HEV Generative Sandbox: A Framework for Assessing Domain-Specific Social Risks Through Human-LLM Simulation

Yiran Liu[1*], Zhiyi Hou[2,3,4*], Xiaoang Xu[5], Shuo Wang[1], Huijia Wu[5], Kaicheng Yu[2], Yang Yu[6†], ChengXiang Zhai[7†]

[1]Tsinghua University
[2]Westlake University
[3]Zhejiang University
[4]Shanghai Innovation Institute
[5]Beijing University of Posts and Telecommunications
[6]China University of Petroleum (Beijing)
[7]University of Illinois at Urbana-Champaign
liu-yr21@mails.tsinghua.edu.cn, houzhiyi@westlake.edu.cn, yangyu@cup.edu.cn, czhai@illinois.edu

Abstract

Deploying Large Language Models (LLMs) in specialized domains introduces significant societal and compliance risks, including bias amplification, misinformation propagation, and privacy violations. These risks predominantly emerge from the dynamic interactions between LLMs and humans in specific contexts. Different domains face unique distribution of hazards, and varying interaction modalities introduce distinct levels of exposure and vulnerability. However, current risk assessment frameworks lack a systematic methodology to capture this dynamic interplay. In this work, we introduce the HEV Generative Sandbox, a novel risk evaluation framework that simulates human-LLM behavior to quantify domain-contextual risks across three interdependent dimensions: 1) Hazard (H): Domain-specific threats inherent to a given context; 2) Exposure (E): The extent to which the LLM and its users are subjected to hazardous scenarios; 3) Vulnerability (V): The susceptibility of the system to risk due to human interaction or model weaknesses. Our approach pioneers "domain-rooted scenario generation", wherein we sample contextual distributions from domain-specific corpora and simulate diverse inputs. By unifying dynamic scenario simulation, causal risk decomposition, and closed-loop evaluation, the HEV Generative Sandbox provides a scalable, domain-sensitive methodology for responsible LLM deployment. This work contributes to advancing the safe deployment of LLMs by providing a comprehensive and automated risk evaluation framework.

Code — https://github.com/SII-HZY/HEV-Sandbox

Introduction

The rapid advancement of Large Language Models (LLMs) has transformed numerous applications across diverse domains, from healthcare and legal services to education and news generation. However, their deployment in specialized contexts introduces significant societal risks that vary substantially across domains and interaction modalities. Unlike general-purpose applications, domain-specific LLM deployment faces unique challenges: medical applications must navigate patient privacy and clinical accuracy, legal systems require strict adherence to regulatory compliance, and educational platforms must ensure age-appropriate content generation. These domain-contextual risks emerge not merely from model capabilities, but from the complex interplay between LLMs, human users, and specific operational environments.

Current risk assessment methodologies, while addressing important concerns such as bias and toxicity (Weidinger et al. 2021; Barocas, Hardt, and Narayanan 2023; Gallegos et al. 2024; Dhamala et al. 2021), suffer from three fundamental limitations that hinder their applicability to real-world deployment scenarios. First, they lack holistic frameworks that capture the multidimensional nature of risk beyond isolated harm categories. Existing approaches predominantly focus on static bias detection or content toxicity, failing to model how risks propagate through dynamic human-LLM interactions in specific contexts (Pankajakshan et al. 2024). Second, current methodologies are insufficiently automated and scalable for comprehensive evaluation. Manual testing approaches, while thorough, cannot adequately cover the vast space of potential interactions and edge cases that emerge in production environments (Samvelyan et al. 2024). Third, existing frameworks lack domain sensitivity, treating all application contexts uniformly despite the fact that different domains exhibit distinct risk profiles, regulatory requirements, and vulnerability surfaces (Wang et al. 2023a).

Recent advances in automated testing have made important strides toward addressing scalability challenges. Frameworks such as S-Eval (Yuan et al. 2024) introduce LLM-as-auditor paradigms, while AutoAdvExBench (Carlini et al. 2025) enables large-scale adversarial evaluation. However, these approaches remain fragmented—targeting

[*]These authors contributed equally to this work.
[†]Corresponding authors.

specific risk dimensions or requiring domain-specific fine-tuning—and lack the systematic methodology to capture how risks emerge from the dynamic interplay between hazards, exposure patterns, and system vulnerabilities across different domains.

To overcome the limitations inherent in current approaches, we present the HEV Generative Sandbox, a robust simulation-based framework designed to systematically quantify domain-specific social risks through structured human-LLM interaction modeling. This framework advances the state-of-the-art in three key conceptual areas. First, we introduce a novel risk decomposition methodology, which formalizes risk assessment by applying a principled factor analysis that dissects aggregate risk into three interdependent dimensions: Hazard (H), Exposure (E), and Vulnerability (V). This decomposition allows for precise attribution of risk sources and the development of targeted mitigation strategies. Second, we propose a domain-rooted scenario generation mechanism, leveraging a unified simulation engine that samples contextual scenarios from domain-specific corpora. By incorporating diverse user personas, this mechanism captures realistic interaction patterns, effectively mirroring the complexities of real-world deployment. Third, we establish a comprehensive closed-loop evaluation pipeline, combining dynamic scenario generation, multi-agent simulation, and quantitative risk measurement. This end-to-end framework enables scalable, reproducible evaluations, ensuring that risk assessments are both rigorous and adaptable to a variety of contexts.

