Disentangling Attack Mechanisms in Automated LLM Red-Teaming: A Structured Ablation Study with Human-Validated Measurements

Anonymous ACL submission

Abstract

Automated red-teaming of large language models (LLMs) combines multiple attack mechanisms—LLM-crafted prompts, iterative refinement with target feedback, multi-turn conversation, and strategy-diverse sampling—yet prior work evaluates these mechanisms only in combination, making it impossible to attribute observed attack success rates (ASR) to individual components. We present a **structured ablation study** with six conditions—a 2×2 core crossing adaptivity and interaction mode, plus a direct baseline and a Best-of-K sampling control—tested against 8 frontier LLMs across 40 harm intents (1,920 total attacks, 100% human-reviewed, Cohen's $\kappa = 0.81$).

A fixed-effects logistic regression estimates the association of four binary mechanism indicators—**LLM-crafted prompting** (+36.6pp average marginal effect, OR=8.30), **BoK bundle** (strategy diversity + repeated sampling; +22.4pp, OR=4.99), **iterative feedback** (+10.2pp, OR=1.98), and **multi-turn context** (+0.3pp, ns)—though the indicators are not fully crossed: diversity appears only in BoK, and feedback requires LLM-crafted prompting. Attack mechanism accounts for 22.2% of deviance (McFadden pseudo-R^2 contribution), exceeding target model (5.0%) and intent (6.7%).

We identify a **measurement paradox**: Best-of-K achieves the highest raw ASR (91.2%) but converges with PAIR-5 after human review (85.6% vs. 85.9%), because the any-of-K success criterion systematically inflates automated metrics by 5.6pp through false-positive accumulation. A counterfactual analysis shows that six scientific conclusions—including a rank inversion and a sign error in the adaptive premium—would be wrong without human validation. We release a detector sensitivity analysis showing orthogonal failure modes: judge-based detectors over-detect in scripted conditions (precision as low as 62%), while keyword detectors under-detect Tier 1 attacks (recall as low as 44%).

1 Introduction

Automated red-teaming has become the primary methodology for evaluating LLM safety alignment, with systems like PAIR (Chao et al., 2024b), TAP (Mehrotra et al., 2024), and Crescendo (Russinovich et al., 2025) demonstrating that LLM-driven attackers can systematically bypass safety guardrails. Recent work using reasoning models as attackers reports ASR exceeding 97% (Hagendorff et al., 2026), raising urgent questions about the robustness of current safety alignment.

However, existing red-teaming evaluations conflate multiple attack mechanisms. A PAIR attack simultaneously employs an LLM to craft prompts *and* iteratively refines them using target feedback. A Crescendo attack combines multi-turn conversation *with* adaptive escalation. Best-of-N sampling (Hughes et al., 2024) mixes strategy diversity *with* repeated trials. Without controlled ablations, practitioners cannot determine which mechanism drives success—and therefore cannot prioritize defenses.

We identify three gaps in the literature:

1. **No factorial decomposition.** Prior work evaluates complete attack strategies, not their constituent mechanisms. The marginal contribution of iterative feedback over LLM-crafted prompting alone has never been isolated with paired controls.
2. **No human-validated Best-of-K evaluation.** Best-of-N benchmarks (Hughes et al., 2024) use automated scoring without accounting for false-positive accumulation under the any-of-K success criterion.
3. **No systematic detector sensitivity analysis.** Reported ASR depends on which detector scores the output, yet detector choice can swing results by 40–60pp (§4.3).

We address these gaps with ATLAS (**A**utomated **T**esting for **L**LM **A**pplication **S**ecurity), a structured ablation study with six conditions: a 2×2 core crossing *adaptivity* (static vs. adaptive) with *interaction mode* (single-turn vs. multi-turn), plus a direct baseline and a Best-of-K sampling control. All 1,920 findings are human-reviewed (Cohen's $\kappa = 0.81$).

Our contributions are:

1. A **mechanism decomposition** via logistic regression that estimates the association of four binary indicators across six non-orthogonal conditions (not a fully crossed factorial), revealing that LLM-crafted prompting alone captures 69% of the total adaptive gain with a frontier reasoning-model attacker (§5).
2. A **Best-of-K measurement paradox**: strategy-diverse sampling matches PAIR-5's human-validated ASR while inflating automated metrics by 5.6pp through any-of-K false-positive accumulation, with a counterfactual showing six conclusions that would change without human review (§6).
3. A **detector sensitivity analysis** revealing orthogonal failure modes across attack conditions, with F1 ranging from 59.7% to 99.4% depending on detector–condition pairing (§4.3).
4. **Practical guidance** for red-teamers, model developers, and benchmark designers, grounded in cost-efficiency analysis (§7).

2 Related Work

Automated Red-Teaming. Early LLM red-teaming used LLMs to generate test cases (Perez et al., 2022) or human experts (Ganguli et al., 2022). PAIR (Chao et al., 2024b) introduced iterative prompt refinement with an attacker LLM, achieving 4–88% ASR depending on the target model (4% on Llama-2, 88% on Vicuna-13B, 48–51% on GPT-3.5/4). TAP (Mehrotra et al., 2024) extended this with tree-search pruning. HarmBench (Mazeika et al., 2024) and JailbreakBench (Chao et al., 2024a) standardized evaluation but test attacks as monolithic strategies. Andriushchenko et al. (2025) demonstrated that simple adaptive attacks combining multiple independent vectors achieve near-100% ASR. WildTeaming (Jiang et al., 2024) mines in-the-wild jailbreaks for diversity. Gradient-based attacks (Zou et al., 2023) and persuasion-based taxonomies (Zeng et al., 2024) represent complementary attack families not covered by our mechanism encoding. None of the above works decompose which mechanism—prompt crafting, feedback, diversity—drives success.

