DeltaLogic: Minimal Premise Edits Reveal Belief-Revision Failures in Logical Reasoning Models

Anonymous authors
Paper under double-blind review

Abstract

Reasoning benchmarks typically evaluate whether a model derives the correct answer from a fixed premise set, but they under-measure a closely related capability that matters in dynamic environments: belief revision under minimal evidence change. We introduce DeltaLogic, a benchmark transformation protocol that converts natural-language reasoning examples into short revision episodes. Each episode first asks for an initial conclusion under premises P, then applies a minimal edit $\delta(P)$, and finally asks whether the previous conclusion should remain stable or be revised. We instantiate DeltaLogic from FOLIO and ProofWriter and evaluate small causal language models with constrained label scoring. On a completed 30-episode Qwen evaluation subset, stronger initial reasoning still does not imply stronger revision behavior: Qwen3-1.7B reaches 0.667 initial accuracy but only 0.467 revision accuracy, with inertia rising to 0.600 on episodes where the gold label should change, while Qwen3-0.6B collapses into near-universal abstention. There, Qwen3-4B preserves the same inertial failure pattern (0.650 initial, 0.450 revised, 0.600 inertia), whereas Phi-4-mini-instruct is substantially stronger (0.950 initial, 0.850 revised) but still exhibits non-trivial abstention and control instability. These results suggest that logical competence under fixed premises does not imply disciplined belief revision after local evidence edits. DeltaLogic therefore targets a distinct and practically important reasoning capability that complements existing logical inference and belief-updating benchmarks.

1 Introduction

Reasoning benchmarks usually evaluate inference from fixed premises. DeltaLogic targets a different capability: local belief revision. When one premise is inserted, deleted, or replaced, the model should update exactly the commitments that the edit warrants and leave the rest stable. This matters for systems that reason over changing documents, dynamic rules, or incrementally updated evidence.

DeltaLogic turns a standard reasoning item into a short revision episode with a known semantic effect. This yields a failure taxonomy that static accuracy cannot expose: inertia (keeping an outdated answer), over-flip (revising under an irrelevant edit), and degenerate abstention. Our contributions are a simple benchmark-construction protocol, metrics that separate initial reasoning from revision discipline, and generative-model evidence that belief revision is a distinct challenge even for current small and near-4B models.

This paper is a measurement contribution, not a new reasoning algorithm. The claim is not that DeltaLogic explains the internal mechanism of belief revision in language models. The claim is narrower and testable: current small LMs that appear competent on static reasoning still fail under minimal premise edits, and those failures can be decomposed into stable, measurable revision modes.

Contributions. (1) We introduce a benchmark-transformation protocol for local belief revision over public logical reasoning datasets. (2) We define a small set of revision-specific metrics that separate stale commitment, unnecessary revision, and abstention. (3) We show

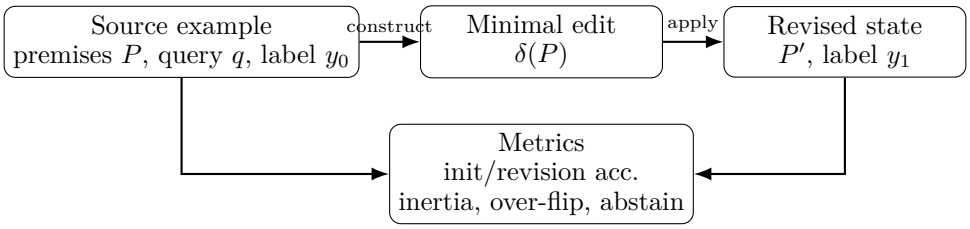

Figure 1: DeltaLogic construction pipeline. A standard reasoning item is turned into a minimally edited revision episode with a known semantic effect, which makes revision errors measurable rather than anecdotal.

on completed runs that stronger static reasoning does not guarantee better local revision, and that models of similar scale can fail through different revision modes.

