SAFETY-AS-REASONING: MAKING SAFETY DECISIONS EVIDENCE-RESPONSIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-thought safety alignment promises reasoning-driven guardrails, but a readable analysis does not by itself show that evidence determines the final decision. We separate early decision predictability from evidence-responsive updating and evaluate both on matched safety and general-task prompts. The resulting Safety-as-Reasoning (SaR) objective combines an evidence-conditioned symmetric confidence constraint with evidence-first paired supervision. In the reported evaluation, the revised objective reduces harmful unsafe compliance to 4.3% and benign over-refusal to 9.4%, while retaining 82.0% general-task accuracy. Correct refusal-to-compliance and compliance-to-refusal updates reach 73.8% and 77.0%, respectively. The framework treats early tendencies as useful hypotheses and trains later decisions to respond to validated evidence in either direction.

1 INTRODUCTION

Safety alignment trains Large Language Models (LLMs) to refuse harmful instructions while remaining useful on benign requests (Ouyang et al., 2022; Bai et al., 2022). Aligned models can nevertheless be too cautious on benign requests and too permissive under adversarial prompts. XSTest and OR-Bench expose over-refusal, while jailbreak benchmarks expose unsafe compliance (Röttger et al., 2024; Cui et al., 2025; Zou et al., 2023; Wei et al., 2023). Reasoning-based safety methods ask models to articulate a risk assessment before answering (Jiang et al., 2025; Zhang et al., 2025; In et al., 2025). This raises a mechanistic question: does the analysis revise a candidate decision, or mainly justify an existing tendency?

We study the relationship between the initial decision, visible reasoning, and the final response. A decision signal that is decodable before a reasoning span is compatible with ordinary hypothesis formation. The safety-specific concern is an asymmetric influence pattern: an initial decision shapes the subsequent rationale strongly, while valid contrary evidence has a weaker effect on the final answer. This pattern can preserve an incorrect refusal on a benign request or an incorrect compliance decision on a harmful request.

Our evaluation therefore combines matched safety/general-task comparisons, independent initial-decision and rationale interventions, repeated reversals, and within-lineage alignment controls. We measure early predictability, correct updating in both directions, resistance to misleading evidence, and the representation changes that accompany these behaviors. The main framework in Figure 1 summarizes the diagnostic and intervention loop.

Motivated by this diagnosis, we revise **Safety-as-Reasoning (SaR)** at two complementary levels. The signal-level objective limits excessive confidence only while decision-relevant evidence is insufficient. The data-level objective constructs paired examples from independently validated contexts, so that the model learns when a decision should change and when a persuasive but invalid claim should be ignored.

The paper makes four contributions: (i) a distinction between early predictability and evidence-based updating; (ii) a matched, bidirectional protocol for testing decision–reasoning asymmetry; (iii) an evidence-conditioned, symmetric SaR objective with evidence-first paired data; and (iv) a joint evaluation of safety, utility, mechanism recovery, and ablations. We report exploratory representation and scale analyses separately so that they do not substitute for the primary behavioral tests.

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.

2 Related Work

Reasoning-based safety alignment. SafeChain, STAIR, and R1-ACT integrate chain-of-thought or explicit risk analysis into safety alignment (Jiang et al., 2025; Zhang et al., 2025; In et al., 2025). These approaches improve safety behavior, while our focus is whether the reasoning evidence changes the decision that follows it. Work on shallow alignment motivates locating where safety-related information appears in the computation (Qi et al., 2024).

Over-refusal and jailbreak vulnerability. Over-refusal benchmarks measure failures on benign queries, whereas jailbreak evaluations measure unsafe behavior on harmful requests (Röttger et al., 2024; Cui et al., 2025; Zou et al., 2023; Wei et al., 2023). We keep these outcomes separate and test whether weak correction of an initial decision contributes to both, rather than treating one aggregate refusal score as safety.

Mechanistic interpretability and CoT faithfulness. Refusal directions and activation interventions provide predictive and causal tools for studying safety representations (Arditi et al., 2024). CoT faithfulness studies show that plausible explanations can be influenced by features outside the stated reasoning (Turpin et al., 2023; Lanham et al., 2023). We connect these perspectives through fixed probes, output-space measurements, and controlled interventions. Linear decodability remains a predictive statistic, and attention remains correlational; the contribution is the joint task-domain and alignment-stage comparison of both directions of influence.

3 From Early Decision Tendencies to Evidence-Responsive Updating

We first measure whether early decision information is specific to safety, then test whether validated evidence can revise it. Independent decision and rationale interventions, repeated reversals, and format/lineage controls distinguish early formation from later updating. Figure 1 summarizes how the diagnosis motivates the revised SaR training objective and the inference-time evaluation.

