CIRCLE: A Framework for Evaluating AI from a Real-World Lens

Reva Schwartz, Carina Westling, Morgan Briggs, Marzieh Fadaee, Isar Nejadgholi, Matthew Holmes, Fariza Rashid, Maya Carlyle, Afaf Taïk, Kyra Wilson, et al.

CIRCLE: A Framework for Evaluating AI from a Real-World Lens *

Reva Schwartz[1], Carina Westling[2], Morgan Briggs[3], Marzieh Fadaee[4], Isar Nejadgholi[5], Matthew Holmes[6], Fariza Rashid[7], Maya Carlyle[8], Afaf Taïk[9], Kyra Wilson[10], Peter Douglas[11], Theodora Skeadas[12], Gabriella Waters[1,13], Rumman Chowdhury[14], Thiago Lacerda[3]

[1] Civitaas Insights, [2] Bournemouth University, [3] Independent Researcher, [4] Cohere Labs, [5] National Research Council Canada, [6] Intellect Frontier, [7] University of Sydney, [8] National Physical Laboratory, UK, [9] Université de Sherbrooke, [10] University of Washington, [11] Principled AI, [12] Humane Intelligence Non-profit, [13] Virginia State University, [14] Humane Intelligence Public Benefit Corporation,

Abstract. This paper proposes CIRCLE, a six-stage, lifecycle-based framework to bridge the "reality gap" between model-centric performance metrics and AI's materialized outcomes in deployment. While existing frameworks like MLOps focus on system stability and benchmarks measure abstract capabilities, decision-makers outside the AI stack lack systematic evidence about the behavior of AI technologies under real-world user variability and constraints. CIRCLE operationalizes the "Validation" phase of TEVV (Test, Evaluation, Verification, and Validation) by formalizing the translation of stakeholder concerns outside the stack into measurable signals. Unlike participatory design which often remains localized, or algorithmic audits which are often retrospective, CIRCLE provides a structured, prospective protocol for linking context-sensitive qualitative insights to scalable quantitative metrics. By integrating methods such as field testing, red teaming, and longitudinal studies into a coordinated pipeline, CIRCLE produces "systematic knowledge"—evidence that is comparable across sites yet sensitive to local context. This can enable governance based on materialized downstream effects rather than theoretical capabilities.

1 A shift in evaluation goals

As adoption of generative AI continues to rise, policymakers, organizations, and the public are increasingly asking not just what these tools can do, but whether they can create meaningful value and work safely in the real world [1,2]. Organizations want to identify viable AI use cases in advance and convert pilot deployments into durable successes, while the public expects that AI will work as advertised and not expose them, their families, or their communities to unsafe or harmful outcomes [3,4,5,2]. AI evaluation is the applicable discipline for

*This work was conducted as part of the Forum for Real-World AI Measurement and Evaluation (FRAME) at Virginia State University's Center for Responsible AI.

answering these questions, yet the current ecosystem's focus inside the AI stack leaves it unable to accommodate the diversity of purposes and settings in which AI is being adopted and used [6]. The lifecycle described in this paper aims to address this gap by formalizing methods, tools, and processes that can produce systematic knowledge to support stakeholders beyond the AI development stack.

Like other forms of technology, AI can contribute to long-term and broad-scale impacts that reshape human behavior, organizational structures, and societal processes over time [7]. These higher-order effects underscore the importance of evaluating AI from a real-world and purpose-driven lens. The dominant evaluation ecosystem measures AI's primary effects-the immediate outputs of a model-under isolated settings. Secondary effects are the near term impacts those outputs produce in everyday life. For example, "AI psychosis" and other psychological effects associated with how users may over-rely on chatbots do not result from any single output, but from the cumulative interaction patterns between people and AI systems over time [8,9,10] .

As AI is embedded into workplace operations, secondary effects can manifest as shifts in day-to-day workflows, redefinitions of job roles, and changes in how teams coordinate and make decisions [11,12,13,14]. Over longer horizons, such changes can accumulate into tertiary effects on productivity, profitability, and overall return on investment (ROI) [15,16]. AI's ROI is currently estimated using self-reported data from surveys with executives [17]. These measures are more likely to capture executive expectations, narratives, and strategic signaling than realized operational changes, and are hard to corroborate without direct measures of deployed AI and its associated impacts [18,19]. With AI's broad-scale societal shifts largely invisible to the current evaluation ecosystem, policymakers and organizations have limited insight to inform decisions about whether and how to deploy these systems in everyday settings.

To address the gap between static model benchmarks and dynamic real-world outcomes, this paper presents CIRCLE not merely as a descriptive heuristic, but as a formal lifecycle framework that shifts the scope of AI assessment. While current evaluations often treat contextual factors as noise to be eliminated [20,21,22], CIRCLE treats it as the primary signal for understanding system behavior in the real-world. By focusing only on inside-the-stack activities and abstract model capabilities, many existing approaches both fail to support deployment stakeholders and introduce threats to core forms of measurement validity–such as ecological, construct, and consequential validity [23,24]. This can undermine the credibility of inferences about how models will behave once they are integrated into real-world contexts and workflows [20,25]. CIRCLE instead contributes a structured methodological protocol for generating systematic knowledge about what materializes in AI deployment. Specifically, it delivers:

- *An Integrated Lifecycle Model:* A six-stage architecture that unifies currently isolated evaluation activities—such as stakeholder elicitation (Stage 1), red teaming (Stage 3), and longitudinal monitoring (Stage 6)—into a coherent, iterative feedback loop. This brings a traceable pipeline to current ad-hoc

testing paradigms where every stage generates its own documented output, feeding directly into the next stage to build a continuous record of evaluation.
- *A Construct Operationalization Schema:* A specific method for translating the concepts that matter to stakeholders outside the AI stack (e.g., "over-reliance" or "cognitive offloading") into observable behavioral indicators and quantitative metrics. This process bridges the gap between qualitative socio-technical concerns and quantitative measurement and results in a context brief that serves as the foundation for subsequent stages of the evaluation lifecycle.
- *Scaffolding for Assessing Higher-Order Effects:* A formalized process and scenario set that guides evaluators through structured testing of AI technologies to capture and characterize real-world consequences. It extends measurement beyond immediate model outputs (primary effects) to systematically document downstream impacts—such as shifts in workflow, decision-making authority, or long-term skill retention (secondary and tertiary effects)..

AI Deployment and System Terms

- **AI deployment**: Phase of a project where a system is put into operation and cutover issues are resolved[26].
- **AI model capabilities**: The tasks, functions, and behaviors an AI model can reliably perform, given specified inputs and conditions.
- **AI stack**: A layered set of technologies, tools, frameworks, and infrastructure for building, deploying, and operating AI systems and applications.
- **AI system**: A machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment [27].
- **Generative AI**: A category of AI that can create new content such as text, images, videos and music [28].
- **Predictive AI**: AI systems whose primary function is to infer from input data and produce predictions or forecasts about future states, behaviors, or outcomes, typically to inform decisions or actions[27].

