AEF-1: Minimum Operating Conditions for Independent Third Party AI Evaluations

Conrad Stosz[1*] **Karson Elmgren**[1] **Charles Foster**[2] **George Balston**[3] **Seth Donoughe**[4]
Samira Nedungadi[4] **Michael Chen**[2] **Jasper Götting**[4] **Sayash Kapoor**[5]
Sarah Schwettmann[1,6] **Rishi Bommasani**[1,7] **Luca Righetti**[2,8] **Sean McGregor**[3]
Rob Reich[1,7] **Arvind Narayanan**[5] **Chris Painter**[2] **Miles Brundage**[3] **Aidan Homewood**[8]
Divya Siddharth[9] **Faisal Lalani**[9] **Jaime Sevilla**[10] **Jacob Steinhardt**[1,11]

[1]Transluce [2]METR [3]AI Verification and Evaluation Research Institute [4]SecureBio
[5]Princeton University [6]Massachusetts Institute of Technology [7]Stanford University [8]GovAI
[9]Collective Intelligence Project [10]Epoch AI [11]UC Berkeley

Abstract

Independent third-party evaluations of AI systems are becoming increasingly central to AI governance approaches pursued by developers, governments, enterprises, and the public. However, the operating conditions of third-party evaluations can be opaque, even though they can significantly impact whether an evaluation is trustworthy and impartial. In particular, it can be challenging to understand whether an evaluator had sufficient access to assess the characteristics of interest, whether they enjoyed meaningful independence, and how transparently they shared their methods and findings. To address this, we present AEF-1, a standard and checklist that third-party evaluators can use to demonstrate how they achieved a set of operating conditions that support a baseline level of independence, access, and transparency during an evaluation.

1 Introduction

As AI systems become increasingly capable and widely deployed, there is a growing need for trustworthy, independent evaluations of their capabilities and risks [1, 2, 3, 4, 5, 6, 7]. AI evaluations are used for a range of purposes by multiple stakeholders, including AI providers evaluating their own systems as a routine part of development [8, 2]. But when customers, regulators, or the public demand high confidence in results, such as when evaluating risks to public safety, independent third parties can provide additional perspectives and approaches to evaluation and avoid the inherent conflicts of interests posed when AI providers evaluate themselves [9, 10, 6, 11].

The trustworthiness of an independent evaluation is the product of many distinct factors, including both technical and operational considerations. In particular, the unsettled science of AI evaluation is the subject of considerable ongoing work, not covered in this document [12, 13, 14, 15]. But the trustworthiness of a third-party evaluation is also significantly affected by the operational conditions under which it was carried out. This includes how much independence a third-party evaluator enjoyed in practice, whether they had sufficient access to the target system, and whether the evaluation was transparent enough to interpret its results accurately [16].

The independence of third-party evaluators is likewise affected affected by a range of factors, such as the extent to which the evaluator is organizationally controlled by or financially dependent upon the AI provider [17, 18]. These factors can be specific to a particular evaluation, such as if a provider narrowly constrains what methods the evaluator can use. Independence can also be affected by other relationships beyond that between the evaluator and provider, such as ties between an evaluator and

Direct correspondence to Conrad Stosz: `conrad@transluce.org`

the provider's direct competitors [19, 18]. Assessing independence therefore requires information about a range of the evaluator's relationships and their dynamics, beyond just whether or not an evaluation was carried out by a third party [16].

AI developers often limit access to their systems, constraining how independent third-parties can evaluate them. Even well-intentioned commercial providers face tradeoffs between preserving evaluators' methodological freedom and protecting their trade secrets and their users' privacy [18]. The necessary level of access is complex and dependent on evolving evaluation methods, with deeper access sometimes affording more reliable and accurate results [20, 21]. For example, methods to directly interpret a model's internal activations are improving over time, but third-party evaluators are generally constrained in practice from using these methods on closed-weight models because of the commercially sensitive nature of accessing these models' internals [21].

Transparency about independent evaluations can help facilitate greater trust [22], both by sharing information about independence and access, but also by helping to ensure that the technical methods and results of the evaluation can be interpreted accurately [16, 21]. This could include for instance ensuring that both positive and negative results are reported consistently and that they are not withheld to conceal poor performance. Methodological transparency can also support scientific rigor, including by aiding replicability and accurate interpretation of results, as well as allowing other third parties to contest or improve upon the evaluation's approach [23, 24, 15, 18].

Done right, independent evaluations can provide a much higher degree of trust in results and can help incentivize market-wide improvements in AI systems. Indeed, such evaluations have become central to AI governance efforts, including with their integration into many AI developers' frontier AI risk frameworks [25], system documentation [26, 27, 28], and voluntary commitments [29, 30]. A range of third-party leaderboards, evaluation reports, and other products have helped deliver on these commitments and fill broader commercial demand for independent evaluation results, covering a broad range of topics like general user preferences for chatbots [31], child safety [32], national security [33], and environmental impact [34]. Governments too have relied heavily on independent evaluations as a core component of their approach to AI risk, including through government testing authorities in the United States [35], United Kingdom [36], the European Union [37], and China [38], as well as the International Network of AI Safety Institutes [39].

2 Scope

This document establishes an initial voluntary standard that evaluators can use to demonstrate how they achieved a baseline set of operating conditions for the independence, access, and transparency of a particular third-party evaluation.

The standard does not cover all aspects of third-party AI evaluation. In particular, it does not address the many critical methodological considerations for conducting scientifically valid AI evaluations [3], nor does it exhaustively cover evaluators' various responsibilities towards system providers, such as responsibilities to act in good faith and avoid causing harm while conducting an evaluation.

This standard also does not apply to all third-party AI evaluations, which can take many forms and apply to many contexts, including those where factors like independence and transparency are less critical. For example, exploratory research collaborations often require less independence, more flexibility, and closer integration between AI providers and evaluators.

