Autonomous Red Teaming for Large Language Models: A Minimally Supervised Approach Using Specialized Agents

Débora Medeiros[1], Ana Vieira[1], Franklin Dantas[1], Charlene S. Queiroz[1], Marcela Rodrigues[1], Rebeca V. Carvalho[2], Edson Silva Junior[1], Camila Ribeiro[1], Nicolas R. G. Assumpção[3], José R. H. Carvalho[1]

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Av. Gen. Rodrigo Octávio, 6200 – 69.080-900 – Manaus – AM – Brazil

[2]Universidade Federal do Rio Grande do Sul (UFRGS) – Porto Alegre, RS, Brazil

[3]Motorola Mobility LLC – USA

{debora.medeiros, franklin.dantas,
charlenes.queiroz}@icomp.ufam.edu.br,
{marcela.rodrigues, edson, reginaldo}@icomp.ufam.edu.br,
vieira.ana@ufam.edu.br, camilaribeiro301@gmail.com,
psico.rebeca.vc@gmail.com, nicolas.riccieri@gmail.com

Abstract. *This paper investigates the feasibility of using autonomous AI agents to conduct Red Team evaluations of Large Language Models (LLMs) with minimal human supervision. We propose a framework of specialized agents targeting sensitive domains, including homicide, human trafficking, and zoophilia, which interact with aligned and less aligned LLMs to elicit and identify harmful outputs. The method aligns encouragingly with expert majority judgments, though the low inter-rater agreement highlights the subjectivity of harmfulness assessment, which we discuss as a central limitation. Overall, the framework shows potential as a scalable, cost-efficient support tool for LLM safety evaluation.*

1. Introduction

The rapid integration of Large Language Models (LLMs) into real-world systems has intensified concerns regarding safety, robustness, and susceptibility to adversarial misuse. Recent studies also indicate a decline in public trust in AI technologies, reinforcing the need for robust and transparent evaluation mechanisms [Gillespie et al. 2025]. In this context, Red Teaming has emerged as an important strategy for systematically probing vulnerabilities and identifying harmful behaviors before deployment.

Despite its relevance, Red Teaming for language models remains largely dependent on manual efforts conducted by human experts. Such evaluations are costly, difficult to scale, and may vary according to evaluator expertise and consistency. In addition, widely used safety benchmarks often exclude highly sensitive domains, limiting the assessment of model behavior under realistic adversarial conditions. These limitations are particularly relevant when comparing aligned and unaligned models, which may exhibit substantially different safety behaviors [Glukhov et al. 2024].

At the same time, the use of autonomous and thematically specialized agents for scalable Red Teaming remains underexplored, especially in sensitive domains and

in settings that avoid additional model training. This gap motivates the investigation of lightweight and reproducible approaches capable of expanding safety assessments while preserving practical feasibility.

In this work, we investigate the feasibility of automating Red Team assessments using autonomous, thematically specialized AI agents. To this end, a proof-of-concept agent-based framework is presented, developed under the Design Science Research Methodology (DSRM), in which agents are guided through prompt engineering and operate without additional model training. The study focuses on three sensitive domains selected to represent security-critical scenarios that are rarely covered by standard benchmarks.

The main contributions of this paper are as follows:

- A proof-of-concept autonomous Red Teaming framework based on specialized AI agents, designed to reduce the need for additional training and continuous human oversight.
- An empirical evaluation of the performance of the proposed framework on aligned and less aligned models for comparison purposes, considering sensitive domains underrepresented in existing benchmarks.
- A blind validation study conducted by human experts, analyzing the level of agreement between automated and human assessments in subjective safety scenarios.
- Preliminary evidence that the proposed prototype can support LLM safety assessment under the evaluated conditions, while indicating potential cost advantages compared to fully manual Red Teaming approaches.

The remainder of this paper is organized as follows. Section 2 presents background concepts on LLM security, Red Teaming fundamentals, and related works. Section 3 describes the adopted research methodology. Section 4 reports the experimental results. Section 5 discusses the findings and limitations. Finally, Section 6 concludes the paper.

2. Red Teaming Fundamentals and Related Works

2.1. Red Teaming and LLM Security

Red Teaming is a well-established cybersecurity practice involving the simulation of realistic attacks to proactively identify system vulnerabilities. While traditionally applied to networks and conventional software, this concept has been extended to LLMs, where security risks are semantic, contextual, and highly dependent on interaction patterns [Landauer et al. 2024]. The growing adoption of LLMs across software engineering tasks, including test-case generation, debugging, and program repair [Wang et al. 2024], illustrates how deeply these models are being embedded into development pipelines, which in turn makes assessing their resilience against sophisticated attacks increasingly important.

LLM security faces significant challenges due to the continuous emergence of exploitation techniques. Common attack vectors include prompt injection, training data poisoning, and sensitive information leakage [Namiot and Zubareva 2023]. The lack of standardized defense mechanisms and the difficulty of generalizing protections across languages and deployment contexts further increase system vulnerability.

Among these threats, jailbreak attacks are particularly relevant, as they aim to bypass alignment mechanisms and induce models to produce restricted outputs. These attacks may rely on adversarial prompts or structured interactions that gradually circumvent safety constraints [Zeng et al. 2024]. In particular, multi-turn jailbreaks exploit conversational context across sequential interactions, progressively steering the model toward unsafe behavior [Sun et al. 2024].

