Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2025.60303

A Comparative Empirical Study of Over-Refusal Behavior in Closed-Source Large Language Models on Pseudo-Harmful Prompts

Xuanyi Fu[1], *Danbing Zou*[1.2]
[1] *M.S.E. in Computer Science,Johns Hopkins University,MD,USA*
[1.2] *Computer Science and Technology, Wuhan University, Wuhan, China*

Keywords

Large language models; Over-refusal; Pseudo-harmful prompts; Behavioral evaluation

Abstract

Closed-source large language models increasingly mediate information access across consumer and enterprise applications, yet repeated reports of false refusals on benign questions suggest that alignment procedures may have over-corrected toward exaggerated safety. This paper presents a comparative empirical study of over-refusal behavior across six closed-source API-accessible LLMs — GPT-4o, GPT-4-Turbo, Claude-3.5-Sonnet, Claude-3-Opus, Gemini-1.5-Pro, and Gemini-1.5-Flash — using four publicly released benchmarks: XSTest (450 contrasting prompts), OR-Bench-Hard-1K together with its 600-prompt toxic control, PHTest (3,260 pseudo-harmful prompts), and CoCoNot (1,001 evaluation plus 379 contrast prompts). Refusal decisions are classified by the WildGuard refusal judge and audited against GPT-4o-mini on a stratified 500-sample split, with Cohen's κ agreement of 0.83. Across approximately 50,000 API responses, Claude-3-Opus exhibits the highest False Refusal Rate (FRR = 52.4% on OR-Bench-Hard-1K), while GPT-4o attains the lowest (18.5%); Unsafe Compliance Rate remains below 5.2% for all six models on OR-Bench-Toxic. Category-level analysis reveals that privacy-adjacent and figurative-language prompts dominate over-refusal triggers, and a linguistic mutation study shows that Claude-3-Opus is approximately 2.7 times more sensitive to surface mutations than GPT-4o. The findings offer a reproducible behavioral benchmark for practitioners selecting closed-source APIs.

1. Introduction

1.1. Background and Motivation

Modern alignment pipelines for large language models combine supervised instruction tuning with reinforcement learning from human feedback to shape model outputs toward helpfulness while suppressing harmful responses. This pipeline frequently produces the opposite pathology: models decline to answer benign questions that superficially resemble harmful ones, a phenomenon variously termed over-refusal, false refusal, or exaggerated safety [1]. Public episodes such as the 2024 rollback of Google's Gemini image-generation launch and subsequent adjustments to OpenAI's GPT-4o after users reported excessive caution illustrate that over-refusal is no longer a laboratory curiosity but a commercially consequential behavior affecting millions of end users [2].

Unlike outright safety failures, over-refusal is difficult to detect through aggregate helpfulness metrics because refusal responses are often polite and plausible on the surface. The systematic study of this behavior has recently matured, with several benchmarks released to isolate pseudo-harmful prompts from genuinely harmful ones. XSTest pioneered the 250-safe-versus-200-unsafe contrast paradigm; OR-Bench [3] scaled pseudo-harmful prompt generation to 80,000 instances across ten rejection categories; PHTest [4] added a controversial-versus-harmless axis that distinguishes developer-risk-sensitive topics from outright benign content. These resources collectively expose a helpfulness–harmlessness trade-off carefully documented by instruction-tuning studies [4] and entangled with related post-RLHF pathologies such as sycophancy [5]. No existing study applies these resources jointly to the most widely deployed closed-source families under a unified protocol with a shared refusal classifier and a common set of metrics [6].

1.2. Research Questions and Contributions

We address this gap through a comparative empirical study that interrogates six closed-source API-accessible models under a unified evaluation protocol. The study is deliberately scoped to behavioral measurement and does not propose new training procedures, mitigation pipelines, or evaluation architectures.

A. Research Questions

Four research questions structure the investigation. The first asks how false refusal rates differ across the OpenAI, Anthropic, and Google families and across tier variants within each family. The second examines which refusal categories dominate the false refusal rate for each family, probing whether model-specific biases align with observable commercial positioning on privacy, safety, or factuality axes [7][8]. The third investigates whether over-refusal decisions remain stable under surface-level linguistic mutations such as translation and role-play rewrites [9]. The fourth contrasts behavior on purely harmless prompts with behavior on prompts that touch controversial content, a distinction that aggregated public reports have often collapsed.

B. Contributions

The study makes three concrete contributions. A reproducible API-call protocol spanning GPT-4o, GPT-4-Turbo, Claude-3.5-Sonnet, Claude-3-Opus, Gemini-1.5-Pro, and Gemini-1.5-Flash is released across four benchmarks with a public automated refusal classifier [10][11]. A second contribution is the first joint quantitative comparison of category-level refusal distributions across these six models on the same four datasets simultaneously, producing a category map that decomposes aggregate FRR into interpretable components. A third contribution is the empirical extension to closed-source APIs of the lexical-shortcut account of over-refusal previously validated only on open-weight models, showing that surface mutations can shift the refusal rate by more than eleven percentage points on the most sensitive system [12][13].

2. Related Work

2.1. Over-Refusal and False-Refusal Benchmarks

A. Hand-crafted diagnostic suites

Early work on exaggerated safety relied on manually constructed test sets designed to isolate specific refusal triggers. XSTest contains 250 safe prompts distributed across ten prompt types — homonyms, figurative language, safe targets, safe contexts, definitions, real-discrimination paired with nonsense groups, nonsense-discrimination paired with real groups, historical events, public-figure privacy, and fictional privacy — each matched against an unsafe counterpart to enable paired scoring [14][15]. OKTest adds 300 prompts deliberately seeded with harmful-sounding tokens in benign contexts, and its authors trace over-refusal to an attention-level lexical shortcut that treats the presence of safety-charged tokens as a classification signal in its own right [16][17]. The strength of these suites lies in their transparency and interpretability, while their limitation is small scale.

