PROOFCOVER: AUDITING PROOF REPLACEMENT UNDER ANSWER-PRESERVING UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

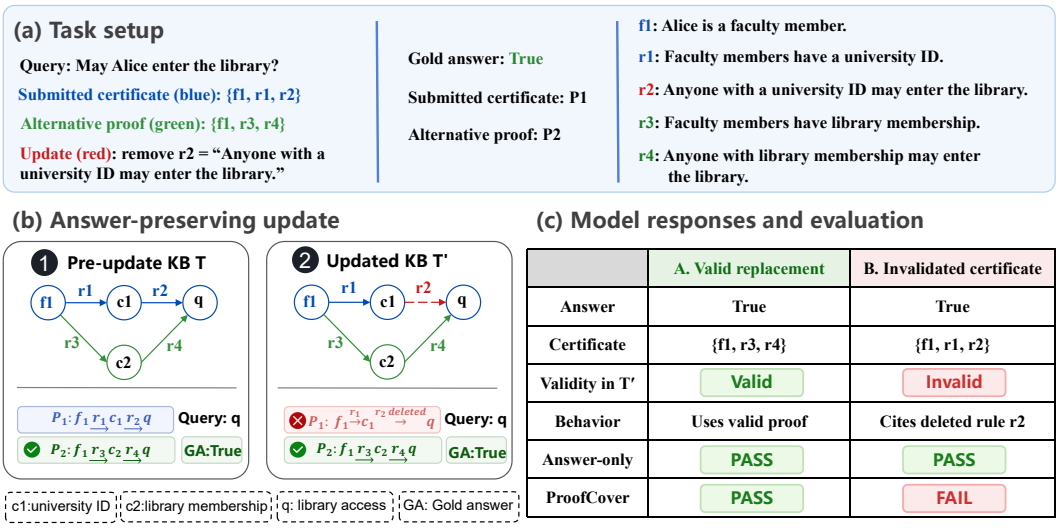
Evaluating rule-based reasoning by final-answer accuracy is standard practice, even when models also return the facts and rules supporting their predictions. However, after the rule world changes, the correct answer may remain the same while the submitted support becomes invalid, causing answer-centric evaluation to award full credit to an invalid certificate. Our key insight is that this is a structural observability gap: answer-only scoring cannot distinguish a valid current-world certificate from an invalid one when both predict the same correct label. Based on this observation, we introduce ProofCover, an evaluation framework and diagnostic benchmark that makes support validity under answer-preserving updates directly measurable. Instead of scoring only the final answer, ProofCover enumerates the complete set of minimal proofs for each query, uses this proof set to construct updates that preserve the correct answer while selectively invalidating the model's submitted certificate, and then applies a deterministic current-world verifier to check whether the model's new certificate remains valid in the updated world. On Formal-500, proof-set-aware construction achieves 100% success on Repair, compared with 16.6%–43.4% for three simple selection baselines. Under Fresh Adaptive Repair, false passes occur in all eight evaluated models: among outputs accepted by answer-only scoring, 15.3%–48.1% (median 27.1%) carry certificates that fail current-world verification. This pattern replicates on RuleTaker-MP, a RuleTaker-derived multi-proof benchmark constructed from official test data, where false-pass rates span 6.2%–22.0% across the same eight models. These results show that answer-only scoring can substantially overestimate support reliability in the evaluated update settings, while highlighting a certificate-validity gap that final-answer accuracy does not measure.

1 INTRODUCTION

Rule-based reasoning benchmarks test whether language models can derive conclusions from explicit facts and rules, with applications in question answering, planning, and knowledge-intensive decision making (Tafjord et al., 2021; Clark et al., 2020; Parmar et al., 2024). These benchmarks typically score models by final-answer accuracy, sometimes additionally requesting proofs or supporting facts and rules. However, real-world rule environments are not static: databases are updated, policies are revised, and facts are retracted over time in deployed settings. A model that correctly answers a query under the current world may cite support that becomes invalid after an update, even when the answer itself remains unchanged.

Existing approaches to evaluating rule-based reasoning primarily assess whether the model predicts the correct answer under a fixed world (Weston et al., 2015; Tafjord et al., 2021; Sinha et al., 2019). While this protocol can verify that a model derives the right conclusion, it cannot detect whether the stated support remains valid after the world changes. Figure 1 illustrates the problem. The original world contains two independent proofs of the query, and the model submits one valid certificate. Removing a cited rule leaves the alternative proof and the True answer intact but invalidates the submitted certificate. An answer-only evaluator awards full credit to both a valid current-world certificate and an invalid one because both predict the same correct label; the two outcomes are observationally indistinguishable under answer-only scoring, creating a structural observability gap.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



	A. Valid replacement	B. Invalidated certificate
Answer	True	True
Certificate	{f1, r3, r4}	{f1, r1, r2}
Validity in T′	Valid	Invalid
Behavior	Uses valid proof	Cites deleted rule r2
Answer-only	PASS	PASS
ProofCover	PASS	FAIL

Figure 1: Answer-preserving support invalidation. The update removes one cited rule while leaving an alternative derivation and the gold answer intact. Answer-only scoring accepts both outputs, whereas current-world certificate verification distinguishes valid from invalid support. Certificates are treated as externally verifiable support claims rather than evidence of latent reasoning.

This creates a specific and reproducible failure mode: *invalid-support false passes*. After an answer-preserving update, a model may submit a certificate that is no longer valid in the updated world — for example through a nonexistent identifier or incomplete support — while still predicting the correct answer. The root cause is that answer-only evaluation observes only the predicted label, not the certificate, so it cannot distinguish outputs whose stated evidence has been invalidated from those whose evidence has been correctly updated. This raises the central question: how can we measure whether a model's stated support remains valid after the world changes, when the answer itself provides no signal?

Our key insight is to use answer-preserving updates to change the proof obligation while keeping the oracle label fixed. This makes support adaptation observable independently of answer adaptation: preserving the correct label is insufficient because the updated world requires a different valid certificate. Complete minimal proof families make this intervention controllable while remaining evaluator-side and hidden from the model.

Based on this insight, we introduce ProofCover, an evaluation framework that turns support maintenance under answer-preserving change into a directly testable property. Given a world-query pair, ProofCover enumerates the minimal proof family, constructs a transition that changes the valid support obligation while preserving the oracle answer, queries the model on the updated world, and checks the returned certificate with a deterministic current-world verifier. The resulting counterfactual pair keeps the conclusion fixed while changing what counts as valid evidence, allowing answer correctness and support maintenance to be measured separately.

Across eight models spanning five families and scales from 4B to 70B, ProofCover finds invalid-support false passes in every evaluated model. On Formal-500, proof-set-aware construction achieves 100% success on Repair, compared with 16.6%–43.4% for three simple selection baselines. Under Fresh Adaptive Repair, 15.3%–48.1% (median 27.1%) of answer-only-accepted outputs carry certificates that fail current-world verification. This pattern replicates on RuleTaker-MP, a RuleTaker-derived multi-proof benchmark constructed from official test data, where false-pass rates span 6.2%–22.0% across the same eight models. Stateful evaluation, which supplies the previous certificate and explicit edit, reduces but does not eliminate false passes in most models, and in some cases redirects outputs to refusals without increasing successful certificate replacement.

Our contributions are as follows:

- **Problem and insight.** We identify a structural observability gap in answer-only evaluation: answer-preserving updates can change the valid proof obligation while leaving the scored label unchanged.
- **Method.** We introduce ProofCover, which uses complete minimal proof families to construct controlled changes in proof obligation and combines them with model-adaptive interventions and deterministic current-world verification.
- **Empirical evidence.** We provide an eight-model evaluation on Formal-500 and RuleTaker-MP showing that proof-set-aware construction reliably realizes intended transitions and that answer-only scoring misses substantial invalid support across model families, scales, and benchmark sources.

2 Related Work

Rule-Based Reasoning and Proof Generation. Rule-based benchmarks test whether models derive queries from explicit facts and rules (Weston et al., 2015; Clark et al., 2020; Sinha et al., 2019; Parmar et al., 2024), while proof-producing extensions require explicit derivations or support structures (Tafjord et al., 2021; Dalvi et al., 2021; Han et al., 2024; Sanyal et al., 2022). multiPRover generates multiple proof graphs (Saha et al., 2021), and LogicGraph studies solver-verified multi-path reasoning with exhaustive minimal proofs (Wu et al., 2026). These works establish the importance of explicit proof structure; ProofCover instead uses the proof family evaluator-side to construct controlled updates and audit whether a submitted certificate remains valid after the world changes.

Verification and Faithfulness. A broad line of work evaluates whether rationales, selected evidence, or chains of thought are sufficient, consistent, or faithful (Nye et al., 2021; Wei et al., 2022; Wang et al., 2023; Lanham et al., 2023; Turpin et al., 2023; Lyu et al., 2024). Neuro-symbolic and repair-oriented systems further combine language models with formal verification, theory resolution, or iterative feedback (Toroghi et al., 2024; Madaan et al., 2023; Shinn et al., 2023; Lightman et al., 2024). These methods ask whether an explanation is reliable or can be improved; ProofCover treats the certificate more narrowly as an externally checkable support claim and tests whether that claim still proves the answer in the current world.

Knowledge Revision and Counterfactual Evaluation. Truth-maintenance, model-editing, contrast-set, and counterfactual work studies how conclusions or behaviors change when facts, assumptions, or stored knowledge are revised (Doyle, 1979; Kaushik et al., 2020; De Cao et al., 2021; Meng et al., 2022; Mitchell et al., 2022; 2021; Zhong et al., 2023; Li et al., 2025; Wu et al., 2024). These interventions typically make the changed behavior itself observable. ProofCover instead holds the oracle answer fixed while changing which certificate is valid, isolating support maintenance as a separate evaluation object. Across these lines of work, two distinctions define our setting: the oracle answer remains invariant while the valid support obligation changes, and the model's submitted certificate is itself evaluated in the updated world. Appendix A19 compares the nearest settings along these axes.

3 Problem Formulation

We study reasoning over finite positive Horn rule environments. A world is $T = (F, R)$, where F is a finite set of facts and R a finite set of one-way rules. Given a query q, the model returns $o = (\hat{y}, C)$, where $\hat{y} \in \{\text{True}, \text{Unknown}\}$ and C is a set of fact and rule identifiers submitted as support. We write $T \models q$ when q is derivable from T; the oracle label is $y(T, q) = \text{True}$ if $T \models q$ and Unknown otherwise, where Unknown denotes non-derivability rather than explicit falsity. The distinction we care about is between getting this answer right and supporting it with a certificate that is valid in the world currently shown to the model.

Let $\text{IDs}(T)$ denote all fact and rule identifiers in T, and let $T[P]$ be the subenvironment induced by an identifier set P. We call P a proof of q when $T[P] \models q$, and a *minimal proof* when no strict subset of P still derives q. The complete family of minimal proofs is $\mathcal{P}(T, q)$. In the running example of Figure 1, the query has two minimal proofs, $P_1 = \{f_S, r_1, r_2\}$ and $P_2 = \{f_S, r_3, r_4\}$. This complete family is an evaluator-side object: the model is never asked to enumerate it and instead submits one certificate C, which need not be minimal.

For a True prediction, current-world certificate validity is

$$V_T(C, q) = \mathbb{1}[C \subseteq \mathrm{IDs}(T) \ \wedge \ T[C] \models q]\,. \tag{1}$$

The first condition rejects deleted, unavailable, or hallucinated identifiers; the second rejects sets of surviving identifiers that nevertheless do not jointly derive the query. For an Unknown prediction, validity requires $C = \emptyset$. Thus a certificate is an externally verifiable behavioral commitment about the stated support, not evidence about the model's latent computation or internal chain of thought.

Now consider an update $T \to T'$ applied after a model has submitted a valid certificate C. We focus on *answer-preserving certificate replacement*: the oracle answer remains True, the old certificate becomes invalid, and at least one alternative proof survives,

$$y(T, q) = y(T', q) = \text{True}, \qquad V_T(C, q) = 1, \quad V_{T'}(C, q) = 0, \qquad \mathcal{P}(T', q) \neq \emptyset. \tag{2}$$

The three requirements separate the phenomenon from ordinary answer-changing edits: the conclusion is unchanged, the particular support commitment is broken, and a valid replacement still exists. In Figure 1, deleting r_2 invalidates P_1 while P_2 continues to prove the same query.

After observing T', the model returns $o' = (\hat{y}', C')$. We record *Success* when $\hat{y}' = \text{True}$ and $V_{T'}(C', q) = 1$, and an *invalid-support false pass*, abbreviated *Stale*, when the answer remains correct but $V_{T'}(C', q) = 0$. We use *Stale* only as a compact outcome label: it denotes any answer-correct certificate that is invalid in the updated world and does not imply literal reuse of the pre-update certificate. Such an output may cite a removed identifier, introduce a nonexistent identifier, or cite only surviving identifiers whose union is insufficient to derive q.

Formally, an answer-only evaluator observes $\hat{y}'$ but not C' and therefore assigns the same score to any two outputs with the same predicted answer. Under the replacement condition, a valid updated certificate and an invalid one can consequently both pass whenever $\hat{y}' = y(T', q)$. Hence answer correctness does not imply $V_{T'}(C', q) = 1$. This is the observability gap targeted by ProofCover: the answer can remain unchanged precisely when the evidence required to justify it must change. A fixed-point formulation, the full replacement obligation, and a proof of answer-only indistinguishability are given in Appendix A2.

4 PROOFCOVER

ProofCover operationalizes the replacement condition in Section 3 as an evaluator-controlled audit. Oracle-side proof information is used only to construct and revalidate controlled world updates; the evaluated model sees the rendered world and returns an answer with one certificate. The framework therefore separates three roles that answer-only evaluation conflates: constructing a support-changing update, eliciting the post-update response, and checking whether the submitted support is valid in the current world. Figure 2 summarizes these routes.

Given a world-query pair (T, q), ProofCover enumerates the complete minimal-proof family $\mathcal{P}(T, q)$, applies a transition operator τ to produce T', reruns the symbolic oracle to confirm the intended answer and proof-set effect, queries the model, and verifies the returned certificate against T':

$$(T, q) \xrightarrow{\text{enum.}} \mathcal{P}(T, q) \xrightarrow{\tau} T' \xrightarrow{\text{query}} (\hat{y}', C') \xrightarrow{\text{verify}} \mathrm{Outcome}. \tag{3}$$

The complete proof family is available only to the constructor and verifier; it is never shown to the evaluated model or used as a target that the model must reproduce.

Controlled proof-set transitions. ProofCover defines operators that isolate proof preservation, removal, creation, and replacement. Let $\mathrm{RuleIDs}(T)$ be the rule identifiers in T, and let $\mathrm{Del}(T, r)$ remove rule r. A standard Repair transition selects r only when it appears in at least one but not every minimal proof, i.e., there exist $P_i, P_j \in \mathcal{P}(T, q)$ with $r \in P_i$ and $r \notin P_j$. It then constructs $T' = \mathrm{Del}(T, r)$ and recomputes the proof family. The edit is accepted only if at least one original proof containing r is destroyed while $\mathcal{P}(T', q) \neq \emptyset$, so the query remains provable through surviving support.

All-Proof-Cut provides the complementary destructive intervention: ProofCover computes a minimum rule-only hitting set H such that $H \cap P \neq \emptyset$ for every $P \in \mathcal{P}(T, q)$, removes all rules in H, and verifies that the updated proof family is empty. Invariant, Unknown-Invariant, and Graft supply preservation and creation controls. Unknown-Invariant adds a rule with no externally reachable

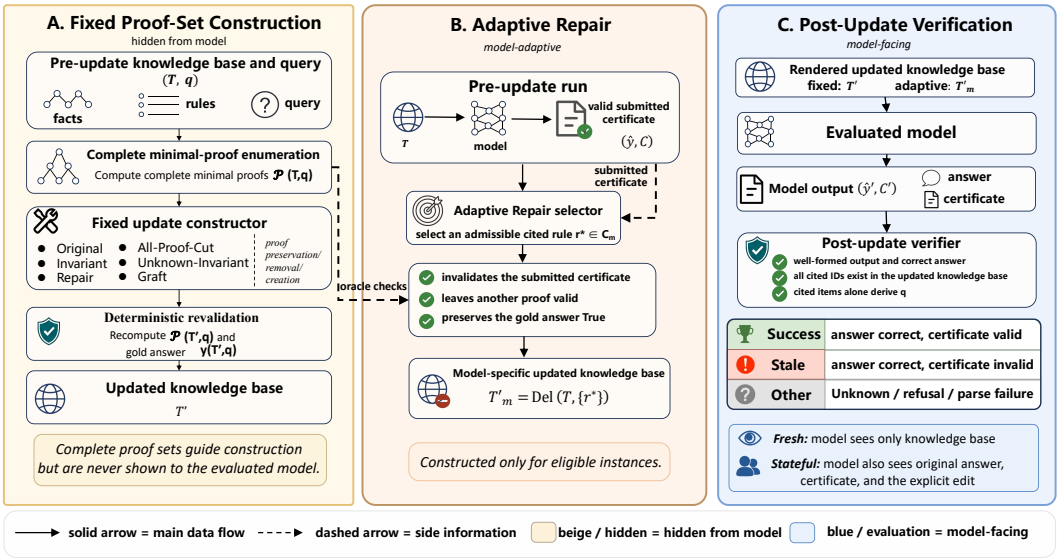


Figure 2: Overview of ProofCover. Fixed proof-set construction is evaluator-side and model-independent; Adaptive Repair conditions on a model's valid submitted certificate; post-update verification separately checks answer correctness and current-world certificate validity. Complete proof sets remain hidden from the evaluated model.

seed and therefore leaves the query non-derivable; Graft introduces a new verifiable route from a reachable fact to the query. In every case, the intended transition is accepted only after deterministic oracle revalidation rather than assumed from the construction procedure.

Model-adaptive repair. A fixed Repair edit can destroy one proof family member yet leave the particular certificate submitted by a model untouched. Adaptive Repair therefore conditions the intervention on that certificate. For model m, ProofCover first obtains an Original output $o_m = (\hat{y}_m, C_m)$. An instance can enter the adaptive subset only when the original oracle label is True, the model answers True, and $V_T(C_m, q) = 1$. ProofCover then searches the cited rules for deletions that invalidate C_m while preserving at least one alternative proof,

$$R_m(T, q, C_m) = \{r \in C_m \cap \mathrm{RuleIDs}(T) : V_{\mathrm{Del}(T,r)}(C_m, q) = 0,\ \mathcal{P}(\mathrm{Del}(T, r), q) \neq \emptyset\}.$$

A fixed total order over canonical rule identifiers selects $r^\star = \min_\prec R_m(T, q, C_m)$, giving the model-specific updated world $T'_m = \mathrm{Del}(T, r^\star)$. The ordering is frozen before model evaluation and is independent of the response later produced on T'_m.

The construction is retained only after all gates are rechecked: the Original oracle label and model answer are True; the Original certificate is valid; the removed rule is cited by that certificate; the same certificate becomes invalid in T'_m; at least one alternative minimal proof survives; and the updated oracle label remains True. These checks ensure that every retained case satisfies the replacement condition rather than merely exposing the model to an arbitrary edit. Because C_m is model dependent, the eligible set E_m is also model dependent; Adaptive Repair is therefore a conditional diagnostic and its rates are always reported with model-specific denominators.

We evaluate two presentation conditions on each accepted intervention. In *Fresh*, the model receives only T'_m. In *Stateful*, it additionally receives its Original answer and certificate, the explicit edit, and an instruction to revalidate its support in the new world. The oracle condition is identical across the two presentations, so their difference isolates whether explicit update context changes certificate replacement behavior.