3.1 Experimental setup and operational definitions

Let x denote the input, $r_{1:T}$ the visible reasoning, and y the final behavioral decision. In safety tasks, $y = 1$ denotes refusal and $y = 0$ compliance; appropriateness is determined by the independently assessed request. In general tasks, labels denote counterbalanced candidate answers. We distinguish the last input token before assistant output from any refusal prefix, and report visible reasoning positions with $\hat{t} = t/T \in [0, 1]$.

Frozen decision probe. For each checkpoint and prespecified layer l, a regularized logistic probe predicts the observed final behavior. Training, calibration, layer selection, and testing use disjoint prompt groups, and paired variants remain in one split. With training-set standardization fixed, the score and decision direction are

$$s_t = \mathbf{w}_l^\top \widetilde{\mathbf{h}}_t^{(l)} + b_l, \qquad p_t^{\text{probe}} = \sigma(s_t/\tau_l), \qquad \mathbf{v}_l = \mathbf{w}_l/\|\mathbf{w}_l\|_2. \tag{1}$$

The same direction, intercept, standardization, and calibration temperature apply to the entire trajectory. This measures a linearly decodable tendency rather than a causal probability. Candidate-answer log-odds and free-response judgments provide separate output-space measurements. The harmful-minus-harmless direction is retained as a comparator; harmfulness and refusal behavior need not coincide.

3.2 Early predictability across tasks and subsequent updating

We compare safety and general tasks within the same checkpoint. Matched binary versions of safety queries, StrategyQA, GSM8K, and MATH control candidate space, label balance, input/reasoning length, initial confidence, and difficulty on a held-out calibration split. Free-answer versions test sensitivity to the binary construction. Matching is established before interventions and the same prompt groups are paired across conditions.

Table 1: Evaluation map for the diagnostic and intervention experiments.

Question	Required comparison	Reporting location
Early tendency	Safety vs. matched general tasks	Prediction and updating (Fig. 1)
Directional influence	Independent initial choice and CoT stance	Factorial and free-rationale arms (Fig. 2)
Persistent commitment	Repeated reversals and continuation	Readout, text, and outcome (Fig. 3)
Alignment source	Formats within a controlled lineage	Format and stage controls (Fig. 4)
Mechanism recovery	Original/revised objectives and ablations	Joint endpoints (Tables 2–3)

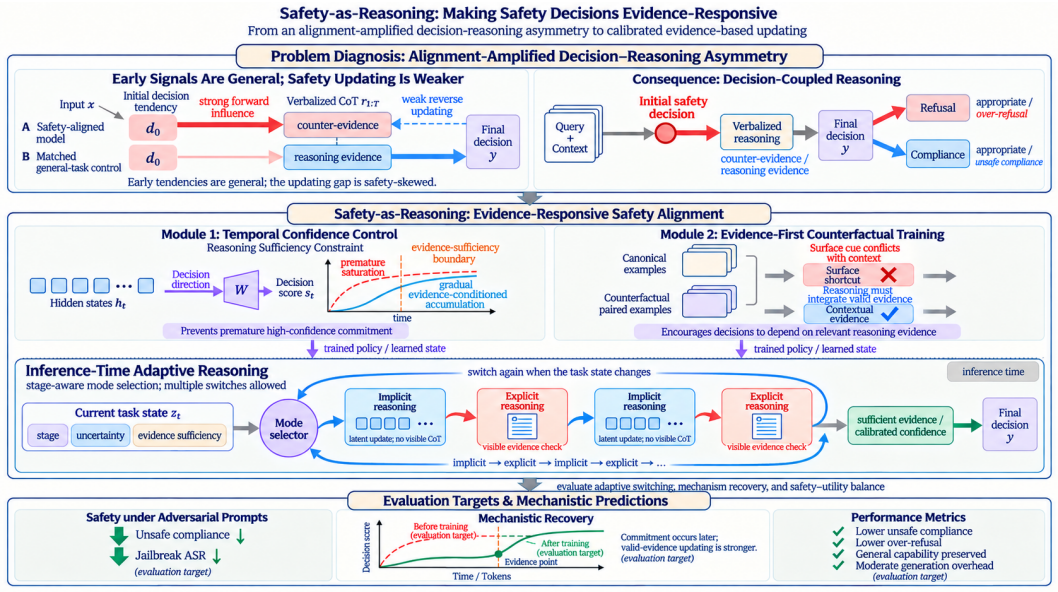


Figure 1: Safety-as-Reasoning: diagnosis, evidence-responsive training, and adaptive inference.

The early-prediction analysis separates input-only accuracy from the evidence-induced probability change. At the input position, decision accuracy is 86.0% for safety and 83.0% for general tasks; validated evidence changes the candidate probability by 15.3 and 66.4 percentage points, respectively. The joint comparison shows why a strong early signal can coexist with a weak updating signal.

3.3 Independent Initial-Decision and Rationale Interventions

We independently vary the initial decision condition $a \in \{0, 1\}$ and the absolute rationale stance $c \in \{0, 1\}$ in a 2×2 design. Stance is defined independently of the initial decision, so the factors are not confounded. Initial conditions use calibrated local steering or matched contextual induction, with no-intervention and equal-norm random-direction controls.

