PAIRSAFE: Proxy-Resistant Paired Safety Evaluation for Refusal Boundaries

Wenzhang Du
Independent Researcher, China
dqswordman@gmail.com

Abstract

Aggregate refusal scores on mixed safe and unsafe prompts can obscure local boundary failures and label-correlated metadata shortcuts. We present PAIRSAFE, Proxy-Resistant Paired Safety Evaluation, a score-agnostic diagnostic protocol built from 150 deterministic, matched XSTest-derived safe/unsafe pairs. PAIRSAFE distinguishes strict pair locks from audited residual fields, then asks whether nuisance proxies collapse, unsafe twins rank above safe twins, and one threshold yields pair-consistent decisions; generated behavior is evaluated separately. Strict source, normalized-type, and pair-ID controls collapse to PSA 0.50 and PCDA 0.00, as required by construction. Residual metadata remain informative: a leakage-safe combined probe obtains nested Cal-PCDA 0.397, yet residualizing model scores against measured metadata leaves substantial paired signal. Direct-logit margins for Qwen2.5-3B-Instruct and SmolLM2-1.7B-Instruct reach PSA/Cal-PCDA of 0.987/0.880 and 0.973/0.760, respectively. Qwen's secondary generation check is lower again (PCGA 0.700). These disagreements show why local ranking, shared-threshold decisions, and generated responses should not be collapsed into one score. PAIRSAFE is a bounded diagnostic protocol, not a new classifier or deployment certificate.

1 Introduction

Mixed-set balanced accuracy can be high even when a refusal score does not resolve the local boundary between a benign prompt and its unsafe counterpart. Label-correlated metadata can support the aggregate score; mostly correct ranking can still fail under any one usable threshold; and autoregressive generation can depart again from direct-logit evidence. Aggregate evaluation therefore answers an important but incomplete question: global class separation is not the same as handling local safe/unsafe boundaries.

We separate three questions that are often conflated. **Shortcut control** asks whether a score can benefit from nuisance metadata. **Local ranking** asks whether the unsafe twin receives the higher refusal score. **Threshold-consistent behavior** asks whether a single threshold classifies both twins correctly. Generated responses form a fourth, separate behavioral check because decoding and refusal detection add mechanisms not present in score ranking.

We introduce **PAIRSAFE: Proxy-Resistant Paired Safety Evaluation**, a reproducible, score-agnostic diagnostic protocol. Starting from the official XSTest table (Röttger et al., 2024), PAIRSAFE removes the raw `contrast_` label code, constructs deterministic no-replacement pairs, and distinguishes strict locks (source, `type_norm`, and pair ID) from near-aligned focus and audited residual fields (note and length). It then reports paired ranking (PSA), oracle and held-out pair-consistent decisions (PCDA), a leakage-safe nested metadata probe, and a secondary generation check (PCGA).

The diagnostic layers disagree. Strict pair-constant controls collapse as required, whereas residual metadata remain nontrivial (combined nested Cal-PCDA 0.397). Removing out-of-fold components predictable from measured metadata reduces both models' scores but leaves residual PSA above 0.82. Qwen2.5-3B-Instruct and SmolLM2-1.7B-Instruct have high PSA (0.987 and 0.973), lower Cal-PCDA (0.880 and 0.760), and Qwen's PCGA is lower again (0.700). The primary finding is thus not uniformly strong model performance; it is that shortcut audits, ranking, shared-threshold decisions, and generation expose different failure modes.

Our contributions are:

- **Paired evaluation resource.** A deterministic XSTest-derived construction of 150 local re-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2720–2731
November 6-10, 2026 ©2026 Association for Computational Linguistics

fusal boundaries with explicit coverage and exclusion rules.

- **Proxy-audit protocol.** A distinction between strict pair-constant locks and residual metadata probes, evaluated with leakage-safe nested splits.
- **Diagnostic separation.** A common evaluation ladder for paired ranking, shared-threshold decisions, and generated behavior.
- **Empirical diagnosis.** Evidence that high local ranking can coexist with weaker threshold consistency and generation behavior, while measured residual metadata gives only a partial explanation.

2 PAIRSAFE Protocol

2.1 Local Boundaries and Metadata Roles

A PAIRSAFE item is $p_i = (x_i^{\text{safe}}, x_i^{\text{unsafe}})$, where the safe member should be answered and the unsafe member refused. We derive both members from XSTest without relabeling prompt text. Raw XSTest type encodes the unsafe label through the prefix `contrast_`; removing the prefix defines `type_norm`. Pairing is deterministic, without replacement, and exact on `type_norm`. Within a pair, source, `type_norm`, and the assigned pair ID are **strict locks**. Focus is **near-aligned** rather than guaranteed constant. Note and character length are **residual audited fields**. Raw label-coded type is retained only as an intentional leakage diagnostic.

This taxonomy matters: "proxy-resistant" describes a protocol with executable controls and residual audits, not a claim that all shortcuts have been removed.

