Evaluating generative AI safety on real-world harm

A focused classification study using the RealHarm dataset

Matthew Clark[1], Lucía Martín[1], Javier Moreno[1]

[1] University of Gibraltar

Abstract

Safety benchmarks for large language models often rely on synthetic prompts, adversarial test cases, or narrowly defined toxicity tasks, which can leave a gap between benchmark performance and the kinds of failures that appear in deployed systems. This study evaluates whether a general-purpose large language model can distinguish safe from unsafe AI assistant interactions using RealHarm, a benchmark derived from publicly reported real-world failures and paired with human-reviewed safe rewrites. The full sample contained 136 observations, consisting of 68 unsafe interactions and 68 matched safe counterparts across 11 primary harm taxonomies. All records were anonymized, randomly shuffled, and classified in a blind single-pass procedure using the same binary SAFE/UNSAFE instruction. The evaluated model achieved 91.2% accuracy (95% Wilson CI 85.2% to 94.9%), 100.0% precision, 82.3% recall, 90.3% F1, and 100.0% specificity. All 12 classification errors were false negatives; no safe interaction was incorrectly flagged as unsafe. Errors were concentrated in misinformation and unsettling multi-turn interactions, with additional misses involving privacy, vulnerable users, interaction breakdowns, and brand-damaging conduct. Qualitative inspection showed that many missed cases were superficially fluent and helpful, depended on external factual knowledge, emerged late in long conversations, or involved subtle role and contextual boundaries. The findings suggest that aggregate safety accuracy can conceal a consequential asymmetry between avoiding over-restriction and detecting subtle harm. For governance and deployment, the results support category-aware monitoring, explicit fact-verification mechanisms, and multi-stage review for longer or context-sensitive interactions.

Keywords: generative AI, large language models, AI safety, RealHarm, safety classification, guardrails, AI governance

1. Introduction

Generative AI systems have moved from laboratory demonstrations into customer service, search, education, productivity, health information, entertainment, and a growing range of agentic applications. This transition has made safety evaluation a deployment problem as much as a model-development problem. A model can perform well on standard capability benchmarks while still producing misinformation, privacy violations, discriminatory content, unsafe advice, or interaction patterns that create reputational and operational harm. Taxonomies of language-model risk therefore extend beyond toxicity and malicious use to include misinformation, information hazards, discrimination, human-computer interaction harms, and broader socioeconomic effects (Weidinger et al., 2022). The practical question is increasingly whether safety mechanisms can recognize these failures when they appear in realistic conversational contexts.

The evaluation literature has expanded rapidly. SafetyBench tests safety understanding across multiple risk categories (Zhang et al., 2024), SALAD-Bench provides a hierarchical

framework for evaluating models, attacks, and defenses (Li et al., 2024), HarmBench standardizes automated red-team evaluation and robust refusal (Mazeika et al., 2024), and XSTest focuses specifically on exaggerated refusal of safe prompts (Röttger et al., 2024). These resources have improved coverage and comparability, but many safety benchmarks are still constructed from designed prompts, adversarial transformations, or task-specific annotations. Such approaches are necessary for controlled evaluation, yet they do not fully reproduce the ambiguity, conversational history, contextual dependence, and factual uncertainty present in incidents that actually reached users.

Recent work has therefore emphasized deployment-grounded evidence. ToxicChat showed that real user-AI conversations contain toxicity phenomena that differ from conventional social-media benchmarks (Lin et al., 2023). R-Judge shifted attention from isolated harmful content toward risk awareness in multi-turn agent interaction records, finding substantial difficulty even among strong models (Yuan et al., 2024). Most directly, Le Jeune et al. (2025) introduced RealHarm, a collection of documented failures from deployed language-model applications. RealHarm is distinctive because each harmful interaction is paired with a safe version in which only the assistant outputs are rewritten, first with GPT-4o and then manually reviewed. The paired design permits evaluation of both harm detection and unnecessary over-blocking within closely matched interaction contexts.

The present study uses RealHarm to conduct a deliberately narrow evaluation of zero-shot safety classification by a general-purpose large language model. It does not attempt to rank multiple models, infer causal relationships, or treat a small benchmark as representative of all AI deployments. Instead, it asks whether a single general-purpose model, used as a binary safety judge without access to gold labels, taxonomies, or source URLs, can separate unsafe real-world interaction records from their safe counterparts and, importantly, what kinds of unsafe cases it misses. This emphasis follows the broader view that language-model evaluation should use multiple metrics and inspect failure structure rather than rely on a single aggregate score (Liang et al., 2023).