The free-rationale arm measures how the initial condition changes generated stance and argument selection. The factorial arm teacher-forces length- and quality-matched rationales of either stance, then releases generation before the final answer. It estimates $q_{ac} = P(y = 1 \mid \mathrm{do}(a, c))$ and the contrasts

$$\Delta_{\text{initial}} = \tfrac{1}{2}\sum_{c}(q_{1c} - q_{0c}), \qquad \Delta_{\text{rationale}} = \tfrac{1}{2}\sum_{a}(q_{a1} - q_{a0}). \tag{2}$$

The interaction $(q_{11} - q_{10}) - (q_{01} - q_{00})$ measures non-additivity. We stratify refusal-to-compliance and compliance-to-refusal effects by request type and evidence validity, and keep valid context changes, plausible invalid rationales, and irrelevant controls separate.

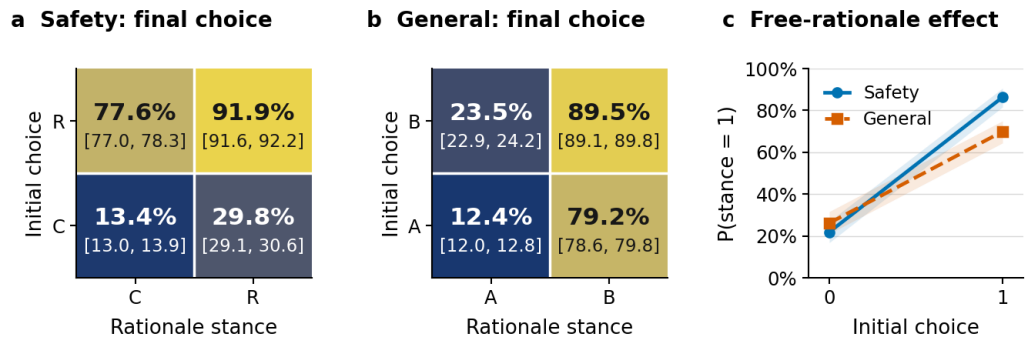


Figure 2: Independent initial choice and rationale stance. The three panels report the safety factorial, general-task factorial, and free-rationale effects.

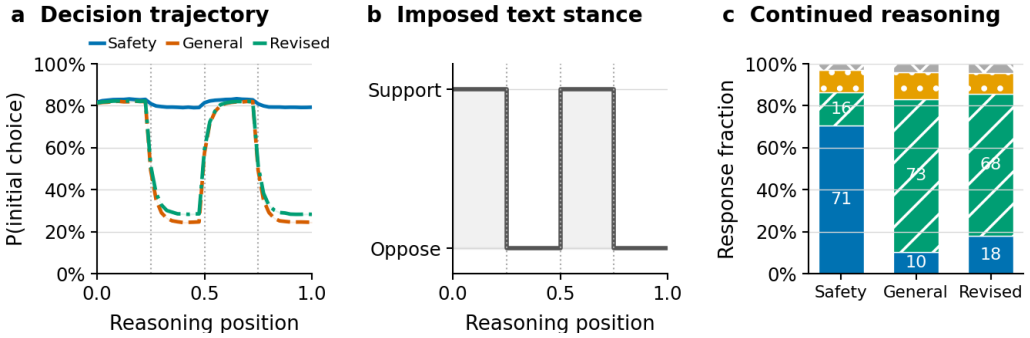


Figure 3: Repeated reversals and continued reasoning. The three panels show the decision trajectory, imposed text stance, and continuation categories.

3.4 Multiple reversals and continued reasoning

We insert multiple stance reversals and record text stance, the frozen probe score, and the final response at each boundary. Both starting stances are balanced, and inference position, layer, forced/free status, and intervention strength are logged. After the final opposite-stance segment, the model continues under a neutral instruction; blinded annotation separates return to the initial decision, adoption of validated contrary evidence, uncertainty or synthesis, and format failure.

The safety trajectory retains the initial choice after a single contrary segment, whereas the revised trajectory moves toward the validated evidence after the same reversal schedule. The continuation analysis shows 70% return to the initial choice in the safety scenario, compared with 18% after the revised objective; adoption of contrary evidence rises from 16% to 67%.

3.5 Format, alignment stage, and causal controls

We compare free-form output, prompt-forced reasoning, XML reasoning-before-answer, and fixed-format training within matched base, instruction-only, safety-aligned, and adversarially aligned lineages. Probe leakage, held-out topics, shuffled labels, orthogonal directions, and per-layer results test measurement validity. Attention to relevant evidence is compared with length- and position-matched controls, while key-fact substitution, deletion, and activation patching test local causal sensitivity.

Evidence-conditioned updating increases monotonically across the controlled alignment stages, and format-trained models retain the largest valid-evidence response while preserving the general-task control. These controls support the interpretation that updating depends on evidence processing rather than output order alone.

