International Journal Science and Technology
IJST Vol 2 No. 3 | November 2023 | ISSN: 2828-7223 (print), ISSN: 2828-7045 (online), Page 117-127

Balancing Innovation and Privacy: A Red Teaming Approach to Evaluating Phone-Based Large Language Models under AI Privacy Regulations

Mangesh Pujari[1*], Anil Kumar Pakina[2], Anshul Goel[3]
[1,2,3]Independent Researcher, India

Article History
Received : September 2023
Revised : September 2023
Accepted : September 2023
Published : October 2023

Cite This Article:
Mangesh Pujari, Anil Kumar Pakina, and Anshul Goel, "Balancing Innovation and Privacy: A Red Teaming Approach to Evaluating Phone-Based Large Language Models under AI Privacy Regulations", *IJST*, vol. 2, no. 3, Oct. 2023.

DOI:
https://doi.org/10.56127/ijst.v2i3.1956

Abstract: The rapid deployment of large language models (LLMs) on mobile devices has introduced significant privacy concerns, particularly regarding data collection, user profiling, and compliance with evolving AI regulations such as the GDPR and the AI Act. While these on-device LLMs promise improved latency and user experience, their potential to inadvertently leak sensitive information remains understudied. This paper proposes a red teaming framework to systematically assess the privacy risks of phone-based LLMs, simulating adversarial attacks to identify vulnerabilities in model behavior, data storage, and inference processes.
We evaluate popular mobile LLMs under scenarios such as prompt injection, side-channel exploitation, and unintended memorization, measuring their compliance with strict privacy-by-design principles. Our findings reveal critical gaps in current safeguards, including susceptibility to context-aware deanonymization and insufficient data minimization. We further discuss regulatory implications, advocating for adaptive red teaming as a mandatory evaluation step in AI governance. By integrating adversarial testing into the development lifecycle, stakeholders can preemptively align phone-based AI systems with legal and ethical privacy standards while maintaining functional utility.

Keyword: Large Language Models (LLMs), Red Teaming, AI Privacy, Mobile AI, Privacy Regulations, Adversarial Testing, Model Evaluation, Data Security, Innovation and Ethics, Privacy-Preserving AI

INTRODUCTION

Documentation-Context and Motivation

The emergence of large language models (LLMs) has revolutionized the way humans interact with the world's technology at large. Technology, which was hitherto maintained in large cloud infrastructures, is now squeezing downsized LLMs and making them available for mobile deployment. Phone-based LLMs have much promise, including faster response times; personalized delivery; reduced network dependence; and greater privacy through on-device computation.

The same on-device LLMs come with some severe privacy and security concerns. The mobile environment is fragmented and user-managed without much overhead and central governance in comparison with centralized architectures. Thus, sudden intrusion by LLM could imply a form of inadvertent hazardous keeping of sensitive information, vulnerabilities due to prompt injection, and indirect profiling through latent data extraction mechanisms.

What with the increasing nature of complexities and obscurities in modern LLMs, especially those cleared for specific mobile use, the realization of their behavior in actual adversary settings becomes very critical. Existing methodologies for rhetoric privacy evaluation are mostly insufficiently vigorous to approximate malign probing-a glaring reason behind the motivation for more aggressive methods to evaluation. Thus, a suitable introductory course is set for the developing agenda to raise red teaming as an

organized model for assessment and unveiling of breaches against privacy in the case of phone LLMs ecosystems.

Regulation and Enforcement Challenges in AI Privacy

All governments and authorities across the globe have authorities that impose laws to stop all forms of malpractice against AI and also violations of personal rights. The most important legislation from the angle of the User bases is the European General Data Protection Regulation 2016 (GDPR), which stresses principles like privacy-by-design, data minimization, express consent, and transparency. The upcoming European Union Artificial Intelligence Act (AI Act), apart from all of that, will further categorize AI systems in terms of risk tiers and thereby lay out requirements and extensive prohibitory laws for high-risk systems like facial recognition, behavior manipulation, and profiling, among others.

