our existing k=16 generations to $k \in \{4, 8, 16\}$ and recompute all metrics without generating new data. Averaging across six models, RMSCE decreases by 22% as k grows ($0.436 \rightarrow 0.339$), confirming that calibration estimates converge with additional samples as expected. Crucially, semantic invariance spread moves in the opposite direction, increasing by 16% over the same range ($0.438 \rightarrow 0.508$; Figure 8). Rather than mitigating the issue, more rollouts sharpen confidence estimates and

thereby expose structural inconsistencies that coarse, low-k estimates had obscured.

5. Learning from Incoherence

The CoherenceBench results show incoherence is pervasive across estimator types and model scales. We now ask whether it can be trained away. **Reinforcement Learning from Exploitation (RLE)** turns the Dutch-book fair-price interpretation of Section 3.1 into a reward: the model is penalised for any price vector an adversary could exploit.

5.1. Method

Training episodes. Episodes use the four templates from Section 4.1 on the same datasets (SimpleQA, MuSiQue (Trivedi et al., 2021), ParaRel (Elazar et al., 2021)). The model quotes a price $c_i \in [0, 1]$ for a contract paying \$1 if a given answer is correct, grounding the fair-price interpretation of Section 3.1 without any verbal calibration prompt.

Reward. Exploitability is the ℓ_1-distance from prices $\mathbf{c}$ to the nearest coherent point in $\mathcal{S}$ (via the minimax theorem), solved as an LP (Huangfu & Hall, 2018):

$$\text{Exploit}(\mathbf{c}, \mathcal{S}) = \min_{\boldsymbol{\omega} \in \mathcal{S}} \|\mathbf{c} - \boldsymbol{\omega}\|_1. \tag{1}$$

The training reward is

$$R(\mathbf{c}, \mathbf{e}) = -\,\text{Exploit}(\mathbf{c}, \mathcal{S}) \;-\; \frac{\lambda}{n} \sum_{i=1}^{n} (c_i - e_i)^2, \tag{2}$$

where $\mathbf{e} \in \{0, 1\}^n$ is ground-truth correctness and λ=0.5.

GRPO fine-tuning. We fine-tune Qwen2.5-7B-Instruct (Yang et al., 2025) via GRPO (Shao et al., 2024) with LoRA (r=16, α=32) (Hu et al., 2021). Per episode, N=8 rollouts are drawn and rewards normalised to advantages $\hat{A}^{(j)} = (R^{(j)} - \mu_R)/(\sigma_R + \epsilon)$. The loss is

$$\mathcal{L} = -\sum_j \hat{A}^{(j)} \sum_t \log \pi_\theta(a_t^{(j)}) + \beta\, \text{KL}(\pi_\theta \| \pi_{\text{ref}}), \tag{3}$$

with β=0.05; ≈10M of 7.6B parameters are trainable.

5.2. Experimental Setup

We train on 800 SimpleQA, 400 MuSiQue, and 400 ParaRel records (base model answers, frozen). Three conditions are compared on a held-out pool: (i) **Base**; (ii) **Brier-Only** (exploit term removed); and (iii) **RLE** (Eq. 2). OOD evaluation uses HLE, an expert-level benchmark with 6% base-model accuracy, unseen during training.

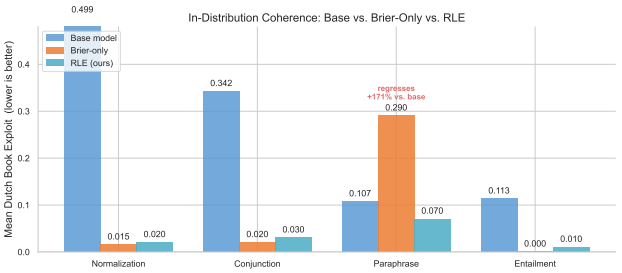

Figure 10. In-distribution exploit per template. Brier-only regresses on paraphrase (+171% vs. base); RLE improves uniformly.

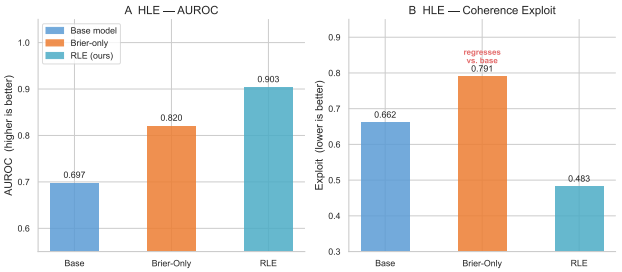

Figure 11. OOD results on HLE. Panel A: AUROC, RLE leads by 10 points. Panel B: exploit, Brier-only regresses vs. base (+19%), RLE improves (−27%).

5.3. Results

RLE beats Brier-only where structural constraints matter. Figure 10 shows in-distribution exploit per template. Both fine-tuned models reduce exploit comparably on normalization, conjunction, and entailment. On **paraphrase**, Brier-only increases exploit from 0.107 to 0.290 (+171% vs. base) while RLE reduces it to 0.070. Brier-only sharpens each price toward its label independently; when two rephrasing of the same fact yield different model answers, this drives prices apart and worsens coherence. The exploit penalty in Eq. (1) detects this gap directly; pointwise Brier loss cannot.

The advantage grows out-of-distribution. On HLE (Figure 11), RLE achieves AUROC 0.903 vs. 0.820 for Brier-only, a 10-point gap. On coherence exploit, Brier-only regresses from 0.662 to 0.791 (+19%) while RLE reduces it to 0.483 (−27%). The same failure mode observed in-distribution is amplified under shift: pointwise calibration without the structural signal degrades coherence on hard, unfamiliar questions.

