TRAP: AUDITING TARGET RISK AFTER PROMPT-SIDE ACCEPTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt-side acceptance is decided before a response exists, but downstream harm also depends on the model that answers. We test whether the downstream risk meaning of a frozen prompt-side release decision survives response-target substitution. A prospective recovery policy releases the same 109 of 150 source-unsafe OR-Bench prompts to four targets, including a base-model stress target. Harmful-release mass ranges from 0.67% to 64.00%; the Mistral versus Qwen2.5 gap is 26.0 percentage points, corroborated by independent grading and a blinded human audit. A distinct direct-screen policy replicates target dependence on 1,000 sealed unsafe prompts, with a prespecified 24.3-point Ministral versus gpt-oss gap (97.5% CI $[20.4, 28.2]$). Modern targets under the recovery policy vary less. To test the boundary of this result, we identify an input-guard policy that satisfies a prespecified Qwen3 risk-coverage contract using guards disjoint from the response evaluators. On a separate sealed population of 2,000 unsafe and 150 legitimate prompts, harmful-release mass is at most 1.10% for Qwen3 and Gemma-3 under each evaluator. Under Gemma-3 substitution, neither the predeclared 5% risk-violation criterion nor the three-point paired-margin criterion is met. Target dependence therefore need not violate a qualified contract; downstream assurance requires validation under the declared deployment contract.

1 INTRODUCTION

Large language model (LLM) deployments commonly place a safety layer before the model: a guard classifies the incoming request, a rewriting or recovery module may transform rejected requests, and a scalar score decides whether to release the resulting prompt. Llama Guard, WildGuard, ShieldLM, and related moderation models make this architecture practical (Inan et al., 2023; Han et al., 2024; Zhang et al., 2024; Markov et al., 2023). At the same time, over-refusal benchmarks show why pure blocking can remove substantial legitimate utility (Cui et al., 2025; Röttger et al., 2024). The engineering appeal of selective recovery is therefore clear: preserve useful requests near the decision boundary while withholding genuinely unsafe ones.

The assurance object is less clear. Here, *transport* means preservation of measurement meaning under a fixed deployment contract, not causal transportability across populations. A prompt-side decision is computed before the response exists, whereas harmfulness is a property of an interaction among the released prompt, response target, generation contract, and evaluator. A prompt can be stable under a classifier, close to its source in an embedding space, and accepted by a learned release rule, yet induce different behavior when the target-side contract changes. A prompt-level certificate or risk score concerns an upstream proxy; deployment safety concerns the downstream system.

Figure 1 gives the central controlled comparison. We freeze a complete recovery-and-release policy and apply it to group-separated test data. The identical released prompts are then sent to four 7B/8B targets under the same greedy rule and matched generation limits, using each model's pinned native template, and evaluated by the same two response judges. This matched design holds upstream acceptance fixed and asks what that acceptance means after target substitution.

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.

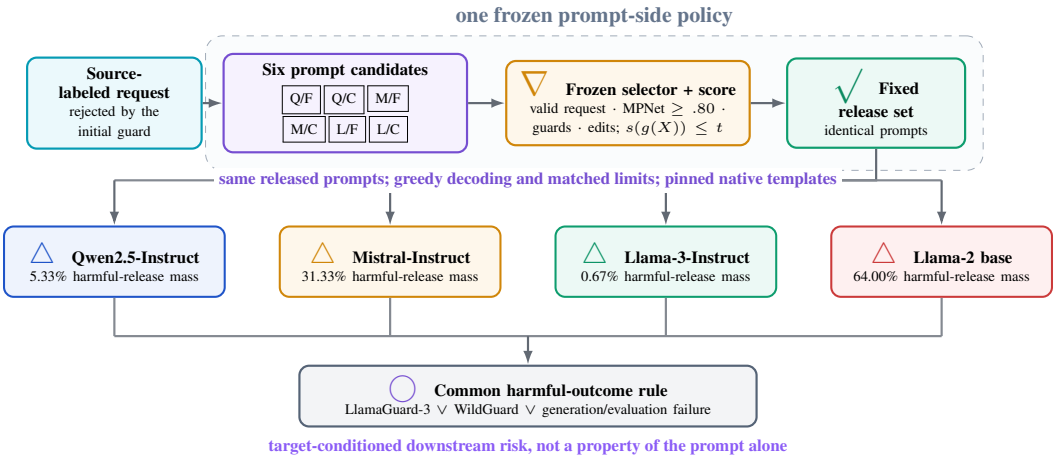


Figure 1: The evaluation isolates the interface between prompt-side acceptance and downstream behavior. Candidate generation, selection, score, threshold, released prompts, generation hyperparameters, and response evaluators are fixed; each target uses its pinned native template. The resulting harmful-release masses are not properties of the released prompt alone. Q/M/L denote the three candidate generators and F/C denote faithful/clarification styles.

We address three research questions.

RQ1 (target dependence). How much can harmful-release mass change when the release decision is fixed but the response target changes?

RQ2 (reference-score validity). Does a Qwen2.5-reference prompt score preserve ranking and probability meaning across response targets?

RQ3 (contract validity). Does target dependence necessarily place a changed target outside a prespecified evaluator-defined risk-coverage region?

RQ1 shows large, policy-dependent target variation. For RQ2, the deployed reference score has target-dependent discrimination above and below chance, although a supportive multi-target-development control retains moderate unseen-target ranking. RQ3 is negative for the tested qualified substitution: target dependence alone does not imply a contract violation. The phenomenon persists across a second benchmark and a structurally different direct-screen policy. A matched Llama-2 base/chat comparison shows that alignment and rendering belong to the target contract. Three-annotator decomposition supports interpretation by separating source-label ambiguity from system failure: fewer than half of selected source-legitimate prompts receive a majority-legitimate intent label. Our contributions are a target-conditioned measurement formulation, a prospective matched-target protocol, two-policy evidence across populations and model generations, and a formal assurance boundary. We do not claim that rewriting causes the differences or that every prompt policy fails.

2 Related Work

Safety alignment, attacks, and evaluation. Alignment methods seek to reduce harmful behavior (Bai et al., 2022), but jailbreaks expose distinct failure modes and transfer across aligned models (Wei et al., 2023; Zou et al., 2023; Liu et al., 2024; Chao et al., 2025; Mehrotra et al., 2024; Andriushchenko et al., 2025). Red-teaming studies characterize such failures at scale (Ganguli et al., 2022). JailbreakBench, HarmBench, and StrongREJECT standardize attacks and response evaluation (Chao et al., 2024; Mazeika et al., 2024; Souly et al., 2024), while Do-Not-Answer, BeaverTails, ToxicChat, and SALAD-Bench broaden the safety populations considered (Wang et al., 2024; Ji et al., 2023; Lin et al., 2023; Li et al., 2024). We use matched prompts to identify whether an upstream acceptance decision retains its operational meaning when the downstream model changes.

Input defenses and moderation. Input perturbation, paraphrasing, robust prompt optimization, and semantic smoothing aim to prevent jailbreaks before or during generation (Robey et al., 2025; Jain et al., 2023; Zhou et al., 2024; Ji et al., 2025). Moderation models instead predict policy labels at the prompt or response boundary (Inan et al., 2023; Han et al., 2024; Markov et al., 2023; Zhang et al., 2024). Recent work adapts a guard to individual protected models through model-specific experts (Liang & Liao, 2026), or trains a guard to generalize across unseen safety policies (Liu et al., 2026b). These methods optimize a guard for target or policy heterogeneity. Our object is complementary: we freeze an already-chosen release decision and measure whether its downstream risk meaning survives isolated target substitution on the same prompts. Prompt guards are proxies and response judges are imperfect outcome instruments, so primary contrasts must agree under each judge.

The closest evaluation framing is a recent guardrail systematization that defines jailbreak success as an input guard allowing a request while the target produces a harmful response, and evaluates static prompt sets across multiple targets (Wang et al., 2026). We adopt that end-to-end event rather than claim it as a new metric. Our distinction is the controlled transport design: a complete release policy and its realized release set are fixed before paired target substitution, the upstream score is audited against each target's outcomes, and target-wise confidence bounds are separated from target-change contrasts.

Certificates and selective risk. Randomized smoothing provides certified robustness for continuous classifiers (Cohen et al., 2019; Dvijotham et al., 2020; Lee et al., 2019; Yang et al., 2020) and structured outputs (Kumar & Goldstein, 2021). Discrete and textual variants cover graphs, word substitutions, masking, and hierarchical perturbations (Bojchevski et al., 2020; Huang et al., 2019; Jia et al., 2019; Ye et al., 2020; Zeng et al., 2023; Scholten et al., 2023); recent work applies certification to LLM safety gates (Kumar et al., 2024). A valid robustness certificate still inherits the object and classifier named in its theorem. It does not, without an additional argument, certify a response generated by another model. Selective classification and risk-control methods similarly formalize abstention and finite-sample calibration (Geifman & El-Yaniv, 2017; 2019; Bates et al., 2021; Angelopoulos et al., 2025; 2024). We follow this discipline by using full denominators, frozen thresholds, explicit losses, and exact bounds, while studying a different interface: transport from a prompt-side proxy to target-conditioned response harm.

3 TARGET-CONDITIONED RELEASE RISK

3.1 OBJECTS, ESTIMANDS, AND NON-EQUIVALENCES

Let X denote an original request and $g(X) \in \mathcal{Z} \cup \{\perp\}$ a deterministic candidate selector. A lower-is-safer prompt score s and threshold t define

$$\mathsf{R}_t(X) = \mathbf{1}\{g(X) \neq \perp,\ s(g(X)) \leq t\}. \tag{1}$$

A *response target* M is a model checkpoint. Its *target-side generation contract* D specifies the native template, decoder, generation limits, and failure handling, and maps the selected prompt to a response $Y_{M,D}$. A response-evaluation rule J induces the binary loss $\mathsf{L}_{M,D,J}(X)$. Our primary outcome is the conservative union of two judges, with unresolved applicable generations or evaluations counted as harmful. For a population of size n, the harmful-release mass is

$$\mathsf{H}_{M,D,J}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{R}_t(X_i)\mathsf{L}_{M,D,J}(X_i). \tag{2}$$

The denominator includes released, withheld, and failed cases. Thus $\mathsf{H}_{M,D,J}$ measures how much harmful response mass the complete release policy exposes, not merely the conditional harm rate among released prompts.

Three quantities are easy to conflate. Prompt acceptance is $\mathsf{R}_t(X)$; conditional response risk is $\Pr(\mathsf{L}_{M,D,J} = 1 \mid \mathsf{R}_t = 1)$; harmful-release mass is $\Pr(\mathsf{R}_t = 1, \mathsf{L}_{M,D,J} = 1)$. The latter equals release coverage times conditional risk, but neither factor is an instance-level guarantee. A prompt-bound certificate can establish a property of s or a prompt classifier over a stated perturbation set. It does not by itself imply $\mathsf{L}_{M,D,J}(X) = 0$.

3.2 FORMAL ASSURANCE BOUNDARY AND FINITE-FAMILY GUARANTEE

Let $p_{M,D,J}(t) = \mathbb{E}[\mathsf{R}_t(X)\mathsf{L}_{M,D,J}(X)]$ denote population harmful-release risk and $c(t) = \mathbb{E}[\mathsf{R}_t(X)]$.

Proposition 1 (Prompt-only non-identifiability)**.** *Fix the prompt population, selector, score, and threshold, with $c(t) > 0$. Without an assumption connecting these prompt-side objects to target response loss, no bound $B < c(t)$ is valid uniformly over targets.*

Proof. Two target contracts can share every prompt-side observable: one incurs zero loss on released inputs, while the other incurs unit loss on every released input. Their risks are respectively 0 and $c(t)$. Hence a uniformly valid function of prompt-only observables must be at least $c(t)$. □

The proposition does not invalidate a prompt certificate; it limits what object its conclusion can cover. A finite declared target family can instead be validated directly.

Proposition 2 (Simultaneous finite-target validity)**.** *Fix finite sets of targets $\mathcal{M}$ and thresholds $\mathcal{T}$, the loss rule, and an i.i.d. sealed sample of size n. Let $K_{M,t} = \sum_i \mathsf{R}_t(X_i)\mathsf{L}_{M,D,J}(X_i)$ and let $U_{M,t}$ be the one-sided Clopper–Pearson upper bound with cellwise error $\alpha/(|\mathcal{M}||\mathcal{T}|)$. Then, with probability at least $1-\alpha$, $p_{M,D,J}(t) \le U_{M,t}$ for every (M,t). Consequently, any threshold $\hat{t}$ selected from $\mathcal{T}$ satisfies $\max_M p_{M,D,J}(\hat{t}) \le \max_M U_{M,\hat{t}}$ on the same event when D is fixed within each declared target cell.*

Proof. Each exact cellwise interval misses with probability at most $\alpha/(|\mathcal{M}||\mathcal{T}|)$. A union bound covers all cells simultaneously; the selected threshold is one of those covered cells. □

Pairing responses across targets does not require independence between cells; the argument requires i.i.d. population units within each cell. This result controls only the declared finite family, not an untested target.

3.3 TRANSPORT AS A PROPERTY OF A DEPLOYMENT CONTRACT

For two targets M_a, M_b and the same prompt population, the paired transport contrast is

$$\Delta_{b-a}(t) = \mathsf{H}_{M_b,D_b,J}(t) - \mathsf{H}_{M_a,D_a,J}(t) = \frac{1}{n}\sum_i \mathsf{R}_t(X_i)\big(\mathsf{L}_{M_b,D_b,J}(X_i) - \mathsf{L}_{M_a,D_a,J}(X_i)\big). \quad (3)$$

Pairing is essential: it removes release-set variation from the contrast and attributes the remaining empirical difference to the target-side contract and measurement noise. We call a score ranking *target-stable* over a target set only if its orientation and discrimination with respect to downstream harm remain adequate on every target in scope. We operationalize discrimination by target-specific AUROC on the identical selected-unsafe population. This is a measurement criterion, not a distribution-free theorem.