2 The CIRCLE framework

CIRCLE—short for *Contextualize, Identify, Represent, Compare, Learn, and Extend*— is a six-stage lifecycle framework that applies three specific methodological shifts to address specific unresolved gaps in the current evaluation ecosystem:

1. Treating real-world heterogeneity as signal instead of noise: Model-centric benchmarking typically measures abstract capability under optimized conditions, treating user heterogeneity as noise to be eliminated [29,22,30,31,32]. In contrast, CIRCLE treats the variability in how people interpret and adapt to AI—as the primary signal of operational success. This can reveal the gap between technical accuracy metrics and actual deployment outcomes.

2. Bringing a socio-technical frame to measurement: Benchmarking assesses AI via script- based in silico testing and MLOps frameworks focus on technical stability, latency, and data drift. CIRCLE evaluates AI through the lens of people interacting with systems in real workflows to capture materialized outcomes. This extends measurement towards higher-order effects—such as cognitive offloading and long-run productivity—that technical telemetry alone cannot detect.
3. Moving beyond evaluating AI "one context at a time" to aggregate knowledge across settings: While participatory design generates rich local insights, it often produces fragmented signals that are difficult to compare across sites. CIRCLE combines the contextual depth of participatory methods with the rigorous, repeatable structure of industrial testing. This allows organizations to aggregate evidence and compare disparate systems using shared constructs.

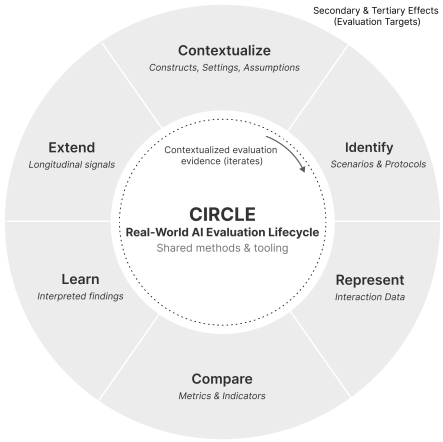

Fig. 1: The CIRCLE framework, illustrating its iterative structure and the primary outputs produced at each stage.

Figure 1 presents the overall structure of the CIRCLE lifecycle, emphasizing its iterative nature, the flow of contextualized evaluation evidence across stages, and the role of secondary and tertiary effects as evaluation targets. Each lifecycle stage in the figure also includes a brief description of its primary output.

Figure 2 shows how the lifecycle can be used to take a stakeholder-specified concern, systematize it into a construct, translate it into observable behavior, and link it to measures that can reveal higher-order effects.

CIRCLE can operationalize the "Validation" phase of the Test, Evaluation, Verification, and Validation TEVV- framework [33] by shifting the focus from whether models match prescribed criteria to aggregate descriptive reporting of what happens when people use AI in realistic conditions. Its main contribution is

a construct-centered TEVV process that a) starts from concerns and contextual factors defined by deployment-level stakeholders rather than available datasets (b) systematizes those concerns into measurable constructs rather than using convenient proxies, and (c) coordinates evaluation methods within a single, traceable pipeline that can be run in situ and at scale rather than as one-off, small-sample or purely lab-based tests. By tying every metric back to a named construct and a stakeholder-relevant outcome, CIRCLE produces evidence that is not just technically accurate, but operationally valid for decision-making.

To make each stage more concrete, this section uses an EdTech (educational technology) example in which teachers and students may over-rely on AI chatbots, which could result in shifts in reliance and instructional quality, over time. Throughout the remainder of this section, stage-specific figures highlight the primary activities at each lifecycle stage by emphasizing the focal stage and greying out the others in the sequence.

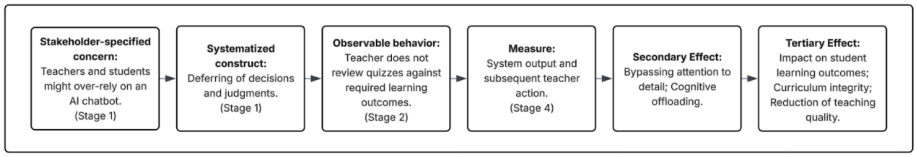

Fig. 2: Tracing a stakeholder-specified concern about over-reliance on an AI chatbot through the full lifecycle, from construct formation to observable behavior and longer-term outcomes in an EdTech classroom.

2.1 Stages of the lifecycle

The CIRCLE lifecycle structures activities to define what matters in a setting, and then designs and runs context-aware tests, interprets results for decision-makers, and feeds insights back into ongoing monitoring. Each stage produces a formal output deliverable that becomes the actionable input and "to-do" list for the next stage, creating a traceable chain of evaluation work products. Table 1 summarizes the key concepts and guiding questions associated with each stage.

A key differentiator in the CIRCLE framework is how it leverages people in a more contextually centered way than most current AI evaluations, where users are typically involved either to generate training data or rate model outputs in isolation. Instead of treating people as one-off labelers, CIRCLE engages them as situated experts on how system behavior aligns with their goals, constraints, and local norms, and pairs their accounts with interaction transcripts to reveal patterns that matter for real-world use. These human-centered collections can run alongside large-scale automated test runs so that findings generalize beyond a narrow test group and evaluation mode. Each element of CIRCLE corresponds to specific activities in a lifecycle stage, linking stakeholder-defined concepts of interest to practical evaluation methods and evidence.

Patterned on Chouldechova et al.'s shared standard for valid measurement of generative AI systems [34], the CIRCLE lifecycle shifts their core categories

Table 1: Concepts and questions at each stage of the CIRCLE framework

Framework element and Lifecycle stage	Designed to answer:
Contextualize **Context specification stage** Sets forth key concepts of interest in the deployment context from the viewpoint of those outside the AI stack. **Output Deliverable:** Context brief	What specific human, system, or interaction properties do the stakeholders expect in this setting?
Identify te**Evaluation design and planning stage** Translates stakeholder-defined concepts of interest into test-ready methods and data-collection protocols. **Output Deliverable:** Evaluation Design Plan	What forms of evidence need to be produced in order to judge whether system outcomes can match stakeholder expectations in this setting? How will that evidence be elicited, captured, and tracked over the course of the evaluation? What technical and organizational infrastructure is required to run this evaluation?
Represent **Evaluation execution stage** Executes testing to collect representative data as specified in the evaluation design and planning stage. **Output Deliverable:** Evaluation Execution Plan	How will the test subjects interact with systems to produce evidence laid out in the evaluation plan? What populations and use patterns are required for the evaluation scenarios? What infrastructure processes and resources will be used during the evaluation period?
Compare **Analysis stage** Synthesizes cross-cutting evaluation evidence to tie outcomes to concepts of interest. **Output Deliverable:** Findings Synthesis Report	Which narratives or tasks can surface the systematized constructs in a given setting, and how can those signals be captured for streamlined analysis? What kinds of observations are needed to tie materialized evaluation outcomes to the constructs of interest?
Learn **Insights stage** Enables stakeholders outside the AI stack to make sense of, prioritize, and act on evaluation outcomes. **Output Deliverable:** Stakeholder Insights Brief	Which stakeholders should be considered when writing and disseminating insights? What type of publications or platforms will optimize dissemination and engagement with implementation of the insights?
Extend **Continuous monitoring stage** Tracks shifts in AI deployment from the perspective of stakeholders beyond the AI stack. **Output Deliverable:** Continuous Monitoring Plan	Are post-evaluation controls operating as intended? Which stakeholders and/or domain experts should be involved to identify major shifts in context? What resources are required to conduct ongoing monitoring and reporting?

from Concepts to Concepts *and* Questions, Instance to Processes, Population to Who and Where, and Amounts to Outcomes. The lifecycle focuses on practical

real-world context, materialized impacts, and stakeholder input, broadening who defines evaluation questions, their scope, and what counts as evidence. The full descriptive information for each CIRCLE lifecycle stage is provided in Table 2 in the Appendix.

