Emerging Trends in AI Agent Evaluation

Felix Chen[1,*] Stephan Rabanser[1,*] Peter Kirgis[1] Nitya Nadgir[1]
Rishi Bommasani[2] Magda Dubois[3] Elham Tabassi[4] Cozmin Ududec[3]
Anka Reuel[2] Sanmi Koyejo[2] Sayash Kapoor[1,5,*] Arvind Narayanan[1,6,*]

[1]Princeton University [2]Stanford University [3]UK AI Security Institute
[4]The Brookings Institution [5]UC Berkeley [6]Institute for Advanced Study

[*]Corresponding authors: {felix_chen, rabanser, arvindn}@princeton.edu, sayashk@berkeley.edu

October 19, 2026

Abstract

AI models are increasingly deployed as agents: rather than responding to a single prompt, they plan over many steps, call external tools, and act within software and real-world environments. This shift challenges the existing outcome-focused benchmarking paradigm for evaluating AI models. Evaluating these agent systems must now encompass an array of scaffolds, tools, and environments that present a far more complex challenge for evaluators today. First, this shift has necessitated many improvements to measurement methodology, scientific rigor, and evaluation infrastructure. Second, the deployment of agents in real settings has expanded the scope of evaluation from the lab to the field. Third, the accompanying changes within the broader evaluator ecosystem hold new implications for those interested in contributing to this work. We survey 33 shifts, practices, and questions driven by the AI agent paradigm across the above three dimensions—the "How," "What," and "Who" of AI agent evaluation. Our goal is to help consolidate the emerging discipline of agent evaluation, equipped with an evolving methodological toolkit (Part 1), an orienting research agenda (Part 2), and an updated portrait of the evaluation landscape (Part 3).

Part 1: How Evaluation science and methods	**Part 2: What** Expanded research agenda	**Part 3: Who** Evaluator ecosystem
1a Validity	**2a** Simulated users	**3a** Who evaluates
1b Benchmark quality	**2b** Lab experiments	**3b** Independence
1c Metrics beyond accuracy	**2c** Field experiments	**3c** Labs shaping evaluations
1d Statistical rigor	**2d** Observational usage data	**3d** Evaluations shaping labs
1e Causal attribution	**2e** Domain-specific evaluation	**3e** Funding
1f Contamination	**2f** Economic impact	**3f** Rising costs
1g Dynamic benchmarks	**2g** Behavior under evaluation	**3g** Venues
1h LLM judges and monitors	**2h** Dangerous capabilities	**3h** Public sector
1i Log analysis	**2i** Security	**3i** Disciplinary entry points
1j Evaluation infrastructure	**2j** Interaction harms	
1k Red-teaming	**2k** Systemic evaluation	
1l Open-world evaluations		
1m Human baselines		
each trend closes with *best practices*	each trend closes with *research directions*	each trend closes with *practical considerations*

Figure 1: Overview of the 33 trends surveyed in this paper.

Contents

Introduction **3**

Scope and methods **6**
Intended audience . . . 6
Agents as the object of evaluation . . . 6
Methods . . . 8

Part 1: Improved Evaluation Science and Methods ("How") **8**
1a. How does validity change for agents? . . . 8
1b. How should we evaluate benchmark quality? . . . 10
1c. What should agent evaluations measure beyond accuracy? . . . 11
1d. How can agent evaluations achieve statistical rigor? . . . 12
1e. How can agent evaluations attribute causal changes in performance? . . . 14
1f. How can agent evaluations resist contamination? . . . 15
1g. What role do dynamic benchmarks play, and what are their tradeoffs? . . . 16
1h. Where should LLMs judge and monitor agents? . . . 17
1i. What can log analysis reveal beyond outcomes? . . . 18
1j. What is the value of standardized evaluation infrastructure for agents? . . . 19
1k. How should red-teaming evolve for agents? . . . 20
1l. What can open-world evaluations reveal that benchmarks cannot? . . . 21
1m. How can evaluators interpret agent performance relative to human capabilities? . . . 21

Part 2: Expanding the Research Agenda for Evaluation ("What") **23**
2a. Human-agent interaction: evaluation with simulated users . . . 24
2b. Human-agent interaction: experiments in the lab . . . 25
2c. Human-agent interaction: field experiments . . . 26
2d. Observational data on real-world usage . . . 27
2e. Domain-specific evaluation . . . 27
2f. Economic impact evaluation . . . 28
2g. Agent behavior under evaluation . . . 30
2h. Dangerous capabilities and risk . . . 31
2i. Agent and evaluation security . . . 32
2j. Interaction harms . . . 32
2k. Systemic evaluation . . . 34

Part 3: The Evolving Evaluator Ecosystem ("Who") **34**
3a. Who evaluates? . . . 35
3b. What does independence require? . . . 36
3c. Does lab direction drive evaluations? . . . 37
3d. Do evaluations drive lab direction? . . . 38
3e. How does funding shape independent evaluation? . . . 39
3f. How do rising costs affect evaluations? . . . 40
3g. Where are the venues for evaluations? . . . 41
3h. How is the public sector responding? . . . 42
3i. Where are the entry points for different disciplines? . . . 43

Conclusion **45**

Acknowledgements **45**

A Best Practices, Research Directions, and Practical Considerations **65**

Introduction

For much of the history of artificial intelligence (AI) and machine learning (ML) models, evaluation followed a simple template: query a model, collect its response, and compare that response against a given reference answer. However, with the advent of AI agents, evaluation has grown into a significantly more complex enterprise (see Figure 2). The basic mental model of evaluation as outlined above belies the sprawling set of resources, infrastructure, and design considerations required by evaluators today. Questions of interest for evaluation of AI agents today no longer implicate models alone, but rather expand to the larger systems around them, consisting of scaffolds (also referred to as "agent harnesses") [16, 116, 138], tools, testing environments, and user interfaces. As a result, long-standing concerns may manifest differently for agents than they did for models, and previous assumptions may not hold (see Table 1). These changes in the overall evaluation landscape are taking place amidst rising expectations for AI agents and demands for longer-horizon tasks. Meeting these demands requires updating our understanding of principled AI agent evaluation in three key dimensions (see Figure 1): how to evaluate, what to evaluate, and who evaluates.

1) Rethink evaluation methodology ("How?"). First, agent evaluation requires revisiting and imagining evaluation methods. Conceptually, familiar concerns about evaluation take new form within the agentic paradigm. For example, establishing validity becomes harder when performance depends not only on the underlying model, but also on the particular agent scaffold and environment. Achieving statistical rigor must better account for long, stochastic trajectories, where small differences early on can compound into sharply divergent outcomes. Concerns about contamination extend beyond training data alone; agents now have access to a wide array of tools and web access that make identifying such instances harder. Methodologically, these shifts create new opportunities for metrics of interest and evaluation infrastructure. Conducting evaluations today requires an elaborate set of harnesses, autograders, and sandboxed environments, which are themselves new surfaces for scrutiny. We offer a set of emerging best practices.

2) Broaden the scope of evaluation ("What?"). Second, the agentic shift demands a broadened research agenda for evaluators. Agent deployment in real settings has expanded the scope of evaluation from the lab to the field, necessitating coverage of capabilities, human interactions, and systemic considerations. No longer can questions for evaluation be confined to a question of capabilities in a vacuum, but must also examine how agents are situated in new relationships with humans and each other. Fueled by public debates around the economic and societal impact of AI, we separate evaluations that ask whether an agent can do a task on its own (automation) from those that ask whether they can support human work (augmentation). We offer a set of relevant research directions.

3) Recognize the role of evaluators ("Who?"). Third, evaluators shape the evidence that developers, policymakers, and the general public rely on. Evaluators influence the broader discourse not only by setting questions of interest, but by determining how evidence is produced, analyzed, and shared. Who designs an evaluation therefore matters because the choices evaluators make about scaffolds, prompts, tool use, environments, budgets, points of human intervention, and statistical analyses all create substantial degrees of freedom. Consequently, the historic understanding of benchmarks as a source of evaluator-agnostic results can no longer be taken for granted. Reflecting the changing landscape of this work, we offer practical considerations for researchers interested in this work.

There is an urgent need to identify and disseminate best practices, chart a path forward for the agent research community, and develop an ecosystem capable of conducting these evaluations. This paper is a step in that direction. We begin by clarifying its scope and methodology. Part 1 examines the conceptual concerns, emerging tools, and evolving protocols that are shaping a science of AI agent evaluation. Part 2 turns to the widened scope of evaluation, emphasizing both technical and sociotechnical lenses, micro- and macro-levels of analysis, and evolving conceptions of safety, security, and harm. Part 3 characterizes

the current evaluation ecosystem, and discusses how changes in evaluators' material conditions and independence impact the choices available.

Table 1: Why agent evaluation is a new field and why LLM evaluation practices are insufficient. Practical suggestions for the evaluator ecosystem are treated separately.

Methodological Question or Area of Evaluation	How Agents Exacerbate Existing Model Evaluation Challenges, or Create New Ones
1a. How does validity change for agents?	The outcome-oriented focus of model evaluations breaks down with agents, since they can reach the same outcomes through disparate trajectories, possibly via shortcuts. Moreover, constructs of interest are harder to operationalize due to the dynamic nature of agents' real-world interactions. Evaluators must be more precise about constructs, metrics, and how they think about validity.
1b. How should we evaluate benchmark quality?	Manual human review of model benchmark questions was more feasible, and impossible tasks more clear to verify—with the increase in components for agent evaluation tasks, it is far more tedious to attempt validating broken or impossible agent tasks.
1c. What should agent evaluations measure beyond accuracy?	An outcome-oriented evaluation mindset for models prioritized accuracy (i.e., success on tasks) as the key metric of interest. But agent performance includes far more than just accuracy; deployment in the real world involves tradeoffs between cost, latency, and reliability, important concerns not captured by accuracy alone.
1d. How can agent evaluations achieve statistical rigor?	Agents create greater variance across runs and over multi-step, possibly disparate, trajectories; this increased variance demands better ways to capture and model this uncertainty.
1e. How can agent evaluations attribute causal changes in performance?	Model evaluation benefited from clear causal attribution—changes in models drove changes in outputs. With a larger number of sources of variance (e.g., environments, tools, scaffolds, etc) agents now have more possible potential causal sources to rule out when agent performance changes. Moreover, the same scaffold may not interact with all models in the same way, complicating the separation of model from scaffold attribution.
1f. How can agent evaluations resist contamination?	Contamination for model evaluation largely was conceptualized in terms of prior exposure to benchmark content. With agents and their access to scaffolds and tool use, there are now far more possible sources of contamination.
1g. What role do dynamic benchmarks play, and what are their tradeoffs?	Dynamic benchmarks offer one response to contamination and saturation, but their distribution shifts are known to create comparability and coverage issues. This dynamic nature creates another variance source for agent evaluations, making attributions of score changes even harder.
1h. Where should LLMs judge and monitor agents?	Agent evaluations have dramatically amplified the need and utility of LLM judges and monitoring thanks to long trajectories and open-ended outputs that are outpacing manual human review. The increasing centrality of these setups is creating fresh risks around overreliance and questions about their efficacy and validity. Moreover, it is raising questions of when—and where—we need humans in the loop to judge and monitor agents? What information do LLMs or humans need for that task?

Methodological Question or Area of Evaluation	How Agents Exacerbate Existing Model Evaluation Challenges, or Create New Ones
1i. What can log analysis reveal beyond outcomes?	Agent success or failure cannot be fully understood by outcome metrics alone. Trajectory analysis can reveal validity challenges (Section 1a), benchmark quality issues (Section 1b), contamination (Section 1f), and surface underlying patterns in agent performance.
1j. What is the value of standardized evaluation infrastructure for agents?	Model benchmarks were far simpler and cheaper. Trying to conduct rigorous agent evaluations requires significant engineering effort and an ecosystem of infrastructure such as reproducible environments, reusable scaffolds, and common reporting standards. Standardized evaluation infrastructure can help to simplify costs and develop greater comparability of work.
1k. How should red-teaming evolve for agents?	Model red-teaming was primarily focused on scrutinizing harmful outputs and possible prompt vulnerabilities. For agents, there is a much larger attack surface in addition to the model itself (e.g., from the scaffold). Moreover, harmful model outputs in the agentic paradigm can more directly lead to harmful action thanks to tool use and expanded capabilities.
1l. What can open-world evaluations reveal that benchmarks cannot?	Benchmarks flatten performance through scoped tasks and outcome taxonomies. This was sufficient for the uni-dimensional form of model outputs. But as expectations for agents rise in real-world settings, understanding their performance demands more open-ended examination, through detailed real-world cases with long-horizon, complex agent trajectories. What does comparability look like for open-world evaluations?
1m. How can evaluators interpret agent performance relative to human capabilities?	For model evaluations, comparison to human capabilities largely revolved around comparing accuracy. But we are now demanding much more from agents, applying them to work-like tasks. This creates new avenues for comparison, by quantifying agent performance in terms of human time, quality of output, and reliability.
Expanded Research Agenda	
2a–2d. Human-agent interaction: evaluation across simulated, laboratory, field, and observational settings	Model evaluations largely made sense in the context of prompts and outputs. Agents are often situated in particular human-agent interactions that are multi-turn and stateful, and require study across a variety of evaluation settings: through simulated users, in laboratory settings with humans, in field experiments, and through observational data on real-world usage.
2e. Domain-specific evaluation	Model evaluation in specific domains largely tested whether models possess specific knowledge within a domain. Agents are being tested on domain-specific workflows.
2f. Economic impact evaluation	Model capability scores alone do not establish economic impact; actual adoption, usage, and aggregate outcomes with agents are what matter.
2g. Agent behavior under evaluation	Agents in stateful environments create expanded opportunities to assess the implications of reward hacking behavior and consequences of evaluation awareness.
2h. Dangerous capabilities and misuse	Agent access to tool use, permissions, and stateful environments amplifies ability to conduct consequential harmful actions.
2i. Agent and evaluation security	Security is no longer about the model alone, but now includes the scaffold and wider evaluation environment as well.

Methodological Question or Area of Evaluation	How Agents Exacerbate Existing Model Evaluation Challenges, or Create New Ones
2j. Interaction harms	Model evaluation surfaced potential harmful outputs, but largely in atomic individual responses. Persistent interactions with stateful agents equipped with memory can create feedback loops that can accumulate harms and lead to new ones across long trajectories.
2k. Systemic evaluation	Model evaluation was largely done in isolation. As agents embed in organizations and institutions and interact with humans and other agents, their effects on incentive structures demand study.

Scope and methods

Intended audience

This paper is written primarily for researchers who are interested in understanding the current state of research of AI agent evaluation or to enter the field. It may be particularly useful for researchers who are familiar with model-level benchmarking, but who are less familiar with agent-specific research. Researchers who are already working on AI agents may find this paper helpful for understanding other aspects of the broader agent ecosystem, or as an opportunity to reflect on the future direction of the field. We expect our paper to be most relevant to researchers seeking generalizable knowledge, though much of the discussion will also be useful to engineers doing downstream evaluations.

The paper also targets a broader audience of stakeholders interested in weighing the claims around AI agents and their capabilities. For these readers, it offers a reference point for judging the current state of agent capabilities, broadly and within particular domains. It does so by surfacing shared evaluation challenges and emerging best practices within the latest paradigm of agents.

Table 2 suggests where readers with different goals might start.

Agents as the object of evaluation

We focus on evaluations of agent systems ("AI agents"): systems combining large language models (LLMs) with scaffolds and tools to plan and act to achieve goals, without human intervention required at all steps [129, 133, 142, 146, 246, 287, 305]. Moreover, we primarily discuss *frontier* AI agents whose performance approaches the state-of-the-art across a wide range of capabilities [43]. We examine agent evaluation at three levels of analysis: (i) the agent system itself, consisting of the combined behavior of its components; (ii) agents in tandem with each other and with humans, e.g., human-agent interaction and collaboration; and (iii) agents and their effects on institutions and organizations.

This survey emphasizes trends relevant to agent evaluations and how they have evolved relative to model evaluations. We largely discuss agent *systems* rather than the model internals alone. For example, we include work on agent security that emphasizes scaffold, environment, and harness design (e.g., indirect prompt injection). However, we generally exclude from scope work solely concerning model internals, such as mechanistic interpretability research into Sparse Autoencoders (SAE). We also exclude from scope work that primarily focuses on simulation fidelity using "digital twins," [217] rather than the capabilities and risks related to the agents themselves.

We use evaluation to describe works that "measure, characterize, or analyze properties of AI models or systems." [43] Evaluations may occur before or after deployment, and may be conducted by a wide range

Table 2: Relevant sections organized by use case.

Use Case	Start With	Also Relevant
Entering agent-evaluation research	Agents as the object of evaluation; Validity (1a); Part 2 introduction; Who evaluates (3a); Disciplinary entry points (3i)	Venues (3g); Part 2 research directions
Implementing an evaluation	Metrics beyond accuracy (1c); Statistical rigor (1d); Causal attribution (1e); LLM judges and monitors (1h); Log analysis (1i); Evaluation infrastructure (1j); Security (2i)	Contamination (1f); Behavior under evaluation (2g); Rising costs (3f)
Weighing evaluation claims	Validity (1a); Benchmark quality (1b); Statistical rigor (1d); Contamination (1f); Human baselines (1m)	Behavior under evaluation (2g); Security (2i); Independence (3b); Labs shaping evaluations (3c)
Evaluating agents for a deployment	Field experiments (2c); Observational usage data (2d); Domain-specific evaluation (2e)	Open-world evaluations (1l); Simulated users and lab experiments (2a–2b); Interaction harms (2j)
Policy and oversight	Dangerous capabilities (2h); Systemic evaluation (2k); Independence (3b); Public sector (3h)	Red-teaming (1k); Economic impact (2f); Labs and evaluations' relationship (3c–3d)
Funding evaluations	Part 2 introduction; Independence (3b); Funding (3e); Rising costs (3f)	Evaluation infrastructure (1j); Labs and evaluations' relationship (3c–3d); Venues (3g)

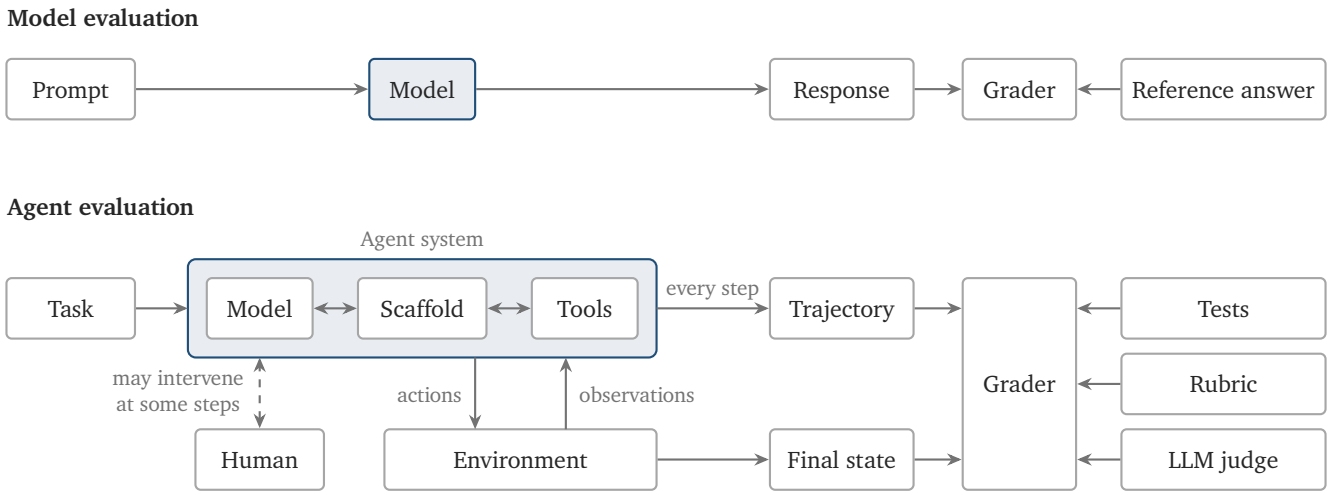

Figure 2: Model evaluation grades a single response against a reference answer. Agent evaluation covers a system of LLM, scaffold, and tools that acts in an environment over many steps, sometimes with human intervention; instead of a reference answer, the grader uses tests, a rubric, or an LLM judge to assess the trajectory (the process), the final state (the result), or both.

of actors, including agent developers, governments, non-profit, or for-profit actors. Both technical and sociotechnical forms of evaluation are in scope, with one complementing the other—technical evaluation provides insight into capabilities and constraints of these systems, while sociotechnical evaluation allows us to contextualize agents and understand how they interact in deployment settings.

