COMMITMENT-AWARE AXIOMATIC COHERENCE: MEASURING NON-VACUOUS CONSISTENCY IN LLM LOGICAL REASONING

Md Muntaqim Meherab[1]
meherab2305101354@diu.edu.bd
[1]Department of Computer Science and Engineering,
Daffodil International University, Bangladesh

ABSTRACT

Large language models (LLMs) are increasingly used for logical tasks, yet they frequently exhibit contradictions across closely related queries. A natural response is to measure *logical coherence* by checking axioms such as negation consistency. However, we show that coherence can be *vacuous*: a model can appear consistent by refusing to commit to either a statement or its negation. We propose **commitment-aware axiomatic coherence**, a lightweight evaluation protocol that complements a standard negation-coherence check with a *commitment* score measuring how much probability mass the model assigns to *entailed* vs. *refuted* outcomes (as opposed to abstention/uncertainty). Using a deterministic log-probability elicitation procedure (YES/NO) and a simple 3-way decision rule (True/False/Uncertain), we evaluate four open LLMs on the public FOLIO v0.0 validation split. Results reveal a clear frontier: some models achieve low contradiction rates primarily by abstaining (low coverage), while others achieve high coverage at the cost of pervasive negation-coherence violations. Our findings argue that reliable logical reasoning evaluation requires reporting both coherence and non-vacuous commitment, not coherence alone.The project is available at https://meherabb.github.io/Commitment/

1 INTRODUCTION

LLMs have become competitive on a range of reasoning tasks, aided by prompting techniques such as chain-of-thought and self-consistency (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022). At the same time, it is now broadly recognized that strong surface performance does not guarantee internal reliability: models can be overconfident, poorly calibrated, or inconsistent across semantically related questions (Guo et al., 2017; Lin et al., 2022; Turpin et al., 2023; Ji et al., 2023). For logical reasoning, inconsistency is not merely cosmetic: if a system alternates between endorsing φ and not endorsing $\neg\varphi$ under the same premises, it undermines downstream uses in science, medicine, and law, where contradictions can be operationally harmful (Ji et al., 2023; Lin et al., 2022).

A tempting evaluation approach is to test whether an LLM respects basic axioms. One example is a negation-consistency constraint: under fixed premises P, the model should not simultaneously assign firm belief to "$P \models \varphi$" and to "$P \models \neg\varphi$". A subtle failure mode, however, is that such axioms can be satisfied trivially by *abstention*. If a model systematically avoids committing to either entailment or refutation, it may incur few explicit contradictions while still being unhelpful as a reasoner.

This paper isolates that failure mode and proposes a simple remedy:

- We formulate a **commitment** score alongside a standard negation-coherence violation score.
- We provide a deterministic black-box elicitation method using normalized YES/NO log-probabilities.
- We empirically demonstrate a **coherence-commitment frontier**: low contradictions can coincide with low usefulness because of abstention.

2 Related Work

LLM reasoning and benchmarks. Progress in reasoning has been catalyzed by prompting and decoding methods (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022) and by benchmark suites such as BIG-bench (Srivastava et al., 2022). Logical reasoning datasets include LogiQA (Liu et al., 2020), rule-based synthetic tasks and the RuleTaker family (Clark et al., 2020), and proof-centric resources such as ProofWriter (Tafjord et al., 2021). FOLIO targets first-order logic reasoning with formal verification support (Han et al., 2024).

Consistency, hallucination, and faithfulness. Contradictions and hallucinations have been studied in generation (Ji et al., 2023) and through sampling-based consistency checks (Manakul et al., 2023). Chain-of-thought rationales can be unfaithful explanations of a model's true decision process (Turpin et al., 2023), motivating evaluation methods that do not depend on natural-language rationales.

Calibration and uncertainty. Expected calibration error (ECE) and related reliability diagnostics are widely used to quantify miscalibration (Naeini et al., 2015; Guo et al., 2017). For LMs, confidence elicitation and self-evaluation have shown that probability estimates can be informative when elicited appropriately (Kadavath et al., 2022; Zhang et al., 2024). Our work is aligned with this direction but focuses on a logic-specific failure mode: *vacuous coherence* via abstention.

3 Problem Setup

Let P be a set of premises expressed in natural language, and let φ be a natural-language conclusion. We evaluate an LLM as a stochastic mapping from prompts to outputs, but we treat it operationally as a belief function over entailment judgments.

We ask two questions under the same P:

$$Q_\varphi: \quad \text{“Is } \varphi \text{ logically entailed by } P\text{?”} \tag{1}$$

$$Q_{\neg\varphi}: \quad \text{“Is } \neg\varphi \text{ logically entailed by } P\text{?”} \tag{2}$$

We enforce a restricted response format; the model must answer YES or NO. Let (φ) denote the model's probability of answering YES to Q_φ, and similarly $(\neg\varphi)$ for $Q_{\neg\varphi}$.

Negation-coherence violation. A basic coherence desideratum is that the model should not simultaneously endorse entailment of φ and entailment of $\neg\varphi$. We quantify the extent of violation by

$$v_{\text{neg}}(\varphi) \;=\; \max\big(0,\; (\varphi) + (\neg\varphi) - 1\big). \tag{3}$$

Commitment. Crucially, low violation does not imply useful reasoning if the model abstains. We therefore define a commitment score:

$$c(\varphi) \;=\; (\varphi) + (\neg\varphi). \tag{4}$$

Intuitively, $c(\varphi)$ measures how much mass the model allocates to the two decisive outcomes (entailed/refuted). If $c(\varphi)$ is small, the model is effectively treating the query as unknown.