Axioms beat labeled data. Both conditions share the same training data and base model; only the exploit term differs. Brier-only regresses on paraphrase in-distribution and on coherence OOD; RLE improves on both. Fitting labelled correctness is insufficient to learn a probability

function. Once enforced via the exploit signal, structural coherence transfers zero-shot to unseen datasets, suggesting RLE teaches the model a property of probability rather than a distributional pattern.

6. Discussion and Conclusion

We argued that calibration alone is insufficient for evaluating LLM confidence: it is satisfied by trivial constant predictors, depends on the evaluation distribution, and is silent about internal consistency. CoherenceBench operationalizes a richer evaluation along three axes—structural coherence, faithfulness, and usefulness—and applying it changes the empirical picture in two ways worth highlighting.

Apparent calibration can reflect saturation, not coherence. Verbal and logit-based estimators concentrate scores near 1.0, which inflates apparent consistency on multiple structural and faithfulness properties. Single-number RMSCE cannot distinguish this regime from genuine coherence; SliCK's broader output range exposes structural violations the model itself harbors, including a 31% entailment monotonicity violation rate and substantial conjunction inconsistency.

Coherence and correctness do not scale together. Within the LLaMA-3 family, prompt semantic invariance worsens with parameter count while RMSCE does not. Faithfulness and calibration are distinguishable axes, and a single metric cannot adjudicate between them.

Axiom-driven post-training as a general principle. RLE suggests a broader pattern: any model output that must satisfy a formal axiomatic structure may be better post-trained against axiom violations than against labelled examples alone. Confidence is one instance; others include utility functions (which must satisfy transitivity and independence to support coherent decision-making (Mazeika et al., 2025)), reward models (which should respect dominance and compositionality), and power-seeking measures (which must be monotone under resource inclusion to be meaningful (Turner et al., 2021)). In each case, a dataset of labelled outcomes trains the function's values but not its structure; the exploit signal trains the structure directly. Whether this advantage holds across these settings, and how axiomatic constraints interact with scale and RLHF alignment (OpenAI et al., 2024), are open questions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. Improving the reliability of confidence estimation in LLMs has broad positive societal implications, enabling safer deployment of these systems in high-stakes applications. There are no specific ethical harms we feel must be highlighted here.

References

Cox, R. T. Probability, frequency and reasonable expectation. *Journal of Symbolic Logic*, 37(2):398–399, 1946. doi: 10.2307/2272983.

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., and Goldberg, Y. Measuring and improving consistency in pretrained language models, 2021. URL https://arxiv.org/abs/2102.01017.

Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., and Herzig, J. Does fine-tuning llms on new knowledge encourage hallucinations?, 2024. URL https://arxiv.org/abs/2405.05904.

Geng, J., Cai, F., Wang, Y., Koeppl, H., Nakov, P., and Gurevych, I. A survey of confidence estimation and calibration in large language models, 2024. URL https://arxiv.org/abs/2311.08298.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Huangfu, Q. and Hall, J. A. J. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018. doi: 10.1007/s12532-017-0130-5. URL https://doi.org/10.1007/s12532-017-0130-5.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023. URL https://arxiv.org/abs/2302.09664.

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words, 2022. URL https://arxiv.org/abs/2205.14334.

Manakul, P., Liusie, A., and Gales, M. J. F. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023. URL https://arxiv.org/abs/2303.08896.

Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Khoja, A., Zhang, O., and Hendrycks, D. Utility engineering: Analyzing and controlling emergent value systems in ais, 2025. URL https://arxiv.org/abs/2502.08640.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Phan, L., Gatti, A., Li, N., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., Hendrycks, D., Han, Z., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., Choi, M., Agrawal, A., Chopra, A., Nattanmai, A., McKellips, G., Cheraku, A., Suhail, A., Luo, E., Deng, M., Luo, J., Zhang, A., Jindel, K., Paek, J., Halevy, K., Baranov, A., Liu, M., Avadhanam, A., Zhang, D., Cheng, V., Ma, B., Fu, E., Do, L., Lass, J., Yang, H., Sunkari, S., Bharath, V., Ai, V., Leung, J., Agrawal, R., Zhou, A., Chen, K., Kalpathi, T., Xu, Z., Wang, G., Xiao, T., Maung, E., Lee, S., Yang, R., Yue, R., Zhao, B., Yoon, J., Sun, X., Singh, A., Peng, C., Osbey, T., Wang, T., Echeazu, D., Wu, T., Patel, S., Kulkarni, V., Sundarapandiyan, V., Le, A., Nasim, Z., Yalam, S., Kasamsetty, R., Samal, S., Sun, D., Shah, N., Saha, A., Zhang, A., Nguyen, L., Nagumalli, L., Wang, K., Wu, A., Telluri, A., Yue, S., Wang, A., Dodonov, D., Nguyen, T., Lee, J., Anderson, D., Doroshenko, M., Stokes, A. C., Mahmood, M., Pokutnyi, O., Iskra, O., Wang, J. P., Levin, J.-C., Kazakov, M., Feng, F., Feng, S. Y., Zhao, H., Yu, M., Gangal, V., Zou, C., Wang, Z., Popov, S., Gerbicz, R., Galgon, G., Schmitt, J., Yeadon, W., Lee, Y., Sauers, S., Sanchez, A., Giska, F., Roth, M., Riis, S., Utpala, S., Burns, N., Goshu, G. M., Naiya, M. M., Agu, C., Giboney, Z., Cheatom, A., Fournier-Facio, F., Crowson, S.-J., Finke, L., Cheng, Z., Zampese, J., Hoerr, R. G., Nandor, M., Park, H., Gehrunger, T., Cai, J., McCarty, B., Garretson, A. C., Taylor, E., Sileo, D., Ren, Q., Qazi, U., Li, L., Nam, J., Wydallis, J. B., Arkhipov, P., Shi, J. W. L., Bacho, A., Willcocks, C. G., Cao, H., Motwani, S., de Oliveira Santos, E., Veith, J., Vendrow, E., Cojoc, D., Zenitani, K., Robinson, J., Tang, L., Li, Y., Vendrow, J., Fraga, N. W., Kuchkin, V., Maksimov, A. P., Marion, P., Efremov, D., Lynch, J., Liang, K., Mikov, A., Gritsevskiy, A., Guillod, J., Demir, G., Martinez, D., Pageler, B., Zhou, K., Soori, S., Press, O., Tang, H., Rissone, P., Green, S. R., Brüssel, L., Twayana, M., Dieuleveut, A., Imperial, J. M., Prabhu, A., Yang, J., Crispino, N., Rao, A., Zvonkine, D., Loiseau, G., Kalinin, M., Lukas, M., Manolescu, C., Stambaugh, N., Mishra, S., Hogg,