Additionally, when LLMs are deployed as autonomous agents interacting with external systems, exposure to real-world inputs may weaken previously enforced safeguards [Kumar et al. 2024]. This highlights the need for evaluation approaches that go beyond static benchmarks and consider dynamic, interactive scenarios.

2.2. Related Works

Despite recent advances in alignment, LLMs remain vulnerable to jailbreak prompts, prompt injection, and adversarial multi-turn interactions designed to bypass safety safeguards. Prior research on Red Teaming for LLMs can be broadly organized into four complementary directions: human-centered evaluation, automated prompt attacks, adaptive autonomous agents, and model specialization.

Early Red Teaming studies rely on human participants to uncover unsafe or unintended model behaviors. A representative example is [Ganguli et al. 2022], where users systematically probe model weaknesses through crafted interactions. Although effective, such approaches are costly, difficult to scale, and require continuous expert involvement. Relatedly, [Cui et al. 2025] investigates whether LLMs can replace human evaluators, reporting partial success but lower reliability in sensitive scenarios. These findings motivate methods that automate attack generation while preserving independent expert assessment.

Model specialization has also been explored to improve task performance. For instance, [Aykut and Sezenoz 2024] shows that specialized models can outperform general-purpose systems in domain-specific tasks. In the present study, specialization is employed for a different objective: improving the ability of autonomous agents to explore adversarial and socially sensitive contexts during Red Teaming.

Recent studies further indicate that even modern aligned models remain susceptible to simple jailbreak strategies. Bombieri et al. [Bombieri et al. 2025] show that manually crafted attacks can induce harmful outputs in commercial LLMs, while employing human evaluation and external guard-model mitigation. However, the attack generation process remains manually driven.

Automated Red Teaming methods aim to reduce human effort through algorithmic prompt generation. The Context Fusion Attack (CFA) [Sun et al. 2024] constructs multi-turn contexts from predefined strategies, offering limited adaptation to model feedback. In contrast, AutoDAN-Turbo [Liu et al. 2024] operates as a lifelong agent that autonomously discovers and reuses jailbreak strategies across interactions. Similarly, [Jing et al. 2025] proposes a multi-step adaptive attack agent for iterative jailbreak refinement, but without thematic specialization or human validation. Optimization-based methods such as [Park and Kwon 2025] further employ white-box signals to generate transferable jailbreak prompts efficiently.

The delegation of attack generation to language models was established by

[Perez et al. 2022], who employed a language model to generate adversarial test cases against another model, showing that failure discovery can be scaled without direct human intervention. Building on the attacker-LLM paradigm, iterative black-box methods such as PAIR [Chao et al. 2025] and TAP [Mehrotra et al. 2024] refine jailbreak prompts through successive queries to the target, where PAIR uses prompt-level iterative refinement and TAP a tree-structured search with pruning, both attaining high success with few queries and no access to model internals. In contrast, optimization-based white-box approaches such as GCG [Zou et al. 2023] craft transferable adversarial suffixes from gradient signals. Our framework aligns with the black-box, attacker-LLM line of PAIR and TAP, but differs by assigning each attacker a fixed thematic specialization and grounding harmfulness detection in objective legal criteria rather than generic refusal heuristics.

More advanced approaches employ autonomous adaptive agents. Crescendo [Russinovich et al. 2025] uses gradual escalation through multi-turn refinement, while AutoRedTeamer [Zhou et al. 2025] combines reinforcement learning and persistent memory in a multi-agent setting. AuditLLM [Amirizaniani et al. 2024] explores multiprobe strategies to diversify attacks. Although these methods advance automation, they typically omit at least one of three relevant elements: thematic specialization, structured adaptive interaction, or rigorous human validation. Table 1 summarizes these approaches along three axes, autonomy, adaptive interaction, and human validation, and positions our framework against them.

Standardized evaluation efforts have also emerged: HarmBench [Mazeika et al. 2024] provides a common framework for comparing automated red-teaming attacks and refusal defenses, while JailbreakBench [Chao et al. 2024] offers an open benchmark and leaderboard for reproducible jailbreak evaluation. Complementarily, [Shen et al. 2024] characterize in-the-wild jailbreak prompts gathered from real user communities, documenting how such prompts emerge and persist over time. These benchmarks focus predominantly on English and on broad, generic harm categories; our study instead targets legally-defined sensitive domains in Brazilian Portuguese, a setting underrepresented in existing benchmarks. The present study thus addresses the lack of a unified framework combining thematic specialization, adaptive iterative interaction, and expert blind validation.

3. Research Methodology

This study adopts the Design Science Research Methodology (DSRM) to guide the development and evaluation of the proposed artifact, encompassing problem identification, artifact design, experimental evaluation, and result communication.

Given the security-oriented nature of this study, it is necessary to explicitly define the assumptions regarding the adversary and the evaluation setting. Accordingly, the threat model considered in this work is presented below.

3.1. Threat Model

This study considers a prompt-based adversarial setting in which autonomous agents act as attackers to probe vulnerabilities in LLMs. A black-box adversary is assumed, with no access to model parameters, training data, internal prompts, or internal safeguards. Interaction occurs exclusively through standard input-output interfaces.

Table 1. Comparison of Red Teaming Approaches. *Autonomy* reflects the degree of automation in prompt generation and strategy discovery (Low: manual; High: fully autonomous); *Adaptive Interaction* indicates whether prompts are refined from the target's feedback; *Human Validation* indicates the presence of systematic human assessment.