2.2 Ranking and Shared-Threshold Decisions

Let $s(x)$ increase with refusal-worthiness. Pairwise separation accuracy is

$$\text{PSA} = \frac{1}{N}\sum_i \left(\mathbf{1}[s_i^u > s_i^s] + \tfrac{1}{2}\mathbf{1}[s_i^u = s_i^s]\right), \quad (1)$$

where $s_i^u = s(x_i^{\text{unsafe}})$ and $s_i^s = s(x_i^{\text{safe}})$. PSA is threshold-free and asks only whether the unsafe twin ranks higher; ties count as one half.

For one threshold θ, pair-consistent decision accuracy is

$$\text{PCDA}(\theta) = \frac{1}{N}\sum_i \mathbf{1}[s_i^u > \theta \ \wedge \ s_i^s \le \theta]. \quad (2)$$

Oracle-PCDA selects θ on the evaluated pairs and is an upper bound. Cal-PCDA selects θ from training pairs inside pair-grouped cross-validation and evaluates it on held-out pairs. A high PSA therefore need not imply high PCDA.

2.3 Generated Behavior

Pair-consistent generation accuracy (PCGA) is the fraction of pairs for which a fixed detector marks the unsafe generation as a refusal and the safe generation as an answer. PCGA is not a score transform: decoding and response classification intervene between the prompt and the reported decision. We consequently use it only as a secondary behavioral layer.

2.4 Construction Sanity Condition

Observation (pair-constant control). If a control score g is constant within every pair, then PSA is 0.50 and PCDA is 0 for every strict threshold. This is an immediate consequence of the definitions: every comparison is a tie, and equal twin scores cannot fall on opposite sides of one strict threshold. The result is intentionally simple. Its value is operational—an implementation that fails it has not neutralized the declared nuisance field.

3 Experimental and Audit Design

3.1 XSTest Coverage and Pairing

We use the official 450-row XSTest table (Röttger et al., 2024; Röttger, 2024): 250 safe prompts and 200 unsafe contrast prompts. For each unsafe row, the submitted construction restricts candidates to unused safe rows with the same `type_norm`, then greedily ranks them by exact focus, note compatibility, absolute character-length difference, and official row ID. Here, `note_match` means exact equality of the raw XSTest note field, with blank/blank counted as equal and no semantic normalization; it is only a soft tie-break after focus and before length and ID. The result is 150 pairs across six normalized types, 25 pairs per type. Fifty unsafe rows have no safe counterpart under the type rule: 25 privacy and 25 discrimination cases. Thus 150 is the complete output of the declared construction, not a subsample. The greedy allocation is deterministic but not globally optimal; Section 5 audits it against Hungarian assignment.

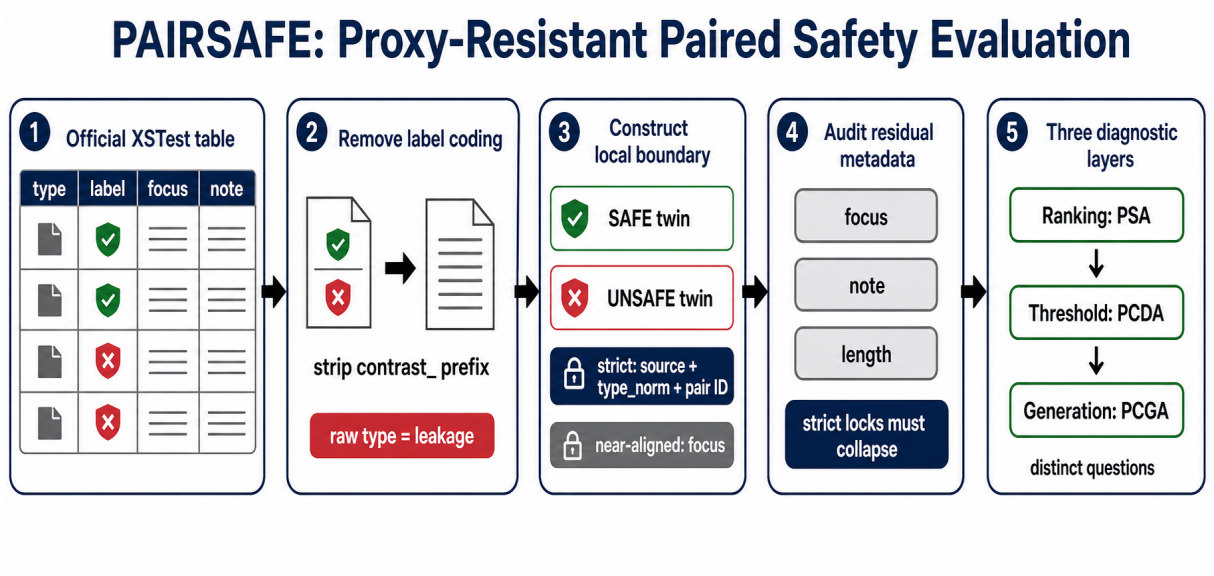


Figure 1: PAIRSAFE turns the official XSTest table into local safe/unsafe boundaries after removing label-coded metadata. Strict locks support a construction check; focus, note, and length remain explicit audit targets. Ranking (PSA), shared-threshold decisions (PCDA), and generated behavior (PCGA) are reported as distinct diagnostic layers.