This document focuses specifically on evaluations where third parties maintain the freedom to define the evaluation's methodology to maximize the robustness and independence of the evaluation, as opposed to third parties simply validating, replicating, or supporting the execution of a provider's own evaluation [2, 6]. This distinction is sometimes similarly referred to as the difference between an "independent audit" versus narrower assessment approaches like "verification" or "validation", which may be a component of a third party assessment, but that generally assess a system's characteristics in the terms set by its developer. Independent validation of a provider's own methods can help increase trust in the relevant results, but given its limited scope and restricted methodology, it represents a lower standard of independence than that covered in this document.

By complying with this document, AI developers and third-party evaluators help demonstrate that a particular third-party evaluation reached a level of independence, access, and transparency that may be demanded in particularly sensitive contexts, such as when an evaluation is used to:

1. Establish that a system provider has adequately addressed risks of acute public concern,
2. Satisfy regulatory mandates for independent third party evaluations or audits, such as requirements to test for system safety, national security risks, or a system's suitability for regulated use cases,
3. Serve as a significant basis for critical internal governance decisions that concern public interest or well-being, such as whether and how a provider should release an AI system that may pose novel risks, or
4. Guide high-risk AI procurement decisions, such as whether a system performs adequately to be used in a safety-critical context.

3 Implementation

The minimum operating conditions in this document are organized into five core principles, each accompanied by a series of specific outcomes that are either required to achieve that principle and fulfill the standard, or are recommended across many circumstances, but do not necessarily apply in all cases. In summary, to comply with this standard, they should have:

1. Sufficient Access and Resources
2. Minimized Conflicts of Interest
3. Analytic Autonomy
4. Transparent Methods and Results
5. Protection of Sensitive Information

Third-party evaluators should demonstrate adherence to this standard by filling out the corresponding checklist (Appendix A) and including it alongside the evaluation results, such as in publications detailing an evaluation or in a report to governance bodies.[1] For any requirement in the checklist that cannot be fulfilled literally, justification should be provided for how the system provider and evaluator were able to achieve the same principle via alternative means. For evaluations that do not meet all of the requirements in Appendix A, third party evaluators can still use the checklist to document the relevant conditions of the evaluation, which specific requirements were not met, and why.

4 Minimum Operating Conditions

Version 1, updated December 4, 2025.

1: Sufficient Access and Resources

Rationale: Third-party evaluators must secure access to the right systems and information from system providers if they are to conduct a thorough, trustworthy evaluation. Evaluation results also often rely heavily on how much the evaluators are able to interact with the evaluated systems, as well as the time, resources, and tools they can apply in the evaluation.

1.1 Technical Access

Requirement: The evaluator secured sufficient technical access to assess the specific system characteristics being evaluated.

Different evaluations may require different levels of access, which may include the following items, among others:

[1] In particular, the checklist could be used by developers to report details of third-party evaluations per the stipulations of the EU AI Act Code of Practice Safety & Security Chapter Measure 7.3(1)(g) and California's Senate Bill 53 22757.12.(c)(2)(C). See https://code-of-practice.ai/?section=safety-security#measure-7-3-documentation-of-systemic-risk-identification-analysis-and-mitigation and https://legiscan.com/CA/text/SB53/id/3270002

1. *Query access.* At a minimum, evaluators must have open-ended black-box access (via a web UI, an API or both, depending on the requirements of the evaluation) that provides timely and flexible queries.[2]
2. *Scaffolding.* Models are generally deployed and used with accompanying tools, safeguards, and other configurations that will be in place for a realistic end user. This scaffolding meaningfully changes a model's performance and properties in the expected deployment context, requiring that the evaluator have access to the expected scaffolding to conduct a realistic, externally valid evaluation.
3. *Safeguard exemptions.* Providers should consider disabling, removing, or minimizing any technical safeguards that would unduly hinder evaluators from producing trustworthy results for the evaluation in question. For example, reliably evaluating the worst-case scenarios for misuse of dangerous capabilities generally requires access to a model version where mitigations that might reduce these capabilities are minimized.[3]
4. *Intermediate system states.* Intermediate system states like model activations and reasoning traces are often important to produce trustworthy evaluation results, particularly when examining how models arrive at final conclusions, such as evaluating for evaluation awareness or scheming. Such access can involve just observing such states, but some evaluation methods may also require being able to directly modify them, such as to establish their causality or investigate mitigations that rely on such modifications.
5. *Finetuning*. For cases where the use case or threat model involves users with the ability to finetune the model, evaluators would need access to a similar ability to conduct a trustworthy evaluation.
6. *Model weights.* Particularly when evaluating a model whose weights are or will be shared externally, ensuring that the evaluator also has access to these weights helps to ensure the results' validity to the deployment context.
7. *Other tools.* System providers should provide access to other tools which are necessary to elicit realistic deployment conditions or otherwise ensure external validity.
8. *User data.* While often exceptionally sensitive and generally requiring careful legal, security, and privacy guardrails, considering a provider's user data can help ground evaluations in realistic deployment conditions, such as determining whether a particular behavior of concern observed in testing has also occurred in production.

1.2 Information

Recommendation: The system provider shared with the evaluator information relevant to carrying out a trustworthy and useful evaluation.

Depending on the context, relevant information may include, among others:

1. At a minimum, clear identification of the systems and specific versions being tested.
2. Details of the model's intended behavior (such as the model specification and system prompt).
3. Information about the training process and data (including test sets).
4. Information necessary for fully eliciting the model's capabilities.
5. Preexisting internal evaluation results that might inform the evaluator's approach.
6. Factors expected to contribute to how well the evaluation would generalize from the evaluation context to the deployment context.

[2] Besides being more time-efficient and convenient than chat access in most cases, API access is necessary for some important kinds of evaluations, such as evaluating a model by deploying it in an agent scaffold.

[3] However, this may not be practical for all categories of misuse, some of which are illegal and/or harmful even in the context of evaluation, such as generating child sexual abuse material, violating the privacy of real individuals, or simulating dangerous activities closely enough to cause real-world harm. Instead, other methods like safe proxies should be used to evaluate these cases responsibly.