Current-world verification. After the update, the model returns $o'_m = (\hat{y}'_m, C'_m)$. The verifier first checks that the response is parseable and uses an allowed label, then applies the current-world

criterion $V_{T'_m}(C'_m, q)$ from Equation 1: all cited identifiers must exist in T'_m and the cited subenvironment alone must derive q. An Unknown prediction must instead carry an empty certificate. These checks induce the Success/Stale/Other classes used in Section 5. For Unknown-Invariant, complete correctness requires Unknown with empty support; for Graft, it requires True with a valid certificate. The same verifier thus evaluates replacement, preservation, and creation against the exact world presented to the model. The complete operators, adaptive gates, algorithms, and verifier order are specified in Appendix A3.

5 Experimental Setup

This section fixes the benchmark, model, and evaluation protocol, then defines eligibility, metrics, and statistical reporting.

Datasets. Our natural-language rule environments are derived from ProofWriter (Tafjord et al., 2021). ProofCover uses the exact symbolic reasoner from Section 4 to compute oracle labels and complete minimal proof sets; every generated transition is retained only after oracle revalidation confirms its intended label and proof-set effect. Formal-500 contains 500 multi-proof True bases, each yielding Original, Repair, and All-Proof-Cut worlds (1,500 fixed-transition records); Adaptive Repair instances are generated separately per model from its submitted Original certificate. The Unknown component contains 855 naturally non-entailed bases, each yielding Original-Unknown, Unknown-Invariant, and Graft worlds (2,565 records). We also generate 700 symbolic and 700 graph-matched natural-language records from simple paired proof motifs; because these motifs are deliberately simple, they serve as mechanism checks and negative controls rather than evidence of structural difficulty. Dataset composition, proof-count/depth distributions, and sampling strata are reported in Appendix A4.

Cross-Benchmark Generalization. To test whether invalid-support false passes replicate beyond ProofWriter, we construct RuleTaker-MP from the official AllenAI RuleTaker V2020.2.5 test split. We first retain the 1,763 official-answer-True queries for which the local symbolic oracle verifies at least two minimal proofs, then deterministically select 900 instances stratified by query depth and proof-set size. The frozen ProofCover protocol is applied verbatim with zero per-benchmark tuning; all 900 frozen cases retain the official True label. Fresh Adaptive Repair results are reported in Section 6.

Models. We evaluate eight models spanning five model families and multiple scales: Qwen3-8B, Qwen3-4B, Qwen3-32B, Llama-3-8B-Instruct, Llama-3.3-70B-Instruct, Phi-4-mini-instruct, Gemma-3-4B-IT, and gpt-oss-20b. All models receive the same task instructions and return an answer and a certificate. The certificate vocabulary is open, allowing the verifier to observe deleted, nonexistent, and incomplete citations rather than ruling them out through the output schema.

Prompts and output protocols were selected on a development set whose source theories are disjoint from the formal evaluation set and were then held fixed. We use deterministic decoding and score only the final answer and submitted certificate; the frozen decoding setup, exact prompts/schema, parser, and per-model artifact manifest are in Appendices A5, A18, and A27.

Adaptive Repair is evaluated in the Fresh and Stateful conditions defined in Section 4. Exact model revisions, numerical precision, software versions, generation limits, and launch commands are reported in the reproducibility appendix.

Metrics. The 500 Formal-500 bases are candidate Original environments for Adaptive Repair, but eligibility is model conditional. For model m, an instance is included only when the model submits a valid Original certificate and a cited rule can be removed so that this certificate becomes invalid while an alternative proof and the True oracle label remain. Let E_m be the eligible set and $N_m = |E_m|$ its model-specific denominator. The denominator may be substantially smaller than 500 because Original certificates can be invalid, contain redundant complete proofs, or resist selective single-rule invalidation. Since E_m depends on the model's certificate, we report model-specific numerators and denominators and do not pool stale rates. Accordingly, these rates are conditional reliability diagnostics within each model's eligible subset, not direct cross-model performance rankings.

For the N_m eligible cases, Success denotes a correct True answer with a valid updated certificate, Stale is the compact label for a correct True answer with an invalid updated certificate, and

Other contains Unknown/refusal outputs and unparseable responses. Success and Stale exhaust the outputs accepted by answer-only scoring; we report the false-pass share of accepted outputs, Stale/(Success + Stale), alongside stale rates over N_m. Because both Success and Stale predict the correct label, $\text{AnswerAccuracy}_m = \text{SuccessRate}_m + \text{StaleRate}_m$. The Stale rate therefore measures answer-correct outputs accepted by answer-only evaluation but rejected by current-world verification.

For Unknown triplets, a base is fully correct only when Original-Unknown and Unknown-Invariant return Unknown with empty certificates and Graft returns True with a valid certificate. We report absolute numerators, denominators, and Wilson 95% confidence intervals. Fresh and Stateful outputs are paired by case identifier; detailed statistical tests and multiplicity correction are reported in the appendix.

6 RESULTS

The audit yields a consistent pattern across construction, post-update behavior, and controls: answer-preserving edits are easy to realize symbolically, but a correct post-update label often coexists with support that no longer proves it. We report the main effects below; Appendix A1 maps each empirical claim to its evidence, and Appendix A6 collects the complete result inventory.

6.1 CONSTRUCTION RELIABILITY AND PROOF-FAMILY CONTROL

Table 1 compares ProofCover with three selection baselines on Formal-500. A construction counts as successful only when oracle revalidation confirms both the intended answer and the intended proof-set effect.

Table 1: Construction reliability on Formal-500. Each task contains 500 base instances. Proof coverage is the fraction of original minimal proofs intersected by the selected edit; Repair requires a partial cut, whereas All-Proof-Cut requires complete coverage.

	Repair: True → True		**All-Proof-Cut: True → Unknown**	
Method	Success	Proof cov.	Success	Proof cov.
ProofCover	**500 (100.0%)**	52.7%	**500 (100.0%)**	100.0%
Random	161 (32.2%)	36.4%	116 (23.2%)	39.6%
Single-proof	217 (43.4%)	77.5%	295 (59.0%)	80.1%
Greedy	83 (16.6%)	92.7%	494 (98.8%)	99.7%

ProofCover succeeds on all 500 Repair and 500 All-Proof-Cut bases. The proof-family statistics explain why Repair is the harder construction: with one edit, ProofCover invalidates 1.49 minimal proofs while retaining 1.25 on average (52.7% proof coverage), whereas Greedy reaches 92.7% coverage and leaves only 0.22 proofs, over-cutting so aggressively that valid Repair success falls to 16.6%. An oracle-scan control nevertheless finds at least one valid Repair target in every world, with 2.29 targets per base on average, so the gap is about selecting the right partial cut rather than target scarcity. The five registered construction seeds are stable, and all methods saturate the deliberately simple synthetic Repair control (Appendices A7 and A23).

6.2 ANSWER-ONLY SCORING HIDES INVALID SUPPORT

Fresh Adaptive Repair is evaluated only after a model has first produced a valid Original certificate and a cited rule can be removed while preserving another proof. Table 2 therefore reports the model-specific denominator together with both the eligible-case stale rate and the false-pass share among outputs that answer-only scoring would accept.

Every model produces answer-correct outputs that fail current-world verification. These certificate-invalid outcomes (Stale) account for 13.4%–38.2% of eligible cases and 15.3%–48.1% (median 27.1%) of answer-accepted outputs. Original certificate validity itself ranges from 9.4% to 78.6%

Table 2: Fresh Adaptive Repair on Formal-500. Eligible sets are model-specific. *Valid orig.* is measured over the shared 500 Original worlds; *FP/accepted* is the share of answer-only-accepted outputs with invalid current-world certificates. Wilson intervals are in Appendix A14.

Model	Valid orig.	Eligible	Success	Stale	Other	Stale/elig.	FP/accepted
Qwen3-8B	153/500	134	68	27	39	20.1%	28.4%
Llama-3-8B	76/500	60	34	17	9	28.3%	33.3%
Qwen3-32B	97/500	82	40	11	31	13.4%	21.6%
Llama-3.3-70B	134/500	98	57	17	24	17.3%	23.0%
gpt-oss-20b	393/500	201	166	30	5	14.9%	15.3%
Phi-4-mini	101/500	86	50	22	14	25.6%	30.6%
Gemma-3-4B-IT	47/500	34	14	13	7	38.2%	48.1%
Qwen3-4B	148/500	126	72	25	29	19.8%	25.8%

Table 3: Cross-benchmark replication on RuleTaker-MP. *Valid orig.* is measured over the shared 900 Original worlds. The frozen ProofCover protocol is transferred without benchmark-specific tuning. *FP/accepted* again uses the answer-accepted denominator.

Model	Valid orig.	Eligible	Success	Stale	Other	FP/accepted
Phi-4-mini	256/900	219	145	41	33	22.0%
Gemma-3-4B-IT	220/900	125	88	21	16	19.3%
Qwen3-4B	330/900	216	146	33	37	18.4%
Qwen3-8B	378/900	236	155	42	39	21.3%
Llama-3-8B	282/900	197	139	38	20	21.5%
gpt-oss-20b	740/900	289	259	17	13	6.2%
Qwen3-32B	298/900	126	76	14	36	15.6%
Llama-3.3-70B	416/900	262	172	23	67	11.8%

over the shared 500 worlds, while 51.2%–87.6% of valid Original certificates contain a selectively removable cited rule. All 500 worlds admit a Repair target, so selectivity enters at the model-certificate level rather than the source-world level (Appendices A21 and A24).

The failure anatomy is more informative than the compact label *Stale*. Across all 162 Fresh certificate-invalid outputs, none re-cites the deleted target rule. Instead, 133 cite only identifiers that still exist but omit a complete proof; under a lenient "contains a surviving proof" check, at most one additional output becomes valid. For the seven non-gpt-oss families, stale certificates retain roughly 46–56% of their Original identifiers and are compact rather than diffuse, with model-level medians of three to five cited items. gpt-oss-20b shows a distinct serialization mode: 27 of its 30 stale certificates use bracket-wrapped nonexistent identifiers. The taxonomy, provenance, size distributions, qualitative cases, and a complete manual audit of the 13 Gemma stale outputs are in Appendices A11, A12, A13, A17, A22, and A28.

The same audit transfers to RuleTaker-MP without benchmark-specific tuning. Table 3 shows nonzero false-pass shares for all eight models, from 6.2% to 22.0%; the effect remains nonzero across populated proof-count and query-depth strata and shows no monotonic complexity trend.

RuleTaker-MP comes from a separate benchmark source, preserves the official True labels, and uses the frozen construction, prompt, and verifier. Complete per-model intervals and complexity strata are reported in Appendix A29.

6.3 Update Context, Creation, and Robustness

Explicit update history changes where failures go, but not in a uniform direction. On the same eligible cases, gpt-oss-20b moves from 166/30/5 Success/Stale/Other under Fresh to 200/0/1 under Stateful, and Qwen3-4B moves from 72/25/29 to 105/20/1. Llama-3.3-70B similarly raises Success

from 57 to 79. Qwen3-32B behaves differently: Stale falls from 11 to 0 while Success also falls from 40 to 30 and Other rises from 31 to 52; Llama-3-8B slightly increases Stale from 17 to 19. Paired McNemar tests with BH correction over the pre-specified 24-test family confirm several large redistributions but no common direction (Appendix A20); certificate-level transition traces and representative success/refusal conversions are in Appendices A12 and A17.

The preservation/creation controls show the same answer–certificate separation from another direction. On 855 Unknown triplets, Graft answer accuracy is roughly 89–99% across models, yet full-triplet certificate accuracy ranges from 4.8% to 97.8%. Under the semantically irrelevant Unknown-Invariant edit, exact retention of a previously valid Unknown output ranges from 31.9% to 99.4%, revealing substantial model-to-model sensitivity even when derivability is unchanged (Appendix A25).

Neither inference aggregation nor a learned judge reliably closes the gap. For Qwen3-8B, five-sample voting raises answer accuracy from 70.9% to 90.3% but increases Stale from 20.1% to 36.6%; paired transition matrices show that voting can move previously successful or refused cases into Stale. gpt-oss-20b is the clear exception, where voting clears the stale mode. Certificate-aware LLM judging increases stale recall—for Qwen3-8B, from 51.9% to 85.2%—but valid-certificate retention falls from 52.9% to 38.2%; no judge reaches the deterministic verifier's perfect stale-recall/valid-retention point (Appendices A8, A9, and A10). A paired symbolic/NL negative control further shows that the simple Repair motifs themselves are non-discriminating across representations, while refusal can be representation-sensitive for one model.

Finally, exploratory internal-state diagnostics ask whether the behavioral replacement failure has a readable representational correlate. Proof identity is linearly decodable at selected residual-stream readouts, but activation patching does not flip the submitted label. A direction-based update diagnostic separates Success from Stale for Qwen3-8B at its best layer (AUC 0.726, Mann–Whitney $p = 0.0025$) but not for Llama-3-8B (best AUC 0.479, $p = 0.594$). Because this line covers only two models and one seed, we treat it as secondary evidence rather than a headline mechanism claim (Appendices A15 and A16).

Across 2,000 world-cluster bootstrap resamples, false-pass intervals remain close to the case-level Wilson intervals, and depth-stratified shares remain nonzero in populated shallow and intermediate strata without a monotonic depth trend. The descriptive effect is therefore not an artifact of treating shared source worlds as independent cases or of one complexity regime (Appendix A26).

7 Discussion, Limitations, and Conclusion

The results separate two capabilities that final-answer accuracy merges: preserving the conclusion and maintaining valid support for that conclusion as the world changes. The dominant Fresh failure is partial support reconstruction rather than literal reuse of a deleted proof, while Stateful prompting and majority voting redistribute outcomes without consistently restoring a complete surviving derivation. Together, these findings make current-world certificate validity a useful evaluation axis in its own right: an explicit support commitment can be checked independently of whether the final label remains correct.

Limitations. The current scope is finite positive Horn reasoning with exact external certificates; extensions to negation, inconsistency, probabilistic inference, and open retrieval remain open. Adaptive subsets are model-conditional and interventions are single-rule, so the reported rates should be interpreted as conditional diagnostics rather than population prevalence. The nearest-work property comparison is in Appendix A19; formal definitions, construction details, and reproducibility material are provided in Appendices A2–A5, A18, and A27.

More broadly, the cross-benchmark replication and robustness controls support treating support validity as a first-class evaluation target beyond any single dataset or decoding choice. Across eight models, false-pass shares reach 15.3%–48.1% on Formal-500 and 6.2%–22.0% on RuleTaker-MP, showing that correct answers can coexist with support that no longer proves them in the current world. More generally, answer-preserving changes expose reliability failures that remain invisible whenever evaluation observes only the final label. ProofCover therefore separates final-answer correctness from the validity of the evidence a model explicitly commits to.

Reproducibility Statement

All evaluation code, prompts, model revisions, decoding settings, and launch commands, together with per-model outcome files and the reconciliation scripts used to derive every reported number, will be released publicly upon acceptance; frozen artifacts underpinning the reported tables and figures are archived with the submission. All results are produced by a deterministic pipeline (fixed seed, greedy decoding, symbolic verifier), so every number is exactly re-runnable from the released artifacts.

AI Use Statement

Generative AI tools assisted with language editing and LaTeX formatting. The authors reviewed all scientific claims, experimental results, citations, and AI-assisted text and take responsibility for the final content of the submission.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018. URL https://arxiv.org/abs/1610.01644.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks, 2023. URL https://arxiv.org/abs/2211.12588.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language, 2020. URL https://arxiv.org/abs/2002.05867.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7358–7370, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.585. URL https://aclanthology.org/2021.emnlp-main.585/.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6491–6506, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.522. URL https://aclanthology.org/2021.emnlp-main.522/.

Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272, 1979. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(79)90008-0. URL https://www.sciencedirect.com/science/article/pii/0004370279900080.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10764–10799. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23f.html.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *Second Agent Learning in Open-Endedness Workshop*, 2023. URL https://openreview.net/forum?id=WSrRF5Wy6v.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1229. URL https://aclanthology.org/2024.emnlp-main.1229/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data, 2020. URL https://arxiv.org/abs/1909.12434.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models, 2025. URL https://arxiv.org/abs/2503.10996.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723, June 2024. doi: 10.1162/coli_a_00511. URL https://aclanthology.org/2024.cl-2.6/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*, 2021.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15817–15831. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/mitchell22a.html.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models, 2021. URL https://arxiv.org/abs/2112.00114.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KS8mIvetg2.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13679–13707, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.739. URL https://aclanthology.org/2024.acl-long.739/.

Swarnadeep Saha, Prateek Yadav, and Mohit Bansal. multiPRover: Generating multiple proofs for improved interpretability in rule reasoning. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3662–3677, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.287. URL https://aclanthology.org/2021.naacl-main.287/.

Soumya Sanyal, Harman Singh, and Xiang Ren. FaiRR: Faithful and robust deductive reasoning over natural language. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1075–1093, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.77. URL https://aclanthology.org/2022.acl-long.77/.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4506–4515, Hong Kong, China, November 2019. Association for

Computational Linguistics. doi: 10.18653/v1/D19-1458. URL https://aclanthology.org/D19-1458/.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Armin Toroghi, Willis Guo, Ali Pesaranghader, and Scott Sanner. Verifiable, debuggable, and repairable commonsense logical reasoning via LLM-based theory resolution. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6634–6652, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.379. URL https://aclanthology.org/2024.emnlp-main.379/.

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, 2023. URL https://arxiv.org/abs/2305.04388.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2550–2575, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.167. URL https://aclanthology.org/2023.findings-emnlp.167/.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks, 2015. URL https://arxiv.org/abs/1502.05698.

Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 33402–33422. Curran Associates, Inc., 2024. doi: 10.52202/079017-1053. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3aa291abc426d7a29fb08418c1244177-Paper-Datasets_and_Benchmarks_Track.pdf.

Yanrui Wu, Lingling Zhang, Xinyu Zhang, Jiayu Chang, Pengyu Li, Xu Jiang, Jingtao Hu, and Jun Liu. Logicgraph : Benchmarking multi-path logical reasoning via neuro-symbolic generation and verification, 2026. URL https://arxiv.org/abs/2602.21044.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.971. URL https://aclanthology.org/2023.emnlp-main.971/.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2023. URL https://arxiv.org/abs/2205.10625.

A1 CLAIM–EVIDENCE MAP

This appendix maps the eight empirical claims to their formal definitions, construction checks, verifier outputs, and statistical controls. "Supported with caveat" denotes a claim whose evidence is subject to the model-conditional eligibility protocol, the stated representation or verifier limitation, or the finite positive-Horn scope.