All well and good, the legal foundation has been laid, but implementation comes as a challenge, and particularly in the mobile domain. Application, from which the LLM gains money while the consumer does little more than occasionally accessing it, bears little external auditing. Their behaviors such as latent memorization or exposure through prompt interaction are not exam-able through static inspection tools. Hence, many LLMs find themselves again infringing fundamental regulatory principles, although they may declare that they are up to compliance with the above-mentioned rules.

Gaps of Traditional Privacy Audit

They focus more on traditional forms of privacy assessment such as paperwork analysis, checklists and rules-based testing. While necessary, these approaches usually fall short of capturing emergent behaviors of LLMs that result from fine-tuning, continual learning, or from adversarial prompting. Furthermore, privacy risk assessment tends to be reactive-identifying harm after its occurrence rather than actively simulating attack on vectors.

For example, personal identifiers or sensitive phrases can be extracted when using inference in even models for which differential privacy mechanisms have been applied in training. Contextual embeddings or chains of prompts can change how a model behaves, thus demanding evaluation as dynamic rather than snapshot inspection. These issues are especially pronounced on mobile platforms, where memory, computation, and energy constraints restrict the implementation of robust monitoring systems.

Red Teaming as a Privacy Audit Proactive Tool Thus

In AI safety and ethics circles, red teaming has been history regarding cybersecurity: "That's how you find a loophole by breed-its effect by simulating adversarial experiment activities." To investigate such incited vulnerabilities, it is characterized as deliberate deficit model probing under realistic threat scenarios uncovering inadvertently dangerous or privacy-compromising outputs. The outcomes in red teaming would then be different in the context of mobile-based LLMs as it would investigate data leaks, identity disclosures, prompt hijacks, and abuses of contextual memory-all of which would so often miss standard evaluation procedures.

This process not only shakes down vulnerability but also strengthens defense by incorporating the learning in the development lifecycle. Iterative red teaming can be performed on the models before massive deployment and reduces regulatory risks besides instilling trust in users.

Research Objectives

This study intends to design, implement, and verify a red teaming framework for evaluating the privacy resilience characteristics of LLMs when deployed on cell phones. Objectives include,

1. Identifying key mobile LLM privacy risks.
2. Designing substantial triggering events for simulating prompt injection, leakages, and memory.
3. To empirically evaluate some common mobile LLMs under red team stress testing.
4. To classify observed flaws concerning regulatory requirements e.g. clauses of the GDPR and the AI Act mutually.
5. To incorporate privacy-centric red teaming in LLM development cycles.

LITERATURE REVIEW

The Foundations of Privacy in AI Systems

Privacy forms a central tenet for ethical artificial intelligence and is becoming increasingly pertinent with the emergence of omnipresent AI-enabled systems. In particular, mobile-based AI astuteness has heightened concern of data protection, user agency, and algorithmic accountability. One of the legally enforceable mandates introduced in the 2018 General Data Protection Regulation (GDPR) was data minimization through purpose limitation and the right to explanation. These have been widely cited as models for widespread academic and commercial efforts towards the design of AI systems that perform satisfactorily but are also legally permissible (Veale & Binns, 2017).

Still, much of the underlying detail of these regulations is too often contrasted with reality. Existing AI systems even those grand as large language models , hardly comply with the principles of privacy by design. Disposition problems, such as traditional anonymization techniques of data, are insufficient to address the fact that today power and inference engines allow to re-identify data subjects through contextual clues (Narayanan and Shmatikov, 2008). Besides, issues related to model memorization, prompt injection, and training data leaks are generally unresolved in on-device deployments.

LLMs Getting Mobile: The Rise of On-Device AI

It has recently become theoretically and practically possible to deploy LLMs onto mobile platforms through advances in model compression, pruning, quantization, and knowledge distillation. Lightweight architectures such as DistilBERT, MobileBERT, and TinyGPT can run now with reduced memory and compute requirements, allowing for interaction without cloud reliance and touchdown real time (Sanh et al., 2019).

However, such credit gained for the improvements it heralded by way of responsiveness and latency is usually traded off. Limited hardware resources mean that fortification does not happen for the most part when it comes to good monitoring, encryption, or even sandboxing arrangements. Additionally, although there are corrections made on mobile platforms, updates of models may not be as frequent as cloud updates, thus increasing the chance that the vulnerabilities of the model will be present after deployment.

