EVIPAIR: PAIRED BEHAVIORAL DIAGNOSTICS FOR EVIDENCE-RESPONSIVE DECISIONS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evidence-based decisions should follow changed premises and retain uncertainty when decisive information is missing. We introduce *EviPair*, a paired behavioral diagnostic combining entailed hints, premise reversals, irrelevant additions, and verified evidence removal on the same underlying problems. Joint correctness tests whether a model follows both premise states; a shared unknown condition tests whether it avoids unsupported assertions. This design measures *Evidential Myopia* as sensitivity to making an already entailed consequence explicit, while separating hint benefit from distraction. Our relational, numerical-aggregation, and natural-language studies cover two open-weight backbones and multiple hosted models, including a seven-family API extension. A screened Qwen2.5 configuration exhibits a 7.42-point hint benefit and a 7.03-point distraction cost in joint correctness. Hosted configurations show varied response profiles, including near-ceiling screened performance. Five-view SFT checkpoints improve premise tracking and uncertainty handling over unadapted backbones on new backgrounds from the same synthetic construction, without test-time gold hints. Matched five-view comparisons find no stable additional benefit from entailment consistency over SFT or paraphrase controls. Together, these results provide a paired evaluation protocol for distinguishing useful evidence interventions from distraction and unsupported certainty.

1 INTRODUCTION

A language model answering from a user's documents should follow the evidence in those documents, including consequences that are not stated directly. Overlooking a consequence can lead to a wrong answer or unnecessary uncertainty; asserting an answer after decisive evidence is removed creates an unsupported claim. Understanding these behaviors matters for grounded question answering and for reducing hallucinations in evidence-based decisions.

We study *Evidential Myopia* through a specific observable gap: a model answers more reliably when an entailed intermediate conclusion is stated explicitly, although the premises already determine the same answer. Consider a ledger that counts approved entries before a deadline. Supplying an entailed subtotal preserves the answer, changing a decisive entry reverses it, and hiding an eligible amount can make it unknown. These interventions let us ask whether better answers reflect reliable responses to the evidence.

Compositionality gaps (Press et al., 2023), overlooked implicit context (Cai et al., 2026), and failures to use retrieved evidence (Ma et al., 2026) motivate this question. A hint gain alone leaves two issues open: whether the decision follows a changed premise, and whether it becomes uncertain when the evidence is insufficient. Comparing relevant and irrelevant additions introduces another ambiguity, because the difference can include harm from the irrelevant text. The unmodified input is therefore an essential baseline.

EviPair links these comparisons on the same decision problem. Two opposite premise states each appear with implicit evidence, a relevant entailed hint, or an irrelevant addition; a seventh input removes decisive evidence. Verified labels and opposite-answer completion witnesses support scoring across all conditions. Joint correctness requires the model to follow both premise states, while the

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.

unknown condition measures unsupported assertions. This design connects an explicitation gain to the behaviors needed for reliable evidence use.

Paired diagnosis. EviPair connects hint benefit to correct answers under premise reversal and warranted uncertainty on the same background. Joint correctness excludes agreement on two wrong answers; the unmodified input separates improvement from distraction. **Empirical characterization.** Relational, ledger, and natural-language studies compare two open-weight backbones and multiple hosted-model families. The two local evaluation cohorts contain 143,104 model responses; an eleven-configuration API extension records 29,341 completed answers. A Qwen2.5 follow-up observes a 7.42-point hint benefit and a 7.03-point distraction cost after a separate behavioral screen, while other settings delineate the effect's range. **Evaluation of adaptation.** Two backbones and three training seeds test evidence-response supervision on new synthetic backgrounds. Matched SFT and paraphrase controls assess the extra consistency penalty; fixed FOLIO evaluation tests transfer beyond the synthetic construction.

2 Related Work

Evidence use and composition. Self-ask measures failure to compose correctly answered subproblems and uses explicit decomposition to improve answers (Press et al., 2023). MixRea evaluates explicit–implicit reasoning and completes overlooked relations (Cai et al., 2026); Proof-of-Use targets grounding in retrieved evidence (Ma et al., 2026). We test explicitation across premise reversals and against two controls, separating improvement from distraction by irrelevant context (Shi et al., 2023).

Abstention and contextual faithfulness. Unanswerable-question benchmarks and selective QA evaluate abstention (Rajpurkar et al., 2018; Kamath et al., 2020); self-evaluation and semantic entropy estimate reliability (Kadavath et al., 2022; Kuhn et al., 2023). Our unknown label instead requires opposite-label compatible completions. Context-faithful prompting and counterfactual QA distinguish supplied context from parametric preference (Zhou et al., 2023; Neeman et al., 2023). False-premise hallucination studies test responses to misleading premises (Yuan et al., 2024), and CounterBench tests counterfactual inference under formal rules (Chen et al., 2026). EviPair links premise changes to answer-preserving hints, an irrelevant-content control, and verified evidence removal on the same background. Context position (Liu et al., 2024), knowledge conflict (Jin et al., 2024), and knowledge priming (Sun et al., 2025; 2024) motivate alternative accounts; our additions occupy a fixed position and follow from the premises.

Behavioral and internal evidence. CheckList supplies general invariance and directional-expectation tests (Ribeiro et al., 2020); CLUTRR, RuleTaker, and ProofWriter provide controlled reasoning constructions (Sinha et al., 2019; Clark et al., 2020; Tafjord et al., 2021). Our specific contribution couples a two-state, three-presentation grid to checked evidence removal, separate behavioral screening, and joint response endpoints. The behavioral target differs from hidden-state patching in CREME (Li et al., 2024) and generated-reasoning faithfulness (Creswell & Shanahan, 2022; Han et al., 2026; Shen et al., 2026).

Inference and training interventions. Chain-of-thought and self-consistency use intermediate generation and sampled paths (Wei et al., 2022; Wang et al., 2023); ParaRel motivates a matched paraphrase-consistency control (Elazar et al., 2021). FRODO trains rationale generation and a separate reasoner using counterfactual rationales and preference objectives (Paul et al., 2024). Our counterfactuals edit source premises, and all main training arms receive the same five labeled views. Our matched comparison measures the extra consistency penalty within a direct-answer model.

3 Diagnosing Evidential Myopia with EviPair

Let E be the supplied evidence, R the allowed rules, and q a statement to judge. We require a nonempty set of compatible worlds,

$$\mathcal{W}_R(E) = \{w : w \models R \wedge E\} \neq \varnothing. \tag{1}$$

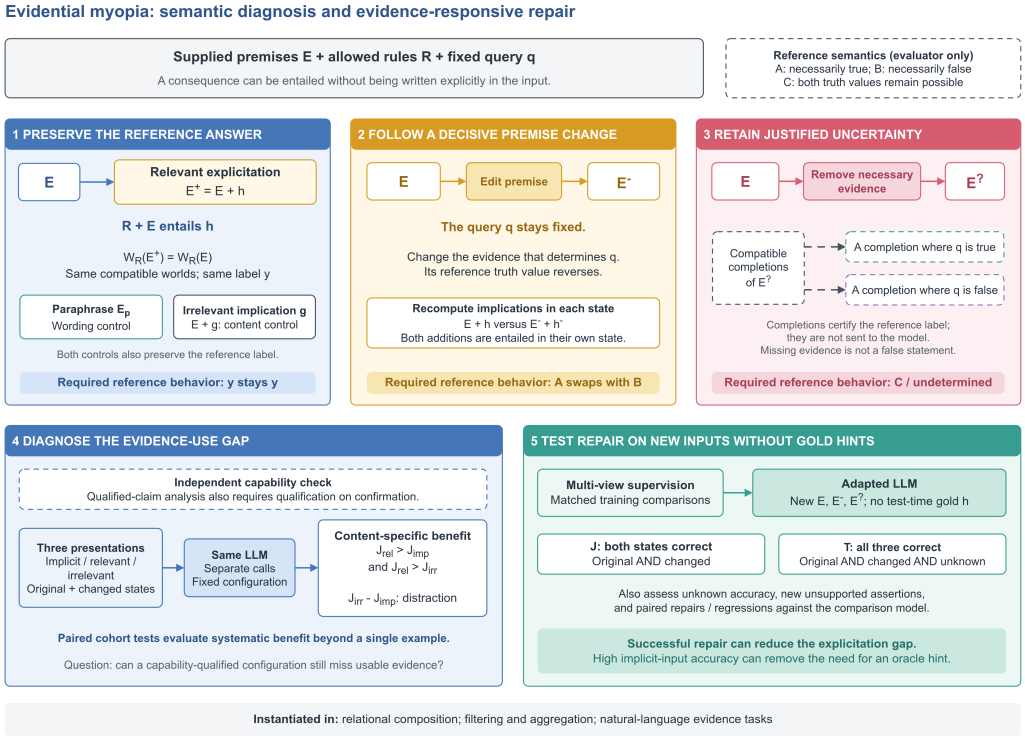


Figure 1: Paired diagnostic: an entailed implication preserves the reference answer, a decisive premise edit reverses it, and insufficient evidence admits both truth values. The figure separates the standalone behavioral screen (qualification) from held-out paired evaluation. Completion witnesses remain evaluator-only; implicit evaluation supplies no gold implication. The shared unknown input receives no explicitation treatment.

The reference label is A when q is true in every compatible world, B when it is false in every compatible world, and C when both truth values are possible. Unknown requires two compatible completions with opposite answers. We examine the explicitation gap together with three response patterns: unchanged definite answers after premise reversal, C responses on determined inputs, and unsupported A/B responses on unknown inputs.

For a relevant intermediate statement h with $R \cup E \models h$, define $E^+ = E \cup \{h\}$. The statement h can itself assert that a relationship does not hold, when that negative statement is entailed by the evidence. Classical entailment gives $\mathcal{W}_R(E^+) = \mathcal{W}_R(E)$, certifying preservation of the reference label (Appendix E). This equality concerns the reference semantics; the textual presentations can differ in computational demands. Explicitation supplies an oracle-verified intermediate consequence for diagnosis; implicit-input evaluation supplies no such hint.

Seven paired evidence conditions. For each background, we construct two premise states $E_{i,0}, E_{i,1}$ that give opposite definite answers to the same query. Each state appears as implicit evidence (imp), evidence with its relevant implication (rel), and evidence with an irrelevant implication (irr). A public-text evaluator checks both reference answers and every added implication in its own premise state. The seventh, shared input removes decisive evidence and has label C. Its two completion witnesses stay with the evaluator. Relevant and irrelevant additions share the sentence template, relation type, and position. The irrelevant statement concerns another subject; its truth polarity is balanced independently of the relevant relation and query label. Polarity is balanced across backgrounds; per-pair token lengths and query overlap are not matched. Each input is answered in a fresh context.

Joint correctness under premise reversal. Joint correctness requires correct answers under both determined states, preventing agreement on two wrong answers from counting as success. For presentation $s \in \{\text{imp}, \text{rel}, \text{irr}\}$, let $c_{i,s,0}$ and $c_{i,s,1}$ indicate correct answers on background i before and after the premise change. We measure

$$J_s = \frac{1}{n} \sum_{i=1}^{n} c_{i,s,0} c_{i,s,1}. \tag{2}$$

We compare relevant hints with both baselines and separately measure distraction:

$$\begin{aligned} \Delta_{\text{imp}} = J_{\text{rel}} - J_{\text{imp}}, \qquad \Delta_{\text{irr}} = J_{\text{rel}} - J_{\text{irr}}, \\ D = J_{\text{irr}} - J_{\text{imp}} = \Delta_{\text{imp}} - \Delta_{\text{irr}}. \end{aligned} \tag{3}$$

A content-sensitive benefit requires positive evidence against both controls; $D < 0$ identifies harm from irrelevant addition. The identity $\Delta_{\text{irr}} = \Delta_{\text{imp}} - D$ explains why the relevant–irrelevant contrast alone can overstate improvement over the original input.

Warranted uncertainty and response profiles. Let $c_{i,?,C}$ indicate a correct C answer on the shared unknown input. Three-state joint correctness is

$$T_s = \frac{1}{n} \sum_{i=1}^{n} c_{i,s,0} c_{i,s,1} c_{i,?,C}. \tag{4}$$

"All three" means T_{imp} unless specified otherwise. We report unknown accuracy and A/B assertions on reference-C inputs separately. The unknown input receives no explicitation treatment. Thus $T_{\text{rel}} - T_{\text{imp}}$ measures the change in the three-state joint endpoint induced by explicitation; the denominator remains all n backgrounds. Completed valid pairs are classified as both correct, an unchanged definite answer, unknown on determined evidence, or reversed wrong answers. Gain and regression counts show which responses change. Invalid outputs count as incorrect; service-missing outcomes remain unknown and receive worst-case bounds. The background, not each question, is the paired unit. Appendix E relates J and T to their marginal accuracies.

Behavioral screening and follow-up selection. A separate screen requires at least 90% observed accuracy for positive, negative, and unknown relation judgments and explicit decisions; passing configurations are *screen-qualified*. For each Qwen backbone, the existing seed-17 SFT checkpoint was selected using the earlier synthetic development split. After the pilot, the screening protocol fixed the candidate order as base model followed by this checkpoint. Screening selected the configuration; it did not select the checkpoint or use implicit-decision outcomes. The selected configurations, new backgrounds, and analysis were frozen before follow-up collection. The Qwen follow-up uses a four-test Holm family: $J_{\text{rel}} - J_{\text{imp}}$ and $J_{\text{rel}} - J_{\text{irr}}$ for each backbone. Relation probes on the follow-up backgrounds provide an auxiliary analysis (Appendix N).

4 Experimental Design

Evaluation coverage. Figure 2 connects the diagnostic and training studies; Appendix Table 7 summarizes their coverage. The two local evaluation cohorts contain 143,104 responses: 119,808 from the nine-condition cohort and 23,296 from the seven-condition cohort. Both evaluate the same 26 configurations from two backbones, including three seeds per adapted arm. The eleven-configuration API study adds 29,341 completed responses from 35,200 planned inputs. Appendix Table 9 gives cohort-level counts; Appendix Table 39 reports every API configuration. Statistical tests use backgrounds as their paired unit.

EviPair materials and splits. EviPair is a versioned suite of author-constructed relational and ledger questions. Its data card and per-cohort manifest distinguish backgrounds, question views, and repeated responses. The core aggregates six relational and ledger cohorts: 2,048 training, 88 development, and 768 evaluation backgrounds, totaling 2,904 backgrounds and 25,576 question texts across views (Appendix C, Table 4). Maternal ancestry and delegated permission are two surface realizations of one relational construction, each determining a resource total. The relational companion contributes 2,048 training, 64 development, and 512 validation backgrounds in total, balanced

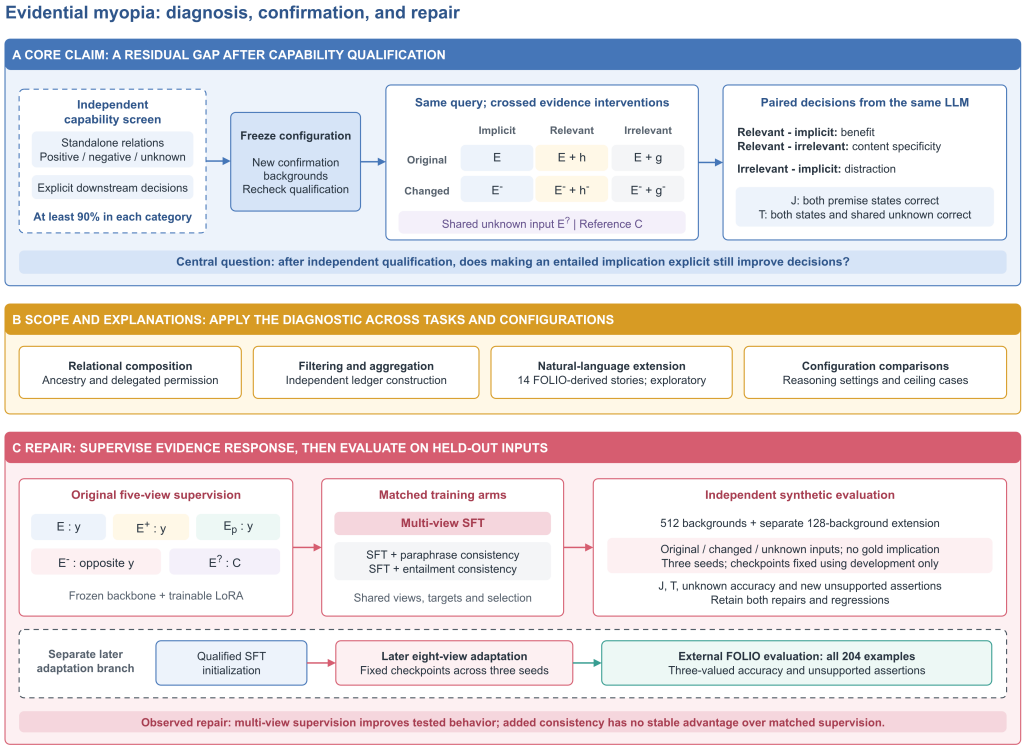


Figure 2: EviPair research framework: a standalone behavioral screen selects configurations for held-out paired evaluation; task and model comparisons assess scope, and supervised adaptation tests evidence-responsive learning. The five-view study compares matched supervision and consistency. A separate eight-view study supplies fixed checkpoints for FOLIO. Qualification denotes empirical behavioral screening.

across the two families, with path lengths two to four and 17–25 entities. Canonicalization removes names and node indices while retaining query roles; duplicate original, changed, or unknown forests are rejected across backgrounds and splits. All views of a background remain together. Templates and depth ranges are shared across splits. The remaining core cohorts are the 128-background relational diagnostic and the ledger's 24 development and 128 validation backgrounds. Later screening and follow-up collections are registered separately.

Unknown cases with matched surface cues. Every input omits one nonterminal relationship: irrelevant to a determined query, necessary to an unknown query. Original and unknown versions match entity lists, record grammar and counts, and endpoint mention counts. A public-text evaluator verifies each label, each implication, and two opposite-label compatible completions for every unknown case. Private graphs, labels, and completion witnesses never enter requests. The nine-condition companion inventory combines five training views, an irrelevant implication, and three standalone relation judgments for training and auxiliary evaluation. The EviPair diagnostic instead uses seven paired decision inputs per background (Appendix Table 14).

Held-out relational API grid. Figure 1 crosses two determined premise states with three presentations and adds one unknown input. Each implication is verified in its own premise state. Before new calls, we froze all seven views on 128 backgrounds (64 per family), disjoint from the 7,872 prior canonical forests, along with configuration and analysis. Every input explicitly distinguishes different names; each configuration has 896 planned answers. This grid is distinct from the Qwen screen (128 backgrounds) and follow-up (256 backgrounds). Exact one-sided background-paired tests compare J_{rel} with J_{imp} and J_{irr}. The initial relational Luna analysis uses a two-test Holm

family. The subsequent GLM, Gemini, and API Qwen extension uses a separate six-test exploratory family (Appendix J). Response-pattern transitions and three-state joint correctness are secondary. Development reuse and localization probes are detailed in Appendix T.

Non-graph range test. A complete ledger requires summing a project's approved entries dated on or before day 14, then applying a threshold to determine a fund total. Every view has one unspecified quantity: ineligible for determined queries, decisive for unknown queries. Public-text evaluation checks the reference labels and opposite-label completions. We freeze 24 development and 128 new validation backgrounds, seven views, four API configurations, and an eight-test Holm family (Appendix L.3).

Models, inference, and missingness. An API configuration denotes a recorded gateway request setting, without independent verification of hidden weights or system prompts. Local models are Qwen3-8B (Yang et al., 2025) and Qwen2.5-7B-Instruct (Qwen et al., 2024), using BF16 greedy generation, 64 output tokens, and Qwen3 thinking disabled. In the original grid, Luna (`gpt-5.6-luna`) uses a 16,384-token cap and effort none; the other HTTP configurations use 4,096 tokens. The separate reasoning comparison uses 16,384 for both settings of every model. DeepSeek disables thinking; Gemini, GLM, and API Qwen leave it unspecified. These are unequal native budgets. Exact settings and the original 16-endpoint inventory are in Appendix J. Completed invalid or truncated answers are task failures; service failures and uncalled cases have unknown correctness. Tests and intervals retain worst-case missing outcomes. Only recorded transport or transient HTTP failures admit bounded retries.

5 Evidence-Response Training

From diagnosis to supervision. The diagnostic requires three behaviors: answer determined questions, change the answer when a decisive premise changes, and return unknown when necessary evidence is missing. We evaluate these requirements using supervised views of each training background and separate LoRA modules on frozen Qwen backbones. Each training background provides five inputs: the original E, its conservative explicitation E^+, a paraphrase E_p, a changed-premise version E^-, and an insufficient-evidence version $E^?$. The last two have their own reference labels; consistency is not imposed across them. The adapted model's implicit-input evaluation supplies no gold intermediate statement.

Does explicit consistency add to supervision? For the five supervised answer sequences, let $\mathcal{L}_{\text{ans}}$ be the mean cross-entropy over the answer and end-of-turn targets. The candidate objective is

$$\mathcal{L}_{\text{ent}} = \mathcal{L}_{\text{ans}} + \lambda \, \mathrm{JS}\Big(\widetilde{P}_\theta(E), \widetilde{P}_\theta(E^+)\Big), \qquad \lambda \geq 0. \tag{5}$$

Here $\widetilde{P}$ partitions the answer-position distribution into A/B/C/OTHER, and both views receive gradients. We use $\lambda = 0.5$. Shared-label cross-entropy already bounds pair disagreement on supervised inputs (Proposition 2), motivating a test of the penalty's additional training effect. Appendix F defines the answer-position surrogate. Evaluation uses free generation and retains invalid and truncated outputs.

Matched supervision comparisons. Five-view SFT (SFT in the five-view tables and figures) minimizes $\mathcal{L}_{\text{ans}}$ on all five views; paraphrase consistency adds $\lambda \, \mathrm{JS}(\widetilde{P}_\theta(E), \widetilde{P}_\theta(E_p))$. All three share data order, targets, forward views, optimizer, update count, and development selection. The No CF/U ablation removes changed-premise and unknown supervision, including all C targets, and changes the loss weighting. It retains five forward views but backpropagates fewer losses. The comparison measures the extra penalty under identical five-view supervision; it does not estimate the advantage of five views over one.

We use rank-16 LoRA (Hu et al., 2022) for 512 updates per seed, with four backgrounds and five views per update. Every 64 updates, mean development accuracy on original, changed, and unknown inputs selects the checkpoint; earlier checkpoints win ties. Validation never enters selection. Appendix Table 17 specifies the complete settings.

6 Evidence Response and Supervised Adaptation

The results address the gap after behavioral screening, its range across tasks and settings, the observed error patterns, and supervised adaptation. All joint counts use background denominators.

6.1 The Gap after Behavioral Screening

Qwen2.5 shows a relevant-hint gain and irrelevant-context cost. The seed-17 Qwen2.5 and Qwen3 SFT checkpoints pass the behavioral screen; both base models fail. On 256 new backgrounds, both SFT configurations also meet the screening criteria on standalone relation probes. The four-test Holm family contains two control contrasts for each backbone (Appendix N).

Configuration	J_{imp}	J_{rel}	J_{irr}	Δ_{imp}	Δ_{irr}	D	Hint G/L	Irr. G/L
Qwen2.5 SFT	237	256	219	$+7.42$	$+14.45$	-7.03	19/0	3/21
Qwen3 SFT	255	256	248	$+0.39$	$+3.13$	-2.73	1/0	0/7
Grok, low	256	256	256	0	0	0	0/0	0/0
Grok, high	256	256	256	0	0	0	0/0	0/0

Table 1: Paired results after behavioral screening, with 256 backgrounds per configuration. Joint counts and gains/losses use background denominators; contrasts are percentage points. All configurations shown have complete responses. Qwen2.5's two control contrasts have Holm-adjusted $p = 5.72 \times 10^{-6}$ and 2.91×10^{-11}; D describes distraction; G/L denotes gains/losses. The Qwen comparison and Grok low/high replication use separate four-test families (Appendix O).

Relevant hints correct 19 Qwen2.5 premise pairs without introducing a regression, raising joint correctness from 237/256 to 256/256. Irrelevant additions correct three pairs and disrupt 21, reducing the count to 219/256 (Table 1). Thus the 14.45-point relevant–irrelevant difference combines a 7.42-point gain over the original input with a 7.03-point distraction cost. Qwen3 and Grok provide near-ceiling and ceiling comparisons. Qwen2.5 answers the shared unknown input correctly on 234/256 backgrounds.

6.2 Scope across Tasks and Model Settings

Relational and ledger decisions. On 128 new ledger backgrounds, Luna's implicit/relevant/irrelevant $J = 3/119/9$: relevant explicitation corrects 116 premise pairs without regression. The corresponding relational counts are $2/120/5$ on a separate 128-background cohort. Both control contrasts pass their declared Holm families. On the relational cohort, 102 of 110 pairs containing unwarranted unknown answers become jointly correct. The shared ledger unknown input produces 24 unsupported assertions, showing that better determined answers can coexist with unsupported certainty. The relational Luna configuration fails the standalone behavioral screen. Gemini reaches ceiling in both constructions, and API Qwen does so in the relational grid (Appendix Figure 5).

Hosted-model comparisons. Table 2 displays all eleven planned API configurations; ten return scientific responses, totaling 29,341 answers. The extension fixes short- and long-path strata with 128 confirmation backgrounds each. Gemini Pro, Gemini Flash, API Qwen Max, and Kimi complete all inputs and meet the positive-, negative-, and unknown-relation and explicit-decision thresholds at both screening and confirmation; their implicit and relevant-explicit J counts are 127–128/128. No stratum supports a positive qualified gap in the fixed 44-test family. GPT-6 Luna has large gains but fails the behavioral screen. Incomplete configurations retain missing-outcome bounds (Appendix S). These results locate the gap in particular configurations alongside larger deficits in settings that fail screening and near-ceiling screened performance.

Natural-language evidence. The same seven conditions are constructed for 14 FOLIO-derived stories. Qwen2.5 SFT gains three and GPT-6 Luna four jointly correct premise pairs; all stories with corrected pairs also receive a correct shared unknown answer. GLM Flash instead ties its relevant and irrelevant joint counts, and Kimi is at ceiling. Claude completes all 98 inputs but violates the original answer-only format; a uniform final-line sensitivity parser recovers all answers correctly.

API configuration	Completed	Correct	Short joint counts	Long joint counts
GPT-6 Luna	3,200	1,944	30/123/33	26/125/25
GLM-5.3	1,888	1,879	—	—
GLM-5.3 Flash	2,786	2,647	—	—
Kimi K3	3,200	3,167	127/127/128	128/128/128
Qwen3.8 Max	3,200	3,199	128/128/128	128/128/128
Qwen3.8-27B	0	—	—	—
Grok 4.7	3,194	3,193	—	—
Gemini 3.8 Flash	3,200	3,197	128/128/128	128/128/127
Gemini 3.1 Pro Preview	3,200	3,200	128/128/128	128/128/128
Claude Fable 5	3,188	3,086	—	—
Claude Opus 5	2,285	1,655	—	—

Table 2: API coverage and paired outcomes. Each configuration has 3,200 planned questions, including screening and confirmation. Correct counts concern completed answers; completed invalid outputs count as wrong. Joint counts are ordered $J_{\mathrm{imp}}/J_{\mathrm{rel}}/J_{\mathrm{irr}}$, use 128 backgrounds per stratum and are shown for service-complete configurations. Dashes mark incomplete or unavailable paired readouts, not zero accuracy; Qwen3.8-27B did not pass interface admission. Native settings differ. Appendix S retains missing-outcome bounds and screening results.

This exploratory extension adds natural-language instances and exposes output interface sensitivity. None of its paired tests passes the declared correction (Appendix R).

6.3 Interpreting the Residual Gap

Within-background relation probes. The auxiliary analysis requires correct relation judgments on both premise states. For Qwen2.5, it retains five gains over implicit input across the 256 backgrounds (adjusted $p = 0.0625$; Appendix N).

Distraction and answer preference. Qwen2.5's implicit A and B accuracy is 245/256 and 244/256; relevant explicitation makes both 256/256, while irrelevant addition yields 239/256 and 231/256 (Appendix N). Thus the residual occurs with both labels. Complementing the query corrects four of the five same-background-qualified failures and introduces one new failure; their A/B patterns are not constant. Ten backgrounds follow the changed query while ignoring premise reversal (Appendix N). These response patterns distinguish query sensitivity from premise tracking.