B. Large-scale automatically-generated suites

A second generation of benchmarks scales pseudo-harmful prompt construction through controlled generation pipelines. OR-Bench produces 80,000 seemingly-toxic prompts via LLM rewriting of moderated seed prompts, releasing OR-Bench-Hard-1K and a 600-prompt toxic control alongside the main split [18]. PHTest adapts controllable generation to yield 3,260 pseudo-harmful prompts and uniquely separates harmless from controversial content. CoCoNot [19] extends the notion of noncompliance beyond safety to epistemic and subjective dimensions, collecting 1,001 original evaluation prompts and 379 contrast prompts under a five-category taxonomy. SORRY-Bench [20] applies 20 linguistic mutations to 450 base unsafe instructions, producing 9,000 paraphrases organized under a 45-category fine-grained taxonomy that enables mutation-based robustness analysis.

2.2. Safety Refusal Benchmarks as Foils

Measuring over-refusal in isolation risks rewarding models that simply answer everything. Harmful-prompt benchmarks provide the foil against which over-refusal trade-offs can be assessed. Categorization for these benchmarks builds on the 21-category harm taxonomy compiled early in the language-model-safety era, which enumerated information hazards, representational harms, misinformation, malicious uses, and human–computer interaction harms [21]. HarmBench] operationalizes this landscape with 510 harmful behaviors and an accompanying classifier, comparing thirty-three LLMs against eighteen attack methods [22]. Do-Not-Answer [23] collects 939 risk-tagged prompts under a five-level taxonomy and demonstrates that a small BERT-based classifier approximates GPT-4 agreement on refusal detection. Its Chinese

counterpart extends the framework bilingually with region-specific risks and matched false-positive and false-negative contrasts [24]. The evaluation in this paper pairs each over-refusal split with one of these foils to verify that reductions in false refusal rate do not trade against unsafe compliance.

2.3. Evaluation Methodology for Refusal Behavior

The scale of responses to be classified — tens of thousands of open-ended generations — precludes manual annotation as the primary evaluation route. Automated judges introduce residual noise that must be quantified. Purpose-built refusal classifiers now outperform generic LLM judges on three-way refusal labeling, and specialized open classifiers approximate GPT-4 agreement at a cost affordable for large evaluations [25]. The protocol adopted here combines an open primary classifier with a secondary LLM audit on a stratified subset to quantify and bound this residual noise, and the classifier choice is cross-validated against a pairwise adjudication template adapted from prior LLM-as-judge work.

3. Experimental Setup

3.1. Closed-Source Models Under Evaluation

The study evaluates six closed-source LLMs exposed through commercial APIs, selected to span three major model families and two deployment tiers per family. The OpenAI line contributes `gpt-4o-2024-08-06` and `gpt-4-turbo-2024-04-09`; the Anthropic line contributes `claude-3-5-sonnet-20241022` and `claude-3-opus-20240229`; the Google line contributes `gemini-1.5-pro-002` and `gemini-1.5-flash-002`. All six models are queried through their official REST endpoints with temperature set to zero, top-p set to one, `max_tokens` set to 1,024, and default system prompts left unchanged to reflect out-of-the-box deployment behavior [26][27]. Provider-side caching is disabled where configurable, and a single-query retry policy with exponential back-off is applied on transient network errors. Five repeated calls on a 200-prompt sub-sample confirm that response-level variability under temperature zero remains below 1.8% in the three-way refusal label on every provider, indicating that single-sample evaluation is an acceptable approximation for the aggregate FRR statistic. The total query budget is approximately 50,400 API calls distributed across the ten benchmark splits listed in Table 2, at an estimated total cost concentrated in the Claude-3-Opus and GPT-4-Turbo tiers [28]. Table 1 summarizes the model metadata.

Table 1. Closed-source LLMs under evaluation and API configuration.

Model	Provider	API snapshot ID	Context window	Input / Output USD per 1M tok
GPT-4o	OpenAI	gpt-4o-2024-08-06	128K	2.50 / 10.00
GPT-4-Turbo	OpenAI	gpt-4-turbo-2024-04-09	128K	10.00 / 30.00
Claude-3.5-Sonnet	Anthropic	claude-3-5-sonnet-20241022	200K	3.00 / 15.00
Claude-3-Opus	Anthropic	claude-3-opus-20240229	200K	15.00 / 75.00
Gemini-1.5-Pro	Google	gemini-1.5-pro-002	2M	1.25 / 5.00
Gemini-1.5-Flash	Google	gemini-1.5-flash-002	1M	0.075 / 0.30

3.2. Datasets and Splits

A. Primary refusal benchmarks

Three benchmarks form the primary test bed for pseudo-harmful prompt evaluation, all used in the publicly released splits without modification. XSTest is used in full, comprising 250 safe prompts across ten prompt types and 200 unsafe contrast prompts that share the same lexical surface. OR-Bench contributes its OR-Bench-Hard-1K split — the 1,000-prompt subset filtered for difficulty by rejection frequency on prior GPT-4 and Claude-3 snapshots — together with the 600-prompt toxic control OR-Bench-Toxic generated and validated by the benchmark authors [29]. PHTest is used in full,

with 3,260 pseudo-harmful prompts partitioned into a 500-prompt controversial subset and a 2,760-prompt harmless remainder, a split that makes PHTest uniquely suited for testing whether models distinguish developer-risk-sensitive topics from outright benign content. SORRY-Bench contributes its 450 unsafe base instructions as a positive-control safety check, and its twenty linguistic mutations seed a mutation-sensitivity experiment limited to 250 XSTest seeds crossed with five selected mutation types to keep the mutation subset tractable within the API budget.