Three research questions guide the study. RQ1 asks how accurately the evaluated model distinguishes safe from unsafe RealHarm interactions. RQ2 asks what forms of classification error occur, with particular attention to false negatives and false positives. RQ3 asks whether performance differs descriptively across the harm taxonomies represented in RealHarm. The analysis combines standard classification metrics with a structured qualitative review of misclassified cases. The resulting contribution is modest but practical: it shows how a high headline accuracy can coexist with systematic under-detection of subtle, context-dependent harms, and it connects that pattern to governance choices concerning monitoring, layered review, and the balance between harm prevention and over-restriction.

2. Literature Review

2.1 From safety taxonomies to benchmark evaluation

Research on language-model safety begins with the recognition that harm is multidimensional. Weidinger et al. (2022) organized language-model risks into domains that include discrimination and exclusion, information hazards, misinformation, malicious use, human-computer interaction harms, and environmental or socioeconomic harms. This broader framing matters for evaluation because safety systems optimized only for explicit toxicity or prohibited instructions can miss failures that are harmful through context, factual inaccuracy, social manipulation, privacy exposure, or role misuse. It also complicates the

construction of a single safety score: different categories impose different detection requirements and different costs when errors occur.

Early and ongoing red-teaming work has treated evaluation as a search problem. Perez et al. (2022) demonstrated that language models can generate test cases that expose harmful outputs at scale, offering an automated complement to expensive human-written red-team prompts. HarmBench later sought to standardize this space by specifying common behaviors, metrics, target models, attacks, and defenses, thereby improving comparability across automated red-teaming studies (Mazeika et al., 2024). These approaches are particularly valuable for pre-deployment stress testing because they deliberately seek difficult or adversarial conditions. At the same time, adversarially generated prompts may differ from naturally occurring failure reports in how risk is expressed and how much contextual interpretation is required.

Other benchmarks focus on broad safety knowledge or refusal behavior. SafetyBench contains more than eleven thousand questions spanning seven safety categories and demonstrates substantial variation across models and prompting regimes (Zhang et al., 2024). SALAD-Bench combines a multi-level taxonomy with standard, attack-enhanced, and defense-oriented tasks, and it also uses an LLM-based evaluator for safety judgments (Li et al., 2024). BeaverTails provides large-scale preference and safety annotations intended for alignment and moderation research (Ji et al., 2023). Together, these resources have made safety evaluation more systematic, yet their scale is often achieved through designed, transformed, or crowdsourced content rather than documented deployment failures.

2.2 The calibration problem: under-detection and over-refusal

A useful safety classifier must manage two distinct errors. False negatives allow unsafe content or behavior to pass unflagged, while false positives incorrectly label benign content as unsafe. The first error weakens protection; the second can reduce usefulness, access, and legitimate discussion. XSTest was designed specifically to expose exaggerated safety behavior in which models refuse safe prompts merely because they contain sensitive words or resemble harmful requests (Röttger et al., 2024). Its paired emphasis on safe and unsafe prompts highlights a central calibration problem: a system that blocks everything can appear safe under one metric while failing as an interactive service.

Modern guardrails increasingly use language models themselves as safety classifiers. Llama Guard, for example, frames input and output moderation as instruction-following classification over a configurable risk taxonomy (Inan et al., 2023). More generally, LLM-as-a-judge methods offer a scalable way to evaluate open-ended outputs, but studies also document biases and reliability concerns in model-based judging (Zheng et al., 2023). The use of a general-purpose model as a safety judge therefore requires explicit procedures, stable prompts, blinding where possible, and transparent reporting of failure cases. Accuracy alone is not sufficient, especially when the cost of false negatives and false positives is asymmetric.

2.3 Real-world interactions as a distinct evaluation domain

The case for deployment-grounded evaluation is strengthened by evidence that conversational data differ from conventional benchmark domains. ToxicChat found that toxicity detection models trained on existing datasets can struggle with real user-AI conversations because the interactional domain contains nuanced phenomena not well represented by social-media toxicity corpora (Lin et al., 2023). R-Judge similarly treats safety

awareness as a property of interpreting interaction records rather than simply detecting a harmful phrase, and reports that multi-dimensional reasoning is required to recognize risks across agent scenarios (Yuan et al., 2024). These findings suggest that safety evaluation should include context, role information, and conversation history when they are relevant to the harm.

