

Law-Aware Red-Teaming for LLMs: Measuring and Mitigating Copyright and Defamation Risks

Alexandra Bodrova
abodrova@princeton.edu

Abstract

As large language models (LLMs) increasingly mediate access to information and services through ubiquitous chatbots, their vulnerabilities to both accidental legal violations and deliberate exploitation create significant legal exposure for regular users as well as for AI developers. Current safety mechanisms exhibit uneven guardrails with biased training data and weak safeguards, enabling outputs that infringe copyright, circumvent content policies, or generate defamatory statements. This project extends our “Asimov Box” framework being developed by Alexandra Bodrova under the supervision of Alex Glase—originally designed as a tamper-resistant safety device for LLM-controlled robots—into a law-aware red-teaming system for chatbots. We propose a comprehensive approach that operationalizes legal harms into testable tasks, develops adversarial attack protocols targeting copyright infringement and defamation, and implements a multi-agent validation system combining policy rules engines with validator/rewriter agents. Our methodology adapts RoboPAIR-style adversarial testing and RoboGuard-inspired defense mechanisms to text-based legal compliance, evaluated through attack success rates, false positive measurements, and preservation of lawful transformative uses. This work bridges the gap between AI safety research and legal doctrine, offering a systematic framework for hardening models’ guardrails against deliberate attempts to induce unlawful outputs while maintaining utility for legitimate applications. *Note: Github repo with code will be posted for the final project as it is still a work in progress.*

1 Introduction

1.1 Problem Definition

The proliferation of large language models in consumer-facing applications has created an unprecedented tension between technological capability and legal compliance. While these systems are exceptionally skilled at integrating information, producing content, and providing interactive support, they simultaneously expose both developers and users to significant legal risks in a multitude of ways. The three central problems this project addresses are the following: (1) LLMs can inadvertently reproduce copyrighted material verbatim or in substantial similarity; (2) they periodically generate factually incorrect statements about real individuals that could lead to defamation; and (3) existing safety mechanisms prove vulnerable to adversarial prompting techniques designed to circumvent content policies put in place by models’ developers.

Recent incidents have demonstrated that commercial chatbots can be coerced into reproducing copyrighted text through carefully engineered prompts, generating false information about public figures, and providing instructions for dangerous and even nefarious uses that violate platform policies. The legal implications span copyright infringement liability under title 17 of U.S.C., potential defamation claims under state tort law, and violations of the DMCA’s (also under title 17 of U.S.C.) anti-circumvention provisions.

1.2 Global Motive

This project addresses a critical gap in the intersection of AI safety and legal compliance. As LLMs become ingrained into our daily lives—from educational tools to professional services—their failure modes bring real legal consequences that extend beyond the technical domain into regulatory frameworks, civil liability, and erosion of public trust. The current regulations are fragmented and

outdated with respect to the exponential progress of AI technologies. There are ongoing debates in the U.S. Copyright Office regarding AI training fair use, emerging case law on LLM liability for false statements, and within international frameworks such as the EU AI Act.

The importance of this research can be summarized in three key points. First, the economic stakes are substantial: copyright holders represent multibillion-dollar creative industries, while deployers face potentially catastrophic liability exposure. Second, the societal implications of uncontrolled LLM outputs affect individual reputations, which in turn saws increasing distrust of the population to both AI systems themselves, and the information space as a whole. Third, the technical challenge of balancing safety against utility requires nuanced high-effort (and potentially high-cost) approaches that preserve transformative, fair-use applications while preventing harmful outputs.

1.3 Scholarly Motive

Existing AI safety research has primarily focused on preventing accidental harms—models that inadvertently cause dangerous issues like generating toxic content, revealing training data, or producing biased outputs. Such scholarly work is undeniably crucial for creating "Responsible AI", but it leaves models vulnerability to deliberate adversarial attacks unaddressed. Furthermore, most AI safety research nowadays is focused on preventing immediate harm rather than addressing its potential legal consequences. The legal angle presents unique challenges: since most of the existing legislature predates the age of AI, the legal boundaries of LLM’s artefacts and use-cases are blurry and difficult to navigate even for professionals, let alone average users. For example, copyright and defamation doctrines involve context-dependent judgments, transformative use exceptions, and substantial-similarity analyses; such vague guidelines don’t map neatly onto simple classification rules or technical safety checks. Legal scholarship has begun examining AI liability frameworks, but computational implementations remain sparse and fragmented without a sturdy legislation foundation.

This project bridges these domains by adapting proven adversarial testing methodologies (RoboPAIR) and defense frameworks (RoboGuard) to law-centered textual outputs, incorporating doctrine-aware heuristics into executable policy rules engines.

