Script as a Safety Signal: Counterfactual Evidence from Romanization

Anonymous ACL submission

Abstract

The *visual form* of text—its script—can function as a safety-relevant signal for large language models, even when the intended request is approximately held fixed. In a counterfactual romanization experiment spanning five typologically diverse languages (Arabic, Japanese, Korean, Chinese, Hindi) on two open-source models, converting harmful prompts from native script to Latin script reduces refusal rates by 41.4 percentage points on average (N=50 per condition; exact McNemar $p < .05$ for all ten cells, Bonferroni-significant for 9/10). The largest effects emerge in Arabic (66 pp), Japanese (62 pp), and Korean (50 pp). Crucially, evaluator choice introduces systematic variance: two independent LLM judges disagree on approximately one in four refusal labels (κ=0.35), whereas human adjudication of this 48-case disagreement subset yields complete agreement (κ=1.000). These findings support treating script as a first-class variable in multilingual safety evaluation and identify a concrete source of measurement instability.

1 Introduction

Current safety alignment evaluations often operationalize risk primarily through the semantic content of a query. Multilingual deployments challenge this assumption: the same harmful request can be expressed in native script, romanized form, mixed script, or translated safety framing, and the *visual form* may alter both model responses and evaluator judgments. We investigate whether script is a causally relevant safety variable in a controlled romanization setting, rather than merely a display artifact.

We address this question through counterfactual romanization: for each harmful prompt, we hold the intended request approximately fixed while manipulating only the script condition. The central finding is that romanization substantially and reproducibly lowers refusal rates across two N=50 model runs traceable from archived artifacts and one diagnostic run. The largest effects appear in Arabic, Japanese, and Korean, with Chinese and Hindi exhibiting smaller yet meaningful gaps. A secondary but equally important finding concerns evaluation reliability: a single LLM judge shows only marginal agreement with keyword-based refusal labels, whereas a multi-annotator audit achieves substantially higher consistency. Together, these findings make script a measurable evaluation variable that should be reported alongside language and task content.

This matters because multilingual safety scores are often interpreted as properties of a model-language pair. If script form changes refusal behavior while language and intended request are approximately fixed, then a single native-script score is not a stable estimate of multilingual safety. It may instead reflect the particular surface form chosen by the benchmark author. The same concern applies to automated judging: if evaluators respond differently to native and romanized text, the measured safety gap can combine model behavior with measurement bias. Our study therefore treats script and evaluator choice as joint sources of validity risk.

Our contributions are:

- **Counterfactual evidence of a script safety gap**: Counterfactual romanization reveals a substantial, reproducible effect—models show much lower refusal rates for romanized prompts than for matched native-script prompts (9/10 Bonferroni-significant cells across two models and five languages).
- **Evaluator reliability as a critical confound**: Independent LLM judges disagree on approximately one in four refusal labels (κ=0.35), revealing that evaluator choice systematically biases multilingual safety assessments.

- **A reproducible audit protocol**: All main tables are deterministically recomputed from audited artifacts with source hashes, explicit provenance boundaries, and a completed human adjudication layer.

2 Related Work

Prior work documents multilingual jailbreak and safety disparities (Shen et al., 2024; Yong et al., 2023; Deng et al., 2023; Emani and R, 2025), and broader multilingual evaluation shows persistent language-coverage gaps (Joshi et al., 2020; Hu et al., 2020; Ruder et al., 2021; Muennighoff et al., 2023; Lee et al., 2025). Script and tokenizer design can itself change representation cost and model behavior (Sennrich et al., 2016; Kudo and Richardson, 2018; Rust et al., 2021; Petrov et al., 2023; Chen et al., 2024), which motivates treating romanization as an intervention rather than only as a translation artifact.

Our safety-evaluation framing builds on toxic-generation, red-teaming, and jailbreak work (Gehman et al., 2020; Ganguli et al., 2022; Perez et al., 2022; Mazeika et al., 2024; Weidinger et al., 2024; Chen et al., 2025; Zou et al., 2023b; Liu et al., 2023; Chao et al., 2023; Wei et al., 2023; Zhao et al., 2025; Mu et al., 2025). Our design follows counterfactual and causal evaluation principles (Pearl, 2009; Feder et al., 2022; Veitch et al., 2021; Goyal et al., 2019) and reports agreement using standard reliability measures (Cohen, 1988; Landis and Koch, 1977; Fleiss, 1971). Alignment techniques such as RLHF, constitutional AI, DPO, and activation steering (Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023; Zou et al., 2023a; Panickssery et al., 2024) motivate our exploratory mitigation analysis, but we separate evaluation evidence from training-provenance claims.

Our work differs from prior multilingual jailbreak studies in two ways. First, we isolate script form through matched native/romanized prompt pairs rather than comparing unrelated prompts across languages. This makes the unit of analysis a paired counterfactual contrast instead of a cross-lingual aggregate. Second, we audit the evaluator itself. Prior safety evaluations often assume that a refusal label is stable once a judge model or keyword rule is chosen; our results show that this assumption can fail in precisely the multilingual settings where script variation is most likely. These two design choices position the paper as a measurement-validity study rather than a new attack benchmark.

3 Method

Counterfactual design. For each prompt p, language ℓ, and model m, we compare a native-script condition with a romanized condition. The treatment is script form, not topic or language identity. Refusal is measured as a binary outcome, and effect sizes are reported as percentage-point deltas and Cohen's h. This design is intentionally conservative: it does not claim that native and romanized strings are semantically identical in every linguistic respect. Instead, it asks whether a minimal surface-form intervention, applied to matched prompt pairs and evaluated with paired tests, produces a stable change in refusal behavior large enough to matter for safety evaluation. The paired design therefore removes many item-level confounds that would affect unpaired cross-language comparisons, while leaving romanization quality and pragmatic shift as explicit limitations rather than hidden assumptions.

Data and models. The canonical E3 table uses 50 prompt pairs per language-condition for GPT-OSS-120B and Qwen3-32B across Arabic, Japanese, Korean, Chinese, and Hindi. We also include a smaller Qwen3-32B-0 diagnostic run with $N = 30$ per language-condition; this diagnostic combines results from separate runs for Arabic ($N = 30$), Chinese ($N = 30$), and the extended Japanese/Korean/Hindi ($N = 30$ each) conditions. Earlier pilot values are retained only as historical provenance and excluded from headline claims because they are outside the canonical E3 v4 audit.

Romanization and metrics. Romanization was generated or checked with automated transliteration tools and manual consistency screens. Because romanization can introduce semantic drift, we report this as a limitation rather than claiming perfect semantic identity. The main metrics are refusal rate, refusal delta, Cohen's h, and a supporting Script Safety Gap that summarizes native-vs-romanized inconsistency. A paired significance audit computes exact McNemar binomial tests on the matched native/romanized outcomes and reports both uncorrected and Bonferroni-adjusted decisions. We treat refusal as an observable evaluation outcome rather than as direct evidence of the model's internal safety representation. This distinction matters: a lower romanized refusal rate

may arise from tokenizer pathways, safety-tuning coverage, instruction-following priors, or evaluator misclassification. The paper therefore reports effect sizes and reliability gates side by side. The main claim is not that romanization is the only cause of multilingual safety failure, but that script form is sufficiently consequential that evaluations which ignore it risk overestimating robustness.

Figure 1 provides a compact map of the paper. Panel A shows the matched native-script/romanized intervention, Panel B summarizes the main E3 refusal-drop result, and Panel C states the reliability and provenance gates that define the claim boundary. Detailed visualizations are in Appendices A–B.

