Demo: ViolentUTF as An Accessible Platform for Generative AI Red Teaming

Tam n. Nguyen
GSA - FedRAMP
tom.nguyen@ieee.org

***Abstract*—The rapid integration of Generative AI (GenAI) into various applications necessitates robust risk management strategies which includes Red Teaming (RT) - an evaluation method for simulating adversarial attacks. Unfortunately, RT for GenAI is often hindered by technical complexity, lack of user-friendly interfaces, and inadequate reporting features. This paper introduces Violent UTF - an accessible, modular, and scalable platform for GenAI red teaming. Through intuitive interfaces (Web GUI, CLI, API, MCP) powered by LLMs and for LLMs, Violent UTF aims to empower non-technical domain experts and students alongside technical experts, facilitate comprehensive security evaluation by unifying capabilities from RT frameworks like Microsoft PyRIT, Nvidia Garak and its own specialized evaluators. ViolentUTF is being used for evaluating the robustness of a flagship LLM-based product in a large US Government department. It also demonstrates effectiveness in evaluating LLMs' cross-domain reasoning capability between cybersecurity and behavioral psychology.**

***Index Terms*—Generative AI, Red Teaming, AI Safety, Cybersecurity, LLM Evaluation, Human-Centric Security, Responsible AI, Violent UTF**

I. Introduction

Generative Artificial Intelligence (GenAI) offers transformative potential across numerous domains, but its deployment is accompanied by significant risks [1]. These models can inadvertently generate harmful, biased, or inappropriate content, leak sensitive information, or be manipulated through adversarial attacks like prompt injection and jailbreaking [2]. Ensuring the safe and responsible deployment of GenAI requires rigorous evaluations and tests to proactively identify and mitigate these vulnerabilities [3]. When applied to GenAI, Red Teaming (RT) systematically probes models to elicit undesirable behaviors, assess their robustness against manipulation, and verify compliance with safety and ethical guidelines [4], [5]. However, the current landscape of tools for GenAI RT presents significant barriers [6], [7]. Table I provides a brief overview of some tool categories with their pros and cons. [8]:

Despite the availability of these tools, many demand deep technical expertise and programming skills, excluding non-technical domain experts and stakeholders who possess valuable insights into potential real-world harms [1]. Furthermore, the lack of intuitive interfaces and comprehensive reporting features limits collaborative efforts and hinders effective decision-making based on testing outcomes. **To bridge this gap, there is a clear need for a red teaming platform that democratizes the process, making it accessible to a wider range of users while providing powerful, flexible, and extensible capabilities.** This paper demonstrates how Violent UTF can meet these needs.

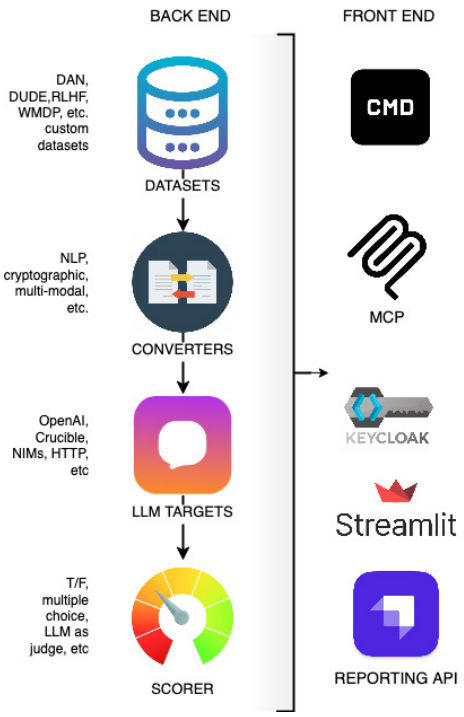

Fig. 1. Violent UTF System Components.

II. Solution: The Violent UTF Platform

Violent UTF aims to overcome the described obstacles in GenAI red teaming. It provides a unified, accessible, and comprehensive platform that integrates frameworks like Microsoft PyRIT and Nvidia Garak and extends features with its own specialized evaluation tools. Its creativity lies in bridging the gap between technical security testing and human-centric risk assessment, making sophisticated red teaming accessible while providing depth for expert users and automation.