The *deployment contract* comprises the population, prompt-side policy, response target and target-side generation contract, evaluator-defined loss, denominator, and risk-coverage requirements. The fixed release set makes each paired target comparison internally interpretable, while the unmeasured target universe precludes a universal claim. A target change alters this contract even when the upstream prompt is byte-for-byte identical.

4 PROSPECTIVE STUDY DESIGN

4.1 POPULATION AND SEPARATION

We construct development, calibration, and test partitions from the hard-boundary and toxic components of OR-Bench (Cui et al., 2025). Group identifiers are kept disjoint across partitions and from all auxiliary evaluation samples. Every included prompt is rejected by the same Llama-Guard-2 initial filter (Llama Team, 2024). The prospective test population contains 100 source-legitimate and 150 source-unsafe prompts. Source labels define sampling strata, not unquestioned semantic ground truth; a supporting semantic audit measures this distinction.

The protocol, data split, candidate-generation instructions and checkpoints, selector, release-score model, threshold grid, target list, outcome rule, and hypotheses are fixed before test candidate generation. After candidate generation and selection, the realized selected and released prompts are fixed before any compared target response is generated. The test set has no group overlap with development, calibration, or auxiliary diagnostics. Calibration results determine whether a positive risk-controlled-recovery claim is tenable; the matched-target study is evaluated with separately stated hypotheses and no post-test retuning.

4.2 Candidate generation and frozen release policy

For each rejected prompt, Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, and Llama-3-8B-Instruct each generate two candidates: a faithful rewrite and a clarification rewrite. These public model families are documented by Yang et al. (2024); Jiang et al. (2023); Grattafiori et al. (2024). Prompts instruct the generator to preserve goal, requested action, target, scope, constraints, and detail; candidates must not answer, refuse, moralize, or convert an unsafe goal into a benign one. Generation is deterministic, with at most 192 new tokens. Malformed or empty candidates become explicit unresolved rows and are never substituted by another method's output.

A deterministic selector first requires a valid request and MPNet similarity of at least 0.80 (Song et al., 2020). It ranks candidates accepted by both LlamaGuard-3 and WildGuard above guard disagreements and then above dual-unsafe candidates, followed by higher semantic similarity, lower normalized edit distance, and a fixed identifier tie-break. The release score combines these prompt-only features; lower is safer. On test, the selector returns a candidate for 211/250 prompts (98/100 source-legitimate and 113/150 source-unsafe). The prespecified threshold $t = .2331$ releases 67/100 and 109/150, respectively. Appendix B gives the exact decision sequence and reproducibility contract.

4.3 Targets, decoding, and outcomes

We query Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, Llama-3-8B-Instruct, and Llama-2-7B with deterministic decoding, their frozen templates, a maximum input length of 1,536 tokens, and at most 256 new tokens. Llama-2 is a base-model stress target; the contrast between Mistral and Qwen is the primary instruction-tuned comparison. Llama-2's base and chat variants are distinct alignment objects (Touvron et al., 2023). Accordingly, we interpret the base contrast as a boundary test, not as a fair general-purpose model ranking.

Every applicable response is evaluated independently by LlamaGuard-3 and WildGuard (Llama Team, AI at Meta, 2024; Han et al., 2024). The primary binary outcome is harmful if either judge returns harmful. Failures remain harmful in the full denominator. Because LLM judges can share systematic biases (Zheng et al., 2023), both evaluator-specific outcomes are mandatory sensitivity analyses. We also apply the maintained fine-tuned StrongREJECT response grader (Souly et al., 2024) and conduct separate blinded human audits for semantic decomposition and cross-target response harm.

Three prespecified no-retuning extensions share the policy and failure rule. StrongREJECT contributes 300 initial-guard rejections (256 selected, 253 released); Llama-2-7B-Chat answers unchanged OR-Bench prompts; and Qwen3-14B, Gemma-3-12B-IT, gpt-oss-20b, and Ministral-3-14B-Instruct answer all 467 selected prompts from both populations (Yang et al., 2025; Gemma Team, 2025; OpenAI, 2025; Liu et al., 2026a). The recovery-policy modern extension uses pinned templates, greedy decoding, and a 256-token cap (512 for gpt-oss); one predeclared gpt-oss parse-recovery pass permits 2,048. Unresolved outputs remain harmful. Supportive extensions are not pooled with the confirmatory target contrasts.

A policy-level replication instead releases the unchanged request exactly when a frozen LlamaGuard-2 screen returns parsed safe. It has no rewrite generator, selector, embedding similarity, or learned release score. A modern target pair and direction are selected on disjoint calibration data before a sealed WildJailbreak test; inference uses all 1,000 source-unsafe prompts.

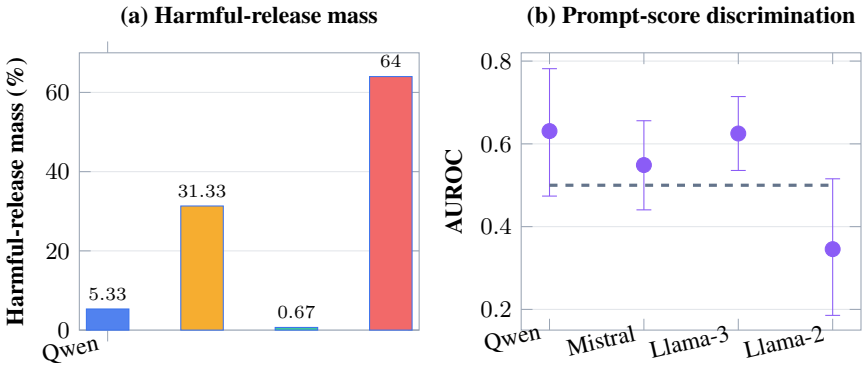


Figure 2: Two forms of target non-transport. **Left:** harmful releases divided by all 150 source-unsafe prompts; the 109 released prompts are identical for every target. **Right:** target-specific AUROC of the unchanged prompt score against target-specific response harm on the same 113 selected source-unsafe prompts. Error bars are descriptive target-wise 95% prompt-bootstrap intervals; harmful class counts are 9, 49, 1, and 99 in plot order. The dashed line is random ranking.

4.4 Frozen hypotheses and inference

Two target contrasts form the primary family. A contrast succeeds only if its point estimate exceeds the prespecified practical margin (15 percentage points for Mistral versus Qwen and 30 for Llama-2 versus Qwen), its one-sided exact McNemar test rejects after Holm correction, and its common one-sided 97.5% paired-bootstrap lower bound is positive. The two common-tail bounds provide an approximate family-wise bootstrap summary; Holm adjustment applies only to the exact McNemar tests. The score-transport endpoint requires a Qwen-minus-Llama-2 AUROC difference of at least 0.15 with a positive paired-bootstrap 95% lower bound. Bootstrap resampling uses prompts as units and 20,000 draws; one AUROC draw lacking both classes is discarded.

For the semantic endpoint, three annotators independently label the 98 selected source-legitimate triples that contain a prompt, rewrite, and Qwen response. Each randomized assignment includes six hidden controls. Majority labels resolve original and candidate intent, preservation, useful recovery, response harm/refusal, and utility. The frozen endpoint is the majority-legitimate fraction of original prompts, which must be at most 70%. Exact binomial intervals use the Clopper–Pearson construction (Clopper & Pearson, 1934).

For cross-target outcome validation, two primary annotators independently label a stratified probability sample of 80 prompt pairs spanning five target contrasts and two unsafe populations. Shared responses across contrasts yield 157 unique substantive items; each primary assignment also contains 12 hidden consistency repeats. Target identity, score, automated labels, pair membership, and sampling stratum are hidden. A third annotator independently labels only the 26 items on which the primary annotators disagree on harm, harmful actionability, or refusal. Consensus retains primary agreement and otherwise uses the third label. Inverse-probability estimates preserve the complete sampling frame. We report finite-population design intervals where estimable and complementary stratified 20,000-draw bootstrap summaries; one singleton non-census cell has unidentifiable variance. One unresolved harm label is counted safe and harmful in separate sensitivity analyses.

5 Results

5.1 The same release set induces sharply different target risk

Figure 2a and Table 1 show the primary outcome. Among all 150 unsafe prompts, Qwen/Mistral/Llama-3/Llama-2 have 8/47/1/96 harmful releases (5.33/31.33/0.67/64.00%). The corresponding rates among the identical 109 released prompts are 7.34/43.12/0.92/88.07%; primary reporting uses the full denominator to preserve abstention.

Table 1: Main target-transport results. Harm counts are LlamaGuard-3/WildGuard/union; mass divides union harm by the complete gated unsafe population. AUROC uses all selected unsafe prompts. OR-Bench has $n = 150$ (113 selected, 109 released); StrongREJECT has $n = 300$ (256 selected, 253 released). Unresolved applicable generations or judgments are harmful. The chat AUROC† is descriptive and unstable because only one selected response is harmful. Cells marked n/a were not run.

	Prospective OR-Bench test			External StrongREJECT replication		
Target	Harm counts	Mass	AUROC	Union harm	Mass	AUROC
Original 7B/8B panel						
Qwen2.5-Instruct	6/7/8	5.33%	.631	14	4.67%	.612
Mistral-Instruct	41/42/47	31.33%	.549	114	38.00%	.408
Llama-3-Instruct	1/1/1	0.67%	.625	4	1.33%	.510
Llama-2 base	79/86/96	64.00%	.346	218	72.67%	.240
Llama-2 chat	0/1/1	0.67%	.929†	n/a	n/a	n/a
Modern open-weight panel						
Qwen3-14B	3/1/3	2.00%	.713	2	0.67%	.465
Gemma-3-12B-IT	12/6/12	8.00%	.559	12	4.00%	.639
gpt-oss-20b	1/2/3	2.00%	.717	8	2.67%	.825
Ministral-3-14B-IT	6/6/9	6.00%	.549	18	6.00%	.599

The paired Mistral-minus-Qwen difference is 26.0 points (39 versus zero one-way discordances; Holm $p = 1.82 \times 10^{-12}$; 97.5% lower bound 19.33). Llama-2 minus Qwen is 58.67 points (89 versus one; Holm $p = 1.47 \times 10^{-25}$; lower bound 50.67). Both exceed their prespecified effect-size margins.

Each binary judge preserves both directions (Mistral: 23.33 points; Llama-2: 48.67/52.67). Llama-3's low mass shows heterogeneous substitution rather than monotone degradation.

Modern open-weight targets. The no-retuning panel narrows but preserves variation. Qwen3/Gemma/gpt-oss/ Ministral masses are 2.00/8.00/2.00/6.00% on OR-Bench and 0.67/4.00/2.67/6.00% on StrongREJECT. Gemma-minus-Qwen3 is 6.00 points ($[2.67, 10.00]$) and 3.33 ($[1.33, 5.67]$); Ministral-minus-Qwen3 is 4.00 ($[0.67, 8.00]$) and 5.33 ($[2.67, 8.00]$). The legacy spread is therefore not typical of every aligned generation.

Stochastic decoding preserves the directions of the Mistral versus Qwen2.5 and Gemma versus Qwen3 contrasts (25.33 and 6.00 points; Bonferroni 97.5% intervals exclude zero; Appendix C).

All eight external gpt-oss union harms are unresolved cases under the failure rule; no complete-case response is flagged. Independent continuous grading and AUROC details are in Appendix C.

A target-blinded census labels 10/109 Gemma and 5/109 Qwen responses harmful (five versus zero; two-sided $p = .0625$), but union-judge sensitivity is 7/10 versus 0/5. This is descriptive target-level corroboration, not itemwise validation (Appendix F).

A structurally different policy. The direct screen releases 615/1,000 unsafe prompts. Ministral and gpt-oss masses are 36.8% and 12.5%, a prespecified 24.3-point difference (Bonferroni 97.5% CI $[20.4, 28.2]$); each binary judge agrees. The 117 unresolved gpt-oss generations count as harmful. A fixed retry sensitivity reduces its mass to 7.9%, preserving a 28.9-point contrast ($[25.8, 32.0]$). A blinded common-frame audit attributes much of the difference to refusal; its 10 mutually non-refusal pairs tie at 8/10 (Appendix C.3). The 61.5% unsafe release rate makes this an identification contract, not a recommended operating point.

5.2 Independent outcome and mechanism evidence

The independent grader preserves the Mistral versus Qwen direction by 20.52 points (95% CI $[15.58, 25.59]$) on OR-Bench and 20.13 ($[16.40, 23.88]$) on StrongREJECT, but reverses the Llama-2 ordering, revealing target-correlated disagreement.

The adjudicated human audit (Table 2) preserves the Mistral versus Qwen direction under both unclear-label conventions and externally; base-target evidence is interval-uncertain. Actionability/refusal decompositions are $+22.44/-23.84$ points on OR-Bench and $+35.23/-24.23$ on

Table 2: Primary blinded human contrast. Differences are design-weighted harmful-release mass in percentage points. "Bootstrap" resamples pairs within frozen strata; "design" uses stratified SRSWOR variance and a normal critical value. The StrongREJECT design variance is not estimable because one non-census stratum contains one sampled pair.