4 Experiments and Results

4.1 E3: Counterfactual Romanization

Table 1 presents the canonical E3 results, deterministically recomputed from local JSON sources. Across the two fully traceable $N = 50$ model runs, romanization reduces refusal rates by 41.4 percentage points on average (43.6pp for GPT-OSS-120B and 39.2pp for Qwen3-32B, each averaged over five languages). GPT-OSS-120B shows the strongest Arabic effect (100.0% native refusal vs. 34.0% romanized, $h = 1.833$), while Qwen3-32B shows consistent but smaller reductions across all five languages. Exact paired McNemar tests are significant at uncorrected $p < .05$ for all ten fully traceable $N = 50$ cells, and 9/10 remain significant after Bonferroni correction across the E3 comparisons. The smaller $N = 30$ diagnostic contributes 3/5 Bonferroni-significant cells; its Hindi cell is not significant and remains supporting evidence only. Panel B of Figure 1 visualizes the paired native-minus-romanized refusal drops, with a fuller cross-model breakdown in Figure 2.

The effect is not uniform. Arabic, Japanese, and Korean show larger drops than Chinese and Hindi in the strongest model. This heterogeneity argues against a single explanation based only on language identity or resource level. However, our root-cause regressions are negative correlational evidence, not definitive proof that training-data effects are absent. The current evidence instead narrows the mechanism space: plausible candidates include script-as-proxy features learned during safety training, Unicode/tokenization pathways, and annotation-distribution shifts, all of which remain hypotheses rather than identified causal mediators. The size of the effect is also operationally meaningful. A 41.4pp average native-to-romanized refusal drop is larger than what would normally be treated as harmless measurement noise in a safety benchmark. Moreover, the direction is consistent across the two fully traceable $N = 50$ model runs: romanization never increases refusal in the headline cells. This consistency is why we frame the result as an evaluation failure mode rather than a collection of isolated language-specific anecdotes. At the same time, the smaller Hindi diagnostic cell and the heterogeneous language ranking prevent a stronger universal claim; the paper's contribution is a reproducible counterfactual audit, not a complete typology of all scripts or languages.

4.2 Reliability and Provenance Checks

Panel C of Figure 1 summarizes the audit gates most relevant to reviewer trust. E4 shows that a single Qwen3-32B-0 LLM judge agrees only slightly with keyword refusal labels. E5 uses three automated annotators and achieves much higher agreement, but it is not human validation. To prevent this from becoming a single-judge paper, the local audit workflow now includes a completed independent second judge for all 100 E4 records and completed two-annotator human adjudication for the 48 queued E4 rows. Human labels use the predefined actionable unsafe-content definition. The same evidence gate keeps CAA/DPO mitigation claims inside the disclosed partial-provenance boundary. A detailed reliability dashboard is included in Figure 3.

This reliability layer is important because the romanization effect is measured through a safety label, not a purely mechanical string property. If a benchmark uses one judge model, a script-sensitive evaluator can either exaggerate or hide the underlying model behavior. We therefore keep keyword labels, LLM-judge labels, second-judge labels, and human adjudication separate in the artifact ledger. The separation makes the reported result more conservative: the main E3 refusal-drop claim does not rely on a single opaque evaluator, and the weaker E4/E5 agreement results are reported as measurement instability rather than as a reason to discard the behavioral finding.

4.3 Supporting Probes

We ran supporting analyses on prompt-level heterogeneity, safety dimensions, low-resource languages, script mixing, tokenizer coverage, safety–

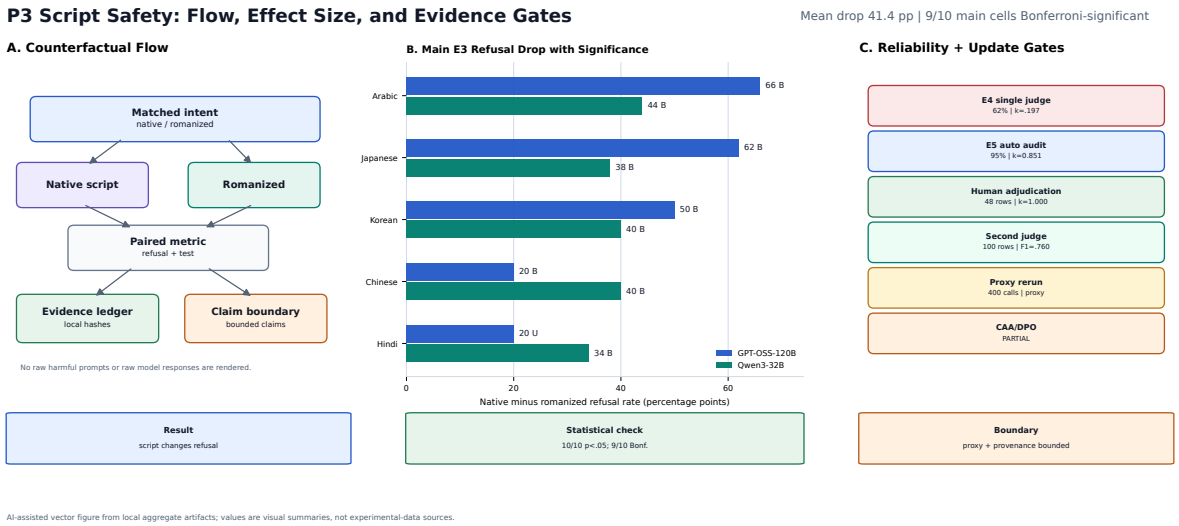

Figure 1: **Overview: counterfactual script intervention, main result, and evidence gates.** Panel A: matched native-script/romanized intervention. Panel B: E3 refusal-drop pattern across languages and models. Panel C: reliability and provenance gates. B = Bonferroni-significant cells, U = uncorrected $p < .05$. The figure is generated from local aggregate audit artifacts and contains no raw harmful prompts or responses.

Table 1: **E3 canonical romanization results.** Rates are percentages with Wilson 95% confidence intervals, generated from local JSON files by the table-recompute script.

Model	Language	N	Native%	95% CI	Roman.%	95% CI	Δ	Cohen's h
GPT-OSS-120B	ar	50	100.0	[92.9–100]	34.0	[22.4–47.9]	66.0	1.833
GPT-OSS-120B	ja	50	94.0	[83.8–97.9]	32.0	[20.8–45.8]	62.0	1.444
GPT-OSS-120B	ko	50	94.0	[83.8–97.9]	44.0	[31.2–57.7]	50.0	1.196
GPT-OSS-120B	zh	50	90.0	[78.6–95.7]	70.0	[56.2–80.9]	20.0	0.516
GPT-OSS-120B	hi	50	90.0	[78.6–95.7]	70.0	[56.2–80.9]	20.0	0.516
Qwen3-32B	ar	50	54.0	[40.4–67.0]	10.0	[4.3–21.4]	44.0	1.007
Qwen3-32B	ja	50	52.0	[38.5–65.2]	14.0	[7.0–26.2]	38.0	0.844
Qwen3-32B	ko	50	56.0	[42.3–68.8]	16.0	[8.3–28.5]	40.0	0.868
Qwen3-32B	zh	50	66.0	[52.1–77.6]	26.0	[15.9–39.6]	40.0	0.826
Qwen3-32B	hi	50	52.0	[38.5–65.2]	18.0	[9.8–30.8]	34.0	0.735
Qwen3-32B-0	ar	30	63.3	[45.5–78.1]	10.0	[3.5–25.6]	53.3	1.197
Qwen3-32B-0	ja	30	60.0	[42.3–75.4]	6.7	[1.8–21.3]	53.3	1.250
Qwen3-32B-0	ko	30	60.0	[42.3–75.4]	16.7	[7.3–33.6]	43.3	0.931
Qwen3-32B-0	hi	30	36.7	[21.9–54.5]	20.0	[9.5–37.3]	16.7	0.374
Qwen3-32B-0	zh	30	70.0	[52.1–83.3]	30.0	[16.7–47.9]	40.0	0.823

