CENTER FOR DEMOCRACY & TECHNOLOGY

Re: VRT Evaluation Criteria Recommendations

The Center for Democracy & Technology (CDT) appreciates the opportunity to provide input for the development of the rubric for the voluntary reporting template (VRT) for demonstrating compliance with the Generative AI Profile of the NIST AI Risk Management Framework. CDT is a nonprofit 501(c)(3) organization that works to advance civil rights and civil liberties in the digital age. Among our priorities, CDT advocates for the responsible and equitable design, deployment, and use of new technologies such as artificial intelligence (AI), and promotes the adoption of robust, technically-informed solutions for the effective regulation and governance of AI systems. CDT is an active member of NIST's AI Safety Institute Consortium.

In developing a proposed rubric for the VRT, we emphasize that it would be insufficient for the VRT to have organizations merely indicate whether they have completed each action in the Generative AI Profile. For many of the actions in the Profile, there is a wide range of practices that could demonstrate that an action has been completed, but these practices differ significantly in how robustly and effectively they address the risk(s) that motivated the suggested action. In order for deployers and end-users to meaningfully make use of the evidence provided via the VRT, they must be able to gauge whether a developer's practices have addressed risks robustly and effectively enough for their needs. As such, for the VRT to be meaningfully helpful, the rubric should evaluate proposals' ability to indicate not just *whether* a developer has completed a given recommended action, but *how* or *to what extent* that developer has done so.

With this in mind, we offer a "stoplight"-based framework to assess developers' evidence of compliance with the Profile. For each action, we suggest that the evidence a developer provides can be placed into one of the following three categories:

- **Green**: The developer's provided evidence demonstrates that they have completed the action via practices that are adequately thorough, robust, and effective.
- **Yellow**: The developer's provided evidence demonstrates that they have completed the action; however, the evidence does not demonstrate that they have completed the action via practices that are adequately thorough, robust, or effective.
- **Red**: The developer's provided evidence fails to demonstrate that they have completed the action.

In order to illustrate how our approach can be used to assess an organization's practices against the GenAI Profile, we provide examples below of the kinds of evidence that would fall into each of these categories for a subset of actions drawn from the Profile. These examples illustrate the considerations that we hope will inform the full VRT rubric. For convenience, we divide these examples into three categories, corresponding to the high-level goals that the actions in question contribute to.

We recommend that these examples be integrated into the VRT rubric and communicated in related materials, so that developers can use effectively use organizations' disclosure to assess their practices and the evidence they provide, and so that reviewers can use it to determine

whether the organization's self-assessment is substantiated by the provided evidence. Along with this document, we have also provided these examples in spreadsheet format, in order to suggest how this framework might be incorporated into the templates that we understand other members of the working group to be developing.

Example Rubric for Evidencing Actions

Transparency & Documentation

GV-1.2-001: Establish transparency policies and processes for documenting the origin and history of training data and generated data for GAI applications to advance digital content transparency, while balancing the proprietary nature of training approaches.

- **Green:** The organization specifies the public datasets used (e.g., CommonCrawl). If they collected additional data beyond publicly available datasets, they explain the collection process and outline the contents in a reasonably specific manner, and they describe the process used to respect robots.txt and other mechanisms for opting out of being included in training data. They detail the methods used to exclude, process, and filter the data, providing justifications for these choices. They provide a high-level breakdown of their training data by category (e.g., by language). If the organization uses synthetic data, they provide a description of how this data was generated.
- **Yellow:** The organization gives broad descriptions of their data sources (e.g., "from public sources," "a mix of public and private data") without offering specific details. If they mention data processing techniques, they do so in general terms and lack detailed explanations.
- **Red:** The organization does not describe their training data or training data processing methods.

MG-3.2-002: Document how pre-trained models have been adapted (e.g., fine-tuned, or retrieval-augmented generation) for the specific generative task, including any data augmentations, parameter adjustments, or other modifications. Access to un-tuned (baseline) models supports debugging the relative influence of the pre-trained weights compared to the fine-tuned model weights or other system updates.