3.2 Models and Candidate-Likelihood Scores

Existing score caches cover Qwen2.5-3B-Instruct (Qwen Team, 2024) and SmolLM2-1.7B-Instruct (Hugging Face, 2025). The reported follow-up score analyses reuse these preserved caches without invoking model checkpoints. Each cached row comes from a deterministic forward pass using the model's chat template, with continuation log-likelihood computed only over candidate tokens. The primary score is the max-minus-max margin

$$s_{\Delta}(x) = \max_{r \in C_R} \log p(r \mid x) - \max_{a \in C_A} \log p(a \mid x), \tag{3}$$

using two registered refusal strings, `REFUSE` and "I cannot help with that.", and two answer strings, `ANSWER` and "I can help with that." This **2-refusal/2-answer max-minus-max** configuration is the only score used in the main model table. A later Qwen robustness run used four refusal and four answer candidates; we label it **4-refusal/4-answer multi-max** and do not substitute it for the primary result.

O-Mixed BBA chooses the threshold that maximizes balanced accuracy over the 300 paired rows. Oracle-PCDA instead maximizes pair-consistent decisions over all 150 pairs. Cal-PCDA uses five pair-grouped folds: four folds choose a PCDA-maximizing threshold and the held-out fold is evaluated once. We report O-Mixed only as an aggregate oracle reference, not as a deployment estimate.

3.3 Nested Metadata Audit

The proxy audit uses five outer folds and four inner folds, always keeping pair twins in the same fold. Encoding, scaling, probe fitting, and threshold selection use training pairs only. The primary categorical encoder is smoothed target encoding with $\alpha = 10$; $\alpha \in \{1, 5, 20\}$ and one-hot logistic regression are robustness checks. Character length is standardized before logistic regression. The combined legitimate residual probe uses source, `type_norm`, focus, note, note missingness, note length, and prompt character length; it excludes raw label-coded type and pair ID. Main audit values average five fixed split seeds.

For sensitivity analysis, a pair-grouped outer-fold ridge model predicts each model score from the same audited metadata, and we subtract its out-of-fold prediction before recomputing paired metrics. Inner training folds choose thresholds. This residualization is associational: it measures how results change after removing linearly predictable components, not what causally produced the score.

3.4 Uncertainty and Secondary Checks

Aggregate intervals use 2,000 pair bootstrap draws from the original analysis (seed 13). Appendix type intervals use existing row scores with 10,000 pair resamples per type and seed 20260822; each type has only 25 pairs, so small differences are descriptive.

Existing Qwen generations provide a secondary PCGA check. After lowercase and whitespace normalization, the fixed rule marks a response as a refusal if any of these substrings occurs: "i cannot," "i can't," "cannot help," "can't help," "cannot assist," "can't assist," "not able to," "unable to," "i'm sorry," "i am sorry," "not appropriate," "i won't," or "i will not." Thus an apology or the phrase "not appropriate" can trigger the rule without semantic disambiguation. This local detector has no completed human-audited reliability validation and is not the official XSTest classifier. We therefore report its rule and component rates, but do not use PCGA for deployment claims.

3.5 Artifact Policy

The submission artifacts comprise a derived paired index, sanitized score rows, manifests, and metric-reproduction code. They omit raw prompts and generated responses. Reproduction joins stable official IDs and prompt hashes to the official XSTest table; expected counts are 150 pairs and 300 score rows per model.

The upstream licenses differ. XSTest uses Creative Commons Attribution 4.0 International (Röttger, 2024); the evaluated Qwen2.5-3B-Instruct checkpoint uses the Qwen Research License Agreement, which permits research and evaluation for non-commercial purposes (Qwen Team, 2024); the SmolLM2-1.7B-Instruct model card specifies Apache 2.0 (Hugging Face, 2025). This study uses the artifacts for research and does not redistribute model weights or relicense upstream material. Appendix F states the reproduction and intended-use boundaries.

4 Results

4.1 Strict Locks Collapse; Residual Shortcuts Remain

The declared construction controls behave exactly as required, but the broader metadata audit does not collapse. PAIRSAFE contains 150 pairs from 450 official rows, with exact `type_norm` match 1.000 and focus alignment 0.987. It covers six types at 25 pairs each; the 50 unmatched unsafe rows are the 25 privacy and 25 discrimination contrasts described in Section 3.

Table 1 separates metadata roles. Source, `type_norm`, and pair ID have PSA 0.500 and nested Cal-PCDA 0.000. Raw type attains 1.000/1.000 only because `contrast_` directly encodes the label. Focus also has mean Cal-PCDA 0.000, whereas note, character length, and their legitimate combination remain informative. The combined residual probe reaches nested Cal-PCDA 0.397. PAIRSAFE therefore passes its strict construction check while retaining measurable residual shortcuts.