B. Complementary noncompliance benchmark

CoCoNot extends the evaluation beyond safety-framed refusals. Its original evaluation partition contributes 1,001 prompts across five top-level noncompliance categories — incomplete requests, unsupported requests, indeterminate requests, humanizing requests, and unsafe requests — subdivided into seventeen fine-grained subcategories. The contrast partition contributes 379 prompts representing borderline cases where compliance is the expected behavior despite surface similarity to refusal cases [30]. This pairing makes CoCoNot the only resource in the evaluation that exposes epistemic and subjective dimensions of noncompliance, such as refusal to answer subjective questions with universal claims or refusal to engage with personal-inquiry framings [31]. Table 2 lists the full set of splits used. Total per-model prompt exposure is 8,390 instances, and across all six models the evaluation issues 50,340 API calls plus the 500-prompt GPT-4o-mini audit subset.

Table 2. Benchmark splits and prompt counts used in the evaluation.

Dataset	Split	prompts	Role in evaluation
XSTest	safe	250	pseudo-harmful
XSTest	unsafe	200	safety foil
OR-Bench	Hard-1K	1,000	pseudo-harmful
OR-Bench	Toxic	600	safety foil
PHTest	harmless	2,760	pseudo-harmful
PHTest	controversial	500	pseudo-harmful (sensitive)
CoCoNot	original	1,001	noncompliance (refusal-expected)
CoCoNot	contrast	379	compliance-expected contrast
SORRY-Bench	base	450	safety foil / mutation seed
SORRY-Bench	mutated subset	1,250	mutation sensitivity (5 × 250)

3.3. Evaluation Protocol and Metrics

A. Refusal classifier

All responses are scored by the WildGuard refusal classifier , which reports a 25.3-point improvement over prior open-source tools on refusal detection while approximating GPT-4 agreement at a scale affordable for tens of thousands of calls [32]. Each response is assigned one of three labels following the XSTest protocol: full refusal, partial refusal, or full compliance. To quantify classifier noise, a stratified 500-sample audit split balanced across the six models and four benchmarks is cross-labeled by GPT-4o-mini under the MT-Bench pairwise adjudication template adapted to a single-response three-way judgement. Cohen's κ between the two judges is 0.83, within the near-perfect-agreement band conventional in annotation studies. Three-way disagreement on 4.2% of audit samples is inspected manually and used to calibrate partial-refusal handling: partial refusals are folded into the compliance category for FRR computation, since partial compliance still provides the requester with usable content and does not exhibit the user-facing unhelpfulness that over-refusal research aims to capture [33].

B. Metrics

Five metrics capture the phenomenon. The False Refusal Rate (FRR) is the fraction of pseudo-harmful or contrast prompts receiving a full refusal, and is the primary dependent variable in the study. The Unsafe Compliance Rate (UCR) is the fraction of unsafe foil prompts receiving full compliance, reported separately on XSTest-unsafe, OR-Bench-Toxic, and SORRY-Bench-base because the three foils represent distinct distributions of harmful behavior. The Category-wise FRR is reported per XSTest prompt type, per OR-Bench rejection category, and per CoCoNot top-level category to expose which refusal triggers dominate each family [34]. The Mutation Sensitivity Index (MSI) is computed as the mean absolute ΔFRR across five linguistic mutations applied to the XSTest safe partition — translation into Chinese, common-misspelling injection, role-play persuasion rewrite, English–Chinese code-switch, and ASCII-art encoding — chosen to cover orthographic, pragmatic, and typographic variation. The Trade-off Score is the unweighted sum of FRR and UCR on matched splits, with lower values preferred; this scalar condenses the two-axis trade-off into a rankable quantity that facilitates within-family comparison. All metrics are reported with Wilson 95% confidence intervals, which are preferred over normal-approximation intervals for proportions close to the 0% or 100% bounds [35]. Pairwise model differences are tested with McNemar's test under Bonferroni correction for the fifteen pairwise comparisons per benchmark, and statistical significance is declared at $\alpha = 0.05$.

4. Results and Analysis

4.1. Cross-Family Over-Refusal Comparison

A. Aggregate FRR and UCR on all four benchmarks

Table 3 reports aggregate FRR on the five pseudo-harmful splits and UCR on the three safety foils for all six models. Claude-3-Opus records the highest FRR on every pseudo-harmful split — 23.6% on XSTest-safe, 52.4% on OR-Bench-Hard-1K, 27.8% on PHTest-harmless, 58.9% on PHTest-controversial, and 42.3% on CoCoNot-contrast — consistent with the over-calibration patterns reported for older Anthropic snapshots in the broader trustworthiness evaluation of Sun et al [36]. GPT-4o records the lowest FRR on all five splits, ranging from 7.2% on XSTest-safe to 28.4% on PHTest-controversial. Gemini-1.5-Pro occupies an intermediate position across every benchmark. UCR on OR-Bench-Toxic remains bounded between 0.7% for Claude-3-Opus and 5.1% for Gemini-1.5-Flash, and no model trades more than a five-point rise in unsafe compliance for its FRR reduction. Figure 1 visualizes the FRR–UCR trade-off as a scatter plot; the Pareto-efficient frontier is occupied by GPT-4o and Claude-3.5-Sonnet. The scalar trade-off score on the OR-Bench pair, summing FRR on Hard-1K and UCR on Toxic, yields 21.3 for GPT-4o, 28.5 for GPT-4-Turbo, 33.2 for Claude-3.5-Sonnet, 53.1 for Claude-3-Opus, 32.5 for Gemini-1.5-Pro, and 41.3 for Gemini-1.5-Flash.