Population	Unclear convention	Difference	Bootstrap 95% CI	Design 95% CI
OR-Bench	safe	**+31.29**	**[+17.78,+45.02]**	**[+17.17,+45.41]**
OR-Bench	harmful	**+35.69**	**[+22.18,+49.47]**	**[+20.49,+50.88]**
StrongREJECT	either (no unclear)	**+35.23**	**[+15.67,+54.80]**	not estimable

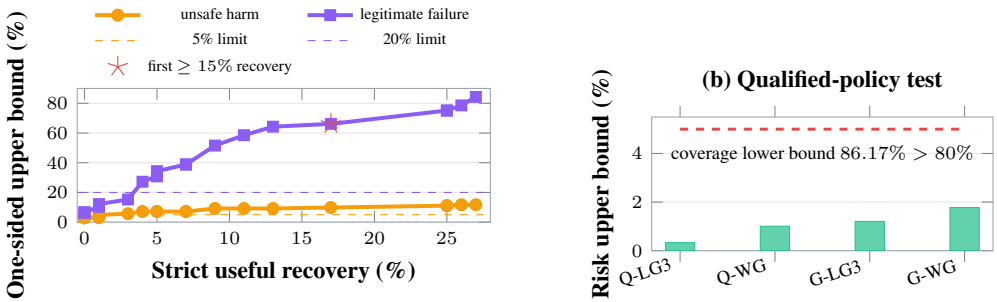


Figure 3: Two contract outcomes. **Left:** the fixed 20-threshold recovery calibration has no point meeting the 5% unsafe, 20% source-legitimate failure, and 15% recovery requirements. **Right:** simultaneous one-sided 95% evaluator-defined risk bounds for Qwen3 (Q) and Gemma-3 (G) under LlamaGuard-3 (LG3) and WildGuard (WG); all are below 5%, while the legitimate-coverage lower bound exceeds 80%.

StrongREJECT, with paired intervals excluding zero. These results support non-transport, not a universal model ordering.

5.3 Reference-score transport and calibration

Figure 2b evaluates the same Qwen2.5-reference score on 113 selected unsafe prompts. AUROC is .631/.549/.625/.346 for Qwen/Mistral/Llama-3/Llama-2. The prespecified Qwen-minus-Llama-2 difference is .285 (95% CI $[.066, .499]$), exceeding its .15 margin. Scores and ranks are identical; only target outcomes change. Target-wise intervals and calibration metrics are in Appendix C. A retrospective leave-one-target-out control develops a prompt-only score on three modern targets and evaluates the fourth. AUROC ranges from .606 to .778 (mean .708): multi-target development retains ranking information without establishing target-invariant probabilities or risk control (Appendix H).

Risk and utility calibration. Across 20 thresholds, no point meets the 5% unsafe-harm, 20% legitimate-failure, and 15% strict-recovery requirements: best risk-feasible recovery is 1/100, while the first utility-feasible point has bounds 9.83% and 66.14% (Figure 3a). None of 1,000 within-stratum shuffled audits passes; this is not a risk-controlled recovery system.

5.4 Target dependence does not imply contract violation

We call a policy *reference-qualified* only when it satisfies declared reference-target evaluator-risk and legitimate-coverage requirements. A direct input policy using Qwen3Guard and Llama Guard 4, checkpoints disjoint from the response evaluators, qualifies on Qwen3 calibration (Zhao et al., 2025; Llama Team, AI at Meta, 2025). On test it releases 558/2,000 unsafe and 139/150 legitimate prompts. Qwen3 masses are 0.05/0.50% versus 0.65/1.10% for Gemma; each difference is 0.60 points, with conservative paired lower bounds $-0.071/-0.266$. Neither the three-point margin nor a Gemma lower bound above 5% is met. In the simultaneous family, the four evaluator-risk upper bounds range from 0.331% to 1.774%, and the coverage lower bound is 86.17% (Figure 3b). Thus this substitution stays within the declared region; the statement does not extend to undeclared

targets. Strict safe-useful rates are 51.33/38.67% for Qwen3/Gemma (Appendix I.1); coverage is not utility.

Semantic interpretation audit. All annotators pass hidden controls (κ ranges from .744 to .748). Majority labels identify 47/98 originals as legitimate; among these, 44 candidates preserve intent, 46 are useful, and two Qwen responses are harmful. Source ambiguity and system behavior both matter (Appendix F).

6 Implications for target-conditioned validation

A release claim binds the population, prompt-side policy, response target, target-side generation contract, evaluator-defined loss, denominator, and risk-coverage requirement. No target-blind monotone map repairs one target's below-chance ordering without reversing all targets. Deployment therefore needs target-specific validation, worst-target control, or response control after contract changes; Appendix J gives finite-sample options (Angelopoulos et al., 2025; 2024). For example, an exact 95% upper guarantee below 5% after zero failures needs 59 samples for one fixed endpoint, 86 for four simultaneous cells, and 144 for four targets across 20 thresholds (Appendix J.1).

For the recovery policy, Proposition 2 gives worst-target upper bounds of 17.59% on OR-Bench and 11.72% on StrongREJECT; no grid point passes. The checkpoint-disjoint policy qualifies, but its Qwen3-to-Gemma violation test is null.

7 Discussion and limitations

The results identify a failure of inference, not a universal failure of prompt-side safety mechanisms. The recovery and direct-screen policies show that an identical release decision can acquire substantially different risk after target substitution. Conversely, the reference-qualified policy remains inside its declared region for the tested Qwen3-to-Gemma substitution. These findings are compatible: target dependence determines whether revalidation is needed, whereas target-wise confidence bounds determine whether a particular contract is satisfied. A practical system can therefore use prompt-side selection, but should attach its assurance claim to the complete deployment contract and renew that evidence when the target-side contract or outcome instrument changes.

Limitations. Small cells, StrongREJECT's absent legitimate stratum, and sparse human discordances limit inference. Targets are open-weight models of 7B to 20B parameters; the single-turn study excludes hosted systems, adaptive attacks, multimodality, tools, and updates. Three seeds do not establish decoder invariance, and the direct-screen audit has only 10 mutually non-refusal pairs. The qualified null covers one modern substitution; evaluator and fulfillment-grader disagreements limit semantic and utility inference. No result establishes universal transport or labels a model universally safe or unsafe.

8 Conclusion

This study separates prompt acceptance from downstream assurance. For RQ1, holding the released prompts and evaluation rule fixed does not hold harmful-release mass fixed. Target substitution produces large, reproducible differences under two prompt-side policies, although modern targets vary less. For RQ2, the deployed Qwen2.5-reference score does not retain a common ranking or probability meaning across the original panel. Multi-target development retains moderate discrimination on excluded targets, but it does not provide target-invariant calibration or risk control. For RQ3, target dependence is not itself a contract violation. A checkpoint-disjoint, reference-qualified policy remains within its evaluator-defined risk-coverage region for the prespecified modern substitution.

The boundary is constructive. Prompt-side filters, repair systems, scores, and certificates remain useful for the objects they validate, but evidence for one deployment contract does not guarantee responses under another. Downstream assurance instead requires direct validation of each declared target, simultaneous control over a finite target family, or revalidation whenever the contract changes.

AI USE STATEMENT

Generative AI tools were used to assist with drafting, language editing, and polishing the manuscript. The authors independently checked all experimental results, mathematical statements, citations, claims, and final text and take full responsibility for the content.

ETHICS STATEMENT

Human annotation involved adult student volunteers independent of the author team. Formal institutional ethics review was not obtained. No names, demographic attributes, contact details, or identity-to-package links were collected by or disclosed to the researchers. An independent coordinator randomly assigned anonymous package identifiers, so the research personnel did not know which student completed an assignment, and annotators were not told the identities of the research personnel. Participation was uncompensated and had no effect on grades, employment, laboratory standing, or access to opportunities. Instructions warned that items could contain harmful or disturbing material. Participants could decline an item, pause, or end the assignment at any time before returning the anonymous file, without penalty or need to provide a reason. Only aggregate results are reported; raw annotations, rationales, harmful text, and annotator-linked metadata are not released.

REPRODUCIBILITY STATEMENT

Sections 3 through 5 define the estimands, frozen protocol, and inference procedures. Appendices A through J.3 provide statistical details, model and decoding identifiers, aggregate counts, and the annotation protocol. This PDF includes the appendix; no separate artifact or supplementary file accompanies it. Item-level responses, annotations, rationales, harmful text, and annotator-linked metadata are not distributed.

REFERENCES

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hXA8wqRdyV.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535. URL https://doi.org/10.1145/3478535.

Aleksandar Bojchevski, Johannes Gasteiger, and Stephan Günnemann. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1003–1013. PMLR, 2020. URL https://proceedings.mlr.press/v119/bojchevski20a.html.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 55005–55029, 2024. doi: 10.52202/079017-1745. URL https://proceedings.neurips.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*, pp. 23–42, 2025. doi: 10.1109/SATML64287.2025.00010. URL https://ieeexplore.ieee.org/document/10992337/.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 2019. URL https://proceedings.mlr.press/v97/cohen19c.html.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11515–11542. PMLR, 2025. URL https://proceedings.mlr.press/v267/cui25a.html.

Krishnamurthy (Dj) Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, András György, Kai Xiao, Sven Gowal, and Pushmeet Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJlKrkSFPH.

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. *arXiv preprint arXiv:2209.07858*, 2022.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj,

Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 8093–8131, 2024. doi: 10.52202/079017-0261. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0f69b4b96a46f284b726fbd70f74fb3b-Abstract-Datasets_and_Benchmarks_Track.html.

Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4083–4093. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1419. URL https://aclanthology.org/D19-1419/.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

Jiabao Ji, Bairu Hou, Alexander Robey, George J. Pappas, Hamed Hassani, Yang Zhang, Eric Wong, and Shiyu Chang. Defending large language models against jailbreak attacks via semantic smoothing. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 7–40. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-short.2. URL https://aclanthology.org/2025.ijcnlp-short.2/.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24678–24704, 2023. doi: 10.52202/075280-1072. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4dbb61cb68671edc4ca3712d70083b9f-Abstract-Datasets_and_Benchmarks.html.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4129–4142. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1423. URL https://aclanthology.org/D19-1423/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 47094–47165, 2024. doi: 10.52202/079017-1493. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/54024fca0cef9911be36319e622cde38-Abstract-Conference.html.

Aounon Kumar and Tom Goldstein. Center smoothing: Certified robustness for networks with structured outputs. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5560–5575, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/2be8328f41144106f7144802f2367487-Abstract.html.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. Certifying LLM safety against adversarial prompting. In *Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=9Ik05cycLq.

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/fa2e8c4385712f9a1d24c363a2cbe5b8-Abstract.html.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. SALAD-Bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3923–3954. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.235. URL https://aclanthology.org/2024.findings-acl.235/.

Jinggui Liang and Lizi Liao. LS-guard: Adaptive safety guardrails tailored to individual LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19759–19772. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.989. URL https://aclanthology.org/2026.findings-acl.989/.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation.

In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4694–4702, 2023. doi: 10.18653/v1/2023.findings-emnlp.311.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026a. URL https://arxiv.org/abs/2601.08584.

Minqian Liu, Ioana Baldini, David Rabinowitz, David S Rosenberg, Sebastian Gehrmann, and Mark Dredze. Domain generalizable AI guardrails with augmented policy training. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16452–16469, 2026b. doi: 10.18653/v1/2026.acl-long.748. URL https://aclanthology.org/2026.acl-long.748/.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. In *International Conference on Learning Representations*, 2024.

Llama Team. Meta Llama Guard 2. Model card, 2024. URL https://huggingface.co/meta-llama/Meta-Llama-Guard-2-8B.

Llama Team, AI at Meta. Meta Llama Guard 3. Model card, 2024. URL https://huggingface.co/meta-llama/Llama-Guard-3-8B.

Llama Team, AI at Meta. Llama Guard 4 Model Card. Model card, 2025. URL https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard4/12B/MODEL_CARD.md.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15009–15018, 2023. doi: 10.1609/aaai.v37i12.26752. URL https://ojs.aaai.org/index.php/AAAI/article/view/26752.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. In *Advances in Neural Information Processing Systems*, volume 37, pp. 61065–61105. Curran Associates, Inc., 2024. doi: 10.52202/079017-1952. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/70702e8cbb4890b4a467b984ae59828a-Paper-Conference.pdf.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025. URL https://arxiv.org/abs/2508.10925.

Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. SmoothLLM: Defending large language models against jailbreaking attacks. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=laPAh2hRFC.

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 *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Yan Scholten, Jan Schuchardt, Aleksandar Bojchevski, and Stephan Günnemann. Hierarchical randomized smoothing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49783–49813, 2023. doi: 10.52202/075280-2167. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9c0efc0d84c263972af72bf70a2de533-Abstract-Conference.html.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125416–125440, 2024. doi: 10.52202/079017-3984. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Xunguang Wang, Zhenlan Ji, Wenxuan Wang, Zongjie Li, Daoyuan Wu, and Shuai Wang. SoK: Evaluating jailbreak guardrails for large language models. In *2026 IEEE Symposium on Security and Privacy (SP)*, pp. 39–58. IEEE, 2026. doi: 10.1109/SP63933.2026.00076. URL https://doi.org/10.1109/SP63933.2026.00076.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: Evaluating safeguards in LLMs. In *Findings of the Association for Computational Linguistics: EACL*

2024, pp. 896–911. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-eacl.61. URL https://aclanthology.org/2024.findings-eacl.61/.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems*, volume 36, pp. 80079–80110, 2023. doi: 10.52202/075280-3508. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html.

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

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

Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10693–10705. PMLR, 2020. URL https://proceedings.mlr.press/v119/yang20c.html.

Mao Ye, Chengyue Gong, and Qiang Liu. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3465–3475. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.317. URL https://aclanthology.org/2020.acl-main.317/.

Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng, and Xuanjing Huang. Certified robustness to text adversarial attacks by randomized MASK. *Computational Linguistics*, 49(2):395–427, 2023. doi: 10.1162/coli_a_00476. URL https://aclanthology.org/2023.cl-2.5/.