Claim	Where in main text	Appendix evidence	Status
C1	Answer-only scoring cannot separate valid from invalid current-world certificates when both answers are correct.	Section 3; A2.	supported
C2	Proof-set-aware construction explicitly characterizes affected and surviving proof families; simple selection baselines are unreliable, while an oracle-scan control establishes Repair feasibility without proof-coverage ranking.	Section 6.1; 5, A23.	supported
C3	Under Fresh Adaptive Repair, every model produces answer-correct but currently-invalid certificates, comprising 15.3–48.1% of accepted outputs.	Section 6.2; 2, 3, A11.	supported
C4	Stateful context redistributes outcomes model-dependently; it neither uniformly repairs nor uniformly helps.	Section 6.3; 26, 27, A20.	supported with caveat
C5	Answer accuracy exceeds certificate accuracy on Graft and full-triplet scope across families and scales.	Section 6.3; 13, 11, 12.	supported
C6	Stale certificates are overwhelmingly incomplete supports over surviving IDs rather than re-citations of removed IDs; gpt-oss-20b adds a distinct format-violation mode.	Section 6.2; A11.	supported
C7	False passes are robust to majority voting and are not reliably detectable by LLM judges.	Section 6.3; A8.	supported with caveat
C8	In the symbolic/NL paired control, motifs are non-discriminating, while refusal is representation-sensitive for one model.	Section 6.3; A8.	supported with caveat

A2 FULL FORMULATION

A world is a finite positive Horn theory $T = (F, R)$. The fact set F consists of ground atoms, each assigned a unique identifier. The rule set R consists of single-conclusive definite clauses, also assigned unique identifiers. A rule has the form

$$r : \ a_1 \wedge \cdots \wedge a_k \to b,$$

where the a_i and b are atoms, possibly after grounding variables over the finite Herbrand universe of the instance. The identifiers of the facts and rules are disjoint, and

$$\mathrm{IDs}(T) = \mathrm{IDs}(F) \cup \mathrm{IDs}(R).$$

For $P \subseteq \mathrm{IDs}(T)$, $T[P]$ denotes the restricted theory containing precisely the facts and rules whose identifiers occur in P, together with all ground instances of the retained rules.

Forward chaining starts from the ground atoms in F. Whenever all premises of a ground instance of a rule in R have been derived, its conclusion is added to the derived set. Since the Herbrand base is finite and the theory is positive, this process reaches a least fixed point, written $\mathrm{cl}(T)$. A derivation of q is a finite acyclic proof tree whose leaves are fact atoms and whose internal nodes apply ground instances of rules. The identifier support of a derivation is the set of fact and rule identifiers used in that tree. Entailment is defined by fixed-point derivability:

$$T \models q \quad \Longleftrightarrow \quad q \in \mathrm{cl}(T).$$

The oracle answer is

$$y(T, q) = \begin{cases} \mathsf{True}, & T \models q, \\ \mathsf{Unknown}, & T \not\models q. \end{cases}$$

Thus Unknown is non-derivability under a closed-world convention over the finite Herbrand base. It does not introduce negated atoms and is distinct from negation-as-failure: the oracle does not infer $\neg q$, and failure to derive q is evaluated only after fixed-point saturation.

A proof support $P \subseteq \mathrm{IDs}(T)$ for q satisfies $T[P] \models q$. The complete minimal proof set is

$$\mathcal{P}(T, q) = \{P \subseteq \mathrm{IDs}(T) : T[P] \models q \ \wedge \ \forall P' \subsetneq P,\ T[P'] \not\models q\}\,.$$

For finite positive Horn theories, $\mathcal{P}(T, q)$ is finite and enumerable. A certificate C is an identifier set supplied by a model and need not be inclusion-minimal. Certificate validity is

$$V_T(C, q) = \begin{cases} 1, & C \neq \varnothing,\ C \subseteq \mathrm{IDs}(T),\ \text{and } T[C] \models q, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The verifier therefore has two distinct rejection modes for a nonempty certificate. Identifier-existence failure occurs when $C \not\subseteq \mathrm{IDs}(T)$, including a correct underlying identifier rendered with additional characters. Entailment failure occurs when every cited identifier exists but $T[C] \not\models q$. For an oracle answer Unknown, the required certificate is $C = \varnothing$; any nonempty certificate is rejected, while an empty certificate is the only representation of the Unknown answer.

Definition (certificate replacement obligation). Let T' be an updated world obtained from T for query q, with $y(T, q) = \mathsf{True}$. The update satisfies a certificate replacement obligation if all of the following hold:

1. the answer is preserved, $y(T', q) = y(T, q) = \mathsf{True}$;
2. at least one original minimal proof is destroyed, namely there exists $P \in \mathcal{P}(T, q)$ such that $T'[P] \not\models q$;
3. at least one distinct alternative remains, namely $\mathcal{P}(T', q) \neq \varnothing$.

For deletion-only updates, these conditions imply

$$\varnothing \neq \mathcal{P}(T', q) \subsetneq \mathcal{P}(T, q)$$

after minimal-support extraction. The obligation distinguishes a semantic answer-preserving transition from a certificate-preserving transition: the truth value remains fixed, while at least one previously sufficient support ceases to be sufficient and another support remains available.

Observation (answer-only indistinguishability). Answer-only evaluation cannot distinguish a valid updated certificate from an invalid certificate when both outputs carry the same answer. Let T' satisfy a certificate replacement obligation for T, q. An output pair (True, C) is answer-equivalent to (True, C') under any scoring rule that observes only the answer component, regardless of whether $V_{T'}(C', q) = 1$. In particular, an output that repeats a certificate valid in T but invalid in T' receives the same answer-only score as an output that supplies a valid certificate for a surviving proof.

Proof. The answer-only projection $\pi(\texttt{label}, C) = \texttt{label}$ removes the certificate component. Since the obligation preserves $y(T', q) = \mathsf{True}$, both the stale and replacement outputs project to the same value True. The distinction is present only in the current-world predicate $V_{T'}$, which is not evaluated by the answer-only projection. □

Table 4: ProofCover transition operators.

Transition	Intended effect
Original	No edit; anchors model behavior in the source world.
Invariant	Adds unreachable near-miss rule; preserves the proof set and True label.
Repair	Removes a nonshared rule; preserves True and requires support replacement.
All-Proof-Cut	Removes a rule-only hitting set; changes True to Unknown.
Original-Unknown	No edit to a naturally non-entailed query.
Unknown-Invariant	Adds an untriggerable near-miss rule; preserves Unknown.
Graft	Adds a reachable-to-query rule; changes Unknown to True.
Adaptive Repair	Removes a model-cited rule while preserving an oracle alternative proof.

Corollary. Among accepted outputs, let S denote outputs with a correct answer and a currently valid certificate, and let St denote outputs with a correct answer and a currently invalid certificate. Answer-only accuracy counts S and St identically. Consequently,

$$\frac{St}{S + St}$$

is exactly the false-pass share among accepted outputs. This quantity is undefined only when $S + St = 0$; all reported adaptive cases have accepted outputs for the corresponding model-conditional denominator.

Alternative-proof survival must be checked against the complete proof set. Consider two minimal proofs

$$P_1 = \{f_1, r_1\}, \qquad P_2 = \{f_2, r_2\},$$

where f_1, f_2 are fact identifiers and r_1, r_2 are rule identifiers. Deleting r_1 destroys P_1 but leaves P_2, so the answer remains True. A procedure that inspects only P_1 cannot establish whether an alternative survives. Conversely, if two minimal proofs share a rule, for example $P_1 = \{f_1, r\}$ and $P_2 = \{f_2, r\}$, deleting the rule r destroys both proofs; inspecting only one proof would fail to detect that the supposed alternative is also removed. The construction therefore enumerates all minimal supports before selecting an edit.

The proof sets used during construction are oracle-side objects. They are computed from the symbolic reasoner and are not model outputs. A certificate is instead a behavioral commitment: it is the identifier set the model supplies for the current world, and its validity is evaluated after the update against the exact current identifiers and current entailment relation.

A3 Full Method Specification

The construction pipeline operates on the finite symbolic representation of each world. It first computes the finite Herbrand base, saturates the theory by forward chaining, and records provenance for every derived atom. For each derivation of q, the pipeline collects the identifiers of its leaves and applied rules. Inclusion-minimal supports are then extracted by removing every support that strictly contains another support. This procedure enumerates all derivation paths relevant to minimal support extraction rather than selecting one proof by search order.

Termination follows from finiteness of the Herbrand base and monotonicity of positive Horn inference. At most $|\mathrm{HB}(T)|$ ground atoms can be newly derived, and each rule instance is checked against a fixed point. Provenance enumeration can have combinatorial worst-case growth in the number of alternative paths, but every support is drawn from the finite set of fact and rule identifiers. The constructed instances contain at most 16 facts, at most 9 rules, and at most 13 minimal proofs, which bounds the enumerated objects in the evaluated corpus.

Cover Repair deletes one rule r that occurs in at least one but not all complete minimal proofs. The resulting world must retain the intended True `label` while changing the proof set. All-Proof-Cut selects a minimum rule-only hitting set of the complete proof set and verifies that the updated query is no longer derivable, hence $y(T', q) = \mathsf{Unknown}$ and $\mathcal{P}(T', q) = \varnothing$. Invariant adds a rule that cannot fire because its premise is unreachable or because it is a self-loop. Unknown-Invariant applies

Algorithm 1 ProofCover fixed-transition construction

Require: World $T = (F, R)$, query q, operator o
Ensure: Updated world T', or REJECT

$\mathcal{P} \leftarrow \text{RE-ENUMERATE}(()T, q)$
$y \leftarrow y(T, q)$
if $o =$ CoverRepair **then**
    Choose $r \in R$ with r occurring in some but not all $P \in \mathcal{P}$
    $T' \leftarrow T \setminus \{r\}$
else if $o =$ AllProofCut **then**
    Choose a minimum rule-only hitting set $H \subseteq R$ for $\mathcal{P}$
    $T' \leftarrow T \setminus H$
else if $o =$ Invariant $\vee$ $o =$ UnknownInvariant **then**
    Add a rule r whose premise is unreachable or whose conclusion is its own premise
    $T' \leftarrow T \cup \{r\}$
else if $o =$ Graft **then**
    Choose a reachable fact a and add $r : a \to q$
    $T' \leftarrow T \cup \{r\}$
else
    return REJECT
end if
$\mathcal{P}' \leftarrow \text{RE-ENUMERATE}(()T', q)$
$y' \leftarrow y(T', q)$
if $o =$ CoverRepair **then**
    if $y' =$ True, $\mathcal{P}' \neq \varnothing$, and $\mathcal{P}' \neq \mathcal{P}$ **then**
        return T'
    end if
else if $o =$ AllProofCut **then**
    if $y' =$ Unknown and $\mathcal{P}' = \varnothing$ **then**
        return T'
    end if
else if $o =$ Invariant $\vee$ $o =$ UnknownInvariant **then**
    if $y' = y$ and $\mathcal{P}' = \mathcal{P}$ **then**
        return T'
    end if
else if $o =$ Graft **then**
    if $y' =$ True and $\mathcal{P}' \neq \varnothing$ **then**
        return T'
    end if
end if
return REJECT

the same non-interference condition to a non-entailed base. Graft adds a rule from a reachable fact to the query and verifies that the query becomes derivable. The Graft construction additionally requires that the added rule creates exactly a new minimal proof; this is checked by re-enumeration rather than inferred from the local rule shape.

The construction baselines use the same verification interface. Single-proof selects from one proof rather than the complete proof set. Random-k samples an edit without proof-set coverage guarantees, and Greedy coverage maximizes local coverage without enforcing the intended partial transition. The resulting construction comparison is reported in 5.

Table 5: Construction success on Formal-500 (500 bases per panel; 5 pre-registered seeds 20260717–20260721 for randomized methods; Successful = mean count of bases realizing the intended proof-set transition, Seed SD = population SD across seeds, CI = conservative envelope of per-seed Wilson 95% intervals. Upper panel: Repair (True→True with certified certificate invalidation); lower panel: All-Proof-Cut (True→Unknown). The panels are different tasks and are not directly comparable.

Method	Successful	Seed SD	Wilson 95% env.	Mean edit	Proof coverage
ProofCover (ours)	500	±0.000	[99.2, 100.0]	1.00	52.7%
Random-k	161	±0.021	[25.6, 39.7]	1.00	36.4%
Single-proof-k	217	±0.007	[37.9, 48.4]	1.00	77.5%
Greedy coverage	83	±0.000	[13.6, 20.1]	1.00	92.7%
Exact minimum cut	500	±0.000	[99.2, 100.0]	1.18	100.0%
Greedy coverage	494	±0.000	[97.4, 99.4]	1.18	99.7%
Random-k	116	±0.013	[18.2, 29.6]	1.18	39.6%
Single-proof-k	295	±0.007	[53.8, 64.4]	1.18	80.1%

Adaptive eligibility and repair. Adaptive Repair is conditioned on the model's output in the Original world. Let C_m be the parsed certificate produced by model m for (T, q). The eligible set is

$$\begin{aligned} E_m = \big\{&\text{the Original output parses under} \\ &\text{the frozen schema, has the oracle } \texttt{label}, \\ &\text{and } V_T(C_m, q) = 1\big\}, \\ N_m = |E_m|.& \end{aligned}$$

Eligibility is therefore model-conditional. The cases in E_m are not pooled across models, and a case can be eligible for one model but not another.

For an eligible case, define the candidate set

$$\begin{aligned} \mathcal{R}_m(T, q, C_m) = \big\{\, r \in R :\ & \\ & r \in C_m, \\ & r \in P \text{ for some } P \in \mathcal{P}(T, q), \\ & r \notin P' \text{ for some } P' \in \mathcal{P}(T, q), \\ & \mathcal{P}(T \setminus \{r\}, q) \neq \varnothing, \\ & y(T \setminus \{r\}, q) = \mathsf{True}\,\big\}. \end{aligned}$$

The candidate rule is

$$r^\star = \min_{\prec} \mathcal{R}_m(T, q, C_m),$$

where $\prec$ is a fixed total order over canonical rule identifiers. The identifiers are the dataset's rule labels, ordered lexicographically; this order is fixed before evaluation.

A candidate is accepted only when all seven gates below hold:

- **G1:** the Original model output parses under the frozen output schema;
- **G2:** the Original `label` is True and agrees with $y(T, q)$;

- **G3:** the Original certificate is valid, $V_T(C_m, q) = 1$;
- **G4:** $r^\star$ is an existing canonical rule identifier and $r^\star \in C_m$;
- **G5:** $r^\star$ occurs in at least one but not all members of $\mathcal{P}(T, q)$;
- **G6:** deleting $r^\star$ preserves the oracle answer and leaves at least one proof, $y(T \setminus \{r^\star\}, q) = \mathsf{True}$ and $\mathcal{P}(T \setminus \{r^\star\}, q) \neq \varnothing$;
- **G7:** re-enumeration confirms a changed proof set, $\mathcal{P}(T \setminus \{r^\star\}, q) \neq \mathcal{P}(T, q)$.

The adaptive construction therefore selects an update from the model's original commitment while preserving the answer and destroying only part of the complete proof set. If the candidate set is empty, the case is ineligible for adaptive evaluation.

Algorithm 2 Adaptive Repair construction

Require: Model m, world T, query q, Original output C_m, fixed order $\prec$
Ensure: Adaptive world T', removed rule $r^\star$, or REJECT

Parse the Original output into `label` ℓ_m and certificate C_m
if the output does not parse, $\ell_m \neq y(T, q)$, or $V_T(C_m, q) = 0$ **then**
  return REJECT
end if
$\mathcal{P} \leftarrow$ RE-ENUMERATE$(()T, q)$
$\mathcal{R}_m \leftarrow \varnothing$
for each rule $r \in R$ **do**
  if $r \in C_m$, r occurs in some but not all $P \in \mathcal{P}$ **then**
    $T_r \leftarrow T \setminus \{r\}$
    $\mathcal{P}_r \leftarrow$ RE-ENUMERATE$(()T_r, q)$
    if $y(T_r, q) = \mathsf{True}$, $\mathcal{P}_r \neq \varnothing$, and $\mathcal{P}_r \neq \mathcal{P}$ **then**
      $\mathcal{R}_m \leftarrow \mathcal{R}_m \cup \{r\}$
    end if
  end if
end for
if $\mathcal{R}_m = \varnothing$ **then**
  return REJECT
end if
$r^\star \leftarrow \min_\prec \mathcal{R}_m$
$T' \leftarrow T \setminus \{r^\star\}$
return $(T', r^\star)$

Fresh and Stateful presentations use the same adaptive world T'. In Fresh evaluation, the model receives the updated world and the frozen forward-verified prompt as a new instance. In Stateful evaluation, the Original interaction is retained and the update is appended verbatim:

$$
\begin{aligned}
&[\text{system}: \text{ You are a precise symbolic reasoner.}] \\
&+ [\text{user}: \text{ original forward_verified prompt for the Original world}] \\
&+ [\text{assistant}: \text{ model's original JSON prediction}] \\
&+ [\text{user}: \text{ WORLD UPDATE: the following IDs were removed: ¡ids¿.} \\
&\qquad\qquad \text{Your previous certificate is historical.} \\
&\qquad\qquad \text{Re-evaluate using only the complete current world below.} \\
&\qquad\qquad + \text{freshly rendered updated world}].
\end{aligned}
$$

Fresh and Stateful outcomes are paired by case_id. The paired Fresh–Stateful comparisons use exact McNemar tests, with the 24 tests consisting of three outcome categories for each of the eight models. The reported multiple-testing correction is the Benjamini–Hochberg procedure over this fixed family.

Current-world verification. Every prediction is evaluated against the world displayed at the point of prediction. The verifier applies the following sequence:

$$\text{parse} \longrightarrow \text{allowed } \texttt{label} \longrightarrow \text{identifier existence} \longrightarrow \text{entailment}.$$

Identifier existence is checked by exact string equality against the displayed identifiers. Entailment is checked by running the symbolic reasoner on the cited identifier set. An Unknown output must contain the empty certificate. A current-world output is classified as Success when its answer and certificate satisfy the current-world contract, as Stale when its answer is correct but its certificate fails current-world validity, and as Other for a refusal, an Unknown output in the adaptive True setting, or an unparseable output. Exact identifier matching makes bracket-wrapped strings such as `[rule3]` fail the existence check against `rule3`; this gpt-oss-20b format mode is analyzed in A11.

All model evaluations use the eight frozen local checkpoints from five families: Qwen3-8B, Qwen3-4B, Qwen3-32B, Llama-3-8B-Instruct, Llama-3.3-70B-Instruct, Phi-4-mini-instruct, Gemma-3-4B-IT, and gpt-oss-20b. The checkpoints are served by vLLM with bf16/fp16 computation; gpt-oss-20b uses native MXFP4 with 20B total and 3.6B active parameters. The random seed is 20260718, temperature is zero, and the context length is 4096. The generation limit is 1024 tokens except for Llama-3-8B-Instruct and Llama-3.3-70B-Instruct, which use 512, and gpt-oss-20b, which uses 2048. The forward_verified prompt variant was selected on a development set with disjoint source theories and then frozen.

A4 Dataset Construction

The source theories are ProofWriter-style finite theories (Tafjord et al., 2021). The pipeline re-derives the oracle labels and complete minimal proof sets with the symbolic reasoner. No source dataset `label` is trusted without oracle revalidation. Each retained world is therefore accompanied by a fixed-point entailment result, a complete minimal proof set when the query is derivable, and a construction-specific verification record.

Formal-500. Formal-500 contains 500 multi-proof True bases satisfying $|\mathcal{P}(T, q)| \geq 2$. Selection is stratified by source configuration. The strata and their base counts are depth-1: 54, depth-2: 121, depth-3: 51, depth-3ext: 59, depth-3ext-NL: 57, and depth-5: 158. Each base contributes three worlds, for 1500 records. The worlds contain 1–16 facts, with median 10, and 2–9 rules, with median 8. The minimal-proof counts are distributed as

$$\{2 : 301,\ 3 : 109,\ 4 : 54,\ 5 : 13,\ 6 : 14,\ 7 : 3,\ 8 : 3,\ 9 : 1,\ 12 : 1,\ 13 : 1\},$$

and the query-depth counts are

$$\{1 : 175,\ 2 : 174,\ 3 : 112,\ 4 : 17,\ 5 : 22\}.$$

Table 6: Benchmark composition. Formal-500 and Unknown bases derive from ProofWriter theories; synthetic sets are deliberately simple negative controls.

