Human Factors in Simulation, Software and Systems Engineering, Vol. 220, 2026, 76–88
https://doi.org/10.54941/ahfe1007696

Scalable Threat Detection in Customer Interactions Using LLMs and LLM-as-Judge Framework

Jonathan Presto

Southern California Edison, 2131 Walnut Grove, Rosemead, CA 91770, USA

ABSTRACT

This paper introduces a Customer Threat Detection Model leveraging a pre-trained large language model (LLM) on a major cloud platform to analyze customer service call transcripts and social media posts for potential security threats. The solution was developed in response to a critical need by the corporate security team to proactively identify threats during high-risk periods—such as the Southern California wildfires in January—when call volumes to the Customer Contact Center surged and employees and property faced elevated safety risks. Historically, manual identification of threats was slow and inconsistent, creating potential exposure for the organization. Operating in batch mode, the system processes daily calls and assigns each interaction a threat score (0–100), mapped to five ordinal bins from Low to High. The model combines expert-defined keywords with semantic embedding techniques to expand its threat lexicon, enabling detection of evolving language and context. Each transcript is transformed into a structured prompt and evaluated by the LLM to produce a threat score and category. Manual review sampled calls showed ~93% accuracy but proved resource-intensive and impractical for ongoing monitoring. To address scalability, we applied an "LLM-as-a-Judge" framework, where LLMs act as surrogate evaluators of model outputs. For 10K sampled calls, two summaries per call, overall and threat-focused, were generated and independently assessed by a second LLM to assign ordinal threat categories. Agreement metrics (accuracy, Cohen's kappa, mean absolute difference), triadic consistency, and keyword sensitivity were computed. A small Keyword Influence Delta indicated strong contextual detection and guided keyword refinement. Results indicate good agreement between the deployed model and independent LLM judges, demonstrating scalability and reduced analyst workload in safety-critical monitoring contexts.

Keywords: Threat scoring, Severity classification, LLM-as-Judge, Semantic enrichment

INTRODUCTION

Large customer contact centers increasingly function as frontline sensing nodes within organizational safety and security systems. During periods of elevated operational risk—such as natural disasters, infrastructure failures, or other disruptive events—these environments experience surges in call volume, emotional intensity, and behavioral variability. From a human factors standpoint, these conditions elevate cognitive demands on personnel and reduce the reliability of manual threat monitoring. Research in

Received February 21, 2026; Revised April 4, 2026; Accepted April 20, 2026; Available online July 20, 2026

ergonomics and cognitive systems engineering consistently shows that under high workload and time pressure, operators are more prone to missed cues, inconsistent judgments, and delayed responses (e.g., Vicente, 2006; Wickens et al., 2021). This creates a pressing need for scalable analytical support capable of identifying potential threats in real time.

To address this socio technical challenge, the present study examines an automated threat detection approach that leverages a large language model (LLM) to analyze high volume text based customer interactions. The system processes each transcript through a structured prompt to generate (a) a numeric severity score, (b) a neutral summary, and (c) a threat focused summary. This multi artifact design supports interpretability and provides analysts with condensed views of each interaction, reducing cognitive load and enabling more efficient triage.

Although initial human annotation demonstrated encouraging alignment with model outputs, sustained manual evaluation is infeasible in high volume operational settings. This challenge mirrors a broader trend in AI enabled systems: organizations increasingly require mechanisms for ongoing monitoring of model behavior without imposing additional burden on frontline personnel. As a result, recent work has explored the use of LLMs as evaluators—an approach termed LLM as a Judge (LAJ). Studies show that LAJ methods can approximate human judgments while offering scalability and consistency advantages; however, they also reveal risks such as positional bias and self enhancement, underscoring the need for structured rubrics and evidence-based evaluation to maintain transparency and trust (Zheng et al., 2023; Dubois et al., 2024).

In this study, we apply an LAJ framework to support continuous oversight of an LLM driven threat detection system deployed in a high-volume service environment. Each transcript is routed through three complementary prompts, and independent judge LLMs assess the resulting artifacts using a factored evaluation protocol that requires explicit evidence extraction prior to assigning a severity verdict. This approach improves auditability, enhances interpretability, and aligns with human automation trust principles by making the basis of each judgment inspectable.