Usefulness via a 3-way decision rule. To connect commitment to decision quality, we map $((\varphi), (\neg\varphi))$ to a 3-way label:

$$\widehat{y}(\varphi) = \begin{cases} \textsc{True} & \text{if } (\varphi) \geq \tau \text{ and } (\varphi) \geq (\neg\varphi) + \delta, \\ \textsc{False} & \text{if } (\neg\varphi) \geq \tau \text{ and } (\neg\varphi) \geq (\varphi) + \delta, \\ \textsc{Uncertain} & \text{otherwise.} \end{cases}$$

We then report: (i) overall 3-way accuracy, (ii) *coverage* (fraction not labeled Uncertain), (iii) accuracy on the committed subset.

4 PROBABILITY ELICITATION

We estimate $(\cdot)$ deterministically using normalized log-probabilities. For each prompt x, we compute

$$\log(\text{YES} \mid x), \qquad \log(\text{NO} \mid x),$$

and normalize over the two options:

$$(x) = \frac{\exp(\log(\text{YES} \mid x))}{\exp(\log(\text{YES} \mid x)) + \exp(\log(\text{NO} \mid x))}. \tag{5}$$

When YES/NO is tokenized into multiple tokens, we sum token-level conditional log-probabilities, matching the implementation in our released notebook.

This approach aligns with a growing body of evidence that LMs can provide meaningful uncertainty estimates when queried in the right format (Kadavath et al., 2022; Guo et al., 2017).

5 EXPERIMENTAL SETUP

Dataset. We evaluate on the public FOLIO v0.0 validation JSONL file (downloaded from the official repository), which contains 204 labeled examples of premise sets, conclusions, and 3-way labels (TRUE/FALSE/UNCERTAIN). FOLIO as a dataset family targets first-order logical reasoning with formal verification (Han et al., 2024).

Models. We evaluate four open models spanning $\sim$1B–3B parameters and different instruction-tuning regimes: TinyLlama-1.1B-Chat, Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, and Phi-2. (We attempted a 7B model under 4-bit quantization, but it exceeded memory limits in our Colab setting; the evaluation code is robust to skipping such models.)

Prompts. We use a stable two-query protocol with a short system message requiring a single-token response (YES or NO). Full prompts are listed in Appendix A.

Thresholds and metrics. We set $(\tau, \delta) = (0.60, 0.10)$ for the 3-way decision rule. We report mean commitment equation 7, mean negation violation equation 6, fraction of examples with $v_{\text{neg}}(\varphi) > 0$, and accuracy/coverage. We compute 95% bootstrap confidence intervals (1,000 resamples) for all reported scalar metrics.

6 RESULTS

Table 1 reports model-level metrics with 95% bootstrap confidence intervals. To make the trade-offs easier to see, Figure 1 plots the per-example coherence-commitment frontier for each model, Figure 2 aggregates the same quantities at the model level, and Figure 3 shows reliability on the committed (non-UNCERTAIN) subset.