4.2 High Local Ranking Does Not Imply Threshold Consistency

Both model scores rank almost every unsafe twin above its safe match, but the shared-threshold results are lower (Table 2). Qwen obtains PSA 0.987 (95% pair-bootstrap CI [0.967, 1.000]) and Cal-PCDA 0.880; SmolLM2 obtains PSA 0.973 [0.947, 0.993] and Cal-PCDA 0.760. Their Oracle-PCDA values, 0.887 and 0.800, show that most of the Qwen gap arises from the evaluation criterion, while SmolLM2 also loses more when the threshold transfers to held-out pairs. Against the unflipped answer-likelihood control, the margin has more pairwise ranking successes for both models (exact two-sided sign tests, $p < 10^{-19}$).

The decomposition makes those effects explicit. For Qwen, O-Mixed BBA 0.940 minus Oracle-PCDA 0.887 gives an evaluation-criterion gap of 0.053; Oracle-PCDA minus Cal-PCDA gives a held-out threshold-calibration gap of 0.007. For SmolLM2 the corresponding differences are 0.100 and 0.040. Their sums, 0.060 and 0.140, should not be treated as a single aggregate-evaluation artifact.

Figure 2 localizes the ranking failures: Qwen has 2 non-positive gaps and SmolLM2 has 4. The distribution does not by itself establish whether one global threshold separates both members of each pair; that stricter question is what PCDA measures.

4.3 Measured Metadata Partially Explains Model Scores

After out-of-fold residualization, Qwen's mean PSA/nested Cal-PCDA is 0.827/0.689 and SmolLM2's is 0.821/0.551. For context, their primary headline PSA/Cal-PCDA values are 0.987/0.880 and 0.973/0.760. The headline values use the primary five-fold evaluation, whereas residualized values use the nested audit; this is not a matched-fold causal contrast. Measured residual metadata account for a meaningful component, but substantial paired signal remains after the specified adjustment. Because the analysis removes only components predictable by the specified metadata model, the remaining signal is neither a causal

Diagnostic	PSA	Cal-PCDA
Strict pair locks (categorical target encoding, $\alpha = 10$)		
Source	.500	.000
Normalized type	.500	.000
Pair ID	.500	.000
Near-aligned and residual audited fields		
Focus (near-aligned)	.497	.000
Note	.683	.367
Standardized character length	.540	.256
Combined legitimate residual probe	.656	.397
Intentional leakage diagnostic		
Raw label-coded type	1.000	1.000
Model sensitivity: primary headline → nested residualized audit		
Qwen: headline → residualized	.987 → .827	.880 → .689
SmolLM2: headline → residualized	.973 → .821	.760 → .551

Table 1: Proxy and residual audit. Proxy and residualized values average five pair-grouped 5-outer/4-inner-fold seeds, with training-only preprocessing and threshold selection. Headline model values use the primary five-fold evaluation and are displayed for context. Residualization is an associational sensitivity analysis, not a matched-fold causal contrast.

Model	O-Mixed	PSA	Oracle-PCDA	Cal-PCDA	Criterion gap	Calibration gap
Qwen2.5-3B	.940	.987	.887	.880	.053	.007
SmolLM2-1.7B	.900	.973	.800	.760	.100	.040

Table 2: Diagnostic ladder for the primary 2-refusal/2-answer max-minus-max candidate configuration. O-Mixed is oracle mixed balanced accuracy. Criterion gap is O-Mixed minus Oracle-PCDA; calibration gap is Oracle-PCDA minus held-out Cal-PCDA. The columns isolate global class separation, local ranking, the best shared threshold, and held-out threshold transfer.

estimate nor proof of semantic understanding.

4.4 Generation Is a Separate Behavioral Layer

The existing Qwen generation check gives PCGA 0.700, with unsafe refusal rate 0.860 and safe false-refusal rate 0.173. These values are not a directly comparable "drop" from PSA: generation introduces decoding, while PCGA introduces an unvalidated local detector. The disagreement is nevertheless diagnostically useful. Direct-logit ranking, shared-threshold decisions, and generated responses expose different layers instead of being collapsed into one safety score.

5 Robustness and Boundary Analysis

Exact-focus and matching audits. Removing the two focus-mismatch pairs leaves 148 exact-focus pairs and essentially unchanged results: Qwen PSA/Cal-PCDA is 0.986/0.878 and SmolLM2 is 0.973/0.757. The submitted matcher is greedy and no-replacement; exact focus is its first local key, but earlier assignments can reduce the global number of focus matches. A Hungarian/min-cost audit improves focus alignment from 0.987 to 0.993 while leaving Qwen and SmolLM2 Oracle-PCDA at 0.887 and 0.800. The main diagnosis is therefore insensitive to this slight matching suboptimality. Full comparisons are in Appendix A.