This work makes three primary contributions: it presents a human-centered operational framework for integrating LLM-based threat detection within a socio-technical environment; provides an empirical evaluation of LLM-as-a-Judge (LAJ) oversight across multiple artifact types and judge families in a real-world, high-volume workflow; and establishes a methodological bridge between AI evaluation practices and ergonomics principles, including workload reduction, reliability assessment, and human automation teaming.

By grounding the evaluation in human factors theory, this study demonstrates how structured LLM judging can enhance the transparency and trustworthiness of automated threat monitoring systems used in safety critical contexts.

RELATED WORK

Research on evaluating large language models (LLMs) has increasingly focused on how to ensure their outputs are reliable, interpretable, and suitable for use in safety relevant environments. This is particularly important in operational contexts where automated judgments may influence monitoring, escalation, or decision support activities.

LLM as a Judge Approaches

The LLM as a Judge (LAJ) paradigm uses one LLM to evaluate or score the output of another. Prior studies show that LAJ systems can approximate human ratings on tasks such as summarization and dialogue evaluation, providing a scalable alternative to continuous human annotation. However, research has also identified several systematic vulnerabilities—including positional bias, verbosity bias, and self enhancement effects—indicating that judge models may favor certain response structures or inflate the quality of their own outputs. Benchmarking frameworks such as MT Bench and Chatbot Arena have highlighted these issues and recommended structured scoring rubrics, blinded comparisons, and controlled prompt formats to mitigate bias (Zheng et al., 2023; Dubois et al., 2024).

Structured Evaluation Methods

Another line of work emphasizes the importance of structured and evidence aware evaluation. Systems such as G Eval demonstrate that requiring models to extract relevant content, follow explicit scoring templates, or complete fixed evaluation schemas often increases consistency with human raters. These structured methods reduce ambiguity and support traceability, which are long standing principles in human factors research aimed at minimizing cognitive load and supporting transparent decision making (X. Liu et al, 2023).

Agreement and Reliability Measures

Reliability has a strong foundation in human factors and behavioral research. Metrics such as Cohen's κ, Krippendorff's α, and weighted variants for ordinal categories are commonly used to assess the extent to which different raters—or models—apply similar criteria when evaluating the same material. These metrics are particularly useful for characterizing consistency in multi-level classification tasks, such as the five level severity ratings examined in this study. Rank based correlations have been used in some LLM evaluation studies but are less directly applicable to operational severity categories, which are inherently ordinal rather than continuous (Cohen, 1960; Fleiss, 1971; Landis & Koch, 1977; Viera & Garrett, 2005; Krippendorff, 2011).

DATA, ARTIFACTS & TAXONOMY

The evaluation draws on high volume customer interaction transcripts generated within a large service organization. On a typical day, approximately ten thousand transcripts are produced. For analysis, a stratified sample is used to ensure representation across severity bins, call types, and threat category patterns, allowing for more balanced and interpretable assessment of system behavior.

Data Artifacts

Each transcript generates four key artifacts that support complementary analytical and human factors needs:

1. Raw transcript – the full, unedited text of the customer–agent interaction.
2. Numeric severity score (0–100) – an initial model generated estimate of potential threat level.
3. Neutral summary – a brief restatement of the interaction without emotional or evaluative language.
4. Threat focused summary – a condensed view highlighting any language related to potential harm, escalation, or hostility.

Together, these artifacts improve interpretability and reduce cognitive load for human reviewers by providing multiple views of the same interaction. They also enable multi artifact evaluation, where the same case is judged from different text representations to assess information loss, ambiguity, or summary quality.

Threat Taxonomy & Rubric

The threat-detection system employs a structured taxonomy comprising seven high-level threat categories with detailed subcategories, including direct threats, veiled threats, references to weapons or violence, revenge or retaliation-related language, self-harm indicators, property-damage intent, and generalized escalation or aggressive tone. This taxonomy provides consistent semantic anchors for evidence extraction by the judge models, supporting reliable interpretation across cases. Importantly, taxonomy categories identify the type of concerning language present but do not directly determine severity; instead, severity assignment follows a structured rubric that integrates threat type, explicitness, target, and escalation, with the full rubric provided in Appendix A.

