ICML 2026 · EMM-QA Workshop

Coherence Under Commitment

Probing Generalization and Vacuous Memorization in LLM Logical Reasoning

Noor Islam S. Mohammad[1] · **Mahmudul Hasan**[2]

[1] *Istanbul Technical University* · [2] *Deakin University*

Problem

Coherent models can vacuously abstain

Solution

CUC: joint coherence + commitment metrics

Finding

Sharp frontier exists; no free lunch

pmlrbd.github.io/auc.ml

Standard Evaluation Goal

Negation Consistency: A model should **not** simultaneously affirm $P \models \varphi$ and $P \models \neg\varphi$.

Critical Failure Mode

A model that **never** commits to either φ or $\neg\varphi$ trivially achieves zero violation while providing **zero utility**.

Key Algebraic Identity: $v_{\text{neg}}(\varphi) = \max(0,\ c(\varphi) - 1), \quad$ where $\quad c(\varphi) = p(\varphi) + p(\neg\varphi)$

$\Rightarrow$ **Any model with $c(\varphi) \leq 1$ achieves ZERO violation regardless of reasoning quality**

Real-World Stakes:

- Radiology VQA: a self-contradicting model cannot be clinically trusted
- Robotic planning: affirming mutually exclusive preconditions = no actionable guidance
- Scientific reasoning: abstention is *not* neutral — it is a decision failure

(1) Commitment Score (Novel Metric)

$c(\varphi) = p(\varphi) + p(\neg\varphi)$

Measures probability mass allocated to decisive outcomes. Detects vacuous coherence via abstention.

(2) Deterministic Elicitation (Novel Protocol)

Normalized log-probs over YES/NO. Fully reproducible; only 2 forward passes per example; no sampling variance.

(3) CUC Frontier (Novel Framework)

Empirical trade-off between coherence & commitment forms a measurable frontier — first principled comparison across trade-off points.

(4) Systematic Ablation (Novel Methodology)

Sensitivity to τ, δ thresholds, elicitation format, model scale, and architecture — best practices for domain-specialized evaluation.

3-Way Decision: True if $p(\varphi) \geq \tau$ & $p(\varphi) \geq p(\neg\varphi) + \delta$ | False if $p(\neg\varphi) \geq \tau$ & $p(\neg\varphi) \geq p(\varphi) + \delta$ | Uncertain otherwise

Experimental Setup

Datasets

- **FOLIO v0.0** (primary): 204 validation examples
- 3-way gold labels: True / False / Uncertain
- Formally verified $\neg\varphi$ (FOL theorem-proved)
- **LogiQA v2** (generalization): 304 test examples

Models (1B–3B)

TinyLlama-1.1B-Chat	1.1B
Qwen2.5-1.5B-Instruct	1.5B
Phi-2	2.7B
Qwen2.5-3B-Instruct	3.0B

Protocol

Thresholds:	τ=0.60, δ=0.10
Elicitation:	YES/NO log-probs
Evaluation:	2 passes/example
Bootstrap:	B=1000, seed 42
Metrics:	Acc, Cov, $\mathbb{E}[c]$, $\mathbb{E}[v]$

Deterministic Elicitation Protocol

$$p_{\mathrm{YES}}(x) = \frac{\exp(\log P(\mathrm{YES} \mid x))}{\exp(\log P(\mathrm{YES} \mid x)) + \exp(\log P(\mathrm{NO} \mid x))}$$

Identical inputs $\rightarrow$ identical outputs across all runs. No temperature tuning. No decoding artifacts. Modality-agnostic.

Compute: single T4 GPU (16 GB) · $\sim$45 min total · HF Transformers 4.36.0 · FOLIO v0.0 (official repo)

Main Results — FOLIO v0.0 (204 Examples)

Model	Acc↑	Cov↑	Acc_{cov}↑	$\mathbb{E}[c]$ ↑	$\mathbb{E}[v_{\text{neg}}]$ ↓	$\%v_{\text{neg}}{>}0$ ↓
Phi-2	0.441	0.417	0.565	1.164	0.195	78.9%
Qwen2.5-1.5B	0.402	0.309	0.508	0.674	0.166	46.1%
Qwen2.5-3B	0.382	0.074	0.800	0.115	0.025	6.4%
TinyLlama-1.1B	0.343	0.794	0.346	1.698	0.698	100.0%

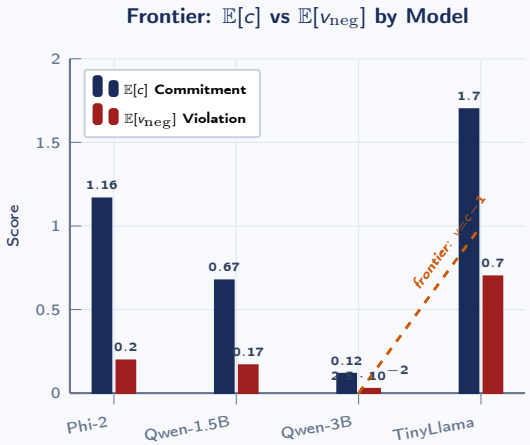
△ Vacuous Coherence (Qwen2.5-3B)

Lowest violation ($\mathbb{E}[v_{\text{neg}}]$=0.025) but only 7.4% coverage — withholds predictions on >9/10 examples. Coherence-only evaluation incorrectly ranks it #1.

