



ISO/IEC JTC 1/SC 27/WG 3 "Security evaluation, testing and specification"

Convenorship: UNE

Convenor: Bañón Miguel Mr



GenAI_Security_Testing_PWI_Proposal_20260916

Document type	Related content	Document date	Expected action
Project / Other	Meeting: VIRTUAL 14 Sep 2026	2026-09-17	INFO

Replaces: N 3165 PWI Security evaluation for GenAI_20260906

PWI on Security evaluation and testing methods for generative AI models

Standardization activities

Generative AI (GenAI) models, particularly large language models (LLMs), have been rapidly deployed across consumer and enterprise applications. While these models demonstrate powerful capabilities, they also introduce a broad range of security threats that differ fundamentally from traditional IT systems. These threats span the entire AI lifecycle, from training data collection and model development to deployment and inference-time interaction.

ISO/IEC FDIS 27090 provides guidance on threat descriptions, consequences, detection and mitigation. ISO/IEC WD 26160 will define AI-specific security functional requirements (SFRs), security assurance requirements (SARs) and evaluation work units. ISO/IEC AWI TS 42119-7 and ISO/IEC AWI TS 42119-8 address red team testing and quality evaluation, respectively. ISO/IEC CD TS 25568 provides guidance on addressing risks in generative AI systems. ISO/IEC WD 25959 defines attack potential tables for deep learning-based technology.

Challenges in using GenAI models in real-world applications

However, these existing and ongoing standards primarily address threat descriptions, risk management frameworks and general testing guidelines. A gap remains in standardized, reproducible and technically sound evaluation and testing methods for verifying the security of GenAI models against specific attack vectors. Specifically, there is a need for methods that define concrete test conditions, sample construction procedures, execution steps, pass/fail criteria and evidence recording requirements. This need is illustrated by the following challenges, described with examples drawn from LLMs as the most widely deployed class of GenAI models:

a) Incompatible testing across different LLMs

LLM vendors publish system cards that report the results of security evaluations (e.g., the robustness of an LLM against prompt injection attacks). However, each vendor reports these results using different evaluation methods and metrics, making it difficult to compare models directly. For example, the system card for OpenAI GPT-5 [1] reports the results of three evaluation categories: browsing prompt injections, tool-calling prompt injections, and coding prompt injections. In contrast, the system card for Anthropic Claude Opus 5 [2] evaluates multiple scenarios designed to assess susceptibility to indirect prompt injection attacks that could induce irreversible harmful actions, such as private data exfiltration, data destruction, system compromise, and unauthorized financial transactions.

A standardized testing methodology would enable users to evaluate different LLMs using a common framework, allowing more meaningful comparisons of their security performance.

[1] <https://arxiv.org/pdf/2601.03267v1>

[2] <https://www-cdn.anthropic.com/b514064af1408018e64b1ad24e7d5e75850b4ffd/Claude%20Opus%205%20System%20Card.pdf>

b) Lack of transparent testing

Developers often fine-tune open-source LLMs to adapt their capabilities to specific domains or applications. However, fine-tuning modifies the model's internal weights, which can unintentionally weaken or alter the security guardrails established by the original model developers.

When deploying a fine-tuned LLM, compliance teams may require evidence that the modified model remains secure. However, developers cannot simply rely on the original system card because the model has been changed. Instead, the original system card can serve as a baseline framework for validating the security of the fine-tuned model. A standardized testing methodology would support this validation by providing a transparent and consistent evaluation process, rather than relying on proprietary testing methods developed by individual organizations.

c) Guidance for test selection

The sheer volume of security testing methods for LLMs makes it impossible for evaluators to test every potential attack within a limited timeframe. Without clear guidance, different evaluators may assess LLMs using completely different approaches, leading to significant inconsistencies in evaluation quality.

To maximize efficiency, for example, evaluators may prioritize test methods that fully leverage available resources to achieve higher attack success rates. For instance, when a high evaluation assurance level is required, the evaluator should select white-box testing methods—utilizing internal information such as system prompts, model architectures, and parameters—to optimize attack vectors. Additionally, evaluators should adopt automated testing tools whenever possible.

Implementing a standardized methodology provides the structured guidance necessary for evaluators to conduct security testing as efficiently and consistently as possible.

Proposed scope

1. The PWI will research and develop security evaluation and testing methods for generative AI models, including adapted or fine-tuned models based on ISO/IEC 15408 and 18045. These evaluation and testing methods will be aligned with the SFRs and SARs defined in ISO/IEC 26160. The PWI will align with ISO/IEC 25959 by providing evaluation activities whose results can be mapped to attack potential ratings. Furthermore, through the development of these evaluation and testing methods, the PWI may identify gaps in the SFRs, SARs, evaluation activities, or attack potential tables currently defined in ISO/IEC 26160 and ISO/IEC 25959. Where such gaps are identified, the PWI intends to provide input to these standards. The outcome of this PWI may be proposals for new work items, new parts of existing standards, or expert inputs to ongoing standards.

2. The PWI will research methodological frameworks, principles and requirements for GenAI security evaluation and testing, addressing:

- evaluator competences, required tooling and model access levels (white-box, grey-box and black-box);
- methodological principles for the selection and justification of test parameters (e.g. sample sizes, iteration counts and decision thresholds);
- methods for the construction, curation and validation of test datasets and behaviour sets, including provenance, coverage, contamination avoidance, representativeness and multilingual/multimodal considerations;
- the specification of test conditions, sample/behaviour construction procedures and execution steps for each evaluation activity;
- attack-success determination using multiple independent verification mechanisms (e.g. rule-based, automated model-based and human review);
- the determination of verdicts for each evaluation activity;
- evidence recording, reproducibility and reporting.

3. In the first stage, the PWI will address evaluation and testing methods for the following threat categories:

- integrity (e.g. direct and indirect prompt injection, adversarial input attacks, data poisoning, backdoor attacks, and reasoning-chain and inference-time attacks);
- confidentiality (e.g. privacy inference attacks, side-channel and system-level information leakage, model exfiltration, training data leakage and system prompt leakage);
- availability (e.g. denial-of-service, resource exhaustion and denial-of-wallet);
- cross-cutting threats (e.g. supply-chain attacks and improper output handling).

Relationship with the PWI on security of AI agents and their orchestration

A parallel PWI [3] addresses security guidelines and evaluation methods for AI agents and their orchestration. The two PWIs agree that the boundary between them is drawn according to the target of evaluation (TOE): this PWI evaluates the generative AI model itself, through its native interfaces and without any agent runtime, while [3] evaluates agent systems, including their execution loops, tool use, memory and inter-agent interactions. Security properties that can be tested by sending prompts or queries to the model alone are in scope of this PWI; properties that arise only when the model is exercised within an agent execution loop are in scope of [3].

[3] PWI on security guidelines and evaluation methods for AI agents and their orchestration (WG3 N3158)

Collaboration with other SC/WG

We propose establishing this work item under SC27/WG3. However, because parts of the proposal fall within the scope of other SC/WGs, particularly SC42, we will submit the approved PWI proposal to the SC27 AHG3 (Security and privacy in AI and Big Data (BD)) following the September SC27/WG3 meeting to ensure proper cross-SC/WGs coordination.

Proposed project editors

Jiazhe Chen (CN), Naruki Kai (JP), Hanyu Li (CN), Heebong Choi (KR), Kwangwoo Lee (KR),
Seira Hidano (JP)