2 Methodology

Let an original reasoning instance consist of premises P, query q, and gold label y_0. A DeltaLogic episode applies a minimal edit operator δ to obtain revised premises $P' = \delta(P)$ with revised gold label y_1. The edit is designed so that its intended semantic effect is deterministic.

Definition 1 (DeltaLogic episode). A DeltaLogic episode is a tuple

$$e = (P, q, y_0, P', y_1, t),$$

where P is the original premise set, q is the hypothesis or query, y_0 is the original gold label, P' is the minimally edited premise set, y_1 is the revised gold label, and t is the edit type.

We use four edit types: support insertion, defeating-fact insertion, support removal, and irrelevant-fact addition. Together they test positive updating, belief retraction, and stability under no-change controls.

DeltaLogic is instantiated from FOLIO and ProofWriter. From FOLIO, we build support-insertion, defeating-fact, and irrelevant-addition episodes. From ProofWriter, we build support-removal and irrelevant-addition episodes using shallow examples with identified support facts. The full construction contains 100 episodes. For this paper we report completed generative-model evaluation subsets that preserve all edit types while remaining feasible on CPU-only hardware: a 30-episode main Qwen subset and a 20-episode near-4B extension.

3 Metrics

We report initial accuracy,

$$\mathrm{Acc}_{\mathrm{init}} = \Pr(\hat{y}_0 = y_0),$$

and revision accuracy,

$$\mathrm{Acc}_{\mathrm{rev}} = \Pr(\hat{y}_1 = y_1).$$

To expose failure modes, let C denote episodes whose gold label should change and U denote episodes whose gold label should remain stable. The inertia rate is

$$\mathrm{Inertia} = \Pr(\hat{y}_1 = \hat{y}_0 \mid e \in C,\ \hat{y}_0 = y_0),$$

The over-flip rate is

$$\mathrm{OverFlip} = \Pr(\hat{y}_1 \neq \hat{y}_0 \mid e \in U,\ \hat{y}_0 = y_0),$$

We also report abstention rate,

$$\mathrm{Abstain} = \Pr(\hat{y}_1 = \mathrm{Uncertain}),$$

because some models appear stable only by retreating into uncertainty.

We also refer to the empirical revision gap, $\mathrm{Acc}_{\mathrm{init}} - \mathrm{Acc}_{\mathrm{rev}}$, as a compact measure of how much performance degrades after the evidence edit.

Table 1: Main DeltaLogic results on the completed generative-model evaluation subsets.

Model	Slice n	Init. ↑	Rev. ↑	Inertia ↓	Over-flip ↓	Abstain ↓
Qwen3-0.6B	30	0.300	0.400	0.400	0.000	1.000
Qwen3-1.7B	30	0.667	0.467	0.600	0.000	0.000
Qwen3-4B	20	0.650	0.450	0.600	0.000	0.000
Phi-4-mini	20	0.950	0.850	0.200	0.100	0.350

4 Experimental Setup

We evaluate Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, and Phi-4-mini-instruct. Prompts end in Label:, and we score the continuations True, False, and Uncertain by average token log-likelihood under the frozen causal LM. This constrained scoring avoids chain-of-thought parsing artifacts and uses no task-specific fine-tuning or verifier model. The completed experiments are split into two evaluation subsets: a 30-episode Qwen main study and a 20-episode near-4B extension. We keep them separate in reporting because the subsets are not identical.

5 Results

Table 1 reports all completed runs. The central finding is that stronger initial reasoning still does not imply stronger revision behavior. On the 30-episode Qwen evaluation subset, Qwen3-1.7B reaches 0.667 initial accuracy but only 0.467 revision accuracy, a revision gap of 0.200, while Qwen3-0.6B reaches 0.400 revision accuracy only through universal abstention. On the completed 20-episode near-4B evaluation subset, Qwen3-4B preserves the same inertial failure pattern, whereas Phi-4-mini-instruct is substantially stronger but not stable.