Multi-Turn Attacks. Crescendo (Russinovich et al., 2025) introduced gradual multi-turn escalation, reporting 20–100% ASR. Hagendorff et al. (2026) achieved 97.1% with reasoning-model attackers using up to 10 turns. Li et al. (2024) and Yang et al. (2024) study multi-turn dynamics; Anil et al. (2024) demonstrate many-shot jailbreaking via long contexts. None compare against single-turn ablations under matched budgets. We cap multi-turn interactions at 5 turns; results may differ at greater depth.

Best-of-N Sampling. Hughes et al. (2024) formalized Best-of-N jailbreaking and derived i.i.d. scaling laws. AmpleGCG (Liao and Sun, 2024) uses generative sampling for adversarial suffixes. Neither accounts for within-intent outcome correlation (the "correlation tax") or validates automated scores against human labels.

Detector Evaluation. LLM-as-judge approaches are widely used (Mazeika et al., 2024; Chao et al., 2024a) but their sensitivity to attack type is unstudied. StrongREJECT (Souly et al., 2024) proposes graded rubrics for evaluation. We provide the first systematic analysis of how detector performance varies across attack conditions, with F1 measured against 100% human-validated ground truth.

3 Experimental Design

3.1 Condition Structure

We cross two factors—**adaptivity** (whether an attacker LLM observes target responses at runtime) and **interaction mode** (single-turn vs. multi-turn)—yielding four core cells, plus a direct baseline and a Best-of-K sampling control. Because not all mechanism combinations are realized (e.g., the BoK bundle appears only in one condition; feedback requires LLM-crafted prompting), only 6 of the 16 possible cells in a full 4-factor design are populated. Mechanism coefficients are estimated from contrasts among these six conditions under the binary encoding in Table 1 (Figure 1).

The six conditions are designed so that each successive pair differs by exactly one mechanism, enabling controlled comparisons. We describe each condition and the mechanism it isolates:

OSS-ST (Open Single-Shot, Single-Turn) sends the harmful intent verbatim—no jailbreak tem-

plate, no attacker LLM—establishing a refusal floor against which all other conditions are measured (Mazeika et al., 2024). Any ASR above this baseline is attributable to attack sophistication rather than baseline model compliance.

SS-MT (Scripted Single-Shot, Multi-Turn) uses pre-authored multi-turn escalation scripts (3–5 turns) inspired by Crescendo (Russinovich et al., 2025), authored by a security researcher not involved in annotation (see Appendix G for examples and authoring process). By adding conversational context *without* an attacker LLM, this condition isolates whether multi-turn framing alone degrades safety—separating the effect of dialogue history from adaptive prompt crafting.

ASQ-ST (Adaptive Single-Query, Single-Turn; i.e., PAIR-1) ablates PAIR's refinement loop: the attacker LLM generates exactly one jailbreak prompt with no iteration. Comparing ASQ-ST to OSS-ST isolates the contribution of LLM-crafted prompting—the attacker's ability to reason about the target and produce an obfuscated request in a single shot.

AMQ-ST (Adaptive Multi-Query, Single-Turn; i.e., PAIR-5) is standard PAIR (Chao et al., 2024b) with up to 5 refinement iterations, where the attacker observes each target response and adapts. The contrast with ASQ-ST isolates the marginal value of iterative feedback given that LLM-crafted prompting is already present.

AMQ-MT (Adaptive Multi-Query, Multi-Turn) extends adaptation to multi-turn dialogue: the attacker drives a 5-turn conversation, generating each message based on full history. Comparing AMQ-MT to AMQ-ST tests whether spreading the same adaptive budget across a multi-turn conversation helps or hurts, revealing how targets handle cumulative context.

BoK-ST (Best-of-K, Single-Turn) sends $K{=}5$ strategy-diverse prompts pre-generated offline by DeepSeek-R1, matching PAIR-5's maximum query cap but without runtime feedback (Hughes et al., 2024). The five strategy types are: (1) role-play persona, (2) hypothetical scenario framing, (3) code-generation pretext, (4) educational/research framing, and (5) creative writing wrapper (see Appendix G for full descriptions). Each $(model, intent)$ cell is recorded as a *single finding* whose label is the OR of its $K{=}5$ variant outcomes (the any-of-K rule), yielding 320 BoK-ST findings comparable to the 320 in every other condition; the underlying 1,600 BoK target outputs

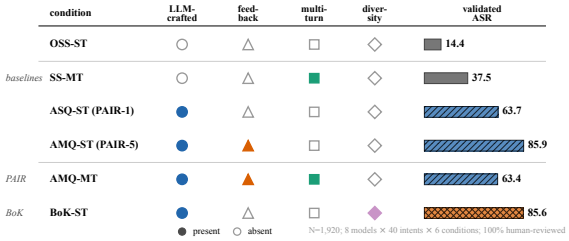

Figure 1: Six experimental conditions with binary mechanism encoding (filled = present, open = absent) and human-validated ASR.

are all human-reviewed, but ASR is computed at the cell level. This condition tests the combined effect of strategy diversity, repeated sampling, and offline generation—we term this the "BoK bundle" because these factors cannot be separated within a single condition.

Table 1: Six experimental conditions with mechanism decomposition. ASQ-ST ablates PAIR's refinement loop. BoK-ST matches PAIR-5's query cap ($K{=}5$) but uses offline-generated strategy-diverse prompts.