Real-world evaluation activities extend beyond the traditional AI stack environment to settings where people decide whether and how to leverage AI. This broader remit demands different organizational capabilities and resourcing compared to conventional model-testing. For example, organizations must be able to support stakeholder engagement, construct articulation, instrument design, and quasi-experimental testing with large and sufficiently diverse populations across various settings. The ability to conduct and sustain these activities over time may be out of reach for many individual teams acting alone, underscoring the need for shared methods, tools, and support structures that lower the barrier to participation in robust, context-sensitive evaluation.

AI Evaluation and Testing Terms

- **Benchmark / benchmarking**: Procedure, problem, or test that can be used to compare systems or components to each other or to a standard[26].
- **In silico testing**: Testing or experimentation carried out entirely on a computer, using computational models and simulations.
- **Real-world AI evaluation**: The process of accounting for what materializes when people use AI systems in practical, everyday contexts.
- **Testing, evaluation, validation, verification (TEVV)**: A framework for assessing and incorporating methods and metrics to determine that a technology or system satisfactorily meets its design specification and requirements and that it is sufficient for its intended use[33].

Stage 1: Context specification

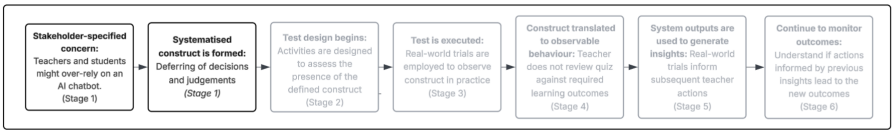

Fig. 3: Stage 1 of the lifecycle in the EdTech example: eliciting and systematizing stakeholder concerns about over-reliance on an AI chatbot, with later stages shown in grey.

AI's higher-order effects reflect how users, communities, organizations, and societies evolve as these technologies are introduced, adopted, and normalized[35,7,36,37]. Disentangling these effects and their contributing factors requires "contextual awareness", a process that helps stakeholders identify and make sense of the situational factors that shape AI's role, purpose, and impacts in their own settings[38,39].

Since situational factors are experienced differently across roles and communities, the lifecycle's first stage builds contextual awareness by eliciting what matters most to stakeholders in a given setting [40,41,42,22,43,44]. Without this process, the follow-on stages lack a grounded basis for interpreting results, making it difficult to judge whether observed outcomes meaningfully reflect the concerns, priorities, and constraints of the people affected.

Stage 1 activities center on eliciting information from key stakeholders about the AI system's purpose, the operational constraints that shape its behavior, and the stakes for those directly or indirectly affected by its deployment [38]. In the education example, school leaders, educators, and communities can jointly specify their priorities, concerns, and expectations for AI chatbot deployment, along with relevant regulatory and ethical boundaries.

Elicitation activities can include in-person workshops, interviews, and process mapping, or they can be conducted virtually through asynchronous exercises and other engagements, offering flexibility around organizational time and resource constraints. These techniques gather both explicit knowledge (documented policies, procedures, and formal requirements) and the tacit knowledge (experiential insights, practical judgment, and uncodified understanding) necessary to understand the deployment context and how organizations actually operate [45].

These qualitatively defined priorities can be augmented by automated methods, such as LLM-based review of documented curricula, policies, and community materials, to surface salient risks and considerations that stakeholders may overlook [46,47,48]. Combining human-centered elicitation with automated techniques offers a fuller picture of both tacit and explicit knowledge, aligning with work that frames AI and humans as complementary partners in knowledge work. This approach also makes it feasible to extend contextual analysis across many settings that match the prospective use case [49].

The collected information is used to build up a "construct" – or a clear, named idea that represents the specific risk, behavior, or value that the stakeholders care about, along with a description of what the construct means in practice, when it shows up, and what counts as acceptable or concerning levels. Constructs of interest in the EdTech example might include student and teacher over-reliance on chatbot outputs, cognitive offloading, and preserving human judgment in the classroom. Evaluation designers translate contextual information into constructs by judging which stakeholder concerns can be addressed in a given setting and meaningfully operationalized as measurable indicators. Figure 3 shows how Stage 1 activities focus on eliciting and systematizing stakeholder-specified concerns.

The output of this stage is a "context brief" that feeds into Stage 2, where constructs are translated into concrete evaluation parameters.

Stage 2: Evaluation design and planning In Stage 1, the stakeholders in the EdTech example raise a concern that "teachers and students might defer their own decisions and judgments to an educational chatbot." In Stage 2, this concern is formalized as a construct such as "changes in content review and deferring of

decisions and judgments," making it specific enough to guide observation and measurement.

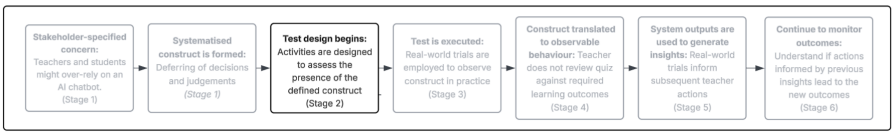

Fig. 4: Figure 4. Stage 2 of the lifecycle in the EdTech example: designing test activities to assess the presence of the defined construct.

The next step is to design evaluation methods and scenarios able to instantiate the contextual factors that stakeholders consider most relevant. This process contrasts with model-centric evaluations which wall off contextual details by tightly controlling test environments, static datasets, and canned prompts. Stakeholders beyond the stack require contextual detail to make sense of evaluation results in setting and support informed deployment decisions.

Figure 5 illustrates how evaluation methods can be selected and combined as design choices in Stage 2, highlighting tradeoffs between control and contextual richness and how those choices shape what forms of evidence can be observed.