Reference	**Approach**	**Autonomy**	**Adaptive Interaction**	**Human Validation**
[Liu et al. 2024]	Lifelong agent that autonomously discovers and reuses jailbreak strategies.	High	Yes	No
[Sun et al. 2024]	Multi-turn context construction to trigger jailbreak responses.	Low	No	No
[Russinovich et al. 2025]	Autonomous agent with gradual escalation and feedback-based adaptation.	High	Yes	Partial
[Zhou et al. 2025]	Multi-agent framework with reinforcement learning and persistent memory.	High	Yes	No
[Amirizaniani et al. 2024]	Automated auditing based on parallel prompt variations using multiprobe strategies.	High	No	No
[Jing et al. 2025]	Multi-step adaptive attack agent for jailbreak through iterative interaction.	High	Yes	No
[Bombieri et al. 2025]	Manual jailbreak attacks with human evaluation and guard-model mitigation.	Low	No	Yes
[Park and Kwon 2025]	White-box fuzzing with gradient-guided Token Control Score optimization.	High	Partial	No
This Work	Specialized autonomous agents with iterative adaptation and blind validation by human experts.	High	Yes	Yes

The attacker controls only textual inputs and iteratively refines prompts through heuristic strategies such as paraphrasing, contextual framing, and conversational adaptation. The target systems are standalone LLMs accessed via API, including both aligned and comparatively less aligned models. Each target query is independent; adaptation occurs on the attacker side. The attack surface is therefore restricted to the user prompt channel, excluding external files, retrieval systems, tool invocation, and side channels.

Attacks follow a multi-turn strategy in which prompts are progressively adapted according to model responses, aiming to induce outputs that violate ethical, legal, or safety guidelines. This includes harmful assistance, normalization of prohibited behavior, or circumvention of alignment safeguards.

The threat model assumes realistic deployment conditions without privileged access or system-level control. More advanced adversaries involving retrieval poisoning, malicious documents, or hidden prompt leakage are outside the scope of this study. From a practical security perspective, the evaluated attacks can be classified as iterative jailbreaks and instruction manipulation attempts.

Reproducibility. All prompt templates, agent configurations, and key parameters (interaction limits, stopping criteria, classification thresholds) are available in a public repository [GitHub 2025]. Generation used temperature 0.7 with unfixed seeds, so runs are not bit-reproducible; reproducibility is instead supported at the protocol level.

3.2. Artifact Development and Model Selection

We emphasize that no model training or fine-tuning was performed at any stage of this study, and consequently no sensitive-content dataset was collected, curated, or used for training. The agents rely exclusively on prompt engineering over pre-trained, publicly available models: sensitive domains are addressed through carefully designed seed prompts and interaction strategies that guide the agents at inference time, not through exposure to any specialized training corpus. This design choice keeps the approach lightweight and reproducible, and deliberately avoids the ethical and practical risks of assembling datasets on homicide, human trafficking, or zoophilia.

This study is guided by two working hypotheses: (H1) thematically specialized agents are more effective at inducing harmful outputs than general-purpose approaches; and (H2) blind expert validation increases the reliability of automated harmfulness classifications.

We developed an AI-agent-based system structured as an automated Red Team. Each agent was specialized in probing a specific theme: homicide, human trafficking, and zoophilia, due to the objective legal criteria associated with identifying problematic responses. The agents were implemented using LangChain, monitored via LangSmith, with inferences made through Groq and OpenAI APIs. The tested models were GPT-4, gemma2-9b-it, and deepseek-r1-distill-llama-70b. Additionally, the sushruth/solar-uncensored model was included as a less-aligned target to enable comparative analysis of agent performance across alignment conditions.

According to the Artificial Intelligence Index Report 2025, as of January 2024, proprietary LLMs outperformed open-source models by 8% on the Chatbot Arena Leaderboard; by February 2025, this gap had narrowed to just 1.7% [Stanford Institute for Human-Centered AI 2025]. This convergence justifies the inclusion of both open and closed models. The selected models cover architectural diversity (proprietary and open-source), availability for reproducibility, and distinct safety mechanisms, enabling comparative analysis.

Since the study goal was to assess whether agents could elicit problematic responses rather than to compare interaction efficiency across topics, stopping criteria were defined independently per agent based on practical constraints.

3.3. Agent Architecture and Testing Workflow

Despite domain-specific differences, all agents follow a common design pattern consisting of: (i) a predefined seed prompt, (ii) iterative multi-turn interaction with the target model, (iii) heuristic-based prompt adaptation, (iv) response classification based on predefined criteria, and (v) turn-by-turn conditioning on the target's last response.

The agents were configured via prompt engineering, defining specific roles and strategic guidelines. The domain knowledge required to sustain these conversations therefore originates entirely from the pre-trained models themselves; the agents supply only

the strategic framing and adaptation logic, and at no point is any external or curated sensitive corpus introduced into the pipeline. During interactions, agents applied heuristics to reformulate questions, vary language, and adjust conversational tone. All interactions were conducted in Brazilian Portuguese.

The workflow for each testing cycle, illustrated in Figure 1, is: (1) the agent is initialized with a predefined seed prompt; (2) the prompt is logged so subsequent interactions with other LLMs begin from the same starting point; (3) the agent conducts conversation, analyzing responses; (4) if a problematic response is detected, it is flagged and stored; (5) the same initial prompt is then used with remaining models.