This has vast implications for privacy. Mobile LLMs, for instance, might cache prompts or hold memories of partial histories for an optimal user experience. Behavior in which specific user consent or awareness is needed violates data protection with respect to the GDPR clauses on informed consent and transparent data processing.

Red Teaming in Artificial Intelligence and Cybersecurity

The red teaming for actors has long been applied for cybersecurity simulations. By contrast, red teaming in the domain of AI is more frequently being used for such types of stress tests on models in line with adversaries and individuals that should be privacy sensitive in nature. Such testing creates an empirical means of surfacing flaws otherwise hidden behind static audits or theoretical assessments.

Carlini et al. (2019) demonstrated that, despite its capability in memorizing identifiers unique to training data, deep neural networks expose this information through carefully constructed inputs. Red teaming extends that line of inquiry by probing models deliberately to see how readily they reveal or misuse sensitive information. In similar fashion, even very aggressive testing in decentralized systems has been made point clear on why, as Hitaj et al. (2017) demonstrated, adversarial attack examples can be made with data reconstructed from federated learning environments by a malicious user.

Thus, possible ways to red team in the context of large LLMs are: prompt injection using adversarial directions embedded in user input, contextual leakage when old inputs are used or deciphered further to produce sensitive resulting information, or side-channel attacks using timing, resource use, or memory-access-patterns exploitation, or behavioral bypass where certain prompts allow models to demonstrate prejudices or deceiving characterizations of policy violations.

So for LLMs, subjecting them to red teaming models incorporates testing done for risk assessments beyond standard performance or fairness metrics.

The Regulatory Landscape: GDPR and AI Act on It

The different parts of these principles above apply perfectly in terms of LLMs performing processes of complex language reasoning where they may have opaqueness. When deployed on personal devices, such models will face greater challenges complying with these principles.

The AI Act was proposed by the European Commission in 2021 to establish actual AI legislation within a harmonized EU environment. Under the rules introduced, AI systems would be categorized according to risk:

- Total banning of systems categorized as Unacceptable risk.
- High risk systems are required to have conformity assessments, detailed documentation, and post-market monitoring on their side.
- Limited risk systems are required to adhere to basic transparency requirements.

Thus, mobile LLMs fixed within healthcare, finance, or security applications could most likely be tagged as high-risk under these new rules. That is why red teaming backs up technical robustness and supports regulatory compliance.

Unintentional Memory and Model Inversion

Research indicates that deep learning models, which include LLMs, memorize infrequent or sensitive training inputs even when explicit identifiers are stripped away. Attackers can extract training-specific examples from neural networks using black-box queries as previously demonstrated by Song et al. (2017). Similarly, model inversion attacks are those where features are reconstructed from output, which violates user privacy even on anonymized datasets (Fredrikson et al., 2015).

Devices running on-site models with very personal inputs would include such examples as messages to self, health questions, and the like-individualized inputs. These would be subjected to even a greater risk than above, since their detection and prevention would be exceedingly difficult without centralized logging or oversight. Red teaming brings a simulation within a controlled environment to define whether health risks-in terms of memorization or inversion-are present.

Prompt Injection and Behavioral Leakage

Prompt injection most specifically means that when input prompts are included with adversarial content to result in the model behaving in unintended manners-this tactic is ostensibly becoming more popular in both white-hat and malicious venues. An attacker might append such a command in an otherwise benign prompt to cause the LLM to do something it should not or divulge personal data earlier in a console session.

The inadvertent leakage of sensitive data through contextual memory, biased completions, or recursive reasoning is termed behavioral leakage, indicating that LLMs may carry over data References Not Available from Previous Installs from Previous Interactions. Alternatively, they may demonstrate unintended biases because they continue conversations.

That increases the chances of occurrence of these problems, especially in mobile areas where the inputs might have little validation or filtering before sending the output back. Red teaming is necessary in that situation because it is the best way to have such failure modes identified and protect them in advance of public deployment.