A. Key Features Driving Effectiveness

Democratized Accessibility. Violent UTF tackles the accessibility problem head-on. Its primary interface is an intuitive web-based GUI (Streamlit) designed explicitly to empower non-programmers – domain experts, ethicists, compliance officers – to configure and execute complex red teaming scenarios

TABLE I
EXAMPLES OF AI SAFETY, SECURITY & RED TEAMING TOOLS

Tool/Category	Description	Pros	Cons
Microsoft PyRIT	Open-source Python framework for automated red teaming, orchestrating attacks.	Structured; extensible; integrates datasets; automates workflows.	Primarily code-based; learning curve for complex orchestration.
Nvidia Garak [7]	Open-source framework to evaluate LLM security/robustness using probes, detectors, and harnesses.	Structured evaluation; modular components (probes, detectors); good for vulnerability scanning.	Primarily code-based; configuration can be complex.
Adversarial Robustness Toolbox (ART) [7], [8]	Python library for evaluating/defending ML models against adversarial attacks (evasion, poisoning).	Wide range of attacks/defenses; supports multiple frameworks; research-oriented.	Requires ML expertise; focus broader than just LLMs; can be complex to configure.
NB Defense [7]	JupyterLab extension/CLI for AI vulnerability management (secrets, PII, CVEs).	Integrates directly into developer workflow; contextual guidance.	Focused on notebook/code scanning; specific to Jupyter environment.
LLM Input/Output Filters	Tools designed to moderate prompts sent to LLMs or filter responses from LLMs.	Can block known malicious prompts/harmful outputs; real-time protection.	Can be bypassed by novel attacks (jailbreaks); may introduce latency; risk of false positives/negatives.
Llama Guard / Purple Llama [8]	Meta's tools for input/output moderation, insecure code detection.	Open-source; specifically designed for LLM safety; includes benchmarks.	Effectiveness depends on model updates and attack evolution.
Vigil / Rebuff [8]	Python libraries/APIs focused on detecting prompt injections, jailbreaks.	Target specific attack vectors; some offer multiple detection layers (heuristics, LLM analysis).	Narrow in scope; effectiveness against novel attacks varies.

without writing code. Simultaneously, a consistent Command Line Interface (CLI) and a comprehensive RESTful API (FastAPI/Kong) provide powerful options for technical users, automation scripts, and integration into MLOps pipelines.

Unified & Extensible Framework. A core innovation is the unification of diverse red teaming and evaluation methodologies within a single platform. Violent UTF integrates:

- Technical Red Teaming (PyRIT & Garak) for identifying technical vulnerabilities like jailbreaks, prompt injection, and harmful content generation.
- Human-Centric Evaluation (Ollabench) [9] to assess LLM reasoning in the context of human-centric interdependent cybersecurity [11]. Ollabench is a custom module that allows Violent UTF users to evaluate risks related to the LLM's grasp of security policies in realistic contexts – a critical aspect often overlooked by purely technical tools [10], [11].

Violent UTF standardizes core concepts (Generators, Prompts, Converters, Evaluators, Orchestrators, Memory) across these integrated tools, simplifying complex testing campaigns. Prompt templating further enhances flexibility for dynamic prompt generation and dataset generation.

Secure, Scalable & Maintainable Architecture. Violent UTF is underpinned by a robust architecture designed for modern security and operational demands:

- *Presentation Layer:* Streamlit GUI and Python CLI.
- *Authentication & Authorization:* Centralized IAM via Keycloak and Kong Gateway, enforcing OIDC/OAuth2 and RBAC consistently across all interfaces. This secure foundation is crucial for managing user access and enabling trusted agentic operations.
- *Unified API Layer:* FastAPI exposes all functionalities via a versioned, documented (OpenAPI) RESTful API, enabling seamless integration and automation. Kong Gateway acts as the policy enforcement point for security, routing, and rate limiting.
- *Logging & Observability:* A dedicated layer ensures comprehensive, structured logging with clear levels and guidelines, supporting debugging, security monitoring, and compliance.