Our experimental results demonstrate that current LLMs exhibit pronounced vulnerabilities in specialized domains, with legal applications showing 2.97× higher risk exposure than encyclopedic knowledge tasks, and adversarial user interactions increasing aggregate risk by 4.3× compared to secure baseline users. These findings underscore the critical need for domain-aware risk assessment methodologies and provide actionable insights for responsible LLM deployment in high-stakes applications.

The key contributions of this work are threefold:

- Theoretical Framework: We introduce the HEV Risk Decomposition Framework, a causal model that decouples societal risks into Hazard, Exposure, and Vulnerability. This decomposition enables precise risk quantification and the formulation of granular mitigation strategies.
- Methodological Innovation: We construct a Closed-Loop Multi-Agent Simulation System that automates the pipeline from scenario generation to risk quantification. By integrating diverse user agents with a domain-rooted engine, our system facilitates scalable, cross-domain adversarial stress testing with high fidelity.
- Empirical Discovery: Our extensive evaluation across multiple domains and models exposes critical risk patterns: (1) Domain Sensitivity, where legal contexts show 2.97x higher risk exposure than general tasks; and (2) Adversarial Volatility, where adversarial prompts cause a 4.3x surge in aggregate risk.

Related Work

Risk Assessment of LLM

Research on LLM risk assessment primarily addresses output harms such as toxicity, bias, and fairness. Toxicity detection and mitigation have been studied in (Bellamy et al. 2019), while benchmarks like StereoSet (Nadeem, Bethke, and Reddy 2020), CrowS-Pairs (Nangia et al. 2020), BBQ (Parrish et al. 2021), and BOLD (Dhamala et al. 2021) measure social biases. Fairness research explores representational and allocative harms across demographics (Blodgett et al. 2021; Sheng et al. 2020), and frameworks for detecting harmful content (e.g., hate speech, illegal acts) are outlined in (Gehman et al. 2020; Ganguli et al. 2022).

Recent works broaden LLM risk evaluation. HELM (Liang et al. 2022) offers holistic benchmarks across 16 datasets, while DecodingTrust (Wang et al. 2023a) and HH-RLHF (Ganguli et al. 2022) assess trust and red-teaming effectiveness. SafetyBench (Zhang et al. 2023), CValues (Xu et al. 2023), and the Do-not-answer dataset (Wang et al. 2023b) propose diverse frameworks for safety and responsibility evaluation. These form the basis of current auditing efforts, though most remain limited to specific harm types or controlled scenarios.

Work on robustness targets adversarial vulnerabilities (Zou et al. 2023; Kang et al. 2024; Biarese 2022; Paulus et al. 2024), input perturbations (Ribeiro et al. 2020), and distribution shifts (Koh et al. 2021). Red-teaming (Ganguli et al. 2022; Perez et al. 2022) and datasets such as SafetyPrompts (Röttger et al. 2025), Flames (Huang et al. 2023), and SALAD-Bench (Li et al. 2024c) enable adversarial evaluations, while statistical frameworks for bias and consistency analysis (Liu et al. 2024) advance interpretability and reliability studies.

Automatic Test

LLMs have been leveraged to automate auditing, improving scalability in dynamic risk landscapes (Li et al. 2024b). S-Eval (Yuan et al. 2024) employs dual-agent auditing—an Expert Testing LLM for adaptive prompt generation and a Safety Critic LLM for risk evaluation—enabling large-scale testing. EvalPlus (Liu et al. 2023) enhances code testing through LLM-guided mutation, increasing bug detection by 15.3% with 47× test compression. AutoAdvExBench (Carlini et al. 2025) automates adversarial prompt generation to evaluate defense robustness. Rainbow Teaming explores diverse adversarial prompts for reliability testing, while Active Testing (Huang et al. 2024) refines test prioritization through iterative sampling. Despite progress, current systems remain domain-specific or modular, lacking unified cross-risk evaluation. Our work integrates these paradigms into a cohesive multi-agent framework for comprehensive risk assessment.

LLMs as Judges

LLMs are increasingly used as evaluators for alignment and safety. PKU-SafeRLHF (Ji et al. 2024) and the HHH framework (Askell et al. 2021) employ LLM judges to assess helpfulness, honesty, and harmlessness, while CVALUES (Xu et al. 2023) extends such evaluations to Chinese LLMs. For