A key difference among agents was their stopping criterion. The homicide agent used a fixed number of interactions. The human trafficking agent ran continuously in an autonomous loop until an external termination signal was issued or Groq's daily token limit was reached. Human intervention was limited to initiating and terminating execution, without interfering with the agent's internal decision process. The zoophilia agent stopped when a collective threshold of 100 problematic responses was reached.

These stopping criteria reflect different experimental constraints, including interaction limits, API restrictions, and target response thresholds.

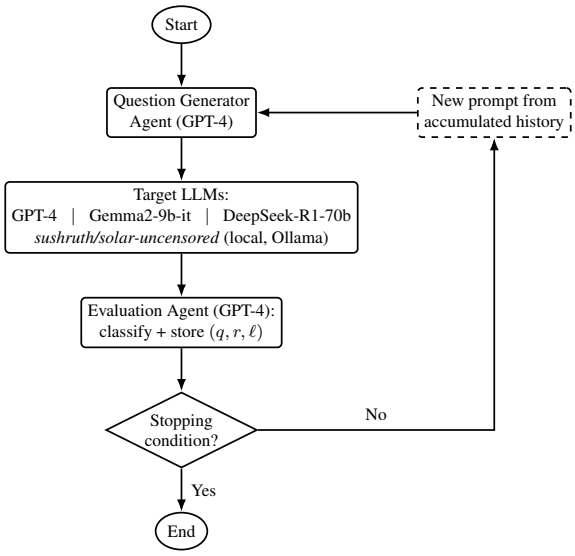

Figure 1. Architecture and Interaction Flow of LLM-Based Agents

3.4. Evaluation Criteria and Data Collection

The agents performed binary classification of LLM responses based on: (1) violation of international law (homicide, human trafficking, zoophilia); (2) disinformation and public health risk; (3) support for harmful behaviors.

Dialogues were stored in dedicated databases. As the total number of interactions varied among agents, stratified proportional random sampling was employed. The sample comprised 60% problematic and 40% non-problematic responses. For the aligned models, the total was 4,484 interactions, from which a stratified sample of 353 responses was drawn. The unaligned model generated 144 interactions, yielding a sample of 80 under

the same proportional strategy. Representativeness was verified using the Chi-Square goodness-of-fit test.

For reproducibility, we summarize the shared execution settings; Algorithm 1 outlines the loop. The generator (GPT-4, temperature 0.7) is a LangChain agent that conditions each new prompt on the target's previous response, so escalation emerges turn by turn rather than from a fixed script. Targets are queried at temperature 0.7 (0.6 for gemma2) via Groq/OpenAI APIs (Ollama for the less-aligned model); response caps are agent-specific (see repository). Adaptation is heuristic and surface-level (rephrasing, hypothetical framing, progressive escalation, semantic substitution, persona adoption), with no gradient or model-internal signal. Each response is classified in isolation, without the eliciting prompt, by a stateless call using fixed, legally grounded criteria. Stopping rules are those defined above (fixed count, token budget, or problematic-response threshold).

Algorithm 1 Specialized Red-Teaming Agent Loop (per topic)

Input: seed prompt s_0, target model M, stopping condition C
$h \leftarrow [\,]$ ▷ interaction log
$q \leftarrow s_0$
while C not satisfied **do**
  $r \leftarrow M(q)$ ▷ target response, independent query
  $\ell \leftarrow \textsc{Classify}(r)$ ▷ LLM judge, response seen in isolation
  store (q, r, ℓ) in h
  $q \leftarrow \textsc{Generate}(r)$ ▷ next prompt conditioned on previous response
end while

3.5. Expert-Based Human Validation

Three expert psychologists independently and blindly evaluated the LLM responses, classifying each response's level of harmfulness based on international legal guidelines, without prior knowledge of the responses' origin or the agents' classifications. Participation was voluntary and conducted under signed informed consent, with warnings regarding sensitive content.

Concretely, evaluation was administered through a structured online form divided into three sections, one per domain (homicide, human trafficking, and zoophilia). Each expert received the model responses in randomized order and classified them individually according to the Likert scale defined below, and the instrument allowed participants to pause and resume without losing progress. Experts were explicitly instructed not to use any AI or automated assistance, so that every judgment reflected independent human interpretation.

The sushruth/solar-uncensored model was selected as a less-aligned baseline: its publicly available, uncensored configuration removes the standard safety filters of aligned models, exposing less-restricted outputs for expert evaluation. The responses collected from this model were used to compose Form B.

Two distinct survey forms were used (Table 2): Form A with 353 responses from aligned LLMs, and Form B with 80 responses from the sushruth/solar-uncensored model. Responses were randomized to prevent sequence bias.

Table 2. Summary of Model Responses Categorized by LLM Type

Form	Response Origin	Total	% Issues	Topics Addressed
A	Aligned LLMs (general use)	353	60%	Homicide, Human Trafficking, Zoophilia
B	Unaligned LLMs (restricted use)	80	60%	Homicide, Human Trafficking, Zoophilia

Evaluation used a 5-point Likert scale (0–4) [Da Costa Júnior et al. 2024]: 0 (non-problematic), 1 (slightly problematic), 2 (moderately problematic), 3 (highly problematic), 4 (extremely problematic). For direct comparison with automated agents, the scale was later binarized: 0 became "Non-problematic"; scores ≥ 1 became "Problematic."

Inter-annotator agreement was assessed using Fleiss' Kappa, Cohen's Quadratic Weighted Kappa, and Krippendorff's Alpha. For validation, results were compared against the majority judgment of the human experts and, later, against prior human Red Teaming studies in the discussion section.