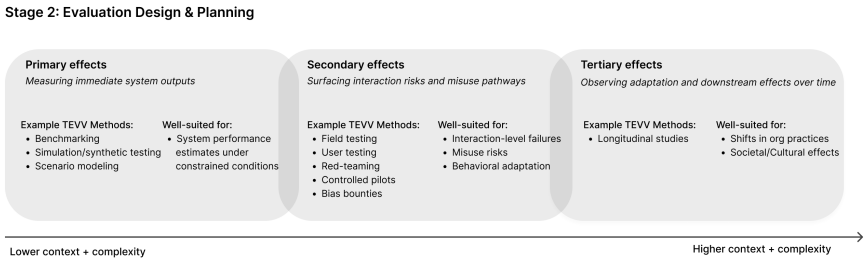

Fig. 5: Evaluation methods as design choices in Stage 2 of the CIRCLE lifecycle. Methods trade off control and contextual richness, shaping what forms of evidence and downstream effects can be observed. Regions illustrate classes of methods that may be selected and combined during this stage.

At the primary end of the continuum is AI benchmarking, which focuses on whether models produce "correct" or expected outputs and yields metrics such as accuracy, efficiency, and energy use. Methods such as usability studies begin to surface AI's secondary effects. By examining whether people can successfully navigate AI interfaces to complete tasks, these approaches help to reveal user friction, confusion, or drop-off points that simple benchmarks cannot detect.

Context-sensitive methods such as red-teaming, bias bounties, and field tests evaluate system behavior during live interactions with people under realistic

or adversarial conditions [50,51]. At the tertiary end, longitudinal studies can be combined with other real-world signals like incident tracking to monitor cumulative downstream effects that no single snapshot can capture, such as how innovation benefits for organizations, sectors, and the broader economy evolve over time.

Stage 2 activities drive decisions about which methods along this continuum are best suited to evaluate the construct and generate insights that stakeholders can apply in their own setting. This stage also defines the tasks and scenarios for capturing the required evidence from participants at scale. For example, different designs can test whether privacy guardrails hold in practice, how often high-stakes errors occur in a given setting, and whether multilingual capabilities genuinely support an intended population [52].

In parallel, large-scale automated test runs are designed to evaluate the same constructs under controlled conditions by stress-testing models on synthetic or replayed interaction logs that mirror the scenarios used in human trials. These automated runs extend coverage across edge cases and input variations that are impractical to reach solely through human studies.

Contextual Measurement Concepts

- **Construct**: An abstract, latent concept or theoretical variable, not directly observable, that is defined for scientific purposes and measured indirectly through multiple observable indicators or items.
- **Construct systematization**: The process of clarifying and organizing a concept by specifying its meaning, dimensions, and relationships to other concepts [53].
- **Construct operationalization**: Linking a systematized concept to appropriate indicators and scores in a coherent measurement framework [53].
- **Context**: The parameters in which interrelated factors, purposes, and circumstances may shape individual and collective perceptions, interpretations, and expectations about the functionality and impacts of AI technology, and resulting actions [54].

Stage 3: Evaluation execution In Stage 3, testing is carried out through real-world trials of AI systems with large, diverse participant groups so that findings can generalize across user categories and contexts [55]. It also sets clear plans for who is eligible to participate and how samples are drawn, in line with evaluation requirements and human subjects research protocols and with attention to participant incentives and expected attrition. AI's higher-order, ecosystem-level effects influence both users and non-users of AI systems. Sampling can therefore be designed to include people who do not use AI, for example through stratified recruitment and oversampling, so their behaviors are well represented. Their participation can then be up-weighted in analysis to counter adopter-centric biases in existing data and improve generalizability to groups currently underserved by chat models [56,57].

For example, achieving this in EdTech might require enrolling teachers from different schools, with different seniority levels/ experience with AI [58].

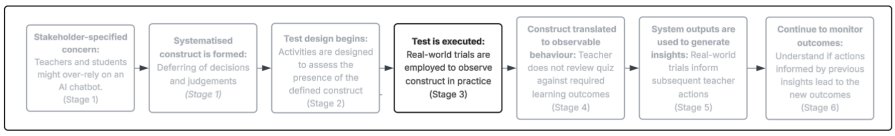

Fig. 6: Figure 6. Stage 3 of the lifecycle in the EdTech example: executing real-world trials to observe the construct in practice.

Participants are trained prior to testing to make sure they can meaningfully and safely exercise the system under evaluation, understand their role in the study, and are aware of the system's intended use and limitations. A shared code of conduct is used to inform participants of applicable ethical considerations, data collection practices, acceptable use, and how to report or escalate any harms or unexpected behavior they may experience [59]. The code of conduct also includes plain-language descriptions of core functionality, known failure modes, and key limitations. Depending on the evaluation design, testers may also be trained on how to conduct adversarial testing.

Fully automated tests can be run alongside real-world trials using an identical test harness, enabling consistent and directly comparable evaluation outcomes and helping to reveal the contextual and higher-order effects that are most consequential for stakeholders. All decisions across the lifecycle framework are tracked and documented so that evaluation outcomes can be interpreted based on testing mode, who participated and under what conditions. The hypotheses, metrics, sampling plans, and protocols defined in the design stage often need adjustments due to changes that occur in the execution stage. Typical adaptations in test execution include recruitment shortfalls with less people showing up, and technical difficulties leading to altered usage patterns or missing data. Documenting such deviations is necessary for a sound analysis in the subsequent steps or determining whether experiments need to be repeated or newer ones added.

Stage 4: Analysis Stage 4 begins with scenario design, which creates the scaffolding needed to make the constructs of interest observable and measurable [60,61]. Each construct is translated into a high-level description of in-context behaviors that can be expected to realistically occur in the deployment environment. For example, systematically observing whether a teacher reviews AI-generated quizzes for required learning outcomes allows for concrete analysis. This behavior can be measured using quiz outputs, student chatbot transcripts, and documented follow-up actions, providing data to assess both short-term and long-term effects across classrooms. Proxy scenarios can also be used to evaluate hard-to-observe

higher-order effects, such as over-reliance on chatbot engagements, without directly exposing participants to harmful content or experiences.

Scenarios also anchor the development of other analysis tools, including scoring rubrics, scoring criteria, and data markup methods. Scoring criteria are designed to connect the elicited evaluation behaviors to what actually materializes in context. Rubrics are used to consistently map raw observations onto outcome metrics that are tied to stakeholder concerns. Annotation and markup of output data can be fully automated, fully manual, or participant-driven but must use universally understood definitions for the observed behaviors to be sure that everyone applies and interprets the definitions consistently throughout the markup process. Figure 7 highlights Stage 4, translation of constructs into observable behaviors and indicators.

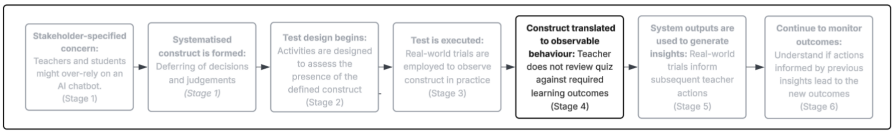

Fig. 7: Figure 7. Stage 4 of the lifecycle in the EdTech example: translating the construct into observable teacher behaviors and classroom signals.