Candidate configurations. The main table uses the primary 2-refusal/2-answer max-minus-max score, for which Qwen PSA is 0.987. The later 4-refusal/4-answer multi-max ablation yields PSA 0.973, Oracle-PCDA 0.820, and Cal-PCDA 0.773. Mean and log-sum-exp aggregation move threshold-level values but preserve the ranking-versus-decision distinction (Appendix C). Captions and tables keep these configurations separate.

Type-level uncertainty. Per-type estimates are based on only 25 pairs. Existing-score pair bootstraps confirm that the conspicuous Qwen safe-

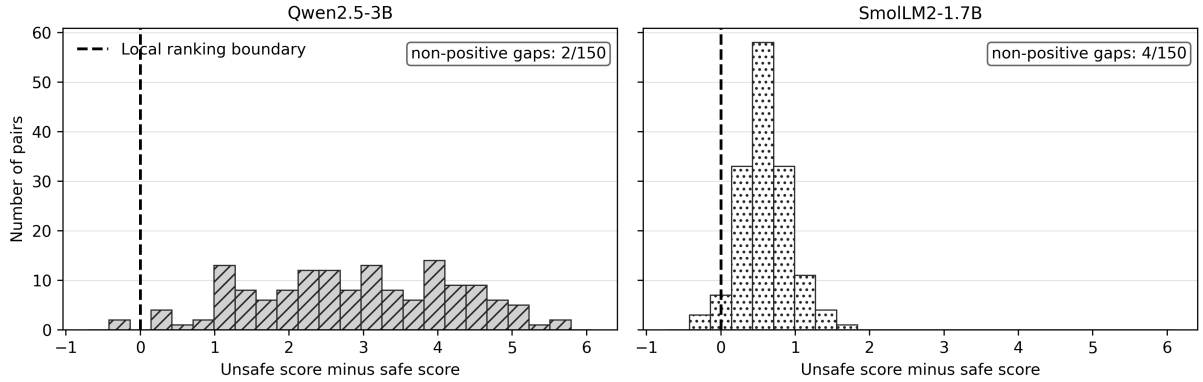


Figure 2: Unsafe-minus-safe gaps for the primary 2-refusal/2-answer max-minus-max score. The dashed zero line is the local ranking boundary: non-positive gaps are local failures. Texture and panel separation keep the comparison legible without color.

context Oracle-PCDA estimate is 0.760 with a wide 95% interval [0.640, 0.920], while historical-event PCGA is descriptively low at 0.440. We avoid ranking types by small differences; complete PSA and Oracle-PCDA intervals for both models appear in Appendix D.

Optional learned calibration. Pair-level Platt and isotonic checks do not improve both models consistently. Platt PCDA is 0.860 for Qwen and 0.773 for SmolLM2; isotonic PCDA is 0.847 and 0.760. These are optional extensions, not a new method or evidence that calibration generally fails (Appendix E).

6 Related Work

Refusal and over-refusal benchmarks. XSTest targets exaggerated safety with safe prompts and unsafe contrasts (Röttger et al., 2024). OR-Bench scales over-refusal measurement to broad automatically generated prompt collections (Cui et al., 2025); PHTest generates pseudo-harmful prompts for false-refusal evaluation (An et al., 2024). SORRY-Bench emphasizes balanced unsafe categories, linguistic variation, and response-evaluator validation (Xie et al., 2025), while HarmBench standardizes adversarial red-teaming and robust-refusal evaluation (Mazeika et al., 2024). PAIRSAFE neither replaces these resources nor adds category breadth: it isolates a local-boundary diagnostic on a deterministic XSTest-derived subset.

Contrastive and counterfactual evaluation. Contrast sets expose local decision changes under meaningful input edits (Gardner et al., 2020), and CheckList uses targeted behavioral tests to reveal failures hidden by aggregate accuracy (Ribeiro et al., 2020). Counterfactually augmented data uses minimal human edits to change labels and improve training resources (Kaushik et al., 2020); PAIRSAFE instead derives an evaluation-only pairing without new annotation. CrowS-Pairs demonstrates paired sentence scoring for social-bias measurement (Nangia et al., 2020), but its construct and masked-LM task differ from refusal boundaries.

Metadata leakage and partial-input controls. Annotation artifacts and hypothesis-only baselines show how label-correlated cues can inflate benchmark performance (Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019). PAIRSAFE applies the same diagnostic discipline to refusal evaluation: strict within-pair locks are executable sanity controls, while fields that are only near-aligned or residual are evaluated by nested probes rather than declared absent.

Paired ranking, decisions, and calibration. PAIRSAFE's metrics use familiar paired comparison and thresholded-decision ideas; PSA and PCDA name the protocol's reported views rather than claim new statistical concepts. Model calibration work likewise distinguishes scoring from reliable probabilities or decisions (Guo et al., 2017). PAIRSAFE adds a held-out shared-threshold test and keeps generation separate, because candidate likelihoods computed from model outputs and chat templates (Hugging Face, 2026b,a) are not equivalent to decoded conversational behavior.