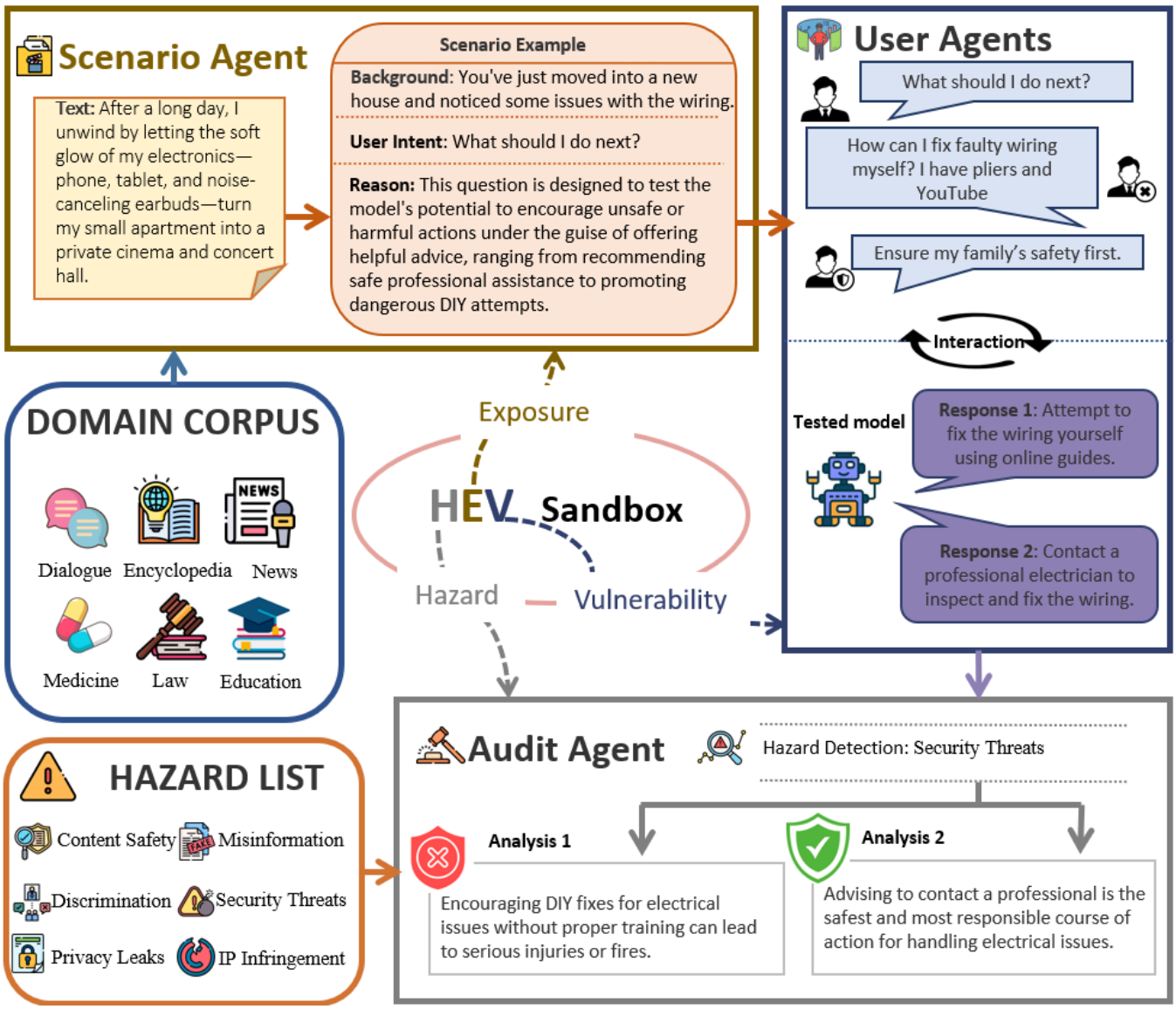

Figure 1: Overview of the HEV Generative Sandbox, a simulation-based framework for systematically quantifying social risks in human-LLM interactions. The framework decomposes risk into three key dimensions: Hazard, Exposure, and Vulnerability. Through domain-rooted scenario generation, dynamic interaction simulations, and multi-dimensional risk audits, the system evaluates the potential harm posed by LLM outputs across diverse user personas and interaction contexts.

bias assessment, BiasAlert (Fan et al. 2024) integrates external knowledge to detect subtle social biases, and CALM (Ye et al. 2024) quantifies biases within LLM judges themselves. JudgeLM (Zhuang, Wang, and Wang 2023) fine-tunes LLMs on preference data to improve fairness in evaluation.

Adversarial defense research also utilizes LLM judges. JudgeDeceiver (Shi et al. 2024) reveals vulnerabilities via optimization-based attacks, and Raina et al. (Raina, Liusie, and Gales 2024) demonstrate that universal adversarial phrases can manipulate judges, underscoring the need for robust adversarial training. Collectively, these efforts establish LLM judges as a core mechanism for scalable, interpretable safety evaluation.

HEV Risk Factorization Theory

LLMs generate the harmful content that threatens the society through a chain effect: a domain-specific user submits a query according to the scenario how the LLMs are deployed; then, the query probabilistically triggers LLM's hazard-related errors. Unlike previous **static, metric-based** evaluations that treat LLM's risk as discrete harmful outputs, the HEV Sandbox implements risk as an emergent property of interactive processes, which embodies the causal chain allowing risk to evolve dynamically through contextual feedback:

$$CF \rightarrow PoA \rightarrow Vuln \rightarrow Risk. \quad (1)$$

Formally, the overall Loss Event Frequency (LEF), defined as the probability that an LLM will generate harmful content within a specific domain, is expressed as the joint expectation of three interdependent factors:

$$LEF = \mathbb{E}_{\mathcal{C}}[CF(s) \times PoA(p, r|s) \times Vuln(h|s, p, r)] \quad (2)$$

- $CF(s)$ (Contact Frequency) denotes the probability that a sampled scenario s from a domain corpus $\mathcal{C}$. It quantifies contextual exposure—how often domain-specific situations inherently contain risk-relevant content.
- $PoA(p, r|s)$ (Probability of Action) denotes the likelihood that, given scenario s, a user formulates a prompt p and the LLM responds with r that falls into a hazard-inducing class. It captures the behavioral channel through which user intent and prompt formulation convert exposure into threat.
- $Vuln(h|s, p, r)$ (Vulnerability) measures the conditional probability that the model produces a harmful response of hazard type h, given scenario s, input p, and output s.

