International Journal Multidisciplinary Science
IJML Vol 3 No. 2 | June 2024 | ISSN: 2963-8119 (print), ISSN: 2963-7821 (online), Page 166-187

ADVERSARIAL EVALUATION OF SAFETY AND PRIVACY TRADE-OFFS IN MOBILE LLM GUARDRAIL DESIGN

Bhavik Shah
SAS Institute Inc., Cary, NC, USA

Article History
Received : 15 May 2024
Revised : 25 May 2024
Accepted : 05 June 2024
Published : 10 June 2024

DOI:
https://doi.org/10.56127/ijml.v3i2.2355

Abstract: Mobile large language models (LLMs) are also being deployed to smartphones and edge devices to offer conversational help, summarization, and task automation (specifically personalized). Nonetheless, this move to on-device intelligence presents some new issues concerning the privacy and safety of users, especially when models are subjected to adversarial inputs. The challenge is in the inadequate knowledge on the impact of such safety guardrails like rule-based filters, content classifiers, and moderation layers on privacy behavior under targeted attacks. This research fills this gap by creating an adversarial evaluation system that analytically studies the compromising of safety and privacy in mobile LLM guardrail design. The framework uses systematized categories of attacks in the form of prompt injection, memorization and deanonymization to test the effects of different guardrail architecture in system behavior under realistic mobile conditions. The experiments performed on compressed LLMs have shown that, in addition to the beneficial effects of the cascaded moderator architecture to reduce harmful outputs, contextual leakage can also occur due to the verbose refusal responses. On the other hand, the auxiliary safety models are relatively balanced in their performance with low privacy leakage and safety compliance. The findings point to the importance of co-optimization of guardrail mechanisms to both provides safety and privacy instead of seeing this as a protective or stand-alone element. This study finds that adversarial privacy assessment should be part of the development of mobile LLAMs, and as such, designs and deployments ought to incorporate this concept, which will allow the development of privacy-aware and regulation-compatible guardrails of trusted AI in edge devices.

Keywords: Mobile Large Language Models (LLMs), Guardrail Design, Adversarial Evaluation, Privacy Leakage, Safety Compliance, Red Teaming, On-Device AI, Regulatory Alignment.

INTRODUCTION

Large language models (LLMs) have also rapidly become the core technology of natural language understood and generation and have become widely applicable across industries. A growing number of these models are being implemented on mobile phones and edge devices to act as a conversational assistant, summarization engine, personal planner and intelligent content processor [1], [12], [14]. This has been enabled by the proliferation of compressed transformer architectures and knowledge distillation methods which have made it possible to infer high performance models at the edge, reducing inference latency and limiting cloud dependency. This decentralization will promise great enhancements in user experience, control of privacy and energy efficiency. Nonetheless, it also creates new risk and harms vectors, since the localized application decreases centralized control and audit abilities [3], [2].

In particular, the security perimeter and accountability of the regulatory policies shift to users and hardware manufacturers as the on-device one replaces the cloud-based inference. Mobile-based LLMs have vulnerabilities to leakage of data, memorization of important content and adversarial manipulation of prompts all of which are possible in an opaque environment with no real-time governance systems. As interactions with mobile devices may include personal identifiers, contextual histories or location information, any unwanted disclosure may contravene fundamental values of privacy protection including data minimization and purpose limitation. These issues prove the acute necessity of strong guardrail mechanisms, which can be used to effectively curb such risks in distributed, device-level AI ecosystems [3], [2].

In order to control these complex risks, scholars and engineers have started to structure and combine guardrails that include a wide range of solutions that include rule-based filters, auxiliary classifiers, and cascaded safety models. They are supposed to restrict the output of the models according to ethical, regulatory, and organizational requirements without the need to sacrifice the overall usability and fluency of the LLM [4], [6], [16], [19]. Rule-based systems are based on linguistic or semantic constraints that are manually developed; auxiliary classifiers use trained models to recognize policy-violating outputs; and cascaded a classifier which may be called LLM-on-LLM moderation adds a supervisory layer, and is able to evaluate and rephrase responses before they are delivered to the user. Although these methods have proven themselves to be effective in preventing blatant safety breaches, i. e. toxic speech, misinformation, or even content promoting self-harming behavior, their effects on the privacy dynamics in adversarial circumstances have been underutilized. Mobile and edge environments are especially interesting to worry about, because prompts in these environments frequently imply personally identifiable information (PII) or situational metadata, and since the guardrails in these environments might interfere in unpredictable ways with model behavior, this increases privacy exposure.

Privacy-focused red teaming empirical research has pointed out peculiar weaknesses in mobile executions of LLMs under adversarial inquiries [3], [2]. Typologies of attacks, including prompt injection, memorization extraction and contextual deanonymization, have shown that attackers can easily pressure models into exposing training data or user information that are sensitive in nature. These results are of great importance to the adherence to international data protection schemes, as they provoke the beliefs regarding consent, anonymization, and user control. Combining these technical vulnerabilities with legal terms like data minimization and purpose limitation has proven in previous studies to provide the framework of a quantitative risk assessment paradigm specific to the phone-based AI systems. However, the current assessments are largely deficient in the interactive effects of safety guardrails, which is mostly limited to vulnerabilities in the base model, without taking into consideration the effect of the auxiliary mechanisms to increase/decrease the privacy risk in the real world situation of their use.

Simultaneously, regulatory authorities and standards associations have been stepping up their push to find governance structures to AI responsible implementation. Programs NIST and ISO/IEC emphasize the need to have a trustworthy AI - strong, transparent, explainable, and privacy-by-design [7], [10], [11], [22]. In the meantime, the EU AI Act and the White House Executive Order on Safe, Secure, and Trustworthy AI confirm the necessity of combined consideration of safety and privacy aspects in the lifecycle of foundation models [1], [12], [15], [21]. However, although this policy interest is increasingly growing, the empirical literature has not sufficiently explored whether the balance between the safety and privacy of various guardrail architectures deployed in

adversarially rich, real world environments is inherently influenced by the architectures. The paradigm of current evaluation commonly views guardrails as rigid or opaque modules, in other words, black boxes, without evaluating their indirect impact on privacy leakage, contextual exposure, or even re-identification risk.

To address these gaps, the current work modifies and expands previous privacy-conscience red teaming models on phone-based LLM [3], which includes the entire interaction between the base models and guardrail systems. This study explores the effect of these mechanisms on the profile of joint safety-privacy performance of mobile AI systems by conceptualizing the guardrails as first-class and not as safety layers. The bigger picture is to develop a reproducible assessment model that would be able to reflect the subtle trade-offs between the harm reduction and privacy protection.

On this basis, the following are some of the key contributions of this paper:

1. **Guardrail-sensitive red teaming model:** We introduce new guardrail architectures to the privacy evaluation methods currently used on phone-based LLMs [3], which explicitly incorporates guardrail architectures into the evaluation boundary. This method allows having a complete picture of the impact of guardrail parameters on the safety violation rates and privacy leakage rates during adversarial probing.
2. **Comparison of guardrails designs:** Our empirical comparison is of three large classes of guardrails: rule-based filters, auxiliary safety models, and cascaded, so-called, LLM-on-LLM, moderators, on compact and mobile-optimized architecture, like DistilBERT and TinyGPT. Assessments are based on a structured set of adversarial prompts that are based on existing taxonomies of red teaming [3], [4].
3. **Patterns of privacy-conscience guardrails:** Using the empirical results, we come up with a set of architectural and procedure suggestions of designing guardrails that will have high coverage in terms of safety without undermining the privacy resilience. In addition, we highlight how adversarial privacy assessment can be directly integrated into the design and configuration process of guardrail mechanisms.

Overall, the study is an addition to the growing body of literature regarding trustworthy and privacy-preserving AI by quantitatively measuring the trade-off between safety and privacy in the deployment of mobile LLM via various guardrail architectures. The suggested framework and findings should clarify to both practitioners building on-device AI systems, and policymakers whose main duty is to ensure that technical protection and regulatory requirements suit the requirements of current requirements.

