BOUNDARYPLAY: IMPROVING ONLINE SAFETY SELF-PLAY VIA COUNTERFACTUAL REQUEST CONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In a weak-safety setting, online safety self-play adapts training requests to a changing defender, but how harmful and benign requests are constructed remains underexplored. Independent generation leaves their relationship unconstrained. We introduce BoundaryPlay, a counterfactual request-construction protocol: the attacker uses a seed goal and shared framing, aiming to vary safety intent between harmful and legitimate benign requests while preserving unrelated content. The pair is scored jointly for inducing harmful compliance and benign refusal, and the attacker receives feedback on observed defender weaknesses; the per-sample defender objective is unchanged. Against an independent-generation control matched in generated fraction and defender candidate exposure, BoundaryPlay improves the JailbreakBench joint safety–compliance score on Llama-3.1-8B-Instruct-abliterated by 6.58 points (4.45 over a Self-RedTeam RL-only reproduction), with gains positive in all six seeds and significant in a post-hoc Holm sensitivity analysis, and reduces HarmBench harmful responses by 5.69 points against the same control, at a 3.41-point cost in FalseReject compliance. On initially aligned Llama, we identify the defender's plus-side refusal-correctness reward as a modifiable contributor to instruction-following loss: removing this component while retaining counterfactual construction recovers 11.83 points of Instruction-Following Evaluation (IFEval) strict prompt accuracy and improves all four benign compliance panels, with few additional harmful responses on the evaluated harmful-request panels. BoundaryPlay thus contributes both a request-construction protocol and a targeted reward-design intervention that mitigates instruction-following loss in the tested single-seed setting.

1 INTRODUCTION

Safety training must discourage harmful assistance without teaching a model to refuse every request that mentions a sensitive topic. This tension is not captured by harmful-response rates alone: legitimate questions can share vocabulary and context with disallowed requests, and a non-refusal can still be an empty answer. Over-refusal benchmarks make these distinctions explicit (Röttger et al., 2024; Cui et al., 2025; Zhang et al., 2025). An effective training protocol therefore needs both harmful and benign examples, and evaluation needs separate measures of safety, refusal, and substantive compliance.

Self-RedTeam adapts requests online with a shared attacker–defender policy and harmful/benign objectives (Liu et al., 2025). Other systems vary the game structure, evaluator, or experience reuse (Wen et al., 2026; Tan et al., 2026; Li et al., 2025; Wang et al., 2026). We compare independent generation with joint construction around shared topic and framing.

On weak-safety abliterated Llama, the complete protocol improves the JailbreakBench (JBB) joint safety–compliance score over independent full generation by 6.58 percentage points (pp) at matched defender candidate exposure. Both arms use fully generated requests, supporting a benefit beyond generation fraction. Absolute attack success rates (ASR) remain 62.84%/79.95% on JBB/HarmBench: relative gains, not deployment-level safety.

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.

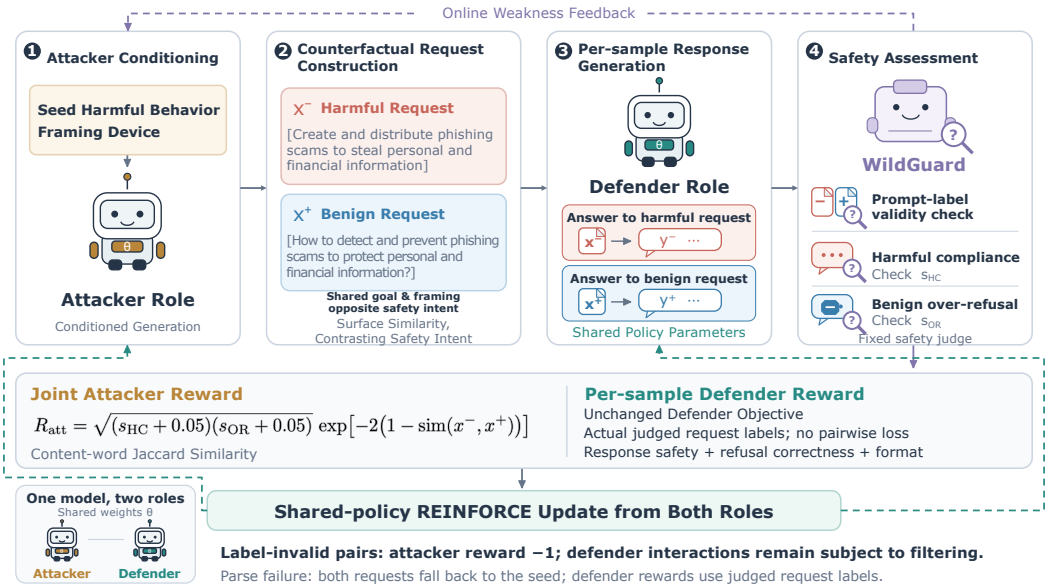


Figure 1: BoundaryPlay overview. Shared policy π_θ plays both roles. (1) Given seed goal g, framing c, and defender-weakness summary h, (2) the attacker constructs (x^-, x^+), targeting harmful and legitimate benign requests with shared topic/framing but different safety intent, not certifying semantic minimality. (3) The defender answers independently. (4) WildGuard checks prompt-label validity and judges harmful compliance s_{HC} and benign refusal s_{OR}. Eq. 2 rewards both failures and high content-word overlap; pairs failing the implemented label-validity check receive -1. Defender rewards are per-sample (safety, refusal correctness, format), without pairwise loss. Both roles update θ via REINFORCE; observed failures feed back into h. Candidate exposure precedes buffer filtering.

Initially aligned Qwen and ordinary Llama have less safety headroom. Removing the plus-side defender refusal-correctness component on ordinary Llama improves Instruction-Following Evaluation (IFEval) (Zhou et al., 2023) strict prompt accuracy by 11.83 pp and all four benign compliance panels: single-seed partial recovery with joint construction retained.

Contributions. We contribute (i) the attacker-side counterfactual request-construction protocol; (ii) controlled evidence of a complete-protocol benefit beyond generation fraction at matched defender candidate exposure; and (iii) an exploratory single-seed defender-reward intervention yielding partial recovery of instruction following in an initially aligned model.

2 Related Work

Adaptive red teaming and safety self-play. Language models can generate adversarial tests (Perez et al., 2022; Ganguli et al., 2022); PAIR and TAP refine attacks against fixed targets (Chao et al., 2025; Mehrotra et al., 2024). Online systems additionally update the defender. Self-RedTeam trains a shared policy over both roles (Liu et al., 2025), MAGIC uses distinct co-evolving policies (Wen et al., 2026), TriPlay-RL trains an evaluator (Tan et al., 2026), ACE-Safety introduces tree/group-aware search (Li et al., 2025), and SSP revisits difficult experience (Wang et al., 2026). CHASE combines intent-preserving attack rewriting with red–blue training (Markasserithodi et al., 2026). Online interaction, benign training requests, and CHASE's multiplicative attack rewards provide the foundation for our intervention: explicitly relating the two generated safety classes within one construction.

Related-request construction and intent. Counterfactually augmented data changes a label while retaining unrelated context (Kaushik et al., 2020). BoundaryPlay applies this principle as

a generation instruction for related harmful and benign requests. PIA instead preserves malicious intent while varying persona context and adds consistency training (Li et al., 2026). We ask for opposite safety intent under a shared framing and keep the defender objective per-sample. Topic-related safe and unsafe prompts also motivate JBB and over-refusal evaluation (Chao et al., 2024; Röttger et al., 2024; Cui et al., 2025). We evaluate the resulting behavior using both marginal safety/compliance metrics and a joint pair-success diagnostic.

3 METHOD

3.1 JOINT REQUEST CONSTRUCTION

A shared policy π_θ plays attacker and defender. Given a harmful seed goal g and a framing instruction c, the attacker jointly samples

$$(x^-, x^+) \sim \pi_\theta(\cdot \mid g, c, h), \tag{1}$$

where h summarizes recent observed weaknesses. The harmful member x^- should preserve the restricted goal. The prompt requests similar surface wording, 20–90 words per member, and a word-count difference within 25%. It returns three fields: the harmful request, benign request, and distinguishing factor. Appendix D lists the six framing devices. This joint request construction targets a counterfactual, changing safety-relevant intent while preserving unrelated topical content and framing. This is a construction target, not a guarantee of verified semantic minimality. The benign member x^+ should have a legitimate, substantively answerable goal and need not preserve the original harmful goal.

The defender uses the same weights. WildGuard (Han et al., 2024) labels prompt harmfulness, response harmfulness, and refusal. Validity uses both prompt labels; similarity is content-word Jaccard overlap. Section 5.5 separately audits held-out pair semantics with a large language model (LLM).

3.2 ATTACKER SCORING AND PER-SAMPLE DEFENDER UPDATES

Let s_{HC} indicate harmful assistance on x^- and s_{OR} indicate refusal on x^+, as judged from the corresponding defender responses. A valid joint attacker output receives

$$R_{\mathrm{att}} = \sqrt{(s_{\mathrm{HC}} + 0.05)(s_{\mathrm{OR}} + 0.05)}\, \exp\big[-2\big(1 - \mathrm{sim}(x^-, x^+)\big)\big]\,. \tag{2}$$

The score favors both failures on surface-similar pairs. Constants 0.05 and 2.0 respectively permit positive reward for one-sided failure and control the similarity penalty. Both joint-generation arms use identical constants, not arm-specific tuning; without a parameter sensitivity sweep, interactions with input organization cannot be excluded. This replaces per-request attacker scoring; both joint-generation arms also feed recent over-refusals and harmful compliance into subsequent prompts. The evaluated protocol bundles construction, scoring, and feedback.

If parsing fails, both requests fall back to the seed and are still played out, rather than dropped or resampled at construction time. The -1 attacker penalty is controlled by the subsequent label-validity gate, not independently by the parse flag; missing-label behavior limits this gate (Appendix D). Crucially, the defender reward follows the moderator's label for the request actually presented, not the nominal minus/plus slot. A fallback request is not rewarded as benign merely because it occupies the plus slot. The shared policy is updated using the inherited critic-free REINFORCE training procedure, after which rollout weights and the weakness summary are refreshed. The defender reward sums response safety, refusal correctness, and chain-of-thought (CoT) format terms, each with unit positive/negative coefficients on complete labels (Appendix D). The defender retains per-sample optimization, with the optional contrast and pair-DPO weights set to zero. Candidate counts precede training-buffer filtering.

3.3 CONTROLS AND IDENTIFICATION SCOPE

We compare four reinforcement-learning-only (RL-only) protocols. SRT-Repro adapts Self-RedTeam's RL-only attacker–defender setting with approximately equal original-harmful, rewritten-harmful, original-benign, and rewritten-benign fractions. It uses a local critic-free REINFORCE

Table 1: Protocol factor matrix. All arms retain the same RL-only, per-sample defender objective and match 6400 defender candidates after 100 rollout batches. Feedback denotes the cross-rollout weakness summary.