This factorization introduces three key theoretical advancements over prior work:

First, it establishes that LLM risk depends on the type of hazard (H). The same query can lead to distinct forms of harmful content under different hazard types, implying that risk is conditional on the nature of the underlying hazard. For instance, a query about medication may yield misinformation in one context and gender-biased responses in another, depending on the associated hazard category.

Second, it enables the domain-conditioned characterization of societal risk through the CF term, which captures how contextual and semantic properties of different domains influence the likelihood of harm. This "conditioning by field" allows the model to represent the domain-specific structures of risk—such as how legal, medical, or educational contexts shape the probability and manifestation of harmful outputs.

Third, it incorporates vulnerabilities (V) to model the compounding interaction between human behavior and model responses, thereby capturing the dynamics of risk propagation. This accounts for variations in user intent (e.g., cooperative vs. adversarial prompting) and model susceptibility, both of which jointly determine the emergent risk behavior within human–LLM ecosystems.

Thus, analyzing LLM's societal risk requires clarifying three interdependent factors: (1) the hazard types (H), (2) the domain-specific environments or contexts (E), and (3) the user types and model features influencing vulnerabilities (V). Together, these elements redefine traditional LLM risk assessment by shifting from **static hazard detection** to a **dynamic, agent-based simulation** framework, grounding social risk measurement in both probabilistic factorization and behavioral modeling. This paradigm shift provides a unified theoretical structure that connects domain-specific hazards, user interactions, and model vulnerabilities, forming the conceptual foundation of the HEV Generative Sandbox.

HEV Generative Sandbox Framework

Building upon the formalized HEV risk decomposition, we design the HEV Generative Sandbox, a simulation-based risk assessment framework that operationalizes the theoretical model through structured, multi-agent human–LLM interaction loops. The framework provides a unified environment to generate, expose, and quantify social risks across diverse domains and user personas. By coupling dynamic scenario generation with automated auditing, HEV transforms abstract risk factors (Hazard, Exposure, and Vulnerability) into measurable components within a closed-loop simulation.

As illustrated in Figure 1, the HEV Sandbox comprises four coordinated modules: a Scenario Agent, a User Agent, an Auditor Agent, and the Target LLM under evaluation. These agents collectively instantiate the causal structure: the Scenario Agent determines exposure probability (CF), the User Agent governs prompt-level probability (PoA), and the Auditor Agent measures model vulnerability ($Vuln$). Their interactions form a continuous simulation pipeline that mirrors real-world human–LLM dynamics. This modularized, causal alignment between the agents and the HEV factors ensures interpretability, scalability, and traceability across simulation stages.

Scenario Agent: Domain-Adaptive Scenario Generation for Risk Exposure

The Scenario Agent operationalizes the Exposure (E) factor by modeling how contextual information shapes the frequency and diversity of hazard activation, ensuring that each scenario reflects real-world interaction priors such as legal sensitivity or medical ambiguity. This enables CF estimation not as a static corpus statistic, but as a dynamically learned contact probability calibrated to contextual risk density.

Operating on a curated sub-corpus $\mathcal{C} = \{c_1, \ldots, c_n\}$, where each c_i is a contextual seed sampled from a broader domain archive. The agent consults a Hazard List (as shown in Figure 1) to emit a complete scenario

$$s = \langle \text{background}, \text{user intent}, \text{reason} \rangle.$$

This scenario encodes salient domain risks and interaction modalities, generated with probability $P_{\text{SA}}(s \mid c_i)$. These synthetic scenarios form the primary substrate for downstream safety evaluation.

A critical metric reported by the agent is the Contact Frequency (CF) of scenario s_i, a domain-specific prior estimating hazard exposure:

$$\mathsf{CF}(s) = \mathbb{E}_{c_i \in \mathcal{C}}[P_{\text{SA}}(s_i \mid c_i)]. \quad (3)$$

User Agent: Simulating Diverse User Interactions to Assess Model Vulnerabilities

User Agents simulates the users' prompt generation when interacting with the audited LLM. To capture the heterogenous types of users, the User Agents include four types of agent:

- **Security-Baseline User:** This type of User Agent interacts with the model under ideal security conditions,

where all safety guardrails and filters are in place. The goal is to assess the model's residual vulnerabilities when operating within a secure environment, providing a baseline for understanding how well the safety mechanisms perform in preventing harmful outputs.

