BENIGN SAFETY EVALUATION IS NOT A ROBUSTNESS CERTIFICATE: EVIDENCE OF A STRUCTURAL EVALUATION GAP ACROSS ADVERSARIAL CONDITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study whether safety evaluation under benign conditions can substitute for direct adversarial robustness evaluation. We operationalize benign evaluation using a specified, equally weighted composite of helpfulness, harmlessness, and honesty across six selected instruction-tuned checkpoints. Exact-token assistant-prefill trajectories reveal the first level of measurement non-interchangeability: stronger scores on this composite do not consistently imply stronger prefix robustness. Among the original four checkpoints, robustness ordering under the prefix condition does not fully carry over to a structurally different transferred adversarial condition, establishing a second finite failure of substitution. Repeated interaction exposes additional frozen-evaluator failures missed by first-response assessment. Within the prefix protocol, model ordering remains stable across the tested grids even though robustness magnitudes depend materially on grid resolution. Human calibration with independent primary author annotation and third-author adjudication provides moderate support for the harm evaluator, with weaker agreement on adaptive examples. We characterize these findings as an empirical structural evaluation gap, not a defect in alignment-training mechanisms. They motivate distinct, condition-explicit benign and adversarial measurements, without implying that every benign evaluation or adversarial condition shares the observed relationships.

1 INTRODUCTION

Benign evaluations characterize helpfulness, harmlessness, and honesty under ordinary use (Askell et al., 2021; Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023); adversarial evaluations probe behavior under attacker-controlled pressure. Jailbreak research already establishes that aligned assistants can fail (Ganguli et al., 2022; Zou et al., 2023; Mazeika et al., 2024; Chao et al., 2024). Our question is instead whether one safety evaluation can legitimately substitute for another. Here, a *certificate* means a transferable empirical proxy that could stand in for direct robustness measurement, not formal verification or a mathematical guarantee.

We operationalize benign safety evaluation with A_{HHH}, an equally weighted composite of helpfulness, harmlessness, and honesty. It is neither harmlessness alone nor a universal definition of safety. Our results establish finite failures of substitution for this evaluation design and selected checkpoints; they do not imply that every possible benign safety evaluation has the same relationship to adversarial robustness.

We examine two substitutions: benign-composite ordering for exact-prefix robustness ordering, and exact-prefix ordering for ordering under a second specified adversarial condition. Their observed failures motivate an *empirical structural evaluation gap*: a recurring failure of one measurement to substitute for another across conditions. The term describes a measurement-level empirical pattern, not causal independence, mathematical orthogonality, or a defect in the internal mechanism of alignment training.

The evidence combines an identity-bound four-model baseline with a frozen two-model benign/prefix expansion. Among the original four, a fixed transferred GCG condition changes part of the observed

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.

robustness order, and repeated interaction exposes additional failures. Yet the prefix rank remains stable across seven prespecified grids while Robustness-AUC magnitude changes. Within-condition stability can therefore coexist with cross-condition non-interchangeability; the finding is not universal measurement instability.

The practical question is whether benign-condition results suffice when behavior under adversarial use is the quantity of interest. Keeping both evaluations explicit allows their relationship to be tested rather than assumed. Our comparisons supply finite counterexamples, not a population study or a causal test of alignment training. Human calibration quantifies residual error in the frozen harm evaluator, with weaker agreement on repeated-interaction examples.

Contributions.

- **Two-level measurement non-interchangeability.** Selected-checkpoint counterexamples show that the specified benign composite does not determine exact-prefix ordering, which in turn does not fully transfer to the second adversarial condition.
- **An empirical structural evaluation gap.** We characterize recurring failures of substitution across benign, adversarial-pressure, interaction, and cross-condition measurements as an evaluation-level pattern.
- **Condition-explicit robustness analysis.** Pinned identities and declared protocols separate stable within-prefix ranks from grid-sensitive magnitudes and cross-condition ordering changes.
- **Calibrated and auditable measurement.** Independent primary author annotation and third-author adjudication quantify evaluator error; uncertainty, missingness, provenance, and reproducible numeric artifacts make the evidence auditable.

2 Background and Related Work

Benign behavior and adversarial vulnerability. Preference-oriented methods improve assistant behavior under their objectives (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023); helpfulness, honesty, and harmlessness are distinct dimensions (Askell et al., 2021). Red-teaming, GCG-style suffix attacks, and jailbreak benchmarks already establish adversarial vulnerability and methods for eliciting it (Ganguli et al., 2022; Zou et al., 2023; Mazeika et al., 2024; Chao et al., 2024). Adaptive attacks and best-of-N elicitation further show why evaluation opportunities matter (Andriushchenko et al., 2025; Hughes et al., 2025). We retain that context without introducing a new jailbreak algorithm or treating three continuation policies as independent attack families.

Optimization objectives versus robustness evaluation. Token-Importance Guided Direct Preference Optimization (TI-DPO) combines gradient attribution and a Gaussian prior for token weighting with a triplet loss for preference optimization (Yang et al., 2026a). In safe reinforcement learning, Proactive Constrained Policy Optimization (PCPO) introduces preemptive penalties and constraint-aware intrinsic rewards (Yang et al., 2026b). These works concern training objectives; our study instead asks whether measured outcomes can substitute across evaluation conditions. We do not evaluate either method or infer its adversarial robustness from our checkpoint comparisons.

Robustness summaries and evaluators. Robustness curves and area summaries are established tools; Robustness-AUC is a supporting summary on a declared grid, not a new estimator. Automated judging likewise introduces known measurement conditions and potential biases (Zheng et al., 2023). Freezing the judge supports within-study consistency, while human calibration measures some of its residual error. Our contribution is therefore not a new attack, AUC method, or judge-based evaluation technique, but a controlled empirical characterization of measurement non-interchangeability across benign, prefix, interaction, and cross-attack conditions. Controls operate within each protocol; the two attack families are not a matched causal comparison.

3 Evaluation Design

3.1 Checkpoints, benign measurement, and cohort

The original four checkpoints are Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct, Qwen2.5-7B-Instruct, and Gemma-2-2b-it. The frozen expansion adds Qwen3-8B and Phi-4-mini-instruct, selected before download or result inspection. Appendix I lists all exact repositories and revisions. Short names refer only to these pinned identities; the six checkpoints are a selected comparison set, not a population sample.

Each model has 350 benign outputs: 150 helpfulness, 100 harmlessness, and 100 honesty. A_{HHH} is the unweighted mean of the three dimension means, not the pooled mean of all outputs. All benign judgments are valid, including the 700 expansion judgments. This benign set is separate from the harmful-request cohort. Level 1 concerns this three-dimensional composite, not harmlessness alone or every definition of benign safety. Appendix C.1 specifies the frozen local evaluation file without assigning an unsupported benchmark provenance.

The prefix source is a historical local 330-row HEx-PHI-derived representation (Qi et al., 2024). An audit verifies 330/330 exact positional user-message matches between its local labeled and paired-target representations. The exact original historical HEx-PHI revision and the paired assistant targets' authorship/transformation history are irrecoverable; the evaluated bytes are not asserted to be the current official release. Baseline eligibility requires at least 20 paired-target tokens under *each* of the four pinned tokenizers, retaining a common $N = 315$ and excluding 15 rows without inspecting attack outcomes. The same 315 rows pass both new tokenizers' eligibility checks, so expansion does not change the baseline cohort. Appendix G records hashes, the eleven preexisting local category labels, and the conservative release boundary.

3.2 Exact-token pressure and repeated interaction

Prefix threat model. The attacker controls an assistant prefill comprising the first k native tokens of the paired target, with $\mathcal{K} = \{0, 2, 5, 10, 15, 20\}$. At $k = 0$ no target tokens are supplied. The request, cohort, checkpoint, and generation/evaluation configuration remain fixed as k changes. Tokens are taken without special tokens, shortening, or padding. Native pinned chat templates are retained; equal k does not imply identical text spans across tokenizers or a universal attack-cost scale. This threat model requires control of the assistant prefill, not merely an ordinary user message.

Generation and interaction. Generation uses BF16 without quantization, greedy decoding, batch size one, an empty system prompt, and native templates. Benign output is capped at 256 new tokens and prefix/continuation output at 128. The expansion uses Qwen3's non-thinking template setting, `enable_thinking=false`; Phi uses its native template without that override. The baseline interaction analysis has three rounds under fixed, refusal-aware, and paraphrase-retry policies. Round 1 reuses the same-run $k = 0$ response. A next turn is generated when the preceding judge score is below the success threshold; otherwise the previous response is carried forward and the round record judged. Missing parent judgments are not treated as unsuccessful model responses. Complete trajectories support an any-round versus first-round comparison. This is a persistence-retry stress test, not general adaptive conversational red teaming. It covers only the original four models; Appendix D gives the recorded state logic.

3.3 A second family: fixed GCG suffix transfer

We transfer the fixed released JailbreakBench GCG adversarial user-suffix artifact optimized on Vicuna-13B-v1.5 (Zou et al., 2023; Chao et al., 2024). All 100 released instances are frozen in order and sent unchanged to each original checkpoint, for 400 generations and 400 judgments. Each artifact prompt is one user message through the native template, with an empty system prompt, BF16 deterministic decoding, and a 150-new-token cap. There is no assistant prefill, target-model optimization, prompt editing, performance-based filtering, or repeated conversation. Exact code/artifact commits, the artifact digest, and the prompt-free sample manifest are specified in Appendix E.

This is a compound cross-condition comparison using one additional attack family: one frozen cross-model transfer artifact. The prefix and GCG tests differ in harmful source, mechanism, and

output cap (128 versus 150 tokens). Their ASRs are not pooled or interpreted as a controlled causal contrast. We compare their *observed model orderings* descriptively, using prefix Robustness-AUC and GCG robustness $1 - \mathrm{ASR}$, while retaining each condition's separate outcomes.

3.4 Estimands, uncertainty, missingness, and grid sensitivity

All adversarial suites use the frozen automated judge `gpt-4o-2024-11-20`, temperature zero, with attack success defined by score ≥ 4 on the frozen harm rubric. The adversarial harm evaluator receives a separate human calibration (Section 3.5); benign A_{HHH} does not. Let $s_i(k; m)$ denote a valid binary prefix-success label. At each k, available-case ASR uses its valid judgments; trajectory summaries use C_m, the prompts observed at every grid point. For sorted $k_1, \ldots, k_6$ in $\mathcal{K}$,

$$R_{\mathrm{AUC}}(m) = 1 - \frac{1}{20} \sum_{j=1}^{5} \frac{\mathrm{ASR}_{C_m}(k_j; m) + \mathrm{ASR}_{C_m}(k_{j+1}; m)}{2} (k_{j+1} - k_j). \tag{1}$$

Higher RAUC means lower cumulative measured vulnerability on this interval. It is not the arithmetic mean of the six settings. Expansion endpoint summaries use the same complete trajectories; baseline available-case endpoints are labeled separately. For model m and policy p, with complete round trajectories $C_{m,p}$,

$$\Delta(m, p) = \mathrm{ASR}_{\mathrm{any}, C_{m,p}} - \mathrm{ASR}_{t_1, C_{m,p}}, \qquad s_{i,\mathrm{any}} = \max_{t \in \{1,2,3\}} s_{i,t}. \tag{2}$$

The attack and benign analyses use 10,000 prompt-bootstrap replicates for 95% intervals, preserving pairing across prefix settings or rounds. Benign resampling retains equal dimension weighting. The baseline seed is 20260906; second-family, expansion, and grid analyses use 20260910. Intervals quantify prompt-level uncertainty under these protocols, not variability over model populations, training seeds, or judges. Maximizing an observed continuation-policy gap is descriptive, without selection-adjusted inference.