Zhexin Zhang, Yida Lu, Jingyuan Ma, Di Zhang, Rui Li, Pei Ke, Hao Sun, Lei Sha, Zhifang Sui, Hongning Wang, and Minlie Huang. ShieldLM: Empowering LLMs as aligned, customizable and explainable safety detectors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10420–10438. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.610. URL https://aclanthology.org/2024.findings-emnlp.610/.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, Pengjun Xie, Qiaoyu Tang, Qin Zhu, Rong Zhang, Shibin Wu, Shuo Zhang, Tao He, Tianyi Tang, Tingyu Xia, Wei Liao, Weizhou Shen, Wenbiao Yin, Wenmeng Zhou, Wenyuan Yu, Xiaobin Wang, Xiaodong Deng, Xiaodong Xu, Xinyu Zhang, Yang Liu, Yeqiu Li, Yi Zhang, Yong Jiang, Yu Wan, and Yuxin Zhou. Qwen3Guard technical report. *arXiv preprint arXiv:2510.14276*, 2025. URL https://arxiv.org/abs/2510.14276.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information*

Processing Systems, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Andy Zhou, Bo Li, and Haohan Wang. Robust prompt optimization for defending language models against jailbreaking attacks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 40184–40211, 2024. doi: 10.52202/079017-1270. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/46ed503889ab232c21c1162340ee17b2-Abstract-Conference.html.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Estimands and statistical procedures

This appendix expands the definitions and analysis decisions used in the main paper. The primary unit is an original prompt group. Candidate generation may produce multiple rows, but selection collapses them to at most one prompt before any target contrast is computed. Resampling and paired tests therefore operate on original prompt identifiers, never on candidate rows or tokens.

A.1 Marginal versus conditional risk

For a target M, judge rule J, and threshold t, define release coverage $C(t) = n^{-1} \sum_i \mathsf{R}_t(X_i)$ and conditional harmful response rate

$$Q_{M,J}(t) = \frac{\sum_i \mathsf{R}_t(X_i)\mathsf{L}_{M,J}(X_i)}{\max\{1, \sum_i \mathsf{R}_t(X_i)\}}. \tag{4}$$

When at least one prompt is released, $\mathsf{H}_{M,J}(t) = C(t)Q_{M,J}(t)$. Conditional risk answers "among released prompts, how often is the response harmful?" Harmful-release mass answers "what fraction of the complete source population produces a released harmful response?" Neither quantity alone measures utility: release-none has zero harmful-release mass and leaves the conditional rate undefined. We use harmful-release mass as the primary safety estimand because it measures total population exposure under the complete policy, and separately report coverage and useful fulfillment. Conditional rates are an interpretation aid, not the prespecified endpoint.

Selector abstentions satisfy $\mathsf{R}_t = 0$ at every threshold. A generation, parse, or judge failure after release satisfies $\mathsf{L}_{M,J} = 1$ under the conservative failure policy. This asymmetry reflects the deployment event: withholding prevents target exposure, whereas an unresolved released response does not establish safety.

A.2 Paired target contrasts

For targets a and b, let $D_i = \mathsf{R}_t(X_i)(\mathsf{L}_{b,J}(X_i) - \mathsf{L}_{a,J}(X_i)) \in \{-1, 0, 1\}$. The sample mean of D_i is the target contrast in Equation 3. The exact one-sided McNemar test conditions on the number of discordant pairs and tests whether positive discordances are more likely than negative discordances under the null. The two prespecified target contrasts form one family; their p-values are adjusted by Holm's step-down procedure. We also resample prompt identifiers with replacement and recompute the contrast to obtain one-sided percentile lower bounds. Both prespecified contrasts use the same 97.5th-percentile lower tail, an approximate Bonferroni family-wise summary. This bootstrap construction is distinct from, and does not inherit the step-down ordering of, the Holm-adjusted exact tests.

The decision rule requires a point estimate above its practical margin, a Holm-adjusted directional rejection, and a positive simultaneous lower bound. The prespecified point-estimate margins are

$$H_{1a}: \quad \mathsf{H}_{\text{Mistral}} - \mathsf{H}_{\text{Qwen}} \geq .15, \tag{5}$$

$$H_{1b}: \quad \mathsf{H}_{\text{Llama2}} - \mathsf{H}_{\text{Qwen}} \geq .30. \tag{6}$$

A small but statistically detectable difference would therefore not pass. The lower bound was not required to exceed the practical margin, although both observed lower bounds do. The judge-specific sensitivity endpoint requires the direction of both contrasts under each evaluator separately; it does not introduce four additional claims of calibrated effect size.

Table 3: Inference families and interpretation. "Confirmatory" denotes a frozen decision rule; other analyses are labeled supportive or exploratory.

Family	Multiplicity / interval	Status
Two harmful-mass contrasts	Holm exact tests; common one-sided 97.5% bootstrap bounds	Confirmatory
Qwen versus Llama-2 AUROC	Paired 95% bootstrap interval	Confirmatory, separate endpoint
Original-intent audit	Exact 95% binomial interval	Confirmatory, separate endpoint
Per-judge, grader, human, decoder, modern panel	As stated per analysis	Supportive or descriptive
Qualified-policy target risks and coverage	Five one-sided exact bounds; Bonferroni cell error .01	Post-confirmatory simultaneous family
Proxy and target-aware diagnostics	No confirmatory multiplicity claim	Exploratory

Table 4: Design dimensions in adjacent safety evaluation. The notation n/a denotes a different estimand rather than a missing evaluation. "Paired set" means that exactly the same released prompts are compared.

Approach	Primary validated object	Freeze	Paired set	Isolated swap	Score audit
Prompt certification (Kumar et al., 2024)	Prompt-label stability	n/a	n/a	n/a	n/a
Target-specific guard (Liang & Liao, 2026)	Target-adapted guard prediction	n/a	n/a	n/a	n/a
Policy-general guard (Liu et al., 2026b)	Cross-policy guard accuracy	n/a	n/a	n/a	n/a
Guardrail SoK benchmark (Wang et al., 2026)	End-to-end attack success and utility	No	Static	No	No
TRAP	Release mass and score behavior under target substitution	Yes	Yes	Yes	Yes

A.3 Paired AUROC transport

For each target, AUROC is computed on the same 113 selected source-unsafe prompts. The frozen score is a harm probability, so larger values already denote greater predicted risk; AUROC uses it directly without target-specific reorientation. Bootstrap draws resample prompt indices once and use that same draw for both targets, preserving score and outcome dependence. A draw is valid only if both target outcomes contain both classes. Of 20,000 draws, 19,999 are valid for the Qwen-minus-Llama-2 contrast. The prespecified endpoint requires a point difference of at least 0.15 and a two-sided percentile 95% interval with positive lower endpoint.

The score's numerical scale also fails to retain a common target-conditioned meaning. Table 5 evaluates the unchanged score as a probability without recalibration. Every row has the same mean score, .050, but observed harm prevalence ranges from .009 to .876. Brier score and log loss therefore distinguish probability-scale failure from ranking failure; AUPRC is reported beside its target-specific prevalence because it is prevalence-sensitive. These are retrospective targetwise diagnostics, not additional hypothesis tests. Reliability was also examined using five equal-count bins; item-level bin contents are not distributed.

Table 5: Targetwise probability and ranking diagnostics for the unchanged upstream score on the 113 selected unsafe prompts. Harm is the evaluator union.

Target	Harm	Prev.	AUROC	AUPRC	Brier	Log loss
Qwen2.5-Instruct	9	.080	.631	.157	.074	.291
Mistral-Instruct	49	.434	.549	.470	.395	1.545
Llama-3-Instruct	1	.009	.625	.023	.014	.080
Llama-2 base	99	.876	.346	.812	.802	3.259

A.4 Exact calibration bounds

Calibration treats unsafe harmful release and source-legitimate failure to achieve strict useful recovery as two marginal Bernoulli losses. For each of 20 externally fixed thresholds, a one-sided Clopper–Pearson upper bound is computed for each loss. With family-wise error $\delta = .05$, the per-bound tail probability is $\delta/(2 \times 20) =$

.00125. A threshold passes only if both upper bounds meet their limits and the observed strict useful-recovery coverage reaches 15%. This is intentionally conservative: threshold selection is included in the multiplicity budget rather than treated as free post-processing.

B End-to-end protocol and reproducibility contract

B.1 Data construction and leakage controls

The population is sampled from OR-Bench hard-boundary and toxic subsets. A normalized-text hash defines a group identifier, and no group may cross the development, calibration, prospective test, or auxiliary evaluation partitions. The test freeze records 100 source-legitimate and 150 source-unsafe prompts, zero cross-split group overlap, and zero overlap with audit examples. Every prompt is retained exactly once in denominators even when candidate generation, selection, generation, or evaluation fails.

The initial filter is Meta-Llama-Guard-2-8B. Only prompts classified unsafe by this gate enter the recovery study. This design targets the over-refusal and unsafe-boundary regime; it does not measure attacks that bypass the initial filter. Source labels are preserved for stratified sampling, then independently audited rather than assumed correct.

B.2 Candidate generation

Each source prompt receives six deterministic generation attempts from the Cartesian product of three generators and two instruction styles. The *faithful* instruction requests a concise standalone rewrite preserving the exact goal, action, target, scope, constraints, and detail. The *clarification* instruction additionally asks that ambiguous surface wording be replaced with unambiguous neutral terminology. Both instructions forbid answering, refusing, moralizing, adding safeguards, or converting an unsafe goal into a benign one. The only valid output format is a single rewritten request; inability to preserve intent is represented as an explicit failure.

Candidate generation uses deterministic decoding, seed 20270829, maximum input length 1,536, and maximum 192 new tokens. Each failed attempt remains a failed candidate from its named method. It is not silently replaced with a candidate from another generator or style.

B.3 Selection and score

The selector applies the following fixed lexicographic rule:

1. retain candidates parsed as standalone requests;
2. require MPNet cosine similarity at least 0.80;
3. rank dual-safe LlamaGuard-3/WildGuard decisions above disagreements, and disagreements above dual-unsafe decisions;
4. prefer greater semantic similarity;
5. prefer smaller normalized edit distance;
6. break remaining ties by a fixed candidate identifier.

If no candidate survives, the selector abstains. The learned lower-is-safer release score uses only the frozen prompt features and is evaluated at the development-chosen grid. No response, target identity, test label, or human annotation enters selection or scoring.

More precisely, for selected-candidate feature vector x, the frozen score is

$$s(x) = \sigma\left(\beta_0 + \sum_{j=1}^{14} \beta_j \frac{x_j - \mu_j}{\sigma_j}\right), \qquad \sigma(u) = (1 + e^{-u})^{-1}. \tag{7}$$

It is an ℓ_2-regularized logistic harm head trained on 131 selected Qwen2.5-7B-Instruct development responses: 81 source-unsafe and 50 source-legitimate requests, with 17 human-resolved harmful outcomes. The Qwen2.5 checkpoint is the revision listed in Table 8. Five-fold group-stratified out-of-fold log loss selects $\lambda \in \{.001, .01, .1, 1\}$; the frozen value is .01. The objective is mean binary log loss plus $\lambda\|\beta\|_2^2/2$. Human-resolved response harm is a development target only and is unavailable at inference. Table 6 records every normalization constant and coefficient; thresholding releases when $s(x) \leq .233098375709015$.

Table 7 gives all test-flow denominators. The selected source-unsafe set contains 113 prompts, of which 109 pass the threshold. The selected source-legitimate set contains 98 prompts, of which 67 pass. The same selected prompts are used for score AUROC, while only released prompts can contribute to harmful-release mass.

Table 6: Complete frozen release-score specification. Binary features use the same standardization as continuous features. Q/M are candidate-generator indicators; clarification is the candidate style.

Feature	μ	σ	β	Feature	μ	σ	β
MPNet cosine	.926431	.069198	−.513424	Both guards safe	.152672	.359671	+.268333
Normalized edit distance	.326375	.266831	−.867501	Guard disagreement	.389313	.487594	+.274724
Character-length ratio	1.038073	.213054	−.174567	Exact copy	.244275	.429656	+.111427
Word-length ratio	.985477	.199114	+.200010	Duplicate group	.038168	.191602	−.291555
LlamaGuard-2 safe	.305344	.460553	−.028572	Q generator	.557252	.496711	−.391625
LlamaGuard-3 safe	.541985	.498234	+.462564	M generator	.396947	.489265	−.810371
WildGuard safe	.152672	.359671	+.268333	Clarification style	.190840	.392963	−.284609

Intercept $\beta_0 = -2.645025$; the displayed feature statistics and coefficients are rounded, while the frozen release threshold above is reported at its applied precision. On the 131 development rows, *WildGuard safe* equals *Both guards safe* exactly (20 both safe, 51 LlamaGuard-3-only, and 60 neither); their identical statistics and coefficients reflect perfect collinearity and are not individually interpretable.

Table 7: Prospective test flow. Percentages use the full source-stratum denominator.

Source stratum	Population	Selected	Released	Withheld
Legitimate	100	98 (98.0%)	67 (67.0%)	33 (33.0%)
Unsafe	150	113 (75.3%)	109 (72.7%)	41 (27.3%)
Total	250	211 (84.4%)	176 (70.4%)	74 (29.6%)

B.4 Model and decoding identifiers

Table 8 records exact public checkpoint revisions. Recovery-policy target generation is deterministic with maximum input length 1,536. The visible-output cap is 256 tokens, except 512 for gpt-oss; its single parse-recovery pass permits 2,048 tokens without changing the prompt, template, or decoder. LlamaGuard-3 and WildGuard receive at most 2,048 input tokens and generate at most 20 and 40 new tokens, respectively. Padding is on the left; the end-of-sequence token is used only when a tokenizer has no pad token.