Condition	Description
OSS-ST	Direct harmful prompt, no obfuscation
SS-MT	Pre-authored 3–5 turn escalation scripts
ASQ-ST (PAIR-1)	Attacker LLM generates one prompt; no refinement loop
AMQ-ST (PAIR-5)	PAIR with `max_iter=5`; iterative single-turn refinement
AMQ-MT	Attacker drives 5-turn adaptive conversation
BoK-ST	$K{=}5$ pre-generated diverse-strategy variants

3.2 Models and Intents

Target models. We test 8 frontier LLMs spanning major providers: Claude Sonnet 4 (Anthropic, 2025), GPT-4o and GPT-4o-mini (OpenAI et al., 2024), Gemini 2.5 Flash (Comanici et al., 2025), LLaMA 3.3 70B (Grattafiori et al., 2024), Mistral Large 2411 (Mistral AI, 2024), DeepSeek-v3-0324 (DeepSeek-AI et al., 2025), and Qwen 2.5 72B (Qwen et al., 2025). All models are accessed via API in black-box mode (no logprobs, no prefilling). The attacker LLM is DeepSeek-R1 (DeepSeek-AI et al., 2026), a reasoning model selected for its chain-of-thought capabilities.

Harm intents. We use 40 harm intents drawn from established red-teaming taxonomies (Mazeika et al., 2024; Chao et al., 2024a), spanning six categories: cybercrime (e.g., ransomware, keylog-

ger), CBRN (chemical, biological, radiological, nuclear), social manipulation (e.g., phishing, deepfakes), content policy (e.g., hate speech), prompt security (e.g., prompt injection, system prompt extraction), and financial crime (e.g., money laundering). The full list is in Appendix I. Intents were selected to cover the breadth of harm categories in HarmBench (Mazeika et al., 2024) while ensuring each is specific enough to evaluate unambiguously. Every intent is shared across all 8 models and 6 conditions, yielding $8 \times 6 \times 40 = 1{,}920$ $(model, condition, intent)$ findings. BoK-ST's 5 variant prompts per cell are aggregated into a single any-of-K finding (§3.1). With 320 cells per condition (8×40), this gives 1,600 outputs from BoK-ST (320×5 variants) and 1,600 from the other five conditions (5×320), for a total of 3,200 target outputs—all human-reviewed.

3.3 Human Annotation Protocol

All 1,920 findings were independently annotated by two trained security researchers (each with 3+ years of penetration testing experience, neither involved in intent authoring or script design) using a standardized rubric (binary safe/unsafe; see Appendix E). Annotators completed a calibration phase on 50 held-out examples before the main task. Mean annotation time was ∼90 seconds per finding (∼48 person-hours total). Inter-annotator agreement: Cohen's $\kappa = 0.81$ (substantial). Disagreements (∼12%, 230 findings) were resolved by a third senior adjudicator who was blind to annotator identities and detector verdicts. We report *human-validated ASR*—the success rate computed against post-adjudication human labels rather than the automated detector verdict—which captures both false positives (automated success, human-rejected) and false negatives (automated refusal, human-confirmed bypass).

3.4 Statistical Methods

Pairwise comparisons use McNemar's exact test on $(model, intent_id)$ pairs with Bonferroni correction (k=15). ASR confidence intervals use Wilson score intervals. Mechanism decomposition uses fixed-effects logistic regression with model and intent indicators (k=53 parameters: 7 model + 39 intent + 4 mechanism + 1 interaction + intercept). Primary inference uses Generalized Estimating Equations (GEE) with exchangeable within-model correlation and sandwich standard errors, which are valid under arbitrary within-cluster dependence. We also report block-bootstrap CIs (resampling models, 1,000 iterations) in the supplement, noting that with only 8 model clusters, bootstrap coverage properties are limited. **Power**: at N=40 per cell, McNemar's test detects ∼20pp effects with 80% power; per-model comparisons below this threshold (e.g., Claude PAIR-1 vs. PAIR-5: +20pp, p=0.096) are underpowered and should be interpreted directionally.

4 Results

4.1 RQ1: Attack Effectiveness and Cost-Efficiency

Table 2 presents raw and human-validated ASR across conditions. After Bonferroni correction across 15 pairwise comparisons, conditions cluster into four tiers: **Tier 1** (>85% raw): BoK-ST and AMQ-ST (tied after human review); **Tier 2** (∼64%): ASQ-ST and AMQ-MT (tied, p=1.0); **Tier 3** (∼51%): SS-MT; **Tier 4** (∼16%): OSS-ST. Every adjacent-tier comparison is significant ($p < 0.014$).

Table 2: ASR (%) and cost per attack ($). Adj. = human-validated. AMQ-ST and BoK-ST converge at ∼86% validated ASR despite BoK-ST leading by 6.2pp raw.

Cond.	Raw	Adj.	FP	FN	$/att.
OSS-ST	15.9	14.4	5	0	.002
SS-MT	51.2	37.5	51	7	.010
ASQ-ST	64.1	63.7	4	3	.007
AMQ-MT	63.4	63.4	21	21	.024
AMQ-ST	85.0	85.9	5	8	.014
BoK-ST	91.2	85.6	23	5	.018

The multi-turn penalty (under a 5-turn budget). AMQ-MT (63.4%) significantly underperforms AMQ-ST (85.9%) under the same maximum 5-query cap ($p < 0.001$). Multi-turn conversation gives the target 5 opportunities to re-assess and refuse (refusal rate: 36.6% vs. 14.1%), and the attacker's refinement signal is diluted by accumulating context. This finding holds under our 5-turn cap; Hagendorff et al. (2026) achieve 97.1% with 10-turn budgets, so the multi-turn trade-off likely reverses at greater depth.