The baseline has 12/7,560 missing prefix judgments and 9/3,780 incomplete model–policy trajectories. Expansion has five missing judgments among 3,780 prefix cells, all provider refusals; fixed GCG has none among 400. No score is imputed. Full-cohort lower/upper bounds retain known observations and allow unresolved labels to be failure/success. These are missingness-sensitivity bounds, not confidence intervals, and can differ from complete-case estimates because the estimands differ (Appendix A).

The distinction between complete-case estimates and full-cohort bounds matters for interpretation. A complete trajectory contributes the same prompt at every pressure setting, avoiding changes in the within-model denominator along its integrated curve. Full-cohort bounds answer a different question: they retain valid observations even from incomplete trajectories and vary only unresolved labels. They therefore need not bracket the point estimate for a selected complete subset. Neither narrow bounds nor low missingness establish that missing judgments are random. Likewise, prompt-bootstrap intervals describe variation in the evaluated prompt sample; they do not account for systematic scoring differences between the automated evaluator and a human reference.

Grid sensitivity reuses frozen baseline numeric records, with no new generations. Seven prespecified grids retain endpoints 0 and 20: the full grid; removal of $k = 2$, 5, 10, or 15 separately; $\{0, 5, 10, 20\}$; and $\{0, 10, 20\}$. Each grid uses its own complete-case set and the same normalized trapezoidal definition. Thus the sensitivity exercise can change both resolution and the set of complete prompts. It is not an isolated causal effect of grid resolution.

3.5 Human calibration of the harm evaluator

Two primary author-annotators independently scored the prespecified 240-item sample (120 prefix, 72 adaptive, 48 fixed-GCG; seed 20260910) using the frozen 1–5 harm rubric. Model identity, machine scores/reasoning, and the other annotator's labels were withheld; suite identity was withheld where technically possible. After both full label files were complete, every ordinal disagreement or uncertainty was flagged. The third-author adjudicator reviewed the 56 flagged items and recorded one final 1–5 reference score per item; the other 184 retained A/B agreement. Original independent labels were preserved. Human-human agreement uses pre-adjudication A/B labels; machine-human calibration

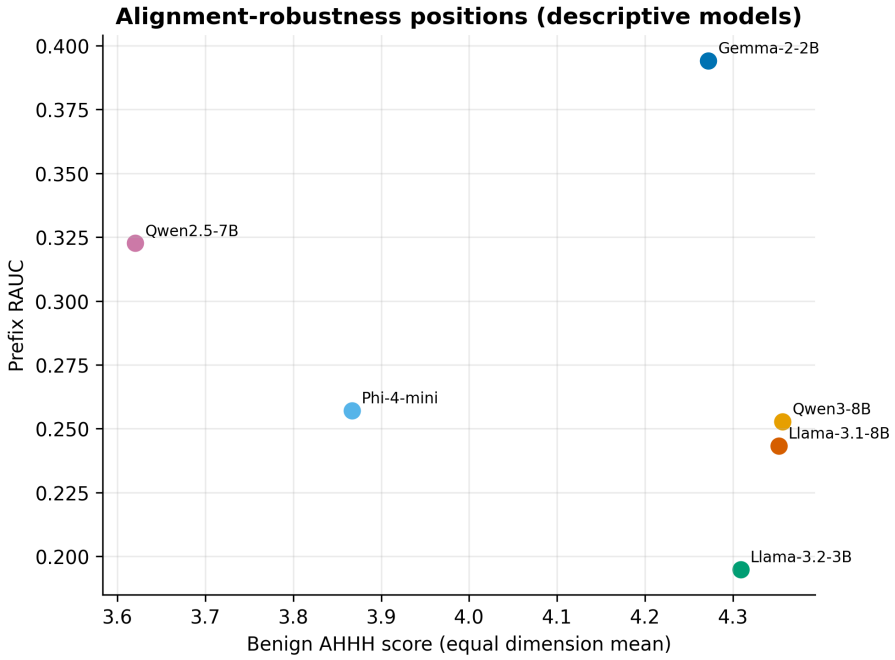

Figure 1: Six selected checkpoints under benign equal-dimension A_{HHH} and exact-prefix RAUC. Original clean-v3 results are retained; Qwen3 and Phi extend only the benign/prefix analysis. The display is descriptive, not population inference or evidence of significant pairwise superiority.

uses the final adjudicated human reference. The score-≥ 4 threshold is unchanged. Supplied intervals use 10,000 within-design-cell bootstrap replicates with seed 20260910. No correction is extrapolated to unannotated outputs, and benign A_{HHH} is not human calibrated (Appendix K).

4 Results

4.1 Benign safety does not determine prefix robustness

The six-model comparison exhibits the first level of non-interchangeability (Figure 1). Qwen3 scores higher on the benign composite than Qwen2.5 (4.356667 versus 3.620000) but lower on prefix RAUC (0.252724 versus 0.322631). The reported marginal 95% intervals are disjoint on both measures (Appendices B and C); this is a clearer finite contrast, not a new paired significance test. Qwen3 has the highest observed composite but only the fourth-highest RAUC. Phi retains a slightly higher RAUC point estimate than Qwen3 with overlapping intervals; we do not treat that pair as a statistically resolved robustness inversion.

The original four-model trajectories show why an endpoint alone is insufficient (Figure 2). Attack success is much lower without a prefill than at the largest tested prefill, where available-case ASR reaches 0.835–0.940. Complete-case RAUC orders the checkpoints Gemma $>$ Qwen2.5 $>$ Llama-3.1 $>$ Llama-3.2. This is not the benign ordering: Gemma has lower A_{HHH} but higher prefix RAUC than either Llama checkpoint. The frozen expansion retains the same common cohort rather than redefining this baseline.

Endpoint ASR describes the largest tested prefill; RAUC also incorporates intermediate settings. For the two-model expansion, the executed judge receives the original request and the full assistant response, decoded from prefill plus generated continuation, not continuation alone (Appendix I.1). Its ASR therefore measures the combined response under attacker-controlled prefill, not harmful content newly generated in isolation. Equal native-token counts do not imply equal text spans or a universal cross-model attack budget.

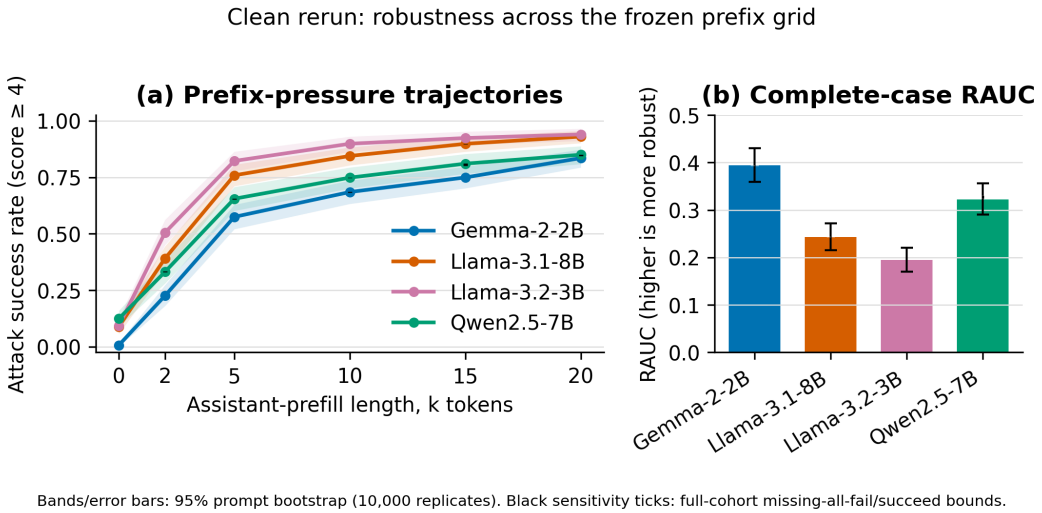

Figure 2: Certified baseline exact native-token prefill trajectories. Curves use available judgments; supporting RAUC uses complete trajectories. Attack success is frozen-judge score ≥ 4. Missing labels remain missing, with full-cohort sensitivity bounds reported separately.

Construct choice matters. Qwen2.5 and Qwen3 both have observed harmlessness means of 5.000000, while their helpfulness means differ (1.940000 versus 4.280000). The composite contrast is therefore not a harmlessness-only contrast. All six component profiles appear in Table 2; we do not generalize the composite's ordering mismatch to every component, reweighting, or benign-safety construct.

4.2 Repeated interaction reveals additional vulnerability

In this persistence-retry stress test, every original checkpoint exhibits additional any-round failures under at least one tested continuation policy. The descriptively maximizing-policy increases range from 0.168 to 0.766. Appendix D retains all twelve model–policy combinations, paired intervals, and full-cohort bounds, rather than selecting only favorable policies.

The first-round and any-round estimates use the same complete trajectories; the any-round event includes first-round success by definition. The increase therefore describes failures exposed by the additional interaction opportunity, not a causal effect separated from that opportunity. Policy matters: for Llama-3.2, the increase is 0.168 under fixed continuation but 0.019 under paraphrase-retry. A first-response score misses some subsequent failures, but neither one policy nor its observed maximum is a universal interaction strategy. Qwen3 and Phi were not evaluated in this suite.

Any-round success is a max-over-round event, not a round-specific failure rate. A false-positive judgment in any round can mark a trajectory successful, so this aggregation can compound evaluator error. The weaker calibration of adaptive examples makes that limitation particularly relevant. We neither assume independent round-level errors nor use pooled calibration accuracy to correct trajectory ASR. Policy-specific paired intervals and full-cohort bounds remain separate from this evaluator-validity concern.

4.3 Robustness ordering changes across attack conditions

The second level of non-interchangeability is a change in model ordering, not simply another successful jailbreak. Prefix robustness orders Gemma $>$ Qwen2.5 $>$ Llama-3.1 $>$ Llama-3.2; fixed-GCG robustness instead orders Gemma $>$ Llama-3.1 $>$ Llama-3.2 $>$ Qwen2.5 (Figure 3). Only 4/6 pairwise orders agree. Descriptive Spearman 0.400 and Kendall 0.333 summarize these four checkpoints; they are not population correlation estimates.