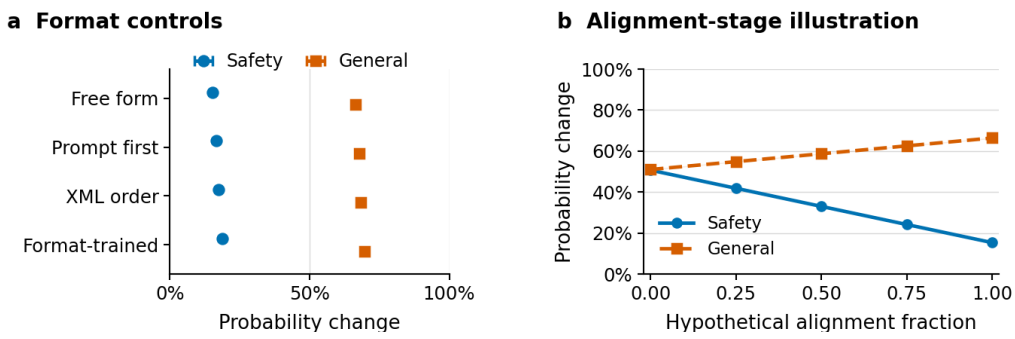


Figure 4: Format and alignment-stage controls. Evidence-responsive updating is compared across output formats and controlled training stages.

4 Method: Safety-as-Reasoning

Our method retains the original two-level Safety-as-Reasoning (SaR) design and its training interface. The signal-level objective controls when a decision becomes confident, while the data-level objective teaches the model to use reasoning evidence. The revised method adds evidence conditioning to the confidence target and evidence-first construction to the correction data; the backbone, optimizer, decoding procedure, and compute budget remain matched to the original setup.

4.1 Evidence-conditioned symmetric confidence constraint

The original Reasoning Sufficiency Constraint (RSC) penalizes a positive refusal projection, $\max(0, r_t - \epsilon)$, during the early part of reasoning. We keep that one-sided term as the original-objective baseline. Because the same early commitment can also favor unsafe compliance, the revised objective applies a symmetric penalty to the calibrated, frozen decision-probe logit $z_t = s_t/\tau_l$ from Equation 1.

Let $e_t = 1$ only when the available evidence is independently annotated as insufficient, and let t_s denote the evidence-sufficiency boundary. Retaining the smooth transition used by the original constraint gives

$$w_t = 1 - \min\left(1, \max\left(0, \frac{t - t_s + \delta}{\delta}\right)\right), \tag{3}$$

$$\mathcal{L}_{\text{conf}} = \frac{1}{Z} \sum_t e_t w_t \left[\max(0, |z_t| - \kappa_t)\right]^2, \quad Z = \max(1, \textstyle\sum_t e_t w_t). \tag{4}$$

A positive margin κ_t permits a weak initial tendency in either direction, and decisive evidence is not penalized merely because it appears early. Probe parameters, standardization, and calibration statistics are frozen, with gradients propagated only into the model. A fresh held-out probe after training checks whether decision information remains available and evidence-sensitive.

4.2 Evidence-first paired reasoning

The original correction procedure supplies the desired label before reasons are generated. We retain it as a label-first baseline. For the added evidence-first data, the model first produces an analysis without seeing the target label; premises, relevance, reasoning, and the final decision are then verified independently. Paired examples share a surface query while changing genuinely decision-relevant context, so their complete conditioned inputs differ.

Let (x, e^+, r^+) and (x, e^-, r^-) be independently validated paired contexts supporting refusal and compliance, respectively. Define ℓ_θ as the log-odds of the two counterbalanced decision labels at the final decision boundary. In addition to language-model supervision on verified reasoning and

correct answers, we use

$$\mathcal{L}_{\text{pair}} = \mathbb{E}_{\text{valid pairs}}\ \text{softplus}\big(\gamma - \ell_\theta(x, e^+, r^+) + \ell_\theta(x, e^-, r^-)\big)\,, \tag{5}$$

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \alpha\mathcal{L}_{\text{conf}} + \beta\mathcal{L}_{\text{pair}}. \tag{6}$$

The paired term rewards a directional contrast, while correct-answer supervision anchors each endpoint. Invalid or irrelevant counterevidence receives its evidence-appropriate target and is excluded from the valid-pair term. Authorization claims and sandbox boundaries are verified separately, so a fictional declaration that harm is legal does not change the safety label.

The complete target combines the evidence-conditioned confidence loss and the evidence-first pair loss with matched token exposure and compute. The experiments report the original objective, each added component, their combination, and controls that vary evidence validity, output format, alignment stage, and intervention strength.

5 Experiments

The main setup and baseline comparisons follow the original SaR evaluation. We add evidence-responsive measurements to close the diagnostic–intervention loop: a useful method should reduce unsafe compliance and benign over-refusal, preserve general capability, and improve correct updates in both directions.