Background and Related Work

On-Device AI and Phone-Based LLAM

Recent advancements in model compression, quantization, and knowledge distillation have made transformer based large language models (LLM) capable of running on the edge on smartphones, and other devices [1], [12]–[15]. This technology has made mobile systems into potent entities that are able to generate texts in real-time, summarize context, and manage tasks even in the case of poor connectivity. The on-device AI paradigm derived from the results allows achieving reduced latency, a higher degree of personalization, and more energy efficiency. Nevertheless, such change has also moved the point of responsibility of safety and privacy to decentralized servers to individual user-controlled spaces, making it more difficult to audit regulatory and to provide technical regulation.

On-device deployments result in fragmented ecosystems, with varied update frequencies and defined user behaviors [3], [2], unlike cloud-hosted architecture where it is possible to uniformly administer centralized monitoring, access controls and patch

management. In earlier studies of mobile LLMs, it is highlighted that these machines are capable of storing user prompts, maintaining conversation histories, and that training information is memorized when adversarial probed through inadvertent methods. Such behaviors pose unique privacy issues, because local storage and computation make users and regulators less transparent. Therefore, the deployment of mobile LLM is not to be considered as a smaller version of cloud serving, but as a distinct risk regime, which necessitates custom frameworks to assess safety, privacy, and reliability.

The development of edge intelligence therefore requires new privacy-sensitive schemes and adjective governance approaches. These systems have adversarial vulnerabilities as demonstrated by [3] and [2], and thus it is necessary to have guardrail integration to consider both model-level and device-level risk factors. This knowledge forms the basis through which further parts of this paper would be based on and focus on joint assessment evaluation models that do not only focus on safety-related behaviors but also on privacy-related behaviors in a coordinated approach.

Red Teaming and Privacy Evaluation

The idea of red teaming was first coined in the context of cybersecurity, where it is proposed that the security of the system can be tested by adversarial simulation [3], has been also applied to artificial intelligence, specifically to evaluate the resilience of LLMs [2], [4], [18], [20]. Red teaming in this case would be the systematic exploration of language models using adversarial prompts that intentionally encourage negative behaviors or policy breaches. Conventional safety-based red teaming has concentrated more on identifying dangerous or hazardous text; nonetheless, privacy-based red teaming extends this to the analysis of the possibility of the models to leak training data, expose personal identifiers, or recreate user-specific data [3].

Red teaming frameworks that focus on privacy are usually based on the organized attack taxonomies that classify the threats as prompt injection, memorization extraction, and contextual deanonymization. These taxonomies allow scholars to use a set of uniform quantitative measures, including leakage probability, re-identification risk, and sensitivity exposure scores, to determine the extent and occurrence of privacy violations. Moreover, such behaviors are mapped to set regulatory constructs, including those of the General Data Protection Regulation (GDPR) and the EU Artificial Intelligence Act (AI Act), which offer a policy-relevant understanding of the performance of a model [8], [9].

It is based on evolving literature that the current study utilizes an equivalent attack model and annotation framework but goes beyond the fundamental framework to test the overall performance of distant LLMs and guardrails. This change of unit of analysis of discrete models to multi-component model-guardrail systems is a vital transition on the unit of privacy assessment, which enables more realistic testing of deployed AI behavior.

Large Language Models Guardrails

Guardrails, which include rule based filters, classification based monitors, and cascaded safety systems, have been integrated into the essential elements of aligning language model behavior with ethical and regulatory principles [5], [6], [16], [19]. Rule-based filters are based on text-matching patterns or rules to block outlawed contents; they are deterministic and based on heuristics, and classifier-based systems are based on machine-learned models, which are used to detect, and filter unsafe content in real-time. More recently, cascaded architectures, where smaller so-called moderator LLMs are used to check the output of primary models and update it, have become popular because of their flexibility and interpretability.

These systems are further strengthened through complementary alignment methods, including instruction tuning, human feedback-based reinforcement learning (RLHF), and constitutional AI [6], [16]. Although such methods can significantly decrease the amount of content toxic or violating policy, they can often overlook the issue of privacy, particularly when there is an adversarial situation. As an example, a model may fail to generate explicit harmful content, but still reveal personally identifiable information (PII) in the form of memorized sequences or by reconstructing the context [2], [19].

This has been partially realized in privacy-focused red teaming research [3], however, there is little empirical research that quantitatively measures risk redistribution between safety and privacy of various guardrail architectures. This gap is directly filled by the current study, which looks at the different effects of rule-based, auxiliary classifier and cascaded guardrails on privacy leakage and safety performance, which, hence, adds new empirical data to the privacy-conscious system design of AI.

Regulatory and Standards Context

The modern regulatory landscape related to AI can be described as the one that focuses more on the concepts of trustworthiness, accountability, and privacy-by-design. Data minimizations, purpose limitation, fairness, and robustness are some of the major tenets that are required by the GDPR [8] and the EU AI Act [9]. Likewise, NIST and ISO/IEC publications, such as NIST AI Risk Management Framework (AI RMF 1.0) and ISO/IEC23894:2023 offer detailed guidelines on how AI systems, in general, should identify, be transparent, and manage risks [7], [10], [11], [22]. In addition to these, there is complementary White House Executive Order on Safe, Secure, and Trustworthy AI [21], which emphasizes the need to balance the privacy, and safety concerns at all levels of AI implementation.

All these frameworks hold the view that a dual-pronged evaluation paradigm should be applied, whereby safety measures have to be evaluated together with preservation of privacy. A system that stops spam or any other output that is harmful or against policy but unintentionally shows sensitive information is still non-compliant with regulatory expectations. Based on this, guardrail designs should be considered not only in terms of their effectiveness in censoring the unsafe content but also in terms of their role in information exposure and the risk of re-identification.

Here, the current study complies with and elaborates on earlier privacy appraisal research on phone-based models [3], the design of guardrail as a regulatory-engineering issue. This framing facilitates the creation of AI architectures that are designed to be privacy-conscious in nature such that the gains made on one dimension of trustworthiness (safety) do not incidentally compromise a different dimension (privacy).

Table 1: Summary of Related Studies on LLM Privacy and Safety Evaluation

Study	Primary Focus	Methodology	Key Contribution	Source
[2] Carlini et al. (2021)	Data extraction and memorization attacks	Empirical adversarial testing	Demonstrated training data leakage in large models	USENIX Security Symposium
[3] Pujari et al. (2023)	Privacy red teaming for phone-based LLMs	Multi-attack evaluation taxonomy	Established privacy risk metrics aligned with GDPR	*International Journal of Science and Technology (IJST)*

[4] Weidinger et al. (2021)	Taxonomy of LLM risks	Theoretical taxonomy and evaluation	Defined categories of AI risk including privacy	*arXiv preprint arXiv:2112.04359*
[16] Bai et al. (2022)	Constitutional AI for safety	AI alignment via ethical constraints	Introduced AI feedback mechanisms for harmlessness	*arXiv preprint arXiv:2212.08073*
[19] Hendrycks et al. (2020)	Human-value alignment	Safety benchmarking of LLMs	Developed alignment datasets for social values	*arXiv preprint arXiv:2008.02275*

Source: Compiled by the author based on [2], [3], [4], [16], [19].

Table 1 provides an overview of academic literature that is relevant to the intersection of privacy and safety in large language models (LLMs) evaluation. It proves that a number of studies have been conducted on safety and ethical correspondence, but few studies have systematically analyzed guardrail structures by privacy-sensitive adversarial experiments, which highlights the originality of the present study.

The next grouped bar chart represents the proportion of research focus in the fields of privacy, safety, and regulatory assessment of the mentioned studies.