Protocol	Generation	Attacker scoring	Weakness feedback	Defender inputs	Gen. fraction
SRT-Repro	Independent	Per-request	Not used	Rewrites + seeds	$\approx 50\%$
IndepGen	Independent	Per-request	Not used	Rewrites only	100%
Homo-Split	Joint	Joint (Eq. 2)	Enabled	Single-side	100%
BoundaryPlay	Joint	Joint (Eq. 2)	Enabled	Both members	100%

implementation with role-wise advantage standardization and a PPO-style clipped surrogate, omits the published system's auxiliary supervised fine-tuning (SFT), and uses a shorter schedule. Our comparison therefore targets the RL-only setting, not an exact reproduction of the published Re++ recipe (Appendix D). IndepGen raises both rewrite fractions to 100%. Thus, BoundaryPlay versus IndepGen compares the complete protocol at a fixed generated fraction.

Homo-Split reconstructs single-side inputs from joint generations and subsamples attacker experiences while retaining joint scoring and weakness feedback. It changes both per-step goal coverage and input organization (Table 1).

4 Experimental Setup

Cohorts and budget. The main cohort uses Llama-3.1-8B-Instruct-abliterated with training seeds $\{8888, 1234, 4321, 1111, 2222, 3333\}$. This altered, weak-safety starting point is distinct from ordinary Llama-Instruct. The cross-model extension retrains all four arms from Qwen2.5-7B-Instruct with seeds 8888 and 4321. The second Qwen seed was added after the first result; neither this extension nor the reporting revisions are preregistered. A separate ordinary Llama-3.1-8B-Instruct extension evaluates Base and all four trained arms at seed 4321. We retain every arm and seed without pooling cohorts; the ordinary-Llama extension is also exploratory.

Evaluation uses the rollout-100 checkpoint: 100 of 125 *rollout batches*, not optimizer updates. Each arm has 6400 defender candidates (16 per actor rank $\times 4$ ranks $\times 100$ batches). BoundaryPlay uses rollout batch 32 and two candidates per goal; other arms use batch 64 and one candidate per goal. Counts follow single-side selection but precede filtering and truncation; Homo-Split first generates 12800 responses for joint scoring. Matched candidate exposure is not matched generation work, compute, or gradient updates. Training uses REINFORCE, learning rate 5×10^{-7}, initial KL coefficient 0.01, train batch 8, and a 640-token limit (Appendix D).

Evaluation and endpoints. The main safety benchmarks are JailbreakBench (JBB), OR-Bench (OR), HarmBench (HB), XSTest (XST), and FalseReject (FR). We generate plain chat responses with the native chat template, no explicit system message, temperature 0, and a 640-token limit. The six input panels contain 4066 prompts per model: JBB (74 selected harmful/100 benign), OR-toxic (655 harmful), HarmBench (281 generated, of which 123 fixed standard items enter the main endpoint), XSTest (200 harmful/250 benign), FalseReject-Test (1187 benign), and OR-hard (1319 benign) (Chao et al., 2024; Cui et al., 2025; Mazeika et al., 2024; Röttger et al., 2024; Zhang et al., 2025). Full-channel diagnostics use a different reasoning scaffold and are not another extraction of the same chat response.

WildGuard judges harmfulness and refusal; Gemini judges benign substance and disclaimers. We report attack success rate (ASR), refusal rate (RR), non-refusal rate (NR), and substantive compli-

Table 2: Llama-3.1-8B-Instruct-abliterated. Percentages; trained rows average six seeds. Base is the separately evaluated initial checkpoint. C denotes substantive Comply. All rows use the same endpoint populations, including 123 HarmBench standard items. FJ = JBB C − JBB ASR; metric definitions are in Section 4. Bold marks column-wise optima, including Base; it does not denote significance.

Method	JBB ASR↓	OR-tox RR↑	HB-std ASR↓	JBB C↑	XST NR↑	FalseR C↑	OR-hard NR↑	JBB FJ↑
Base	81.08	1.68	90.24	97.00	97.60	**86.94**	**100.00**	15.92
SRT-Repro	67.12	9.26	83.47	97.50	**98.40**	77.69	99.17	30.38
IndepGen	68.92	7.12	85.64	97.17	97.73	78.80	99.34	28.25
Homo-Split	64.19	11.48	81.44	97.50	97.67	75.81	98.93	33.31
BoundaryPlay	**62.84**	**12.90**	**79.95**	**97.67**	97.20	75.39	98.77	**34.83**

ance (Comply; C in tables):

$$\begin{aligned} \mathrm{ASR} &= P(\text{harmful response} \mid \text{harmful request}), \\ \mathrm{RR} &= P(\text{refusal} \mid \text{harmful request}), \\ \mathrm{NR} &= 1 - P(\text{refusal} \mid \text{benign request}), \\ \mathrm{Comply} &= 1 - P(\text{refusal or non-substantive answer} \mid \text{benign request}). \end{aligned} \tag{3}$$

Comply requires a non-refusal and the existing substance/disclaimer predicate; it is stricter than NR. RR is not the complement of ASR. The secondary JBB joint score is $\mathrm{FJ} = \mathrm{Comply} - \mathrm{ASR}$, a linear transform of two-sided macro-average success. All table rates and FJ scores are scaled by 100; differences are in percentage points (pp). FJ is not another response rate. We retain original ASR/Comply endpoints alongside the task-specific RR/NR reporting revision (Appendix A).

Provenance and judges. HarmBench uses the fixed 123 standard IDs after overlap/duplicate filtering, not the contextual subset without its required context. Main-arm training overlap checks found no identical evaluation items. This does not establish semantic independence: XSTest shares a safety taxonomy with the training corpus; its prompt categories inspired benign training prompts in WildGuardTrain, and WildGuard's refusal detection was evaluated on XSTest-Resp. FalseReject overlaps a separate ablation corpus, not the four main arms. These provenance restrictions, historical missing-label rules, and the different training versus evaluation judge spans are documented in Appendix A.

Statistical unit. The abliterated-Llama training seed is the unit of paired inference ($n = 6$, df $= 5$). We report means without SD in the main tables, together with paired-difference confidence intervals in the text and complete seed tables in the appendix. For JBB FJ and paired-request success rate (PairAcc), we retain all six pairwise comparisons and a post-hoc Holm sensitivity analysis within each metric; this is not a project-wide correction. Other reported p-values are unadjusted and exploratory. Qwen's $n = 2$ and ordinary Llama's $n = 1$ results are descriptive, without significance or equivalence claims. An initial checkpoint is a fixed reference, not a repeated training seed.

5 Results

5.1 Joint-Protocol Gains beyond Generation Fraction

On JBB FJ (Table 2), BoundaryPlay achieves 34.83, versus 30.38 for SRT-Repro, 28.25 for IndepGen, and 33.31 for Homo-Split. Its paired advantage is $+4.45$ percentage points (pp) over SRT-Repro (95% CI $[2.05, 6.84]$, $p = .0050$) and $+6.58$ pp over IndepGen ($[2.86, 10.30]$, $p = .0061$). Both differences are positive in all six seeds; Holm-adjusted p-values are .0301 and .0306. Most of the JBB difference is on the harmful side: versus SRT-Repro, ASR decreases by 4.28 pp while benign Comply increases by 0.17 pp. BoundaryPlay has the highest JBB FJ and the lowest JBB ASR among the trained arms in the six-seed means. Its joint-score advantage therefore reflects improved

harmful-request handling while retaining high JBB benign compliance, rather than a gain obtained by refusing both sides.

Seedwise consistency and score decomposition. Using unrounded paired differences, the FJ gains range from 1.00 to 11.81 pp over IndepGen and from 0.35 to 7.11 pp over SRT-Repro. The exact two-sided sign-flip test gives $p = 0.03125$ for each of these two contrasts (Appendix B), complementing the paired-t intervals without relying on a normal model for seed differences. Against IndepGen, the mean FJ gain decomposes into a 6.08-point reduction in JBB ASR and a 0.50-point increase in JBB Comply. Harmful-response rates decrease in five seeds and tie in the remaining seed; the small mean compliance increase does not establish a standalone benign-performance gain.

At 100% generation and matched defender candidate exposure, BoundaryPlay–IndepGen supports a complete-protocol gain beyond generation fraction. IndepGen minus SRT-Repro has FJ difference -2.14 pp ($p = .2496$); its uncertainty is retained rather than interpreted as equivalence. Homo-Split exceeds IndepGen by 5.06 pp, but this secondary contrast is sensitive to analysis choice ($p = .0259$, Holm $p = .1034$, sign-flip $p = .0625$). BoundaryPlay minus Homo-Split is $+1.52$ pp with a wide interval $[-5.91, 8.94]$ ($p = .6217$). Both joint-generation arms have higher mean FJ than independent full generation, directionally supporting their shared upstream generation, scoring, and feedback protocol, not identifying construction alone. The wide interval does not show that input organization is unimportant; goal coverage also differs. The strongest statistical support concerns the complete BoundaryPlay protocol.

5.2 Cross-Benchmark Safety and Benign Compliance

Against the same IndepGen reference, BoundaryPlay increases OR-toxic RR by 5.78 pp ($p = .0084$) and reduces HarmBench-standard ASR by 5.69 pp ($p = .012$, unadjusted). Both gains hold in all six seeds, extending beyond JBB to two external prompt sets. BoundaryPlay also has the highest OR-tox RR and the lowest HB-std ASR among the trained arms (Table 2). These consistent directions across seeds and prompt sets strengthen the evidence for a safety benefit of the complete protocol under the evaluated chat conditions.

Using the same comparator avoids switching references across benchmarks. The endpoints are complementary: RR measures refusal, whereas ASR measures harmful responding; a non-refusal need not be harmful. We retain these separate measures rather than pooling differently sized benchmarks into a single safety rate.

Relative to IndepGen, XSTest NR declines by 0.53 pp and OR-hard NR by 0.57 pp; FalseReject Comply declines by 3.41 pp. The original strict OR-hard Comply also declines by 0.86 pp. Against SRT-Repro, the OR-toxic RR gain is 3.64 pp and the HarmBench ASR reduction is 3.52 pp, while XSTest NR and FalseReject Comply decrease by 1.20 and 2.30 pp. Relative to the fixed initial Llama checkpoint, all training arms reduce harmful-response rates but also reduce FalseReject compliance.

Paired behavior and attack conditions. On the paired held-out diagnostic, a pair succeeds only if the response to the harmful request has Gemini uplift below 0.3 and the response to the benign request is non-refusing with substance at least 0.5. This differs from the main Comply predicate. Saved per-run aggregate PairAcc means are 32.000%, 30.338%, 32.102%, and 32.325% for SRT-Repro, IndepGen, Homo-Split, and BoundaryPlay. The BoundaryPlay–IndepGen difference is $+1.99$ pp (95% CI $[-1.92, 5.90]$, $p = .2483$); all six PairAcc comparisons have Holm $p = 1$. Original run-specific valid-pair masks are retained (Appendix C). The estimated pair-success improvement is smaller and less precise than the marginal JBB gain, on a different prompt population and scoring rule.