The edit-type breakdown reveals the real structure of the problem. Qwen3-1.7B succeeds on support-insertion episodes with perfect revision accuracy and remains perfectly stable on the ProofWriter irrelevant-addition control. Yet it fails completely on support-removal and defeating-fact edits. This asymmetry is scientifically important. Adding a new premise that explicitly supports the query is easier than revising a previously justified belief once its support disappears or once a defeating fact arrives. The harder case requires reasoning about the loss or reversal of support rather than simply reacting to explicit positive evidence.

Qwen3-0.6B exhibits a different pathology. Its revised abstention rate is 1.000, so its apparent stability is not genuine revision discipline. It gets support-removal episodes correct only because the gold revised label in that regime is also Uncertain. The smaller model therefore avoids stale commitments by retreating into blanket uncertainty, whereas the larger Qwen models preserve sharper beliefs but fail to revise them when they should change. Qwen3-4B does not materially improve over Qwen3-1.7B on inertia.

Phi-4-mini-instruct behaves differently. It is the strongest completed model in the paper, but it still shows a non-zero over-flip rate and a substantial abstention rate. The main scientific takeaway is therefore not simply that a stronger model solves the problem; rather, DeltaLogic distinguishes between multiple revision regimes. The Qwen family is primarily inertial, while Phi-4-mini is more revision-capable but still hedges and occasionally revises unnecessarily. The edit breakdown makes this concrete: Qwen3-4B remains at 0.000 revision accuracy on ProofWriter support-removal episodes, while Phi-4-mini reaches 0.667 on the same edit type.

The important observation is not just that one model scores higher. It is that scale alone does not induce the right local update rule. Qwen3-4B is larger than Qwen3-1.7B but preserves essentially the same inertia profile. That is evidence against a naive "more scale fixes revision" story and in favor of the paper's narrower claim: minimal belief revision is a distinct capability worth measuring directly.

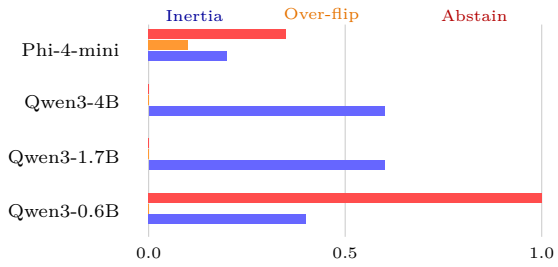

Figure 2: Failure-mode comparison across completed runs. The Qwen family remains inertia-dominated, while Phi-4-mini trades lower inertia for non-zero over-flip and abstention.

6 Related Work

DeltaLogic is closest to three neighboring lines. First, Belief-R studies adaptation under new evidence and motivates the general update-versus-stability tension (Wilie et al., 2024). DeltaLogic differs by using smaller local edits with deterministic semantic effects. Second, logical-reasoning benchmarks such as FOLIO, ProofWriter, LogicBench, and LogicGame measure inference from fixed premises (Han et al., 2022; Tafjord et al., 2021; Parmar et al., 2024; Gui et al., 2025). DeltaLogic is layered on top of them as a transformation protocol for revision rather than static inference. Third, revision-oriented question answering and correction benchmarks such as ReviseQA study answer updating under changing evidence, but do not focus on logically controlled minimal premise edits. DeltaLogic asks a narrower question: after one evidence edit, does the model update exactly the commitments that should change and keep the rest stable?

7 Limitations

DeltaLogic is a controlled benchmark built from transformed FOLIO and ProofWriter examples rather than open-world interactive agents, so the results trade ecological breadth for precise edit semantics. The current submission also reports completed but unevenly sized model evaluations because causal-LM scoring on the available CPU-only setup is expensive. Accordingly, the results support the existence and structure of the failure mode rather than a definitive leaderboard. Finally, the benchmark targets single-step revision and label correctness, not longer revision chains or free-form justification quality.