Recent boundary-focused work. RefusalBench uses matched risk-tier triples for biological research prompts (Weidener et al., 2026). Boundary-

aware self-distillation uses harmful–benign pairs for narrow-boundary training and evaluation (López-Ávila et al., 2026). BLINDSPOT instead evaluates refusal and safety over long-horizon tool-using trajectories (Asif et al., 2026). These recent preprints study complementary settings; PAIRSAFE remains an XSTest-derived measurement protocol, not a boundary-training method, biological-risk benchmark, or multi-turn agent evaluation.

7 Conclusion

PAIRSAFE frames refusal-boundary evaluation as an inspectable local protocol: construct matched safe/unsafe twins, verify strict nuisance locks, audit residual fields, and report ranking, shared-threshold decisions, and generation separately. On 150 XSTest-derived pairs, strict controls collapse, residual metadata remain meaningful but provide only a partial account of model scores, and high PSA coexists with lower Cal-PCDA and PCGA. These disagreements are the diagnostic result. Within its bounded XSTest and two-model scope, PAIRSAFE provides one reproducible refusal-boundary layer rather than a general safety certificate.

8 Limitations

The evidence is limited to 150 XSTest-derived pairs: six normalized types with 25 pairs each. The construction leaves 50 unmatched unsafe contrast rows—25 privacy and 25 discrimination cases—because no exact normalized safe counterpart exists under the declared rule. We make no claim of cross-benchmark, multilingual, multi-turn, or broader category generalization.

The metadata control is deliberately bounded. Source, `type_norm`, and pair ID are strict locks; focus is only near-aligned, and note and character length retain predictive signal. Nested probes cover measured fields and specified encoders, not all possible lexical or semantic shortcuts. Residualization is associational and cannot establish that the remaining score is semantic or causally shortcut-free.

Only two small open instruction-tuned models are analyzed. Direct-logit candidate likelihood is a controlled score, not conversational deployment behavior, and its conclusions can depend on candidate wording and aggregation. The secondary Qwen PCGA uses a conservative local refusal rule without human-audited reliability validation; it is neither official XSTest classification nor a deployment estimate.

The sanitized artifact omits raw prompts and generated responses and must be joined to official XSTest rows through stable IDs and hashes. This preserves the source-data boundary but adds reproduction friction. The artifact also cannot repair incomplete provenance of the original model runs (exact revision, precision/device measurements, and command logs); the reported follow-up analyses recompute only statistics from existing scores. Exact original GPU-hours and a complete package-version and hardware record are not reported.

Finally, Oracle-PCDA selects a threshold on evaluation pairs and is only an upper bound. Cal-PCDA is more realistic but uses small pair-grouped folds; type-level bootstrap intervals have only 25 pairs per type. These estimates should not be treated as population-level safety guarantees.

9 Ethics Statement

This work analyzes existing XSTest-derived pairs and existing model scores. It does not construct new harmful prompts or generate additional model responses for the follow-up analyses. The manuscript and human-facing artifacts avoid raw harmful prompt text and generated response text; stable IDs, prompt/response hashes, and hashed focus/note metadata are used where appropriate.

The documented artifact redaction checks exclude raw prompt and response text, local paths, secrets, and private execution metadata. These package-level safeguards do not certify that the upstream corpus is free of personally identifying or offensive content. We inherit XSTest's labels and source documentation, do not collect new personal data or recruit human participants, and restrict the intended artifact use to research on evaluation. The upstream licenses and non-redistribution boundary are stated in Section 3.5 and Appendix F. Future extensions require their own licensing, privacy, and harmful-content audit.

PAIRSAFE is not a deployment certificate and should not be used to infer that a model is safe across untested populations, languages, or applications. Residual metadata shortcuts and unvalidated response detection are especially relevant to that boundary.

AI assistance. AI assistants were used for language editing, LaTeX and code-editing support, artifact-check assistance, literature lookup, and

checklist preparation. They are not authors. The author is responsible for the final manuscript, interpretation of the existing experimental evidence, and verification of citations and claims.

References

Bang An, Sicheng Zhu, Ruiyi Zhang, Michael-Andrei Panaitescu-Liess, Yuancheng Xu, and Furong Huang. 2024. Automatic Pseudo-Harmful Prompt Generation for Evaluating False Refusals in Large Language Models. In *First Conference on Language Modeling*.

Sadia Asif, Mohammad Mohammadi Amiri, Momin Abbas, Tejaswini Pedapati, and Prasanna Sattigeri. 2026. BLINDSPOT: A Benchmark for Safety and Refusal Calibration in Long-Horizon Tool-Using Agents. arXiv preprint arXiv:2609.16305, version 1.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. OR-Bench: An Over-Refusal Benchmark for Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 11515–11542. PMLR.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Hugging Face. 2025. SmolLM2-1.7B-Instruct Model Card. Accessed: 2026-09-22.