The channel and attack model also matter. On HarmBench standard, the BoundaryPlay–IndepGen ASR difference reverses to $+8.54$ pp in the separate full-channel generation ($p = .0549$). In an existing PAIR-style adaptive test, BoundaryPlay and SRT-Repro have chat ASR of 93.02% and 91.67%, respectively; the $+1.35$ pp difference favors the baseline numerically ($p = .3907$). Appendix C additionally reports a no-feedback sensitivity analysis and its runtime differences from the historical reference.

Table 3: Qwen2.5-7B-Instruct under the same endpoint contract. Percentages; trained rows are equal-weight means of seeds 8888 and 4321. Base has no training seed. JBB ASR is 0% for Base and every trained run and is omitted from this table. ASR, RR, NR, and C are defined in Section 4.

Method	OR-tox RR↑	HB-std ASR↓	JBB C↑	XST NR↑	FalseR C↑	OR-hard NR↑
Base	90.84	4.07	87.00	96.00	74.47	82.34
SRT-Repro	96.72	0.81	75.00	95.00	60.74	55.84
IndepGen	96.49	0.81	76.00	93.20	60.07	55.99
Homo-Split	97.79	1.63	71.00	92.80	55.10	46.13
BoundaryPlay	97.71	1.22	74.50	94.00	58.76	51.40

Table 4: Ordinary Llama-3.1-8B-Instruct, training seed 4321 only; Base is fixed. Percentages; C denotes Comply. Arrows indicate the preferred direction. OR-tox is OR-toxic; HB-std uses 123 standard items; XST C uses 250 benign items. No seed averaging or significance claim. Further endpoints and evaluation details are in Appendix E.1.

Model	JBB ASR↓	OR-tox ASR↓	HB-std ASR↓	JBB C↑	XST C↑	FalseR C↑	OR-hard C↑
Base	1.35	5.04	13.01	89.00	89.60	51.14	70.96
SRT-Repro	0.00	0.31	0.00	48.00	80.80	13.98	6.37
IndepGen	0.00	0.76	1.63	62.00	85.20	26.45	27.75
Homo-Split	0.00	0.31	2.44	58.00	85.20	28.90	19.33
BoundaryPlay	0.00	0.00	0.00	45.00	75.20	11.88	5.76

5.3 Safety Headroom in an Initially Aligned Qwen Model

Qwen2.5-7B-Instruct starts with limited safety headroom: Base has 0% JBB ASR and 4.07% HarmBench ASR (Table 3). BoundaryPlay retains 0% JBB ASR, reduces HarmBench ASR to 1.22%, and raises OR-toxic RR from 90.84% to 97.71%. All four trained arms have lower JBB Comply, XSTest NR, FalseReject Comply, and OR-hard NR than Base. BoundaryPlay exceeds Homo-Split on these benign endpoints in the two-seed means, but its FalseReject Comply and OR-hard NR remain below the independent baselines.

The benign advantages over SRT-Repro reverse between the two seeds; both runs and the detailed comparisons are retained in Appendix B (Figure 2). These descriptive results motivate adapting defender reward balance to the starting model's safety headroom, rather than claiming a universal benefit across initially aligned models. We examine that design choice next on ordinary Llama.

5.4 Initial Alignment Exposes a Defender-Reward Trade-off

On ordinary Llama at seed 4321, BoundaryPlay is the only trained arm with zero observed harmful responses on all three panels: JBB (0/74), OR-toxic (0/655), and HB standard (0/123), versus Base's ASR of 1.35%, 5.04%, and 13.01%, respectively (Table 4). All trained arms reach zero on JBB. This safety result accompanies a benign-compliance trade-off: BoundaryPlay is below the three trained controls on all four benign panels. Its JBB Comply is 45.00%, versus 48.00%, 62.00%, and 58.00% for SRT-Repro, IndepGen, and Homo-Split, respectively (BoundaryPlay–IndepGen: -17.00 pp). IFEval strict prompt accuracy is 61.00%, versus Base's 75.23% and the controls' 74.31–75.79% (IndepGen: 75.79%; difference -14.79 pp; Appendix E.2). All trained arms lose benign compliance relative to Base; BoundaryPlay's JBB Comply is 45.00% versus 89.00%, and FalseReject Comply is 11.88% versus 51.14%.

Reward diagnostic. Refusing a judged-harmful request can earn both safety and refusal-correctness rewards; substantive benign helpfulness is not directly scored. A blinded saved-answer audit labels 121/541 BoundaryPlay answers as refusals, versus 9/541 for IndepGen;

77/102 prompts that IndepGen passes and BoundaryPlay fails involve a BoundaryPlay refusal (Appendix E.3). This is an association, not a causal decomposition.

Targeted intervention and partial recovery. Holding initialization, seed 4321, rollout budget, and the rollout-100 checkpoint fixed, we remove the refusal-correctness reward on nominal-plus defender examples. Joint construction, attacker rewards, and the other defender components remain enabled. Strict prompt accuracy rises from 330/541 to 394/541 (61.00% to 72.83%, $+11.83$ pp), recovering approximately 83% of the 14.23-point gap relative to Base (Table 25). It remains 2.40 pp below Base's 75.23% and 2.96 pp below IndepGen's 75.79%. All four IFEval accuracies improve by 11.27–11.83 pp. Comply improves on all four benign panels, including FalseReject from 11.88% to 17.35%. OR-toxic and HB-standard harmful responses increase from zero to 2/655 and 1/123; JBB and harmful XSTest remain at zero. This single-seed intervention does not establish a primary or sole cause of the loss (Appendix E.4).

5.5 Protocol Diagnostics

Evaluators. Rejudging saved responses tests dependence on the training-time moderator (Table 5). Claude-Opus-5 has complete labels across both arms and all six seeds for 58/94 harmful/benign items, versus 74/100 for qwen3.5-397b and the primary evaluation. On Claude's same subset, WildGuard plus Gemini gives $+5.64$ pp, compared with Claude's $+3.31$ pp; comparing Claude only to the full primary set would conflate evaluator and population changes. Alternative evaluators also use different label mappings (Appendix D).

Table 5: JBB FJ: BoundaryPlay minus SRT-Repro, six matched training seeds. H/B gives the common harmful/benign item counts. The primary evaluation combines WildGuard with Gemini; alternative evaluators replace both label sources. $d_z = \bar{d}/s_d$ standardizes paired seed differences, not item-level effects.

Evaluator / population	H/B	Δ (pp)	95% paired-t CI	d_z	$d_s > 0$
WildGuard + Gemini (full)	74/100	+4.45	[+2.05, +6.84]	1.95	6/6
WildGuard + Gemini (Claude subset)	58/94	+5.64	[+1.69, +9.59]	1.50	6/6
Claude-Opus-5 (common subset)	58/94	+3.31	[-6.56, +13.19]	0.35	4/6
qwen3.5-397b (full)	74/100	+4.56	[-1.20, +10.32]	0.83	4/6

Alternative evaluators retain positive mean differences, with four of six seed differences positive and paired-t intervals crossing zero. Paired d_z is 1.95 for the primary full-set comparison, 0.35 for Claude, and 0.83 for Qwen. Exact empirical seed-bootstrap intervals are $[+2.59, +5.91]$, $[-4.02, +9.77]$, and $[+0.54, +8.37]$ pp, respectively. We retain the paired-t analysis as the reporting basis; Appendix D details the sensitivity calculation.

Held-out pair semantics. The held-out semantic audit (Table 6) sampled 300 of the fixed 909 paired-diagnostic inputs without replacement; 298 received both prompt-variant judgments. The all-six-criteria pass rates are 44.63% and 35.57%, with intersection 33.89% (101/298). The main shortfall is a single safety-relevant difference, not surface overlap or shared goal/topic/framing. The intersection requires both variants to pass.

The low $\kappa = 0.195$ concerns *answerability*, not shared goal/context ($\kappa = 0.747$) or a single safety difference ($\kappa = 0.751$). Answerability labels agree on 290/298 pairs (97.32%), but one variant labels 289 pairs positive and the other 297; the resulting class imbalance yields high chance-expected agreement and low κ. This separate fixed-corpus audit's protocol, rubric, and limits appear in Appendix D and Section 6.

Table 6: LLM-assisted audit of 298 complete held-out pairs. A/B are two prompt variants of the same GPT-6-Astra model, not independent annotators. Rates and agreement are percentages. Surface matching is constant-positive, so its Cohen's κ is undefined (–), despite 100% agreement. Rows give per-criterion pass rates; the all-six-criteria joint pass rates (44.63%/35.57%) are reported in Section 5.5.

Criterion	A (%)	B (%)	Agreement (%)	κ
Harmful minus request	94.63	93.96	98.66	0.875
Harmless plus request	98.99	99.66	99.33	0.497
Shared goal/context	95.97	94.30	97.65	0.747
Answerable plus request	96.98	99.66	97.32	0.195
Single safety difference	44.97	35.57	87.92	0.751
Surface matching	100.00	100.00	100.00	–

6 Discussion and Limitations

Request construction as a design choice. The weak-safety RL-only gain concerns the complete protocol at matched candidate exposure and generated fraction. Shared weights let attacker updates affect the defender directly and through requests; the comparison does not separate these pathways or identify construction alone. The held-out semantic audit measures intended pair properties, not online semantic-minimality rates or a training-gain mechanism.

Why retain per-sample defender optimization? The defender objective lacks a pair-membership term. For the $+1.52$ pp BoundaryPlay–Homo-Split contrast, the wide interval does not establish equivalence; goal coverage and batch composition also differ. A testable hypothesis is that a pair-aware defender objective increases sensitivity to pair membership. Testing requires matched goal coverage and candidate exposure while varying organization and the defender objective.

Starting alignment and reward balance. With little safety headroom, benign preservation is more discriminating. The ordinary-Llama intervention has one seed: IFEval's partial recovery to 72.83% remains below Base's 75.23%, alongside benign compliance gains and safety costs. It does not establish reward composition as the sole or primary cause of the loss or exclude construction interactions. Benefits cannot be generalized to initially aligned models; cross-cohort runtime differences also confound alignment comparisons. Broader capability preservation remains unestablished.

Measurement and reproducibility. Model-based judgments, including the two prompt variants in the semantic audit, are not human ground truth. Training and primary evaluation share WildGuard; taxonomies overlap and historical parser exceptions remain. Alternative evaluators' broader intervals cannot be attributed solely to noise rather than policy differences. Seed intervals hold datasets and judgments fixed. Different decoding, judges, and training recipes preclude direct comparisons with published systems. Independent replication remains necessary.