7. Information on relevant patterns of real-world usage, such as about known performance patterns or instances of relevant misuse.[4]
8. Knowledge related to the system's propensity to produce misleading results, such as through reward hacking or contamination.
9. Knowledge of system vulnerabilities or other security issues that might, for instance, allow attackers to bypass system safeguards or degrade system performance.
10. Information that might contradict the evaluation result.

1.3 Computational Resources

Requirement: The evaluator had access to sufficient computational resources to complete a thorough evaluation.

This might include, for example:

1. Multiple evaluation runs to ensure statistical robustness by accounting for sampling variance.
2. Exploratory probing to investigate issues related to the target characteristic to produce a thorough evaluation.
3. Sufficient query and token limits to elicit high performance from the model.

1.4 Time

Requirement: The evaluator had adequate time to carry out a thorough evaluation.

Time allotted to evaluations should generally be proportionate to confidence needed in the results as well as the degree of novelty of the system, in terms of its capabilities, structure, or otherwise. It should also account for the expected speed of evaluation given the type of access and evaluation methods used.[5] Based on prior industry experience, at least 20 business days are often necessary to carry out the various stages of evaluations when assessing substantially novel systems or system characteristics, and possibly more depending on the type of evaluation.[6]

Where the evaluator explicitly agrees to a timeline with the provider, it should:

1. Be set sufficiently in advance to allow the evaluator to plan business operations for the evaluation.
2. Allow for evaluators to carry out each necessary evaluation stage, such as to debug system access, design an appropriate evaluation, carry it out, analyze the results, and make adjustments as necessary.
3. Allow the provider time to consider and act on feedback from the evaluator before any major resulting decisions are made, such as a decision whether or not to release the system.

1.5 Safe harbor

Recommendation: The system provider provided legal safe harbor for actions by the evaluator that are within the agreed-upon scope of the evaluation.

[4] For example, OpenAI provided METR several assurances for their evaluation of GPT-5, detailed here: https://evaluations.metr.org/gpt-5-report/#assurance-checklist-summary

[5] For example, evaluations conducted manually through a web UI tend to be slower than those conducted through an API.

[6] This aligns with the EU CoP Safety & Security Chapter Appendix 3.4 suggestion of at least 20 business days for most systemic risks and model evaluation methods, see https://code-of-practice.ai/?section=safety-security#appendix-3-4-qualified-model-evaluation-teams-and-adequate-resources. As one example precedent, METR's evaluation of GPT-5 took three weeks, see https://evaluations.metr.org/gpt-5-report/#metr%E2%80%99s-access-to-gpt-5

In some evaluations, default legal agreements such as a system provider's standard terms of service may be sufficient to satisfy this criteria. In others, such as adversarial testing to bypass a system's safeguards against misuse, the provider may need to waive certain terms of service or other legal rights to permit the evaluator's actions. In all cases, evaluators must be able to operate under the expectation that they can carry out an agreed-upon evaluation thoroughly and in good faith, without fear of adverse legal action for doing so, in particular to prevent any threat of retaliation over the content or presentation of the results.

2: Minimized Conflicts of Interest

Rationale: If an evaluator faces a financial incentive or related pressures to produce findings that are not accurate or that are not reasonably complete, that is a conflict of interest that can undermine both independent judgment and external trust. The appearance of a conflict of interest, or appearing to conceal one, can also degrade external trust, regardless of whether the individuals involved feel themselves to be conflicted. As a result, it is important for organizations to both take concrete steps to ensure they have minimized conflicts, as well as to provide broader transparency about their relationships.

2.1 Contingent compensation

Requirement: The evaluator did not receive compensation contingent on the results of the evaluation.

2.2 Organizational control

Requirement: The system provider did not exercise organizational or financial control over the evaluator.

This includes ensuring at least the following conditions, among others:

1. The system provider does not own either voting shares or a large proportion of shares overall in the evaluator organization.
2. The system provider does not control any of the evaluator's board seats.

2.3 CoI policy

Requirement: The evaluator has published a conflict of interest policy, and it was applied to the evaluation.

Such policies should generally cover at least:

1. The organization's reporting requirements for factors that could reasonably contribute to a conflict of interest.
2. The circumstances under which individuals, subcontractors, or other entities involved in the evaluation would be recused from involvement in the evaluation to avoid conflicts of interest.
3. Any restrictions on funding or payment received from relevant system providers.

2.4 CoI disclosure

Requirement: The evaluator clearly disclosed potential conflicts of interest relevant to the evaluation.

To ensure adequate transparency into potential conflicts of interest, the evaluator must make disclosures at least in the following circumstances, among others:[7]

1. The evaluator was paid by the system provider or its direct competitors to conduct the evaluation.

[7] These categories of conflicts of interest are broadly similar to those in regulations on credit rating agencies and financial auditing. See: https://www.law.cornell.edu/cfr/text/17/240.17g-5 and https://pcaobus.org/oversight/standards/ethics-independence-rules/details/ET101

2. A meaningful fraction of the evaluator's funding (either revenue or donations) comes from a system provider, its employees, or its direct competitors.
3. The system provider, its employees, or its direct competitors own equity in the evaluator organization.
4. Evaluator staff who carried out the evaluation own equity in the system provider or its direct competitors.
5. Individuals working on the evaluation also work for the system provider or its direct competitors, based on belonging to both the groups defined below in (a) and (b) simultaneously:
    (a) Individuals involved with the evaluation include staff working on the evaluation, their management chain, or any others with influence over the methods, design, execution, review, approval, or presentation of the evaluation and its results.
    (b) Individuals working for a system provider include any who serve as an employee, contractor, board member, or advisor to a system provider (this does not include those working for the evaluator as part of a contract for the evaluation in question).

These disclosures should be included when the evaluator disseminates the evaluation results to other organizations, including via private reports or public release.

2.5 Recusals

Requirement: The evaluator recused any individuals with a significant financial interest in the system provider from carrying out the evaluation.