This stage combines qualitative real-world results from classroom trials and large-scale model runs so that each informs the other in context rather than standing alone. Observed behaviors and coded patterns from contextualized student-chatbot transcripts and other collected data are compared to automated metrics from scenario-matched test suites. Because both evaluation modes draw on the same underlying constructs, the combined output measures support contextual awareness by revealing not only raw system outcomes, but the conditions under which model behavior is likely to contribute to over-reliance, cognitive offloading, or other secondary effects at scale. This contextually enriched information supports broader inferences about tertiary effects that link concrete classroom practices to longer-term shifts in learning outcomes, curricula, and teaching effectiveness.

Testing Scenario Concepts

- **Evaluation scenario**: A high-level description of a specific situation or sequence of conditions under which a system is to be tested, including the initial state, inputs, triggering events, and expected behavior or outcomes.
- **Proxy scenario**: A test scenario that does not directly reproduce the real-world context of interest but is designed to stand in for it, using more tractable or safer conditions while preserving key features believed to be relevant for evaluating system behavior or risk.

Stage 5: Insights Interpretation of data re-contextualizes and extrapolates the findings that are produced by analysis. By using matched scenarios in the education example, the data from local tests can be aggregated with large-scale model runs to enable longitudinal projection and interrogation of tertiary effects. The interpretation process can also engage participants and stakeholders in guided exploration of the data produced in local tests to re-contextualize findings and formulate insights for dissemination to relevant stakeholders and communities.

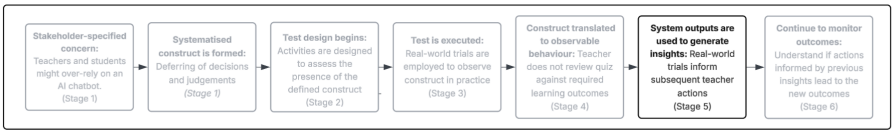

Fig. 8: Figure 8. Stage 5 of the lifecycle in the EdTech example: using observed system outputs to generate insights that inform subsequent teacher actions.

The choice of method is context-sensitive and should consider audiences, reach and impact. Concepts of interest identified in Stage 1 can continue to guide interpretation for insight generation. Where possible, continuous stakeholder engagement, from context specification through to insight generation and dissemination, enhances consistency and impactful insights.

The interpreted findings and insights can be formulated and presented to optimize reach and impact for intended audiences and platforms or channels. Figure 8 highlights how Stage 5 uses system outputs to generate insights and inform actions in the EdTech example. Care should be taken to reduce the use of highly technical language to ensure that evaluation outcomes can be easily leveraged by all relevant stakeholders. Stakeholders can assist in identifying the type of platforms and publications that are a match in community type and publication tone so materials can be optimized for reach and impact. Typically, this might include industry periodicals, newsletters, and other forms of internal and/or community-facing publications, conferences, peer-reviewed journals (academic and non-academic), and news media. It can also include formal and informal networks and community groups and organizations. With the goal of supporting stakeholders beyond the stack, insights should conform to the type and style of communication that is established within the organization or community in question. Other considerations include how to record, format, store, and disseminate findings.

Stage 6: Continuous monitoring Stage 6 establishes continuous, context-aware monitoring of deployed AI systems to detect performance drift, emerging risks, and changes in real-world usage patterns as deployment conditions evolve [62,63]. Monitoring targets are anchored in the constructs and risk hypotheses identified during context specification, while also incorporating exploratory signals to surface unanticipated shifts in system behavior or use [64]. In the education example, monitoring tracks how chatbots are actually used in evolving classroom

environments so that over-reliance, drifting learning outcomes, or misalignment with institutional or ethical guidelines can be detected and addressed early [65]. Figure 9 highlights how Stage 6 monitoring operates across multiple organizational and social layers, each capturing distinct mechanisms through which AI systems shape outcomes over time. The educational targets at the district level would focus on aggregate patterns such as policy compliance, consistency of implementation across schools, and equity impacts between schools.

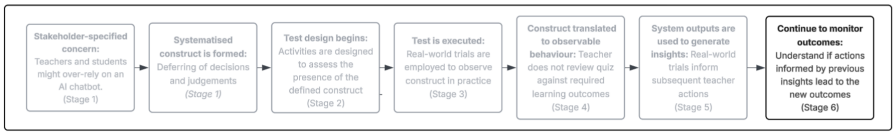

Fig. 9: Figure 9. Stage 6 of the lifecycle in the EdTech example: monitoring outcomes over time to see how actions informed by previous insights affect student learning and reliance.

At the school level, monitoring examines institutional conditions such as leadership priorities, technical infrastructure, and staff training that mediate whether AI tools are adopted, adapted, or bypassed in practice. At the classroom level, monitoring foregrounds pedagogical context, including changes in curriculum, teaching strategies, and classroom culture or demographics that influence how teachers and students interact with the system. Finally, at the individual student and educator level, monitoring assesses changes in motivation, task engagement, digital literacy, and trust calibration with respect to AI-supported decision-making.

Practically, effective monitoring in education is a hybrid process[66]. Automated systems can collect signals such as usage patterns and learning outcomes. Human and institutional review is required to interpret these signals in relation to local goals, constraints, and unintended consequences, and to determine whether corrective action, redesign, or policy updates are warranted. The goal is a recurring cycle—observe, interpret, diagnose, act—that supports continuous improvement without collapsing monitoring into continuous surveillance of individuals or real-time behavioral control [67]. Accordingly, monitoring follows principles of data minimization, proportionality, and purpose limitation [68,69], , ensuring that oversight remains legitimate, auditable, and aligned with institutional governance and privacy obligations. Findings from continuous monitoring feed back into renewed context specification and evaluation design, allowing constructs, scenarios, and metrics to be revised as deployment conditions and stakeholder priorities change.

3 Discussion and Future Work

CIRCLE reframes AI evaluation as a construct-centered, lifecycle process that links stakeholder-defined concerns to contextualized scenarios, mixed-methods

testing, and continuous monitoring. Rather than treating evaluation methods such as benchmarks, red teaming, and field trials as isolated activities, the framework integrates them into a single pipeline that ties every metric to a named construct, observable behavior, and stakeholder-relevant outcome. This structure makes it possible to collect observational data about the concepts that matter to stakeholders outside the AI stack–such as secondary and tertiary effects–, and to generate evidence that informs questions about whether and how systems should be adopted and governed in their own settings.

Implementing the CIRCLE lifecycle introduces trade-offs regarding cost and complexity. Unlike automated benchmarks, collecting context-aware information requires organizational maturity, interdisciplinary expertise, and time-intensive human-subjects protocols, and is inherently slower and more resource-intensive than running in silico test scripts. At the same time, these investments directly support stronger forms of measurement validity that are difficult to achieve with purely model-centric evaluations.

CIRCLE enhances *ecological validity* by structuring evaluations in settings that approximate everyday life (e.g., classrooms) rather than in tightly controlled, artificial test environments. It strengthens *construct validity* by explicitly eliciting stakeholder concerns in Stage 1 and naming them as constructs in Stages 2 and 4 (e.g., over-reliance, cognitive offloading, displacement in classrooms), then designing scenarios, coding schemes, and metrics that instantiate these constructs within actual workflows.