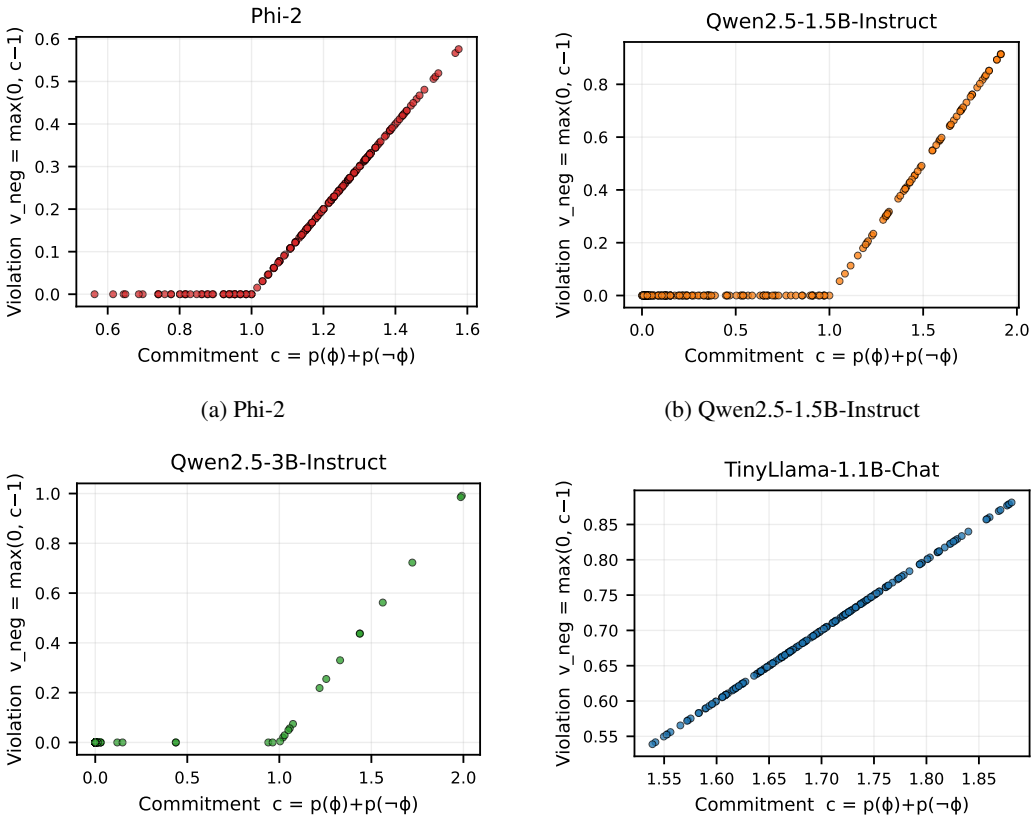
6.1 A COHERENCE-COMMITMENT FRONTIER

The results reveal a sharp trade-off between non-vacuous commitment and negation coherence (Figures 1 and 2). At a high level, models that commit more often (higher coverage and higher $\mathbb{E}[c]$) also incur more negation-coherence violations, while the most "coherent" models can achieve low violations simply by declining to take a stance.

Vacuous coherence by abstention. Qwen2.5-3B exhibits very low violation ($\mathbb{E}[v_{\text{neg}}] \approx 0.025$) but also extremely low commitment ($\mathbb{E}[c] \approx 0.115$) and low coverage ($\approx 0.074$). When the model does commit, it is often correct ($\text{Acc}_{\text{cov}} \approx 0.80$), but it avoids making decisions most of the time. This is the signature of *vacuous coherence*: coherence metrics look favorable because the model rarely assigns substantial mass to either entailment or refutation.

Model	n	**Acc**↑	**Cov**↑	**Acc$_{\text{cov}}$**↑	$\mathbb{E}[c]$↑	$\mathbb{E}[v_{\text{neg}}]$↓
Phi-2	204	0.441 [0.373, 0.510]	0.417 [0.348, 0.480]	0.565 [0.458, 0.667]	1.164 [1.136, 1.190]	0.195 [0.175, 0.215]
Qwen2.5-1.5B	204	0.402 [0.338, 0.471]	0.309 [0.250, 0.373]	0.508 [0.386, 0.629]	0.674 [0.590, 0.762]	0.166 [0.129, 0.205]
Qwen2.5-3B	204	0.382 [0.319, 0.451]	0.074 [0.039, 0.108]	0.800 [0.571, 1.000]	0.115 [0.067, 0.167]	0.025 [0.010, 0.044]
TinyLlama-1.1B	204	0.343 [0.279, 0.412]	0.794 [0.735, 0.848]	0.346 [0.275, 0.422]	1.698 [1.687, 1.709]	0.698 [0.687, 0.709]

Table 1: Commitment-aware coherence results on FOLIO v0.0 validation (204 examples). Acc: overall 3-way accuracy. Cov: coverage (fraction predicted as true/false). Acc$_{\text{cov}}$: accuracy on covered examples. $\mathbb{E}[c]$ is mean commitment; $\mathbb{E}[v_{\text{neg}}]$ is mean negation-coherence violation. Brackets denote 95% bootstrap CIs.



(a) Phi-2

(b) Qwen2.5-1.5B-Instruct

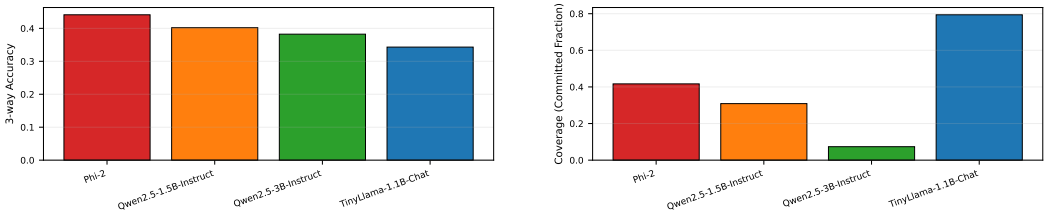
(c) Qwen2.5-3B-Instruct

(d) TinyLlama-1.1B-Chat

Figure 1: **Per-example coherence—commitment frontier.** Each point corresponds to one example; the x-axis is commitment $c = p(\varphi) + p(\neg\varphi)$ and the y-axis is violation $v_{\text{neg}} = \max(0, c - 1)$.

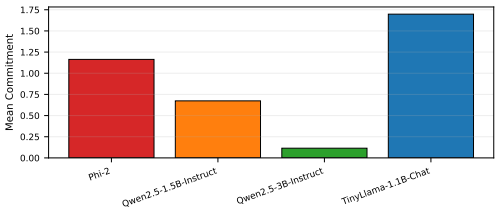
High coverage with pervasive contradictions. TinyLlama shows the opposite regime: high coverage (≈ 0.79) and very high commitment ($\mathbb{E}[c] \approx 1.70$), but massive negation violations ($\mathbb{E}[v_{\text{neg}}] \approx 0.70$) and $v_{\text{neg}} > 0$ on essentially all examples. In practice, this means the model frequently assigns high probability to *both* $P \models \varphi$ and $P \models \neg\varphi$, producing contradictions at scale.