Recused staff must not define or carry out evaluations of the provider's system(s) or modify or approve relevant findings. At a minimum, individuals must recuse themselves if they hold direct equity in an evaluated system provider, either personally or through a spouse, spousal equivalent, or dependent.[8]

However, this does not include:

1. Staff with only indirect holdings in the system provider which are not directly managed by the individual, such as through an index or mutual fund.
2. Staff whose involvement is limited to support services, such as providing logistical support or helping to implement infrastructure or tools, provided these activities are directly and effectively overseen by qualified leadership who have transparency into the relevant financial interests.

2.6 Agreements

Recommendation: The evaluator disclosed any separate agreements with the system provider that significantly impact the independence and trustworthiness of the evaluation.

Such disclosure should cover at least the nature of the agreements and how they could impact the evaluation. These disclosures should be included in any documentation produced by the evaluator for dissemination to other organizations, including publications as well as private reports to the organizational board or government bodies.

3: Analytic Autonomy

Rationale: To maximize the robustness and independence of an evaluation, third-party evaluators must control the methodology–which heavily impacts results and how trustworthy they are–as well as retaining the ability to present their results accurately and free from undue influence from system providers.

3.1 Scoping

[8] In line with the Public Company Accounting Oversight Board's definition of an "immediate family member", see https://pcaobus.org/about/rules-rulemaking/rules/section_3#rule3501

Recommendation: The evaluator retained flexibility to define which specific properties of the system to evaluate.

Scoping evaluations too narrowly can undermine trustworthiness by limiting the evaluation to misrepresentative subcategories of a broader system characteristic, such as an evaluation that claims to target mental health risks broadly but that only examines suicidal ideation and ignores other risks like eating disorders and violent behavior. Narrow scope can also exclude overlapping system behaviors that may significantly contribute to that characteristic, such as a system's general sycophancy contributing to specific negative mental health outcomes.

Even in cases where system providers and evaluators agree to scope an evaluation to a general category, evaluators should retain flexibility where possible to define which subcategories are most meaningful to evaluate and which related properties are necessary to usefully characterize the agreed upon category. As this may become clear only during the course of an evaluation, evaluators should retain the ability to adjust their focus along these lines throughout the evaluation.

3.2 Evaluation autonomy

Requirement: The evaluator had autonomy in deciding the methods of the evaluation.

This includes at least the following specific methodological characteristics, among others:

1. Determining the appropriate metrics, including summary statistics and sampling strategies, such as best of N, pass@k, etc.
2. How to best elicit the target properties, such as defining the inputs used to prompt the system, deciding how to scaffold agents, what tools to provide them, and if and how jailbreaking is necessary.
3. Defining task completion, scoring rubrics, and other definitions of success or failure.

3.3 Direct access

Recommendation: The evaluator ran the evaluations themselves via direct system access.

Evaluators should generally maintain direct technical control of the evaluation, rather than relying on staff of the system provider to serve as intermediaries. However, this condition may not always be justified, especially for evaluation methods dealing with highly sensitive types of access, such as those that involve production user data or direct access to proprietary model weights. In such cases evaluators may also seek other ways to validate that their evaluations are carried out consistent with their intent and without undermining the evaluation's integrity, such as contractual guarantees that providers will not use their methods other than to carry out evaluation activities on the evaluator's behalf.

3.4 Editorial control

Requirement: The evaluator retained editorial control over how they present the results of their evaluation.

This editorial control must apply whenever the evaluator discloses its results to any given audience, and it must include at least freedom to decide the following elements, among others:

1. Determining how to characterize what they evaluated and how.
2. Describing the system's observed performance as well as supporting evidence such as scores, capability levels, failure modes, specific input and output examples, and any other relevant facts.
3. Describing how the results appear to relate to relevant red-lines or rule-out thresholds.
4. How to make baseline comparisons, such as to human, random, and marginal risk references.
5. Commenting on what was in and out of scope for the evaluation, and what conclusions can and cannot be drawn from the evaluation given its scope.
6. Describing how any lack of access to relevant systems or information limits the conclusions that can be drawn from the evaluation.

7. Describing the general nature of any redactions or other limitations on sharing information about the evaluation due to its sensitivity.

However, the system provider and other relevant third parties can and should inform the evaluation, provide feedback on results, and present their own dissenting views where necessary. Note also that satisfying this condition does *not* mean that system providers are prohibited from protecting confidential information shared with the evaluator, nor does it prevent evaluators and system providers from agreeing to limit disclosure only to specific audiences (e.g., for evaluations of non-public systems).

4: Transparent Methods and Results

Rationale: For evaluations to serve their purpose in informing governance, their results must not be misrepresented, unduly delayed, or selectively withheld depending on the findings, and they must be provided in sufficient detail to demonstrate their scientific validity and to enable replicability and critique.

4.1 Methodological transparency

Requirement: The evaluator shared sufficient methodological details to allow for independent review of the results.

Evaluators should disclose alongside their results enough details that their methodological approach and results can be scrutinized by external parties, ideally allowing for direct replication of results. Evaluators are permitted to withhold testing datasets to maintain the integrity of their evaluations, but should release representative examples of the data used in the evaluation at a minimum. Evaluators should also limit disclosure of information hazards that would cause harm if disclosed, such as specific methods to generate child sexual abuse material or produce dangerous materials, in line with a responsible disclosure policy as described in item 5.4.

4.2 Disclosure rights

Recommendation: The system provider granted upfront any necessary rights to disclose results to the evaluation's intended audiences.

Doing so will help to avoid the appearance or reality that disclosure is contingent on results being favorable.

For evaluations of risk, target audiences should generally include at least the relevant staff or affiliates of the system provider charged with governance (e.g., the provider's board of directors and/or any safety board) and any recipients mandated by applicable regulations.

When evaluating a system before it is widely released, evaluators and providers should consider agreeing upfront if and how the evaluation results will be released once the system is released. This might include, for instance, defining any necessary caveats to communicate differences between the version of the model that was evaluated and the version that was ultimately released.

When evaluating public systems via publicly available access methods, evaluators will most often have broad freedom to disclose results by default, with no action necessary to satisfy this recommendation.

4.3 No contingent release

Recommendation: The intended audiences for an evaluation were not narrowed based on the results.