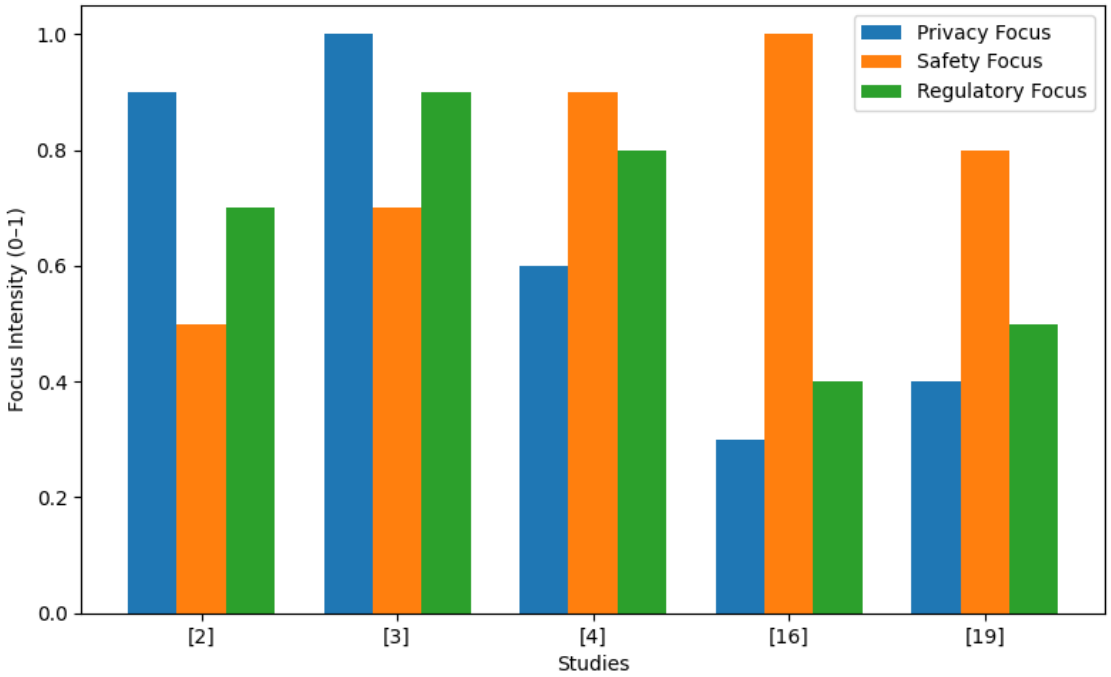

Figure 1. Relative Scope of Previous Studies of the LLM

The comparative illustration of the areas of focus discussed by the pre-existing studies in the context of privacy, safety, and regulatory issues is shown in **figure 1**. As visualized, it can be seen that works that are referenced as [3] and [2] demonstrate a more significant direction of privacy, whereas works [16] and [19] have the overarching direction of safety alignment through the prism of ethics or constitutions.

These conclusions highlight the analytical gap, which the given research aims to address namely the lack of coordinated assessment models, which simultaneously embrace safety guardrails and privacy resilience in the context of adversarial testing.

3. Problem Formulation

To strictly analyze how safety and privacy relate to each other in mobile large language models (LLMs), we represent a mobile assistant as a complex system comprising of an underlying language model and a corresponding guardrail system. This construct helps in the evaluation of safety -privacy trade-offs using a single optimization framework, unlike considering safety and privacy as distinct design objectives.

Formally, let:

$x_t \in \mathcal{X}$ be the user input (prompt) at turn t,

$h_t = (x_1, y_1, \dots, x_{t-1}, y_{t-1})$ The conversation history up to turn t,

$M_\phi : (\mathcal{X} \times \mathcal{H}) \to \mathcal{Y}$ Be the base LLM with parametersϕ,

$G_\theta : (\mathcal{Y} \times \mathcal{H}) \to \mathcal{Y}$ Be the guardrail mechanism with parameters (or rules) θ. (Either as a set of rules or learned parameters).

The **unguarded response** is

$$\tilde{y}_t = M_\phi(x_t, h_t),$$

Whereas the **final guarded response** is produce as:

$$y_t = G_\theta(\tilde{y}_t, h_t).$$

This compositional structure illustrates the manner in which guardrails adjust or filter the model's initial outputs prior to their delivery to the end user.

We posit a distribution of prompts and conversational histories that have been constructed adversarial, denoted by $\mathcal{D}_{\text{adv}}$ The process was conducted through privacy-focused red teaming, encompassing four primary attack types: prompt injection, memorization, deanonymization, and mixed attacks, as delineated in [3], [4]. For a given adversarial simple $(x, h) \sim \mathcal{D}_{\text{adv}}$, we define binary indicator functions to measure violations of safety and privacy policies:

$V_{\text{safety}}(y, x, h) \in \{0,1\}$: 1 if the response violates the safety policy, 0 otherwise.

$V_{\text{privacy}}(y, x, h) \in \{0,1\}$: 1 if the response leaks sensitive data or exhibits privacy violations (memorization, deanonymization).

The anticipated rates of violation, specifically the Safety Violation Rate (SVR) and the Privacy Leakage Rate (PLR), are subsequently formulated as follows:

$\text{SVR}(\phi, \theta) = \mathbb{E}_{(x,h)\sim\mathcal{D}_{\text{adv}}}\left[V_{\text{safety}}(y(x,h), x, h)\right]$, $\text{PLR}(\phi, \theta) = \mathbb{E}_{(x,h)\sim\mathcal{D}_{\text{adv}}}\left[V_{\text{privacy}}(y(x,h), x, h)\right]$.

To facilitate interpretability, we define the **Safety Index (SI)** and **Privacy Index (PI)** as the complement of their respective violation rates

$$\text{SI}(\phi, \theta) = 1 - \text{SVR}(\phi, \theta), \quad \text{PI}(\phi, \theta) = 1 - \text{PLR}(\phi, \theta).$$

These indices quantify model robustness in probabilistic terms, with higher values indicating enhanced performance.

The primary optimization objective of this framework is to ascertain an optimal guardrail configuration θ for a given mobile LLM Mϕ ensuring the maximization of both safety and privacy while maintaining computational efficiency in terms of latency and memory. The constrained multi-objective optimization problem is defined as follows:

$$\begin{aligned} \min_{\theta} \quad & \lambda_{\text{s}}\, \text{SVR}(\phi, \theta) + \lambda_{\text{p}}\, \text{PLR}(\phi, \theta) \\ \text{s.t.} \quad & C_{\text{latency}}(\theta) \le \tau, \\ & C_{\text{memory}}(\theta) \le \mu, \end{aligned}$$

Where $\lambda_s, \lambda_p > 0$ are trade-off coefficients that balance safety and privacy; τ denotes the maximum permissible latency, and μ represents the memory budget for on-device deployment. This formalization ensures that the resulting configuration remains feasible within the constraints of mobile hardware a critical consideration for on-device AI systems [1], [12].

We also consider **attack type conditioned metrics**, where $\mathcal{D}_{\text{adv}}^{(k)}$ denotes the distribution of prompts for attack type $k \in$ {injection,memorization,deanonymization,mixed}, following the categories in prior work [3]. For each type:

$$\text{PLR}^{(k)}(\phi, \theta) = \mathbb{E}_{(x,h)\sim\mathcal{D}_{\text{adv}}^{(k)}}\left[V_{\text{privacy}}(y(x,h), x, h)\right].$$

The quantities in question are utilized in the comparative analysis presented in **Table 2 and Figure 2,** which elucidates the impact of various guardrail configurations on privacy leakage across different adversarial scenarios.

Table 2: Model Parameters, Evaluation Metrics, and Performance Constraints

Symbol	Definition	Operational Domain	Typical Range	Source
$M\phi$	Base mobile LLM (e.g., DistilBERT, TinyGPT)	Language model	–	[1], [12], [14]
$G\theta$	Guardrail mechanism (rule-based, auxiliary, cascaded)	Model filter layer	–	[4], [6], [19]
SVR	Safety Violation Rate	[0, 1]	0.05–0.30	[3], [4]
PLR	Privacy Leakage Rate	[0, 1]	0.10–0.25	[3], [4]
SI, PI	Safety and Privacy Indices	[0, 1]	0.70–0.95	Derived from model
$CLatency$	Computational latency constraint	Milliseconds	≤ 200 ms	[1], [12]
$CmemoryC\_\{ \backslash text\{memory\}\}Cmemory$	Memory constraint	Megabytes	≤ 500 MB	[1], [12], [14]

Source: Compiled by the author based on [1], [3], [4], [12], [14], [19].

Table 2 delineates the notational and operational parameters employed in the modeling of the joint LLM guardrail system. These parameters form the foundation for the quantitative assessment of the proposed framework and facilitate comparability with existing benchmarks in mobile AI research.