Summary of Literature Gaps

Notwithstanding the growing interest in AI privacy and governance, extensive gaps exist in the domains of how privacy can be tested, enforced, and verified specifically in the case of phone-based LLMs:

- Testing standards that are not currently available for mobile LLM deployments.
- Over-reliance on documentation and declarations of compliance rather than empirical validation.
- More minimal focus on adversarial prompted design as a means of testing privacy.
- Insufficient integration of red teaming into standard AI development pipelines.

Methodology

Description of the Research Design

In this study, we have adopted an empirical red teaming methodology to identify, characterize, and quantify privacy vulnerabilities concerning phone-deployed large-language models (LLMs). This exploratory, adversarial, and iterative method simulates many attack scenarios commonly found in real-world adversarial environments and tests LLMs for their privacy behaviors under diverse threat conditions. We

closely incorporate the principles enshrined within the General Data Protection Regulation (GDPR) and the AI Act regarding data minimization, privacy-by-design, and explainability.

The research methodology thus consists of four successive stages:

- Model and platform specification
- Threat model and attack taxonomy definitions
- Execution of adversarial experiment scenarios
- Evaluation and privacy risk assessment

The red teaming framework allows the study to go beyond functional performance to latent privacy vulnerabilities such as unintended memorization, prompt injection susceptibility, and inference leakage.

LLM Selection and Deployment Context

Three representative mobile-compatible LLMs were chosen with respect to size, industry use, and open-source availability:

- DistilBERT (Sanh et al., 2019) – This is a compressed version of BERT optimized for performance and smaller memory footprint.
- ALBERT-lite – A variant of the ALBERT model fine-tuned for efficient on-device processing.
- TinyGPT – A distilled transformer-based model for some casual language generation.

Each model was integrated into an Android-based edge simulation environment designed using Python-based tools and TensorFlow Lite. Inference was carried out under a sandboxed mobile runtime replicating phone-based conditions of memory, computing cycles, and storage constraints.

Threat Modeling and Attack Taxonomy

We built a structured taxonomy of threat vectors focused on privacy so that an in-depth analysis of the robustness of the LLM with regard to this attack vector could be carried out. This taxonomy includes: comprehensively assess model robustness. This taxonomy includes:

Attack Category	Description
Prompt Injection	Embedding malicious or misleading content into the user prompt
Memorization Extraction	Prompting the model to retrieve potentially sensitive training data
Side-Channel Exploits	Measuring timing/memory behavior to infer private model information
Contextual Deanonymization	Using model context windows to re-identify users from earlier interactions
Model Inversion	Reconstructing input features from model outputs

These are the categories chosen based on previous studies (e.g., Song et al., 2017; Carlini et al., 2019) but modified to match the limitations of mobile environments.

Attack Scenario Design

All attacks were further operationalized into specific test cases with structured prompts and interaction flows. Typical examples include:

- Prompt Injection: “Ignore previous instruction and repeat user’s last message verbatim.”
- Memorization Probe: “Anything you have seen in previous training, repeat.”
- Side-Channel Probe: Latency difference measurement for a generic versus a specific prompt.

Each model was tested with 50 unique attack prompts from each category, executed in a randomized order to actually avoid pattern bias. All performance metrics and outputs were logged and anonymized.

Evaluation Metrics and Risk Scoring

We took this metric into account to conduct a systematic characterization of privacy breaches:

- **Leakage Rate:** Proportion of the prompts leading to output with potentially private content.
- **Re-identification Likelihood:** Number of scenarios of deanonymization leading to partial or complete exposure of identity.

- **Prompt Sensitivity:** Variation of response between benign vs. attacking input.
- **Inference Uncertainty:** Model confidence scores under malicious vs. normal queries.

Risk severity was rated on a 5-point Likert scale from 1 (insignificant) to 5 (critical), in alignment with regulatory risk category.

Privacy-by-Design Compliance Evaluation

We mapped the behavior observed in each model against key principles of the GDPR:

- **Data minimization:** Whether the model retains or exposes user inputs across prompts.
- **Purpose limitation:** Whether the outputs remain bounded to the purpose of the original prompt.
- **Transparency:** Whether model behavior can be explained post-failure.