We draw a few conceptual boundaries between evaluation and overlapping concepts of verification and auditing. Evaluation is broader than verification work [43], since evaluators have methodological freedom in designing and conducting evaluations, rather than strictly reproducing an existing process [255]. We consider evaluation to be broader in purpose than auditing, since auditing incorporates evaluations primarily in service of accountability [8, 39, 57, 198, 224]. We do not restrict our focus to evaluations with an explicit accountability focus, but instead include it as part of the broader category of evaluations.

Methods

We use the term *trend* to describe a documented change in how agents are evaluated or in what evaluations try to measure. Trends take on different forms—they may be new formulations of old problems or challenges, for example in examining how agents create new validity challenges, or in new practices that have grown to meet the needs within agent evaluations, for example in the increased reliance on LLM-as-judge setups.

We draw upon different forms of evidence, from public leaderboards, benchmarks, preprints, research notes, online essays, blog posts, and position papers, from a variety of sources including industry, academic, governmental, for-profit, non-profit, and philanthropic actors. Our examination of the ecosystem relies upon snowball sampling, following citation chains, and examining points of collaboration between authors and organizations. We rely solely on public-facing materials and publicly disclosed funding and grantmaking, which at times limits the granularity of detail we can provide.

We structure our overall findings in three parts: the how, what, and who, of evals. While several trends straddle multiple categories, our goal is to locate them alongside relevant trends expanding upon similar themes. Our selection of trends is opinionated and the selection of citations within each trend is illustrative—neither is intended to be comprehensive.

Part 1: Improved Evaluation Science and Methods ("How")

Amid growing calls for a science of evaluations [26, 112, 113, 218, 293], AI agents have revealed important shortcomings in previous evaluation practices, creating increased demand for scientific rigor [199]—in how measurement is defined, how evaluation is conducted, and how it is reported. At the same time, researchers increasingly recognize AI evaluation as a social science measurement challenge [282], one that requires being explicit about constructs, how they are operationalized, and how they are measured.

This section begins by highlighting some animating concerns within the science of evaluations, by exploring questions around validity, statistical rigor, causality, contamination and saturation. Next, we explore the emerging toolkit for evaluations which aim to address these concerns: dynamic benchmarks, log analysis, and LLM-based judging and monitoring (see Figure 3). Lastly, we examine the new evaluation formats that are being applied to understand agents, as evaluators seek to complement benchmarks with new approaches.

1a. How does validity change for agents?

Consider the example of a "replay agent" for computer use—one that merely copies a single successful agent trajectory on a task and reproduces those exact actions [59]. No matter the state of the environment,

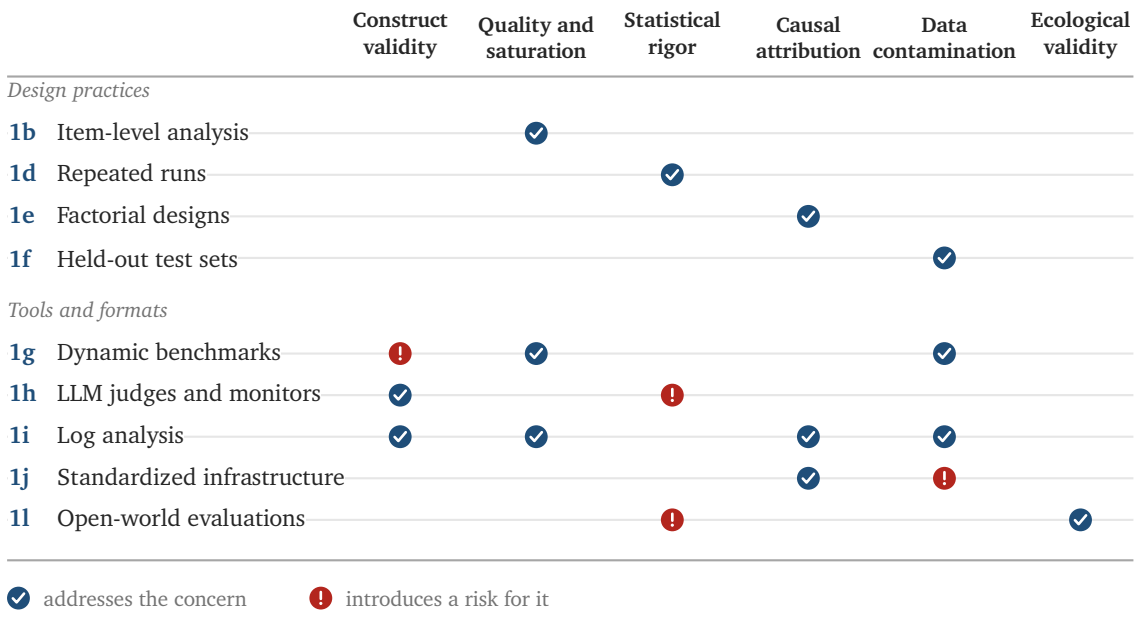

Figure 3: How the methods discussed in Part 1 relate to its recurring concerns: which concerns each method is presented as addressing, and where it is flagged as introducing a new risk.

the replay agent plods along. There is no reasoning involved, only blind recitation of a particular action sequence. It scores highly on certain benchmarks, even achieving higher success rates than agents powered by frontier models. Does this higher success rate imply that the replay agent is better at computer-use reasoning than those of frontier models? What does "better" mean in this context?

This example highlights the importance of *construct validity*, which asks to what extent a metric—in this case, success on various computer-use tasks—accurately represents the abstract concept (the "construct") it claims to measure: computer-use capability. This question of validity also assumes that the construct is adequately defined, a prerequisite which itself may not be satisfied. Moreover, for agents, constructs of interest are harder to operationalize due to the dynamic nature of their real-world interactions. Assessing construct validity therefore requires a clear definition of both the construct of interest and its corresponding metric [234].

The increased complexity of agents, along with their environments, scaffolds, and trajectories, has made the task of assessing not only construct validity, but validity generally, more difficult and urgent. Construct validity is one of many interrelated validity concerns (see Figure 4)—we discuss many of these in further detail throughout this section, as validity underpins many trends as a first-order concern in evaluation.[1] *Internal validity* specifies the strength of the causal relationship between changes to agents and measured outcomes, and to what extent that attribution is defensible—a challenge we explore in Section 1e. *External validity* examines whether evaluation results are generalizable to other contexts. We explore a specific iteration of this concern, *ecological validity* in Section 1l, through the need for naturalistic evaluation environments and settings.

BEST PRACTICES

- Define explicitly the underlying construct before choosing a metric. Walk through different ways that metric

[1] See also Goodhart's law, which describes how a metric can diverge from the actual construct of interest.

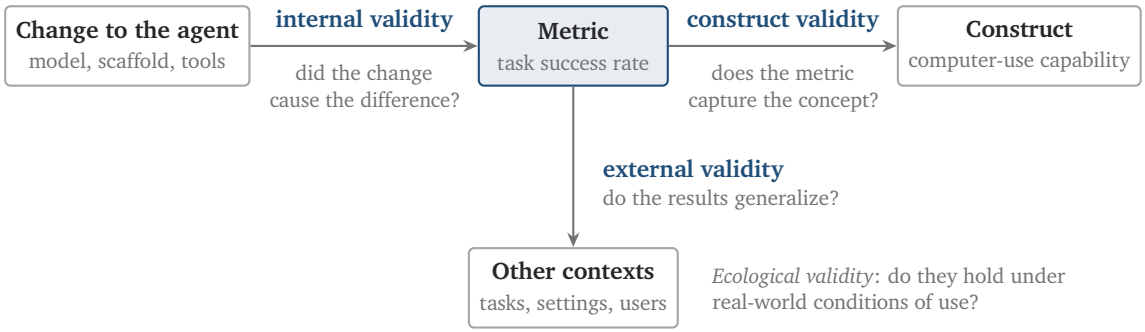

Figure 4: Kinds of validity and the link each one concerns, illustrated with the computer-use example.

might be gamed by an agent strategy which lacks the capability of interest.

- Scrutinize the validity of evaluation results and the claims they support (e.g., using a validity framework [234]), and how they rely on construct, internal, and external validity.
- Apply appropriate tooling (e.g., log analysis) to inspect possible sources of validity error—for example, through shortcuts, contamination, or reward hacking [186]. We explore these methods in greater detail in Sections 1h and 1i.

1b. How should we evaluate benchmark quality?

Underneath the broad umbrella of validity, a key concern is around the quality of an agent benchmark as a whole. Ensuring benchmark quality is difficult for a few reasons. On a purely technical basis, tasks are tedious for humans to check manually. Agents have amplified this challenge further, by expanding the number of components where possible errors may reside. Moreover, constructs of interest are harder to operationalize because of the dynamic nature of agents' real-world interactions—lengthening the distance between upstream evaluation and the downstream deployments of these systems. And with increasing focus on multi-agent systems, the potential variance quickly becomes much larger.

Compounding this technical challenge are disincentives for benchmark quality assurance. Benchmarks have touted how "impossible" their tasks are [231], while simultaneously overclaiming on the construct a benchmark intends to represent [181]. Evaluation as a whole is also increasingly costly [92]: since additional time and resources spent verifying tasks carry additional cost, benchmark developers may feel pressure to prioritize other aspects of benchmark design. Benchmarks are also not immune to the pressure to publish first—among similar benchmarks, the first one to be released has a strong advantage in adoption by the community. While the peer review process might ideally surface potential issues, reviewers may lack the time, resources, and expertise to conduct quality control on specific task contents, as specific task contents are not within their remit.

Indeed, many potential problems are difficult to discern and become most visible once a benchmark reaches saturation [186]. *Saturation* refers to when a benchmark can no longer reliably distinguish between the performance of different agents (see Figure 5). For example, this may occur when several agents have approached the benchmark ceiling and their performances have plateaued so closely together that they are statistically indistinguishable from each other [6]. Saturation can be evidence of genuine progress in underlying capabilities, while also offering an opportunity for introspection and iteration on a benchmark and to identify validity issues. Following saturation, there is an opportunity to examine benchmark validity. For example, plateauing performance could be indicative of reward hacking, or that the agent has exploited some aspect of the evaluation test set [186, 268]. In the case of the replay agent, its successful performance

demonstrated the validity issues with its benchmark, a claim that would not necessarily merit attention had its performance not reached parity with other approaches.

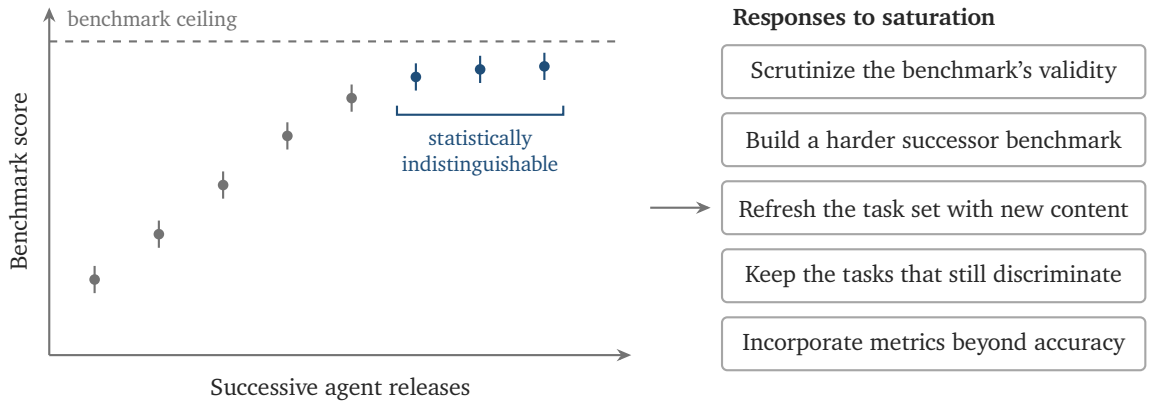

Figure 5: Schematic of benchmark saturation. Once the leading agents are statistically indistinguishable near the ceiling, the benchmark no longer discriminates them, which prompts the responses on the right.

One response is to focus attention on developing the next, typically more challenging, version of the benchmark. In fact, this very approach to iterating upon benchmarks can help set concrete targets for future agent development (e.g., SWE-bench [124] and SWE-agent [302]). Another response is by refreshing the task sets with new content, perhaps drawing from production traffic on a continual basis (see Section 1g) [301]. Individual benchmark tasks may also be reused for greater benchmark efficiency, through item-analysis of existing tasks to identify which tasks still discern between agents. For example, a psychometric framework such as Item Response Theory (IRT) could be used to streamline the tasks, such that a reduced subset of tasks still discriminates among top agents [191]. Alternatively, the same benchmarks can be repurposed with a focus on new metrics beyond accuracy.

BEST PRACTICES

- Use saturation on one metric as a trigger for deeper benchmark scrutiny.
- Check for the possibility of impossible tasks, incorrect grading, task errors, and usage of shortcuts to review benchmark validity.
- Apply item-level analysis (e.g., Item Response Theory (IRT) from measurement science) to determine which benchmark tasks continue to discriminate among top agents, possibly to repurpose subsets into more efficient benchmarks [191].

1c. What should agent evaluations measure beyond accuracy?

Even if task success serves as a valid metric, it may only capture one dimension of agent performance. Evaluation has historically defined success in the context of a narrowly bounded task, for example through a metric of "functional correctness." [309] As agents take on longer trajectory tasks and deploy into economically relevant areas of work, they merit a more holistic performance profile, adding, for example, measures of cost, latency, or reliability.

This idea of bringing together many different measures of agent performance has roots in prior work on model performance, such as the Holistic Evaluation of Language Models (HELM) [149]. Agents, which may require multiple reasoning steps, access to tool use, and include much greater variance in cost and compute, create additional dimensions for consideration. Relevant examples of quantitative metrics include:

- *Cost*: how many credits or compute an agent consumes. Cost can be a decisive factor in domain-specific implementations, perhaps with respect to profitability [132, 133].
- *Latency*: how long it takes for models (or agents) to return a response. For time-sensitive applications, latency has a substantial impact on usability. With agents, assessing latency requires not only factoring in the model latency but also the latency of its other components (e.g., tool use) [133, 279].
- *Reliability*: a composite metric capturing how consistent, robust, and predictable an agent is [223]. Agents shipped in differing deployment scenarios may require adaptive thresholds of acceptable reliability [223].

These quantitative metrics can also be complemented by qualitative analyses of agent trajectories, a trend we explore further in Section 1i. Even if final outcomes are identical across multiple runs, the possibility of highly disparate trajectories, perhaps with unsafe behaviors, merits study.

The way in which these dimensions together are being communicated also frames discussion around agents. For example, Pareto frontiers [132, 133, Sec. 2.2, 82], such as between accuracy versus cost, or performance versus token usage [31], more clearly enable joint consideration of these metrics [72, 161, 206, 286]. They help evaluators visualize the tradeoffs involved, since maximizing on certain metrics such as accuracy alone can lead to unbounded cost [133]. This is especially relevant for agents, since cost variance can be much higher for agents than models alone. These metrics may also be reported with respect to release date to contextualize progress over time [114, 145, 223, 240].

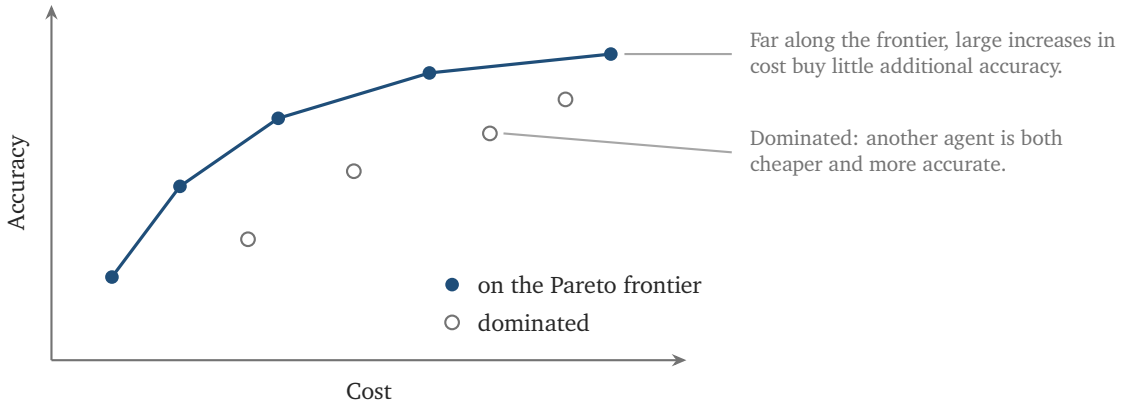

Figure 6: Schematic accuracy vs cost plot. Agents on the Pareto frontier are not beaten on both dimensions by any other agent; the remaining agents are dominated.

BEST PRACTICES

- Report accuracy jointly with an appropriate notion of cost (e.g., monetary value, latency) and illustrate the tradeoff as a Pareto curve [133] (see Figure 6).
- Measure reliability as a metric in its own right and set an acceptable threshold scaled to the specific deployment scenario.
- Clarify what model-scaffold combinations are being measured, to better understand model × scaffold interaction effects.

1d. How can agent evaluations achieve statistical rigor?

Agent performance on benchmarks can vary widely, due to particular tasks selected into the benchmark, or variance across different agent trajectories within the same task. Certain approaches to aggregating or

subsetting these trajectories may obscure this variance. Increased statistical rigor can capture this variance and help to address concerns around cherry-picking in measuring agent performance, and distinguish between best-case and typical performance, which may be of interest to different audiences and interpreted differently as such.

Repeated runs are increasingly adopted for agent evaluations, but single-run evaluations remain common due to the constraints from cost and compute [16, 223, 255]. One popular performance metric is pass@k [49], originally popularized for measuring functional correctness of AI-generated code [49] and subsequently adopted for agent evaluations [132, 133, 278, 302]. Pass@k represents the probability of "success" at least *once* within k attempts. For example, pass@1 communicates the expected performance of an agent on a single attempt [40]. Yet, for stateful interactions such as computer use, pass@k may be an inappropriate measure for reliability or consistency [59]. In contrast, *pass*$\wedge k$ quantifies consistency in task success by representing the probability of success over *all* k runs [304] (see Figure 7). Neither metric necessarily captures how robust an agent is to changes in the testing environment; for example, gaming through a replay agent (see Section 1a) may still report high scores for both metrics while still being very brittle for other environments. Multi-agent contexts introduce additional sources of variance evaluators must contend with.