The system provider should not withhold results from intended audiences, even if they indicate poor performance, unmitigated risks, or other concerning findings. Any expected disclosure to the system provider's relevant internal governance bodies should also proceed regardless of whether an evaluated model is ultimately deployed.

The above does not apply when disclosure is withheld specifically to follow pre-defined responsible disclosure practices as described in condition 5.4.

4.4 No misrepresentation

<u>Recommendation:</u> **The system provider did not misrepresent the evaluation's findings.**

This includes at least the following forms of misrepresentation, among others:

1. Directly misstating what the evaluator found. However, evaluators should not seek to constrain others from *disagreeing* with their methodology or findings.
2. Using the fact of the evaluation, or partial results of the evaluation, as evidence of a system's capabilities or risks without making the full scope of the methodology and results available to the same audience.

To help avoid misrepresentation, system providers should generally allow evaluators to review representations of the evaluation, especially for use in prominent publications and other impactful disclosures.

4.5 Timely disclosure

<u>Recommendation:</u> **The system provider or other external parties did not unduly delay the evaluation from being disclosed based on the content of the results.**

Disclosure of the evaluation should not be subject to delay by the system provider or other external parties based on the content of the findings, except to follow pre-defined responsible disclosure practices as described in condition 5.4, which may delay sharing of particular risks and vulnerabilities.

Agreements to delay disclosure for reasons other than the content of the results may also be appropriate, such as to provide the system provider advanced notice of publication or to avoid public release of results on an unreleased system until it is released.

4.6 Redactions

<u>Requirement:</u> **The system provider did not have authority to redact results to conceal concerning findings.**

Any authority that is granted to the system provider or another external party to review the results of an evaluation and redact sensitive information from it or block reporting altogether as a result must only be used as is necessary to protect privacy, ethical restrictions, trade secrets, or information hazards, or to ensure factual accuracy about the terms of the agreement or the provider's actions. It must not be used not to conceal general poor performance, unmitigated system issues, or other concerning findings or to prevent results from being shared based on disagreements over how to interpret them. Disclosure of system vulnerabilities should be governed by a responsible disclosure policy, as described in condition 5.4.

4.7 Redaction disclaimer

<u>Recommendation:</u> **The evaluator clearly disclosed the redaction authorities granted to the system provider or other external entities.**

In particular, any authorities granted to the system provider or other external parties to edit or redact information from the evaluator's publications should be noted in a disclaimer on the publication. This disclaimer should always be included, and it should specify whether the authorities were exercised or not, as well as describing what was redacted and why where possible. The evaluator should also retain the ability to comment on the appropriateness of redactions and how material they are to understanding the evaluation.

5: Protection of Sensitive Information

Rationale: Both system providers and evaluators possess significant sensitive information, including intellectual property and potentially hazardous information, improper disclosure of which could harm their own and/or public interests. Adequately protecting sensitive information shared or obtained in the course of an evaluation is therefore both an ethical responsibility, as well as a practical precondition for third parties to maintain the integrity of their evaluation methods and to maintain sufficient system access to be able to carry out meaningful evaluations.

5.1 Publication terms

Requirement: The evaluator had the system provider's permission prior to releasing any results based on non-public information or systems.

Where evaluations are based only on information and access that is publicly available, such as when using a publicly-released model under its general terms of service, evaluators often need not consult with the system provider regarding publication, except for instance where evaluation methods reveal particularly sensitive information, such as extracting sensitive personally identifiable information, or else system vulnerabilities that should be responsibly disclosed, as described in condition 5.4.

5.2 Evaluation integrity

Recommendation: The evaluations methods were not gamed or leaked.

To prevent this, evaluators and system providers should do at least the following, among others:

1. The evaluator should minimize how much information they share that the system provider might foreseeably use to game their evaluations.
2. The system provider should protect the confidential information they receive for the purposes of the evaluation.
3. The system provider should agree not to view, retain, share, use, or train on system interactions that include held out evaluation data (e.g., transcripts, prompts, or answer keys), except with written consent from the evaluator (e.g., for refusal training).[9]
4. The system provider should make a good-faith attempt to provide their system in a manner that will provide accurate results, including aligning the evaluated version of the system to real-world deployments, except where differences are disclosed to the developer and are reasonably necessary (e.g., because the production configuration is not yet complete at the time of evaluation, or where the provider removes safeguards to help the evaluator assess worst case scenarios).
5. The system provider and evaluator should cooperate to help identify and address system behavior that may undermine the evaluation, such as through overfitting, reward-hacking, or sandbagging.

Evaluators should also consider adding canaries to evaluation data to enable later detection of models that have been trained on evaluations.

5.3 Protecting confidential information

Requirement: The evaluator implemented measures to protect any confidential information they received from the system provider.[10]

Examples:

[9] This is in line with the requirement per the EU CoP Safety & Security Chapter Appendix 3.5, which requires signatories to not "undermine the integrity of external model evaluations by storing and/or analysing inputs and/or outputs from test runs without express permission from the evaluators", see https://code-of-practice.ai/?section=safety-security#appendix-3-5-independent-external-model-evaluations

[10] Note that doing so would help establish evaluators as adequately qualified per the EU CoP Safety & Security Appendix 3.5, see https://code-of-practice.ai/?section=safety-security#appendix-3-5-independent-external-model-evaluations

1. Internal measures such as staff non-disclosure agreements, access controls, and cybersecurity protections to limit access to confidential information.
2. An agreement to withhold publishing results generated from a non-public system until that system is released.

5.4 Responsible disclosure policy

Requirement: The evaluator established and followed a responsible disclosure policy.