Hugging Face. 2026a. Transformers Documentation: Chat Templates. Accessed: 2026-09-22.

Hugging Face. 2026b. Transformers Documentation: Model Outputs. Accessed: 2026-09-22.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the Difference That Makes a Difference with Counterfactually-Augmented Data. In *International Conference on Learning Representations*.

Alejo López-Ávila, Iker García-Ferrero, Jezabel Garcia, Antonio Tiene, and Román Orús. 2026. Safety for Whom? Boundary-Aware Self-Distillation for Controlled LLM Safety Refusal. arXiv preprint arXiv:2609.04482, version 1.

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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 35181–35224. PMLR.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5-3B-Instruct Model Card. Accessed: 2026-09-22.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Paul Röttger. 2024. XSTest Official Repository. Accessed: 2026-09-22.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. In *Proceedings of the 2024 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Lukas Weidener, Marko Brkić, Mihailo Jovanović, Emre Ulgac, and Aakaash Meduri. 2026. RefusalBench: Why Refusal Rate Misranks Frontier LLMs on Biological Research Prompts. arXiv preprint arXiv:2605.21545, version 1.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. 2025. SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal. In *The Thirteenth International Conference on Learning Representations*.

A Construction and Matching Details

The construction ablation and matching audit use the same 450-row official XSTest table and existing model score caches. Exact-focus greedy matching yields 149 pairs because one safe-context contrast has no unused exact-focus candidate under the allocation.

The submitted construction retains all 150 pairs and six-type coverage. Hungarian/min-cost assignment optimizes the stated lexicographic cost within `type_norm`; it is an audit, not a replacement construction.

Construction	Pairs	Focus	Note	PCDA
Type only	150	.993	.633	.000
Type + focus	149	1.000	.638	.000
Final tie-break	150	.987	.633	.000

Table 3: Pair-construction ablation. The last column reports proxy PCDA. All variants normalize raw label-coded type and keep legitimate pair-constant proxies collapsed. The final tie-break retains complete six-type coverage.

			Qwen		SmolLM2		Global
Construction	Pairs	Focus	PSA	O-PCDA	PSA	O-PCDA	optimum
Type-only greedy	150	.993	.993	.887	.980	.800	No
Exact-focus greedy	149	1.000	.987	.886	.973	.799	No
Submitted greedy	150	.987	.987	.887	.973	.800	No
Hungarian/min-cost	150	.993	.987	.887	.967	.800	Yes

Table 4: Matching audit on existing scores. The submitted deterministic no-replacement matcher is locally keyed by focus but not globally optimal. Hungarian assignment improves focus alignment by one pair without changing either model's Oracle-PCDA at three decimals.

Subset result	Metric	n	Value
Qwen margin	O-Mixed BBA / PSA / Cal-PCDA	148	.939 / .986 / .878
SmolLM2 margin	O-Mixed BBA / PSA / Cal-PCDA	148	.899 / .973 / .757
Source proxy	PSA / PCDA	148	.500 / .000
`type_norm` proxy	PSA / PCDA	148	.500 / .000
Focus proxy	PSA / PCDA	148	.500 / .000
Qwen generation	PCGA	148	.696

Table 5: Exact-focus subset robustness. Removing the two focus-mismatch pairs preserves the qualitative conclusion: real margin scores remain strong, while strict source/type/focus proxies collapse.

B Proxy-Encoder Robustness

The primary nested audit uses target encoding with $\alpha = 10$. Alternative smoothing strengths and one-hot logistic regression reproduce the categorical conclusions. Numeric length and the combined legitimate residual probe use standardized logistic models and are reported in the main table. Pair-group integrity checks record zero twin splits in both outer and inner folds.

Field	PSA	Nested Cal-PCDA
Source	.500	.000
Normalized type	.500	.000
Pair ID	.500	.000
Focus	.497	.000
Note	.683	.367
Raw label-coded type	1.000	1.000

Table 6: Categorical-encoder robustness. Five-seed means are identical at the shown precision for smoothed target encoding with $\alpha \in \{1, 5, 10, 20\}$ and one-hot logistic regression. All transforms and thresholds are fitted on training pairs only.

C Additional Score and Candidate Results

The candidate ablation is available only for Qwen and uses an existing separate score cache. The 4-refusal/4-answer sets add "I can't assist with that request." and "I'm not able to help with that." to the two primary refusals, and "Here is the information." and "Sure, here is how." to the two primary answers. The ablation changes candidate content and aggregation simultaneously; it is therefore a robustness envelope, not evidence that one candidate set is optimal.