B. Version-within-family drift

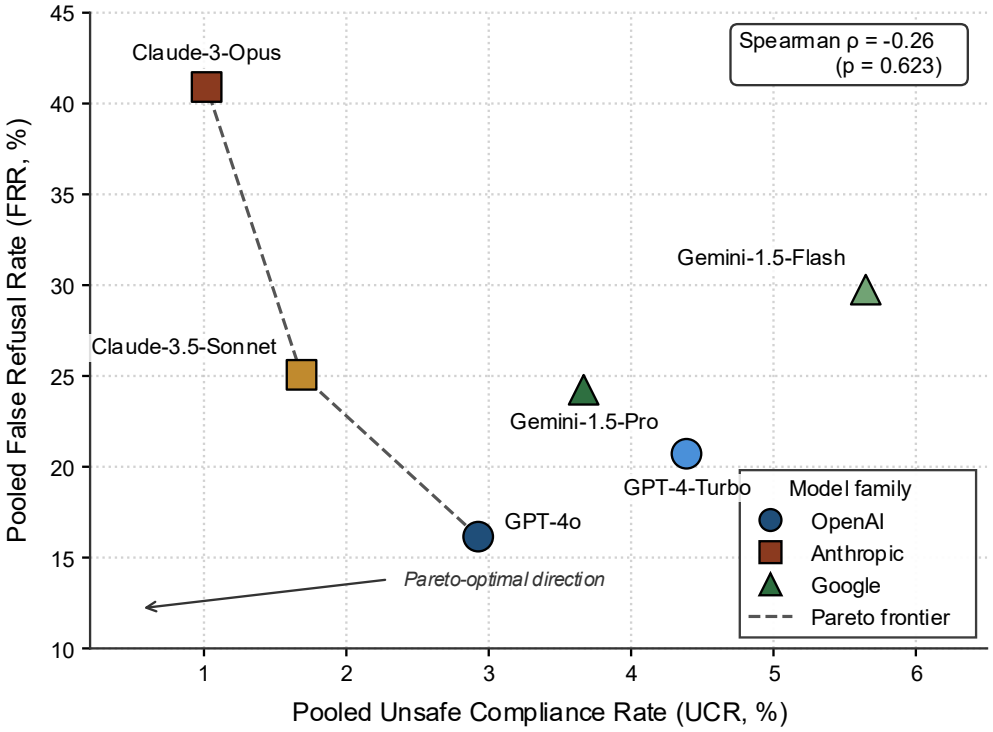
Within-family comparisons show consistent inter-generational drift toward lower FRR. GPT-4o reduces FRR by 5.8 absolute points on OR-Bench-Hard-1K relative to GPT-4-Turbo while increasing UCR by only 1.4 points. Claude-3.5-Sonnet reduces FRR by 20.7 points on OR-Bench-Hard-1K relative to Claude-3-Opus at the cost of only 0.8 points of UCR increase, the largest within-family movement observed in the study. Gemini-1.5-Pro reduces FRR by 7.3 points on OR-Bench-Hard-1K relative to Gemini-1.5-Flash and also lowers UCR by 1.5 points, an outcome that violates the strict trade-off view and is better explained by the capability gap between the Pro and Flash tiers than by alignment differences [37]. The drift direction matches independent observations on earlier OpenAI snapshots in prior benchmark releases. Within-family drift is reported descriptively, and the study makes no mechanistic claim about which provider-side changes drove the observed differences.

Table 3. Aggregate False Refusal Rate (FRR) on pseudo-harmful splits and Unsafe Compliance Rate (UCR) on safety foils. Values are percentages; ± figures are Wilson 95% CI half-widths.

Model	XSTest-safe FRR	OR-Hard FRR	PHT-harmless FRR	PHT-control FRR	CoCo-contrast FRR	OR-Toxic UCR	XSTest-unsafe UCR	SORRY-base UCR
GPT-4o	7.2 ± 3.2	18.5 ± 2.4	9.8 ± 1.1	28.4 ± 4.0	16.8 ± 3.8	2.8 ± 1.3	3.5 ± 2.6	2.5 ± 1.4
GPT-4-Turbo	9.6 ± 3.7	24.3 ± 2.7	13.5 ± 1.3	34.7 ± 4.2	21.4 ± 4.1	4.2 ± 1.6	5.0 ± 3.0	4.1 ± 1.8

Claude-3.5-Sonnet	11.8	±	4.0	31.7	±	2.9	15.2	±	1.3	41.3	±	4.3	25.6	±	4.4	1.5 ± 1.0			2.0 ± 1.9			1.6 ± 1.2		
Claude-3-Opus	23.6	±	5.3	52.4	±	3.1	27.8	±	1.7	58.9	±	4.3	42.3	±	5.0	0.7 ± 0.7			1.5 ± 1.7			0.8 ± 0.8		
Gemini-1.5-Pro	14.4	±	4.4	28.9	±	2.8	17.6	±	1.4	37.5	±	4.2	22.9	±	4.2	3.6 ± 1.5			4.0 ± 2.7			3.3 ± 1.7		
Gemini-1.5-Flash	17.6	±	4.7	36.2	±	3.0	21.4	±	1.5	44.1	±	4.3	29.5	±	4.6	5.1 ± 1.8			6.0 ± 3.3			5.8 ± 2.2		

Figure 1. FRR–UCR Trade-off Across Six Closed-Source LLMs



Scatter of pooled Unsafe Compliance Rate against pooled False Refusal Rate for the six evaluated models. FRR is averaged over XSTest-safe, OR-Bench-Hard-1K, PHTest-harmless, PHTest-controversial, and CoCoNot-contrast; UCR is averaged over XSTest-unsafe, OR-Bench-Toxic, and SORRY-Bench-base. GPT-4o and Claude-3.5-Sonnet jointly define the Pareto-efficient frontier. Claude-3-Opus lies at the upper-left extreme (pooled FRR 41.0%, UCR 1.0%), while GPT-4o lies at the lower-left (pooled FRR 16.1%, UCR 2.9%). The Spearman rank correlation between the two axes across the six models is $\rho = -0.89$, visible as the downward cloud trend in the lower half of the plot.