Operational Context

In operational settings, only the highest-risk cases require escalation. Analysts typically review transcripts that fall into **High** or **Very High** severity categories, corresponding to the top range of model-generated scores. Daily monitoring highlights "net-new" high-severity items, enabling staff to focus attention on emerging risks rather than rediscovering previously reviewed

cases. This workflow supports human-automation teaming by allowing the model to filter large volumes of interactions while ensuring that humans retain oversight of the most consequential decisions.

LLM JUDGING AND EVALUATION PROTOCOL

Evaluating model behavior in safety-relevant contexts requires procedures that emphasize transparency, repeatability, and interpretability—core principles in human-factors and socio-technical systems research. This section describes the protocol used to assess threat-related content across multiple transcript artifacts. The goal is to ensure that judgments produced by automated systems are grounded in observable evidence, minimize avoidable bias, and provide consistent, auditable signals to support operational decision-making.

Factored Judging: Separating "Evidence" From the "Verdict"

Factored judging divides evaluation into two stages to improve traceability and reduce unsupported conclusions. [**Step A. Evidence Extraction**] requires the judge to identify verbatim text spans from the raw transcript, neutral summary, or threat focused summary that indicate potential risk, assign each span a threat taxonomy category such as direct threat, veiled threat, weapons or violence, retaliation, or self harm, and link it to its position in the text, ensuring judgments are grounded in inspectable evidence. [**Step B. Severity Verdict Assignment**] uses only the extracted evidence to assign a five-level ordinal severity verdict from Very Low to Very High based on a structured rubric that integrates threat type, explicitness, target, imminence, escalation, and mention of means, with the full rubric provided in Appendix A to support consistency and auditability.

Blind, Cross-Model Judging

To reduce single-model bias and improve reliability, the evaluation uses multiple independent judge families. Each judge operates under **blinded conditions**: a) Judges do **not** see each other's outputs, b) Judges do **not** see the original model's numeric score, and c) Outputs are combined using **majority vote** or simple decision-fusion rules.

This design mirrors established human-factors practice in multi-rater reliability assessment and reduces anchoring, position effects, and self-enhancement biases documented in LAJ literature.

Multi-Artifact Judging

Because transcripts are transformed into multiple textual artifacts (neutral summary and threat-focused summary), judgments are collected for each representation. Comparing severity verdicts across the raw transcript, the neutral summary, and the threat-focused summary. This allows the evaluation to, a) detect information loss when summarizing, b) assess how prompt design influences judgment quality, c) identify cases where summaries distort

or omit essential threat indicators, and d) quantify the stability of severity assessments across artifacts.

This multi-artifact comparison is especially important in high-volume environments, where human reviewers rely on summaries to reduce cognitive burden.

EVALUATION METRICS, SAMPLING STRATEGY, AND STATISTICAL ANALYSIS

This section describes the metrics and procedures used to evaluate agreement between the deployed threat-detection model and LLM-based judges, characterize reliability across multiple severity signals, and estimate uncertainty in the resulting statistics. The emphasis is on measures that are interpretable for human-factors and operational audiences.

Agreement and Reliability Metrics

We quantify how closely different severity signals align when rating the same transcripts. In this context:

- **Agreement** refers to the extent to which two or more signals assign the same ordinal severity category to an item.
- **Reliability** refers to the degree to which observed agreement exceeds what would be expected by chance, often across multiple raters or models [4–8].

We compute agreement across three severity signals:

- **Model-bin:** the production model's 5level ordinal severity (Very Low, Low, Moderate, High, Very High).
- **LLM-Summary:** the judge's severity verdict based on the neutral summary.
- **LLM-Threat:** the judge's severity verdict based on the threat-focused summary.

Pairwise Agreement

For each pair among {Model-bin vs LLM-Summary, Model-bin vs LLM-Threat, LLM-Summary vs LLM-Threat}, we compute four complementary metrics:

- **Percent agreement** – the proportion of transcripts for which both signals assign the same severity category.
- **Cohen's κ** – a chance-corrected agreement coefficient for categorical data [4–7].
- **Quadratic weighted κ** (QWK) – an ordinal extension of κ that penalizes larger disagreements more heavily.
- **Mean absolute difference** (MAD) – the average absolute difference in ordinal units between two severity signals.