Native reasoning settings. Higher reasoning settings increase implicit joint correctness on the same backgrounds, with reported completion tokens rising about 326-fold for DeepSeek Flash and 406-fold for Pro. Full results, token use, and missing-response bounds appear in Appendix L.1.

6.4 Supervised Adaptation on New Backgrounds

All 26 configurations in the five-view validation are complete, with checkpoints frozen before dispatch (Appendix Figure 14 and Table 42). Qwen2.5's three SFT seeds raise both-state correctness from 1/512 to 469–480/512 and unknown correctness from 213/512 to 482–496/512. Absolute unsupported counts fall from 299 to 16–30 out of 512, while seven, seven, and nine cases become newly unsupported. Qwen3's corresponding changes are 5 to 502–506 and 7 to 500–509. These SFT/base readouts are secondary; Appendix Table 43 gives every seed.

On original-input accuracy, entailment consistency gains an average of 1.50 points over SFT and 0.65 over paraphrase for Qwen2.5; Qwen3's mean is below both baselines. Unknown and changed-premise effects also vary by seed. The added term does not meet the prespecified superiority and risk conjunction: seedwise tests do not pass, and new-assertion bounds exceed 2%. Matched SFT already provides the large observed improvement, leaving limited headroom for the original five-point target. Appendix V reports the seedwise outcomes; Appendix M preserves the original criteria and their status.

The No CF/U ablation leaves zero correct unknown answers for Qwen2.5 and 0/2/2 for Qwen3 across the three seeds.

Replication on a new seven-condition cohort. The local seven-condition extension repeats this comparison on 128 new backgrounds for all 26 configurations. SFT raises implicit both-state correctness from 0 to 117–121 for Qwen2.5 and from 1 to 126–127 for Qwen3; both three-seed conjunctions pass the ten-test Holm family. Entailment consistency has no superiority over either matched baseline (Appendix U). An exploratory eight-view development study has small mean gains with mixed seedwise signs (Appendix P); the fixed external test below assesses those checkpoints beyond the synthetic construction.

Fixed external evaluation. On all 204 examples of the public FOLIO v0.0 validation release (Han et al., 2024), nine fixed adapted Qwen2.5 checkpoints and one unadapted reference produce 2,040 outputs. The adapted checkpoints come from the separate eight-view study: three arms with three seeds each. Their mean accuracies are 61.11% for answer SFT, 58.01% for paraphrase, and 57.84% for semantic consistency. Every adapted checkpoint exceeds the unadapted reference's 50.00%. The semantic mean is 3.27 points below SFT and 0.16 below paraphrase; two of three seeds are negative against both controls. The external test shows no average pairing advantage (Appendix Q).

7 Discussion and Limitations

Implications for evidence use. A relevant hint can help while irrelevant context distracts, and improved determined answers can coexist with unsupported assertions on missing evidence. EviPair measures these responses together. The supervised study turns the diagnostic requirements into training targets evaluated without gold hints; matched supervision measures the consistency term's additional contribution.

Scope and future work. The clearest gap after screening occurs in one Qwen2.5 setting. Ceiling results identify configurations for which the present diagnostic is no longer demanding. The screen establishes standalone behavioral competence on the tested probes, not correct intermediate representations or internal derivations within the full decision context. Synthetic splits share templates and depths; the 14-story intervention study and unchanged FOLIO evaluation examine different aspects of natural-language behavior. A shallow structural predictor reaches 38.87% against a 33.33% baseline, indicating residual surface cues without attributing the paired hint effect (Appendix C). Held-out templates and depths, larger natural cohorts, and single-view and class-balanced training controls are priorities for extending the study. Native reasoning settings differ in realized token use; matched-compute and internal interventions would address efficiency and mechanism (Appendix B).

8 Conclusion

EviPair couples entailed hints, premise reversals, distraction, and warranted uncertainty on the same problems. The Qwen2.5 follow-up shows an explicitation gap after standalone screening, while broad API coverage identifies near-ceiling settings. Five-view SFT checkpoints improve over unadapted backbones on new backgrounds from the same synthetic construction; matched five-view comparisons find no stable additional consistency benefit. The framework supports evaluation of accurate evidence-based decisions as information changes or becomes incomplete.

AI USE STATEMENT

We used ChatGPT to aid and polish writing, retrieve and discover related work, support research ideation and execution, and draft sections of the paper. Research assistance included refining hypotheses and experimental designs, developing synthetic-data and evaluation code, assisting with mathematical analysis and proofs, and interpreting results. The manuscript describes the reference checks and validation procedures used for the reported experiments. The authors are responsible for the final content and findings.

REPRODUCIBILITY STATEMENT

Sections 4 and 5 describe input construction, statistical units, and training controls. Appendices E, F, and G state the mathematical assumptions, derivations, and counterexamples. Appendix J gives submitted inference settings and offline reproduction archives for the closed API results; Appendix I provides traceable examples. The supplied code and raw captures retain incomplete outcomes and recompute the reported statistics. The original claim criteria and later local extension are recorded separately in Appendices M and U. The original local validation is complete; the later crossed study is reported separately.

The portable local-inference replay covers all 792 development controls, including input and output tokens, thinking boundaries, and matched SFT comparisons (Appendix L.2). The ledger generator reconstructs all frozen development and validation inputs and their public-text references independently of model collection.

REFERENCES

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(85):2399–2434, 2006. URL https://jmlr.org/papers/v7/belkin06a.html.

Yuanqing Cai, Ziyi Huang, Minhao Liu, Lixin Duan, Wen Li, and Yanru Zhang. MixRea: Benchmarking explicit-implicit reasoning in large language models, 2026. URL https://arxiv.org/abs/2605.20128. Preprint, version 1.

Yuefei Chen, Vivek K. Singh, Jing Ma, and Ruixiang Tang. CounterBench: Evaluating and improving counterfactual reasoning in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30350–30358, 2026. doi: 10.1609/aaai.v40i36.40287. URL https://ojs.aaai.org/index.php/AAAI/article/view/40287.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 3882–3890. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/537. URL https://doi.org/10.24963/ijcai.2020/537. Main track.

Antonia Creswell and Murray Shanahan. Faithful reasoning using large language models, 2022. URL https://arxiv.org/abs/2208.14271.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410. URL https://aclanthology.org/2021.tacl-1.60/.

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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1229. URL https://aclanthology.org/2024.emnlp-main.1229/.

Yunseok Han, Yejoon Lee, and Jaeyoung Do. RFEval: Benchmarking reasoning faithfulness under counterfactual reasoning intervention in large reasoning models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.17053v3.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16867–16878. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.1466/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.503. URL https://aclanthology.org/2020.acl-main.503/.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.09664.

Zhaoyi Li, Gangwei Jiang, Hong Xie, Linqi Song, Defu Lian, and Ying Wei. Understanding and patching compositional reasoning in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9668–9688. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.576. URL https://aclanthology.org/2024.findings-acl.576/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Shengjie Ma, Chenlong Deng, Jiaxin Mao, Jiadeng Huang, Teng Wang, Junjie Wu, Changwang Zhang, and Jun Wang. Proof-of-use: Mitigating tool-call hacking in deep research agents, 2026. URL https://arxiv.org/abs/2510.10931v2. Version 2; original preprint posted in 2025.

William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.07923.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.559. URL https://aclanthology.org/2023.acl-long.559/.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.882. URL https://aclanthology.org/2024.findings-emnlp.882/.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in 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. 1256–1272. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0987b8b338d6c90bbedd8631bc499221-Paper.pdf.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, 2024. URL https://arxiv.org/abs/2412.15115.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. FaithCoT-Bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6c7154e394e24c69409256ccf8bf0804-Paper-Conference.pdf.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. 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. 31210–31227. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

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/.

Chen Sun, Renat Aksitov, Andrey Zhmoginov, Nolan Andrew Miller, Max Vladymyrov, Ulrich Rueckert, Been Kim, and Mark Sandler. How new data pollutes LLM knowledge and how to dilute it, 2024. URL https://openreview.net/forum?id=LA0ep4kKD0. Earlier manuscript version.

Chen Sun, Renat Aksitov, Andrey Zhmoginov, Nolan Andrew Miller, Max Vladymyrov, Ulrich Rueckert, Been Kim, and Mark Sandler. How new data permeates LLM knowledge and how to dilute it. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NGKQoaqLpo.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

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 *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Zhenyu Wu, Qingkai Zeng, Zhihan Zhang, Zhaoxuan Tan, Chao Shen, and Meng Jiang. Large language models can self-correct with key condition verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12846–12867. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.714. URL https://aclanthology.org/2024.emnlp-main.714/.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Hongbang Yuan, Pengfei Cao, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. Whispers that shake foundations: Analyzing and mitigating false premise hallucinations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2670–2683. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.155. URL https://aclanthology.org/2024.emnlp-main.155/.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14544–14556. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.968. URL https://aclanthology.org/2023.findings-emnlp.968/.

A NOTATION AND STUDY PROCEDURES

Symbol or term	Meaning
E, R, q	Supplied premises, allowed rules, and the fixed query.
$\mathcal{W}_R(E)$	Nonempty set of worlds compatible with premises and rules.
A/B/C	Query true in all / false in all / mixed across compatible worlds.
h, g	Relevant entailed implication / entailed but query-irrelevant addition.
$E^+, E_p, E^-, E^?$	Relevant explicit, paraphrased, changed-premise, and insufficient-evidence inputs.
$\mathrm{imp}, \mathrm{rel}, \mathrm{irr}$	Implicit, relevant-explicit, and irrelevant-explicit presentations.
J_s, T_s	Fractions of backgrounds correct in both determined states / both states and the shared unknown input. Tables sometimes report their numerators, with denominators stated.
Unknown accuracy	Correct C responses divided by all unknown inputs.
Unsupported assertion	A or B on a reference-C input; invalid output is separately retained.
New unsupported assertion	A/B on a reference-C input when the comparator does not assert A/B (five-view protocol); the eight-view protocol specifically counts comparator-C to A/B changes.
$\widetilde{P}, \lambda$	A/B/C/OTHER answer-position distribution and consistency coefficient.
J in the tangent appendix	Parameter Jacobian, distinct from behavioral joint correctness J_s.

Table 3: Notation and statistical units. The five-view entailment-consistency arm and the eight-view semantic-consistency arm belong to different studies; their names are not interchangeable.

Diagnostic procedure. For each background, the public-text evaluator verifies two determined premise states with opposite labels for the same query. It checks the relevant and irrelevant additions separately in each state. Removing a decisive link must admit two compatible completions with opposite query labels; these witnesses remain outside the model request. The seven diagnostic inputs are the two states crossed with three presentations, plus one shared unknown input. Standalone capability questions use a separate prompt and the fixed screening cohort. Freeze the selected configuration, new confirmation inputs, scoring, and hypothesis family before collecting confirmation answers. Generate each response independently, preserve invalid and missing outcomes, then compute background-level paired contrasts, J, T, and unknown risk. Surface-cue matching is a construction check, not a statistical proof of absence of lexical shortcuts.

Five-view adaptation procedure.

1. Generate disjoint training, development, and validation backgrounds. For each training background construct $E, E^+, E_p, E^-, E^?$ and verify each reference label. Original, explicit, and paraphrased views share a label; the changed and unknown views receive their own labels.
2. For each backbone, seed, and arm, initialize a separate LoRA adapter. Use the same seeded order and four backgrounds per update, with all five forward views and 512 updates per seed.
3. Average cross-entropy on the answer and end-of-turn targets over the supervised views. SFT adds no penalty; entailment and paraphrase arms add 0.5 JS between E, E^+ or E, E_p, respectively. Both paired outputs receive gradients. Changed and unknown predictions are not forced to agree with the original prediction.
4. Every 64 updates, evaluate original, changed, and unknown development inputs. Select the highest mean accuracy, breaking ties toward the earlier checkpoint. Bind selected adapter hashes before validation.
5. Evaluate implicit inputs by free generation without gold intermediate implications. Preserve all fixed seeds and endpoints, including regressions and new unsupported assertions. The joint No CF/U ablation removes two supervised views and all C targets; it is not a class-balanced factorial test.

The eight-view study uses a different loss and development protocol, including a label permutation for changed-premise consistency (Appendix P); it is not implemented by silently extending this five-view procedure.

B Interpretation and Extension Priorities

The seven-condition design identifies observable responses to evidence interventions. The following distinctions specify what further experiments would be needed for stronger causal or method-attribution claims.

From a qualified gap to an internal-use mechanism. Behavioral screening is an empirical screen, not observation of an internal representation during the downstream decision. The same-background relation mask reduces Qwen2.5's relevant–implicit gain to five backgrounds, with corrected $p = 0.0625$ (Appendix N). Explicitation can reduce derivation, memory, or interference demands; the present interventions do not separate these causes. Hidden-state interventions would be required to test a proposed internal mechanism. Asymptotic expressivity results (Merrill & Sabharwal, 2024) do not imply impossibility on our fixed short paths.

From matched templates to absence of surface shortcuts. Relevant and irrelevant additions match sentence template, relation type, and position, but can differ in names, query overlap, negation, and token length. Polarity is independently balanced across the cohort. Original and unknown inputs match record grammar, counts, and endpoint mentions, while semantic checks establish opposite-label completions. The post-hoc shallow-feature audit in Appendix C detects residual structural cues. Broader lexical audits and held-out template/depth transfer would test remaining shortcuts. The current synthetic splits primarily test new backgrounds within the specified generator.

From a repair to its specific source of improvement. All main training arms share five supervised views, so their contrast isolates the consistency term. A single-view SFT arm would be needed to attribute the gain to multi-view exposure itself. The joint No CF/U ablation removes all C targets as well as two semantic views; a class-balanced factorial ablation would separate these changes. Generated relation notes have no established advantage over deliberation or review (Appendix L); uniformly compute-matched baselines would be needed for an efficiency claim. Equal output caps do not equalize realized computation or reveal provider-side execution. FOLIO supplies an external accuracy test; its unchanged examples cannot supply the synthetic paired J/T endpoints.

C EviPair Dataset and Statistical-Family Registry

A named resource built from the recorded materials. EviPair (Paired Evidence-Response Evaluation) organizes existing, author-constructed synthetic questions into a versioned submission snapshot. It creates no new experimental observations. The core contains 2,904 backgrounds and 25,576 unique question texts, including training and development companions (Table 4). Later screen-qualified and API cohorts retain separate registries. FOLIO-derived interventions are adapted external material; unchanged FOLIO is a separate external benchmark. Neither is reclassified as an original natural-language dataset.

Model inputs and evaluator-only records. The export preserves each original input string. The input file stores an ID and text; only text enters a tested model. IDs, splits, conditions, gold labels, construction graphs and completion witnesses remain outside the prompt. Evaluator sidecars retain original IDs and source locations, while the manifest binds source and exported SHA-256 hashes. Nine-view companions contain the five training views, an irrelevant addition and three relation probes; they are not interchangeable with the seven-view crossed diagnostic.

The export re-evaluates all 25,576 public texts with the existing reference implementations; all agree with their stored labels and oracle records. There are no repeated input texts across the six core cohorts. The recorded relational marked-forest signatures are distinct across backgrounds (8,256 signatures); the ledger has 152 distinct canonical backgrounds. All 10,752 relational completion-witness texts are reconstructed and checked. The 304 ledger witness records retain opposite labels,

Component	Original split	Backgrounds	Views	Questions
Relational companion	Train	2,048	9	18,432
Relational companion	Development	64	9	576
Relational companion	Validation	512	9	4,608
Relational diagnostic	Independent	128	7	896
Ledger diagnostic	Development	24	7	168
Ledger diagnostic	Validation	128	7	896
Core total	Mixed roles	2,904	—	25,576

Table 4: EviPair core inventory. Training used five of the nine companion views. The total is not a count of independent test problems; model settings, training seeds and response repetitions do not increase these counts.

and ledger reference oracles are recomputed. These are construction and integrity checks, not an independent proof of the reference implementation or general task diversity.

Post-hoc shallow-feature audit. In a post-hoc diagnostic, we fitted two fixed shallow Gini trees on original relational training backgrounds and evaluated the untouched original validation split. Each background contributes implicit, changed-premise and unknown questions: 6,144 training and 1,536 validation questions, balanced over A/B/C. The maximum depth is four, minimum leaf size is 32, and each feature considers at most 32 training-derived thresholds. No hyperparameter is selected using validation answers. Length features are character and whitespace-word counts. The larger feature set adds line, punctuation, negation/missing-marker and record/entity counts, plus omitted-source positions parsed from public text. Identifiers, condition labels, raw names and oracle metadata are not features.

Predictor	Accuracy	Macro-F1	C recall	C precision
Training majority	33.33	0.167	0.00	0.00
Length only	32.94	0.246	16.41	32.94
Surface counts and missing position	38.87	0.318	70.12	39.58

Table 5: Exploratory surface audit on 512 validation backgrounds. Accuracy, C recall and C precision are percentages. Balanced accuracy equals accuracy because all three reference classes have 512 questions.

The structural predictor exceeds the 33.33% baseline, revealing residual shallow cues; high unknown recall has low precision. The length-only predictor does not exceed that baseline. These results do not establish shortcut-free data or explain the tested LLMs' responses. Full confusion matrices, family-specific readouts, fitted trees and fixed-feature code accompany the data card. This is a post-hoc diagnostic, not a newly preregistered hypothesis or a held-out-template test. Original data and LLM scores remain unchanged.

Statistical families and analysis status. Table 6 consolidates the reported inferential families. A prospectively frozen follow-up can be designed after an earlier pilot; that status does not make the entire research program preregistered. Conjunctions use the largest constituent one-sided paired p-value before Holm correction, as specified by the original protocol. Primary, auxiliary, exploratory and risk-bound families are not pooled retrospectively.

Original C2 risk bounds use 12 one-sided binomial intervals with $\alpha = 0.05/12$; the local crossed extension uses 18 with $\alpha = 0.05/18$. They are distinct simultaneous-coverage requirements, not extra Holm tests. The API-scale adverse-missingness sensitivity was added during collection and retains the original 44-slot family, without replacing the service-complete primary rule. Development selection, query-complement probes, transition summaries and the surface audit do not become primary confirmation through this registry. Study-specific correction does not control selection across the full sequence of research decisions.

Study / status	Family size	Scope and correction
Original C1 / original criteria	18 model slots	Conjunction of both control contrasts; Holm across fixed model slots.
Original five-view C2 / original criteria	4 comparisons	Two backbones by two baselines; three-seed conjunction, then Holm.
Qwen screened follow-up / post-pilot freeze	4 tests	Two selected configurations by two controls; exact paired tests and Holm.
Relational API grid / prospective	2; then 6 tests	Original two contrasts and separate three-configuration follow-up family; Holm within each.
Ledger / prospective	8 tests	Four configurations by two controls; Holm.
Local crossed extension / post-pilot freeze	10 hypotheses	Four base explicitation contrasts and six three-seed training conjunctions; Holm.
Grok screened follow-up / prospective	4 primary; 4 auxiliary	Two settings by two controls; primary and relation-mask families separate, each Holm.
Native settings / prospective, unequal compute	3 primary; 12 secondary	Three higher/lower comparisons; six settings by two explicitation controls separately; Holm.
API-scale / prospective	44 primary; 44 auxiliary	Eleven configurations by two strata by two controls; separate Holm families.
Natural interventions / exploratory	10; then 22 tests	First batch and expanded campaign retain separate Holm families.
FOLIO fixed external / descriptive	No new primary family	Fixed-checkpoint accuracy and per-seed differences; no claimed global confirmatory significance.
Surface audit / post-hoc exploratory	No hypothesis tests	Fixed train/validation diagnostic; no significance claim.

Table 6: Registry of the principal reported comparison families. The original named appendices retain exact eligibility, effect thresholds, failure status, auxiliary analyses and missingness rules. Family sizes are not additional independent samples.

D Study Provenance and Evidence Coverage

Question	Evaluation coverage	Comparison
Gap after screening	Qwen: 2,560 screening and 5,120 follow-up answers; Grok: 6,400 answers across two settings.	Relevant, implicit, and irrelevant joint correctness.
Task and setting scope	Relational and ledger grids; 29,341 completed answers in the 11-configuration API extension; 14 natural-language stories.	Paired hints, premise changes, and missing evidence.
Supervised adaptation	Two backbones; three seeds per adapted arm; 143,104 answers across 512- and 128-background cohorts.	Five-view SFT, entailment and paraphrase consistency, and No CF/U.
External transfer	All 204 FOLIO validation examples, ten fixed configurations, 2,040 answers.	Accuracy after synthetic-data checkpoint selection.

Table 7: Experimental coverage organized by the question each study answers. The natural intervention set is exploratory; FOLIO transfer uses unchanged benchmark inputs. Neither repeated views nor reused stories add independent samples.

Table 8 identifies which studies support each part of the argument. Prospectively frozen follow-ups were designed after earlier pilots but fixed before their new answers were collected. They are not independent preregistrations of the entire research program. An original prespecified test retains its original criteria, whether passed or not. A secondary readout can be calculated from a complete frozen study without becoming a prespecified primary endpoint. The headline qualified gap and local training counts use complete validation outputs and verified selected checkpoints. Incomplete API branches retain their full denominators and contribute only the explicitly marked ranges or descriptive comparisons.

Study	Evidence status	Role in the paper
Qwen screen-qualified follow-up, 256 backgrounds	Post-pilot, prospectively frozen; complete	Residual gap after an independent screen; frozen four-test family.
Grok follow-up, 256 backgrounds	Separately frozen follow-up; complete	Separately frozen ceiling replication; no positive gap.
Luna relational and ledger grids, 128 backgrounds each	Prospectively frozen; complete	Content-sensitive input effects in lower-reasoning settings; no internal-use identification.
Local validation, 512 backgrounds	Confirmatory comparison; complete	Consistency-versus-matched-training test; SFT-versus-base effects are secondary readouts.
Local crossed validation, 128 backgrounds	Prospectively frozen extension; complete	New-background SFT repair and consistency comparisons across three seeds.
FOLIO, all 204 validation examples	Fixed external evaluation; complete	Transfer of checkpoints selected using synthetic development data.
API reasoning comparison, 128 backgrounds	Frozen comparison; two service failures	Configuration sensitivity on reused backgrounds, with conservative missingness ranges.
Label confusion and query complement	Completed secondary / exploratory	Diagnose response patterns; do not redefine confirmatory tests.
Natural intervention set, 14 stories	Completed exploratory	Descriptive test on curated natural-language premises.
Gradient, learning-curve, and eight-view development studies	Completed exploratory	Analyze and select on development data; not independent validation.
11-model API extension	Collection ended; partial coverage	29,341 completed responses; qualified complete configurations are near ceiling.
Earlier interrupted API and local executions	Interrupted / excluded	Unverified answers and invalid adapters never supply headline measurements.
Earlier R2 unknown-input condition	Excluded for validity	Surface cue exposed the label; not used for evidence-boundary claims.

Table 8: Provenance and inferential role. Studies with different materials, settings, or execution epochs are not pooled into one validation cohort. Details and the original hypothesis history follow in their named appendices.

Study	Backgrounds	Configs.	Views	Records
Local original validation	512	26	9	119,808
Local crossed validation	128	26	7	23,296
Qwen capability screen	128	4	5	2,560
Qwen screened follow-up	256	2	10	5,120
Grok capability screen	128	2	5	1,280
Grok screened follow-up	256	2	10	5,120
API relational grid	128	4	7	3,584
API reasoning comparison	128	6	7	5,376
Natural-language interventions	14	13	7	1,274
Fixed FOLIO evaluation	204	10	1	2,040

Table 9: Selected evaluation cohorts. Backgrounds are per configuration; the reasoning comparison reuses the relational grid with fresh responses. Records include terminal service failures: five in the API grid, two in the reasoning comparison, and four in the natural-language study (three GLM Flash and one Qwen Max), leaving 1,270 completed natural-language answers. Screening and follow-up are separate stages; the two local validation rows reuse the same 26 configurations. Configurations include checkpoints and settings. These rows are not a total of independent questions. Appendix S separately reports the terminal 11-model extension: 29,341 completed responses within its fixed 35,200-question plan, with all missing outcomes retained.

Execution interruptions are summarized in Appendix T. Its retained update counts and missing captures distinguish physical execution from the logical training trajectory and the validation answers used above.

Configuration	Same A/B pair	C on determined pair	A/B on unknown
Luna, none	16	110	22
Qwen2.5, unadapted	78	50	79
Qwen3, unadapted	122	5	123

Table 10: Three observable failure patterns on the same 128 relational backgrounds under implicit presentation. The first two columns concern determined premise pairs; the last concerns the separate unknown input. Thus columns are not an exhaustive, disjoint partition of backgrounds. Configurations use their recorded native settings.

E Proofs of the Semantic and Distributional Statements

Conservative explicitation. Every world satisfying R and E^+ satisfies R and E, because $E \subseteq E^+$. Conversely, every world satisfying R and E satisfies the entailed statement h, and hence R and E^+. Equality of the compatible-world sets preserves the truth classification of the unchanged query. This elementary set-inclusion argument assumes classical entailment and the nonempty compatible-world set required in the main text.

Distributional motivation. Let P and Q be answer distributions for E and E^+ with the same correct label y^*. For $r(P) = 1 - P(y^*)$, the defining property of total variation gives

$$|r(P) - r(Q)| \le \mathrm{TV}(P, Q), \qquad r(P) \le r(Q) + \mathrm{TV}(P, Q). \tag{6}$$

Consistency can therefore transfer reliable explicit-task behavior, but it can also align two wrong distributions. Equation 6 supplies a design constraint, not a generalization guarantee or a theorem about greedy accuracy. Similarly, if changed evidence E^- has a different correct label and both inputs receive their respective correct labels with probabilities at least $1 - \epsilon_0$ and $1 - \epsilon_1$, then

$$\mathrm{TV}(P_E, P_{E^-}) \ge \max(0, 1 - \epsilon_0 - \epsilon_1). \tag{7}$$

Correct inference must distinguish semantic changes. The following paragraphs give the short proofs and the insufficient-evidence construction. These elementary bounds are not presented as independent mathematical innovations.

Joint correctness and marginal accuracy. If a_s, b_s are determined-state accuracies and u is unknown accuracy, the intersection bounds give

$$\max(0, a_s + b_s - 1) \le J_s \le \min(a_s, b_s), \qquad \max(0, J_s + u - 1) \le T_s \le \min(J_s, u).$$

Consequently $\max(0, a_s + b_s + u - 2) \le T_s \le \min(a_s, b_s, u)$. A marginal-to-joint accuracy drop can follow mechanically from the marginals; its size alone does not identify a distinct failure mechanism. The response partition and paired transitions provide the additional behavioral detail.

Risk transfer. For distributions on a common discrete answer space, $\mathrm{TV}(P, Q) = \sup_S |P(S) - Q(S)|$. Taking $S = \{y^*\}$ gives $|P(y^*) - Q(y^*)| \le \mathrm{TV}(P, Q)$. Substituting $r(P) = 1 - P(y^*)$ proves Equation 6. The statement concerns the specified distribution and scoring event, not unconstrained greedy completions under a coarsened first-token partition.

Separation after a semantic change. Let the original and changed inputs have distinct correct labels y_0, y_1. By assumption, $P(y_0) \ge 1 - \epsilon_0$ and $Q(y_0) \le 1 - Q(y_1) \le \epsilon_1$. Consequently $\mathrm{TV}(P, Q) \ge P(y_0) - Q(y_0) \ge 1 - \epsilon_0 - \epsilon_1$. Combining this with nonnegativity proves Equation 7.

Why the unknown cases are not false cases. The unknown construction omits the final necessary relationship of the queried path. The public rules allow that unrecorded relationship to point to either of two distinct endpoints. In the delegated permission family the endpoints have opposite terminal permissions. In the ancestry family one endpoint is the queried ancestor and the other is distinct. Both completions preserve the recorded relationships and acyclicity, but give opposite truth values to the same relation query. The conditional resource totals consequently take the two different allowed values. Since the claimed total is one of them, the decision query also has both truth values across compatible completions. The evaluator verifies the two completed public texts for every unknown case; it does not infer unknownness from a missing name or a dedicated sentence form.