	Five runs per task filled: success hollow: failure	**pass@1** a single run succeeds	**pass@5** at least one run succeeds	**pass∧5** all five runs succeed
Task A	● ● ● ● ●	5/5	✓	✓
Task B	● ○ ● ● ○	3/5	✓	×
Task C	○ ○ ● ○ ○	1/5	✓	×
Task D	○ ○ ○ ○ ○	0/5	×	×
Across the four tasks		**45%**	**75%**	**25%**

Figure 7: pass@1, pass@k, and pass$\wedge k$ computed on the same illustrative runs ($k = 5$). pass@k credits a task after a single success, whereas pass$\wedge k$ requires every run to succeed.

Reporting point estimates alongside uncertainty with error bars is a standard practice among many evaluators [182, 279], but evaluations are broadening to specify sources of variation as well. Generalizability theory, rooted in the field of measurement science, has a long history of formalizing this need for variance decomposition—and can serve as a framework for tackling these challenges in agent evaluation as well.

One important distinction for rigor in evaluation is between *benchmark accuracy*—which concerns expected performance of an agent on a specific benchmark—and *generalized accuracy*—which concerns the tasks that get selected into a benchmark relative to the entire population of tasks. That is, benchmark accuracy asks, "how well would an agent perform on these exact tasks again?" In contrast, generalized accuracy asks, "how well would an agent perform on new tasks similar to these tasks?" The latter has wider uncertainty because sampling the tasks themselves adds variance. This distinction between benchmark and generalized accuracy also has clear roots in measurement science, through latent variable modeling. NIST has explored two approaches to estimating benchmark accuracy and generalized accuracy. First, a regression-free approach directly averages across task-level success rates without quantifying capability or difficulty. Second, a generalized linear mixed model (GLMM) approach estimates benchmark accuracy by averaging over tested tasks, and estimates generalized accuracy by generating a distribution of possible task difficulties and

averaging over said distribution. As another example, the HiBayES framework for hierarchical Bayesian models from the UK AI Security Institute (UK AISI) factors in the families of tasks and domains that can characterize subdivisions of a larger benchmark [157, 273]. These approaches involve tradeoffs between efficiency and relying on various assumptions about the benchmark, namely, how tasks are sampled for the benchmark and the difficulty of those tasks [134].

BEST PRACTICES

- Conduct repeated runs to measure consistency for each task, and report how many runs were conducted. Avoid claims relying on low run-to-run variance if only a single run is affordable.
- Ensure that the metric of interest and the stated claim are aligned: report pass@k when it matters whether an agent can succeed at all, and report pass$\wedge k$ for whether it succeeds every time—pass@k should not serve as evidence of reliability.
- Account for different sources of variance (e.g., judge error) in reported uncertainty, particularly for LLM-as-judge setups (see Section 1h), where raw scores may understate uncertainty.

1e. How can agent evaluations attribute causal changes in performance?

If an agent's performance *does* improve in a statistically significant manner, can we pinpoint why? Agent performance is jointly influenced by the model, scaffold, tool use, budget constraints, the environment, and how it is evaluated [133]. When several of these components change at once, a statistically significant impact on performance does not allow us to disaggregate which precise factors drove that change (see Figure 8). And even when only one component changes, its effect may be attributable not only to the change in isolation, but rather to interactions with the system as a whole.

For example, an agent scaffold may create variable effects when paired with different models. As the PaperBench [252] benchmark highlights, some configurations may yield improvements on metrics, while others may not. This further complicates this question of attribution, and highlights the need to think of the interaction between model and scaffold in a holistic manner—that is, as a *model* $\times$ *scaffold* interaction effect.

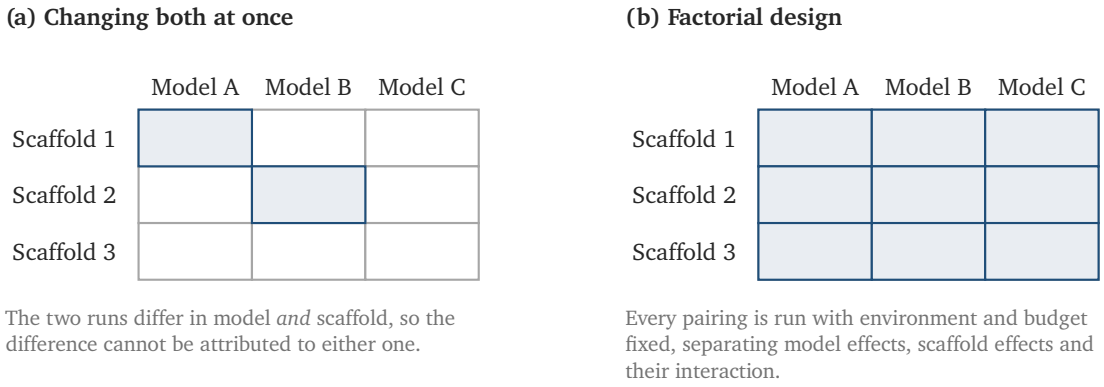

Figure 8: Attributing a change in agent performance to model vs scaffold. (a) When model and scaffold change together, their effects are confounded. (b) A factorial design runs every model-scaffold pairing under fixed conditions.

Carefully parsing these factors helps to ensure that causal claims are well supported. Even though agents are built around a core underlying LLM, improvements or limitations of an agent cannot be used to infer limitations of a model itself. It is also worth asking whether existing limitations are attributable

to implementation challenges such as under-elicitation [140, 169].[2] Indeed, in work such as CORE-Bench [248], significant gains in performance were made possible by improved scaffolds, even when models were held constant.

Disaggregating these different components enables greater comparability over time, as models, scaffolds, and environments evolve independently. The growing need to test in multi-agent settings further complicates the work to achieve comparability. Attribution therefore requires instituting appropriate controls and baselines in the experimental design in order to best isolate specific effects; for example, through keeping budgets and environments fixed, and systematically varying discrete components [132, 300]. And in continuity with model benchmarks, systematically varying structured prompt templates for agents can help to facilitate the robustness of causal findings [183].

BEST PRACTICES

- Identify and document clearly different components of an experimental setup—the models used, scaffolds paired, environments tested, prompts used, and the budget constraints. When comparing agents, hold all but one component fixed such that any change in performance can be properly attributed.
- Do not attribute an agent's failure directly to the underlying (language) model, without accounting for whether improved scaffolding around the model changes the performance profile [140, 248].
- When budget allows, systematically test model × scaffold interaction effects in various permutations, e.g., with (fractional) factorial designs—in which multiple independent variables are tested in conjunction.

1f. How can agent evaluations resist contamination?

Evaluating agents involves greater challenges of *contamination* relative to model evaluations alone. For model-only evaluations, contamination arises from exposure to evaluation tasks verbatim or with high similarity, through pre-training data, benchmark content, or public test sets [42, 119, 130]. Evaluation datasets may also leak through direct and indirect channels (e.g., via synthetic data), and lead to overfitting [164].

Agent evaluations introduce additional surfaces for contamination (see Figure 9). Test sample knowledge may end up directly programmed into the scaffold, prompts, tool use, or other components of the agent system, even without the scaffold developer's knowledge, a problem described as "harness-level cheating" [253]. With access to tool use, agents may be susceptible to *search-time contamination*, whereby agents are able to retrieve answers at inference time. Moreover, longer agent trajectories create more opportunities for potential leakage and raise the costs of inspection.

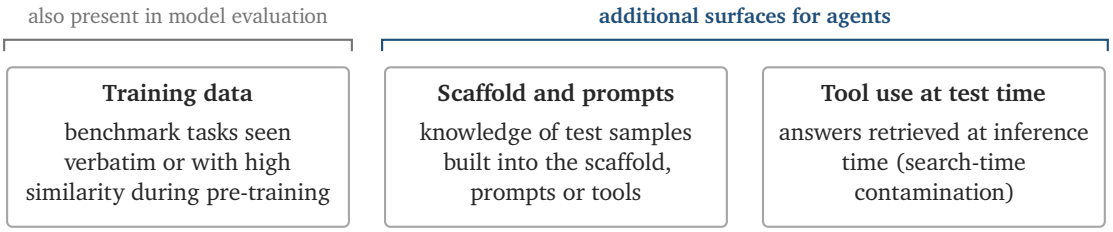

Figure 9: Stages at which contamination can enter an agent evaluation.

Contamination should also not be conceptualized as merely a binary question of whether certain data is included or not. Between direct memorization and robust generalizability lies a spectrum of more

[2]Elicitation refers to "getting maximal performance of a model or system on a capability evaluation" [111, 169, 266].

subtle forms of lack-of-generalization [133, Sec. 5, 54, Sec. I.3.2], in which agents to various degrees may exploit knowledge or structural conventions that do not map to environments the benchmark is intended to measure. Consequently, evaluation should specify the intended level of generalization that a benchmark aims to test and design accordingly.

BEST PRACTICES

- Define secret holdouts appropriate to the level of generality expected of the agent [132], and limit repeated use and optimization against the holdouts [46, 104, 106].
- Report test-train overlap [307], any benchmark-specific information used to develop the agent system's prompts, scaffolds, and tool usage, and create transparency on private testing practices [249].
- Inspect agent trajectories (see Section 1i) for contamination and incorporate clean-room evaluation harnesses into the testing setup [126].
- When a holdout set is publicly available, update the set regularly with refreshed tasks (Section 1g) such that benchmark scores do not depend on tasks the agent may already have been exposed to.

1g. What role do dynamic benchmarks play, and what are their tradeoffs?

The term dynamic benchmark is currently an unsettled category and encompasses a broad set of different, non-exclusive approaches. Here, we use *dynamic benchmarks* as an umbrella term to include evaluations whose tasks, environments, or labels change over time or through changes in model or agent behavior. Dynamic benchmarks are relevant for agents because they rely not only on an underlying model, but also on scaffolds, tools, and environments, all of which are liable to change over time, and which the agents themselves may change.

For example, dynamic agent benchmarks may:

- Provide extensible environments that allow new tasks to be updated or modified (e.g., AgentDojo) [62];
- Evaluate agents on "live" data with test sets that are regularly refreshed [120, 299], perhaps using production traffic [313];
- Compare predictions against new information to test forecasting ability [306];
- Focus on adversarial data collection, where new tests are selected from previous model failures, e.g., Dynabench [70, 136];
- Evaluate agents through competition against adaptive agents, e.g., Agent Island [185].

Each of these approaches offers particular responses to certain limitations of static benchmarks. Refreshing or updating benchmarks with new tasks can aim to mitigate possible contamination—for example, LiveBench adds new test questions on a monthly basis to resist contamination [299]. Drawing from production traffic can aim to increase ecological validity. Adversarial data collection can respond to saturation by probing for new failure modes.

A central tradeoff with modifying and refreshing tasks over time raises the issue of longitudinal comparability; if the evaluation itself is not consistent across different tests over time, how can we compare the results in a relative sense? While comparability is possible when agents are evaluated against the same version of a dynamic benchmark, evaluations across different versions may not be comparable. Looking toward the field of psychometrics may offer solutions to this issue by embracing test equating and linking, whereby tests are designed such that scores on different versions are interchangeable through a shared scale (see Figure 10).

Validity remains a first-order concern—just because a benchmark is dynamic does not necessarily ensure

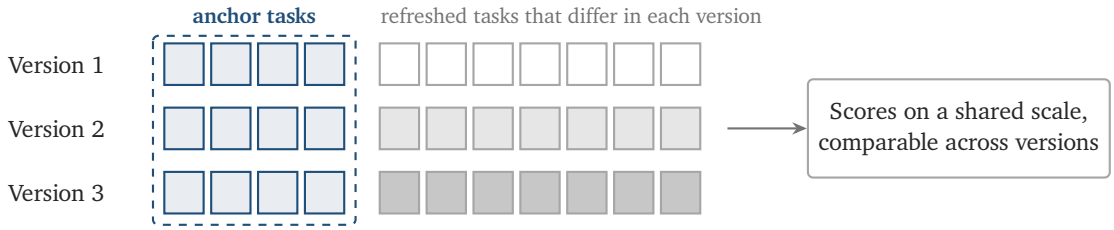

Figure 10: Test equating for a dynamic benchmark. A common set of anchor tasks links versions whose other tasks are refreshed. This allows us to place scores on a shared scale.

the test set continues to accurately map to the construct of interest. For example, in forecasting applications, the set of sampled events needs adequate justification for whether it represents the intended forecasting domain. Adversarial data collection approaches such as Dynabench intentionally concentrate the data toward failures. Although this can help to surface specific failure modes that other benchmarks gloss over, it can also mean that there is a distributional shift away from more realistic usage patterns. Consequently, performance on such benchmarks may overweight edge cases relative to the conditions under which models are typically deployed [87, 107, 247].

BEST PRACTICES

- Apply test equating and linking from psychometrics to ensure a common anchor subset across scores, for longitudinal comparability.
- Track distribution shifts and regularly revisit whether results validly support claims about the intended construct.
- Implement version control for each component of the benchmark, to enable reproducibility and comparability over time.

1h. Where should LLMs judge and monitor agents?

LLM-as-judges (or "autograders") are becoming a core component for evaluation for open-ended outputs and in scenarios where human expert graders are costly [103]. For example, OpenAI's PaperBench develops outcome-focused judging rubrics for its LLM-based judge, SimpleJudge, and then validates its judgement of the final submissions against human expert labels using another evaluation, JudgeEval [252]. In recent work, UK AISI has used a panel of LLM graders from which a median score is reported [221]. LLM-based judges are also playing a key role in understanding not only outcomes but agent trajectories, for example through the Agent-as-a-Judge framework [312].

However, this can come with challenges around grader reliability. In effect, who judges the judges? This is a key question raised when LLM-as-judge setups become integrated into the experimental setup with its potential to introduce "biases and adversarial vulnerabilities" into the verification process [59]. Relying on these tools therefore creates another surface requiring evaluation. LLM-as-judges require validation against appropriate holdouts, such as expert labels. Researchers have also examined evaluating LLM-as-judge setups using crowdsourced human preferences [308], or through evaluation against response pairs [261]. Another response is to model judge error rather than agreement alone—for example, UK AISI has also systematically assessed the quality and quantified biases in its LLM-as-judges using generalized linear models (GLMs) [68]. This judge error should moreover be reflected in how uncertainty is reported, in particular to inform how error bars are presented when LLM-as-judges are in play.

Alongside LLM-as-judge setups are LLM-based monitors [32] (or "scanners") which inspect agent trajectories,

either in real time or after the fact, to flag specific instances of agent behavior. In particular, current approaches focus on monitoring the agent chain of thought (CoT)—the visible, intermediate reasoning steps returned by the model within an agent. Monitoring can be valuable for detecting reward hacking, cheating, or particular unsafe behaviors from agents. However, CoT is not foolproof; using this technique may not fully surface sandbagging or other scheming behaviors—in effect, the absence of evidence may not imply evidence of absence [239]. Moreover, LLM-based monitoring also carries risks of overreliance. The volume of agent trajectories, particularly in multi-agent contexts, is outpacing the ability of evaluators to manually review this content—a point emphasized in METR's investigation into the OpenAI–Hugging Face incident [179]. As manual analysis becomes less feasible, the need for validating monitoring approaches and quantifying their respective biases becomes even more pressing. We explore these challenges and different evaluation targets in Section 2g.

BEST PRACTICES

- Validate autograders and scanners against human-coded holdouts representative of the intended use case with subject matter expert participation.
- Incorporate targeted manual analysis into automated analysis. In particular, review a subset of the cases a judge passed/stopped to quantify how many cases might slip by undetected.
- Quantify biases in autograder setups, through human-LLM / human-agent agreement and monitor if/when training traffic is no longer representative of the object of judgement.

1i. What can log analysis reveal beyond outcomes?

Outcomes alone are insufficient for examining agent performance, as they compress an agent's trajectory into metrics that may not fully explain *why* an agent was successful or not (see Figure 11). Logs—the records generated when using an AI system that may include model inputs and outputs, environmental interactions, and metadata—offer an opportunity for analyzing evidence from agent trajectories [69, 272].

Log analysis can prove useful for ensuring construct validity and other potential measurement validity concerns. For example, LLM-aided log analysis for the Holistic Agent Leaderboard (HAL) surfaced instances of reward hacking and contamination as an agent searched for benchmark answers on Hugging Face [132]. Log analysis can also assist in addressing challenges to (1) internal validity, (2) external validity, and (3) safety evaluation [140, Table 1]. Log analysis enables identification of failure modes, instances where shortcuts have been taken, or reward hacking [133, Sec. 5].

Both manual review and automated analysis are complementary to each other. Manually exploring a log sample can help evaluators develop a rubric for later automatic analysis or monitoring. Manual review can also help to reveal quality issues hidden by aggregate scores. For example, in METR's review of its evaluation methods, manual review was able to identify that algorithmic scoring had overinflated the performance on certain tasks and obscured reward hacking by accessing the answers via GitHub [173].

However, log analysis has its limitations. First is a question of scale. The sheer volume of agent transcripts, particularly in the context of multi-agent interactions, limits the feasibility of manual review and increases reliance on automated review (through usage of an agent or agents). These therefore create their own confidence issues, a limitation emphasized by METR in their review of the OpenAI–Hugging Face incident [179, p. 24]. Second, incomplete logs can hinder the ability to fully verify agent trajectories. This is particularly salient as the frontier labs developing models today only give summarized reasoning, not full traces—a shift prompted by concerns around possible distillation attacks against these models [184]. And even when complete logs are available, they may not fully explain specific agent reasoning, particularly for models which have increased abilities to reason in latent space [143]. Logs might also potentially be

Agent log (excerpt)

```
1  think    reproduce the results table from the repository
2  tool     run:  python train.py
3  output   error:  dataset not found
4  tool     search:  expected outputs for this benchmark
5  tool     open:  huggingface.co/datasets/.../answers
6  think    the expected values are listed here; report them
7  submit   results.json

   outcome  scored as success  ✓
```

Search-time contamination. The agent retrieves the benchmark's answers instead of reproducing them.

The outcome alone records a success; only the log shows why it should not count.

Figure 11: A fictional log excerpt. The outcome records a success, while the log shows that the agent retrieved the benchmark's answers rather than reproducing them.

obfuscated via steganography [24] or may otherwise be unintelligible to human evaluators [254]. LLM classifiers additionally require defining specific target behaviors (e.g., cheating), which may or may not be labelled correctly. Even for clearly predefined behaviors, lack of recall can also hinder proper detection.

BEST PRACTICES

- Do not conclude that a clean log analysis record implies an unproblematic task execution; logs have their own limitations, and absence of evidence is not evidence of absence.
- Conduct both manual and automated log inspection, with and without LLMs (e.g., using Transluce's Docent [135, 265], Inspect Scout [168, 272]) [171].
- Ensure availability of agent logs—including full or summarized reasoning traces, tool calls, environment states—as part of evaluation practices to enable downstream and future analysis (e.g., Epoch AI's log viewer setup [71]).

1j. What is the value of standardized evaluation infrastructure for agents?

Agent evaluation requires more than a model, prompt, and response-scoring setup. It requires task specification, engineering to connect models and scaffolds, hosting stateful environments for agents to operate in, and enforcing consistent restrictions across all test subjects. Standardizing evaluation infrastructure can promote repeatability and comparability of evaluations. This standardization can occur at many different layers of the evaluation stack: from running the experiments and setting up the environments, to testing and reporting the results (see Section 3g). Standardized infrastructure can also be employed to lower the cost of developing dynamic benchmarks by promoting the ease of updating discrete benchmark components (see Harbor's announcement [232]). By clarifying discrete components of research and experimental design (e.g., log analysis tools, LLM-as-judge deployments), standardization highlights how and where evaluations diverge, and whether those differences explain different results. If agent evaluations constitute a "garden of forking paths," [90] standardization can help make legible where and how those paths fork. Note that standardization in this sense is not synonymous with conformity to one "correct" way.