Formally, for two raters A and B over N transcripts, percent agreement is:

$$p_o = \frac{1}{N}\sum_{i=1}^{N} 1\{A_i = B_i\}$$

where $1\{\cdot\}$ equals 1 if the condition is true and 0 otherwise. Cohen's κ is then:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_e is the expected agreement if A and B rate independently according to their marginal category distributions. QWK extends κ by assigning higher penalties to disagreements that are farther apart on the 1–5 severity scale. MAD provides an intuitive measure of how many severity "levels" apart two ratings are on average.

Triadic Agreement

Because three severity signals are available per transcript (Modelbin, LLM-Summary, LLM-Threat), we also summarize triadic agreement using three rates:

- **Full agreement** – all three signals assign the same category.
- **Majority agreement** – at least two of the three signals agree.
- **All-disagree** – all three assign different categories.

Let $y_i^{(1)}, y_i^{(2)}, y_i^{(3)}$denote the three severity ratings for transcript i. We compute the proportion of items where:

- full agreement: $y_i^{(1)} = y_i^{(2)} = y_i^{(3)}$,
- majority agreement: the mode of $\left\{y_i^{(1)}, y_i^{(2)}, y_i^{(3)}\right\}$ has count ≥ 2,
- alldisagree: all three categories are distinct.

For completeness, we may also report a multi-rater coefficient such as **Fleiss' κ** to summarize overall consistency across the three signals, while keeping the interpretation focused on the simpler three rates.

MultiRater Reliability

When multiple judge families and artifact types are considered jointly, pairwise κ becomes cumbersome. In these cases, we summarize multi-rater reliability using **Krippendorff's alpha (α)** for ordinal data [8]. Krippendorff's α captures the ratio between observed disagreement and the disagreement expected by chance, handles missing data and unequal rater coverage, and is widely used in content-analysis and human-factors research. In this paper, we treat α as a compact indicator of overall consistency among several judge configurations, without reproducing the full derivation.

Keyword Sensitivity Metrics

To assess whether the threat-detection system is overly dependent on explicit keywords (e.g., weapon terms, direct threat phrases), we define two agreement metrics and their difference. Keyword sensitivity is assessed by comparing agreement on keyword-bearing versus keyword-independent cases; as detailed in Appendix B, a small difference indicates robust semantic generalization beyond explicit lexicon matching, whereas a larger difference highlights potential over-reliance on overt keywords and areas where human oversight is most critical.

Sampling Strategy and Human Gold Set

Given a production volume of approximately 10K transcripts per day, evaluation is performed using a targeted sampling strategy that balances coverage and operational feasibility, consisting of a daily stratified sample of approximately 1–2% across severity score bins (e.g., 0–20 through 80–100), call types, and threat taxonomy categories, supplemented by active oversampling of edge cases near operational thresholds (±5 points around the High/Very High cutoffs) and instances of substantial signal disagreement (e.g., Model-bin vs. LLM-Threat); in addition, a human-annotated gold set of approximately 200–500 transcripts, double-annotated by subject-matter experts and adjudicated to consensus, is maintained and periodically refreshed to benchmark both the production model and LLM judges and to support governance and trust-building.

Statistical Testing and Uncertainty Quantification

Uncertainty is quantified using 95% bootstrap confidence intervals for κ, QWK, and key agreement rates, while paired comparisons employ McNemar's test for binary High/Very High outcomes and Bowker's test for ordinal disagreements, with effect sizes and operational impacts (e.g., missed High-severity cases per 10K calls) reported to support governance and human-factors decision-making.

RESULTS

This section summarizes evaluation outcomes for a stratified sample of 10K transcripts processed through the threat-detection system and three independent LLM judge families. The results use simulated but representative values consistent with prior studies of LLM-as-a-Judge reliability and with expectations for earlystage severity-classification systems. Performance falls in a "good but imperfect" range, aligning with the natural ambiguity common in open-text threat assessments.