F A CONDITIONAL LEARNING-DYNAMICS EXPLANATION

Logical equivalence does not force a parameterized predictor to respond identically to two strings. A possible source of persistent disagreement is that the parameter directions learned from the two inputs have different effects on their predictions. To make this explanation precise, we first separate an exact local statement from a simplified dynamical model.

Learning dynamics and regularization. Neural tangent kernels express output changes through parameter Jacobians (Jacot et al., 2018). Gradient Starvation analyzes how coupled feature-learning dynamics can suppress predictive features under cross-entropy training (Pezeshki et al., 2021). We do not identify that phenomenon with our API errors. Our narrower calculation concerns equivalent inputs with different parameter sensitivities. Its quadratic pair penalty is a graph-Laplacian regularizer, an established construction (Belkin et al., 2006); the calculation motivates a diagnostic and does not establish novelty of that regularizer.

Transfer depends on the parameter geometry. For two differentiable losses ℓ_i and ℓ_j, one small gradient step on ℓ_j changes the other loss by

$$\ell_i(\theta - \eta\nabla\ell_j) = \ell_i(\theta) - \eta\langle\nabla\ell_i, \nabla\ell_j\rangle + O(\eta^2). \tag{8}$$

The remainder assumes a locally bounded Hessian of ℓ_i. Sharing a correct label does not specify the sign or size of this inner product. Successful learning on the explicit input can therefore coexist with weak transfer to its implicit counterpart. Neither high explicit accuracy nor the behavioral gap alone measures this geometry.

An exactly solvable surrogate. Let $f = (f_r, f_e)^\top$ be two scalar outputs with the same target y. Linearize their dependence on the trainable parameters, hold their Jacobian J fixed, and write $K = JJ^\top$. Consider

$$\mathcal{Q}_\rho(f) = \tfrac{1}{2}\|f - y\mathbf{1}\|_2^2 + \tfrac{\rho}{2}(f_r - f_e)^2, \qquad L = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{9}$$

Under parameter gradient flow, the output error $e = f - y\mathbf{1}$ satisfies

$$\dot{e} = -K(I + \rho L)e. \tag{10}$$

Here ρ is a weight in the scalar quadratic model, not the JS weight λ in the implemented training objective.

Proposition 1 (Coupling unequal but learnable views)**.** *Assume $K = \mathrm{diag}(a, b)$ with $0 < a < b$ and $\rho > 0$. The slow exponential rate of Equation 10 is*

$$\mu_-(\rho) = \frac{(1+\rho)(a+b) - \sqrt{(1+\rho)^2(a+b)^2 - 4ab(1+2\rho)}}{2}, \tag{11}$$

and

$$a < \mu_-(\rho) < \frac{2ab}{a+b}, \qquad \lim_{\rho\to\infty} \mu_-(\rho) = \frac{2ab}{a+b}. \tag{12}$$

In the K^{-1} error norm, $\|K^{-1/2}e(t)\|_2 \le e^{-\mu_-(\rho)t}\|K^{-1/2}e(0)\|_2$.

With separate supervision and equal initial errors, the two errors are $e_r(0)e^{-at}$ and $e_e(0)e^{-bt}$. Thus a gap can arise without a difference in targets or supplied facts. The pair penalty improves the worst-mode convergence envelope under the proposition's assumptions. It does not guarantee that every coordinate improves at every finite time. If both views instead have rate a, their common-error mode still decays at a; consistency does not improve that mode. If the raw Jacobian is zero, the raw prediction cannot move in this fixed-feature model at all.

Connection to the implemented consistency term. We supervise an A/B/C token and an end-of-turn token. At the answer position, let $p_\theta(t \mid x)$ denote next-token probability. The four-category surrogate retains all non-answer mass:

$$\widetilde{P}_\theta(x) = \left(p_\theta(A \mid x), p_\theta(B \mid x), p_\theta(C \mid x), 1 - \sum_{a\in\{A,B,C\}} p_\theta(a \mid x) \right).$$

Two reasoning-prefixed outputs can agree in OTHER while their final decisions differ. This answer-first surrogate is therefore not a measure of final-answer agreement for verbalized reasoning. The original five-view study fixes $\lambda = 0.5$; the later eight-view study separately selects and fixes its coefficients on development data (Appendix P).

For $p = \mathrm{softmax}(z)$ and $F(p) = \mathrm{diag}(p) - pp^\top$, a local expansion gives

$$\mathrm{JS}(\mathrm{softmax}(z + \delta/2), \mathrm{softmax}(z - \delta/2)) = \tfrac{1}{8}\delta^\top F(p)\delta + O(\|\delta\|^4). \tag{13}$$

The same expansion applies to effective logits for our four output categories. Near agreement, JS therefore penalizes prediction differences in a probability-dependent metric. This provides a local connection to a pair penalty, not an equivalence between the implemented five-view CE–JS loss with AdamW and Equation 10.

The analysis yields falsifiable conditions: a useful pairing must connect views with a common target and a learnable discrepancy; equal slow modes, zero sensitivity, or an incorrect semantic pairing can defeat the argument. For example, pairing opposite scalar targets biases their equilibrium even when both are supervised. We therefore do not equate untransformed original and counterfactual distributions, or a determined distribution with an unknown case. The later semantic-change loss first permutes the labels of its counterfactual pair (Appendix P). Appendix G gives the proofs and an analytic illustration. Establishing this explanation in the trained LLMs still requires parameter-level measurements; the current behavioral results do not establish a training cause.

Consistency under nearly fitted labels. The matched SFT baseline already supervises both inputs of each consistency pair. This gives a further elementary constraint on the candidate objective.

Proposition 2 (A supervision bound on pair disagreement)**.** *For one five-view training background, let $\mathcal{L}_{\mathrm{ans}}$ be the mean of the five two-token answer losses defined in Section 5. If the implicit and explicit inputs share their canonical label, then, using natural logarithms,*

$$\mathrm{JS}\Big(\widetilde{P}_r, \widetilde{P}_e\Big) \le \min\{\log 2,\ 10(\log 2)\mathcal{L}_{\mathrm{ans}}\}. \tag{14}$$

Proof. For any two distributions P, Q, write $t = \mathrm{TV}(P, Q)$. Their common probability mass has weight $1 - t$, so they can be decomposed as $P = (1 - t)R + tP'$ and $Q = (1 - t)R + tQ'$, with disjoint residuals P', Q'. Joint convexity of JS, obtained from the log-sum inequality, gives $\mathrm{JS}(P, Q) \le t \log 2$; the cases $t = 0, 1$ follow directly. For common label y, the triangle inequality through the point mass δ_y then gives

$$\mathrm{TV}(P, Q) \le (1 - P(y)) + (1 - Q(y)) \le -\log P(y) - \log Q(y).$$

Writing ℓ_v^{seq} for the mean answer-token and end-token loss, nonnegative end-token losses give

$$-\log P(y) - \log Q(y) \le 2(\ell_r^{\mathrm{seq}} + \ell_e^{\mathrm{seq}}) \le 10\mathcal{L}_{\mathrm{ans}}.$$

Combining these inequalities gives Equation 14. □

This loose bound improves on the universal $\log 2$ ceiling only when $\mathcal{L}_{\mathrm{ans}} < 0.10$ nats. It applies per training background; averaging gives an average bound, not a per-example guarantee from a training window mean. Fitting shared labels forces agreement in the zero-loss limit, but the penalty can still alter finite-time optimization and generalization. Neither this bound nor small training disagreement certifies unseen accuracy or proves that the regularizer is redundant at every finite loss.

G Proofs and Boundaries of the Tangent-Model Analysis

Gradient-flow equation. For the affine predictor $f_\theta = f_{\theta_0} + J(\theta - \theta_0)$ and common target $y\mathbf{1}$, the output gradient of the quadratic objective is $(I + \rho L)e$, because $L\mathbf{1} = 0$. Parameter gradient flow gives $\dot\theta = -J^\top(I + \rho L)e$, hence $\dot e = J\dot\theta = -K(I + \rho L)e$. This is exact for the stated affine model.

Proof of Proposition 1. The generator is similar to the symmetric positive-definite matrix

$$S_\rho = K^{1/2}(I + \rho L)K^{1/2} = \begin{pmatrix} a(1+\rho) & -\rho\sqrt{ab} \\ -\rho\sqrt{ab} & b(1+\rho) \end{pmatrix}.$$

Its trace is $(1+\rho)(a+b)$ and its determinant is $ab(1+2\rho)$, giving Equation 11. The upper-left entry of $S_\rho - aI$ is $a\rho > 0$ and its determinant is $a\rho(b-a) > 0$. Sylvester's criterion therefore gives $S_\rho - aI \succ 0$ and $\mu_- > a$.

For $v = (a^{-1/2}, b^{-1/2})^\top$, the pair-penalty component has zero quadratic form, and the Rayleigh quotient of S_ρ evaluated at v is $2/(a^{-1}+b^{-1}) = 2ab/(a+b)$. This bounds μ_- above. Equality at finite ρ would require v to be a smallest-eigenvalue eigenvector. But $S_\rho v = Kv$, which is not proportional to v when $a \neq b$, so the bound is strict. Rationalizing the smaller root gives

$$\mu_-(\rho) = \frac{2ab(1+2\rho)}{(1+\rho)(a+b) + \sqrt{(1+\rho)^2(a+b)^2 - 4ab(1+2\rho)}}.$$

Dividing numerator and denominator by ρ proves the stated limit. Finally, $u = K^{-1/2}e$ obeys $\dot{u} = -S_\rho u$, so the spectral theorem gives the claimed norm bound. □

The general spectral statement. For any positive-definite fixed K and nonnegative ρ, $S_\rho = K + \rho K^{1/2} L K^{1/2}$ is increasing in the positive-semidefinite order. Its smallest eigenvalue is therefore nondecreasing. This need not be a strict improvement: a least-eigenvalue mode annihilated by $K^{1/2}LK^{1/2}$ can remain unchanged. A spectral bound also does not order individual errors at every time or establish generalization to unseen inputs.

For two correlated views, write $K = \left(\begin{smallmatrix} a & c \\ c & b \end{smallmatrix}\right) \succ 0$. The trace and determinant of S_ρ are $a + b + \rho(a + b - 2c)$ and $(ab - c^2)(1 + 2\rho)$. Rationalizing its smaller eigenvalue therefore gives

$$\lim_{\rho\to\infty} \mu_-(\rho) = \frac{2(ab-c^2)}{a+b-2c} = \frac{2}{\mathbf{1}^\top K^{-1}\mathbf{1}}. \tag{15}$$

The same value is its Rayleigh quotient at $K^{-1/2}\mathbf{1}$, hence an upper bound at every ρ. In the diagonal case this is less than $2a$, so the worst-mode acceleration over rate a is less than twofold; that factor does not hold for all correlated kernels. In particular, $\lambda_{\min}(K) \leq \min(a, b)$ even when $c > 0$, rather than necessarily exceeding a. The entries of K are output-Jacobian inner products; the loss-gradient measurements below do not estimate this kernel directly.

Failure cases. If $K = aI$, the common mode $\mathbf{1}$ is an eigenvector with eigenvalue a for every pair weight. If the raw Jacobian is zero, the corresponding row and column of K are zero, and the raw output is fixed under the linearized dynamics. Finally, consider $K = I$ and opposite scalar targets $y = (1, -1)^\top$. An erroneously imposed pair penalty minimizes $\frac{1}{2}\|f - y\|^2 + \frac{\rho}{2}(f_1 - f_2)^2$ at $f = (I + \rho L)^{-1}y = y/(1 + 2\rho)$. It introduces a nonzero target error even though zero supervised error is possible without the penalty.

Local JS expansion. Write $p_\pm = \mathrm{softmax}(z \pm \delta/2)$ and $p = \mathrm{softmax}(z)$, with finite logits. The softmax derivative is $F(p)$, so $p_\pm = p \pm F(p)\delta/2 + O(\|\delta\|^2)$. Their mixture is $m = p + O(\|\delta\|^2)$. The second-order categorical KL expansion gives

$$\mathrm{KL}(p_\pm \| m) = \tfrac{1}{8}\delta^\top F(p)\,\mathrm{diag}(p)^{-1}F(p)\delta + O(\|\delta\|^3).$$

Since $F(p)\,\mathrm{diag}(p)^{-1}F(p) = F(p)$, averaging the two KL terms gives the quadratic term in Equation 13. For fixed finite z, JS is smooth and unchanged under $\delta \mapsto -\delta$; its odd Taylor terms vanish, giving the fourth-order remainder stated there. The individual KL expansions need not be even. This expansion is local to agreement; it does not characterize large, confidently wrong discrepancies.

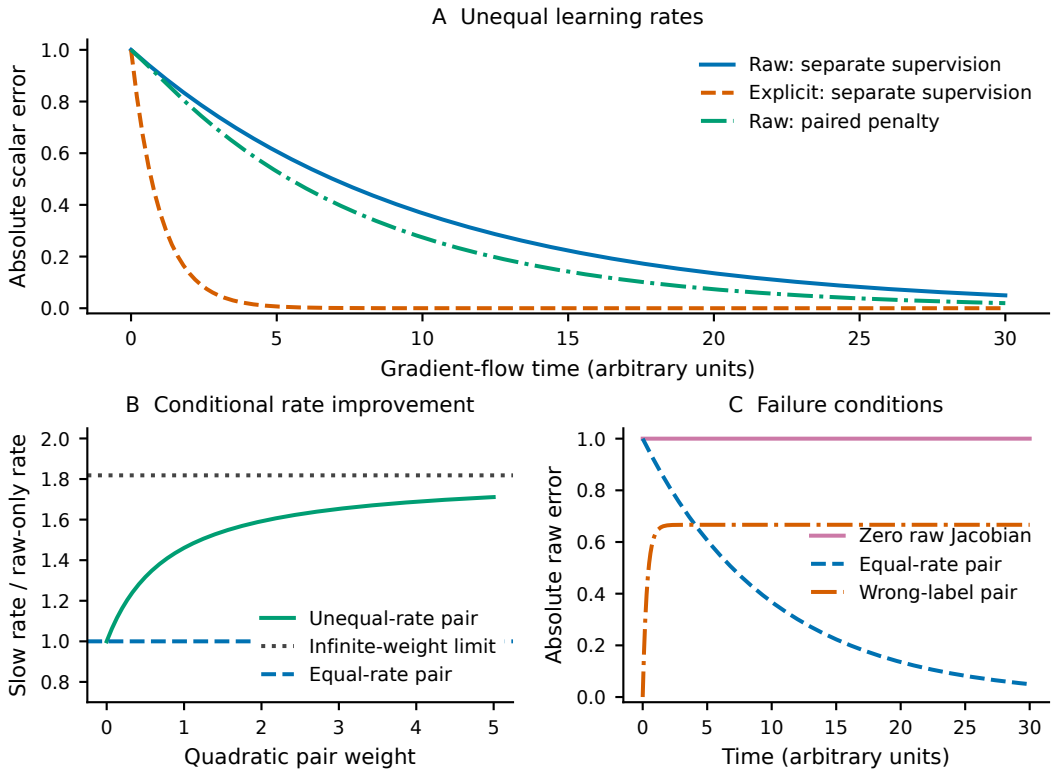


Figure 3: Analytic illustration of the conditional model, not language-model measurements. A: $a = 0.1$, $b = 1$, equal initial errors of one, and pair weight $\rho = 0.5$. B: the slow rate relative to separate raw supervision, with its limiting bound. C: separate counterexamples: a zero raw Jacobian retains its initial error; an equal-rate common mode is unchanged by the penalty; an incorrect pair with targets $(1, -1)$, $K = I$, $\rho = 1$, and initially correct outputs acquires error $2(1 - e^{-3t})/3$. Time is in arbitrary gradient-flow units. Curves are deterministic analytic solutions, so no sampling intervals are attached.

H Development Gradient and Finite-Step Readouts

Training arm	Explicit / implicit CE		Relation JS / unknown CE	
	Step 64	Selected	Step 64	Selected
SFT	10	6	51	22
Entailment	3	0	47	21
Paraphrase	13	1	64	15

Table 11: Negative gradient inner products out of 64 development backgrounds at each checkpoint, Qwen2.5 seed 17. No zero-norm pair is excluded. The first pair enters Equation 8; the second measures a component conflict, not the full update. Relation-JS gradients in the SFT and paraphrase arms are offline probes of fixed checkpoints, not losses active during those runs.

This post-hoc diagnostic uses the same 64 development backgrounds at six already selected Qwen2.5 checkpoints: the first scheduled checkpoint (step 64) and the originally selected checkpoint for SFT, entailment, and paraphrase, all at seed 17. Selection of checkpoints, all seven loss components, five update directions, three step sizes, and eight finite-step backgrounds was fixed before collecting the diagnostic. The eight backgrounds are four per family chosen by a fixed hash without conditioning on errors. No checkpoint file is updated and no validation answer is used.

We differentiate the five answer losses and the two JS terms with dropout disabled. Let g_d denote one of the SFT, entailment, paraphrase, explicit-only, or unknown-self directions. The common step size for a background and fraction f is $\eta = f \max(\|\theta\|, 1)/\max_d \|g_d\|$, with $f \in \{10^{-4}, 10^{-3}, 10^{-2}\}$. Every direction is evaluated at that same step size, restoring the parameters after each perturbation. A sign is called resolved only when both the predicted and observed changes exceed $10^{-5}(1 + |\ell_{\text{baseline}}|)$. Repeated baseline differences, own-objective changes, and all unresolved or contradictory outcomes are retained. The gradients concern raw loss directions, not a replay of AdamW with its momentum and preconditioner. BF16 forward computation also limits finite-step resolution. Low-precision rounding and higher-order terms can both contribute to unresolved or disagreeing signs. Without a matched higher-precision repeat or directional-curvature measurement, the observed step-size pattern does not identify either as the primary cause.

Checkpoint	10^{-4}	10^{-3}	10^{-2}
SFT, step 64	1/0/7	2/1/5	6/1/1
SFT, selected	0/0/8	0/0/8	4/1/3
Entailment, step 64	1/1/6	2/0/6	8/0/0
Entailment, selected	4/3/1	7/0/1	8/0/0
Paraphrase, step 64	0/0/8	0/0/8	0/1/7
Paraphrase, selected	1/2/5	5/1/2	7/1/0

Table 12: Explicit-only update: signs of predicted versus observed changes in implicit CE. Each entry is agreement / disagreement / unresolved, with fixed denominator eight. Larger steps improve resolution in some checkpoints but do not uniformly validate the first-order prediction.

The complete offline export replays 384 gradient records and 48 finite-step background records, including all five directions and all three fractions. It verifies native tokenization, Gram symmetry and positivity, direction coefficients and norms, probability-derived JS, finite-step predictions, residuals, and saved resolution flags. Direct independent reads of the raw Gram entries and observed changes reproduce Table 11 and the table above. The measurements support the existence of local gradient misalignment despite equal labels; they do not identify the cause of the API behavior or establish that adding consistency necessarily worsens unknown-case accuracy.

I Audited Examples of Evidence Response

We illustrate the lexicographically smallest independent background identifier in each family, retaining all seven views. This mechanical selection was made after analysis and is illustrative, not a prespecified statistical endpoint. All fourteen saved responses were replayed against the original requests, raw response bytes, logical references, and scoring rules. The machine-readable export at `artifacts/entailment-paper-examples-v1/examples.json` contains the complete system and user inputs, visible answers, proofs, and source hashes. The paths below summarize relevant records for explanation; the evaluated prompts also contain the other records, names, and public rules.

In the ancestry example, the original path is Jade $\rightarrow$ Rowan $\rightarrow$ Cedar. Heath and Cedar are distinct, so the condition for adding six units to the 18-unit pool is false. The relevant addition states that Heath is not Jade's two-step maternal ancestor. Changing Rowan's mother to Heath makes the condition true and the claimed 18-unit total false; the relevant addition now states the positive relationship. The irrelevant addition instead concerns Heath's relationship to Moss. In the missing-evidence view, Rowan's mother is unspecified, allowing totals of either 18 or 24.

In the delegated permission example, the original path is Clover $\rightarrow$ Teal $\rightarrow$ Una, where Una has terminal status DENIED. The pool starts at 13 units and gains 12 exactly when Clover has permission, so the claimed 13-unit total is true. The relevant addition states that Clover has no permission. Redirecting Teal to Onyx, whose terminal status is GRANTED, changes the total to 25; the relevant addition states that Clover has permission. The irrelevant addition states that Zephyr has no permission. The missing-evidence view leaves Teal unspecified, so both totals are possible. In both families, filler records balance the missing-evidence construction; the full exported inputs retain those changes.

View	Reference	Maternal ancestry	Delegated permission
Original, implicit	A	B	B
Original, relevant	A	A	A
Original, irrelevant	A	C	A
Changed, implicit	B	C	C
Changed, relevant	B	B	B
Changed, irrelevant	B	C	A
Missing evidence	C	C	A

Table 13: Exact visible Luna answers for the first independent background in each family. A denotes entailed, B contradicted, and C undetermined. Relevant explicitation resolves both determined questions in each example. The unmodified missing-evidence delegated permission question still elicits an unsupported assertion.

Condition	Measurement role
Original	Complete downstream decision from implied relationships
Relevant explicitation	Effect of supplying a verified implication
Irrelevant implication	Added-text and redundancy control
Paraphrase	Sensitivity to meaning-preserving wording changes
Counterfactual	Correct answer changes after a premise changes
Insufficient evidence	Preserve uncertainty when a necessary link is unspecified
Three relation questions	Standalone inference on original, changed, and insufficient evidence

Table 14: Nine conditions per background. The three relation questions include supported, refuted, and unknown labels; an always-supported answer cannot pass the capability check.

J Inference and Training Settings

Table 15 records the original evaluation controls, as verified from saved requests for the API configurations and the frozen local decoder. The names identify requested endpoints, not independently authenticated hidden checkpoints. For Flash the requested and observed response identifier in the audited exemplar is `deepseek-flash`; the internal inventory key is `deepseek-v4-flash`. An omitted field is not evidence for a particular server default. In particular, the Sol CLI requests contain no explicit output cap, even though its separate HTTP registry entry lists a budget. The CLI envelope and reasoning context are audited separately from the HTTP clients.

Requested configuration	Reasoning	Output cap	Sampling
`gpt-5.6-luna`	effort none	16,384	not submitted
`deepseek-flash`	thinking disabled	4,096	temperature 0
`deepseek-v4-pro`	thinking disabled	4,096	temperature 0
`gpt-5.6-sol` (CLI)	effort low	not submitted	not submitted
`glm-5.3`	not submitted	4,096	temperature 0
`gemini-3.1-pro-preview`	not submitted	4,096	temperature 0
`qwen3.8-max`	not submitted	4,096	temperature 0
Qwen2.5-7B-Instruct	standard template	64	greedy
Qwen3-8B	thinking disabled	64	greedy

Table 15: Submitted inference controls for the original grid and earlier branches. The separate reasoning comparison in Table 19 sets 16,384 tokens for both settings of Luna, Flash, and Pro. Caps are in tokens: `max_completion_tokens` for Luna, `max_tokens` for the other HTTP configurations, and `max_new_tokens` locally. Sol uses reasoning context `all_turns`. These are native configurations with unequal budgets.

All original-grid task requests instruct the model to return one letter and provide no tools. The common system input defines the total as the starting amount plus its one conditional contribution, and includes three task-independent label examples. HTTP calls provide that system input and the complete task as two messages. Local inputs apply each model's chat template; Qwen3 explicitly disables thinking. Sol uses the captured CLI envelope with the same task instructions. Service-

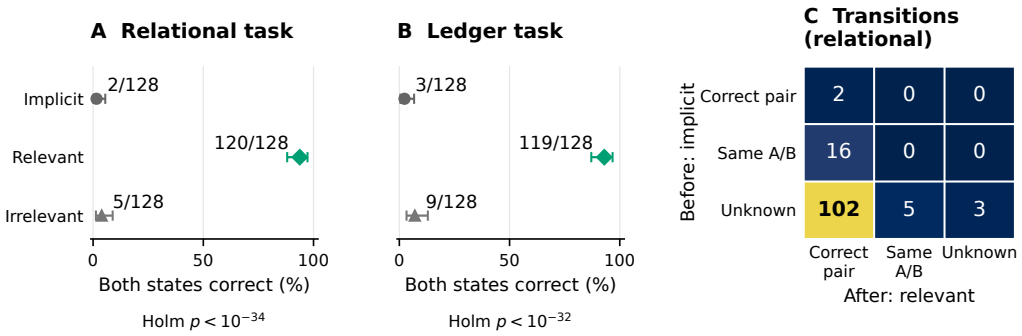


Figure 4: Detailed Luna response changes on two independent constructions. Panels A/B use exact 95% intervals. Panel C counts relational transitions from implicit to relevant presentation: 102 of 110 pairs containing an unknown answer on determined evidence become correct at both ends. The separate unknown inputs are untreated.

API configuration	J_{imp}	J_{rel}	J_{irr}	Unknown
Luna, none	2	120	5	106
Gemini 3.1 Pro Preview	128	128	128	128
API Qwen3.8 Max	128	128	128	128
GLM-5.3	125–128	128	127–128	126–127

Table 16: Independent seven-condition grid on the same 128 new backgrounds. The first three columns count correct answers at both premise states. Ranges identify possible counts under service missingness, not confidence intervals; GLM also has one truncated unknown answer counted as wrong. No explicitation intervention is applied to unknown cases. Native settings differ, so this is not an equal-compute model ranking.

reported usage is retained as telemetry, not equated with visible prompt length or matched inference effort. Full request and record hashes for the illustrated configurations accompany the example export.

Additional API follow-up. After the Luna result and local development comparisons, we froze a follow-up for `glm-5.3`, `gemini-3.1-pro-preview`, and `qwen3.8-max`. Each receives all 896 unchanged inputs from the same 128-background grid; there is no item selection by earlier model errors or reuse of model answers. The two both-state contrasts per configuration form one six-test Holm family. This follow-up is prospective for those configurations, not registration before the Luna study. Shared inputs do not increase the independent sample size. The requested Qwen ID is `qwen3.8-max`, and the returned ID is `qwen3.8-2.4t-a95b`. These endpoint IDs do not authenticate hidden weights.

The initial Gemini and API Qwen branches each close with 896 strictly audited, scorable responses and no missing outcomes. Gemini uses 896 physical science requests; Qwen uses 904, including eight recorded transport-error attempts under the frozen retry rule. Each has one separate health request. Both configurations are at ceiling in all seven conditions, so this construction provides no observed explicitation benefit for either. Zero observed unsupported assertions in 128 cases does not establish the original 2-percentage-point risk bound.

GLM's initial branch stops after 32 requests: 29 completed answers and three literal gateway error responses. Its remaining 864 inputs and the three failed services are missing in the initial analysis; all six tests remain in the family, with worst-case completion of missing outcomes. A separately recorded continuation dispatches only the 864 never-started cases with one worker, preserving every original answer and failure. Continuing past isolated service failures until three consecutive failures is an explicit execution-policy departure from the original first-failure stop. It closes all 864 dispatches with 862 completed services and two further failures. The combined result therefore retains 891 completed services and five missing outcomes among the original 896 inputs: four literal

gateway errors and one final transport failure. Its 904 physical science requests include eight extra bounded transport attempts; two health requests are separate. The initial analysis remains available alongside the combined analysis.

All 764 observed determined answers are correct. Both-state correctness is 125/128 with three missing outcomes for implicit inputs, 128/128 for relevant explicitation, and 127/128 with one missing outcome for irrelevant explicitation. Both contrasts have zero observed gains and losses and Holm-adjusted $p = 1$. Missing-outcome effect ranges are $[0, 2.34]$ and $[0, 0.78]$ percentage points; their conservative 95% paired intervals are $[-3.37, 7.41]$ and $[-3.37, 4.88]$, respectively. The apparent count differences cannot be treated as observed repairs. On unknown inputs, 126/128 answers are correct, one service is missing, and one response exhausts the 4,096-token cap with no visible answer. That truncation is a completed task failure, not service missingness. No definite assertion is observed, but the unresolved unsupported-assertion count remains in $[0, 2]$. The six-test family is unchanged.

Local training configuration. Table 17 specifies the five-view adaptation branch in Figure 2. LoRA adapts every decoder block's query, key, value, and output attention projections and the MLP gate, up, and down projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`). The backbone and biases are frozen. Stored LoRA tensors use FP32, while backbone loading and forward autocasting use BF16; the loss and four-category log probabilities are evaluated in FP32. Attention uses PyTorch SDPA, and gradient checkpointing is non-reentrant. The recorded scheduling amendment disables autograd multithreading; its scope and failures are described in Appendix T.