Component	Bases	Records	Content
Formal-500	500	500	multi-proof True bases; Original/Repair/All-Proof-Cut (1500 records)
Unknown triplets	855	855	naturally non-entailed bases; Original-Unknown/Unknown-Invariant/Graft (2565 records)
Synthetic symbolic	100	100	7 worlds × 100 = 700 records, paired proof motifs
Synthetic NL	100	100	graph-matched NL renderings of the same 100 bases × 7 worlds = 700

Table 6 gives the collection-level accounting: the collection or split, the number of bases, the number of worlds per base, and the resulting number of records. The base count refers to distinct source theories before world expansion; the world count counts the Original and constructed variants associated with each base; the record count is the number of model-evaluation records after expansion.

Table 7 reports the formal distributional constraints. Its configuration or stratum column identifies the source configuration, its base-count column gives the number of selected bases, its fact and rule columns describe the world-size support, and its proof-count and query-depth columns report the corresponding symbolic complexity distributions. These columns separate source-stratum balance from proof multiplicity and query derivation depth; neither dimension is inferred from model behavior.

Table 7: Formal-500 Original-base distributions. Minimal-proof count $|\mathcal{P}(T, q)|$ and query dependency depth by ProofWriter source configuration.

	$\lvert\mathcal{P}(T,q)\rvert$										
Count	2	3	4	5	6	7	8	9	12	13	total
Bases	301	109	54	13	14	3	3	1	1	1	500

	query depth $qdep$					
Depth	1	2	3	4	5	total
Bases	175	174	112	17	22	500

Table 8: Formal-500 sampling strata by ProofWriter source configuration. Facts per base: min 1 / median 10 / max 16; rules per base: min 2 / median 8 / max 9.

Configuration	Bases	Share
depth-5	158	31.6%
depth-2	121	24.2%
depth-3ext	59	11.8%
depth-3ext-NL	57	11.4%
depth-1	54	10.8%
depth-3	51	10.2%

Table 8 gives the configuration-level construction accounting. The configuration column identifies the source setting, the base column counts retained bases, the world or record column counts expanded instances, and the fact, rule, minimal-proof, and query-depth columns report the symbolic properties used for revalidation. The table is descriptive rather than a model result: all listed quantities are computed before model inference.

The repair panel uses single-rule deletion with a True-to-True target. ProofCover realizes all 500 intended transitions, whereas Single-proof, Random-k, and Greedy coverage realize 217, 161, and 83, respectively. The corresponding rates are 100.0%, 43.4%, 32.2%, and 16.6%. Under Repair edits, mean proof coverage is 52.7% for ProofCover and 92.7% for Greedy coverage, demonstrating that the latter frequently over-cuts the intended partial transition. The matched cut panel uses a True-to-Unknown target: Exact-min-cut realizes 500 cases, Greedy 494, Single-proof 295, and Random-k 116. On the synthetic symbolic Repair control, all methods realize 100%, so that control does not discriminate structural proof-set handling.

Unknown-855. Unknown-855 contains 855 naturally non-entailed bases. Each base is expanded into the triplet

$$\{\mathsf{Original\text{-}Unknown}, \mathsf{Unknown\text{-}Invariant}, \mathsf{Graft}\},$$

yielding 2565 records. The Original-Unknown member provides a non-entailed anchor on which refusal or explicit Unknown behavior can be measured. Unknown-Invariant adds an irrelevant rule that cannot fire and tests whether the model preserves the non-entailed status under an answer-preserving structural change. Graft adds a rule from a reachable fact to the query and tests proof creation from the same base.

Each Unknown-855 world is re-verified by complete re-enumeration. The Invariant rule must leave $\mathcal{P}(T, q)$ unchanged. The Graft rule must create exactly a new minimal proof, and the resulting query must be derivable. The triplet is therefore defined by symbolic transition constraints rather than by the surface form of the inserted rule.

The M3 evaluation reports both answer and certificate behavior on the full triplets and on Graft. Full-triplet answer accuracy ranges from 7.6% for Llama-3-8B-Instruct to 97.9% for gpt-oss-20b. On Graft, the answer-minus-certificate gaps are 28.0 percentage points for Llama-3-8B-Instruct, 19.1 for Qwen3-8B, 32.7 for Qwen3-4B, 65.3 for Gemma-3-4B-IT, 19.9 for Phi-4-mini-instruct, 4.0 for Qwen3-32B, 1.7 for Llama-3.3-70B-Instruct, and 0.1 for gpt-oss-20b. Parse errors total 18/2565 for gpt-oss-20b, 2 for Qwen3-8B, and 0 for every other model.

Model	Family	Params	Quantization	Seed	Temperature	num_ctx	num_predict
Qwen3-8B	Qwen	8B	FP16	20260718	0	4096	1024
Qwen3-4B	Qwen	4B	BF16	20260718	0	4096	1024
Qwen3-32B	Qwen	32B	FP16	20260718	0	4096	1024
Llama-3-8B-Instruct	Llama	8B	FP16	20260718	0	4096	512
Llama-3.3-70B-Instruct	Llama	70B	FP16	20260718	0	4096	512
Phi-4-mini-instruct	Phi	not specified	BF16	20260718	0	4096	1024
Gemma-3-4B-IT	Gemma	4B	BF16	20260718	0	4096	1024
gpt-oss-20b	GPT-OSS MoE	20B/3.6B active	MXFP4 native	20260718	0	4096	2048

Table 9: Frozen decoding configuration. The GPT-OSS entry gives total and active MoE parameter counts.

Synthetic paired control. The synthetic control contains 100 paired bases, each rendered symbolically and as a graph-matched natural-language version. Each base has seven worlds, producing 700 symbolic and 700 graph-matched NL records. The bases use trivial two-proof motifs. The control is restricted to mechanism checks and negative controls; it was never used to claim structural difficulty. Repair-world agreement is 100% in both representations for all models, showing that the motifs do not discriminate between the two representations under the Repair transition. Original-Unknown answer behavior is representation-discordant for Qwen3-8B on 81/100 bases, with NL-only Unknown outputs and exact McNemar $p = 8.3 \times 10^{-25}$; every other model has 0/100 discordant cases with $p = 1$. Thus refusal on these trivial motifs is representation-sensitive for one model despite non-discriminating Repair behavior.

Certificate checking is a deterministic, symbolic instance of entailment verification (Thorne et al., 2018; Dalvi et al., 2021) restricted to the positive Horn fragment, where validity is exact rather than learned.

Adaptive repair files. The adaptive world file formal_500_adaptive.jsonl stores base_id, removed_ids, surviving_proofs, and `label`. base_id preserves the pairing to the source base, removed_ids records the selected update, surviving_proofs records the oracle-side proof set after the update, and `label` records the revalidated current-world oracle answer. Model certificates are not substituted for surviving_proofs; the latter are generated by the symbolic reasoner and are used only to define and verify the transition.

A5 Experimental Protocol and Reproducibility

All experiments use the Formal-500 construction benchmark and the frozen `forward_verified` prompt variant. Formal-500 contains 500 ProofWriter-derived multi-proof True bases, stratified by query depth and theory depth, with three worlds per base and hence 1500 records. The bases contain between 1 and 16 facts and between 2 and 9 rules. The Unknown evaluation uses 855 naturally non-entailed bases in the Original-Unknown, Unknown-Invariant, and Graft conditions, yielding 2565 records. The synthetic control contains 100 paired bases evaluated in seven worlds under both symbolic and graph-matched natural-language representations.

Decoding. All models were served from local Hugging Face checkpoints with vLLM. The decoding configuration was greedy ($temperature = 0$) with seed 20260718 and $num_ctx = 4096$. The per-model output limits and checkpoint formats are shown below; the parameter count for Phi-4-mini-instruct is not specified in the frozen experiment record and is therefore not inferred from its model name.

There is one decoding seed per model. The reported uncertainty intervals are Wilson 95% confidence intervals over evaluated cases, not seed-replicate intervals. Greedy decoding makes seed-induced variation plausibly small in this setting, but the protocol does not establish that it is negligible and does not support a claim about between-seed variance.

Serving and certificate serialization. Each model was exposed through a vLLM OpenAI-compatible endpoint. The output schema constrained the certificate to a JSON array with `uniqueItems=true` and `maxItems=|facts|+|rules|`. Certificate items were typed as

arbitrary strings rather than being restricted to the identifiers present in the rendered world. This open vocabulary is intentional: the verifier must be able to observe a deleted identifier, an identifier that never existed, or an incomplete citation rather than having the schema reject the certificate before verification.

The optional `id_shape_constraint` pattern was not enabled in the reported runs. In particular, it would have masked the bracket-format failures of gpt-oss-20b, in which identifiers copied from the display format such as `[rule3]` were returned instead of the verbatim identifier `rule3`. Those failures are part of the observed failure taxonomy in Section A11; excluding them at generation time would change the measured behavior rather than merely improve parsing.

Broad answer-centric evaluation suites document the evaluation setting this protocol departs from (Hendrycks et al., 2021; **?**; **?**); contamination analysis is a general concern for frozen benchmarks (Oren et al., 2024).

Prompt selection and scoring. The `forward_verified` prompt was selected deterministically on a development set with disjoint source theories and then frozen for all reported evaluation cases. Only the final answer and certificate were scored. Intermediate reasoning, if produced by a model, was not used by the verifier or by the outcome labels.

For a world $T = (F, R)$, let $\mathrm{IDs}(T)$ be the identifiers of its facts and rules. For an identifier set P, $T[P]$ denotes the subtheory induced by P. Let $\mathcal{P}(T, q)$ denote the admissible proof-support sets for query q, and let $V_T(C, q)$ be the deterministic certificate predicate

$$V_T(C, q) \;=\; \big(C \subseteq \mathrm{IDs}(T)\big) \wedge \big(T[C] \models q\big).$$

The answer `label` is denoted $y(T, q)$. For model m, E_m is the model-specific set of cases passing the eligibility gates in Section A3, and $N_m = |E_m|$. This notation makes explicit that the evaluated case sets are model-conditional.

The Fresh condition evaluates a response against the original world. The Stateful condition appends the model's original response to a new user message and then asks for reevaluation after the specified edit. The message construction is exactly

```
[system: "You are a precise symbolic reasoner."]
```

followed by the original `forward_verified` user prompt and the model's original JSON prediction. The appended update is exactly

```
[user: "WORLD UPDATE: the following IDs were
removed: <ids>. Your previous certificate is
historical. Re-evaluate using only the complete
current world below." + freshly rendered
updated world].
```

Thus the Stateful condition preserves the original conversational context while supplying a newly rendered current world.

After eligibility gating, outcomes are reported as Success, Stale, and Other. Success requires an answer and certificate that pass the current-world verification; Stale denotes an answer-preserving but invalid certificate under the updated world; Other contains Unknown refusals and unparseable outputs. The false-pass share is

$$\frac{\mathrm{Stale}}{\mathrm{Success} + \mathrm{Stale}}.$$

Intervals are Wilson 95% intervals computed over cases within each model; models are never pooled. Fresh-versus-Stateful outcome comparisons use paired exact McNemar tests over the 24-test family consisting of eight models and three outcomes, with Benjamini–Hochberg correction. The complete adjusted family is carried in Table 27; the Stateful outcome summary is in Table 26.

What would change the conclusion. The protocol studies single-rule edits, so it does not establish behavior under interacting or multi-rule updates. The adaptive sets are model-conditional and are not pooled, which limits cross-model comparisons to the frozen eligibility-defined cases. Finally,

the prompt was selected on a development set and then frozen; a different prompt-selection process could change absolute rates. These limitations would matter most if multi-edit updates reduced stale certificates, if a shared non-adaptive case set reversed the model ordering, or if prompts selected without development-set information eliminated the answer-certificate gap.

A6 Complete Results

The construction table (Table 5) separates the ability to preserve the target answer from the amount of proof material removed. ProofCover preserves the Repair answer across the full panel while retaining deliberately partial proof coverage, whereas Greedy coverage reaches substantially higher coverage and therefore over-cuts relative to the intended construction. The contrast between 52.7% and 92.7% is consequently evidence that answer preservation does not require deleting nearly all alternative support. The small seed standard deviations for the randomized methods indicate that this construction contrast is not driven by one of the five registered seeds. The synthetic symbolic Repair control remains a negative control: all methods preserve the answer there.

A7 Seed Stability and Proof-Family Effects in Construction

Table 10 reports construction results over 500 bases for seeds 20260717–20260721. In the repair panel, ProofCover succeeds on 500/500 bases for every seed, with minimum and maximum both 500 and seed standard deviation 0.000. Its construction guarantee is therefore seed-invariant on these artifacts.

The randomized baselines vary within comparatively narrow seed ranges but remain well below ProofCover. Random-k repair has mean success 161, a per-seed range of 147–177, and standard deviation 0.021; Single-proof-k has mean 217, range 211–220, and standard deviation 0.007. Greedy coverage has mean 83 and range 83–83.

The proof-family columns measure the repair objective directly. With mean edit size 1.00, ProofCover invalidates 1.49 complete proofs per base while retaining 1.25, corresponding to 52.7% proof coverage. Greedy coverage reaches 92.7% but leaves only 0.216 proofs surviving, illustrating that answer preservation does not require removing nearly all alternative support and that maximizing coverage can violate the repair contract.

The cut panel exhibits the complementary pattern. Exact minimum cut succeeds on 500/500 bases for every seed, with mean edit size 1.18, 2.75 invalidated proofs, 0.00 surviving proofs, and 100.0% coverage; Greedy coverage attains mean success 494 and 99.7% coverage, while Random-k attains mean 116, range 108–128, and 39.6% coverage, and Single-proof-k attains mean 295, range 291–301, and 80.1% coverage. These seed results concern dataset construction only; all statistical analyses of model behavior use frozen artifacts.

Table 10: Construction behavior across the five pre-registered seeds on Formal-500 (500 bases per method). Success is the mean count of bases realizing the intended transition; Min/Max give the per-seed spread. Proof-family columns report the mean number of complete proofs invalidated and surviving per base. Repair targets a True→True edit that destroys part of the proof family; cut targets a True→Unknown edit.

Regime	Method	Success	Min	Max	Edit	$\overline{\text{Inval.}}$	$\overline{\text{Surv.}}$
Repair	ProofCover (ours)	500	500	500	1.00	1.49	1.25
Repair	Random-k	161	147	177	1.00	0.99	1.76
Repair	Single-proof-k	217	211	220	1.00	2.08	0.67
Repair	Greedy coverage	83	83	83	1.00	2.53	0.22
Cut	Exact minimum cut	500	500	500	1.18	2.75	0.00
Cut	Greedy coverage	494	494	494	1.18	2.73	0.01
Cut	Random-k	116	108	128	1.18	1.09	1.65
Cut	Single-proof-k	295	291	301	1.18	2.17	0.58

The two M3 tables expose a separation between identifying the correct Unknown-triplet answer and producing a certificate that satisfies the corresponding proof contract. The answer-only ranking is

Table 11: Per-model answer accuracy on the 855 Unknown bases, by world, with total parse errors over 2,565 records. Original-Unknown and Unknown-Invariant are correct when the model answers Unknown; Graft when it answers True; Full triplet requires all three worlds correct.

Model	PE	Orig-Unk	Unk-Inv	Graft	Full triplet
Qwen3-8B	2	89.7%	47.6%	97.9%	45.3%
Qwen3-4B	0	79.2%	49.8%	92.6%	42.2%
Qwen3-32B	0	96.8%	95.1%	89.1%	82.5%
Llama-3-8B-Instruct	0	30.1%	10.4%	97.1%	7.6%
Llama-3.3-70B-Instruct	0	69.0%	54.7%	96.8%	46.9%
Phi-4-mini-instruct	0	86.9%	56.8%	87.0%	44.3%
Gemma-3-4B-IT	0	44.7%	32.2%	91.7%	15.8%
gpt-oss-20b	18	99.1%	99.2%	99.3%	97.9%

Table 12: Per-model certificate-conditional accuracy on the 855 Unknown bases (answer correct and certificate requirement satisfied: empty for Unknown, valid for True).

Model	PE	Orig-Unk	Unk-Inv	Graft	Full triplet
Qwen3-8B	2	89.7%	47.6%	78.8%	37.4%
Qwen3-4B	0	79.2%	49.8%	59.9%	31.9%
Qwen3-32B	0	96.8%	95.1%	85.1%	79.1%
Llama-3-8B-Instruct	0	30.1%	10.4%	69.1%	4.8%
Llama-3.3-70B-Instruct	0	69.0%	54.7%	95.1%	46.2%
Phi-4-mini-instruct	0	86.9%	56.8%	67.1%	37.7%
Gemma-3-4B-IT	0	44.7%	32.2%	26.4%	5.5%
gpt-oss-20b	18	99.1%	99.2%	99.2%	97.8%

therefore not a reliable proxy for certificate accuracy: gpt-oss-20b is close to the ceiling on both views, while Llama-3-8B is weakest, but several intermediate models retain a material answer-minus-certificate gap. The gap is largest on the Graft condition for the weaker models and nearly absent for the strongest models, which rules out an interpretation in which every certificate failure is simply an answer-classification error. It also motivates treating the certificate as a separately audited output.

The transition table shows that the dominant signatures preserve the Unknown answer through the first two triplet positions while changing the certificate outcome at the final position, summarized by the recurring $U \to U \to T$ pattern. This concentrates the failure in the certificate leg rather than in a wholesale reversal of the answer `label`. The stronger gpt-oss-20b result has a different residual: its triplet behavior is near ceiling, but 18 records are parse errors. Thus the transition signatures should be read jointly with parser outcomes, rather than as evidence that all non-ceiling behavior reflects logical uncertainty.

Success and Other cases differ only weakly and inconsistently in proof count, depth, original certificate size, and surviving minimum proof size. These topology features therefore provide a mixed signal rather than a clean predictor. In particular, the stale outcome is not explained by a simple instance-difficulty gradient based on the available proof and depth statistics. This is why the stale analysis treats the dominant failure as certificate construction and closure checking, rather than attributing it to a uniformly harder subset of theories. The conclusion remains model-conditional because the eligible sets and their feature distributions are model-specific.

A8 Controls: Inference Strategy and Evaluator Choice

Inference strategy (Experiment A). Experiment A applies majority vote with $K = 5$ to identical adaptive cases (Wang et al., 2023), so the comparison changes inference aggregation without changing the evaluated theories or edits. The stale-rate movements are $20.1\% \to 36.6\%$ for Qwen3-8B, $28.3\% \to 43.3\%$ for Llama-3-8B-Instruct, $13.4\% \to 11.0\%$ for Qwen3-32B, $17.3\% \to 17.3\%$ for Llama-3.3-70B-Instruct, and $14.9\% \to 1.0\%$ for gpt-oss-20b. Four of the five models therefore obtain no stale reduction under voting. For Qwen3-8B, answer accuracy rises from 70.9% to 90.3%,

Table 13: Matched answer/certificate comparisons on 855 Unknown triplets. Triplet columns require agreement across all three worlds.

	Graft world		Full triplet	
Model	Answer	Certificate	Answer	Certificate
Qwen3-8B	97.9%	78.8%	45.3%	37.4%
Llama-3-8B-Instruct	97.1%	69.1%	7.6%	4.8%
Qwen3-32B	89.1%	85.1%	82.5%	79.1%
Llama-3.3-70B-Instruct	96.8%	95.1%	46.9%	46.2%
gpt-oss-20b	99.3%	99.2%	97.9%	97.8%
Phi-4-mini-instruct	87.0%	67.1%	44.3%	37.7%
Gemma-3-4B-IT	91.7%	26.4%	15.8%	5.5%
Qwen3-4B	92.6%	59.9%	42.2%	31.9%

Table 14: Behavioral transition signatures over the Unknown triplet (Original-Unknown $\rightarrow$ Unknown-Invariant $\rightarrow$ Graft, answer-level). The last column aggregates all remaining signatures including parse errors.