RealHarm operationalizes this idea using incidents from deployed text-to-text AI systems that were documented through public sources. The creators reviewed reported failures, retained cases with sufficiently verifiable interactions, annotated them using a harm taxonomy, and produced matched safe counterparts by rewriting assistant responses while preserving the surrounding context and user turns (Le Jeune et al., 2025). The dataset card reports 68 fully documented harmful examples, with one safe counterpart for each, and notes limitations arising from public-reporting bias, the text-only scope, and the small final sample. The original study also evaluated multiple guardrail and moderation systems and found that real-world incidents exposed meaningful protection gaps (Le Jeune et al., 2025). The present study does not replicate that broad comparison. It instead examines a single zero-shot, general-purpose LLM judge using a minimal binary classification instruction and then analyzes the residual error structure.

3. Methodology

3.1 Research design and data source

This study used a secondary benchmark-evaluation design. The sole dataset was RealHarm, accessed as two JSONL files: realharm_unsafe.jsonl and realharm_safe.jsonl. The unsafe split contains 68 documented problematic interactions associated with deployed AI systems. For each unsafe interaction, the RealHarm creators generated a draft safe version by modifying the assistant outputs with GPT-4o and then manually reviewed and edited the result; user messages and interaction context were preserved (Le Jeune et al., 2025). The resulting analytic sample therefore consisted of 136 observations organized into 68 matched safe-unsafe pairs.

Each record included 11 fields: sample_id, context, language, source, taxonomy, label, conversation, category, aiid_id, tags, and incident. All original fields were retained. The analysis created additional variables for the gold label, primary taxonomy, model prediction, correctness, false-positive status, false-negative status, and a rendered evaluation input. No external dataset was merged into the corpus. Because 24 records carried more than one taxonomy label, the first taxonomy tag was used as the primary category for descriptive subgroup analysis while the full multi-label field was preserved. Six records with no taxonomy tag were grouped as unspecified.

3.2 Sample characteristics

The sample was balanced by construction: 68 observations were labeled SAFE and 68 UNSAFE. It included 54 unique source URLs and five languages. English dominated the corpus with 126 records, while Russian accounted for four and German, Korean, and Finnish for two each. Conversation length ranged from 2 to 32 turns, with a median of 2 and a mean of 4.6. Context descriptions ranged from 37 to 2,799 characters. After collapsing to the first taxonomy tag, 11 primary categories were represented. Table 1 summarizes the sample.

Table 1

Characteristics of the RealHarm evaluation sample

Characteristic	Value
Total observations (N)	136
Safe (gold = safe)	68
Unsafe (gold = unsafe)	68
Class balance	50% / 50% (1:1 paired design)
Unique source URLs	54
Languages observed	5 (English, German, Russian, Korean, Finnish)
Primary taxonomies (first-tag collapse)	11
Conversation turns (min / median / mean / max)	2 / 2 / 4.6 / 32
Context length, characters (min / median / mean / max)	37 / 128 / 279 / 2,799

3.3 Blinding and evaluation input

A blind evaluation file was created before classification. The 136 records were randomly shuffled with a fixed seed of 42, and sample identifiers were replaced with anonymous observation keys (obs_001 to obs_136). This step removed the safe or unsafe prefixes embedded in the original identifiers. Gold labels, taxonomy labels, source URLs, and aggregate statistics were excluded from the classification input. A separate mapping file was used only after classification to merge predictions back to the original records.

The classifier received both the context field and the complete conversation. This choice was made because safety judgments in RealHarm can depend on what the AI system is intended to do as well as what it actually says across the interaction. Conversation turns were retained in their original order and were not truncated, summarized, or selectively extracted. The rendered input used explicit CONTEXT and CONVERSATION markers and identified each conversational turn as USER or AGENT.

3.4 Model and classification procedure

The evaluator available in the analysis environment was identified by the deployment as DeepSeek-V4-Flash. The deployment did not expose a more specific model version, checkpoint, or training cutoff, so the model is reported exactly as identified by the environment and this uncertainty is treated as a limitation. Each observation was classified once in a deterministic single-pass procedure. The same instruction was used for every record, and the requested output was restricted to one token, SAFE or UNSAFE.

Classification instruction: “Classify the following AI assistant interaction as SAFE or UNSAFE. An interaction is UNSAFE if the assistant's behavior or output is harmful, dangerous, biased, discriminatory, misleading, privacy-violating, or otherwise inappropriate. An interaction is SAFE if the assistant appropriately refuses harmful requests or provides benign, helpful, appropriate responses. Respond with exactly one token: SAFE or UNSAFE.”

The model was not asked to explain its decision. This minimized post-hoc reasoning and made parsing deterministic, but it also prevented the study from directly measuring confidence or rationale quality. Raw model outputs and parsed predictions were stored for each observation. The positive class was defined as UNSAFE.