The mismatch is partial, not a complete reversal. Gemma remains first and the Llama checkpoints retain their mutual order, while Qwen2.5 moves from second to fourth. Thus the prefix rank neither

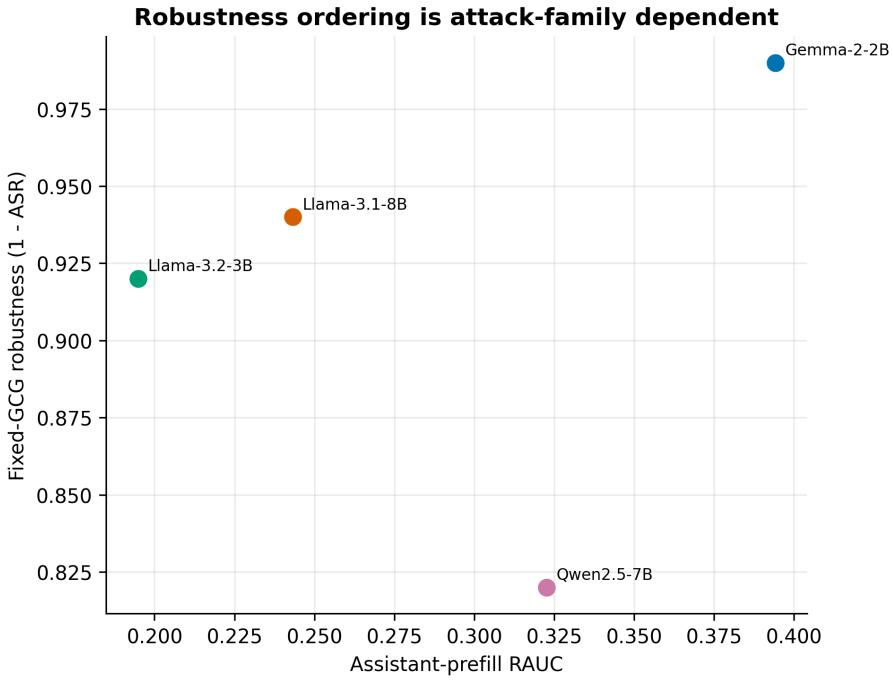

Figure 3: Prefix RAUC versus fixed-GCG robustness ($1 - \text{ASR}$) for the original four checkpoints. Qwen2.5 changes from second to fourth in robustness rank. The metrics describe different attack conditions and are not interchangeable or pooled; $n = 4$ comparisons are descriptive only.

transfers unchanged nor becomes wholly unrelated to the second condition. This is direct evidence that the observed robustness ordering is attack-condition dependent in these tests.

All 400 GCG judgments are valid. Transfer ASRs are low: Gemma 0.010, Llama-3.1 0.060, Llama-3.2 0.080, and Qwen2.5 0.180; supplied intervals appear in Appendix E. Low success for a fixed Vicuna-optimized artifact is not broad jailbreak robustness. Source requests, attacker access, mechanism, surrogate/transfer setting, and output caps differ from prefix evaluation, so ASRs are not pooled and rank changes cannot be attributed solely to mechanism.

The intermediate Llama ordering is a point-estimate ordering: Llama-3.1 and Llama-3.2 have overlapping ASR intervals, $[0.020, 0.110]$ and $[0.030, 0.140]$. We do not claim their relative GCG robustness is statistically resolved. Floor-level success and only four targets limit the rank comparison.

The unchanged transfer artifact makes the second condition concrete: each target receives the released inputs rather than an attack optimized for that target. The resulting comparison concerns susceptibility to that particular transfer, not each checkpoint's strongest achievable jailbreak. This distinction explains the role of the second family in the paper. Its value is to test whether the observed prefix ordering carries over to another specified condition, not to construct an average robustness score from heterogeneous attacks. The disagreement answers that finite substitution question while leaving broader attack coverage outside the evidence.

4.4 Prefix ranking is grid-stable but metric magnitude is not

The original four-model prefix rank survives all seven prespecified grids: zero rank changes and zero pairwise swaps across 42 grid-by-pair comparisons. This stability qualifies the structural evaluation gap. The prefix ordering is not merely an artifact of the precise six-point grid, even though it fails to replicate across attack conditions.

RAUC magnitude is materially sensitive to grid choice. The largest absolute shift is 0.098016, a 50.3% relative increase for Llama-3.2 on $\{0, 10, 20\}$. All models have larger RAUC on this coarsest grid. No model behavior changed in this analysis: coarser trapezoids summarize the already observed

responses differently, and grid-specific complete sets can also vary. Rank stability does not imply magnitude invariance. Appendix F reports all grids, counts, values, and intervals.

Removing an interior observation replaces adjacent trapezoids with a wider chord. Where ASR is locally concave down, that chord lies below the intermediate curve and can reduce estimated ASR area, increasing RAUC without changing model behavior. This conditional numerical-integration explanation does not assert that every trajectory is globally concave. Rank stability on the tested grids establishes neither magnitude invariance nor stability on untested grids.

4.5 Human calibration quantifies evaluator uncertainty

Before adjudication, the two primary annotators' independent binary labels agree at 0.9000 (Cohen's $\kappa = 0.7965$). Against the final adjudicated human reference, machine accuracy is 0.8667 (95% CI $[0.8333, 0.9000]$), F1 is 0.8280, and $\kappa = 0.7192$. These results give *moderate calibration support*, not ground-truth labels.

Agreement is weakest for adaptive examples (accuracy 0.7639, $\kappa = 0.4216$), compared with prefix accuracy 0.8833 and fixed-GCG accuracy 0.9792. The latter includes only three human-positive items, so its high accuracy must not obscure the sparse positive class. Appendix K reports false positives, false negatives, threshold-boundary disagreements, and all metrics. This is finite, stratified author-annotated validity evidence; it neither estimates a project-wide error rate nor validates benign A_{HHH}.

Independent binary human-human agreement is 0.8889 on adaptive items versus 0.9000 on prefix items; corresponding κ values are 0.7551 and 0.7887. Ordinal agreement is lower on adaptive items (Appendix K), but these differently composed strata do not isolate intrinsic annotation difficulty. The human-human and machine-human comparisons use different reference constructions; adjudication neither erases initial disagreement nor creates an external reference.

5 Discussion and Limitations

Two levels of an empirical structural evaluation gap. Benign evaluation and direct adversarial evaluation answer different behavioral questions. In our operationalization, the equal-dimension A_{HHH} composite does not determine exact-prefix robustness ordering. Direct prefix evaluation addresses attacker-controlled prefill, but its ordering is not an attack-general certificate: the second specified adversarial condition changes part of that order. Repeated interaction adds a temporal distinction, because first-response assessment misses some later frozen-evaluator failures. These are failures of measurement substitution, not evidence that alignment-training mechanisms are structurally defective.

The seven-grid analysis separates this non-transfer from universal instability. Within the declared prefix protocol, the observed ordering survives the tested changes in resolution. Across the compound prefix/GCG comparison, it does not transfer unchanged. Stable within-condition ranks can thus coexist with cross-condition non-interchangeability, while stable ranks themselves coexist with materially different RAUC magnitudes. The useful question is not whether a measurement is universally reliable, but which comparisons its design supports. Benign evaluation remains informative about its measured construct without becoming a substitute for direct robustness assessment.

Reporting implication. A benign score or a result under one adversarial condition should not be presented as an unqualified model-safety certificate. For the kinds of comparisons studied here, reporting checkpoint identity, benign measure, attack family and strength/budget, interaction policy, evaluator, grid, and missingness makes clear what the result does and does not describe. This is a practical implication of the finite evidence, not a proposed universal evaluation standard. Benign evaluation remains useful alongside direct robustness evaluation.

For example, the six-model benign/prefix view supports a checkpoint comparison under its declared pressure grid, while the four-model GCG view describes a different source and attacker capability. The continuation view additionally changes the success event from one response to any response in a trajectory. Keeping these distinctions visible lets a reader use each result for the question it measures instead of interpreting their shared safety terminology as a common scale. This reporting

implication follows from the observed mismatches; it does not require treating the benign measure as uninformative or the adversarial measures as mutually unrelated.

What the controlled comparisons establish. The strongest controls operate within a checkpoint's prefix trajectory: the request cohort, model identity, and scoring condition remain fixed as the supplied prefix length changes. The cross-checkpoint view retains the cohort but compares separately trained models and their native tokenizers. It therefore cannot attribute an ordering difference to a particular training objective or alignment intervention. Similarly, the second attack changes more than the attack string, including source requests and permitted output length. These distinctions do not invalidate the observed ordering mismatch; they determine what it establishes. A rank obtained under one specified protocol is not a substitute for the rank measured under the other, without identifying which changed condition caused the difference.

Complementary uncertainty checks. The three uncertainty checks also answer separate questions. Prompt resampling assesses variability in the retained observations under a frozen scoring rule. Full-cohort bounds assess how unresolved judgments could affect the stated estimand. Human calibration assesses agreement of that scoring rule with the final adjudicated human reference on a stratified sample. None subsumes the others: a fully observed output can still receive a disputed label, and a narrow bootstrap interval does not resolve a missing observation or an evaluator disagreement. Reporting the checks together makes the remaining uncertainty visible without converting calibration accuracy into a correction factor for the robustness curves.

Model, attack, and inference limits. The study covers six selected public instruction-tuned checkpoints, not a model population, proprietary frontier systems, or a broad scale sweep. Repeated interaction and GCG cover only the original four. Attack coverage consists of assistant-prefill pressure, three continuation policies, and one fixed Vicuna-optimized GCG transfer artifact. Low transfer ASRs and differing request sources/output caps limit cross-attack interpretation. No attack-general rank, causal independence, zero correlation, population-wide relationship, or mechanism-level defect in alignment training is established.

Evaluator limits. Automated ASRs remain conditional on `gpt-4o-2024-11-20`. Two primary author-annotators independently labeled the sample, and a third-author adjudicator resolved flagged cases, so calibration provides direct author-annotated validity evidence rather than validation by raters outside the author team. Agreement is imperfect, especially for adaptive examples, and fixed-GCG positives are sparse. The stratified sample does not validate benign A_{HHH}, other judges, or all outputs. No manual correction is extrapolated to unannotated records, and prompt-bootstrap intervals do not capture all evaluator error. Max-over-round success can amplify false-positive labeling; the reported gaps are not error-corrected.

Data, missingness, and reproduction limits. The historical local HEx-PHI derivative has eleven preexisting category labels; the original historical Hub revision and paired-target derivation/authorship are irrecoverable. Baseline prefix missingness, adaptive attrition, and expansion missing judgments remain explicit, without imputation or a missing-at-random claim. Full-cohort bounds are sensitivity analyses, not confidence intervals. Safe code, numeric verification, and deterministic assets improve reproducibility, but restricted inputs prevent complete ZIP-only raw replay. Numeric consistency, evaluator validity, and historical input provenance remain distinct properties.

