OWASP Top 10 for LLMs - AI Red Teaming & Evaluation Guidelines Initiative

Revision History

superseded by v4

OWASP Top 10 for LLMs - AI Red Teaming Methodologies, Guidelines and Best Pra…

Description	Date	Notes
Initial Version OWASP Top 10 for L…	6.7.24	
V2 OWASP Top 10 for L…	6.10.24	Added more focus (as per comments from Steve, Jason & Scott).
V3	6.14.24	Added more details, mainly from insightful comments from John Sotiropoulos

Abstract

Generative AI Red Teaming (GRT) is still a black art (as of 6.7.24) ! AI Red Teaming/GRT is slightly different from traditional Red Teaming[1] - Evaluations are an integral part of GRT. The methodologies, the tests, the interpretation of the results and the remediation tactics are all still neither well defined nor standardized.As the Seol report[4] says "*Several technical methods (including benchmarking, red-teaming and auditing training data) can help to mitigate risks, though all current methods have limitations, and improvements are required*." Moreover there are no guidelines to map OWASP Top 10 for LLMs between Responsible AI policy primitives (specific to an organization), threat modeling frameworks and vulnerability standards. This initiative aims at making all the above a little easier.

This project aims to answer the "*how*" to our "*what*" question i.e. OWASP Top 10 for LLMs answers the question "*What are the top risks that I should worry about when I deploy my application that has LLM components ?*" *[Sandy and Rachel had asked this*

question in one of our calls] . This initiative answers the "*how*" question in a systematic way - for the AI Red Teaming specifically and LLM Evaluations. It will cover the triad of security (*protecting the operators and adhering to traditional CIA principles*), safety (*ensuring user protection*), and trust (*building user confidence*)

Discussions

Ever since NIST coined the term AI Red Teaming, questions have been raised, given that Red Teaming has a certain connotation in traditional cybersecurity. AI Red Teaming has similarities to traditional but it also adds a few more mechanisms.

AI Red Teaming is a systematic, adversarial approach, employed by human testers, to identify issues/problems in systems that have Generative AI components viz.unsafe material, Inaccuracies, Out-of-scope responses & identify risks unknown at the time of development testing, that come to light from live usage/discovery of new vulnerabilities/new benchmarks. I have a GitHub repository collecting papers, metrics, benchmarks et al at https://github.com/xsankar/AI-Red-Teaming. Very initial stages.

The key recommendations from the GRT from the AI Village at DEFCON31[6] is very informative and forms the background for this initiative.

- *Red teaming for policy serves a different purpose from red teaming at companies and should seek to augment, not replace or compete with existing corporate red teaming practices.*
- *Red teaming provides empirical evidence for evaluating standards and requirements, including providing an understanding of what guardrails are doing and not doing effectively. Therefore it serves as both a test of a model as a whole as well as a test of the model safeguards.*

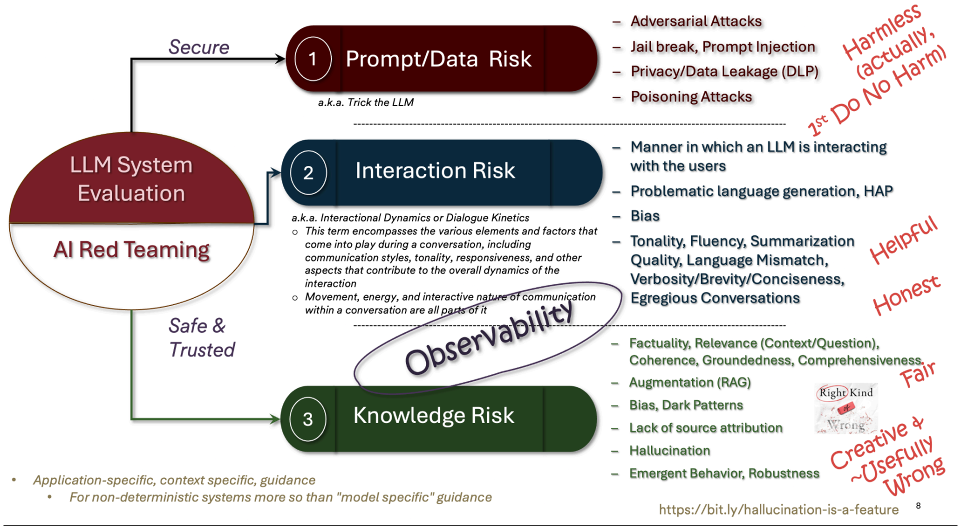
The four main differences between the AI Red Teaming and traditional Red Teaming are:

1. **Conceptual -** Traditional Red Teaming test the systems (that take deterministic inputs) with random inputs; rarely knowledge level testing. But AI Red teaming deals with LLMs that take a broader spectrum of inputs - The "aperture" of what is possible for an LLM is infinitely wide. So, we need to infer the system's knowledge by testing the responses against known knowledge graphs & implicatures. Hence, the datasets = knowledge prompts + plausible contextual responses; AI Red Teaming is predominantly a de-risking activity; not pen testing
2. **Focus** - While traditional Red Teaming aims to identify vulnerabilities in physical security, network security, and information systems, AI Red Teaming has

additional goals viz. the safety (of the user) e.g., bias, toxicity et al as well as trust(by the user) e.g. hallucinations, relevance et al.

3. **Outcome & Remediation** - Traditional Red Teaming outcomes usually lead to recommendations for strengthening defenses; while AI Red Teaming results in recommendations for improving the robustness of AI models (fine tuning, improving RAG performance and so forth), in addition to defenses like guardrails.
    - AI/ML Engineers are a major users of the AI Red Teaming results.
    - LLMsOps can use the results to configure guardrails.
    - Developers can then use the information to retrain/augment the models or develop "guardrail" rules to mitigate risk.
    - In that sense, many times, AI Red Teaming is an assessment - based on the context as well as advancements in the model architectures, training methods, datasets and so forth.
4. **Contextual** - The AI Red Teaming is contextual and use case specific. As *some of the characteristics depend on the context a catch-all guardrail suppressing outputs/hallucination is not a valid solution. For example,while hallucination is not good for deterministic apps, it is an essential component for creative apps like drug discovery, protein folding, marketing, recommendation and so forth (See https://bit.ly/hallucination-is-a-feature and https://bit.ly/gen-ai-org-surgery)*

Extending the above discussion, let us look at the 3 dimensions of AI Red Teaming that we will address in this initiative. Figure 1 shows a summary of the three dimensions along with the system characteristics/expectations.



LLM System Evaluation

AI Red Teaming

Secure

1 Prompt/Data Risk

a.k.a. Trick the LLM

- Adversarial Attacks
- Jail break, Prompt Injection
- Privacy/Data Leakage (DLP)
- Poisoning Attacks

2 Interaction Risk

a.k.a. Interactional Dynamics or Dialogue Kinetics

- *This term encompasses the various elements and factors that come into play during a conversation, including communication styles, tonality, responsiveness, and other aspects that contribute to the overall dynamics of the interaction*
- *Movement, energy, and interactive nature of communication within a conversation are all parts of it*

- Manner in which an LLM is interacting with the users
- Problematic language generation, HAP
- Bias
- Tonality, Fluency, Summarization Quality, Language Mismatch, Verbosity/Brevity/Conciseness, Egregious Conversations

Safe & Trusted

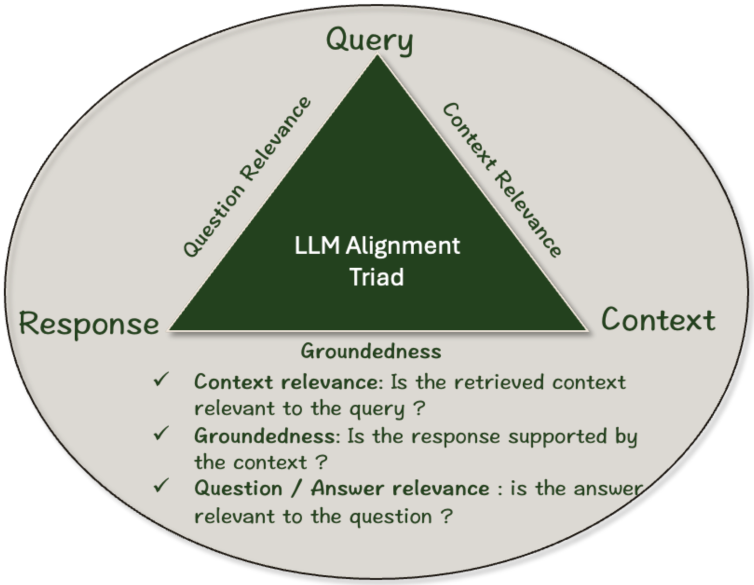
3 Knowledge Risk

- Factuality, Relevance (Context/Question), Coherence, Groundedness, Comprehensiveness,
- Augmentation (RAG)
- Bias, Dark Patterns
- Lack of source attribution
- Hallucination
- Emergent Behavior, Robustness