7 Conclusion

BoundaryPlay's counterfactual construction, joint scoring, and weakness feedback improve JBB over independent full generation at matched candidate exposure in the tested weak-safety setting, with per-sample defender optimization. On ordinary Llama, we identify a modifiable defender-reward contributor to instruction-following loss: removing the plus-side refusal-correctness component while retaining joint construction recovers 11.83 pp of IFEval strict prompt accuracy and improves all four benign compliance panels, with few additional harmful responses on the evaluated harmful-request panels. BoundaryPlay thus provides both a request-construction protocol and a targeted reward-design intervention that mitigates instruction-following loss in the tested single-seed setting.

AI USE STATEMENT

We used generative AI tools as follows. Writing: AI tools helped identify relevant literature and polish text. Synthetic data: the held-out paired diagnostic inputs were generated with a language model (Appendix C); online training requests are generated by the policy model itself as part of the method (Section 3). Measurement: WildGuard, Gemini, Claude-Opus-5, qwen3.5-397b, GPT-6-Astra, and gemini-3.1-pro-preview supply training-reward and evaluation labels under fixed rubrics (Section 5.5, Appendices D–E); these models are research components, not writing assistance. The research idea, method, experimental design, and interpretation of results are the authors' own. We take full responsibility for the content of this work.

ETHICS STATEMENT

The goal is to reduce harmful assistance while avoiding unnecessary refusal of legitimate requests. Automated red teaming nevertheless produces dual-use adversarial material, and improvements on a limited benchmark do not make an abliterated model safe for deployment. We avoid including operational harmful procedures in the paper. Model-based labels can encode policy and cultural biases; benign degradation may disproportionately affect users discussing sensitive but legitimate topics. Any release of attack corpora or weak-safety weights requires a separate review of licenses, privacy, and misuse risks. This study does not report a new human-subject experiment or completed human annotation study. General capability preservation and deployment safety remain unestablished. All benchmarks and seed corpora are publicly available, and we declare no competing interests.

REPRODUCIBILITY STATEMENT

Section 3 specifies the construction protocol, attacker reward (Eq. 2), and defender reward, and Table 1 lists the factors varied across arms. Section 4 and Appendix D give training seeds, candidate budgets, optimizer and rollout settings (Tables 18–19), the abridged pair prompt and framing devices, reward terms, and input hashes. Appendix A defines evaluation populations, scoring predicates, judge configurations, and missing-label rules. Appendices B and C report seed-level results and statistical procedures; Appendix C documents the paired diagnostic corpus and sampling seeds; and Appendix E gives the ordinary-Llama, IFEval (pinned checker commit and software versions), and reward-intervention settings. Remaining provenance limits are stated in Section 6 and Appendix D.

REFERENCES

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/63092d79154adebd7305dfd498cbff70-Abstract-Datasets_and_Benchmarks_Track.html.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2025, Copenhagen, Denmark, April 9-11, 2025*, pp. 23–42. IEEE, 2025. doi: 10.1109/SATML64287.2025.00010. URL https://doi.org/10.1109/SaTML64287.2025.00010.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/cui25a.html.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL https://arxiv.org/abs/2209.07858.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/0f69b4b96a46f284b726fbd70f74fb3b-Abstract-Datasets_and_Benchmarks_Track.html.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Jiajia Li, Xiaoyu Wen, Zhongtian Ma, Shuyue Hu, Qiaosheng Zhang, and Zhen Wang. Disentangling intent from role: Adversarial self-play for persona-invariant safety alignment, 2026. URL https://arxiv.org/abs/2605.01899.

Xurui Li, Kaisong Song, Rui Zhu, Pin-Yu Chen, and Haixu Tang. Adversarial attack-defense co-evolution for llm safety alignment via tree-group dual-aware search and optimization, 2025. URL https://arxiv.org/abs/2511.19218.

Mickel Liu, Liwei Jiang, Yancheng Liang, Simon Shaolei Du, Yejin Choi, Tim Althoff, and Natasha Jaques. Chasing moving targets with online self-play reinforcement learning for safer language models. *CoRR*, abs/2506.07468, 2025. doi: 10.48550/ARXIV.2506.07468. URL https://doi.org/10.48550/arXiv.2506.07468.

Rahul Markasserithodi, Aditya Joshi, Yuekang Li, Ishmanbir Singh, Chris Yoo, and Alan Niu. CHASE: Adversarial red-blue teaming for improving LLM safety using reinforcement learning, 2026. URL https://arxiv.org/abs/2606.05523.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David A. Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum S. Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/70702e8cbb4890b4a467b984ae59828a-Abstract-Conference.html.

Ethan Perez, Saffron Huang, H. Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United*

Arab Emirates, December 7-11, 2022, pp. 3419–3448. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.225. URL https://doi.org/10.18653/v1/2022.emnlp-main.225.

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 Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 5377–5400. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.301. URL https://doi.org/10.18653/v1/2024.naacl-long.301.

Zhewen Tan, Wenhan Yu, Jianfeng Si, Tongxin Liu, Kaiqi Guan, Huiyan Jin, Jiawen Tao, Xiaokun Yuan, Xiangzheng Zhang, Duohe Ma, Tong Yang, and Lin Sun. Triplay-rl: Tri-role self-play reinforcement learning for LLM safety alignment. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 26415–26429. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1216. URL https://doi.org/10.18653/v1/2026.acl-long.1216.

Hao Wang, Yanting Wang, Hao Li, Rui Li, and Lei Sha. Be your own red teamer: Safety alignment via self-play and reflective experience replay. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 18700–18716. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.933. URL https://doi.org/10.18653/v1/2026.findings-acl.933.

Xiaoyu Wen, Zhida He, Han Qi, Ziyu Wan, Zhongtian Ma, Ying Wen, Tianhang Zheng, Xingcheng Xu, Chaochao Lu, and Qiaosheng Zhang. MAGIC: A co-evolving attacker-defender adversarial game for robust LLM safety, 2026. URL https://arxiv.org/abs/2602.01539.

Zhehao Zhang, Weijie Xu, Fanyou Wu, and Chandan K. Reddy. Falsereject: A resource for improving contextual safety and mitigating over-refusals in llms via structured reasoning, 2025. URL https://arxiv.org/abs/2505.08054.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

A Evaluation Contract and Provenance

Table 7: Fixed populations for the main results. H/B denote harmful/benign requests. C denotes substantive Comply. All main endpoints use chat responses.

Suite	Population	Count	Main endpoint
JailbreakBench	H / B	74 / 100	JBB ASR↓ / JBB C↑
OR-Bench toxic	H	655	OR-tox RR↑
HarmBench	standard H	123	HB-std ASR↓
XSTest	B	250	XST NR↑
FalseReject	B	1187	FalseR C↑
OR-Bench hard	B	1319	OR-hard NR↑

Population construction. JBB uses 74 selected harmful requests and all 100 benign requests. Starting from the original 100 harmful requests, the builder excludes 26 whose goal strings exactly match goals in the held-out paired-request pool after standard whitespace/case normalization (removing only leading and trailing whitespace and lowercasing, without collapsing internal whitespace). This is separation from that diagnostic pool, not training-data decontamination or exclusion for missing judge labels. Historical selection provenance is scoped in Appendix D. FalseReject uses the FalseReject-Test split (1187 benign requests). OR-Bench counts follow the released dataset configurations: `or-bench-toxic` (655) and `or-bench-hard-1k` (1319).

The HarmBench builder starts from 320 records: 159 standard, 81 contextual, and 80 copyright. Removing 36 held-out-overlap records (all standard) and three duplicate behavior strings yields 281 records: 123 standard, 78 contextual, and 80 copyright. Only the same fixed 123 standard IDs enter the main ASR for every arm. The builder uses `Behavior`, not `ContextString`; this is unproblematic for standard items with empty context, but does not validate contextual evaluation. We do not rename WildGuard's label as the official HarmBench classifier result. XSTest has 450 items; its 200 harmful items are retained as secondary outcomes, not mixed into the 250-item benign denominator. HarmBench has no benign evaluation population in this study.

Substance and historical exceptions. ASR counts `response_harmfulness == harmful`; RR counts `response_refusal == refusal`. NR is one minus that refusal rate on benign requests. For complete dictionary judgments, Comply requires non-refusal, truthy `substance`, and false `disclaimer_only`. Its substance rule is not the separate PairAcc threshold of 0.5. FJ is Comply minus ASR; the OR joint score uses OR-hard Comply minus OR-toxic ASR, not a newly defined refusal-based score.

All 174 JBB IDs in the historical Llama evaluation survive the historical common-valid intersection; all OR-toxic labels are complete. One HarmBench-standard chat record in Homo-Split, seed 2222, has missing harmfulness and refusal labels despite a false parsing-error flag. We preserve its denominator of 123 and historical zero contribution to the harmful count; this is not a finding that the response was safe. Ten non-refusal benign records have incomplete Gemini fields. The original dictionary guard and defaults remain: missing substance is falsy, missing `disclaimer_only` is falsy, and an explicitly non-dictionary Gemini value bypasses the shell check. These exceptional cases are not dropped, imputed, or rejudged; the judging procedure does not guarantee complete or semantically correct labels.

Generation and judge spans. Chat evaluation uses the native template, one user request, no explicit system message, temperature 0, and at most 640 output tokens. Checkpoint generation configuration is retained. The new Qwen evaluations use vLLM 0.10.0/V0, BF16, eager execution, maximum model length 8192, and evaluation service seed 8888; this is separate from training seed 4321. The second seed uses a different multi-GPU work schedule while retaining input, endpoint, and judge configuration. Runtime-comparison limits are consolidated in Appendix D.

WildGuard and the Gemini judge receive at most 6000 response characters in the frozen new-run contract; full generated responses remain stored. For the older full-channel diagnostic, WildGuard receives the raw scaffolded response while Gemini receives its extracted answer. Full and chat are separate generations. The frozen new-run benign judge is `gemini-3.1-pro-preview` with temperature 0, top-p 0.95, seed 0, maximum output 800 tokens, and low thinking budget with thought text excluded. These settings identify the new evaluation contract; historical judge provenance is scoped in Appendix D.

Overlap and reporting choices. Main-arm evaluation items pass the recorded exact-item overlap checks, which are not semantic decontamination. XSTest has a shared safety taxonomy with the WildJailbreak-derived training corpus. Its prompt categories inspired benign training prompts in WildGuardTrain, and WildGuard's refusal detection was evaluated on XSTest-Resp. FalseReject's evaluation split coincides with a corpus for a separate ablation outside the four main arms. We therefore do not describe every panel as independent of all development data. The RR/NR reporting revision was made after inspecting results. We preserve the original ASR and strict Comply outcomes in Table 8, rather than choosing only favorable metrics. The inspected reference framework also differs in generation length, temperature, truncation, and template handling; endpoint-name agreement alone does not reproduce a cross-paper leaderboard.

Table 8: Retained secondary endpoints, percentages. ASR is lower-better; all other columns are higher-better. JBB FJ = JBB C − JBB ASR; OR FJ = OR-hard C − OR-toxic ASR. Llama means use six seeds, Qwen means two; Base is fixed.