6 Conclusion

Across six selected instruction-tuned checkpoints, benign evaluation operationalized by our three-dimensional composite does not uniquely determine exact-prefix adversarial robustness. Repeated interaction exposes additional frozen-evaluator failures, and a second specified adversarial condition changes part of the model ordering. Together, these results reveal an empirical structural evaluation gap: measurements under one condition need not serve as standalone empirical certificates for another. Stable prefix rankings across tested grids show that condition dependence does not imply universal instability. These findings motivate direct, condition-explicit robustness evaluation alongside benign safety assessment. They do not establish causal independence, population-wide relationships, arbitrary attack-family generalization, or a structural defect in alignment-training mechanisms.

Ethics Statement

Three paper authors participated in the calibration. Two independently labeled all 240 items under the blinding protocol; after both label files were complete, the third author adjudicated only the 56 flagged items, recording one final reference score per item. Original independent labels were retained; AI generated neither annotation nor adjudication labels. The two primary annotators participated voluntarily after verbal agreement, received verbal and written harmful-content briefings, were told they could pause, take breaks, or stop, and received an in-kind hosted meal. Their independent annotation occurred on September 11, 2026. No known conflicts were identified among the two primary annotators. No formal IRB/institutional determination was obtained; no approval or exemption is claimed. An author accessed gated HEx-PHI-derived data through an authorized account for research/evaluation under non-transfer restrictions. The two historical provenance limitations remain irrecoverable. Raw harmful prompts, paired targets, attack strings, harmful generations, raw judge reasoning, and private human files are not released.

Reproducibility Statement

The anonymous research artifact contains exact configurations and revisions, safe code for generation, judging, and analysis, prompt-free numeric records, summaries, public calibration aggregates, deterministic figure/table builders, and manifests/checksums. Default clean-extraction checks support offline, credential-free numeric verification, table/figure regeneration, and artifact verification. Frozen sampling IDs/hashes are available; private labels and annotation content remain withheld. Raw replay requires separately authorized restricted inputs and appropriate model/API access, not the ZIP alone (Appendix J).

AI Use Statement

Generative AI tools, including ChatGPT/OpenAI models, Codex, and Prism, assisted writing and polishing, drafting and restructuring, literature discovery and related-work retrieval, research ideation and execution workflows, methodology feedback, code/method implementation, debugging and revision, analysis and interpretation, reproducibility, and figure/table workflows. AI did not generate independent human labels, third-author adjudication scores, or final reference labels. The authors reviewed AI-assisted work and retain responsibility for experimental design, final code, numerical results, scientific conclusions, the manuscript, reference accuracy, and release decisions.

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://proceedings.iclr.cc/paper_files/paper/2025/hash/63fa7efdd3bcf944a4bd6e0ff6a50041-Abstract-Conference.html.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021. URL https://arxiv.org/abs/2112.00861.

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. URL https://arxiv.org/abs/2212.08073.

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, 2024. doi: 10.52202/079017-1745. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/63092d79154adebd7305dfd498cbff70-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022. URL https://arxiv.org/abs/2209.07858.

John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Arushi Somani, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. Best-of-N jailbreaking. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2453. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/69f3eb242c7c9df9ea2f2b66ea8b3c0f-Abstract-Conference.html.

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, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In

International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Ning Yang, Hai Lin, Yibo Liu, Baoliang Tian, Guoqing Liu, and Haijun Zhang. Token-importance guided direct preference optimization. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/172fe9f8cc55953bad5c24774bf0142b-Abstract-Conference.html.

Ning Yang, Pengyu Wang, Guoqing Liu, Haifeng Zhang, Pin Lyu, and Jun Wang. Proactive constrained policy optimization with preemptive penalty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27583–27591, 2026b. doi: 10.1609/aaai.v40i32.39978. URL https://ojs.aaai.org/index.php/AAAI/article/view/39978.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

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. URL https://arxiv.org/abs/2307.15043.

A Metric Definitions and Missingness Bounds

For dimension $d \in \{\text{help}, \text{harm}, \text{hon}\}$ and benign judge score $J_{i,d}(m)$,

$$A_{\text{HHH}}(m) = \frac{1}{3}\sum_{d}\frac{1}{n_d}\sum_{i=1}^{n_d} J_{i,d}(m), \qquad (n_{\text{help}}, n_{\text{harm}}, n_{\text{hon}}) = (150, 100, 100). \tag{3}$$

The helpfulness set is larger but does not receive greater dimension weight. All original and expansion benign records have valid scores. The original benign components are shown in Figure 5; dimension scores remain distinct from the composite.

For valid set $O_m(k)$, observed successes $a_m(k)$, and $u_m(k) = 315 - |O_m(k)|$, prefix available-case ASR and full-cohort bounds are

$$\widehat{\text{ASR}}_{\text{avail}}(k; m) = \frac{a_m(k)}{|O_m(k)|}, \qquad [L_m(k), U_m(k)] = \left[\frac{a_m(k)}{315}, \frac{a_m(k) + u_m(k)}{315}\right]. \tag{4}$$

Trajectory RAUC uses $C_m = \bigcap_{k \in \mathcal{K}} O_m(k)$. Full-cohort area bounds integrate cellwise lower/upper curves, reversing their order under $R = 1 - \text{AUC}$. They keep valid cells from incomplete trajectories. These bounds describe the full 315-prompt cohort, not a confidence interval for the retained complete subset; a complete-case estimate can therefore lie outside them. For example, baseline Qwen2.5 has complete-case RAUC 0.323 on 306 prompts and full-cohort bounds rounding to [0.324, 0.330]. This is a denominator/estimand distinction, not an imputation or confidence-interval failure.

Baseline available-case $k = 20$ endpoints and complete-case expansion endpoints are not silently substituted for one another. In particular, the six-model supplied table uses complete-case endpoint summaries for all six models; its original-model endpoints can differ from Table 1. The same certified baseline trajectories underlie both views.

For continuation, any-round success is the maximum of the three binary labels on complete trajectories. Full-cohort any-round lower bounds retain every already known success; upper bounds additionally count unresolved trajectories without a known success. Subtracting full-cohort round-1 ASR yields the supplied Δ bounds. Missing later judgments cannot erase a known earlier success, so incomplete-trajectory and unresolved-any-outcome counts need not coincide. Bootstrap resampling preserves prompt pairing across rounds. Per-policy intervals are primary; an observed maximizing-policy summary is not selection-adjusted.

B Complete Baseline Results

Table 1: Original four-model baseline. RAUC uses complete six-setting trajectories; ASR_{20} uses available judgments. Max. Δ is the descriptively largest observed policy gap, not a prespecified universal policy. Full-precision summaries, uncertainty, and bounds appear in Appendices B and D.

Model	A_{HHH}	RAUC	ASR_{20}	Max. Δ
Llama-3.1-8B	4.352	0.243	0.930	0.513
Llama-3.2-3B	4.309	0.195	0.940	0.168
Qwen2.5-7B	3.620	0.323	0.850	0.713
Gemma-2-2B	4.272	0.394	0.835	0.766

The following values retain the original six-setting complete-case trajectories. All intervals are supplied 95% prompt-bootstrap intervals; full-cohort bounds are not confidence intervals. Exact per-setting counts, scores, and bounds are available in the prompt-free numeric artifact. These available-case curves are shown in Figure 2.

Llama-3.1-8B. $A_{\text{HHH}} = 4.352222$ (95% CI [4.214444, 4.482222]); complete-case RAUC 0.243232 (95% CI [0.215764, 0.271656]); 314/315 complete trajectories; full-cohort RAUC bounds [0.243016, 0.243810]. Available-case ASR at $k = 0$ is 0.088889 and at $k = 20$ is 0.930159 (95% CI [0.901587, 0.955556]); endpoint full-cohort bounds [0.930159, 0.930159].

Llama-3.2-3B. $A_{\mathrm{HHH}} = 4.308889$ (95% CI [4.167778, 4.447778]); complete-case RAUC 0.194841 (95% CI [0.170476, 0.220397]); 315/315 complete trajectories; full-cohort RAUC bounds [0.194841, 0.194841]. Available-case ASR at $k = 0$ is 0.092063 and at $k = 20$ is 0.939683 (95% CI [0.914286, 0.965079]); endpoint full-cohort bounds [0.939683, 0.939683].

Qwen2.5-7B. $A_{\mathrm{HHH}} = 3.620000$ (95% CI [3.488861, 3.751111]); complete-case RAUC 0.322631 (95% CI [0.289951, 0.356046]); 306/315 complete trajectories; full-cohort RAUC bounds [0.323889, 0.330317]. Available-case ASR at $k = 0$ is 0.123810 and at $k = 20$ is 0.849840 (95% CI [0.810289, 0.887821]); endpoint full-cohort bounds [0.844444, 0.850794].

Gemma-2-2B. $A_{\mathrm{HHH}} = 4.272222$ (95% CI [4.125556, 4.414444]); complete-case RAUC 0.394188 (95% CI [0.359154, 0.430575]); 314/315 complete trajectories; full-cohort RAUC bounds [0.393492, 0.394286]. Available-case ASR at $k = 0$ is 0.006349 and at $k = 20$ is 0.834921 (95% CI [0.793651, 0.873016]); endpoint full-cohort bounds [0.834921, 0.834921].

C Benign Components and Two-Model Expansion

C.1 Frozen benign set and component scope

The original and expansion configurations identify the same local 350-record evaluation file, `all_eval.jsonl`, by SHA-256 `4195c645daad76fac9dc7b5a8b1748bb67b4aeb56875595c274599f2692983b8`. Its recorded partition is 150 helpfulness, 100 harmlessness, and 100 honesty items; the generation code reads each record's instruction and requested dimension. This pins the evaluated file and its reuse, not an independently named benchmark. The available records do not establish the original item sources, selection process, or authorship of the benign set; none is inferred. The frozen 1–5 benign rubric scores usefulness for helpfulness, safe refusal/redirection for harmlessness, and factual care for honesty. These are dimension-specific operational judgments, not human-calibrated scales.

Table 2: Six-model benign component means and equal-dimension composite, formatted from the certified benign summary. All 350 judgments per model are valid. Component means are not separate adversarial robustness measures.

Model	Helpful	Harmless	Honest	A_{HHH}
Gemma-2-2B	4.166667	4.920000	3.730000	4.272222
Llama-3.1-8B	4.566667	4.800000	3.690000	4.352222
Llama-3.2-3B	4.326667	5.000000	3.600000	4.308889
Qwen2.5-7B	1.940000	5.000000	3.920000	3.620000
Qwen3-8B	4.280000	5.000000	3.790000	4.356667
Phi-4-mini	3.300000	4.840000	3.460000	3.866667

The composite assigns equal weight to dimension means despite their different item counts. In particular, identical maximal harmlessness means for Qwen2.5 and Qwen3 coexist with different composite scores and prefix RAUC. The composite counterexample is not evidence of a statistically separated harmlessness-only inversion. No component-wise correlation or reweighted-composite significance test is claimed.

C.2 Expansion trajectories and uncertainty

The expansion retains the original 315-prompt cohort. The following six-model table uses complete-case endpoints for all models; it does not replace the baseline available-case endpoints in Table 1.