T., Bosio, C., Coppola, B. P., Salazar, J., Jin, J., Sayous, R., Ivanov, S., Schwaller, P., Senthilkumar, S., Bran, A. M., Algaba, A., Van den Houte, K., Van Der Sypt, L., Verbeken, B., Noever, D., Kopylov, A., Myklebust, B., Li, B., Schut, L., Zheltonozhskii, E., Yuan, Q., Lim, D., Stanley, R., Yang, T., Maar, J., Wykowski, J., Oller, M., Sahu, A., Ardito, C. G., Hu, Y., Kamdoum, A. G. K., Jin, A., Vilchis, T. G., Zu, Y., Lackner, M., Koppel, J., Sun, G., Antonenko, D. S., Chern, S., Zhao, B., Arsene, P., Cavanagh, J. M., Li, D., Shen, J., Crisostomi, D., Zhang, W., Dehghan, A., Ivanov, S., Perrella, D., Kaparov, N., Zang, A., Sucholutsky, I., Kharlamova, A., Orel, D., Poritski, V., Ben-David, S., Berger, Z., Whitfill, P., Foster, M., Munro, D., Ho, L., Sivarajan, S., Hava, D. B., Kuchkin, A., Holmes, D., Rodriguez-Romero, A., Sommerhage, F., Zhang, A., Moat, R., Schneider, K., Kazibwe, Z., Clarke, D., Kim, D. H., Dias, F. M., Fish, S., Elser, V., Kreiman, T., Vilchis, V. E. G., Klose, I., Anantheswaran, U., Zweiger, A., Rawal, K., Li, J., Nguyen, J., Daans, N., Heidinger, H., Radionov, M., Rozhoň, V., Ginis, V., Stump, C., Cohen, N., Poświata, R., Tkadlec, J., Goldfarb, A., Wang, C., Padlewski, P., Barzowski, S., Montgomery, K., Stendall, R., Tucker-Foltz, J., Stade, J., Rogers, T. R., Goertzen, T., Grabb, D., Shukla, A., Givré, A., Ambay, J. A., Sen, A., Aziz, M. F., Inlow, M. H., He, H., Zhang, L., Kaddar, Y., Ängquist, I., Chen, Y., Wang, H. K., Ramakrishnan, K., Thornley, E., Terpin, A., Schoelkopf, H., Zheng, E., Carmi, A., Brown, E. D. L., Zhu, K., Bartolo, M., Wheeler, R., Stehberger, M., Bradshaw, P., Heimonen, J., Sridhar, K., Akov, I., Sandlin, J., Makarychev, Y., Tam, J., Hoang, H., Cunningham, D. M., Goryachev, V., Patramanis, D., Krause, M., Redenti, A., Aldous, D., Lai, J., Coleman, S., Xu, J., Lee, S., Magoulas, I., Zhao, S., Tang, N., Cohen, M. K., Paradise, O., Kirchner, J. H., Ovchynnikov, M., Matos, J. O., Shenoy, A., Wang, M., Nie, Y., Sztyber-Betley, A., Faraboschi, P., Riblet, R., Crozier, J., Halasyamani, S., Verma, S., Joshi, P., Meril, E., Ma, Z., Andréoletti, J., Singhal, R., Platnick, J., Nevirkovets, V., Basler, L., Ivanov, A., Khoury, S., Gustafsson, N., Piccardo, M., Mostaghimi, H., Chen, Q., Singh, V., Khánh, T. Q., Rosu, P., Szlyk, H., Brown, Z., Narayan, H., Menezes, A., Roberts, J., Alley, W., Sun, K., Patel, A., Lamparth, M., Reuel, A., Xin, L., Xu, H., Loader, J., Martin, F., Wang, Z., Achilleos, A., Preu, T., Korbak, T., Bosio, I., Kazemi, F., Chen, Z., Bálint, B., Lo, E. J. Y., Wang, J., Nunes, M. I. S., Milbauer, J., Bari, M. S., Wang, Z., Ansarinejad, B., Sun, Y., Durand, S., Elgnainy, H., Douville, G., Tordera, D., Balabanian, G., Wolff, H., Kvistad, L., Milliron, H., Sakor, A., Eron, M., Favre, A., Shah, S., Zhou, X., Kamalov, F., Abdoli, S., Santens, T., Barkan, S., Tee, A., Zhang, R., Tomasiello, A., De Luca, G. B., Looi, S.-Z., Le, V.