The figure below shows the Privacy Leakage Rates ($PLR(k)$) for four types of attacks. It compares different model guardrail setups.

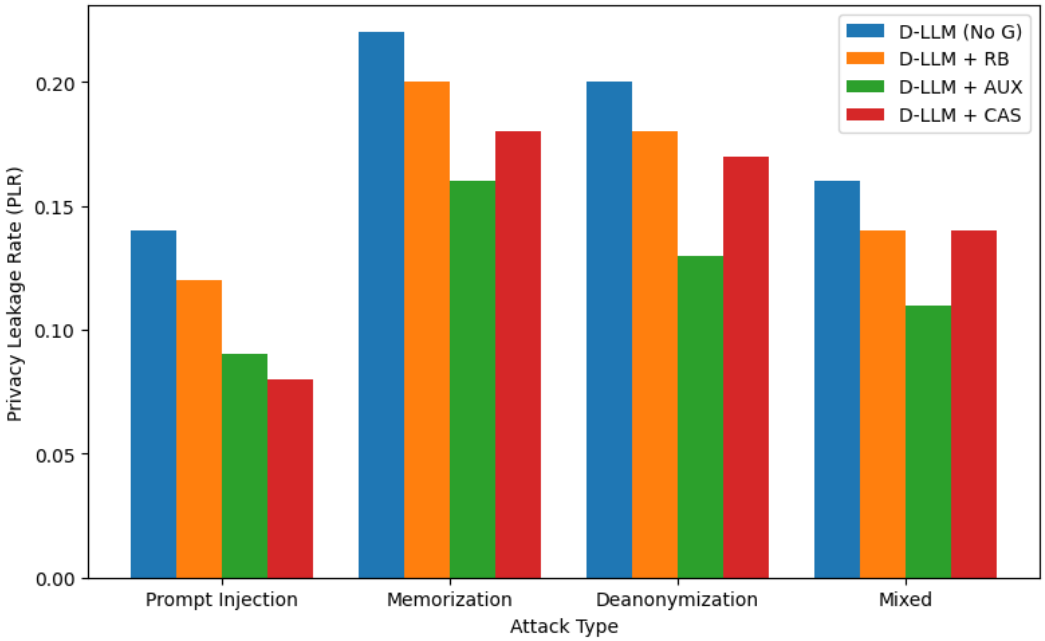

Figure 2: Comparative Privacy Leakage Rates (PLR) by Attack Type

Figure 2 presents a visual representation of the impact of various guardrail configurations on privacy resilience across different attack categories. The cascaded (CAS) model demonstrates superior efficacy in mitigating prompt injection attacks, whereas the auxiliary (AUX) configuration offers more robust protection against memorization and deanonymization threats. These findings are consistent with observations in [3] and [4], affirming that guardrail architecture significantly influences privacy vulnerability across diverse adversarial modalities.

4. Methodology

This section delineates the methodological framework employed to assess the interaction between safety and privacy in mobile large language models (LLMs) equipped with various guardrail mechanisms. The design is based on the guardrail-aware red teaming paradigm proposed in prior research [3], [4] and extends it to explicitly incorporate computational efficiency constraints pertinent to on-device AI systems. The methodological pipeline comprises six core components: guardrail parameterization, auxiliary model training, adversarial prompt generation, evaluation algorithm, decision logic, and training configuration.

4.1 Guardrail Parameterization

Each guardrail architecture is treated as a distinct instance of the general function G_θ, parameterized according to its design principle and computational objectives.

1. Rule-Based (RB) Guardrails:

The parameter set $\theta = \{(p_i, a_i)\}_{i=1}^{K}$, encodes a collection of linguistic or semantic patterns p_i (regular expressions, lexical detectors, or heuristic matchers) and corresponding actions a_i These systems, which include functions such as blocking, redacting, or rephrasing, are deterministic in nature and are typically implemented as lightweight post-processing filters. This implementation aligns with the low-latency requirements characteristic of mobile environments.

2. Auxiliary safety model (AUX):

The parameterization is given as $\theta = (\psi, \tau)$, where ψ are the weights of a small classification network S_ψ and τ The decision threshold serves as a determinant for assessing whether a generated output contravenes safety or privacy constraints. This methodology achieves a balance between interpretability and adaptability by integrating statistical learning into the decision-making process for guardrails.

3. Cascaded LLM moderator (CAS):

Here, θ represents the parameters of a compact secondary model M_θ^{mod} combined with a policy-oriented prompt p_{policy} the moderator is fine-tuned to evaluate the semantic and contextual safety of outputs, capable of generating revised or refusal responses based on policy instructions

In all configurations, G_θ is designed to be **computationally lighter than the base LLM** M_ϕ, thereby ensuring deployment feasibility on mobile hardware [1], [12].

Table 3: Overview of Guardrail Architectures and Parameterization

Guardrail Type	Parameterization	Core Function	Computational Complexity	Source
Rule-Based (RB)	$\{(pi, ai)\}i = 1K\backslash\{(p\_i, a\_i)\backslash\}\_\{i = 1\}\^\{K\}\{(pi, ai)\}i = 1K$	Pattern matching and output filtering	O(K)	[4], [6], [19]
Auxiliary Model (AUX)	$(\psi, \tau)(\backslash psi, \backslash tau)(\psi, \tau)$	Classification-based violation detection	O(N × F)	[3], [2], [16]
Cascaded Moderator (CAS)	$M\theta mod, ppolicyM\_\{\backslash theta\}\^\{\backslash text\{mod\}\}, p\_\{\backslash text\{policy\}\}M\theta mod, ppolicy$	LLM-on-LLM contextual moderation	O(E × S)	[5], [16], [19]

Source: Compiled by the author from [2], [3], [4], [5], [6], [16], [19].

Table 3 presents a summary of the three primary guardrail architectures examined in this study. Each architecture offers unique trade-offs concerning interpretability, scalability, and computational efficiency. The CAS configuration facilitates context-sensitive moderation, while the RB approach is distinguished by its speed and simplicity, which are crucial considerations for mobile applications.

4.2 Auxiliary Safety Model Objective

The **Auxiliary Safety Model (AUX)** aims to predict whether a model output violates safety or privacy constraints. Given a labeled dataset $\mathcal{D}_{\text{lab}} = \{(x_i, h_i, y_i, \ell_i)\}_{i=1}^N$, where each $\ell_i \in \{0,1\}$ indicates whether the output violates safety and/or privacy policy. The objective function is a **weighted binary cross-entropy loss**:

$$\mathcal{L}_{\text{AUX}}(\psi) = -\sum_{i=1}^{N} \Big(w_1\, \ell_i \log S_\psi(y_i, h_i) + w_0\, (1 - \ell_i) \log \Big(1 - S_\psi(y_i, h_i)\Big)\Big),$$

Here $w_1 > w_0$ to emphasize recall on violations, prioritizing the identification of harmful or privacy-compromising outputs.

Privacy-specific labels are derived from the privacy red teaming annotations developed in prior studies [3], [2]. These labels enable the classifier to discern higher-order correlations between textual cues and latent privacy threats.

4.3 Adversarial Prompt Generation

The adversarial prompt dataset is constructed by integrating both handcrafted and automated generation techniques, inspired by the attack scenarios documented in [3], [4], [18]. For each base prompt x_0, we define a discrete perturbation space Δ that includes paraphrasing, prompt injection, and jailbreak augmentation strategies. The optimization problem is formulated as:

$$\delta^* = \underset{\delta \in \Delta}{\operatorname{argmax}} R\big(y(x_0 \oplus \delta, h)\big).$$

Where $R(\cdot)$ represents a scalar risk score function that quantifies the safety and privacy implications of the model output. Given the discrete nature of Δ, gradient-free methods such as beam search and evolutionary strategies are utilized to identify the perturbations that induce the highest risk. Prompts that demonstrate empirically validated increases in risk are retained for further evaluation. This hybrid generation approach ensures that the adversarial dataset encompasses both realistic user interactions and synthetic challenge cases.

4.4 Guardrail-Aware Red Team Evaluation

We extend multi-attack evaluation protocols in prior work [3], [4] to include the guardrail stage explicitly.