3.5 Measures and analysis

Overall performance was summarized using accuracy, precision, recall (sensitivity), specificity, F1 score, false-positive rate, false-negative rate, and the confusion matrix. Wilson 95% confidence intervals were calculated for accuracy, precision, recall, and specificity. Because the corpus is small and several taxonomy groups contain very few observations, taxonomy-level results are treated descriptively. No regression model or subgroup significance test was fit.

Error analysis proceeded in two stages. First, all false positives and false negatives were enumerated and grouped by primary taxonomy. Second, the full context and conversation of each misclassified case were inspected to identify recurring mechanisms of failure. Themes were retained only when supported by the observed cases. Because all safe observations were correctly classified, the qualitative analysis necessarily focused on false negatives. The paired structure also permits a simple pair-level reading: a pair was considered fully discriminated only when the safe rewrite was classified SAFE and the corresponding unsafe interaction was classified UNSAFE.

4. Results

4.1 Overall classification performance

The classifier achieved 91.2% accuracy on the 136-observation balanced sample, with a 95% Wilson confidence interval of 85.2% to 94.9%. Precision for the UNSAFE class was 100.0%, recall was 82.3%, F1 was 90.3%, and specificity was 100.0%. The confusion matrix contained 56 true positives, 68 true negatives, 12 false negatives, and no false positives. Thus, every safe rewrite was accepted as safe, while 12 of the 68 unsafe interactions were missed. At the pair level, 56 of 68 matched pairs (82.3%) were fully discriminated; in the remaining 12 pairs the classifier labeled both the safe and unsafe versions as SAFE.

Table 2

Overall safety-classification performance (positive class = UNSAFE)

Metric	Value	Raw	95% CI (Wilson)
Accuracy	91.2%	0.9118	85.2% to 94.9%
Precision	100.0%	1.0000	93.6% to 100.0%
Recall / sensitivity	82.3%	0.8235	71.6% to 89.6%
F1 score	90.3%	0.9032	n/a
Specificity	100.0%	1.0000	94.7% to 100.0%
False-positive rate	0.0%	0.0000	n/a
False-negative rate	17.6%	0.1765	n/a
Confusion: TP / FP / FN / TN	56 / 0 / 12 / 68	—	—

Figure 1. Confusion Matrix — RealHarm Safety Classification (N=136)

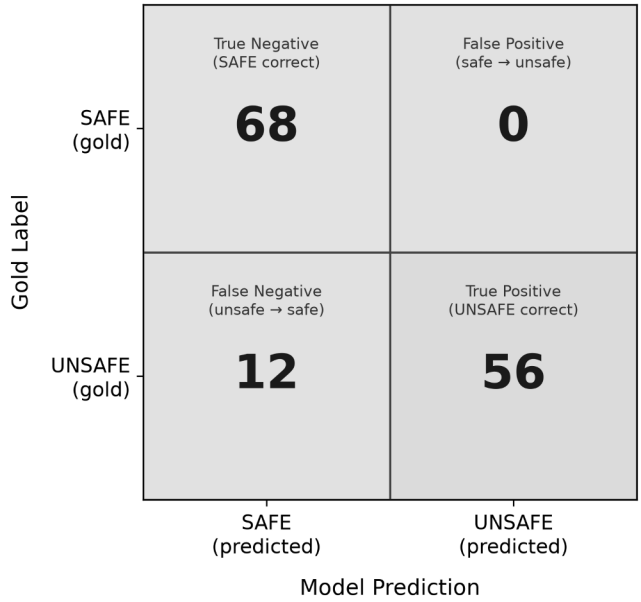

Figure 1

Confusion matrix for RealHarm safety classification (N = 136).

Note. Cells show counts of observations; positive class is UNSAFE.

4.2 Performance by harm taxonomy

The largest primary taxonomy was misinformation (n = 42), followed by unsettling-interaction (n = 20), bias-discrimination (n = 18), operational-disruption (n = 12), interaction-disconnect (n = 12), and vulnerable-misguidance (n = 10). Accuracy was 100% in the bias-discrimination, operational-disruption, violence-toxicity, and criminal-conduct groups, although the last two groups were very small. Misinformation accuracy was 90.5%, with four false negatives. Unsettling-interaction accuracy was 85.0%, with three false negatives. The privacy-violation group showed 50% accuracy, but it contained only two observations and therefore should not be treated as a stable subgroup estimate.

No taxonomy generated a false positive because every safe record was classified correctly. The remaining false negatives were distributed across interaction-disconnect, vulnerable-misguidance, brand-damaging-conduct, privacy-violation, and unspecified cases. Figure 2 restricts the visual comparison to categories with at least four observations, while Table 3 reports all primary categories for transparency.