- *Application-specific, context specific, guidance*
    - *For non-deterministic systems more so than "model specific" guidance*

Observability
Harmless (actually, 1st Do No Harm)
Helpful
Honest
Fair
Right Kind of Wrong
Creative & ~Usefully Wrong
https://bit.ly/hallucination-is-a-feature
8

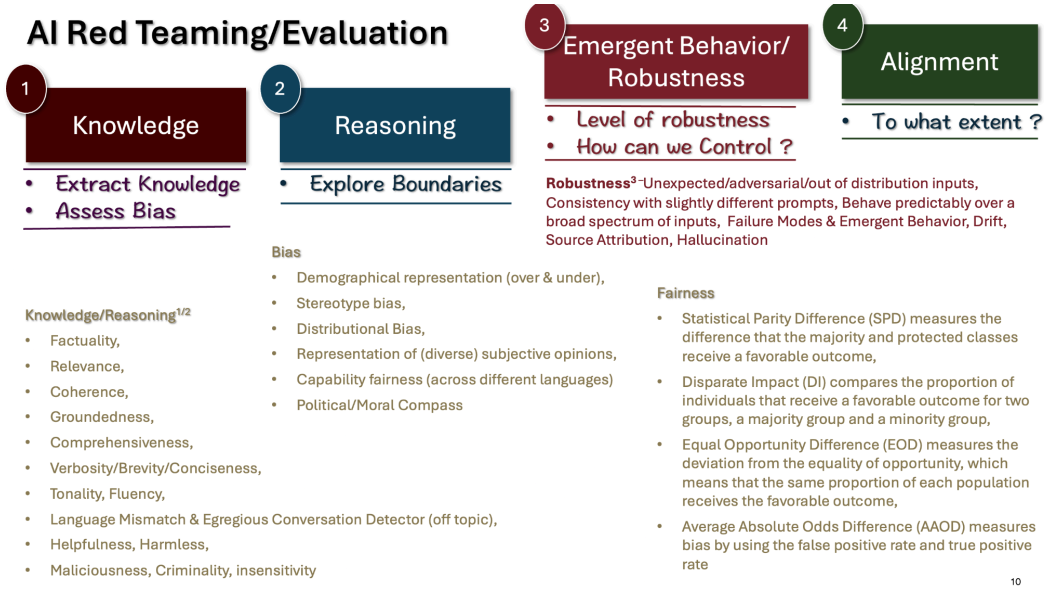
1. Adversarial Attacks/Vulnerability Scanning (security) to assess Prompt/Data Risk
    - As John Sotiropoulos eloquently commented in v2, there is a difference between vulnerability scanning and evaluations - vulnerabilities will emerge regardless of our model improvements (like RAg and finetuning) and they are more into the adversarial attack realm - the security and the CIA triad.
2. Evaluations to access Interaction Risks (Safety of the users) viz. Bias/Toxicity/misinformation
    - Malfunctioning general-purpose AI can also cause harm, for instance through biased decisions with respect to protected characteristics like race, gender, culture, age, and disability-Seol Report[4]
    - Toxic language and misinformation have multiple risks to an organization
3. Evaluations to assess Knowledge Risk (Trust by the users)
    - Red teaming will contribute to evaluations and we need methodologies to make it repeatable and productive. Conversely, Red Teaming will reuse evaluation suites (see UK AISIS Inspect) as part of its baseline assessment but it will go further than evaluating a model and will encompass application semantics. The RAG/LLM Alignment triad is an example where an evaluation methodology adds application semantics to LLM evaluation - because there is the raw LLM plus the retrieval augmentation mechanisms.



- ✓ **Context relevance**: Is the retrieved context relevant to the query ?
- ✓ **Groundedness**: Is the response supported by the context ?
- ✓ **Question / Answer relevance** : is the answer relevant to the question ?

Initiative Goals

1. **Generative AI Red Teaming Methodology, Guidelines & Best Practices**: A canonical methodology and process for Generative AI Red Teaming, including (but not limited to) LLM Red Teaming
2. **Standardised Evaluations to boost trust**: Metrics, Benchmarks, Datasets, Frameworks, Tools and Prompt Banks (as applicable) for LLM evaluation for a "Standardised Evaluations to boost trust" (as John Sotiropoulos puts it). As an example, the following two diagrams list some of the evaluations we will attempt. As far as possible we will leverage publicly available papers, datasets et al plus vendor products. See Github [3] & [7] for an initial list.