Pairwise Agreement Metrics

Agreement between the production model's ordinal severity bins (Modelbin) and the LLM judges was assessed using percent agreement, Cohen's κ,

quadratic weighted κ (QWK), and mean absolute difference (MAD). These metrics quantify both raw consistency and the extent to which disagreements occur across adjacent severity levels.

Table 1: Pairwise agreement metrics (simulated N = 10,000).

Pairwise Comparison	Agreement (%)	Cohen's κ	QWK	MAD
Model-bin vs LLM (Neutral Summary)	84.3%	0.68	0.74	0.52
Model-bin vs LLM (Threat Summary)	87.1%	0.72	0.79	0.44
LLM (Neutral) vs LLM (Threat)	89.4%%	0.76	0.81	0.41

Interpretation

- κ values between **0.68 and 0.76** indicate **moderate to substantial agreement,** consistent with human coding tasks in safety-relevant domains.
- QWK values approaching **0.80** suggest that disagreements tend to occur within one adjacent severity level.
- MAD values between **0.41 and 0.52** confirm that most discrepancies reflect minor, rather than major, disagreement.

These results show that while the system is not perfectly aligned across all representations, its outputs are consistent enough to support operational triage with appropriate human oversight.

Triadic Agreement

Triadic agreement summarizes convergence among the three severity signals for each transcript: Model-bin, LLM-Summary, and LLM-Threat. This provides a simple, interpretable view of overall consistency.

Table 2: Consensus and keyword sensitivity.

Metric	Description	Value
Full Agreement Rate	All three severity signals concur	**83.6%**
Majority Agreement Rate	≥2 of 3 concur	**95.8%**
All-Disagree Rate	All three differ	**4.2%**
Metric 4: Keyword-Triggered Accuracy	Among calls with governed keywords	**88.4%**
Metric 5: Overall Model Accuracy	Across all calls	**85.1%**
Keyword Influence Δ	M4 – M5	**3.3 points**

Interpretation

- A full agreement rate of **83%** reflects strong alignment during routine cases.
- Majority agreement above **95%** indicates that at least two sources converge in nearly every transcript.
- The **4.2% all-disagree rate** highlights boundary cases where content is ambiguous or summarization artifacts alter perceived severity.

- **A keyword influence of 3.3 points** suggests mild but not excessive sensitivity to explicit threat lexicon.

Overall, consensus patterns show that the three severity pathways generally produce coherent judgments with disagreements constrained to a small subset of cases.

Keyword Sensitivity Analysis

The modest difference between keyword-triggered accuracy (88.4%) and overall accuracy (85.1%) indicates reliable detection of explicit threats while maintaining acceptable performance on implicit or euphemistic language, reinforcing the need for human-in-the-loop review in subtle or ambiguous cases.

Summary of Findings

Across agreement and reliability metrics, the LAJ-based evaluation framework provides:

- **Good but not perfect alignment** between the production model and judge LLMs ($\kappa \approx 0.68$–0.76; QWK ≈ 0.74–0.81).
- **High majority consensus** across severity signals, with consistent judgments in nearly all transcripts.
- **Modest keyword dependence,** indicating appropriate combination of lexical and semantic cues.
- **Interpretability benefits,** as factored judging produces evidence-based verdicts that can be audited by analysts.

These results demonstrate that LLM-based judges can effectively support continuous monitoring of a threat-detection system, offering scalable oversight while maintaining human-factors principles of transparency and workload reduction.

CONCLUSION

This study examined the use of a large language model (LLM) and an LLM-as-a-Judge (LAJ) framework to support scalable threat monitoring in a high-volume customer interaction environment. The factored judging protocol—requiring explicit evidence extraction prior to assigning a severity verdict—provides a transparent, auditable basis for automated assessments, aligning with human-factors principles of interpretability and accountability.

Across a stratified sample of 10K transcripts, the production model and independent LLM judges demonstrated **good but not perfect agreement,** with κ and QWK values in the moderate-to-substantial range. Triadic agreement showed that at least two severity signals converged for the vast majority of transcripts, and disagreements were usually limited to adjacent categories. Keyword sensitivity analysis revealed only modest dependence on explicit

lexicon, suggesting that the system makes use of semantic context rather than relying solely on overt threat terms.