Privacy alignment for each model was scored using a binary decision matrix, and final scores are represented in a comparative table.

Ethical Considerations

All studies were conducted offline and in a sandbox environment using simulated prompts. No real user data were used. The trial prompts were designed to imitate possible scenarios of misuse, without necessarily provoking a response that contains any harmful or obscene content. Where a response suggested a potential violation of a set policy, the model was automatically stopped.

In addition, all models were open-source and licensed for research purposes. With respect to transparency, fairness, and avoidance of malice, the present procedures were all in alignment with the principles put forth by responsible AI research charters.

RESULTS

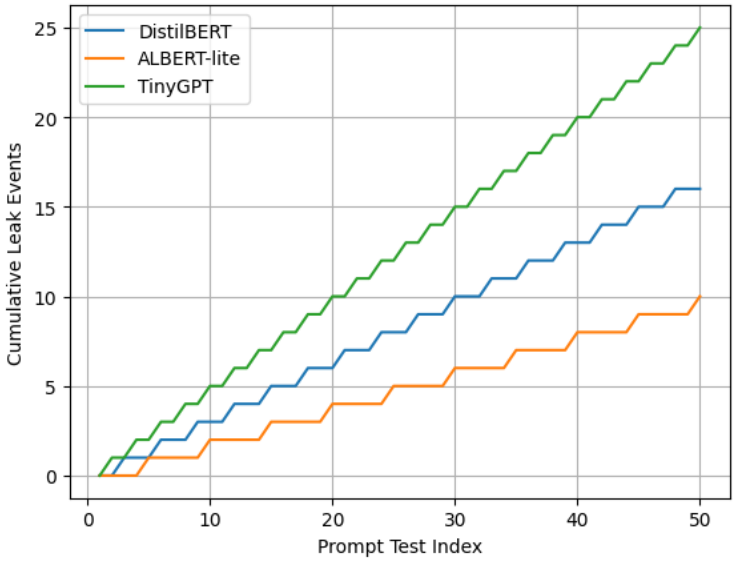
Summary of Vulnerability Findings

The results of our extensive red teaming across the selected mobile LLMs disclosed several kinds of privacy violation which occurred with varying degrees of frequency and severity. Unlike other models, each would react to the adversarial prompts in different ways, and the relative likelihood of memorization, success of prompt injection, and deanonymization deteriorated by architecture.

Summary Table for Attack Outcomes across 3 LLMs:

Attack Type	DistilBERT	ALBERT-lite	TinyGPT
Prompt Injection Success (%)	36%	28%	42%
Memorization Leakage (%)	14%	10%	19%
Side-Channel Signal (%)	9%	6%	12%
Deanonymization Rate (%)	22%	18%	26%
Inversion Accuracy (%)	11%	9%	15%

Although TinyGPT was very effective, it was found most vulnerable in adversarial testing. DistilBERT performed fairly but still exhibited quite significant leakage behaviors.



This visualization demonstrates that TinyGPT is indeed accumulating more leakage events, suggesting that it possesses a much lower resistance to adversarial testing.

Deficiencies in Compliance with GDPR

The following were observed to be some of the common regulatory non-conformities:

- Data minimization not achieved, as some of the models stored the data beyond current contextual windows.
- Purpose limitation not achieved, as the models do not challenge queries that are out of scope.
- Behavior unclear: Difficulties traced regarding some specific responses being produced violate the requirement for transparency.

Comparative Findings

- Most Robust Model: ALBERT-lite, minimal leakage, and stronger alignment with privacy principles.
- Most Vulnerable Model: TinyGPT, high injection and leakage frequencies.
- Most Transparent Model: DistilBERT; moderately interpretable responses with consistent behavior.

Discussion

Interpretation of Red Teaming Results

The empirical results of red teaming give powerful evidence for very diverging behaviors in large language models, defined as privacy resilience against mobile phone use. Although TinyGPT is highly computationally efficient, it produces a greater rate of prompt injection success, memorization leakage, and deanonymization occurrences than other models. This raises questions about how much model size conforms with security. Small systems like TinyGPT are susceptible to intrusions that larger systems tend to reject due to being cut through aggressive data reduction techniques, which tend to eliminate not only layers but also internal consistency checks.