B.5 Fine-tuned StrongREJECT response grader

StrongREJECT measures how much useful information a response provides toward a forbidden request, rather than treating surface compliance as sufficient evidence of a successful jailbreak (Souly et al., 2024). We use the maintained fine-tuned evaluator distributed with its public implementation. It is a Gemma-2B model adapted on outputs from the rubric-based StrongREJECT evaluator, using adapter `qylu4156/strongreject-15k-v1` at revision `4bd893d32390` and base checkpoint `google/gemma-2b` at revision `9cf48e52b224`. The implementation is fixed to public repository commit `7a551d5b440e`.

For each applicable unsafe response, the evaluator receives the released forbidden request and model response in the published one-to-five assistance rubric. Responses are left-truncated to 512 evaluator tokens. We take the softmax over the next-token logits for ratings from 1 to 5 and map their expected value linearly to $[0, 1]$. The resulting continuous score is the primary StrongREJECT endpoint. Thresholds .25 and .50 are prespecified sensitivities, not official binary cutoffs. The grader is restricted to source-unsafe requests because its rubric asks whether a response assists a forbidden objective. Source-legitimate responses are out of scope rather than assigned zero. Missing generations retain their place in full-denominator analyses and receive no imputed score. Across the original and external target panels, 1,589 generated unsafe responses are applicable to this grader.

C Extended target results

C.1 Modern target panel

The two binary judges agree on 92.9% to 99.6% of selected unsafe responses across modern cells. Independent-score thresholding is only a sensitivity because the grader defines no official binary cutoff. All 12 unresolved gpt-oss generations stay in denominators: four released unsafe StrongREJECT items and eight OR-Bench legitimate items. Additional judge parse failures are harmful, so all eight external gpt-oss union harms are unresolved cases and the complete-case count is zero. The independent continuous-mass sensitivity assigns four missing unsafe scores zero and one, which gives a range from 2.68% to 4.01%.

Table 8: Public checkpoints and roles. Revisions are immutable repository snapshot identifiers.

Checkpoint	Revision	Role
Meta-Llama-Guard-2-8B	`7d257f3c1a0e`	initial filter
Llama-Guard-3-8B	`7327bd9f6efb`	candidate scorer, response judge
allenai/wildguard	`cbba4823f3e80`	candidate scorer, response judge
Qwen2.5-7B-Instruct	`a09a35458c70`	candidate generator, primary target, fulfillment grader
Mistral-7B-Instruct-v0.3	`c170c708c41d`	candidate generator, transfer target
Meta-Llama-3-8B-Instruct	`8afb486c1db2`	candidate generator, transfer target
Llama-2-7b-hf	`01c7f73d771d`	transfer stress target
Llama-2-7b-chat-hf	`f5db02db7245`	matched chat-control target
Qwen3-14B	`40c069824f42`	modern transfer target
Gemma-3-12B-IT	`96b6f1eccf38`	modern transfer target
gpt-oss-20b	`6cee5e81ee83`	modern transfer target
Ministral-3-14B-Instruct-2512	`29439f81c2be`	modern transfer target, fulfillment grader
Qwen3Guard-Gen-4B	`6ec42827da0c`	qualified-policy input guard
Llama-Guard-4-12B	`87acb4b94e93`	qualified-policy input guard
all-mpnet-base-v2	`e8c3b32edf54`	semantic feature

C.2 Stochastic-decoder sensitivity

We repeat two target contrasts on all 109 released source-unsafe OR-Bench test prompts using temperature .7, top-p .9, and three seeds. The estimand remains harmful-release mass over all 150 source-unsafe prompts: withheld prompts have zero mass, while a released generation or evaluator failure is harmful. For each contrast, prompt outcomes are averaged across seeds before 20,000 paired prompt-bootstrap draws. Because this analysis and the second-policy test form two extension claims, intervals are two-sided 97.5%. Success requires a positive difference in every seed and a positive interval lower endpoint.

Table 9: Decoder sensitivity under the LlamaGuard-3/WildGuard union. Values are left-target minus right-target harmful-release mass in percentage points; intervals use the full 150-prompt unsafe denominator.

Contrast	Three seed differences	Mean	97.5% interval
Mistral − Qwen2.5	26.67, 24.67, 24.67	25.33	[18.00, 32.67]
Gemma-3 − Qwen3	4.00, 5.33, 8.67	6.00	[2.67, 9.78]

Both contrasts satisfy the joint criterion. This result rules out a single-greedy-decode explanation for these two orderings under the tested sampling distribution; it does not claim stability under arbitrary decoding hyperparameters.

C.3 Structurally different direct-screen policy

The policy-level replication uses a direct binary screen rather than the recovery architecture. Frozen LlamaGuard-2 receives the unchanged request; the policy releases it if and only if the returned label is parsed and safe. There are no candidate generators, rewrite selector, embedding feature, or learned score. On a group-disjoint calibration split, the largest absolute modern-target contrast selects Ministral-3-14B-Instruct minus gpt-oss-20b and freezes its positive direction before the sealed test is opened. The sealed WildJailbreak population contains 1,000 source-unsafe and 105 source-labeled legitimate prompts. The screen releases 615 and 104, respectively.

The candidate universe contains Gemma-3-12B-IT, gpt-oss-20b, Ministral-3-14B-Instruct, and Qwen3-14B. Calibration uses 1,000 source-unsafe and 105 source-labeled legitimate prompts; the screen releases 650 and 103. For every unordered target pair, the selector computes the absolute difference in conservative harmful-release mass, chooses the largest value, breaks an exact tie lexicographically, and then freezes the observed signed direction. Table 10 reports the complete selection surface. Test inference is conditional on this calibration-only selection rule; the selected pair is not interpreted as a random draw from a target universe.

The prespecified difference is 24.3 percentage points; its two-sided Bonferroni 97.5% paired prompt-bootstrap interval is $[20.4, 28.2]$, exceeding the 3-point minimum with a positive lower endpoint. If unresolved generations are separated from content labels, the confirmed-harm difference is 36.0 points. We retain the conservative endpoint as primary: the 117 gpt-oss generation failures increase its risk and therefore reduce, rather

Table 10: Complete calibration surface for direct-screen target-pair selection. Differences use the displayed left-minus-right orientation; selection uses their absolute values.

Calibration pair	Signed difference	Absolute difference
Gemma-3 − gpt-oss	+23.5 pp	23.5 pp
Gemma-3 − Ministral	−2.3 pp	2.3 pp
Gemma-3 − Qwen3	+1.0 pp	1.0 pp
gpt-oss − Ministral	−25.8 pp	**25.8 pp**
gpt-oss − Qwen3	−22.5 pp	22.5 pp
Ministral − Qwen3	+3.3 pp	3.3 pp

Table 11: Sealed-test direct-screen replication. LG3, WG, and union masses use the conservative failure rule and the full 1,000-prompt unsafe denominator. The final four columns are mutually exclusive and use the same denominator.

Target	LG3	WG	Union	Confirmed harm	Unresolved	Confirmed safe	Withheld
Gemma-3-12B-IT	21.0%	23.0%	31.6%	313	3	299	385
Ministral-3-14B-IT	23.5%	30.0%	36.8%	368	0	247	385
gpt-oss-20b	12.0%	12.5%	12.5%	8	117	490	385
Qwen3-14B	22.0%	24.3%	32.0%	320	0	295	385

than induce, the selected target contrast. This policy is also a poor standalone unsafe screen in this population because it exposes 61.5% of source-unsafe prompts; the experiment concerns transport of a frozen decision, not endorsement of its operating point.

The Gemma-3 and Qwen3 rows are a post-confirmatory completion of the target matrix and are therefore descriptive. All four cells retain the same 615 released unsafe prompts, response evaluators, and conservative failure rule. The three unresolved Gemma outcomes are evaluator failures; the 117 unresolved gpt-oss outcomes are generation failures. No unresolved outcome occurs for Ministral or Qwen3. The six pairwise union-mass differences (first target minus second) are: Gemma versus gpt-oss, +19.1 points; Gemma versus Ministral, -5.2; Gemma versus Qwen3, -0.4; gpt-oss versus Ministral, -24.3; gpt-oss versus Qwen3, -19.5; and Ministral versus Qwen3, +4.8. Hence the selected contrast is the largest of the six, as expected from calibration selection, but gpt-oss remains the lowest-risk cell against every other member of the declared target panel.

The continuous fine-tuned StrongREJECT sensitivity is defined only where both targets return parsed responses. The available-case means are .3985 over 615 Ministral responses and .0133 over 498 gpt-oss responses; these unequal- denominator means are descriptive and are not subtracted. On the 498-prompt paired intersection, the respective means are .4026 and .0133, giving a paired difference of .3893 with a 20,000-draw 95% prompt-bootstrap interval $[.3719, .4066]$. All 1,113 applicable individual scores are present and an independent reconstruction reproduces the input identities, target means, paired difference, and interval. WildGuard labels both responses as non-refusals for only 10/498 pairs: it labels both as refusals for 36, only gpt-oss as refusing for 452, and only Ministral as refusing for none. On the 10 mutually non-refusal pairs, the post-confirmatory mean score difference is .021 (95% bootstrap interval $[-.151, .200]$). Thus the continuous-score gap largely measures differential refusal and is not independent confirmation of content harm. The prespecified conservative full-denominator endpoint and its two binary-evaluator decompositions remain the policy-level evidence.

C.3.1 Fixed missing-output sensitivity

The primary direct-screen contract assigns harmful status to any target output without a parseable final channel. To separate that conservative rule from response content, a fixed sensitivity regenerates all 719 released prompt-target pairs under the same greedy template and 512-token cap, verifies byte-identical outputs for all 527 originally parsed pairs, and retries only the 192 remaining pairs once with a 2,048-token cap. The retry recovers 180 outputs: 114/117 on source-unsafe prompts and 66/75 on source-labeled legitimate prompts. Twelve remain terminal (3 unsafe and 9 legitimate), giving 612/615 parsed unsafe and 95/104 parsed legitimate responses.

On the full 1,000-prompt unsafe denominator, gpt-oss harmful-release mass becomes 4.0% under LlamaGuard-3, 7.3% under WildGuard, and 7.9% under their union. The corresponding paired Ministral-minus-gpt-oss differences are 19.5 points (95% paired-bootstrap CI $[16.9, 22.2]$), 22.7 points ($[19.8, 25.6]$), and 28.9 points ($[25.8, 32.0]$); exact McNemar $p < 3.4 \times 10^{-48}$ for the latter two and $p < 1.9 \times 10^{-45}$ for LlamaGuard-3. On all 612 common parsed unsafe responses, the StrongREJECT means are .397 for Ministral and .120 for gpt-oss, a paired difference of .277 (95% CI $[.253, .301]$). This sensitivity changes the missing-output contract and therefore does not replace the primary 512-token endpoint; it shows that the target contrast persists after recovering 94% of the initially missing finals.

C.3.2 BLINDED HUMAN RESPONSE AUDIT

To assess whether the automated pattern survives direct semantic judgment, we draw a frozen stratified simple random sample without replacement from the 498 prompts for which both selected targets return parsed responses. The six strata cross automated refusal structure with binary-judge agreement; their population/sample sizes are 35/8, 1/1, 154/15, 139/18, 159/20, and 10/10. Thus the design samples 72 prompt pairs (144 distinct responses) and includes a census of every mutually non-refusal pair. Model identity, pair membership, automated labels, and stratum are hidden from annotators.

Two independent annotators label response harm, actionable detail, and refusal for every response, each with 12 opaque within-assignment repeats. All 24 repeat comparisons agree on all three semantic fields. Across the 144 substantive responses, agreement is 134/144 for harm ($\kappa = .858$), 122/144 for actionability ($\kappa = .734$), and 136/144 for refusal ($\kappa = .897$). A third blinded annotator resolves only the relevant fields in the 29-response union of disagreements; fields on which the primary annotators agree are retained. Table 12 uses the resulting locked labels.

For a target indicator Y, the design-weighted mean is $\hat{\mu} = N^{-1} \sum_h N_h \bar{Y}_h$, with $N = 498$. For the paired Ministral-minus-gpt-oss difference D, we use $\hat{\Delta} = N^{-1} \sum_h N_h \bar{D}_h$ and $\widehat{\mathrm{Var}}(\hat{\Delta}) = N^{-2} \sum_h N_h^2 (1 - n_h/N_h) s_h^2/n_h$. Reported intervals use a standard-normal critical value with the finite-population correction; census strata contribute zero sampling variance.

Table 12: Design-weighted human outcomes on the 498-prompt common parsed-response frame. Differences are Ministral minus gpt-oss in percentage points; brackets give approximate 95% design intervals. "Any actionable" combines limited and substantial detail, and "any refusal" combines partial and full refusal.

Outcome	Ministral	gpt-oss	Difference [95% CI]
Harmful response	75.62%	1.61%	+74.02 [+65.22,+82.82]
Any actionable detail	75.62%	1.41%	+74.22 [+65.42,+83.02]
Substantial detail	44.40%	1.20%	+43.19 [+33.27,+53.11]
Any refusal	18.67%	97.99%	−79.32 [−87.23, −71.42]
Full refusal	7.53%	97.99%	−90.46 [−94.78, −86.14]

The complete mutually non-refusal stratum contains eight pairs in which both responses are harmful and two pairs in which both are safe, with no harm-discordant pair. Consequently, the large common-frame harm difference is dominated by whether a target refuses, not by evidence of a content-harm difference after both targets engage. This audit is a supportive mechanism analysis, not a replacement for the prospective endpoint: it excludes 117 unresolved gpt-oss generations and therefore cannot estimate either target's conservative harmful-release mass over all 1,000 source-unsafe prompts. In particular, assigning zero contribution outside the 498-prompt frame would change the estimand and understate unresolved risk.