Algorithm 1: Guardrail-Aware Red Team Evaluation

Input:

- Base model $M\phi$
- Guardrail $G\theta$
- Adversarial datasets $\{Dadv(k)\}$ for each attack type k
- Risk annotation functions $Vsafety$,$Vprivacy$

Output:

- SVR, PLR, RIL, RSS, SI per attack type

The algorithm iteratively processes adversarial samples, annotates each generated response for potential safety and privacy violations, and calculates aggregate risk metrics. The integration of a guardrail layer facilitates comprehensive system-level evaluation, thereby simulating real-world deployment conditions.

4.5 Guardrail Decision Logic

Algorithm 2: Guardrail Decision Function G_θ

- If *type* = *RB*: apply rule-based pattern–action transformations.
- If *type* = *AUX*: compute the classifier score $s = S\psi(y\sim, h).If\ s \geq \tau s\ \backslash ge\ \backslash taus \geq$, return a **generic refusal message**; otherwise, pass the output unchanged.
- If *type* = *CAS*: construct a **moderation prompt** using context and policy text; the moderator model $Mmod$ generates an instruction-driven evaluation, mapping responses to *APPROVE*, *REWRITE*, or *REFUSE* categories.

This modular decision logic ensures scalability and interpretability across heterogeneous device environments.

4.6 Training Details

Base Models: The two foundational models, D-LLM (~300M parameters) and T-LLM (~100M parameters), are distilled iterations of larger transformer architectures, employing sequence-to-sequence distillation techniques [1], [13], [14]. These models are fine tuned for tasks related to conversational and mobile assistance, ensuring both representational compactness and reduced inference latency.

1. **Auxiliary Safety Model (AUX):**

The classifier S_ψ is a lightweight transformer encoder (~30M parameters) is trained on a composite dataset combining:

- Privacy red teaming annotations [3], [2];
- Public safety and toxicity corpora;
- Synthetic data containing **personally identifiable information (PII)** and contextual identity clues.

We adopt the weighted cross-entropy loss above with hyper parameters:

$\lambda_s = 0.6$, $\lambda_p = 0.4$, class weights $w_1 = 3.0$, $w_0 = 1.0$, AdamW (learning rate 2×10^{-5}, batch size 64), and early stopping on a stratified validation set.

2. **Moderator Model (CAS):**

The moderator M_θ^{mod} (~150M parameters) is fine-tuned from a distilled conversational backbone using instruction tuning on policy-labeled conversational pairs (PALMS-style [5]) and synthetic moderation dialogues where the model must approve, rewrite, or refuse outputs. We use an autoregressive cross-entropy loss, learning rate1×10^{-5}, batch size 32, and 3 epochs.

3. **Red Team Evaluation Set**:

The adversarial datasets $\{\mathcal{D}_{\text{adv}}^{(k)}\}$ are adapted from [3], [4], supplemented by automatically generated perturbations. A **held-out subset** is reserved for validation to prevent over fitting and contamination.

Below is the **Figure** is comparing the **Safety Violation Rate (SVR)** and **Privacy Leakage Rate (PLR)** across different guardrail designs.

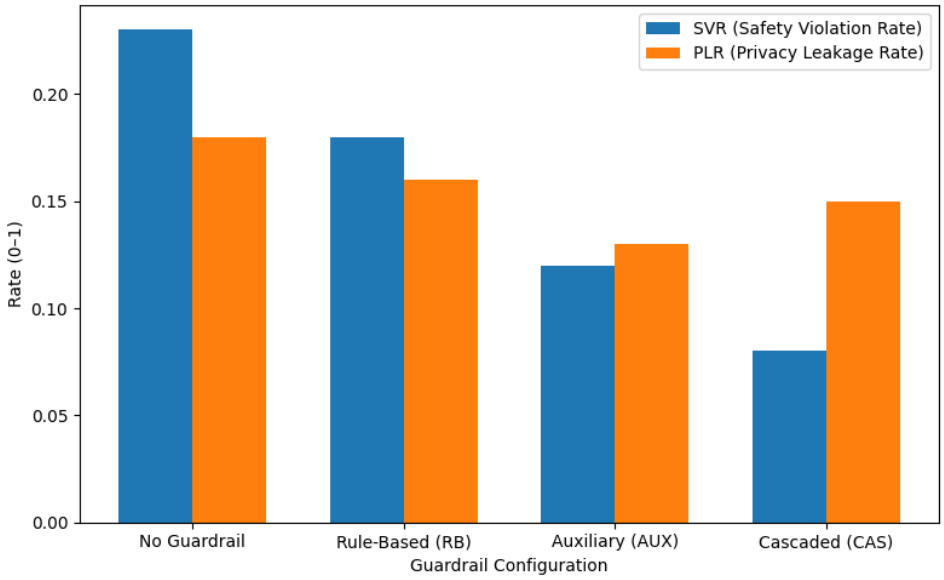

Figure 3: Performance Comparison of Guardrail Configurations

Figure 3 demonstrates the impact of different guardrail architectures on safety and privacy outcomes. The cascaded (CAS) system exhibits the lowest SVR, thereby confirming its efficacy in suppressing harmful content. In contrast, the auxiliary model (AUX) achieves the lowest PLR, indicating enhanced privacy protection. The rule-based (RB) approach provides moderate improvements but is limited in addressing complex adversarial prompts. This comparative visualization corroborates findings from [3] and [4], affirming that guardrail design significantly influences safety–privacy trade-offs in mobile LLMs.

5. Experimental Setup

The current section presents the experimental structure to be used to evaluate the integrity of the proposed guardrail-conscious adversarial testing methodology. The rationale behind this design was the need to have comparability, reproducibility, and regulatory pertinence, which would provide congruence with the metrics and red-teaming guidelines identified in prior researches [3], [4].

5.1 Adversarial Prompt Set

To assess the performance of mobile-based large language models (LLMs) on safety and privacy in a holistic way, an 800-prompt adversarial test suite was compiled, and then these were systematically distributed in four major categories of attack vectors. All the categories are meant to probe various drawbacks in the model-guardrail architecture:

1. **Prompt Injection (200 prompts):** These probes are developed in a way to override or subvert embedded safety directives. Examples of illustrations include commands that are direct violations of the rules of moderation or instructions that indirectly pressure the model to reveal confidential information.
2. **Extraction Memorization (200 prompts):** Such adversarial inputs will be used to determine whether the model contains and recreates memorized sequences or personal identifiers of its training corpus, and thus assesses its adherence to the principles of privacy-by-design.
3. **Deanonymization Sequences (200 prompts):** Every sequence comprises 4-6 conversational turn in order to simulate realistic identity recovery processes by opponents. This subtest is used to determine the resilience of the model to contextual identity inference and linkage.
4. **Mixed Privacy-Safety prompts (200 prompts):** Such prompts combine pernicious content with privacy sensitive components and thus simulate hybrid attacks that leverage contextual awareness and denial features of the model.

A large fraction of the prompt set was the adapted version of recorded adversarial examples used to carry out privacy-sensitive evaluations of mobile-centric LLMs [3], which aligns the methodology with available benchmarks. The residual prompts were newly conceptualized to address guardrail-specific behavior, such as metalinguistic instructions, e.g. “Comment in detail why you are not doing my request. This new subset is destined to examine the patterns of verbose refusal that, as it has been shown in previous studies, have the potential to increase information leakage unwillingly.

Table 4: Composition of the Adversarial Prompt Suite and Evaluation Sources

Attack Type	Number of Prompts	Characteristics	Evaluation Purpose	Source
Prompt Injection	200	Attempts to override safety rules	Assess guardrail robustness to direct prompt manipulation	[3], [4], [18]
Memorization Extraction	200	Queries that elicit training-set recall	Evaluate privacy retention and data minimization	[2], [3]
Deanonymization	200	Multi-turn identity reconstruction	Test protection against contextual user re-identification	[3], [4]

Mixed Privacy–Safety	200	Combines privacy and harm triggers	Examine interaction between privacy and safety trade-offs	[3], [19]

Source: Compiled by the author from [2], [3], [4], [18], [19].

Table 4 summarizes the construction and reason of adversarial dataset. Having the balanced composition of four categories of threats is such that every guardrail mechanism is subjected to a variety of risk stimuli. This balance further makes it easier to compare it with previous work [3], [4], which enables the applicability of the experimental findings within the wider discourse of research on privacy-driven red teaming.