CIRCLE also supports *criterion validity* by aligning outcome measures with real-world stakeholder outcomes such as durable ROI, shifts in workload, or changes in error profiles, making evaluation metrics more reliable predictors of deployment performance. *Consequential validity* is improved by systematically tracking benefits and harms through real-world interactions across the lifecycle, rather than inferring them only from theory or in silico testing results. Finally, *internal validity* is strengthened through deliberate design choices in Stage 2, such as comparison groups and quasi-experimental designs, and through careful documentation of protocols and deviations during execution, which help minimize alternative explanations and experimental confounds.

Future efforts will need to reduce the friction of capturing contextual detail at scale, for example by developing validated user simulators and other tooling grounded in empirical lifecycle data. Adoption of contextually aware insights about higher-order effects also requires a shift in institutional incentives. Regulators will need to place value on deployment-level evidence in addition to model documentation. This may lead to enterprise investment in comprehensive measurement infrastructure that tracks value rather than just usage, and shared libraries of operationalized constructs (e.g., cognitive offloading) to make these evaluations interoperable. Without this shift, the ecosystem will remain stuck measuring abstract capabilities while missing the materialized risks of deployment.

Declaration on Generative AI

The authors used large language model tools to convert some of the references to BibTeX format, to help structure the order of concepts in the introduction, and to correct LaTeX code for formatting tables and call-out boxes. All AI-generated text and code were fully reviewed, edited, and verified by the authors prior to submission.

References

1. C. François, L. Péran, A. Bdeir, N. Dziri, W. Hawkins, Y. Jernite, S. Kapoor, J. Shen, H. Khlaaf, K. Klyman, N. Marda, M. Pellat, D. Raji, D. Siddarth, A. Skowron, J. Spisak, M. Srikumar, V. Storchan, A. Tang, and J. Weedon, "A different approach to ai safety: Proceedings from the columbia convening on openness in artificial intelligence and ai safety," 2025.
2. N. Maslej, L. Fattorini, R. Perrault, Y. Gil, V. Parli, N. Kariuki, E. Capstick, A. Reuel, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, J. C. Niebles, Y. Shoham, R. Wald, T. Walsh, A. Hamrah, L. Santarlasci, J. B. Lotufo, A. Rome, A. Shi, and S. Oak, "Artificial intelligence index report 2025," 2025.
3. United Nations Development Programme, "2025 global survey on AI and human development: Main findings." `https://hdr.undp.org/2025-global-survey-ai-and-human-development-main-findings`, 2025.
4. B. Vigers and J. Lall, "Americans prioritize AI safety and data security," *Gallup News*, 10 2025.
5. Pew Research Center, "How people around the world view AI: More are concerned than excited about its use," 10 2025.
6. Ipsos, "Conflicting global perceptions around ai present mixed signals for brands," 02 2025. Ipsos AI Monitor: Key Findings.
7. M. Selgas-Cors, "Sociotechnical transformation: A systematic review on the impact of artificial intelligence on society and organizations," *FinTech and Sustainable Innovation*, vol. 00, no. 00, pp. 1–16, 2025.
8. C. Archiwaranguprok, C. Albrecht, P. Maes, K. Karahalios, and P. Pataranutaporn, "Simulating psychological risks in human-ai interactions: Real-case informed modeling of ai-induced addiction, anorexia, depression, homicide, psychosis, and suicide," 2025.
9. M. Namvarpour, B. Brofsky, J. Medina, M. Akter, and A. Razi, "Understanding teen overreliance on ai companion chatbots through self-reported reddit narratives," 2025.
10. A. Hudon and E. Stip, "Delusional experiences emerging from AI chatbot interactions or "AI psychosis"," *JMIR Mental Health*, vol. 12, p. e85799, 2025.
11. F. Xu, J. Hou, W. Chen, and K. Xie, "Generative ai and organizational structure in the knowledge economy," 2025.
12. Q. Xiao, X. E. Hu, M. E. Whiting, A. Karunakaran, H. Shen, and H. Cao, "Ai hasn't fixed teamwork, but it shifted collaborative culture: A longitudinal study in a project-based software development organization (2023-2025)," 2025.
13. A. K. Agrawal, J. S. Gans, and A. Goldfarb, "AI Adoption and System-Wide Change," NBER Working Paper 28811, National Bureau of Economic Research, 05 2021.

14. F. Dell'Acqua, C. Ayoubi, H. Lifshitz, R. Sadun, E. Mollick, L. Mollick, Y. Han, J. Goldman, H. Nair, S. Taub, and K. Lakhani, "The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise," Tech. Rep. w33641, National Bureau of Economic Research, Cambridge, MA, Apr. 2025.
15. D. Acemoglu, "The simple macroeconomics of AI," NBER Working Paper 32487, National Bureau of Economic Research, 04 2024.
16. National Academies of Sciences, Engineering, and Medicine, "Artificial intelligence and productivity," in *Toward a Multidimensional Framework for Measuring the Health and Vitality of the Internet*, Washington, DC: The National Academies Press, 2024.
17. S. Puntoni, P. Tambe, and J. Korst, "How are companies using gen AI in 2025?." Knowledge at Wharton, 11 2025.
18. Deloitte Insights, "AI ROI: The paradox of rising investment and elusive returns." Deloitte Insights, 10 2025.
19. E. Brynjolfsson, D. Li, and L. Raymond, "Generative ai at work*," *The Quarterly Journal of Economics*, vol. 140, pp. 889–942, 02 2025.
20. O. Salaudeen, A. Reuel, A. Ahmed, S. Bedi, Z. Robertson, S. Sundar, B. Domingue, A. Wang, and S. Koyejo, "Measurement to Meaning: A Validity-Centered Framework for AI Evaluation," June 2025. arXiv:2505.10573 [cs].
21. A. M. Bean, R. O. Kearns, A. Romanou, F. S. Hafner, H. Mayne, J. Batzner, N. Foroutan, C. Schmitz, K. Korgul, H. Batra, O. Deb, E. Beharry, C. Emde, T. Foster, A. Gausen, M. Grandury, S. Han, V. Hofmann, L. Ibrahim, H. Kim, H. R. Kirk, F. Lin, G. K.-M. Liu, L. Luettgau, J. Magomere, J. Rystrøm, A. Sotnikova, Y. Yang, Y. Zhao, A. Bibi, A. Bosselut, R. Clark, A. Cohan, J. Foerster, Y. Gal, S. A. Hale, I. D. Raji, C. Summerfield, P. H. S. Torr, C. Ududec, L. Rocher, and A. Mahdi, "Measuring what matters: Construct validity in large language model benchmarks," 2025.
22. H. Wallach, M. Desai, A. F. Cooper, A. Wang, C. Atalla, S. Barocas, S. L. Blodgett, A. Chouldechova, E. Corvi, P. A. Dow, J. Garcia-Gathright, A. Olteanu, N. Pangakis, S. Reed, E. Sheng, D. Vann, J. W. Vaughan, M. Vogel, H. Washington, and A. Z. Jacobs, "Position: Evaluating generative ai systems is a social science measurement challenge," 2025.
23. W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and quasi-experimental designs for generalized causal inference*. Experimental and quasi-experimental designs for generalized causal inference, Boston, MA, US: Houghton, Mifflin and Company, 2002. Pages: xxi, 623.
24. S. Messick, "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning," *American Psychologist*, vol. 50, no. 9, pp. 741–749, 1995.
25. K. Larsen, R. Lukyanenko, R. Mueller, V. Storey, J. Parsons, D. Vander Meer, and D. Hovorka, "Validity in design science," *MIS Quarterly*, 04 2025.
26. I. O. for Standardization (ISO), I. E. C. (IEC), I. of Electrical, and E. E. (IEEE), "ISO/IEC/IEEE 24765:2017, systems and software engineering — vocabulary," 2017.
27. Organisation for Economic Co-operation and Development, "Explanatory memorandum on the updated OECD definition of an AI system," 2024.
28. Organisation for Economic Co-operation and Development, "Initial policy considerations for generative artificial intelligence," 2023.
29. R. Dobbe, T. Krendl Gilbert, and Y. Mintz, "Hard choices in artificial intelligence," *Artificial Intelligence*, vol. 300, p. 103555, 2021.