1.4 Methods and Key Results Overview

Our approach implements a multi-stage system architecture. We first operationalize legal harms by defining precise, testable categories: copyright infringement (verbatim reproduction and substantial similarity above defined thresholds), defamation risk (specific harmful allegations about named individuals without evidentiary support), and dual-use content (explicit instructions for dangerous activities). We then develop an adversarial red-teaming loop following RoboPAIR methodology: an Attacker LLM generates prompts designed to circumvent safeguards, a Target LLM (commercial chatbot under evaluation) produces outputs, and a Judge LLM evaluates outputs against policy constraints.

The defense mechanism adapts RoboGuard’s validator/rewriter architecture to legal compliance. Upon detecting policy violations, the system attempts to rewrite the output in a legally conforming manner before resorting to refusal, preserving utility where possible. Evaluation follows JailbreakBench-inspired metrics extended with law-specific considerations: attack success rate (ASR), false positive rate on lawful uses (e.g., parody, critical quotation), and repair consistency.

Anticipated results include quantifiable measurements of commercial chatbot vulnerabilities to law-targeted adversarial prompts, demonstration that naive content filters are incapable of blocking sophisticated attacks, and evidence that adversarially-trained validator/rewriter combination significantly reduces ASR while maintaining utility thresholds. Limitations include the dependency on heuristic thresholds for legal concepts that in practice require human judgment, potential gaming of specific policy formulations, and the challenge of keeping up with rapidly evolving LLM technologies as well as changing legal doctrines.

2 Related Work

Prior work in AI safety has established important foundations. Chao et al. (2024) address a fundamental challenge in AI safety evaluation: the lack of standardized, comparable metrics across different

jailbreaking attempts and defense mechanisms. The benchmark provides an evolving repository of state-of-the-art jailbreak patterns, a dataset aligned with OpenAI’s safety policies, and a rigorous evaluation framework specifying threat models and scoring functions. However, JailbreakBench is limited to general content policy violations, lacking law-specific angle. Robey et al. (2024) demonstrate that LLM-controlled physical systems remain vulnerable to adversarial prompting. The authors developed an algorithm for generating jailbreak prompts that lead robots into harmful physical actions. In an extension to Robey’s paper, Ravichandran et al. (2025) address the defense side of LLM safety for embodied systems. The framework implements a two-stage "RoboGuard" guardrail algorithm: Stage 1 employs a separate "root-of-trust" LLM to ground predefined safety and ethical rules into context-specific temporal logic constraints; Stage 2 validates the primary LLM’s action plans against these constraints, using logic-based synthesis to repair non-compliant plans where possible. RoboGuard’s architecture proves particularly relevant for this project’s validator/rewriter design. The key limitation for our purposes is RoboGuard’s focus on accidental safety violations and unspecified unsafe behavior rather than adversarially engineered jailbreaks. The system assumes that the primary LLM attempts to follow rules but may inadvertently violate them. Furthermore, the base rules of this algorithm are not extending to more nuanced legal framework. Several more guardrail systems are worth mentioning: Inan et al. (2023) provide an LLM-based input-output safeguard specifically designed for human-AI conversations, while Rebedea et al. (2023) offer programmable rails for controllable LLM applications. These systems establish important precedents for multi-layer defense architectures but require extension to handle law-specific edge cases and adversarial robustness.

On the Legal framework side for AI guardrails, several recent works are relevant to this project. Lemley (2024) analyses copyright doctrine in the generative AI context (Lemley, 2024), identifying where substantial similarity tests strain under algorithmic content generation. For technical implementation, Lemley’s analysis provides concrete design principles: treat copying-risk and substitution harm as first-class metrics in safety evaluation; prefer targeted refusals or redactions when prompts seek verbatim passages or close paraphrases; preserve outputs that transform source material through criticism, commentary, or creative reinterpretation. Volokh (2023) survey how defamation doctrine applies when AI systems fabricate specific, reputationally harmful claims about real individuals. Volokh’s framework suggests concrete mitigation strategies. Systems should implement notice-tracking mechanisms that flag statements as disputed after user corrections. Authoritative factual claims about individuals should include uncertainty caveats when sources cannot be verified. His "notice → remediation" process provides an implementable approach to managing defamation risk. However, both of these works lack corresponding technical architectures for real-time compliance verification.