-K., Kolt, N., Pan, J., Rodman, E., Drori, J., Fossum, C. J., Muennighoff, N., Jagota, M., Pradeep, R., Fan, H., Eicher, J., Chen, M., Thaman, K., Merrill, W., Firsching, M., Harris, C., Ciobâcă, S., Gross, J., Pandey, R., Gusev, I., Jones, A., Agnihotri, S., Zhelnov, P., Mofayezi, M., Piperski, A., Zhang, D. K., Dobarskyi, K., Leventov, R., Soroko, I., Duersch, J., Taamazyan, V., Ho, A., Ma, W., Held, W., Xian, R., Zebaze, A. R., Mohamed, M., Leser, J. N., Yuan, M. X., Yacar, L., Lengler, J., Olszewska, K., Di Fratta, C., Oliveira, E., Jackson, J. W., Zou, A., Chidambaram, M., Manik, T., Haffenden, H., Stander, D., Dasouqi, A., Shen, A., Golshani, B., Stap, D., Kretov, E., Uzhou, M., Zhidkovskaya, A. B., Winter, N., Rodriguez, M. O., Lauff, R., Wehr, D., Tang, C., Hossain, Z., Phillips, S., Samuele, F., Ekström, F., Hammon, A., Patel, O., Farhidi, F., Medley, G., Mohammadzadeh, F., Peñaflor, M., Kassahun, H., Friedrich, A., Perez, R. H., Pyda, D., Sakal, T., Dhamane, O., Mirabadi, A. K., Hallman, E., Okutsu, K., Battaglia, M., Maghsoudimehrabani, M., Amit, A., Hulbert, D., Pereira, R., Weber, S., Handoko, Peristyy, A., Malina, S., Mehkary, M., Aly, R., Reidegeld, F., Dick, A.-K., Friday, C., Singh, M., Shapourian, H., Kim, W., Costa, M., Gurdogan, H., Kumar, H., Ceconello, C., Zhuang, C., Park, H., Carroll, M., Tawfeek, A. R., Steinerberger, S., Aggarwal, D., Kirchhof, M., Dai, L., Kim, E., Ferret, J., Shah, J., Wang, Y., Yan, M., Burdzy, K., Zhang, L., Franca, A., Pham, D. T., Loh, K. Y., Robinson, J., Jackson, A., Giordano, P., Petersen, P., Cosma, A., Colino, J., White, C., Votava, J., Vinnikov, V., Delaney, E., Spelda, P., Stritecky, V., Shahid, S. M., Mourrat, J.-C., Vetoshkin, L., Sponselee, K., Bacho, R., Yong, Z.-X., de la Rosa, F., Cho, N., Li, X., Malod, G., Weller, O., Albani, G., Lang, L., Laurendeau, J., Kazakov, D., Adesanya, F., Portier, J., Hollom, L., Souza, V., Zhou, Y. A., Degorre, J., Yaln, Y., Obikoya, G. D., Michael Pokorny, R., Bigi, F., Boscá, M. C., Shumar, O., Bacho, K., Recchia, G., Popescu, M., Shulga, N., Tanwie, N. M., Lux, T. C. H., Rank, B., Ni, C., Brooks, M., Yakimchyk, A., Quinn Liu, H., Cavalleri, S., Häggström, O., Verkama, E., Newbould, J., Gundlach, H., Brito-Santana, L., Amaro, B., Vajipey, V., Grover, R., Wang, T., Kratish, Y., Li, W.-D., Gopi, S., Caciolai, A., de Witt, C. S., Hernández-Cámara, P., Rodolà, E., Robins, J., Williamson, D., Raynor, B., Qi, H., Segev, B., Fan, J., Martinson, S., Wang, E. Y., Hausknecht, K., Brenner, M. P., Mao, M., Demian, C., Kassani, P., Zhang, X., Avagian, D., Scipio, E. J., Ragoler, A., Tan, J., Sims, B., Plecnik, R., Kirtland, A., Bodur, O. F., Shinde, D. P., Labrador, Y. C. L., Adoul, Z., Zekry, M., Karakoc, A., Santos, T. C. B., Shamseldeen, S., Karim, L., Liakhovitskaia, A., Resman, N., Farina, N., Gonzalez, J. C., Maayan, G., Anderson, E., De Oliveira Pena, R., Kelley, E., Mariji, H., Pouriamanesh, R., Wu, W., Finocchio, R., Alarab, I., Cole, J., Ferreira, D., Johnson, B., Safdari, M., Dai, L., Arthornthurasuk, S., McAlister, I. C., Moyano, A. J., Pronin, A., Fan, J., Ramirez-Trinidad, A., Malysheva, Y., Pottmaier, D., Taheri, O., Stepanic, S., Perry, S., Askew,