Middle regimes. Phi-2 and Qwen2.5-1.5B occupy intermediate positions, with moderate coverage and non-trivial violation rates. They illustrate that the frontier is not an artifact of a single model

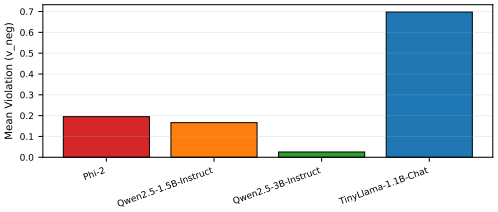


(a) 3-way accuracy

(b) Coverage (non-UNCERTAIN)



(c) Mean commitment



(d) Mean negation violation

Figure 2: **Aggregate comparisons.** Accuracy/coverage quantifies usefulness; commitment and negation-violation expose abstention vs. contradiction regimes.

family: even among relatively small open models, increasing decisiveness tends to come with a measurable increase in contradiction risk.

6.2 WHY COMMITMENT MATTERS FOR EVALUATION

A purely coherence-driven evaluation would conclude that Qwen2.5-3B is the "best" model among those tested, because it exhibits the lowest negation violation. Our commitment-aware view produces a different conclusion: Qwen2.5-3B is coherent largely because it abstains. If the downstream application requires decisive entailment/refutation judgments (as in diagnosis or legal reasoning), that behavior is not acceptable, even if it reduces contradictions.

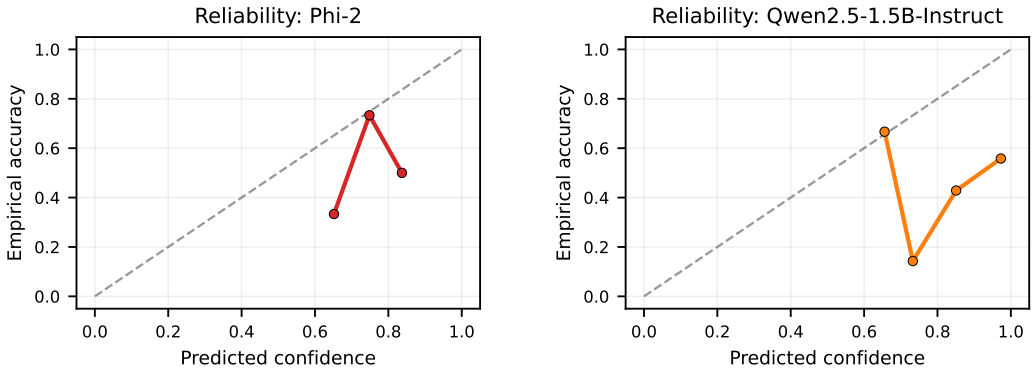
Figure 2 makes this tension explicit: the lowest violations coincide with the lowest coverage. Figure 3 adds a second cautionary note: even when a model does commit, confidence in the covered subset can remain imperfectly calibrated. This mirrors concerns in the broader LLM reliability literature: systems that appear safe under narrow metrics can fail under interaction, distribution shift, or alternative elicitation strategies (Lin et al., 2022; Ji et al., 2023; Turpin et al., 2023). Commitment is a compact way to expose the abstention failure mode in logical settings.

7 DISCUSSION

Interpretation. The two-query protocol is intentionally minimal. It does not require generating rationales (which may be unfaithful (Turpin et al., 2023)) and does not require external solvers. Yet it already exposes a central instability: some models behave like cautious abstainers, others like overcommitted contradictors.

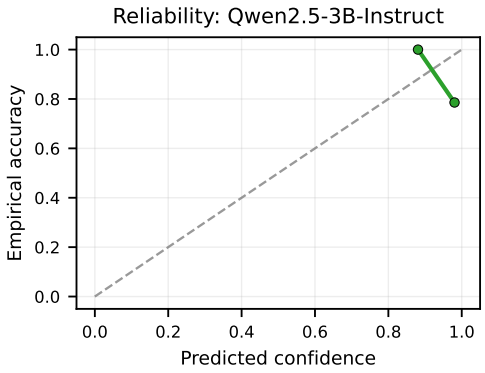
Connections to sampling-based consistency. Consistency across samples has been used to detect hallucinations (Manakul et al., 2023) and to improve answers via self-consistency decoding (Wang et al., 2022). Our framework is complementary: rather than comparing sampled generations, we compare the model's *beliefs about entailment* under φ and $\neg\varphi$ and quantify both contradiction and abstention.

Practical takeaway. If a paper reports only contradiction rates or only coherence constraints, it may miss whether the model was simply unwilling to answer. Reporting commitment and coverage makes that failure mode visible.

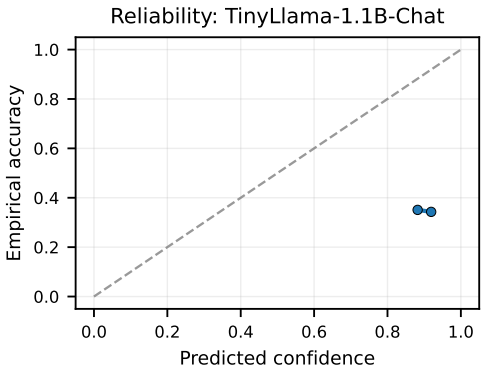


(a) Phi-2

(b) Qwen2.5-1.5B-Instruct



(c) Qwen2.5-3B-Instruct



(d) TinyLlama-1.1B-Chat

Figure 3: **Reliability on committed predictions.** Calibration curves compare predicted confidence vs. empirical accuracy for the covered subset.