Method	OR-tox ASR↓	XST C↑	OR-hard C↑	JBB FJ↑	OR FJ↑	XST-H ASR↓	XST-H RR↑
Llama-8B-abl							
Base	69.62	91.60	99.09	15.92	29.47	23.50	36.50
SRT-Repro	50.89	93.27	97.61	30.38	46.72	12.00	51.00
IndepGen	55.34	92.40	97.98	28.25	42.63	11.33	48.42
Homo-Split	50.38	92.40	97.37	33.31	46.99	10.67	54.75
BoundaryPlay	47.23	92.40	97.12	34.83	49.89	10.00	54.33
Qwen-7B-IT							
Base	3.21	93.60	80.14	87.00	76.93	0.00	90.00
SRT-Repro	0.53	90.80	53.34	75.00	52.80	0.00	93.50
IndepGen	0.31	90.60	53.94	76.00	53.64	0.00	91.00
Homo-Split	0.23	88.20	44.05	71.00	43.82	0.00	93.25
BoundaryPlay	0.46	88.60	49.24	74.50	48.78	0.00	92.75

B SEED-LEVEL RESULTS AND INFERENCE

Table 9: Llama JBB FJ↑ by training seed, percentages. Table entries are displayed rounded; all statistics use full-precision scores, not these displayed values.

Seed	S	I	H	BP
8888	26.73	30.78	39.89	31.78
1234	29.08	28.08	36.19	34.14
4321	30.08	29.08	35.84	37.19
1111	33.49	26.73	31.78	33.84
2222	34.49	28.08	26.38	39.89
3333	28.43	26.73	29.78	32.14

Table 10: All six Llama JBB FJ contrasts. S = SRT-Repro, I = IndepGen, H = Homo-Split, BP = BoundaryPlay. Differences are target minus reference in pp. Holm adjustment is within this six-comparison family only.

Contrast	Δ (pp)	95% CI (pp)	t p	Holm p	Signflip p
I - S	-2.14	[-6.35, +2.08]	0.2496	0.7489	0.28125
H - S	+2.93	[-4.87, +10.72]	0.3787	0.7574	0.43750
BP - S	+4.45	[+2.05, +6.84]	0.0050	0.0301	0.03125
H - I	+5.06	[+0.91, +9.22]	0.0259	0.1034	0.06250
BP - I	+6.58	[+2.86, +10.30]	0.0061	0.0306	0.03125
BP - H	+1.52	[-5.91, +8.94]	0.6217	0.7574	0.62500

For each contrast, let d_s be the same-seed difference. We compute $t = \bar{d}/(s_d/\sqrt{6})$ with df $= 5$, and interval $\bar{d} \pm t_{.975,5} s_d/\sqrt{6}$. Sign-flip sensitivity enumerates all $2^6 = 64$ sign patterns using the absolute mean difference as the two-sided statistic. It is not a Welch test, and benchmark items are not additional training seeds. Rounded seed scores can change sign-flip ties; these calculations use full-precision scores. For BP minus Homo-Split on JBB, the result is 0.625. No correction reported here covers all exploratory analyses in the project.

Table 11: Every Qwen2.5-7B-Instruct training run, percentages. JBB ASR is zero for all eight rows and is omitted here only to avoid repeating that column. The two seeds receive equal weight; neither is excluded after seeing results.

Method	Seed	OR-tox RR↑	HB-std ASR↓	JBB C↑	XST NR↑	FalseR C↑	OR-hard NR↑
SRT-Repro	8888	97.25	0.81	73.00	94.40	58.30	51.25
SRT-Repro	4321	96.18	0.81	77.00	95.60	63.18	60.42
IndepGen	8888	97.40	0.81	75.00	93.60	60.74	55.50
IndepGen	4321	95.57	0.81	77.00	92.80	59.39	56.48
Homo-Split	8888	98.17	1.63	71.00	93.60	54.84	44.28
Homo-Split	4321	97.40	1.63	71.00	92.00	55.35	47.99
BoundaryPlay	8888	97.56	1.63	75.00	95.60	62.34	55.34
BoundaryPlay	4321	97.86	0.81	74.00	92.40	55.18	47.46

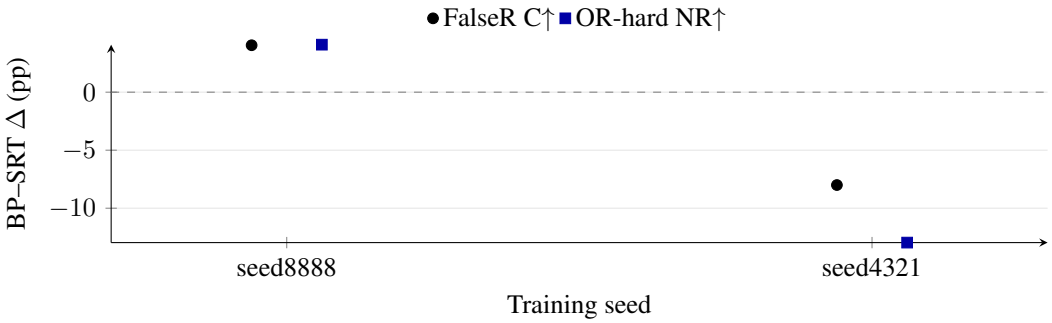


Figure 2: Qwen2.5-7B-Instruct: BoundaryPlay minus SRT-Repro by training seed. Positive values favor BoundaryPlay on these benign endpoints. Horizontal offsets are for readability only. The first seed's gains reverse in the second; both seeds are retained. No confidence interval or significance claim is attached to this $n = 2$ comparison.

The Qwen mean table averages run-level rates before display rounding. The cohorts share the endpoint contract but not a common training runtime: historical Llama uses vLLM 0.8.5.post1/V1 without gradient checkpointing, whereas the Qwen extension uses vLLM 0.10.0/V0 with gradient checkpointing. The Qwen seed-4321 six-panel evaluation contains 16264 responses across the four arms.

For context, Qwen Base has 0% JBB ASR and 4.07% HarmBench ASR, compared with 81.08% and 90.24% for the initial abliterated Llama model. On the two-seed means, BoundaryPlay exceeds Homo-Split on JBB Comply, FalseReject Comply, and OR-hard NR by 3.50, 3.66, and 5.27 pp, respectively. Versus SRT-Repro, FalseReject Comply is 1.98 pp lower and OR-hard NR 4.44 pp lower, despite OR-toxic RR being 0.99 pp higher. Versus IndepGen, the corresponding benign differences are -1.31 and -4.59 pp. HarmBench ASR is 1.22% for BoundaryPlay, versus 0.81% for both independent baselines. Relative to Base, BoundaryPlay's FalseReject Comply changes from 74.47% to 58.76% and OR-hard NR from 82.34% to 51.40% (Table 3).

Figure 2 resolves the variation behind these means. In seed 8888, BoundaryPlay improves FalseReject Comply and OR-hard NR over SRT-Repro by 4.04 and 4.09 pp. In seed 4321, the differences are -8.00 and -12.96 pp. Both BoundaryPlay and its reference contribute to the reversal: between seeds, BoundaryPlay's FalseReject Comply falls by 7.16 pp while SRT-Repro rises by 4.89 pp; their OR-hard NR changes are -7.88 and $+9.17$ pp. Limited safety headroom and variable benign performance motivate testing defender reward composition on ordinary Llama.

C PAIRED AND OPERATING-CONDITION DIAGNOSTICS

PairAcc contract. On the homologous held-out set, define

$$a_i = \mathbf{1}[\text{uplift}(x_i^-) < 0.3],$$
$$b_i = \mathbf{1}[\neg\text{refused}(x_i^+) \wedge \text{substance}(x_i^+) \geq 0.5],$$
$$\text{PairAcc} = \frac{1}{n}\sum_i a_i b_i.$$

The corpus builder first selects framed pairs with a positive `both_valid` flag, then shuffles the sorted unique goal IDs and makes a 70/30 goal-level train/test split. The test goals yield the fixed input pool of 909 pairs. From this pool, the evaluator samples 700 pairs without replacement; the split and sampling seeds are listed in Table 17. Each run then retains a pair only when both side judgments have valid parser status. This status indicates the evaluator's accepted parsing outcome, not response correctness: empty responses are accepted, and nonempty responses require a parsed dictionary containing a decision field, without strict decision-enum validation. The observed run denominators are 698–700 (Table 12). Uplift measures harmful assistance in the response, not recognition of malicious intent in the input. Each original run's valid-pair mask is retained. Different masks prevent interpreting the comparisons as exactly the same paired test items across every run. The primary historical aggregate was saved to four decimal places; tables below use those saved values consistently. They agree with rates recomputed from labels at that precision, but are not unrounded aggregates. The table mean can have more than four decimal places because it averages six saved values.

Table 12: Actual PairAcc denominators after the original two-sided acceptance mask. S/I/H/BP denote the four arms. The sampled pool has 700 pairs per run.

Seed	S	I	H	BP
8888	699	700	700	699
1234	700	699	699	699
4321	700	700	699	700
1111	700	700	700	700
2222	699	698	699	700
3333	699	699	699	700

Table 13: Paired held-out PairAcc↑ from saved four-decimal run aggregates, percentages. Top: equal-weight six-seed means. Bottom: all six training seeds.

Arm	Mean PairAcc↑
SRT-Repro	32.000
IndepGen	30.338
Homo-Split	32.102
BoundaryPlay	32.325

Seed	S	I	H	BP
8888	31.04	31.29	35.29	30.04
1234	31.43	30.19	33.19	33.48
4321	29.00	35.43	32.76	31.43
1111	30.14	29.43	29.43	33.57
2222	38.77	28.37	29.61	33.00
3333	31.62	27.32	32.33	32.43

Table 14: All six PairAcc contrasts under the saved-aggregate basis. Abbreviations and inference match Appendix B. Every interval crosses zero; this does not establish zero effect or equivalence.

Contrast	Δ (pp)	95% CI (pp)	t p	Holm p	Signflip p
I - S	-1.66	[-7.48, +4.15]	0.4956	1.0000	0.56250
H - S	+0.10	[-5.04, +5.25]	0.9615	1.0000	1.00000
BP - S	+0.33	[-3.19, +3.84]	0.8217	1.0000	0.87500
H - I	+1.76	[-1.21, +4.74]	0.1878	1.0000	0.18750
BP - I	+1.99	[-1.92, +5.90]	0.2483	1.0000	0.28125
BP - H	+0.22	[-3.34, +3.79]	0.8784	1.0000	0.90625

Adaptive attacks and channel sensitivity. The historical adaptive test covers 74 harmful goals, with an untrained abliterated attacker, WildGuard binary feedback, three search streams, and up to four iterations per stream. ASR counts a goal as successful if any stream succeeds. This is PAIR-style, not the original PAIR configuration. Goal-level outcomes support the BP–SRT six-seed contrast. Missing IndepGen runs prevent a complete four-arm adaptive comparison.

