AGM-BENCH: DO LARGE LANGUAGE MODELS REVISE BELIEFS RATIONALLY?

Ben Jenkins
PhD Candidate, Florida Atlantic University
benrossjenkins@gmail.com

ABSTRACT

Large language models (LLMs) are increasingly deployed in settings that require updating conclusions as new information arrives, from multi-turn dialogue to agentic workflows with evolving evidence. Yet virtually all evaluations of LLM logical reasoning focus on *static* problems: given fixed premises, derive a conclusion. We introduce AGM-BENCH, the first benchmark grounded in the AGM theory of belief revision, which tests whether LLMs update their beliefs *rationally* when confronted with new, potentially contradictory information. AGM-BENCH operationalizes six classical rationality postulates, namely *Success*, *Consistency*, *Inclusion*, *Vacuity*, *Extensionality*, and *Preservation*, as well as the Darwiche–Pearl postulates for iterated revision, across 2,400 synthetic reasoning scenarios of controlled logical complexity. We evaluate seven frontier LLMs and find that: (1) all models satisfy *Success* and *Consistency* at high rates, but systematically violate *Inclusion* (minimal change) and *Preservation* (stability of unrelated beliefs); (2) under iterated revision, models exhibit severe *belief inertia* (retaining retracted information) and *collateral damage* (retracting beliefs not logically affected by the new evidence); and (3) reasoning-trained models (o3-mini, DeepSeek-R1) show improved single-step revision but *degrade faster* under iteration than standard chat models. Our results reveal a fundamental gap between LLM reasoning and rational belief dynamics.

1 INTRODUCTION

Logical reasoning is a cornerstone capability for reliable AI systems, and significant progress has been made in evaluating and improving the deductive, inductive, and abductive abilities of large language models (Cheng et al., 2025; Parmar et al., 2024). However, the vast majority of this work evaluates reasoning in a *static* setting: the model is presented with a fixed set of premises and asked to derive a conclusion. Real-world reasoning is fundamentally *dynamic*: agents must update, retract, and revise their beliefs as new information becomes available.

Consider a medical diagnosis system that initially concludes a patient has condition A based on test results, but must revise this conclusion when a new test contradicts a key premise. Or a legal reasoning assistant that must update its analysis when a precedent is overturned. In both cases, the *quality* of the update matters as much as the initial reasoning. An irrational revision, one that discards too much prior knowledge, fails to incorporate the new evidence, or introduces contradictions, can be more harmful than no revision at all.

The gold standard for rational belief revision in formal epistemology is the AGM theory (Alchourrón et al., 1985; Gärdenfors, 1988), which specifies a set of postulates that any "rational" revision operator should satisfy. These postulates capture intuitions such as: new information should be successfully incorporated (*Success*); the revised beliefs should be consistent if the new information is consistent (*Consistency*); and the revision should change as little as possible (*Inclusion*/minimal change). Extensions by Darwiche & Pearl (1997) address *iterated* revision, specifying how an agent's disposition toward future revisions should itself change.

Despite the centrality of belief dynamics to both logic and AI, no benchmark systematically tests whether LLMs satisfy these rationality postulates. Existing consistency evaluations (Calanzone et al., 2025; Lin et al., 2025; Liu et al., 2025) focus on static properties (negation consistency,

transitivity) across isolated question pairs. Work on defeasible reasoning (Rudinger et al., 2020; Allaway et al., 2025; Bao et al., 2025) tests whether models can override defaults, but does not measure whether the *overall revision behavior* is rational. We aim to close this gap.

Contributions. We make the following contributions:

1. We introduce AGM-BENCH, a benchmark of 2,400 scenarios that operationalizes six basic AGM postulates and four Darwiche–Pearl iterated revision postulates as testable behavioral criteria for LLMs (§3).
2. We provide a formal mapping from the AGM framework to a multi-turn prompting protocol that isolates belief revision behavior from confounds such as instruction-following and in-context recency effects (§4).
3. We evaluate seven frontier models and report systematic postulate violations, identifying *Inclusion* and *Preservation* as the primary failure modes, and documenting a novel "iterated degradation" phenomenon in reasoning-trained models (§5).

2 BACKGROUND: THE AGM FRAMEWORK

We briefly review the AGM theory; for comprehensive treatments, see Gärdenfors (1988) and Hansson (1999).