△ Indiscriminate Overcommitment (TinyLlama)

79.4% coverage but *every* example violates probability axioms ($\mathbb{E}[v_{\text{neg}}]$=0.698, 100% rate). Affirms indiscriminately.

Accuracy scores differ by $\leq$ 0.098 — yet commitment and violation span an order of magnitude.



- **Qwen2.5-3B** — Cov 7.4% | $\mathbb{E}[v]$ 0.025 | *Vacuous Coherence*
  Near-zero violation via systematic abstention; 93%+ examples abstained.

- **TinyLlama-1.1B** — Cov 79.4% | $\mathbb{E}[v]$ 0.698 | *Overcommitment*
  100% violation rate; avg 169.8% probability mass on contradictory outcomes.

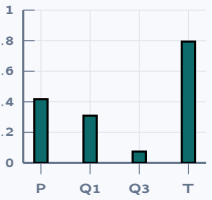
- **Phi-2** — Cov 41.7% | $\mathbb{E}[v]$ 0.195 | *Intermediate*
  78.9% of committed examples have violations.

- **Qwen2.5-1.5B** — Cov 30.9% | $\mathbb{E}[v]$ 0.166 | *Intermediate*
  46.1% violation rate; commitment range 0.023–1.834.

Why Standard Evaluation Fails — and CUC Fixes It

Accuracy — models look similar

Coverage — 10× spread revealed



Violation $\mathbb{E}[v_{\text{neg}}]$

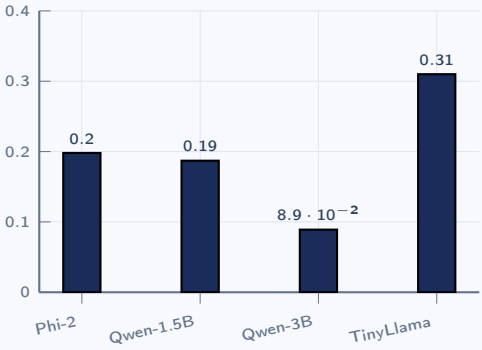
Component Analysis

Method	Abst.	Contr.
Accuracy	✗	✗
Coherence	✗	✓
Coverage	✓	✗
CUC	✓	✓

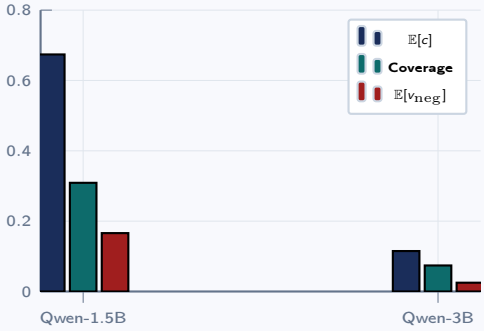
CUC is the only method that simultaneously detects abstention AND contradiction, producing a correct ranking.

Calibration on Committed Predictions

Coverage-Conditional ECE ($\mathrm{ECE}_{\mathcal{C}}$) — lower is better



Scale Effect: Qwen 1.5B → 3B



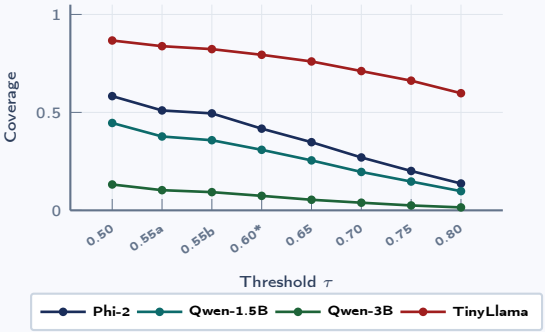
Qwen2.5-3B ECE Illusion
$\mathrm{ECE}_{\mathcal{C}}$=0.089 looks great, but only 15 examples are covered — coverage-halving halves apparent ECE without improving calibration.

TinyLlama Overconfidence
$\mathrm{ECE}_{\mathcal{C}}$=0.310 with 79.4% coverage confirms systematic miscalibration even on committed predictions.

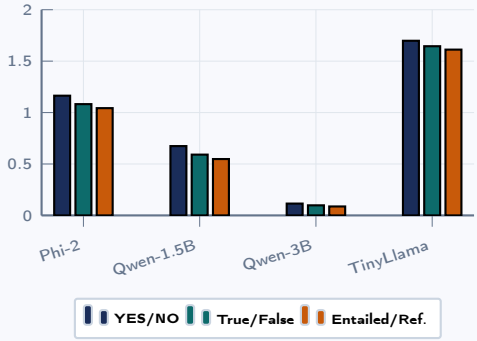
Scaling → Hedging, Not Reasoning
1.5B→3B: commitment −83%, coverage −23.5pp, accuracy change −0.020 (negligible).