- **Green:** If the model has been fine-tuned, the organization provides a detailed explanation of how the fine-tuning data, including prompts, completions, and preference data, as applicable, were collected. If people wrote the completions, the organization specifies how these contributors were sourced and provides the instructions they received. If the completions were generated using a language model, the organization describes the process for doing so, including whether any content or quality filters were applied and how synthetically generated completions were reviewed, as applicable. The organization outlines the distribution of prompts in their fine-tuning datasets, ideally listing the tasks or topics covered and indicating the proportion of the dataset devoted to each. They describe the technical methods used for fine-tuning, mentioning any intensive or unconventional approaches. If the model has been given access to external tools (e.g., a retrieval-augmented generation database), the organization specifies these tools.
- **Yellow:** The organization notes which adaptation strategies they used (e.g., supervised fine tuning, reinforcement learning with human feedback), but does not justify their choices or explain these techniques in detail.
- **Red:** The organization does not describe their adaptation methods.

Consulting with Relevant Stakeholders

GV-1.3-004: Obtain input from stakeholder communities to identify unacceptable use, in accordance with activities in the AI RMF Map function.

- **Green:** The organization demonstrates that it has considered and consulted a diverse range of stakeholders, including direct stakeholders (those who use, develop, or work with the system), indirect stakeholders (those affected by the system but not directly using it), and excluded stakeholders (those unable to use the system), commensurate with the organization's resources. The consultation process is interactive, involving genuine dialogue rather than simply soliciting narrow forms of feedback from impacted stakeholders. These interactions are non-extractive, ensuring stakeholders have a meaningful opportunity to influence the technologies that affect them. Individuals who are consulted are given context on the purpose of their input and the nature of the system they are being consulted about. Organizations that are consulted are given credible ex ante assurances that their input will be acted upon. Consulted individuals and organizations are given the opportunity to provide specific, granular input.
- **Yellow:** The organization limits its consultations to stakeholders who are directly impacted by the product, such as users or clients, and further restricts these consultations to those within a narrow geographic area or within constrained demographic groups. Instead of exploring concerns interactively, the organization presents stakeholders with a narrow set of predefined issues to endorse, offering no opportunity for open input, or only gives stakeholders the opportunity to comment on high-level, vague principles or goals.
- **Red:** Absent, or the organization solely leverages large language models (LLMs) as sock puppets to simulate stakeholder engagement.

Evaluation & Assessment

MS-2.6-002: Assess existence or levels of harmful bias, intellectual property infringement, data privacy violations, obscenity, extremism, violence, or CBRN information in system training data.

- **Green:** The organization demonstrates that it has assessed most or all of the referenced categories and provides a reasonable explanation for any category it has not evaluated. It specifies the precise evaluations used, the methods employed in the assessment, and the results obtained. Where there is a justifiable reason not to disclose precise results, the organization describes in detail how it conducted these assessments and discloses its findings with an appropriate level of specificity. When assessing the presence of potentially harmful information, the organization evaluates "helpful-only" versions of the model under assessment (i.e., versions that have not been fine-tuned to refuse harmful prompts). When appropriate, the organization uses evaluation methods that involve fine-tuning the model on domain-specific data. The organization devotes an appropriate amount of computing resources and technical expertise to conducting these assessments, commensurate with the organization's resources, potentially including the development of novel assessment datasets or methods.
- **Yellow:** The organization evaluates only a limited subset of the categories. It acknowledges conducting assessments for some or all categories, but fails to provide detailed descriptions of the evaluation methods or results. The organization alleges that high scores on widely-used harm benchmarks or ordinary users' inability to elicit harms indicate that the system is incapable of harm of the type being assessed for.
- **Red:** The organization provides no evidence of evaluation for the categories.

MS-2.13-001: Create measurement error models for pre-deployment metrics to demonstrate construct validity for each metric (i.e., does the metric effectively operationalize the desired concept): Measure or estimate, and document, biases or statistical variance in applied metrics or structured human feedback

processes; Leverage domain expertise when modeling complex societal constructs such as hateful content.