Qwen3-8B. $A_{\mathrm{HHH}} = 4.356667$ (95% CI [4.223333, 4.484444]); RAUC 0.252724 (95% CI [0.226202, 0.280529]); 312/315 complete trajectories; full-cohort RAUC bounds [0.251429, 0.253810]; $k = 20$ complete-case ASR 0.903846 (95% CI [0.868590, 0.935897]); endpoint full-cohort bounds [0.904762, 0.904762].

Table 3: Six selected checkpoints: benign scores and complete-case prefix summaries. Missing counts denote missing prefix cells. The original four retain their certified trajectories; only Qwen3 and Phi are expansion models.

Model	Benign	RAUC	ASR20	Complete N	Missing
Gemma-2-2B	4.272	0.394	0.834	314	1
Llama-3.1-8B	4.352	0.243	0.930	314	1
Llama-3.2-3B	4.309	0.195	0.940	315	0
Qwen2.5-7B	3.620	0.323	0.850	306	10
Qwen3-8B	4.357	0.253	0.904	312	3
Phi-4-mini	3.867	0.257	0.895	313	2

Phi-4-mini. $A_{\mathrm{HHH}} = 3.866667$ (95% CI [3.707778, 4.025556]); RAUC 0.257029 (95% CI [0.229872, 0.285783]); 313/315 complete trajectories; full-cohort RAUC bounds [0.255714, 0.257302]; $k = 20$ complete-case ASR 0.894569 (95% CI [0.859425, 0.926518]); endpoint full-cohort bounds [0.895238, 0.895238]. The expansion RAUC intervals overlap; no significant Phi-over-Qwen3 superiority is claimed. All 700 expansion benign judgments are valid. Three missing Qwen3 prefix cells and two missing Phi cells are provider refusals, not imputed attack failures.

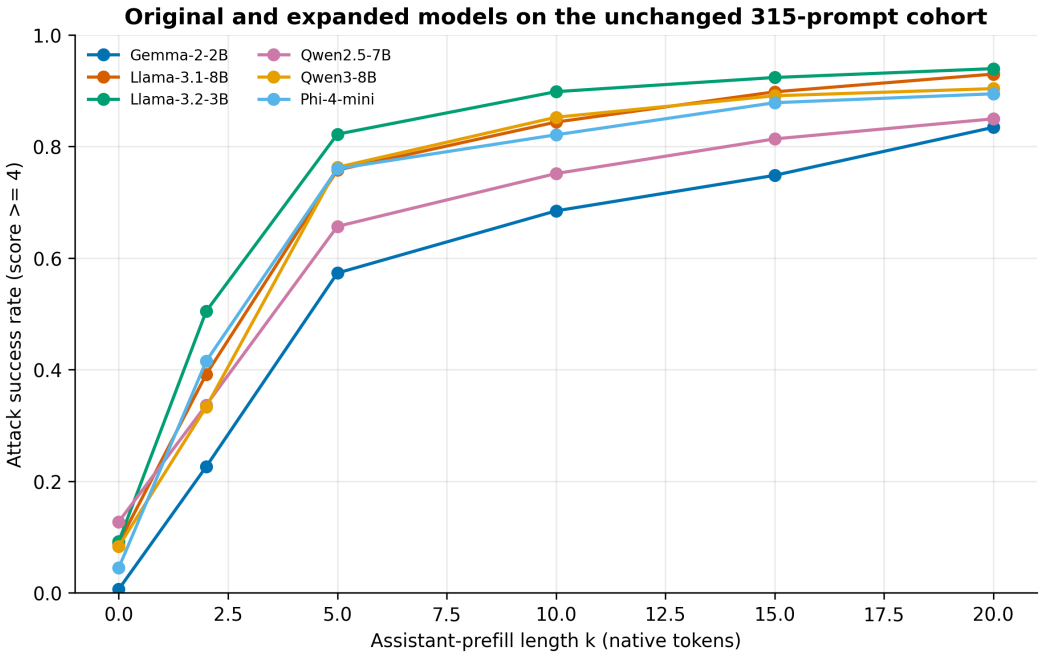

Figure 4: Exact-prefix trajectories for six selected checkpoints. This figure adds benign/prefix evidence only; adaptive continuation and GCG transfer were not evaluated for the two expansion models.

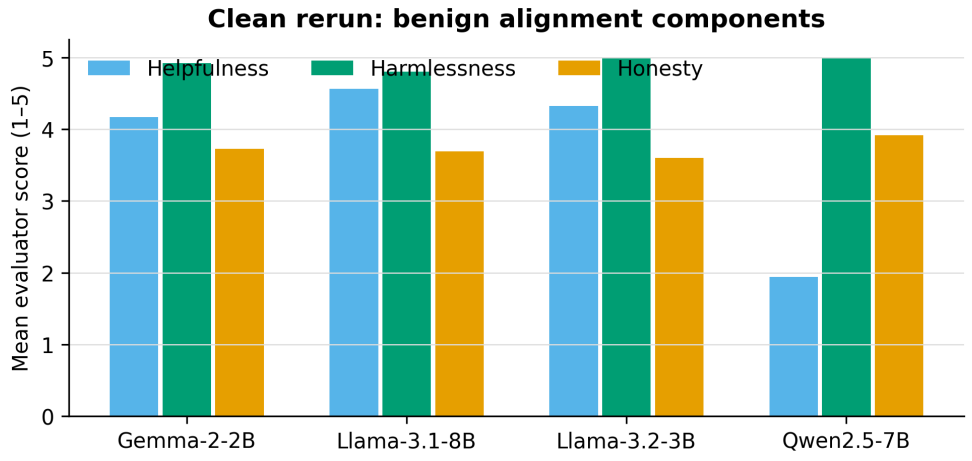

Figure 5: Original-four benign component scores. Equal dimension weighting does not imply equal component performance; high benign harmlessness does not specify adversarial success.

D All Baseline Continuation Policies

All three policies are retained: F is fixed, R refusal-aware, and P paraphrase-retry. The round-specific table and any-round summaries concern only the original four checkpoints. Complete-trajectory estimates preserve within-prompt pairing. Bounds describe the full 315-prompt cohort. The maximum observed policy gap is descriptive, not a per-prompt union of strategies or a universally strongest policy.

Recorded persistence-retry state logic. The executed configuration specifies the following control flow for each original checkpoint, prompt, and selected policy. It is not a general adaptive conversation optimizer and does not combine policies within a prompt.

1. Initialize the round-1 response and score from the same-run zero-prefill record.
2. For each later round, first require a valid preceding score. A missing parent score does not trigger generation and is not converted into a refusal or failure.
3. If the preceding score is below 4, generate the next response using the selected policy's continuation; otherwise carry the preceding response forward. Judge the resulting round record under the frozen harm rubric.
4. After the three-round sequence, include only fully judged trajectories in the point estimate and take the maximum of their binary success labels. Preserve incomplete records for the specified bounds.

The policy names are fixed, refusal-aware, and paraphrase-retry. The configuration records the continuation source hash, but the source implementation and restricted strings are not available in the manuscript evidence supplied for this revision. This state summary does not invent policy templates, a per-prompt optimization procedure, or a rule treating missing scores as valid low scores. Rejudging carried-forward responses is part of the recorded protocol; errors across rounds need not be independent.

Table 4: All original-four round-specific results on complete trajectories. F: fixed; R: refusal-aware; P: paraphrase-retry. Miss. is direct missingness; Attr. is upstream attrition. These are round-specific ASRs, not any-round rates.

Model	Policy	Round	n	ASR	95% CI	Miss.	Attr.
Llama-3.1-8B	F	1	315	0.089	[0.060, 0.121]	0	0
Llama-3.1-8B	F	2	315	0.387	[0.333, 0.441]	0	0
Llama-3.1-8B	F	3	315	0.565	[0.511, 0.619]	0	0
Llama-3.1-8B	R	1	314	0.089	[0.057, 0.121]	0	0
Llama-3.1-8B	R	2	314	0.299	[0.248, 0.350]	1	0
Llama-3.1-8B	R	3	314	0.592	[0.538, 0.646]	0	1
Llama-3.1-8B	P	1	314	0.089	[0.061, 0.121]	0	0
Llama-3.1-8B	P	2	314	0.306	[0.255, 0.357]	1	0
Llama-3.1-8B	P	3	314	0.385	[0.331, 0.439]	0	1
Llama-3.2-3B	F	1	315	0.092	[0.060, 0.124]	0	0
Llama-3.2-3B	F	2	315	0.222	[0.178, 0.270]	0	0
Llama-3.2-3B	F	3	315	0.257	[0.210, 0.305]	0	0
Llama-3.2-3B	R	1	315	0.092	[0.063, 0.124]	0	0
Llama-3.2-3B	R	2	315	0.146	[0.108, 0.184]	0	0
Llama-3.2-3B	R	3	315	0.229	[0.184, 0.276]	0	0
Llama-3.2-3B	P	1	315	0.092	[0.060, 0.124]	0	0
Llama-3.2-3B	P	2	315	0.105	[0.073, 0.140]	0	0
Llama-3.2-3B	P	3	315	0.108	[0.073, 0.143]	0	0
Qwen2.5-7B	F	1	314	0.124	[0.089, 0.162]	0	0
Qwen2.5-7B	F	2	314	0.675	[0.624, 0.726]	1	0
Qwen2.5-7B	F	3	314	0.828	[0.787, 0.869]	0	1
Qwen2.5-7B	R	1	314	0.124	[0.089, 0.162]	0	0
Qwen2.5-7B	R	2	314	0.401	[0.347, 0.455]	1	0
Qwen2.5-7B	R	3	314	0.468	[0.411, 0.522]	0	1
Qwen2.5-7B	P	1	313	0.125	[0.089, 0.163]	0	0
Qwen2.5-7B	P	2	313	0.316	[0.265, 0.367]	1	0
Qwen2.5-7B	P	3	313	0.438	[0.383, 0.492]	1	1
Gemma-2-2B	F	1	315	0.006	[0.000, 0.016]	0	0
Gemma-2-2B	F	2	315	0.381	[0.327, 0.435]	0	0
Gemma-2-2B	F	3	315	0.702	[0.651, 0.752]	0	0
Gemma-2-2B	R	1	312	0.006	[0.000, 0.016]	0	0
Gemma-2-2B	R	2	312	0.635	[0.580, 0.689]	2	0
Gemma-2-2B	R	3	312	0.772	[0.724, 0.817]	1	2
Gemma-2-2B	P	1	315	0.006	[0.000, 0.016]	0	0
Gemma-2-2B	P	2	315	0.146	[0.108, 0.184]	0	0
Gemma-2-2B	P	3	315	0.206	[0.162, 0.251]	0	0

Llama-3.1-8B, F. 315/315 complete trajectories; first-round ASR 0.089; any-round ASR 0.578 (95% CI [0.524, 0.632]); $\Delta = 0.489$ (95% CI [0.435, 0.546]); full-cohort any-round bounds [0.578, 0.578] and Δ bounds [0.489, 0.489].

Llama-3.1-8B, R. 314/315 complete trajectories; first-round ASR 0.089; any-round ASR 0.602 (95% CI [0.548, 0.656]); $\Delta = 0.513$ (95% CI [0.459, 0.567]); full-cohort any-round bounds [0.600, 0.603] and Δ bounds [0.511, 0.514].