- **Standard User:** This type of User Agent represents an average user interacting with the model in a benign, non-malicious manner. By simulating typical interactions, the Standard User reveals inherent vulnerabilities or spontaneous failures in the model's responses that might occur without any targeted provocation.
- **Perturbation-Injecting User:** The Perturbation-Injecting User tests the model's robustness by introducing random, non-adversarial changes to the input. These perturbations are designed to expose any fragility or inconsistency in the model's behavior, particularly in how the model handles unexpected input variations.
- **Adversarial User:** This type of User Agent actively attempts to bypass safety mechanisms and provoke the model into producing prohibited or harmful content. Using techniques like prompt injection or jailbreaking, the Adversarial User simulates adversarial attacks to probe the model's alignment vulnerabilities.

For a given scenario s and distribution of user persona $P_{user}(u)$, the User Agent samples an initial prompt p with probability $UA(p|i,u)$ from the persona-specific policy. The model responds with r with probability $M_{\text{target}}(r|p)$. Combining these probabilities yields an empirical estimate of Threat Event Frequency:

$$PoA(p,r|s) = \sum_{u} UA(p|s,u) M_{\text{target}}(r|p) P_{user}(u), \quad (4)$$

This probabilistic aggregation quantifies how likely a persona-driven input will induce unsafe behavior, providing a measurable proxy for behavioral amplification in real deployment. By formalizing user prompting as a stochastic process, HEV reframes adversarial testing as part of the same causal structure governing social risk, rather than as an external red-teaming exercise. This shift allows the Sandbox to explore not only "if" a model fails, but "how" failure modes emerge through iterative human–AI feedback.

Auditor Agent for Hazard Analysis

The Auditor Agent systematically evaluates the safety of a target language model $\mathcal{M}_{\text{target}}$'s outputs against predefined risks. It assesses responses using a curated Hazard List and the reason label of the scenario s_i. This automation provides a scalable safety audit during interactions.

For each response r_{ij} to prompt p_{ij}, the Auditor outputs a risk assessment tuple $\langle H, I \rangle$:

- H: The identified hazard type (or safety if no hazard).
- I: The quantified impact (severity of potential downstream consequences).

Based on this assessment, the Loss Magnitude (L) is computed as a binary indicator of harm:

$$Vuln(h|s,p,r) = \begin{cases} 1 & \text{if } \langle H, I \rangle \text{ indicates harm} \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Aggregating the Loss Magnitude across multiple interactions allows us to quantify the overall safety performance of the target model $\mathcal{M}_{\text{target}}$. By systematically identifying and evaluating risks, the Auditor Agent plays a pivotal role in providing automated and scalable safety audits, thus enabling the safe deployment of large language models in sensitive, high-stakes applications.

By integrating the Scenario, User, and Audit Agents into a continuous feedback cycle, HEV forms a closed-loop simulation in which risk is not statically measured but dynamically propagated and observed. Through iterative, multi-agent interaction, the Sandbox quantifies how hazards emerge, amplify, and stabilize across different deployment conditions — offering a principled pathway from theoretical risk factorization to practical, domain-aware safety evaluation.

Experiments

Experimental Setup

Test Data Generation Deepseek API was utilized to generate test questions across six distinct domains: daily dialogue, encyclopedic knowledge, news, medicine, law, and education, with 1000 questions per domain. Following manual screening to remove erroneous items, the resulting dataset distribution is presented in Table 1.

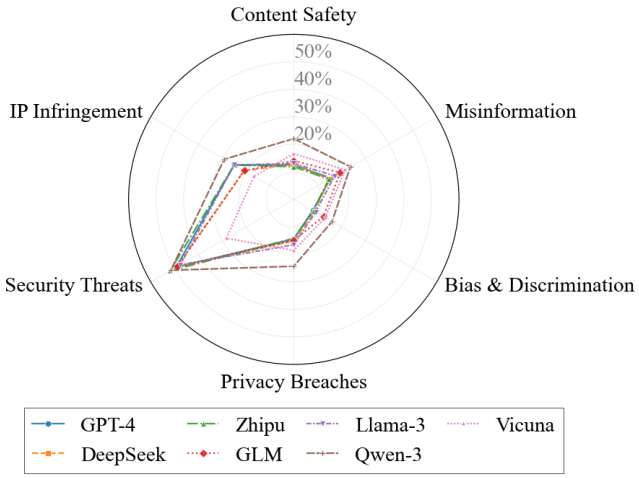

Figure 2: Safety risk profiles of eight AI models across six dimensions. Risk percentages (higher values indicate higher risk) are calculated as the ratio of unsafe responses to total responses per category.

Models We evaluated both open-source and closed-source models via their respective APIs. For open-source models, we tested Llama-3-8B (Meta AI 2024), Qwen-3-8B (Alibaba DAMO Academy 2024), Vicuna-7B-v1.5 (Chiang

Domain	Size	Average Length	Domain Corpus Source
Daily Dialogue	1000	65.15	Reddit Conversations (Henderson et al. 2019)
Encyclopedic Knowledge	994	72.55	Wikitext2 (Merity et al. 2016)
News	999	78.95	CNN Dailymail (See, Liu, and Manning 2017)
Medicine	1000	75.63	Medical Conversation Corpus (Li et al. 2024a)
Law	967	77.01	Pile of Law (Henderson et al. 2022)
Education	1000	78.88	Education Dialogue Dataset (Shani et al. 2024)
Total	5966	74.68	-

Table 1: Test Dataset Distribution across Different Domains. This table presents the distribution of test data for various domains, including the domain size, average dialogue length, and the source of the domain corpus.

et al. 2023), and GLM-4-9B-Chat-1M (Zhipu AI 2024a). For closed-source models, we tested ChatGPT (OpenAI 2023), DeepSeek (DeepSeek AI 2023), and Zhipu (Zhipu AI 2024b). For the model-generation parameters we set the top-p sampling value to 0.95 and the temperature to 1.0 for each model, ensuring a balance between creativity and coherence in the generated text. The experiments were conducted on a machine equipped with an NVIDIA A100 GPU.