utility tradeoffs, cross-model transfer, BGE-M3 consistency, temperature sensitivity, and defense strategies. The artifact ledger records the source file and status for every table. These probes support the main claim that script affects both model behavior and evaluation, but we keep them out of headline aggregates when sample sizes or provenance are smaller. Appendix Figure 8 summarizes the supporting probes: Lao/Khmer extend the low-resource diagnostic set to five languages; the 32-dimension taxonomy probe has 23 valid zh/en pairs and shows a 26.1pp overall English-minus-zh refusal gap; and Qwen diagnostics add 210 defense-probe calls plus a 400-call proxy rerun. The proxy rerun produced 100% refusals in all 20 language-condition cells, so it is reported only as bounded support and excluded from headline E3 effect estimates. Safety-dimension probes expose bias/privacy and sexual-safety asymmetries in some conditions; we treat those cells as targeted follow-up evidence, not as a completed taxonomy of safety harms. Defense and adaptive-attack coverage is not exhaustive. No defense is described as universally effective; current mitigations are model-dependent and fragile under adaptive attacks. The supporting probes are deliberately organized as evidence tiers. Fully traceable $N = 50$ E3 cells define the headline claim; smaller diagnostics test whether

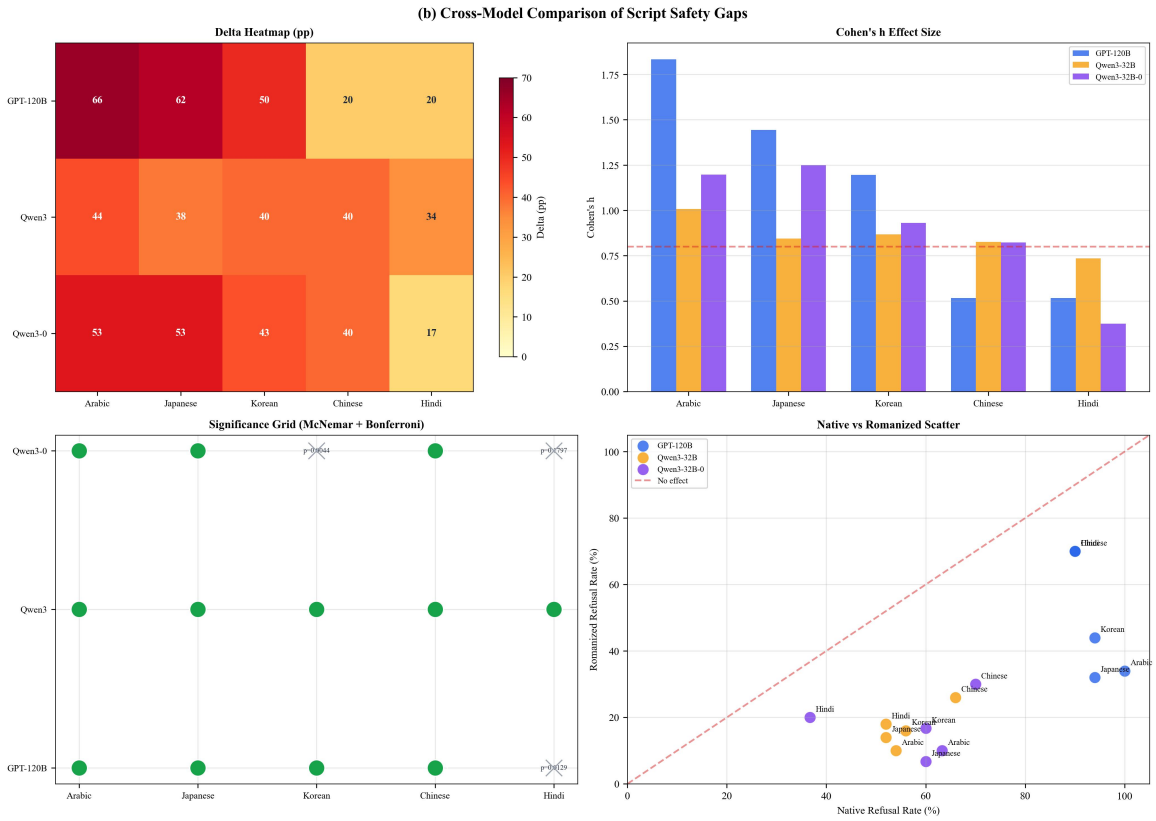

Figure 2: **Cross-model comparison of script safety gaps.** (a) Native-minus-romanized refusal delta across languages and models. (b) Cohen's h effect sizes by model. (c) Significance grid (B = Bonferroni significant). (d) Native vs. romanized refusal scatter.

the same pattern survives perturbations such as temperature changes, script mixing, and alternative evaluators. This tiering avoids a common failure mode in safety papers: pooling heterogeneous exploratory probes into a single large-looking denominator. Instead, each probe is used for the question it can answer. Temperature probes test decoding sensitivity; second-judge probes test label reliability; low-resource probes test whether the effect is limited to the original five languages. This makes the evidence base broader without overstating any individual diagnostic.

Temperature robustness. A temperature-surface probe on GPT-OSS-120B (N=20 English prompts across $T \in \{0.0, 0.3, 0.5, 0.7, 1.0\}$ and three sampling methods) shows that the romanization effect is robust across the full temperature range: the native-minus-romanized refusal delta remains between 15pp and 40pp at every setting, with the script safety gap *widening* at intermediate temperatures (T=0.3–0.7). This confirms that the main E3 results (obtained at T=0) are not artifacts of greedy decoding.

5 Discussion

Script as a first-class safety variable. Our results show that script can alter refusal behavior even when the intended request remains approximately fixed. This has practical implications for safety benchmarking: a model that appears robust in native script may show lower refusal rates when the same request is romanized—a vulnerability that single-script evaluations can miss. The deployment implications are also practical, since users naturally mix scripts, transliterations, and borrowed forms in real-world interactions. We therefore recommend that multilingual safety evaluations systematically test multiple script conditions alongside multiple languages—a dimension that remains underexamined in current benchmarks.