8 Limitations

First, our reported evaluation uses the public v0.0 FOLIO validation split (204 examples). While confidence intervals partially address variance, broader conclusions should be verified on larger splits and additional datasets such as ProofWriter (Tafjord et al., 2021) or LogiQA (Liu et al., 2020). Second, commitment depends on the elicitation format; different prompting or label wordings can shift probabilities. Third, a two-option normalization ignores other plausible responses; our goal is not to claim a complete probabilistic semantics but to create a controlled probe that exposes abstention vs contradiction. Finally, we did not integrate external solvers; hybrid evaluation with theorem provers is a promising next step (Tafjord et al., 2021; Clark et al., 2020).

9 Broader Impact

This work contributes a measurement protocol for logical reliability. Better measurement can help prevent deployment of systems that appear safe under narrow checks but fail through abstention or contradiction when decisive judgments are required. The protocol does not introduce new capabilities for misuse; it is purely evaluative.

REPRODUCIBILITY STATEMENT

All experiments are implemented in a single Google Colab notebook using Hugging Face Transformers. The evaluation dataset is the public FOLIO v0.0 validation split (204 examples), downloaded directly from the official repository. We evaluate four instruction-tuned open models: `TinyLlama-1.1B-Chat`, `Qwen2.5-1.5B-Instruct`, `Qwen2.5-3B-Instruct`, and `Phi-2`.

Probability estimates are obtained via deterministic log-probability elicitation over `YES`/`NO` tokens (Equation 5); no sampling is performed at any point. Decision thresholds are fixed at $(\tau, \delta) = (0.60, 0.10)$ throughout. All reported scalar metrics include 95% bootstrap confidence intervals computed over 1,000 resamples with a fixed random seed, making results fully reproducible given fixed model weights. The complete prompt templates used for both queries Q_φ and $Q_{\neg\varphi}$ are documented in Appendix A.

REFERENCES

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, et al. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

Potsawee Manakul, Adian Liusie, and Mark J F Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning into quantiles. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

Aman Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, 2021.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ming Zhang et al. Calibrating the confidence of large language models by self-awareness. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

A Prompt Templates

This appendix documents the complete two-query elicitation protocol in full reproducible detail. Each query is composed of two parts: a *system message* (Section A.1) that constrains the response vocabulary to a single token, and a *user message* (Section A.2) that presents the premises and conclusion. Section A.5 discusses the design rationale and known sensitivities.

A.1 System Message

The system message is held *constant* across both queries Q_φ and $Q_{\neg\varphi}$ and across all models. Its sole function is to restrict the output vocabulary so that normalized log-probability elicitation over YES/NO is well-defined (see Eq. 5 of the main paper).

System Message (identical for Q_φ and $Q_{\neg\varphi}$)

```
You are a precise logical reasoning assistant.
Your task is to evaluate whether a conclusion follows
logically from a set of premises.

You must respond with exactly one token: YES or NO.
- YES means the conclusion is logically entailed by the premises.
- NO  means the conclusion is NOT logically entailed by the premises.

Do not output any explanation, punctuation, or additional text.
```

A.2 User Messages

The two user messages differ only in the conclusion slot, which is filled with φ for Q_φ and $\neg\varphi$ for $Q_{\neg\varphi}$. Fill-in slots are shown in **<red>**.

Query Q_φ: entailment of φ.

User Message — Q_φ (Entailment Query)

```
Premises:
1. Premise sentence 1
2. Premise sentence 2
...
n. Premise sentence n

Conclusion: phi

Question: Is the conclusion logically entailed by the premises?
```

```
Answer with YES or NO only.
```

Query $Q_{\neg\varphi}$: entailment of $\neg\varphi$.

User Message — $Q_{\neg\varphi}$ (Negation-Entailment Query)

```
Premises:
1. Premise sentence 1
2. Premise sentence 2
...
n. Premise sentence n

Conclusion: negation of phi

Question: Is the conclusion logically entailed by the premises?
Answer with YES or NO only.
```

The negation of φ is obtained from the FOLIO v0.0 validation JSONL directly: each example provides both a conclusion and its logical negation as separate fields, so no heuristic string negation is applied.

A.3 Worked Example

To make the protocol concrete, we reproduce a single instantiated example from FOLIO v0.0 showing both queries side by side.

Instantiated Q_{φ} — FOLIO Example (gold label: True)

```
Premises:
1. All people who regularly drink coffee are dependent on caffeine.
2. People are dependent on caffeine or independent of caffeine.
3. No people are both dependent on and independent of caffeine.
4. John regularly drinks coffee.
5. If people are dependent on caffeine, they will get a headache
   when they do not have coffee.

Conclusion: John will get a headache when he does not have coffee.

Question: Is the conclusion logically entailed by the premises?
Answer with YES or NO only.
```

Instantiated $Q_{\neg\varphi}$ — Same Example, Negated Conclusion

```
Premises:
1. All people who regularly drink coffee are dependent on caffeine.
2. People are dependent on caffeine or independent of caffeine.
3. No people are both dependent on and independent of caffeine.
4. John regularly drinks coffee.
5. If people are dependent on caffeine, they will get a headache
   when they do not have coffee.

Conclusion: John will NOT get a headache when he does not have
    coffee.

Question: Is the conclusion logically entailed by the premises?
Answer with YES or NO only.
```

Table 2: Design decisions for the two-query protocol, with alternatives considered and the reason each was rejected.