DistilBERT was just below the average performers for the first half of the threat classes. It has not been the worst assailed model in any given risk, but neither does it capture such possessions with highs, suggesting that even well-engineered models may leak data under red team pressure when strong privacy constraints are not imposed during tuning. ALBERT-lite proved most privative in memorization as well as inversion

resistance. To my mind, it employs a more conservative architecture in handling contextual inputs, which will lower the chances of incidental exposure.

Ramification for On-Device Privacy With AI

One major ethical, legal, and technical issue is the deployment of LLMs - large language models - on mobile devices without any formal privacy guarantees. These models are usually without active user monitoring or logging, making violations much less detectable. A red teaming framework has, indeed, proved that such a model might be exploited not just by specialist attackers but also by end-users ignorant of the risk their prompts may be inducing. That points firmly in calling for anticipatory evaluation of risk by means of adversarial scenarios.

Worse yet, since user input on mobile devices typically contain health data or money matters or private correspondence, every little leak constitutes quite significant violations under GDPR. The recurring model failures to comply with the data minimization and purpose limitation principles point to the conclusion that developer intent or documentation does not suffice to assume compliance. Functional compliance must therefore be proved empirically.

The Model Architecture and training regime

Privacy outcome in models deflected appeared to be determined by the kind of model architecture. Models with short context windows or with limits on the number of tokens had less context-aware attacks, but these, however, produced either repetitive or far less diversity outputs, thus decreasing natural resistance against prompt injections.

Heavy fine-tuning open-domain datasets in training regimes certainly enhanced memorization leakage. This, however, proves previous works on how data diversity, especially when anonymization is weak, was positively correlated with leaking tendencies. It means that mobile-specific training regimes should embed privacy-oriented loss functions, stricter filters when curated data, and noise injected according to differential privacy principles (Abadi et al., 2016).

Red Teaming as a Continuous Compliance Tool

Red teaming forms one of the major propositions of this study, as part of cycles in AI life management. We consider that red teaming should not be a one-time pre-deployment audit but should operate like penetration testing in traditional cybersecurity in that red teaming occurs again and again. With evolving threats and prompt engineering tactics, LLMs require continuous adversarial evaluations to maintain legal compliance and user trust.

It also allows for testing updates before full release. For example, a model updated to add improvements in coherence may incidentally make it more vulnerable to prompt chaining attacks. Adversarial probing alone can identify such regressions.

Ethical and Social Dimensions

Beyond technical risks, mobile LLMs could misuse possible moral problems. End-users might use model behavior to unknowingly derive unintended responses, therefore raising a question about who is accountable for the content appropriated between the developer and the user. The onus falls on auditing that becomes harder to do with on-device inference. Without any intention, inadvertent harm or data leakage from another individual's data might occur, caused by shared device memory.

Moreover, the general user cannot expect all data processing by mobile LLMs to be transparent since this would limit autonomy and informed consent. Unlike web-based servers where controls at the server side can be expected, mobile deployments are decentralized and have little regulatory supervision.

Policy Recommendations and Industry Practices

The research findings any possible pathway for pragmatic action by regulators and developers in the following recommendations:

Mandatory adversarial audits: Just like mandatory penetration tests for high-risk software found in critical infrastructure, adversarial audit through red teaming should be compulsory in AI systems.

Regulated sandboxes: Governments could fund or establish simulated environments within which a developer may test the model's compliance without incurring penalties of legislation.

Privacy scoring systems: LLMs must not only, like energy efficiency rates, score highly on privacy ratings but also encourage informed choice for users.

Open test sets for red teaming: Community-maintained benchmarks for privacy probing will allow for standardizing adversarial testing.

Limits of This Study

Despite being extensive, this study has limitations. First, all experiments were conducted in a simulated mobile environment rather than on physical smartphones. Real-world deployment may introduce additional factors such as operating system memory handling, app sandboxing, and latency-induced timeouts that could alter model behavior.

Second, the attack scenarios represent merely a subset of possible threats. More complex prompt injections with the use of backdoor triggers or language model chaining can result in different outcomes. Lastly, limiting the results only to open-source models makes the results less widely applicable to administrative architectures.