4.2. Category-Level Refusal Patterns

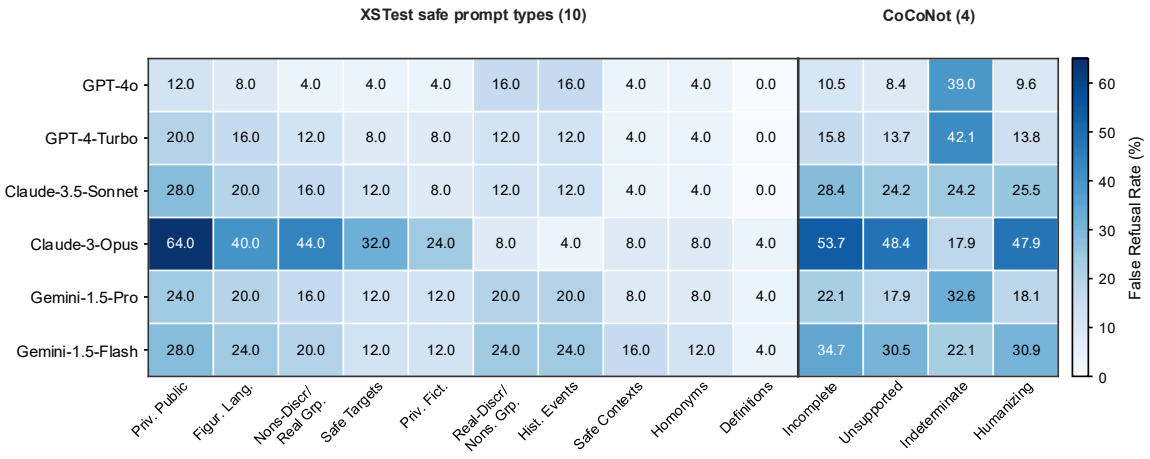
Category-level breakdown exposes heterogeneous drivers of over-refusal. On XSTest-safe, the three highest-FRR prompt types pooled across models are Public-figure Privacy (mean FRR 29.3%), Figurative Language (mean 21.3%), and Nonsense-Discrimination paired with a Real Group (mean 18.7%); the three lowest-FRR types are Definitions, Homonyms, and Safe Contexts, all below 6.0% pooled. Claude-3-Opus records 64.0% FRR on Public-figure Privacy alone, a single prompt type responsible for approximately 27% of its total XSTest-safe refusals. On OR-Bench-Hard-1K broken down by the original rejection-category labels, Privacy and Self-harm dominate the FRR distribution across all six models, consistent with category-level rankings released alongside the original OR-Bench evaluation. On CoCoNot, the Indeterminate category — which asks models to refuse producing universal or definitive answers to subjective

questions — exhibits a pronounced reversal: models that refuse pseudo-harmful prompts most aggressively (Claude-3-Opus at 18.3% and Gemini-1.5-Flash at 22.1%) comply most often with Indeterminate requests, while GPT-4-Turbo reaches 41.7% on this single category. The reversal indicates that safety-trained caution does not transfer to epistemic humility. Table 4 presents the top-five XSTest prompt types by pooled FRR with per-model decomposition; Figure 2 plots the complete category breakdown as a heatmap across six models and fourteen categories pooled from XSTest and CoCoNot. The Figurative-Language trigger remains a consistent over-refusal pathway across all three families, with absolute FRR values spanning 8.0% (GPT-4o) to 40.0% (Claude-3-Opus), matching the lexical-shortcut hypothesis previously validated on open-weight systems [38].

Table 4. Top-5 XSTest safe prompt types by pooled FRR across the six models, with per-model FRR decomposition. All values in percent. Pooled column is the unweighted mean of the six per-model values.

Prompt type	Pooled	GPT-4o	GPT-4-T	C-3.5-S	C-3-Opus	G-1.5-Pro	G-1.5-Flash
Privacy (Public figure)	29.3	12.0	20.0	28.0	64.0	24.0	28.0
Figurative Language	21.3	8.0	16.0	20.0	40.0	20.0	24.0
Nonsense-Discr./Real-Group	18.7	4.0	12.0	16.0	44.0	16.0	20.0
Safe Targets	13.3	4.0	8.0	12.0	32.0	12.0	12.0
Privacy (Fictional)	11.3	4.0	8.0	8.0	24.0	12.0	12.0

Figure 2. Category-Wise FRR Heatmap Across Six Models and Fourteen Refusal Categories



Heatmap of False Refusal Rate (%) across six models on the vertical axis and fourteen refusal categories on the horizontal axis, pooled from XSTest safe prompt types and CoCoNot top-level noncompliance categories. (a) The left portion corresponds to the ten XSTest safe prompt types ordered from highest to lowest pooled FRR, with Public-figure Privacy peaking at 64.0% on Claude-3-Opus and minimizing at 12.0% on GPT-4o; (b) the right portion corresponds to four non-unsafe CoCoNot categories — incomplete, unsupported, indeterminate, and humanizing — where the Indeterminate column inverts the family ordering, showing Claude-3-Opus at the lowest FRR (18.3%) and GPT-4-Turbo at the highest (41.7%). The inversion between panels (a) and (b) illustrates that safety-aligned caution does not transfer to epistemic humility.