Reading guide. The symbols below are reused throughout: $\mathcal{K}$ denotes the agent's belief set, φ and ψ denote formulas (new evidence or probe queries), $\circledast$ is the revision operator, and Cn is deductive closure. On a first read, one may skim the postulate statements for their English glosses and return to the formal conditions as needed; Appendix A lists symbols in one place.

Belief sets. A *belief set* $\mathcal{K}$ is a deductively closed set of propositions: $\mathcal{K} = \mathrm{Cn}(\mathcal{K})$, where Cn denotes the logical closure operator. Intuitively, $\mathcal{K}$ represents everything an agent believes, including all logical consequences of its explicit beliefs.

Revision. Given a belief set $\mathcal{K}$ and new information φ, the *revision* $\mathcal{K} \circledast \varphi$ produces a new belief set that incorporates φ. The AGM postulates constrain this operator:

Postulate 1 (Success). *$\varphi \in \mathcal{K} \circledast \varphi$. The new information is always accepted.*

Postulate 2 (Consistency). *If φ is consistent, then $\mathcal{K} \circledast \varphi$ is consistent.*

Postulate 3 (Inclusion). *$\mathcal{K} \circledast \varphi \subseteq \mathrm{Cn}(\mathcal{K} \cup \{\varphi\})$. Revision does not introduce beliefs beyond what follows from the old beliefs and the new information.*

Postulate 4 (Vacuity). *If $\neg\varphi \notin \mathcal{K}$, then $\mathcal{K} \circledast \varphi = \mathrm{Cn}(\mathcal{K} \cup \{\varphi\})$. If the new information does not contradict current beliefs, revision reduces to simple expansion.*

Postulate 5 (Extensionality). *If $\varphi \equiv \psi$, then $\mathcal{K} \circledast \varphi = \mathcal{K} \circledast \psi$. Logically equivalent inputs yield the same revision.*

Postulate 6 (Preservation). *If $\psi \in \mathcal{K}$ and ψ is consistent with φ, then $\psi \in \mathcal{K} \circledast \varphi$. Beliefs compatible with the new information are retained.*

These are simplified from the standard eight AGM postulates (K*1–K*8) for expository clarity; see Appendix B for the complete formulation.

Iterated revision. The basic AGM postulates constrain a single revision step. Darwiche & Pearl (1997) proposed additional postulates for sequences of revisions $(\mathcal{K} \circledast \varphi) \circledast \psi$, capturing intuitions about how the agent's *epistemic state* (not just belief set) should evolve. A key postulate is:

Postulate 7 (DP1: Recency). *If $\psi \models \varphi$, then $(\mathcal{K} \circledast \varphi) \circledast \psi = \mathcal{K} \circledast \psi$. If the second piece of evidence entails the first, the first revision becomes irrelevant.*

Postulate 8 (DP2: Irrelevance). *If $\psi \models \neg\varphi$, then $(\mathcal{K} \circledast \varphi) \circledast \psi = \mathcal{K} \circledast \psi$. If the second evidence contradicts the first, the first revision is fully overridden.*

These postulates are particularly relevant for LLMs, which process information sequentially and may exhibit recency bias or excessive anchoring to earlier context.

3 THE AGM-BENCH BENCHMARK

3.1 DESIGN PRINCIPLES

We design AGM-BENCH around three principles: (1) **formal grounding**, where each test instance directly operationalizes a specific rationality postulate; (2) **controlled complexity**, where reasoning depth is parameterized so we can study how violations scale; and (3) **isolation**, where we control for confounds such as world knowledge by using abstract entities and relations.

3.2 SCENARIO GENERATION

Each scenario consists of a *knowledge base* $\mathcal{K}$ (a set of natural-language premises), a *revision input* φ (new information), and a set of *probe queries* designed to test specific postulates. We generate scenarios using a three-step pipeline:

Step 1: Logical skeleton. We generate propositional logic programs of controlled depth $d \in \{1, 2, 3, 4\}$ using a grammar over abstract predicates and entities (e.g., "Every blicket is a dax," "All daxes are feps"). This follows the methodology of ProofWriter (Tafjord et al., 2021) but extends it with revision-specific structures.

Step 2: Revision point. We select a fact $f \in \mathcal{K}$ and construct a revision input φ such that φ contradicts f (for contradiction-triggered revision) or is compatible with $\mathcal{K}$ (for vacuity tests). We ensure that the logical consequences affected by the revision are precisely computable.

Step 3: Probe queries. For each postulate, we generate targeted queries:

Table 1: Probe query design for each AGM postulate. Each query type tests a specific rationality constraint on the model's revised beliefs.