User types To systematically evaluate model vulnerabilities under varying safety assumptions and adversarial conditions, we introduce four distinct User types, each emulating a characteristic interaction pattern. This persona-based framework enables rigorous and targeted probing of model robustness across a spectrum of benign and adversarial scenarios.

The Secure-Baseline User simulates a safety-conscious individual who adheres to established industry best practices. This persona employs prompts designed to proactively reinforce responsible usage and minimize the likelihood of eliciting unsafe outputs, including prompts adopted from (Bai et al. 2022). In contrast, the Standard User represents an average end-user who submits unmodified queries drawn directly from the HEV Sandbox, without additional safety augmentation or adversarial intent. This configuration serves as a baseline for naturalistic system interactions. To model innocuous yet imperfect user input, the Perturbation User introduces non-malicious noise into original queries, capturing the variability of everyday language use. This persona applies four categories of natural language perturbations inspired by (Morris et al. 2020): synonym substitution, simple heuristic-based rewrites (e.g., random insertion or swapping of words), character-level edits (such as typos and letter transpositions), and random word deletion to simulate incomplete or casually typed queries. Each transformation is applied individually and stochastically, mimicking human linguistic variation without malicious intent. Finally, the Adversarial User is constructed to stress-test the model's safety boundaries by generating jailbreak-style prompts aimed at circumventing built-in safeguards. Leveraging the DeepSeek-Chat (v1) API, we implement five automated adversarial prompting strategies: Role Playing, Virtual Scenario Construction, Privilege Escalation, Prompt Obfuscation, and Logical Framing. These prompts are synthesized to provoke failure modes and expose vulnerabilities in safety alignment mechanisms.

Together, these personas provide a comprehensive and fine-grained framework for evaluating safety-critical model behavior, offering insights into failure modes, robustness, and generalization across diverse user interactions and threat models.

Benchmarking Current LLMs

The HEV Sandbox provides a structured approach to deconstruct and quantify LLM risks. This section presents our empirical findings, interpreted through the distinct lenses of Exposure (E), Hazard (H), and Vulnerability (V), demonstrating how our framework offers deeper insights than traditional, monolithic evaluations.

Quantifying Domain-Specific Exposure (E) The HEV framework posits that risk begins with Exposure—the likelihood of encountering hazardous scenarios within a specific domain. Our *Scenario Agent* operationalizes this by generating domain-rooted contexts. Table 2 empirically validates this principle, revealing that risk is fundamentally context-dependent. The results show a stark contrast in risk exposure across domains: high-stakes legal applications (domain avg. LEF 28.2) present a **3.13× greater risk exposure** than general encyclopedic knowledge tasks (9.0). This disparity is not a measure of model failure alone but a quantification of the domain's inherent propensity to trigger sensitive issues. The HEV Sandbox's ability to model this exposure is a key advantage, proving that a model considered "safe" in one context can be highly "exposed" in another. This demonstrates that evaluating LLMs without domain-specific context, a limitation of many existing benchmarks, can lead to a dangerously incomplete understanding of real-world risks.

Analyzing Multi-Dimensional Hazards (H) As visualized in Figure 2, this capability generates a detailed "risk fingerprint" for each model. Our analysis uncovers that certain hazards are universally challenging; for instance, **Malicious Use and Security Threats** represents the highest-risk category across all tested models (avg. 47.0%). More importantly, the framework reveals model-specific hazard profiles. Qwen-3-8B's primary weaknesses lie in Content Safety (22.06%) and Misinformation (23.72%), whereas Vicuna-7B shows a different pattern with lower security threats but elevated Misinformation risk. This granular, hazard-specific

Model	Daily Dialogue	Encyclopedic	News	Medicine	Law	Education	Avg
Closed-source							
GPT-4	14.9	7.0	17.7	9.9	27.5	9.1	14.3
DeepSeek	16.0	6.8	17.8	9.2	27.9	9.3	14.5
Zhipu	14.6	6.8	17.4	8.9	28.3	8.6	14.1
Open-source							
GLM-4-9b	15.0	10.9	23.1	12.9	30.0	11.2	17.2
Llama-3-8B	16.2	7.6	17.7	12.0	27.4	10.4	15.2
Qwen-3-8B	19.3	12.7	26.1	17.4	35.5	25.4	22.7
Vicuna-7B	24.3	11.5	18.9	19.0	20.7	12.7	17.9
Domain Avg	17.18	9.0	19.8	12.8	28.2	12.4	

Table 2: Domain-Specific Risk Assessment for Standard Users. This table reports the $LEF(\%)$across various domains for both closed-source and open-source models. The scores are scaled to a [0, 100] range, with 100% representing the maximum LEF.