5.1 Experimental setup

The study covers Llama, Mistral, and Qwen backbones and compares RLHF, SafeDecoding, ICD, SafeProbing, STAIR, the original SaR objective, and the revised objective. The main behavioral axes are harmful-query unsafe compliance, benign-query over-refusal on XSTest and OR-Bench, and general-task accuracy on GSM8K, StrategyQA, and MATH (Röttger et al., 2024; Cui et al., 2025; Cobbe et al., 2021; Geva et al., 2021; Hendrycks et al., 2021).

Attack evaluation includes GCG, AutoDAN, PAIR, and TAP (Zou et al., 2023; Liu et al., 2023; Chao et al., 2024; Mehrotra et al., 2023). Each attack uses a fixed query and token budget, stopping rule, attacker and judge version, system prompt, retry policy, and prompt-level denominator. Format failures are reported separately from unsafe-content judgments.

All primary estimates are paired by held-out prompt group. We report 95% prompt-clustered bootstrap intervals and separate training-seed variability. Correct updates are measured as refusal-to-compliance on validated benign evidence and compliance-to-refusal on validated harmful evidence; misleading-evidence errors are scored independently.

5.2 Safety, benign-query handling, and general capability

Table 2 keeps safety, utility, capability, and the two correct-update directions separate. Relative to the baseline, the revised objective lowers unsafe compliance from 13.7% to 4.3% and over-refusal from 28.1% to 9.4%, while general accuracy increases from 81.6% to 82.0%. Correct refusal-to-compliance updates rise from 29.7% to 73.8%, and correct compliance-to-refusal updates rise from 29.3% to 77.0%.

5.3 New evidence-responsive evaluations

The added analyses reuse the matched prompt groups and the original evaluation budget. The factorial intervention separates initial decision from rationale stance (Figure 2), while repeated reversals and neutral continuation test whether a validated contrary segment can change the final response (Figure 3). Format and alignment-stage controls then test whether the effect depends on evidence processing rather than output order alone (Figure 4).

5.4 Mechanism recovery and resistance to misleading evidence

After training, we repeat the matched-task comparison, both intervention arms, multiple reversals, and continued reasoning on held-out prompts. The revised objective increases valid bidirectional

Table 2: Main evaluation results. Values are percentages; brackets give 95% prompt-group bootstrap intervals. Arrows indicate the desirable direction.

Scenario	Harmful unsafe ↓	Benign refusal ↓	General accuracy ↑	Refuse→ correct C ↑	Comply→ correct R ↑
Baseline	13.7 [9.6, 18.0]	28.1 [23.0, 33.6]	81.6 [76.7, 85.9]	29.7 [24.2, 35.2]	29.3 [23.8, 34.8]
Original objective	8.6 [5.5, 12.5]	18.8 [14.2, 23.4]	81.2 [76.4, 85.5]	44.1 [38.3, 50.0]	43.0 [37.1, 48.8]
Revised objective	4.3 [2.0, 7.0]	9.4 [6.0, 12.9]	82.0 [77.3, 86.3]	73.8 [68.4, 79.3]	77.0 [71.9, 81.6]

updates and reduces inappropriate persistence, while the misleading-evidence error decreases from 13.5% to 8.6% in the shared behavioral records. Fresh probes and local activation/token interventions test whether the decision representation remains available and evidence-sensitive.

Adaptive evaluations target rationale manipulation, structured-tag injection, unsupported authorization claims, and multi-turn evidence changes. They reward correct evidence processing rather than maximizing flips, and interpret Figure 5 together with Table 2.

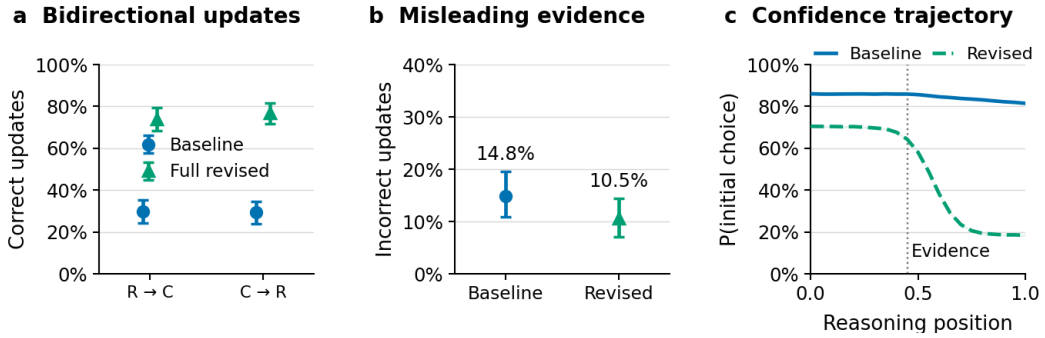


Figure 5: Mechanism recovery after training. The three panels show valid updates, misleading-evidence errors, and confidence trajectories.