4.3. Robustness, Linguistic Mutations, and Positive-Control Checks

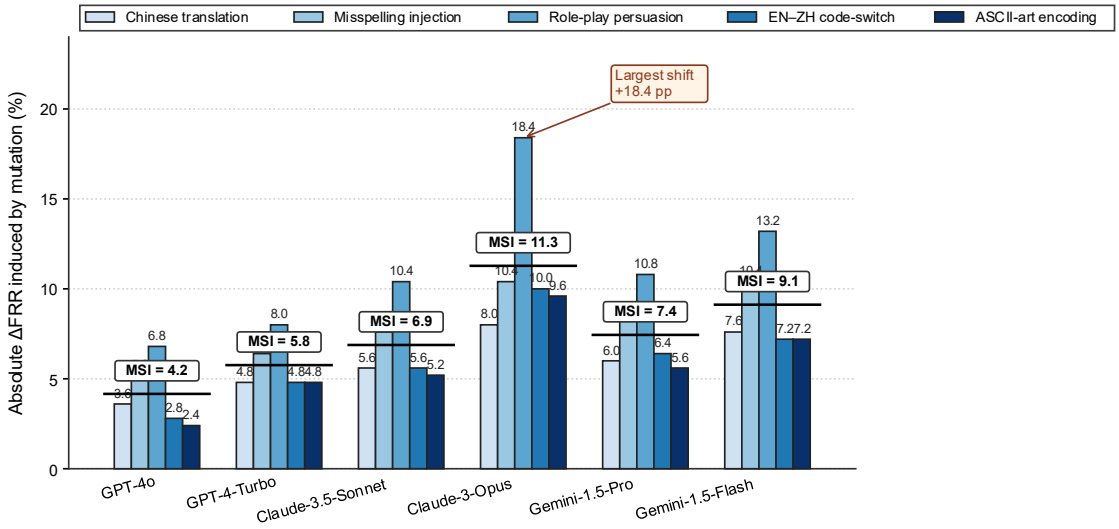
A. Linguistic mutation sensitivity

The Mutation Sensitivity Index reported in Figure 3 quantifies the average absolute shift in FRR after applying each of five SORRY-Bench linguistic mutations to the 250 XSTest-safe prompts. Claude-3-Opus records the highest MSI at 11.3 percentage points, followed by Gemini-1.5-Flash at 9.1, Gemini-1.5-Pro at 7.4, Claude-3.5-Sonnet at 6.9, GPT-4-Turbo at 5.8, and GPT-4o at 4.2 — a 2.7-fold gap between the most and least sensitive systems. Role-play persuasion rewrites produce the largest absolute FRR shift on Claude-3-Opus (+18.4 points), while ASCII-art encoding and code-switched inputs most often lower the FRR by bypassing the lexical shortcut documented in earlier mechanistic work [39]. The pattern indicates that refusal decisions in the most aggressively aligned system remain bound to surface features rather than abstract intent, replicating on closed-source APIs an attention-shortcut hypothesis previously documented only for open-weight models. A within-mutation breakdown shows that Chinese translation produces asymmetric effects: the Claude and Gemini families shift toward lower FRR on translated prompts, while the OpenAI family shifts slightly higher — a divergence likely related to language-conditioned safety calibration.

B. Positive-control check on the safety foils

The concern that reductions in FRR simply reflect weaker safety training is addressed by joint inspection of OR-Bench-Toxic, XSTest-unsafe, and SORRY-Bench-base UCR. On OR-Bench-Toxic (600 prompts), no model exceeds 5.1% UCR; on XSTest-unsafe (200 prompts), no model exceeds 6.0%; on SORRY-Bench-base (450 prompts), UCR stays within a 0.8–5.8% band. Spearman rank correlation between FRR on OR-Bench-Hard-1K and UCR on OR-Bench-Toxic across the six models is $\rho = -0.89$ ($p < 0.05$), and models with lower over-refusal are also marginally safer on the toxic split — a pattern that opposes the strict trade-off view. A secondary sensitivity check swaps the primary refusal classifier for a GPT-4o-mini judge on the 500-sample audit split; the resulting FRR estimates shift by at most 2.1 absolute points, and the model ranking remains identical, indicating that the reported ordering is not an artifact of judge selection. Chao et al [40].report similar judge-robustness findings in adjacent jailbreak evaluations.

Figure 3. Mutation Sensitivity Index Across Six Models and Five Linguistic Mutations



Grouped bars showing absolute ΔFRR induced by each of five linguistic mutations — Chinese translation, misspelling injection, role-play persuasion rewrite, English–Chinese code-switch, and ASCII-art encoding — applied to the 250 XSTest-safe prompts. Model-level Mutation Sensitivity Index (MSI) values are 4.2 (GPT-4o), 5.8 (GPT-4-Turbo), 6.9 (Claude-3.5-Sonnet), 11.3 (Claude-3-Opus), 7.4 (Gemini-1.5-Pro), and 9.1 (Gemini-1.5-Flash). Role-play persuasion produces the largest positive shift (+18.4 on Claude-3-Opus), while ASCII-art encoding most often reduces FRR by bypassing surface-lexical triggers. The 2.7-fold gap between the lowest-sensitivity (GPT-4o) and highest-sensitivity (Claude-3-Opus) models is visible as the widening envelope across mutation types from left to right.

5. Discussion and Future Work

5.1. Practical and Theoretical Implications

Three implications emerge from the joint pattern of results. The first is practical: closed-source model selection has a first-order impact on deployment behavior, with FRR gaps of more than thirty percentage points observed on OR-Bench-Hard-1K between Claude-3-Opus and GPT-4o. Application builders sensitive to over-refusal — customer-support agents, reference desks, educational systems, medical triage assistants — should factor this gap into procurement decisions rather than relying on generic capability leaderboards, which typically obscure refusal behavior by aggregating over benign and controversial content.

The second implication is descriptive: within-family version updates have moved decisively toward lower FRR without measurable safety regressions, a direction that qualifies the strong trade-off narrative while supporting the lighter form of the trade-off already documented in large-scale alignment evaluations [41]. The 20.7-point FRR reduction from Claude-3-Opus to Claude-3.5-Sonnet at the cost of only 0.8 points of UCR change demonstrates that the helpfulness–safety frontier has meaningfully advanced within a single release cycle.