Model	U→U→T	U→T→T	T→T→T	U→U→U	Other
Qwen3-8B	387 (45.3%)	360 (42.1%)	85 (9.9%)	17 (2.0%)	6 (0.7%)
Qwen3-4B	361 (42.2%)	253 (29.6%)	172 (20.1%)	59 (6.9%)	10 (1.2%)
Qwen3-32B	705 (82.5%)	30 (3.5%)	11 (1.3%)	92 (10.8%)	17 (2.0%)
Llama-3-8B-Instruct	65 (7.6%)	169 (19.8%)	590 (69.0%)	17 (2.0%)	14 (1.6%)
Llama-3.3-70B-Instruct	401 (46.9%)	162 (18.9%)	224 (26.2%)	26 (3.0%)	42 (4.9%)
Phi-4-mini-instruct	379 (44.3%)	254 (29.7%)	105 (12.3%)	100 (11.7%)	17 (2.0%)
Gemma-3-4B-IT	135 (15.8%)	184 (21.5%)	384 (44.9%)	54 (6.3%)	98 (11.5%)
gpt-oss-20b	837 (97.9%)	0 (0.0%)	0 (0.0%)	3 (0.4%)	15 (1.8%)

but the gain is assigned to Stale rather than Success. Only gpt-oss-20b, the strongest model in this control, clears stale at the reported level. The result rules out single-sample decoding noise as the general explanation: improving aggregation does not substitute for maintaining a certificate against the updated world, in contrast to the gains reported for decomposition and program-based inference (Nye et al., 2021; Zhou et al., 2023; Chen et al., 2023; Gao et al., 2023; Yao et al., 2023; **?**)

A9 Paired Outcome Transitions under Majority Voting

Table 17 reports paired transitions from single-sample decoding, denoted $K = 1$, to majority voting with $K = 5$ on identical adaptive cases. Rows retain the $K = 1$ outcome and columns record the corresponding $K = 5$ outcome. This pairing distinguishes changes in outcome from changes in the evaluated case set. The matrices therefore show whether voting preserves, repairs, or introduces certificate failures at the case level.

For Qwen3-8B, 53 Success cases remain Success, 15 Success cases become Stale, and 5 Stale cases become Success. Refusal cases also contribute 12 transitions into Stale, while 13 remain Refusal. Voting consequently moves more mass into Stale from Success and Refusal than it moves from Stale into Success. Llama-3-8B shows the same directional concern in a sharper form: 10 Success cases become Stale, no Stale cases become Success, and 9 Refusal cases remain Refusal. These transitions show that aggregation can select an answer with an invalid certificate even when a single sample had produced a valid certificate.

The Qwen3-32B matrix is more mixed. It contains 34 Success-to-Success transitions, 1 Success-to-Stale transition, and 5 Success-to-Refusal transitions. Its Stale row contains no Stale-to-Success transition, with 7 cases remaining Stale and 4 becoming Refusal. Llama-3.3-70B preserves 56 Success cases as Success and changes only 1 Success case to Stale; its Stale row contains 1 Stale-to-Success transition and 16 cases that remain Stale. These rows do not establish a common direction for voting: lower transition into Stale does not imply systematic repair of existing stale certificates.

Table 15: Mean feature values for Success vs. Other (Stale+refusal) eligible cases on Formal-500 adaptive subsets (Mann–Whitney tests stored in appendix_data.json). Columns: proof count, query depth, original certificate size, surviving minimal-proof size (Success/Other means).

Model	$\lvert\mathcal{P}\rvert$	qdep	$\lvert C_{\text{orig}}\rvert$	surv. min size
Qwen3-8B	3.10/2.74	1.46/1.67	6.34/5.92	2.81/3.09
Qwen3-4B	2.76/2.89	1.62/1.57	4.85/5.46	2.86/3.09
Qwen3-32B	2.75/2.60	1.50/1.50	5.38/4.12	2.90/3.02
Llama-3-8B-Instruct	2.85/2.92	1.15/1.35	3.97/4.46	2.26/2.88
Llama-3.3-70B-Instruct	2.84/2.49	1.26/1.66	3.61/3.59	2.67/3.22

Table 16: Inference-strategy control: majority voting over five samples on the same model-specific adaptive cases. Voting redistributes outcomes in a model-dependent way (stale rate Fresh → K5: 20.1→36.6, 28.3→43.3, 13.4→11.0, 17.3→17.3, 14.9→1.0): for four of five models voting does not reduce false passes, and where it raises answer accuracy (Qwen3-8B) the gain is absorbed by Stale rather than Success. Only the strongest model clears most stale outputs under voting. False passes are not an artifact of single-sample decoding noise.

Model	Strategy	n	Answer acc.	Success	Stale	Refusal / no-cons.
Qwen3-8B	Single (K=1)	134	70.9%	50.7%	20.1% (27)	39
	MajorityVote (K=5)	134	90.3%	53.7%	36.6% (49)	13
Llama-3-8B	Single (K=1)	60	85.0%	56.7%	28.3% (17)	9
	MajorityVote (K=5)	60	81.7%	38.3%	43.3% (26)	11
Qwen3-32B	Single (K=1)	82	62.2%	48.8%	13.4% (11)	31
	MajorityVote (K=5)	82	58.5%	47.6%	11.0% (9)	34
Llama-3.3-70B	Single (K=1)	98	75.5%	58.2%	17.3% (17)	24
	MajorityVote (K=5)	98	75.5%	58.2%	17.3% (17)	24
gpt-oss-20b	Single (K=1)	201	97.5%	82.6%	14.9% (30)	5
	MajorityVote (K=5)	201	92.5%	91.5%	1.0% (2)	15

gpt-oss-20b is the only model for which aggregation clears the stale failure mode in the paired matrix. It has 28 Stale-to-Success transitions and no Stale-to-Stale transitions. Its Success row contains 154 Success-to-Success transitions, while 2 Success cases become Stale and 10 become Refusal. The result is consistent with voting repairing stale certificates for this model, but it remains a property of this model and these paired cases rather than a general guarantee of majority voting.

Refusal-persistent blocks remain visible in the $R \to R$ cells: 13 for Qwen3-8B, 9 for Llama-3-8B, 25 for Qwen3-32B, 24 for Llama-3.3-70B, and 3 for gpt-oss-20b. Voting cannot resolve a case that single-sample decoding consistently leaves at Refusal. This cell-level view complements the marginals in the Experiment A table. Aggregation neither consistently improves nor consistently degrades certificate validity; its direction is model-conditional, so majority voting cannot substitute for maintaining and verifying the certificate against the current world.

Evaluator choice (Experiment B). Answer-only evaluation has zero stale recall by construction because it does not inspect the certificate or the updated world. Static-Original-Cert likewise does not consult the update and accepts all residual stale cases, with residual stale counts of 66, 26, 42, and 41 on the four judged models. The two LLM-as-a-Judge variants recover at most 88% of stale cases, while their valid-certificate retention falls into a low range reaching 38–44%; the most aggressive false rejection discards 42 of 68 valid certificates. Only the deterministic current-world verifier is 100/100 by construction. Self-verification, tool-interactive critique, and learned step verifiers improve answer reliability (Weng et al., 2023; Gou et al., 2023; Manakul et al., 2023; **?**), but as the table shows a judge still trades stale recall against valid retention. The control therefore identifies symbolic current-world checking as necessary: a learned critic can recover part of the signal, but cannot reliably preserve valid certificates while detecting stale ones.

Table 17: Paired outcome transitions from single-sample decoding ($K = 1$, Fresh) to five-sample majority voting ($K = 5$) on the identical adaptive cases, per model. Rows are the single-sample outcome, columns the voted outcome; entries are case counts over the model-specific eligible set.

Model	$K = 1$	Success	Stale	Refusal/no-cons.	Total
Qwen3-8B	Success	53	15	0	68
	Stale	5	22	0	27
	Refusal	14	12	13	39
Llama-3-8B	Success	23	10	1	34
	Stale	0	16	1	17
	Refusal	0	0	9	9
Qwen3-32B	Success	34	1	5	40
	Stale	0	7	4	11
	Refusal	5	1	25	31
Llama-3.3-70B	Success	56	1	0	57
	Stale	1	16	0	17
	Refusal	0	0	24	24
gpt-oss-20b	Success	154	2	10	166
	Stale	28	0	2	30
	Refusal	2	0	3	5

Table 18: LLM-as-a-Judge control on the same Fresh outputs (rows: model; first line Judge-Answer, second line Judge-Certificate). Judges reach at most 88% stale recall while false-rejecting 32–42% of valid certificates; the deterministic ProofCover verifier is 100/100 by construction. The gap quantifies why current-world checking must be symbolic rather than judged.

Model	n	Stale recall	Valid retention	False rejects
Qwen3-8B	134	52%	53%	32
		Cert: 85%	*Cert:* 38%	*Cert:* 42
Llama-3-8B	60	59%	50%	17
		Cert: 88%	*Cert:* 44%	*Cert:* 19
Qwen3-32B	82	45%	82%	7
		Cert: 45%	*Cert:* 75%	*Cert:* 10
Llama-3.3-70B	98	41%	51%	28
		Cert: 59%	*Cert:* 56%	*Cert:* 25

A10 Judge Confusion Detail

Table 19 gives the confusion counts for the two LLM-judge variants on Fresh eligible outputs from the models evaluated in Experiment B. The entries distinguish accepted valid outputs, rejected stale outputs, accepted stale outputs, and rejected valid outputs. Stale recall measures the fraction of stale outputs rejected by the judge, while valid retention measures the fraction of valid outputs accepted. The two metrics expose different error costs: rejecting stale certificates improves recall, whereas rejecting valid certificates lowers retention.

Judge-Certificate substantially increases stale recall relative to Judge-Answer for every model. For Qwen3-8B, recall rises from 51.9% to 85.2%, while valid retention falls from 52.9% to 38.2%. Llama-3-8B shows a change from 58.8% to 88.2% in stale recall and from 50.0% to 44.1% in valid retention. For Qwen3-32B, stale recall changes from 45.5% to 90.9%, while valid retention changes from 82.5% to 67.5%. Llama-3.3-70B changes from 41.2% to 88.2% in stale recall and from 50.9% to 70.2% in valid retention.

The confusion counts show how these metric changes arise. For Qwen3-8B, Judge-Answer accepts 13 stale outputs and rejects 32 valid outputs; Judge-Certificate accepts 4 stale outputs but rejects 42 valid outputs. For Qwen3-32B, the corresponding stale errors change from 6 accepted stale outputs to 1, while rejected valid outputs change from 7 to 13. The increased sensitivity to certificate defects therefore carries a model-dependent cost in valid outputs discarded. Llama-3.3-70B gives

a particularly asymmetric Judge-Answer example: the judge misses 10 of 17 stale outputs while discarding 28 of 57 valid outputs.

No judge variant reaches the deterministic verifier's $(100, 100)$ point for stale recall and valid retention. The best stale recall shown is 90.9%, paired with valid retention of 67.5%. Judge-Certificate therefore approaches the verifier more closely on stale detection for some models, but its acceptance decision still rejects valid certificates. Judge-Answer preserves more valid outputs in some rows, yet it misses a substantial portion of stale certificates. The observed frontier is consequently a recall-retention trade-off rather than an approximation of exact verification.

These results support the claim that learned critics cannot substitute for symbolic current-world checking in this setting. The deterministic verifier attains $(100, 100)$ by construction because it evaluates the certificate against the updated rules rather than inferring validity from the answer or certificate presentation. The LLM judges instead make asymmetric errors whose magnitude and direction vary across models. Their outputs can provide an additional signal, but the confusion cells do not support treating either judge variant as a certificate-validity oracle.

Table 19: Judge confusion detail on the model-specific eligible Fresh cases (Experiment B). "Accept valid" counts accepted certificates that pass the deterministic current-world check; "reject stale" counts rejected certificates that the deterministic check flags as stale; the complementary cells are the corresponding errors. Stale recall is $\mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$ and valid retention is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ under the (accept, reject) framing of the main text.

Model	Variant	n	Accept valid	Reject stale	Accept stale	Reject valid	Stale rec.	Valid ret.
Llama-3.3-70B	Judge-Answer	98	29	7	10	28	41.2%	50.9%
Llama-3.3-70B	Judge-Certificate	98	32	10	7	25	58.8%	56.1%
Llama-3-8B	Judge-Answer	60	17	10	7	17	58.8%	50.0%
Llama-3-8B	Judge-Certificate	60	15	15	2	19	88.2%	44.1%
Qwen3-32B	Judge-Answer	82	33	5	6	7	45.5%	82.5%
Qwen3-32B	Judge-Certificate	82	30	5	6	10	45.5%	75.0%
Qwen3-8B	Judge-Answer	134	36	14	13	32	51.9%	52.9%
Qwen3-8B	Judge-Certificate	134	26	23	4	42	85.2%	38.2%

Table 20: Paired symbolic/NL mechanism check (same 100 bases, graph-matched renderings): Repair answer agreement by representation and Original-Unknown answer discordance with exact McNemar p. High NL-only Unknown rates on trivial motifs show refusal behavior is representation-sensitive on the negative-control set; Repair agreement confirms the motifs do not separate construction methods.

Model	Rep. both	sym. only	NL only	neither	Unk- discordant	McNemar p
Qwen3-8B	100	0	0	0	81	8.3e-25
Qwen3-32B	100	0	0	0	0	1
Llama-3-8B-Instruct	100	0	0	0	0	1
Llama-3.3-70B-Instruct	100	0	0	0	0	1
gpt-oss-20b	100	0	0	0	0	1

Representation control. The paired symbolic and graph-matched natural-language Repair motifs agree for every model, as intended for a non-discriminating negative control. On Original-Unknown, Qwen3-8B has discordant answers on 81 of 100 pairs, with Unknown produced only in the natural-language representation and an exact McNemar $p = 8.3 \times 10^{-25}$; every other model has zero discordance with $p = 1$. The result shows that refusal on trivial motifs can be representation-sensitive for one model. These motifs therefore test representation robustness and are not evidence that one representation is intrinsically more difficult.

A11 Failure Taxonomy

The taxonomy is defined on the verifier's observed certificate, not on an interpretation of the model's hidden reasoning. Let $T = (F, R)$ be the original world and T' the current world after deletion. A

	Qwen3-8B	Llama-3-8B	Qwen3-32B	Llama-3.3-70B	gpt-oss-20b	Phi-4-mini	Gemma-3-4B	Qwen3-4B
Fresh Success S	68	34	40	57	166	50	14	72
Stale: removed_id_cited	0	0	0	0	0	0	0	0
Stale: nonexistent_id	1	0	1	0	27	0	0	0
Stale: incomplete_support	26	17	10	17	3	22	13	25

Table 21: Fresh Success counts and verifier-observed stale categories on the frozen adaptive cases. Other contains refusals and parse failures and is not included in the three stale rows.

certificate C is classified as `removed_id_cited` when it contains an identifier that was present in T but was removed in forming T'. It is classified as `nonexistent_id` when a cited identifier is not an identifier in the current world and is not accounted for by the removed-ID category; bracket-wrapped strings such as `[rule3]` are distinct from the verbatim identifier `rule3`. It is classified as `incomplete_support` when every cited identifier exists in the current world but $T'[C] \not\models q$. Refusals and outputs that cannot be parsed into the expected answer-and-certificate structure are Other.

The first headline result is that `removed_id_cited` is zero for all eight models. No model re-cites the deleted rule it was shown being removed. The naive stale explanation, in which the model simply ignores the edit and reuses the dead rule, is therefore not supported. Instead, the models usually recompute a locally plausible support set whose members survive the edit but whose combined theory no longer entails the query: the failure is a near-miss proof caused by missing global closure verification.

The second result is that seven of eight models are essentially 100% incomplete-support failures among their stale certificates. The exact incomplete-support counts over stale cases are $26/27$ for Qwen3-8B, $17/17$ for Llama-3-8B-Instruct, $10/11$ for Qwen3-32B, $17/17$ for Llama-3.3-70B-Instruct, $22/22$ for Phi-4-mini-instruct, $13/13$ for Gemma-3-4B-IT, and $25/25$ for Qwen3-4B. These certificates cite real, surviving, and often relevant items, but those items no longer form a complete derivation of the query. The gap is therefore in verifying proof closure after the update, not in tracking the identity of the deleted rule.

gpt-oss-20b exhibits a distinct format-compliance mode. Of its 30 stale certificates, 27 contain bracket-wrapped identifiers such as `[rule3]` copied from the displayed `[id]` format. The underlying identifier often corresponds to a surviving proof item, but the returned string fails the verbatim-identifier existence contract. The remaining three stale certificates are incomplete-support cases. This is a serialization failure rather than evidence of a reasoning failure, and it inflates the apparent false-pass count under the strict verifier. The reported strict-checking false-pass share of 15.3% is the floor used for the exact identifier contract; under lenient identifier normalization, the apparent stale component would drop toward the incomplete-only reading of approximately 1.5%, namely $3/201$. The `id_shape_constraint` was deliberately disabled because enabling it would have hidden precisely this observed mode.

Stale certificates are small rather than bloated. Table 23 shows that, across models, the median is between three and five cited items and the maximum is at most 14. The failure is consequently not explained by a tendency to emit excessively large certificates that accidentally include unsupported material; it is more consistent with selecting a compact but insufficient support set.

The Gemma audit in Appendix A28 is the extreme case. All 13 Gemma stale certificates are incomplete-support failures, and the model has the smallest eligible set and the lowest original validity. This concentration indicates that a single weakness in certificate construction can drive both poor original validity and stale behavior. It does not, however, imply that eligibility size alone predicts stale failure across models, since the adaptive sets remain model-conditional.

A12 Certificate Provenance under Fresh and Stateful Updates

Table 22 decomposes stale certificates into removed citations, nonexistent citations, and incomplete support. Across all models, the dominant failure is incomplete support. Fresh evaluation contains 133 incomplete certificates among 162 stale certificates. Combining Fresh and Stateful evaluation gives 219 incomplete certificates among 256 stale certificates. Removed-identifier citations account

for only 7 cases across every model and condition. Thus, stale certificates generally preserve citations that remain syntactically available after the update but no longer provide a derivation of the updated query.

The same pattern appears within the model rows. Qwen3-8B has 26 incomplete certificates among 27 Fresh stale cases and 11 among 14 Stateful stale cases. Qwen3-4B has 25 among 25 Fresh stale cases and 17 among 20 Stateful stale cases. Qwen3-32B has 10 among 11 Fresh stale cases. Llama-3-8B, Llama-3.3-70B, Phi-4-mini, and Gemma-3-4B-IT likewise have Fresh incomplete counts of 17 among 17, 17 among 17, 22 among 22, and 13 among 13, respectively. Their Stateful rows remain dominated by incomplete support whenever stale certificates occur.

gpt-oss-20b is the exception in Fresh evaluation. Its stale count is 30, of which 27 are classified as nonexistent citations and 3 as incomplete support. The nonexistent strings use bracket-format identifiers such as `[rule3]` rather than `rule3`. This behavior is therefore a serialization failure in the citation representation, rather than evidence that the model reasoned from a logically deleted rule. The Stateful row for gpt-oss-20b has no stale certificates. Qwen3-32B also has no Stateful stale certificates, while its refusal count is 52, so the disappearance of stale certificates there includes conversion into refusals rather than universal certificate repair.

Stateful evaluation reduces the stale count for 7 of 8 models. Llama-3-8B is the exception, with its stale count changing from 17 in Fresh evaluation to 19 in Stateful evaluation. The reduction therefore cannot be interpreted uniformly as improved support maintenance: in some rows, the model avoids an invalid certificate by refusing to produce one. This distinction matters because a refusal and a valid certificate have different operational meanings even when both remove the stale outcome.