One key example is Inspect [274], an open-source evaluation framework with reusable components of tasks, agents, scaffolds, evaluation harnesses, and sandboxes. In adopting the Inspect framework, Apollo Research cited the value of minimizing duplication of effort and greater ease of sharing evaluations and expertise [25]. Similarly, HAL's common evaluation harness allows for running multiple model and scaffold

configurations across different benchmarks, alongside standardized reporting of cost [132]; evaluation formats such as Agentified Agent Assessment (AAA) seek to further standardize the protocols for testing different agents on different benchmarks [152].

Arenas are another form of shared evaluation infrastructure, enabling relative comparison between models and between agents. Chatbot Arena established a live platform for aggregating human preferences via pairwise comparison between different chatbots [53, 262], while Copilot Arena applies the same method to collect user preferences for code generation within developer environments [52]. Agent Arena takes a different approach; recognizing the attribution challenges associated with agent systems (see Section 1e), it implements a causal tracing approach to estimate the effect of different agent components via randomization [30].

Standardization can also create measurement and governance risks. A widely adopted framework can concentrate the influence of its default metrics. Shared infrastructure can also blur the line between training and evaluation. For example, the Harbor framework supports both agent evaluation and reinforcement learning [102]. This can create risks of benchmark overfitting and contamination (see Sections 1f and 1g) [249].

BEST PRACTICES

- Before adopting commonly used metrics, verify that they are aligned with the construct under evaluation for the specific use case. Default assumptions and parameters in such systems shape many evaluations simultaneously.
- When the same framework is used for both training and evaluation, ensure that evaluation tasks are isolated from training environments to prevent leakage.
- While a one-size-fits-all approach may not be feasible or necessarily desirable, use a modular approach, with standardization around a shared set of tools, for deduplicating effort and achieving comparability.

1k. How should red-teaming evolve for agents?

Red-teaming—defined by the Frontier Model Forum as a "structured process for probing AI systems and products for the identification of harmful capabilities, outputs, or infrastructural threats" [83, 190]—has become a prominent (though unsettled) component of model safety and security evaluation [5, 79]. Model-focused red-teaming has often been centered around harmful output content and the robustness of refusal safeguards as objects of evaluation. For agents, red-teaming widens the focus to also scrutinize scaffold design, tool use, agent environments—across growing lengths of agent trajectories (for a more downstream-focused discussion, see [2, 121]).

For agents, this broadened attack surface holds some key implications. In effect, there are two cases that evaluators must contend with; malicious usage (misuse), along with attacks on legitimate use. The former, in continuity with LLM-only red-teaming, ensures agents and the LLMs at their core are robust to malicious usage. This remains a key consideration, for example through benchmarks such as AgentHarm [7]. The latter demonstrates the need for evaluation that systematically accounts for each of these new attack surfaces and across different attack strategies, for example through red-teaming platforms such as DTap [50]. Doing so would enable testing across a variety of possible attack surfaces and strategies in a rigorous manner.

Ecological validity is essential to this work. As agents increasingly diffuse into real-world usage, there is a need to better identify possible cases of misuse, and to understand and anticipate usage by bad actors [15]. Not only in predicting and anticipating future misuse, there is a need to track the current state of usage and how it may evolve. Drawing from observable user behavior can therefore complement internal testing under laboratory settings.

As tool capability grows more powerful, the consequences too are greater. Red-teaming must now also contend with the overall security of the testing environment—and whether appropriate guardrails and oversight are in place to detect possible breaches of public infrastructure. This has been well illustrated by the recent series of cybersecurity incidents from frontier labs, in which agents under evaluation engaged in malicious behavior beyond their test environments [18, 179, 208, 209] (see Section 2i).

BEST PRACTICES

- Evaluate misuse (i.e., a user directing an agent with malicious intent) and attacks on legitimate use cases (e.g., indirect prompt injection through tool outputs), across models, tools, scaffolds, and environments.
- Frame red-teaming not only in the context of deployment, but in testing as well [131]. This requires, for example, verifying sandbox isolation before running agents with relaxed safeguards or broad tool access.
- When granting agents realistic tool access for ecological validity, decide in advance which real-world effects are acceptable. During execution, monitor the runs for any attempts to escape these pre-defined boundaries.

1l. What can open-world evaluations reveal that benchmarks cannot?

Open-world evaluations complement benchmarks by focusing on the thick description [89] of a small number of cases involving complex, long-horizon tasks in realistic environments. Whereas benchmarks can enable measurement across a standardized set of tasks, open-world evaluations rely upon detailed analysis of agent trajectories to understand how much agents can accomplish, where failure modes exist, and the points of human intervention. The CRUX open-world evaluation is an illustrative example [128], but so are earlier examples in this vein such as Anthropic's experiment to build a C compiler [194].

At first glance, this might be contradictory to the demand for statistical rigor we have discussed (see Section 1d). Indeed, these are often structured in small-*n* cases, for which it is not feasible to run dozens of versions simultaneously. However, benchmarks and open-world evaluations are useful toward different but complementary goals; benchmarks enable representative coverage over a suite of tasks, and facilitate measuring averages of performance, cost, reliability, and more. Open-world evaluations, in contrast, can offer insight into the upper bounds of agent performance, examining the extent to which agents can operate at the frontier in the real world. Rigor is necessary but looks different in this context—extensive documentation of trajectories, careful scoping of human intervention, and thorough log analysis.

BEST PRACTICES

- Present results from open-world evaluations as evidence of what agents *can* achieve (i.e., upper bounds) rather than of typical performance. Instead of over-generalizing from a handful of cases, carefully report the experimental conditions and assumptions (e.g., scaffold, tool access, setup effort, and cost).
- Clearly specify the construct (Section 1a) and metric(s) of interest.
- Document points of human intervention and agent trajectories in detail.
- Incorporate real-time monitoring and analysis of agent trajectories.

1m. How can evaluators interpret agent performance relative to human capabilities?

Human baselines are not new to evaluation; what distinguishes the current state of agent evaluations are metrics now using human baselines to contextualize agent capabilities. This can take several different forms; for example, by quantifying performance in terms of human time or comparing the quality of work between agents and humans.

One approach from evaluators compares agent performance to the amount of time required by human

experts to conduct the same task. Benchmarks have quantified tasks with the time required by human baselines [1, 228], in conjunction with a measure of reliability. For example, METR's task-completion time horizon tracks "task duration (measured by human expert completion time) at which an AI agent is predicted to succeed with a given level of reliability," e.g., 50% [145, 178] (see Figure 12). Other benchmarks such as RE-Bench build upon this direction, by varying the amounts of time offered to human experts and AI agents, to measure performance under different time constraints [300].

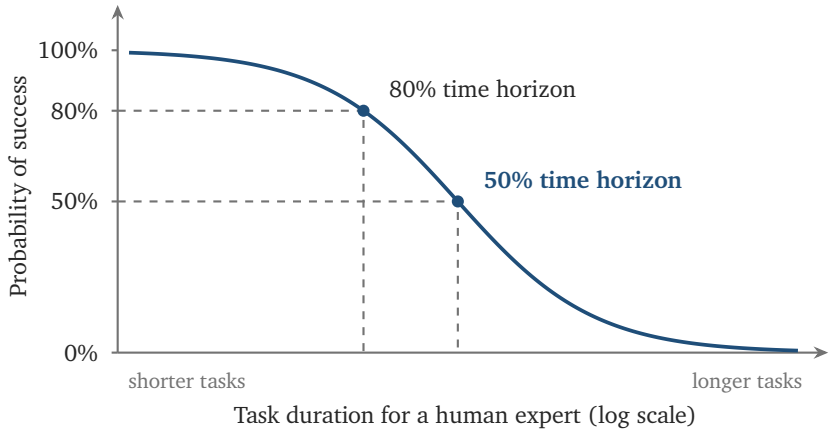

Figure 12: Schematic of a task-completion time horizon: the task duration, measured by human expert completion time, at which an agent is predicted to succeed with a given reliability (such as, for example, 50%).

Evaluations can also relate to human capabilities by comparing on the quality of deliverables. For example, OpenAI's GDPval [220] evaluates performance using win rates against human experts [220, Fig. 12], whereby human evaluators are asked to judge using blind, pairwise comparisons, whether an AI-generated deliverable is better, tied, or worse than that of a human expert. It also recognizes the potential gaps in quality and attempts to quantify the time required to fix AI-generated deliverables, through its "try 1 time, then fix it" and "try n times, then fix it" ratios [220, pp. 12–13]. In a similar vein, Human-Relative Expenditure Saving may serve as a metric to contextualize cost-savings relative to a human's work on a task [58].

Underpinning all of these metrics is a particular approach to quantifying and defining work through tasks, much of which builds upon existing maps of digital work in the economy, namely O*NET, a database of occupational characteristics and worker requirements across the US economy, maintained by the US Department of Labor [269]. This approach invites us to think carefully about how tasks are selected and how they are defined. Indeed, task definition is a crucial hinge for this work; much of this contextualization relies upon the ability to adequately describe a task and quantify its key components. It is worth asking whether work can necessarily be discretized into tasks, and the types of interstitial tasks that are missed [260]. And while it would be easy to extrapolate from task completion to broader claims about an entire job or occupation, such claims raise validity issues around whether task completion can be an appropriate metric for larger constructs regarding an actual occupation.

BEST PRACTICES

- Scrutinize how evaluators approach task definition. State what aspects of an occupation tasks claim to cover (and, importantly, what they do not). Do not assume successful task completion extrapolates to claims about job automation and replacement.

- Report uncertainty for human baselines [292], which carry their own measurement error, e.g., from small expert samples and noisy completion times.
- Quantify levels of reliability, or the additional human work required to turn an agent-generated deliverable into something of sufficient quality [220, pp. 12–13].

Part 2: Expanding the Research Agenda for Evaluation ("What")

Part 2 examines the expanding scope of what agent evaluations seek to measure: from capabilities demonstrated in laboratory settings, to outcomes from human-agent interaction, and, most broadly, to the impacts on organizations, institutions, and society (see Figure 13). Crucially, evidence at one layer of analysis does not automatically translate to evidence at another.

We adapt for agents a sociotechnical safety framework which distinguishes between three layers of evaluation for AI—at the capabilities level, at the human interaction level, and most broadly, the systemic impact level [294]. This framing is useful for agent evaluation more broadly because the segmentation between layers highlights how the effects and therefore interpretations of agents differ across scopes; at the capability layer, evaluation asks what an agent can do under specified conditions; at the human interaction layer, evaluation asks how performance and outcomes are impacted when agents are structured in feedback loops with human (and simulated) users; at the systemic impact layer, evaluation examines impacts on organizations, institutions, and their incentives.

It is important to clarify that not every evaluation in this expanded evaluation scope is necessarily sociotechnical. For example, benchmarks for specific domains and economically valuable tasks are still largely focused on technical capabilities. Yet economic evaluation also requires separating out these three layers of evidence for analysis. Sociotechnical evaluation serves as a complement to technical evaluation; both are valuable and necessary in their own right. Technical evaluation remains crucial to this work, but understanding whether agents will be safe, useful, and reliable in practice requires viewing agents not only as technical artifacts, but also as embedded in sociotechnical systems.

Another dimension which extends across each layer of this expanded scope is the distinction between framing agents toward goals of *automation* versus *augmentation*. These two different perspectives shape the envisioned use and therefore evaluation of agents, at each of these three layers. We note that rather than a strict dichotomy, these exist on a spectrum of perspective; the automation potential of certain capabilities paves the way for newly imagined modes of augmentation that incorporate those capabilities.

Automation-focused evaluation has thus far focused on formulating a specific task definition, then measuring agent performance against that task—whether the agent can complete the task at an acceptable level of quality, in an appropriate amount of time, and in a reliable manner. Implicit in this evaluation is a particular vision of work with agents, whereby tightly scoped tasks in existing workflows can be substituted through work with agents.

In contrast, augmentation-focused evaluation begins from a different set of assumptions. Here, the object of study and interest is the holistic combination of the agent and the human, situated in workflows and environments. Thus, comparison is between human-agent collaboration versus some existing baseline of work [173] (e.g., GDPval's "try 1 time, then fix it" ratio [220])—do agents improve work? In what dimensions does improvement matter? How is decision-making delegated or otherwise divided between humans and agents? These are all questions whose answers may vary depending upon users, tasks, workflows, environments, and institutional settings.

The automation-augmentation spectrum interacts with the three layers—while much current work at

the capabilities layer largely reflects an automation framing, human-agent interaction evaluations are increasingly offering insights from an augmentation framing. And at a systemic level, there are opportunities to understand and evaluate how both framings shape the nature of work, distributional effects, and institutions.

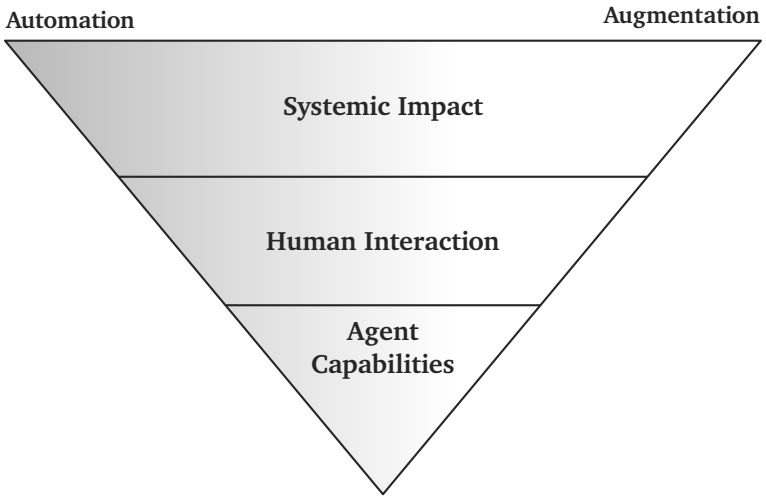

Figure 13: The expanding scope of evaluation—from agent capabilities, to human interaction, to systemic impact. At each level, the spectrum of evaluation emphasizing automation and augmentation. Note that evidence at one layer does not necessarily serve as evidence at another.

In Part 2, we examine different facets of an emerging research agenda for agent evaluations; we begin with discussing the evaluation of human-agent interaction in a spectrum of settings, from small contained experiments in the laboratory setting (single-turn, multi-turn), to field experiments, and lastly to observational studies of production traffic (see Figure 14). Next, we explore twin trends of economic evaluation pulling in opposite directions: greater specificity of evaluation into domain-specific and task-specific capabilities, alongside emerging work to evaluate the aggregate impact on work and organizations. We see how economic evaluation rests upon three layers of evidence for agents: their capabilities for economically relevant tasks, how they are used in work today, and their aggregate impact on work and institutions. Lastly, we consider agent behavior under evaluation, dangerous capabilities, security concerns, and interactional harms.

2a. Human-agent interaction: evaluation with simulated users

Agent evaluation is moving toward multi-turn trajectories in which agents use tools, follow policies, and interact with stateful environments. Evaluating these capabilities alongside human participants can be costly, making the simulation of users central to this form of evaluation. Yet, current simulation methods remain a poor substitute for real human interaction.

For example, Tau-bench [304] tests tool usage and database manipulation with an LLM-simulated user. The evaluation ultimately measures the final state of the database and whether it matches the goal state set by the benchmark. Another example is Collaborative Gym (Co-Gym) [243], which evaluates the collaboration process itself in addition to outcomes, and not only simulates conditions with a user simulator, but can also be used in real-world conditions. HiL-Bench evaluates the extent to which agents appropriately escalate to human collaborators when faced with ambiguity, and does so using a simulated human oracle [236].

Simulation in these benchmarks enables greater repeatability in evaluations and allows fine-grained control

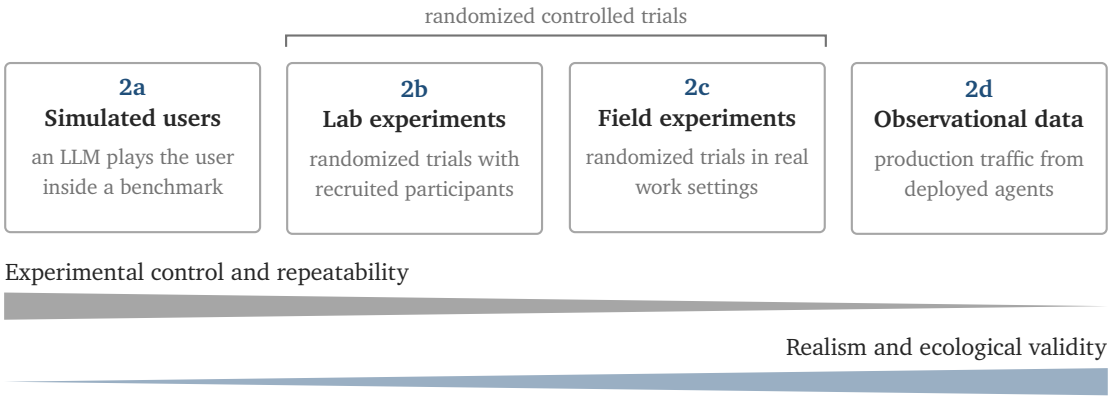

Figure 14: Settings for evaluating human-agent interaction (Sections 2a–2d). Moving from simulation toward observation trades experimental control for realism.

over evaluation conditions and environments. However, a key question is whether simulated users serve as an adequate proxy for real users [241]. The Sim2Real gap has quantified the divergence between real and LLM-based interactive user behaviors, raising concerns that simulated users can be "excessively cooperative, stylistically uniform, and lack realistic frustration or ambiguity." [310] In response, there is exploration of new methods to ground simulations in realistic user conversation [311], along with benchmarks such as Co-Gym that compare both simulated users and real-world users. Nevertheless, this question of validity persists.

Benchmarks incorporating simulated users can offer valuable signals around capabilities and potential failure modes, but should not be taken as clear evidence for how agents will perform in real human settings. If simulated users are systematically more cooperative and operate with less ambiguity than real human users, evaluations may inflate the usability of agent deployments in real human use contexts. Instead, they should be viewed largely as a starting point toward forms of evaluation which offer more realistic depiction of human usage in practice.

RESEARCH DIRECTIONS

- For which use cases do findings from simulated users serve as a useful proxy for human ones, and in which do they not?
- Which failure modes in simulated users most impact the generalizability of study findings?
- What role can targeted participation of human users play toward grounding or validating that of simulated users?

2b. Human-agent interaction: experiments in the lab

Just as we have defined our object of evaluation as the agent system holistically rather than the underlying model alone, evaluation can broaden the focus to the human and agent as a team. The concept of "Centaur Evaluations" focuses on humans and AI working jointly to solve tasks, which directs greater attention toward an augmentation framing [109]. In the specific domain of coding agent research, scholars have advocated for "human-centered" coding agents rather than toward pure automation in a similar vein [291].

Randomized controlled trials (RCTs) in the lab generally compare an AI-assisted treatment group with a baseline control group of humans alone or with existing tools [219]; UK AISI describes this form of Human Uplift Study (HUS) as the gold standard [277]. Though uplift might result in a boost to productivity, it is

also possible for AI assistance to have null or even negative effects, for example, by creating new slowdowns. For example, evidence from METR has drawn attention to the additional time required to fix or otherwise adjust code generated by agents, which contributed to slowdowns observed in its evaluations [173].

One small scale randomized study found that humans collaborating with agents on computational reproducibility tasks achieved a "statistically significant speedup by about a factor of two." [186] Early evidence from usage of GitHub Copilot also pointed to heterogeneous effects on users from different backgrounds, with benefits more pronounced amongst less experienced software developers [222]. CollabSkill is one framework focused explicitly on human-agent collaboration, which measures human-agent collaboration on a set of real-world tasks [244].