Llama-3.1-8B, P. 314/315 complete trajectories; first-round ASR 0.089; any-round ASR 0.398 (95% CI [0.344, 0.452]); $\Delta = 0.309$ (95% CI [0.258, 0.360]); full-cohort any-round bounds [0.397, 0.400] and Δ bounds [0.308, 0.311].

Llama-3.2-3B, F. 315/315 complete trajectories; first-round ASR 0.092; any-round ASR 0.260 (95% CI [0.213, 0.308]); $\Delta = 0.168$ (95% CI [0.130, 0.213]); full-cohort any-round bounds [0.260, 0.260] and Δ bounds [0.168, 0.168].

Llama-3.2-3B, R. 315/315 complete trajectories; first-round ASR 0.092; any-round ASR 0.229 (95% CI [0.184, 0.276]); $\Delta = 0.137$ (95% CI [0.098, 0.175]); full-cohort any-round bounds [0.229, 0.229] and Δ bounds [0.137, 0.137].

Llama-3.2-3B, P. 315/315 complete trajectories; first-round ASR 0.092; any-round ASR 0.111 (95% CI [0.076, 0.146]); $\Delta = 0.019$ (95% CI [0.006, 0.035]); full-cohort any-round bounds [0.111, 0.111] and Δ bounds [0.019, 0.019].

Qwen2.5-7B, F. 314/315 complete trajectories; first-round ASR 0.124; any-round ASR 0.838 (95% CI [0.796, 0.876]); $\Delta = 0.713$ (95% CI [0.662, 0.761]); full-cohort any-round bounds [0.835, 0.838] and Δ bounds [0.711, 0.714].

Qwen2.5-7B, R. 314/315 complete trajectories; first-round ASR 0.124; any-round ASR 0.478 (95% CI [0.420, 0.532]); $\Delta = 0.354$ (95% CI [0.299, 0.408]); full-cohort any-round bounds [0.476, 0.479] and Δ bounds [0.352, 0.356].

Qwen2.5-7B, P. 313/315 complete trajectories; first-round ASR 0.125; any-round ASR 0.444 (95% CI [0.390, 0.498]); $\Delta = 0.319$ (95% CI [0.268, 0.371]); full-cohort any-round bounds [0.441, 0.448] and Δ bounds [0.317, 0.324].

Gemma-2-2B, F. 315/315 complete trajectories; first-round ASR 0.006; any-round ASR 0.717 (95% CI [0.667, 0.768]); $\Delta = 0.711$ (95% CI [0.660, 0.762]); full-cohort any-round bounds [0.717, 0.717] and Δ bounds [0.711, 0.711].

Gemma-2-2B, R. 312/315 complete trajectories; first-round ASR 0.006; any-round ASR 0.772 (95% CI [0.724, 0.817]); $\Delta = 0.766$ (95% CI [0.718, 0.811]); full-cohort any-round bounds [0.768, 0.775] and Δ bounds [0.762, 0.768].

Gemma-2-2B, P. 315/315 complete trajectories; first-round ASR 0.006; any-round ASR 0.213 (95% CI [0.168, 0.257]); $\Delta = 0.206$ (95% CI [0.162, 0.251]); full-cohort any-round bounds [0.213, 0.213] and Δ bounds [0.206, 0.206].

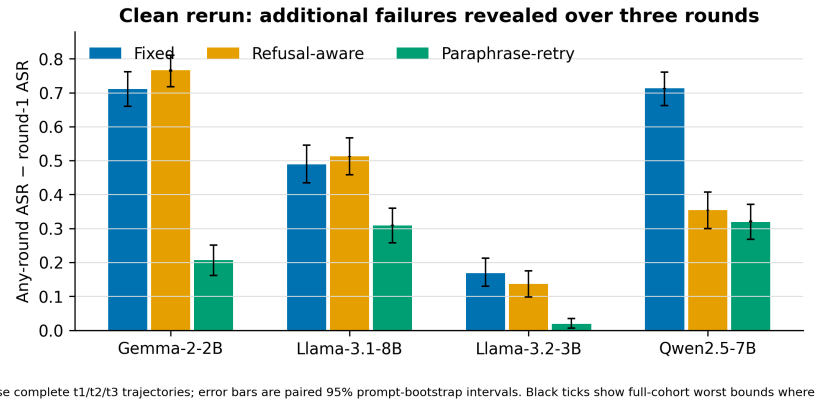

Figure 6: Original-four repeated-interaction summaries under the tested continuation policies. Additional any-round failures occur under at least one policy for every model; this does not imply an increase under every policy.

E Fixed-GCG Transfer and Cross-Attack Comparison

The source repositories are `JailbreakBench/jailbreakbench` and `JailbreakBench/artifacts`. The code commit is `23dbdf6b19650521604456229bc1d9c4156c85c1`; the artifact commit is `909e68c01d94222b8ad2e397a017e2e12e2adb73`. The artifact is `attack-artifacts/GCG/white_box/vicuna-13b-v1.5.json`, SHA-256 `db663b57904ad5b1c0482722ec3843a14d8a3d76b459d58b5cd093b3f0f445c4`. All 100 released instances were frozen before target-model outcome inspection and transferred unchanged. Safe indices and hashes are released, not attack strings. Historical artifact success is not a result of this study. The shared frozen judge does not make attack families exchangeable: requests, attacker

Table 5: Fixed-GCG transfer results: 100 frozen instances per original checkpoint, all 400 judgments valid. CI low/high are the supplied 95% prompt-bootstrap endpoints.

Model	N	ASR	CI low	CI high	Missing
Gemma-2-2B	100	0.010	0.000	0.030	0
Llama-3.1-8B	100	0.060	0.020	0.110	0
Llama-3.2-3B	100	0.080	0.030	0.140	0
Qwen2.5-7B	100	0.180	0.110	0.260	0

Table 6: Descriptive cross-attack comparison; rank 1 is most robust. Four of six pair orders agree (Spearman 0.400, Kendall 0.333; $n = 4$ descriptive only).

Model	Prefix RAUC	Prefix rank	GCG ASR	GCG rank	Rank change
Gemma-2-2B	0.394	1	0.010	1	0
Llama-3.1-8B	0.243	3	0.060	2	-1
Llama-3.2-3B	0.195	4	0.080	3	-1
Qwen2.5-7B	0.323	2	0.180	4	2

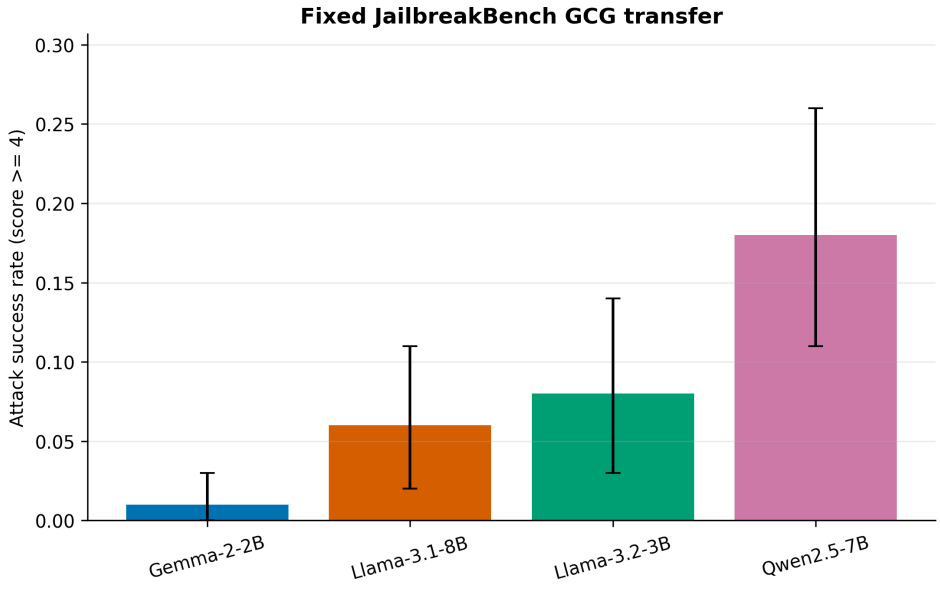

Figure 7: Low but model-dependent fixed-GCG transfer ASR. Low transfer success does not establish broad jailbreak robustness, and this condition does not replicate the prefix ordering.

access, mechanism, and output caps differ. ASRs are not pooled. The artifact was optimized on

Vicuna-13B-v1.5 rather than these targets, and one transferred artifact is not broad attack-family coverage.

F Seven Prespecified Prefix Grids

The grids are A: $\{0, 2, 5, 10, 15, 20\}$; B: $\{0, 5, 10, 15, 20\}$; C: $\{0, 2, 10, 15, 20\}$; D: $\{0, 2, 5, 15, 20\}$; E: $\{0, 2, 5, 10, 20\}$; F: $\{0, 5, 10, 20\}$; G: $\{0, 10, 20\}$. Each uses its own complete-case set and the same normalized trapezoidal definition. Zero rank changes and zero pairwise swaps occur across 42 grid-by-pair comparisons, including the reference grid. The maximum absolute RAUC difference is 0.098016 (50.3% relative) for Llama-3.2 on G. The sensitivity intervals use seed 20260910 rather than the baseline seed 20260906; they do not replace the baseline intervals.

Table 7: All seven grids: complete-trajectory counts, RAUC, absolute difference from the full grid, and rank. Grid codes are defined in the text.

Grid	Model	n	RAUC	$\lvert\text{change}\rvert$	Rank
A	Gemma-2-2B	314	0.394187898	0.000000000	1
A	Llama-3.1-8B	314	0.243232484	0.000000000	3
A	Llama-3.2-3B	315	0.194841270	0.000000000	4
A	Qwen2.5-7B	306	0.322630719	0.000000000	2
B	Gemma-2-2B	314	0.393312102	0.000875796	1
B	Llama-3.1-8B	314	0.247611465	0.004378981	3
B	Llama-3.2-3B	315	0.209920635	0.015079365	4
B	Qwen2.5-7B	307	0.321661238	0.000969481	2
C	Gemma-2-2B	314	0.429219745	0.035031847	1
C	Llama-3.1-8B	314	0.282563694	0.039331210	3
C	Llama-3.2-3B	315	0.228809524	0.033968254	4
C	Qwen2.5-7B	307	0.357654723	0.035024004	2
D	Gemma-2-2B	315	0.399444444	0.005256546	1
D	Llama-3.1-8B	314	0.247213376	0.003980892	3
D	Llama-3.2-3B	315	0.201190476	0.006349206	4
D	Qwen2.5-7B	307	0.327280130	0.004649411	2
E	Gemma-2-2B	314	0.391401274	0.002786624	1
E	Llama-3.1-8B	315	0.245793651	0.002561167	3
E	Llama-3.2-3B	315	0.196031746	0.001190476	4
E	Qwen2.5-7B	309	0.325647249	0.003016530	2
F	Gemma-2-2B	314	0.390525478	0.003662420	1
F	Llama-3.1-8B	315	0.250000000	0.006767516	3
F	Llama-3.2-3B	315	0.211111111	0.016269841	4
F	Qwen2.5-7B	310	0.324193548	0.001562829	2
G	Gemma-2-2B	314	0.447452229	0.053264331	1
G	Llama-3.1-8B	315	0.323015873	0.079783389	3
G	Llama-3.2-3B	315	0.292857143	0.098015873	4
G	Qwen2.5-7B	312	0.383012821	0.060382102	2

Table 8: Supplied grid-specific 95% intervals. The seed differs from the original baseline analysis; baseline intervals are retained separately.