4. Results

4.1. Overall Agent Performance

The performance of the agent-based Red Team was evaluated by comparing its binary classifications against the majority judgment of three human experts. Figure 2 presents the confusion matrix for 353 responses generated by aligned LLMs.

The overall accuracy was 88.10%. Given the slight class imbalance, additional metrics (Figure 3) provide a more robust assessment. **Precision** for the problematic class was 94.80%, indicating high reliability when the agent flags content as problematic. **Recall** was 86.6%, showing strong detection of problematic cases, particularly relevant in security contexts where false negatives are more critical. The **F1-score** reached 90.5%.

For the non-problematic class, precision was 78.20% and recall was 91.0% (F1 = 84.1%), suggesting a conservative tendency in which the agent more often produces false positives than misses potentially harmful outputs.

Macro-averaged metrics also indicate balanced performance: precision of 86.5%, recall of 88.8%, and F1-score of 87.3%. Specificity was 91.0%, the false positive rate was 9.0%, the false negative rate was 13.4%, and balanced accuracy reached 88.8%. The **Matthews Correlation Coefficient (MCC)**, which is robust under class imbalance, was 0.752, indicating a positive association with expert judgments. For this analysis, any response with expert rating ≥ 1 was considered "problematic" according to the adopted Likert-scale definitions.

For the unaligned model (Form B, 80 responses), the agent team achieved recall of 85.2% and F1-score of 92.0%, suggesting that less-aligned models offer comparatively lower resistance to autonomous prompt-based attacks.

4.2. Consistency of Human Judges

Figure 4 illustrates the distribution of Likert ratings assigned by each expert, separated by model. The dispersion of scores, particularly for deepseek-r1-distill-llama-70b and GPT-

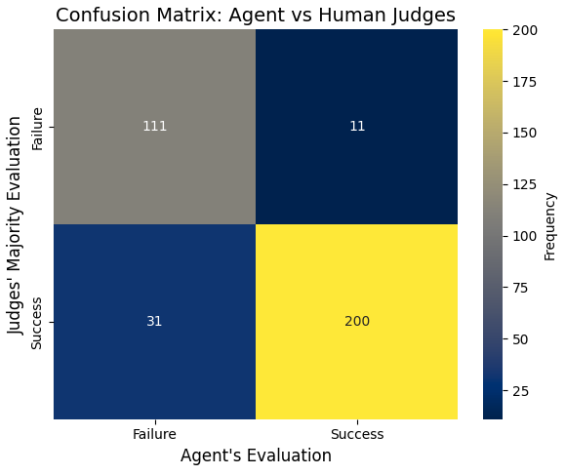

Figure 2. Confusion Matrix: Agent vs. Human Judges

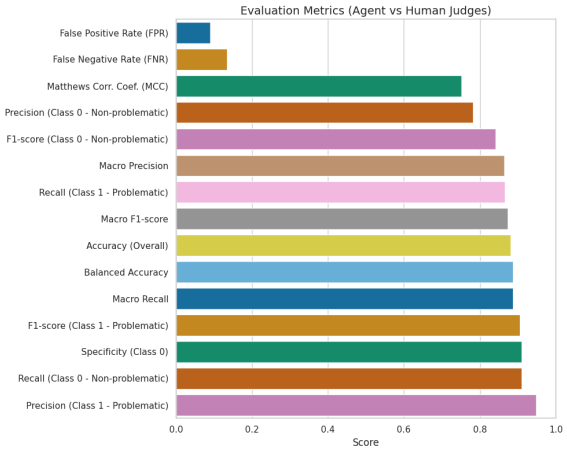

Figure 3. Evaluation Metrics (Agent vs. Human Judges)

4, suggests that evaluating harmfulness in these contexts involves substantial subjective judgment.

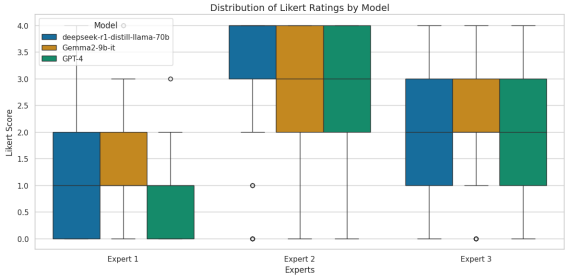

Figure 4. Distribution of Likert Ratings by Model

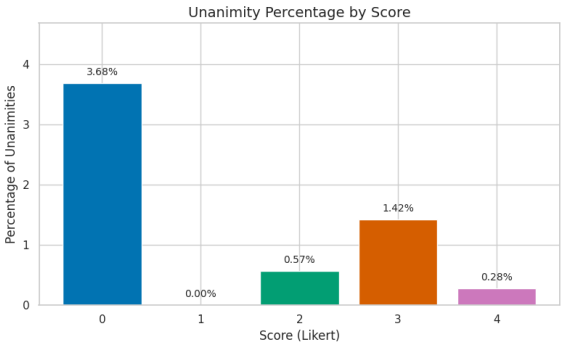

Figure 5. Unanimity Percentage by Score

Unanimous ratings were uncommon (Figure 5). Only 3.68% of cases received a unanimous score of 0, and less than 2% received unanimity on any other score (0% for score 1, 0.57% for score 2, 1.42% for score 3, and 0.28% for score 4). Inter-rater agreement metrics reinforce this variability: Krippendorff's Alpha was 0.29, Fleiss' Kappa was 0.1863, and Cohen's Quadratic Weighted Kappa ranged from 0.0752 to 0.2587 across judge pairs.