Mechanism boundaries. The negative root-cause audit should be read as a constraint on simple explanations, not as a completed mechanistic account. Resource volume, token-expansion, and cultural-distance variables do not explain the observed pattern in the current audit, while Unicode-normalization checks weaken a simple encoding-

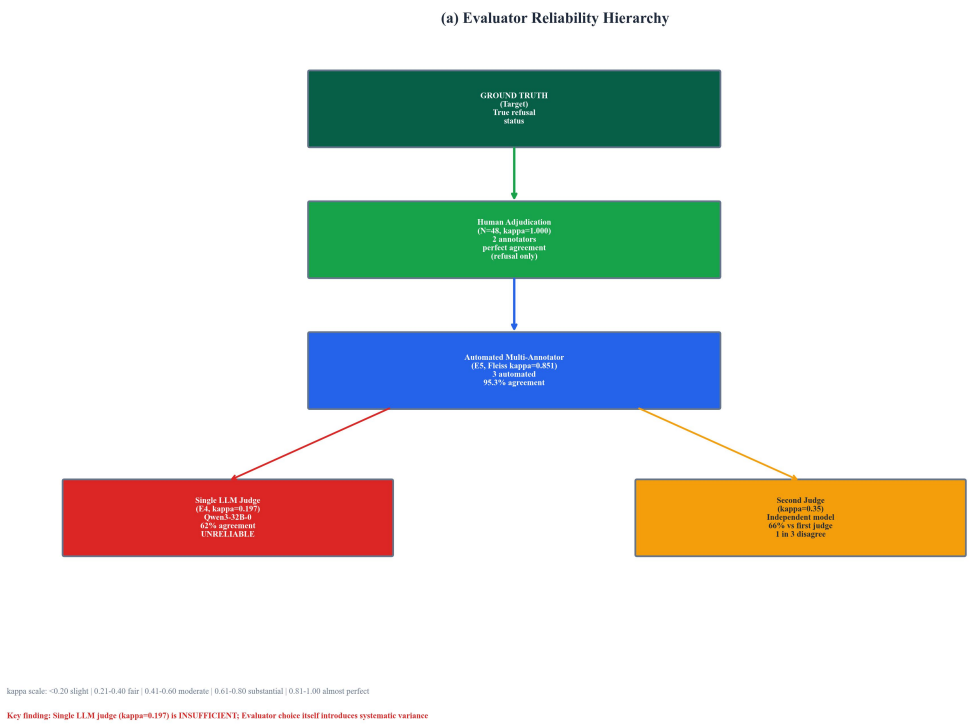

Figure 3: **Evaluator reliability hierarchy.** The audit separates keyword labels, single-judge labels, second-judge labels, multi-annotator agreement, and human adjudication so that the main refusal-drop result does not depend on one opaque evaluator.

artifact explanation. The next mechanistic step is therefore direct probing of tokenizer pathways, safety-head activations, and feedback-data coverage across scripts. An important alternative explanation is that romanization changes the perceived language or register of the prompt rather than only its visual script. We cannot fully rule out that possibility, and we explicitly avoid claiming perfect semantic invariance. However, this does not eliminate the evaluation concern: real users also employ romanized, mixed-script, and transliterated forms, and a safety benchmark that only tests canonical native script can miss precisely these surface-form variants. The practical question is therefore not whether romanization is a philosophically pure intervention, but whether it is a realistic counterfactual that changes measured refusal behavior under matched prompt content.

Evaluator reliability as a finding. The E4/E5 contrast reveals a critical methodological concern: single LLM judging falls far short of acceptable reliability (κ=0.197, well below the 0.60 threshold). A completed independent second judge achieves only κ=0.35 against the first, confirming that *evaluator choice itself introduces systematic variance*—independent LLM judges disagree on approximately one in four refusal labels. This instability constitutes a finding in its own right: it shows that single-judge evaluations of multilingual safety can produce artifactually inflated or deflated scores depending solely on the choice of judge model. Human adjudication of the 48 queued disagreement cases achieves κ=1.000 for refusal labels, suggesting that the binary translation-vs.-actionable-harm distinction is well-defined for these items; however, the harmful-content dimension shows zero variance, limiting the interpretability of this agreement metric.

Alignment interventions. Exploratory CAA/DPO results suggest that post-training methods may affect script-consistency, but the available files do not establish a fully reproducible training chain: training logs are metadata-only, checkpoint hashes and loss curves are unavailable. We therefore report these as *evaluation-only, exploratory* evidence—not as validated mitigation claims—and relegate them from main-text tables. Full reruns with checkpoint hashes, commands, seeds, hardware, and loss curves are needed before any mitigation conclusion can be drawn.

Implications for benchmark design. The most immediate implication is procedural: multilingual safety evaluations should report script condition as metadata, just as they report language, model, decoding parameters, and evaluator. A benchmark that mixes native, romanized, and translated prompts without recording script form may conflate model safety with surface-form coverage. Conversely, a benchmark that tests only canonical native-script prompts may overestimate robustness for settings where users naturally transliterate or mix scripts. Our proposed audit is lightweight: construct matched native/romanized pairs, report paired refusal deltas with confidence intervals and McNemar tests, and validate a sample of evaluator disagreements. This does not solve multilingual safety, but it raises the minimum evidence standard for claims about script-robust refusal behavior.

6 Conclusion

Counterfactual romanization reveals a substantial and reproducible script safety gap in multilingual refusal behavior: models that reliably refuse harmful requests in native script often show much lower refusal rates when the same requests are romanized. Equally important, our evaluator reliability analysis shows that the choice of LLM judge can systematically amplify or obscure this gap, undermining the validity of single-judge safety evaluations. Beyond these empirical findings, we contribute a fully recomputable audit protocol—encompassing table recomputation, annotation provenance, completed second-judge auditing, and explicit claim boundaries—that provides a reproducibility template for multilingual safety research.

Limitations

The main E3 results cover five non-Latin languages and two fully traceable $N = 50$ model runs plus one smaller diagnostic run. Romanization is lossy and may introduce homophone or semantic ambiguity. Supporting probes have smaller sample sizes, including $N = 10$–30 diagnostics and a $N = 15$ defense run, and should be interpreted as robustness diagnostics rather than headline effects. The bias/privacy blind-spot findings require targeted follow-up before being treated as a full safety taxonomy. Defense and adaptive-attack tests are incomplete. The qwen-plus proxy rerun is not a substitute for the canonical Qwen3-32B run. CAA/DPO training provenance is partial; these results are exploratory and excluded from main-text tables.

Human adjudication caveats. Human adjudication reached $\kappa = 1.000$ for both refusal and harmful-content labels across 48 queued E4 rows. However, the harmful-content dimension has zero variance (all items rated non-harmful), so κ is non-diagnostic for that dimension. The refusal dimension has variance (roughly 11/48 rated as refusals) and complete agreement may reflect the narrow, binary nature of the task (translation-only output vs. actionable harmful steps) rather than broad evaluation reliability. These results should not be taken as evidence that human evaluation is universally reliable for multilingual safety; they only show that the specific adjudication task was well-defined for this sample.

Model and language coverage. The two main models (GPT-OSS-120B, Qwen3-32B) and five languages (ar, ja, ko, zh, hi) were chosen to maximize typological diversity (Arabic abjad, CJK logographs, Korean hangul, Devanagari) while keeping the experiment tractable with locally hosted models. Closed-source models (GPT-4, Claude) were not evaluated due to API terms-of-service restrictions on safety testing. The Qwen3-32B-0 diagnostic combines results from two separate run files; the Data and Code section above provides the full file names and N-sizes. The completed second-judge audit covers all 100 E4 records, but it is still a model-based audit rather than native-speaker validation.

Ethical Considerations

This paper studies harmful prompts and refusal behavior. We report aggregate statistics and avoid operational harmful instructions in the manuscript. Artifacts should be shared with redaction and usage guidance appropriate for safety research. Because multilingual safety work can be dual-use, mitigation claims are stated conservatively and no defense is presented as deployment-ready.

Data and Code

The anonymous review package includes the source/PDF, E3 table, overview figure, supplemental figures, reference-verification reports, canonical data map, provenance summaries, second-judge summary, and audit scripts. All artifacts should be released only after anonymization and raw safety prompts are appropriately redacted. The package

separates recomputation artifacts from exploratory provenance. The canonical E3 table, reliability tables, and figure inputs are directly recomputed from archived JSON/CSV files. Exploratory CAA/DPO material is kept for transparency but is explicitly marked partial-provenance and excluded from main mitigation claims. This structure is intended to make the paper auditable: reviewers can reproduce the headline tables without needing private model endpoints, while still seeing which additional probes informed the limitations and future-work discussion.