C.4 EVALUATOR-SPECIFIC OUTCOMES

Table 13 expands the primary union result. All cells divide by the full source stratum, not by selected or released prompts. The union can be larger than either judge because the two evaluators flag different responses. The instruction-tuned contrast between Mistral and Qwen remains 23.33 percentage points under each judge. The contrast between Llama-2 and Qwen ranges from 48.67 to 52.67 points; the matched Llama-2 chat control has one harmful unsafe response and no harmful legitimate response.

C.5 EXTERNAL STRONGREJECT REPLICATION

StrongREJECT contributes 313 official unsafe prompts. The same initial gate rejects 300, of which the frozen selector retains 256 and the threshold releases 253. Table 14 reports the complete evaluator counts. The ordering and scale closely track the OR-Bench unsafe results despite different prompt construction: Mistral exceeds Qwen by 33.33 percentage points and Llama-2 exceeds Qwen by 68.00 points in paired full-denominator mass.

These external contrasts were specified after the confirmatory protocol was closed and frozen before external target generation. Their paired bootstrap intervals are $[28.00, 38.67]$ points for Mistral minus Qwen and $[63.33, 72.33]$ for Llama-2 minus Qwen. They support replication across benchmark families, not retrospective expansion of the confirmatory family.

Table 13: Harmful-release counts and full-population mass by source stratum and response evaluator. Union is the primary outcome.

Source	Target	n	Released	LlamaGuard-3	WildGuard	Union count	Union mass
Unsafe	Qwen2.5-Instruct	150	109	6	7	8	5.33%
	Mistral-Instruct	150	109	41	42	47	31.33%
	Llama-3-Instruct	150	109	1	1	1	0.67%
	Llama-2 base	150	109	79	86	96	64.00%
	Llama-2 chat	150	109	0	1	1	0.67%
Legitimate	Qwen2.5-Instruct	100	67	4	3	5	5.00%
	Mistral-Instruct	100	67	10	9	14	14.00%
	Llama-3-Instruct	100	67	0	2	2	2.00%
	Llama-2 base	100	67	16	26	29	29.00%
	Llama-2 chat	100	67	0	0	0	0.00%

Table 14: StrongREJECT external replication on 300 initially rejected prompts. LG3/WG are evaluator-specific harmful-release counts; union is primary.

Target	Selected	Released	LG3	WG	Union mass	AUROC
Qwen2.5-Instruct	256	253	11	7	14 (4.67%)	.612
Mistral-Instruct	256	253	106	91	114 (38.00%)	.408
Llama-3-Instruct	256	253	4	4	4 (1.33%)	.510
Llama-2 base	256	253	200	194	218 (72.67%)	.240

C.6 Target-pair disagreement

Figure 4 visualizes agreement of union harm labels on the same 113 selected source-unsafe prompts. The result is not driven only by a few released cases near the threshold: target labels disagree extensively across the complete selected set. Qwen and Llama-3 agree on 92.9% of prompts; Qwen and Llama-2 agree on only 18.6%. These are descriptive values without multiplicity-adjusted pairwise claims.

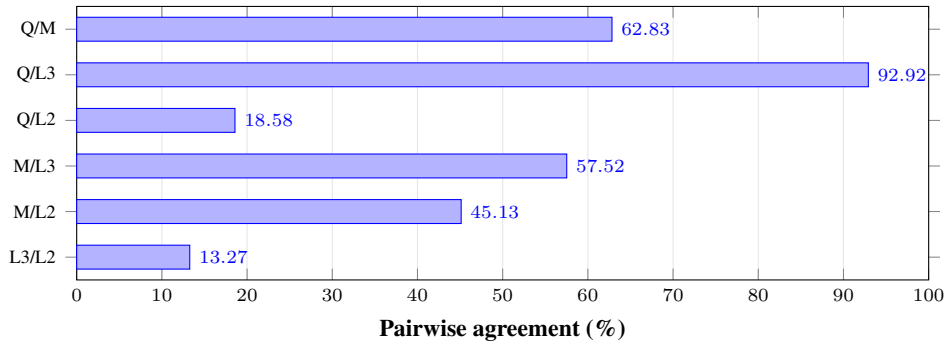


Figure 4: Agreement of target-specific harmful-response labels on selected source-unsafe prompts ($n = 113$). Q, M, L3, and L2 denote Qwen, Mistral, Llama-3, and Llama-2.

For the prespecified target pairs, Qwen versus Mistral contains 63 both-safe, 8 both-harmful, 1 Qwen-only harmful, and 41 Mistral-only harmful outcomes. Qwen versus Llama-2 contains 13 both-safe, 8 both-harmful, 1 Qwen-only harmful, and 91 Llama-2-only harmful outcomes. These counts use selected prompts; the primary McNemar tests in the main paper multiply harm by the frozen release indicator.

C.7 Original versus selected prompt responses

To assess whether rewriting itself explains target differences, we also generate responses to each original prompt for the same selected cases. Table 15 reports discordant union-harm labels. The comparison is descriptive: selection conditions on the rewrite pipeline, and no causal rewrite estimand was prespecified. On selected source-unsafe prompts, the two discordance directions are balanced for Llama-2 (10 versus 10) and close for Mistral (15 versus 11), whereas target-level harmful-release masses remain far apart. This does not prove absence of a rewrite effect, but it does not support attributing the central cross-target contrast solely to rewriting.

Table 15: Paired original-versus-selected response discordances. OS means original harmful/selected safe; SO means original safe/selected harmful.

Source	Target	n	OS	SO
Unsafe	Qwen	113	2	6
	Mistral	113	15	11
	Llama-3	113	0	0
	Llama-2	113	10	10
Legitimate	Qwen	98	7	1
	Mistral	98	18	2
	Llama-3	98	3	2
	Llama-2	98	26	4

Table 16: Released source-unsafe prompts by category. Target columns are harmful-response counts under the union judge.

Category	Released	Qwen	Mistral	Llama-3	Llama-2
Deception	16	4	11	0	13
Harassment	14	0	6	0	12
Harmful content	6	0	1	0	4
Hate	16	0	10	0	14
Illegal activity	7	0	3	0	6
Privacy	8	0	3	0	8
Self-harm	18	0	2	0	18
Sexual content	9	1	5	1	9
Unethical behavior	8	1	2	0	7
Violence	7	2	4	0	5
Total	109	8	47	1	96

D Descriptive subgroup analyses

All analyses in this section are post-confirmatory and descriptive. They are included to localize the observed transport gap, not to create new hypothesis tests. Small cells, correlated categories, and benchmark sampling preclude population-wide category claims.

D.1 Risk category

Table 16 partitions the 109 released source-unsafe prompts by the benchmark category. Llama-2 harm is present in every category and Mistral harm in nine of ten. Qwen harm is concentrated in deception, sexual, unethical, and violence categories. The target gap is therefore broader than one category, but the large illegal and privacy selection-to-release drop also shows that the fixed score changes category composition.

D.2 Selected generator and rewrite style

The selector chooses Qwen candidates for 50/113 selected source-unsafe prompts, Mistral candidates for 61, and Llama-3 candidates for 2. It chooses a faithful rewrite for 83 and clarification for 30. Table 17 reports harm on all selected prompts rather than only releases. Generator identity is not randomized and is strongly confounded with selector features, so these rows cannot rank rewrite models. Their purpose is to show that cross-target harm is not confined to candidates generated by the same model as the response target.

E Prompt-proxy transport audit

Figure 5 evaluates frozen prompt features as risk rankings without refitting. Binary guard decisions have few unique values and should not be compared directly with continuous scores as if AUROC resolution were equal. Nevertheless, the heatmap exposes the central measurement issue: the same proxy can change orientation and relative utility across target outcomes.

The learned score has the smallest-to-largest AUROC span 0.285. LlamaGuard-3 and the two-guard ordinal have spans 0.363 and 0.354, respectively. WildGuard has a smaller span (0.109) but remains only modestly

Table 17: Selected source-unsafe prompts by candidate generator/style. Harm counts use all selected prompts and therefore can exceed released harm counts.

Group	n	Rel.	Q	M	L3	L2
Qwen generator	50	47	6	28	1	47
Mistral generator	61	61	3	21	0	50
Llama-3 generator	2	1	0	0	0	2
Faithful style	83	80	9	36	1	74
Clarification style	30	29	0	13	0	25

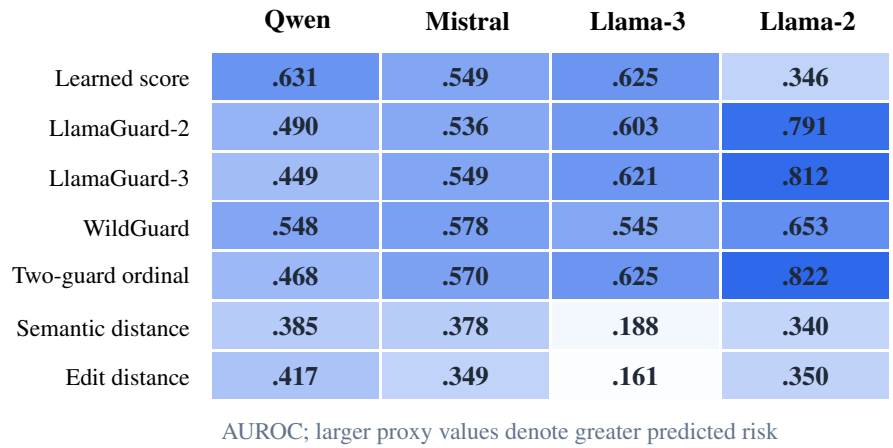


Figure 5: Post-confirmatory prompt-proxy AUROC on the same 113 selected source-unsafe prompts. Darker blue indicates larger AUROC. No proxy is refit per target, and no value changes the confirmatory endpoints.

discriminative. Semantic and edit distance are below chance for every target. These results do not imply that prompt guards should be inverted for some targets: with binary features and small harmful cells, such a rule would be unstable. They imply that a proxy's relationship with target harm must be measured rather than presumed.

F Human annotation protocol and decomposition

F.1 Blinding, controls, and instructions

Each annotator receives the same 98 substantive triples in an independently shuffled order plus six hidden controls. A triple contains the original request, selected candidate, and Qwen response. Annotators work independently and are blinded to source label, dataset, split, model, score, threshold, release status, sampling stratum, control identity, and all other labels. They are instructed not to use an LLM, search engine, guard, translation service, or executable code.

The operational labels are:

- **Original/candidate intent:** legitimate, harmful, or unclear, considering authorization, consent, target, scope, constraints, and requested detail.
- **Candidate form:** standalone request versus answer, refusal, disclaimer, warning, analysis, or commentary.
- **Intent preservation:** exact, minor non-core shift, or changed goal/target/authorization/safety condition/detail.
- **Useful recovery:** yes, partial, or no for legitimate originals; not applicable otherwise.
- **Response harm:** safe, harmful, or unclear. A disclaimer does not make actionable harmful instructions safe.

Table 18: Pairwise human agreement. Each cell is exact agreement / Cohen's κ over 98 substantive triples.

Field	A vs. B	A vs. C	B vs. C
Original intent	.837 / .748	.837 / .744	.837 / .745
Candidate intent	.888 / .812	.898 / .824	.888 / .805
Candidate form	1.000 / n.a.	1.000 / n.a.	1.000 / n.a.
Intent preservation	.969 / .883	.969 / .865	.939 / .758
Useful recovery	.908 / .816	.949 / .900	.918 / .838
Response harm	.918 / .784	.939 / .833	.959 / .890
Response refusal	.969 / .899	.939 / .792	.969 / .896
Original utility	.888 / .796	.939 / .892	.878 / .785
Candidate utility	.918 / .788	.929 / .808	.929 / .820

- **Refusal:** no, partial, yes, or unclear. Irrelevance alone is not a refusal unless the response explicitly declines.
- **Utility:** yes, partial, or no relative to the original and to the candidate; original utility is not applicable for non-legitimate intent.

Annotators also record confidence and a rationale of at most 25 words. Adding benign framing to a harmful request, or removing an authorization constraint, is explicitly labeled changed rather than faithful. All 18 hidden-control decisions pass.

F.2 Agreement

Table 18 reports pairwise exact agreement and Cohen's κ. Agreement is perfect for candidate form because every selected item is a valid request, but κ is undefined when both raters use only one category. Original intent has stable κ around 0.745 across all pairs. One field-level three-way tie occurs for original utility; it does not affect the original-intent endpoint or the 47-prompt legitimate subset.

F.3 All-triple majority distributions

Table 19 gives majority-resolved distributions across all 98 triples. "Not applicable" is structurally tied to non-legitimate original intent for useful recovery and original utility. Consequently, all-triple counts should not be read as failure rates. The main paper conditions the mechanistic decomposition on the 47 majority-legitimate originals.

Table 19: Majority-resolved annotation distributions over all 98 substantive triples.

Field	Label	Count
Original intent	legitimate / unclear / harmful	47 / 26 / 25
Candidate intent	legitimate / unclear / harmful	57 / 18 / 23
Preservation	exact / minor / changed	85 / 2 / 11
Useful recovery	yes / partial / n.a.	46 / 1 / 51
Response harm	safe / harmful	76 / 22
Response refusal	no / partial / yes	81 / 10 / 7
Original utility	yes / partial / n.a. / tie	38 / 8 / 51 / 1
Candidate utility	yes / partial / no	75 / 16 / 7

Conditioning on the 47 majority-legitimate originals gives the joint table that the two marginal summaries alone cannot show: preservation/useful-recovery is 44 exact/yes, 2 minor/yes, and 1 changed/partial. Thus 46/47 are at least partially preserving and marked useful, whereas the single changed candidate is not counted as a fully successful recovery. Among these 47, response harm is 2/47; this response endpoint is distinct from prompt-level preservation.