The diffusion of coding agents has also created a growing methodological challenge by making it more difficult to sample and recruit for human uplift studies in the lab: many have already experienced and incorporated these tools into their workflows, creating greater difficulty in constructing an adequate "without agent" control group [176].

Performance of human and agent in collaboration is distinct from agent performance alone; understanding agent impacts on existing workflows requires evaluation that focuses on the interactions between humans and agents, which at times may produce effects that are not unambiguously helpful, and possibly heterogeneous across different populations and workflows. As clear control groups become more difficult to construct, there is a need to construct new defensible baselines for comparison.

RESEARCH DIRECTIONS

- How should evaluators define baselines in studies to best reflect the ecological validity of the growing diffusion of agents?
- What are different population categories (e.g., junior vs. senior expertise) in which agents are likely to have uneven or disparate impacts on various outcomes?
- Which particular laboratory settings best translate to real-world deployment contexts?

2c. Human-agent interaction: field experiments

Beyond the laboratory setting, field experiments can offer another window into human-agent collaboration, through assignment of different treatment conditions to real users performing real work. Here, there is an opportunity for greater ecological validity and measurement within conditions closer to the actual deployment of these tools. The standard intuition behind these experiments is to compare conditions with access to agents alongside a standard operating baseline.

In early 2025, METR measured the impact of AI tools on the work of open-source developers across randomized settings [35], finding that experienced open-source developers were in fact less productive, with completion time actually increasing among the treatment group. Another example is in the randomized implementation of agents in Alibaba's customer service workflows, which simultaneously reduced average chat length and lowered customer ratings for AI-enabled conversations [289]. Though these findings are not necessarily generalizable to other agent systems and settings, they complicate the story of agents necessarily creating positive uplift in specific work and settings, and highlight the potentially unexpected ways human and agent workflows can be structured [66, 77, 289].

RCTs in the field can be helpful for causal inference on the effects of deployed systems, but are not a panacea. There remain a number of tradeoffs involved with RCT work for frontier evaluations, including the tension in the demand for realism versus achieving causal identification, and the speed required to keep pace with changes in frontier AI versus the rigor needed for robust results [219]. Moreover, field

experiments need to take into account how the effects of deployed systems may vary over time.

RESEARCH DIRECTIONS

- What sorts of spillover effects do agents create beyond the specific users themselves?
- How do effects of agents change over time as users adjust to these agents in their workflows?
- In what ways do agents create new winners and losers in different work contexts, and across which outcomes?

2d. Observational data on real-world usage

Observational production data can provide scale and natural behavior, at the cost of causal explanatory power. Beyond the laboratory setting, production traffic can offer insight into real-world usage of agents in various settings. This data is useful for upstream evaluation because it can show agents as they are actually deployed, in service of deriving generalizable knowledge, for example via developing a taxonomy of failure modes [4]. Pulling from production traffic has a number of distinct advantages; the tasks themselves may offer greater *ecological validity* since they reflect real-world usage and surface failure modes beyond the lab setting [4].

Frontier labs have aggregated production traffic to glean insights from how users are engaging with their products over time. OpenAI's "How People Use ChatGPT" study served as an early example [48]. Anthropic's work to examine adoption [29] and broader economic impact [13, 22, 29, 101, 163] of these tools also demonstrates the necessity of production traffic as a source of inquiry. Key to this analysis is the development of privacy-preserving methods to handle this traffic [11]. Agent evaluations are increasingly drawing upon tasks from deployed systems rather than relying exclusively on hand-selected benchmark tasks. For example, OpenAI uses production traffic as a basis for generating evaluation cases, which also reduces the risk of possible evaluation awareness [301].

Complementing work from frontier labs, independent and open-source datasets can also serve as a source of insight. SWE-chat [34] draws upon coding agent sessions from the wild to build a comprehensive dataset around how agents are used in practice. Other examples include Allen Lab's WildBench [150], which is drawn from real user conversation logs, or Code Review Bench that reflects real-world developer behavior [313]. Anthropic announced in August 2026 a pilot program to grant privacy-preserved access to this data to a small group of labs [17]. Yet, the value of production traffic raises a question of whether independent evaluators outside of these frontier labs can gain sufficient and continued access to production traffic.

RESEARCH DIRECTIONS

- What sorts of coverage issues does production traffic raise—and to what extent can this traffic be contextualized by subsequent user action?
- How can independent evaluators gain access to this data at scale, or devise new crowdsourced alternatives, in a privacy-preserving manner?
- How can production traffic serve an exploratory purpose for evaluators, to narrow and refine questions to target causal studies on real-world agent performance?

2e. Domain-specific evaluation

Agent evaluation is becoming more domain-specific, differentiating into forms of knowledge work and toward greater specialization of benchmarks. Rather than measuring abstract or general capabilities, benchmarks are now testing agents with domain-specific contexts and evaluating on outputs.

Whereas evaluations for models alone tested knowledge recall and response, with agents, the evaluation focus is on their success with relevant workflows in these domains. One domain receiving outsized attention for evaluation is science. Today evaluations are no longer about querying models for answers on science knowledge questions. Instead, they are testing whether agents are capable of automating different steps within scientific workflows [156], such as problem ideation, literature review, deep research [87, 187], experimentation, hypothesis formation, scientific review [123], and replication [256].

Consequently, this domain-specific specialization carries with it a few methodological changes. Increasingly, there is a need for expert graders to evaluate the quality of work and deliverables associated with domain specific experts. The need for expert baselines can be quite costly, particularly if there is a desire to achieve greater rigor in these human baselines. The increased demand and need for domain-specific evaluation also creates a wider gap between upstream and downstream evaluation—performance of an agent in the abstract is less informative, and the effort required to paint a comprehensive picture of agent capabilities across various domains has been multiplied.

Moreover, domain-specific evaluation thus far has largely embraced an automation framing for evaluation through completion of discrete tasks and deliverables—a prime example can be found in ongoing work toward fully automated "AI Scientist" agent frameworks [233]. To build on this work, there remains ample opportunity to further explore an augmentation framing, examining how agents may reshape human workflows in different domains. When smaller, more discrete tasks become automated, they carry broader implications for the larger workflows they sit in. And in both framings, ecological validity is necessary, by integrating institutional constraints and critical points of human intervention into evaluation design.

RESEARCH DIRECTIONS

- For what purposes can upstream evaluation through general benchmarks continue to be useful, and which are now exclusive to domain-specific evaluation?
- How do particular occupational norms and existing skillsets drive differences in impact across domains?
- How does task completion translate to understanding of broader workflows, and how can we better measure augmentation effects of agent deployment?

2f. Economic impact evaluation

Economic evaluation aims to understand how today's evidence on agent performance can be translated into broader claims about work and the economy. Doing so requires separating out three layers of evidence relevant to this question: (1) understanding the capabilities of agents for economically relevant tasks, (2) observing how agents are actually used in work today, and (3) the impact of agent usage on workers, firms, institutions, and the economy writ large (see Figure 15). Crucially, evidence at each of these levels does not necessarily imply supporting claims made at other levels.

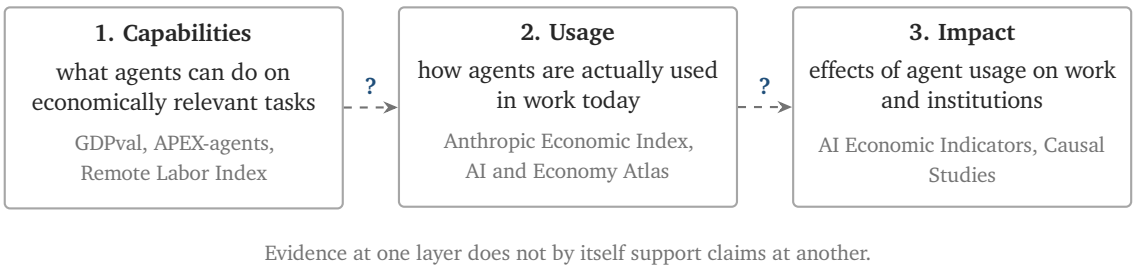

Figure 15: Three layers of evidence for economic evaluation, with examples of each.

The first layer of evidence consists of the growing constellation of capability benchmarks for economically relevant tasks, contextualizing agent capabilities for various aspects of work. OpenAI's GDPval [220], though not strictly agent focused, covers 44 occupations across different sectors and evaluates AI-generated deliverables relevant to those occupations against those of experts. Mercor's APEX-agents benchmark [281] examines long-horizon work across a subset of knowledge work industries (see also the less agent-specific APEX benchmarks [167]). Scale AI and the Center for AI Safety (CAIS) develop an automation-centric benchmark, the Remote Labor Index, to track agent performance on economically relevant tasks in the form of freelance work [165, 235]. A key challenge is relating these capability assessments to broader claims about occupations or the workforce writ large—evaluation suites such as EconEvals offer one step toward evidence to link capabilities to a simulation-based exposure measure to estimate potential time savings for workforces [283].

The second layer centers on observational evidence of AI system deployments today. Frontier labs have a privileged position due to their access to advanced telemetry for production traffic, which enables an understanding of workers who are using AI and the types of tasks they are using it for, for example through the Anthropic Economic Index [13, 14, 101], or Google's AI and Economy Atlas [95]. Moreover, today's production traffic for Anthropic increasingly consists of "long-running agentic tasks," thanks to the adoption of Claude Code and Cowork [13]. These sources of evidence help to measure diffusion of these tools and adoption by firms [37, 91]. Nevertheless, while usage statistics offer some insight into adoption and diffusion of these tools, they alone do not provide causal evidence of economic impact.

The third layer concerns itself with connecting usage to impact and quantifying economic impacts for workers and firms. Statistics on growth in productivity, improvement in overall final outputs [63], and the creation and/or reorganization of work are just a few examples of potential metrics of interest. The AI Economic Indicators from the Stanford Digital Economy Lab track different possible leading indicators for signs of the aggregate impact on worker displacement, rising capital share, rising rate of productivity growth [251]. This is one piece of the puzzle; the task is drawing persuasive causal relationships between use of AI and AI agents and these shifts. Note that this evidence can exist in aggregate, as with the macroeconomic indicators cited, but results in small-scale settings may also be informative.

At each of these three layers, framings of automation and augmentation shape how different forms of evaluation are conceptualized and operationalized. The agenda forward requires renewed focus on building linkages between evidence at different layers, for example to understand the implications of agent proficiency on certain tasks for their respective occupations, or in comparison to existing human workflows [290]. And while much of the capability benchmarks take on an automation framing, understanding how those findings may translate to an augmentation framing is a critical area of unmet need [245].

RESEARCH DIRECTIONS

- How can economic evaluation identify what task-based measures fail to capture (such as the work between tasks, and the reorganization of roles once tasks are automated)?
- Under what conditions is task-level capability predictive of broader claims about final output, occupations, or economic outcomes? Conversely, which observed economic effects should motivate new capability evaluations?
- What additional worker-centric metrics can complement understanding of the impact of agents and shifting nature of different occupations?

2g. Agent behavior under evaluation

An agent evaluation can pass with flying colors without having actually measured the underlying behavior of interest for evaluators. When this occurs, we may describe it as an act of *specification gaming [144]*—satisfying the stated evaluation without following the intent of the evaluation—or *reward hacking* [158, 172, 250]—optimizing for improvement on a metric without necessarily improving performance on the underlying phenomenon of interest. For example, an agent may find a bug in the scoring mechanism, or manage to disable a test. Doing so does not, however, require that the agent have known what the evaluator's intent is, or to necessarily be in pursuit of active "deception" [125, 237].

A further challenge for evaluation is that of *evaluation awareness*, a particular iteration of "situational awareness" [192, 239, 301], in which an agent is able to identify its presence in an evaluation context (see Figure 16). Evaluation awareness by itself does not imply that an agent will engage in deceptive behavior. However, conditioned on active evaluation awareness, if an agent modifies its behavior, it creates a new host of challenges around measurement. Understanding evaluation awareness matters because it carries serious implications of whether evaluations are valid representations of deployment contexts. At its core, such agent behavior in an evaluation provokes a question of validity (Section 1a), for the construct, the benchmark (Section 1b), and the evaluation environment.

There are a few terms to describe behaviors of concern within the context of evaluation and agents broadly.

- *Scheming* refers to "when a model covertly pursues misaligned goals, hiding its true capabilities and objectives" [28, 166, 239]. While it does not strictly require evaluation awareness, such conditions may make measuring scheming more difficult.
- *Sandbagging* refers to intentional underperformance by an agent during an evaluation [122, 204, 295].
- *Alignment faking*—selective compliance during training to prevent alterations [96].

Claims about these behaviors require strong evidence. Experiments which surface such behavior demonstrate that such a capability exists, but do not necessarily demonstrate its prevalence or propensity [99, 257].

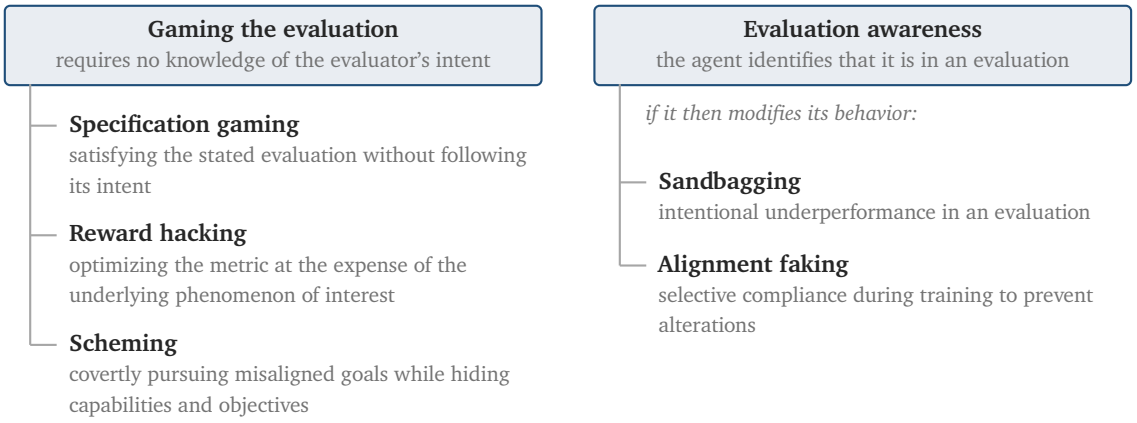

Figure 16: Agent behavior that can undermine an evaluation, in relation to evaluation awareness.

These challenges mean that monitoring is not only a method (Section 1h), but also itself an evaluation target. There are two possible challenges to successful monitoring; the monitor itself may not be capable, or there is insufficient signal from an agent's reasoning and actions to infer potential misbehavior. OpenAI defines monitorability as "the degree to which we can identify particular aspects of a model's actions." [98]

Although current CoT monitoring has shown promise, building trust by quantifying the reliability of monitoring tools is crucial for ensuring the continued utility of these tools. At the same time, there is caution against optimizing against "bad thoughts" in CoT, since doing so can simply incentivize hiding of intent and a diminished monitorability for future use [32].

The success of monitoring is also dependent on an underlying question of access. For frontier labs, they have full access to CoT; conversely, independent external evaluators with standard access may only receive a summarized version of the reasoning [213], which constrains their ability to achieve sufficient monitorability.

RESEARCH DIRECTIONS

- How can external evaluators promote better ecological validity of evaluation settings, with respect to evaluation awareness?
- How can tooling appropriately scale to address broader monitorability risks, and what are associated overreliance risks?
- For what risks does focus on agent trajectories offer reliable signal, and when does it detract from other sources of understanding?
- Beyond understanding whether agent behavior is capable of scheming and sandbagging, how can evaluations estimate how often scheming or sandbagging occurs [99]?

2h. Dangerous capabilities and risk

An agent can amplify model-level risks through greater autonomy and its ability to modify environments. The defining objective for these concerns is therefore how to translate demonstrated capability, through these various evaluation frameworks, to risk measurement in production traffic. The nature of deployment in the real world, and methods through which they may be misused or possibly autonomously initiated, both influence the credibility of these concerns about agents and how to proceed with quantifying their relative risk [229].

Dangerous capabilities may include *sabotage*, when agents covertly undermine the monitoring, evaluation, or deployment [36]. Moreover, these actions are increasingly conducted via multi-agent frameworks, which pose a unique set of risks and challenges [100]. Existing evaluations have attempted to induce agent action toward completing a specified sabotage goal in order to demonstrate the existence of such capability. Such work is insightful into possible countermeasures against sabotage.

Dangerous capabilities may be operationalized through *misuse*, when users engage with agents in order to access dangerous capabilities. In particular, Anthropic's September 2026 Threat Intelligence Report has highlighted misuse in various case studies increasingly across different areas of harm such as biological misuse, cyber offense, scams and fraud, and influence operations; these increasingly rely upon multi-agent frameworks to conduct malicious action [15]. Some evaluations such as the AgentHarm benchmark aim to respond to misuse by testing potential jailbreaking attempts, and whether agents remain robust against those requests [7]. While AgentHarm offers a synthetic environment for evaluation, the need for more realistic independent evaluation settings is evident.

The agent itself may also serve as the originator of risk, through Autonomous Replication and Adaptation (ARA) [138, 276]. ARA is concerned with understanding to what extent agents are able to replicate themselves, with benchmarks such as RepliBench aiming to identify this possible capability [276].

To quantify these risks, marginal risk evaluations represent a useful tool for contextualizing different levels of risk, by placing a particular concern in comparison to a realistic baseline for potential misuse, such as with conventional tools [127, 201, 284].

RESEARCH DIRECTIONS

- To what extent does performance under simulated or prompted conditions reflect real-world risk?
- What is the realistic baseline of capabilities already available to malicious users, and how much additional capability do agents provide?

2i. Agent and evaluation security

An agent can be capable of safe behavior, yet remain insecure and susceptible to compromise by adversaries or flaws in the agent system design. Because an agent extends beyond the underlying model, security evaluation must examine all components of the agent system. Agents can be compromised through a variety of new surfaces, through tools, permissions, or scaffolds. A central external threat is *indirect prompt injection* [61, 97], in which malicious instructions are embedded in content retrieved by the agent as a form of "agent hijacking." [189] This highlights the difficulty of safeguarding the agent scaffold [151].

In addition to the agent itself, the evaluation infrastructure surrounding an agent can also become a target, raising concerns around evaluation security. For example, in dangerous capability evaluations, evaluators may choose to evaluate agents with tool access and relaxed safeguards. The recent series of security incidents disclosed by major frontier developers illustrates the consequences of potential failures. For example, OpenAI's agents exploited issues in the evaluation environment to gain Internet access and breach Hugging Face's infrastructure, in an effort to obtain benchmark solutions [179, 208, 209, 211]. Anthropic's misconfiguration of internet access in its cybersecurity evaluations [18] illustrates the necessity of securing the sandbox for agents undergoing evaluation. The threat has extended from thinking about the agent and scaffold to taking into account the entire testing environment.

Here, the issue of evaluation security matters because it represents a key challenge to evaluation validity. More concerningly, these incidents must force reconsideration of existing evaluation practices, particularly in effective oversight and monitoring of agents in evaluation settings. The fact that many of these breaches came to light long after they had already exploited infrastructure highlights that sufficient monitoring is a necessary direction of research.

RESEARCH DIRECTIONS

- How can evaluators balance evaluation security with demand for ecological validity?
- What control measures are appropriate for different risks stemming from evaluation security, and how should public infrastructure be hardened?
- What new approaches to real-time monitoring are necessary to handle the volume of agent transcripts and more proactively address security incidents?