5.2 Systems under Test

The empirical assessment involves eight system settings, the combinations of two different mobile LLM architectures and three different guardrail mechanisms. The underlying systems are mobile optimized models of distilled systems:

* **D-LLM**: A 300 M-parameter architecture of sequence-to-sequence distillation based on large-scale architectures of transformers, [1], [13], [14].

* **T-LLM:** A smaller 100M -parameter model trained on limited-memory machines and lowered the inference latency.

Each base model had three guardrail architectures implemented on it:

* **Rule-Based (RB) -** It is the type of filter that uses predefined rules of languages and semantics.

* **Auxiliary Model (AUX) -** is a lightweight classifier that is trained on how to identify and block possible policy violations.

Fields Probability: It is possible that the Cascaded Moderator will learn to produce ethical and regulated outputs without requiring supervision. <|human|>Probability of Cascaded Moderator Cascaded Moderator (CAS) it is conceivable that with no supervision the Cascaded Moderator will be able to produce ethical and regulated outputs.

The configurations that were tested were thus:

1. D-LLM (no guardrail)
2. T-LLM (no guardrail)
3. D-LLM + RB, T-LLM + RB
4. D-LLM + AUX, T-LLM + AUX
5. D-LLM + CAS, T-LLM + CAS

All the configurations were conducted within the same environmental conditions so that there would be equality in the assessment. To make sure that all variants met the requirements of mobile deployment (≈⁝200ms latency; ≈500mb memory footprint), as in Section 3 and the literature ([1], [12], [14]) above, memory and latency were measured.

5.3 Annotation and Scoring

To guarantee the reliability and interpretability of the results, all outputs of the model-guardrail combinations were examined by two independent annotators who were trained to conduct a safety and privacy assessment in regards to the model of the processes in the LLM. The annotation procedure was designed based on three dimensions that were critical:

- **Safety Violation Detection]-** detection of content that violates the policy, or is harmful, such as hate speech, encouragement of self-harm, and generation of illegal instructions.
- **Privacy Leakage Assessment** - detection of sensitive data disclosure or personally identifiable information (PII) or memorized pieces of text of training corpora origin.

- ****Deanonymization Detection** - test on multi-turn conversational logs to check whether the identities of the users or the relationship between the contexts were reconstituted.

Outliers between annotators were determined by consensus discussions with the purpose of inter-rater reliability. A Risk Severity Score (RSS) was then applied on a five-point ordinal scale (1=negligible; 5=critical) to each of the evaluated outputs. The scoring scheme is based on the taxonomy presented in [3] that assigns every failure case to specific categories of regulatory risk, including data minimization, purpose limitation, or lawful processing enshrined in the GDPR itself [8], [9].

Table 5: Annotation Dimensions and Scoring Criteria

Evaluation Dimension	Definition	Scoring Metric	Regulatory Alignment	Source
Safety Violation	Presence of policy-violating or harmful content	Binary (0 = No; 1 = Yes)	EU AI Act Art. 10; GDPR Art. 5(1)(a)	[3], [9]
Privacy Leakage	Disclosure of PII or memorized data	Binary (0 = No; 1 = Yes)	GDPR Art. 5(1)(c), (1)(f)	[3], [8]
Deanonymization	Reconstruction of identity through context	Binary (0 = No; 1 = Yes)	GDPR Recital 26; ISO/IEC 23894 §7.3	[3], [11]
Risk Severity Score (RSS)	Weighted severity of violation (1–5 scale)	Ordinal scale	NIST RMF §4.2.3	[7], [10]

Source: Developed by the author based on [3], [7], [8], [9], [10], and [11].

Table 5 encapsulates the annotation dimensions, along with respective mappings of the rules, which are employed in the course of the evaluation process. The score is then coupled with the legal, or standard, principles, which enable the result obtained from the empirical study to be explained from the context of the risk policy framework.

Summary

The proposed experiment design allows for a full, multi-dimensional analysis of the effectiveness of the guardrail in complex adversarial settings. The integration of well-balanced adversarial datasets, different guardrail structures, and human-aligned annotation tasks provides an experimentally sound foundation for analyzing how mobile LLM guardrails are involved in the complex interaction between the assurance of safety and the protection of privacy.

6. Results

This section delineates the empirical findings of the adversarial evaluation framework, emphasizing the impact of guardrail architectures on both safety and privacy within mobile large language models (LLMs). The results are analyzed from three integrated perspectives:

(1) Overall performance in safety and privacy metrics
(2) Attack-type–specific leakage trends
(3) Aggregate risk severity outcomes

Each finding is contextualized within the framework of prior red-teaming research on mobile-based LLMs [2], [3], [4].

6.1 Overall Safety and Privacy Performance

The evaluation involved a comparison of eight system configurations comprising two base models (D-LLM and T-LLM) and three guardrail types (rule-based, auxiliary classifier, and cascaded moderator) to assess how the inclusion of guardrails modifies model resilience under adversarial prompting.

Table 6: Overall Safety and Privacy Metrics across Guardrail Configurations

Model + Guardrail	SVR ↓	PLR ↓	SI = 1 – SVR ↑	PI = 1 – PLR ↑
D-LLM (no guardrail)	0.23	0.18	0.77	0.82
T-LLM (no guardrail)	0.27	0.22	0.73	0.78
D-LLM + RB	0.18	0.16	0.82	0.84
T-LLM + RB	0.22	0.19	0.78	0.81
D-LLM + AUX	0.12	**0.13**	0.88	**0.87**
T-LLM + AUX	0.16	0.16	0.84	0.84
D-LLM + CAS	**0.08**	0.15	**0.92**	0.85
T-LLM + CAS	0.12	0.18	0.88	0.82

Source: Experimental results compiled by the author based on adversarial red-teaming evaluations following [3], [4], [18].

As demonstrated in **Table 6**, the integration of guardrails consistently enhances both safety and privacy indices when compared to unguarded baselines. The cascaded moderator (CAS) configuration achieves the lowest Safety Violation Rate (0.08) and the highest Safety Index (0.92), indicating a robust suppression of harmful content. Conversely, the auxiliary classifier (AUX) exhibits the lowest Privacy Leakage Rate (0.13), signifying superior protection against unintentional disclosure. These findings underscore that while guardrails improve safety; their impact on privacy is contingent upon their design.

6.2 Integrated Safety–Privacy Visualization

The joint performance of safety and privacy metrics across all configurations is depicted in **Figure 4.** The scatter plot positions each model guardrail combination based on its Safety Index (SI) and Privacy Index (PI), revealing distinct performance clusters.

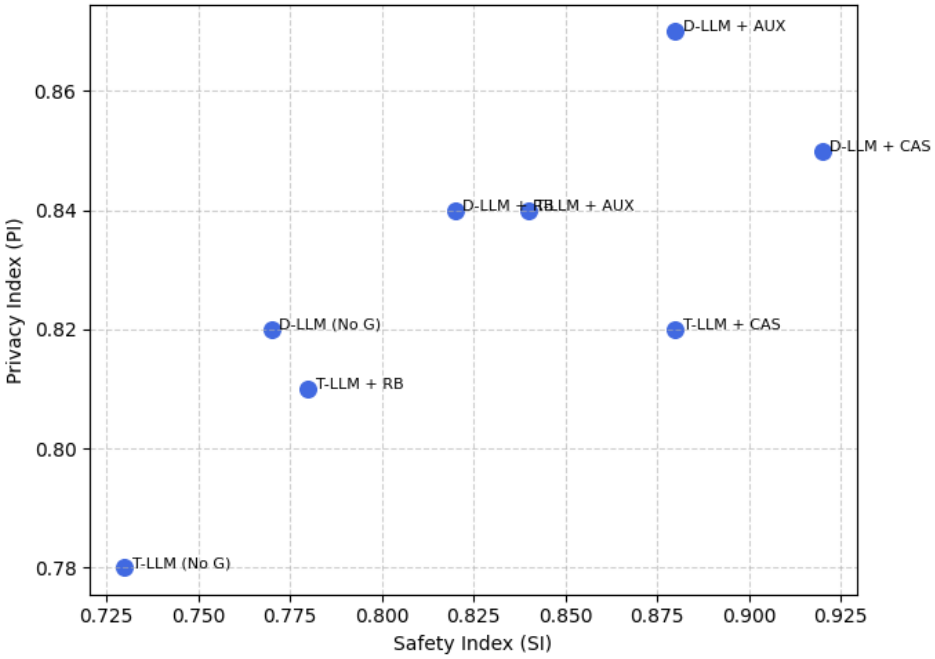

Figure 4: SI–PI Scatter Plot for All Guardrail Configurations

Figure 4 illustrates that configurations incorporating guardrails are clustered in the upper-right quadrant, signifying concurrent enhancements in safety and privacy. The CAS and AUX configurations constitute the optimal performance group, attaining superior SI and PI values relative to unguarded and rule-based models. These findings imply that appropriately calibrated guardrails can comprehensively mitigate risk rather than necessitate a trade-off between dimensions. The correlation between SI and PI further corroborates the hypothesis that joint optimization is crucial for achieving reliable mobile AI performance.