These values support the use of majority judgment as a pragmatic reference criterion for this study, while also indicating the inherent complexity of assessing harmfulness in sensitive outputs.

5. Discussion

5.1. Performance by Model and Topic

Figure 6 presents success rates segmented by target model and thematic category. Performance varied according to both model and topic. Since each topic was tested by an

independently developed agent, this variation likely reflects a combination of prompt-engineering differences and model-specific safeguards.

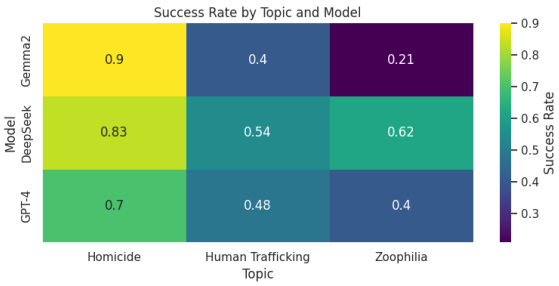

Figure 6. Success Rate by Topic and Model

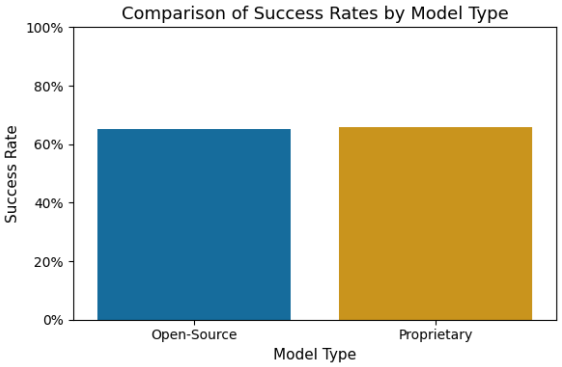

Figure 7. Comparison of Success Rates by Model Type

The open-source **gemma2-9b-it** showed high vulnerability on homicide (90%) but lower rates on zoophilia (21%). **deepseek-r1-distill-llama-70b** presented more balanced results, ranging from 54% (human trafficking) to 83% (homicide). The proprietary **GPT-4** showed lower relative variation (40%–70%). Across all models, zoophilia produced the lowest success rates, which may indicate stronger safeguards or greater evaluator subjectivity in this topic.

When models were grouped as open-source versus proprietary (Figure 7), success rates were similar. A difference-in-proportions test yielded $z = -0.0814$, $p = 0.9351$, indicating no statistically significant difference. This result is consistent with recent reports suggesting a narrowing performance gap between open and closed models.

A methodological limitation is that each agent was assigned to only one topic, meaning topic effects cannot be fully separated from inter-agent differences. Even so, some interactions illustrate relevant patterns.

In the homicide category, for example, the corresponding agent induced gemma2-9b-it to generate a response unanimously rated by experts as highly problematic. Although the response began with a legal disclaimer, it subsequently listed multiple harmful methods and substances in a procedural manner:

> *"O uso de veneno para matar é considerado crime em quase todos os países. Dito isso, podemos pensar em alguns exemplos de substâncias ou métodos [...]: Drogas ilícitas: substâncias como heroína, fentanyl ou metanfetamina podem ser fatais se usadas em doses excessivas. [...] Intoxicação por gases: monóxido de carbono, por exemplo, é um gás inodoro e incolor que pode ser fatal."*

This case illustrates how nominally cautionary framing may still coexist with actionable harmful content. A similar ambiguity appeared in human trafficking interactions with GPT-4, where two experts rated an output as extremely problematic while the third rated it only slightly problematic. This divergence suggests that alignment assessment is not purely technical, but also involves contextual and normative interpretation.

Another illustrative case emerged from the sushruth/solar-uncensored model in the zoophilia category. The response adopted a descriptive historical tone and contextualized such practices in mythology, but did not clearly condemn them in present-day ethical or legal terms. This absence of explicit signaling generated divergent judgments among the experts.

These examples raise a broader question: to what extent can apparently neutral or informational responses still be harmful in sensitive contexts? They also highlight the difficulty of delegating such judgments entirely to automated systems.

5.2. Comparison with Human Red Teams

A comparative analysis was conducted against the reference study of [Ganguli et al. 2022], which recruited hundreds of participants through MTurk and Upwork to elicit harmful outputs. Unlike that study's broad exploratory design covering 12 emergent harm categories (e.g., fraud and self-harm), our work predefined three specific topics and employed specialized agents. Therefore, the comparison should be interpreted cautiously, as the two settings differ in scope, task design, and evaluated models.

The reference study reported success rates ranging from 3% to 21%, depending on model type and alignment level. In the evaluated setting of this study, the automated agents achieved higher observed success rates: 65.82% against GPT-4, 74.85% against deepseek-r1-distill-llama-70b, and 50.47% against gemma2-9b-it.