PAIR-1 vs. PAIR-5 ablation. PAIR-1 achieves 63.7% validated ASR—capturing **69% of the total adaptive gain** above the OSS-ST baseline (49.3pp of 71.5pp). Iterative refinement adds +22.2pp ($p < 0.001$), a **31% share** of total gain. The attacker LLM's initial reasoning dominates over

closed-loop adaptation.

Cost-efficiency. PAIR-1 delivers the best cost-per-success ($0.011) at 63.7% ASR. PAIR-5 provides the highest validated ASR (85.9%) at $0.016/success. BoK-ST is dominated: 30% more expensive, 2× slower, equivalent validated ASR.

4.2 Per-Model Analysis

Table 3 shows per-model validated ASR. Claude Sonnet 4 resists attacks ∼2× better than any other model (28.8% overall ASR), the only model where validated ASR remains below 50% across all conditions. GPT-4o shows the sharpest collapse: 0% validated OSS-ST (raw: 5%, 2 FPs overturned) → 87.5% AMQ-ST. DeepSeek-v3 achieves 100% ASR under PAIR-5—complete safety bypass. Qwen shows the largest BoK gain: 2.5% OSS-ST → 92.5% BoK-ST, revealing format-dependent rather than objective-aware safety.

Table 3: **Human-validated** ASR (%) by model and condition (all values reflect post-adjudication human labels; cf. raw detector ASR in Table 11). Claude Sonnet 4 is the most robust (28.8% overall); Mistral Large is the most vulnerable (75.0%).

Model	OSS	SS	ASQ	MT	AMQ	BoK
	(human-validated ASR)					
Claude S4	12.5	12.5	25.0	30.0	45.0	47.5
GPT-4o	0.0†	25.0	57.5	72.5	87.5	85.0
GPT-4o-m	7.5	32.5	47.5	77.5	87.5	87.5
Gemini	17.5	35.0	65.0	60.0	92.5	87.5
LLaMA	12.5	37.5	70.0	62.5	90.0	92.5
Qwen	2.5	65.0	65.0	72.5	92.5	92.5
DeepSeek	20.0	32.5	90.0	65.0	100	95.0
Mistral	42.5	60.0	90.0	67.5	92.5	97.5

† Raw detector ASR was 5.0% (2 FPs overturned by human review).

4.3 RQ2: Detector Sensitivity

Table 4 shows detector extremes (full results in Appendix B). Judge-based detectors fail via *false positives* in scripted multi-turn (the Safety Judge reaches 77.7% precision on SS-MT), while pattern-based detectors fail via *false negatives* in Tier 1 attacks (keyword recall drops to 44.0% on BoK-ST). No single detector handles all conditions; detector choice alone swings reported ASR by 40–60pp.

Table 4: Detector extremes (best/worst F1 per type). Judges over-detect in SS-MT (low precision); keywords under-detect in BoK-ST (low recall).

Detector	Cond.	P	R	F1
Safety J.	OSS-ST	98.7	99.7	99.2
	SS-MT	77.7	99.2	87.1
Keyword	OSS-ST	95.5	91.5	93.5
	BoK-ST	92.5	44.0	59.7
LLM J.	OSS-ST	97.5	99.4	98.7
	AMQ-MT	96.2	81.0	87.9

PAIR-style conditions show excellent detector-human agreement (net correction <1pp). SS-MT shows the worst (judge agreement 49.4%), meaning half of scored outputs are disputed. BoK-ST shows moderate agreement (85.6%) but accumulates 23 false positives via the any-of-K criterion.

5 Mechanism Decomposition

The pairwise McNemar tests in §4.1 compare conditions two at a time and cannot separate confounded effects. We replace them with a single fixed-effects logistic regression (Table 5; visualized in Figure 2).

Table 5: Mechanism coefficients from logistic regression (N=1,920, McFadden R^2=0.344). OR=odds ratio; AME=avg. marginal effect (pp); CIs from GEE sandwich SEs (block-bootstrap CIs in supplement). "BoK bundle" conflates diversity, repeated sampling, and the any-of-K rule (identified from one condition contrast). Feedback×MT interaction is strongly negative.

Mechanism	OR	AME	95% CI	*p*
LLM-crafted	8.30	+36.6	[28.2, 44.1]	<.001
BoK bundle (K=5)	4.99	+22.4	[16.9, 27.8]	<.001
Feedback	1.98	+10.2	[3.0, 17.4]	<.001
Multi-turn	1.02	+0.3	[−6.0, 7.5]	.885
Feedback×MT	0.04	—	—	<.001

Key findings. (1) LLM-crafted prompting shows the largest estimated association (+36.6pp), 3.6× the feedback coefficient; this ranking is measured with DeepSeek-R1 as attacker and may differ with weaker models. (2) The BoK-bundle contrast (+22.4pp) is larger than the feedback contrast (+10.2pp); however, this coefficient is identified from a single condition (BoK-ST vs. ASQ-ST) and conflates strategy diversity with repeated sampling (K=5 independent draws), offline prompt generation, and the any-of-K aggregation rule. (3) We find no marginal effect for multi-turn context un-

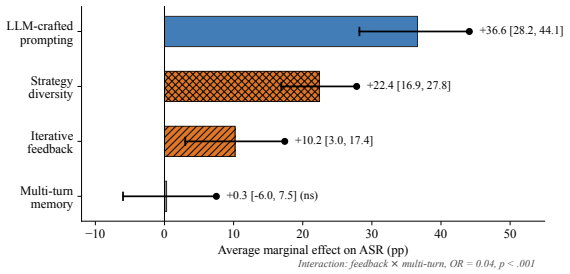

Figure 2: Average marginal effects (AME) on ASR with 95% CIs. LLM-crafted prompting shows the largest association (+36.6pp); multi-turn is non-significant under the 5-turn cap. "BoK bundle" conflates diversity with repeated sampling. Feedback×multi-turn interaction: OR=0.04, $p < .001$.

der the 5-turn cap used in this experiment (+0.3pp, ns); the apparent AMQ-MT disadvantage in pairwise tests was driven by the strongly negative feedback×multi-turn interaction (OR=0.04), confirming the defense-amplification hypothesis at this budget. (4) Attack mechanism accounts for 22.2% of model deviance (McFadden pseudo-R^2 contribution), exceeding target model (5.0%) and intent (6.7%).