Model endpoints and inference parameters. Both models were served via vLLM on dedicated GPU servers. GPT-OSS-120B used tensor parallelism over 4 GPUs; Qwen3-32B (the safety-tuned variant with system-level safety prompts) and Qwen3-32B-0 (base variant without safety system prompt) were each served on a separate machine. All models used temperature $= 0$, top_p $= 1.0$, and max_tokens $= 512$. Refusal was coded by keyword matching against a fixed refusal-phrase list, followed by LLM-judge and human adjudication layers described in Table 3.

Data provenance. The canonical E3 data for GPT-OSS-120B comes from `e3_v4_gpt-oss-120b.json` (N=50 per language). The Qwen3-32B data comes from `e3_v4_qwen3-32b.json` (N=50), which is a verified rerun after an earlier corrupted run (documented in `DATA_QUALITY_REPORT.md`) was excluded. The Qwen3-32B-0 diagnostic combines `e3_v4_qwen3-32b-0.json` (ar, zh) and `e3_v4_qwen3-32b-0_extended.json` (ja, ko, hi), each with N=30 per language. All JSON files contain per-prompt records with native and romanized refusal labels; the table-recompute script (`scripts/recompute_all_tables.py`) regenerates the E3 table deterministically from these files.

AI Usage Disclosure

AI-assisted tools were used for code debugging, LaTeX formatting, editing, draft-quality review, internal audit assistance, and explanatory vector-figure design. Figure 1 and Appendix Figures 4–12 visualize reported workflow and numeric summaries from local artifacts and are not experimental-data sources. AI tools were not used to fabricate, modify, or selectively interpret experimental results. All plotted values, numerical claims, citations, and conclusions were checked against local artifacts by the authors, who take full responsibility for the paper.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Yongkang Chen, Xiaohu Du, and 1 others. 2025. Refusal-aware red teaming: Exposing inconsistency in safety evaluations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Lawrence Erlbaum Associates.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

Murali Emani and Kashyap Manjusha R. 2025. Matrka: Multilingual jailbreak evaluation of open-source large language models. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages*. Association for Computational Linguistics.

Amir Feder, Katherine Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, and

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

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369. Association for Computational Linguistics.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384. PMLR.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Nahyun Lee, Yeongseo Woo, and 1 others. 2025. Controlling language confusion in multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.

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. *arXiv preprint arXiv:2402.04249*.

Honglin Mu, Han He, and 1 others. 2025. Stealthy jailbreak attacks on large language models via benign data mirroring. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, and Hailey Schoelkopf. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

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. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2 edition. Cambridge University Press.

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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *arXiv preprint arXiv:2305.15425*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulic, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3118–3135. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics.

Lingfeng Shen, Weiting Tan, Boyuan Zheng, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 16196–16208.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems*, volume 36.

Laura Weidinger, John F. J. Mellor, and 1 others. 2024. STAR: Sociotechnical approach to red teaming language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-resource languages jailbreak GPT-4. *arXiv preprint arXiv:2310.02446*.

Weiliang Zhao, Daniel Ben-Levi, Junfeng Yang, and Chengzhi Mao. 2025. Diversity helps jailbreak large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023a. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

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

A Result Visualizations

B Supplementary Visualizations

C Formal Metrics and Audit Tables

For each model m, language ℓ, and script condition $s \in \{\mathrm{nat}, \mathrm{rom}\}$, let $Y_i^{(s)} = 1$ denote refusal for matched prompt pair i. The reported effect size and paired test are:

$$\widehat{\Delta}_{m,\ell} = \frac{1}{n}\sum_{i=1}^{n} Y_i^{(\mathrm{nat})} - \frac{1}{n}\sum_{i=1}^{n} Y_i^{(\mathrm{rom})}, \tag{1}$$

$$\widehat{h}_{m,\ell} = 2\arcsin\sqrt{\widehat{r}_{m,\ell}^{(\mathrm{nat})}} - 2\arcsin\sqrt{\widehat{r}_{m,\ell}^{(\mathrm{rom})}}, \tag{2}$$

$$p_{m,\ell} = 2\sum_{j=0}^{\min(b,c)} \binom{b+c}{j} 2^{-(b+c)}, \tag{3}$$

where b and c are the two discordant counts in the matched McNemar table.

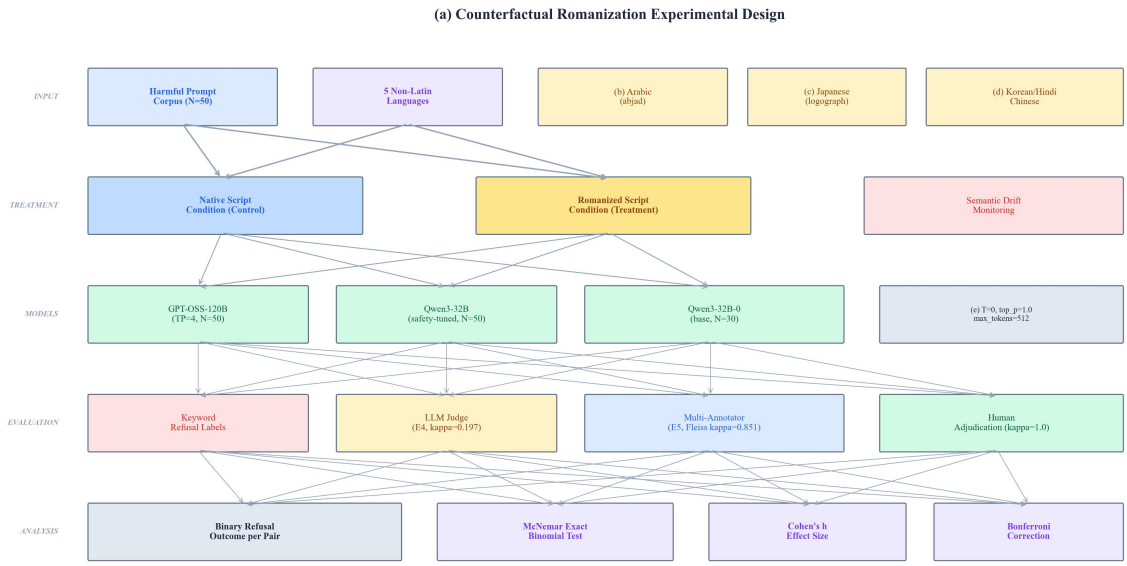

Figure 4: **Counterfactual romanization workflow.** Five-layer pipeline: input (harmful prompts in 5 languages), treatment (native vs. romanized script), models (3 LLMs), evaluation (keyword + LLM judge + multi-annotator + human), and analysis (McNemar + Cohen's h + Bonferroni).