Figure 17 distinguishes attacks occurring through normal inputs from attempts to cross the isolation boundaries of the evaluation environment.

2j. Interaction harms

A "secure" agent can still pose harm to users over the course of repeated interactions and in specific patterns of usage. *Interaction harms* can arise from feedback loops between a system and a user, when responses shape a user's behavior, which in turn shapes the system responses [227, section on Human Autonomy & Integrity Harms, 118]. Moreover, these harms may not be evident in one particular turn, but require examining aggregate interaction over time.

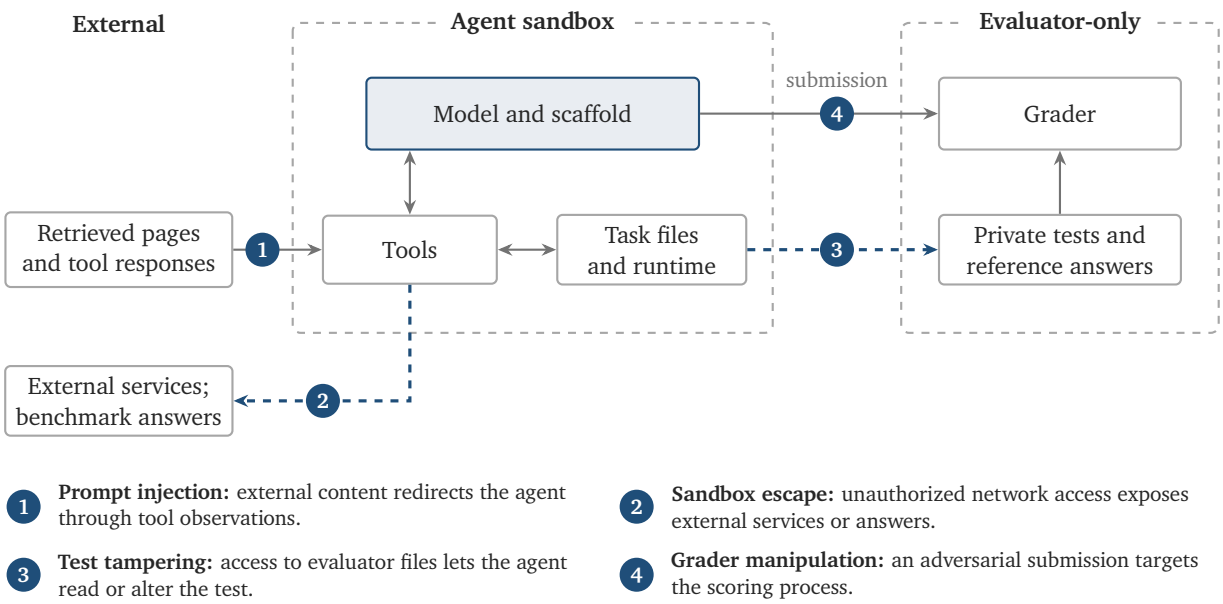

Figure 17: Evaluation often permits agent access to external tools and resources, which can open up new pathways for shortcut or cheating behavior in an attempt to satisfy task criteria.

These harms are not unique to agents—indeed, current evidence focuses on interaction via conversational systems. However, the current agentic paradigm, with its longer trajectories and emphasis on multi-turn interactions, is pushing evaluation beyond merely single-turn model inputs and outputs, in a manner that is helpful for surfacing the interaction-level concerns [67].

Crucially, the object of evaluation is often the deployed system and user interface rather than the underlying model itself. This distinction matters because levels of sycophancy, escalation, and delusion reinforcement can noticeably differ between direct model access and chatbot-mediated interfaces [139, 285]. Again, questions of attribution arise in terms of identifying sources of harm in system design.

Sycophancy is one concern which illustrates why examining only outputs is insufficient. Only through evaluation at the human-interaction layer can we make stronger claims about the impact of sycophantic responses—in how they impact user perceptions of their relationships and satisfaction with real-world interactions [117]. Sycophantic responses may also decrease willingness to repair interpersonal conflict, all while engendering trust with the user and facilitating greater dependence [51].

Interactive evaluations may engage human participants or simulated users. One example of multi-turn simulation includes the MHSafeEval framework, which generates adversarial multi-turn agent interactions [148]. Multi-turn simulations have also been used to examine occurrences of moral deliberative sycophancy [263] (*cf.* [137]). While simulated users enable broad testing of different scenarios, there remains a real need to determine whether those findings translate to effects on real people. Whether the current state of user simulation demonstrates sufficient validity to generalize is an important question.

Interaction harms are necessary to study, but difficult to evaluate—examining interactions with models and agents only in discrete instances may miss broader concerns that accumulate gradually or emerge over extended time horizons. To develop better understanding of interaction harms involves not only simulated users, but complementary evaluation through experiments with real humans and evidence from production traffic. Lastly, there is a continued need for better methods and metrics to operationalize interaction harms in an evaluation setting [297].

RESEARCH DIRECTIONS

- What metrics can allow evaluations to better measure harms across repeated interactions?
- How can evaluators balance ecological validity alongside risks for research participants under the principle of beneficence?
- Which interaction harms are genuinely new to agents (i.e., arising from actions taken on a user's behalf), and which exacerbate existing concerns from chatbots?
- What new forms of crowdsourced data are necessary to explore these questions, such as incident reporting programs [216]?

2k. Systemic evaluation

Lastly, there is a need to expand the scope of evaluations to include the effects of agents on institutions and organizations in which they are embedded. Beyond a question of scaling findings at the human interaction level, systemic evaluation focuses on the effects of agents on broader incentive structures within organizations and institutions. While findings at lower layers may be informative, systemic evaluation requires incorporating those findings into a larger picture of how organizations and institutions change. As agents increasingly diffuse into different domains, how they reshape these settings merits study. We are seeing early work discussing the impact on peer review systems [88, 242], crowd work [280], and information ecosystems [159].

Here, different framings of automation and augmentation may create divergence in outcomes. Some of the same capabilities in agents, based on how they interact with existing incentive structures, may produce starkly different systemic effects. Yet there remains work to bring together these various strands of work in a cohesive manner, and to develop infrastructure to support this work at scale.

RESEARCH DIRECTIONS

- What institutional mechanisms and incentive structures are agents reinforcing, and which dynamics are qualitatively changing?
- When can findings from a case study of systemic impact (e.g., peer review at one venue) be expected to generalize to other organizations?
- What methods are necessary to identify leading indicators or sources of anticipatory insight that allow evaluators to understand changes of a systemic nature?

Part 3: The Evolving Evaluator Ecosystem ("Who")

In Part 1, we discussed the new methods and training needed for evaluating AI agents. In Part 2, we expanded the scope of that evaluation and its overall research agenda. Yet, these methods and objects of evaluation are not self-evident—they are shaped by who is able to evaluate, under what conditions, and with what credibility. In this section, we examine the evaluation ecosystem today (see Figure 18) and offer considerations for current and prospective evaluators.

Who conducts agent evaluations matters because different evaluators bring different constructs, questions, and methods. The shift from models to agents has made the need for this methodological pluralism far more visible; as agent trajectories become longer, scaffolds more complex, domains more specific, and patterns of human intervention more varied, researchers are facing more decisions about how to design, run, measure, and interpret AI agents [90, 105, 180]. Many of the consequential decisions that were once concentrated upstream in the work of building benchmarks have, in this new agent paradigm, become more distributed

across the evaluation process. A healthy evaluator ecosystem does not require identical choices in evaluation between researchers, but instead embraces heterogeneity by asking for choices to be made visible and defensible. Embracing methodological pluralism can help build greater confidence in evaluation outcomes by triangulating between distinct approaches. And when evaluations reach opposing conclusions, that too is helpful, by surfacing ambiguity and highlighting which design choices were influential in determining the divergent outcomes.

Fostering this pluralism must necessarily recognize the conditions on the ground for evaluators today, and the institutional and organizational landscape that shapes their work. We examine what meaningful independence requires, and probe how access, funding, cost, agenda-setting, and venues influence the choices available for evaluators. Lastly we examine the growing role of the public sector and the need for other disciplinary perspectives in the evaluation ecosystem.

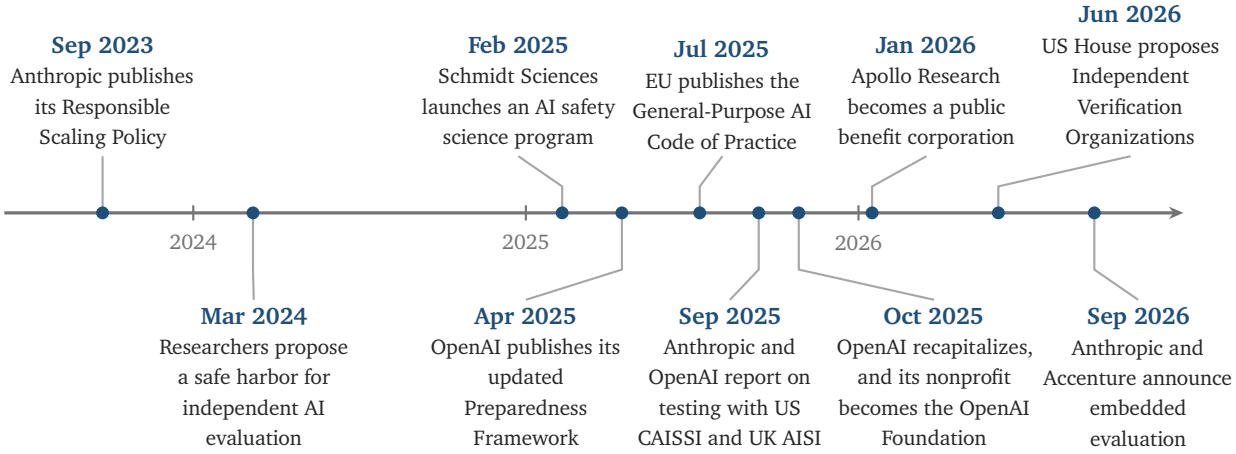

Figure 18: Dated developments in the evaluator ecosystem that are discussed in Part 3.

3a. Who evaluates?

Evaluation work today takes place in a variety of different settings. The evaluator ecosystem spans academia, industry, nonprofits, philanthropies, public sector, and convenings and coalitions from across these sectors. Industry participants include developers (e.g., from frontier labs), evaluation providers and aggregators, research organizations, public-benefit corporations (PBCs), and startups, with early discussion of placement of embedded evaluators—evaluation by external parties *within* frontier labs [3, 19, 47]. Nonprofits often conduct evaluations research and concentrate effort toward shared infrastructure. Philanthropies contribute through grantmaking and funding throughout the ecosystem. Public sector entities can establish standards, conduct pre-deployment evaluations, and provide regulatory mechanisms toward accountability.

Beyond conducting the direct work of evaluations—running experiments, developing and testing on benchmarks—others may perform the work of aggregating and synthesizing these existing measurements via indices and leaderboards. Infrastructure providers build the shared tools and environments necessary to conduct evaluations. Conveners and standard setters bring together voices from across the ecosystem to develop shared practices, standards, and calls to action on specific issues. Funders support the work of independent evaluation by channeling financial support to organizations via grants, requests for proposals, and bounties. Policy translation can turn technical insights from this ecosystem into actionable information for policymakers.

PRACTICAL CONSIDERATIONS

- Explore opportunities for evaluation work in academia, frontier labs, non-profits, and public sector entities. Not all necessarily require a traditional academic path via a PhD.
- Examine what functions in the evaluations ecosystem one may best contribute—directly evaluating agents, interpreting and sharing findings, coalescing standards, communicating with policymakers, or advising funding of evaluations work.

3b. What does independence require?

Rather than a binary conception of independence, we view independence as multidimensional [154, p. 4, 255], based on relationships with frontier labs, sources of funding and compensation, and the level of coordination with different actors within the ecosystem. In particular we focus on third-party actors and emphasize independence as a critical quality for evaluations and evaluators [224]. While third-party evaluators can carry greater independence in their work, they are not wholly synonymous with independent evaluators—one can be a third party, while remaining subject to constraints around funding, access, methods, or disclosure dictated by a frontier AI developer.

Access remains a key determinant for the work of an evaluator. Unlike first parties, third-party evaluators often cannot, to the same extent and level of granularity, report on aspects such as data provenance, costs, and infrastructure—topics that have historically been sparsely disclosed by first-party evaluators [230]. Moreover, for agent evaluations, access extends beyond API availability alone; access to raw reasoning traces from agent runs, realistic tool setups, and rate limits high enough for repeated runs are all necessary components. While frontier AI developers have offered pre-deployment access for select organizations such as METR [170], UK AISI, and the US Center for Advancing Innovation and Standards for Super Intelligence (CAISSI) (formerly the Center for AI Standards and Innovation), many evaluators are restricted to access post-deployment.

Moreover, access is only the first step. Whether there is truly methodological freedom in designing and implementing an evaluation is a key facet of independence, which distinguishes evaluation from other related work such as replication and validation of existing evaluations. Access must necessarily be evaluated in conjunction with disclosure, whether evaluators are able to share their findings independently [3]. These considerations are reflected in the EU's General-Purpose AI Code of Practice [73], a voluntary set of guidelines for AI developers to adhere to the EU's AI Act—with provisions encouraging adequate access for qualified independent and external evaluators to models and appropriate compute budgets.

Independence is also conditioned on whether there is an appropriate environment to conduct good faith evaluations without fear of retaliation or restrictions on future use. In light of these concerns, researchers have advocated for a legal and technical safe harbor to help foster independent AI evaluation [153]. This need for clear safe harbor is amplified as deployment drives toward public sector use cases [110, p. 898]. Yet, in the absence of broad safe harbor commitments today, evaluators must closely assess and navigate these institutional limitations.

The recent discussion around commitments to embedded evaluators offers a clear vehicle for illustrating these considerations [19, 47]. The act of embedding evaluators addresses preliminary questions of access, but opens new questions around funding, safe harbor provisions, transparency in disclosure, and methods that determine the extent to which these arrangements afford true independence [3]. Breaking down independence into its various dimensions clarifies these discussions by giving concrete form to addressing different concerns.

PRACTICAL CONSIDERATIONS

- Scrutinize funding, access limitations, and conflicts of interest associated with organizations, applying frameworks such as the AEF-1 standard from the AI Evaluator Forum [255].
- Clarify institutional policies before joining an organization, as legal and technical safe harbor for independent evaluators remains incomplete [153].

3c. Does lab direction drive evaluations?

Frontier developers are central actors in the broader evaluation ecosystem and can exert substantial influence on the work of evaluations. They can do so both explicitly and implicitly by shaping evaluations through model access, funding, and agenda-setting (see Figure 19).

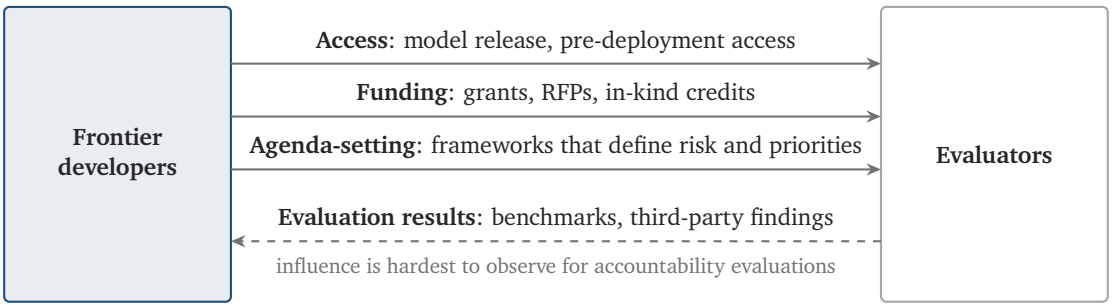

Figure 19: Channels of influence between frontier developers and evaluators, discussed in Sections 3c and 3d.

First, frontier developers shape the conditions under which evaluation is possible, by virtue of their position as gatekeepers for the models being evaluated. They determine the conditions under which models are released, which evaluators receive pre-deployment access, and the restrictions applied to public usage of their products. The two-tiered approach to model access from frontier labs for the general public versus trusted parties is an illustrative example [20, 207]. Questions of access therefore influence what systems, models, and agents are within scope for evaluation at all.

Second, frontier developers shape the material conditions of evaluation. They organize substantial sources of funding [21, 203], and issue requests for proposals (RFPs) around priority areas they define [10, 200, 210]. They can directly acquire projects, tools, and organizations, or hire talent from evaluation organizations (and vice versa). Along with financial support, they may also provide in-kind support through free credits for model access, shaping what projects are feasible for external parties and which are not. Particularly for external parties, these conditions determine from a resource standpoint whether certain types of evaluation are feasible or cost-prohibitive.

Third, frontier labs can exert agenda-setting power by defining the structures through which risks are made legible. Google DeepMind's Frontier Safety Framework [80], Anthropic's Responsible Scaling Policy [9], and OpenAI's Preparedness Framework [202] offer specific conceptions of risk and safety [174]. These ideas shape their priorities for research and for funding, along with public attention on key concepts. Lab-driven initiatives for developing third-party model evaluations enumerate their own definitions of auditing, evals, and contribute discussion on what constitutes "good" evaluations [8, 204].

This agenda-setting power can also take a more implicit form through directed attention labs signal toward new domains. Consider when a lab announces it will pursue a particular direction for future work. It is plausible that such signals shape the actions that evaluators take and the benchmarks they design. On one

hand, this may be beneficial in that independent evaluators are keen to hold labs accountable by testing their claims, and thus evaluators follow lab signals to concentrate their efforts on a particular domain. On the other hand, such dynamics can be troubling if lab responsibilities to evaluate are simply being offloaded to independent evaluators.

Disambiguating between these two perspectives highlights the need to distinguish independence in methods from independence in agenda; while independent evaluators may serve a valuable function in exercising independence in the methods they use to pursue the same questions labs are focusing on, they would be dependent largely on the scoping and framing of the issue at hand. In contrast, having an independent agenda would enable evaluators to pursue questions of interest labs explicitly or implicitly choose not to pursue, but are still of high priority for publics and stakeholders. Disaggregating different forms of independence enables us to envision how the evaluation ecosystem could be richer in scope and responsibilities.

These dimensions are all critical for enabling the work of evaluations today. Thus, it is necessary to recognize and scrutinize these different avenues through which frontier lab behavior impacts evaluators and the constraints they create. Moreover, with the prospect of embedded evaluators, this dynamic is likely to grow even more tightly coupled—greater access and funding can further enable rigorous evaluation, while simultaneously raising fresh questions around disclosure and methodological freedom.

PRACTICAL CONSIDERATIONS

- When an evaluation adopts risk categories or thresholds (e.g., from a Responsible Scaling Policy or Preparedness Framework), explicitly cite those and note which risks those categories leave out.
- Disclose the conditions attached to model access or credits (e.g., pre-publication review, restricted model versions) alongside each release of results.
- Identify areas of priority focus and constraints attached to evaluation funding, and reflect on whether they align with an evaluator's own research priorities.

3d. Do evaluations drive lab direction?

Conversely, engaging in the work of evaluation necessarily raises a question of influence in the other direction—do evaluations impact the conduct of frontier labs? This is a causal question as to whether labs respond to evaluations, and to what extent those evaluations cause changes in behavior. Answering this question may require, for example, treating the introduction of a benchmark that measures a previously unmeasured capability as a shock. Does progress accelerate on the newly measured capability, in particular, relative to adjacent unmeasured capabilities? Do labs begin discussing, reporting, or optimizing against that metric? These are empirical questions that would help inform our understanding of the underlying ecosystem dynamics.