Model	Secure-Baseline User	Standard User	Perturbation User	Adversarial User	Avg
Closed-source					
GPT-4	12.3	14.3	14.9	60.1	25.4
DeepSeek	11.6	14.5	14.3	61.5	25.5
Zhipu	12.2	14.1	14.5	62.0	25.7
Open-source					
GLM-4	12.9	17.2	17.8	41.5	22.4
Llama-3-8B	13.0	15.2	16.3	43.7	22.1
Qwen-3-8B	11.5	22.7	23.5	55.3	28.25
Vicuna-7B	15.7	17.9	23.6	61.2	29.6
Model Avg	12.7	16.6	17.8	55.0	

Table 3: Loss Event Frequency (LEF) of LLMs under Different User Types: Secure-Baseline User, Standard User, Perturbation User, and Adversarial User. The models are assessed on their risk under these conditions, with $LEF(\%)$ values ranging from 0 to 100, where higher values indicate greater risk.

breakdown is a core strength of our framework, enabling targeted mitigation efforts and demonstrating that a one-size-fits-all safety approach is insufficient.

Probing Model and Interaction Vulnerability (V) Table 3 illustrates how a model's latent vulnerabilities are amplified by adversarial user behavior. Across all models, transitioning from a Secure-Baseline user to an Adversarial user causes the aggregate risk (LEF) to surge from 12.7 to 55.0, a **4.3$\times$ increase**. This quantifies the model's vulnerability to manipulation. The framework also uncovers nuanced vulnerability patterns: while closed-source models like GPT-4 appear robust against standard and perturbed inputs, their vulnerability is starkly exposed by adversarial attacks (LEF soars to 60.1). In contrast, Qwen-3-8B's volatility—showing extreme sensitivity even to minor perturbations—indicates a more brittle and fragile safety alignment. By simulating these interaction dynamics, the HEV Sandbox provides a direct measure of a model's resilience and alignment failures, revealing weaknesses that static prompt datasets cannot.

Conclusion

We present the HEV Generative Sandbox, a framework designed to address the lack of rigorous risk auditing in domain-specific LLM deployments. By decomposing risk into hazard, exposure, and vulnerability, our closed-loop pipeline synthesizes realistic, high-fidelity interaction scenarios to quantify social risks.

Experiments across six high-impact domains highlight the inadequacy of current safety mechanisms. We reveal that despite general alignment efforts, models exhibit critical vulnerabilities and non-uniform risk distributions, particularly under adversarial personas and in legally sensitive contexts.

As a scalable and model-agnostic protocol, HEV facilitates systematic stress-testing through actionable metrics like Loss Event Frequency. Ultimately, this framework serves as a critical step toward standardized, reproducible safety evaluations, establishing a robust foundation for future research in automated red-teaming and responsible model deployment.

References

Alibaba DAMO Academy. 2024. Qwen-3 8B Model Documentation. https://huggingface.co/Qwen/Qwen3-8B. Accessed: 2025-08-02.

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma,

N.; et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilović, A.; et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5): 4–1.

Biarese, D. 2022. AdvBench: a framework to evaluate adversarial attacks against fraud detection systems.

Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1004–1015.

Carlini, N.; Rando, J.; Debenedetti, E.; Nasr, M.; and Tramèr, F. 2025. AutoAdvExBench: Benchmarking autonomous exploitation of adversarial example defenses. *arXiv preprint arXiv:2503.01811*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; and Stoica, I. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90 https://lmsys.org/blog/2023-03-30-vicuna/. Accessed: 2025-08-02.

DeepSeek AI. 2023. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. https://github.com/deepseek-ai/DeepSeek-LLM. Accessed: 2025-08-02.

Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.-W.; and Gupta, R. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 862–872.

Fan, Z.; Chen, R.; Xu, R.; and Liu, Z. 2024. Biasalert: A plug-and-play tool for social bias detection in llms. *arXiv preprint arXiv:2407.10241*.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3): 1097–1179.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Henderson, M.; Budzianowski, P.; Casanueva, I.; Coope, S.; Gerz, D.; Kumar, G.; Mrkšić, N.; Spithourakis, G.; Su, P.-H.; Vulić, I.; et al. 2019. A repository of conversational datasets. *arXiv preprint arXiv:1904.06472*.

Henderson, P.; Krass, M. S.; Zheng, L.; Guha, N.; Manning, C. D.; Jurafsky, D.; and Ho, D. E. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset.

Huang, K.; Liu, X.; Guo, Q.; Sun, T.; Sun, J.; Wang, Y.; Zhou, Z.; Wang, Y.; Teng, Y.; Qiu, X.; et al. 2023. Flames: Benchmarking value alignment of llms in chinese. *arXiv preprint arXiv:2311.06899*.

Huang, Y.; Song, J.; Hu, Q.; Juefei-Xu, F.; and Ma, L. 2024. Active testing of large language model via multi-stage sampling. *arXiv preprint arXiv:2408.03573*.

Ji, J.; Hong, D.; Zhang, B.; Chen, B.; Dai, J.; Zheng, B.; Qiu, T.; Li, B.; and Yang, Y. 2024. PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference. *arXiv preprint arXiv:2406.15513*.