Setting	Value
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, zero weight decay
Learning rate	Peak 10^{-4}; 16-update linear warmup, then cosine schedule ending at zero
Gradient clipping	Global norm 1.0
LoRA	Rank 16, $\alpha = 32$, dropout 0.05, no bias adaptation
Consistency weight	0.5; both paired outputs receive gradients
Update unit	Four backgrounds, five separate forward views each
Training exposure	2,048 backgrounds; 512 updates; one epoch per seed
Seeds	17, 29, 43, with the same seeded order across arms
Checkpoint selection	Every 64 updates; mean original/changed/unknown development accuracy; earliest tie
Maximum sequence length	1,024 tokens; overlength input causes an error, not truncation
Canonical token IDs	A/B/C: 32/33/34; end-of-turn target: 151645

Table 17: Frozen settings for the original five-view study. The SFT, paraphrase, and entailment arms share answer supervision on all five views. The ablation removes changed-premise and unknown supervision while retaining the entailment penalty and all five forwards.

Local generation retains repetition penalties of 1.05 for Qwen2.5 and 1.0 for Qwen3, for both base and adapted configurations. Greedy token selection uses the scores after this processing.

The saved adapter headers contain 40,370,176 trainable parameters for Qwen2.5 and 43,646,976 for Qwen3. One epoch processes 5,584,162 and 5,625,122 forward input tokens respectively, excluding development generation and replayed work. Maximum observed forward lengths are 650 and 654 tokens. Each input contains the prompt and canonical answer; its two supervised positions predict that answer and the end-of-turn token. Examples are processed separately without padding or sequence packing.

The documented environment uses Python 3.11.14, PyTorch 2.6.0+cu124, Transformers 4.57.3, PEFT 0.14.0, Accelerate 1.12.0, and Safetensors 0.7.0. The two model queues each occupy one NVIDIA GeForce RTX 4090 D with 24 GB of memory. The settings and counts above are extracted from the frozen implementation, saved tokenization, and adapter records. Failed or replayed work is retained separately from the nominal one-epoch exposure.

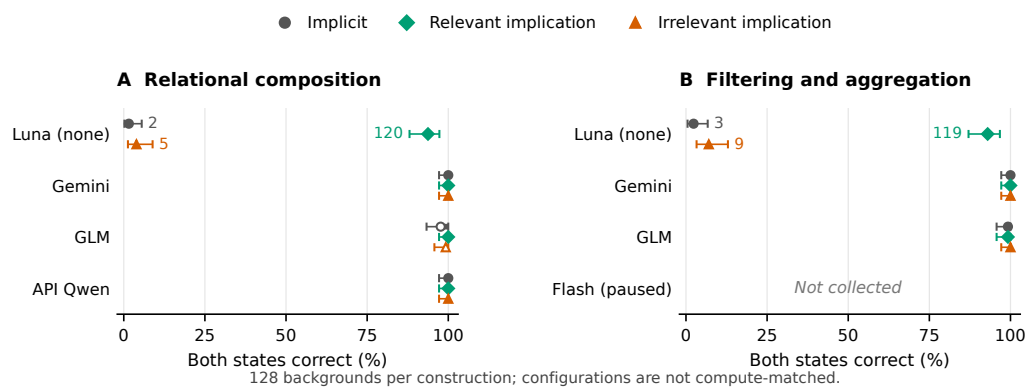


Figure 5: Independent input-intervention studies, 128 backgrounds per construction. Markers show both-state correctness; whiskers are exact 95% intervals conservatively extended for service missingness. Thick segments mark missing-outcome identification ranges. Luna labels are counts out of 128. Flash's paused ledger cohort remains unobserved.

Offline reproduction of the closed API results. The accompanying reproduction snapshot contains all 5,504 distinct public inputs used by the independent Luna grid and the three corrected API branches. It preserves 14,702 audited responses and 18 missing outcomes out of 14,720 planned responses; shared R3 inputs do not create additional independent backgrounds. The script `reproduce.py` recomputes logical reference labels from the public texts, verifies both completions of unknown cases, checks actual requests, decodes the saved response bytes, and recalculates the reported statistics using the unchanged scorer and statistical routines. It requires Python and SciPy but no GPU, model weights, credentials, or network connection. Incomplete captures and their excluded derived records remain available and are never promoted to scored answers. A separate follow-up archive adds all 2,688 records from GLM, Gemini, and API Qwen on the same 896 grid inputs, including five failed services and one truncated response. It reproduces both the initial and combined six-test analyses, preserves the reviewed runtime departure, and audits four health requests in this archive (one Gemini, one API Qwen, and two GLM, including its reviewed continuation). Both archives run offline; neither substitutes for local validation or establishes the original C1/C2 claims.

Offline local-validation replay. The full local archive contains all 119,808 responses across 26 configurations of Qwen2.5 and Qwen3. The earlier Qwen2.5-only snapshot remains available. The replay checks public reference labels, native input tokenization, output decoding, scoring, and paired statistics offline. Native tokenizers are included; model weights and retraining are outside this replay. The later local crossed study is not included in this archive. A recorded Qwen3 continuation preserves an initial memory failure before any question attempt and binds successful continuation launch/exit receipts; no answer is repeated or changed.

K Input-Intervention Scope Across Constructions

Larger deficits in a lower-reasoning configuration. Luna at effort none reaches 2/128, 120/128, and 5/128 both-state correctness with implicit, relevant, and irrelevant presentations, respectively (Figure 5A). Relevant explicitation repairs 118 pairs without regressions: +92.19 points, conservative 95% paired interval $[81.80, 96.59]$. Both control comparisons pass Holm correction (Appendix Table 16). Of 110 implicit pairs containing an unknown answer despite determined evidence, 102 become correct at both states; irrelevant addition leaves 109 such pairs. The intervention mainly repairs unwarranted uncertainty while tracking changed premises. The shared unknown condition still has 22 unsupported assertions, leaving 98/128 correct in all three states. This configuration fails relation qualification, so the large oracle-assisted gain characterizes the failure pattern rather than the residual effect established above.

Luna's benefit extends to filtering and aggregation. The independent ledger cohort yields implicit/relevant/irrelevant counts of 3/119/9 (Figure 5B). The 116 repairs without regressions give +90.63 points over implicit input, 95% paired interval $[79.86, 95.53]$; both contrasts pass the eight-test Holm family. The shared unknown condition has 24 unsupported assertions and 96/128 backgrounds pass all three states. The same diagnostic pattern therefore extends from graph composition to filtering and aggregation.

Other API configurations reach ceiling. Gemini 3.1 Pro Preview and API Qwen3.8 Max (`qwen3.8-max`) each answer all 896 relational questions correctly: both contrasts have zero gains or losses, Holm $p = 1$, and 95% paired intervals $[-3.37, 3.37]$ points. GLM retains five service failures and one truncated unknown answer; its other 890 answers are correct (Appendix J). On the independent ledger, Gemini is again perfect, while GLM scores 893/896 with $J_{\text{imp}}/J_{\text{rel}}/J_{\text{irr}} = 127/127/128$ and 127/128 unknown answers. Large explicitation gains are therefore configuration-dependent on both constructions.

L API Reasoning Controls and Development Follow-ups

A small strategy and cost comparison. We select 16 backgrounds per task family by fixed hash order from the earlier 128-background relational grid and collect fresh answers to the original, changed-premise, and unknown questions (96 inputs). Luna uses direct none, native medium, or none with short reasoning and plurality voting. Each sample uses temperature 0.7 and must end with a unique `Final: A/B/C` line; ties favor the earliest valid sample. Medium's newly reported completion-token count sets that input's sampling budget. Each path requests at most 256 tokens and at most the remaining budget, with a maximum of 16 paths. Invalid paths still consume their reported tokens. Sampling stops independently of correctness.

Strategy	J/32	T/32	Unsupported/32	Calls	Input tokens	Output tokens
Direct none	6	1	21	96	51,697	480
Native medium	25	25	0	96	51,697	5,255
Short-reasoning vote	32	30	2	141	76,157	7,788

Table 18: Fixed 32-background Luna strategy comparison. J requires both determined answers; T additionally requires the unknown answer. Counts and costs include every planned scientific input and generated path; three separate interface checks are excluded. These are fresh responses on previously studied materials, not an independent new task sample.

All 333 scientific responses are service-complete, without retries. Of 141 sampled paths, 139 meet the output format; every input has at least one valid path. The prespecified primary comparison, vote versus medium on T, has six repairs and one regression: $+15.63$ points, conservative paired 95% interval $[-12.17, 38.90]$, two-sided exact McNemar $p = 0.125$. It does not establish superiority. Medium makes no unsupported unknown assertion; vote makes two. Summed request latencies are 235.66, 375.25, and 509.07 seconds for direct, medium, and all voting paths, respectively; they are request-time sums, not parallel wall-clock runtimes.

Actual token matching fails: 94 of 141 paths report more completion tokens than requested. Only 17/96 inputs fall within the predetermined 90–110% medium-token range, and no background matches on all three inputs. Voting uses 1.482 times the medium completion tokens in aggregate and also rereads more input. We retain all inputs and report a cost–performance comparison, without attributing the difference to a compute-matched reasoning mechanism.

L.1 Native reasoning configuration comparison

We next fix a within-model comparison on the same 128 backgrounds, collecting 896 fresh answers per setting. Public messages and the 16,384-token output cap are identical within each model; only the native reasoning fields change. Luna compares none with medium; DeepSeek compares disabled with enabled/high. The primary endpoint is implicit both-state correctness, with one paired exact test per model and Holm correction over all three. These are new answers on previously studied materials, with unequal actual inference computation.

API model	Reasoning	J_{imp}	J_{rel}	J_{irr}	Unknown
Luna	none	1	118	4	114
	medium	118–119	127–128	127	126
DeepSeek Flash	disabled	19	64	7	13
	enabled/high	128	128	128	127
DeepSeek Pro	disabled	0	9	0	67
	enabled/high	128	128	128	128

Table 19: Within-model reasoning comparison, $n = 128$ backgrounds. J_{imp}, J_{rel}, and J_{irr} count both-state correctness for implicit, relevant, and irrelevant presentations. Unknown counts correct C answers. Luna medium has two service-missing responses; ranges retain their unknown outcomes. All answers are fresh, including Luna none; this is a repeat on the same backgrounds, so its counts need not equal the earlier grid's counts.

Table 19 shows improvements of 91.41–92.19 percentage points for Luna, 85.16 for Flash, and 100 for Pro. Conservative 95% paired intervals are $[80.82, 96.59]$, $[73.34, 91.47]$, and $[93.27, 100]$; all three Holm-adjusted p values are below 2×10^{-33}. Unsupported assertions on unknown inputs fall from 14 to 2, 115 to 1, and 61 to 0, respectively. Relevant explicitation beats both controls in every lower setting in the separate twelve-test family. Under higher settings, neither DeepSeek model has remaining determined errors, and neither Luna contrast passes correction. This demonstrates a configuration effect. Reported output-token totals rise about 326-fold for Flash and 406-fold for Pro; no compute-matched sampling comparison separates token expenditure from reasoning structure. These results do not establish that a visible rationale causally determines the answer. The following tables give token use and latency; subsequent sections retain DeepSeek validation, and exploratory automatic-repair and non-graph studies.

A 32-background budget follow-up is reported in Appendix L; provider overruns prevented the intended realized-token match.

Protocol and complete accounting. The within-model comparison uses the complete earlier seven-condition material, with three models and two settings. Each setting contributes 896 new logical outcomes; none reuses an earlier model answer. The protocol, inputs, configuration fields, scoring, and two multiplicity families were bound before dispatch. Setting order was fixed within the common case order. The primary family tests higher versus lower implicit both-state correctness for three models. The secondary family tests relevant explicitation versus each of the two controls in all six configurations. It contains twelve hypotheses, including the ceiling configurations. The 128 backgrounds, not 5,376 responses, define the independent sample size. This follow-up was specified after the original grid findings and is not a new blind task sample.

All 5,376 logical outcomes are recorded. Luna medium retains two final service failures after bounded transport retries; the other 5,374 outcomes are service-complete. The primary intervals and tests retain missing outcomes under adverse completion. All 2,688 request pairs were checked to contain identical public inputs and request bodies after removing only the intended native reasoning fields. Submitted configuration fields do not reveal the provider's hidden execution. This separate reasoning experiment has six health requests, one per model–setting pair; they are distinct from the four in the API follow-up archive. They are excluded from the scientific denominators and usage in Table 20.

Automatic repair is not uniformly better than deliberation. Answer-conditioned verification has prior precedent: ProCo recovers a masked condition to check an answer (Wu et al., 2024). Our relation-note procedure is a distinct development control and does not implement that algorithm. A separate exploratory comparison uses 24 existing development backgrounds, selected by family, depth, and truth before collection. Four public-input-only methods are tested: direct answering, one-call visible deliberation, two-call generic review, and a generated relation note followed by a decision. The two staged methods share 512-token limits per stage; direct and deliberation calls use 1,024. Only the original public problem and the generated note enter the second-stage prompt. Gold labels, private graphs, reference deductions, and oracle-selected paths are excluded. This matches two-call output limits, not realized compute or answer format across all four methods. Exact mes-

Model	Setting	Output tokens	Mean seconds	Physical requests
Luna	none	4,486	2.26	896
Luna	medium	51,636*	4.47	900
DeepSeek Flash	disabled	896	4.77	896
DeepSeek Flash	enabled/high	291,768	6.40	896
DeepSeek Pro	disabled	896	5.25	896
DeepSeek Pro	enabled/high	363,433	11.17	896

*Luna medium has usage for 894 attempts and lacks usage for six. The total excludes unobserved usage rather than imputing zero.

Table 20: Returned completion-token totals and mean logical-call elapsed time for the reasoning comparison. Tokens are service-reported usage, not FLOPs or equivalent computation. Latency includes logical-call retries and local transport overhead; it is not isolated model inference time.

sage templates, including both stages, appear in Appendix X. These controls do not implement multi-path self-consistency, an executable deductive verifier, or the counterfactual preference training of FRODO (Paul et al., 2024); superiority to those alternatives is untested.

	Luna, none		Flash, disabled	
Method	Both ends	All three	Both ends	All three
Direct	0	0	3	1
Visible deliberation	24	24	19	10
Generic review	17	15	24	23
Relation note	17	17	23	21

Table 21: Automatic-repair development counts out of 24 backgrounds. Each model has 336 recorded outcomes and 480 physical scientific requests, including the separate 48 oracle-implication comparisons. No method dominates both deliberation and generic review across these two configurations.

DeepSeek Pro's queued development collection was cancelled before any science or health request under the user's API pause; its unobserved outcomes remain in the original planned family. No new validation outcomes are reported. A Luna development check reuses identical generated notes but makes fresh final calls with the note before or after the final question. Relation-note both-state scores are 19 versus 18; generic-review scores are 15 versus 14. Both contrasts have one-sided exact $p = 0.5$ (unadjusted). This check does not establish note placement as an explanation of the repair failures.

L.2 Local inference and matched adaptation controls

We evaluate all 24 development backgrounds selected for automatic repair, without selecting errors. Each configuration answers the original, changed-premise, and unknown questions, giving 72 fresh generations. The five unadapted configurations total 360 answers. Inputs are preserved up to a 2,048-token admission limit; every configuration permits 4,096 output tokens. Qwen2.5 uses greedy decoding. All Qwen3 configurations share temperature 0.6, top-p 0.95, top-k 20, and min-p 0, with a fixed model/case-specific random seed. These matched Qwen3 settings follow the native-thinking sampling recommendation, not an optimization of each non-thinking baseline. Native thinking changes only the chat-template switch; visible deliberation requests a final line `FINAL: A/B/C`.

Only tokens after the closing `</think>` can supply a native final answer. Unclosed thinking and nonterminal generation are failures; answer-like text inside a reasoning block is never scored. Full tokens, decoded text, termination, and final-answer parsing are replayed independently. Table 22 retains all planned backgrounds.

The three exploratory both-state comparisons against direct answering use one Holm family. Adjusted p values are 4.88×10^{-4} for Qwen2.5 deliberation, 1.79×10^{-7} for Qwen3 deliberation, and 2.38×10^{-7} for Qwen3 native thinking. These are development comparisons on one fixed sampling

Model	Setting	Original	Changed	Unknown	Both	All three	Unsupported
Qwen2.5	Direct	7	6	11	0	0	13
	Visible deliberation	14	14	3	11	2	16
	SFT, seed 17	21	24	23	21	20	1
	SFT, seed 29	24	24	23	24	23	1
	SFT, seed 43	22	24	22	22	22	2
Qwen3	Direct	9	13	3	0	0	21
	Visible deliberation	24	24	11	24	11	11
	Native thinking	24	23	20	23	19	3
	SFT, seed 17	24	24	24	24	24	0
	SFT, seed 29	24	24	24	24	24	0
	SFT, seed 43	24	24	24	24	24	0

Table 22: Local development controls, 24 backgrounds per row. "Unknown" counts correct C answers; the last column counts unsupported definite answers on the same unknown inputs. SFT uses all three existing adapters with the corresponding direct-input tokens, generation settings, and random seeds. No new training or validation-based selection occurs.

stream. Qwen2.5 deliberation introduces seven new unsupported assertions; Qwen3 deliberation introduces one, while native thinking introduces none on these backgrounds.

Qwen2.5 deliberation generates 19,042 tokens, compared with 144 for direct answers, and includes one truncation. Strict parsing retains 14 invalid `Final:` prefixes. A separately reported post-hoc case-insensitive prefix analysis raises both-state correctness to 14, leaves all-three at 2, and counts 20 unsupported assertions. Four formerly invalid unknown outputs are definite assertions, so formatting does not explain the uncertainty regression. Original scores remain unchanged. Qwen3 deliberation generates 22,657 tokens and has two invalid unknown final answers; native thinking generates 109,106 tokens with two unclosed 4,096-token truncations. Its cap therefore does not measure unconstrained native capability.

All three previously selected SFT adapters per model are then evaluated under the matching direct-input and generation settings, adding 432 development answers without training. Each adapter generates 144 tokens. All six input/output and adapter-binding audits pass. SFT retains substantial three-state improvement under these controls, but the inference results show that adaptation is not the only way to repair an unadapted model. The original 512-background and later 128-background validations remain the independent evidence for training generalization.

A non-graph development task shows a similar configuration boundary. A natural-language ledger requires filtering approved entries by project and date, summing their quantities, and testing a threshold that controls the final total. All seven views preserve one unspecified quantity: it belongs to an ineligible entry in determined cases and to a decisive eligible entry in unknown cases. Opposite-label completions and an independent public-text parser check this distinction. No parser output is supplied as a model tool. On 24 development backgrounds, implicit/relevant/irrelevant both-state counts are 0/24/1 for Luna, 0/12/0 for Flash, 24/24/24 for GLM, and $[23, 24]/24/24$ for Gemini. Correct unknown answers are 20, 19, 24, and 24, respectively. Gemini's range reflects one service-missing answer, not a confidence interval; GLM completes all 168 questions correctly. These are exploratory synthetic-ledger findings. The separately admitted validation cohorts appear in Appendix L.3; neither stage measures performance on natural business records.

Earlier validation shows different response patterns. The corrected DeepSeek validation branches each retain 4,604 strictly audited responses and four unknown in-flight outcomes among 4,608 planned questions. Only never-attempted questions were collected after the interrupted processes had ended and the existing raw records were audited. Table 23 reports the 512-background denominators. Relevant explicitation improves Flash's original accuracy by 23.83 percentage points (122 gains, no losses; 95% paired interval $[18.86, 28.33]$). Pro's benefit lies in $[11.91, 12.11]$ points under missing outcomes (67 observed gains, six losses; conservative interval $[7.17, 16.64]$). The respective benefits over irrelevant implications are 28.52 and $[15.43, 15.63]$ points. These branch contrasts have smaller magnitudes than Luna's; they do not establish the original C1 claim.

Configuration	Original	Explicit	Irrelevant	Both ends	Unknown
DeepSeek Flash	262	384	238	68	69–71
DeepSeek Pro	140–141	202	122–123	0	256

Table 23: Correct outcomes out of 512 corrected validation backgrounds. Ranges bound the number correct under missing responses; they are not confidence intervals. "Both ends" is a post-hoc joint analysis of the original and changed-premise views, not the separately frozen semantic grid.

A post-hoc response analysis finds 361 unchanged definite pairs for Flash, 69 containing unknown on determined evidence, 68 correct pairs, and 14 reversed but incorrect pairs. Pro has 250 unchanged definite pairs, 261 containing unknown, and one pair with a missing response. None can be correct at both ends: the observed changed-premise answer in the incomplete pair is already wrong. Thus the zero joint count does not assume that a missing answer is wrong. The unknown condition elicits 441 verified unsupported definite assertions for Flash, with two missing responses, and 256 for Pro. These findings describe distinct behavioral patterns on independent backgrounds; the response-mode analysis itself was not a prespecified R3 primary endpoint.

L.3 Independent ledger validation

After all four development collections closed, the fixed validation inputs were admitted without changing the material generator, model inventory, request settings, worker counts, scoring, or eight-test family. Luna, Gemini, and GLM were dispatched. DeepSeek Flash is held under the user's pause and retains 896 unobserved outcomes in the original planned denominator. It is not an observed negative result or a response-based model exclusion.

Luna's complete cohort contains 896 service-complete answers and 900 physical scientific requests, with bounded transport retries retained. Original, changed-premise, relevant, changed-relevant, irrelevant, changed-irrelevant, and unknown correct counts are 61, 53, 124, 123, 54, 52, and 104, each out of 128. The both-state relevant comparisons have 116 and 110 gains, respectively, and no losses. The main text reports their conservative paired intervals and eight-test Holm adjustment. Public reference labels, all visible answers, and the exact paired tails were also independently recomputed. Gemini and GLM subsequently close all 896 questions each, with 896 physical scientific requests per model. All Gemini responses are correct. GLM has three scored errors: one changed-premise answer, one relevant-explicit answer, and one unsupported B on an unknown input. Its implicit, relevant, and irrelevant both-state counts are 127, 127, and 128. The two models' four Holm-adjusted comparisons are all 1. The original eight-test family remains intact, including the paused Flash cohort's unobserved outcomes; this is not a completed four-model validation. The independent positive explicitation result is confined to Luna none. Gemini and GLM provide ceiling or near-ceiling observations, not replications of a large benefit; Flash provides no observed validation outcome.

Configuration	J_{imp}	J_{rel}	J_{irr}	Unknown	Unobserved answers
Luna, none	3	119	9	104	0
Gemini 3.1 Pro Preview	128	128	128	128	0
GLM-5.3	127	127	128	127	0
DeepSeek Flash, paused	—	—	—	—	896

Table 24: Independent ledger validation. Joint and unknown counts use 128 backgrounds; each complete configuration has 896 answers. Dashes denote unobserved outcomes, not zero correct answers. The original eight-test family includes the paused configuration.

M Original Prespecified Claims

These requirements were fixed for the earlier C1 and five-view C2 claims. They remain unchanged. The screen-qualified confirmation is reported in the main text as a prospective follow-up result, not as a replacement for this historical cross-family criterion. Its selection, materials, and analysis were frozen before confirmation, while the later crossed test and eight-view transfer study have separate protocols and endpoints.

C1 eligibility requires at least 58 known-correct answers among the 64 development backgrounds in *each* of the explicit, original-relation, counterfactual-relation, and unknown-relation conditions. This is an operational development threshold, not a certification of population accuracy. Every admitted model receives the full validation set regardless of its development effect. We use the conjunction of one-sided exact paired tests for both explicitation comparisons, with Holm correction over the fixed 18 local and API models. The intended cross-model claim requires support from at least one local base model and two distinct commercial model families.

C2 compares the candidate with SFT and paraphrase consistency for both local models. Each seed has a separate background-paired test; their conjunction requires improvement in all three seeds. Holm correction covers the four model–baseline comparisons. The candidate must exceed both baselines by an average of at least five percentage points in original-question accuracy, with positive direction in every seed and corrected statistical support. Family-specific effects are reported without redefining the primary endpoint after validation.

A new unsupported assertion occurs on an unknown case when the candidate outputs A or B and the baseline does not. We report a one-sided exact binomial upper bound for each model, baseline, and seed, using $\alpha = 0.05/12$ for simultaneous coverage. All bounds must be at most two percentage points before claiming the prespecified evidence-boundary risk requirement. Counterfactual accuracy is a separate outcome; consistency alone cannot establish it.

N Screen-Qualified Follow-up after Supervised Adaptation

This follow-up was designed after inspecting the initial experiments and frozen before collecting its new confirmation answers. The narrower configuration-level interpretation does not replace the original cross-family criterion in Appendix M. All new original, changed, and unknown marked canonical forests are disjoint from the earlier training, development, validation, and crossed-test backgrounds. Each task family contributes half of every split.

A separate behavioral screen. On 128 new backgrounds, each configuration receives one positive, one negative, and one unknown relation judgment, plus two explicit decisions. The fixed selection order is the base model followed by its existing seed-17 SFT checkpoint, which was selected using the earlier development split. Each of four empirical accuracies must reach 90%; this is an operational screen, not a simultaneous population-accuracy certificate. Table 25 reports all four configurations. Both base configurations fail the screen, and both SFT configurations pass. The screen contains no implicit-decision question used to select a large explicitation gap.

Model	Configuration	Positive	Negative	Unknown	Explicit
Qwen2.5	Base	69/128	2/128	67/128	142/256
	SFT, seed 17	121/128	121/128	117/128	256/256
Qwen3	Base	108/128	1/128	41/128	123/256
	SFT, seed 17	128/128	126/128	128/128	256/256

Table 25: Separate behavioral screen. Positive, negative, and unknown columns are standalone relation questions; explicit counts both premise states. The sample unit remains one of 128 backgrounds.

New confirmation backgrounds. Each selected model receives ten questions on every one of 256 further backgrounds: the two-premise by three-presentation grid, one unknown decision, and three standalone relation questions. The model checkpoints, materials, selection, and four-test Holm family were frozen before confirmation. Under the original C1 rule, both explicitation differences must be at least five percentage points as point estimates and pass a one-sided exact background-paired test after correction. The capability requirements must also hold on confirmation. We report both models regardless of this historical effect-size requirement. No question was missing or infrastructure-failed.

Qwen2.5's confirmation relation counts are 247/256 positive, 240/256 negative, and 239/256 unknown, with 512/512 explicit decisions. Qwen3's corresponding counts are 255/256, 254/256, 254/256, and 512/512. Both therefore retain the empirical qualification. Their implicit, explicit,

and irrelevant J counts are respectively $(237, 256, 219)$ and $(255, 256, 248)$. For Qwen2.5, the two gains are 19 and 37 backgrounds with no losses, giving differences of 7.42 and 14.45 points and Holm-adjusted $p = 5.72 \times 10^{-6}$ and 2.91×10^{-11}. Qwen3 gains one and eight backgrounds (0.39 and 3.13 points), with adjusted $p = 0.5$ and 0.0078125; neither reaches the original five-point requirement. Thus the new screen-qualified claim is supported for the Qwen2.5 configuration and is not a uniform result across both models. Unknown-decision correctness is 234/256 and 253/256; implicit three-state joint correctness is 222/256 and 252/256.

Label-conditioned decisions and irrelevant-content effects. Table 26 replays the saved responses by reference label, pooling the two determined premise states. Each A/B row contains 256 questions from the same 256 backgrounds; the C row uses the shared unknown input. This is a descriptive decomposition of the opened cohort, with no additional hypothesis family. For Qwen2.5, irrelevant addition repairs three originally incorrect joint decisions and disrupts 21 correct ones, retaining 216 correct and 16 incorrect pairs. For Qwen3 it repairs none, disrupts seven, and retains 248 correct and one incorrect pair.

		Qwen2.5 predictions			Qwen3 predictions		
Presentation	Reference	A	B	C	A	B	C
Implicit	A	245	6	5	255	0	1
Implicit	B	8	244	4	0	255	1
Relevant	A	256	0	0	256	0	0
Relevant	B	0	256	0	0	256	0
Irrelevant	A	239	12	5	248	4	4
Irrelevant	B	20	231	5	0	254	2
Unknown	C	11	11	234	2	1	253

Table 26: Label-conditioned confusion counts for the screen-qualified SFT checkpoints. All rows sum to 256, with zero format or service failures. Unknown inputs are shared across presentations, not additional explicit or irrelevant treatments.