The aggregate provenance counts identify support closure as the principal failure mode under answer-preserving updates. A citation can survive the edit while the certificate ceases to entail the query, and this failure is more prevalent than explicit recitation of removed identifiers. The finding is consistent with the Gemma audit section, where certificate validity depends on whether the cited support still closes the proof after the update. These data support a model-conditional account of false passes: the central risk is failure to re-establish derivational support, with serialization errors forming a separate case for gpt-oss-20b.

Table 22: Provenance of invalid certificates (Stale outputs) on the model-specific eligible adaptive cases, separately for the Fresh and Stateful conditions. Columns give the count of stale certificates citing a removed identifier, citing a nonexistent identifier, or forming incomplete support in the current world. Counts of the three categories sum to Stale for every row.

Model	Cond.	n	Stale	Removed-cit.	Nonexistent-cit.	Incomplete
Qwen3-8B	Fresh	134	27	0	1	26
Qwen3-8B	Stateful	134	14	2	1	11
Qwen3-4B	Fresh	126	25	0	0	25
Qwen3-4B	Stateful	126	20	3	0	17
Qwen3-32B	Fresh	82	11	0	1	10
Qwen3-32B	Stateful	82	0	0	0	0
Llama-3-8B-Instruct	Fresh	60	17	0	0	17
Llama-3-8B-Instruct	Stateful	60	19	0	0	19
Llama-3.3-70B-Instruct	Fresh	98	17	0	0	17
Llama-3.3-70B-Instruct	Stateful	98	15	1	0	14
Phi-4-mini-instruct	Fresh	86	22	0	0	22
Phi-4-mini-instruct	Stateful	86	17	1	0	16
Gemma-3-4B-IT	Fresh	34	13	0	0	13
Gemma-3-4B-IT	Stateful	34	9	0	0	9
gpt-oss-20b	Fresh	201	30	0	27	3
gpt-oss-20b	Stateful	201	0	0	0	0

A13 Stale-Certificate Size Distributions

Table 23 reports the stale-certificate size distributions: Qwen3-8B has 27 stale certificates with sizes 3–14 and median 4; Qwen3-4B has 25 with sizes 2–7 and median 5; Qwen3-32B has 11 with sizes

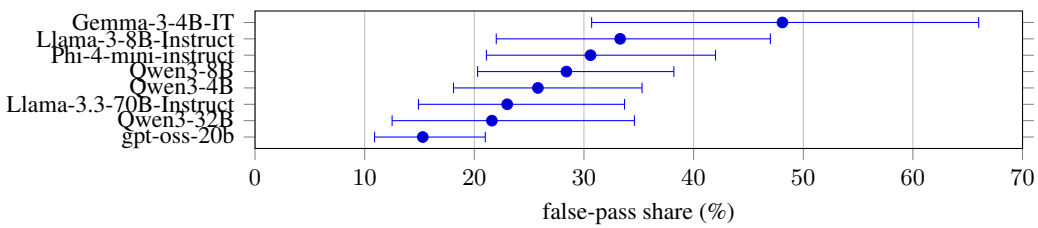


Figure 3: False-pass share under Fresh Adaptive Repair (Wilson 95% CIs); the dashed line marks the median (27.1%).

2–8 and median 3; Llama-3-8B has 17 with sizes 2–6 and median 5; Llama-3.3-70B has 17 with sizes 1–7 and median 3; Phi-4-mini has 22 with sizes 2–8 and median 4; Gemma-3-4B-IT has 13 with sizes 2–8 and median 3; and gpt-oss-20b has 30 with sizes 2–7 and median 4. Stale certificates are therefore compact, with medians of 3–5: models submit short cited supports rather than long redundant lists. The one extreme is Qwen3-8B's maximum of 14 items.

In conjunction with the provenance results in Table 22, where incomplete support dominates, these compact certificates indicate that the failure is a missing needed identifier rather than a proof buried in noise: with only 2–8 items cited, omitting the single surviving minimal premise is sufficient to break derivability. This rules out a "certificate overflow" explanation. gpt-oss-20b's sizes characterize bracket-format serialization failures rather than incomplete-support cases.

Table 23: Number of cited items in Fresh stale certificates, per model. n is the number of stale certificates; sizes count the identifiers the model placed in its submitted certificate.

Model	n	Min	Median	Max
Qwen3-8B	27	3	4	14
Qwen3-4B	25	2	5	7
Qwen3-32B	11	2	3	8
Llama-3-8B-Instruct	17	2	5	6
Llama-3.3-70B-Instruct	17	1	3	7
Phi-4-mini-instruct	22	2	4	8
Gemma-3-4B-IT	13	2	3	8
gpt-oss-20b	30	2	4	7

A14 False-Pass Shares with Interval Estimates

Table 24 reports false-pass shares separately for each model. The point estimates range from 15.3–48.1%: gpt-oss-20b has 30 stale certificates among 196 answer-accepted outputs, whereas Gemma-3-4B-IT has 13 among 27. This range is per-model and is not a pooled estimate.

Uncertainty differs substantially with the number of answer-accepted outputs. The Wilson 95% interval for gpt-oss-20b is 10.9–21.0%, a width of approximately 10 percentage points, while the interval for Gemma-3-4B-IT is 30.7–66.0%, spanning 35 percentage points. Thus, the headline upper bound is accompanied by a wide interval and should not be read as a precise point claim.

Every model's interval overlaps the interval of at least one other model. The intermediate point estimates—21.6%, 23.0%, 25.8%, 28.4%, 30.6%, and 33.3%—therefore do not support a finely resolved ranking.

The false-pass share answers a specific counterfactual: among outputs that answer-only scoring would accept, what fraction would deterministic current-world verification reject because the certificate is stale? It does not estimate the population stale rate, because the denominator conditions on eligibility as defined in Section A21; the eligible counts in Table 24 range from 34 to 201.

Table 24: Fresh false-pass share per model: $\mathrm{Stale}/(\mathrm{Success} + \mathrm{Stale})$ over answer-accepted eligible outputs, with the numerator/denominator counts and Wilson 95% intervals. The share is the fraction of answer-only-accepted outputs whose certificate fails the current-world check; it is reported per model and never pooled.

Model	Stale	Answer-acc.	Share [Wilson 95%]	Eligible
Qwen3-8B	27	95	28.4% $[20.3, 38.2]$	134
Qwen3-4B	25	97	25.8% $[18.1, 35.3]$	126
Qwen3-32B	11	51	21.6% $[12.5, 34.6]$	82
Llama-3-8B-Instruct	17	51	33.3% $[22.0, 47.0]$	60
Llama-3.3-70B-Instruct	17	74	23.0% $[14.9, 33.7]$	98
Phi-4-mini-instruct	22	72	30.6% $[21.1, 42.0]$	86
Gemma-3-4B-IT	13	27	48.1% $[30.7, 66.0]$	34
gpt-oss-20b	30	196	15.3% $[10.9, 21.0]$	201

A15 Mechanistic Analysis: Reading and Rewriting Internal Proof State

The behavioral taxonomy shows that models track the deletion: no fresh certificate cites a removed identifier, yet many submitted certificates provide incomplete support. A secondary internal-line analysis on Qwen3-8B and Llama-3-8B asks whether proof content remains represented after the update and whether that content can be rewritten into a valid replacement. The results are mechanistic diagnostics for the behavioral pattern, not additional headline estimates; the behavioral results in Appendix A20 do not depend on them.

Figure 4 shows probe accuracy across layers and readout positions. The proof-identity probes (Alain & Bengio, 2018) recover which proof the model is producing from residual-stream readout positions. Their accuracy varies by layer and position, but the qualitative conclusion is stable: proof identity is linearly readable.

Figure 5 evaluates whether activation patching changes the submitted label outcome. Activation patching in the causal-abstraction sense (Geiger et al., 2021) does not flip the submitted `label` outcome when proof-content activations are transplanted. Thus, the internal state is readable but not locally writable in the tested intervention. This is consistent with the behavioral finding that exposing the edit does not by itself produce a valid replacement certificate.

Figure 6 displays layerwise shifts for stale and successful proof updates. The Δ-IRS analysis measures the representational shift induced by the update separately for each pair. Stale cases show weaker late-layer separation of the alternative-proof direction than successful cases, which is consistent with a failure to replace the deleted support.

Figure 7 compares old-proof and alternative-proof directions across stale and successful cases. The corresponding geometric diagnostic compares post-update states with the old-proof and alternative-proof directions. Across the displayed comparisons, stale cases retain evidence related to the old support while showing weaker organization around the alternative proof than successful cases.

Figure 8 displays the distribution of pre-update and post-update representations. The representation diagnostics provide a layerwise distributional comparison of pre- and post-update states. Taken together, the internal line is consistent with the audit story: the old proof remains readable, while writing the new proof is the bottleneck. These analyses use two models, a single seed, and correlational diagnostics; they do not carry a headline claim, and all behavioral numbers stand without them.

A16 Figure Gallery for the Mechanistic Diagnostics

The figures below collect exploratory, secondary diagnostics. No headline behavioral claim depends on these analyses: the internal comparisons cover two models under a single seed, with 55 paired Success/Stale cases for Qwen3-8B and 32 for Llama-3-8B.

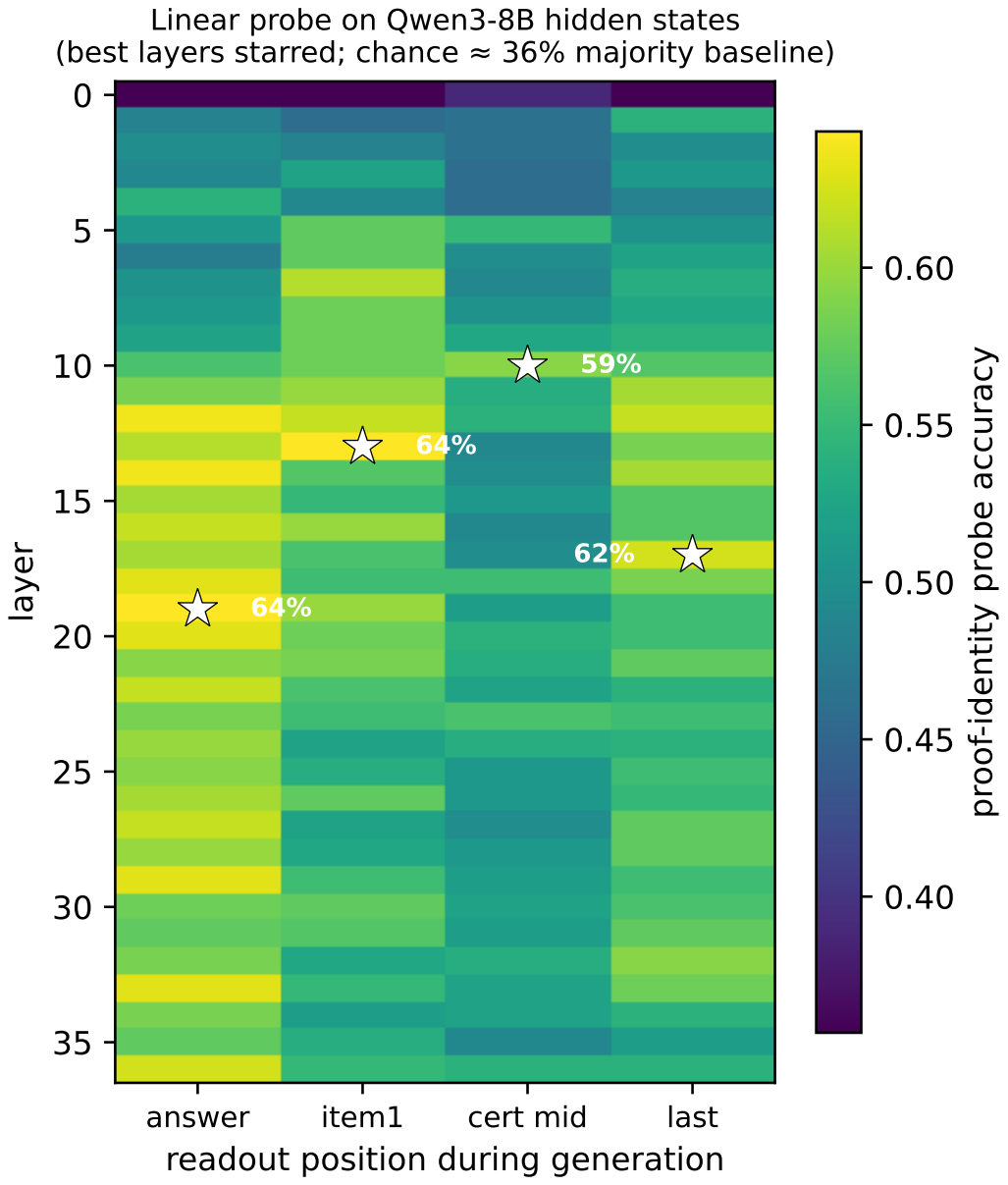


Figure 4: Proof-identity linear-probe accuracy across residual-stream layers and readout positions for Qwen3-8B and Llama-3-8B. The reader should compare recoverability across the layer-by-position grid: accuracy varies with both layer and position, supporting the qualitative conclusion that the identity of the proof being produced is linearly readable from the residual stream. This is a secondary diagnostic on two models and does not establish a behavioral success rate.

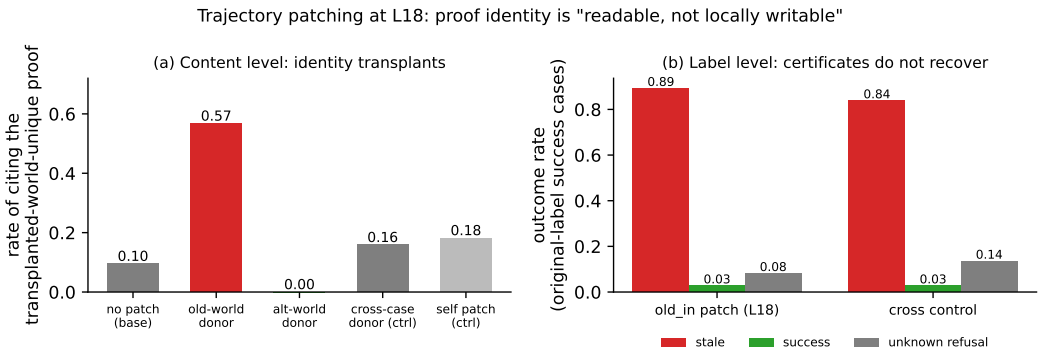


Figure 5: Activation-patching results for proof-content representations in Qwen3-8B and Llama-3-8B. The reader should inspect whether transplanting proof-content activations changes the submitted `label`. The absence of a `label` flip supports the distinction between readable proof content and locally writable proof state: the content can be decoded, but its transplantation does not directly rewrite the answer outcome.

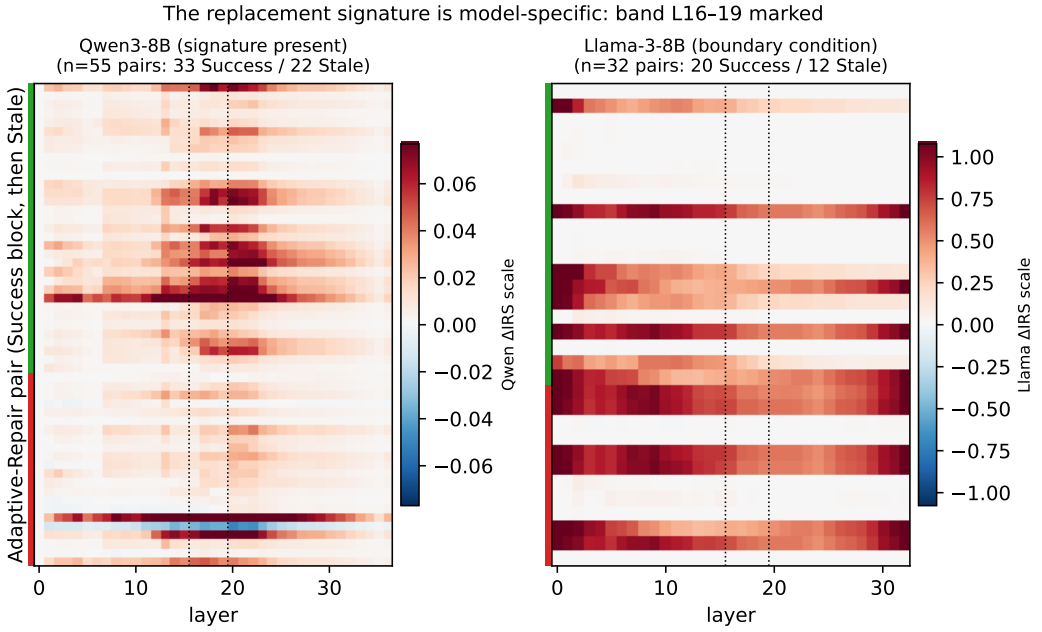


Figure 6: Per-pair layerwise Δ-IRS after the world update for Qwen3-8B and Llama-3-8B. The reader should compare the late-layer separation associated with the alternative-proof direction in stale and successful cases. Stale cases show weaker separation of that direction at late layers, providing an internal correlate of failed replacement rather than an independent behavioral claim.

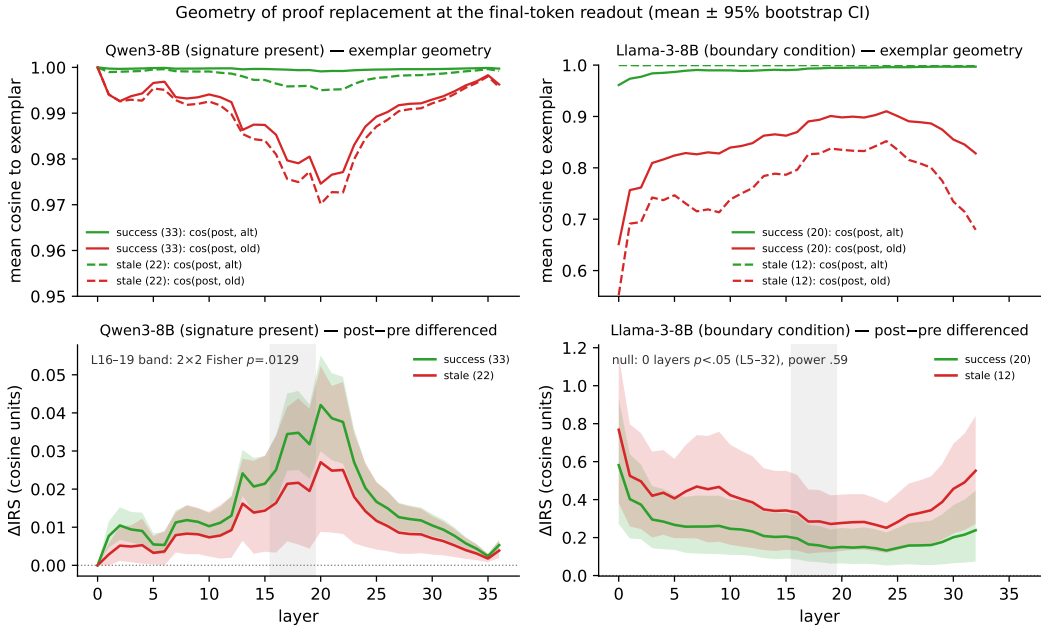


Figure 7: Post-update, alternative-proof, and old-proof cosine geometry together with Δ-IRS comparisons for successful and stale cases in Qwen3-8B and Llama-3-8B. The reader should compare the relative geometry of the alternative and old proof directions across the two outcome groups. The panel supports the qualitative interpretation that stale cases retain an internal relation to the old support without establishing the alternative-proof representation as clearly as successful cases.