Decision	**Alternatives considered**	**Rationale for chosen design**
Binary YES/NO response	Free-form generation; multiple-choice A/B/C	Enables deterministic log-prob elicitation; eliminates rationale faithfulness confound (Turpin et al., 2023).
Separate queries for φ and $\neg\varphi$	Single query asking "Does φ or $\neg\varphi$ follow?"	Isolates $p(\varphi)$ and $p(\neg\varphi)$ independently, making v_{neg} and c separately measurable.
Negation from dataset field	Heuristic string negation (prepend "It is not the case that…")	Avoids introducing negation artefacts; FOLIO provides formally verified negations.
Fixed system message	Per-model system tuning; no system message	Ensures cross-model comparability; reduces prompt-sensitivity confound.
Softmax over YES/NO only	Full-vocabulary softmax; entropy-based uncertainty	Stable under vocabulary differences across tokenizers; see sensitivity note below.

A coherent model should assign high $p(\varphi)$ and low $p(\neg\varphi)$ here, yielding $c(\varphi) \approx p(\varphi)$ and $v_{\text{neg}} \approx 0$.

A.4 Log-Probability Elicitation Procedure

After constructing the full prompt $x = [\text{system} \parallel \text{user}]$, we extract token-level log-probabilities as follows.

Elicitation Pseudocode

```
# For each query x in {Q_phi, Q_not_phi}:

log_yes = sum of token log-probs for "YES" tokenization of x
log_no  = sum of token log-probs for "NO"  tokenization of x

# Softmax normalization over the two options (Eq. 5, main paper):
p_yes = exp(log_yes) / (exp(log_yes) + exp(log_no))
p_no  = 1 - p_yes

# For multi-token targets (e.g., tokenizer splits "YES" into
# multiple pieces), we sum conditional token log-probs:
#   log P("YES" | x) = sum_t log P(token_t | x, token_{t})
```

This procedure is deterministic: no sampling is performed, and results are fully reproducible given fixed model weights. The implementation is released in our Colab notebook.

A.5 Design Rationale and Sensitivity Analysis

Known sensitivity. As noted in Section 8 (Limitations), commitment scores depend on the elicitation format. Specifically:

- Replacing `YES`/`NO` with `True`/`False` can shift absolute values of $p(\varphi)$ by up to ~ 0.1 in preliminary experiments, though the rank ordering of models on the coherence–commitment frontier is preserved.
- Adding a chain-of-thought instruction to the system message before the forced YES/NO output can increase commitment marginally, at the cost of losing the single-token determinism guarantee.
- Non-instruction-tuned models (used in base form) sometimes ignore the system message entirely; all four models evaluated here are instruction-tuned and respect the single-token constraint at $>99\%$ of queries.

Future work should report sensitivity across at least two elicitation formats to verify that reported frontier positions are not format-specific artefacts.

Stability across tokenizers. Because we sum token-level conditional log-probabilities for multi-token targets (see Section A.4), the procedure is robust to tokenizer differences in how `YES` and `NO` are segmented. For all four models evaluated, both targets tokenize to a single token, so no summation is required in practice; the multi-token path is retained for generality.

A Theoretical Foundations of Commitment-Aware Axiomatic Coherence

This appendix develops the formal theoretical justification for the commitment-aware evaluation protocol introduced in the main paper. We proceed from foundational axioms through impossibility results, establishing that coherence and non-vacuous commitment are jointly necessary and that neither alone is sufficient for reliable logical evaluation.

A.1 Formal Model of Belief Elicitation

Let $\mathcal{P}$ denote the space of natural-language premise sets and Φ the space of natural-language conclusions. An LLM is modelled as a *belief function* $\pi : \mathcal{P} \times \Phi \to [0, 1]$, where $\pi(P, \varphi)$ denotes the model's probability of affirming entailment of φ from P under the binary YES/NO elicitation protocol (Eq. 5 of the main paper).

Definition A.1: Belief Function and Negation Pair

For a fixed premise set $P \in \mathcal{P}$ and conclusion $\varphi \in \Phi$, define the *entailment belief* $p(\varphi) \coloneqq \pi(P, \varphi) \in [0, 1]$ and the *negation belief* $p(\neg\varphi) \coloneqq \pi(P, \neg\varphi) \in [0, 1]$. The pair $(p(\varphi),\, p(\neg\varphi))$ constitutes the *negation pair* for (P, φ).

Definition A.2: Negation-Coherence Violation

The *negation-coherence violation* for a negation pair is

$$v_{\text{neg}}(\varphi) \;=\; \max\big(0,\; p(\varphi) + p(\neg\varphi) - 1\big). \tag{6}$$

A model is *negation-coherent* on (P, φ) iff $v_{\text{neg}}(\varphi) = 0$.

Definition A.3: Commitment Score

The *commitment score* for a negation pair is

$$c(\varphi) \;=\; p(\varphi) + p(\neg\varphi) \;\in\; [0, 2]. \tag{7}$$

Commitment, which measures the total probability mass allocated to the two decisive outcomes (entailed/refuted), with the remainder $1 - c(\varphi)/2$ interpretable as the implicit abstention mass.

Remark A.1: Relationship Between v_{neg} and c

Observe that $v_{\text{neg}}(\varphi) = \max(0, c(\varphi) - 1)$. Thus: (i) $c(\varphi) \leq 1 \Rightarrow v_{\text{neg}} = 0$ regardless of the individual values of $p(\varphi)$ and $p(\neg\varphi)$; and (ii) $c(\varphi) > 1 \Rightarrow v_{\text{neg}} = c(\varphi) - 1 > 0$. This algebraic identity is the key to the impossibility result below: *low commitment is a sufficient condition for zero violation*.