This includes at least the following measures, among others:

1. First disclosing novel system vulnerabilities[11]–such as meaningfully novel weaknesses in system safeguards or prompt injection methods–to the system provider, as well as other providers known to be vulnerable where feasible, before making them public. Vulnerability disclosure should generally follow at least these principles:
   (a) It may not be appropriate to publicly disclose vulnerabilities in safeguards against misuse if they cannot reasonably be fixed and are likely to be exploited, such as a method to generate child sexual abuse material using a widely disseminated open-weight model. It may be more appropriately to treat such vulnerabilities as information hazards, as in bullet 2 below.
   (b) Otherwise, to protect users and the broader public, evaluators should generally not withhold information about vulnerabilities in publicly available systems for longer than 60 days after disclosure to the system provider.
   (c) Where a system provider does not itself follow a reasonable responsible disclosure policy that offers safe harbor and instead may, for instance, retaliate against an evaluator for disclosing system vulnerabilities, prior notice to the provider before publication is not required.
   (d) When considering whether and how to disclose a particular probabilistic attack, like a particular string variation that can successfully trigger a jailbreak or prompt injection, evaluators should balance the marginal utility of the discovered information to defenders, evaluators, and potential attackers. Given the difficulty of fully fixing such flaws, it may not meaningfully advance security to disclose all discovered attack variations, and doing so may also greatly undermine evaluators' ability to evaluate system security over time.
2. Redacting other particularly sensitive details from their results, including sensitive personal information and information that presents a severe information hazard (such as CBRN development or the generation of child sexual abuse material) and for which no near term mitigation is possible. This may also involve redacting such details in materials provided to the system provider.

References

[1] Yoshua Bengio. International AI safety report. January 2025.

[2] Frontier Model Forum. Technical Report on Third-Party Assessments, April 2025.

[3] Shayne Longpre, Kevin Klyman, Ruth E. Appel, Sayash Kapoor, Rishi Bommasani, Michelle Sahar, Sean McGregor, Avijit Ghosh, Borhane Blili-Hamelin, Nathan Butters, Alondra Nelson, Amit Elazari, Andrew Sellars, Casey John Ellis, Dane Sherrets, Dawn Song, Harley Geiger, Ilona Cohen, Lauren McIlvenny, Madhulika Srikumar, Mark M. Jaycox, Markus Anderljung, Nadine Farid Johnson, Nicholas Carlini, Nicolas Miailhe, Nik Marda, Peter Henderson, Rebecca S. Portnoff, Rebecca Weiss, Victoria Westerhoff, Yacine Jernite, Rumman Chowdhury, Percy Liang, and Arvind Narayanan. In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI, March 2025.

[4] Matthias Samwald, Yoshua Bengio, Marietje Schaake, Marta Ziosi, Daniel Privitera, Anka Reuel, Alexander Zacherl, Nitarshan Rajkumar, and Markus Anderljung. Code of Practice for General-Purpose AI Models.

[11] For more detail on responsible disclosure of system flaws, consider existing approaches such as in [3]

[5] U.S. National Telecommunications and Information Administration. AI Accountability Policy Report: Independent Evaluations. https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/ai-system-evaluations/independent-evaluations, March 2024.

[6] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel E. Ho. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance, June 2022.

[7] Lydia T. Liu, Inioluwa Deborah Raji, Angela Zhou, Luke Guerdan, Jessica Hullman, Daniel Malinsky, Bryan Wilder, Simone Zhang, Hammaad Adam, Amanda Coston, Ben Laufer, Ezinne Nwankwo, Michael Zanger-Tishler, Eli Ben-Michael, Solon Barocas, Avi Feller, Marissa Gerchick, Talia Gillis, Shion Guha, Daniel Ho, Lily Hu, Kosuke Imai, Sayash Kapoor, Joshua Loftus, Razieh Nabi, Arvind Narayanan, Ben Recht, Juan Carlos Perdomo, Matthew Salganik, Mark Sendak, Alexander Tolbert, Berk Ustun, Suresh Venkatasubramanian, Angelina Wang, and Ashia Wilson. Bridging Prediction and Intervention Problems in Social Systems, November 2025.

[8] Miranda Bogen. Assessing AI: Surveying the Spectrum of Approaches to Understanding and Auditing AI Systems. January 2025.

[9] Markus Anderljung, Everett Thornton Smith, Joe O'Brien, Lisa Soder, Benjamin Bucknall, Emma Bluemke, Jonas Schuett, Robert Trager, Lacey Strahm, and Rumman Chowdhury. Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework, November 2023.

[10] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, J. B. Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, April 2020.

[11] Anka Reuel, Avijit Ghosh, Jenny Chim, Andrew Tran, Yanan Long, Jennifer Mickel, Usman Gohar, Srishti Yadav, Pawan Sasanka Ammanamanchi, Mowafak Allaham, Hossein A. Rahmani, Mubashara Akhtar, Felix Friedrich, Robert Scholz, Michael Alexander Riegler, Jan Batzner, Eliya Habba, Arushi Saxena, Anastassia Kornilova, Kevin Wei, Prajna Soni, Yohan Mathew, Kevin Klyman, Jeba Sania, Subramanyam Sahoo, Olivia Beyer Bruvik, Pouya Sadeghi, Sujata Goswami, Angelina Wang, Yacine Jernite, Zeerak Talat, Stella Biderman, Mykel Kochenderfer, Sanmi Koyejo, and Irene Solaiman. Who Evaluates AI's Social Impacts? Mapping Coverage and Gaps in First and Third Party Evaluations, November 2025.

[12] Patricia Paskov, Michael J. Byun, Kevin Wei, and Toby Webster. Preliminary suggestions for rigorous GPAI model evaluations. Technical report, May 2025.

[13] Laura Weidinger, Deb Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, and Sanmi Koyejo. Toward an Evaluation Science for Generative AI Systems. https://www.nae.edu/19579/19582/21020/337862/338231/Toward-an-Evaluation-Science-for-Generative-AI-Systems#about_author338231.

[14] Evan Miller. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations, November 2024.

[15] Tegan McCaslin, Jide Alaga, Samira Nedungadi, Seth Donoughe, Tom Reed, Rishi Bommasani, Chris Painter, and Luca Righetti. STREAM (ChemBio): A Standard for Transparently Reporting Evaluations in AI Model Reports, September 2025.

[16] Leon Staufer, Mick Yang, Anka Reuel, and Stephen Casper. Audit Cards: Contextualizing AI Evaluations, August 2025.