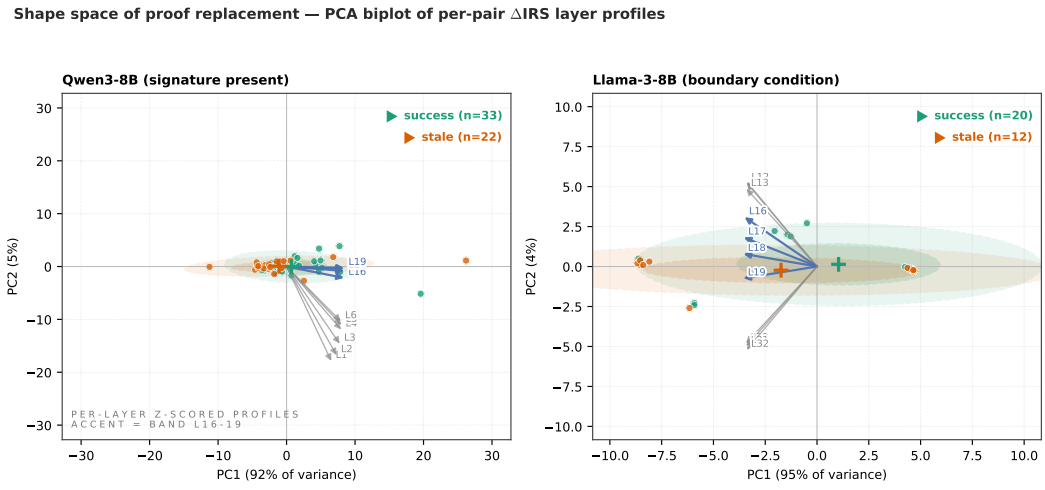


Figure 8: Biplot representation of pre-update and post-update state differences for Qwen3-8B and Llama-3-8B. The reader should inspect the distributional displacement between the two states and the extent to which the displayed groups separate after the update. The panel provides a layerwise distributional comparison of pre- and post-update representations and is descriptive rather than a causal or behavioral result.

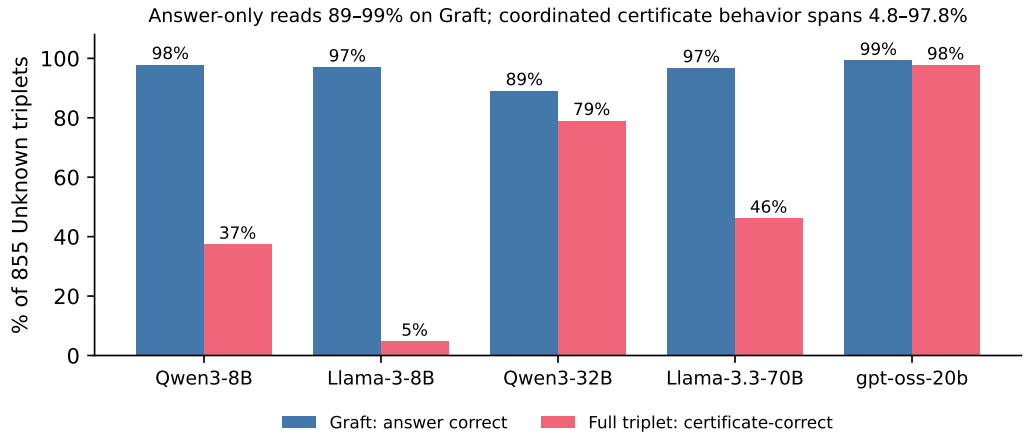


Figure 9: Answer accuracy and full-triplet certificate accuracy for Graft on Unknown triplets across five external models.

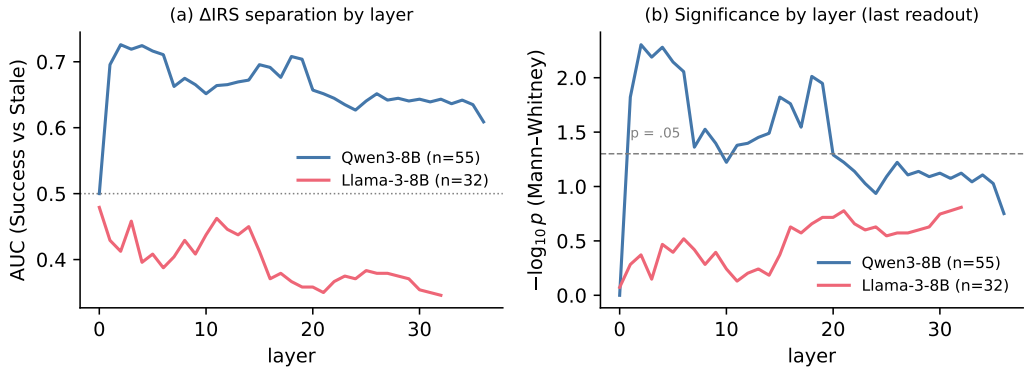


Figure 10: Per-layer AUC of the last-token Δ-IRS statistic for separating Success from Stale outcomes.

Figure 9 is a behavioral summary rather than an internal-state diagnostic. Graft answer accuracy remains between 89–99%, while full-triplet certificate accuracy ranges from 4.8–97.8%, directly displaying the gap between producing the answer and supplying a valid certificate.

Figure 10 evaluates the difference between post-update and pre-update residual states projected onto the alternative-proof and old-proof directions. Qwen3-8B peaks at AUC 0.726 on layer 2 with Mann–Whitney $p = 0.0025$, whereas Llama-3-8B peaks at AUC 0.479 with $p = 0.594$ and shows no separation. The AUC of 0.726 records a distributional difference over 55 pairs, not evidence of a mechanism.

Figure 11 decomposes each paired displacement into $\Delta_{\text{alt}} - \Delta_{\text{old}}$ on the horizontal axis after rotation and the common mode $\Delta_{\text{alt}} + \Delta_{\text{old}}$. The separate model panels and marginal densities expose the observed distributions without establishing that either direction causally controls certificate validity.

Figure 12 shows Qwen3-8B separation emerging in the middle band, consistent with its layer-2 AUC peak. The Llama-3-8B distributions largely overlap. This asymmetry is the supported summary: proof-state readability is observed for Qwen3-8B in this diagnostic but is not established as model-independent.

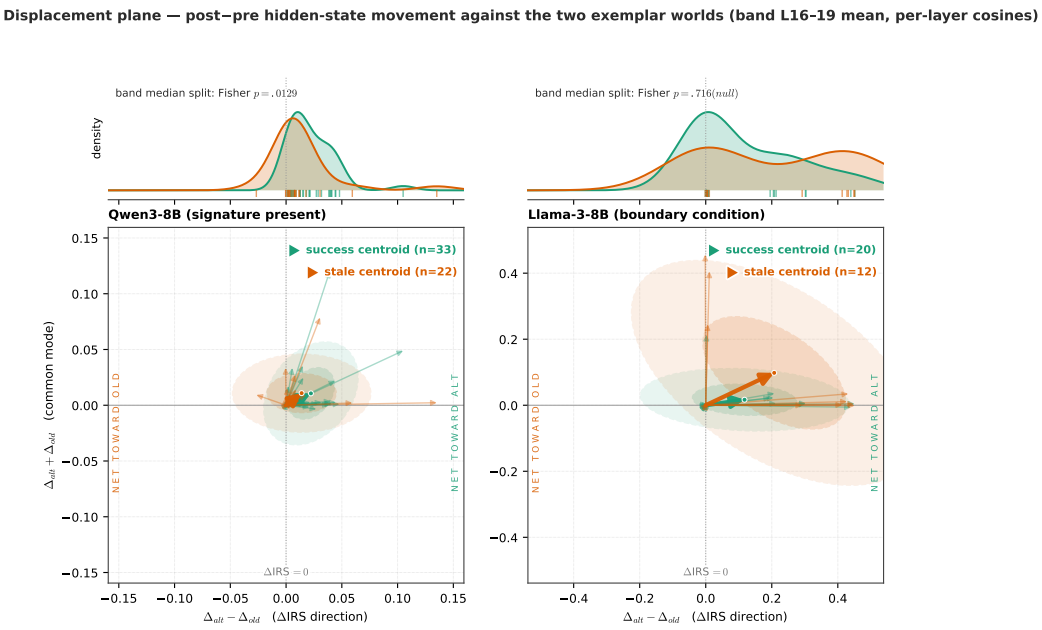


Figure 11: Post–pre residual displacement decomposed into the rotated Δ-IRS direction and common mode, with marginal densities for each model.

A17 Qualitative Failure Cases

The following compact cards report the released case identifiers and the information fixed in the audit artifacts. Where the frozen fact block does not reproduce a case-specific query, displayed rule text, certificate, or identifier list, that omission is stated explicitly rather than filled by reconstruction.

Case 1: Qwen3-8B, incomplete support. `pw:AttNoneg-OWA-D2-1042:13804`. *Query:* "Anne is red."

Relevant displayed-rule subset: `[rule3], [rule5]`.

Removed IDs: `rule6`.

Submitted stale certificate: `[rule3, rule5, triple1, triple2]`.

Surviving minimal proof: `[rule5, triple3]`.

Diagnosis: All cited items exist after the edit but no longer chain to the query; the model retained two dead premises, `triple1` and `triple2`, with `rule3`, alongside a partial surviving route that is missing `triple3`. This is *incomplete_support*.

Case 2: gpt-oss-20b, bracket-format failure. `pw:AttNoneg-OWA-D1-689:13618`. *Query:* "Dave is rough."

Relevant displayed facts and rule: "Dave is white."; `rule3`: "If something is white then it is rough."

Removed IDs: `rule1`.

Submitted stale certificate, raw JSON:

```
{"answer":"True","certificate":["[triple6]","[rule3]"]}
```

Surviving minimal proof: `[rule3, triple6]`.

Diagnosis: Semantically, the model found a valid surviving proof, but copied the display brackets into the identifiers, so the strings failed exact identifier existence. The strict verifier therefore scores the output as Stale. This is the gpt-oss-only format-compliance mode.

Case 3: Gemma-3-4B-IT, incomplete support; representative of all 13 Gemma stale cases. *Artifact locator:* the Gemma audit in Appendix A28; the frozen fact block does not provide a per-instance case identifier.

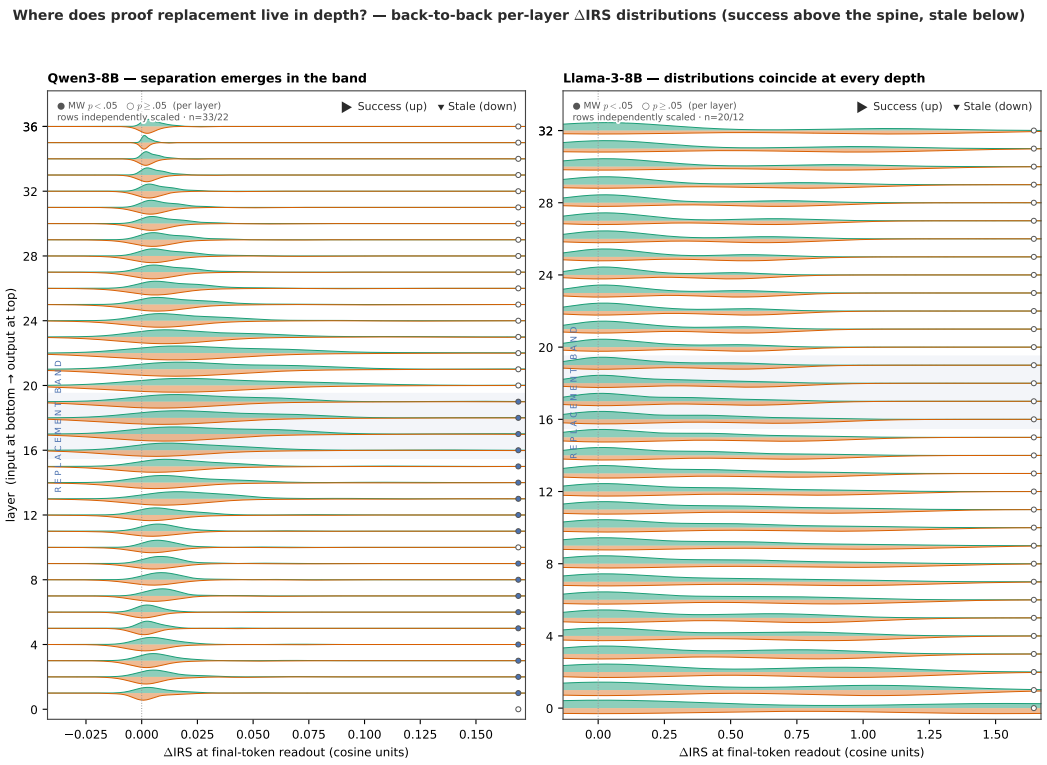


Figure 12: Layerwise ridge distributions of last-token Δ-IRS for Success and Stale cases, with the input layer at the bottom.

Query, displayed-rule subset, and removed IDs: case-specific fields are in the released Gemma audit artifact but are not reproduced in the frozen fact block.

Submitted stale certificates: every stale certificate cites only surviving IDs and nevertheless fails to derive the query; certificate sizes range from 2 to 8, with median 3.

Surviving minimal proof: case-specific proof contents are not enumerated in the frozen fact block.

Diagnosis: The certificates preserve identifier existence but fail derivability after the edit, which is *incomplete_support*; this finding represents all 13 Gemma stale cases.

Case 4: Qwen3-32B, stateful over-refusal. *Artifact locator:* the Qwen3-32B fresh-to-stateful transition matrix; no per-instance case identifier or case-specific query is provided in the frozen fact block.

Query, relevant displayed rules, removed IDs, and submitted stale certificate: case-specific fields are not enumerated in the frozen fact block.

Surviving minimal proof: an alternative proof existed, but its identifiers are not enumerated in the frozen fact block.

Diagnosis: A Fresh Stale output became a Stateful Unknown refusal: the model no longer vouched for the old support but did not construct the available replacement. This is *stateful over-refusal*, contributing to Stateful Other rising from 31 to 52 while stale fell from 11 to 0.

Case 5: Qwen3-4B, stateful success conversion. *Artifact locator:* the Qwen3-4B fresh-to-stateful transition matrix; no per-instance case identifier or case-specific query is provided in the frozen fact block.

Query, relevant displayed rules, removed IDs, and submitted stale certificate: case-specific fields are not enumerated in the frozen fact block.

Surviving minimal proof: the alternative proof was easy to re-derive from the displayed updated world; its identifiers are not enumerated in the frozen fact block.

Diagnosis: Re-eliciting the certificate in the updated conversation converts Fresh Stale to Stateful Success in 18 of 25 stale cases and converts refusal to success in 22 cases. The same conditioning collapses refusals from 29 to 1, exhibiting *stateful success conversion* when replacement derivation is easy.

Case 6: gpt-oss-20b, stateful stale clearance. *Artifact locator:* the gpt-oss-20b fresh-to-stateful transition matrix; no per-instance case identifier or case-specific query is provided in the frozen fact block.

Query, relevant displayed rules, removed IDs, and submitted stale certificate: case-specific fields are not enumerated in the frozen fact block.

Surviving minimal proof: the updated conversation re-elicits the proof without the bracket-wrapped identifier habit; the case-specific proof identifiers are not enumerated in the frozen fact block.

Diagnosis: All 30 Fresh Stale cases resolve, with 30 stale-to-success transitions and zero stale-to-stale transitions. The disappearance of the bracket-format failure under stateful re-elicitation is supported by $q_{\mathrm{BH}} = 2.2\mathrm{e}{-8}$ and is classified as *stateful stale clearance*.

A18 Prompts, Schemas, and Parsing

The frozen forward-verified prompt is rendered with the complete current world and uses the displayed square-bracket identifiers as the identifier contract.

```
Determine whether the QUERY follows from the current world.
Before returning JSON, build a forward-chaining ledger
internally:
1.  Initialize the ledger with displayed facts and their
IDs.
2.  Repeatedly fire every rule whose premises are already in
the ledger, recording the rule ID and producer IDs.
3.  Backtrack from the QUERY through the ledger to extract
one complete derivation with all terminal facts and rules.
4.  Perform a final verification that the extracted IDs
alone derive the QUERY and that the last rule concludes the
QUERY.
If the QUERY never enters the ledger, answer Unknown.  Do
not output the internal ledger.
 Return JSON only with exactly this schema:
{"answer":"True|Unknown",
"certificate":["fact_or_rule_id",...]}
The certificate must use only current IDs and should be a
smallest sufficient support.
 Copy every certificate ID exactly from the square brackets
below.  Use the displayed IDs verbatim.
 A valid True certificate MUST include every fact ID used
as a premise and every rule ID used in the derivation; rules
alone are not sufficient.
 If the answer is Unknown, there is no proof:  certificate
MUST be an empty list [].
 FACTS
[fact_id] text ...
 RULES
[rule_id] text ...
 QUERY
<query>
```

The system message is fixed as follows.

```
"You are a precise symbolic reasoner."
```

The stateful conversation is constructed by retaining the original interaction and appending the update together with a freshly rendered current-world prompt.

```
user = the fresh prompt for the Original world;
assistant = the model's own original JSON prediction;
user = "WORLD UPDATE: the following IDs were removed:
<removed ids>.  Your previous certificate is historical.
Re-evaluate using only the complete current world below."
followed by the fresh prompt rendered for the updated world.
```

The output schema enforced by vLLM structured outputs is:

```
{"answer":  enum True/Unknown, "certificate":  array of
strings, uniqueItems=true, maxItems = number of displayed
items}
```

Certificate items have no pattern constraint. The vocabulary is therefore open by design, allowing deleted, nonexistent, and malformed identifier strings to reach the verifier. The optional `id_shape_constraint` was never enabled in the reported runs.

Parsing first locates a JSON object with a regular expression and then requires the `answer` field to be either `True` or `Unknown`. A parse failure or disallowed `label` is classified as Other. A `True` output with a nonempty certificate is passed to the verifier as V'_T. An `Unknown` output with an empty certificate is Unknown-correct. An `Unknown` output with a nonempty certificate is Other under the protocol, so an invalid Unknown-with-support output counts against the model rather than receiving refusal credit.

The identifier contract uses strict string equality because certificate membership is a claim about the exact displayed objects in the current world. Whitespace trimming, bracket removal, punctuation repair, case folding, and other format normalization are therefore excluded. This distinction materially affects the reported range in the gpt-oss bracket mode (Appendix A11): strings such as `[rule3]` are not the same identifier as `rule3`, even when the underlying displayed rule exists and participates in a surviving proof.

A19 Nearest-Work Comparison

Table 25 compares ProofCover with the nearest concrete resources along the protocol properties most relevant to our claim. The table is descriptive: a dash means that the property is not the evaluation target of that work.

Table 25: Property-level comparison with nearest work. AP = answer-preserving.

Work	Multi-proof structure	Support-changing update	AP update	Submitted cert. targeted	Current-world cert. check
multiPRover (Saha et al., 2021)	✓	–	–	–	–
LogicGraph (Wu et al., 2026)	✓	–	–	–	–
LLM-TRes (Toroghi et al., 2024)	–	repair	–	–	–
Knowledge-update eval. (Mitchell et al., 2021; Zhong et al., 2023)	–	✓	not required	–	–
ProofCover	✓	✓	✓	✓	✓

multiPRover and LogicGraph make multiple derivations explicit, but evaluate proof generation or path coverage rather than certificate maintenance after a world update. LLM-TRes supports verifiable logical repair, while knowledge-update evaluations test behavior under changed information, but neither protocol targets a model-submitted certificate while holding the oracle answer fixed and guaranteeing surviving alternative support. Among these nearest resources, ProofCover combines those properties with deterministic current-world certificate verification; this combination, rather than multi-proof enumeration alone, defines the evaluation setting studied here.