L., Rodrguez, R. A. H., Minissi, A. M. R., Lorena, R., Iyer, K., Fasiludeen, A. A., Clark, R., Ducey, J., Piza, M., Somrak, M., Vergo, E., Qin, J., Borbás, B., Chu, E., Lindsey, J., Jallon, A., McInnis, I. M. J., Chen, E., Semler, A., Gloor, L., Shah, T., Carauleanu, M., Lauer, P., Huy, T. D., Shahrtash, H., Duc, E., Lewark, L., Brown, A., Albanie, S., Weber, B., Vaz, W. S., Clavier, P., Fan, Y., Poesia Reis e Silva, G., Tony Lian, L., Abramovitch, M., Jiang, X., Mendoza, S., Islam, M., Gonzalez, J., Mavroudis, V., Xu, J., Kumar, P., Goswami, L. P., Bugas, D., Heydari, N., Jeanplong, F., Jansen, T., Pinto, A., Apronti, A., Galal, A., Ze-An, N., Singh, A., Jiang, T., of Arc Xavier, J., Agarwal, K. P., Berkani, M., Zhang, G., Du, Z., de Oliveira Junior, B. A., Malishev, D., Remy, N., Hartman, T. D., Tarver, T., Mensah, S., Loume, G. A., Morak, W., Habibi, F., Hoback, S., Cai, W., Gimenez, J., Montecillo, R. G., Łucki, J., Campbell, R., Sharma, A., Meer, K., Gul, S., Gonzalez, D. E., Alapont, X., Hoover, A., Chhablani, G., Vargus, F., Agarwal, A., Jiang, Y., Patil, D., Outevsky, D., Scaria, K. J., Maheshwari, R., Dendane, A., Shukla, P., Cartwright, A., Bogdanov, S., Mündler, N., Möller, S., Arnaboldi, L., Thaman, K., Siddiqi, M. R., Saxena, P., Gupta, H., Fruhauff, T., Sherman, G., Vincze, M., Usawasutsakorn, S., Ler, D., Radhakrishnan, A., Enyekwe, I., Salauddin, S. M., Muzhen, J., Maksapetyan, A., Rossbach, V., Harjadi, C., Bahaloohoreh, M., Sparrow, C., Sidhu, J., Ali, S., Bian, S., Lai, J., Singer, E., Uro, J. L., Bateman, G., Sayed, M., Menshawy, A., Duclosel, D., Bezzi, D., Jain, Y., Aaron, A., Tiryakioglu, M., Siddh, S., Krenek, K., Shah, I. A., Jin, J., Creighton, S., Peskoff, D., EL-Wasif, Z., P, R., Richmond, M., McGowan, J., Patwardhan, T., Sun, H.-Y., Sun, T., Zubić, N., Sala, S., Ebert, S., Kaddour, J., Schottdorf, M., Wang, D., Petruzella, G., Meiburg, A., Medved, T., ElSheikh, A., Hebbar, S. A., Vaquero, L., Yang, X., Poulos, J., Zouhar, V., Bogdanik, S., Zhang, M., Sanz-Ros, J., Anugraha, D., Dai, Y., Nhu, A. N., Wang, X., Demircali, A. A., Jia, Z., Zhou, Y., Wu, J., He, M., Chandok, N., Sinha, A., Luo, G., Le, L., Noyé, M., Perełkiewicz, M., Pantidis, I., Qi, T., Purohit, S. S., Parcalabescu, L., Nguyen, T.-H., Winata, G. I., Ponti, E. M., Li, H., Dhole, K., Park, J., Abbondanza, D., Wang, Y., Nayak, A., Caetano, D. M., Wong, A. A. W. L., del Rio-Chanona, M., Kondor, D., Francois, P., Chalstrey, E., Zsambok, J., Hoyer, D., Reddish, J., Hauser, J., Rodrigo-Ginés, F.-J., Datta, S., Shepherd, M., Kamphuis, T., Zhang, Q., Kim, H., Sun, R., Yao, J., Dernoncourt, F., Krishna, S., Rismanchian, S., Pu, B., Pinto, F., Wang, Y., Shridhar, K., Overholt, K. J., Briia, G., Nguyen, H., Quod Soler Bartomeu, D., Pang, T. C., Wecker, A., Xiong, Y., Li, F., Huber, L. S., Jaeger, J., De Maddalena, R., Lù, X. H., Zhang, Y., Beger, C., Kon, P. T. J., Li, S., Sanker, V., Yin, M., Liang, Y., Zhang, X., Agrawal, A., Yifei, L. S., Zhang, Z., Cai, M., Sonmez, Y., Cozianu, C., Li, C., Slen, A., Yu, S., Park, H. K., Sarti, G., Briański, M., Stolfo, A., Nguyen, T. A., Zhang, M., Perlitz, Y., Hernandez-Orallo, J., Li, R., Shabani, A., Juefei-Xu, F., Dhingra, S., Zohar, O., Nguyen, M. C., Pondaven, A., Yilmaz, A., Zhao, X., Jin, C., Jiang, M., Todoran, S., Han, X., Kreuer, J., Rabern, B., Plassart, A., Maggetti, M., Yap, L., Geirhos, R., Kean, J., Wang, D., Mollaei, S., Sun, C., Yin, Y., Wang, S., Li, R., Chang, Y., Wei, A., Bizeul, A., Wang, X., Arrais, A. O., Mukherjee, K., Chamorro-Padial, J., Liu, J., Qu, X., Guan, J., Bouyamourn, A., Wu, S., Plomecka, M., Chen, J., Tang, M., Deng, J., Subramanian, S., Xi, H., Chen, H., Zhang, W., Ren, Y., Tu, H., Kim, S., Chen, Y., Marjanović, S. V., Ha, J., Luczyna, G., Ma, J. J., Shen, Z., Song, D., Zhang, C. E., Wang, Z., Gendron, G., Xiao, Y., Smucker, L., Weng, E., Lee, K. H., Ye, Z., Ermon, S., Lopez-Miguel, I. D., Knights, T., Gitter, A., Park, N., Wei, B., Chen, H., Pai, K., Elkhanany, A., Lin, H., Siedler, P. D., Fang, J., Mishra, R., Zsolnai-Fehér, K., Jiang, X., Khan, S., Yuan, J., Jain, R. K., Lin, X., Peterson, M., Wang, Z., Malusare, A., Tang, M., Gupta, I., Fosin, I., Kang, T., Dworakowska, B., Matsumoto, K., Zheng, G., Sewuster, G., Villanueva, J. P., Rannev, I., Chernyavsky, I., Chen, J., Banik, D., Racz, B., Dong, W., Wang, J., Bashmal, L., Gonçalves, D. V., Hu, W., Bar, K., Bohdal, O., Patlan, A. S., Dhuliawala, S., Geirhos, C., Wist, J., Kansal, Y., Chen, B., Tire, K., Yücel, A. T., Christof, B., Singla, V., Song, Z., Chen, S., Ge, J., Ponkshe, K., Park, I., Shi, T., Ma, M. Q., Mak, J., Lai, S., Moulin, A., Cheng, Z., Zhu, Z., Zhang, Z., Patil, V., Jha, K., Men, Q., Wu, J., Zhang, T., Vieira, B. H., Aji, A. F., Chung, J.-W., Mahfoud, M., Thi Hoang, H., Sperzel, M., Hao, W., Meding, K., Xu, S., Kostakos, V., Manini, D., Liu, Y., Toukmaji, C., Yu, E., Demircali, A. E., Sun, Z., Dewerpe, I., Qin, H., Pflugfelder, R., Bailey, J., Morris, J., Heilala, V., Rosset, S., Yu, Z., Chen, P. E., Yeo, W., Jain, E., Chigurupati, S., Chernyavsky, J., Reddy, S. P., Venugopalan, S., Batra, H., Park, C. F., Tran, H., Maximiano, G., Zhang, G., Liang, Y., Shiyu, H., Xu, R., Pan, R., Suresh, S., Liu, Z., Gulati, S., Zhang, S., Turchin, P., Bartlett, C. W., Scotese, C. R., Cao, P. M., Wu, B., Karwowski, J., and Scaramuzza, D. A benchmark of expert-level academic questions to assess ai capabilities. *Nature*, 649(8099):1139–1146, January 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-09962-4. URL http://dx.doi.org/10.1038/s41586-025-09962-4.

