Public Comment on NIST AI 800-2 Initial Public Draft

Practices for Automated Benchmark Evaluations of Language Models

Submitted by: Empty Set LLC Contact: Brice Love, Co-Founder Email: brice@emptysetllc.com

Date: March 2026 **Submitted to:** ai800-2@nist.gov **Subject:** NIST AI 800-2 ipd **Prior Submissions:** NIST-2025-0035 RFI (March 6, 2026, Tracking: mmf-286f-aneg); NCCoE Concept Paper on AI Agent Identity and Authorization (March 14, 2026)

I. Summary

Empty Set LLC respectfully submits this comment on NIST AI 800-2 ipd, "Practices for Automated Benchmark Evaluations of Language Models." We previously submitted comments to CAISI on the Security Considerations for Artificial Intelligence Agents RFI (March 6, 2026) and the NCCoE Concept Paper on AI Agent Identity and Authorization (March 14, 2026). Both submissions argued that behavioral safety constitutes a distinct and unaddressed domain requiring independent evaluation, credentialing, and integration with authorization and insurance infrastructure.

NIST AI 800-2 provides a strong and timely framework for automated benchmark evaluation. The nine practices are well-structured, the agent evaluation coverage is substantive, and the document's candid acknowledgment that "automated benchmark evaluations cannot meet all AI evaluation objectives" reflects appropriate epistemic discipline.

We offer five observations where the document's framework could be strengthened, particularly as evaluation practices increasingly inform downstream decisions beyond internal development—including third-party credentialing, regulatory compliance, and insurance underwriting.

II. Comments

1. Behavioral Safety as a Distinct Evaluation Domain

Observation: The document addresses capabilities evaluation, security benchmarks (AgentDojo, CVE-Bench), and robustness testing, but does not recognize behavioral safety as a formalized evaluation domain.

Why this matters: An AI agent can pass every capabilities benchmark and security evaluation while still exhibiting unsafe behavioral patterns—escalating crisis interactions instead of deferring to human oversight, operating outside its delegated authority scope, or failing to disclose its non-human status in consumer-facing contexts. These are not capability failures or security vulnerabilities. They are behavioral failures that exist on an independent axis.

The judicial record confirms this distinction. In *Garcia v. Character Technologies, Inc.* (M.D. Fla., May 2025), the court found design defects not in an agent's technical capabilities but in its behavioral parameters—absence of crisis escalation, anthropomorphic manipulation, and engagement-maximizing design. In *Gavalas v. Google LLC* (N.D. Cal., filed March 2026), the allegations center

on a frontier model that was technically secure and functionally capable but behaviorally unsafe in the specific context of user distress.

Recommendation: Practice 1.1 (Define Evaluation Objectives) should acknowledge behavioral safety as a distinct evaluation objective, separate from capabilities and security. This would prompt evaluators to consider whether their benchmark selection (Practice 1.2) adequately covers the behavioral dimension, or whether supplementary non-automated evaluation methods are needed.

2. Evaluation for Third-Party Credentialing and Certification

Observation: The document identifies its target audience as "technical staff at organizations evaluating AI systems, including AI deployers, developers, and third-party evaluators." However, the practices are framed exclusively around internal evaluation and voluntary assessment. The document does not address how these practices apply when evaluation results are used for downstream credentialing, certification, or conformance determinations.

Why this matters: Third-party evaluation for certification purposes introduces requirements that internal evaluation does not:

- **Reproducibility under adversarial conditions.** When evaluation results determine whether an agent receives a credential, there is an incentive for model developers to optimize for benchmark performance rather than genuine capability. Certification-grade evaluation must account for this dynamic.
- **Defined pass/fail thresholds.** Internal evaluation can tolerate ambiguous results. Credentialing requires defensible thresholds that translate evaluation data into binary or tiered determinations.
- **Chain of custody for evaluation artifacts.** When evaluation results inform third-party decisions (coverage, deployment authorization, regulatory compliance), the provenance and integrity of evaluation data become critical. Practice 3.2 (Share Details of Evaluation) addresses transparency but not evidentiary integrity.
- **Evaluator independence.** The document does not distinguish between self-evaluation and independent third-party evaluation—a distinction that is essential for credentialing legitimacy.

