Evaluation Integrity for AI Systems

Md Mahadi Hassan

Research Vision

A benchmark score is valuable only when success on the benchmark reflects the underlying capability it was designed to test. For modern language models, that connection is increasingly difficult to guarantee. As models are trained on internet-scale corpora, benchmarks that are published online can enter model's future training data. A model may therefore perform well because it has encountered the evaluation before not because it can solve the underlying task. At the same time, once an evaluation becomes a known target, model developers have an incentive to optimize performance on the evaluation itself. But even when a score is not affected by prior exposure or benchmark-specific optimization, the evaluation itself may capture only a narrow part of the intended capability. Success on isolated tasks may tell us little about how a model would handle complex, evolving situation where interaction with other actors and accounting for the consequences of earlier actions as additional constraint. Together, these forces weaken the connection between benchmark performance and evidence of underlying capability. This problem now extends beyond research practice as regulators, auditors, and organizations increasingly use benchmark results to assess and compare AI systems.

I frame this challenge as **evaluation integrity**, which has two complementary requirements: **measurement validity**, whether an evaluation captures the capability it intends to measure, and **attribution integrity**, whether observed success reflects that capability rather than familiarity with or optimization for the evaluation. When a benchmark is still private, I study how to estimate a model's performance on it without exposing the benchmark to the model provider, so the benchmark can remain unexposed throughout the evaluation process. When a benchmark is already exposed, I study how to determine whether exposure has inflated the measured performance and if so, how we can recover the benchmark's ability to distinguish generalizable capability from recall. Across both conditions, I study how to preserve evaluation integrity over repeated use as models, training data, and sources of exposure continue to change. Beyond preserving the integrity of existing benchmarks, I want to rethink the evaluation task itself as a generative task. Rather than relying on a set of representative tasks, I want to develop methods for generating evolving worlds that captures the essential mechanism form real world that are needed to measure a particular capability.

Research Foundations

Grounding Evaluation in Observed Behavior. My research on evaluation grew out of studying state reconciliation in infrastructure-as-code (IaC): whether automation moves a computing environment from its current state to the intended one. In an empirical study of 5,110 state-reconciliation defects, I used qualitative analysis to develop a defect taxonomy and derived heuristics for LLM prompts that uncovered **nine previously unknown defects** in the Ansible compiler (FSE 2024) [1]. The lesson extended beyond IaC: correctness should be established from the state a system actually produces, not from the apparent plausibility of the artifact intended to produce it.

That lesson became relevant to AI evaluation as LLMs began to be used for code generation and automation, where plausible code offers no guarantee of the intended system behavior. Existing IT automation benchmarks did not adequately verify whether generated solutions produced the desired state. I therefore led the development of ExITBench, **an execution-based benchmark** that evaluates LLMs on real-world IT automation tasks by executing their generated solutions in controlled environments and verifying the resulting system state (Findings of ACL 2026) [2]. Even with ten attempts, the strongest model we evaluated succeeded on only 23.9% of tasks. Rather than superficial coding mistakes, main failures concentrated around state-reconciliation reasoning (42.1%) and specialized knowledge of automation modules (26.2%).

These projects established a foundation of my research on evaluation integrity: capability claims should be grounded in the behavior that matters, not plausible outputs. But observing the right behavior is only the first step; it does not yet tell us whether success reflects capability rather than prior exposure.

When Benchmark Scores Can Be Trusted. Execution can verify the intended outcome, but even a correct result may provide only a partial view of the capability we think we are measuring. In collaborative work on program

reasoning (ACL 2026), we showed that success on one form of reasoning does not reliably generalize to another [3]. The gap persisted with larger models and reasoning-focused post-training, suggesting that scale alone does not make these forms of reasoning interchangeable. This revealed that even a valid benchmark score may support only a narrow capability claim. But even that claim assumes that success reflects the model's underlying ability rather than prior exposure to the benchmark. This led me to ask: *How can we estimate a black-box model's performance on a private benchmark without exposing the test cases whose privacy we want to preserve?*

I addressed this question through developing the **Private Evaluation Protocol (PEP)** (AACL-IJCNLP 2026), which estimates performance on a private benchmark from limited interaction with a black-box model [4]. On structured benchmarks, PEP estimated target-model accuracy within 5.88 percentage points on average while keeping the private remainder hidden from the model provider. Reasoning-heavy tasks were more challenging, suggesting that limited observations may not always capture enough of a model's behavior to predict its performance on unseen examples. PEP therefore established that private evaluation is possible without surrendering the benchmark, but also revealed the central tension that remains: protecting the benchmark limits how much behavior we can observe, while reliable evaluation depends on observing enough.

Future Research Agenda

My prior work shows that reliable evaluation requires more than observing whether a model succeeds: we must also understand what that success establishes and whether it can still be attributed to the capability under test. My research will maintain evaluation integrity across the full life-cycle of a benchmark, from preserving private evaluations to recovering exposed ones and maintaining both over time. Beyond fixed test cases, I will investigate how to preserve evaluation integrity when evaluations are generated rather than predefined. I will develop methods to construct evolving worlds that capture the mechanisms needed to measure a specific capability, while limiting opportunities for models to succeed through memorization or optimization against the environment generator.