A20 Full Stateful Comparison and Statistical Family

Table 26: Fresh and Stateful outcome distributions on the same model-specific eligible cases. Entries are Success/Stale/Other counts; Other includes Unknown refusals and unparseable outputs.

Model	Eligible	Fresh S/T/O	Stateful S/T/O
Qwen3-8B	134	68/27/39	72/14/48
Llama-3-8B-Instruct	60	34/17/9	41/19/0
Qwen3-32B	82	40/11/31	30/0/52
Llama-3.3-70B-Instruct	98	57/17/24	79/15/4
gpt-oss-20b	201	166/30/5	200/0/1
Phi-4-mini-instruct	86	50/22/14	53/17/16
Gemma-3-4B-IT	34	14/13/7	17/9/8
Qwen3-4B	126	72/25/29	105/20/1

Table 27 reports the paired Fresh-vs-Stateful McNemar tests for each outcome and model. The BH-FDR family comprises all $8 \times 3 = 24$ tests; inferences are based on the BH-adjusted q-values.

Table 27: Paired McNemar exact tests (Fresh vs. Stateful) per model and outcome, with BH-adjusted q-values over the 24-test family.

	Success		Stale		Other	
Model	p	q_{BH}	p	q_{BH}	p	q_{BH}
Qwen3-8B	0.659	0.832	0.0146	0.035	0.108	0.216
Llama-3-8B	0.167	0.308	0.824	0.860	0.0039	0.010
Qwen3-32B	0.087	0.190	0.00098	0.0029	0.00051	0.0017
Llama-3.3-70B	2.7e-5	0.00011	0.774	0.860	1.9e-6	0.0000092
gpt-oss-20b	1.1e-9	2.2e-8	1.9e-9	2.2e-8	0.500	0.706
Phi-4-mini	0.711	0.853	0.302	0.517	0.804	0.860
Gemma-3-4B-IT	0.549	0.732	0.388	0.620	1.000	1.000
Qwen3-4B	1.1e-6	0.0000064	0.473	0.706	7.5e-9	6.0e-8

A21 Original-World Certificate Validity and Eligibility

Table 28 reports valid original certificates over the shared set of 500 Formal-500 originals and the adaptive eligible count that follows. Eligibility requires a valid original certificate, so the adaptive sample is conditioned on certificate validity before any update is applied. The resulting count is therefore downstream of the model's ability to produce a valid certificate for the original world, rather than a measure based only on whether the original answer is correct.

gpt-oss-20b has 393 valid original certificates out of 500, corresponding to 78.6%, and yields 201 eligible adaptive cases. Qwen3-8B has 153 valid originals, or 30.6%, and yields 134 eligible cases. Qwen3-4B has 148 valid originals, or 29.6%, and yields 126 eligible cases. Llama-3.3-70B has 134 valid originals, or 26.8%, and yields 98 eligible cases. These counts demonstrate that even comparatively strong answer-level behavior does not determine the number of cases available for certificate auditing.

The remaining models produce smaller eligible sets. Phi-4-mini has 101 valid originals, or 20.2%, and yields 86 eligible cases. Qwen3-32B has 97 valid originals, or 19.4%, and yields 82 eligible cases. Llama-3-8B has 76 valid originals, or 15.2%, and yields 60 eligible cases. Gemma-3-4B-IT has 47 valid originals, or 9.4%, and yields 34 eligible cases. These model-specific counts explain the differing sample sizes in the adaptive tables, including the transition matrices and stale-mode summaries.

Conditioning on valid original certificates concentrates the adaptive analysis on cases where the model has already demonstrated certificate competence in the original world. The resulting false-pass rates are therefore conditional failure rates for this selected population; they do not estimate

failure prevalence over the full set of source worlds. Cases in which the model could not initially certify its answer are excluded before the update is evaluated, and no direction of bias for the unconditioned population is assumed.

The ordering also illustrates why model size or answer-level accuracy cannot be used to infer the adaptive sample count. Qwen3-8B has 153 valid originals, exceeding the 134 valid originals for Llama-3.3-70B, despite the larger model's higher answer-level accuracy. Eligibility is consequently model-conditional by construction. Comparisons across adaptive tables should retain the reported per-model denominators and should not be read as estimates for a common, unconditioned population.

Table 28: Original-world certificate validity over the shared 500 Formal-500 originals and the resulting adaptive eligibility count per model. "Valid orig." is the count of originals whose submitted certificate passes the original-world check; eligibility for the adaptive repair evaluation requires a valid original certificate, hence the correlation between the two columns.

Model	Valid orig. (of 500)	Rate	Adaptive eligible
Qwen3-8B	153	30.6%	134
Qwen3-4B	148	29.6%	126
Qwen3-32B	97	19.4%	82
Llama-3-8B-Instruct	76	15.2%	60
Llama-3.3-70B-Instruct	134	26.8%	98
Phi-4-mini-instruct	101	20.2%	86
Gemma-3-4B-IT	47	9.4%	34
gpt-oss-20b	393	78.6%	201

A22 Stale-Certificate Decomposition

This appendix section decomposes every Fresh Stale output (correct True answer, invalid updated certificate) into its certificate-provenance components, to test whether stale certificates simply reuse the now-deleted rule, and whether relaxing to lenient certificate semantics would change the conclusions. Table 29 reports the counts per model.

Table 29: Decomposition of Fresh Stale certificates. *target* = cites a removed rule; *halluc.* = cites an identifier absent from the updated world; *clean* = cites only surviving identifiers yet no surviving proof is fully contained in the cited set; *unchanged* = certificate identical to the valid Original certificate; *lenient-valid* = at least one surviving proof is fully contained in the cited set. The rightmost column is the mean fraction of the Original certificate retained in the Stale certificate.

Model	Stale	target	halluc.	clean	unchanged	lenient-valid	overlap
Qwen3-8B	27	0	1	26	6	1	0.54
Llama-3-8B	17	0	0	17	3	0	0.49
Qwen3-32B	11	0	1	10	2	0	0.49
Llama-3.3-70B	17	0	0	17	4	0	0.48
gpt-oss-20b	30	0	27	3	0	0	0.05
Phi-4-mini	22	0	0	22	8	0	0.50
Gemma-3-4B-IT	13	0	0	13	6	0	0.46
Qwen3-4B	25	0	0	25	8	0	0.56

Two findings follow. First, no Stale certificate in any model reuses the deleted rule: the Repair operator reliably forces models away from the invalidated support, so stale failures are not a certificate-reuse artifact. Second, except for gpt-oss-20b (27/30 stale certificates cite hallucinated identifiers), stale certificates cite only identifiers that survive the update; the failure is that the cited subset no longer contains any complete surviving proof — genuinely incomplete support with valid identifiers. Lenient certificate semantics (accepting any surviving proof fully contained in the cited set) rescue at most 1 of 162 stale outputs, so the strict-support verifier is not the source of the measured false passes. The overlap column shows stale certificates retain on average 46–56% of the original cer-

tificate identifiers for seven models, but only 5% for gpt-oss-20b, whose stale outputs are dominated by identifier hallucination rather than partial reuse.

A23 Oracle-Scan Construction Control

A natural control is to scan candidate rule deletions with the same exact symbolic oracle but without using proof-coverage ranking to choose among valid targets. For each Formal-500 Original world and each candidate rule, we delete the rule and evaluate the Repair contract under the frozen Horn semantics: the answer must remain True, at least one original minimal proof must be invalidated, and at least one proof must survive. The stored proof families implement these exact oracle checks; they are not used to rank the scan's valid candidates.

Across all 500 bases, every base has at least one valid Repair target (100% applicability), with a mean of 2.29 valid targets per base. The scan consumes a mean of 7.36 oracle calls per base (mean 2.95 calls to the first valid target) and runs in under 0.01 s total over the full set, so it is computationally negligible. However, the scan only enumerates the valid-target set; it does not by itself produce a construction. Ranking valid targets by proof-invalidation coverage recovers exactly the ProofCover Repair selection on every base (agreement 500/500). Relative to the frozen construction baselines over five seeds (random 32.2%, single-proof 43.4%, greedy-coverage 16.6% construction success), the scan-based selection attains 100% construction success at the same one-rule budget. This control shows that 100% Repair feasibility is not unique to ProofCover's coverage-based ranking. ProofCover's added value is the explicit characterization and ranking of affected versus surviving proof families, which makes the intervention semantics auditable.

A24 World-Level Operator Applicability and Eligibility Selectivity

Reviewers may ask whether the 100% construction-success claim and the adaptive stale rates depend on a selective applicability filter. We separate world-level applicability from model-conditional eligibility.

World level. The Repair operator requires a rule whose deletion invalidates at least one proof while preserving at least one proof. Over the 500 Formal-500 Original worlds, all 500 admit such a rule (100%), with a mean of 2.29 candidate targets per world, uniformly across all four depth–polarity families (AttNeg, AttNoneg, RelNeg, RelNoneg each at 100%). World-level operator applicability is therefore not a selection filter for this dataset.

Model level. For each model, eligibility further requires that the model's valid Original certificate actually cite such a target rule. This is model-conditional and does select cases: eligibility over valid Original certificates ranges from 51.2% (gpt-oss-20b) to 87.6% (Qwen3-8B), 73.1–85.2% for the remaining models. The exclusions are almost entirely certificates that do not cite any selectively-removable rule (no-target-in-certificate), not parse failures.

Structural selectivity. Table 30 compares proof-set structure between eligible and excluded cases per model. Eligible cases have mean proof counts of 2.56–2.93, versus 2.20–2.73 proofs for excluded cases. The eligible pool is therefore not structurally biased toward easier or harder worlds; the selectivity is confined to whether the model's own certificate happens to cite a selectively-removable rule, which is a property of model behavior rather than world structure. This supports reading the adaptive stale rates as model-conditional reliability on cases where the model demonstrably could certify originally, not as artifacts of a skewed world pool.

A25 Irrelevant-Addition Control on Unknown Triplets

As a separate control for generic update sensitivity, we measure response retention under an irrelevant update that does not change the query's derivability. The Unknown-Invariant world from the M3 Unknown benchmark adds an untriggerable rule while leaving the queried statement unprovable,

Table 30: Proof-set structure of eligible versus excluded (no-target-in-certificate) cases. Mean proofs is the average number of distinct minimal proofs of the query.

Model	eligible n	eligible mean proofs	excluded mean proofs
Qwen3-8B	134	2.93	2.53
Llama-3-8B	60	2.88	2.69
Qwen3-32B	82	2.67	2.73
Llama-3.3-70B	98	2.69	2.58
gpt-oss-20b	201	2.81	2.64
Phi-4-mini	86	2.74	2.20
Gemma-3-4B-IT	34	2.56	2.69
Qwen3-4B	126	2.82	2.59

so a valid Unknown response with an empty certificate remains correct. This control uses a different benchmark component from Adaptive Repair and therefore does not constitute a matched causal control for the targeted deletion. We condition on cases where the model validly refuses on the Original-Unknown world (correct Unknown answer with empty certificate) and measure how often the model still validly refuses on the Invariant world, where nothing about the queried statement's provability has changed.

Table 31: Irrelevant-addition control on Unknown triplets. Denominator = cases with a valid Unknown response on Original-Unknown; numerator = cases where the model again returns correct Unknown with an empty certificate on Unknown-Invariant.

Model	valid Unknown	retained	retention
Qwen3-8B	767	404	52.7%
Llama-3-8B	257	82	31.9%
Qwen3-32B	828	797	96.3%
Llama-3.3-70B	590	427	72.4%
gpt-oss-20b	847	842	99.4%
Phi-4-mini	743	479	64.5%
Gemma-3-4B-IT	382	189	49.5%
Qwen3-4B	677	420	62.0%

Retention varies from 31.9% (Llama-3-8B) to 99.4% (gpt-oss-20b), showing substantial model-to-model variation in robustness even to irrelevant edits. Because these cases come from the separate Unknown component, the comparison is descriptive rather than a paired estimate of the causal effect of targeted support invalidation. It nevertheless provides a useful reference for generic update sensitivity and motivates interpreting Adaptive Repair rates conditionally rather than as pure measures of update-specific failure.

A26 Inferential Robustness: World-Cluster Bootstrap and Depth Stratification

All eight models are evaluated on the same 500 fixed worlds, so per-case outcomes are a repeated-measures vector over a shared design rather than 500 independent draws. We complement the Wilson intervals reported in the main text with a world-level bootstrap (2,000 resamples of source-base IDs) of each model's false-pass share among answer-accepted outputs (Success+Stale), and with per-depth (qdep) stratified shares. Table 32 reports both.

The bootstrap intervals are close to the case-level Wilson intervals, so the descriptive conclusion is not driven by treating the accepted outputs as unclustered cases. Depth stratification likewise yields nonzero false-pass shares in the populated shallow and intermediate strata, without supporting a monotonic depth trend.

Table 32: World-level bootstrap 95% percentile intervals for the Fresh false-pass share, with per-query-depth (qdep) stratification. Denominators contain answer-accepted outputs (Success+Stale).

Model	stale/n	rate	world-bootstrap 95%
Qwen3-8B	27/95	28.4%	[20.0, 37.9]
Llama-3-8B	17/51	33.3%	[21.6, 47.1]
Qwen3-32B	11/51	21.6%	[11.8, 33.3]
Llama-3.3-70B	17/74	23.0%	[13.5, 32.4]
gpt-oss-20b	30/196	15.3%	[10.2, 20.9]
Phi-4-mini	22/72	30.6%	[19.4, 41.7]
Gemma-3-4B-IT	13/27	48.1%	[29.6, 66.7]
Qwen3-4B	25/97	25.8%	[16.5, 35.1]

A27 Artifact Manifest

Table 33 records the per-model quantization, `num_predict` setting, seed (`20260718`), record counts for the original files (500 per model), the Fresh and Stateful adaptive files matching each model's eligible set, the M3 files (2,565 records each), the synthetic control files, present for the five July-wave models only, and the UTC completion timestamps. The first five models (Qwen3-8B, Qwen3-32B, Llama-3-8B, Llama-3.3-70B, and gpt-oss-20b) completed on 2026-07-21; the second-wave models (Phi-4-mini-instruct, Gemma-3-4B-IT, and Qwen3-4B) completed on 2026-09-21. Each model directory also contains a `SHA256SUMS.txt` manifest over its files, per-run summaries, and the mechanistic-analysis files (probe, patching, and IRS results) where applicable. Every table number in this appendix is recomputable from these per-model files with the released reconciliation scripts.

The two run waves differ only in model, hardware placement, and date; they do not differ in prompt, seed, decoding, or verifier. The synthetic representation control was not re-run for the second-wave models, so those rows are marked rather than imputed.

Table 33: Frozen per-model artifacts. Counts are JSONL records in the released outcome files. Originals: 500 Formal-500 bases; Adaptive Fresh/Stateful: eligible adaptive repair cases per condition; M3: Unknown-triplet records (855 bases $\times$ 3 worlds); Synth.: paired symbolic and natural-language control records. †: the three backbones added in the second experimental wave did not re-run the synthetic representation control, which is reported for the five original models in Table 20. Completed: UTC timestamp recorded in each run manifest; all runs share seed 20260718 and the frozen `forward_verified` prompt.

Model	Quant.	`num_pred`	Seed	Orig.	Fresh	Stateful	M3	Synth. †	Completed
Qwen3-8B	FP16	1024	20260718	500	134	134	2,565	1,400	2026-07-21T15:49:35Z
Qwen3-4B	BF16	1024	20260718	500	126	126	2,565	–	2026-09-21T08:36:55Z
Qwen3-32B	FP16	1024	20260718	500	82	82	2,565	1,400	2026-07-21T08:19:09Z
Llama-3-8B-Instruct	FP16	512	20260718	500	60	60	2,565	1,400	2026-07-21T15:49:36Z
Llama-3.3-70B-Instruct	FP16	512	20260718	500	98	98	2,565	1,400	2026-07-21T08:19:09Z
Phi-4-mini-instruct	BF16	1024	20260718	500	86	86	2,565	–	2026-09-21T07:37:45Z
Gemma-3-4B-IT	BF16	1024	20260718	500	34	34	2,565	–	2026-09-21T08:18:58Z
gpt-oss-20b	MXFP4	2048	20260718	500	201	201	2,565	1,400	2026-07-21T15:51:20Z

A28 Stale-Certificate Audit for Gemma-3-4B-IT

We manually audited all 13 Fresh Stale outputs for Gemma-3-4B-IT (the model with the highest false-pass share and the smallest eligible set, 34). In all 13 cases the model predicts True and submits a certificate that cites only identifiers that survive the update (no removed rules, no nonexistent identifiers), yet the cited subset no longer suffices to derive the query — i.e., the model constructs genuinely incomplete support in the updated world rather than reusing deleted rules or hallucinating new ones. The certificates range from 2 to 8 cited items (median 3). This uniformity indicates the false passes reflect a systematic certificate-construction failure, not a parsing or identifier-mapping

Table 34: Complete RuleTaker-MP Adaptive Repair results (Fresh). Stale rate uses Eligible; false-pass share uses answer-accepted outputs (Success+Invalid Support).

Model	Orig. valid	Eligible	Success	Invalid supp.	Stale rate	False-pass share
Phi-4-mini-instruct	256	219	145	41	18.7 ([14.1, 24.4])%	22.0 ([16.7, 28.5])%
Gemma-3-4B-IT	220	125	88	21	16.8 ([11.3, 24.3])%	19.3 ([13.0, 27.7])%
Qwen3-4B	330	216	146	33	15.3 ([11.1, 20.7])%	18.4 ([13.4, 24.8])%
Qwen3-8B	378	236	155	42	17.8 ([13.4, 23.2])%	21.3 ([16.2, 27.6])%
Llama-3-8B-Instruct	282	197	139	38	19.3 ([14.4, 25.4])%	21.5 ([16.1, 28.1])%
gpt-oss-20b	740	289	259	17	5.9 ([3.7, 9.2])%	6.2 ([3.9, 9.6])%
Qwen3-32B	298	126	76	14	11.1 ([6.7, 17.8])%	15.6 ([9.5, 24.4])%
Llama-3.3-70B-Instruct	416	262	172	23	8.8 ([5.9, 12.8])%	11.8 ([8.0, 17.1])%

Table 35: False-pass share by proof-set size and query depth (Fresh, answer-accepted denominator).

Stratum	Share	95% CI
proof count=2	15.9%	[13.3, 18.8]
proof count=3-4	17.1%	[13.8, 21.1]
proof count=5+	16.0%	[12.3, 20.5]
query depth=0-1	13.7%	[11.7, 15.9]
query depth=2-3	25.2%	[20.7, 30.2]
query depth=4-5	13.3%	[7.4, 22.8]

artifact; the small eligible set (34) is driven by the same weakness in Original certificate validity (47/500). The outcome distribution is consistent with the lower Graft certificate accuracy (26.4%) observed in the independent Unknown evaluation.

A29 RuleTaker-MP Cross-Benchmark Results

Table 34 reports the complete per-model Fresh Adaptive Repair outcomes on RuleTaker-MP, including Wilson intervals for stale rates and false-pass shares. Table 35 stratifies answer-accepted false-pass shares by proof-set size and query depth. The proof-count strata are nearly flat (15.9–17.1%), while query depth is non-monotonic: 13.7% at depth 0–1, 25.2% at depth 2–3, and 13.3% at depth 4–5. We therefore treat complexity as a descriptive stratification rather than evidence of monotonic degradation.