8 Conclusion

Belief revision under minimal premise edits is a distinct reasoning capability that current evaluations only partially capture. DeltaLogic makes that capability measurable by converting standard reasoning instances into short revision episodes with known semantic effects. Across the completed runs in this paper, small causal LMs fail in at least three distinct ways: clinging to stale commitments, collapsing into uncertainty, or revising mostly correctly but with residual abstention and occasional over-flip. The practical implication is straightforward: getting the original answer right is not enough. A reliable reasoning model must also know how to update that answer precisely when the evidence changes, and DeltaLogic provides a direct probe of that ability.

References

Parmar, M., Patel, N., Varshney, N., Nakamura, M., Luo, M., Mashetty, S., Mitra, A., and Baral, C. LogicBench: Towards Systematic Evaluation of Logical Reasoning Ability of Large Language Models. ACL 2024. https://aclanthology.org/2024.acl-long.739/

Wilie, B., Cahyawijaya, S., Ishii, E., He, J., and Fung, P. Belief Revision: The Adaptability of Large Language Models Reasoning. EMNLP 2024. https://aclanthology.org/2024.emnlp-main.586/

Gui, J., Liu, Y., Cheng, J., Gu, X., Liu, X., Wang, H., Dong, Y., Tang, J., and Huang, M. LogicGame: Benchmarking Rule-Based Reasoning Abilities of Large Language Models. Findings of ACL 2025. https://aclanthology.org/2025.findings-acl.77/

Yan, Y., et al. ReviseQA: A Benchmark for Belief Revision in Question Answering. ICML 2025.

Han, S., et al. FOLIO: Natural Language Reasoning with First-Order Logic. 2022. https://arxiv.org/abs/2209.00840

Tafjord, O., Dalvi, B., and Clark, P. ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language. EMNLP 2021.

A Appendix: Algorithmic Details

Benchmark construction. Each DeltaLogic episode is constructed from a base example (P, q, y) with premise set P, query q, and gold label $y \in \{\text{True}, \text{False}, \text{Uncertain}\}$. We first select examples whose semantic status can be modified by one controlled edit. We then generate either a changed episode, in which the label is intended to change, or a control episode, in which the label should remain stable.

1. Sample a base example from FOLIO or ProofWriter.
2. Derive an edit operator e from a fixed family: support insertion, defeating-fact insertion, support removal, or irrelevant-fact addition.
3. Apply e to obtain revised premises $P' = e(P)$.
4. Compute the revised label y' using dataset metadata and edit semantics.
5. Emit a four-turn episode: (i) answer the original query from P; (ii) optionally explain briefly which premise matters most; (iii) observe the revised premise set P'; (iv) answer the same query again and state whether the answer should change.

Label-scoring inference. The current experiments use closed-label scoring rather than unconstrained generation. For each model, prompt state, and candidate label $a \in \{\text{True}, \text{False}, \text{Uncertain}\}$, we score the continuation by conditional log-likelihood,

$$s(a \mid x) = \sum_{t=1}^{|a|} \log p_\theta(a_t \mid x, a_{<t}),$$

where x is the serialized dialogue context. The predicted label is

$$\hat{a} = \arg \max_{a \in \{\text{True},\text{False},\text{Uncertain}\}} s(a \mid x).$$

Initial accuracy is measured on the pre-edit state; revision accuracy is measured on the post-edit state. Inertia is the rate at which a model preserves the old label when the gold label changes. Over-flip is the rate at which a model changes its answer on control episodes where the gold label should remain stable. Abstention is the prediction rate for Uncertain.

Practical implementation. On causal language models, each label score requires a full forward pass over the dialogue prefix plus label tokens. This is why the generative-model experiments are more expensive than discriminative NLI baselines. All reported runs use deterministic preprocessing, fixed random seeds for benchmark construction, and locally cached model and dataset artifacts to avoid network variance.