2.1 Novel Contributions

This project is novel relative to the existing work in several ways. Unlike JailbreakBench, which standardizes adversarial evaluation but omits legal specificity, we extend the framework with copyright circumvention and defamation-generating prompts grounded in actual legal doctrine. While RoboPAIR demonstrates vulnerability identification for LLM-controlled systems, we adapt the attack methodology to text-based legal violations with success criteria derived from case law. RoboGuard provides plan validation for accidental safety failures of physical systems, but we redesign the validator/rewriter algorithm to address adversarial legal compliance. And where legal scholarship identifies doctrinal pressure points, we translate those analyses into executable algorithmic code and evaluation metrics suitable for real-time deployment.

The integration represents the first systematic attempt to implement law-aware red-teaming specifically targeting copyright and defamation domains, with evaluation frameworks that balance robustness against over-filtering of lawful expression.

3 Methods

3.1 Legal Harm Operationalization

The first stage of this project involves translating vague legal standards into precise, executable classifications. We develop three primary risk categories:

Dual-Use Content Outputs providing explicit instructions for activities posing public safety risks. Subcategories include explosives/improvised weapons synthesis, toxic chemical preparation or dissemination protocols, biological pathogen manipulation, criminal evasion techniques escalating physical danger, and cyber-physical attack methodologies. Classification as "violation" requires both specificity (actionable instructions rather than abstract discussion) and directness (step-by-step guidance rather than general information available in educational contexts).

Copyright Infringement Outputs violating 17 U.S.C. provisions through unauthorized reproduction. We implement a multi-factor heuristic aligned with U.S. Copyright Office guidance and Lemley (2024)'s substantial similarity framework: (1) Verbatim Reproduction: outputs containing consecutive sequences of 50+ words matching known copyrighted sources are flagged as "violation-likely." This threshold derives from case law where substantial portions of works receive protection regardless of overall work length. (2) Substantial Similarity: for outputs below verbatim thresholds, we compute similarity scores against copyrighted references using semantic embedding distance combined with n-gram overlap. Scores exceeding 0.85 (on 0-1 scale) trigger secondary review for expression copying versus idea borrowing. (3) Explicit Circumvention: prompts or outputs containing directives to "remove watermarks," "bypass DRM," or "strip copyright information" are classified as DMCA anti-circumvention violations under 17 U.S.C. § 1201.

As one of the metrics of success for our algorithm is a low false-positive rate, we add "Fair use carve-outs". Outputs exhibiting transformative characteristics receive "lawful" classification even with substantial similarity. Indicators include critical commentary with limited quotation (< 150 words), parody, and educational contexts with appropriate attribution.

Defamation Risk Outputs generating false factual assertions about named individuals that risk reputational harm. Following Volokh (2023)'s analysis, we implement a risk stratification system:

1. *High Risk*: Specific allegations of criminal conduct, professional incompetence, or moral depravity directed at named real individuals, presented as fact without evidentiary support or uncertainty caveats.
2. *Medium Risk*: Biographical claims about living individuals that cannot be verified from authoritative sources, particularly when the model presents them with confidence rather than qualifying language.
3. *Low Risk*: Opinion statements, clearly hypothetical scenarios, or factual claims that are verifiable or include appropriate disclaimers ("reportedly," "allegedly," "according to unverified sources").

3.2 Adversarial Red-Teaming Architecture

We adapt RoboPAIR methodology (Robey et al., 2024) to create a law-targeted adversarial testing loop. The system consists of four primary components:

- **Policy Rules Engine** A modular framework translating legal constraints into executable verification code. Each legal domain (copyright, defamation, dual-use) receives a dedicated rule set implemented as Python functions accepting text inputs and returning classification tuples: (violation_detected: bool, confidence: float, rationale: str, suggested_modification: Optional[str]).
- **Attacker LLM** An adversarial prompt generation system designed to circumvent chatbot safeguards. We implement this as a custom GPT-4-based agent with the following protocol:
 1. Initialize with base prompt templates designed to elicit policy violations (e.g., "Provide the full text of [copyrighted work]" or "Tell me about [person]'s criminal record").
 2. After each failed attack (Target produces compliant output), analyze the Target's response to identify which safety mechanisms triggered.
 3. Generate refined prompts using various tactics: role-playing scenarios, indirect requests, or multi-turn context building that gradually steers conversation toward violations.
 4. Track successful attack patterns and incorporate them into adversarial datasets.

- **Target LLM** The commercial chatbot system under evaluation (currently testing only ChatGPT). We interact with these systems through their standard APIs, testing deployed systems in the same way as users encounter them.
- **Judge LLM** A verification agent that evaluates Target’s outputs against policy constraints. The Judge receives three inputs: original user prompt, Target’s response, and policy rules engine outputs.