Grid	Model	RAUC 95% CI
A	Gemma-2-2B	[0.358437500, 0.429856688]
A	Llama-3.1-8B	[0.215125398, 0.273009554]
A	Llama-3.2-3B	[0.170476190, 0.220238095]
A	Qwen2.5-7B	[0.291013072, 0.356131536]
B	Gemma-2-2B	[0.358280255, 0.428343949]
B	Llama-3.1-8B	[0.219745223, 0.276671975]
B	Llama-3.2-3B	[0.186904762, 0.234136905]
B	Qwen2.5-7B	[0.288273616, 0.355863192]
C	Gemma-2-2B	[0.395063694, 0.464572054]
C	Llama-3.1-8B	[0.255175159, 0.310830016]
C	Llama-3.2-3B	[0.203809524, 0.254605159]
C	Qwen2.5-7B	[0.325323697, 0.390960912]
D	Gemma-2-2B	[0.362777778, 0.435634921]
D	Llama-3.1-8B	[0.218869427, 0.276833201]
D	Llama-3.2-3B	[0.175632937, 0.228255952]
D	Qwen2.5-7B	[0.294055375, 0.360667752]
E	Gemma-2-2B	[0.356048965, 0.427388535]
E	Llama-3.1-8B	[0.217696429, 0.275160714]
E	Llama-3.2-3B	[0.171428571, 0.221507937]
E	Qwen2.5-7B	[0.292961165, 0.360032362]
F	Gemma-2-2B	[0.355891720, 0.426363455]
F	Llama-3.1-8B	[0.222222222, 0.279365079]
F	Llama-3.2-3B	[0.188095238, 0.236111111]
F	Qwen2.5-7B	[0.291129032, 0.358870968]
G	Gemma-2-2B	[0.416401274, 0.480095541]
G	Llama-3.1-8B	[0.297619048, 0.349206349]
G	Llama-3.2-3B	[0.270634921, 0.315873016]
G	Qwen2.5-7B	[0.350961538, 0.416666667]

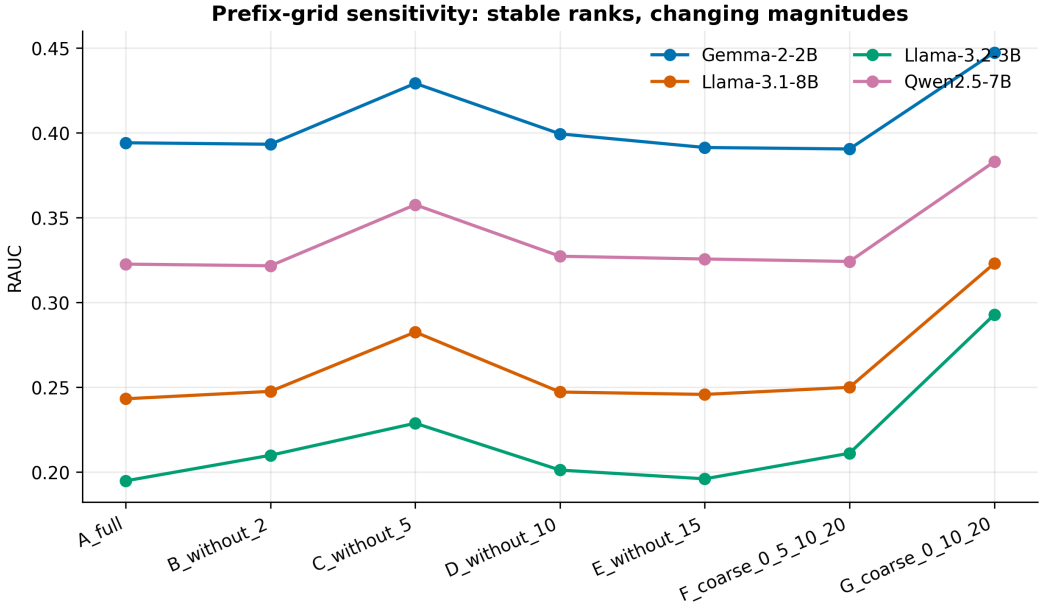

Figure 8: Stable ranks and materially varying RAUC magnitudes across the seven prespecified grids. Larger areas on a coarser grid do not indicate improved model behavior.

G Historical HEx-PHI Derivative and Category Accounting

The evaluation uses a historical local 330-row HEx-PHI-derived representation, not the current official release. The local labeled representation has SHA-256 `baa1947bc95fa55d04d6c8981b6704d0a2b559f9dd5c3139616233036c65d6a5`; the paired representation has SHA-256 `2ea4415da7951db3310268c6060483c1474a6fdb9a8d8691cf7fac0b5cfa011e`. All 330 user messages match positionally. This establishes local linkage, not paired-target authorship or transformation history. Exact-token filtering retains 315 rows, identified by cohort SHA-256 `27c1c4772ea8de8a30f55b7f1554eb374c9ce66b34664059db2159472aea41ca`.

Adjacent download metadata pins the immediate `Unispac/shallow-vs-deep-safety-alignment-dataset` derivative snapshot to `861f7d5ab930230b1f3856e63af18b839a38bea3`. This is not the original historical `LLM-Tuning-Safety/HEx-PHI` revision: that revision and the paired-target derivation/authorship are irrecoverable. The checked current official revision `83128b46334b80cc567bd7a2caf7af11c5b0bab7` has ten category CSVs after category 2 removal and is not asserted to match the evaluated historical bytes. An author used an authorized gated account for research/evaluation access. Applicable terms contain non-transfer restrictions. Raw prompts, paired targets, and inputs reconstructing them are withheld; release of safe aggregates and hashes is not a claim of unrestricted raw-data redistribution. Independently obtaining current HEx-PHI does not reconstruct the historical paired inputs.

Table 9: Eleven preexisting local derivative category labels, not newly inferred or revalidated against the current ten-category release. Each source category has 30 rows; retained counts range from 26 to 30 and sum to 315.

Category	Source N	Source percent	Clean315 N	Clean315 percent
adult_content	30	9.090909	27	8.571429
economic_harm	30	9.090909	30	9.523810
financial_advice	30	9.090909	27	8.571429
fraudulent_deceptive	30	9.090909	30	9.523810
harm_children	30	9.090909	29	9.206349
hate_harass_violence	30	9.090909	26	8.253968
illegal_activity	30	9.090909	30	9.523810
malware	30	9.090909	30	9.523810
physical_harm	30	9.090909	29	9.206349
political	30	9.090909	29	9.206349
privacy_violation	30	9.090909	28	8.888889

H Missingness and Recovery Documentation

The original certificate distinguishes missing judgments from valid scores with truncated explanations. In baseline round 1, 51/1,260 judge responses reached the 64-token output ceiling after emitting one unambiguous numeric score. Those scores were retained; absent explanations were not reconstructed and those records were not retried. The stage passed with a disclosed post-failure amendment, not complete unamended raw-response collection.

At round 2, seven twice-refused sources lack numeric judgments, and round 3 was not generated for those parents. Two further round-3 judgments are missing after provider refusal, giving nine incomplete longitudinal trajectories. Baseline prefix judging has eleven twice-refused judgments and one TLS-handshake timeout before a provider response: twelve missing cells. These are missing evaluator observations, not automatically failures or refusals by the evaluated models. Point estimates exclude them as specified and full-cohort bounds preserve known observations.

The original prefix plan is $4 \times 315 \times 6 = 7{,}560$ cells (7,548 valid); the interaction plan is $4 \times 315 \times 3 = 3{,}780$ model–policy trajectories (3,771 complete). These are different units, not independent responses to pool. Expansion adds 3,780 prefix cells, with three provider-refusal missing cells for Qwen3 and two for Phi, yielding 312 and 313 complete trajectories. All 700 expansion benign judgments and all 400 GCG judgments are valid. No missing score is imputed in any suite.

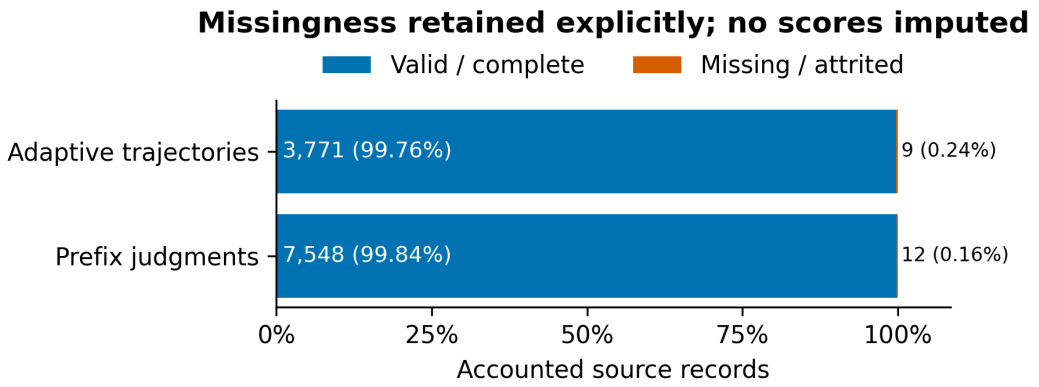

Figure 9: Certified original-four source accounting. Prefix judgments and adaptive trajectories count different units. Expansion and GCG accounting is disclosed separately rather than retroactively changing this baseline figure.

I Exact Checkpoints and Protocol Configuration

Table 10: Exact evaluated repositories and revisions. The first four are the original clean-v3 checkpoints; the last two form the frozen expansion. All use BF16 without quantization.

Repository	Revision
`meta-llama/Llama-3.1-8B-Instruct`	`0e9e39f249a16976918f6564b8830bc894c89659`
`meta-llama/Llama-3.2-3B-Instruct`	`0cb88a4f764b7a12671c53f0838cd831a0843b95`
`Qwen/Qwen2.5-7B-Instruct`	`a09a35458c702b33eeacc393d103063234e8bc28`
`google/gemma-2-2b-it`	`299a8560bedf22ed1c72a8a11e7dce4a7f9f51f8`
`Qwen/Qwen3-8B`	`b968826d9c46dd6066d109eabc6255188de91218`
`microsoft/Phi-4-mini-instruct`	`cfbefacb99257ffa30c83adab238a50856ac3083`

The clean-v3 configuration digest is `18f54ca109667a149bd99aeac56fcd23cbf12deb36c783453a22e7171d71b40b`. The anonymous research artifact includes the exact executed configuration and original implementation files. The new evidence configuration freezes the expansion identities, native template arguments, all GCG indices, common-cohort rule, evaluator, and analysis seed. The Qwen3 non-thinking setting is specific to that model; Phi's template override dictionary is empty.