Capability on the same background. The prespecified auxiliary mask R_b equals one when both standalone determined-relation judgments are correct on background b. We compare $R_b J_{\mathrm{rel}}$ with $R_b J_{\mathrm{control}}$ over all 256 backgrounds, using a separate four-test Holm family. This does not select examples by their final-decision errors. Qwen2.5 has 233 such backgrounds; its gains over implicit and irrelevant controls are five and 22, with no losses and corrected $p = 0.0625$ and 9.54×10^{-7}. Qwen3 has 254 such backgrounds, zero gains over implicit and seven over irrelevant inputs, with corrected $p = 1$ and 0.0234375. The Qwen2.5 result therefore establishes a configuration-level gap while leaving a stronger inference about correct derivation followed by failed use less secure against the implicit control. Standalone correctness is also not a direct observation of the representation present during the complete decision.

Query complement on the same confirmation backgrounds. Keeping the premises and resource rule fixed, we complement the queried total. If the two possible totals are b and $b + w$, replacing the query value c by $2b + w - c$ reverses both determined reference answers while preserving C for the unknown case. This gives four determined conditions: original-query original/changed premises, then complemented-query original/changed premises. Correct patterns are ABBA or BAAB. Each model receives 768 additional questions and reuses 1,280 records on the same 256 backgrounds. All additional responses are complete and format-valid. This is a post-hoc diagnostic on already opened material, with the original answer-format checkpoints; it adds no confirmation test.

The five original Qwen2.5 $R_b = 1$ failures of J yield AABA, BBBA, BBBA, AAAB, and AABB in the four-condition order. Four become jointly correct under the complemented query, while another qualified background newly fails. Their behavior therefore cannot be explained by a constant answer letter across these conditions. In the complete cohort, J changes from 237 to 239 with five repairs and three regressions; within $R_b = 1$, it changes from 228 to 231 with four repairs and one regression. Qwen3 has no original qualified J failure, but gains one under the complemented query (ABBC, with reference ABBA); its full-cohort J is 255 before and 254 after the query change.

Four-condition pattern	Qwen2.5 All	Qwen2.5 $R_b = 1$	Qwen3 All	Qwen3 $R_b = 1$
All four correct	234	227	254	253
Query-only response	10	1	0	0
Other mixed A/B responses	5	5	0	0
C in at least one condition	7	0	2	1
Background denominator	256	233	256	254

Table 27: Query-complement patterns. R_b requires both standalone determined-relation judgments to be correct. Query-only responses change with the query but remain unchanged across premise states within each query. Constant, fully reversed, premise-only, and invalid patterns have zero counts. The four conditions remain one paired background.

Unknown correctness changes from 234 to 238 for Qwen2.5, comprising seven repaired assertions and three new assertions. Qwen3 remains at 253, with the same three backgrounds producing unsupported definite answers whose A/B labels reverse with the query. Query responsiveness therefore does not by itself establish warranted evidence use. The original qualification, denominators, and auxiliary Holm results are unchanged.

A secondary output-interface diagnostic. The same frozen checkpoints also generate a T/F/U intermediate state followed by an A/B/C answer on the three implicit decision states. This supplies no gold state at inference. The checkpoints were trained to emit one answer letter, and this new two-label interface produces 275/768 invalid outputs for Qwen2.5 and 535/768 for Qwen3; three-state final correctness is 25/256 for each. Among valid outputs, Qwen2.5 also has 84 correct-state/wrong-answer and 57 wrong-state/correct-answer completions. These are observable disagreements in this interface, not an internal causal mechanism. The format failures are reported separately and do not establish that an interface-matched intermediate-evidence repair is ineffective. All raw outputs, prompt tokens, capability selections, and paired analyses are retained in the independent follow-up artifacts.

O Grok Gateway Follow-up of the Screen-Qualified Test

We extend the screen-qualified design to the API request model `grok-4.7`, whose saved responses identify it as `grok-4.7-build`. These are gateway-reported identifiers, not an independent verification of hidden weights. The two fixed settings request `reasoning_effort=low` and `high`, with temperature zero, `max_tokens=4096`, no tools, and at most four concurrent requests. They are two settings of one endpoint. Both receive the same public messages and previously generated capability and confirmation materials used for the local test. The scientific contrasts, test families, and request settings were fixed before Grok received these materials; the researchers had already seen the local-model outcomes. The usage-rule amendment below was fixed after partial screening and before all confirmation calls.

Screening and fixed confirmation design. Each setting answers five questions on each of 128 capability backgrounds: positive, negative, and unknown standalone relations, and explicit decisions at both premise states. The four empirical accuracy requirements are each 90%; qualification does not depend on finding an implicit-decision error. Both settings answer all 640 questions correctly. The observed counts for each are 128/128 in every relation category and 256/256 explicit decisions, with no invalid or truncated answers. A conservative token-overrun check described below changes high's unknown-relation count to 127/128 and leaves all other counts unchanged; both settings still qualify.

Each qualified setting's confirmation design contains ten questions on each of the same 256 further backgrounds used for the local confirmation: the six determined decisions, one unknown decision, and three standalone relation judgments. These are 2,560 answers per setting, with the background as the statistical unit. The primary analysis compares relevant-explicit J with implicit and irrelevant J, retaining all four low/high by control slots in one Holm family. A separate four-test family compares the same-background products R_bJ. All effect sizes are reported, including zero and

negative differences. Positive directional evidence requires qualification in both stages, positive differences against both controls, and both adjusted one-sided paired tests below 0.05. The original five-percentage-point C1 criterion is retained as a separate historical label; it does not determine inclusion in the paper.

Observed token use and response retention. The gateway did not always respect the requested token limit: one high screening response reported 4,520 completion tokens, including 4,519 reasoning tokens, despite the submitted limit of 4,096. It returned a valid single-letter answer with finish reason `stop`. The original collector stopped under its then-current usage rule after retaining 640 low and 463 high responses. Before any confirmation request, an explicit amendment set the observed-completion review limit to 8,192 while leaving every scientific request unchanged. All 1,103 responses were retained with links to the original bytes and classifications; only the 177 uncalled high screening questions were subsequently sent. The original failed execution remains recorded separately.

Qualification must survive counting every response above 4,096 reported completion tokens as incorrect. Confirmation reports the observed answers, flags such responses, and gives identification ranges allowing affected answers to be wrong; these ranges are distinct from sampling intervals. Actual token totals, transport failures, and health checks are reported separately. Submitted limits do not establish equal realized compute or a verified billing ceiling.

Complete confirmation outcomes. Both settings answer all 2,560 confirmation questions correctly, including 256/256 in each standalone relation category and 512/512 explicit decisions. Implicit, explicit, and irrelevant J are each 256/256; unknown correctness, implicit three-state correctness, and both-relation correctness are also 256/256. Each task family contributes 128/128 correct backgrounds to every joint presentation. No final response is invalid, truncated, or service-missing.

Every primary and same-background relation-qualified contrast has zero gains, zero losses, and 256 jointly correct pairs. All eight adjusted p values in the two separate four-test families equal one. The conservative two-sided 95% interval for each difference is $[-1.70, 1.70]$ percentage points; these intervals are not simultaneous. Descriptive high-minus-low differences are also zero. Figure 6 displays every fixed primary contrast. These observations establish neither a positive C1 effect nor equivalence on other tasks or inference settings.

The high setting has 13 further token-overrun answers in confirmation: four unknown-relation questions and nine unknown decisions, all in the maternal-ancestry family. The largest reports 5,301 completion tokens. Counting every overrun answer as wrong leaves its capability screen at 127/128 unknown relations and its confirmation capability at 252/256; both remain qualified. Allowing arbitrary correctness on the nine affected unknown decisions gives unknown and three-state counts in $[247, 256]$, and unsupported assertions in $[0, 9]$. None of the determined decision pairs is affected: all J counts remain exactly 256 and both primary differences remain zero. These are token-policy sensitivity ranges, not sampling intervals.

Stage	Setting	Answers	Extra attempts	Completion	Reasoning
Screen	Low	640	0	250,482	249,842
	High	640	0	489,780	489,140
Confirmation	Low	2,560	2	1,061,742	1,059,182
	High	2,560	8	2,282,936	2,280,376

Table 28: Recorded Grok generation and reported token use. Screen rows include retained original responses. Extra attempts are bounded transport retries in the retained/continued collection; the earlier 11 unanswered channel attempts are additional. Unknown usage is not assigned zero.

There are 6,421 physical scientific requests: 6,400 answered requests, 11 earlier channel failures without answers, and ten additional transport attempts during confirmation. The 1,103 retained responses are included once. Known scientific usage totals 11,291,794 prompt tokens, including 6,705,152 reported cached tokens, and 4,084,940 completion tokens, including 4,078,540 reasoning tokens. All four usage fields are unknown for 21 attempts. Four generation health checks are separate from scientific denominators: two returned HTTP 403 and two returned valid answers with 173 and 140 completion tokens. Actual currency charges were not available. The original channel

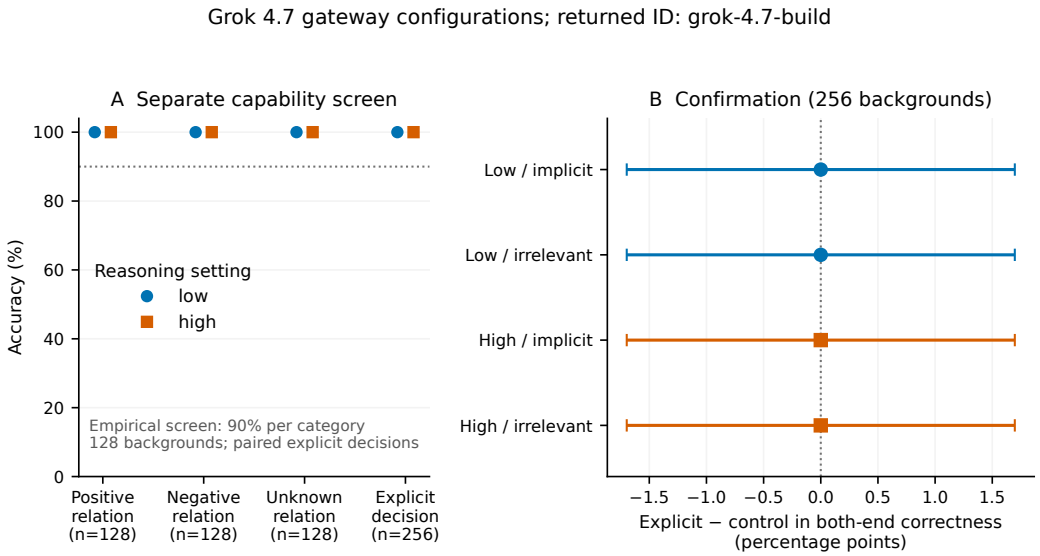


Figure 6: Complete Grok capability screen and confirmation. Left: empirical capability accuracy, without sampling intervals; the explicit-decision questions are paired within 128 backgrounds. Both settings qualify even if the one high-screen token-overrun answer is counted as wrong. Right: all four explicit-minus-control differences on 256 confirmation backgrounds, with conservative paired 95% intervals, not simultaneous. Each difference is zero; all corresponding Holm-adjusted p values are one.

failure and strict-usage halt retain their nonzero exits; successful completion refers to the explicitly amended collection.

P Full-Budget Development of Semantic-Change Consistency

This separate eight-view development study examines semantic-change consistency after the screen-qualified initialization. It does not replace the five-view independent validation above. Each model starts from its fixed qualified seed-17 adapter with a fresh optimizer and uses 2,048 training backgrounds, 512 updates, and eight checkpoints at intervals of 64 updates. The three development splits contain 128 backgrounds each: ID, structural shift, and expression shift. All checkpoint selection uses these development splits; confirmation remains unopened.

Training and ID development use paths of two to four links. Structural development uses five or six links while retaining the same 17, 21, or 25 entities, so it does not jointly increase the entity-count range. Expression development keeps two-to-four-link paths and changes the record sentences to held-out phrasings, such as “X’s mother is Y” and “X passes authority to Y.” The rules and resource queries retain their semantics. All original, changed, and unknown canonical forests are disjoint across these splits and from the prior study. Within each family and split, relation truth and the final A/B label are independently balanced. These are controlled structural and template shifts, not unrestricted natural-language variation.

The prospective transfer-confirmation protocol uses 512 new backgrounds for each transfer slice and 1,024 for ID protection, with seeds 17, 29, and 43. Its primary endpoint is T, with a mean improvement of at least three percentage points over each matched control and positive direction in every seed. Each model–control comparison uses the largest of its three exact paired p values, followed by Holm correction across four comparisons. The ID difference must have a simultaneous lower bound of at least -1 point, and the new-assertion rate a simultaneous upper bound of at most 2%. Here a new assertion changes a baseline C to A or B; its denominator includes all transfer backgrounds. The two bound families divide a total error budget of 0.05 equally; each covers twelve model–control–seed comparisons. Each transfer slice also has a -1-point mean-difference guard. These requirements concern a different endpoint from the original five-point original-question crite-

rion in Appendix M. Development point-estimate screening does not establish these confirmation bounds.

Matched objectives and interfaces. The eight supervised views are implicit, explicit, paraphrased, changed-premise, explicit changed-premise, paraphrased changed-premise, unknown, and paraphrased unknown. Answer SFT emits an A/B/C answer and EOS; the three state-format arms emit T/F/U, a final answer, and EOS. All state-format arms share eight-view cross-entropy. Their state probabilities retain OTHER as a fourth category. Writing $P_I, P_E, P_P, P_F, P_{FE}, P_{FP}, P_U, P_{UP}$ for these distributions and S for the T/F swap that leaves U and OTHER fixed, the semantic term is

$$\frac{1}{3}\left[\mathrm{JS}(P_I, P_E) + \mathrm{JS}(P_F, P_{FE}) + \mathrm{JS}(P_I, SP_F)\right].$$

The matched paraphrase term averages $\mathrm{JS}(P_I, P_P)$, $\mathrm{JS}(P_F, P_{FP})$, and $\mathrm{JS}(P_U, P_{UP})$. Both regularized arms use the full fixed grid $\lambda \in \{0.1, 0.5\}$. Each model therefore has six complete runs, including the two SFT interfaces. The twelve runs provide 96 training checkpoints and 110,592 unique retained development answers. Figure 7 retains all 96 checkpoint values and the six selected-endpoint pairings.

Model	Arm	λ/step	ID T	Struct. T	Expr. T	Transfer T
Qwen2.5	Semantic	0.5/448	128	103	128	231
	Paraphrase	0.5/384	128	99	125	224
	State SFT	0/448	126	86	127	213
	Answer SFT	0/448	125	102	128	230
Qwen3	Semantic	0.5/448	126	123	128	251
	Paraphrase	0.5/192	122	119	128	247
	State SFT	0/320	127	121	127	248
	Answer SFT	0/256	126	122	125	247

Table 29: Complete-budget seed-17 development endpoints. Each split has 128 backgrounds; transfer combines structure and expression, giving 256. T requires all three final answers to be correct. SFT endpoints were fixed earlier; paraphrase uses its strongest full-grid point, and semantic uses the protected development selection. These selected development values are not independent confirmation results.

Positive increments and their paired composition. The selected semantic arm gains 18 and three transfer backgrounds over state SFT, corresponding to 7.03 and 1.17 percentage points for Qwen2.5 and Qwen3. Relative to the strongest paraphrase point, its net gains are seven and four backgrounds (2.73 and 1.56 points). Table 30 separates repairs from regressions. The two state-format controls and semantic candidate still share ten failures on Qwen2.5 and two on Qwen3. The fixed answer-SFT reference gives net gains of one and four backgrounds, motivating an interface-matched test of answer-space consistency.

Model	Semantic compared with	Repairs	Regressions	Net /256
Qwen2.5	State SFT	26	8	+18
	Paraphrase	20	13	+7
	Answer SFT	12	11	+1
Qwen3	State SFT	4	1	+3
	Paraphrase	7	3	+4
	Answer SFT	7	3	+4

Table 30: Paired changes in three-answer joint correctness on the same 256 transfer backgrounds per model. A repair or regression is a background-level event, irrespective of how many of its three answers change.

Unknown evidence and remaining errors. The Qwen2.5 candidate repairs all seven transfer assertions from paraphrase while introducing two elsewhere. Qwen3 repairs all four and makes no

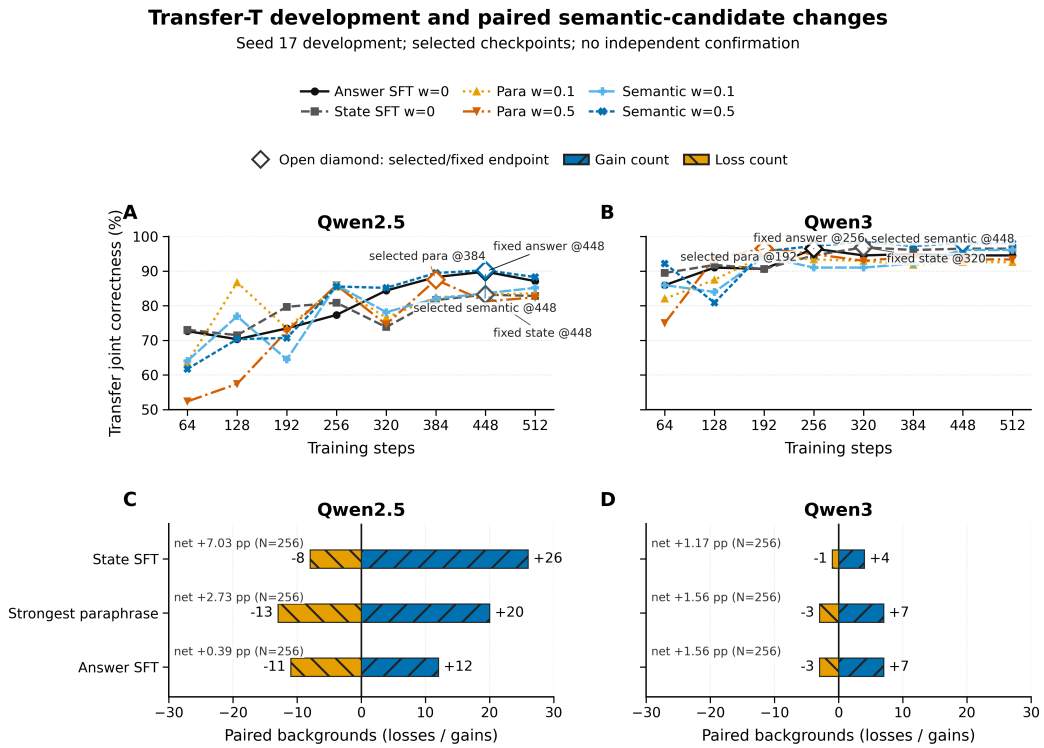


Figure 7: Full-budget seed-17 development. Top: all eight checkpoints of six training configurations per model; the two initial-adapter references are omitted. T requires correct final answers on the implicit, changed-premise, and unknown views. Open diamonds mark the previously fixed SFT references, the strongest paraphrase point, and the protected semantic selection. Qwen3's semantic maximum at step 320 fails the ID protection; step 448 is selected. Bottom: paired repairs and regressions of the semantic candidate against the three references in Table 30, using the same 256 transfer backgrounds per model. These are descriptive development results from one training seed.

transfer assertion, but retains two ID assertions. Thus the transfer protection does not imply absence of new risk on every split. On structural development, Qwen2.5 has 37 determined inputs answered U and two other known-state errors; Qwen3 has eight U errors across five depth-six backgrounds. None of the six selected state-format endpoints has a correct-state/wrong-final-answer response. This locates the observed errors in state prediction within that interface, without identifying an internal derivation. It motivates testing the same semantic constraints directly in the answer space with matched answer-format supervision.

Consistency in the answer space. We next apply the same three semantic pairs to the answer-position A/B/C/OTHER distributions, with S exchanging A and B and preserving C and OTHER. All eight views retain answer/EOS cross-entropy, and both branches of each consistency pair receive gradients. The initialization, supervised tokens, optimizer, and 512-update budget match answer SFT. Both coefficients are complete for both models: four runs, 32 checkpoints, and 36,864 semantic-arm development answers. All retained answers are format-valid and untruncated. Figure 8 shows every checkpoint, including those excluded by protection criteria.

Model	Objective	λ/step	ID T	Struct. T	Expr. T	Transfer T
Qwen2.5	Semantic, selected	0.1/448	126	111	127	238
	Fixed answer SFT	0/448	125	102	128	230
	Strongest paraphrase	0.1/384	126	110	123	233
Qwen3	Semantic, peak	0.1/384	127	120	128	248
	Fixed answer SFT	0/256	126	122	125	247

Table 31: Answer-space consistency and controls on the same seed-17 development splits. Split denominators are 128; transfer has 256 backgrounds. The Qwen3 peak is shown to explain its exclusion and is not an admitted candidate. Its structural difference is $-2/128$, below the -1-point guard.

Qwen2.5 has one protected point among 16: $\lambda = 0.1$ at step 448, with 238/256 transfer backgrounds correct versus 230/256 for fixed answer SFT. The gain of 3.13 percentage points comprises 16 repairs and eight regressions. Structural correctness increases by nine backgrounds, expression correctness decreases by one, and ID correctness increases by one (Table 31). Three previously correct unknown answers become assertions on the 256 transfer backgrounds. The higher unprotected score of 244/256 at step 384 is excluded: expression correctness falls by four backgrounds and seven unknown answers become assertions. The protected candidate was fixed before its matched answer-space paraphrase grid.

The complete matched grid gives a strongest paraphrase control of 233/256 at $\lambda = 0.1$, step 384. The locked semantic candidate gains 5/256 (1.95 percentage points), with 13 repairs and eight regressions (Figure 9). Its ID difference is zero, structural difference $+1/128$, and expression difference $+4/128$; two previously correct unknown answers become assertions. These satisfy the prespecified development guards. The family breakdown is heterogeneous: $-3/128$ on delegated permission and $+8/128$ on maternal ancestry. All 16 control points and 18,432 new answers are retained, with no invalid or truncated answers. The full runs share a 512-update budget; the selected paraphrase checkpoint uses an earlier prefix than the semantic and SFT references at step 448. This is an estimated single-seed development gain. Fixed seed-29 and seed-43 replications assess its stability without reselecting coefficients or checkpoints. Effect sizes are reported independently of the original three-point strong-confirmation criterion.

Qwen3 has no protected point among the 16 checkpoints across both coefficients. Every point fails the structural guard. Its unprotected maximum of 248/256 has eight repairs and seven regressions relative to fixed answer SFT, with two fewer structural backgrounds correct. At this point, determined structural questions answered C increase from four to eleven. The aggregate one-background increase therefore does not establish a qualifying repair. Both model-specific outcomes retain the same protection rules and complete development denominators.

Replication at fixed coefficients and checkpoints. We retain the seed-17 choices for two further training seeds, 29 and 43: answer SFT uses $\lambda = 0$, step 448; paraphrase consistency uses $\lambda = 0.1$, step 384; semantic consistency uses $\lambda = 0.1$, step 448. All start from the same previously qualified

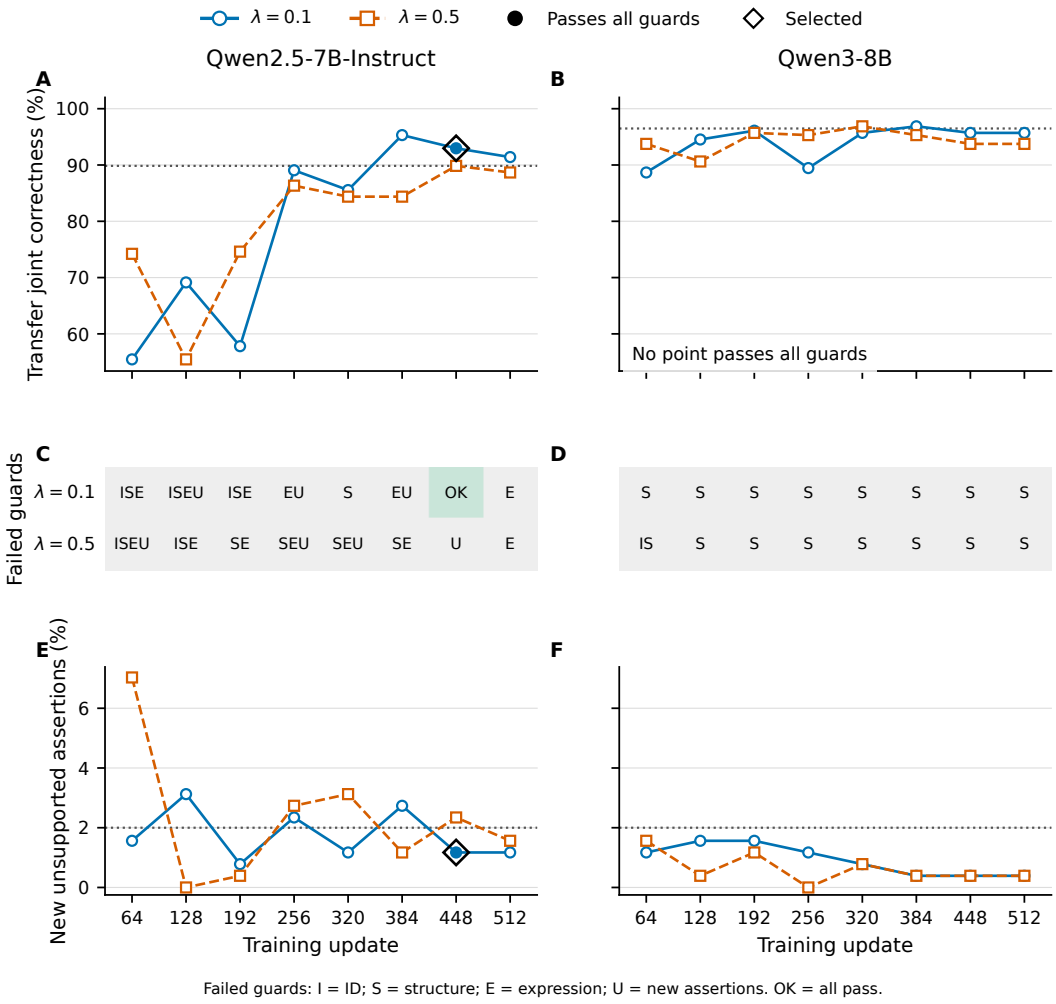


Figure 8: Complete answer-space semantic grids, seed 17. Top: transfer T for all eight checkpoints at each coefficient; dotted horizontal lines show the previously fixed answer-SFT references. Filled markers pass all development guards, and the diamond marks the protected Qwen2.5 selection. Middle: failed guards at each point—I, S, and E denote an ID, structural, or expression difference below -1 percentage point; U denotes new transfer unknown assertions above 2%. OK indicates all guards pass. Bottom: rates of baseline-C answers changed to A or B, with the 2% limit shown. Each model reuses the same 256 transfer backgrounds across its 16 checkpoints. These are descriptive development comparisons, without independent-confirmation intervals or additional independent samples.

SFT adapter, with fresh optimizers, and run the full 512-update budget. The later seeds do not reselect a coefficient or checkpoint. Table 32 reports every fixed endpoint.

Seed	Arm	ID	Structure	Expression	Transfer
17	Answer SFT	125	102	128	230
	Paraphrase	126	110	123	233
	Semantic	126	111	127	238
29	Answer SFT	125	99	126	225
	Paraphrase	123	100	125	225
	Semantic	126	90	126	216
43	Answer SFT	125	87	126	213
	Paraphrase	125	94	126	220
	Semantic	127	106	127	233

Table 32: Complete fixed-seed development replication on Qwen2.5. Entries count three-state joint correctness T: ID, structure, and expression each use 128 backgrounds; transfer pools the latter two, $n = 256$. The same backgrounds are reused across arms and seeds.

Mean transfer correctness is 86.98% for answer SFT, 88.28% for paraphrase consistency, and 89.45% for semantic consistency. The semantic differences average $+2.47$ and $+1.17$ percentage points, respectively. Seeds 17 and 43 improve over both controls, whereas seed 29 declines by 3.52 points against each. Most seed variation occurs on the structural shift; expression counts remain near ceiling. Table 33 preserves both repairs and regressions, as well as newly unsupported answers on unknown inputs.