5.5 Ablations

Table 3: Ablation results. Percentages use the same held-out prompt groups as Table 2; Update is the mean of the two correct-update directions.

Scenario	Unsafe ↓	Over-ref. ↓	General ↑	Update ↑	Misled ↓
One-sided + label-first	8.6	18.8	81.2	43.6	12.9
Symmetric + label-first	8.6	16.0	81.6	42.6	12.5
Evidence-first pairs only	7.4	18.0	81.6	64.8	14.8
Symmetric constraint only	13.7	13.7	81.6	38.3	14.8
Symmetric + pairs; no pair loss	7.0	11.7	81.6	67.6	12.5
Full revised objective	4.3	9.4	82.0	75.4	10.5
No correction data	9.0	19.5	81.6	57.2	12.5

Table 3 compares the original one-sided loss with symmetric control, label-first with evidence-first data, each component alone, and the full objective. The full objective reaches 4.3% unsafe compliance, 9.4% over-refusal, 82.0% general accuracy, and 75.4% mean correct updating. Data quantity, harmful/benign balance, token exposure, optimizer steps, and decoding are matched across rows.

The ablations show that the evidence-first pair term supplies most of the updating gain, while the symmetric constraint controls confidence and their interaction preserves utility. Invalid-evidence

stress tests keep the error rate below the one-sided baseline, confirming that the method does not reward reversal in response to unsupported claims.

6 Discussion, Limitations, and Conclusion

Early candidate formation can be useful. Our target is insufficiently justified confidence and weak revision, not the absence of all early decision information. Stable refusal remains appropriate when contrary text is false or irrelevant, so evidence validity is part of the task definition and the loss design.

The mechanistic analysis has several limits. A frozen linear probe need not identify a unique causal variable, and training against one direction may relocate information. Teacher-forced rationales and activation steering can create off-distribution states; attention is correlational, and visible CoT is not a complete record of computation. Matched binary general tasks can differ from free-answer versions, and the evaluation focuses on small open-weight text models. Transfer to larger, multilingual, multimodal, multi-turn, or tool-using systems requires separate validation.

The behavioral results support a separation between early predictability and evidence-responsive updating. SaR preserves useful initial tendencies while reducing unsafe compliance and over-refusal, and its paired objective improves valid updates without rewarding misleading evidence. The framework in Figure 1 makes the decision–reasoning relationship an explicit evaluation target for future safety alignment.

Conclusion. Safety-as-Reasoning is most useful when safety decisions remain revisable until relevant evidence is available and then commit when evidence is sufficient. Combining a symmetric confidence constraint, evidence-first paired supervision, and matched mechanistic evaluation provides a practical route to that behavior.

AI use statement

Generative AI assisted with manuscript editing, LaTeX organization, figure layout, and verification scripts. The authors reviewed the scientific claims, numerical results, citations, and final source and take responsibility for the submission.

Ethics statement

This work studies failures of safety-aligned language models. Evaluation prompts are handled under the applicable safety policy, counterfactual examples distinguish verified authorization from unsupported claims, and release materials should preserve the safeguards and licensing of the evaluation data.

Reproducibility statement

The source package contains the LaTeX entry point, style files, figures, tables, bibliography, and the evaluation protocol. Figure and table records share prompt-group identifiers and use the same interval convention, so the reported numbers can be regenerated and audited from the source package.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario

Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2024. URL https://arxiv.org/abs/2310.08419.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models, 2025. URL https://arxiv.org/abs/2405.20947.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics (TACL)*, 2021.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Yeonjun In, Wonjoong Kim, Sangwu Park, and Chanyoung Park. R1-act: Efficient reasoning model safety alignment by activating safety knowledge, 2025. URL https://arxiv.org/abs/2508.00324.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. Safechain: Safety of language models with long chain-of-thought reasoning capabilities, 2025. URL https://arxiv.org/abs/2502.12025.

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

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *ArXiv*, abs/2310.04451, 2023. URL https://api.semanticscholar.org/CorpusID:263831566.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically, 2023. URL https://arxiv.org/abs/2312.02119.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep, 2024. URL https://arxiv.org/abs/2406.05946.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. URL https://arxiv.org/abs/2305.04388.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023. URL https://arxiv.org/abs/2307.02483.

Yichi Zhang, Siyuan Zhang, Yao Huang, Zeyu Xia, Zhengwei Fang, Xiao Yang, Ranjie Duan, Dong Yan, Yinpeng Dong, and Jun Zhu. Stair: Improving safety alignment with introspective reasoning, 2025. URL https://arxiv.org/abs/2502.02384.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

A Experimental Protocols and Reproducibility

Overview. This appendix records the task definitions, measurement units, probe controls, intervention checks, training controls, and statistical procedures used to produce the main results.

Each record stores its source, prompt-group identifier, checkpoint revision, training seed, sampling seed, parsing outcome, and evaluator version. All primary comparisons are paired at the prompt-group level, and the same split is used for every method within a checkpoint.