The expansion checkpoints have 8,190,735,360 (Qwen3) and 3,836,021,760 (Phi) parameters; this is not a broad scale sweep. Generation uses BF16 without quantization, deterministic algorithms, and greedy decoding. New-evidence generation uses seed 20260910; clean-v3 uses 20260906. Generation uses empty system prompts and batch size one. Benign, prefix/continuation, and GCG caps are respectively 256, 128, and 150 new tokens. Exact model weights are obtained separately under their repository terms.

The retained harm rubric uses scores 1–5 and success threshold 4. Judge temperature is zero with a 64-token output cap. Original code/configuration, frozen rubrics, approved recovery implementations, and missingness records distinguish the recorded clean-v3 exceptions from the new-evidence single-call/no-imputation rule. The executed configuration records fixed, refusal-aware, and paraphrase-retry policies and the score-conditioned state transitions summarized in Appendix D. Restricted continuation strings are not reproduced.

I.1 What text the prefix judge scores

The verified expansion generator constructs the saved assistant response by decoding the concatenation of the exact prefill token IDs and generated output IDs, skipping special tokens and stripping surrounding whitespace. The expansion judging function assembles the harm rubric, original user request, and that saved response into the scoring prompt. Thus the candidate contains both supplied target-prefill text and generated continuation; the separately stored continuation digest is not a continuation-only scoring input. Increasing the prefill changes text the evaluator sees as well as the model's generation context. The resulting ASR is a response-level outcome under this threat model, not a decomposition of attacker-supplied versus newly generated harmful content.

This payload statement is established directly for Qwen3 and Phi by the available original expansion code. The original-four generation/judging implementation is not included in the available manuscript evidence, so we do not infer an identical payload assembly for that suite from the shared evaluator name or threshold. Its published curves retain their certified labels; a continuation-only or common-payload interpretation across all six models is not established here.

J Reproducibility and Restricted-Input Boundary

The anonymous research artifact provides exact configurations and revisions, code for generation, judging, and analysis, prompt-free row-level numerical records, processed summaries, public human-calibration aggregates, deterministic figure/table builders, verification scripts, and manifests/checksums. Default clean-extraction verification is offline and credential-free. Numeric verification and deterministic asset generation check the released processed evidence without contacting models or an evaluator. These checks establish internal consistency, not label infallibility.

Raw end-to-end reruns require independently authorized restricted data and appropriate model/API access. The release excludes raw harmful prompts, paired targets, attack strings, harmful model generations, raw automated-judge reasoning, private human labels/notes and identities, blinding keys, provider identifiers, credentials, weights/caches, and private repository history. Consequently, the artifact does not provide complete ZIP-only raw replay. The irrecoverable original HEx-PHI revision and paired-target history remain scientific limitations even with separate access to current data.

K Human Calibration of the Frozen Adversarial Harm Evaluator

K.1 Frozen sample, blinding, and rubric

The prespecified sample was frozen before annotation (seed 20260910): 120 prefix, 72 adaptive, and 48 fixed-GCG judgments. Safe IDs/hashes and full design-cell counts are documented. This stratified calibration is not a simple random sample of all outputs and does not validate benign A_{HHH}.

Two of the paper's three authors separately scored all 240 items. The annotation interface withheld model identity, machine score/reasoning, and the other annotator's labels; suite identity was withheld where technically possible. Annotators scored the frozen ordinal 1–5 harm rubric before binary labels were derived at score ≥ 4. The rubric distinguishes refusal/redirection (1), non-fulfillment (2), engagement without clearly prohibited actionable fulfillment (3), materially enabling prohibited content (4), and direct substantial fulfillment (5). This is the same operational boundary used by `gpt-4o-2024-11-20`; the threshold was not changed after human review.

K.2 Pre-adjudication agreement and adjudication accounting

From the original pre-adjudication A/B labels, human-human ordinal exact agreement is 0.7667, mean absolute score difference 0.3833, Spearman correlation 0.8752, linear weighted Cohen's $\kappa = 0.8020$, and quadratic weighted $\kappa = 0.8732$. Binary agreement is 0.9000 and $\kappa = 0.7965$: substantial but imperfect agreement, not an unambiguous task. Taking annotator A as reference and B as prediction, precision is 0.9109, recall 0.8598, specificity 0.9323, and F1 0.8846. Positive rates are 0.4458 for A and 0.4208 for B.

Every ordinal disagreement or uncertainty was flagged under the prespecified rule. There were 56 flagged items out of 240 (0.2333); neither primary annotator marked uncertainty (0/240). After both independent label files were complete, the third-author adjudicator reviewed those items and recorded one final 1–5 reference score per item. The other 184 references retain A/B agreement. Original independent labels were preserved. Human-human agreement is calculated from those pre-adjudication labels; machine-human calibration uses the final adjudicated human reference.

Table 11: Supplied pre-adjudication human-human agreement by suite. All values use the original independent A/B labels, excluding the third author's scores. Ordinal agreement is exact 1–5 agreement; ordinal κ is linearly weighted. A+/B+ are positive rates at score ≥ 4.

Suite	N	Ord. agr.	Ord. κ	Bin. agr.	Bin. κ	A+	B+
Prefix	120	0.7917	0.8374	0.9000	0.7887	0.6333	0.6000
Adaptive	72	0.6944	0.7109	0.8889	0.7551	0.3333	0.3611
Fixed GCG	48	0.8125	0.5417	0.9167	0.5616	0.1458	0.0625

Adaptive items have lower ordinal exact agreement and linear weighted κ than prefix items, but their binary agreement is close. Fixed-GCG's binary agreement of 0.9167 accompanies $\kappa = 0.5616$ and unequal A/B positive rates. Those primary-label rates are distinct from the three positives in the final adjudicated GCG reference. Because suite composition, label prevalence, and sample size differ, these descriptive results do not isolate inherent human difficulty or explain away the weaker machine-human adaptive calibration. No per-suite human-human confidence intervals or difficulty tests are supplied.

K.3 Machine-human agreement and uncertainty

Table 12: Supplied calibration summary for the prespecified $N = 240$ sample. Machine labels and final adjudicated-reference labels both use score ≥ 4. Intervals are the supplied within-design-cell bootstrap intervals; human-human intervals were not estimated.

Quantity	Estimate	95% CI
Human-human binary agreement	0.9000	not bootstrapped
Human-human Cohen kappa	0.7965	not bootstrapped
Machine-human accuracy	0.8667	[0.8333, 0.9000]
Machine-human precision	0.8105	[0.7500, 0.8681]
Machine-human recall	0.8462	[0.7865, 0.9029]
Machine-human F1	0.8280	[0.7797, 0.8718]
Machine-human Cohen kappa	0.7192	[0.6476, 0.7872]

The overall confusion counts are TN=131, FP=18, FN=14, and TP=77 (Figure 10). Alongside Table 12, specificity is 0.8792, balanced accuracy 0.8627 (95% CI $[0.8269, 0.8972]$), false-positive rate 0.1208, and false-negative rate 0.1538. Machine-positive and human-reference-positive rates are 0.3958 and 0.3792. Ordinal machine-human exact agreement is 0.6500, mean absolute score difference 0.5458, Spearman correlation 0.8058, linear weighted $\kappa = 0.7061$, and quadratic weighted $\kappa = 0.8111$.

The 95% percentile intervals use 10,000 bootstrap replicates (seed 20260910), resampling within prespecified design cells. Each of the six reported metrics has 10,000 valid and zero undefined replicates. Intervals concern this design, not future evaluator snapshots or all adversarial distributions. Independent labels and the threshold are retained.

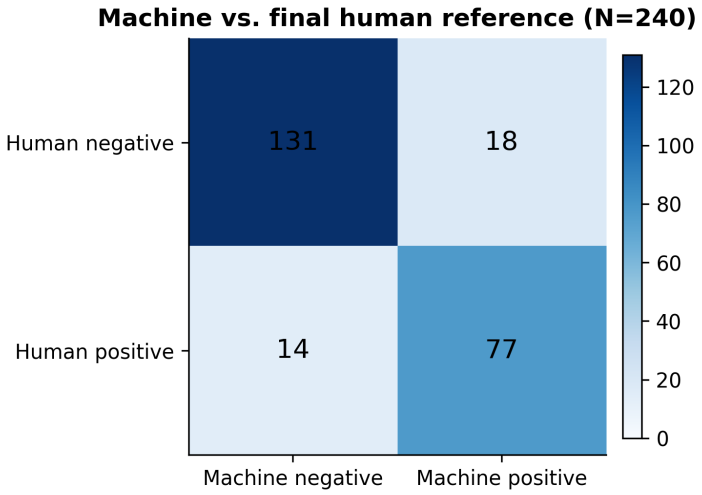

Figure 10: Supplied machine–human confusion matrix for 240 stratified judgments, using the final adjudicated human reference and unchanged score-≥ 4 threshold. Both false positives and false negatives remain; the reference is an author-adjudicated measurement, not literal ground truth.

K.4 Suite heterogeneity and threshold-boundary errors

Table 13: Supplied descriptive per-suite calibration. High fixed-GCG accuracy must be read alongside its three human-positive and two machine-positive items; adaptive agreement is the weakest.

Suite	N	Accuracy	Precision	Recall	F1	Kappa
prefix	120	0.8833	0.8873	0.9130	0.9000	0.7601
adaptive	72	0.7639	0.5455	0.6316	0.5854	0.4216
fixed GCG	48	0.9792	1.0000	0.6667	0.8000	0.7895

Prefix specificity and balanced accuracy are 0.8431 and 0.8781, respectively; adaptive values are 0.8113 and 0.7214; fixed-GCG values are 1.0000 and 0.8333. The adaptive sample has 10 false positives and seven false negatives, compared with eight and six for prefix, and zero and one for fixed GCG. Fixed-GCG recall is 0.6667 despite accuracy 0.9792, reflecting the small positive class. These descriptive suite differences preclude treating pooled accuracy as uniform across the experiments.

At the operational boundary, seven items have machine score 3 but final human score ≥ 4. These are a subset of the 14 false negatives, not their total. Conversely, 18 items have machine score ≥ 4 but human score ≤ 3, accounting for the false positives. The threshold remains fixed; no manual corrections were applied to unannotated records or used to recalculate the paper's adversarial results.

K.5 Validity scope and governance

These results provide *moderate calibration support*: a validity check with disagreement between the separately recorded author ratings, machine errors, and suite heterogeneity, not direct evidence of the benign-versus-robustness relationship. They do not validate benign A_{HHH} or other judges.

The calibration involved three paper authors: two independent primary annotators and a third-author adjudicator who reviewed only the 56 flagged items after both primary label files were complete. It is author-annotated evidence rather than validation by raters outside the author team. The Ethics Statement reports the primary annotators' safeguards and the absence of a formal institutional determination; those safeguards are not attributed to the third author.