Seed	Semantic versus	Repairs	Regressions	ΔT (pp)	New A/B
17	Answer SFT	16	8	$+3.13$	3
	Paraphrase	13	8	$+1.95$	2
29	Answer SFT	13	22	-3.52	4
	Paraphrase	10	19	-3.52	4
43	Answer SFT	30	10	$+7.81$	2
	Paraphrase	20	7	$+5.08$	1

Table 33: Paired development changes on the same 256 transfer backgrounds. New A/B counts a control's correct C becoming an unsupported definite answer under semantic consistency; it is distinct from a regression in T. No minimum effect size determines inclusion. These are descriptive counts, not independent-confirmation tests.

The complete audit reconstructs 58,752 answers: 55,296 from all eight checkpoints of the six later-seed runs, plus 3,456 from the three fixed seed-17 endpoints. No answer is invalid, truncated, or service-failed. Each of the 51 checkpoint endpoints has its prompt tokens, generated tokens, and three development splits independently reparsed. The paired comparisons are reconstructed from background-level outputs, and all nine selected weight files match their frozen checkpoint receipts. These repetitions characterize training-seed variation on development material; they neither multiply the number of task backgrounds nor open the separate C2 confirmation.

A host interruption during seed-29 paraphrase training required an explicitly recorded recovery. Across the six later-seed runs, the logical budget is 3,072 updates, with 3,109 recorded physical updates and an upper bound of 3,110 because the pre-interruption step 38 has an unknown completion state. Original process exits remain unknown. The successful recovery and remaining run exits are recorded separately; equal logical budgets do not imply equal physical execution costs.

Stopping the teacher branch. We additionally test whether removing the consistency gradient from its target branch preserves a useful correction signal. Each semantic pair uses the same Jensen–Shannon value, with the explicit view serving as the stopped target for its corresponding implicit view; in the premise-flip pair, the permuted changed-premise distribution is the stopped target. Both coefficients retain the eight-view answer/EOS supervision, original initialization, and 512-update logical budget. This is a standard stop-gradient control (Xie et al., 2020), not a new self-distillation mechanism.

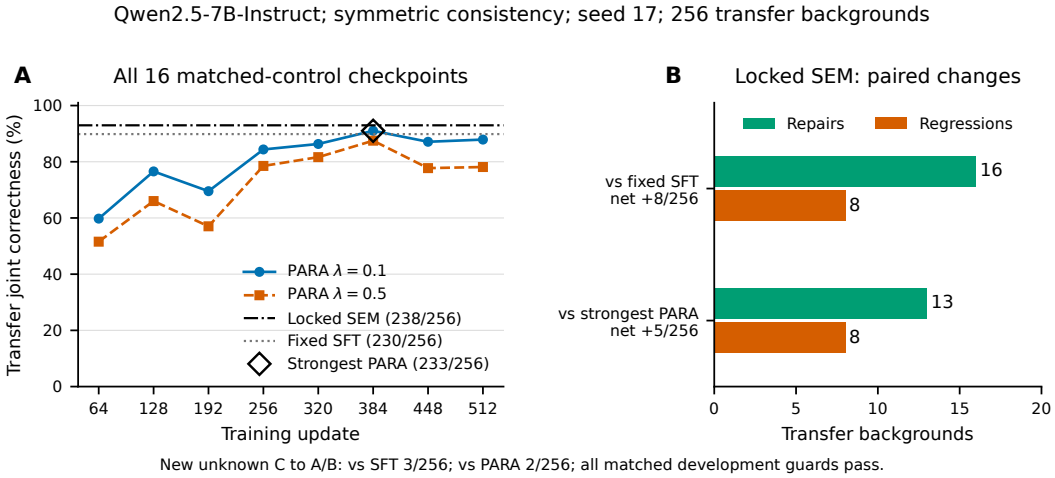


Figure 9: Complete matched answer-space control for Qwen2.5, seed 17. Left: all eight checkpoints at each paraphrase coefficient; the open diamond marks the strongest of all 16 points. Horizontal lines show the previously locked semantic candidate and fixed answer SFT. Right: paired repairs and regressions of that candidate relative to both references. All comparisons reuse the same 256 transfer backgrounds; the counts are descriptive development results. The semantic candidate is unchanged after observing the controls.

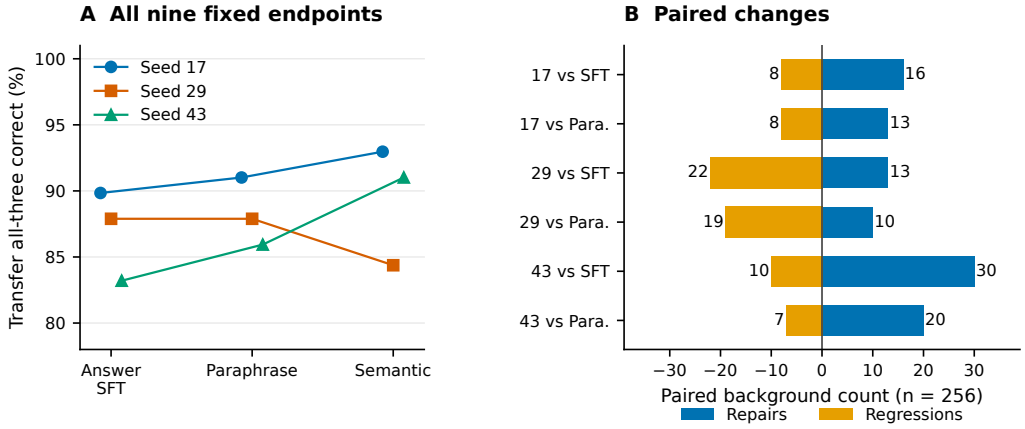


Figure 10: All fixed Qwen2.5 seed endpoints and paired transfer changes. Left: the same semantic candidate has a higher three-seed mean but lower seed-29 correctness than both controls. Right: every repair and regression contributing to the six differences. All comparisons reuse 128 structural and 128 expression development backgrounds; the figure adds no inferential test or independent sample.

The complete Qwen3 grid contains 16 checkpoints and 18,432 new development answers. Together with 1,152 cached SFT-reference answers, all 19,584 raw records and attempts were independently reparsed, with no invalid or truncated output. Every checkpoint has lower pooled transfer correctness than the fixed SFT reference, 247/256. The highest stopped-teacher point is 243/256 at $\lambda = 0.5$, step 320: ID, structure, and expression counts are 124/128, 115/128, and 128/128, compared with 126/128, 122/128, and 125/128 for SFT. All 16 points lose structural correctness (Figure 11). Thus this change does not improve the tested grid even before a minimum positive effect criterion is applied.

A host interruption required one explicitly recorded continuation from the complete step-256 optimizer, scheduler, and RNG state. The two runs contain 1,024 logical updates and 1,080 recorded physical updates; an unclosed pre-interruption update gives an upper bound of 1,081. All earlier artifacts and unknown process exits remain preserved. These execution costs are separate from the logical training-budget comparison.

Answer-position follow-up. We also examine the earlier unprotected semantic checkpoint using 32 development backgrounds: eight repairs, seven regressions, one jointly incorrect case, and 16 stable-correct controls relative to fixed answer SFT. The cohort was selected from existing outcomes before collecting eight-view logits at three fixed checkpoints (the qualified initializer, answer SFT, and the original semantic checkpoint). We independently reconstructed the readouts from all 768 saved full-vocabulary logit arrays; no parameters were updated. On the seven regression backgrounds, the original semantic checkpoint's highest-probability vocabulary token is correct for 0/7 implicit inputs and 6/7 explicit inputs; the corresponding mean gold-token probabilities are 0.068 and 0.835. Fixed answer SFT is correct on 7/7 of each view. This error-conditioned readout documents residual sensitivity to relevant explicitation at a failed checkpoint. It measures the first answer position, not a new full-generation accuracy estimate or a causal effect of an optimizer update.

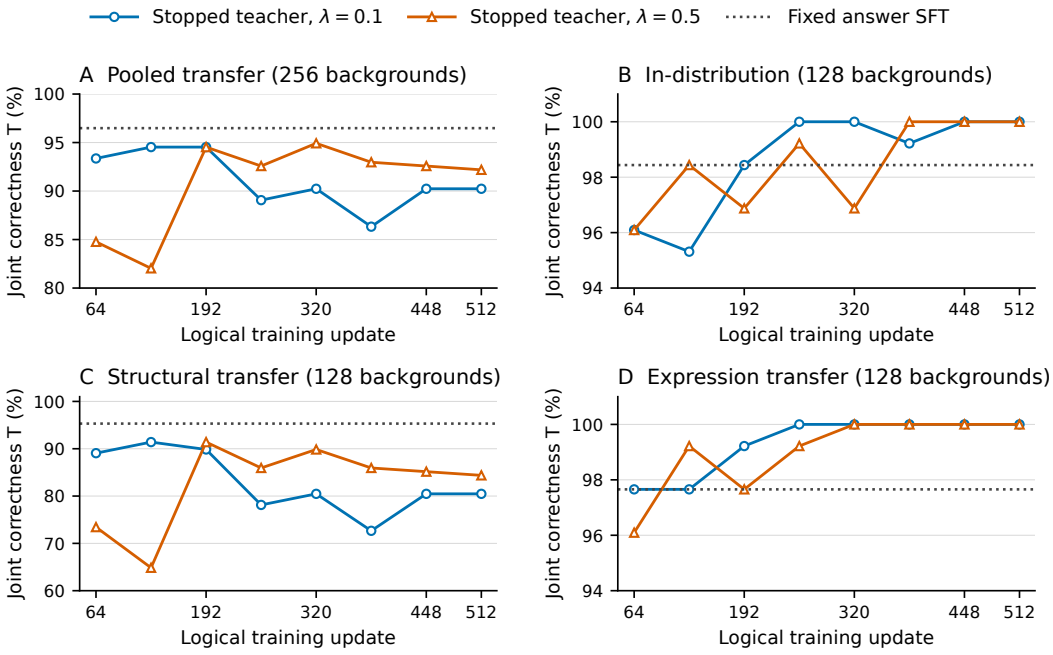


Figure 11: Complete stopped-teacher semantic grid on Qwen3, seed 17. Each curve contains all eight checkpoints at its fixed coefficient. Dotted lines show the previously fixed answer-SFT reference. Transfer combines the same 128 structural and 128 expression backgrounds at every checkpoint; ID uses a separate 128. Panel axes have different vertical ranges. The curves are descriptive development results, and checkpoints do not add independent samples. All 16 points have lower pooled transfer T and lower structural T than their SFT references.

Q FIXED EXTERNAL EVALUATION ON FOLIO

The external evaluation uses the public v0.0 validation release of FOLIO (Han et al., 2024), pinned to the authors' repository commit `5d7bb84c7edab3fb358e057d2807f19cf5cf5e2d`. The source is `data/v0.0/folio-validation.jsonl` in `https://github.com/Yale-LILY/FOLIO`. All 204 examples are retained: 72 True, 63 False, and 69 Uncertain, mapped to A, B, and C. This older public release is distinct from the corrected data and the publication's private test set. We do not revise its labels, filter difficult examples, or compare its accuracy directly with published test-set rankings. Pretraining exposure to this public material is unknown. We score against its released annotations; the exact public-text reference evaluator for our generated tasks is not applied here.

Models and input interface. The design fixes nine Qwen2.5 checkpoints: seeds 17, 29, and 43 for answer SFT, answer paraphrase consistency, and answer semantic consistency. The respective coefficient/checkpoint pairs are $0/448$, $0.1/384$, and $0.1/448$, selected using seed-17 synthetic development and unchanged for the later seeds. All nine share the prior qualified SFT initialization; their complete training budgets are 512 updates. The unadapted base model provides one additional descriptive reference. The complete denominator is ten configurations by 204 examples, or 2,040 generated answers.

Every configuration receives the same zero-shot interface: a system message defines A as necessarily true, B as necessarily false, and C as undetermined, and asks for one letter. The user message numbers the original natural-language premises and presents the original conclusion. Formal-logic annotations, reference labels, demonstrations, retrieval, and tools are excluded. The common BF16 model and tokenizer use greedy decoding with 64 new tokens. All 204 inputs pass the fixed 4,096-token input limit without truncation (the longest has 298 tokens). A non-letter or truncated completion is incorrect and separately counted; model errors do not trigger answer retries.

The system message is the following literal string:

```
Use only the given premises. Answer A if the conclusion must be true, B
↪  if it must be false, or C if its truth cannot be determined. Output
↪  exactly one letter and nothing else.
```

The user message has the following template. The first placeholder contains every original premise in its original order, one numbered line per premise (`1.`, `2.`, and so on); the second contains the unchanged conclusion. The displayed blank line is retained.

```
Premises:
{numbered_premises}

Conclusion:
{conclusion}
```

Paired analysis and its scope. Accuracy uses all 204 examples. Class-specific counts and unsupported A/B answers on all 69 Uncertain examples are reported separately for every configuration. The main descriptive contrasts are semantic minus SFT and semantic minus paraphrase, both per seed and averaged over the three fixed seeds. The release contains no story identifiers. We therefore group by identical complete natural-language premise lists, yielding 73 text groups, and explicitly treat these as a grouping proxy.

For each contrast, 10,000 paired bootstrap draws resample these groups with replacement (random seed 20260922), retaining every example and all three seeds within a selected group. Each draw computes the case-weighted mean difference, rather than averaging group accuracies equally. The 2.5th and 97.5th percentiles form a descriptive interval conditional on these checkpoints and this grouping; they are not simultaneous intervals or evidence from 612 independent examples. Every sign and every seed is retained, with no minimum effect required for reporting. This evaluation tests ordinary three-valued accuracy outside the generated tasks; it has no matched three-state T construction and does not replace the original C2 confirmation or identify an internal mechanism.

Complete results. All ten configurations complete 204/204 examples with no invalid format, truncation, service failure, or answer retry. The independent raw audit reconstructs all 2,040 input prefixes, token sequences, scores, and paired intervals. Table 34 gives all configuration and class counts.

Model or arm	Seed	All /204	True /72	False /63	Unknown /69
Unadapted	–	102	42	44	16
Answer SFT	17	118	51	44	23
Paraphrase	17	119	50	43	26
Semantic	17	125	50	41	34
Answer SFT	29	127	55	48	24
Paraphrase	29	118	44	45	29
Semantic	29	114	39	33	42
Answer SFT	43	129	56	45	28
Paraphrase	43	118	45	38	35
Semantic	43	115	39	47	29

Table 34: All fixed FOLIO results, scored against the public v0.0 annotations. Entries are correct counts; denominators appear in the headers. Since every output is a valid letter, unsupported assertions on Unknown examples equal 69 minus the last column. The base model is a descriptive reference and was not the common initialization of the later adaptation stage.

The semantic three-seed mean is 57.84%, compared with 61.11% for answer SFT and 58.01% for paraphrase consistency. Its mean paired differences are -3.27 percentage points against SFT, with descriptive 95% interval $[-6.63, 0.16]$, and -0.16 against paraphrase, interval $[-2.82, 2.69]$. Seed 17 improves by 3.43 and 2.94 points, respectively; seeds 29 and 43 decline against both controls (Table 35; Figure 12). The intervals condition on the three fixed training seeds and 73 text groups. They do not establish equivalence or a positive average external benefit.

An exploratory natural-language addition probe. A preliminary probe uses 64 target conclusions from 59 stories in the public FOLIO v0.0 training release. It appends either a true-labelled conclusion from the same story or one from another story, withholding target labels and formal annotations. Same-story origin alone does not verify that the added statement is a relevant intermediate for the target, and the two additions use different source labels. This is an addition probe rather than a replication of the controlled semantic intervention. The Qwen2.5 screen-qualified SFT checkpoint completes all 192 inputs: unchanged, same-story, and other-story accuracies are 33/64, 36/64, and 32/64, with 20, 20, and 19 unsupported A/B answers on the 25 uncertain targets. Luna scores 40/64, 37/64, and 38/64, with no unsupported assertion on those targets. These descriptive counts neither establish an external explicitation benefit nor a C1 confirmation. Repeated targets from a story are not independent backgrounds; the public release may have appeared in pretraining. No model-effect claim is made for the incomplete GLM/Gemini batches.

Seed	Semantic versus	Repairs	Regressions	Difference (pp)
17	Answer SFT	16	9	$+3.43$
	Paraphrase	16	10	$+2.94$
29	Answer SFT	18	31	-6.37
	Paraphrase	15	19	-1.96
43	Answer SFT	8	22	-6.86
	Paraphrase	14	17	-1.47

Table 35: Complete FOLIO paired changes on the same 204 examples per seed. Each difference is (repairs minus regressions)/204; no configuration or example is excluded according to its observed effect.

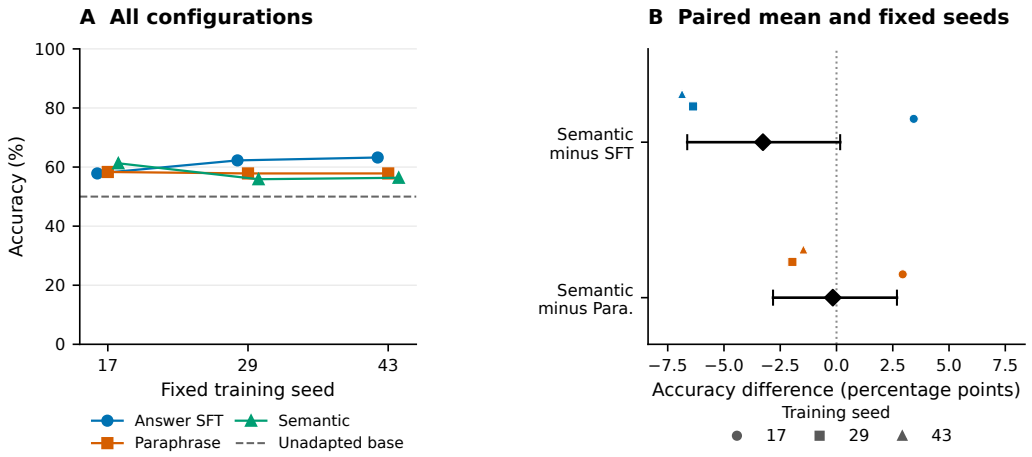


Figure 12: Fixed external FOLIO evaluation. Left: every seed and arm, with the unadapted model as a descriptive reference. Right: semantic-minus-control differences for each seed and their mean; whiskers are the fixed paired premise-group bootstrap intervals. Intervals are conditional on these seeds, not simultaneous, and use text groups as a proxy for unavailable story IDs.

The semantic arm averages 35/69 correct unknown judgments, compared with 25/69 for SFT and 30/69 for paraphrase. This descriptive class pattern coexists with lower correctness on determined statements; uncertainty handling alone does not establish a better overall decision rule. Each configuration uses 31,330 prompt tokens and 408 generated tokens (one letter plus EOS for each example). An initial CPU-affinity error prevented the owner from starting and produced no model answers; its logs are retained separately. After the scheduling correction, all ten configurations ran once under the unchanged scientific admission.

R NATURAL-LANGUAGE SEVEN-CONDITION EXTENSION

We construct seven matched inputs from each of 14 distinct stories in the public FOLIO v0.0 training release. Original and changed premises each have implicit, relevant-explicit, and irrelevant-explicit presentations; a shared removed-evidence input requires uncertainty. These are manually constructed FOLIO-derived interventions, not unchanged benchmark examples. Each relevant addition is verified as a premise consequence, and the unknown case has two compatible completions with opposite target labels. A Boolean encoding checks the reviewed interpretation; it does not certify English semantics. Some stories share templates, limiting population inference.

The retained stories come from an outcome-unseen 16-story candidate set; two ambiguous cases are excluded before collection (a rule-direction issue and a modal-negation issue). Required premise edits and all exclusions are recorded. Ten later candidates are not admitted. The material freeze occurred after the originally planned noon deadline; this departure is retained. Eight disjoint interface stories cover all three output labels. Interface admission requires terminal, parseable service responses, not correct answers. No scientific input or model is selected by its observed effect. API output caps are 4,096 tokens; local Qwen2.5 uses the previously screen-qualified SFT checkpoint with BF16 greedy decoding and a 64-token cap.

Configuration	J_{imp}	J_{rel}	J_{irr}	u	$T_{\mathrm{imp}}/T_{\mathrm{rel}}/T_{\mathrm{irr}}$
Qwen2.5 SFT	11	14	11	11	8/11/8
Luna (5.6, none)	10	12	8	13	9/11/7
Grok-4.7 low	12	13	12	14	12/13/12
Gemini-3.1 Pro Preview	13	13	13	14	13/13/13

Table 36: The complete first batch of FOLIO-derived seven-condition results. All entries are correct story counts out of 14, with 98 scientific inputs per configuration. imp, rel, and irr denote implicit, relevant-explicit, and irrelevant-explicit presentations. The shared unknown input determines u and contributes to every T. GLM-5.3 has no scientific result: its interface returned the unaccepted identity `glm-5.3:dev`, stopping collection before scientific dispatch. It remains in the planned cohort and multiplicity family.

Relevant explicitation improves J by 3/14 for Qwen2.5 and 2/14 for Luna, with no regressions relative to implicit presentation. Against the irrelevant addition, the corresponding repairs are 3/14 and 4/14, also without regressions. Grok has two repairs and one regression relative to implicit presentation; Gemini has no net change. All 392 scientific inputs are service-complete; one Gemini output is unscorable and is counted wrong. None of the ten planned one-sided paired tests passes Holm correction (the smallest adjusted p is 0.625; unavailable GLM tests receive $p = 1$). Descriptive whole-story bootstrap intervals for $J_{\mathrm{rel}} - J_{\mathrm{imp}}$ are $[0, 42.86]$, $[0, 35.71]$, $[-14.29, 28.57]$, and $[0, 0]$ percentage points in table order. These small-sample results extend the observed behavioral pattern to constructed natural-language interventions; they do not establish the original cross-family C1 criterion or an internal-knowledge claim.

Which errors are repaired? The two Luna-5.6 repairs and four GPT-6 Luna repairs replace unwarranted unknown answers on determined premise pairs with both-state correctness. Qwen2.5's three repairs comprise two such cases and one unchanged definite answer after a premise reversal. Every repaired story in these configurations also answers the shared, genuinely insufficient-evidence input correctly; this is a joint outcome, not an effect on an untreated unknown input. The irrelevant addition repairs none of these Luna-5.6 or Qwen2.5 failures and only one of the four GPT-6 Luna failures. Conversely, both Grok repairs also occur with the irrelevant addition. Its small aggregate gain therefore provides weaker evidence of content-specific repair. These are descriptive transitions on the full 14-story cohort, not independent replications per question or an internal-mechanism test.

For a concrete Qwen2.5 example, laboratory membership implies a discounted price, which implies 1080p resolution; the final rule maps that resolution to lack of type-C support. The model answers A under both the original and reversed final rule. Adding only the intermediate resolution statement produces B then A, while removing the entity's lab-membership premise yields C. The added sentence does not state the requested port-support answer. This recorded case illustrates the inter-

vention's semantic distinction; it was selected for illustration after collection, not used to estimate prevalence by itself.

Additional API configurations. A separately authorized eleven-configuration expansion retains the same 14 stories, public inputs, generation caps, and original scorer. Interface admission now checks service completion and identity; parseability is recorded but does not select configurations. The fixed family contains 22 contrasts, including unavailable tests as $p = 1$. In its completed GPT-6 Luna configuration, $J_{\text{imp}}/J_{\text{rel}}/J_{\text{irr}} = 8/12/8$ and unknown correctness is 14/14. Relevant explicitation repairs four stories without regression against either control, a 28.57-point gain; each unadjusted one-sided paired $p = 0.0625$. This is an additional positive behavioral observation, not a cross-family confirmation.

Configuration	Recorded	$J_{\text{imp}}/J_{\text{rel}}/J_{\text{irr}}$	u	$T_{\text{imp}}/T_{\text{rel}}/T_{\text{irr}}$	F/S
GPT-6 Luna	98/98	8/12/8	14	8/12/8	0/0
GLM-5.3 route	98/98	6/7/9	10	5/7/8	31/0
GLM-5.3 Flash	98/98	10/14/14	14	10/14/14	0/3
Kimi K3	98/98	14/14/14	14	14/14/14	0/0
Qwen3.8 Max	98/98	11/11/11	14	11/11/11	0/1
Qwen3.8-27B	0/98	—	—	—	—
Gemini-3.8 Flash	98/98	13/14/13	14	13/14/13	0/0
Claude Opus 5	98/98	0/0/0	0	0/0/0	98/0
DeepSeek Flash	98/98	0/0/0	9	0/0/0	63/0
DeepSeek V4 Pro	98/98	0/0/0	8	0/0/0	73/0
DeepSeek Vision Exp	0/98	—	—	—	—

Table 37: Expanded API results on the same 14 stories. Every J, T, and u is a count out of 14. Recorded includes failed scientific inputs; F/S separates format or truncation failures from service failures. Dashes denote no scientific estimate. Kimi retains its failed original interface check. Claude's all-zero strict scores reflect explanatory output; the uniform final-line sensitivity analysis below recovers all 98 answers correct. Configurations and views do not increase the story count.

The completed `deepseek-flash` and `deepseek-v4-pro` configurations produce 63 and 73 unscorable outputs among 98 scientific inputs, respectively, so both have $J_{\text{imp}}/J_{\text{rel}}/J_{\text{irr}} = 0/0/0$ under the original format rule. These counts do not establish uniformly incorrect reasoning. After inspecting the first two invalid records in case-ID order from each configuration, we freeze a post-hoc sensitivity parser: preserve originally accepted answers, otherwise extract only a terminal final line containing a single A/B/C after stripping outer Markdown and an optional Answer/Final-answer prefix. Apply it uniformly to all completed configurations, without another model call. The two DeepSeek J triples become $10/9/12$ and $6/6/7$; unknown correctness becomes 12/14 and 13/14, with five and 14 outputs still unscorable. GPT-6 Luna is unchanged. This format sensitivity analysis does not replace original scoring or support confirmatory testing, and it supplies no consistent DeepSeek explicitation benefit. Claude Opus 5 completes all 98 inputs with explanatory outputs, none accepted by the original answer-only rule. The same frozen final-line parser extracts every answer correctly, yielding $J_{\text{imp}}/J_{\text{rel}}/J_{\text{irr}} = 14/14/14$ and 14/14 unknown judgments. This is a format-sensitive ceiling observation, not evidence of uniformly incorrect reasoning from its strict-score row.

Qwen3.8-27B and the previously registered DeepSeek Vision Exp have no scientific results after interface identity errors. GLM-5.3 closes with $J_{\text{imp}}/J_{\text{rel}}/J_{\text{irr}} = 6/7/9$, 10/14 unknown judgments, and 31 unscorable scientific outputs; the gateway mixes the reported IDs `glm-5.3` and `glm-5.3:dev`, so this is a route-configuration result, not an isolated stable-model estimate. GLM-5.3 Flash closes with $J_{\text{imp}}/J_{\text{rel}}/J_{\text{irr}} = 10/14/14$ and 14/14 unknown judgments, retaining three service failures in the implicit conditions. Its implicit J has missingness bounds $[10, 13]/14$, and relevant explicitation does not outperform the irrelevant addition. Kimi's original interface batch stops after seven successful responses and one midstream upstream error, before any science. A separate prospective continuation admits those seven connections and collects only previously unasked science inputs, retaining the failed check. It completes all 98 inputs correctly, yielding $J_{\text{imp}}/J_{\text{rel}}/J_{\text{irr}} = T_{\text{imp}}/T_{\text{rel}}/T_{\text{irr}} = 14/14/14$; Kimi therefore supplies a ceiling boundary. Qwen3.8 Max, returned as `qwen3.8-2.4t-a95b`, records all 98 inputs with one service failure

in the relevant-explicit condition. Its observed $J_{\mathrm{imp}}/J_{\mathrm{rel}}/J_{\mathrm{irr}} = 11/11/11$, with 14/14 unknown judgments; each relevant-control difference has a missingness interval of $[0, 1]/14$. Gemini-3.8 Flash completes all 98 inputs with $J_{\mathrm{imp}}/J_{\mathrm{rel}}/J_{\mathrm{irr}} = 13/14/13$ and unknown correctness 14/14. Its single repair removes an unwarranted unknown answer; the irrelevant addition does not repair that story. Each contrast has unadjusted $p = 0.5$. All requested configurations are terminal and retained in cohort accounting. The expanded campaign has 882 scientific and 76 interface records, including the original Kimi interface failure; its 22 adjusted p values are all 1. Service-incomplete configurations receive no confirmatory test ($p = 1$). Across the first and expanded batches, 1,385 records are preserved: 1,274 scientific inputs and 111 interface checks. Configurations use the same 14 stories, and repeated GLM interface checks do not add an independent model. An offline replay reconstructs public requests, raw API answers, local token outputs, both statistical families, and the uniform format sensitivity.