[17] Lara Thurnherr, Robert Trager, Amin Oueslati, Christoph Winter, Cliodhna Ní Ghuidhir, Joe O'Brien, Jun Shern Chan, Lorenzo Pacchiardi, Anka Reuel, Merlin Stein, Oliver Guest, Oliver Sourbut, Renan Araujo, Seth Donoughe, and Yi Zeng. Who Should Develop Which AI Evaluations?

[18] Elliot Jones, Mahi Hardalupas, and William Agnew. Under the radar? Examining the evaluation of foundation models. https://www.adalovelaceinstitute.org/report/under-the-radar/, July 2024.

[19] Hritik Bansal and Pratyush Maini. Peeking Behind Closed Doors: Risks of LLM Evaluation by Private Data Curators, February 2025.

[20] Benjamin S Bucknall and Robert F Trager. Structured Access for Third-Party Research on Frontier Ai Models: Investigating Researchers' Model Access Requirements. October 2023.

[21] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-Box Access is Insufficient for Rigorous AI Audits. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2254–2272, Rio de Janeiro Brazil, June 2024. ACM. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3659037.

[22] Rishi Bommasani, Daniel Zhang, Tony Lee, and Percy Liang. Improving Transparency in AI Language Models: A Holistic Evaluation. February 2023.

[23] Sayash Kapoor, Emily M. Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A. Bail, Odd Erik Gundersen, Jake M. Hofman, Jessica Hullman, Michael A. Lones, Momin M. Malik, Priyanka Nanayakkara, Russell A. Poldrack, Inioluwa Deborah Raji, Michael Roberts, Matthew J. Salganik, Marta Serra-Garcia, Brandon M. Stewart, Gilles Vandewiele, and Arvind Narayanan. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances*, 10(18):eadk3452, May 2024. doi: 10.1126/sciadv.adk3452.

[24] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, September 2023. ISSN 2666-3899. doi: 10.1016/j.patter.2023.100804.

[25] Frontier Model Forum. Issue Brief: Components of Frontier AI Safety Frameworks, November 2024.

[26] OpenAI. GPT-5 System Card, August 2025.

[27] Anthropic. Claude Sonnet 4.5 System Card, September 2025.

[28] Google. Gemini 2.5 Pro - Model Card. June 2025.

[29] UK Department for Science, Innovation & Technology. Frontier AI Safety Commitments, AI Seoul Summit 2024. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024.

[30] The White House. Voluntary AI Commitments, September 2023.

[31] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference, March 2024.

[32] Common Sense Media. AI Risk Assessment ChatGPT 5. October 2025.

[33] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao,

Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning, May 2024.

[34] HuggingFace. AI Energy Score. https://huggingface.github.io/AIEnergyScore/.

[35] National Institute of Standards and Technology. Pre-Deployment Evaluation of OpenAI's o1 Model. *NIST*, December 2024.

[36] UK AI Security Institute. Early lessons from evaluating frontier AI systems. https://www.aisi.gov.uk/blog/early-lessons-from-evaluating-frontier-ai-systems, October 2024.

[37] The European Commission. AI Office contributes to the third-joint testing exercise of the International Network of AI Safety Institutes. https://digital-strategy.ec.europa.eu/en/news/ai-office-contributes-third-joint-testing-exercise-international-network-ai-safety-institutes, July 2025.

[38] Shanghai AI Lab, Xiaoyang Chen, Yunhao Chen, Zeren Chen, Zhiyun Chen, Hanyun Cui, Yawen Duan, Jiaxuan Guo, Qi Guo, Xuhao Hu, Hong Huang, Lige Huang, Chunxiao Li, Juncheng Li, Qihao Lin, Dongrui Liu, Xinmin Liu, Zicheng Liu, Chaochao Lu, Xiaoya Lu, Jingjing Qu, Qibing Ren, Jing Shao, Jingwei Shi, Jingwei Sun, Peng Wang, Weibing Wang, Jia Xu, Lewen Yan, Xiao Yu, Yi Yu, Boxuan Zhang, Jie Zhang, Weichen Zhang, Zhijie Zheng, Tianyi Zhou, and Bowen Zhou. Frontier AI Risk Management Framework in Practice: A Risk Analysis Technical Report, July 2025.

[39] International Network of AI Safety Institutes. International Joint Testing Exercise: Agentic Testing.

Appendix A Implementation Checklist

Evaluation results claiming to implement this standard should include the following checklist to demonstrate compliance.[12] Each condition in the checklist is marked as either required to comply with these minimum operating conditions, or else recommended but not applicable in all cases.

[12] Template checklists for .docx and LaTeX are available at https://aievaluatorforum.org/initiatives/minimum-operating-conditions.

AEF-1 Checklist

This checklist summarizes how this evaluation addressed the requirements and recommendations for independent AI evaluations established by the voluntary standard AEF-1 *Minimum Operating Conditions for Independent Third Party AI Evaluations*, which are intended to help ensure the independence, transparency, and access of third-party AI evaluations. See https://aievaluatorforum.org/initiatives/minimum-operating-conditions for more details, including more detailed definitions of each condition below. This is version 1 of this checklist format (updated December 4, 2025).