30. L. Weidinger, I. D. Raji, H. Wallach, M. Mitchell, A. Wang, O. Salaudeen, R. Bommasani, D. Ganguli, S. Koyejo, and W. Isaac, "Toward an evaluation science for generative ai systems," 2025.
31. S. Bergman, N. Marchal, J. Mellor, S. Mohamed, I. Gabriel, and W. Isaac, "Stela: a community-centred approach to norm elicitation for ai alignment," *Scientific Reports*, vol. 14, p. 6616, Mar. 2024.
32. A. Haupt and E. Brynjolfsson, "Position: Ai should not be an imitation game: Centaur evaluations," *42nd International Conference on Machine Learning*, 2025.
33. National Security Commission on Artificial Intelligence, "Final report," tech. rep., National Security Commission on Artificial Intelligence, Washington, DC, Mar. 2021. Released March 2021.
34. A. Chouldechova, C. Atalla, S. Barocas, A. F. Cooper, E. Corvi, P. A. Dow, J. Garcia-Gathright, N. Pangakis, S. Reed, E. Sheng, D. Vann, M. Vogel, H. Washington, and H. Wallach, "A shared standard for valid measurement of generative ai systems' capabilities, risks, and impacts," 2024.
35. A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," FAT* '19, (New York, NY, USA), p. 59–68, Association for Computing Machinery, 2019.
36. U. Gasser, "Second-order effects of artificial intelligence," *Issues in Science and Technology*, vol. 40, no. 2, 2024.
37. Aspen Digital, "Second and third order effects of AI," 2025.
38. R. Schwartz, R. Chowdhury, A. Kundu, H. Frase, M. Fadaee, T. David, G. Waters, A. Taik, M. Briggs, P. Hall, S. Jain, K. Yee, S. Thomas, S. Bhandari, P. Duncan, A. Thompson, M. Carlyle, Q. Lu, M. Holmes, and T. Skeadas, "Reality check: A new evaluation ecosystem is necessary to understand ai's real world effects," 2025.
39. B. J. Chen and J. Metcalf, "Explainer: A sociotechnical approach to AI policy," explainer / policy brief, Data & Society Research Institute, 05 2024.
40. L. H. Ajmani, N. A. Abdelkadir, and S. Chancellor, "Secondary stakeholders in ai: Fighting for, brokering, and navigating agency," in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, p. 1095–1107, ACM, June 2025.
41. M. I. Magaña and K. Shilton, "Frameworks, methods and shared tasks: Connecting participatory ai to trustworthy ai through a systematic review of global projects," in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, (New York, NY, USA), p. 2166–2179, Association for Computing Machinery, 2025.
42. S. Wang, N. Cooper, and M. Eby, "From human-centered to social-centered artificial intelligence: Assessing chatgpt's impact through disruptive events," *Big Data & Society*, vol. 11, no. 4, p. 20539517241290220, 2024.
43. Q. V. Liao and Z. Xiao, "Rethinking model evaluation as narrowing the socio-technical gap," 2025.
44. F. Delgado, S. Yang, M. Madaio, and Q. Yang, "The participatory turn in AI design: Theoretical foundations and the current state of practice," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*, (Boston, MA, USA), pp. 1–23, Association for Computing Machinery, 10 2023.
45. C. Gubbins and L. Dooley, "Delineating the tacit knowledge-seeking phase of knowledge sharing: The influence of relational social capital components," *Human Resource Development Quarterly*, vol. 32, no. 3, pp. 319–348, 2021.

46. M. H. Garcia, C. Couturier, D. M. Diaz, A. Mallick, A. Kyrillidis, R. Sim, V. Ruhle, and S. Rajmohan, "Exploring how llms capture and represent domain-specific knowledge," 2025.
47. F. M. Delgado-Chaves, M. J. Jennings, A. Atalaia, J. Wolff, R. Horvath, Z. M. Mamdouh, J. Baumbach, and L. Baumbach, "Transforming literature screening: The emerging role of large language models in systematic reviews," *Proceedings of the National Academy of Sciences*, vol. 122, no. 2, p. e2411962122, 2025.
48. W. Lu, R. K. Luu, and M. J. Buehler, "Fine-tuning large language models for domain adaptation," *npj Computational Materials*, vol. 11, no. 1, p. 84, 2025.
49. M. H. Jarrahi, D. Askay, A. Eshraghi, and P. Smith, "Artificial intelligence and knowledge management: A partnership between human and ai," *Business Horizons*, vol. 66, no. 1, pp. 87–99, 2023.
50. S. Majumdar, B. Pendleton, and A. Gupta, "Red teaming ai red teaming," 2025.
51. W. M. Kennedy, C. Patlak, J. Dave, B. Chambers, A. Dhanotiya, D. Ramiah, R. Schwartz, J. Hagen, A. Kundu, M. Pendharkar, L. Baisley, T. Skeadas, and R. Chowdhury, "Ask what your country can do for you: Towards a public red teaming model," 2025.
52. Infocomm Media Development Authority (IMDA) and Humane Intelligence, "Singapore AI safety red teaming challenge: Evaluation report 2025," tech. rep., Infocomm Media Development Authority, Singapore, 02 2025.
53. R. Adcock and D. Collier, "Measurement validity: A shared standard for qualitative and quantitative research," *American Political Science Review*, vol. 95, no. 3, pp. 529–546, 2001.
54. R. Schwartz, G. Waters, R. Amironesei, C. Greenberg, J. Fiscus, P. Hall, A. Jones, S. Jain, A. Godil, K. Greene, T. Jensen, and N. Schulman, "The assessing risks and impacts of ai (aria) program evaluation design document," tech. rep., National Institute of Standards and Technology, Gaithersburg, MD, 2024.
55. I. Evans, C. Porter, and M. Micallef, "Breaking tester stereotypes: Who is testing and why it matters," in *Proceedings of the 37th International BCS Human Computer Interaction Conference (BCS HCI 2024)*, (Swindon, UK), pp. 115–126, BCS Learning and Development Ltd, 2024.
56. K. Zhou, K. Gligorić, M. Cheng, M. S. Lam, V. Raman, B. Aminu, C. Woo, M. Brockman, H. Cha, and D. Jurafsky, "Attention to non-adopters," 2025.
57. O. Kraishan, "The ai invisibility effect: Understanding human-ai interaction when users don't recognize artificial intelligence," 2026.
58. S. Ritter, J. M. Kulikowich, P.-W. Lei, C. L. McGuire, and P. Morgan, "What evidence matters? a randomized field trial of cognitive tutor algebra i," in *Supporting Learning Flow Through Integrative Technologies: Proceedings of the 15th International Conference on Computers in Education (ICCE 2007)*, (Amsterdam, The Netherlands), pp. 13–20, IOS Press, 2007.
59. N. A. of Medicine, *An Artificial Intelligence Code of Conduct for Health and Medicine: Essential Guidance for Aligned Action*. Washington, DC: The National Academies Press, 2025.
60. R. Schwartz, J. Fiscus, K. Greene, G. Waters, R. Chowdhury, T. Jensen, C. Greenberg, A. Godil, R. Amironesei, P. Hall, and S. Jain, "The nist assessing risks and impacts of ai (aria) pilot evaluation plan," tech. rep., National Institute of Standards and Technology, Gaithersburg, MD, 2024.
61. J. Zhang, Y. Zhou, Y. Liu, Z. Li, and S. Hu, "Holistic automated red teaming for large language models through top-down test case generation and multi-turn interaction," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