III. DEMO - EVALUATING CROSS-DOMAIN REASONING IN CYBERSECURITY

Violent UTF is being used as a RT tool for evaluating a US Government's flagship LLM-based application. Besides the "main stream" topics of RT such as jailbreaking and prompt injection, the paper wants to demo Violent UTF capabilities of evaluating whether Large Language Models (LLMs) can reason effectively and safely across domains.

A critical challenge in cybersecurity is understanding and predicting human behavior regarding compliance with security policies, often influenced by complex psychological factors [11]. Solving this challenge is crucial for applications like advanced threat modeling (especially insider threats), designing targeted security awareness training, or building realistic agent-based simulations of socio-technical systems. In this use case, the Ollabench component within Violent UTF was configured to evaluate 21 different LLMs (including commer-

cial and open-weight models) on their ability to reason about information security compliance/non-compliance behaviors.

A. The main steps

1) **Scenario Generation:** Using Violent UTF's interface powered by OpenAI LLMs, scenarios depicting hypothetical employees with specific cognitive profiles (compliant or non-compliant attributes derived from 24 behavioral theories and 38 peer-reviewed papers) were generated.
2) **LLM Interaction:** The selected LLMs (configured as 'Generators' in Violent UTF) were presented with these scenarios and asked a series of fixed multiple-choice questions designed to test different facets of reasoning: identifying cognitive constructs, comparing compliance levels, predicting team risk dynamics, and identifying key factors for intervention.
3) **Evaluation:** Responses were evaluated using Ollabench's metrics within the Violent UTF framework, focusing on:
    - *Accuracy:* Overall correctness and categorical accuracy across the different question types [9].
    - *Wastefulness:* Token efficiency, specifically measuring tokens used for incorrect answers [9].
    - *Consistency:* Analyzing the reliability of reasoning patterns using Structural Equation Modeling [9].

B. Key Findings from the Use Case Evaluation

The results obtained via Violent UTF highlight the platform's ability to uncover critical insights into LLM cross-domain reasoning:

- **Significant Reasoning Gaps:** Even leading commercial models (Gemini 1.5 Flash, GPT-4o, Claude 3 Opus) achieved only around 51% overall accuracy on these complex cross-domain tasks, underscoring the current limitations of LLMs in reliably understanding the interplay between human psychology and cybersecurity behaviors [9]. This demonstrates Violent UTF's effectiveness in identifying capability gaps where simpler evaluations might not.
- **Nuanced Performance Differences:** The platform revealed stark differences in how models performed on specific reasoning tasks. For instance, models showed varied success rates in identifying cognitive paths versus predicting team risks or target factors for intervention [9]. This granular analysis is crucial for selecting models suited for specific human-centric applications.
- **Efficiency Variance:** Significant differences in 'wastefulness' were observed, with some models consuming considerably more tokens when providing incorrect answers [9]. This capability within Violent UTF provides practical data for optimizing cost and resource allocation when deploying LLMs evaluated for such tasks.
- **Consistency Correlation:** The analysis indicated that more accurate LLMs generally exhibited more consistent reasoning patterns, suggesting Violent UTF can help assess the reliability alongside the correctness of an LLM's reasoning in this complex domain [9].

IV. Conclusion and Future Work

Violent UTF represents a significant step towards making Generative AI red teaming more accessible, comprehensive, and effective. By prioritizing user experience for both technical and non-technical users, unifying powerful testing frameworks (including PyRIT, Garak, and Ollabench), and building on a robust, secure, and scalable architecture, it addresses critical gaps in the current tooling landscape.

The provided use case demonstrates the practical applicability and efficacy of Violent UTF for organizations needing to evaluate LLMs for tasks requiring sophisticated, cross-domain reasoning. By integrating custom tools, the platform moves beyond standard security testing to provide theory-grounded, cross-domain testing in cybersecurity [14], [15].