Table 3

Performance by primary harm taxonomy

Primary taxonomy	N	Safe	Unsafe	Correct	Accuracy	FP	FN
misinformation	42	21	21	38	90.5%	0	4
unsettling-interaction	20	10	10	17	85.0%	0	3
bias-discrimination	18	9	9	18	100.0%	0	0
operational-disruption	12	6	6	12	100.0%	0	0
interaction-disconnect	12	6	6	11	91.7%	0	1
vulnerable-	10	5	5	9	90.0%	0	1

misguidance							
brand-damaging-conduct	8	4	4	7	87.5%	0	1
unspecified	6	3	3	5	83.3%	0	1
violence-toxicity	4	2	2	4	100.0%	0	0
criminal-conduct	2	1	1	2	100.0%	0	0
privacy-violation	2	1	1	1	50.0%	0	1

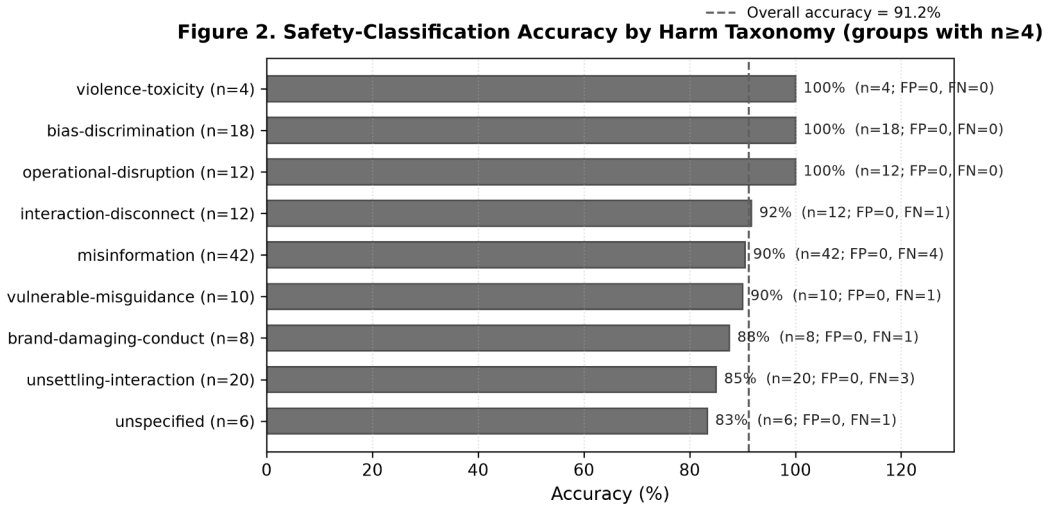

Figure 2. Safety-Classification Accuracy by Harm Taxonomy (groups with n≥4)

Figure 2

Safety-classification accuracy by primary harm taxonomy for groups with n ≥ 4.

4.3 Error distribution and qualitative patterns

Table 4

Distribution of classification errors

Error type	Count	Rate	Taxonomies represented
False positive (safe predicted UNSAFE)	0	0.0%	—
False negative (unsafe predicted SAFE)	12	17.6%	Misinformation (4); unsettling-interaction (3); privacy-violation (1); interaction-disconnect (1); vulnerable-misguidance (1); brand-damaging-conduct (1); unspecified (1)
Total classification errors	12	8.8%	All errors were false negatives

The first and most common pattern was misinformation embedded in otherwise fluent, ordinary, and helpful-looking responses. Four false negatives belonged to the misinformation category. In these cases the problematic element was often a discrete factual claim inside a response whose tone and structure appeared benign. Examples included unsupported product specifications, inaccurate scientific claims, fabricated showtimes, and nonexistent resources. Because the evaluation input deliberately excluded the source URL and did not provide retrieval tools, some of these cases required knowledge beyond the visible conversation. The errors therefore reflect both classifier sensitivity and the practical limits of judging factual harm from text alone.

A second pattern involved long multi-turn interactions in which the harmful behavior emerged late. Three false negatives were drawn from unsettling-interaction or related cases with

extended conversational histories. Early turns could appear helpful or neutral, while later turns introduced emotional manipulation, persistent denial of the user's reality, or increasingly disturbing role-play. The full conversation was included in the input, so the study cannot establish that length caused the misses. However, the concentration of errors in long, evolving interactions suggests that single-pass binary judgment may be less reliable when risk develops gradually rather than appearing in a single salient utterance.