AI Red Teaming/Evaluation

1 Knowledge
- Extract Knowledge
- Assess Bias

2 Reasoning
- Explore Boundaries

3 Emergent Behavior/ Robustness
- Level of robustness
- How can we Control ?

4 Alignment
- To what extent ?

Robustness[3]–Unexpected/adversarial/out of distribution inputs, Consistency with slightly different prompts, Behave predictably over a broad spectrum of inputs, Failure Modes & Emergent Behavior, Drift, Source Attribution, Hallucination

Knowledge/Reasoning[1/2]
- Factuality,
- Relevance,
- Coherence,
- Groundedness,
- Comprehensiveness,
- Verbosity/Brevity/Conciseness,
- Tonality, Fluency,
- Language Mismatch & Egregious Conversation Detector (off topic),
- Helpfulness, Harmless,
- Maliciousness, Criminality, insensitivity

Bias
- Demographical representation (over & under),
- Stereotype bias,
- Distributional Bias,
- Representation of (diverse) subjective opinions,
- Capability fairness (across different languages)
- Political/Moral Compass

Fairness
- Statistical Parity Difference (SPD) measures the difference that the majority and protected classes receive a favorable outcome,
- Disparate Impact (DI) compares the proportion of individuals that receive a favorable outcome for two groups, a majority group and a minority group,
- Equal Opportunity Difference (EOD) measures the deviation from the equality of opportunity, which means that the same proportion of each population receives the favorable outcome,
- Average Absolute Odds Difference (AAOD) measures bias by using the false positive rate and true positive rate

10

3. **Artifacts addressing the intended audience for this initiative**

Deliverable	**Audience**
Top-line guide to AI red teaming	Security Architects and CISOs new to the concepts of leveraging red Teaming for AI
Gen AI Red Teaming guidelines, internal governance checklists, regulatory checklists (*.docx*) (*Could be*	Regulators, CISO, Chief AI Officer or CTO and Governance CxO Gen AI security architects

*multiple documents. As Scott cautions, need to watch for scope cree*p)	
Gen AI Red teaming Best Practices (*.docx*)	LLMOps Security Engineers working on Generative AI systems, focusing on security, safety, and trust aspect
Gen AI Red Teaming/LLM evaluation metrics, benchmarks, datasets, frameworks, tools and prompt banks (as applicable) and result interpretations (*.docx, GitHub*)	Developers, LLMOps Security Engineers

Expected Outcomes

- **An AI Red Teaming Methodology that organizations can use for their development, operations, governance and regulatory processes**: A well articulated methodology for AI red Teaming improves the common understanding between the various constituents in the Generative AI ecosystem. The requirement of the details and content varies by the audience and so achieving a contextual common understanding is not easy. Our addition of best practices will definitely help the organizations.
- **Standard set of LLM evaluations**: The LLM evaluation requires broader artifacts spanning ***metrics, benchmarks, datasets, frameworks, tools and prompt banks (as applicable).*** A canonical collection and a toolset gives the practitioners a head start. They can, of course, customize it depending on the use case and organizational policies.
- **Audience-specific, context-specific artifacts**: We will have tailored and customized templates and profiles, thus making this domain (and OWASP Top 10 LLMs) accessible, approachable and more importantly consumable by a wide variety of audiences - folks who have high information overload and low attention span. As Sandy says "*new consumption model where the content is immediately interesting and people can understand what you are writing about quickly*

because let's face it - we're all speed reading - there is just way too much content"!

References:

1. The Role of AI Red Teaming in Cybersecurity https://bit.ly/ai-red-teaming
2. What's the Difference Between Traditional Red-Teaming and AI Red-Teaming? https://www.cranium.ai/traditional_vs_ai_red_teaming/
3. AI Red Teaming Resources https://github.com/xsankar/AI-Red-Teaming
4. https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai
5. https://ukgovernmentbeis.github.io/inspect_ai/eval-suites.html
6. https://drive.google.com/file/d/1JqpbIP6DNomkb32umLoiEPombK2-0Rc-/view
7. https://github.com/xsankar/Awesome-LLM-Eval-MetricMinds

Additional Goals/Future:

These might be pursued either after the initial goals are met or as separate initiatives.