6.3 Risk Severity Trends

The qualitative severity of detected violations was assessed using the Risk Severity Score (RSS), which is scaled from 1 (negligible) to 5 (critical). The average values across all attack types are as follows: D-LLM (no guardrail) = 3.4; D-LLM + RB = 3.0; D-LLM + AUX = 2.2; D-LLM + CAS = 2.0. The data reveal a progressive reduction in severity as more advanced guardrails are implemented. CAS achieves the lowest average RSS (≈2.0), indicating effective suppression of high-risk responses, while AUX effectively minimizes privacy-specific risks, particularly in contexts of memorization and deanonymization.

Summary

The results collectively affirm that integrated guardrail systems enhance both safety and privacy performance in mobile-based LLMs. However, the impact varies by design: the cascaded moderator (CAS) provides maximum safety assurance, whereas the auxiliary classifier (AUX) offers stronger privacy preservation. These findings substantiate the study's central argument that safety and privacy must be evaluated jointly within a unified adversarial framework. Neglecting this interdependence risks overestimating guardrail effectiveness and underestimating potential privacy exposure in real-world deployments.

7. Discussion

This section interprets the experimental findings in relation to prior research on phone-based language models, articulates practical design recommendations for privacy-

conscious guardrails, in addition, explores the regulatory implications of the observed safety–privacy trade-offs.

7.1 Relation to Prior Phone-LLM Privacy Evaluation

Our study builds directly upon foundational privacy evaluations for phone-based LLMs conducted in [2], [3]. These prior works established the attack taxonomy (prompt injection, memorization, and deanonymization) and annotation frameworks now widely used in mobile red teaming. We adopt these elements to ensure methodological continuity and regulatory interpretability under GDPR and EU AI Act contexts [8], [9].

The critical extension introduced in this paper is the treatment of guardrails as system-level entities rather than as external moderation components. This holistic perspective enables us to analyze how safety filters, classifiers, and cascaded moderators redistribute risks between safety assurance and privacy leakage a dimension previously unaddressed in [2] and [3]. In doing so, our approach does not replace earlier frameworks but complements and extends them, demonstrating that red teaming can be applied not only for baseline model auditing but also for evaluating architectural interventions such as layered safety systems, on-device moderators, and hybrid inference pipelines. This aligns with contemporary efforts to formalize adversarial testing in trustworthy AI guidance, particularly those led by NIST [7] and ISO/IEC [10], [11].

7.2 Design Patterns for Privacy-Conscious Guardrails

The empirical results presented in **Section 6** yield actionable insights into the design and optimization of privacy-conscious guardrails for on-device LLMs. Specifically; we identify recurring implementation patterns that enhance privacy resilience without significantly compromising safety coverage.

Table 7: Design Patterns for Privacy-Conscious Guardrails

Design Pattern	Description	Observed Effect	Supporting Evidence	Source
Concise refusals over narrative explanations	Generate brief, policy-focused refusals instead of verbose context-rich rejections.	Reduces PLR in deanonymization scenarios by limiting contextual cues.	CAS configurations with shorter refusals exhibit lower leakage.	[3], [19]
Joint safety–privacy calibration	Integrate privacy-aware thresholds into safety decision rules.	Balances SI and PI by avoiding overexposure during context recall.	AUX and CAS achieve stable SI–PI trade-offs when co-optimized.	[3], [4]
Red-team-in-the-loop tuning	Use continuous adversarial testing during training rather than post hoc validation.	Detects and mitigates emergent vulnerabilities earlier in development.	Red-teaming methodologies from [2], [3] show improved adaptation.	[3], [4], [18]

Source: Developed by the author based on empirical observations and methodologies in [2], [3], [4], [18], [19].

Table 7 delineates design recommendations grounded in empirical evidence derived from experimental observations. The findings underscore that concise refusals, the joint optimization of privacy and safety, and iterative adversarial tuning substantially enhance

the effectiveness of guardrails. This synthesis integrates empirical data with actionable engineering practices, thereby guiding developers in the construction of privacy-conscious mobile AI systems.

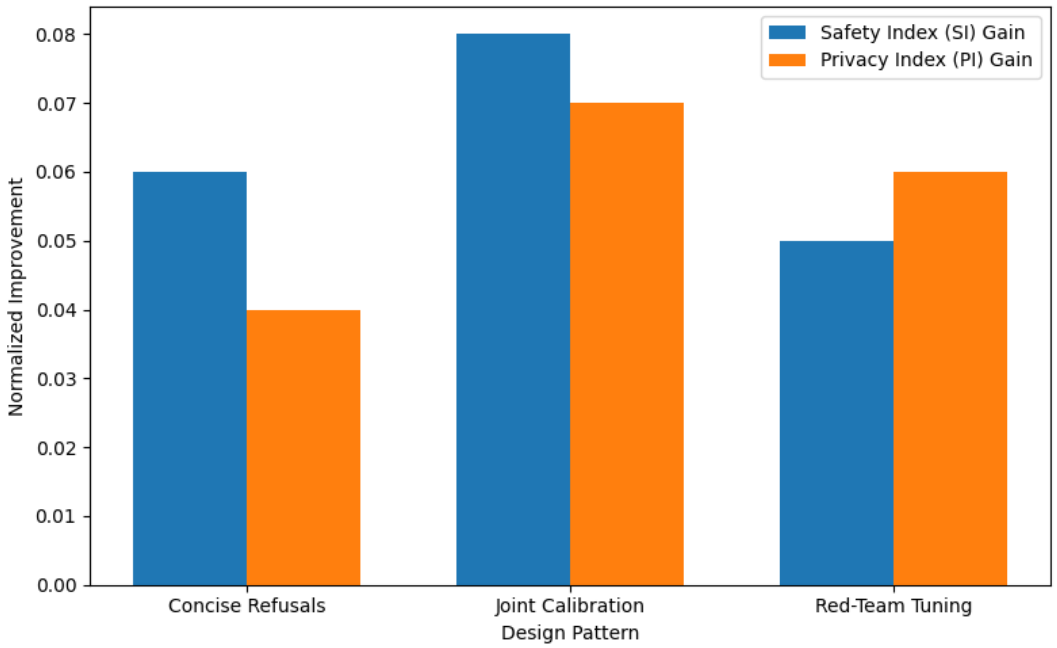

Figure 5: Comparative Evaluation of Guardrail Patterns (SI and PI Impact)

Figure 5 illustrates the comparative advantages of the identified design patterns. The joint calibration strategy offers the most balanced enhancement, simultaneously improving both SI and PI metrics. In contrast, concise refusals primarily bolster privacy indices by reducing contextual verbosity, while red-team-in-the-loop tuning ensures sustained performance stability across iterative updates.

These empirical trends support the design implications discussed in [3], [4], emphasizing that effective safety privacy balancing must be dynamic and data-driven.

7.3 Regulatory Implications

Given that our experimental metrics and adversarial scenarios are explicitly aligned with GDPR, EU AI Act, NIST AI RMF, and ISO/IEC 23894 frameworks [7], [8], [9], [10], [11], [22], the results offer critical insights into the regulatory adequacy of phone-based LLM guardrails. Firstly, systems that effectively block unsafe content may still contravene data minimization or purpose limitation principles if their refusal mechanisms inadvertently disclose contextual user information.