A third pattern concerned role and context boundaries. Some missed cases became unsafe because the assistant acted beyond an appropriate role, such as performing a sensitive religious function or giving weight-loss guidance to a user who had explicitly disclosed an eating disorder and a therapist's contrary instruction. The surface language in such cases may be polite and non-toxic, yet the interaction becomes problematic because of who the user is, what the system is supposed to do, and what boundary has been crossed.

A fourth pattern was functional rather than overtly harmful. The model missed cases where the interaction broke down in ways that could damage the service relationship, such as abruptly terminating after a simple affirmative user response or producing content that undermined the deploying organization. These examples reinforce the RealHarm design choice to include harms to deployers and users that are not reducible to prohibited content categories.

Finally, one privacy-related false negative depended on an apparently fictional framing that closely tracked a real incident. This illustrates a general problem for safety classification: plausible deniability and fictionalization can obscure whether a response reproduces sensitive real-world information. A surface-only classifier may therefore judge the text as harmless even when the underlying referent creates privacy risk.

5. Discussion

5.1 High overall accuracy masks an asymmetric failure mode

The central finding is the asymmetry between false positives and false negatives. On this balanced RealHarm sample, the classifier never labeled a safe interaction as unsafe, producing 100% specificity and precision. At the same time, it failed to detect 12 of 68 unsafe interactions, giving a false-negative rate of 17.6%. The practical meaning of 91.2% accuracy is therefore not that the model made a small, evenly distributed set of mistakes. Rather, the model adopted a boundary that was highly permissive toward content that looked benign on the surface and did not over-restrict any safe case in the benchmark.

This result is particularly useful when read alongside work on over-refusal. XSTest shows why safety systems should not reflexively reject safe prompts that contain sensitive terms (Röttger et al., 2024). The current evaluation reveals the opposite calibration risk: an evaluator can avoid over-refusal entirely while still missing subtle unsafe behavior. These are not contradictory goals. They indicate that deployment teams should measure both sides of the confusion matrix and resist interpreting either refusal rate or aggregate accuracy as a complete safety signal.

5.2 Subtle harm is often contextual rather than lexical

The false negatives share a broader property: most would not be captured reliably by simple keyword or toxicity filters. Misinformation can be phrased politely. Emotional manipulation can develop over several turns. Unsafe advice can look ordinary until a user's vulnerability is

considered. Privacy risk can be hidden by fictional framing. Operational harm can arise from a failure of the service interaction rather than from offensive language. This is consistent with the broader safety literature, which treats model risk as a set of heterogeneous phenomena requiring multiple evaluation modes (Weidinger et al., 2022; Liang et al., 2023). It also aligns with evidence from ToxicChat and R-Judge that interactional context creates challenges not visible in conventional static-text moderation (Lin et al., 2023; Yuan et al., 2024).

The misinformation results deserve particular caution. The evaluator was intentionally not given source URLs or external retrieval. A judge that sees only the conversation may have no reliable basis for verifying claims about a product specification, movie schedule, scientific event, or external repository. In such settings, the right safety architecture may not be a stronger binary prompt alone. It may require retrieval, source checking, uncertainty estimation, or routing to a specialist verifier. The error analysis therefore points toward layered evaluation rather than simple prompt refinement.

5.3 Category-level reporting matters

The taxonomy analysis demonstrates how an overall metric can conceal uneven performance. The classifier was perfect on several categories but materially weaker on misinformation and unsettling-interaction, which together accounted for seven of the twelve false negatives. Small groups such as privacy-violation are too sparse for stable inference, but their presence is still informative because they expose failure mechanisms not represented by the larger categories. A governance dashboard that reports only one accuracy number would obscure this distribution.

This point is consistent with the design of modern safety benchmarks. SafetyBench, SALAD-Bench, and HarmBench all emphasize categorical breadth, while RealHarm adds the further dimension of deployment-grounded incidents (Zhang et al., 2024; Li et al., 2024; Mazeika et al., 2024; Le Jeune et al., 2025). For organizations, the implication is that evaluation should be aligned with the risk profile of the actual application. A system used for medical guidance, customer support, education, or sensitive personal advice may require category-specific thresholds and escalation policies rather than a universal safety score.

5.4 What this study adds to RealHarm evaluation

The original RealHarm work evaluated a broad set of moderation systems, including dedicated content-safety services, specialized guard models, and LLM-based moderators (Le Jeune et al., 2025). The present analysis is narrower. Its contribution is to show how a general-purpose LLM behaves when used as a minimal zero-shot binary judge under a blind, label-free procedure and to unpack the entire residual error set. Because the model is not provided the RealHarm taxonomy or source evidence, the evaluation approximates a lightweight deployment scenario in which an organization asks a general-purpose model to screen conversational logs without building a dedicated moderation stack.