Postulate	**Probe Design**
Success	Ask whether φ holds after revision. Expected: *Yes*.
Consistency	Ask both q and $\neg q$ for conclusions q derivable from φ. Expected: not both *Yes*.
Inclusion	Ask about a novel conclusion $q \notin \mathrm{Cn}(\mathcal{K} \cup \{\varphi\})$. Expected: *No* (model should not "hallucinate" new beliefs).
Vacuity	When φ is compatible with $\mathcal{K}$, ask about conclusions of $\mathcal{K} \cup \{\varphi\}$. Expected: unchanged plus φ's consequences.
Extensionality	Present logically equivalent φ and ψ in separate trials; compare responses. Expected: identical belief sets.
Preservation	Ask about a belief $\psi \in \mathcal{K}$ that is logically independent of φ. Expected: *Yes* (unrelated beliefs should survive revision).

Iterated revision scenarios. For the Darwiche–Pearl tests, we construct three-turn sequences: (1) establish $\mathcal{K}$, (2) revise with φ, (3) revise with ψ, where the relationship between φ and ψ is controlled (entailment for DP1, contradiction for DP2, independence for DP3/DP4). We include chains of up to four sequential revisions to test long-range coherence.

3.3 DATASET STATISTICS

AGM-BENCH contains 2,400 scenarios:

- **Single-step revision**: 1,200 scenarios (200 per postulate $\times$ 6 postulates), evenly split across depths $d \in \{1, 2, 3, 4\}$.
- **Iterated revision**: 800 scenarios (200 per Darwiche–Pearl postulate $\times$ 4 postulates), with chain lengths 2–4.
- **Control set**: 400 scenarios testing static reasoning (no revision) to establish a baseline for each model's deductive ability.

All ground-truth answers are computed by a symbolic pipeline independent of any LLM: each scenario's propositional skeleton is encoded in conjunctive normal form (CNF), and satisfiability and entailment checks are carried out with the Z3 SMT solver (De Moura & Bjørner, 2008), ensuring that label correctness does not depend on any model's output.

4 Evaluation Protocol

4.1 Multi-Turn Prompting

We implement belief revision as a structured multi-turn conversation:

1. **Establishment turn:** Present the initial knowledge base $\mathcal{K}$ as a list of facts and rules. Ask the model to confirm its understanding and answer preliminary queries to establish its initial belief state.
2. **Revision turn:** Present the new information φ with an explicit instruction: "*You have just learned the following new fact, which you should treat as definitely true. Update your beliefs accordingly.*" This ensures the model understands it should perform revision, not merely consider a hypothetical.
3. **Probe turn:** Ask the postulate-specific queries. Each query is a binary yes/no question about a specific proposition.

For iterated revision, we repeat steps 2–3 for each new piece of evidence. To mitigate ordering effects, we randomize the order of probe queries and average over two random orderings per scenario.

4.2 Scoring

For each postulate, we compute a *satisfaction rate*: the fraction of scenarios in which the model's responses are fully consistent with the postulate's requirements. We report macro-averaged rates across depths as well as per-depth breakdowns.

Formal scoring definitions. Let $R(q)$ denote the model's response to query q after revision:

- **Success score:** $\nvdash[R(\varphi) = \textit{True}]$.
- **Consistency score:** $\nvdash[\neg(R(q) = \textit{True} \wedge R(\neg q) = \textit{True})]$ for all tested q.
- **Inclusion score:** $\nvdash[R(q_{\text{novel}}) = \textit{False}]$ for hallucination probes q_{novel}.
- **Preservation score:** $\nvdash[R(q_{\text{indep}}) = \textit{True}]$ for independent beliefs q_{indep}.

4.3 Models Evaluated

We evaluate seven frontier models representing diverse architectures and training paradigms:

- **Standard chat models:** GPT-4o (Achiam et al., 2023), Claude 3.7 Sonnet (Anthropic, 2025), Llama 3.3 70B (Grattafiori et al., 2024), Qwen 2.5 72B (Yang et al., 2024).
- **Reasoning-trained models:** o3-mini (OpenAI, 2024), DeepSeek-R1 (Guo et al., 2025), Claude 3.7 Sonnet with extended thinking (Anthropic, 2025).

All models are accessed via their respective APIs with temperature 0 (greedy decoding) for reproducibility.

5 Experiments and Results

5.1 Single-Step Revision

Table 2 presents the postulate satisfaction rates for single-step revision across all models.