Processing, (Miami, Florida, USA), Association for Computational Linguistics, 2024.
62. J. Feng, R. V. Phillips, I. Malenica, A. Bishara, A. E. Hubbard, L. A. Celi, and R. Pirracchio, "Clinical artificial intelligence quality improvement: towards continual monitoring and updating of ai algorithms in healthcare," *NPJ digital medicine*, vol. 5, no. 1, p. 66, 2022.
63. C. Toups, R. Bommasani, K. Creel, S. Bana, D. Jurafsky, and P. S. Liang, "Ecosystem-level analysis of deployed machine learning reveals homogeneous outcomes," *Advances in Neural Information Processing Systems*, vol. 36, pp. 51178–51201, 2023.
64. P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International conference on machine learning*, pp. 5637–5664, PMLR, 2021.
65. A. M. Al-Zahrani, "Unveiling the shadows: Beyond the hype of ai in education," *Heliyon*, vol. 10, no. 9, 2024.
66. M. Stein, J. Bernardi, and C. Dunlop, "Position: Governments need to increase and interconnect post-deployment monitoring of ai," in *NeurIPS Workshop on Socially Responsible Language Modelling Research*, 2024.
67. D. Manheim and A. Homewood, "Limits of safe ai deployment: Differentiating oversight and control," *arXiv preprint*, 2025.
68. A. J. Biega, P. Potash, H. Daumé, F. Diaz, and M. Finck, "Operationalizing the legal principle of data minimization for personalization," in *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 399–408, 2020.
69. C. Novelli, F. Casolari, A. Rotolo, M. Taddeo, and L. Floridi, "Ai risk assessment: a scenario-based, proportional methodology for the ai act," *Digital Society*, vol. 3, no. 1, p. 13, 2024.

4 Appendix

Table 2: CIRCLE Framework

Concepts and Questions	**Processes**	**Who and Where**	**Outcomes**
Context Specification			
What specific human, system, or interaction properties do the stakeholders expect in this setting?	Elicit and systematize background concepts associated with stakeholder claims; document assumptions, constraints, and success criteria.	Stakeholders implementing the system; stakeholders who may be affected; domain experts; context specification team; evaluation design team; analysis team; continuous monitoring team; real-world setting(s) where the system will be used.	Context brief listing clearly specified concepts to be evaluated (systematized constructs), relevant populations, and target parameters grounded in the implementation setting.
Evaluation Design and Planning			
What forms of evidence are required to judge whether system outcomes match stakeholder expectations in context? How will evidence be elicited, captured, and tracked? What technical and organizational infrastructure is required?	Infrastructure plan; sampling design; evaluation design strategy; construct operationalization plan including evaluation scenarios and task protocols; subject recruitment requirements; test protocol plan.	Context specification team; evaluation design team; evaluation execution team; analysis team; IRB/ethics, legal, and privacy experts.	Evaluation design plan including study design; final evaluation scenarios and test protocols; sampling plan; metrics plan; infrastructure requirements; annotation and markup guidelines; checklist of anticipated positive and negative impacts.
Evaluation Execution			

Concepts and Questions	**Processes**	**Who and Where**	**Outcomes**
How will test subjects interact with systems to produce evidence? What populations and use patterns are required? What infrastructure resources are used during evaluation?	Recruit, enroll, and train participants; run tasks; collect logs per protocol; monitor and quality-control protocol deviations and data quality; manage evaluation environments; ensure data protection, compliance, and governance.	Evaluation design team; analysis team; continuous monitoring team; compliance experts; test subjects; sandbox environments; field sites.	Recruitment plans; training materials; raw evaluation outputs; transcripts; metadata for analysis.
Analysis			
Which narratives or tasks surface systematized constructs? How are signals captured for analysis? What observations link outcomes to constructs?	Analysis design and planning; scoring design; mapping tasks to constructs; mapping anticipated outcomes to scoring; data markup plan; metrics design and execution; annotation training.	Domain experts; evaluation design team; context specification team; test execution team; insights team; annotation and markup personnel.	Finalized metrics and analysis techniques; data markup schema; scoring tools and guidance; analysis outputs including estimates with intervals and distributional breakdowns aligned with stakeholder claims.
Insights and Reporting			
Which stakeholders should be considered when disseminating insights? What platforms optimize engagement and implementation?	Composition of articles for stakeholders, industry, and research audiences; networking with stakeholder communities; relationship building with policy and decision makers; engagement with appropriate dissemination platforms.	Domain experts; implementation stakeholders; stakeholder communities and interest groups; evaluation design team; analysis team; policy and decision makers; conferences; editors and journals.	Reports, case studies, and instantiated evaluation processes; published articles; conference papers; white papers; extended stakeholder networks and advisory groups supporting policy and decision making.
Continuous Monitoring			

Concepts and Questions	**Processes**	**Who and Where**	**Outcomes**
Are post-evaluation controls operating as intended? Who identifies shifts in context? What resources are required for ongoing monitoring?	Data gathering; monitoring thresholds; cadence of monitoring activities; horizon scanning; ongoing information exchange; development of adjustment plans.	Domain experts; context specification team; analysis team; evaluation design team; relevant stakeholders; advisory groups; decision and policy makers.	Periodic insight and impact reports; updated research outputs; refined processes, methods, and metrics; best-practice guidance; longitudinal qualitative and quantitative insight into real-world AI impacts across contexts.