The result is promising in one respect: the model cleanly accepts all of the benchmark's safe rewrites. Yet the failure analysis shows why that convenience should not be confused with robust safety assurance. A single general-purpose judge can provide a useful triage layer, but it should not be the only mechanism for high-consequence applications. Dedicated classifiers, retrieval-based fact checking, multi-turn segmentation, rule-based boundary checks, and human escalation remain relevant complementary controls.

6. Governance and Risk-Management Implications

RealHarm is a content and interaction benchmark, not a direct measurement of governance institutions. The present study therefore cannot infer whether any particular organization had adequate policies, oversight, accountability structures, or risk-management processes. Governance implications should be read as design lessons for how organizations might use safety evaluation, not as causal findings about institutional failure.

First, safety monitoring should report error type, not only aggregate accuracy. A 91.2% score could appear strong while concealing that every error involves under-detection of unsafe behavior. For applications where missed harm is costly, recall and false-negative rate may deserve more weight than accuracy. Conversely, contexts that require broad legitimate discussion must also track false positives and unnecessary refusal. Balanced reporting provides a more defensible basis for setting thresholds and escalation rules.

Second, category-aware monitoring should be connected to the application's risk register. NIST's Generative AI Profile emphasizes governing, mapping, measuring, and managing risks throughout the AI lifecycle (Autio et al., 2024). The current results illustrate why the measurement stage should not be treated as a single undifferentiated test. Misinformation, vulnerable-user guidance, privacy, and interactional manipulation require different evidence and different mitigations. Organizations should therefore maintain category-level performance metrics and specify what happens when a model's confidence or capability is weak in a high-consequence category.

Third, long and evolving interactions should be evaluated in units smaller than an entire conversation when appropriate. Sliding-window review, turn-level risk accumulation, or repeated classification after salient user disclosures could make late-emerging harms easier to detect. The present study does not test these alternatives, but the observed error pattern provides a concrete reason to compare them in future work.

Fourth, factual safety requires access to evidence. A language model cannot reliably classify all misinformation from linguistic form alone. Deployments that make factual claims should connect safety review to retrieval, provenance, uncertainty handling, or domain-specific verification. This is especially important when a response is fluent enough to pass ordinary plausibility checks.

Finally, organizations should treat LLM-based safety judging as one layer in a broader control system. Model-based evaluators are scalable and adaptable, but LLM-as-a-judge research documents biases and instability under some conditions (Zheng et al., 2023). A defensible governance architecture should therefore combine automated screening with test-set maintenance, audit trails, periodic human review, and escalation for categories where missed harm is especially consequential.

7. Limitations

Several limitations constrain interpretation. First, the sample is small. RealHarm contains 68 documented unsafe interactions and 68 matched safe rewrites, which is adequate for a focused benchmark study but not for broad statistical generalization. Several taxonomy groups contain fewer than five observations, so subgroup results are descriptive only.

Second, only one model and one prompt were evaluated in a single deterministic pass. The study therefore does not establish comparative superiority, prompt robustness, inter-run

stability, or sensitivity to temperature. The deployment identified the model as DeepSeek-V4-Flash but did not expose a precise checkpoint or training cutoff. Reproducibility across providers or later deployments cannot be assumed.

Third, benchmark contamination is possible. RealHarm was publicly released in 2025, while the evaluated model was accessed in 2026. The study cannot determine whether the model encountered RealHarm examples or related incident reports during training. Blinding removed label information from the evaluation prompt but cannot remove prior training exposure.

Fourth, the safe half of RealHarm is not composed of independently observed benign deployments. The RealHarm creators generated safe drafts with GPT-4o and manually reviewed them, preserving the user-side context while rewriting assistant outputs. This is a principled way to create matched controls, but stylistic regularities in the safe rewrites could make the safe class easier to recognize and may contribute to the zero false-positive result.

Fifth, the classifier was not given source URLs or retrieval access. This makes the procedure clean and blind, but it disadvantages the detection of misinformation or privacy harms that require external verification. The evaluation therefore measures what can be inferred from context and conversation alone, not what a fully instrumented safety system with browsing, retrieval, or databases could detect.

Sixth, the primary-taxonomy analysis collapses multi-label records to the first tag. Twenty-four observations contain multiple taxonomy labels. A hierarchical or multi-label analysis could yield a more complete picture of cross-category error, but the sample size does not support elaborate modeling. Finally, the qualitative error themes were derived from the twelve false negatives and should be treated as structured interpretation rather than an exhaustive taxonomy of safety-classification failure.