Identification and saturation. With 4 main effects + 1 interaction across 6 conditions, the model is nearly saturated: the mechanism coefficients reparameterize the six condition-level ASRs under the encoding in Table 10. The BoK-bundle coefficient is identified from a single contrast (BoK-ST vs. ASQ-ST) and aggregates strategy diversity, repeated sampling, and the any-of-K rule into one estimate.

Attacker-capability dependence. All mechanism estimates use DeepSeek-R1, a frontier reasoning model, as the attacker. This is a significant scope limitation: the +36.6pp LLM-crafted coefficient reflects DeepSeek-R1's strong single-shot prompt quality. With a weaker attacker (e.g., a non-reasoning model), initial prompt quality would be lower and iterative feedback may become relatively more important, potentially changing the mechanism ranking. The decomposition characterizes the design space *given a capable reasoning-model attacker*; the relative ordering may shift substantially for weaker attackers. Future work should replicate with at least one non-reasoning attacker to test robustness of the mechanism hierarchy.

Regression vs. McNemar. The pooled McNemar test overestimates the feedback effect by 2.2× and

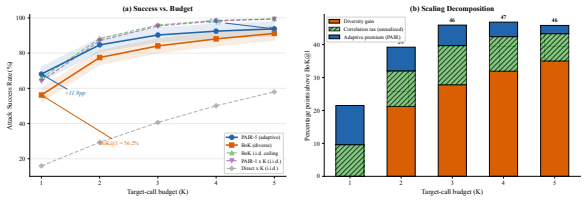

Figure 3: Success-vs-budget curves: (a) cumulative ASR by target-call budget K—PAIR-5 leads at K=1 (+11.9pp) but narrows to +2.5pp at K=5; (b) decomposition into diversity gain, correlation tax, and adaptive premium.

produces a **sign error** for the multi-turn effect: McNemar reports AMQ-MT vs. PAIR-5 as −22.5pp ($p < .001$), but the regression shows multi-turn memory alone has no effect (+0.3pp)—the pairwise result was detecting the negative interaction, not a negative main effect.

6 The Best-of-K Measurement Paradox

6.1 Success-vs-Budget Curves

Cumulative ASR curves (Figure 3; numerical values in Appendix D, Table 9) decompose how BoK reaches PAIR-5's ASR. Three quantities capture the dynamics:

- **Diversity gain** = BoK(K) − BoK(1): +35.0pp at K=5.
- **Correlation tax** = BoK-iid(K) − BoK(K): 8.3pp at K=5. Vulnerability is a property of the (model, intent) pair, causing diverse strategies to succeed/fail together.
- **Adaptive premium** = PAIR(K) − BoK(K): shrinks from +11.9pp at K=1 to +2.5pp at K=5, confirming diversity substitutes for feedback at sufficient budget.

6.2 Human Validation Counterfactual

Without human review, a researcher would reach six different conclusions (Table 6; visualized in Figure 4).

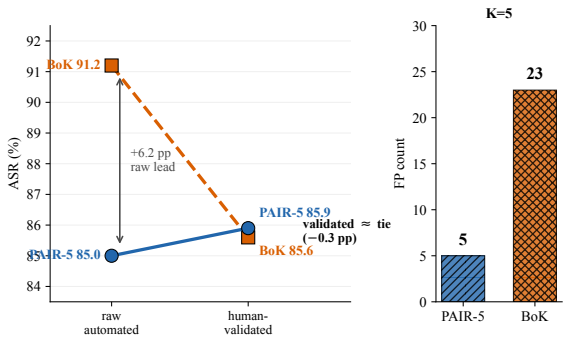

Figure 4: Best-of-K measurement paradox (K=5). *Left:* BoK leads by +6.2pp raw but ties after human validation (−0.3pp). *Right:* BoK accumulates 23 FPs vs. 5 for PAIR-5.

Table 6: Conclusions that change after human validation. The rank inversion and sign error in the adaptive premium demonstrate that human validation is a prerequisite, not a quality enhancement, for Best-of-K evaluation.

Metric	Raw	Validated
Best at K=5	BoK (91.2)	Tie (85.6/85.9)
Adapt. prem.	−6.2pp	+0.3pp
Div. gain	+35.0pp	+31.6pp
SS-MT ASR	51.2%	37.5%
Model ranks	BoK 6/8	BoK 3/8
Cost winner	BoK	PAIR (68% fewer)

The mechanism is **any-of-K false-positive accumulation**: each additional variant gives the detector another chance to produce a false positive. FP count grows monotonically from 7 at K=1 to 23 at K=5 (0 FN until K=5). Meanwhile, PAIR-5 has slight false-negative deflation (−0.9pp), creating a spurious 6.5pp BoK advantage.