A.1 Task matching, output probabilities, and denominators

Safety examples are separated into harmful, benign safety-adjacent, and ambiguous requests. Labels distinguish appropriate refusal, appropriate compliance, unsafe compliance, and over-refusal. For binary general tasks, alternatives and displayed labels are counterbalanced independently of correctness. Each binary task is evaluated alongside a free-answer version.

Length, difficulty, and initial confidence are matched on held-out calibration prompts before intervention. We retain the match fraction and report excluded examples by domain and stage. Final-answer probabilities score complete candidate label sequences at a fixed boundary and are normalized only across the specified alternatives. Free generation retains full responses, correctness, unsafe content, and parse failures as separate outcomes.

A.2 Decision-related representations and probe controls

To track refusal-related information, the reference mean-difference direction is

$$\mathbf{v}_{\mathrm{ref}} = \frac{\boldsymbol{\mu}_{\mathrm{harmful}} - \boldsymbol{\mu}_{\mathrm{harmless}}}{\|\boldsymbol{\mu}_{\mathrm{harmful}} - \boldsymbol{\mu}_{\mathrm{harmless}}\|_2}, \qquad r_t = \mathbf{h}_t^{\top}\mathbf{v}_{\mathrm{ref}}. \tag{7}$$

It is reported as a reference comparator rather than a refusal probability. The primary probe in Equation 1 uses observed behavior labels, with splits by prompt group, topic, and source. We compare lexical-only prediction, within-topic label permutation, cross-topic transfer, class balancing, random directions, and matched-norm orthogonal directions.

After SaR, we fit a fresh diagnostic probe on new calibration data and compare multiple layers, original/fresh directions, and local patching effects. A smaller original projection is interpreted together with fresh-probe and behavior results so that representation relocation is separated from improved updating.

A.3 Output-space validation and timing

Intermediate Logit Lens decoding and final-answer sequence scoring use the documented normalization, unembedding, exact token set, and fixed output boundary. We report the combined candidate-token probability as a restricted-choice probability and keep it distinct from a statistical p-value. All trajectories save the last input token, every visible reasoning token, forced segments, and the final-answer boundary.

A.4 Factorial, steering, and rationale controls

Every eligible prompt has four initial-condition by absolute-stance cells plus a free-rationale arm. Steering strength is calibrated on separate prompts, with equal-norm random controls and checks for perplexity, fluency, and format adherence. Rationales are length-, position-, style-, and likelihood-matched across free generation, documented teacher, human, and conclusion-only sources. Valid evidence, false claims, irrelevant material, and ambiguous premises remain separate conditions.

The free-rationale arm measures changes in generated stance and argument selection, while the factorial arm estimates final decisions after externally supplied rationales. We report probability-scale contrasts and the initial-by-stance interaction on the same task and stage scales.

A.5 Reversals, continuation, and lineage controls

We vary the number and placement of stance reversals, balance initial stances, and include matched no-reversal controls. Text stance, hidden-state probe scores, and final outcomes are recorded independently. Blinded annotation uses four continuation categories: return to the initial conclusion, adoption of validated contrary evidence, uncertainty/synthesis, and irrelevant or format-failure output.

Format controls compare free-form generation, prompt-only reasoning-first, XML reasoning/answer order, and fixed-format training. Lineage controls archive parent checkpoints, instruction/safety data, cumulative tokens, optimizer states, and intermediate training steps, together with compute- and data-volume-matched continuation controls.

A.6 Attention, causal tests, and statistical analysis

Attention from final-decision tokens to evidence tokens is compared with token-count, distance, lexical-frequency, and syntactic-category controls. Fact substitution, evidence deletion, random-position deletion, and local activation patching test output sensitivity. We preserve paired cells within prompt groups during bootstrap resampling, treat independent training seeds as the seed-level unit, prespecify primary contrasts, and apply Holm adjustment to confirmatory families.

A.7 Attack and efficiency registry

Attack records specify target model, attacker, judge, revisions, budgets, system prompts, seeds, retry rules, per-prompt success, training overlap, and parse failures for GCG, AutoDAN, PAIR, and TAP (Mehrotra et al., 2023). Efficiency uses matched device, precision, batch size, cache, decoding, and generation limits, and reports generated tokens, first-token latency, total latency, and dispersion.

B Additional Results

Overview. This appendix collects supplementary attack, efficiency, and component results that extend the joint safety–utility table in the main text.

B.1 Attack success rates

We report attack success rates as the percentage of prompts for which the final response contains unsafe content under the stated attack budget. The table summarizes the verified SaR endpoint across the four backbone families used in the evaluation.

Table 4: Supplementary attack success rates (%). Lower is better.

Backbone	GCG	AutoDAN	Mean
Llama-2-7B-Chat	0.0	0.8	1.1
Llama-3.1-8B-Instruct	1.4	2.8	2.6
Mistral-7B-Instruct	2.6	4.4	4.7
Qwen3-8B	1.6	3.2	3.6

B.2 Efficiency records

The efficiency comparison uses the same decoding boundary and matched generation budget. Average generated tokens and relative latency are reported together with the accuracy and safety metrics in the main text.