Table 2: Postulate satisfaction rates (%) for single-step belief revision, averaged across all depths. Higher is better. The best result per postulate is **bolded**; satisfaction rates below 70% are highlighted.

Model	Succ.	Cons.	Incl.	Vac.	Ext.	Pres.	Avg.
GPT-4o	96.2	91.4	62.8	78.3	73.1	64.7	77.8
Claude 3.7 Sonnet	**98.1**	93.7	71.2	82.6	78.4	70.3	82.4
Llama 3.3 70B	94.8	88.2	58.4	74.1	68.9	59.2	73.9
Qwen 2.5 72B	95.6	89.9	61.1	76.8	71.3	62.4	76.2
o3-mini	97.4	**95.2**	**76.8**	**86.4**	**82.7**	**73.9**	**85.4**
DeepSeek-R1	97.8	94.1	74.3	84.9	80.2	71.6	83.8
Claude 3.7 + Think	97.6	93.9	73.6	83.1	79.8	72.1	83.4

Key findings. Several patterns emerge:

Success and Consistency are largely satisfied. All models incorporate new information at >94% rates and avoid overt contradictions at >88% rates. This suggests that the basic mechanics of updating, accepting a new fact, is well-handled by current LLMs.

Inclusion is the hardest postulate. All models score below 77% on Inclusion, meaning they frequently introduce beliefs that do not follow from the combined old beliefs and new evidence. This "revision hallucination" effect is distinct from standard factual hallucination: the model generates logically unwarranted conclusions during the update process itself.

Preservation is systematically violated. Models frequently retract beliefs that are logically independent of the revision input. When told "Blickets are not daxes" (contradicting an earlier premise), models often also lose track of completely unrelated facts like "All feps are grobs." This *collateral damage* effect is most pronounced in weaker models (Llama: 59.2%) but persists even in the strongest (o3-mini: 73.9%).

Reasoning-trained models are uniformly better. o3-mini, DeepSeek-R1, and Claude with extended thinking all outperform their standard counterparts, with o3-mini achieving the highest average score (85.4%). The improvement is most notable on Inclusion (+14.0 pp over GPT-4o) and Vacuity (+8.1 pp).

5.2 Scaling with Reasoning Depth

Figure 1 illustrates how postulate satisfaction degrades as logical depth increases.

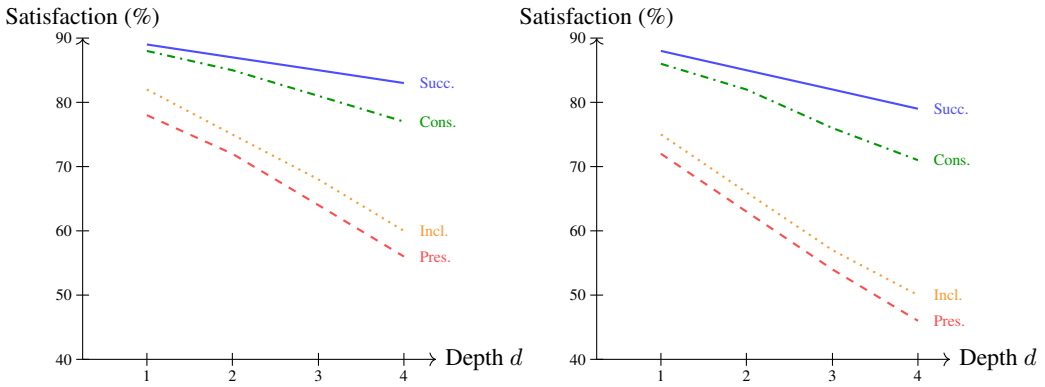
The most striking finding is the steep decline in *Preservation*: at depth 4, even o3-mini only preserves independent beliefs 56% of the time (down from 78% at depth 1). This suggests that when the logical structure is complex, models adopt an overly aggressive "reset" strategy, discarding beliefs they cannot confidently trace through the dependency chain.

5.3 Iterated Revision

Table 3 reports results for the Darwiche–Pearl iterated revision tests.

Belief inertia. The most common failure mode for DP2 (Irrelevance) is *belief inertia*: after revising with φ and then with ψ (where $\psi \models \neg\varphi$), models frequently retain consequences of φ that should have been retracted. In 34% of DP2 failures for GPT-4o, the model explicitly states the retracted proposition as still true, suggesting that the earlier revision "imprinted" on the model's context in a way that resists subsequent contradiction.