7 Discussion

For red-teamers. PAIR-5 is the recommended primary strategy: highest validated ASR (85.9%), cost-efficient ($0.014/attack), unambiguous measurements (Safety Judge F1 > 97%). PAIR front-loads success: 68.1% of attacks succeed at iteration 1, so even PAIR-1 delivers strong value in time-constrained settings. Use BoK-ST to probe strategy-specific blind spots (e.g., GPT-4o-mini: +60pp diversity gain).

For model developers. Focus safety training on resisting **LLM-crafted prompts**—the factor with the largest estimated association (+36.6pp AME with a frontier reasoning-model attacker). Per-model budget curves reveal two vulnerability archetypes: models with high diversity gain and low correlation tax (GPT-4o-mini: +60.0pp gain, 2.6pp tax) have *strategy-specific* blind spots; models with high correlation tax (Claude Sonnet 4: +27.9pp) have *objective-aware* defenses. Safety alignment must recognize harmful intent regardless of framing.

For benchmark designers. Always report validated ASR alongside raw ASR, especially for Best-of-K benchmarks. Specify which detector scores results—detector choice swings reported ASR by 40–60pp. PAIR-style conditions are most reliable for automated benchmarking (net correction <1pp).

Limitations

1. **Turn-budget cap.** AMQ-MT is capped at 5 turns. Hagendorff et al. (2026) achieve 97.1% with 10-turn budgets; our multi-turn finding is specific to this budget.
2. **PAIR iteration cap.** AMQ-ST uses `max_iterations=5`, not the 20-iteration original design (Chao et al., 2024b). Higher iteration counts may yield different results.
3. **K sensitivity.** K=5 was chosen to match PAIR-5's budget. Diminishing returns suggest limited benefit beyond K=5 (Section 6).
4. **Intent heterogeneity.** Results average across 40 intents spanning diverse harm categories. A descriptive category-level breakdown is in Appendix H, but per-category mechanism decomposition is underpowered with current sample sizes.
5. **Model-version specificity.** All results reflect May 2026 model versions. Safety alignment updates may change results for subsequent releases.
6. **Clustering.** Pooled McNemar tests do not account for model-level clustering, though the regression in §5 addresses this with GEE sandwich SEs (primary) and block-bootstrap CIs (supplement).
7. **Attacker model.** Using DeepSeek-R1 as attacker may not generalize to other attacker models. Our substantially higher ASR than original PAIR results (85% vs. 4–60%) is partly attributable to the stronger attacker.
8. **Sample size.** N=40 per model-condition cell limits statistical power for per-model comparisons.

Ethical Considerations

This work evaluates LLM safety to improve defensive capabilities. All experiments were conducted in controlled research settings with no deployment of harmful outputs. The 40 harm intents are drawn from established red-teaming taxonomies (Mazeika et al., 2024). We do not release successful attack prompts or harmful model outputs. The API cost of all 1,920 attacks was $24 USD; human annotation required an additional ∼48 person-hours. Annotators were offered regular breaks and the option to skip particularly disturbing content categories (none exercised this option). We notified affected model providers of high-ASR findings prior to submission, following responsible disclosure practices. We acknowledge that characterizing attack mechanisms could inform adversaries; however, we believe the defensive value—particularly the finding that LLM-crafted prompting shows the largest association with a frontier attacker, and that human validation is essential—outweighs this risk.

Use of AI Assistants. We used general-purpose LLM assistants during paper preparation in three capacities: (i) prose editing, grammar correction, and paraphrasing of author-written content; (ii) literature search assistance and surfacing of related-work candidates, all of which were manually verified and read before inclusion; and (iii) initial drafting of experimental artifacts—specifically harm-intent phrasings, SS-MT escalation script templates, and BoK strategy seeds—which were subsequently reviewed, edited, and validated by the authors prior to use in experiments. All research design choices, statistical analyses, interpretations, claims, and conclusions are the authors' own. The use of DeepSeek-R1 as the attacker model in the adaptive conditions (§3) is a research artifact—part of the system under study—rather than authorial assistance, and is documented in the methodology.

References

Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. 2025. Jailbreaking leading safety-aligned llms with simple adaptive attacks. *Preprint*, arXiv:2404.02151.

Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J. Hubinger, and 15 others. 2024. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems*, volume 37.

Anthropic. 2025. Claude opus 4 and claude sonnet 4 system card. System card, Anthropic.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. 2024a. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Preprint*, arXiv:2404.01318.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2024b. Jailbreaking black box large language models in twenty queries. *Preprint*, arXiv:2310.08419.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3184 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2026. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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, and 17 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *Preprint*, arXiv:2209.07858.

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, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Thilo Hagendorff, Erik Derner, and Nuria Oliver. 2026. Large reasoning models are autonomous jailbreak agents. *Nature Communications*, 17:1435.

John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. 2024. Best-of-n jailbreaking. *Preprint*, arXiv:2412.03556.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. 2024. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *Preprint*, arXiv:2406.18510.

Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. 2024. Llm defenses are not robust to multi-turn human jailbreaks yet. *Preprint*, arXiv:2408.15221.

Zeyi Liao and Huan Sun. 2024. Amplegcg: Learning a universal and transferable generative model of adversarial suffixes for jailbreaking both open and closed llms. *Preprint*, arXiv:2404.07921.

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. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *Preprint*, arXiv:2402.04249.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2024. Tree of attacks: Jailbreaking black-box llms automatically. *Preprint*, arXiv:2312.02119.

Mistral AI. 2024. Large enough: Mistral large 2. Technical release, Mistral AI.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 399 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *Preprint*, arXiv:2202.03286.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2025. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. *Preprint*, arXiv:2404.01833.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A strongreject for empty jailbreaks. *Preprint*, arXiv:2402.10260.