Recommendation: The document should include a discussion—even if brief—of how the nine practices apply differently when evaluation results are consumed for credentialing or certification purposes. This need not prescribe specific certification frameworks, but acknowledging the use case would help third-party evaluators (an explicitly identified audience) apply the practices correctly.

3. Continuous and Periodic Re-Evaluation

Observation: The framework treats evaluation as a point-in-time activity. An evaluator defines objectives, selects benchmarks, runs evaluations, and reports results. There is no discussion of temporal evaluation dynamics—how often evaluations should be repeated, what triggers re-evaluation, or how results change over time as models are updated, fine-tuned, or deployed in new contexts.

Why this matters: Language models and AI agents are not static artifacts. Models receive updates, fine-tuning, and post-deployment alignment modifications. Agent scaffolding evolves. The operational environment changes. An evaluation conducted at deployment may not reflect the system's behavior six months later.

For applications where evaluation results inform ongoing decisions—insurance coverage renewals,

regulatory compliance maintenance, or credential validity—the absence of guidance on evaluation cadence and re-evaluation triggers is a meaningful gap.

Recommendation: Consider adding guidance on: - Triggers for re-evaluation (model updates, fine-tuning, scaffolding changes, significant shifts in operational context) - Evaluation cadence for high-stakes applications - How to design evaluation protocols (Practice 2.1) that are efficient for repeated execution, not just initial assessment

4. Downstream Consumers of Evaluation Results

Observation: Practice 3.3 (Report Qualified Claims) provides valuable guidance on how to communicate evaluation results accurately. However, it focuses on communication among technical audiences—researchers, developers, and evaluators. The document does not address how evaluation results are consumed by non-technical stakeholders who increasingly rely on them for consequential decisions.

Why this matters: Evaluation results are entering decision chains that extend well beyond AI development teams:

- **Insurance underwriters** are using evaluation data to price AI operator liability coverage. Armilla AI requires independent third-party verification; Relm Insurance tiers coverage based on verified governance structures. The statistical uncertainty discussed in Practice 3.1 has different significance when it informs a coverage decision than when it informs a development iteration.
- **Procurement officers** are incorporating AI evaluation results into vendor selection criteria. Federal agencies evaluating AI systems under OMB M-24-10 need evaluation results in formats compatible with their risk management processes.
- **State regulators** in jurisdictions with AI oversight mandates (Colorado SB 24-205, California SB 243) may accept evaluation results as evidence of compliance—if those results are reported in a manner that maps to statutory requirements.

Recommendation: Practice 3.3 should acknowledge that qualified claims may be consumed by non-technical stakeholders and suggest that reporting formats account for these audiences. This could include guidance on translating statistical results into risk-relevant summaries, mapping evaluation outcomes to regulatory or compliance frameworks, and clearly communicating the scope and limitations of evaluation results to stakeholders who may not understand confidence intervals or benchmark-specific caveats.

5. Commendation: Agent Evaluation Coverage

Observation: The document's treatment of agent evaluation—including agent scaffolding as an evaluation protocol setting, agent budget controls (tokens, time, cost), stopping conditions for open-ended agent tasks, and agent-specific benchmarks (AgentDojo, CVE-Bench)—is thorough and timely.

This coverage is especially valuable because the autonomous AI agent ecosystem is evolving faster than evaluation methodology. Many organizations are deploying agents without established evaluation practices. NIST AI 800-2's guidance on how to design, bound, and interpret agent evaluations fills a genuine need.

We note that as agent evaluation standards mature, they will form the methodological foundation

for third-party behavioral safety credentialing—the practice of independently verifying that an AI agent meets defined behavioral standards before it is authorized for deployment. The evaluation practices described in this document are directly applicable to that emerging use case.

III. Closing

NIST AI 800-2 is a strong foundation. The observations above are intended to extend its utility to the growing set of stakeholders—beyond AI developers—who rely on evaluation results for consequential decisions. We appreciate the opportunity to comment and CAISI's continued leadership in establishing rigorous, practical AI evaluation standards.

Empty Set LLC remains available to support CAISI's work on AI evaluation, behavioral safety, and credentialing standards.

Contact: Brice Love, Co-Founder Empty Set LLC brice@emptysetllc.com