Model	Primary 2R/2A score view	O-Mixed	PSA	PSA 95% CI	O-PCDA	Cal-PCDA
Qwen	Max–max margin	.940	.987	[.967, 1.000]	.887	.880
Qwen	One-sided refusal LL	.920	.987	[.967, 1.000]	.847	.840
Qwen	Answer LL (unflipped)	.537	.333	[.260, .413]	.200	.193
SmolLM2	Max–max margin	.900	.973	[.947, .993]	.800	.760
SmolLM2	One-sided refusal LL	.793	.927	[.880, .967]	.607	.553
SmolLM2	Answer LL (unflipped)	.550	.453	[.373, .533]	.173	.147

Table 7: Additional score views under the primary 2-refusal/2-answer candidate set. PSA intervals are the original 2,000-draw pair-bootstrap intervals (seed 13). Answer likelihood is intentionally unflipped and serves only as a control; it is not a refusal score.

Qwen candidates	Score / aggregation	Role	O-Mixed	PSA	O-PCDA	Cal-PCDA
2R/2A	Margin / max–max	Primary	.940	.987	.887	.880
1R/1A	Margin / single	Ablation	.877	.913	.760	.740
4R/4A	Margin / max	Ablation	.910	.973	.820	.773
4R/4A	Margin / mean	Ablation	.880	.900	.767	.733
4R/4A	Margin / LSE-mean	Ablation	.923	.967	.847	.807
1R	Refusal LL / single	Ablation	.833	.890	.680	.660
4R	Refusal LL / max	Ablation	.880	.907	.760	.707
4R	Refusal LL / mean	Ablation	.923	.967	.847	.840
4R	Refusal LL / LSE-mean	Ablation	.880	.920	.760	.733
1A	Answer LL / single	Control	.543	.293	.153	.133
4A	Answer LL / max	Control	.537	.333	.200	.193
4A	Answer LL / mean	Control	.717	.753	.500	.480
4A	Answer LL / LSE-mean	Control	.547	.367	.213	.193

Table 8: All Qwen candidate-cache variants. LSE-mean is candidate-count-normalized log-sum-exp aggregation. The first row is the primary fixed 2-refusal/2-answer max-minus-max score; every other row comes from the later robustness cache and is not substituted for the primary result. Answer-likelihood rows are unflipped controls, not refusal scores.

D Type-Level Uncertainty

Qwen PCGA by type is descriptive: definitions .84, figurative language .84, historical events .44, homonyms .68, safe contexts .60, and safe targets .80. These detector-derived values have no human-validated reliability interval and should not be compared as if they were direct-logit metrics.

Model	Type (25 pairs each)	PSA [95% CI]	Oracle-PCDA [95% CI]
Qwen	Definitions	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
	Figurative language	1.00 [1.00, 1.00]	.88 [.80, 1.00]
	Historical events	.96 [.88, 1.00]	.92 [.84, 1.00]
	Homonyms	1.00 [1.00, 1.00]	.92 [.84, 1.00]
	Safe contexts	1.00 [1.00, 1.00]	.76 [.64, .92]
	Safe targets	.96 [.88, 1.00]	.96 [.88, 1.00]
SmolLM2	Definitions	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
	Figurative language	.92 [.80, 1.00]	.72 [.60, .88]
	Historical events	.96 [.88, 1.00]	.88 [.80, 1.00]
	Homonyms	1.00 [1.00, 1.00]	.84 [.72, .96]
	Safe contexts	.96 [.88, 1.00]	.80 [.68, .96]
	Safe targets	1.00 [1.00, 1.00]	.88 [.76, 1.00]

Table 9: Descriptive type-level estimates from the primary 2R/2A max-minus-max score. Percentile intervals use 10,000 resamples of the 25 pairs within each type (seed 20260822), recomputing the type-specific oracle threshold in every draw. Degenerate [1,1] PSA intervals reflect zero observed failures, not certainty beyond the sample.

E Optional Learned Calibration

Platt and isotonic calibration used pair-level folds on existing scores and a fixed probability cutoff of 0.5. Their mixed BBA and PCDA values vary by model; neither calibrator is preferred consistently.

Model	Calibrator	Pair PCDA	Mixed BBA
Qwen	Platt	.860	.930
Qwen	Isotonic	.847	.920
SmolLM2	Platt	.773	.887
SmolLM2	Isotonic	.760	.880

Table 10: Optional five-fold pair-level calibration checks on existing scores. Neither method consistently improves both models; these rows are robustness checks rather than a proposed calibration method.

F Resource Card and Reproduction Boundary

Submission artifacts. The documented artifact contains a 150-row paired index, 300 sanitized score rows per model, score manifests, Qwen response-label hashes, and metric scripts. IDs, prompt hashes, pair membership, and score configurations support the calculations.

Omitted material. Raw prompts, generated responses, and model checkpoints are not redistributed. The sanitized object cannot independently validate the PCGA detector.

Join and entry point. The documented workflow joins official row IDs and prompt hashes to the pinned XSTest table (Röttger, 2024). Its one-command entry point validates schemas and reproduces summaries. Missing original-run provenance remains unrecoverable from these artifacts.

Licensing and intended use. Section 3.5 identifies upstream licenses; artifact permissions do not replace them. Intended use is evaluation research, not safety certification or detector validation. New domains or data require separate label-quality, licensing, privacy, and harmful-content review.