Iterated degradation in reasoning models. A surprising finding is that reasoning-trained models, while starting from a higher baseline, *degrade faster* under iteration than standard models. Figure 2 shows satisfaction rates as chain length increases from 2 to 4.



(a) o3-mini (b) GPT-4o

Figure 1: Postulate satisfaction rates as a function of reasoning depth d for o3-mini (left) and GPT-4o (right). *Success* degrades gracefully, but *Preservation* and *Inclusion* show steep declines at depth ≥ 3, indicating that models struggle to identify which beliefs are logically affected by the revision when the dependency chain is long.

Table 3: Darwiche–Pearl postulate satisfaction rates (%) for iterated revision (chain length = 2). Rates below 65% are highlighted.

Model	**DP1** (Recency)	**DP2** (Irrelevance)	**DP3** (Indep. Pres.)	**DP4** (Indep. Chg.)	**Avg.**
GPT-4o	71.3	58.6	54.2	52.8	59.2
Claude 3.7 Sonnet	74.8	63.1	59.7	56.3	63.5
Llama 3.3 70B	64.2	51.4	47.8	45.6	52.3
Qwen 2.5 72B	68.1	55.3	51.1	49.2	55.9
o3-mini	**79.6**	**68.4**	62.3	58.7	**67.3**
DeepSeek-R1	77.2	64.8	59.6	55.1	64.2
Claude 3.7 + Think	76.4	64.1	60.2	57.4	64.5

We hypothesize that reasoning models' extended chain-of-thought creates *longer dependency chains* in context, making it harder to surgically update specific beliefs without disturbing others. The additional reasoning tokens, while helpful for initial problem-solving, may act as "anchors" that resist revision, a form of the sunk-cost fallacy in computational reasoning.

5.4 Error Analysis

We conduct a qualitative analysis of 200 randomly sampled failure cases (across all models and postulates) and identify four recurring error patterns:

(1) Scorched-earth revision (41% of failures). The model, upon receiving contradictory information, discards not just the contradicted premise but large swaths of related (and unrelated) knowledge. This primarily drives Preservation violations.

(2) Revision hallucination (27%). The model introduces novel conclusions that follow from neither the original beliefs nor the new evidence, e.g., inventing a new relationship between entities not mentioned in the revision. This drives Inclusion violations.

(3) Belief inertia (22%). The model fails to fully propagate the consequences of the revision, retaining downstream beliefs that depend on the now-retracted premise. This is the primary driver of iterated revision failures.

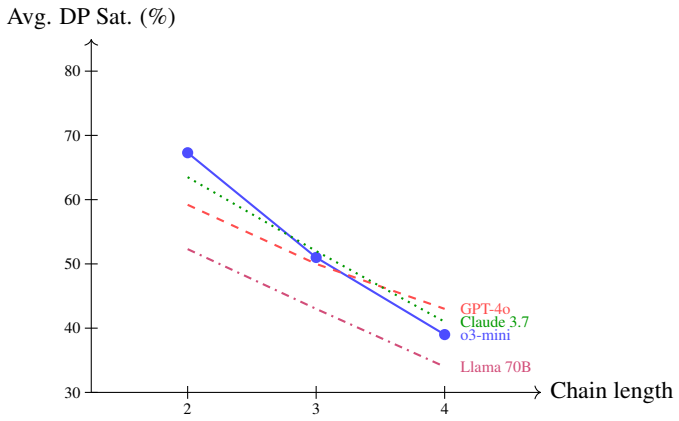


Figure 2: Average Darwiche–Pearl satisfaction rate vs. revision chain length. Reasoning-trained models (o3-mini) start higher but converge with standard models (GPT-4o) by chain length 4, exhibiting steeper relative degradation.

(4) Surface-form sensitivity (10%). The model gives different revision outcomes for logically equivalent inputs presented in different surface forms (e.g., "Not all Xs are Ys" vs. "There exists an X that is not a Y"), violating Extensionality.

6 Related Work

LLM logical reasoning. The survey by Cheng et al. (2025) provides a comprehensive taxonomy dividing the field into logical question answering and logical consistency. For question answering, benchmarks like FOLIO (Han et al., 2024), ProofWriter (Tafjord et al., 2021), and LogicBench (Parmar et al., 2024) test deductive reasoning over fixed premise sets. For consistency, Calanzone et al. (2025) introduces a neuro-symbolic loss for compositional consistency, and Liu et al. (2025) proposes metrics for transitivity, commutativity, and negation invariance. Lin et al. (2025) show that frontier LLMs still lack self-consistency on simple relational tasks and propose metrics to quantify inconsistency. All of these evaluate *static* consistency; none test dynamic revision.