Across all configurations, the sensitivity parser additionally recovers one Gemini Pro output without changing its joint counts, and one GLM-5.3 output raising only J_{imp} from six to seven. All other non-DeepSeek/non-Claude counts are unchanged. Service failures remain missing in both analyses. The expanded API campaign reports 670,440 prompt and 241,182 completion tokens across 953 attempts with known usage, plus five attempts without usage; interface costs are included, and actual currency charges are unavailable.

S CAPABILITY-CONDITIONED API EXTENSION

Question and prospective scope. The small natural-language extension above motivates a larger test of configuration-level behavioral C1 on independently generated backgrounds. We fix two strata before collection: short paths of two to four links and long paths of five or six links. Each configuration receives 64 capability backgrounds and 128 further confirmation backgrounds per stratum. The two task families contribute equally, with the intermediate truth value, final claim, and irrelevant implication's truth independently balanced. The generator excludes all 22,848 recorded structural signatures from earlier materials. New capability, confirmation, and interface groups also have disjoint signatures; all public views pass the existing semantic oracle, including opposite-label completions for unknown questions.

The short capability depth counts are 32/16/16, and the long counts are 32/32. Confirmation counts are 48/48/32 and 64/64, respectively. Contexts contain 17, 21, or 25 entities. Their distributions differ between strata: short confirmation has 48/48/32 backgrounds at these entity counts, whereas long confirmation has 64/32/32. We therefore interpret each within-background presentation contrast, without attributing between-stratum differences solely to path length. These remain controlled synthetic tasks; the extension does not replace external natural-language validation.

Fixed configurations and questions. The roster contains GPT-6-Luna, GLM-5.3 and GLM-5.3-Flash, Kimi K3, Qwen3.8-Max and Qwen3.8-27B, Grok 4.7, Gemini 3.8 Flash and 3.1 Pro Preview, and Claude Fable 5 and Opus 5. Settings follow the preceding connected configurations: GPT uses `none` reasoning, Grok uses `low`, and the other submitted fields and returned identities are retained in the protocol. Reported routing names are not independent weight-identity verification; GLM's accepted stable/development routing is reported as such.

Each capability background has five questions: original, changed-premise, and unknown relation judgments, plus explicit decisions in both determined premise states. Each confirmation background has ten: the six crossed determined decisions, the shared unknown decision, and three standalone relation judgments. This gives 640 capability and 2,560 confirmation questions per configuration, or 35,200 planned scientific questions across the fixed roster. Eight separate interface questions per model test identity and terminal response delivery. Admission requires at least four valid terminal responses, regardless of their answer correctness or output format. Models failing the capability criterion still receive the confirmation questions for descriptive diagnosis, subject to service availability and the fixed deadline.

Qualification, scoring, and endpoints. Within each stratum, empirical qualification requires at least 90% accuracy in each of positive relations, negative relations, unknown relations, and explicit decisions, with no missing service outcomes. Original and changed-premise relation questions are bucketed by their reference answer. A configuration must qualify on the independent screen and on the corresponding confirmation measurements. These observed thresholds are not simultaneous certificates of population accuracy.

Before any new model call, the primary output parser accepts a single A/B/C, its one-field JSON or fenced representation, or a terminal answer whose final nonempty line is a single label with optional `Answer`/`Final` prefix and Markdown decoration. This uniform rule prospectively adopts the preceding study's format sensitivity; literal single-letter output remains a separate sensitivity. Truncated or unparseable completed responses are incorrect; service failures remain distinguished from task errors.

The primary endpoint is background-paired $J_{\mathrm{rel}} - J_{\mathrm{imp}}$ and $J_{\mathrm{rel}} - J_{\mathrm{irr}}$ in each stratum. All 11 configurations, two strata, and two controls remain in a fixed 44-test one-sided exact paired-test family with Holm correction. Unavailable, service-incomplete, or unqualified strata receive $p = 1$ in that family. Support for $\mathrm{C1_B}$ requires qualification and a positive, adjusted-significant difference against both controls. There is no minimum effect size governing whether a result is reported. The historical cross-family C1 criterion is unchanged. A separate 44-test auxiliary family applies the same contrasts to $R_b J$, retaining every planned background in the denominator. Pooled strata, task families, and error-transition types are descriptive. We also report T, shared-unknown accuracy, unsupported assertions, and identification ranges for service missingness. Shared unknown questions receive no explicitation treatment, so they cannot identify an effect of explicitation on uncertainty.

Conservative missingness sensitivity. During collection, after observing service errors but before any complete confirmation cohort, we additionally record a worst-case sensitivity. It does not replace the original service-complete primary rule. Capability missingness is counted as incorrect; each of the four category accuracies must still meet the same empirical threshold in both stages. For each presentation contrast, service-missing explicit answers are filled as incorrect and service-missing control answers as correct. This yields the smallest paired improvement compatible with the recorded answers. We apply the same exact test and the same fixed 44-test Holm family to these worst-case vectors. Any surviving directional evidence is reported with this during-collection status and alongside the unchanged primary verdict.

Collection and route handling. Requests use eight workers per model, with shared limits of 64 per host and 16 per credential route, a submitted 4,096-token output limit, and bounded retries only for unanswered transport failures or eligible HTTP errors. Wrong answers and nonempty HTTP-200 responses are not asked again. Raw bytes, attempts, failures, token reports, and planned denominators are retained; reported token totals do not establish equal realized computation or verified currency charges. In the initial collection, new dispatch ended at 10:00 China Standard Time on September 25, 2026, with a noon hard drain deadline. After HTTP 429/503 and read timeouts appeared on the shared host, an operational governor reduced its in-flight concurrency ceiling from 64 to 16 without interrupting issued requests or changing any model question. The separate GPT route was unchanged. Timing and service failures therefore remain part of the execution record rather than equal-budget evidence.

The original Fable collector stopped when one capability response added the namespace prefix `anthropic/`. Before continuation, a separate amendment admitted exactly that alias while preserving its original identity-failure record. The 16 completed capability inputs are retained, and only unasked inputs are continued. Qwen27B's original interface returned upstream error sentinels; a separate finite interface attempt requests the already specified `:free` alias. Neither amendment changes scientific questions, answer scoring, effect thresholds, or statistical family size. The original literal-identity collection and the amended route results are reported separately; no alternate model substitutes for either request.

Completed GPT-6-Luna configuration. All 640 capability and 2,560 confirmation responses are service-complete and parseable. At effort `none`, the short-path confirmation counts are $J_{\mathrm{imp}} = 30$, $J_{\mathrm{rel}} = 123$, and $J_{\mathrm{irr}} = 33$ out of 128; long-path counts are 26, 125, and 25. Relevant explicitation repairs 94 and loses one short-path implicit pair, and repairs 99 with no loss on long paths. The corresponding relevant-minus-implicit differences are 72.66 and 77.34 percentage points; against irrelevant additions they are 70.31 and 78.13 points. These are substantial observed configuration effects, but not screen-qualified C1 evidence.

The independent screen has short-path positive/negative/unknown relation counts of 34/41/43 out of 64, compared with 127/128 explicit decisions. Long-path counts are 22/33/28 and 126/128. Both strata therefore fail the 90% qualification rule before confirmation; confirmation relation measurements also fail it. Unknown decision correctness is only 29/128 and 42/128, respectively. Hence implicit/explicit/irrelevant three-state counts are 5/26/10 on short paths and 8/41/7 on long paths. A large gain on determined decisions coexists with derivation errors and failures to retain uncertainty. All four primary C1 test slots remain at $p = 1$ under the frozen eligibility rule, rather than promoting this configuration because its unqualified paired effect is large. Among the implicit-pair repairs, 56 short-path and 53 long-path backgrounds previously gave the same definite label before and after the decisive premise edit. Another 27 and 35 contained unwarranted unknown answers; 11 in each stratum reversed the labels in the wrong direction. The one short-path regression instead becomes an unchanged definite pair. This decomposition shows that the gain includes repaired premise responsiveness, without converting it into evidence of independent derivation competence.

Initial collection closure and missing outcomes. The initial collection and the two named route amendments terminated with 14,888 scientific question records out of 35,200 planned questions. Of these records, 14,155 contain completed service responses and 733 retain service failures; 20,312 planned questions were not dispatched. These categories are distinct from completed but incorrect or unparseable answers. GPT-6-Luna is the only service-complete configuration. Nine other configurations stopped after the fixed sequence of sustained service failures; Qwen3.8-27B failed both

permitted interface attempts and has no scientific responses. Table 38 retains the full roster and planned denominators.

Request configuration	Capability	Confirmation	Service failures	Unasked
GPT-6-Luna	640	2,560	0	0
GLM-5.3	640	408	64	2,152
GLM-5.3-Flash	640	1,128	265	1,432
Kimi K3	640	320	45	2,240
Qwen3.8-Max	640	400	61	2,160
Qwen3.8-27B	0	0	0	3,200
Grok 4.7	640	664	41	1,896
Gemini 3.8 Flash	640	832	127	1,728
Gemini 3.1 Pro Preview	640	800	38	1,760
Claude Fable 5	640	480	37	2,080
Claude Opus 5	640	896	55	1,664

Table 38: Initial collection coverage, including the named route amendments. Capability and confirmation columns count terminal records, including service failures, against fixed planned counts of 640 and 2,560 per configuration. Service failures count scientific questions without a completed response after their allowed attempts; interface failures are excluded from this column. Fable's single namespace-only identity failure is retained in the original archive and decoded under the explicit alias amendment in this table. No scientific input is counted twice.

Every incomplete configuration has identification ranges spanning both positive and negative effects for each relevant-versus-control contrast in both strata. Thus its observed pair counts cannot establish either an explicitation benefit or its absence. All 44 primary and 44 auxiliary adjusted test slots equal one under the frozen qualification and service-completeness rules. The separate worst-case sensitivity also provides no qualified directional evidence. These missing-data outcomes are not evidence that the affected models lack relation capability. The complete GPT-6-Luna result instead fails qualification on observed relation answers, as reported above.

The retained initial and route-amendment archives contain 15,426 physical requests, including interface checks and permitted transport retries. Reported use totals 15,793,157 input and 2,596,312 completion tokens; 1,180 attempts lack complete usage reports. Two Kimi responses report completion counts above the submitted 4,096-token limit. These totals are observed reports rather than complete billing or matched actual computation. The raw-response replay and full per-stratum tables retain format sensitivity, unknown-case outcomes, repair/regression pairs, and missingness ranges for every planned configuration.

Final service recovery and complete accounting. All collection processes have terminated. Separately recorded recoveries cover the non-Claude routes and the final Claude/Grok continuation. They retain the original inputs, settings, scoring, and 44-test families. The first service-completed answer to each scientific question is final, including wrong, unparseable, or truncated responses. Only service-missing and previously unasked questions are eligible for later dispatch. The initial-collection results above remain a separate historical snapshot.

The five-epoch raw replay verifies 189,493 file bindings and independently recomputes the paired statistics. The final fixed 35,200-question roster contains 29,341 completed responses, 126 terminal service failures, four interrupted requests without terminal results, and 5,729 unasked inputs. The last two categories jointly constitute 5,733 inputs without a terminal record. Of the completed responses, 27,167 are correct and 2,174 are wrong; 90 of the wrong answers are unparseable or truncated under the prospectively fixed parser. Interface checks and retries do not increase these scientific denominators. Table 39 reports every model.

Qualified complete configurations are at or near ceiling. Gemini Pro, Gemini Flash, Qwen Max, and Kimi complete both stages and pass the independent and confirmation capability requirements in both strata. Their implicit and relevant-explicit joint counts range from 127 to 128 out of 128 (Table 40). All 44 adjusted primary and all 44 adjusted auxiliary values equal one. The separate worst-case missingness analysis also supports no positive qualified contrast. Thus the enlarged study

Configuration	Responses	Correct	Wrong	Invalid	Service	No record
GPT-6 Luna	3,200	1,944	1,256	0	0	0
GLM-5.3	1,888	1,879	9	5	15	1,297
GLM-5.3 Flash	2,786	2,647	139	31	47	367
Kimi K3	3,200	3,167	33	20	0	0
Qwen3.8 Max	3,200	3,199	1	0	0	0
Qwen3.8-27B	0	0	0	0	0	3,200
Grok 4.7	3,194	3,193	1	0	6	0
Gemini 3.8 Flash	3,200	3,197	3	0	0	0
Gemini 3.1 Pro Preview	3,200	3,200	0	0	0	0
Claude Fable 5	3,188	3,086	102	33	12	0
Claude Opus 5	2,285	1,655	630	1	46	869
Total	29,341	27,167	2,174	90	126	5,733

Table 39: Final scientific accounting across all authorized service epochs. Each configuration has 3,200 planned questions (640 capability and 2,560 confirmation). Responses equal correct plus wrong; invalid is a subset of wrong. Service denotes a terminal failure without a completed answer. No record includes four dispatched-but-interrupted GLM-5.3 requests; the other 5,729 inputs were unasked. A zero-response row has no measured accuracy. Different native settings and task views preclude a model ranking.

establishes a scope boundary for the qualified local gap, while the large GPT-6 Luna improvement remains an unqualified configuration effect. It does not provide another positive $\mathrm{C1_B}$ replication.

Configuration	Stratum	$J_{\mathrm{imp}}/J_{\mathrm{rel}}/J_{\mathrm{irr}}$	Unknown	$T_{\mathrm{imp}}/T_{\mathrm{rel}}/T_{\mathrm{irr}}$
Gemini 3.1 Pro Preview	Short	128/128/128	128	128/128/128
	Long	128/128/128	128	128/128/128
Gemini 3.8 Flash	Short	128/128/128	128	128/128/128
	Long	128/128/127	127	127/127/126
Qwen3.8 Max	Short	128/128/128	128	128/128/128
	Long	128/128/128	128	128/128/128
Kimi K3	Short	127/127/128	127	126/126/127
	Long	128/128/128	123	123/123/123

Table 40: All service-complete configurations that pass both capability checks. Counts use 128 confirmation backgrounds per stratum. These are the complete set meeting the frozen eligibility conditions; the unqualified complete GPT-6 Luna configuration is reported above.

Terminated branches with missing responses. Fable dispatches all 3,200 questions and passes its independent capability screen. Its 12 missing confirmation responses prevent complete-cohort qualification; observed short/long $J_{\mathrm{imp}}/J_{\mathrm{rel}}/J_{\mathrm{irr}}$ are 116/108/119 and 122/117/118. Even the worst-case qualification sensitivity establishes no positive effect. Opus completes its capability stage but fails qualification on observed answers. Sustained service failures stop its final continuation; 46 confirmation questions retain service failures and 869 remain unasked. Its partial joint counts cannot be interpreted as complete-model estimates. Grok passes both capability measurements but retains six missing decision responses; all observed determined responses are correct. The GLM recovery branches retain the user's explicit stop, including the four interrupted requests, and Qwen27B never passes interface admission. Process termination therefore does not imply complete scientific coverage.

The five collection epochs retain 31,790 physical requests, including interface checks, failed attempts, and retries. Available usage reports sum to 32,747,559 prompt and 6,848,173 completion tokens; 2,247 attempts lack complete usage. Cached-token counts are already included in reported prompt counts. These are recorded service totals, not verified billing or matched computation. The final replay, question ledger, and initial snapshot remain separately identifiable; no failed or superseded physical attempt is erased.

T DEVELOPMENT PROBES, EARLIER PILOT, AND EXECUTION LIMITS

Localization design. After observing poor standalone relation performance, we prepared a separate development probe on 24 backgrounds, selected without using their outcomes: two per family, path length, and original relation truth value. It checks every path edge with true and false direct-relation questions, includes a missing-edge control, and asks the same composed relation after retaining only the necessary path premises. This gives 240 new questions and reuses 168 previously recorded full-context answers. Premise selection uses the evaluator's graph and is an oracle diagnostic, not an automatic retrieval method. The probe can distinguish behavioral patterns associated with reading, composition, and full-context integration; it is not independent validation and does not change C1 eligibility.

Failure is observable before multi-hop composition. The public rules explicitly state "Each person has exactly one mother" and, for delegated permission, "Each case has either one delegation target or one terminal permission status." They preserve missing links as unspecified; the latter task explicitly forbids default grant or denial. False direct-edge queries contradict a recorded unique target, rather than being false merely because an edge is absent. Both families use open-world uncertainty constrained by the stated functionality rules. Failure to apply those rules and context interference remain competing explanations; the probe does not isolate either. All 240 new questions in the 24-background development diagnostic were service-complete and audited. Luna correctly affirms all 72 directly stated edges and answers all 24 missing-edge controls correctly, but refutes only 17/72 false edges, returning unknown on the other 55. Refutation is 0/36 in maternal ancestry and 17/36 in delegated permission. No background has every positive and negative direct-edge answer correct. Thus a multi-hop failure alone cannot explain this diagnostic pattern. On the same backgrounds, retaining only the necessary path premises yields 15/24 original and 15/24 changed-relation answers correct, compared with 10/24 each in the previously recorded full context. Unknown-relation performance changes from 17/24 to 16/24. Context reduction therefore helps some outcomes without uniformly solving the inference task. The 72 edge questions are correlated within backgrounds and are not treated as 72 independent experimental units.

A subsequent adaptive diagnostic uses 12 development backgrounds, one per family, depth, and relation-truth stratum. We explicitly distinguish the entity names, then either restate the local uniqueness rule or retain only one relevant ordinary record. All 108 new answers were service-complete and audited. Direct affirmation and missing-edge accuracy remain 12/12 in every version. Direct refutation is 5/12 in the reused original answers, 7/12 after name clarification, 6/12 with the added local uniqueness wording, and 11/12 with minimal context. Relative to name clarification, context reduction has four gains and no losses; the local wording has no gains and one loss. These small, adaptively motivated comparisons are descriptive. They suggest that the difficulty persists beyond name clarity and depends on context, but do not identify whether or where a uniqueness rule is represented internally. Oracle premise selection is not an automatic repair.

The earlier pilot has both replication and a validity limitation. Luna's earlier R2 validation contains 512 new backgrounds and 4,608 completed, scorable responses. Original, explicit, and irrelevant conditions yield 89/512, 504/512, and 65/512 correct answers. Original and changed relation questions yield 259/512 and 265/512. Thus the earlier behavioral difference replicates on new backgrounds, but again does not satisfy the capability prerequisite. The R2 unknown condition used a dedicated registration sentence that exposed its label; we exclude it from evidence-boundary conclusions. The original and explicitation contrasts did not contain that sentence, but we retain the material defect and do not pool R2 with the corrected study.

The same pilot also identifies a performance boundary. In R2 development, Sol is correct on 62/64 original problems and all 63 service-complete explicit problems, with one explicit service failure. The earlier pilot configuration Astra (requested `gpt-6-astra`, not a model in the later crossed comparison) is correct on 64/64 original and 64/64 explicit problems; all its service-complete development responses are correct. These configurations provide little evidence of a positive explicitation effect in this task. Their native low settings differ from Luna's none setting. A ceiling here does not establish absence of evidence-use failures on other tasks.

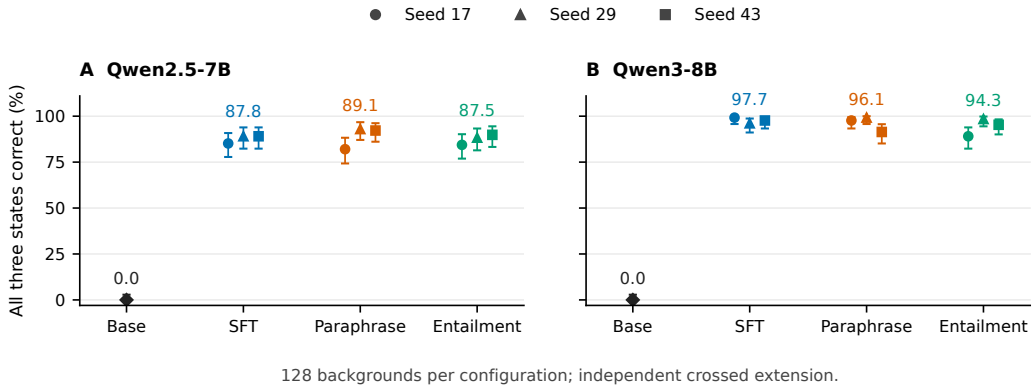


Figure 13: Local crossed extension, 128 new backgrounds per configuration. Markers, intervals, and mean labels follow Figure 14. These inputs and generations are separate from the original 512-background validation; all-three scores use the implicit presentation.

Execution interruptions and recovery. A host-wide interruption left several in-flight API outcomes unknown and required replay from durable local checkpoints. Replayed training updates retain distinct execution records. Further local runs encountered native crashes and checkpoint object errors. An independently checked scheduling amendment retains non-reentrant gradient checkpointing while executing backward callbacks on the main thread; affected executions are recorded separately. The root cause remains unresolved, and short runtime checks do not establish numerical identity or host reliability. A subsequent server reboot left some recent adapter files empty despite surviving checksum receipts. All seven already selected completed-run adapters remained valid. One incomplete Qwen2.5 seed-29 run had no valid matching weight–optimizer recovery point and was replayed from its original seed, retaining the 472 earlier update records. Later checkpoint writing explicitly syncs adapter files and directories and retains a full optimizer/RNG snapshot at each scheduled checkpoint. This persistence amendment changes neither the training objective nor checkpoint selection and does not establish that the underlying host fault has been resolved. API continuation dispatches only never-attempted questions; incomplete raw captures and failed identity controls remain missing rather than being replaced by new answers. A later native-client envelope change and a raw-response integrity failure curtailed two earlier-pilot branches. Such missingness may be nonrandom; point estimates on surviving responses are insufficient for a confirmatory claim.

U Frozen Local Semantic-Grid Extension

After the Luna grid closed, we fixed a local extension using the same 128 independent backgrounds and seven views. Both local models contribute their base configuration and all twelve selected adapters: four existing arms and three fixed seeds, for 26 configurations and 23,296 new generations. No new development cases, training objectives, or checkpoint-selection rules are introduced. All selected adapters for a model are bound before its first new call. Existing training and the original 512-background validation have resource priority; the extension waits for their completion.

The primary endpoint is correct response at both premise states. A single Holm family contains ten hypotheses: explicit versus implicit and irrelevant presentation for each base model; and implicit-pair comparisons of SFT versus base, entailment versus SFT, and entailment versus paraphrase for each model. Each training comparison is the conjunction of the three seed-specific paired exact tests, using the largest p-value. Seeds do not multiply the background sample size. Missing outcomes are assigned against the claimed improvement. No new minimum-effect threshold is imposed, and the original C2 criteria are unchanged. The eighteen corresponding seed-specific new-assertion bounds use one-sided Clopper–Pearson coverage with $\alpha = 0.05/18$; a sample of 128 cannot establish a small risk bound merely by observing zero new assertions. Mode transitions, family summaries, the supervision ablation, and paraphrase versus SFT are secondary descriptive analyses. This extension was specified after Luna's results and before local outcomes on these inputs; it is not a retrospective claim of registration before every experiment.

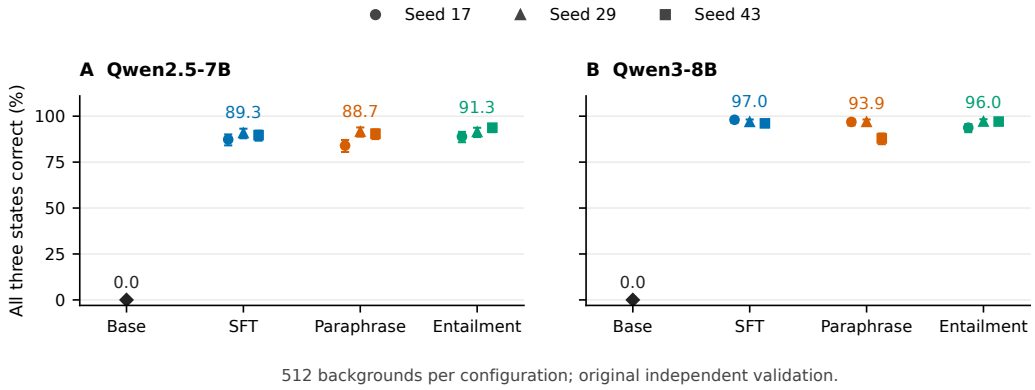


Figure 14: Five-view supervised adaptation on 512 held-out backgrounds per configuration. Each marker and exact 95% interval belongs to one fixed training seed; base models are unadapted. For adapted models, labels show the three-seed mean in percent. All-three correctness requires original, changed-premise, and unknown answers to be correct on the same background. Seeds do not multiply the sample size.

Complete local extension. All 23,296 generations close with verified public input tokens and decoded outputs. Table 41 lists all fixed seeds. Qwen2.5's pretrained relevant-implication comparison improves over implicit by 29.69 points and over irrelevant by 28.12 points (Holm-adjusted $p = 2.91 \times 10^{-11}$ and 1.02×10^{-10}). Qwen3's pretrained model does not improve under explicitation. SFT improves implicit both-state correctness in all three seeds of both models, with adjusted conjunction $p = 5.42 \times 10^{-35}$ and 2.35×10^{-37}. The four entailment-over-baseline comparisons all have adjusted $p = 1$. Mean entailment-minus-SFT point differences are -0.26 and -1.82 percentage points for Qwen2.5 and Qwen3. This later joint endpoint and sample differ from the original 512-background original-answer comparison.

Model	Arm	J_{imp}	J_{rel}	J_{irr}	Unknown	T_{imp}
Qwen2.5	Base	0	38	2	49	0
	SFT	117/121/119	128/128/127	107/119/90	118/120/121	109/114/114
	Entailment	114/121/121	120/127/125	106/122/115	120/120/122	108/113/115
	Paraphrase	113/125/124	127/128/128	86/119/108	114/121/122	105/119/118
	No CF/U	88/94/71	127/128/126	70/66/22	0/0/0	0/0/0
Qwen3	Base	1	0	0	5	0
	SFT	127/127/126	128/128/128	123/125/125	128/124/126	127/123/125
	Entailment	120/128/125	125/128/128	114/127/123	121/126/124	114/126/122
	Paraphrase	128/127/123	128/128/128	112/117/115	125/128/120	125/127/117
	No CF/U	118/108/108	127/127/128	112/83/96	0/1/1	0/1/1

Table 41: Complete local seven-condition extension, counts out of 128 backgrounds. J_{imp}, J_{rel}, J_{irr} are implicit, explicit, and irrelevant pair counts; T_{imp} is implicit all-three correctness. Slash-separated values are seeds 17, 29, and 43, not additional independent samples. SFT denotes five-view SFT; No CF/U removes changed and unknown supervision.

V Local Validation and Development Outcomes

Closed local validation branches. All twelve adapters of each local model were selected before that model's first validation dispatch. Table 43 reports the base model and all four arms at each of three seeds. Each contributes 4,608 strictly audited responses; none is missing, truncated, or invalid. These are independent R3 backgrounds, not the later seven-view extension. Both models' original validation branches are complete, as is the separately reported local crossed study.

Training arm	Original	Both ends	Unknown	All three
Qwen2.5-7B-Instruct				
Pretrained model	29.49	0.20	41.60	0.00
SFT	94.79	92.84	95.18	89.26
Entailment consistency	96.29	94.79	95.57	91.34
Paraphrase consistency	95.64	93.03	94.27	88.67
No CF/U	79.49	63.80	0.00	0.00
Qwen3-8B				
Pretrained model	48.83	0.98	1.37	0.00
SFT	99.41	98.50	98.31	97.01
Entailment consistency	98.24	97.20	98.57	95.96
Paraphrase consistency	98.44	96.81	96.22	93.88
No CF/U	95.38	90.62	0.26	0.26

Table 42: Independent local validation, accuracy in percent. Adapted rows average three fixed seeds evaluated on the same 512 new backgrounds; seeds do not multiply the independent sample size. "All three" requires correct original, changed-premise, and unknown answers on one background.