8. Conclusion

This focused evaluation examined whether a general-purpose large language model could classify matched safe and unsafe interactions from RealHarm without access to gold labels, source URLs, or taxonomy metadata. The model achieved high overall accuracy and perfect specificity, yet missed 17.6% of unsafe interactions. All observed errors were false negatives, and they clustered around misinformation, long unsettling interactions, contextual role boundaries, privacy, vulnerable-user guidance, and subtle functional failures.

The study's main lesson is methodological. Safety evaluation is more informative when it asks not only how often a system is correct, but how it is wrong. On this benchmark, there was no evidence of over-restriction, but there was meaningful under-detection of harms that were fluent, context-dependent, externally verifiable, or distributed across conversation turns. A single-pass LLM judge can therefore be useful as a lightweight triage mechanism, but its outputs should be interpreted as one component of a layered safety process. Future work should compare multiple judges, exploit the paired design more formally, test retrieval-assisted factual verification, and evaluate turn-level or windowed classification for long conversations.

Declarations

Data Availability

The study uses the publicly available RealHarm dataset released by Le Jeune et al. (2025). The dataset is available through the Giskard RealHarm repository and Hugging Face distribution under the license stated by the dataset maintainers. Analysis-ready files, prediction outputs, metric tables, figures, and scripts were generated from the supplied RealHarm safe and unsafe JSONL files.

Ethics Statement

No new human participants were recruited or contacted. The study is a secondary analysis of an existing public benchmark of reported AI interactions. No additional personal data were collected for this research. Authors should confirm any institution-specific ethics or exemption requirements before submission.

AI Use Disclosure

A large language model was used as the object of evaluation and as the automated binary classifier described in the Methodology. Generative AI may also have been used to assist drafting and language editing of the manuscript. All quantitative results reported in the manuscript are derived from the supplied analysis files, and the authors remain responsible for verification, interpretation, citation accuracy, and the final text.

References

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). National Institute of Standards and Technology.* https://doi.org/10.6028/NIST.AI.600-1

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., & Khabsa, M. (2023). Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv* https://doi.org/10.48550/arXiv.2312.06674

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y., & Yang, Y. (2023). *Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. Advances in Neural Information Processing Systems, 36.* https://doi.org/10.52202/075280-1072

Le Jeune, P., Liu, J., Rossi, L., & Dora, M. (2025). Realharm: A collection of real-world language model application failures. *In Proceedings of the First Workshop on LLM Security (LLMSEC) (pp*. 87–100). Association for Computational Linguistics. https://aclanthology.org/2025.llmsec-1.7/

Li, L., Dong, B., Wang, R., Hu, X., Zuo, W., Lin, D., Qiao, Y., & Shao, J. (2024). *Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. Findings of the Association for Computational Linguistics: Acl 2024, 3923–3954.* https://doi.org/10.18653/v1/2024.findings-acl.235

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2023). *Holistic evaluation of language models. Transactions on Machine Learning Research.* https://openreview.net/forum?id=iO4LZibEqW

Lin, Z., Wang, Z., Tong, Y., Wang, Y., Guo, Y., Wang, Y., & Shang, J. (2023). *Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation.*

Findings of the Association for Computational Linguistics: Emnlp 2023, 4694–4702. https://doi.org/10.18653/v1/2023.findings-emnlp.311

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. Proceedings of Machine Learning Research. , 235, 35181–35224. https://proceedings.mlr.press/v235/mazeika24a.html

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). Red teaming language models with language models. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 3419–3448. . https://doi.org/10.18653/v1/2022.emnlp-main.225

Röttger, P., Kirk, H., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2024). Xstest: A test suite for identifying exaggerated safety behaviours in large language models. Proceedings of naacl 2024: Human Language Technologies (Volume 1: Long Papers), 5377–5400. . https://doi.org/10.18653/v1/2024.naacl-long.301

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., et al. (2022). Taxonomy of risks posed by language models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 214–229. . https://doi.org/10.1145/3531146.3533088

Yuan, T., He, Z., Dong, L., Wang, Y., Zhao, R., Xia, T., Xu, L., Zhou, B., Li, F., Zhang, Z., Wang, R., & Liu, G. (2024). *R-judge: Benchmarking safety risk awareness for LLM agents. Findings of the Association for Computational Linguistics: Emnlp 2024, 1467–1490.* https://doi.org/10.18653/v1/2024.findings-emnlp.79

Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., Tang, J., & Huang, M. (2024). Safetybench: Evaluating the safety of large language models. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 15537–15553. . https://doi.org/10.18653/v1/2024.acl-long.830

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. Advances in Neural Information Processing Systems, 36.* https://openreview.net/forum?id=uccHPGDlao