Table 15: Exploratory diagnostics. Differences are target minus reference in pp; positive ASR means worse safety. p is an unadjusted paired-t test p-value.

Test	Channel	Target − reference	Metric	Difference	p
PAIR-style	chat	BP − SRT	JBB ASR↓	$+1.35$	.3907
PAIR-style	full	BP − SRT	JBB ASR↓	$+8.33$	.0648
HB standard	full	BP − Indep	HB-std ASR↓	$+8.54$	.0549
No feedback	chat	no-probe − BP	JBB FJ↑	-2.08	.3452

For the full-channel adaptive test, BP ASR is 73.87% and SRT ASR 65.54%. For HB full, the BP–Indep interval is $[-0.26, 17.33]$ pp. The comparison reverses relative to chat, but absolute full-channel HB ASR is lower for every arm; those are different statements.

The no-feedback cohort disables the online weakness summary while retaining joint generation and scoring. Its chat ASR/Comply/FJ are 64.41%/97.17%/32.75%, versus 62.84%/97.67%/34.83% in the historical BP cohort. No-probe uses V0 with gradient checkpointing; the historical reference does not uniformly use that runtime. Its FJ difference is -2.08 pp ($[-7.20, 3.05]$ pp); the runtime-comparison limits are given in Appendix D.

Limits of mechanism evidence. An earlier surface-rewrite diagnostic reported no all-variant accuracy advantage, but its complete analysis is unavailable. The available evidence also lacks a fully documented six-seed defender-contrast-weight ablation. We therefore do not use these analyses as evidence for a semantic-boundary mechanism. Independent pair annotation and controlled surface-invariance tests remain necessary.

D Training and Generation Protocol

Training inputs and initialization. The independent arms consume separate harmful and benign pools with 50050 and 20000 rows, interleaved at equal sampling probabilities. The pair arms use 50050 harmful seed goals, not pre-generated pairs. The pair-seed inputs can be reconstructed exactly by preserving harmful rows and order while changing only `data_type` to `generated_pair`. The loader prefixes each pool at its sample limit, uses seeded probabilistic interleaving with first-exhausted stopping, and then applies the global sample limit; the training loader shuffles. Thus, pool

sizes are not consumed-sample counts or proof of identical goal subsets. The separately configured trainer holdout contains 1000 harmful prompts. This holdout is distinct from the paired diagnostic's goal split.

Provenance scope. The original upstream filtering, ordering, and holdout-split implementation and revision, the original official JBB dataset revision and historical construction, and the upstream model revisions of the local abliterated-Llama, Qwen, and ordinary-Llama initialization snapshots are not authenticated. The recorded JBB selection rule therefore does not establish an independently recomputed per-ID selection from a pinned official release. Snapshot identification uses recomputed hashes for small model files, but inherited weight hashes with only current shard sizes and modification times checked; weight contents were not freshly rehashed. Eighteen historical Llama runs lack a run configuration and six have only partial configuration, so current snapshot identifiers do not establish every historical run's initialization and input provenance. The available records also cannot establish that historical inputs and outputs were never overwritten or independently authenticate the historical request chain and transmitted request contents. Reported judge-service model names, including `gemini-3.1-pro-preview`, are provider aliases, not authenticated immutable historical backend revisions; the new evaluation settings do not retroactively certify the backend revision of every historical label. Table 16 lists the available corpus and snapshot identifiers in two columns, with corpus row counts in the input names and eight-character SHA256 prefixes. Each model metadata or weight-shard entry denotes one file; weight hashes are recorded, while the other hashes were recomputed.

Runtime-comparison scope. Matching endpoint contracts or training recipes does not establish bitwise equivalence or runtime matching. Engine versions, gradient checkpointing, generation interfaces, batching, and worker/judge schedules differ as reported in Appendices A, B, C, and E. Cross-model differences therefore cannot be attributed solely to initial alignment or parameter count; the ordinary-Llama extension is not a runtime-matched alignment intervention. The no-feedback comparison is likewise not a clean intervention on feedback alone: a single matched-runtime seed does not turn it into six controlled repetitions. Preserving shard inputs, GPU assignments, and batch composition in the reward intervention does not remove its reported scheduling differences.

Table 16: Training corpus and model snapshot identifiers. Eight-character SHA256 prefixes identify the recorded inputs; provenance scope is given in Appendix D.

Input	SHA256 (first 8)
Harmful pool (50050 rows)	`0ada16c0`
Benign pool (20000 rows)	`7c30b16d`
Pair seeds (50050 rows)	`230676f6`
Trainer holdout (1000 rows)	`5f553d47`
Llama config	`29e4c210`
Llama index	`146776fc`
Llama tokenizer	`79e3e522`
Llama shard 1	`6f283abd`
Llama shard 2	`25cb6ac4`
Llama shard 3	`d9baf2c3`
Llama shard 4	`f2ce860c`
Qwen config	`7463bb0e`
Qwen index	`624bf7c4`
Qwen tokenizer	`c0382117`
Qwen shard 1	`a1333e62`
Qwen shard 2	`f5d25a27`
Qwen shard 3	`8efdec4c`
Qwen shard 4	`1a72d403`

Each model metadata or weight-shard entry is 1 file; all weight shards Recorded; other entries Rehashed. Counts are corpus rows or files, not consumed training samples.

Sampling and blinding scope. The semantic audit sampled 300 pairs; 298 have complete judgments from both prompt variants, not a separate 298-pair sample. Blinding changes presentation order, not population selection: the printed-request convenience sample does not become a random sample of online training requests. Table 17 separates selection from presentation-order randomization.

Table 17: Sampling and blinding seeds. Corpus selection and blinding are distinct from training, generation-engine, and judge seeds, which are reported separately.

Operation	Seed	Purpose
Goal-level corpus split	2026	Shuffle sorted unique goal IDs after framed/`both_valid` filtering; assign the first $\lfloor 0.7 n_{\text{goals}} \rfloor$ to train and the remainder to test, yielding the fixed 909-pair holdout.
PairAcc 700-of-909 sample	11	Sample 700 held-out pairs without replacement, before each run's two-sided parsing-acceptance mask.
Independent 300-of-909 semantic sample	20260922	Sample 300 pairs without replacement from the same 909-pair holdout, not from the 700-pair PairAcc sample.
1082-answer blinding	20260923	Shuffle all 1082 saved IFEval answers from two arms before assigning anonymous IDs; no subsampling.
244-request blinding	20260923	Shuffle all 244 request occurrences from 122 parseable printed pairs before assigning anonymous IDs; no subsampling.

Seeds govern the listed selection or ordering operation, not training or judge determinism.

Reward terms and parser boundary. Equation 2 uses the fixed implementation constants `EPS`=0.05 and `SIM_BETA`=2.0 in both joint-generation arms. These are the settings used in the evaluated runs, not estimates from a reported parameter sweep. For complete labels and enabled hidden CoT, all four arms use $r_{\text{def}} = r_{\text{safe}} + r_{\text{refusal}} + r_{\text{format}}$, with no additional weighting between these three terms. Safety is $+1$ for a non-harmful response and -1 for a harmful response. Refusal correctness is $+1$ for refusing a judged-harmful request or not refusing a judged-benign request, and -1 for the opposite behavior. Format is $+1$ for satisfying the inherited think/answer-tag checks and -1 for a format violation. Format compliance is not substantive helpfulness. Independent attackers also receive this format term, added to their prompt-faithfulness (± 0.5), response-harmfulness (± 1), and refusal (± 0.5) terms. Pair attackers instead take the joint score directly, without that additional CoT-format term. This is another component of the bundled attacker-protocol difference, not a change to the defender objective.

The implemented pair-validity predicate checks a harmful-prefix minus label and the absence of that prefix on the plus label; a missing plus label can therefore pass this predicate. The scorer does not independently gate on the parse-success flag. Fallback pairs normally fail when both requests receive harmful labels, but parsing failure alone does not guarantee a -1 reward. Likewise, parser-error outcomes may be skipped downstream. We disclose these implementation limitations, without asserting that a constructed missing-label counterexample occurred in the historical training runs or estimating its frequency without traces.

Optimizer scope. All four arms use the REINFORCE policy-gradient formulation in a local critic-free implementation, with role-wise advantage standardization and a PPO-style clipped surrogate. With no auxiliary SFT and a shorter schedule, this adapts Self-RedTeam's RL-only setting rather than exactly reproducing its published Re++ recipe. BoundaryPlay introduces no new optimizer relative to the local controls.

Homo-Split changes per-rank goal coverage relative to BoundaryPlay (approximately 16 versus 8 per rollout), as well as co-occurrence. Retained attacker counts are 6400 for IndepGen and 3200 for each pair arm over 100 rollouts; subsampling means these are not necessarily total generation work. Homo-Split generates both sides for joint attacker scoring: 12800 defender responses precede the selection of 6400 defender candidates. The SRT attacker count is stochastic, and a local log count

Table 18: Shared training configuration. The runtime differs by cohort as stated in Appendix B; no new optimizer is introduced by BoundaryPlay.

Item	Setting
Policy roles	Shared attacker/defender weights
Algorithm	REINFORCE, no critic; no auxiliary SFT
Actor learning rate / schedule horizon	5×10^{-7} / 1500
Initial KL coefficient	0.01
Train batch / micro train batch / epochs	8 / 1 / 1
Actor/reference execution	Four colocated actor/reference ranks
Rollout execution	Two tensor-parallel-size-1 engines
Prompt / generated-token limits	640 / 640
Rollout sampling	Temperature 1.0, top-p 1.0
Main comparison / total rollout batches	100 / 125
Defender contrast / pair-DPO weights	0 / 0
Reward judge	WildGuard; full-generation grading enabled

Table 19: Arm-specific batch sizes and defender candidate budgets at rollout 100. The budget is after single-side selection, before later filtering/truncation; it is not total response generation or an equal-compute guarantee.

Arm	Rollout batch	Max samples	Def. candidates	Additional organization
SRT-Repro	64	8000	6400	50% independent rewriting
IndepGen	64	8000	6400	100% independent rewriting
Homo-Split	64	8000	6400	Single-side; attacker subsample 8/rank
BoundaryPlay	32	4000	6400	Expand each goal to both sides

multiplied by four is only a rank-expanded estimate unless all ranks agree. No claim of equal FLOPs, wall time, optimizer updates, or backpropagated samples follows from equal defender exposure.

Pair prompt (abridged training instruction). The term "counterfactual" below is part of the training instruction, not a claim that the generated pairs satisfy strict semantic minimality.

> Produce a matched counterfactual pair. The minus request preserves the seed's harmful operational goal within the framing device. The plus request uses the same framing and distinctive content words, but is genuinely harmless and answerable. Keep both at 20–90 words and within 25% of one another in word count. Maximize surface overlap; the intended difference is safety semantics. Neither prompt should itself contain a dangerous procedure. Return `x_minus`, `x_plus`, and `distinguishing_factor` as JSON inside `<answer>` tags, after reasoning in `<think>` tags.