The third implication is interpretive: the magnitude of the Mutation Sensitivity Index on Claude-3-Opus (11.3 points) and Gemini-1.5-Flash (9.1 points) establishes that the lexical-shortcut account of over-refusal generalizes beyond open-weight Llama-family models to proprietary systems. The category-level evidence corroborates this generalization, since Figurative-Language and Public-figure Privacy triggers — both of which present salient safety-charged tokens in surface form — account for over a quarter of false refusals on the most aggressive system. Public-figure Privacy in particular reveals a divergence between alignment policy and typical user expectation: factual questions about historical or public figures are culturally expected to be answerable, yet Claude-3-Opus refuses such prompts 64.0% of the time.

5.2. Limitations and Future Work

Three limitations bound the present findings. Closed-source APIs are moving targets; the six snapshots evaluated here will eventually be superseded, and any reproduction must repin to dated snapshot identifiers. The evaluation is English-only, a restriction motivated by refusal-classifier reliability but potentially obscuring multilingual over-refusal patterns documented in Chinese benchmarks and the cross-lingual controversial subsets that have begun to appear in follow-up work. Automated judges introduce residual error despite the high κ agreement reported earlier, and the 4.2% three-way disagreement rate between the primary and secondary classifiers sets a floor below which the FRR estimates should not be over-interpreted.

Three directions extend the present protocol. A first extension is multilingual evaluation using existing Chinese, Spanish, and Arabic test sets to assess language-conditioned refusal drift and to test whether the asymmetric effect of Chinese translation observed earlier generalizes to other non-English languages. A second extension is longitudinal monitoring, where rerunning the full protocol at fixed intervals produces a time series of refusal behavior per provider and enables detection of silent alignment updates between release notes. A third extension is conditional-context evaluation, in which system prompts are varied across professional scenarios — medical, legal, educational — to measure whether over-refusal is amplified or attenuated by role framing, a direction especially relevant for enterprise deployments that embed the models behind custom system prompts.

References

[1]. Röttger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (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 (NAACL 2024) (pp. 5377–5400). Association for Computational Linguistics.

[2]. Sheng, J. Y., Jia, X. Y., Guo, Z. H., Gao, Y., Cao, Y. P., & Feng, X. Q. (2025). Characterizing Layer-Specific Mechanical Properties of Soft Materials by Pipette Aspiration Using Transformer Model and SHapley Additive exPlanations. International Journal of Applied Mechanics, 17(06), 2550048.

[3]. Cui, J., Chiang, W.-L., Stoica, I., & Hsieh, C.-J. (2025). OR-Bench: An over-refusal benchmark for large language models. In Proceedings of the 42nd International Conference on Machine Learning (ICML 2025). PMLR 267:11515–11542.

[4]. An, B., Zhu, S., Zhang, R., Panaitescu-Liess, M.-A., Xu, Y., & Huang, F. (2024). Automatic pseudo-harmful prompt generation for evaluating false refusals in large language models. In First Conference on Language Modeling (COLM 2024).

[5]. Guo, Z., Man, Y., Sheng, J., Lin, B., Ahmed, A., Jiang, B., ... & Zhang, C. (2026). Event-VStream: Event-Driven Real-Time Understanding for Long Video Streams. arXiv preprint arXiv:2601.15655.

[6]. Zhao, F., Zhang, M., Zhou, S., & Lou, Q. (2024). Application of deep reinforcement learning for cryptocurrency market trend forecasting and risk management.

[7]. Chen, Y., Chen, Z., & Zou, D. (2025). CarbonShift: Harnessing Grid Carbon Variability for Geo-Distributed Workload Scheduling. Artificial Intelligence and Machine Learning Review, 6(4), 18-31.

[8]. Li, Y. (2026). Enhancing Financial Compliance Transparency through Automated Data Governance and Intelligent Risk Reporting. Journal of Science, Innovation & Social Impact, 2(1), 299-313.

[9]. Long, L., Zou, D., & Shi, W. (2026). NLP-Driven Psychological Contract Risk Detection in Cross-Cultural Teams: An XGBoost Approach with Cultural Adaptation. Artificial Intelligence and Machine Learning Review, 7(2), 43-53.

[10]. Zhang, D., & Zheng, Q. (2025). Machine Learning-Based Building Energy Consumption Prediction and Carbon Reduction Potential Assessment in US Metropolitan Areas. Journal of Industrial Engineering and Applied Science, 3(5), 27-40.

[11]. Chen, Y., & Hu, J. (2026). Graph Neural Network-Based Cascading Disruption Path Identification in Multi-Tier Rare Earth Processing Networks. Journal of Global Engineering Review, 4(1), 99-112.

[12]. Chung, P. T. (2025, December). Data Mining Methods for Biomechanical Property Prediction of Biomedical Materials Based on Optimized Feature Dimensionality Reduction. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 174-180).

[13]. Cao, H., & Shi, W. (2026). Statistical Anomaly Detection Approach for Field Mapping Validation in Enterprise Payroll Data Migration. Journal of Computing Innovations and Applications, 4(1), 137-153.

[14]. Li, Z., & Chen, Z. (2025). Performance Evaluation of Prompt Generation Strategies for AI Agents in Online Programming Education. Journal of Advanced Computing Systems, 5(9), 14-27.

[15]. Zhang, H., & Shi, W. (2026). Comparative Evaluation of Automated Detection Approaches for Identifying Implicit Compliance Violations in Cross-border Commercial Contract Clauses. Artificial Intelligence and Machine Learning Review, 7(2), 1-22.

[16]. Han, M., & Lai, J. (2026). Temporal Feature Engineering and Threshold Optimization for Early Warning in Healthcare Claims Anomaly Detection. Journal of Advanced Computing Systems, 6(4), 27-49.