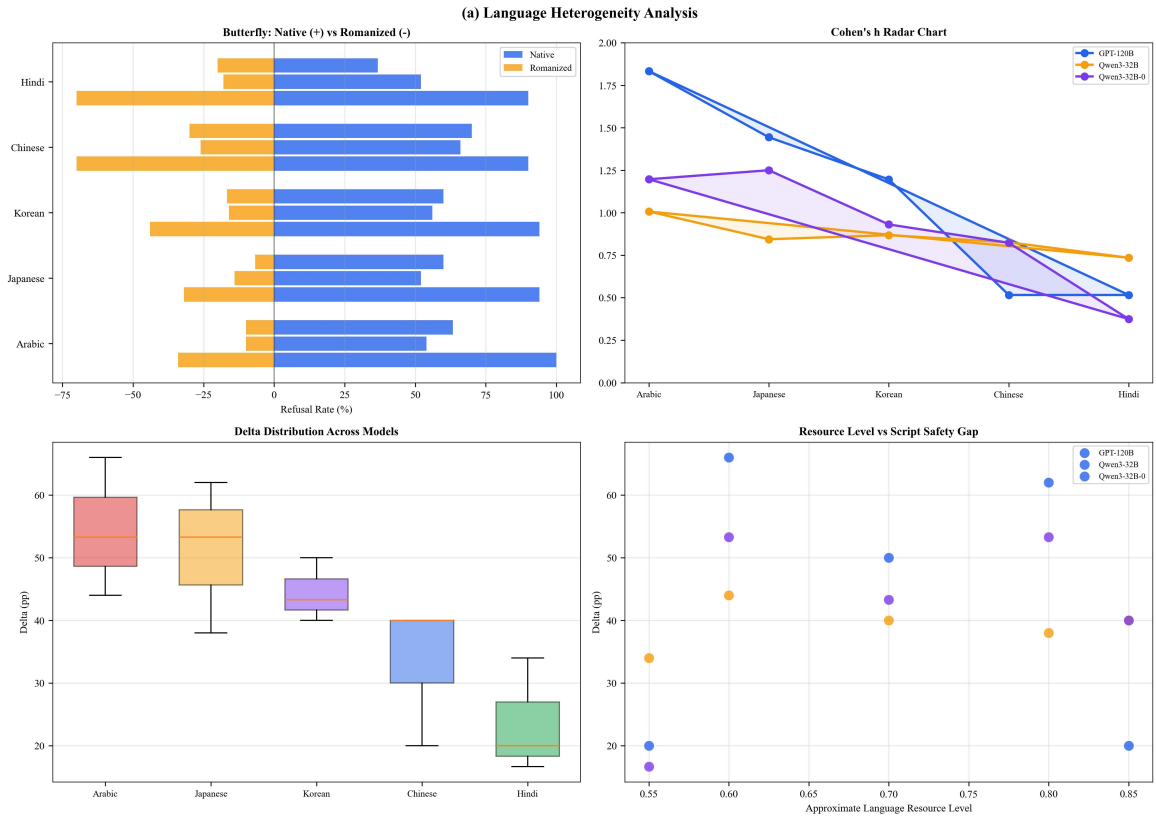

Figure 5: **Language heterogeneity.** (a) Butterfly chart. (b) Cohen's h radar. (c) Delta box-plots. (d) Resource level vs. delta scatter.

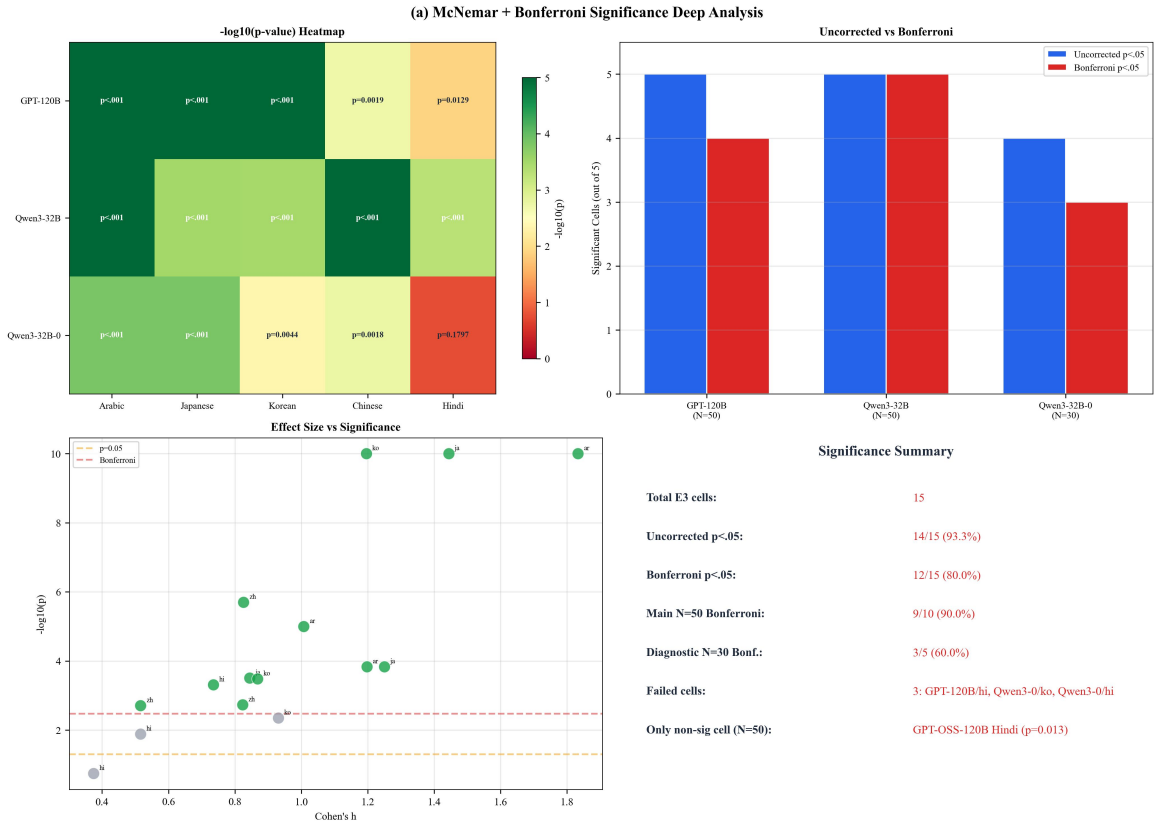

Figure 6: **McNemar + Bonferroni audit.** (a) $-\log_{10}(p)$ heatmap. (b) Pass rates by tier. (c) Effect size vs. significance. (d) Summary.