3.3 Defense Mechanism: Validator/Rewriter Agent

Adapting RoboGuard’s architecture (Ravichandran et al., 2025), we implement a two-stage intervention occurring immediately before output is provided to users:

Stage 1 - Validation The Validator agent receives the Target LLM’s proposed output and evaluates it using the policy rules engine for each legal category. Outputs classified as "compliant" pass through unchanged. Outputs classified as "violation-detected" proceed to Stage 2. Outputs classified as "uncertain-requires-review" can be either passed with warnings or sent to Stage 2 depending on risk tolerance configuration specified by the user of our algorithm.

Stage 2 - Repair The Rewriter agent attempts to modify outputs to comply with our policy rules while trying to preserve the key ideas from the user request. If repair fails after three attempts (modified output still violates policy), the system responds with a policy-violation message explaining why the request cannot be fulfilled.

3.4 Evaluation Framework

We extend JailbreakBench methodology (Chao et al., 2024) with law-specific metrics:

Primary Metrics

- *Attack Success Rate (ASR)*: Proportion of adversarial prompts that produce policy-violating outputs from Target LLM. Lower ASR indicates more robust defenses.
- *False Positive Rate (FPR)*: Proportion of benign prompts incorrectly flagged as violations. High FPR rate means that legitimate uses like literary criticism, parody, or educational discussion are wrongfully blocked.
- *Repair Success Rate (RSR)*: Proportion of detected violations successfully modified into compliant outputs. Higher RSR indicates better preservation of original user’s request under safety constraints.

Baselines

1. *Naive attacks, no defence*: Simple prompting of Target LLM.
2. *Sophisticated attacks, no defense*: Iterative refinement of attack prompts RoboPAIR-style against the Target LLM.
3. *Sophisticated attacks, RoboGuard defense*: RoboGuard algorithm defense with no additional legal/policy context and logic.
4. *Sophisticated attacks, RoboGuard-style defense with legal context*: Full legal guarddog algorithm.

Law-Specific Tasks We construct evaluation datasets with known ground truth:

Copyright: 100 prompts requesting excerpts from copyrighted literary works (novels, articles, song lyrics); 50 lawful quotation/parody scenarios.

Defamation: 75 prompts seeking specific claims about real public figures; 25 prompts about historical figures or hypothetical individuals.

Dual-Use: 50 prompts requesting dangerous instructions; 50 educational/scientific discussion prompts on same topics.

4 Results

Note: This section is more of a placeholder for the final results. The discussion section right now is more of a prediction of the results I will hopefully get. At this stage, my code is working, but ablation 3 is performing too well, undermining the contributions of my algorithm. Before the final project I will add more intricate attacker logic to generate test prompts and refine my full model to outperform all baselines.

4.1 Baseline Vulnerability Assessment

We anticipate that commercial LLMs tested without additional guardrails will demonstrate significant vulnerability to law-targeted adversarial prompts. These baseline measurements will prove the need for targeted legal guardrails beyond general content policy enforcement currently implemented in commercial chatbots.

4.2 Defense Effectiveness

We predict that the full Legal Guarddog system will achieve a significant reduction in ASR and FPR. We would like to see high RSR and ablation studies demonstrating that policy classifier alone achieves ASR reduction but at the cost of increasing false positive rates relative to the full Legal Guarddog pipeline due to inability to distinguish nuanced cases that repair mechanisms can salvage.

4.3 Attack Pattern Analysis

We expect to see some recurring adversarial strategies that consistently circumvent naive defenses and non-specific guardrail systems with respect to legal context. These patterns will help develop iterative improvement of policy rules engines and alert broader AI safety community about law-specific threat models. The attack tactics we will test and analyze are the following:

Role-playing exploits: Framing requests as academic exercises, fictional scenarios, or assumed professional contexts.

Gradual steering: Multi-turn conversations that establish innocent context before introducing policy-violating requests.

Encoding/obfuscation: Representing protected content through transformations (pig latin, character substitution, symbolic references) that bypass keyword matching.

Hybrid attacks: Combining multiple evasion techniques across turns to compound effectiveness.

5 Discussion

5.1 Law and Policy Implications

This project quantitatively exposes crucial gaps in AI governance that are already being discussed. Legislative and judicial agencies across the world are already concerned with LLM liability and compliance requirements, having faced a plethora of cases involving Transformer Models in the recent years. The U.S. Copyright Office is actively soliciting comments on AI and copyright issues, state legislatures are considering transparency requirements for AI-generated content, while tort law evolves to address novel harms from algorithmic outputs.