Defeasible and non-monotonic reasoning. Rudinger et al. (2020) introduced defeasible NLI, and subsequent work has evaluated LLMs on default logic (Xiu et al., 2022), formal defeasible logic programs (Bao et al., 2025), and property inheritance with exceptions (Allaway et al., 2025). While defeasible reasoning involves retracting conclusions, these benchmarks test whether the model can *perform* defeasible inference, not whether its overall revision behavior satisfies formal rationality postulates.

Belief revision in AI. The AGM framework (Alchourrón et al., 1985) and its extensions (Darwiche & Pearl, 1997; Katsuno & Mendelzon, 1991; Boutilier, 1996; Nayak et al., 2003) have been foundational in knowledge representation. However, the connection to LLMs has been largely unexplored. Dalvi et al. (2022) studies continual learning from user feedback in a QA system, which is conceptually related but does not formalize the revision process in AGM terms. The dynamic epistemic logic tradition (van Benthem, 2011) provides complementary formal tools that could further extend our framework.

Chain-of-thought faithfulness. Recent work has shown that CoT reasoning is often unfaithful to models' actual computation (Arcuschin et al., 2025), with models arriving at correct answers via incorrect intermediate steps. Our finding that reasoning-trained models exhibit steeper iterated degradation may be related: the extended reasoning traces, while locally helpful, create rigid narrative structures that resist the surgical modifications required for rational revision.

LLM unlearning. A major trend in late 2025 and early 2026 has been *LLM unlearning*—the process of forcing a model to "forget" specific data. Our Preservation failure findings directly connect to this community: the "scorched-earth" revision we observe (models discarding unrelated knowledge when retracting a contradicted premise) is precisely the same problem unlearning researchers face—the catastrophic forgetting of knowledge that should remain intact when surgically removing target information (Gandikota et al., 2025; Hu et al., 2025). AGM-BENCH thus provides a formal, postulate-grounded lens on a failure mode central to both belief revision and unlearning.

7 DISCUSSION AND FUTURE WORK

Why do LLMs fail at rational revision? We conjecture that the core issue is architectural: transformer-based LLMs process the entire context holistically via attention, rather than maintaining a structured, modifiable belief store. When asked to revise a belief, the model must effectively "rederive" its entire belief state from the concatenation of original premises and the revision instruction. This makes surgical, minimal-change revision inherently difficult, as the model lacks a mechanism to identify precisely which prior conclusions depend on the retracted premise.

Toward revision-aware LLMs. Our results suggest several avenues for improvement: (1) *Neuro-symbolic hybrid approaches* that maintain an explicit belief store alongside the LLM, using a symbolic reasoner to compute the minimal revision and the LLM for natural-language generation. (2) *Revision-specific training*, using AGM-BENCH-like scenarios as fine-tuning data with AGM-compliant ground truth. (3) *External memory architectures* that represent beliefs as modifiable records rather than frozen context tokens.

System-level and neuro-symbolic baselines. Our study evaluates end-to-end LLMs; we do not report scores for alternative designs (e.g., explicit symbolic memory modules, retrieval-augmented stores, or a neuro-symbolic stack with a hard belief database). Running such systems on AGM-BENCH—using the same probes and Z3-derived labels—is natural follow-up work and would clarify how much of the observed gap is architectural versus purely neural.

Extensions: selective reuse and mixed retention. Current scenarios stress whether independent prior beliefs survive revision and whether iterated updates respect Darwiche–Pearl constraints. They do not yet systematically require *mixed-retention* behavior: surgically retracting some dependent conclusions while reusing other intact premises to support *new* downstream inferences after the update. Designing probe suites that force this selective reuse would tighten the link to deployed assistants that must both forget and continue reasoning from residual structure.

Limitations. Our benchmark uses synthetic, abstract reasoning scenarios to enable precise ground-truth computation. Whether the patterns we observe transfer to naturalistic domains (medical, legal, scientific reasoning) remains an open question. Additionally, our propositional fragment does not capture the full richness of first-order or modal belief revision. We view AGM-BENCH as a foundation that can be extended along both dimensions. While this evaluation covers the most advanced reasoning models available as of early 2026, the performance of rapidly emerging frontier architectures (e.g., GPT-5-nano) remains an important area for immediate follow-up studies.