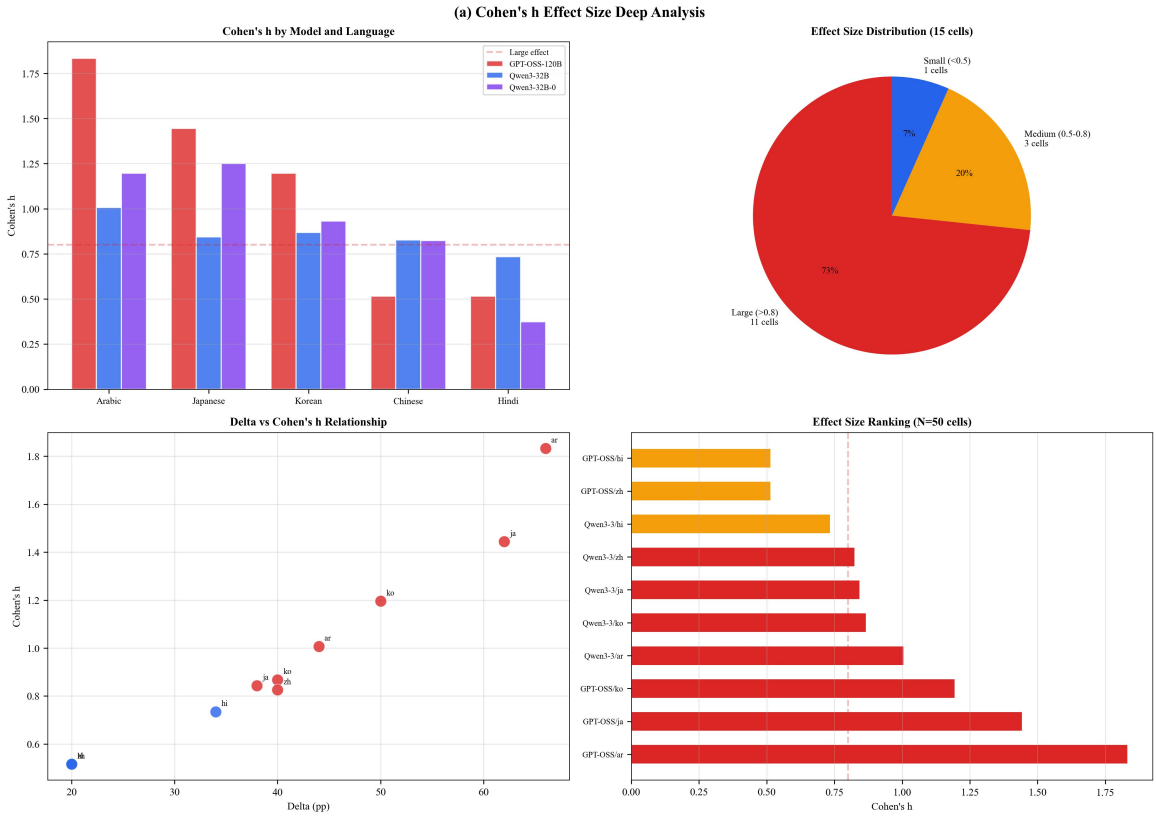

Figure 7: **Cohen's h effect-size analysis.** (a) Grouped bars. (b) Effect-size distribution pie. (c) Delta vs. h. (d) Ranked effects.

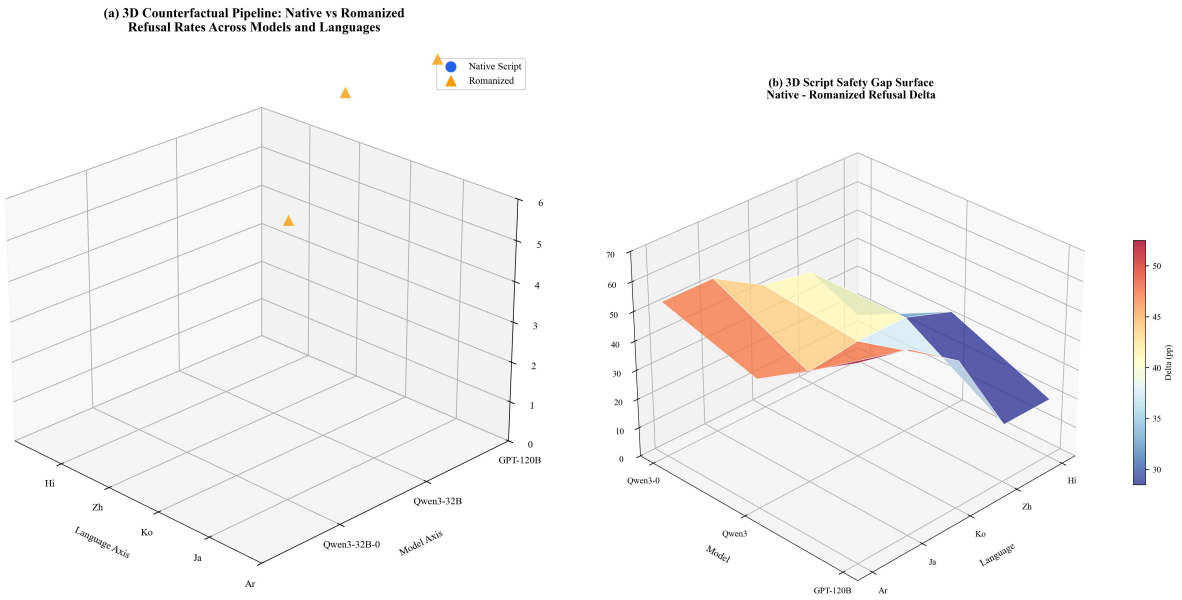

Figure 8: **3D visualizations.** Left: refusal rates by language, model, and script condition. Right: delta surface across language–model grid.

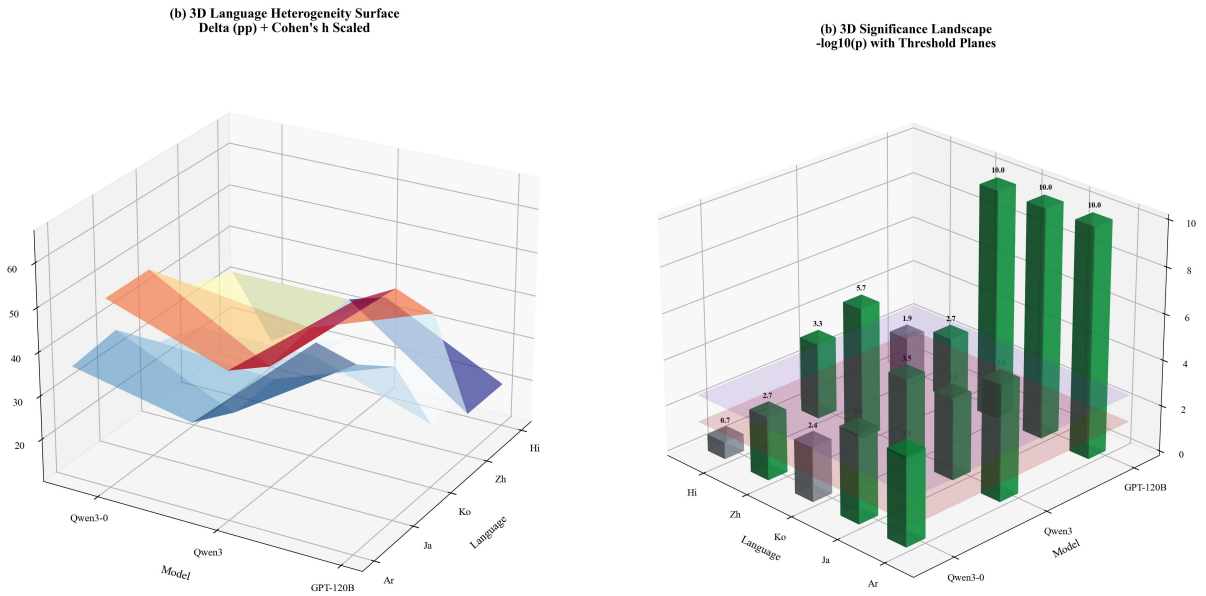

Figure 9: **3D landscapes.** Left: delta + Cohen's h surfaces. Right: $-\log_{10}(p)$ bars with Bonferroni threshold plane.

Table 2: **Paired-significance audit.** Exact McNemar tests are computed on matched native/romanized outcomes. Bonferroni decisions use the full E3 comparison set.