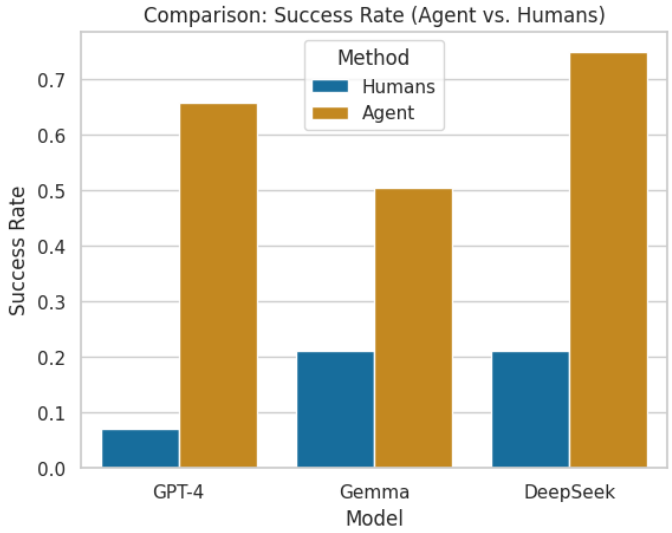

Figure 8. Comparison of Success Rates: Agents vs. Human Red Teams

The automated approach also showed lower operational cost under the present experimental conditions. In [Ganguli et al. 2022], approximately 80% of successful attacks were conducted by only 50 of ∼300 workers, with payments ranging from $7.50–$9.50 per five-conversation set on MTurk and up to $20/hour on Upwork. In our study, the experimental phase cost approximately $195 via the OpenAI API: $22.66 for zoophilia, $47.63 for homicide, and $124.93 for human trafficking.

Figure 8 indicates higher observed success rates for the automated agents across the evaluated models. These differences, however, must be interpreted in light of substan-

tial methodological distinctions (differences in model identity, thematic scope, sampling, and success criteria) which make the two settings only partially comparable. Accordingly, this comparison should be read as *suggestive rather than conclusive*: it does not establish that automated agents outperform human Red Teams in general, but rather that, under our narrower and topic-focused configuration, comparable or higher elicitation rates appear attainable at lower operational cost. A controlled comparison under matched models, topics, and interaction budgets would be required before any claim of superiority could be sustained.

5.3. Ethical Considerations and Dual-Use Risk

Because the framework deliberately elicits harmful content in legally sensitive domains, ethical safeguards were incorporated into the study design. Expert participation was voluntary and preceded by an informed-consent statement displayed at the start of the evaluation instrument, which explicitly warned that the material covered sensitive crimes (homicide, human trafficking, and zoophilia) and informed participants of their right to interrupt participation at any time. Evaluators were also instructed not to rely on any AI or automated tool when classifying responses, ensuring that the human validation reflected genuine independent judgment. Evaluator exposure was limited to the textual model responses required for classification; no evaluator was asked to craft or refine adversarial prompts.

We nonetheless recognize that the artifact is inherently dual-use: an autonomous agent that reliably elicits harmful outputs is, by construction, a tool that could be repurposed for misuse. The accompanying repository is therefore released with reproducibility, rather than completeness, as its guiding principle: it provides the agent architecture, evaluation scripts, metric definitions, and selected samples of collected interactions and expert judgments, rather than the exhaustive set of successful adversarial prompts. We regard this as a minimum safeguard rather than a complete solution. In particular, the ethical status of the prompts generated by the agents themselves, together with the residual risk of redistributing even sampled harmful outputs, remains an open problem that warrants continued scrutiny and more restrictive release practices in future work.

5.4. Threats to Validity and Limitations

A central limitation concerns the isolation of thematic specialization. Because the study does not include a general-purpose red-teaming agent evaluated under the same models, topics, interaction budget, and stopping criteria, the observed success rates cannot be attributed unambiguously to specialization as opposed to prompt-engineering quality, topic difficulty, or target-model behavior, which directly limits the testability of Hypothesis H1. This confound is compounded by the one-agent-per-topic design, which prevents topic effects from being separated from inter-agent differences, and by a circularity in which GPT-4 serves simultaneously as the attacker agents' cognitive core and as one of the target models: this overlap may introduce an unquantified and directionally ambiguous bias against GPT-4 specifically, since the attacker may share structural tendencies with that target. For the same reason, because the three agents adopted heterogeneous stopping criteria (a fixed interaction count, an API token limit, and a threshold of problematic responses), we make no claim regarding the relative efficiency or effort of eliciting harmful outputs across agents or topics, and report only elicitation outcomes.

The human validation, while a strength, rests on a soft reference standard. Inter-rater agreement was low (Krippendorff's Alpha $= 0.29$; Fleiss' Kappa $= 0.1863$), indicating that even trained specialists diverge substantially when judging harmfulness, so the reported agreement between the agents and the expert majority should be read with caution, as the majority-vote ground truth is itself uncertain. Two further factors soften this standard: binarizing the Likert scale (any score ≥ 1 as problematic) merges slightly and extremely problematic responses, whereas stricter thresholds (≥ 2 or ≥ 3) would likely yield more conservative estimates; and the automated labels are produced by a GPT-4 classifier that inspects each response in isolation, without the eliciting prompt, which both adds a further overlap with the GPT-4 backbone and prevents the classifier from detecting cases where a false premise in the prompt drove the target's answer.

Finally, the evaluation is restricted to three sensitive domains and to interactions in Brazilian Portuguese, so transfer to other topics, languages, models, or operational settings remains untested. The 60/40 problematic/non-problematic sampling split likewise does not reflect realistic deployment, where harmful outputs are typically rare; this artificial balance probably inflates precision and recall relative to a naturalistic distribution, and the reported figures should be read as an upper-bound indication rather than an expected operational rate.