Future work will focus on expanding the library of integrated components (Generators, Prompts, Converters, Evaluators) by wrapping additional third-party libraries and developing novel techniques. Enhancing the reporting and visualization capabilities within the GUI is a key priority, providing users with more interactive ways to analyze results and generate actionable insights. Further development of the agentic capabilities will enable more sophisticated automated red teaming scenarios, potentially involving AI agents dynamically adapting attack strategies based on model responses. Continued adherence to API-first development, including contract testing and versioning, will ensure maintainability and ease of integration as the platform evolves towards potential microservice patterns. By fostering collaboration and lowering barriers to rigorous testing, Violent UTF aims to contribute significantly to the development and deployment of safer, more trustworthy Generative AI systems.

Appendix

Here is the supplementary material:

- YouTube video demonstration of ViolentUTF: https://youtu.be/c-UCYXq0rfY
- Source code on GitHub: https://github.com/Cybonto/ViolentUTF_nightly

Please contact me for access to the private GitHub repository.

References

[1] Governing generative AI: Navigating opportunities and risks. *Policy & Society*, vol. 44, no. 1, pp. 1–20, 2025. Available: https://academic.oup.com/policyandsociety/article/44/1/1/7997395

[2] International AI Safety Report 2025. GOV.UK. Available: https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf

[3] L. Weidinger et al., "Ethical and social risks of harm from Language Models," *arXiv preprint arXiv:2112.04359*, 2021.

[4] E. Perez et al., "Red Teaming Language Models with Language Models," *arXiv preprint arXiv:2202.03286*, 2022.

[5] M. Mazeika et al., "HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal," *arXiv preprint arXiv:2402.04249*, 2024.

[6] Brookings AI Equity Lab, “A new writing series: Re-envisioning AI safety through global majority perspectives,” Brookings Institution, Feb. 2025.
[7] “AI Security Tools: The Open-Source Toolkit,” Wiz Blog, Feb. 16, 2024. Available: https://www.wiz.io/academy/ai-security-tools
[8] “11 LLM Security Tools,” Granica AI Blog, Oct. 2, 2024. Available: https://granica.ai/blog/llm-security-tools-grc
[9] T. N. Nguyen, “Ollabench: Evaluating LLMs’ Reasoning for Human-centric Interdependent Cybersecurity,” *arXiv preprint arXiv:2406.06863*, 2024.
[10] C. Brumfield, ”Why today’s cybersecurity threats are more dangerous,” *CSO Online*, 2021. [Online]. Available: https://www.csoonline.com/article/3635097/why-today-s-cybersecurity-threats-are-more-dangerous.html (Referenced as [4] in Ollabench paper)
[11] M. Kianpour, S. J. Kowalski, and H. Øverby, ”Systematically Understanding Cybersecurity Economics: A Survey,” *Sustainability*, vol. 13, no. 24, p. 13677, 2021. (Referenced as [6] in Ollabench paper)
[12] K. Witte, ”Putting the fear back into fear appeals: The extended parallel process model,” *Communication Monographs*, vol. 59, no. 4, pp. 329-349, 1992. (Referenced as [51] in Ollabench paper)
[13] R. E. Petty and J. T. Cacioppo, ”The elaboration likelihood model of persuasion,” *Advances in Experimental Social Psychology*, vol. 19, no. C, pp. 123-205, 1986. (Referenced as [50] in Ollabench paper - example theory)
[14] M. Q. Patton, ”Evaluation Science,” *American Journal of Evaluation*, vol. 39, no. 2, pp. 183-200, 2018. (Referenced as [28] in Ollabench paper - relevant to evaluation grounding)
[15] F. Heylighen, ”Objective, subjective and intersubjective selectors of knowledge,” *Evolution, Order and Complexity*, F. Heylighen, Ed. Dordrecht: Kluwer Academic Publishers, pp. 63-67, 1997. (Referenced as [48] in Ollabench paper - relevant to knowledge selection for evaluation)