Does this evaluation satisfy all the minimum requirements of AEF-1? (provide details below)	*Yes/No*

1: Secure Sufficient Access and Resources

Condition		Fulfilled?	Notes/Evidence
1.1	**Requirement:** The evaluator secured sufficient technical access to assess the specific system characteristics being evaluated. *Note: for the below sub-elements of condition 1.1, only 1.1.1 is required. The rest may be important to a valid evaluation, but will depend on the nature of the evaluation methods.*	*Yes/No*	
	The evaluator had query access.	*Yes/No*	
	The evaluator had access to the system's scaffolding.	*Yes/No*	
	The evaluator had exemptions from system safeguards.	*Yes/No*	
	The evaluator had access to intermediate system states.	*Yes/No*	
	The evaluator had finetuning access.	*Yes/No*	
	The evaluator had access to model weights.	*Yes/No*	
	The evaluator was granted access to other tools for elicitation or otherwise supporting external validity.	*Yes/No*	
	The evaluator had access to relevant user data.	*Yes/No*	
1.2	**Recommendation:** The system provider shared with the evaluator information relevant to carrying out a trustworthy and useful evaluation.	*Yes/No*	
1.3	**Requirement:** The evaluator had access to sufficient computational resources to complete a thorough evaluation.	*Yes/No*	

Condition		Fulfilled?	Notes/Evidence
1.4	**Requirement:** The evaluator had adequate time to carry out a thorough evaluation.	*Yes/No*	
1.5	**Recommendation:** The system provider provided legal safe harbor for actions by the evaluator that are within the agreed upon scope of the evaluation.	*Yes/No*	
2: Minimized Conflicts of Interest			
2.1	**Requirement:** The evaluator did not receive compensation contingent on the results of the evaluation.	*Yes/No*	
2.2	**Requirement:** The system provider did not exercise organizational or financial control over the evaluator.	*Yes/No*	
2.3	**Requirement:** The evaluator has published a conflict of interest policy, and it was applied to the evaluation.	*Yes/No*	(Should include a link to the policy on a publicly available website)
2.4	**Requirement:** The evaluator clearly disclosed potential conflicts of interest relevant to the evaluation. *Note: For each of the below sub-elements of condition 2.4, indicate whether the answer is "yes" or "no" in the "Notes/Evidence" column, and add accompanying information if the answer is "yes".*	*Yes/No*	
	Was the evaluator paid by the system provider or its direct competitor to conduct the evaluation?	*Yes/No*	
	Does a meaningful fraction of the evaluator's funding come from the system provider, its employees, or its direct competitors?	*Yes/No*	
	Do the system provider, its employees, or its direct competitors own equity in the evaluator organization?	*Yes/No*	
	Do evaluator staff who carried out the evaluation own equity in the system provider or its direct competitors?	*Yes/No*	
	Do any evaluation staff working on the evaluation simultaneously work for the system provider or its direct competitors?	*Yes/No*	
	Did the evaluator have any other conflicts of interest relevant to the evaluation?	*Yes/No*	

Condition		Fulfilled?	Notes/Evidence
2.5	**Requirement:** The evaluator recused any individuals with a significant financial interest in the system provider from carrying out the evaluation.	*Yes/No*	
2.6	**Recommendation:** The evaluator disclosed any separate agreements with the system provider that significantly impact the independence and trustworthiness of the evaluation.	*Yes/No*	
3: Analytic Autonomy			
3.1	**Recommendation:** The evaluator retained flexibility to define which specific properties of the system to evaluate.	*Yes/No*	
3.2	**Requirement:** The evaluator had autonomy in deciding the methods of the evaluation.	*Yes/No*	
3.3	**Recommendation:** The evaluator ran the evaluations themselves via direct system access.	*Yes/No*	
3.4	**Requirement:** The evaluator retained editorial control over how they present the results of their evaluation.	*Yes/No*	
4: Transparent Methods and Results			
4.1	**Requirement:** The evaluator shared sufficient methodological details to allow for independent review of the results.	*Yes/No*	
4.2	**Recommendation:** The system provider granted upfront any necessary rights to disclose results to the evaluation's intended audiences.	*Yes/No*	
4.3	**Recommendation:** The intended audiences for an evaluation were not narrowed based on the results.	*Yes/No*	
4.4	**Recommendation:** The system provider did not misrepresent the evaluation's findings.	*Yes/No*	
4.5	**Recommendation:** The system provider or other external parties did not unduly delay the evaluation from being disclosed based on the content of the results.	*Yes/No*	
4.6	**Requirement:** The system provider did not have authority to redact results to conceal concerning findings.	*Yes/No*	

Condition		Fulfilled?	Notes/Evidence
4.7	**Recommendation:** The evaluator clearly disclosed the redaction authorities granted to the system provider or other external entities.	*Yes/No*	
5: Protection of Sensitive Information			
5.1	**Requirement:** The evaluator had the system provider's permission to release any results based on non-public information or systems.	*Yes/No*	
5.2	**Recommendation:** The evaluation methods were not gamed or leaked.	*Yes/No*	
5.3	**Requirement:** The evaluator implemented measures to protect any confidential information they received from the system provider.	*Yes/No*	
5.4	**Requirement:** The evaluator established and followed a responsible disclosure policy.	*Yes/No*	

Appendix B Terminology

For the purposes of this document, these terms are defined as follows:

Conflict of interest A situation that creates a risk that evaluator, its employees, or other agents working on its behalf to carry out an evaluation of a partiuclar system could be influenced by an incentive to not provide as rigorous and accurate an evaluation as possible.

Evaluator An organization that conducts assessments of an AI system provider's systems or processes.

System provider An organization which develops or operates AI models or systems.

System vulnerability A weakness in a system that could be exploited by a malicious actor to harm the confidentiality, integrity, or availability of the system and its data, or which could be used to bypass safeguards intended to constrain the system's behavior to prevent misuse.

Information hazard Information about the misuse of AI systems or the general pursuit of harmful acts that, if disclosed, would foreseeably cause more harm than benefit, such as specific details of how to produce dangerous chemical, biological, radiological, or nuclear (CBRN) materials or child sexual abuse material (CSAM). Relevant factors in determining whether disclosure would cause more harm than benefit include:

1. Whether broader knowledge of the issue is more likely to help people protect themselves, or to help malicious actors cause harm. For example, knowledge that a system's behavior is harmful to mental health is generally much more likely to allow consumers to avoid unsafe systems than to aid malicious actors, whereas knowledge about how a particular system can aid criminal activity is more likely to attract malicious actors to use that system.
2. Whether the relevant issue can be remediated, either by direct system improvements or broader interventions. For example, a system's security flaws can almost always be remediated either by improvements to the system, or else by encouraging users to prefer more secure alternatives. Conversely, there is less that can be done to prevent users from accessing an open-weight AI model that is useful for making CSAM after it has been released.

Third-party A separate organization that is not under the control of the system provider.