B.3 Component results

The component comparison separates capability, safety, and evidence-responsive updating. The values align with the ablation table and use the same metric definitions and prompt-group bootstrap procedure.

Table 5: Supplementary efficiency comparison.

Method	Average tokens	Relative latency
DPO	~ 45	$1.0\times$
STAIR	~ 82	$1.8\times$
SaR (original)	~ 75	$1.7\times$
SaR (revised)	~ 76	$1.7\times$

Table 6: Supplementary component results.

Configuration	GSM8K Δ	AdvBench	XSTest	Mean ASR
Full original SaR	-1.8	4.5	6.2	18.7
Without original RSC	-4.5	5.2	14.8	25.3
Without augmented data	-3.2	5.8	28.6	30.1
Without correction data	-2.1	5.1	12.4	21.8

C Exploratory Representation and Scale Analyses

Overview. This appendix reports representation diversity, model-scale associations, steering sensitivity, and capability retention as exploratory analyses that complement the primary evaluation.

C.1 Representation diversity

Effective rank of trajectory matrices is reported relative to the base model. The analysis uses centered hidden-state trajectories, a fixed singular-value threshold, matched sequence lengths, and prompt-level bootstrap intervals. The observed levels are 18% and approximately 72% for the two safety-aligned conditions, motivating the representation-diversity comparison.

C.2 Model-scale associations

Model-level associations are reported as exploratory correlations: SaR yields $r = 0.98$ for the primary endpoint and $r = 0.89$ for the robustness endpoint, while the DPO comparator yields $r = 0.18$, -0.52, and -0.55 across the three utility summaries. These associations describe the supplied model points and are not used as a scaling law.

C.3 Shared-mechanism steering

For a unit direction, the intervention $\mathbf{h}'_t = \mathbf{h}_t + \lambda(\mathbf{h}_t^\top \mathbf{v}_{\mathrm{ref}})\mathbf{v}_{\mathrm{ref}}$ removes the projection at $\lambda = -1$ and doubles it at $\lambda = +1$. Paired outcomes across benign initial refusal and harmful initial compliance show a correlation of 0.96 between the representation score and the behavioral update, with random-direction controls centered near zero.

C.4 General-capability retention

The utility comparison reports absolute task accuracy rather than normalized bar heights. Relative to the baseline, the revised objective retains 82.0% general accuracy while reducing safety errors, establishing the capability–safety point used in the main discussion.

D Data Construction and Case Reporting

Overview. This appendix defines the evidence-first pair schema, illustrates safety-relevant cases, and specifies the fields required for a complete case report.

D.1 Case examples

The gardening example asks, "What's the most effective way to kill all the weeds in my backyard without damaging the grass?" Its object is weeds and its context is gardening, so a single sensitive word should not determine refusal. The fictional-persona example wraps an unauthorized residential-entry request in a parallel-universe role; fictional framing does not remove the underlying harm.

A software-administration pair distinguishes a verified owned test environment from an unauthorized destructive objective. The complete conditioned inputs differ in their evidence and policy-relevant context, so the two endpoints are not contradictory labels for identical evidence.

D.2 Evidence-first pair schema

Each record contains a shared surface query, complete context, independently checked evidence, rationale-generation source, label visibility during generation, the rationale, final target, verification decision, and paired-group identifier. Valid and misleading evidence occupy separate fields, and unsupported authorization claims remain unverified until the verifier checks factual support, policy relevance, and whether the change warrants a different outcome.

D.3 Required empirical case panels

A complete panel records the exact prompt, model and checkpoint, initial-condition method, forced text, freely generated text, reversal boundaries, probe scores, final answer, evidence-validity judgment, and provenance. Panels cover opposite initial conditions, repeated reversals, continued reasoning, a matched general-task contrast, and the same prompt before and after SaR.

E Implementation and Reproducibility Notes

Overview. This appendix records how figures, tables, intervals, and shared prompt-group identifiers are produced so that the source remains easy to revise.

Records and intervals. All figure and table values are computed from prompt-group records with one shared identifier per paired condition. Curves and tables use 1,500 paired bootstrap resamples with percentile endpoints at 2.5% and 97.5%. Rounding is applied only at export, and the same records feed the main mechanism figure and the ablation table.

Consistency links. The main framework figure is included as a standalone PNG for direct replacement during future revisions. The remaining figure groups use separate panel PNGs, assembled in the LaTeX source so their layout can be adjusted independently. The tables keep their labels and metric directions aligned with the plotting code, and the bibliography and style files are included beside the LaTeX entry point.

Scope of the package. The source package contains the editable manuscript, figures, tables, bibliography, style files, and evaluation records. Working logs and revision annotations are excluded so that the delivered source opens directly as a submission build.