Ramsey, F. P. Truth and probability. In Braithwaite, R. B. (ed.), *The Foundations of Mathematics and other Logical Essays*, chapter 7, pp. 156–198. McMaster University Archive for the History of Economic Thought, 1926. URL https://EconPapers.repec.org/RePEc:hay:hetcha:ramsey1926.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the

limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, 2023. URL https://arxiv.org/abs/2305.14975.

Trivedi, H., Balasubramanian, N., Khot, T., and Sabharwal, A. Musique: Multi-hop questions via single-hop question composition. *CoRR*, abs/2108.00573, 2021. URL https://arxiv.org/abs/2108.00573.

Turner, A. et al. Avoiding side effects in complex environments, 2021. URL https://arxiv.org/abs/2006.06547.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Wang, Z., Dong, Y., Delalleau, O., Zeng, J., Shen, G., Egert, D., Lin, J. J., Ping, W., Sun, Y., and Chang, M.-W. Helpsteer2-preference: Complementing ratings with preferences, 2024. URL https://arxiv.org/abs/2410.01257.

Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., and Fedus, W. Measuring short-form factuality in large language models, 2024. URL https://arxiv.org/abs/2411.04368.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms, 2024. URL https://arxiv.org/abs/2306.13063.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Ye, J., Gu, J., Zhao, X., Yin, W., and Wang, G. Assessing the creativity of llms in proposing novel solutions to mathematical problems, 2024. URL https://arxiv.org/abs/2410.18336.

Zhou, H., Liu, F., Gu, B., Zou, X., Huang, J., Wu, J., Li, Y., Chen, S. S., Zhou, P., Liu, J., Hua, Y., Mao, C., You, C., Wu, X., Zheng, Y., Clifton, L., Li, Z., Luo, J., and Clifton, D. A. A survey of large language models in medicine: Progress, application, and challenge, 2024. URL https://arxiv.org/abs/2311.05112.

A. Further Confidence Function Theory Clarification

This section clarifies the semantics of the confidence function $c([x],[y])$, its relationship to correctness, and why the experimental protocol provides a valid approximation for evaluating its properties.

A.1. Well-Defined Probability Space

We model each prompt x as inducing a latent random variable over semantic answer classes:

$$Z_x \in \mathcal{Y}_x := \{[y] : y \text{ is a valid generation for } x\},$$

where $[y]$ denotes a semantic equivalence class under $\simeq$.

We assume that correctness is defined at the level of equivalence classes: there exists a (possibly set-valued) subset $\mathcal{Y}_x^\star \subseteq \mathcal{Y}_x$ such that any $[y] \in \mathcal{Y}_x^\star$ constitutes a correct answer. In practice, we assume that all correct answers belong to a single equivalence class, induced by semantic clustering.

The confidence function is then defined as:

$$c([x],[y]) := P(Z_x = [y]),$$

which is a valid probability distribution over $\mathcal{Y}_x$ satisfying:

$$\sum_{[y]\in\mathcal{Y}_x} c([x],[y]) = 1.$$

Under this definition, the probability of answering correctly is:

$$P(\text{correct} \mid x) = \sum_{[y]\in\mathcal{Y}_x^\star} c([x],[y]).$$

When $\mathcal{Y}_x^\star$ contains a single class (as enforced by our semantic clustering), this reduces to:

$$P(\text{correct} \mid x) = \max_{[y]} c([x],[y]) = \bar{c}([x]).$$

Thus, $\bar{c}([x])$ admits a direct probabilistic interpretation as the model's marginal probability of correctness.

A.2. Relation to Estimators

The true confidence function c is not directly observable. Instead, each estimator $\hat{c}(x, y)$ provides a proxy for $c([x], [y])$ based on different signals:

- **Logit-based** and **verbal** estimators approximate $P(\text{correct} \mid x, y)$ directly from the model's internal state.
- **SliCK** approximates $c([x], [y])$ via Monte Carlo sampling:

$$\hat{c}(x, y) \approx \frac{|\{j : y_j \in [y]\}|}{k},$$

  which is a consistent estimator of the underlying class probability under i.i.d. sampling from the model.

Thus, SliCK provides an empirical approximation to the full distribution $c([x], \cdot)$, while output-based estimators provide pointwise estimates conditioned on a single generation.