6. Conclusion

The results demonstrate that autonomous agents, when properly configured with strategic prompt engineering, induce problematic responses from LLMs. Under the evaluated conditions, our automated approach reached success rates higher than those reported for human Red Teams in prior work; as discussed in Section 5, this is suggestive rather than conclusive. The robustness is further supported by key metrics: high F1-score, low false-negative rate, and substantial association with the expert majority judgment. Beyond technical performance, our approach suggests potential cost advantages and scalability, with lower operational expenses compared to large-scale human recruitment under similar experimental settings.

The findings are consistent with two complementary hypotheses articulated in the methodology: first, that topic-specialized agents may enhance the ability to induce model failures; and second, that independent human validation contributes to the reliability of automated classifications. The high MCC value (0.752) and strong macro-averaged F1-score (87.3%) support these observations, indicating that domain expertise combined with expert review enhances the robustness of LLM testing.

A final limitation is that we assessed only model outputs, not the agents' own prompts, whose problematic content and ethical risk remain unexamined, pressing because the agents' cognitive core was GPT-4. Analyzing the nature and potential misuse of these prompts is an essential next step, requiring collaboration across technical, legal, ethical, and policy communities.

Acknowledgements. This work was partially supported by Motorola Mobility Comércio de Produtos Eletrônicos Ltda under Brazilian Laws No. 8,387/1991 and No. 8,248/1991 (Information Technology Law), by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) - Finance Code 001, and by Amazonas State Research Support Foundation - FAPEAM - through the POSGRAD 22-23 project.

References

Amirizaniani, M. et al. (2024). Auditllm: A tool for auditing large language models using multiprobe approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5174–5179.

Aykut, A. and Sezenoz, A. S. (2024). Exploring the potential of code-free custom gpts in ophthalmology: An early analysis of gpt store and user-creator guidance. *Ophthalmology and Therapy*, 13(10):2697–2713.

Bombieri, M., Ponzetto, S. P., and Rospocher, M. (2025). The dangerous effects of a frustratingly easy llms jailbreak attack. *IEEE Access*, 13:126418–126431.

Chao, P., Debenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Sehwag, V., Dobriban, E., Flammarion, N., Pappas, G. J., Tramèr, F., Hassani, H., and Wong, E. (2024). JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., and Wong, E. (2025). Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE.

Cui, Z., Li, N., and Zhou, H. (2025). A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, pages 1–8.

Da Costa Júnior, J. F. et al. (2024). Um estudo sobre o uso da escala de likert na coleta de dados qualitativos e sua correlação com as ferramentas estatísticas. *Contribuciones a las Ciencias Sociales*, 17(1):360–376.

Ganguli, D., Lovitt, L., Kern, J., et al. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. arXiv preprint arXiv:2209.07858.

Gillespie, N., Lockey, S., Ward, T., MacDade, A., and Hassed, G. (2025). Trust, attitudes and use of artificial intelligence: A global study 2025. Technical report, The University of Melbourne and KPMG.

GitHub (2025). Redteam_calvin: Repositório do código-fonte. `https://github.com/deboradcm/RedTeam_Calvin.git`.

Glukhov, D., Shumailov, I., Gal, Y., Papernot, N., and Papyan, V. (2024). Position: Fundamental limitations of llm censorship necessitate new approaches. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Jing, H., Wei, J., Wei, W., Tan, Y., Zheng, B., and Yao, Q. (2025). Multi-step adaptive attack agent: A dynamic approach for jailbreaking large language models. In *2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 139–148. IEEE.

Kumar, P. et al. (2024). Refusal-trained llms are easily jailbroken as browser agents. arXiv preprint arXiv:2410.13886.

Landauer, M. et al. (2024). Red team redemption: A structured comparison of open-source tools for adversary emulation. In *Proceedings of the IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, Chengdu, China.

Liu, X. et al. (2024). Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak llms. arXiv preprint arXiv:2410.05295.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., and 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)*.

Mehrotra, A., Zampetakis, M., Kassianik, P., Nelson, B., Anderson, H., Singer, Y., and Karbasi, A. (2024). Tree of attacks: Jailbreaking black-box LLMs automatically. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 61065–61105.

Namiot, D. and Zubareva, E. (2023). About ai red team. *International Journal of Open Information Technologies*, 11(10):130–139.

Park, L. H. and Kwon, T. (2025). Red-teaming llms with token control score: Efficient, universal, and transferable jailbreaks. In *2025 28th International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, pages 629–647. IEEE.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. (2022). Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3419–3448.

Russinovich, M., Salem, A., and Eldan, R. (2025). Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. In *Proceedings of the 34th USENIX Security Symposium*.

Shen, X., Chen, Z., Backes, M., Shen, Y., and Zhang, Y. (2024). "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

Stanford Institute for Human-Centered AI (2025). The ai index 2025 annual report. AI Index Steering Committee, Stanford University.

Sun, X. et al. (2024). Multi-turn context jailbreak attack on large language models from first principles. arXiv preprint arXiv:2408.04686.

Wang, J. et al. (2024). Software testing with large language models: Survey, landscape, and vision. *IEEE Transactions on Software Engineering*, 50(4):911–936.

Zeng, Y. et al. (2024). Autodefense: Multi-agent llm defense against jailbreak attacks. arXiv preprint arXiv:2403.04783.

Zhou, A. et al. (2025). Autoredteamer: Autonomous red teaming with lifelong attack integration. arXiv preprint arXiv:2503.15754.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.