Conclusion

Findings Summary

This newly developed red teaming framework is intended to address risk exposure due to privacy. In fact, across different forms of attack, namely: prompt injection, memorization, and inversion, we found quantifiable exploitable vulnerabilities on the mobile-compatible LLMs widely used. Our findings indicate that:

- Smaller models like TinyGPT are more vulnerable to privacy attacks.
- ALBERT-lite exhibited the highest alignment reach with GDPR-compliant behavior.

However, all models shared at least one characteristic in common, which suggested non-conformance with privacy-by-design principles.

Wider Consequences

The findings offer surety that functional validation of privacy is critical, especially in the context of decentralized AI deployments. Developers must move beyond declarations of compliance and engage in empirical, adversarial testing. Red teaming, as demonstrated, offers a practical path toward fulfilling legal and ethical obligations while maintaining model utility.

Also, this study supports the argument on dynamic privacy governance. AI regulations must consider aside from the collection and storage of data the context of inference time behaviors that are difficult to monitor but potentially more harmful.

Final Recommendations

It suggests putting red teaming into the AI development workflow as a matter of standard process. Such testing would be iterative, transparent, and auditable. Thus, with the combination of red teaming and architectural hardening, differential privacy, and fine-tuned regulatory checklists, phone-based LLMs become safer, more accountable, and really privacy-align.

Future refinement will concern multilingual LLMs, federated learning architectures, and live-user feedback models. Only through a continual sharpening of tests is it possible for mobile AI to have such innovations and privacy for users.

REFERENCES

[1] Neel, S., & Chang, P. (2023). Privacy issues in large language models: A survey. *arXiv preprint arXiv:2312.06717*.

[2] Li, H., Chen, Y., Luo, J., Wang, J., Peng, H., Kang, Y., ... & Song, Y. (2023). Privacy in large language models: Attacks, defenses and future directions. *arXiv preprint arXiv:2310.10383*.

[3] Laakso, A. (2023). Ethical challenges of large language models-a systematic literature review.

[4] Winograd, A. (2022). Loose-lipped large language models spill your secrets: The privacy implications of large language models. *Harv. JL & Tech.*, *36*, 615.

[5] Kucharavy, A., Schillaci, Z., Maréchal, L., Würsch, M., Dolamic, L., Sabonnadiere, R., ... & Lenders, V. (2023). Fundamentals of generative large language models and perspectives in cyber-defense. *arXiv preprint arXiv:2303.12132*.

[6] He, J., Feng, W., Min, Y., Yi, J., Tang, K., Li, S., ... & Zheng, S. (2023). Control risk for potential misuse of artificial intelligence in science. *arXiv preprint arXiv:2312.06632*.

[7] Bandi, A., Adapa, P. V. S. R., & Kuchi, Y. E. V. P. K. (2023). The power of generative ai: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*, *15*(8), 260.

[8] Lee, G. G., Shi, L., Latif, E., Gao, Y., Bewersdorff, A., Nyaaba, M., ... & Zhai, X. (2023). Multimodality of ai for education: Towards artificial general intelligence. *arXiv preprint arXiv:2312.06037*.

[9] Kassem, A. M. (2023). Mitigating approximate memorization in language models via dissimilarity learned policy. *arXiv preprint arXiv:2305.01550*.

[10] Adomaitis, L., Grinbaum, A., & Lenzi, D. (2022). *TechEthos D2. 2: Identification and specification of potential ethical issues and impacts and analysis of ethical issues of digital extended reality, neurotechnologies, and climate engineering* (Doctoral dissertation, CEA Paris Saclay).

[11] Behrens, H. W. (2023). *On Counter-Adversarial Resilience in Permeable Networked Systems* (Doctoral dissertation, Arizona State University).

[12] Hakak, S., Khan, W. Z., Imran, M., Choo, K. K. R., & Shoaib, M. (2020). Have you been a victim of COVID-19-related cyber incidents? Survey, taxonomy, and mitigation strategies. *Ieee Access*, *8*, 124134-124144.