A.3. Why Structural Properties Are Meaningful

All structural properties evaluated in CoherenceBench follow directly from the probability interpretation of c:

- **Normalization** enforces that $c([x], \cdot)$ is a valid probability distribution over answer classes.
- **Conjunction Consistency** follows from the product rule:

$$P(A \cap B) = P(A)\, P(B \mid A),$$

  where A and B correspond to correctness events for sub-questions.
- **Entailment Monotonicity** follows from set inclusion: if correctness on x implies correctness on x', then

$$P(\text{correct on } x) \leq P(\text{correct on } x').$$

Violations of these properties therefore indicate that the estimator cannot be interpreted as a coherent probability distribution over semantic outcomes.

A.4. Why the Experimental Protocol Is Valid

Our experimental design approximates these properties using sampled rollouts and semantic clustering:

- Sampling ($k = 16$) provides an empirical approximation to the latent distribution over equivalence classes.
- LLM-based clustering induces a partition of the output space that operationalizes $\simeq$.
- Aggregation over rollouts approximates expectations under c.

Importantly, the same protocol is applied uniformly across all estimators. Thus, while the approximation may be imperfect, differences in measured properties reflect differences in estimator behavior rather than artifacts of the evaluation pipeline.

Table 1. Models evaluated in this work. **SQ** = SimpleQA, **MQ** = MuSiQue, **PR** = ParaRel.

Model	OpenRouter ID	Params	Type	Datasets
LLaMA-3.2-1B	`meta-llama/llama-3.2-1b-instruct`	1B	Instruct	SQ, PR
LLaMA-3.2-3B	`meta-llama/llama-3.2-3b-instruct`	3B	Instruct	SQ, MQ, PR
LLaMA-3-8B	`meta-llama/llama-3-8b-instruct`	8B	Instruct	SQ, MQ, PR
LLaMA-3.3-70B	`meta-llama/llama-3.3-70b-instruct`	70B	Instruct	SQ, MQ, PR
Gemma-3n-E4B	`google/gemma-3n-e4b-it`	~4B	Instruct	SQ, PR
Gemma-3-27B	`google/gemma-3-27b-it`	27B	Instruct	SQ, MQ, PR
Gemini-2.5-FL	`google/gemini-2.5-flash-lite`	—	Instruct	SQ, MQ, PR
Qwen3-14B	`qwen/qwen3-14b`	14B	Instruct	SQ, MQ, PR
DeepSeek-R1	`deepseek/deepseek-r1-distill-qwen-32b`	32B	Chain-of-thought	SQ, MQ, PR

A.5. Interpretation of Results

Under this framework:

- Output-based estimators that assign high confidence to many mutually exclusive classes violate normalization and cannot correspond to any underlying probability distribution.
- SliCK, by approximating the full distribution over classes, enables direct evaluation of structural properties and exposes violations arising from the model itself rather than the estimator.
- Faithfulness metrics test whether $\hat{c}$ respects the equivalence relation defining the domain of c; violations imply that $\hat{c}$ is not a well-defined function over semantic classes.

Overall, the experiments should be interpreted as testing whether a given estimator induces a function that can be consistently interpreted as a probability distribution over semantic outcomes. Structural violations therefore reflect a breakdown of this interpretation.

A. Experimental Setup Details

A.1. Models

We evaluate nine publicly available instruction-tuned and reasoning models accessed via the OpenRouter API. Table 1 lists each model together with its parameter count, architecture family, and the datasets on which results are reported.

A.2. Datasets

We use three benchmarks from the SliCK evaluation suite:

- **SimpleQA**: Factual single-hop questions with unambiguous answers. We sample 200 questions (seed 42).
- **MuSiQue**: Multi-hop reasoning questions over Wikipedia. Each of the 200 sampled questions is decomposed into a *full* version (all hops) and a *hop-1* version (first hop only), yielding 120 conversations total. The pair structure is used for the Entailment Monotonicity benchmark.

Table 2. Hyperparameters used across all experiments.

Stage	Parameter	Value
4*Generation	Rollouts per question (k)	16
	Sampling temperature	1.0
	Max. output tokens	8,192
	Random seed	42
3*Summarization	Judge model	`qwen/qwen3-30b-a3b`
	Sampling temperature	0.0
	Max. output tokens	4,096
3*Verbal confidence	Judge model	`qwen/qwen3-30b-a3b`
	Sampling temperature	0.0
	Max. output tokens	4,096
API	Concurrent requests	15

- **ParaRel**: Factual probes expressed as semantically equivalent rephrasings of the same underlying relation. We sample 42 conversations covering 5 unique facts, used exclusively for the Prompt Semantic Invariance benchmark.

A.3. Hyperparameters

B. RLE Training Details

B.1. Data Preparation

The QA pool is built once from the frozen base model before training. For each question the base model generates an answer (greedy, max 256 tokens), which is then judged for correctness by an OpenRouter LLM judge (`qwen/qwen3-14b`, temperature 0) using a structured prompt returning {`grade: A/B/C`}. ParaRel uses exact-normalised string match instead. The pool is cached to disk; subsequent runs load from cache.

B.2. Hyperparameters

Table 3. RLE training hyperparameters.