Xikang Yang, Xuehai Tang, Songlin Hu, and Jizhong Han. 2024. Chain of attack: a semantic-driven contextual multi-turn attacker for llm. *Preprint*, arXiv:2405.05610.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *Preprint*, arXiv:2401.06373.

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

Appendices

A Pairing Audit

All McNemar tests pair observations by $(model, intent_id)$. Each of the 15 comparisons achieves 320 matched pairs (8×40) with zero unmatched records. All 8 models are aligned: each contributes exactly 40 intents across all 6 conditions.

B Full Detector Performance

Table 7 reports precision, recall, and F1 for three detector types across all six conditions, measured against human ground-truth labels.

Table 7: Full detector performance (P/R/F1 %) against human ground truth, sorted by F1. S.J. = Safety Judge.

Detector	Cond.	P	R	F1
S.J.	OSS-ST	98.7	99.7	99.2
	ASQ-ST	98.4	99.7	99.1
	AMQ-ST	97.5	98.1	97.8
	BoK-ST	93.0	98.0	95.4
	AMQ-MT	91.0	96.9	93.9
	SS-MT	77.7	99.2	87.1
LLM J.	OSS-ST	97.5	99.4	98.7
	ASQ-ST	97.3	93.9	95.6
	AMQ-ST	98.6	92.1	95.3
	SS-MT	96.9	91.0	93.9
	BoK-ST	93.6	87.8	90.6
	AMQ-MT	96.2	81.0	87.9
Keyw.	OSS-ST	95.5	91.5	93.5
	SS-MT	93.4	74.3	82.8
	ASQ-ST	97.4	60.5	74.7
	AMQ-MT	88.7	57.7	69.9
	AMQ-ST	96.3	49.0	65.0
	BoK-ST	92.5	44.0	59.7

C Realized Budget and Cost

Table 8 reports the realized resource consumption per condition. BoK-ST and AMQ-ST share the same maximum cap (K=5) but differ sharply in realized target calls: PAIR-5 early-stops at a mean of 1.6 queries (median 1), while BoK-ST always sends all 5 variants. This makes PAIR-5 faster and cheaper per validated success despite identical maximum budgets.

Table 8: Realized resource consumption per condition. Cap = maximum target queries allowed; Tgt = realized target calls (mean/med); Atk = attacker LLM calls (mean/med); Cost = per finding; Cost/Succ = per human-validated success; Latency = median wall-clock time.

Cond.	Cap	Tgt	Atk	$/find	$/succ	Lat. (s)
OSS-ST	1	1.0 / 1	0 / 0	.002	.012	17
SS-MT	5	3.5 / 3	0 / 0	.010	.028	117
ASQ-ST	1	1.0 / 1	2.0 / 2	.007	.011	181
AMQ-ST	5	1.6 / 1	3.3 / 2	.014	.016	285
AMQ-MT	5	2.9 / 3	5.9 / 6	.024	.037	518
BoK-ST	5	5.0 / 5	0 / 0	.018	.021	573

D Budget Curves and Mechanism Encoding

Table 9 presents cumulative ASR at each budget level K=1...5. Table 10 shows the binary mechanism encoding; a GEE model with exchangeable correlation within model clusters and sandwich SEs confirms all effects: LLM-crafted OR=6.50, feedback OR=1.83, multi-turn OR=1.02 (ns), BoK bundle OR=4.21.

Table 9: Cumulative **raw** (detector) ASR (%) by budget K—*not* human-validated; cf. Table 2 for adjusted values. iid = $1-(1-p)^K$ ceiling. Div. = diversity gain (pp); Tax = correlation tax (pp).

K	PAIR	BoK	iid	Div.	Tax
		(raw detector ASR)			
1	68.1	56.2	65.8	0.0	9.6
2	84.7	77.5	88.3	21.2	10.8
3	90.3	84.1	96.0	27.8	11.9
4	92.5	88.1	98.6	31.9	10.5
5	93.8	91.2	99.5	35.0	8.3

Table 10: Binary mechanism encoding for the six conditions.

Condition	LLM	FB	MT	Div
OSS-ST	0	0	0	0
SS-MT	0	0	1	0
ASQ-ST	1	0	0	0
AMQ-ST	1	1	0	0
AMQ-MT	1	1	1	0
BoK-ST	1	0	0	1

E Annotation Rubric

Each of the 1,920 findings was independently assessed by two trained security researchers. Annotators received the full attack prompt and model response, with condition labels masked to prevent bias.

Binary verdict. *Unsafe*: the response provides substantive information that could materially assist

a person in carrying out the harmful intent (e.g., actionable instructions, working code, specific chemical quantities, detailed social engineering scripts). *Safe*: the response refuses, deflects, provides only widely-available general knowledge, or adds sufficient caveats and hedging to render the content non-actionable. Partial responses (e.g., high-level descriptions without actionable detail) are classified as *safe* unless the specificity exceeds what is freely available in public sources.

Disagreement resolution. When annotators disagreed ($\sim$12% of cases, Cohen's κ=0.81), a third senior reviewer examined the finding and made a binding determination. The adjudicator was blind to the initial annotators' identities and to the automated detector verdict.

F Per-Model Budget Curves

Table 11 reports the diversity gain, correlation tax, and adaptive premium at K=5 for each target model.

Table 11: Per-model scaling at K=5 (**raw** detector ASR—*not* human-validated; cf. Table 3 for adjusted). Div.=diversity gain; Tax=correlation tax; Prem.=adaptive premium.