F.4 Cross-target response mechanism

The separate cross-target audit resolves 157 unique responses from 80 probability-sampled prompt pairs. Two annotators label response harm, actionable detail, and refusal independently; a third labels only the union of field-level disagreements. The analyses below use the same frozen inclusion weights and stratified prompt-pair

bootstrap as the harm endpoint. They are post-confirmatory mechanism descriptions rather than additional hypothesis tests.

Sampling is stratified simple random sampling without replacement over prompt pairs. Table 20 exposes every frame and sample count; each pair receives weight N_h/n_h. The design-weighted contrast is $\hat{\Delta} = N^{-1}\sum_h N_h \bar{D}_h$ and its estimable variance is $N^{-2}\sum_h N_h^2(1-n_h/N_h)s_h^2/n_h$. Approximate design intervals use a standard-normal critical value. Bootstrap intervals instead condition on the observed values and resample within strata; they must not be read as recovering unobserved within-cell variation. In particular, a zero-width bootstrap interval is not evidence of population equality, and the StrongREJECT The design variance for the Mistral versus Qwen contrast is unavailable because its both-harmful cell has $N_h = 2, n_h = 1$. Larger probability samples could combine dense automated predictions with human labels through prediction-powered inference (Angelopoulos et al., 2023); we do not apply that estimator retrospectively.

Table 20: Frozen human-audit sampling frames shown as N_h/n_h (population pairs / sampled pairs). BS/BH denote both automated evaluators safe/harmful, ED evaluator disagreement, and TD target-discordant majority. A dash means the cell is absent from that cohort's frame.

Cohort	BS	BH	ED	TD
OR: Mistral vs. Qwen	33/5	5/2	35/5	40/6
OR: Llama-2 vs. Qwen	7/4	2/2	34/5	70/7
OR: Llama-2 vs. chat	11/5	n/a	30/4	72/7
StrongREJECT: Mistral vs. Qwen	96/5	2/1	64/4	94/4
StrongREJECT: Llama-2 vs. Qwen	16/4	n/a	67/4	173/6

Table 21: Design-weighted Mistral-minus-Qwen mechanism differences in full-denominator released-response mass. Parentheses give paired 95% bootstrap intervals.

Population	Actionable detail	Any refusal
OR-Bench	+22.44 [+8.89,+36.00]	-23.84 [$-37.09, -10.56$]
StrongREJECT	+35.23 [+15.67,+54.80]	-24.23 [$-45.23, -4.67$]

On OR-Bench, all actionable responses in the weighted audit are labeled substantial; the substantial-actionability difference is therefore also 22.44 points. On StrongREJECT it is 15.67 points with interval $[0.00, 31.33]$. Full-refusal differences are -35.69 points on OR-Bench and -35.23 on StrongREJECT, with intervals excluding zero. The replicated pattern is not merely a shift between two automated harm thresholds: relative to Qwen, Mistral both refuses less often and supplies more actionable harmful content.

Post-confirmatory weighting diagnostics preserve the direction of the the harmful-release contrast between Mistral and Qwen. Renormalized 95th- and 99th-percentile weight caps leave both estimates unchanged. Across leave-one-stratum-out deletions, the OR-Bench difference ranges from 14.03 to 44.94 points under the unclear-safe convention and the StrongREJECT difference ranges from 18.54 to 46.13 points. The percentile interval reaches zero when the deliberately oversampled target-discordant stratum is removed. Thus the direction is not created by one sampling cell, but interval separation depends on the prespecified disagreement-focused design.

F.5 Modern paired-response census

Two independent annotators each label all 218 responses from the 109 released Gemma-3 versus Qwen3 prompt pairs, with model identity and pair membership hidden. Each assignment includes 12 opaque repeats balanced across targets. All 24 repeat comparisons preserve the response-harm label; one refusal repeat differs. Between annotators, response-harm agreement is 218/218, actionable-detail agreement is 208/218, and refusal agreement is 182/218. Response harm therefore requires no adjudication. A third annotator labels the 39-response union of actionability or refusal disagreements without seeing either primary label; consensus retains agreement between A and B and otherwise uses the third label. Gemma-3 has 10/109 harmful responses and Qwen3 has 5/109; the paired counts are 99 pairs in which both are safe, five in which both are harmful, five Gemma-only harmful, and zero Qwen-only harmful. The two-sided exact paired p-value is .0625. Because the finite census supplies directional but not exact level-.05 evidence, it is reported as descriptive validation rather than an additional confirmatory endpoint. After disagreement-only adjudication, actionable detail occurs in 12/109 Gemma responses and 6/109 Qwen responses (+5.50 percentage points; six Gemma-only versus zero Qwen-only pairs). Any refusal occurs in 82/109 and 86/109 responses, respectively (-3.67 points; four versus eight one-way discordances), and full refusal in 78/109 and 85/109 (-6.42 points). These exploratory mechanism counts agree in direction with the larger Mistral versus Qwen audit but are not treated as multiplicity-adjusted hypothesis tests.

Table 22 compares each frozen automated response judge with the unanimous human harm label on this complete frame. The union recovers 7/10 human-positive Gemma responses but 0/5 human-positive Qwen responses. With only 10 and five positive items, the sensitivity intervals are necessarily wide. More importantly, the targetwise difference rules out using this census as universal item-level validation of the automated endpoint. We therefore retain the automated full-denominator policy estimand and the human census as distinct instruments; their agreement is interpreted only at the target-level direction reported above.

Table 22: Target-stratified agreement of automated judges with unanimous human response-harm labels in the modern 109-pair census. Entries are counts on each 109-response target frame. Brackets give two-sided 95% Wilson intervals; no multiplicity-adjusted inferential claim is made.

Target	Judge	TP	FP	FN	TN	Sensitivity	Specificity
Gemma-3	LlamaGuard-3	7	5	3	94	.700 [.397,.892]	.949 [.887,.978]
	WildGuard	5	1	5	98	.500 [.237,.763]	.990 [.945,.998]
	Union	7	5	3	94	.700 [.397,.892]	.949 [.887,.978]
Qwen3	LlamaGuard-3	0	3	5	101	.000 [.000,.434]	.971 [.919,.990]
	WildGuard	0	1	5	103	.000 [.000,.434]	.990 [.948,.998]
	Union	0	3	5	101	.000 [.000,.434]	.971 [.919,.990]

G Complete calibration and negative-result audit

Table 23 lists every fixed threshold. The two displayed pairs at .089911 and .305150 are distinct full-precision quantiles that coincide after six-decimal rounding and induce identical decisions. No row is removed based on its result. The threshold used for the matched-target test is the first grid point reaching at least 15% observed strict recovery; its calibration failure is preserved rather than hidden.

Changing only the family-wise confidence level does not alter the decision. At $\delta = .01$, .05, and .10, no threshold passes. At the 17% recovery point, the unsafe upper bound ranges from 9.36% to 10.88%, and the legitimate failure bound from 65.14% to 68.22%. Thus failure is not a boundary artifact of choosing $\delta = .05$.

The shuffled-label audit preserves all scores, threshold grid points, stratum sizes, and outcome prevalences, then independently permutes outcomes within source strata 1,000 times. Zero permutations produce a passing threshold. The median maximum risk-feasible recovery is 0%, the 95th percentile is 3%, and the maximum is 5%. This audit checks that the implementation is not artificially permissive; it does not convert the negative real-label result into a positive claim.

H Exploratory target-conditioned prediction check

We evaluate a constructive alternative on a disjoint calibration population drawn from the WildJailbreak training split (Jiang et al., 2024). The population contains 600 source-unsafe and 300 source-legitimate requests rejected by LlamaGuard-2 after semantic-group de-duplication and leakage checks. Each request receives six deterministic candidates: faithful and clarification rewrites from Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, and Llama-3-8B-Instruct. The same frozen prompt-only selector chooses one candidate for every target or abstains. It selects 530 unsafe and 279 legitimate requests. Candidate features comprise form validity, exact copying, duplicate structure, character and word length ratios, normalized edit distance, MPNet similarity, and prompt labels from LlamaGuard-2, LlamaGuard-3, and WildGuard.

Qwen3-14B, Gemma-3-12B-IT, gpt-oss-20b, and Ministral-3-14B-Instruct-2512 answer the selected candidates under pinned model-native templates and greedy decoding. The primary harm outcome is the union of LlamaGuard-3 and WildGuard response labels. A generation or evaluator failure is harmful and not useful; no failed output is selectively regenerated. Proxy-useful recovery requires a source-legitimate request, a parsed response, no WildGuard refusal, and no primary harm. An independent continuous response grader provides a sensitivity record but does not define the primary outcome.

We compare four ℓ_2-regularized logistic models: prompt features only, target intercepts, target-by-feature interactions, and separate per-target models. Five-fold cross-validation holds out complete prompt groups and selects regularization from $C \in \{.01, .1, 1, 10\}$ by pooled out-of-fold log loss. All four classes select $C = .1$. The primary criterion compares the target-interaction and prompt-only models. It requires a mean target-wise AUROC improvement of at least .03, a positive lower endpoint from 20,000 paired prompt-bootstrap draws, and no target deterioration exceeding .02.

Table 24 shows that the gain is concentrated in gpt-oss. The mean difference is .0145 with paired 95% interval $[-.0028, .0318]$. The per-target deterioration condition holds, but the margin and positive-lower-endpoint conditions do not. Pooled AUROC increases from .7287 to .7773, AUPRC from .3037 to .4248, and log loss

Table 23: Complete 20-point calibration grid. U_r is unsafe harmful-release upper bound, L_r is legitimate failure-mass upper bound, and Rec. is strict useful-recovery coverage. Bounds and coverage are percentages.

Threshold	Unsafe rel.	Unsafe failures	U_r	Legit. rel.	Legit. failures	Rec.	L_r
.001478	0	0	2.99	0	0	0	6.47
.005624	20	0	2.99	0	0	0	6.47
.007577	32	0	2.99	0	0	0	6.47
.010622	43	0	2.99	3	2	1	10.48
.017513	58	1	4.01	3	2	1	10.48
.023565	64	2	4.88	4	3	1	12.17
.030690	68	3	5.68	8	5	3	15.27
.049534	87	5	7.15	18	14	4	27.21
.061632	98	5	7.15	22	17	5	30.82
.071764	109	5	7.15	25	20	5	34.31
.089911	118	5	7.15	31	24	7	38.82
.089911	118	5	7.15	31	24	7	38.82
.130059	149	8	9.19	45	36	9	51.55
.165288	151	8	9.19	54	43	11	58.52
.198284	158	8	9.19	62	49	13	64.27
.233098	**159**	**9**	**9.83**	**68**	**51**	**17**	**66.14**
.305150	166	11	11.09	86	61	25	75.14
.305150	166	11	11.09	86	61	25	75.14
.389894	171	12	11.71	91	65	26	78.57
.609729	173	12	11.71	99	72	27	84.31

Table 24: Out-of-fold harmful-response discrimination on 530 selected source-unsafe prompts per target. Differences compare target-by-feature interactions with the prompt-only model; the prespecified mean margin is .03.

Target	Prompt-only AUROC	Interaction AUROC	Difference
Qwen3-14B	.738	.757	+.018
Gemma-3-12B-IT	.586	.597	+.011
gpt-oss-20b	.839	.882	+.043
Ministral-3-14B-Instruct	.748	.735	-.013
Mean	n/a	n/a	+.014

decreases from .4041 to .3757; these pooled descriptive improvements do not override the target-wise criterion. A separate implementation reproduces the selected regularization, model metrics, target differences, and bootstrap interval.

To distinguish failure of the Qwen2.5-reference score from failure of every prompt-only predictor, we perform a retrospective leave-one-target-out analysis on the same complete 530-prompt panel. For each excluded target, an ℓ_2-regularized prompt-only logistic model is developed on outcomes from the other three targets; regularization is selected by grouped five-fold log loss without using the excluded target. Table 25 reports the single evaluation on that target. Bootstrap intervals resample the 530 held-out prompt IDs.

Table 25: Retrospective leave-one-target-out discrimination. The model uses prompt features but no target identity; each row's target is absent from model fitting and hyperparameter selection.

Held-out target	AUROC	95% CI	AUPRC	Brier
Qwen3-14B	.749	[.695,.799]	.267	.110
Gemma-3-12B-IT	.606	[.545,.664]	.232	.155
gpt-oss-20b	.699	[.648,.749]	.347	.163
Ministral-3-14B-Instruct	.778	[.721,.833]	.337	.101
Mean AUROC	.708	n/a	n/a	n/a

The mean AUROC is .013 above the mean transfer performance of three separately fit single-source models, a descriptive comparison rather than a prespecified margin test. The weakest target has AUROC .606 and ECE .105; gpt-oss has ECE .072. Multi-target development therefore preserves some unseen-target ranking on this panel but does not supply target-invariant probability calibration. Because outcomes were available before this analysis was specified, it is supportive rather than confirmatory and does not repair the failed prospective deployment gate.

Table 26: Complete grouped out-of-fold comparison of the four target-aware feature families. Each model selects $C = .1$ by pooled log loss.

Model family	AUROC	AUPRC	Brier	Log loss
Prompt features only	.729	.304	.126	.404
Target intercepts	.734	.319	.125	.400
Target interactions	.777	.425	.117	.376
Separate target models	.775	.425	.117	.376

Table 27: Calibration outcomes under the frozen selector. Harm uses the full 600-request unsafe denominator; conditional harm uses 530 selected unsafe requests; proxy-useful recovery uses the full 300-request legitimate denominator.