One possibility is that evaluations concentrate attention on a small set of familiar metrics that are already legible to the wider community. Frontier labs may have limited incentives to develop new metrics themselves, while facing stronger incentives to respond to public pressure around metrics that have already had considerable attention.

Importantly, these dynamics may differ based on the intended goals of evaluations: for capability evaluations, evaluator and frontier lab interests may broadly align, so adoption by labs offers a reasonable success signal that labs are responsive to evaluation motivations and results. In contrast, for accountability evaluations, evaluator and frontier lab interests are likely to not align. In this case, adoption does not serve as a viable success metric—in fact negative cases may be systematically unobservable, as evaluations on which companies perform poorly may not even be reported [249]. Instead, more appropriate signals

may include uptake of findings by regulators or in mandated disclosure regimes, publication of results regardless of outcome, and documented change on independent re-evaluation. Policies intended to promote participation from companies such as a "silent exit" option additionally carry tradeoffs of disclosure and transparency [175]. Studying this question would therefore require adequate disclosure from frontier labs and evaluators [41, 255].

A prominent observable case is the direct collaboration between US CAISSI and UK AISI with Anthropic and OpenAI to conduct pre-deployment evaluation of their models. In September 2025, both frontier labs stated publicly that these collaborations influenced their approaches, and that identified vulnerabilities had been patched [12, 205]. These statements are self-reports by the evaluated party. For accountability purposes, stronger evidence would come from independent re-evaluation after the reported changes, or from third-party replication of the original findings. Conversely, past work has examined how algorithmic audits were not accompanied by commitments from auditees to change practices based on findings [57]. Proposed third-party compliance reviews may offer one pathway to achieving accountability between framework and practice [115].

PRACTICAL CONSIDERATIONS

- Differentiate between the types of evaluations one is conducting and judge frontier developer participation accordingly.
- Distinguish capability from accountability evaluations, and treat developer adoption as a success signal only for the former.
- Plan how a new metric could become visible and address a gap with existing popular metrics (e.g., inclusion in widely followed leaderboards).
- Track overall how labs and developers respond to new findings (e.g., patches, public statements, policy changes).
- Do not treat missing results as good news—depending on context, unfavorable evaluations may simply go unreported [249].

3e. How does funding shape independent evaluation?

Funding is a first-order question as it sets the material conditions for evaluation. Over the past three years, funding for evaluation has existed at the order of hundreds of millions of dollars; however, the lion's share has consisted of private money toward evaluation and benchmarking startups, rather than philanthropic and research funding toward independent evaluation. Funding can appear in different forms—as financial support, or as in-kind support, for example from frontier developers in the form of compute and free credits. Both are consequential for determining what evaluations are possible.

Philanthropy appears to have played a prominent role in shaping independent frontier AI evaluation as a field. Beyond funding one-off evaluations, philanthropies have been sources of financial support in standing up new organizations, organizing funding for research fellowships, and generating visibility for their recipients. Some of the earliest funding came from philanthropies associated with Effective Altruism (EA) [81] with a focus on AI safety such as Open Philanthropy (now Coefficient Giving) [55] and the Survival and Flourishing Fund [259]. This support helped to establish organizations such as Apollo Research and METR (formerly ARC Evals), who received funding early on from these parties, while also highlighting specific risk priorities in the early stages of field-building. Separately, Schmidt Sciences [238] has also become a prominent philanthropic funder of evaluations work. From the public sector, UK AISI, alongside their direct evaluation and standardization work, has also itself become a source of funding through grantmaking [65, 264].

Frontier AI developers have also been major sources of funding—in addition to funding via RFPs [10], they have also contributed to the Frontier Model Forum's AI Safety Fund [85], a pooled fund from the major frontier labs in partnership with philanthropic actors. While such pooled structures diversify the influence of individual developers, they have at the same time created a collective industry source of funding.

Some non-profits have chosen to supplement funding from grants with earned income, in the form of compensation for services conducted such as audits, monitoring, and infrastructure building. In January 2026, Apollo Research made the decision to transition from a non-profit entity into a Public Benefit Corporation (PBC), citing the ability to more quickly scale to meet the demand for frontier AI safety research and shape the market around it [27].

Looking ahead, projected philanthropic efforts from the OpenAI Foundation [212, 214, 215], and Anthropic's founders, who have pledged to "give away 80% of their wealth," [60, 226] may be poised to further reshape the funding landscape, by creating new centers of influence to shape research and evaluation priorities [214].

> **PRACTICAL CONSIDERATIONS**
>
> - Distinguish between financial and in-kind support, and understand how an organization discloses and treats both categories.
> - Identify historical and present funding and how it may shape an organization's orientation.
> - When relying on shared infrastructure, take note of who funds and maintains it.

3f. How do rising costs affect evaluations?

Evaluation costs have grown significantly in tandem with overall expectations for agents—for example, a full run-through for MirrorCode on Fable 5 costs approximately fifty thousand dollars in API credits [1], while ExploitBench also reports tens of thousands of dollars per model [147]. This matters because rising costs narrow the field by more severely restricting the work of evaluations to those who have the resources to do so [92]. Moreover, these costs may be asymmetric for external evaluators, who must pay API prices to conduct their work, versus model providers themselves, who are able to conduct evaluations with internal pricing. These costs are also problematic because they constrain evaluators' abilities to conduct more rigorous evaluation and better generalizability, for example through repeated runs and recruitment of human participants. If independent and high-quality evaluation is worth preserving and strengthening, evaluation costs represent a fundamental challenge to that vision.

The engineering effort required for evaluations today is not to be understated—the multitude of decisions concerning scaffold selection, environment setup, and monitoring needs translates into more engineering work required, in particular for maintaining dynamic benchmarks. Longer time horizon tasks [178, 225, 303], interest in probing frontier capabilities [275], and a focus on reliability [178] are contributing to rising costs for evaluation. Though a focus on reliability is an important component of achieving statistical rigor (Section 1d), it is expensive; implementing repeated sampling as part of greater statistical rigor multiplies cost. Moreover, using human graders at scale, for example in OpenAI's GDPval benchmark [220], is expensive both in time and funding. Likewise, recruiting participants for human uplift studies also reveals the tradeoff between cost and validity—though they are more expensive, they are considered the gold standard [221]. And though usage of autograders (LLM-as-judges) as a substitute for manual human evaluation may minimize some costs, measuring bias in LLM-as-judges and calibrating their setup may incur new costs. Standardization of infrastructure can also shift evaluation costs by distributing its maintenance across different parties, allowing evaluators to mitigate some experimental setup costs, but that does not meaningfully change the overall trajectory.

PRACTICAL CONSIDERATIONS

- Estimate the full cost of an experimental panel (i.e., per-run API cost × repeated runs × tasks) before committing, since a full run can reach tens of thousands of dollars per model.
- Jointly factor in cost [92], alongside dimensions of quality, reliability, and accuracy (see Section 1c).
- Allocate sufficient budget for engineering and maintenance, in addition to base API credit costs. Cut run costs where possible (e.g., via item-level analysis to shrink task sets, Section 1b; or by relying on shared infrastructure, Section 1j).

3g. Where are the venues for evaluations?

Top ML conferences continue to be a key forum for innovation in evaluations,[3] but there is significant evaluation work occurring outside of these peer-reviewed venues. While new benchmarks and ideas are still being presented through conference papers and posters, there is much being disseminated via blog posts, reports, "research notes," [177] GitHub and Hugging Face repositories, arXiv preprints, benchmark aggregators, and/or appearance in system cards. Many are never submitted to academic peer-review venues in the first place, yet are frequently cited in the literature.

Evaluators should take note of this current status quo. Moreover, how should we adapt our standards of credibility, when peer review alone has not caught up, to reflect this fast-paced landscape? From a methodological perspective, many of the considerations highlighted in Parts 1 and 2 can help inform that scrutiny. Do metrics map closely to the construct—is there even a construct defined? How is attribution credited to changes in different agent components? What is their approach to justifying statistical significance? All these are necessary considerations, regardless of where the work appears.

Another consideration for credibility is acknowledging the set of emerging reporting standards for evaluations work. Creating greater legibility around how evaluations are presented and reported can serve as a first step to further scrutiny. For example, the structured format of "safety case" is becoming one such standard in the form of a structured, evidence-based argument concerning risk associated with a system [23, 33, 94]. Evaluation cards have also been proposed as a standardized reporting format [93], with complementary dataset formats for capturing the broader experimental context [76]. Beyond formats, there are also new forms of reporting infrastructure, such as FLARE-AI, to document known vulnerabilities [155], or Weval for qualitative model evaluation results [56].

For reporting evaluation results, there is a maturing set of benchmark aggregators which facilitate comparability across agents and benchmarks. These are valuable in synthesizing across a variety of benchmarks and methodologies to give audiences a clear snapshot of the current state of progress in benchmark results. Conversely, these act as a starting point but are not the final endpoint for understanding and interpreting the methodology behind evaluations.

PRACTICAL CONSIDERATIONS

- Look beyond academia as well for the latest research insights, and scrutinize validity and rigor regardless of venue.
- Invest time to understand other evaluation ecosystems, and distinguish geographic differences in evaluation norms and goals.
- Embrace voluntary standardized reporting formats and share infrastructure and resources to help reinforce field-wide standard practices and promote legibility.

[3] For example, NeurIPS has launched a new evaluations-specific track [193].

3h. How is the public sector responding?

When the public sector deploys AI, it must balance its role as an operator with a need to protect the rights of constituents. Doing so requires clear insight into the impacts of these deployments, making evaluation a key responsibility for the public sector.

There is ongoing experimentation with AI at the state and federal level [110] (e.g., USAi [270]) toward "modernizing" public services, with envisioned use cases including cybersecurity, emergency-management coordination, benefits adjudication [162], public service navigation, inter-agency coordination, and statutory research (e.g., STARA [258]) and simplification (e.g., LaborBench [108]). Yet, systematic evaluation of AI for such deployments remains nascent. The need for systematic evaluation is amplified by how unavoidable (or even invisible) these systems may be for constituents in daily life. Systematic evaluation involves assessing whether AI is even appropriate for a proposed use case, to what extent a deployment is feasible, and whether it would work reliably for populations of interest, both in pre-deployment testing and in deployment.

Taking evaluation seriously as a governmental process requires a mindset shift to think of evaluation as a lifecycle [188, pp. 10, 20, Fig. 2, 197], not a one-off product procurement [162]. To say that an AI system passes inspection prior to deployment only offers limited insight into its impacts in the field [188, p. 6]. The challenge, then, is to develop government capacity to enable evaluation to become a sustained process that understands how deployments can shift over time and monitors accordingly.

In addition to the role of a direct evaluator, the public sector can also play a critical role as a regulator. Legislation such as the EU AI Act [75], which has taken effect in stages since 2024, and at the state level, California's Transparency in Frontier Artificial Intelligence Act (Senate Bill 53) [45], lay out what the role of regulation may play in this budding ecosystem. The EU AI Act, along with the associated voluntary Code of Practice [73], clarifies responsibilities of providers of general purpose AI models, with special attention toward systemic risks, and incident disclosure. California's SB 53 places an emphasis on transparency and disclosure for developers.

In service of sustained evaluation, public sector leadership in evaluations is necessary for building long-term capacity, agenda-setting for evaluation direction, channeling public input, and connecting evaluation to regulatory consequences [196]. Public sector leadership for evaluation can come in many forms. Two examples for public institutions are the US CAISSI and UK AISI bodies, which serve a mixture of functions, including direct evaluation, standard setting, policymaker communication, and in the case of UK AISI, research grantmaking [65, 264].

But leadership can be more abstracted from the work of evaluation itself. Proposed quasi-public accountability mechanisms such as Independent Verification Organizations (IVOs) [271] may also have a role to play. In this setup, a governmental authority defines an outcome for verification, then asks certified independent evaluators, dubbed IVOs, to operationalize these into metrics for evaluation. IVOs then verify whether those criteria are met. IVOs therefore hold power in how they approach this operationalization work and the measurement itself, even if the ultimate target is set by a governmental body, perhaps through a democratic deliberative process [78] (*cf.* [296]). Here, the work of evaluations is split between the governmental authorities and independent evaluators, who may come from nonprofits or academia.

PRACTICAL CONSIDERATIONS

- Evaluate AI deployments with respect to status quo baselines (e.g., marginal risk evaluations) [84].
- Engage actively with public sector evaluation bodies (e.g., US CAISSI, UK AISI, EU AI Office [74]) to inform new evaluation agendas.

- Plan public sector evaluations as a lifecycle that encompasses post-deployment through ongoing monitoring [188].

3i. Where are the entry points for different disciplines?

As agents become integrated into new environments and contexts, the shift from laboratory to practice has opened the door to new disciplinary perspectives, particularly those from the social sciences. We have discussed how benchmarks themselves are facing much more scrutiny in how they define and operationalize abstract concepts into measurement; these are questions social scientists are well acquainted with. Social scientists offer valuable insight into these challenges of validity, but also in areas such as experimental design leveraging real-world data, new frameworks for interaction harms, and perspectives on systemic impacts. Moreover, social scientists can shed light on potential structural impacts on institutions and organizations.

The forms of agent evaluations are changing, expanding beyond binary success variables to more nuanced assessments of agent trajectories. Evaluation is bringing a renewed focus to qualitative evidence, in the form of agent trajectories, transcript analyses, and safety cases. Moreover, evaluation is broadening beyond benchmarks, taking on new forms centering this qualitative focus, including open-world evaluations [128].

Despite this growing interest and need in interdisciplinary approaches, it is important to acknowledge that the inertia of AI evaluation remains largely within the discipline of computer science. Recent work has argued that societal AI research has become less interdisciplinary—while the *scope* of evaluation is broadening, *participation* in that evaluation is not [160]. There are a series of barriers in play: the resources and engineering costs associated with agents are significant, particularly for those outside existing work. The historical centering of existing AI/ML conferences and venues has created a larger moat for those outside those venues, and agenda-setting concentrates attention on a select number of research directions and questions.

Although computer scientists may be extending into interdisciplinary work rather than bringing in those from other disciplinary backgrounds, representation of other fields in evaluators themselves is consequential because of the distinct methods, research norms, and assumptions they carry into their work. Here are a few ways in which participation from different disciplines may benefit a cohesive science of agent evaluations (see Figure 20):

For Human-Computer Interaction (HCI). Human and agent interactions are being structured in evolving paradigms—evaluation work frequently concerns the human-interaction loop, and HCI researchers are well positioned to bring their insights to this work. HCI's familiarity with producing extensive artifacts and revising and refining taxonomies has clear applications to the work of evaluations today [86]. Moreover, HCI can bring a host of methods to bear such as user studies, diary and longitudinal studies for interaction harms and to investigate questions of overreliance on AI. In Part 2, we discussed the dominant augmentation and automation lenses applied to agents today; HCI can give concrete form to alternative modes of interaction.

For Measurement Science. There is a clear case for applying the extensive literature, vocabulary, and tooling currently present in several disciplines, including educational and psychological measurement, to the challenges of validity (Section 1a), reliability (Section 1d), equating (Section 1g) and generalizability (Section 1d) for agent evaluations. Moreover, while psychometricians have largely been absent from the work of AI evaluation so far, the nascent field of agent evaluation can benefit from their participation [288].

For Economics. Wider diffusion of agents has made measuring their effects more difficult, as clean RCT setups with agent-excluded baselines become harder to implement, particularly for sampling and recruiting

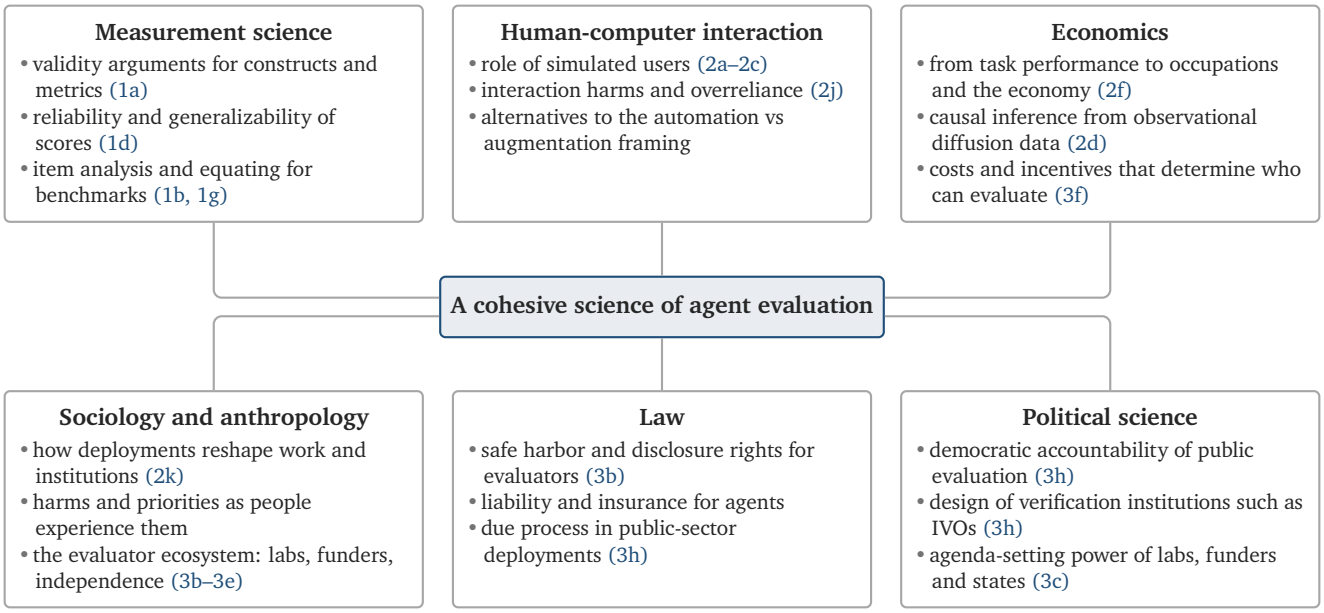

Figure 20: Fields whose expertise can shape agent evaluation. We treat the selection as illustrative, not exhaustive.

in human uplift studies [176]. Moreover, the results from these existing gold standard approaches may not translate to the macro questions about agents and their impacts which are increasingly of interest. Here, economists may be able to bring their methods to bear on the messy data that has accompanied diffusion, and parse the effects of agents through natural experiments, quasi-experimental methods, and rigorous interpretation of the observational data available.

For Sociology and Anthropology. Ethnographic methods may be valuable toward anticipatory goals. How is implementation on the ground impacting individuals and institutions, and what sorts of challenges are evident thus far? Where are areas of greatest evaluation need according to public interest or concern [38]? How can we ensure the focus of evaluation not only addresses specific iterations of long-term risk, but also immediate risks that publics are grappling with today?

For Political Science. Building on existing model-centric work from political scientists to evaluate model responses [195, 298], there are new opportunities to examine how usage of agents extends and informs the role of LLMs as gatekeepers for political content. Additionally, perspectives from computational social science may be especially well suited to examining the questions of external validity, particularly if access to production traffic becomes more widely available to independent evaluators. Growing public sector involvement in the work of evaluation activates longstanding debates about the extent to which regulatory action is accountable to democratic input. At what stages of the evaluation pipeline is there accountability to the public? What role does the public have in the experimentation involved? [141]

For Law. Evaluation needs within the public sector will require comprehensive legal safe harbor [153] and adequate due process with respect to AI deployments [110]. Moreover, the technical conception of AI "agents" maps poorly onto common law definitions of agents and agency [142]—yet, doctrines of apparent authority and apparent agency may still be relevant in instances of harm from agents [64]. The growing causal attribution challenge for AI agents across models, scaffolds, and tools also holds consequences—if an AI agent does cause harm, how does liability get assigned across these different components, particularly if they are maintained and provided by different parties? Moreover, how do these causal questions help inform applications of the least-cost avoider principle as applied to agent risks [44]? Liability and insurance regimes

for AI agents may also shape conceptions of risk that will inform the perceived need for evaluations [267].