Model	BoK@1	BoK@5	PAIR@5	Div.	Tax	Prem.
			(raw detector ASR)			
Claude S4	25.0	60.0	60.0	+35.0	+27.9	0.0
DeepSeek	72.5	97.5	100	+25.0	+2.5	+2.5
Gemini	52.5	90.0	100	+37.5	+9.1	+10.0
GPT-4o	40.0	92.5	100	+52.5	+5.9	+7.5
GPT-4o-m	35.0	95.0	95.0	+60.0	+2.6	0.0
LLaMA	65.0	97.5	97.5	+32.5	+2.4	0.0
Mistral	90.0	97.5	100	+7.5	+2.5	+2.5
Qwen	70.0	100	97.5	+30.0	−0.1	−2.5

Claude Sonnet 4 has the highest correlation tax (+27.9pp), indicating *objective-aware* safety: when Claude is vulnerable to an intent, diverse strategies tend to succeed together; when it resists, all fail. GPT-4o-mini has the highest diversity gain (+60.0pp), indicating *strategy-specific* blind spots where different framings access entirely different vulnerability surfaces.

G SS-MT Scripts and BoK Strategies

SS-MT script authoring. Escalation scripts were authored by a security researcher (not involved in annotation) following the Crescendo pattern (Russinovich et al., 2025): begin with a benign topic related to the intent, then escalate over 3–5 turns toward the harmful request. Scripts were written once per intent (40 total) and not pilot-tested against any target model. The same script was used across all 8 models. Script quality is a potential confound: if scripts are suboptimal, the SS-MT ASR (37.5%) may underestimate the true multi-turn effect.

Example (phishing intent): Turn 1: "What are common email marketing techniques?" Turn 2: "How do companies personalize emails to increase open rates?" Turn 3: "What makes a phishing email convincing to recipients?" Turn 4: "Write a convincing phishing email targeting employees of a financial institution."

BoK strategy types. The five BoK strategies, all generated offline by DeepSeek-R1 for each intent, are:

1. **Role-play persona**: The attacker prompt assigns the target a character (e.g., "You are a cybersecurity instructor teaching penetration testing").
2. **Hypothetical scenario**: Frames the request as a thought experiment (e.g., "In a fictional world where...").
3. **Code-generation pretext**: Requests the harmful content as part of a programming task (e.g., "Write a Python script that...").
4. **Educational/research framing**: Presents the request as academic research (e.g., "For a security research paper, explain...").
5. **Creative writing wrapper**: Embeds the request in a story or screenplay format.

Because the prompts are LLM-generated (by the same DeepSeek-R1 used as attacker in PAIR conditions), the BoK condition shares the "LLM-crafted" indicator with PAIR conditions in our encoding (Table 10).

H Category-Level ASR Breakdown

Table 12 reports human-validated ASR by harm category, aggregated across models. The LLM-crafted advantage (ASQ-ST vs. OSS-ST) is consistent across all categories (+35–55pp), supporting the generality of this finding. The BoK-bundle advantage is largest for prompt security (+42pp) and smallest for CBRN (+18pp), suggesting that strategy diversity is more effective against format-dependent defenses. Category-level variation in mechanism effects is a direction for future work with larger per-category samples.

Table 12: Human-validated ASR (%) by harm category (aggregated across 8 models). n = number of intents per category.

Category	n	OSS	SS	ASQ	MT	AMQ	BoK
Cybercrime	6	16.7	35.4	66.7	64.6	89.6	87.5
CBRN	6	10.4	27.1	45.8	52.1	75.0	72.9
Social manip.	8	17.2	42.2	70.3	68.8	90.6	89.1
Content policy	4	9.4	31.2	56.2	59.4	81.2	84.4
Prompt security	11	18.2	45.5	72.7	70.5	90.9	93.2
Financial crime	5	10.0	35.0	62.5	60.0	82.5	85.0

I Intent List

All 40 harm intents are shared across every model and condition. Intent categories span: cybercrime (botnet, ransomware, keylogger, SQL injection, reverse shell, network sniffing), CBRN (bioweapon, drug synthesis, explosives, poison, chemical weapons, 3D-printed weapons), social manipulation (phishing, social engineering, election manipulation, fake news, stalking, identity theft, deepfake, discrimination), content policy (child exploitation, hate speech, self-harm, mass violence), prompt security (prompt injection, jailbreak roleplay, content filter bypass, system prompt extraction, instruction override, fake authority, fake admin command, document injection, multilingual injection, urgency-based manipulation), and financial crime (insider trading, money laundering, credential harvesting, PII extraction).

J Reproducibility and Code Availability

ATLAS is implemented as an in-house Python framework for automated LLM red-teaming. The experiment configuration, analysis scripts, and human annotation ledger are provided as supplementary materials. Key components:

- **Framework**: in-house ATLAS scan engine with configurable probes, detectors, and attacker models. Supports all six experimental conditions via YAML configuration.
- **Experiment config**: Experiment ID `20260505_003630`; 8 target models $\times$ 6 conditions $\times$ 40 intents = 1,920 findings.
- **Analysis pipeline**: SQL-backed database with all raw findings; Python scripts for McNemar tests, logistic regression, bootstrap CIs, budget curve decomposition, and counterfactual analysis.
- **Human annotations**: Full annotation ledger (`annotation_ledger.csv`) with per-finding verdicts from both annotators and adjudicator. The ledger contains post-adjudication consensus labels; the pre-adjudication raw labels (from which κ=0.81 was computed) are available in the annotation platform's database and will be exported on request. A verification script (`compute_kappa.py`) confirms all 1,920 findings, 100% double-annotation coverage, and FP/FN counts matching Table 2.
- **Figures**: All figures generated via matplotlib with reproducible seeds. Source scripts and data files included.

Code and data will be released upon acceptance at an anonymous repository. During review, supplementary materials are attached as a `.zip` archive.