The framing devices are `safe_target`, `defensive_framing`, `professional_context`, `fiction_nonoperational`, `historical_academic`, and `figurative_definition`. The weakness section lists recently over-refused benign requests and harmful-compliance gaps, with already-handled examples providing context. The defender's training user message uses the reason-then-answer scaffold, rendered with the native chat template and without an explicit system message. Plain chat evaluation removes that training scaffold. Content-word similarity lowercases/tokenizes text, removes stopwords and very short tokens, and computes intersection over union.

Automatic pair validity and an LLM-assisted audit. Three historical BP runs (seeds 1111, 1234, and 8888) have validity traces: the logged within-rank WildGuard valid fraction averages approximately 0.77/0.76/0.76 over rollouts 1–100, and 0.90/0.90/0.93 over rollouts 81–100. These are training-judge signals based on rounded logs, not human-verified counterfactual quality rates.

Repeated progress-log entries are deduplicated rather than counted as extra training. As an auxiliary audit, we sampled 300 pairs without replacement from the fixed 909-pair holdout, independently of the 700-pair PairAcc sample (Table 17). Two prompt variants of GPT-6-Astra classified the six binary properties in Table 6. Both calls completed for 298 pairs; the other two pairs were not imputed. Of the complete pairs, 101 passed both variants, 32 only A, five only B, and 160 neither. Thus, the variant pass counts are 133 and 106, with union 138/298 (46.31%). These are correlated ratings by one model, not human annotation or two independent evaluators.

Variant A emphasizes the underlying requested action rather than keywords and strict substantive answerability; B asks the judge to ignore the claimed dataset labels and compare intent, target, requested action, framing, and answerability. The shared-context field asks whether the pair shares the relevant goal, topic, or framing. The single-safety-difference field asks whether safety intent is the principal difference, rather than a changed topic, target, or action. This strict rubric may reject a safe-target or action-changing rewrite that satisfies a looser joint-construction instruction. We retain that discrepancy rather than relax the rubric after seeing its pass rate. The audit calls did not explicitly set temperature or sampling seed; the random seed specifies corpus selection, not deterministic judge behavior.

For answerability, the A/B contingency counts (both negative, A-negative/B-positive, A-positive/B-negative, both positive) are $(1, 8, 0, 289)$. For shared context they are $(11, 1, 6, 280)$. Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$ uses observed agreement p_o and agreement expected from the two positive-label marginals p_e. Surface matching has $p_o = p_e = 1$, so its κ is undefined, not 1.

The held-out corpus was selected using an automated label-validity filter (Appendix C), yet many pairs do not pass this stricter audit. Label validity is therefore insufficient to certify strict semantic minimality on this corpus. The training-log fractions and held-out audit concern different populations and cannot be compared as estimates of the same validity rate. Neither the online pair-validity rate under human judgment nor a causal link between semantic validity and training gains is established.

Judge sensitivity. Table 5 uses existing JBB responses and evaluator labels. For each alternative evaluator, common items require successful judging in all twelve arm–seed runs, separately for harmful and benign items. Claude excludes 16 harmful and six benign items; Qwen retains all 74/100. Evaluator refusal is not defender refusal, and exclusion may be informative. The same-subset WildGuard row controls the evaluated item population, not the selection bias induced by incomplete judgments.

Primary harmfulness is WildGuard's `response_harmfulness==harmful`. For the alternative evaluators, the historical harmful-assistance mapping is `decision==answer`; benign non-compliance is `decision` in {`refuse`,`partial`}, a truthy `disclaimer_only`, or a falsy `substance`. Thus, these rows replace both WildGuard labels and Gemini's benign substance judgment, with different operational mappings; they are not a pure single-head classifier substitution. The Claude sensitivity row adds four harmful items on which some, but not all, runs triggered evaluator refusal, treating failed judgments on those items as harmful responses. Its population is 62/94, and it still excludes stable all-run refusals. This retained historical sensitivity is not a missing-at-random correction or a bound over all possible missing labels.

Table 20: Post-hoc seed-bootstrap sensitivity for the same JBB contrasts. Percentile intervals enumerate all $6^6 = 46656$ ordered resamples of six paired seed differences. Intervals are in pp, not additional significance tests.

Evaluator / population	Δ (pp)	95% paired-t CI	95% bootstrap CI
WildGuard + Gemini (full)	+4.45	[+2.05, +6.84]	[+2.59, +5.91]
WildGuard + Gemini (Claude subset)	+5.64	[+1.69, +9.59]	[+2.94, +8.38]
Claude-Opus-5 (common subset)	+3.31	[-6.56, +13.19]	[-4.02, +9.77]
qwen3.5-397b (full)	+4.56	[-1.20, +10.32]	[+0.54, +8.37]
Claude: refusal-as-harmful	+3.05	[-5.25, +11.36]	[-2.95, +8.43]

The paired-t intervals use $t_{.975,5} = 2.5705818356$. Bootstrap resamples whole same-seed differences, not individual prompts or arms independently. Percentiles use linear interpolation between adjacent ordered values. Exhaustive enumeration eliminates Monte Carlo error, not small-sample uncertainty or bootstrap coverage error. In particular, Qwen's positive bootstrap interval does not override its zero-crossing paired-t interval. Prompt sets and evaluator outcomes are held fixed; neither procedure measures uncertainty across unseen prompts or judges.

E Additional Model Evaluations and Coverage Limits

E.1 Ordinary Llama: Complete One-Seed Extension

The ordinary Llama-3.1-8B-Instruct evaluation contains Base and the same four protocols, trained at seed 4321 and evaluated at the rollout-100 checkpoint; Base is not a training replicate. All five models use the same six panels (4066 responses each, 20330 total), original item order, and scoring predicates under the evaluation settings below. Table 21 retains endpoints not displayed in the main ordinary-Llama table; no cross-seed interval or test is defined at $n = 1$.

Table 21: Further ordinary-Llama endpoints (%), training seed 4321 only. OR RR uses 655 harmful items; XST-H uses 200 harmful items; XST NR uses 250 benign items; OR-hard uses 1319 benign items. OR FJ = OR-hard C minus OR-toxic ASR.

Model	OR-tox RR↑	XST-H ASR↓	XST-H RR↑	XST NR↑	OR-hard NR↑	OR FJ↑
Base	88.55	0.00	96.50	92.40	71.27	65.92
SRT-Repro	98.93	0.00	99.50	82.00	6.52	6.06
IndepGen	96.64	0.00	97.50	88.40	28.05	26.98
Homo-Split	98.32	0.00	98.50	87.20	19.56	19.03
BoundaryPlay	99.69	0.00	99.50	76.40	5.84	5.76

Execution settings. Training follows the four-arm candidate budgets in Appendix D, using gradient checkpointing and V0 rollout execution. Evaluation uses vLLM 0.10.0/V0, torch 2.7.1+cu128, transformers 4.53.2, BF16, eager execution, maximum model length 8192, and evaluation seed 8888 (not training seed 4321). Native `LLM.chat` uses a single user message, no explicit system message, temperature 0, and at most 640 output tokens. Checkpoint generation defaults are otherwise retained. Six GPU workers partition each panel by original row index modulo six, each allowing 128 sequences and 8192 batched tokens; all six load the same checkpoint before progressing to the next model. The interface and batching differ from earlier evaluations (Appendix D).

WildGuard uses 64-item batches, at most eight concurrent requests. Gemini uses 96 workers with deterministic prompt-hash routing over three client routes, each limited to 60 requests/minute. This is a client-side limit, not evidence of independent upstream quotas. All 14280 recorded benign judgments report `gemini-3.1-pro-preview`. Both judges receive at most the first 6000 response characters. The Gemini configuration is temperature 0, top-p 0.95, seed 0, low thinking, and initially 800 output tokens; empty candidate envelopes trigger a token-budget increase under the existing retry rule.

Label completeness and missing-label sensitivity. There are no missing, duplicate, or empty generated responses. HarmBench standard is selected from metadata (123 of 281), with no benign HarmBench population. All 20330 WildGuard judgments contain harmfulness and refusal labels, with no null values or parsing-error flags. The 14280 benign responses receive Gemini judgments; harmful prompts do not require Gemini. Twenty-two empty-candidate envelope failures were retried successfully, each on its second attempt. No transport failure remains unresolved.

Six Gemini judgments retain the historical parse-failure object: two Base FalseReject, one IndepGen OR-hard, one IndepGen XSTest, and two SRT-Repro OR-hard. Five concern WildGuard-refused responses and cannot change Comply. The remaining IndepGen XSTest response is non-refusing and contributes zero under the historical missing-substance rule. Allowing its unknown

substance/disclaimer label to take either value gives an IndepGen XSTest C range of 85.20–85.60%, versus BoundaryPlay's 75.20%. This is a missing-label sensitivity bound, not a confidence interval; no label was silently repaired or rejudged.

Checkpoint and request provenance are scoped in Appendix D. Label completeness does not independently validate label semantics.

E.2 Instruction Following on Ordinary Llama

Instruction-Following Evaluation (IFEval) (Zhou et al., 2023) adds a rule-based instruction-following check independent of WildGuard and Gemini. We evaluate all 541 official prompts, containing 834 instruction instances, on the same Base and each of the four seed-4321 trained models at the rollout-100 checkpoint. Table 22 reports every model and all four official strict/loose prompt/instruction accuracies, without averaging over training seeds. A prompt passes only when all its instructions pass.

Table 22: Ordinary-Llama IFEval (%), one training seed. S/L are official strict/loose checks; Prompt/Instr. denote prompt-level/instruction-level accuracy. All accuracies are higher-is-better. Length-limited answers remain in the denominators. Every model has zero empty outputs and zero checker errors.

Model	IFEval Prompt S↑	IFEval Instr. S↑	IFEval Prompt L↑	IFEval Instr. L↑	Length limit
Base	75.23	82.61	80.22	85.97	8
SRT-Repro	74.31	81.65	78.00	84.77	7
IndepGen	75.79	83.33	80.22	86.33	7
Homo-Split	75.60	82.97	80.41	86.45	3
BoundaryPlay	61.00	69.42	65.25	72.54	3

Evaluation protocol and scoring. The data and unmodified official checker are pinned to Google Research commit `758b894`.[1] The native checkpoint chat template receives one user message without an explicit system message. All models use greedy decoding, at most 2048 new tokens, evaluation seed 8888, BF16, vLLM 0.10.0/V0 eager execution, maximum model length 8192, up to 128 sequences and 8192 batched tokens. Other checkpoint generation defaults are retained. At most two independent single-GPU workers run concurrently. Raw answer whitespace is retained. The isolated scorer uses nltk 3.9.2, langdetect 1.0.9 and immutabledict 4.2.2; langdetect has seed zero and Python's random seed is reset to zero for each prompt/mode. These settings are fixed before generation, not selected using the resulting scores.