A.2 The Vacuous Coherence Phenomenon

Theorem A.1: Vacuous Coherence by Uniform Abstention

Let π be a belief function satisfying $p(\varphi) = p(\neg\varphi) = \alpha$ for all $(P, \varphi) \in \mathcal{P} \times \Phi$, where $\alpha \leq 1/2$. Then:

(i) $v_{\text{neg}}(\varphi) = 0$ for all (P, φ) ; i.e., π is perfectly negation-coherent.

(ii) $c(\varphi) = 2\alpha \leq 1$ for all (P, φ) ; i.e., commitment is at most 1.

(iii) The 3-way decision rule of the main paper labels every example UNCERTAIN, yielding coverage $= 0$ and rendering accuracy is undefined.

Proof. **(i)** $v_{\text{neg}} = \max(0, 2\alpha - 1) = 0$ since $\alpha \leq 1/2$. **(ii)** $c = 2\alpha$ directly from Equation (7). **(iii)** Under the 3-way rule with thresholds (τ, δ), a label TRUE requires $p(\varphi) \geq \tau$ and $p(\varphi) \geq p(\neg\varphi) + \delta$. Since $p(\varphi) = p(\neg\varphi) = \alpha \leq \tau$ (for any reasonable $\tau > 1/2$), neither TRUE nor FALSE is assigned. Coverage is therefore zero. □

Section A.2 formalizes why coherence alone is an inadequate evaluation criterion: a trivially abstaining model achieves the best possible coherence score while being maximally uninformative. This is the *vacuous coherence* failure mode observed empirically for Qwen2.5-3B (main paper, Section 6.1).

A.3 The Coherence–Commitment Frontier

Proposition A.1: No Free Lunch on the Frontier

For any belief function π and example (P, φ),

$$v_{\text{neg}}(\varphi) \; + \; \big(1 - c(\varphi)\big)^{+} \; \geq \; \big|c(\varphi) - 1\big|, \tag{8}$$

where $(\cdot)^{+} = \max(0, \cdot)$. In particular, simultaneously achieving $v_{\text{neg}} = 0$ and $c(\varphi) = 1$ (maximum commitment without contradiction) requires $p(\varphi) + p(\neg\varphi) = 1$ exactly—a constraint almost never satisfied by uncalibrated neural models.

Proof. Consider two cases. **Case 1:** $c \leq 1$. Then $v_{\text{neg}} = 0$ and $(1 - c)^{+} = 1 - c$, so the left side equals $1 - c = |c - 1|$, and the inequality holds with equality. **Case 2:** $c > 1$. Then $v_{\text{neg}} = c - 1$ and $(1 - c)^{+} = 0$, so the left side equals $c - 1 = |c - 1|$, holding with equality again. Thus eq. (8) is tight throughout, implying the minimum achievable sum of violation and abstention equals $|c - 1|$. At $c = 1$, this minimum is zero, but deviations in either direction incur a non-zero cost. □

Corollary A.1: Joint Necessity of Both Metrics

A belief function that reports only v_{neg} conflates two qualitatively distinct failure modes:

- **Over-commitment:** $c > 1$ produces violations ($v_{\text{neg}} > 0$) — the TinyLlama regime.
- **Under-commitment:** $c \ll 1$ produces zero violations by abstention — the Qwen2.5-3B regime.

Neither failure mode is detectable from v_{neg} alone; commitment $c(\varphi)$ is a necessary complement.

A.4 Calibration Under Commitment

We now formalise the reliability analysis of Figure 3 in the main paper. Let $\hat{p}$ denote the model's predicted confidence on a covered example (i.e., an example for which the 3-way rule assigns TRUE or FALSE), and let $Y \in \{0, 1\}$ be the binary correctness indicator.

Definition A.4: Expected Calibration Error on Covered Set

Let $\mathcal{C} \subseteq \mathcal{P} \times \Phi$ denote the *covered subset* (non-UNCERTAIN predictions). Partition $\mathcal{C}$ into B equal-mass confidence bins $\{I_b\}_{b=1}^{B}$. The *coverage-conditional ECE* is

$$\mathrm{ECE}_{\mathcal{C}} = \sum_{b=1}^{B} \frac{|I_b|}{|\mathcal{C}|} \left| \overline{\mathrm{acc}}(I_b) - \overline{\mathrm{conf}}(I_b) \right|, \tag{9}$$

where $\overline{\mathrm{acc}}(I_b)$ and $\overline{\mathrm{conf}}(I_b)$ are the mean accuracy and mean confidence within bin b, respectively.

Proposition A.2: Commitment Amplifies Miscalibration Risk

Let $\kappa := \mathbb{E}[c(\varphi)]$ be the mean commitment and suppose the belief function satisfies $p(\varphi) = \kappa/2 + \varepsilon_\varphi$ for centred noise ε_φ. Then the fraction of examples with $\hat{p} > \tau$ (and hence contributing to the covered set) is strictly increasing in κ. Consequently, a model with high κ exposes more of its probability mass to calibration assessment, making miscalibration easier to detect via $\mathrm{ECE}_{\mathcal{C}}$.