1. **Mapping & Traceability to OWASP Top 10 for LLMs between Responsible AI primitives, threat modeling frameworks and vulnerability standards**: Mapping and guidelines between various frameworks like NIST RMF, MITRE ATLAS (which is a framework based on ATT&CK) and others. Organizations also have their own internal Responsible AI initiatives which will need to be mapped as well. We will have a canonical Responsible AI policy primitives and map them to the frameworks thus making the full mapping a consumable reference for CSIOs and security professionals

   Expected Outcomes : **Understanding of the various standards mapping for secure, safe and trusted Gen AI systems:** By mapping the OWASP Top 10 for LLMs to the broader initiatives like the NISt RMF, NIST AISIC, MITRE ATLAS, Responsible AI et al, the enterprise practitioners will have a comprehensive understanding of the LLM security landscape and can apply to their own organization.

   Note from John Sotiropoulo : *The value is not mapping. Few people except those compiling compliance documentation or vendor reports use mappings. What is missing is a guide of how standards fit together and as per my update on the last call i am working - as your standards alignment lead - on a Standards Compass draft that will release for review*

Addendum - Call for Action:

Sandy has written a compelling call for action for this initiative ! It captures the essence of this body of work.

Invitation to Contribute to AI Red Teaming & Evaluation Initiative

Hello hackers, tech wizards and code sorcerers

We've got an opportunity for you to flex your hacker muscles and dive into the murky waters of Large Language Model (LLM) vulnerabilities. We're putting together a team to map and tackle the OWASP Top Ten vulnerabilities for LLM applications, and we want you on board. Yes, you, the one with the hoodie and the suspiciously fast typing speed.

Why Should You Bother?

Show Off Your Skills: This is your chance to be the star of the show, to let your brilliance shine! Your expertise is needed for an audience of CISOs, risk advisors, developers, and threat analysts focused on defending LLM for everyday people and organizations.

Join the Cool Kids Club: Collaborate, challenge, and charm your way through complex problems with like-minded geniuses. It's like The Avengers, but for nerds.

Make a Real Impact: Your work will help prevent LLM badness happening to everyday people, businesses, and organizations who would like to use LLMs but have no idea about the gotchas. Think of it as superhero work, but without the spandex.

What’s This All About?

Our mission is to crack the "how" of AI Red Teaming and LLM Evaluations. We’re not AI doomers only talking about the problems, but we do want to help people decipher the noise from known AI issues that could cause real harm. That’s where you come in! We need your help quantifying the issues, why they happen, and how they can be prevented.

AI Red Teaming Methodology: Your contributions would be used to create a standardized approach for AI Red Teaming.

Standardized Evaluations: Metrics, benchmarks, datasets, frameworks, tools, and prompt banks – let’s create the gold standard.

Audience-specific, context-specific artifacts: Specialized templates and profiles, which will improve how people understand and do AI Red Teaming using the OWASP Top 10 for LLMs.

How Can You Contribute ?

1. **Find the Exploitable Bits:**Help us uncover and document real-world examples of LLM vulnerabilities.
2. **Share:** Share how you test and the method to your madness.
3. **Create the Tools:** Contribute to developing tools and benchmarks that will be the envy of AI Red Teamers everywhere.

Not all heroes wear capes; some debug code (and *Red Team LLMs*)!

For more details and to get started contact Krishna Sankar/Sandy Dunn at ksankar42@gmail.com/sandy.dunn@owasp.org .

Reference

Red Teaming vs Pen Testing

Scope and Objectives:

- Penetration testing aims to identify as many vulnerabilities as possible within a specific scope, often focusing on technical infrastructure.
- Red teaming is more targeted and objective-oriented, focusing on testing an organization's overall security posture, including detection and response capabilities.

Methodology:

- Penetration testing follows a more structured and methodical approach, often with the support and awareness of the client's IT team.

- Red teaming employs a more comprehensive and stealthy approach, mimicking real-world adversaries and using various tactics like social engineering and physical infiltration.

Focus:

- Penetration testing focuses on finding and exploiting vulnerabilities.
- Red teaming emphasizes stealth, evasion, and testing the organization's ability to detect and respond to threats.

Maturity Level:

- Penetration testing is suitable for organizations at various stages of security maturity.
- Red teaming is more appropriate for organizations with mature security programs that have already addressed basic vulnerabilities.

https://www.cyderes.com/blog/penetration-testing-vs-red-teaming
https://www.pwc.com/mt/en/publications/technology/red-teaming-and-penetration-testing.html
https://www.cobalt.io/blog/red-teaming-vs.-pentesting