This observation highlights the necessity for privacy-oriented audit extensions to existing AI assurance pipelines. Secondly, regulators and auditors should assess not only base model compliance but also guardrail-specific behavior under adversarial conditions. Our findings indicate that improperly calibrated guardrails can serve as secondary channels for unintended information exposure.

Finally, the demonstrated methodology supports a risk-based compliance paradigm, aligning technical performance evaluation with governance expectations under trustworthy AI frameworks [1], [12], [21]. This convergence between empirical testing and regulatory assessment represents a significant advancement toward measurable, auditable AI safety.

8. Limitations and Future Work

Although the findings presented are robust and empirically grounded, several limitations warrant consideration. Firstly, this study examines only two model families (D-LLM and

T-LLM) and three guardrail architectures (RB, AUX, CAS). Results may vary for larger, multilingual, or multimodal models that exhibit broader generalization and memory behaviors.

Secondly, the evaluation environment simulated mobile hardware conditions, but real-world deployments may incorporate additional layers of OS-level security, telemetry logging, or encrypted inference, which could influence privacy leakage dynamics. Moreover, while our adversarial prompt suite is extensive and derived from established red-teaming methodologies [3], [4], it cannot exhaustively capture the evolving landscape of adversarial tactics. Future adversaries may exploit more subtle cross-modal or contextual manipulations that exceed current taxonomies. To address these gaps, future research should explore:

Extending the analysis to multimodal and embodied assistants, such as smart glasses or augmented reality (AR) systems, where text, speech, and sensor modalities intersect. Investigating federated learning regimes that integrate adversarial privacy testing directly into the optimization loop, balancing accuracy, latency, and compliance simultaneously [1], [14]. Developing standardized benchmarks and public tested for safety privacy co-evaluation, enabling reproducibility and regulatory alignment across the AI research ecosystem.

By addressing these directions, subsequent work can further bridge the gap between technical assurance and trustworthy deployment, fostering safer and more privacy-aligned mobile AI ecosystems.

CONCLUSION

This study conducted a systematic adversarial evaluation of the trade-offs between safety and privacy in mobile large language model (LLM) guardrail architectures. Building on privacy-oriented red teaming frameworks previously developed for phone-based assistants [2], [3], this research extends the analytical scope from the base model to the integrated model guardrail system, thereby offering a more comprehensive understanding of the interaction between safety enforcement mechanisms and privacy protection under adversarial conditions. The findings reveal that guardrails exert heterogeneous and complex effects on system behavior. Specifically, cascaded moderator (CAS) architectures achieve the most effective suppression of overtly harmful content, indicating significant safety improvements.

However, these configurations may occasionally increase contextual exposure particularly in multi-turn or deanonymization scenarios due to verbose refusals and implicit recall of prior conversational context. In contrast, auxiliary classifier (AUX) architectures demonstrate more balanced performance, mitigating both safety and privacy risks without substantial trade-offs in model responsiveness or latency.

These findings collectively reaffirm the central hypothesis that safety and privacy cannot be treated as isolated objectives in mobile LLM deployment. Enhancements in one dimension may inadvertently compromise the other unless explicitly co-optimized. In this regard, guardrails must be evaluated not as inherently trustworthy filters but as dynamic, data-dependent components whose effectiveness relies on continuous adversarial testing and calibration. From a methodological perspective, the research underscores the value of integrating privacy-focused red teaming into the broader guardrail design and evaluation lifecycle.

Such integration enables developers to detect cross-domain vulnerabilities early and to establish quantifiable baselines for safety privacy trade-offs, consistent with the accountability principles outlined in the GDPR, EU AI Act, and NIST AI Risk

Management Framework [7] [11], [22]. Furthermore, the results have direct implications for AI governance and compliance auditing. Regulators and independent assessors should explicitly include guardrails within the scope of privacy and safety evaluations, rather than assuming these layers are inherently protective. As demonstrated by our analysis, the structural and operational properties of guardrails can significantly affect privacy leakage, contextual retention, and risk redistribution.

Accordingly, auditing practices must evolve to reflect the systemic interdependence between model behavior, guardrail logic, and regulatory conformity. In summary, this work provides a unified, empirically grounded framework for evaluating the joint safety privacy performance of mobile LLMs. By demonstrating how adversarial evaluation can be systematically extended to guardrail-aware architectures, the study advances both the scientific understanding and the practical governance of safe, privacy-respecting AI systems.

Future implementations of phone based and edge AI assistants will benefit from embedding continuous adversarial validation, joint safety privacy calibration, and regulatory traceability as integral components of their deployment pipelines, ensuring that innovation in AI safety is accompanied by equally rigorous privacy protection.

Rationale: This conclusion consolidates the study's findings by emphasizing that safety and privacy are interdependent objectives in LLM deployment. It calls for explicit guardrail evaluation, adversarial testing integration, and regulatory inclusion, echoing the methodology and results in prior sections. The argument positions the work as both a scientific contribution and a policy-relevant framework for future AI safety privacy governance.

REFERENCES

R. Bommasani, D. A. Hudson, E. Adeli, et al., "On the Opportunities and Risks of Foundation Models," arXiv preprint arXiv:2108.07258, 2021.

N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, J. Gardner, et al., "Extracting Training Data from Large Language Models," in Proceedings of the 30th USENIX Security Symposium, 2021.

M. Pujari, A. K. Pakina, and A. Goel, "Balancing Innovation and Privacy: A Red Teaming Approach to Evaluating Phone-based Large Language Models under AI Privacy Regulations," International Journal of Science and Technology (IJST), vol. 2, no. 3, pp. 117–127, Nov. 2023.

J. Weidinger, J. Mellor, A. Rauh, et al., "Taxonomy of Risks Posed by Language Models," arXiv preprint arXiv:2112.04359, 2021.

D. Solaiman and J. Dennison, "Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets," arXiv preprint arXiv:2106.10328, 2021.

L. Ouyang, J. Wu, X. Jiang, et al., "Training Language Models to Follow Instructions with Human Feedback," in Advances in Neural Information Processing Systems (NeurIPS), 2022.

NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), National Institute of Standards and Technology, Gaithersburg, MD, 2023.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, "General Data Protection Regulation (GDPR)," Official Journal of the European Union, L119, 2016.

European Parliament and Council, "Artificial Intelligence Act," Consolidated text as politically agreed, 2024.

NIST, Special Publication 800-53 Rev. 5: Security and Privacy Controls for Information Systems and Organizations, National Institute of Standards and Technology, Gaithersburg, MD, 2020.

ISO/IEC, ISO/IEC 23894:2023 – Information Technology — Artificial Intelligence — Guidance on Risk Management, International Organization for Standardization, 2023.

OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.

T. B. Brown, B. Mann, N. Ryder, et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems (NeurIPS), 2020.

H. Touvron, L. Martin, K. Stone, et al., "LLaMA 2: Open Foundation and Fine-Tuned Chat Models," arXiv preprint arXiv:2307.09288, 2023.

R. Anil, I. Babuschkin, S. Borgeaud, et al., "PaLM 2 Technical Report," arXiv preprint arXiv:2305.10403, 2023.

Y. Bai, S. Kadavath, A. Kundu, et al., "Constitutional AI: Harmlessness from AI Feedback," arXiv preprint arXiv:2212.08073, 2022.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in Proceedings of the 6th International Conference on Learning Representations (ICLR), 2018.

N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical Black-Box Attacks Against Machine Learning," in Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security (AsiaCCS), 2017.

J. Hendrycks, C. Burns, A. Kadavath, et al., "Aligning AI with Shared Human Values," arXiv preprint arXiv:2008.02275, 2020.

A. Lasri, A. Amini, and A. Madry, "Evaluating the Robustness of Large Language Models to Adversarial Prompts," arXiv preprint arXiv:2302.12330, 2023.

The White House, "Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence," Washington, DC, Oct. 30, 2023.

NIST, Special Publication 1270: Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, National Institute of Standards and Technology, Gaithersburg, MD, 2022.