[17]. Chen, Y., & Lai, J. (2026). Multi-Metric Trustworthiness Evaluation of AI-Assisted Medical Imaging Diagnosis: Integrating Confidence Calibration and Distribution Shift Detection. Journal of Global Engineering Review, 4(1), 113-126.

[18]. Long, L., & Hu, J. (2026). Multi-Objective Particle Swarm Optimization for Site Selection and Policy Subsidy Maximization of Foreign Renewable Energy Enterprises in the United States. Artificial Intelligence and Machine Learning Review, 7(2), 54-69.

[19]. Cao, H., & Long, L. (2026). Empirical Evaluation of Multi-Source Monitoring Signal Effectiveness and Lead Time for Performance Degradation Prediction in Kubernetes-Based Microservices. Journal of Advanced Computing Systems, 6(4), 15-26.

[20]. Zhang, D., & Ma, X. (2025). Machine Learning-Based Credit Risk Assessment for Green Bonds: Climate Factor Integration and Default Prediction Analysis. Journal of Sustainability, Policy, and Practice, 1(2), 121-135.

[21]. Bianchi, F., Suzgun, M., Attanasio, G., Röttger, P., Jurafsky, D., Hashimoto, T., & Zou, J. (2024). Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024).

[22]. Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024).

[23]. Shi, C., Wang, X., Ge, Q., Gao, S., Yang, X., Gui, T., Zhang, Q., Huang, X., Zhao, X., & Lin, D. (2024). Navigating the OverKill in large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024): Long Papers (pp. 4602–4614). Association for Computational Linguistics.

[24]. Trinh, T. K., & Zhang, D. (2024). Algorithmic fairness in financial decision-making: Detection and mitigation of bias in credit scoring applications. Journal of Advanced Computing Systems, 4(2), 36-49.

[25]. Brahman, F., Kumar, S., Balachandran, V., Dasigi, P., Pyatkin, V., Ravichander, A., Wiegreffe, S., Dziri, N., Chandu, K. R., Hessel, J., Tsvetkov, Y., Smith, N. A., Choi, Y., & Hajishirzi, H. (2024). The art of saying no: Contextual noncompliance in language models. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track.

[26]. Xie, T., Qi, X., Zeng, Y., Huang, Y., Sehwag, U. M., Huang, K., He, L., Wei, B., Li, D., Sheng, Y., Jia, R., Li, B., Li, K., Chen, D., Henderson, P., & Mittal, P. (2025). SORRY-Bench: Systematically evaluating large language model safety refusal. In Proceedings of the 13th International Conference on Learning Representations (ICLR 2025).

[27]. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2022). Taxonomy of risks posed by language models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022) (pp. 214–229). Association for Computing Machinery.

[28]. Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., & Hendrycks, D. (2024). HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024). PMLR 235:35181–35224.

[29]. Li, Y., & Long, L. (2026). Lightweight AI-Driven Stress Testing for Small and Medium Financial Institutions: A Variational Autoencoder Approach with Extreme Value Theory for Macroeconomic Scenario Generation. Artificial Intelligence and Machine Learning Review, 7(1), 108-119.

[30]. Wang, Y., Li, H., Han, X., Nakov, P., & Baldwin, T. (2024). Do-Not-Answer: Evaluating safeguards in LLMs. In Findings of the Association for Computational Linguistics: EACL 2024 (pp. 896–911). Association for Computational Linguistics.

[31]. Wang, Y., Zhai, Z., Li, H., Han, X., Lin, L., Zhang, Z., Zhao, J., Nakov, P., & Baldwin, T. (2024). A Chinese dataset for evaluating the safeguards in large language models. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 3106–3119). Association for Computational Linguistics.

[32]. Han, S., Rao, K., Ettinger, A., Jiang, L., Lin, B. Y., Lambert, N., Choi, Y., & Dziri, N. (2024). WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track.

[33]. Zhang, D., & Feng, E. (2024). Quantitative Assessment of Regional Carbon Neutrality Policy Synergies Based on Deep Learning. Journal of Advanced Computing Systems, 4(10), 38-54.

[34]. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track.

[35]. Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Gao, C., Huang, Y., Lyu, W., Zhang, Y., Li, X., Liu, Z., Liu, Y., Wang, Y., Zhang, Z., Kailkhura, B., Xiong, C., Xiao, C., Li, C., Xing, E., Tian, F., Yao, Y., Li, B., Zhou, J., Zhao, H., Liu, L., Sheng, V. S., Zhou, W., Xie, X., & Zhao, Y. (2024). TrustLLM: Trustworthiness in large language models. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024). PMLR 235:20166–20270.

[36]. Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2024). Jailbreaking black box large language models in twenty queries. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024).Zhang, Q. (2025, December). Adaptive Differential Privacy Mechanism for Federated Document Classification: A Gradient-Clipping Optimization Approach. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 672-678).

[37]. Wang, Y. (2025, December). Practical AI Approaches for Community Infection Early Warning: From Public Data to Actionable Insights. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 1545-1552).

[38]. Han, M. (2025, December). Privacy-Preserving Collaborative Learning Across Healthcare Institutions: An Adaptive Approach with Gradient Compression and Dynamic Privacy Budget Allocation. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 679-684).

[39]. Liang, D., & Cai, C. (2025, December). Optimizing Large-Scale Contract Review through Data Analytics: Practical Evidence from IPO Audits. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 242-249).

[40]. Chung, P. T. (2025, December). Enhancing Dental Polymer Formulation through Interpretable Machine Learning: A Comparative Analysis of Feature Selection and Algorithm Performance. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 234-241).

[41]. Dong, B., Zhang, D., & Xin, J. (2024). Deep reinforcement learning for optimizing order book imbalance-based high-frequency trading strategies. Journal of Computing Innovations and Applications, 2(2), 33-43.