Qwen2.5 configuration	Original	Changed	Both	Unknown	All three
Base	151	155	1	213	0
SFT (seed 17)	481	488	469	482	447
Entailment (seed 17)	485	483	472	488	455
Paraphrase (seed 17)	483	478	458	475	430
No CF/U (seed 17)	426	435	352	0	0
SFT (seed 29)	489	491	480	496	465
Entailment (seed 29)	497	503	494	481	468
Paraphrase (seed 29)	494	496	488	491	469
No CF/U (seed 29)	430	441	362	0	0
SFT (seed 43)	486	494	477	484	459
Entailment (seed 43)	497	495	490	499	480
Paraphrase (seed 43)	492	498	483	482	463
No CF/U (seed 43)	365	411	266	0	0

Table 43: Correct outcomes on 512 independent validation backgrounds. "Both" requires original and changed-premise correctness; "all three" also requires a correct unknown answer. Checkpoints are selected on development before validation. Additional seeds do not increase the background denominator.

At seed 17, SFT's original benefit over base is 64.45 percentage points, with an individual conservative 95% paired interval $[57.97, 69.93]$. Unsupported assertions on unknown inputs fall from 299 to 30, but seven are new; their individual one-sided 95% upper bound is 2.55%. This secondary base comparison is outside the original twelve-bound C2 risk family. Its both-state gains are 468 with no losses. Response-mode analysis repairs 259 of 285 unchanged definite pairs and 208 of 224 pairs containing unknown answers despite determined evidence.

At seed 17, entailment versus SFT has 19 original-answer gains and 15 losses; versus paraphrase it has 20 gains and 18 losses. The unadjusted component p values already exceed 0.05, so taking a maximum across seeds and then Holm correction cannot make these comparisons pass. The three-state joint benefits contain 39 gains/31 losses against SFT and 55 gains/30 losses against paraphrase. Unknown outcomes contain 21 repairs/15 new assertions and 28 repairs/15 new assertions respectively. Both original $\alpha = 0.05/12$ upper bounds are 5.50%. Entailment also loses five net changed-premise answers relative to SFT. Thus some joint outcomes improve while the original C2 requirements fail; joint scores do not replace the prespecified original-answer endpoint.

At seed 29, SFT has 343 original-answer gains and five losses over base; unknown answers have 290 repairs and seven new assertions. Entailment versus SFT has 16 gains/eight losses on original answers, 16/four on changed answers, 23/nine on both-state correctness, and eight/23 on unknown answers. Its all-three total rises by only three, with 29 gains and 26 losses. The original-answer

component p is 0.0758 and the original simultaneous risk upper bound is 7.48%. Thus increased correctness on determined questions coexists with more failures on insufficient evidence in this second seed.

The first validation ablation, which removes changed-premise and unknown supervision together, answers all 512 unknown questions with definite labels (289 A and 223 B), despite 426 correct original answers and 352 correct pairs. Relative to full entailment training at the same seed, it has 19 original gains and 78 losses, 14 both-state gains and 134 losses, and 488 new unsupported assertions. This extends the development failure below to new backgrounds, while retaining the combined ablation's label-frequency and loss-weight confounds.

At seed 43, entailment versus SFT has 20 original gains/nine losses, 11 changed-premise gains/ten losses, and 24 unknown repairs/nine new assertions. All-three correctness gains 38 and loses 17 backgrounds. Against paraphrase, it has 13 original gains/eight losses, nine changed gains/twelve losses, and 24 unknown repairs/seven new assertions; all-three correctness gains 32 and loses 15. The original risk upper bounds are 3.92% and 3.37%, respectively. This seed improves unknown accuracy relative to both baselines, whereas seed 29 reduces it; the method does not uniformly trade unknown accuracy for determined accuracy.

Across the three seeds, original point gains average 1.50 percentage points against SFT and 0.65 against paraphrase. Each seed uses the same 512 backgrounds; these averages are descriptive and receive no independent-sample interval. The corresponding three-seed conjunction values are 0.304 and 0.436. The six original risk bounds range from 3.37% to 7.72%, so all exceed 2%. Changed-premise differences are $-5, +12, +1$ against SFT and $+5, +7, -3$ against paraphrase. Thus the fixed Qwen2.5 C2 criteria are not met, while the exact four-comparison Holm-adjusted p values are all 1 in the complete two-model analysis.

The headroom calculation is a post-hoc arithmetic diagnostic. If a baseline's mean original accuracy is $\bar{a}$, no candidate's mean gain can exceed $1 - \bar{a}$. SFT leaves 5.21 percentage points (an accuracy difference of $80/1536$), and paraphrase leaves 4.36 points ($67/1536$). The denominator averages three sets of 512 outcomes, not 1,536 independent backgrounds. The latter makes a five-point gain impossible even for a perfect candidate against the observed baseline. This design limit does not turn small gains into confirmed superiority or waive the risk bounds.

The second combined ablation yields 430/512 original, 441/512 changed, 362/512 both-state, and 0/512 unknown correctness; its unknown answers are 250 A and 262 B. The third yields 365/512, 411/512, 266/512, and 0/512, respectively, with 248 A and 264 B. Relative to full entailment at seed 43, the ablation has 11 original gains/143 losses, ten both-state gains/234 losses, and 499 new unsupported assertions. All three ablation seeds therefore retain determinate correctness while failing every unknown question on these new backgrounds.

Complete Qwen3 validation. Table 44 retains all thirteen configurations. Compared with SFT, entailment changes the original count by $-13, -5, 0$ across seeds; compared with paraphrase, by $-11, 0, +8$. The mean differences are -1.17 and -0.20 percentage points. Unknown-count differences against SFT are $-7, +6, +5$, so a small positive unknown average coexists with lower original and three-state joint means. New unsupported assertions against SFT number 9/6/5; against paraphrase, 10/6/3. Their original simultaneous upper bounds are 3.92/3.08/2.79% and 4.19/3.08/2.17%, all above 2%. Even perfect original answers could improve SFT's observed mean by only 0.59 points and paraphrase's by 1.56, below the original five-point criterion. These are complete original-criterion failures, not missing-seed results.

What the negative C2 result explains. Three distinct questions should not be conflated. First, supervision repairs the base models: the large SFT/base improvements are reproducible on new backgrounds. Second, the added consistency term does not meet the frozen superiority-and-risk conjunction. Qwen2.5's small average gains coexist with seed-dependent changes in uncertainty, while Qwen3's mean original-answer effects are negative; every simultaneous new-assertion bound exceeds 2%. Third, the observed strong baselines leave insufficient headroom for several of the original five-point comparisons. This last arithmetic restriction explains why those comparisons could not pass even with perfect candidate accuracy, but does not explain the observed regressions or establish that the method would succeed under a different endpoint.

There is also a distinction between the objective's information and its optimization effects. The matched SFT control already receives every supervised view and its reference label. The consistency term supplies a coupling penalty, rather than additional labeled evidence; the supervision bound in Proposition 2 constrains same-label disagreement as training error vanishes. It does not prove that the penalty is useless at finite training time, nor predict its generalization benefit. The measured gradient conflicts are compatible with competing updates but do not identify the cause of C2's failure. The later eight-view study tests a different coupling and endpoint: its mixed development seeds and negative FOLIO mean do not rescue the original claim. Thus the supported result is supervised behavioral repair, with no established stable gain attributable specifically to the added consistency objective.

Qwen3 configuration	Original	Changed	Both	Unknown	All three
Base	250	251	5	7	0
SFT (seed 17)	509	506	505	509	502
Entailment (seed 17)	496	502	489	502	480
Paraphrase (seed 17)	507	505	501	503	496
No CF/U (seed 17)	496	499	483	0	0
SFT (seed 29)	510	508	506	500	496
Entailment (seed 29)	505	504	501	506	497
Paraphrase (seed 29)	505	505	500	506	496
No CF/U (seed 29)	483	483	455	2	2
SFT (seed 43)	508	504	502	501	492
Entailment (seed 43)	508	507	503	506	497
Paraphrase (seed 43)	500	498	486	469	450
No CF/U (seed 43)	486	479	454	2	2

Table 44: Qwen3 correct outcomes on the same 512 independent backgrounds. All 59,904 answers are audited, with no missing, truncated, or invalid output. The joint-supervision ablation leaves 0/2/2 unknown answers correct.

Complete development results. Table 45 retains the paired development outcomes for completed configurations. Each row uses the same 64 backgrounds; additional training seeds do not increase this denominator.

Detailed matched development outcomes. For Qwen2.5 seed 17, SFT changes original accuracy from 19/64 to 59/64, and both-state correctness from 0/64 to 59/64. Entailment consistency reaches 62/64 and 61/64 respectively, but paraphrase consistency reaches 63/64 and 62/64. Relative to SFT, entailment consistency gains five and loses three backgrounds on all-three correctness, for a net gain of two. All three trained Qwen2.5 configurations have 61/64 unknown answers correct; equal totals do not exclude new unsupported assertions on different instances. Qwen3 SFT reaches 64/64 on all three outcomes, whereas entailment consistency yields 63/64 original, 62/64 changed, and 63/64 unknown accuracy. Qwen3 paraphrase consistency also reaches 64/64 on all three main development outcomes, while answering only 59/64 irrelevant-addition cases correctly. For Qwen3 seed 29, SFT and entailment reach 64/64 on all three main conditions, versus paraphrase's 63/64 original, 64/64 changed, and 62/64 unknown accuracy. Their nine-view totals are 575, 570, and 566 out of 576, respectively. Both SFT seeds score 575/576; the sole error moves from a changed-relation question for seed 17 to a paraphrase for seed 29. The third Qwen3 SFT seed also scores 64/64 on all three main conditions, but 569/576 across the nine views: its seven remaining errors concern one paraphrase and six standalone relation questions. Qwen3 entailment at seed 43 scores 64/64 original, 63/64 changed, and 64/64 unknown, with one changed-premise regression relative to SFT. Its nine-view total is nevertheless higher, 571/576 versus 569/576. The aggregate and target-decision readouts describe different scopes and are both retained.

Further Qwen2.5 development comparisons. For seed 29, entailment consistency reaches 64/64 original, 63/64 changed, and 63/64 unknown accuracy, versus SFT's 63/64, 64/64, and 62/64. Both have 63 correct determined pairs, with one gain and one loss between them. On unknown cases, entailment repairs two SFT errors but introduces one new unsupported assertion. All-three correctness is 62/64 versus 61/64. Thus the second seed again combines small gains with instance-level

Configuration	Original	Changed	Both	Unknown	All three
Luna	12	8	3	50	3
Sol	64	62	62	44	43
DeepSeek Flash	29	33	6	11	0
DeepSeek Pro	18	17	1	36	0
Qwen2.5 base	19	22	0	28	0
Qwen3 base	32	33	1	0	0
Qwen2.5 SFT	59	63	59	61	56
Qwen2.5 paraphrase	63	63	62	61	60
Qwen2.5 entailment	62	62	61	61	58
Qwen2.5 No CF/U	50	57	44	0	0
Qwen3 SFT	64	64	64	64	64
Qwen3 paraphrase	64	64	64	64	64
Qwen3 entailment	63	62	62	63	61
Qwen3 No CF/U	63	62	61	0	0
Qwen2.5 SFT (seed 29)	63	64	63	62	61
Qwen2.5 paraphrase (seed 29)	63	64	63	63	62
Qwen2.5 entailment (seed 29)	64	63	63	63	62
Qwen2.5 No CF/U (seed 29)	51	57	44	0	0
Qwen3 SFT (seed 29)	64	64	64	64	64
Qwen3 paraphrase (seed 29)	63	64	63	62	62
Qwen3 entailment (seed 29)	64	64	64	64	64
Qwen3 No CF/U (seed 29)	57	58	51	1	1
Qwen2.5 SFT (seed 43)	63	64	63	61	61
Qwen2.5 paraphrase (seed 43)	62	63	61	60	59
Qwen2.5 entailment (seed 43)	62	64	62	63	61
Qwen2.5 No CF/U (seed 43)	51	51	38	0	0
Qwen3 SFT (seed 43)	64	64	64	64	64
Qwen3 entailment (seed 43)	64	63	63	64	63
Qwen3 paraphrase (seed 43)	64	63	63	63	62
Qwen3 No CF/U (seed 43)	57	58	51	1	1

Table 45: Correct answers or correct joint outcomes out of 64, all on development data. Trained rows use the selected checkpoint and seed 17 unless otherwise marked; additional seeds are further development runs, not independent method validation. "Both" requires original and changed-premise correctness; "all three" also requires the unknown answer. Every included configuration has 576 (64 backgrounds $\times$ 9 conditions) strictly replayed responses across the original nine conditions. The analysis was motivated by observed development errors and is not confirmatory.

regressions. The same-seed paraphrase baseline reaches 63/64, 64/64, and 63/64 on those conditions, and matches entailment's 62/64 all-three correctness. That equal joint total contains two gains and two losses between the methods; their equal unknown accuracy likewise hides one gain and one loss. The third SFT seed also has 61/64 all-three correctness, matching seed 29's total but with three gains and three losses on different backgrounds. At seed 43, entailment scores 62/64 original, 64/64 changed, and 63/64 unknown, versus SFT's 63/64, 64/64, and 61/64. Original answers have one gain and two losses; unknown answers have two gains and no losses. Both have 61/64 all-three correctness, with two gains and two losses between them. The third paraphrase seed yields 62/64 original, 63/64 changed, and 60/64 unknown accuracy, with 59/64 all-three correctness. Relative to SFT, that joint total contains two gains and four losses; relative to entailment, it contains one gain and three losses. Its unknown accuracy is three cases below entailment, with no gains and three losses. These comparisons use the same development backgrounds and do not establish held-out method superiority.

Combined supervision ablation. Without changed-premise and unknown supervision, Qwen2.5 seeds 17 and 29 reach 50/64 and 51/64 original accuracy, respectively, with both scoring 57/64 changed and 0/64 unknown. Seed 43 reaches 51/64 original, 51/64 changed, and 0/64 unknown, with 38/64 both-state correctness. Its equal original total versus seed 29 hides eight gains and eight losses; the lower joint score contains ten gains and sixteen losses. Qwen3 seed 17 reaches 63/64, 62/64, and

0/64; seed 29 reaches 57/64, 58/64, and 1/64, with 51/64 both-state correctness. Relative to its seed-17 ablation, the latter joint score has one gain and eleven losses. The single correct unknown answer shows that removing C supervision does not make every C response impossible, although unknown-case performance remains poor. The ablation changes training answer-label counts from 4,096 A, 4,096 B, and 2,048 C to 3,072 A, 3,072 B, and zero C. Each retained view's cross-entropy weight also changes from $1/5$ to $1/3$. These combined changes cannot isolate a counterfactual-specific mechanism. The separately frozen local semantic-grid extension evaluates all existing arms and seeds after the original validation queue; its complete results are reported in Appendix U.

Recorded optimization traces. To examine whether the consistency penalty changed optimization, we analyze all nine runs belonging to the three completed matched SFT/entailment/paraphrase triples available at analysis time. Table 46 uses fixed windows of logical updates, not windows selected by performance. Each entry averages stochastic, pre-update training minibatches; these are training trajectories rather than losses on a common fixed evaluation set. The same seed uses the same background order across arms, and their recorded first-update component losses agree. Later windows contain different backgrounds from earlier windows.

Model / seed	Arm	Early relation JS	Late relation JS	Late answer CE
Qwen2.5 / 17	SFT	0.1993	0.0291	0.0693
	Entailment	0.0446	0.0035	0.0663
	Paraphrase	0.2001	0.0290	0.0591
Qwen2.5 / 29	SFT	0.1463	0.0086	0.0310
	Entailment	0.0287	0.0104	0.0350
	Paraphrase	0.1483	0.0121	0.0351
Qwen3 / 17	SFT	0.1521	0.0016	0.0126
	Entailment	0.0301	0.0036	0.0194
	Paraphrase	0.1421	0.0050	0.0209

Table 46: Descriptive training losses from the completed matched triples. Early and late windows are updates 1–64 and 449–512 respectively, with four backgrounds per update. Relation JS is measured on implicit–explicit four-category distributions; answer CE averages the five supervised views. These are not independent validation metrics or checkpoint-selection scores.

The entailment arm has lower early relation disagreement in all three comparisons. SFT also reaches much smaller late disagreement than early disagreement, and its late value is below entailment's for Qwen2.5 seed 29 and Qwen3 seed 17. Thus the added objective has an observable early effect, while low disagreement can also arise under shared-label supervision. The traces alone do not establish a cause of API-model failures or a generalization benefit from minimizing this surrogate.

Recovery records determine the logical trajectory: at each restart we retain the recorded prefix through its restored step and discard later steps only from the reconstructed path. All raw records remain preserved. The nine logical 512-update trajectories use 4,608 updates; the 648 additional physical update records are replayed work, not extra examples or seeds. An independent read verifies the selected paths and recomputes all 72 window summaries. This analysis is post hoc and does not change the training or validation protocols.

Development learning curves. Figure 15 uses all eight recorded checkpoint-selection evaluations for the same nine-run cohort. Each checkpoint evaluates the same 64 development backgrounds on the original, changed-premise, and unknown conditions. These fixed inputs differ from the training minibatches in Table 46, but they selected checkpoints and do not provide independent validation. We replay all 13,824 saved answers against the reference labels, check their input token sequences against the fully audited selected-checkpoint runs, and reproduce the frozen checkpoint-selection rule. Seven evaluations from the earlier, abandoned Qwen2.5 seed-29 SFT execution remain preserved but do not contribute points to the retained trajectory.

At step 64, Qwen3 entailment has 2/64 correct original–changed pairs versus SFT's 0/64, while unknown accuracy is 0/64 versus 60/64. For Qwen2.5 seed 17 at step 256, entailment has more correct pairs than SFT (49/64 versus 35/64) but fewer correct unknown answers (47/64 versus 56/64). The full trajectories retain both gains and reversals: relative development outcomes vary with the config-

uration, checkpoint, and outcome. Lower early training disagreement therefore does not establish uniformly faster learning of reliable decisions.

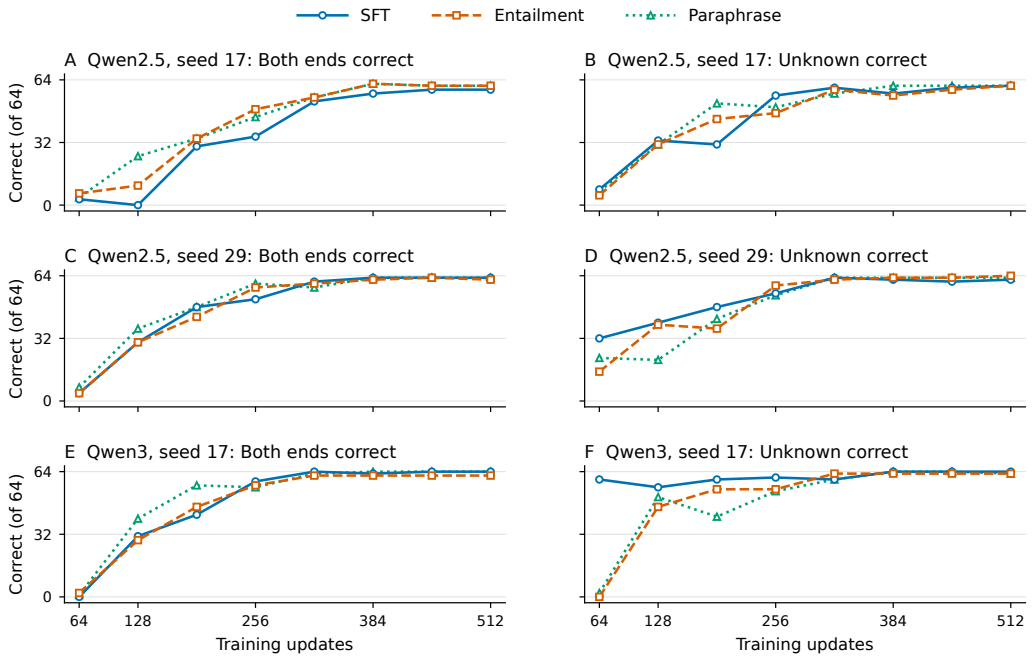


Figure 15: Development learning curves at all eight saved checkpoints in the fixed nine-run cohort. Left: original and changed-premise questions both correct. Right: an evidence-insufficient question correctly answered C. Every point uses the same 64 checkpoint-selection backgrounds; within a seed, training arms share the background order. Lines connect recorded counts. Updates follow the retained logical trajectory, not wall-clock cost. These descriptive curves and their repeated evaluations do not constitute independent validation or additional independent samples.

W Additional Marginal and Capability Results

A large behavioral difference survives the material correction. On all 64 corrected development backgrounds, Luna answers 12 original problems correctly, compared with 62 after relevant explicitation and 10 after an irrelevant implication. The differences are 78.13 and 81.25 percentage points. All 576 responses across the nine conditions were service-complete and scorable. However, original, counterfactual, and unknown relation questions have only 24, 29, and 44 correct answers. This branch fails the prespecified capability condition. It supports a large representation-dependent behavioral difference, while leaving relation derivation as an important explanation.

Different errors can hide behind marginal accuracy. Appendix Table 45 reports paired and unknown outcomes on the same closed development backgrounds. Qwen3 base has 63 unchanged definite pairs and one correct pair. Its pair count equals the lower bound implied by its 32/64 and 33/64 marginal accuracies, so the low joint score alone adds no new statistical fact. The unchanged answers specify the observed pattern. Luna instead has 56 pairs containing an unknown answer to determined evidence; four unchanged definite pairs, three correct pairs, and one reversed but incorrect pair account for the remainder. Sol has 62 correct pairs, one unchanged definite pair, and one pair containing unknown, but its unknown-case answers include 20 unsupported definite assertions. These descriptive patterns motivate the independent crossed intervention, rather than a universal ranking or a shared internal-failure explanation.

Corrected independent backgrounds retain the difference. Luna's corrected validation branch accounts for all 4,608 planned questions: 4,598 outcomes pass the unchanged strict audit, six

have incomplete raw captures, and four were interrupted in flight. Missing answers are not reissued or scored from unaudited records. On the fixed denominator of 512 backgrounds, the original condition has 84 verified correct answers and one missing outcome; relevant explicitation has 501 correct answers and no missing outcome; the irrelevant control has 66 correct answers and two missing outcomes. The corresponding benefit identification intervals are $[81.25, 81.45]$ and $[84.57, 84.96]$ percentage points. Worst-case one-sided exact paired tests give $p = 5.91 \times 10^{-126}$ and $p = 4.91 \times 10^{-129}$; the conservative 95% interval for the original-input benefit is $[76.21, 85.15]$ percentage points. These are branch-level contrasts, not a completed cross-model multiplicity analysis. The development capability condition already failed; C1 remains unestablished despite the large difference.

The following tables retain the earlier capability and marginal-accuracy results. They are separate from the new 128-background crossed study.

Model	Original	Explicit	Irrelevant	Rel. base	Rel. changed	Rel. unknown
Luna	12	62	10	24	29	44
Qwen2.5 base	19	38	17	21	19	30

Table 47: Correct answers out of 64 in closed corrected-development branches. Each model has 576 service-complete responses across all nine conditions. These are development results; neither row passes the capability condition. Luna uses its native none setting, and Qwen2.5 uses the specified local greedy configuration. No equal-compute or model-ranking claim is made.

Condition	Verified correct	Missing	Accuracy interval (%)
Original	84	1	16.41–16.60
Relevant explicitation	501	0	97.85–97.85
Irrelevant implication	66	2	12.89–13.28
Paraphrase	79	0	15.43–15.43
Changed premise	101	0	19.73–19.73
Insufficient evidence	451	4	88.09–88.87
Relation, original	228	1	44.53–44.73
Relation, changed	240	1	46.88–47.07
Relation, unknown	352	1	68.75–68.95

Table 48: Corrected Luna validation, with 512 planned backgrounds per condition. Accuracy identification intervals assign each missing outcome either incorrect or correct; they are not confidence intervals. The ten missing outcomes remain in the denominators.

X VERBATIM INFERENCE PROMPTS

The strings below reconstruct the messages used in the automatic-repair development comparison. We checked all 960 saved scientific request bodies for its two completed models against these templates. The local inference controls reuse the direct or deliberation messages, adding only the model's chat template and the previously reported decoding settings. Visual line wrapping below does not insert characters; the supplied text files and JSON retain exact newlines. The only placeholders are `public_text`, the complete original public problem, and `generated_note`, the prior model output. Neither is replaced by a reference answer or oracle proof.

Common direct/decision system message S.

```
Judge the final statement using only the supplied facts and rules. A
↪ means the statement is supported in every situation consistent with
↪ the facts. B means it is false in every such situation. C means both
↪ truth and falsity remain possible. Missing information is not
↪ automatically false. Return exactly one letter: A, B, or C. Do not
↪ output an explanation or use tools.
```

```
In resource-total questions, the starting amount plus the specified
↪  single conditional contribution is the complete definition of the
↪  final total. There are no other contributions or withdrawals.
Label examples (these are examples, not facts for the current problem):
Facts: a value is exactly 2. Statement: the value is 2. Answer: A.
Facts: a value is exactly 2. Statement: the value is 3. Answer: B.
Facts: a value is either 2 or 3, with no further information. Statement:
↪  the value is 2. Answer: C.
For the actual problem below, return exactly one letter A, B, or C. Do
↪  not output an explanation or use tools.
```

Analysis system prefix S_a. The explanatory stages use the following complete prefix, including its terminal newline. Concatenate two further newline characters and either the deliberation suffix D or relation suffix R shown below.

```
Judge the final statement using only the supplied facts and rules. A
↪  means the statement is supported in every situation consistent with
↪  the facts. B means it is false in every such situation. C means both
↪  truth and falsity remain possible. Missing information is not
↪  automatically false. Do not use tools.

In resource-total questions, the starting amount plus the specified
↪  single conditional contribution is the complete definition of the
↪  final total. There are no other contributions or withdrawals.
Label examples (these are examples, not facts for the current problem):
Facts: a value is exactly 2. Statement: the value is 2. Answer: A.
Facts: a value is exactly 2. Statement: the value is 3. Answer: B.
Facts: a value is either 2 or 3, with no further information. Statement:
↪  the value is 2. Answer: C.
```

Deliberation suffix D.

```
Reason through the supplied problem using only its stated facts and
↪  rules.
Check whether the evidence supports, refutes, or leaves the final
↪  statement undetermined.
Give a concise working analysis, then end with a separate line exactly
↪  FINAL: A, FINAL: B, or FINAL: C.
Do not assume that any particular answer is wrong. Do not use tools.
```

Relation-note suffix R.

```
Identify the condition that controls the resource contribution in the
↪  supplied problem.
Derive that condition from the original records before considering the
↪  final numerical statement.
Trace the relevant links in order, keeping entity identities and
↪  direction explicit.
If the chain reaches a different stated endpoint, use the stated
↪  uniqueness and distinct-name rules.
If a necessary link is unspecified and either outcome remains possible,
↪  preserve that uncertainty.
Write a concise evidence note and state whether the condition is
↪  supported, refuted, or undetermined.
Do not answer the final resource-total question yet. Use only the
↪  supplied public text, with no tools.
```

Message assembly and second stage. Direct answering uses system S and user `public_text`; visible deliberation uses system $S_a + \texttt{\textbackslash n\textbackslash n} + D$ and the same user message. Generic review first uses that deliberation template. Relation notes instead use $S_a + \texttt{\textbackslash n\textbackslash n} + R$. Both two-stage methods then use system S and the identical user template:

```
{public_text}
```

```
The following working note was generated by a model and can contain
↪  errors.
Check it against the original facts and rules. Keep a supported answer;
↪  correct an unsupported one.
Do not treat the note as an additional fact. Answer the original final
↪  statement with exactly A, B, or C.
<working_note>
{generated_note}
</working_note>
```

Direct and one-call deliberation permit 1,024 output tokens; each stage of review and relation notes permits 512. The public problem is repeated in full at the second stage; the first stage's note is marked as fallible model output, not an extra premise. The relation-note condition does not ask the first stage to decide the final resource-total question.