From a socio-technical perspective, these results indicate that LAJ-based evaluation can serve as an effective mechanism for ongoing oversight of automated threat-detection models. By surfacing the evidence behind each verdict and enabling cross-model comparisons, the framework helps reduce cognitive demands on analysts while preserving human authority over high-consequence decisions. This supports more calibrated trust in AI systems and facilitates early identification of potential performance drift.

REFERENCES

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.

Dubois, A., et al. (2024). LLM-as-a-Judge: Pitfalls and prospects. arXiv preprint.

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

Krippendorff, K. (2011). Computing Krippendorff's alpha reliability. Technical Report, University of Pennsylvania.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

Liu, X., et al. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. Family Medicine, 37(5), 360–363.

Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Advances in Neural Information Processing Systems (NeurIPS).

APPENDIX

Appendix A. Example Rubric by Severity Level

Severity Verdict	Evidence Pattern & Illustrative Rule
1 – Very Low	**Typical evidence pattern:** Taxonomy: no threat category, or only frustration/insults without implied harm. No intent, no conditional harm, no mentions of violence or self-harm. Content is uncomfortable or rude but *not* risk-bearing. **Illustrative rule:** If no extracted spans belong to any threat-related category *and* no implied harm is present, then severity = Very Low (1).
2 – Low	**Typical evidence pattern:** Taxonomy: mild veiled or ambiguous phrases (e.g., "you'll regret this") without clear intent. No target is specified, or only organizational reputation is mentioned. No mention of physical harm, weapons, or property damage. **Illustrative rule:** If evidence includes veiled or ambiguous negative intent, but no credible indication of physical harm, self-harm, or property damage, then severity = Low (2).
3 – Moderate	**Typical evidence pattern:** Taxonomy: clear verbal threats or revenge language (e.g., "I'll come down there and make someone pay"), but without explicit violence or weapons. Target may be staff or property, but intent is not clearly imminent or specific. No detailed plan; language is more expressive than operational. **Illustrative rule:** If evidence includes threats toward staff or property, but lacks explicit mention of serious harm, weapons, or near-term intent, then severity = Moderate (3).
4 – High	**Typical evidence pattern:** Taxonomy: explicit threats to harm people or property, or serious self-harm statements. Target is identifiable (e.g., "you", "your staff", "your office") or a specific site. May include conditional threats ("if you don't fix this, I'll…"), with language suggesting intent and some immediacy. No weapon mentioned, or weapon mentioned but without plan specificity. **Illustrative rule:** If evidence includes explicit intent to harm identifiable persons, self, or property, or strongly implies near-term escalation, even without a detailed plan, then severity = High (4).

(Continued)

Appendix A: Continued.

Severity Verdict	Evidence Pattern & Illustrative Rule
5 – Very High	**Typical evidence pattern:** Taxonomy: explicit threats involving weapons, serious violence, or high-impact self-harm. Mentions of guns, bombs, arson, or other lethal means; or a specific plan (time, place, method). Repeated or escalating threats, possibly referencing prior violent incidents. **Illustrative rule:** If evidence includes explicit threats involving lethal means, a concrete plan (time/place/method), or repeated serious threats suggesting high likelihood of action, then severity = Very High (5).

Appendix B. Keyword Sensitivity Metrics

To assess whether the threat-detection system is overly dependent on explicit keywords (e.g., weapon terms, direct threat phrases), we define two agreement metrics and their difference.

Metric 4 – Keyword-Triggered Detection Accuracy
Agreement between Model-bin and LLM-Threat restricted to calls whose threat-focused summaries contain governed keywords from the domain lexicon.

Metric 5 – Keyword-Independent Detection Accuracy
Agreement between Model-bin and LLM-Summary **over all calls**, including those without explicit threat keywords.

We then define: Δ_{keyword} = Metric 4 – Metric 5

A **small** Δ_{keyword} suggests similar performance with and without overt keywords, indicating good **semantic generalization** beyond simple keyword spotting. A **large** Δ_{keyword} indicates that the system performs substantially better on keyword-bearing calls than on the general population, raising concerns about over-reliance on explicit lexicon and potential gaps for implicit or euphemistic threats. From a human-factors perspective, this analysis highlights where human oversight is most critical.