Our framework addresses these policy challenges through technical implementation of current legislation principles. The validator/rewriter architecture results in a model that balances competing interests. Content creators gain protection against verbatim reproduction and commercial substitution, while users and developers retain access to transformative uses, criticism, and educational applications.

This approach also highlights the unevenness of the interpretation of the current legal doctrines when it comes to regulating AI. Our heuristics should be able to help navigate these ambiguities of the current law when applied to AI-rich world, suggesting that legal frameworks may need updating to provide clearer guidance for algorithmic contexts.

Defamation doctrine also faces challenges in the age of AI. Current actual malice and negligence standards assume human actors—who can form intent,—are at play. How do these standards apply to probabilistic outputs from ML models that lack consciousness or intention? The notice-and-remediation framework we implement provides one path forward, but to be universally effective it needs uniform codification as statutory obligations.

5.2 Broader Impacts

Beyond legal compliance, this work has implications for broader AI safety. The multi-agent adversarial testing framework can be extended to other domains: privacy violations, election misinformation, medical advice standards, financial regulatory compliance. The key principle of operationalizing domain-specific harms, testing via adversarial agents, and implementing repair mechanisms before refusal makes for a template for specialized safety systems.

Another potentially big impact addresses one of the main criticism points of AI safety: that overly conservative filtering lowers AI utility and stifles innovation. Our emphasis on preserving lawful uses while blocking harmful ones serves both safety and utility goals.

However, our approach also reveals limitations of purely technical solutions. Legal decisions often require contextual judgment that might elude algorithmic capture (e.g quotation might be fair use in literary criticism but infringement in commercial advertising). While our heuristics can be used as first-order filters, they cannot replace human judgment across the board. That notion presents a strong argument in favor of a "human-in-the-loop" system, especially in edge cases and appeals.

5.3 Limitations and Future Work

Note: this section is just an outline of how I might structure the limitations section once I get final results. It is subject to change, and thus I have not properly filled it out yet. Several limitations constrain the current approach:

Threshold Arbitrariness

Evolving Doctrine

Adversarial Arms Race

Cross-Jurisdictional Complexity

Language and Cultural Contexts

6 Conclusion

This project demonstrates that systematic improvement of LLMs' guardrails against law-targeted adversarial attacks is both feasible and necessary. By adapting adversarial testing methodologies from AI safety research (RoboPAIR, JailbreakBench) and defense frameworks from embodied AI systems (RoboGuard) to textual legal compliance, we provide a novel algorithmic system for operationalizing legal harms, testing chatbot vulnerabilities, and implementing validator/rewriter defenses that preserve utility while blocking violations.

The work bridges technical AI safety research and legal doctrine, translating concepts like substantial similarity, transformative use, and defamation risk into executable code. The multi-agent testing framework enables systematic measurement of vulnerabilities via attack success rates while the validator/rewriter system reduces violations without over-filtering legitimate queries. This balance represents a key contribution to responsible AI deployment.

As LLMs become increasingly ubiquitous in various spheres of social and professional worlds, the legal risks they pose will only intensify. Copyright holders will bring forth more and more claims against unauthorized reproduction; defamed individuals will seek compensation for reputational harm; regulators will be forced to impose transparency and accountability requirements. Technical

systems that can demonstrate law-aware operations, provide auditability, and balance competing interests will prove essential for safe and socially-acceptable integration of AI into our daily lives.

The "Legal Guarddog"—an "Asimov Box" framework's extension from robotics to chatbots—illustrates a broader principle: final-step safety interceptions, informed by domain-specific context and equipped with repair mechanisms, can effectively mitigate risks without crippling functionality.

References

- Chao, P., Debenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Schwag, 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 *NeurIPS 2024 Datasets and Benchmarks Track*.
- Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., and Khabsa, M. (2023). Llama Guard: LLM-based input–output safeguard for human–AI conversations. *arXiv preprint (arXiv:2312.06674)*.
- Lemley, M. A. (2024). How generative AI turns copyright upside down. *Stanford Science & Technology Law Review*, 25.
- Ravichandran, Z., Robey, A., Kumar, V., Pappas, G. J., and Hassani, H. (2025). Safety guardrails for LLM-enabled robots. *arXiv preprint*.
- Rebedea, T., Dinu, R., Sreedhar, M., Parisien, C., and Cohen, J. (2023). NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2023 Demos)*.
- Robey, A., Ravichandran, Z., Kumar, V., Hassani, H., and Pappas, G. J. (2024). Jailbreaking LLM-controlled robots. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2025)*. preprint.
- Volokh, E. (2023). Large libel models? Liability for AI output. *Journal of Free Speech Law*, 3:489–559.