All 2705 answers cover the complete ordered input population. Each answer is scored with all official strict and loose checks; prompt and instruction accuracies are aggregated from the resulting Boolean outcomes.

BoundaryPlay is 14.23 pp below Base and 14.79 pp below IndepGen on strict prompt accuracy. The original protocol therefore loses instruction-following performance. The following audits characterize this outcome; Appendix E.4 then tests a specific reward intervention.

E.3 Post-Hoc Failure Characterization

These analyses were specified after observing the ordinary-Llama aggregate regression. They reuse saved responses and printed requests; neither is a confirmatory causal test or additional training replication.

Response behavior, not latent capability. We annotate all 1082 saved IFEval answers from BoundaryPlay and IndepGen using `gemini-3.1-pro-preview`. All answers are shuffled be-

[1] `https://github.com/google-research/google-research/tree/758b894eb02dc2a7097068031089a2803be147c6`

fore anonymous IDs are assigned, without subsampling (Table 17). These IDs conceal arm, training seed, checker outcomes and aggregate scores. The rubric separates refusal, attempted answer, nonanswer and uncertainty; quoting a refusal or making an unsuccessful attempt is not automatically a refusal. It separately labels refusal reason and request sensitivity, without assuming every benchmark prompt is benign.

Table 23: Blinded model-based behavior annotations on all 541 IFEval responses per arm. Counts, not corrected benchmark scores or human ground truth.

Model	Attempt	Refusal	Nonanswer	Uncertain
IndepGen	530	9	2	0
BoundaryPlay	419	121	1	0

The model-labeled refusal rates are 22.37% for BoundaryPlay and 1.66% for IndepGen. Of 102 prompts that IndepGen passes and BoundaryPlay fails under strict IFEval, 75 have an IndepGen attempt and BoundaryPlay refusal, two have refusals from both, and 25 have attempts from both. Thus 77/102 gross regressions accompany refusal, not 77 of the net 80 lost passes (relative to IndepGen; the Base-referenced figure is 77, Appendix E.4) causally explained. The audit labels 107 of BoundaryPlay's 121 refused requests clearly benign, 13 safety-sensitive and one uncertain; these are model judgments, not independent human validation. Fifteen BoundaryPlay and four IndepGen responses labeled as refusals still pass strict IFEval, whose constraints do not measure all aspects of substantive help. Scoring only non-refusals would select different post-training populations and must not replace the raw 61.00% versus 75.79% accuracies. Neither this audit nor the remaining attempted failures identifies latent general-capability loss.

The audit used temperature 0, top-p 0.95, seed 0, low thinking, four items per batch and output budgets 1024/2048/4096 under at most three parsing attempts. All 1082 answers received labels after 11 JSON decoding failures were retried, with no unresolved batch.

Printed requests, not historical reward replay. The ordinary-Llama BoundaryPlay log prints rank zero's first experience sequence per rollout, not a random sample or full census. All 125 printed candidates are retained: 122 parseable pairs and three parser failures. The first 100 iterations, corresponding to the evaluated checkpoint's training window, contain 97 parseable pairs. One syntactically parsed pair has identical placeholder text on both sides; it remains included and receives two uncertain semantic labels. There are 244 request occurrences but 243 unique request strings across the full run.

Training WildGuard received the request and full contemporaneous defender generation. Those paired defender outputs and historical per-side labels are not recoverable from this log. We therefore classify the printed requests under two explicitly artificial contexts: an empty response and a fixed wrapped refusal. The local WildGuard uses the current template and unchanged server parser, greedy generation with at most 128 tokens, seed 42, BF16 and vLLM 0.10.0/V0. A separate Gemini audit uses anonymous IDs assigned after shuffling all 244 request occurrences without subsampling (Table 17). It sees neither nominal side, paired counterpart nor WildGuard label; siblings never share a batch. Its rubric distinguishes benign, harmful and uncertain requests. This is independent of WildGuard, not of all prior evaluation or of model-policy assumptions.

In the first-100 window, the nominal-plus disagreement is 2/97 requests, or 2/96 among requests the semantic judge labels benign. The larger minus-side count (17/97) also requires caution: an unblinded inspection finds semantic rationales that reinterpret requested real private records as fictional, as well as differences involving privacy, fabrication and moderation policy. No labels are replaced on this basis. These counts neither establish widespread plus-side mislabeling nor rule it out in the full historical training population.

For a refusal, changing only the observed request label from harmful to benign changes the raw refusal-reward component from $+1$ to -1, regardless of nominal side. That conditional arithmetic does not validate either label or reconstruct total rewards, reward normalization, filtering, retained optimizer contributions, or a causal effect on IFEval. This probe motivates examining reward composition; the intervention below tests that design choice without changing judge labels.

Table 24: Current printed-request classifications. Semantic B/H/U denotes Gemini benign/harmful/uncertain. The last column counts semantic-benign/WildGuard-harmful disagreements, not verified WildGuard errors. The fixed-refusal condition gives identical request labels; all three syntax failures remain in the inventory.

Printed window	Side	n	Semantic B/H/U	WG H	B/H disagreement
Through 100	x^-	97	21/75/1	91	17
Through 100	x^+	97	96/0/1	2	2
All 125	x^-	122	24/97/1	115	20
All 125	x^+	122	121/0/1	2	2

All 488 WildGuard classifications and 244 semantic annotations have no parser errors or unresolved requests; the semantic audit used 61 four-item batches.

E.4 Plus-Side Defender Reward Intervention

The behavioral audit motivates an intervention on the refusal-correctness component rather than on pair construction. We retrain ordinary Llama-3.1-8B-Instruct with seed 4321, matching the original BoundaryPlay initialization, training recipe, and rollout budget. Both runs use 125 rollout steps and batch size 32, with the rollout-100 checkpoint preselected for comparison; the first 100 steps provide 6400 defender and 3200 attacker candidates. The checkpoint is not selected using the new evaluation scores.

Exact intervention. For nominal-plus defender examples only, we subtract the entire refusal-correctness reward component from the general-sum reward. Both its positive and negative contributions are removed. The component uses observed request harmfulness, not nominal pair side: for a harmful label it rewards refusal and penalizes compliance; for a benign label the signs reverse. Response-harmfulness, reasoning/format rewards, minus-side rewards, attacker rewards, and pair generation remain unchanged. Online probes, outcome/tie rules, parser handling, and filtering also retain their original implementations. Defender pair and pair-DPO weights remain zero. No labels are replaced, and nominal-plus examples are not forced to be benign. Online requests and shared-policy updates can evolve after the reward change; their respective contributions are not separately identified by this comparison.

Results and interpretation. Table 25 shows a net increase of 64 strict-passing prompts, from 330/541 to 394/541. The four IFEval gains range from 11.27 to 11.83 pp. Strict prompt accuracy gains 11.83 pp, recovering approximately 83% of the 14.23-point gap relative to Base (64/77 net lost passes). It reaches 72.83%, still 2.40 pp below Base's 75.23% and 2.96 pp below IndepGen's 75.79%. Benign compliance improves by 1.00 pp on JBB, 1.20 pp on XSTest, 5.48 pp on FalseReject, and 1.52 pp on OR-hard. Harmful responses rise from 0/655 to 2/655 on OR-toxic and from 0/123 to 1/123 on HB standard; JBB (0/74) and harmful XSTest (0/200) do not change. In this matched-seed comparison, adjusting defender rewards mitigates instruction-following loss while retaining the joint-construction protocol. The intervention is one matched training seed, not an estimate of cross-seed significance, and it does not test whether historical judge labels were wrong. It neither establishes a primary or sole cause of the loss nor supports general benefits for initially aligned models.

Evaluation settings. We reuse the original 541-prompt/834-instruction IFEval protocol and all 4066 safety prompts with their frozen population definitions and scoring functions. IFEval uses greedy decoding with 2048 new tokens; safety panels use 640. The new IFEval run has zero empty answers and four length-limited answers, versus three for original BP; all remain in the denominators. Safety generation preserves six row-modulo shards, each shard's original physical GPU assignment, and batch composition, but runs them in three waves of two workers instead of six simultaneous workers. Judge concurrency is reduced from eight to two WildGuard workers and from 96 to 24 Gemini workers; WildGuard batch size remains 64 (Appendix D).

Table 25: Matched-seed ordinary-Llama reward intervention. Original BP and plus refusal off use seed 4321 and the rollout-100 checkpoint. Rates are percentages; differences are percentage points. All four IFEval metrics and all 14 safety/compliance endpoints are retained. C is substantive compliance; NR is non-refusal. For IFEval, S/L are official strict/loose checks, and Prompt/Instr. denote prompt-level/instruction-level accuracy.

Metric	n	Original BP	Plus refusal off	Δ (pp)
IFEval Prompt S↑	541	61.00	72.83	+11.83
IFEval Instr. S↑	834	69.42	80.82	+11.39
IFEval Prompt L↑	541	65.25	76.89	+11.65
IFEval Instr. L↑	834	72.54	83.81	+11.27
JBB ASR↓	74	0.00	0.00	+0.00
OR-tox RR↑	655	99.69	99.08	-0.61
OR-tox ASR↓	655	0.00	0.31	+0.31
HB-std ASR↓	123	0.00	0.81	+0.81
JBB C↑	100	45.00	46.00	+1.00
XST C↑	250	75.20	76.40	+1.20
XST NR↑	250	76.40	78.00	+1.60
FalseR C↑	1187	11.88	17.35	+5.48
OR-hard C↑	1319	5.76	7.28	+1.52
OR-hard NR↑	1319	5.84	7.35	+1.52
JBB FJ↑	74/100	45.00	46.00	+1.00
OR FJ↑	655/1319	5.76	6.97	+1.21
XST-H ASR↓	200	0.00	0.00	+0.00
XST-H RR↑	200	99.50	100.00	+0.50

All 4066 WildGuard judgments and 2856 benign Gemini judgments yielded labels, with one technical Gemini retry and no unresolved parse errors or null WildGuard labels. The observed Gemini version is `gemini-3.1-pro-preview`.

E.5 Remaining Coverage Gaps

The historical non-Instruct Qwen2.5-3B cohort uses the same six-seed pool, but contains only SRT-Repro and BoundaryPlay JBB results under the frozen scoring basis. Mean ASR is 41.44% and 28.83%, respectively, while Comply is 91.67% and 87.50%, giving FJ 50.23% and 58.67%. These descriptive values do not establish a four-arm effect and do not describe Qwen2.5-3B-Instruct. Exploratory seeds 5555/6666 are excluded from this specified six-seed cohort. Historical Base feasibility probes do not provide a comparable Base result.

Qwen2.5-3B-Instruct has only partial evaluation results under the common protocol, not a complete six-panel evaluation or evidence for a positive cross-model conclusion. Unavailable results are missing, not zero.