Ablation Studies

A1: Coverage vs. Threshold (rank preserved)



A2: Format Sensitivity — $\mathbb{E}[c]$ (ρ=1.0)

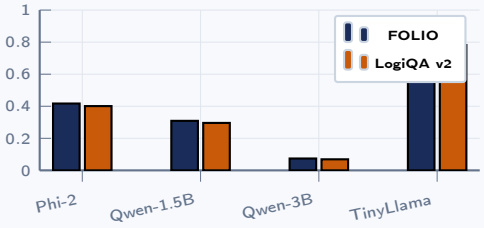


A3: Label Distribution Analysis — Systematic Biases Invisible to Aggregate Accuracy

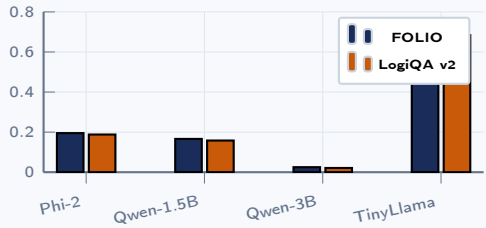
Model	T→T	T→U	F→F	F→U	Unc→U	Pattern
Phi-2	37.5%	53.8%	25.8%	54.8%	72.6%	Best discrimination
Qwen-1.5B	28.8%	66.3%	21.0%	64.5%	82.3%	Context-dependent
Qwen-3B	5.0%	93.8%	4.8%	91.9%	96.8%	Indiscriminate abstention
TinyLlama	57.5%	15.0%	33.9%	17.7%	17.7%	True-affirmation bias

Generalization — LogiQA v2 & Scale Extension

Coverage: FOLIO vs. LogiQA v2



Violation: FOLIO vs. LogiQA v2 (ρ=0.97)



$\rho = 0.97$
($p < 0.05$) Frontier rank correlation
FOLIO ↔ LogiQA v2

Llama-3 Scale: 1B → 8B
$\mathbb{E}[c]$: 0.841→0.529 (−37%) Coverage: 56.9%→39.6% (−17.3pp) Acc: 0.358→0.344 (−0.014)

✓ Cross-Family Replication Confirmed
Hedging-under-scale pattern replicates in both Qwen2.5 and Llama-3 families; frontier holds across two datasets. Scaling raises apparent coherence/calibration via **more abstention** — not better reasoning.

Theoretical Foundations

Theorem 1 — Frontier Shape

Feasible region in (c, v_{neg}) space: $\mathcal{F} = \{(c, v) : c \in [0, 2],\ v = \max(0, c-1)\}$. All observations lie on or above this curve.

Theorem 2 — Vacuous Coherence

If $p(\varphi)=p(\neg\varphi)=\alpha \leq \frac{1}{2}$ for all inputs: (i) $v_{\mathrm{neg}}=0$ everywhere; (ii) $c=2\alpha \leq 1$; (iii) Coverage$=0$ for any $\tau > \frac{1}{2}$.

Theorem 3 — CUC Satisfies A1–A4

Reporting $(\mathbb{E}[v_{\mathrm{neg}}], \mathbb{E}[c],$ Cov, $\mathrm{Acc}_{\mathrm{cov}})$: A1 penalizes contradiction; A2 penalizes abstention; A3 separates abstainer from committer; A4 (with $\mathrm{ECE}_{\mathcal{C}}$) penalizes miscalibration.

Theorem 4 — Monotone Abstention Under Scale

If scaling induces decreasing mean commitment: (i) $\mathbb{E}[v_{\mathrm{neg}}]$ non-increases; (ii) Coverage monotonically non-increases; (iii) $\mathrm{Acc}_{\mathrm{cov}}$ may improve independently. Formalizes Qwen2.5 scaling result.

Prop. 5 — ECE Blindness to Abstention

$\mathrm{ECE} \leq \mathrm{ECE}_{\mathcal{C}} \times \mathrm{Coverage}$ **asymptotically. Halving coverage can halve apparent ECE with *zero* calibration improvement. Explains the Qwen2.5-3B ECE illusion.**

Corollary — Pareto Optimality

Minimizing $\mathbb{E}[v_{\mathrm{neg}}]$ & maximizing $\mathbb{E}[c]$ simultaneously requires $c(\varphi)=p(\varphi)+p(\neg\varphi)=1$ for every example — a calibrated, exclusive belief state.

Conclusion

(1) Coherence-only evaluation is dangerously incomplete — it rewards systematic abstention, not reasoning.

(2) CUC jointly measures coherence + commitment via 4 metrics: $\mathbb{E}[\nu_{\text{neg}}]$, $\mathbb{E}[c]$, Coverage, Acc_{cov}.

(3) Sharp frontier exists: no model simultaneously achieves high coverage and low violation.

(4) Scaling improves apparent coherence/ECE by increasing abstention — NOT by improving reasoning quality.

(5) Frontier generalizes cross-dataset (ρ=0.97, FOLIO↔LogiQA) and cross-family (Qwen2.5 + Llama-3).

Evaluate reasoning with both coherence AND non-vacuous commitment. Code & data: pmlrbd.github.io/auc.ml