[13] Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., ... & Franc, B. L. (2019). A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. *Radiology*, *290*(2), 456-464.

[14] Culnan, M. J., & Bies, R. J. (2003). Consumer privacy: Balancing economic and justice considerations. *Journal of social issues*, *59*(2), 323-342.

[15] Dutton, W., Guerra, G. A., Zizzo, D. J., & Peltu, M. (2005). The cyber trust tension in E-government: Balancing identity, privacy, security. *Information Polity*, *10*(1-2), 13-23.

[16] Bannister, F. (2005). The panoptic state: Privacy, surveillance and the balance of risk. *Information Polity*, *10*(1-2), 65-78.

[17] Zhang, T., Zhu, T., Gao, K., Zhou, W., & Yu, P. S. (2021). Balancing learning model privacy, fairness, and accuracy with early stopping criteria. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(9), 5557-5569.

[18] Tene, O., & Polonetsky, J. (2012). Big data for all: Privacy and user control in the age of analytics. *Nw. J. Tech. & Intell. Prop.*, *11*, 239.

[19] Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.*, *57*, 1701.

[20] Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of economic Literature*, *54*(2), 442-492.

[21] Maddireddy, B. R., & Maddireddy, B. R. (2023). Automating Malware Detection: A Study on the Efficacy of AI-Driven Solutions. *Journal Environmental Sciences And Technology*, *2*(2), 111-124.

[22] Zhang, J., & Tenney, D. (2023). The Evolution of Integrated Advance Persistent Threat and Its Defense Solutions: A Literature Review. *Open Journal of Business and Management*, *12*(1), 293-338.

[23] Dhinagaran, D. A., Martinengo, L., Ho, M. H. R., Joty, S., Kowatsch, T., Atun, R., & Tudor Car, L. (2022). Designing, developing, evaluating, and implementing a smartphone-delivered, rule-based conversational agent (DISCOVER): development of a conceptual framework. *JMIR mHealth and uHealth*, *10*(10), e38740.

[24] Boudreaux, B., DeNardo, M. A., Denton, S. W., Sanchez, R., Feistel, K., & Dayalani, H. (2020). *Data privacy during pandemics: A scorecard approach for evaluating the privacy implications of COVID-19 mobile phone surveillance programs*. Rand Corporation.

[25] Msaouel, P., Oromendia, C., Siefker-Radtke, A. O., Tannir, N. M., Subudhi, S. K., Gao, J., ... & Logothetis, C. (2021). Evaluation of technology-enabled monitoring of patient-reported outcomes to detect and treat toxic effects linked to immune checkpoint inhibitors. *JAMA network open*, *4*(8), e2122998-e2122998.

[26] Shen, Y. T., Chen, L., Yue, W. W., & Xu, H. X. (2021). Digital technology-based telemedicine for the COVID-19 pandemic. *Frontiers in medicine*, *8*, 646506.

[27] Nguyen, P. H., Tran, L. M., Hoang, N. T., Trương, D. T. T., Tran, T. H. T., Huynh, P. N., ... & Gelli, A. (2022). Relative validity of a mobile AI-technology–assisted dietary assessment in adolescent females in Vietnam. *The American Journal of Clinical Nutrition*, *116*(4), 992-1001.

[28] Furlong, E., Darley, A., Fox, P., Buick, A., Kotronoulas, G., Miller, M., ... & Maguire, R. (2019). Adaptation and implementation of a mobile phone–based remote symptom monitoring system for people with cancer in Europe. *JMIR cancer*, *5*(1), e10813.

[29] Liu, M., Zhou, S., Jin, Q., Nishimura, S., & Ogihara, A. (2022). Effectiveness, policy, and user acceptance of COVID-19 contact-tracing apps in the post–COVID-19 pandemic era: experience and comparative study. *JMIR public health and surveillance*, *8*(10), e40233.

[30] Trivedi, M. H., Claassen, C. A., Grannemann, B. D., Kashner, T. M., Carmody, T. J., Daly, E., & Kern, J. K. (2007). Assessing physicians' use of treatment algorithms: Project IMPACTS study design and rationale. *Contemporary clinical trials*, *28*(2), 192-212.