Evaluation Without Exposure. PEP showed that behavior observed on accessible examples can predict performance on hidden ones. My future work asks a broader question: *under what conditions does behavior observed on accessible tasks provide reliable evidence about behavior on a protected evaluation?* One challenge is **what behavior to compare**: for generation, code, tool use, and agentic tasks, matching answers is insufficient, so I will develop task-specific measures of behavioral agreement. Another is **which behavior to observe**: rather than reveal examples at random, I will identify the interactions that are most informative about a target model and use them to build higher-fidelity surrogates. A more ambitious question is **how far that evidence can transfer**. I will map benchmarks by the skills and task structures they require and test whether behavior on nearby public benchmarks can predict performance on a protected benchmark or strengthen its surrogate. When such transfer is weak, targeted interaction with the protected benchmark remains necessary.

Diagnosing and Repairing Exposed Benchmarks. For an exposed benchmark, the central question is not merely whether the model saw the evaluation, but whether exposure changed its behavior and made the score overstate its capability. The same exposure may leave different behavioral signatures across tasks and modalities, and exposure need not create a measurable advantage. Signals such as likelihood, accuracy, confidence, and output diversity may vary in meaning across settings, making it difficult to distinguish familiarity from task difficulty, calibration, or ordinary generalization. I will identify which signals indicate exposure-driven advantages and which remain informative across benchmarks. Repair raises a parallel question: *how should an exposed benchmark be transformed when the effects of exposure differ across evaluation modalities?* A transformation that preserves meaning in text may have no meaningful analogue in code or vision. I will study how to preserve meaning while removing exposure-driven advantages across modalities, whether successful mitigation strategies transfer, and where they fail. Most importantly, I will test whether the modality-specific signatures identified during diagnosis explain which repair strategies succeed.

Maintaining Benchmark Integrity Over Time. Evaluation integrity must survive more than a single model evaluation. My central question is: *how should an evaluation adapt as the models it measures continue to change without giving up the integrity it was designed to preserve?* For a private benchmark, each new model

may be harder to approximate with the information already revealed. The challenge is to maintain surrogate fidelity and estimation accuracy across successive models while revealing as little additional benchmark data as possible. Because every newly revealed example is an irreversible loss of privacy, I want to understand when existing evidence is sufficient, when additional revelation becomes necessary, and how to minimize the cumulative information released over the benchmark's lifetime.

A repaired benchmark faces a different maintenance problem. As it is reused and new models are trained, familiarity with the repaired evaluation can accumulate again. The challenge is to recognize when its diagnostic value has begun to erode and determine whether further repair is sufficient or replacement is necessary. I plan to connect these two challenges through **state reconciliation**: characterize the benchmark's current integrity state, compare it with the state required for reliable evaluation, and intervene only when that gap becomes meaningful. This raises the longer-term question of whether benchmark integrity can be maintained indefinitely under repeated evaluation.

Evaluating AI in Evolving Worlds. Most AI benchmarks evaluate complex capabilities through predefined tasks and expected responses. While useful, these evaluations often provide only a shallow proxy for how models behave in real-world situations, where decisions depend on changing conditions, interactions with other actors, and consequences of earlier actions. Therefore, I envision a shift toward treating evaluation itself as a generative task: *constructing evolving simulated worlds designed to measure specific capabilities*. These worlds would contain configurable actors, beliefs, objectives, constraints, and environmental dynamics, allowing us to generate fundamentally different situations through controlled parametric changes. The goal is not merely to produce varied situations, but to ensure that success reflects the intended capability rather than familiarity with the generator. Evaluation would then examine how models reason, interact, make decisions, and adapt as situations evolve.

The central challenge is not reproducing the entire real world, but understanding what makes a simulated world sufficient to measure a particular capability. Given a capability we want to evaluate, which environmental mechanisms of real world situations must be preserved, and which can be simplified? *Can we systematically identify the smallest set of mechanisms that preserves the validity of a capability measurement? How can we generate controlled variations that distinguish genuine capability from memorization or optimization of the generation process?* Most importantly, *how can we establish that performance in these generated worlds provides reliable evidence of capability beyond the simulation?*

Building the Research Program

Turning this vision into a research program requires both general evaluation methods and knowledge of the domains being studied. My group will focus on developing methods, implementing evaluation tools, and testing their assumptions across different settings. My prior work on execution-based evaluation, model behavior, and benchmark integrity provides a foundation for this effort. My prior collaborations span learning sciences, ECE, and mathematics, giving me a foundation for working across disciplinary boundaries. I will build on this experience through collaborations with researchers in software systems, social sciences, and other fields to ground evaluations in the behavior and constraints allowing my group to pursue general evaluation methods. I will pursue **NSF CISE Future CoRe** funding for foundational evaluation methods and **SaTC** support where security and trustworthiness are central, while developing an **NSF CAREER** proposal around *reliable AI capability measurement*. The program will involve students in method development, controlled evaluation experiments, and interdisciplinary studies.

References

[1] M. M. Hassan, J. Salvador, S. K. K. Santu, and A. Rahman, "State reconciliation defects in infrastructure as code," *Proc. ACM Softw. Eng.*, vol. 1, no. FSE, Jul. 2024.

[2] M. M. Hassan, J. Salvador, A. Rahman, and S. Karmaker, "Large language models for IT automation tasks: Are we there yet?" in *Findings of the Association for Computational Linguistics: ACL 2026*, Jul. 2026, pp. 11 534–11 573.

[3] E. Hasanov, M. M. Hassan, S. Karmaker, and A. Yadavally, "The path not taken: Duality in reasoning about program execution." Association for Computational Linguistics, Jul. 2026, pp. 16 165–16 180.

[4] M. M. Hassan, S. Rahaman, Y. S. Uddin, and S. Karmaker, "Private evaluation protocol: Contamination-resistant black-box evaluation via public-slice surrogates," in *Proceedings of AACL-IJCNLP*, 2026.