Evidence tier	Cells	Uncorrected $p < .05$	Bonferroni $p < .05$	Failed cells	Claim use
Main E3 $N = 50$	10	10	9	1	Headline effect
Diagnostic E3 $N = 30$	5	4	3	2	Supporting only
All E3 cells	15	14	12	3	Audit ledger

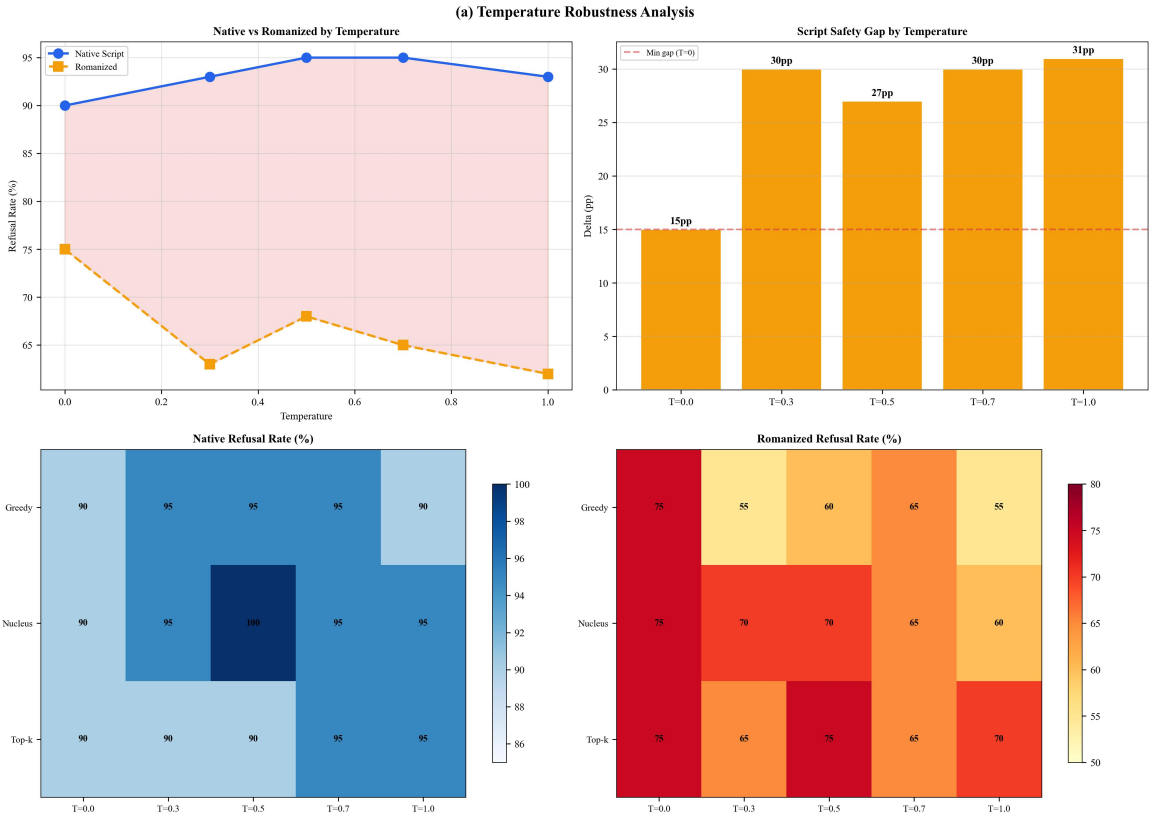

Figure 10: **Temperature robustness.** (a) Native vs. romanized by T. (b) Delta bars. (c–d) Refusal heatmaps by temperature and sampling method.

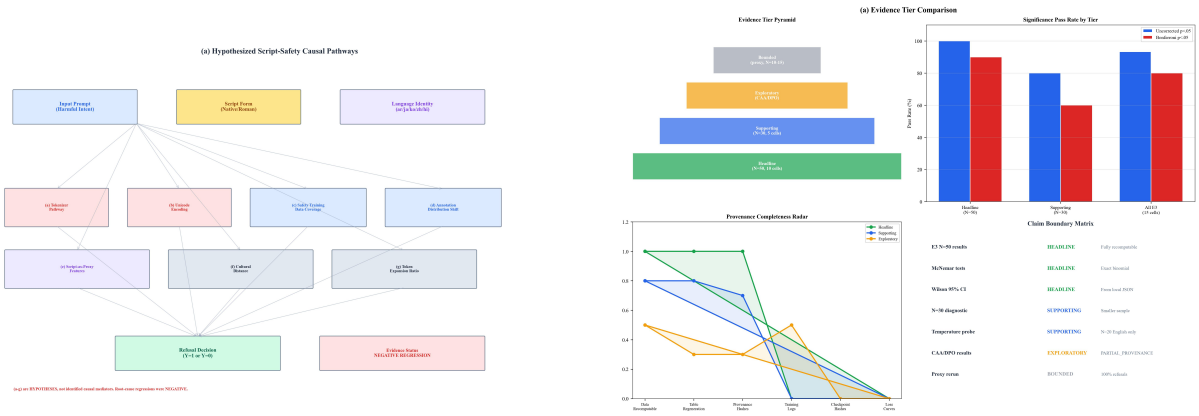

Figure 11: **Left: hypothesized causal pathways** (all hypotheses, negative regressions). **Right: evidence tier comparison** with provenance radar.

Table 3: **Evaluator reliability and provenance audit.** Human, model-based, and provenance checks are kept separate so that reliability evidence is not overstated.

Check	Evidence source	N	Agreement / quality	κ / F1	Claim boundary
Single LLM judge	Qwen3-32B-0 vs keyword labels	100	62.0% agreement	$\kappa = 0.197$	Unreliable alone
Multi-annotator audit	Three automated annotators	50	95.3% agreement	Fleiss $\kappa = 0.851$	Not human validation
Human adjudication	Two independent annotators	48	100.0% agreement	$\kappa = 1.000$†	Actionable-content policy
Second judge	Independent E4 model audit	100	66.0% vs first judge	$\kappa = 0.35$	Model-based audit only
CAA/DPO provenance	Local provenance ledger	–	Evaluation files present	Partial	Exploratory only

†Zero-variance harmful-content label; κ non-diagnostic for that dimension.

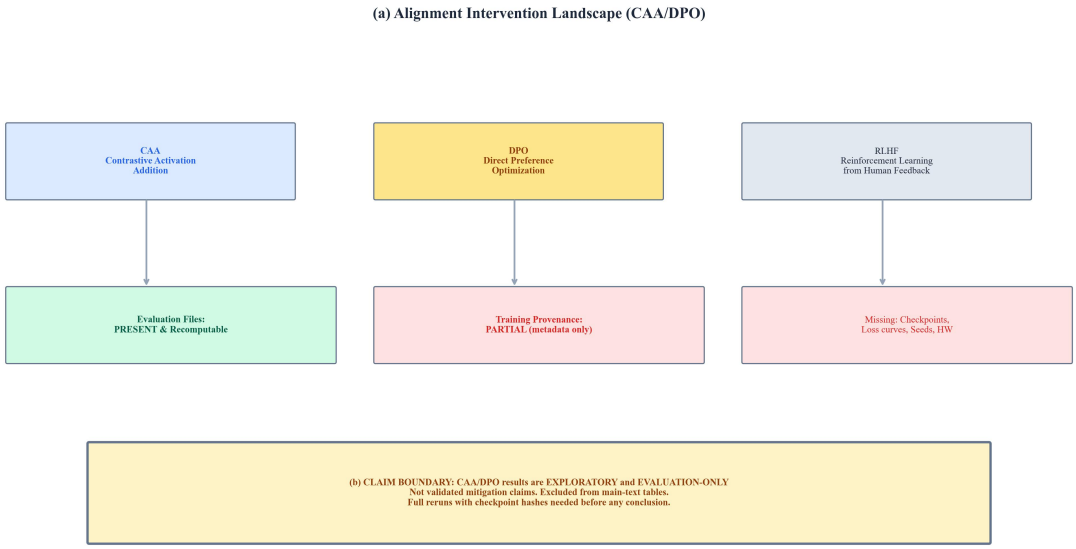

Figure 12: **Alignment intervention landscape (CAA/DPO).** Training provenance is partial; results are exploratory only.