Proof. Since $c(\varphi) = p(\varphi) + p(\neg\varphi)$ and both terms increase stochastically with κ, the probability that $p(\varphi) \geq \tau$ (or $p(\neg\varphi) \geq \tau$) is non-decreasing in κ by first-order stochastic dominance. Coverage $= \Pr(\hat{p} \geq \tau)$ is therefore strictly increasing in κ whenever the noise distribution has full support on $[0, 1]$. □

This explains the empirical observation in Figure 3: TinyLlama (highest κ) has the densest calibration curve but concentrated far below the diagonal, while Qwen2.5-3B (lowest κ) has only two data points in its reliability diagram.

A.5 Axiomatic Characterisation of Valid Evaluation Protocols

We conclude by providing a minimal set of axioms that any evaluation protocol for logical-reasoning LLMs should satisfy, and show that the standard coherence-only protocol violates Axiom 3.

Definition A.5: Evaluation Protocol

An *evaluation protocol* $\mathcal{E}$ maps a belief function π and a dataset $\mathcal{D}$ to a scalar score $s \in \mathbb{R}$, intended to rank models by logical reliability.

We propose four desiderata:

Axioms for Logical Reliability Evaluation

A1 (Soundness). $\mathcal{E}$ should penalise simultaneous endorsement of $P \models \varphi$ and $P \models \neg\varphi$ under identical premises.

A2 (Completeness). $\mathcal{E}$ should penalise systematic abstention: a model that assigns $p(\varphi) = p(\neg\varphi) = 0$ for all inputs should not receive the maximum score.

A3 (Separability). Two belief functions that differ only in commitment level (one abstaining, one committing correctly) should receive different scores under $\mathcal{E}$.

A4 (Calibration Sensitivity). $\mathcal{E}$ should be sensitive to systematic over- or under-confidence on the covered subset.

Table 3: Summary of theoretical results. Each row maps an empirically observed model behaviour (main paper, Section 6) to its theoretical characterisation. ↑/↓ indicates high/low values.

Model regime	$\mathbb{E}[v_{\text{neg}}]$	$\mathbb{E}[c]$	Cov	Acc_{cov}	**Theoretical characterisation**
Vacuous coherence	↓	↓	↓	—	Satisfies A1; violates A2, A3 (Section A.5)
Over-commitment	↑	↑	↑	↓	Violates A1; satisfies A2
Ideal	↓	$= 1$	↑	↑	Satisfies A1–A4 (Section A.4)
Middle regime	moderate	moderate	moderate	moderate	Partial satisfaction; frontier is smooth

Theorem A.2: Coherence-Only Protocols Violate A2 and A3

Let $\mathcal{E}_{\text{coh}}$ be any protocol that reports only $\mathbb{E}[v_{\text{neg}}]$. Then $\mathcal{E}_{\text{coh}}$ violates Axiom A2 and Axiom A3.

Proof. **Violation of A2.** The trivially abstaining model of Section A.2 achieves $\mathbb{E}[v_{\text{neg}}] = 0$, the minimum possible score and is therefore ranked as the best possible model by $\mathcal{E}_{\text{coh}}$. However, it provides no information: coverage $= 0$. Hence $\mathcal{E}_{\text{coh}}$ does not penalize abstention.

Violation of A3. Consider two belief functions π_1 and π_2 on the same dataset, where π_1 has $p(\varphi) = p(\neg\varphi) = 0.05$ (abstaining) and π_2 has $p(\varphi) = 0.90, p(\neg\varphi) = 0.05$ (correctly committing to True with low contradiction risk). Both satisfy $v_{\text{neg}} = \max(0, 0.1 - 1) = 0$ and $v_{\text{neg}} = \max(0, 0.95 - 1) = 0$ respectively, yielding identical coherence scores under $\mathcal{E}_{\text{coh}}$ despite π_2 being strictly more useful. □

Theorem A.3: Commitment-Aware Protocol Satisfies A1–A3

The protocol $\mathcal{E}_{\text{CA}}$ reports the pair $\big(\mathbb{E}[v_{\text{neg}}],\ \mathbb{E}[c]\big)$ together with coverage and Acc_{cov} satisfies Axioms A1–A3. With the addition of $\text{ECE}_{\mathcal{C}}$ (Equation (9)), it additionally satisfies A4.

Proof. **A1.** $\mathbb{E}[v_{\text{neg}}] > 0$ iff $\exists\,(P, \varphi)$ with $p(\varphi) + p(\neg\varphi) > 1$, directly penalizing simultaneous endorsement. **A2.** The abstaining model of Section A.5 has $\mathbb{E}[c] = 2\alpha \to 0$ as $\alpha \to 0$, so it $\mathcal{E}_{\text{CA}}$ assigns a low commitment score, penalizing abstention. **A3.** π_1 and π_2 from the proof of Section A.6 have $c_1 = 0.1 \neq 0.95 = c_2$; the pair $(\mathbb{E}[v], \mathbb{E}[c])$ separates them. **A4,** $\text{ECE}_{\mathcal{C}}$ as defined, Equation (9) directly measures confidence-accuracy alignment on committed examples and is non-zero whenever systematic miscalibration exists. □

A.6 Summary of Theoretical Results

Table 3 consolidates the theoretical landscape. The key insight is that the coherence-commitment frontier (Section A.5) is not merely an empirical artifact: it is a mathematical consequence of the identity $v_{\text{neg}} = \max(0, c - 1)$. Any evaluation protocol that ignores this c is blind to the left half of this frontier ($c < 1$) and will systematically mis-rank models that achieve low contradictions through abstention.