Parameter	Value
Base model	Qwen2.5-7B-Instruct
LoRA rank r	16
LoRA α	32
LoRA dropout	0.05
Target modules	q, k, v, o, gate, up, down projections
Trainable parameters	≈10M / 7.6B (0.13%)
GRPO group size N	8 rollouts per episode
Episodes per step	4
Total steps	1,000
Brier weight λ	0.5
KL coefficient β	0.05
Learning rate	2×10^{-5}
Gradient clip	1.0
Optimiser	AdamW ($w_d = 0.01$)
Price max tokens	16
Answer max tokens	256
Price temperature	1.0
Answer temperature	0.0 (greedy)
Training pool	800 SimpleQA + 400 MuSiQue + 400 ParaRel
Eval pool (held-out)	200 SimpleQA + 100 MuSiQue + 100 ParaRel
Training seed	42
Eval seed	99
Wall time	≈11 hours (single GPU)

B.3. Contract Prompt

Each confidence query is framed as a derivatives contract to ground the Dutch-book fair-price interpretation:

RLE Contract Prompt

```
A derivatives counterparty will buy or
sell at your stated price a contract
paying $1 if the following answer is
correct for the given question.
Question: {question}
Answer: {answer}
State your probability price as a single
decimal between 0.000 and 1.000.
Price:
```

B.4. Exploitability LP Formulation

Each coherence set $\mathcal{S}$ defines a convex polytope; the LP variables are $[\mathbf{u}, \boldsymbol{\omega}] \in \mathbb{R}^{2n}$ where $u_i \geq |c_i - \omega_i|$ is linearised via two inequality constraints per coordinate.

[noitemsep,topsep=2pt]

- **Normalization**: $\sum_i \omega_i = 1, \omega_i \geq 0$. Closed form for n=2: $|c_0 + c_1 - 1|$.
- **Conjunction**: Fréchet polytope with $\omega_{AB} \leq \omega_A$, $\omega_{AB} \leq \omega_B, \omega_A + \omega_B - \omega_{AB} \leq 1$.
- **Paraphrase**: Isotonic equality $\omega_i = \omega_j$; closed form $\sum_i |c_i - \bar{c}|$.
- **Entailment**: Isotonic cone $\omega_0 \leq \omega_1 \leq \cdots$. Closed form for n=2: $\max(0, c_0 - c_1)$.

All LPs are solved via HiGHS (Huangfu & Hall, 2018) through `scipy.optimize.linprog`.

C. Full Benchmark Results

Tables **??**–5 report per-model scores for all benchmarks. Both the SliCK (frequency-based, **F**) and Verbal (**V**) estimators are shown where applicable. Where data permit, values are accompanied by a ± standard deviation: per-item s.d. for deviation-based metrics (Norm, Gen-SI, CC), mean within-fact s.d. for Prompt SI, and binomial s.e. for Ent-Mono violation rate. "—" denotes that the benchmark is not applicable to that dataset or that the model did not produce any correct answers (precluding AUROC computation).

D. Prompts

D.1. Summarization Prompt

The following prompt is sent to the judge model (`qwen/qwen3-30b-a3b`) once per rollout to cluster model responses into semantic equivalence classes. Ex-

Table 4. ParaRel results. SI = Prompt Semantic Invariance (mean confidence spread across semantically equivalent rephrasings ± mean within-fact standard deviation). Lower is better.

Model	SI_F↓	SI_V↓
LLaMA-3.2-1B	0.252 ± 0.074	0.207 ± 0.076
LLaMA-3.2-3B	0.280 ± 0.137	0.545 ± 0.257
LLaMA-3-8B	0.498 ± 0.182	0.314 ± 0.115
LLaMA-3.3-70B	0.625 ± 0.207	0.642 ± 0.216
Gemma-3n-E4B	0.677 ± 0.272	0.123 ± 0.049
Gemma-3-27B	0.541 ± 0.195	0.170 ± 0.056
Gemini-2.5-FL	0.584 ± 0.197	0.287 ± 0.099
Qwen3-14B	0.522 ± 0.185	0.363 ± 0.130
DeepSeek-R1-32B	0.323 ± 0.116	0.311 ± 0.121

Table 5. MuSiQue results. CC = Conjunction Consistency SliCK (mean absolute deviation ± per-question s.d.); Ent-Mono = Entailment Monotonicity violation rate ± binomial s.e. Lower is better for both. "—" indicates no MuSiQue data for that model.

Model	CC_F↓	Ent-Mono↓
LLaMA-3.2-1B	—	—
LLaMA-3.2-3B	0.320 ± 0.344	0.143 ± 0.094
LLaMA-3-8B	0.244 ± 0.214	0.400 ± 0.083
LLaMA-3.3-70B	0.286 ± 0.235	0.282 ± 0.072
Gemma-3n-E4B	—	—
Gemma-3-27B	0.273 ± 0.225	0.225 ± 0.066
Gemini-2.5-FL	0.247 ± 0.240	0.237 ± 0.069
Qwen3-14B	—	—
DeepSeek-R1-32B	—	—

isting cluster labels from prior rollouts are accumulated and passed in the {summaries} field so that later rollouts reuse established clusters.

Summarization Prompt

```
You are an expert data annotation
assistant.  You will be given:
1.  A question.
2.  A response to the question.
3.  Several summaries the final answer
could fall under.
Categorize the final answer as one of
the provided summaries, or create a
new one if none fits.  New summaries
must be as short and specific as
possible---never a full sentence.  For
numeric answers, always include the
number.
Do not recompute or verify the answer
yourself.
Question:  {question}
Response:  {response}
Summaries:  {summaries}
Output JSON only:
{"reasoning":  <reasoning>, "summary":
<summary>}
If the model says it does not know,
return "Unknown".
```

D.2. Verbal Confidence Prompt

After each rollout is generated, the message below is appended to the conversation and sent back to the *same model under evaluation* (not the judge) at temperature 0 to elicit a scalar confidence score.

Verbal Confidence Prompt

```
Provide the probability that your guess
is correct.  Give ONLY the probability,
no other words or explanation.
For example:
Probability:  <a number between 0.0 and
1.0, nothing else>
```

The response is parsed by attempting (in order): a direct `float()` cast; regex matching for `Probability:\s*([0-9.]+)`; and extraction of the first numeric token. Values outside $[0, 1]$ or unparseable responses are recorded as `None` and excluded from all downstream aggregation.