Evidence as certain input. The prompting protocol asks models to treat each revision input as *definitively true*, in line with AGM *Success* and with our Boolean ground truth. This abstracts away settings where evidence is uncertain, unreliable, or must be arbitrated against other sources—precisely the regimes where epistemic decision-making, not only logical revision, matters. Extending the benchmark with graded or competing evidence would complement (not replace) the present logical core. Unconditional acceptance pairs naturally with SAT-based labels; modeling uncertainty would require a different semantic layer than propositional satisfiability alone.

8 CONCLUSION

We introduced AGM-BENCH, the first benchmark grounded in the AGM theory of belief revision for evaluating LLMs. Our evaluation of seven frontier models reveals that while LLMs handle basic

belief update mechanics (accepting new information, avoiding overt contradictions), they systematically fail at the subtler rationality requirements: changing minimally, preserving unrelated beliefs, and maintaining coherence across iterated revisions. The gap between current LLM behavior and rational belief dynamics is substantial, and narrows only partially with reasoning-specific training. We hope AGM-BENCH will catalyze research at the intersection of formal belief revision theory and large language model development, ultimately leading to AI systems that update their knowledge as rationally as they derive it.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Carlos E Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985.

Emily Allaway et al. Evaluating defeasible reasoning in LLMs. In *Proceedings of NAACL*, 2025.

Anthropic. The Claude model family. Technical report, Anthropic, 2025.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, et al. Chain-of-thought reasoning in the wild is not always faithful. In *Workshop on Reasoning and Planning for Large Language Models*, 2025.

Qiming Bao et al. Exploring formal defeasible reasoning of large language models: A chain-of-thought approach. *Knowledge-Based Systems*, 2025.

Craig Boutilier. Iterated revision and minimal change of conditional beliefs. *Journal of Philosophical Logic*, 25(3):263–305, 1996.

Aaron Calanzone, Aldo Pacchiano, and Jianfeng Gao. LoCo-LMs: Teaching LLMs to be logically consistent. *arXiv preprint arXiv:2502.01550*, 2025.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, et al. Empowering LLMs with logical reasoning: A comprehensive survey. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. Towards teachable reasoning systems: Using a dynamic memory of user feedback for continual system improvement. In *Proceedings of EMNLP*, 2022.

Adnan Darwiche and Judea Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, 89 (1-2):1–29, 1997.

Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient SMT solver. In *Tools and Algorithms for the Construction and Analysis of Systems*, volume 4963 of *Lecture Notes in Computer Science*, pp. 337–340. Springer, 2008.

Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. Erasing conceptual knowledge from language models. *arXiv preprint arXiv:2410.02760*, 2025.

Peter Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, 1988.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, He Zhang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, 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.

Sven Ove Hansson. *A Textbook of Belief Dynamics: Theory Change and Database Updating*. Kluwer Academic Publishers, 1999.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Hirofumi Katsuno and Alberto O Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52(3):263–294, 1991.

Zhenru Lin, Jiawen Tao, Yang Yuan, and Andrew Chi-Chih Yao. Existing LLMs are not self-consistent for simple tasks. *arXiv preprint arXiv:2506.18781*, 2025.

Yinhong Liu, Zhijiang Guo, Tianya Liang, Ehsan Shareghi, Ivan Vulić, and Nigel Collier. Aligning with logic: Measuring, evaluating and improving logical preference consistency in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38518–38539. PMLR, 2025.

Abhaya C Nayak, Maurice Pagnucco, and Pavlos Peppas. Dynamic belief revision operators. *Artificial Intelligence*, 146(2):193–228, 2003.

OpenAI. Learning to reason with LLMs. Technical report, OpenAI, 2024.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, et al. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13679–13707, 2024.

Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandni Chandra, and Yejin Choi. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4661–4675, 2020.

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)*, pp. 3621–3634, 2021.

Johan van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2011.

Yi Xiu, Zhong Xiao, and Yang Liu. LogicNMR: Probing the non-monotonic reasoning ability of pre-trained language models. In *Findings of EMNLP*, pp. 3616–3626, 2022.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

A NOTATION

Symbol	Meaning
$\mathcal{K}$	Belief set (deductively closed), before revision
$\mathcal{K} \circledast \varphi$	Belief set after revising with new information φ
$\mathrm{Cn}(\cdot)$	Deductive (logical) closure
φ, ψ	Formulas: revision inputs or probe queries
$\models$	Logical entailment (classical)

Table 4: Notation used in the main text for AGM belief revision.

B FULL AGM POSTULATES

For completeness, we state the standard eight AGM postulates for revision (K*1–K*8), following Alchourrón et al. (1985). Let $\mathcal{K}$ be a belief set and φ a sentence.