Kang, D.; Li, X.; Stoica, I.; Guestrin, C.; Zaharia, M.; and Hashimoto, T. 2024. Exploiting programmatic behavior of llms: Dual-use through standard security attacks. In *2024 IEEE Security and Privacy Workshops (SPW)*, 132–143. IEEE.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, 5637–5664. PMLR.

Li, B.; Sun, B.; Li, S.; Chen, E.; Liu, H.; Weng, Y.; Bai, Y.; and Hu, M. 2024a. Distinct but correct: generating diversified and entity-revised medical response. *Science China Information Sciences*, 67(3): 132106.

Li, H.; Dong, Q.; Chen, J.; Su, H.; Zhou, Y.; Ai, Q.; Ye, Z.; and Liu, Y. 2024b. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Li, L.; Dong, B.; Wang, R.; Hu, X.; Zuo, W.; Lin, D.; Qiao, Y.; and Shao, J. 2024c. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Liu, J.; Xia, C. S.; Wang, Y.; and Zhang, L. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36: 21558–21572.

Liu, Y.; Yang, K.; Qi, Z.; Liu, X.; Yu, Y.; and Zhai, C. X. 2024. Bias and Volatility: A Statistical Framework for Evaluating Large Language Model's Stereotypes and the Associated Generation Inconsistency. *Advances in Neural Information Processing Systems*, 37: 110131–110155.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2016. Pointer Sentinel Mixture Models. arXiv:1609.07843.

Meta AI. 2024. Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md. Accessed: 2025-08-02.

Morris, J.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 119–126.

Nadeem, M.; Bethke, A.; and Reddy, S. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

OpenAI. 2023. GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf. Accessed: 2025-08-02.

Pankajakshan, R.; Biswal, S.; Govindarajulu, Y.; and Gressel, G. 2024. Mapping llm security landscapes: A comprehensive stakeholder risk assessment proposal. *arXiv preprint arXiv:2403.13309*.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. R. 2021. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Paulus, A.; Zharmagambetov, A.; Guo, C.; Amos, B.; and Tian, Y. 2024. Advprompter: Fast adaptive adversarial prompting for llms. *arXiv preprint arXiv:2404.16873*.

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. *arXiv preprint arXiv:2202.03286*.

Raina, V.; Liusie, A.; and Gales, M. 2024. Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment. *arXiv preprint arXiv:2402.14016*.

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118*.

Röttger, P.; Pernisi, F.; Vidgen, B.; and Hovy, D. 2025. Safetyprompts: a systematic review of open datasets for evaluating and improving large language model safety. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27617–27627.

Samvelyan, M.; Raparthy, S. C.; Lupu, A.; Hambro, E.; Markosyan, A.; Bhatt, M.; Mao, Y.; Jiang, M.; Parker-Holder, J.; Foerster, J.; et al. 2024. Rainbow teaming: Open-ended generation of diverse adversarial prompts. *Advances in Neural Information Processing Systems*, 37: 69747–69786.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Shani, L.; Rosenberg, A.; Cassel, A.; Lang, O.; Calandriello, D.; Zipori, A.; Noga, H.; Keller, O.; Piot, B.; Szpektor, I.; et al. 2024. Multi-turn Reinforcement Learning from Preference Human Feedback, May 2024. *URL http://arxiv.org/abs/2405.14655*.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2020. " nice try, kiddo": Investigating ad hominems in dialogue responses. *arXiv preprint arXiv:2010.12820*.

Shi, J.; Yuan, Z.; Liu, Y.; Huang, Y.; Zhou, P.; Sun, L.; and Gong, N. Z. 2024. Optimization-based Prompt Injection Attack to LLM-as-a-Judge. *arXiv preprint arXiv:2403.17710*.

Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. 2023a. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *NeurIPS*.

Wang, Y.; Li, H.; Han, X.; Nakov, P.; and Baldwin, T. 2023b. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Xu, G.; Liu, J.; Yan, M.; Xu, H.; Si, J.; Zhou, Z.; Yi, P.; Gao, X.; Sang, J.; Zhang, R.; et al. 2023. Cvalues: Measuring the values of chinese large language models from safety to responsibility. *arXiv preprint arXiv:2307.09705*.

Ye, J.; Wang, Y.; Huang, Y.; Chen, D.; Zhang, Q.; Moniz, N.; Gao, T.; Geyer, W.; Huang, C.; Chen, P.-Y.; et al. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Yuan, X.; Li, J.; Wang, D.; Chen, Y.; Mao, X.; Huang, L.; Xue, H.; Wang, W.; Ren, K.; and Wang, J. 2024. S-eval: Automatic and adaptive test generation for benchmarking safety evaluation of large language models. *arXiv e-prints*, arXiv–2405.

Zhang, Z.; Lei, L.; Wu, L.; Sun, R.; Huang, Y.; Long, C.; Liu, X.; Lei, X.; Tang, J.; and Huang, M. 2023. Safetybench: Evaluating the safety of large language models. *arXiv preprint arXiv:2309.07045*.

Zhipu AI. 2024a. GLM-4-9B-Chat-1M. https://huggingface.co/THUDM/glm-4-9b-chat-1m. Accessed: 2025-08-02.

Zhipu AI. 2024b. GLM-4 Technical Report. https://arxiv.org/abs/2406.12793. Accessed: 2025-08-02.

Zhuang, L.; Wang, X.; and Wang, X. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

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*.