- **Green:** The organization provides a clear definition of each variable they intend to measure, including both its conceptual and operationalized forms. They explain how the operationalized definition effectively captures the conceptual idea, drawing on relevant literature or expert input, and acknowledge any potential gaps between the two. To enhance measurement robustness, they use multiple metrics for the same conceptual variable and carefully analyze their similarities and differences. They perform reliability estimation to understand how variance might manifest in real-world situations.
- **Yellow:** The organization outlines the benchmarks they have used but fails to justify their selection or explore the limitations of these measures. The organization offers operationalized definitions or mathematical formulations for the metrics, but does not clarify how these choices align with the intended measurement goals.
- **Red:** The organization does not explain or justify their measurement approach.

MS-2.11-002: Conduct fairness assessments to measure systemic bias. Measure GAI system performance across demographic groups and subgroups, addressing both quality of service and any allocation of services and resources. Quantify harms using: field testing with sub-group populations to determine likelihood of exposure to generated content exhibiting harmful bias, AI red-teaming with counterfactual and low-context (e.g., "leader," "bad guys") prompts. For ML pipelines or business processes with categorical or numeric outcomes that rely on GAI, apply general fairness metrics (e.g., demographic parity, equalized odds, equal opportunity, statistical hypothesis tests), to the pipeline or business outcome where appropriate; Custom, context-specific metrics developed in collaboration with domain experts and affected communities; Measurements of the prevalence of denigration in generated content in deployment (e.g., sub-sampling a fraction of traffic and manually annotating denigrating content).

- **Green:** The organization measures a robust range of subgroups (extending beyond those legally mandated) based on documented engagement with affected communities and across various metrics. It provides a clear and well-founded explanation of why these metrics appropriately represent bias or fairness within the context of their model or product. The organization employs multiple forms of assessment, including red teaming exercises involving affected individuals, and responds substantively to the findings from these exercises and fairness evaluations.
- **Yellow:** The organization disaggregates measurements by only a single or limited set of groups and relies on a single metric to assess fairness or bias, without attempting to identify more relevant indicators of potential harmful bias in their system's context. Furthermore, the organization fails to engage with affected communities to inform this effort.
- **Red:** The organization does not attempt to measure fairness or bias.

MS-3.3-001: Conduct impact assessments on how AI-generated content might affect different social, economic, and cultural groups.

- **Green:** The organization's impact assessments draw on relevant literature, input from expert organizations, and conversations with affected communities. These assessments identify potential risks, detail how they have been or will be mitigated or monitored, and transparently acknowledge any risks that remain unaddressed.
- **Yellow:** The organization superficially analyzes potential impacts and dismisses the risks as inconsequential without justification. The organization theorizes non-empirically about potential impacts, but fails to verify these theories via relevant literature or quantitative or qualitative data. The organization notes some risks, but fails to take clear steps to address them or offer recommendations for system deployers on how to manage these risks.
- **Red:** The organization did not conduct an impact assessment.

MP-1.2-002: Verify that data or benchmarks used in risk measurement, and users, participants, or subjects involved in structured GAI public feedback exercises are representative of diverse in-context user populations.

- **Green:** When relevant, the organization compares benchmark data with real-world usage data, whether public or internal, to support any conclusions about real-world users drawn from the benchmark. They carefully consider, given their domain of application, which user populations should be included in the measurements and at what proportions. The organization ensures that these populations are directly represented in the data rather than relying on proxies such as geographic location. The organization either verifies that the system under assessment has not been exposed to benchmark data during training or qualifies their results by explicitly noting the possibility of benchmark contamination.
- **Yellow:** The organization assumes that large benchmark datasets are inherently diverse, and does not evaluate this assumption explicitly. They also rely on the assumption that one type of diversity, such as geographic diversity, will accurately represent other relevant forms of diversity. They use proxy measures for assessing diversity, without considering the potential weaknesses of this approach.
- **Red:** The organization does not consider the representativeness of evaluation data.