K*1 (Closure) $\mathcal{K} \circledast \varphi = \mathrm{Cn}(\mathcal{K} \circledast \varphi)$.

K*2 (Success) $\varphi \in \mathcal{K} \circledast \varphi$.

K*3 (Inclusion) $\mathcal{K} \circledast \varphi \subseteq \mathrm{Cn}(\mathcal{K} \cup \{\varphi\})$.

K*4 (Vacuity) If $\neg\varphi \notin \mathcal{K}$, then $\mathrm{Cn}(\mathcal{K} \cup \{\varphi\}) \subseteq \mathcal{K} \circledast \varphi$.

K*5 (Consistency) $\mathcal{K} \circledast \varphi = \mathcal{K}_\perp$ only if $\vdash \neg\varphi$ (i.e., revision yields inconsistency only if φ is itself inconsistent).

K*6 (Extensionality) If $\vdash \varphi \leftrightarrow \psi$, then $\mathcal{K} \circledast \varphi = \mathcal{K} \circledast \psi$.

K*7 (Superexpansion) $\mathcal{K} \circledast (\varphi \wedge \psi) \subseteq \mathrm{Cn}((\mathcal{K} \circledast \varphi) \cup \{\psi\})$.

K*8 (Subexpansion) If $\neg\psi \notin \mathcal{K} \circledast \varphi$, then $\mathrm{Cn}((\mathcal{K} \circledast \varphi) \cup \{\psi\}) \subseteq \mathcal{K} \circledast (\varphi \wedge \psi)$.

In the main text, we group K*3–K*4 under *Inclusion* and *Vacuity* respectively, and test Preservation (which follows from K*3, K*4, and K*6 jointly) as a separate, practically important criterion.

C Darwiche–Pearl Postulates

Let Ψ denote an epistemic state (from which a belief set $\mathcal{K}(\Psi)$ can be extracted), and $\Psi \circledast \varphi$ the revised epistemic state. The four Darwiche–Pearl postulates (Darwiche & Pearl, 1997) are:

DP1 If $\psi \models \varphi$, then $\mathcal{K}((\Psi \circledast \varphi) \circledast \psi) = \mathcal{K}(\Psi \circledast \psi)$.

DP2 If $\psi \models \neg\varphi$, then $\mathcal{K}((\Psi \circledast \varphi) \circledast \psi) = \mathcal{K}(\Psi \circledast \psi)$.

DP3 If $\varphi \in \mathcal{K}(\Psi \circledast \psi)$, then $\varphi \in \mathcal{K}((\Psi \circledast \varphi) \circledast \psi)$.

DP4 If $\neg\varphi \notin \mathcal{K}(\Psi \circledast \psi)$, then $\neg\varphi \notin \mathcal{K}((\Psi \circledast \varphi) \circledast \psi)$.

Intuitively, DP1–DP2 state that superseding evidence fully overrides the intermediate revision. DP3–DP4 state that a prior commitment to φ should not be undermined by an independent subsequent revision.

D Example Scenario

We provide a concrete example of an AGM-Bench scenario testing the Preservation postulate at depth 2.

Initial knowledge base $\mathcal{K}$:

Fact 1: Every blicket is a dax.
Fact 2: Every dax is a fep.
Fact 3: Every toma is a grob.
Fact 4: Wug is a blicket.
Fact 5: Zib is a toma.

Preliminary queries (establishing initial beliefs):

Q1: "Is Wug a fep?" Expected: Yes (via Facts 1, 2, 4).
Q2: "Is Zib a grob?" Expected: Yes (via Facts 3, 5).

Revision input φ:

"You have just learned the following new fact, which you should treat as definitely true: Wug is not a dax."

This contradicts the consequence of Fact 1 + Fact 4. A rational revision should retract "Wug is a dax" and its consequences ("Wug is a fep"), but preserve all beliefs about tomas and grobs.

Probe queries:

Success probe: "Is it true that Wug is not a dax?" Expected: Yes.
Preservation probe: "Is Zib a grob?" Expected: Yes (logically independent).
Inclusion probe: "Is Wug a toma?" Expected: No (not derivable from $\mathcal{K} \cup \{\varphi\}$).

Common failure (Preservation violation): In our experiments, 3 of 7 models answered "I'm not sure" or "No" to the Preservation probe, having apparently "forgotten" or doubted the Zib-grob chain after the Wug-related revision.