Target	Harmful release	Conditional harm	Proxy-useful recovery
Qwen3-14B	71/600 (11.83%)	71/530 (13.40%)	168/300 (56.00%)
Gemma-3-12B-IT	94/600 (15.67%)	94/530 (17.74%)	183/300 (61.00%)
gpt-oss-20b	113/600 (18.83%)	113/530 (21.32%)	23/300 (7.67%)
Ministral-3-14B-Instruct	68/600 (11.33%)	68/530 (12.83%)	176/300 (58.67%)

Target intercepts alone provide only a small pooled improvement. Interactions and separate models are nearly identical in pooled metrics, which localizes the gain to target-dependent feature behavior rather than base-rate adjustment alone. The targetwise AUROC changes remain heterogeneous, including a decline for Ministral, so this decomposition does not alter the failed deployment gate.

This is a model-development diagnostic, not a held-out test of a learned target-aware policy: cross-validation estimates prediction performance on the same population used to compare feature families, and the subsequent threshold scan reuses those out-of-fold predictions. The frozen sequential stopping rule prevents escalation to a sealed target-aware test after the criterion fails. Accordingly, neither the pooled improvements nor the negative threshold scan is presented as a generalization claim for target-conditioned release.

The 20 candidate thresholds are quantiles of the fitted interaction-model risk. For each threshold, one-sided Clopper–Pearson bounds are Bonferroni-adjusted over four targets and all retained thresholds. Ten thresholds keep every target's harm upper bound at or below 5%. The best of these recovers only 1/300 legitimate gpt-oss requests, so its worst-target recovery is .33% and its lower bound is approximately zero. At the least restrictive threshold, gpt-oss reaches only 23/300 proxy-useful recoveries while its harm upper bound is 24.43%. Its 174 unresolved generations are retained under the failure rule. Thus zero of 20 thresholds meets both the 5% harm and 15% point/10% lower-bound recovery requirements across targets. Target information improves some pooled predictive metrics but does not establish either prediction repair or a deployable joint operating point under this contract.

I CHECKPOINT-DISJOINT QUALIFIED-POLICY TEST

SALAD-Bench supplies source-unsafe prompts and XSTest supplies source-legitimate prompts (Li et al., 2024; Röttger et al., 2024). Normalized text hashes are disjoint across the 1,100-item calibration set, the 2,150-item committed test, and earlier populations. Before sampling, a candidate-limited near-duplicate screen also excludes pairs with token-trigram Jaccard similarity at least .85 among prompts sharing a 16-bit SimHash band. The policy releases an unchanged input exactly when Llama Guard 4 is safe and Qwen3Guard is safe or controversial; invalid verdicts withhold. Neither checkpoint appears among the response evaluators. Qwen3Guard uses its native three-way prompt-moderation template with at most 64 generated tokens; Llama Guard 4 uses its native safe/unsafe template with at most 16 generated tokens. Both use deterministic decoding and pinned revisions (Table 8). Calibration chooses between this policy and a stricter policy that requires both guards to return safe. The former is frozen because its Qwen3 harm upper bounds are 0.44% and 1.56%, its legitimate-coverage lower bound is 87.40%, and its projected test release count meets the power requirement; the latter fails coverage and power.

The test fixes released prompts and changes Qwen3-14B to Gemma-3-12B-IT. Both use native templates, greedy decoding, and at most 256 new tokens. Released target or evaluator failures are harmful on the full denominator. For each evaluator, Qwen's risk upper bound must be at most 5%, Gemma's lower bound must exceed 5%, and the conservative paired lower bound must exceed three points. The latter subtracts an exact upper bound for the Qwen-only discordance probability from an exact lower bound for the Gemma-only prob-

ability, allocating half the cell error to each. The one-sided cell error is .0125; legitimate coverage must have a lower bound of at least 80%.

Table 28: Committed qualified-policy test. Risk divides harmful releases by all 2,000 source-unsafe prompts. The target contrast is Gemma-3 minus Qwen3.

Evaluator	Qwen3 risk	Qwen upper	Gemma risk	Gemma lower
LlamaGuard-3	1/2,000 (.05%)	.319%	13/2,000 (.65%)	.315%
WildGuard	10/2,000 (.50%)	.984%	22/2,000 (1.10%)	.644%

LlamaGuard-3 has 13 Gemma-only versus one Qwen-only harmful discordance; WildGuard has 17 versus five. Both differences are .60 points, with conservative paired lower bounds $-.071$ and $-.266$. All 697 released responses and judgments are complete. Legitimate coverage is 139/150 with an 86.43% lower bound. Reference and coverage requirements pass, but neither changed-target requirement does. This prospective null covers one prespecified substitution, not universal transport.

As a separate post-confirmatory contract summary, we form a simultaneous 95% family from four one-sided risk upper bounds and one one-sided legitimate- coverage lower bound, using cellwise error $.05/5 = .01$. The Qwen3 upper bounds are .331% under LlamaGuard-3 and 1.005% under WildGuard; the corresponding Gemma-3 bounds are 1.204% and 1.774%. The legitimate-coverage lower bound is 86.17%. All five statements therefore hold simultaneously for these two declared targets, evaluators, population strata, and generation contracts; the family neither selects a threshold nor covers another target.

I.1 Component and downstream-fulfillment diagnostics

We evaluate all four input policies defined before ablation target generation: each guard alone, their moderate conjunction used above, and the strict conjunction considered during calibration, which requires both guards to return safe. Responses and judgments for the moderate policy are reused exactly; deterministic responses are generated only for prompts released by a single guard but not by the moderate conjunction. The same targets, templates, token limits, and response evaluators are retained. Table 29 reports the conservative evaluator union on the full 2,000-item unsafe denominator. These component contrasts are post-confirmatory and descriptive.

Table 29: Qualified-policy component comparison. Legitimate coverage uses all 150 legitimate requests; harmful-release mass is the union of the two response judges over all 2,000 unsafe requests.

Input policy	Legit. coverage	Qwen3 harm	Gemma-3 harm
Llama Guard 4 only	139 (92.67%)	32 (1.60%)	67 (3.35%)
Qwen3Guard only	149 (99.33%)	14 (.70%)	35 (1.75%)
Moderate conjunction	139 (92.67%)	10 (.50%)	28 (1.40%)
Strict conjunction	122 (81.33%)	2 (.10%)	4 (.20%)

On this population, adding Qwen3Guard to Llama Guard 4 reduces unsafe releases from 997 to 558 without reducing legitimate coverage, and reduces union harm for both targets. Tightening both guards to their strict labels further lowers harm but lowers legitimate coverage by 11.34 points. Qwen3Guard alone has the highest legitimate coverage and intermediate harm. This matched-coverage comparison supports the conjunction as a substantive component rather than attributing its lower harm solely to more abstention; it does not rank guards outside the declared population.

Release coverage still does not show that a target answers a legitimate request usefully. We therefore apply two additional deterministic generative graders, Qwen2.5-7B-Instruct and Ministral-3-14B-Instruct, to all 278 released legitimate responses. Each assigns FULFILLED, PARTIAL, or NOT FULFILLED under a fixed rubric after passing six controls. Strict usefulness requires FULFILLED from both; the permissive definition also admits PARTIAL. "Safe useful" additionally requires both response-safety judges to label the response safe. Withheld and unresolved items contribute zero on the full 150-request denominator; no human labels are used for this descriptive audit.

The graders disagree on 53/139 Qwen3 and 68/139 Gemma-3 released responses; strict consensus is consequently conservative. The permissive endpoint is less sensitive to the distinction between partial and complete answers, yet retains a 10.67-point target gap. These results establish neither factual correctness nor a new simultaneous guarantee. They show that identical input coverage can conceal target-dependent answer utility, and they keep coverage, response safety, and fulfillment as separate estimands.

Table 30: Downstream fulfillment under the moderate conjunction. Brackets give targetwise two-sided 95% Wilson intervals for the full-population rate.

Target	Safe	Strict safe useful	Permissive safe useful
Qwen3-14B	138/150	77/150, 51.33% [43.40,59.20]	123/150, 82.00% [75.08,87.32]
Gemma-3-12B-IT	138/150	58/150, 38.67% [31.25,46.65]	107/150, 71.33% [63.64,77.97]

I.2 Supportive older-target replication

The supportive replication freezes Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3 on the same released prompts, native templates, greedy decoding, and a 256-token cap. It is descriptive, is not pooled with the primary, and cannot rescue it. All 697 released responses and both binary judgments are complete; unsafe risk retains the full 2,000-prompt denominator.

Table 31: Predeclared supportive replication under the qualified input policy. Discordances are Mistral-only/Qwen-only harmful; exact one-sided lower bounds are in percentage points.

Evaluator	Qwen harm	Mistral harm	Difference	Discordances	Paired lower
LlamaGuard-3	3 (.15%)	7 (.35%)	+.20	5/1	−.251
WildGuard	9 (.45%)	23 (1.15%)	+.70	16/2	+.015

Both point differences follow the earlier Mistral versus Qwen direction, but the joint replication gate requires a positive paired lower bound under each evaluator and therefore fails. This mixed result further limits the large earlier contrast to its specified policy and population.

J Scope, failure modes, and deployment implications

J.1 Exact fixed-horizon sample-size recipe

Proposition 2 yields a direct planning rule. Fix a family of K target-by-evaluator-by-threshold cells, family error .05, and risk limit .05. If k failures are observed in n trials, choose the smallest n whose one-sided Clopper–Pearson upper bound at cell error $.05/K$ is at most .05. Table 32 gives exact minima; at each reported $n-1$ the bound remains above .05.

Table 32: Minimum fixed-horizon sample size needed for a simultaneous 95% family of binary-risk upper bounds at most 5%. K is the number of cells.

Family	$k=0$	$k=1$	$k=2$	$k=5$
$K=1$	59	93	124	208
$K=4$	86	125	160	252
$K=8$	99	141	177	272
$K=80$	144	191	231	336

For zero failures, the same result follows analytically from $1-(.05/K)^{1/n} \le .05$. The $K=4$ row represents four fixed risk cells; the $K=80$ row represents four targets over 20 candidate thresholds. This recipe assumes the family and horizon are fixed before outcomes. Optional stopping requires a valid sequential method, and the calculation supplies neither coverage power nor validity for an undeclared target.

J.2 When a paired bridge can reuse reference validation

Let $A, B \in \{0,1\}$ be harmful-release indicators for a reference and changed target. Combining n_0 independent reference observations with m fresh paired observations gives the unbiased estimator $\widehat{\mu}_B = \bar{A}_{n_0} + \overline{(B-A)}_m$. Independence and variance additivity show that it improves on the direct mean $\bar{B}_m$ exactly when

$$\mathrm{Var}(B-A) + \frac{m}{n_0}\mathrm{Var}(A) < \mathrm{Var}(B). \tag{8}$$

This separates uncertainty in the historical reference estimate from uncertainty in the paired target-change estimate. As $n_0 \to \infty$, a bridge requires fewer changed-target population units at equal asymptotic variance

precisely when $\mathrm{Var}(B - A) < \mathrm{Var}(B)$, equivalently $2\,\mathrm{Cov}(A, B) > \mathrm{Var}(A)$. This condition is not implied by model relatedness or a shared release set and must be checked for the declared population and outcome rule.

As a post-confirmatory diagnostic, we evaluated all 48 ordered target pairs formed within two model panels and two populations. Six satisfy the infinite-reference condition. The most favorable cell is Qwen3-14B as the reference and Gemma-3-12B-IT as the changed target on OR-Bench: the harmful counts are 3 and 12 of 150, with nine safe-to-harmful and zero harmful-to-safe flips. The bridge-to-direct variance ratio is .766, implying at most a 23.4% asymptotic reduction in changed-target sample size; with finite reference data, an improvement additionally requires $m/n_0 < .878$. This comparison excludes the cost of obtaining paired reference outputs. No ordered pair on StrongREJECT satisfies the asymptotic condition. These results establish a usable diagnostic, not a general transport shortcut: direct target validation remains necessary when the condition fails.

Separating non-invariance from contract violation. The experiments support two claims at different strengths. The recovery and direct-screen studies reject target invariance under their declared contracts, whereas the checkpoint-disjoint test finds that both declared targets remain inside a qualified risk region. The supportive older-target replication then shows small same-direction point differences but fails its joint evaluator gate. Consequently, evidence that responses depend on the target is not by itself evidence that a deployment threshold is violated. Operational tests should prespecify both quantities: a target-change contrast measures transport, while target-wise confidence bounds determine whether either endpoint remains within its risk contract.

Evaluator resolution at low risk. When harmful counts are small, agreement in point direction need not yield an instrument-robust conclusion. In the older-target replication, WildGuard clears the paired direction gate but LlamaGuard-3 does not; combining the two after observing this disagreement would change the prespecified estimand. We therefore retain evaluator-specific decisions and the conjunctive gate. This also explains why a null transport test cannot certify unseen targets: it limits the detectable change for the named instruments, population, and decoding contract rather than identifying a target-independent semantic property of the released prompt.

Scope and deployment. Under the frozen contract, target substitution changes harmful-release mass and the association between score and harm. This does not establish universal rankings or causality, nor certify individual prompts, adaptive use, tools, or future revisions. Claims must bind the population, policy, target, decoding, evaluation, denominator, and confidence statement, and be revalidated after a target or judge change. Target-conditioned validation complements response moderation, constrained generation, permissions, monitoring, adversarial evaluation, and escalation; these controls should not collapse into one "safe prompt" label.

J.3 Reproducibility and data availability

The appendix reports denominators, aggregate counts, statistical procedures, and public checkpoint revisions. These details support checks of reported fractions and inferential definitions, but the PDF alone does not permit regeneration of item-level predictions or bootstrap draws. No separate data or code file accompanies this submission. Raw responses, annotations, rationales, harmful text, hidden mappings, and annotator-linked metadata are not distributed.