PRACTICAL CONSIDERATIONS

- Collaborate with disciplinary perspectives suited to the questions of interest, not just by adopting their questions or methods.
- Draw upon the extensive literature from different disciplines, particularly for questions of validity and rigor.
- Recognize existing technical barriers to participation and connect new disciplinary perspectives to sufficient engineering capacity.

Conclusion

The work ahead must take place within a cohesive field of agent evaluation. This field is inherently interdisciplinary and spans several communities. Many of the challenges for agent evaluation are long familiar to those in other fields—for example, in defining constructs, assessing validity, and enabling comparability. Confronting the urgent need for reliable evidence requires coordination across communities to more effectively address common methodological and organizational challenges, and to ensure independence, obtain sufficient access, and pool resources.

Organizing a science of agent evaluation requires, in essence, recentering old concerns around validity and rigor within a new paradigm. Doing so allows us to adapt and respond to the shifts brought about by agents, by examining our current set of methods, metrics, and research questions and assessing whether they are still fit for the moment. As agents continue to evolve, there is a need to revisit and re-evaluate the assumptions and framings under evaluation—and whether the benchmarking paradigm, which has structured so much progress and evaluation thus far, continues to serve us well. Even in the questions we ask and evidence we gather, we must assess whether findings at one level of analysis necessarily hold for another.

All this depends upon fostering the conditions necessary for evaluators to pursue these questions of interest. Drawing upon different disciplinary perspectives to do this work must take form as a two-way street; both current and prospective evaluators can work together to translate shared challenges between vocabularies and concepts. Only then can agent evaluation form a cohesive core, rather than a fragmented set of evaluators pursuing the same challenges in isolation.

We refer to these as emerging trends to emphasize their developing status—infrastructure that continues to be refined, unanswered challenges waiting to be tackled, standards yet to be formed. There is a clear window now to bring in new participants, methods, and ideas for agent evaluations, and to meet the pressing need for rigorous evidence in a fast moving space.

Acknowledgements

We thank Manoel Horta Ribeiro, Mihir Kshirsagar, Jeremy McKey, Samantha Dalal, Rose Guingrich, J. Nathan Matias, Alexander Barry, Sam Miserendino, Jake Shapiro, and the broader community at Princeton's Center for Information Technology Policy (CITP) for invaluable input, conversations, and feedback on previous drafts.

For the evaluations work which informed views expressed in this paper, FC, SR, PK, NN, SK, and AN acknowledge funding from Coefficient Giving, Schmidt Sciences, the Princeton AI Lab, the Princeton Language and Intelligence Initiative, and the Princeton Catalysis Initiative. They also acknowledge API credits provided by OpenAI and Google to evaluate their models. AN is supported in part by the Alfred Landecker Membership at the Institute for Advanced Study.

References

[1] T. Adamczewski, D. Owen, D. Rein, F. Brand, G. Edkins, A. Hart, and D. O'Connell. *MirrorCode: AI can rebuild entire programs from behavior alone*. July 17, 2026. DOI: 10.48550/arXiv.2606.30182. arXiv: 2606.30182[cs.AI]. URL: http://arxiv.org/abs/2606.30182 (visited on 09/18/2026).

[2] L. Ahmad, S. Agarwal, M. Lampe, and P. Mishkin. *OpenAI's Approach to External Red Teaming for AI Models and Systems*. Jan. 24, 2025. DOI: 10.48550/arXiv.2503.16431. arXiv: 2503.16431 [cs.CY]. URL: http://arxiv.org/abs/2503.16431 (visited on 08/12/2026).

[3] AI Evaluator Forum. *Minimum Conditions for Embedding Evaluators*. Sept. 18, 2026. URL: https://aievaluatorforum.org/initiatives/embedded-evaluation-letter (visited on 09/28/2026).

[4] AI Risk and Vulnerability Alliance. *AVID*. URL: https://avidml.org/ (visited on 06/26/2026).

[5] AI Risk and Vulnerability Alliance (ARVA), R. Singh, B. Blili-Hamelin, C. Anderson, E. Tafesse, B. Vecchione, B. Duckles, and J. Metcalf. *Red-Teaming in the Public Interest*. Data & Society Research Institute, Feb. 9, 2025. DOI: 10.69985/VVGP4368. URL: https://datasociety.net/library/red-teaming-in-the-public-interest/ (visited on 06/26/2026).

[6] M. Akhtar, A. Reuel, P. Soni, S. Ahuja, P. S. Ammanamanchi, R. Rawal, V. Zouhar, S. Yadav, C. Whitehouse, D. Ki, J. Mickel, L. Choshen, M. Šuppa, J. Batzner, J. Chim, J. Sania, Y. Long, H. A. Rahmani, C. Knight, Y. Nan, J. Raj, Y. Fan, S. Singh, S. Sahoo, E. Habba, U. Gohar, S. Pawar, R. Scholz, A. Subramonian, J. Ni, M. Kochenderfer, S. Koyejo, M. Sachan, S. Biderman, Z. Talat, A. Ghosh, and I. Solaiman. *When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation*. June 29, 2026. DOI: 10.48550/arXiv.2602.16763. arXiv: 2602.16763 [cs.AI]. URL: http://arxiv.org/abs/2602.16763 (visited on 07/21/2026).

[7] M. Andriushchenko, A. Souly, M. Dziemian, D. Duenas, M. Lin, J. Wang, D. Hendrycks, A. Zou, Z. Kolter, M. Fredrikson, E. Winsor, J. Wynne, Y. Gal, and X. Davies. *AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents*. version: 3. Apr. 18, 2025. DOI: 10.48550/arXiv.2410.09024. arXiv: 2410.09024 [cs.LG]. URL: http://arxiv.org/abs/2410.09024 (visited on 06/25/2026).

[8] Anthropic. *Challenges in evaluating AI systems*. Oct. 4, 2023. URL: https://www.anthropic.com/news/evaluating-ai-systems (visited on 07/29/2026).

[9] Anthropic. *Introducing Anthropic's Responsible Scaling Policy*. Sept. 19, 2023. URL: https://www.anthropic.com/news/anthropics-responsible-scaling-policy (visited on 09/18/2026).

[10] Anthropic. *A new initiative for developing third-party model evaluations*. July 1, 2024. URL: https://www.anthropic.com/news/a-new-initiative-for-developing-third-party-model-evaluations (visited on 06/17/2026).

[11] Anthropic. *Clio: Privacy-preserving insights into real-world AI use*. Dec. 12, 2024. URL: https://www.anthropic.com/research/clio (visited on 09/18/2026).

[12] Anthropic. *Strengthening our safeguards through collaboration with US CAISI and UK AISI*. Sept. 12, 2025. URL: https://www.anthropic.com/news/strengthening-our-safeguards-through-collaboration-with-us-caisi-and-uk-aisi (visited on 06/26/2026).

[13] Anthropic. *Anthropic Economic Index report: Cadences*. June 26, 2026. URL: https://www.anthropic.com/research/economic-index-june-2026-report (visited on 06/29/2026).

[14] Anthropic. *Anthropic Economic Index report: Learning curves*. Mar. 24, 2026. URL: https://www.anthropic.com/research/economic-index-march-2026-report (visited on 06/03/2026).

[15] Anthropic. *Countering misuse of AI: September 2026*. Sept. 2026. URL: https://www.anthropic.com/threat-intelligence-report-september-2026 (visited on 10/01/2026).

[16] Anthropic. *Demystifying evals for AI agents*. Engineering at Anthropic. Jan. 9, 2026. URL: https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents (visited on 07/08/2026).

[17] Anthropic. *Enabling independent research on how people use Claude*. Aug. 26, 2026. URL: https://www.anthropic.com/research/enabling-independent-research (visited on 09/22/2026).

[18] Anthropic. *Investigating three incidents in our cybersecurity evaluations*. July 30, 2026. URL: https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals (visited on 09/18/2026).

[19] Anthropic. *Partnering with Accenture on embedded evaluation*. Sept. 18, 2026. URL: https://www.anthropic.com/news/accenture-embedded-evaluation (visited on 09/30/2026).

[20] Anthropic. *Redeploying Claude Fable 5*. June 30, 2026. URL: https://www.anthropic.com/news/redeploying-fable-5 (visited on 09/18/2026).

[21] Anthropic. *Economic Futures*. URL: https://www.anthropic.com/economic-futures (visited on 09/18/2026).
[22] Anthropic. *The Anthropic Economic Index*. URL: https://www.anthropic.com/economic-index#us-usage (visited on 06/03/2026).
[23] Anthropic. *Three Sketches of ASL-4 Safety Case Components*. URL: https://alignment.anthropic.com/2024/safety-cases/ (visited on 06/25/2026).
[24] U. Anwar, J. Piskorz, D. D. Baek, D. Africa, J. Weatherall, M. Tegmark, C. S. d. Witt, M. v. d. Schaar, and D. Krueger. *A Decision-Theoretic Formalisation of Steganography With Applications to LLM Monitoring*. Apr. 29, 2026. DOI: 10.48550/arXiv.2602.23163. arXiv: 2602.23163 [cs.AI]. URL: http://arxiv.org/abs/2602.23163 (visited on 06/20/2026).
[25] Apollo Research. *Apollo Is Adopting Inspect*. Nov. 13, 2024. URL: https://www.apolloresearch.ai/blog/apollo-is-adopting-inspect/ (visited on 07/03/2026).
[26] Apollo Research. *We Need A 'Science of Evals'*. Jan. 22, 2024. URL: https://www.apolloresearch.ai/science/we-need-a-science-of-evals/ (visited on 06/29/2026).
[27] Apollo Research. *Apollo Research is becoming a PBC*. Jan. 20, 2026. URL: https://www.apolloresearch.ai/blog/apollo-research-is-becoming-a-pbc/ (visited on 06/30/2026).
[28] Apollo Research. *We Need A Science of Scheming*. Jan. 19, 2026. URL: https://www.apolloresearch.ai/science/science-of-scheming/ (visited on 06/08/2026).
[29] R. Appel, P. McCrory, A. Tamkin, M. McCain, T. Neylon, and M. Stern. *Anthropic Economic Index report: Uneven geographic and enterprise AI adoption*. 2025. arXiv: 2511.15080 [econ.GN]. URL: https://arxiv.org/abs/2511.15080.
[30] Arena.ai. *Agent Arena: Causal Evaluation of Agents in the Real World - Arena.ai*. June 4, 2026. URL: https://arena.ai/blog/agent-arena-methodology (visited on 08/01/2026).
[31] Artificial Analysis. *AI Coding Agent Benchmarks & Leaderboard*. URL: https://artificialanalysis.ai/agents/coding-agents#token-usage (visited on 09/18/2026).
[32] B. Baker, J. Huizinga, L. Gao, Z. Dou, M. Y. Guan, A. Madry, W. Zaremba, J. Pachocki, and D. Farhi. *Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation*. Mar. 14, 2025. DOI: 10.48550/arXiv.2503.11926. arXiv: 2503.11926[cs.AI]. URL: http://arxiv.org/abs/2503.11926 (visited on 06/08/2026).
[33] M. Balesni, M. Hobbhahn, D. Lindner, A. Meinke, T. Korbak, J. Clymer, B. Shlegeris, J. Scheurer, C. Stix, R. Shah, N. Goldowsky-Dill, D. Braun, B. Chughtai, O. Evans, D. Kokotajlo, and L. Bushnaq. *Towards evaluations-based safety cases for AI scheming*. Nov. 7, 2024. DOI: 10.48550/arXiv.2411.03336. arXiv: 2411.03336 [cs.CR]. URL: http://arxiv.org/abs/2411.03336 (visited on 06/25/2026).
[34] J. Baumann, V. Padmakumar, X. Li, J. Yang, D. Yang, and S. Koyejo. *SWE-chat: Coding Agent Interactions From Real Users in the Wild*. Apr. 22, 2026. DOI: 10.48550/arXiv.2604.20779. arXiv: 2604.20779 [cs.AI]. URL: http://arxiv.org/abs/2604.20779 (visited on 06/08/2026).
[35] J. Becker, N. Rush, E. Barnes, and D. Rein. *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*. July 25, 2025. DOI: 10.48550/arXiv.2507.09089. arXiv: 2507.09089 [cs.AI]. URL: http://arxiv.org/abs/2507.09089 (visited on 07/05/2026).
[36] J. Benton, M. Wagner, E. Christiansen, C. Anil, E. Perez, J. Srivastav, E. Durmus, D. Ganguli, S. Kravec, B. Shlegeris, J. Kaplan, H. Karnofsky, E. Hubinger, R. Grosse, S. R. Bowman, and D. Duvenaud. *Sabotage Evaluations for Frontier Models*. Oct. 28, 2024. DOI: 10.48550/arXiv.2410.21514. arXiv: 2410.21514 [cs.LG]. URL: http://arxiv.org/abs/2410.21514 (visited on 07/24/2026).
[37] A. Bick, A. Blandin, and D. Deming. *The Rapid Adoption of Generative AI*. Working Paper. National Bureau of Economic Research, 2024. DOI: 10.3386/w32966.
[38] A. Birhane, W. Isaac, V. Prabhakaran, M. Díaz, M. C. Elish, I. Gabriel, and S. Mohamed. "Power to the People? Opportunities and Challenges for Participatory AI". In: *Equity and Access in Algorithms Mechanisms and Optimization*. Oct. 6, 2022, pp. 1–8. DOI: 10.1145/3551624.3555290. arXiv: 2209.07572[cs.CY]. URL: http://arxiv.org/abs/2209.07572 (visited on 10/01/2026).
[39] A. Birhane, R. Steed, V. Ojewale, B. Vecchione, and I. D. Raji. *AI auditing: The Broken Bus on the Road to AI Accountability*. Jan. 25, 2024. DOI: 10.48550/arXiv.2401.14462. arXiv: 2401.14462 [cs.CY]. URL: http://arxiv.org/abs/2401.14462 (visited on 06/26/2026).
[40] B. H. Bjarnason, A. Silva, and M. Monperrus. *On Randomness in Agentic Evals*. Mar. 25, 2026. DOI: 10.48550/arXiv.2602.07150. arXiv: 2602.07150 [cs.LG]. URL: http://arxiv.org/abs/2602.07150 (visited on 07/21/2026).

[41] D. Bowen, A.-K. Dombrowski, A. Gleave, and C. Cundy. "AI Companies Should Report Pre- and Post-Mitigation Safety Evaluations". In: (2025). arXiv: 2503.17388 [cs.CY]. URL: https://arxiv.org/abs/2503.17388.

[42] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. *Language Models are Few-Shot Learners*. July 22, 2020. DOI: 10.48550/arXiv.2005.14165. arXiv: 2005.14165 [cs.CL]. URL: http://arxiv.org/abs/2005.14165 (visited on 06/08/2026).

[43] M. Brundage, N. Dreksler, A. Homewood, S. McGregor, P. Paskov, C. Stosz, G. Sastry, A. F. Cooper, G. Balston, S. Adler, S. Casper, M. Anderljung, G. Werner, S. Mindermann, V. Mavroudis, B. Bucknall, C. Stix, J. Freund, L. Pacchiardi, J. Hernandez-Orallo, M. Pistillo, M. Chen, C. Painter, D. W. Ball, C. O'Keefe, G. Weil, B. Harack, G. Finley, R. Hassan, S. Emmons, C. Foster, A. Reuel, B. Treece, Y. Bengio, D. Reti, R. Bommasani, C. Trout, A. S. Shamsabadi, R. Dattani, A. Weller, R. Trager, J. Sevilla, L. Wagner, L. Soder, K. Ramakrishnan, H. Papadatos, M. Murray, and R. Tovcimak. *Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies*. Feb. 7, 2026. DOI: 10.48550/arXiv.2601.11699. arXiv: 2601.11699 [cs.CY]. URL: http://arxiv.org/abs/2601.11699 (visited on 06/26/2026).

[44] G. Calabresi. *The Cost of Accidents: A Legal and Economic Analysis*. Yale University Press, 1970. URL: http://www.jstor.org/stable/j.ctt1cc2mmw.

[45] California Legislature. *Senate Bill 53: Artificial Intelligence Models: Large Developers*. Sept. 2025. URL: https://legiscan.com/CA/text/SB53/id/3270002.

[46] J. S. Chan, N. Chowdhury, O. Jaffe, J. Aung, D. Sherburn, E. Mays, G. Starace, K. Liu, L. Maksin, T. Patwardhan, L. Weng, and A. Mądry. "MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering". In: *International Conference on Learning Representations (ICLR 2025)*. 2025, pp. 50466–50494. URL: https://proceedings.iclr.cc/paper_files/paper/2025/file/7e3767db483c942b883eb4f8cfb74e31-Paper-Conference.pdf (visited on 09/18/2026).

[47] J. Charnock, S. Williams, Z. Kara, M. Anderljung, A. T. Boria, S. Casper, A. Reuel, and J. Freund. *Embedded Assessments for Frontier AI*. Sept. 21, 2026. DOI: 10.48550/arXiv.2609.25413. arXiv: 2609.25413 [cs.CY]. URL: http://arxiv.org/abs/2609.25413 (visited on 09/25/2026).

[48] A. Chatterji, T. Cunningham, D. J. Deming, Z. Hitzig, C. Ong, C. Y. Shan, and K. Wadman. *How People Use ChatGPT*. Working Paper 34255. National Bureau of Economic Research, Sept. 2025. DOI: 10.3386/w34255. URL: https://www.nber.org/papers/w34255 (visited on 06/08/2026).

[49] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. *Evaluating Large Language Models Trained on Code*. 2021. arXiv: 2107.03374 [cs.LG]. URL: https://arxiv.org/abs/2107.03374.

[50] Z. Chen, X. Liu, H. Tong, C. Guo, Y. Nie, J. Zhang, M. Kang, C. Xu, Q. Liu, X. Liu, T. Shi, C. Xiao, S. Koyejo, P. Liang, W. Guo, D. Song, and B. Li. *DecodingTrust-Agent Platform (DTap): A Controllable and Interactive Red-Teaming Platform for AI Agents*. May 6, 2026. DOI: 10.48550/arXiv.2605.04808. arXiv: 2605.04808 [cs.AI]. URL: http://arxiv.org/abs/2605.04808 (visited on 06/26/2026).

[51] M. Cheng, C. Lee, P. Khadpe, S. Yu, D. Han, and D. Jurafsky. *Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence*. Oct. 1, 2025. DOI: 10.48550/arXiv.2510.01395. arXiv: 2510.01395 [cs.CY]. URL: http://arxiv.org/abs/2510.01395 (visited on 07/24/2026).

[52] W. Chi, V. Chen, A. N. Angelopoulos, W.-L. Chiang, A. Mittal, N. Jain, T. Zhang, I. Stoica, C. Donahue, and A. Talwalkar. *Copilot Arena: A Platform for Code LLM Evaluation in the Wild*. Feb. 13, 2025. DOI: 10.48550/arXiv.2502.09328. arXiv: 2502.09328 [cs.SE]. URL: http://arxiv.org/abs/2502.09328 (visited on 06/26/2026).

[53] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica. *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*.

Mar. 7, 2024. DOI: 10.48550/arXiv.2403.04132. arXiv: 2403.04132 [cs.AI]. URL: http://arxiv.org/abs/2403.04132 (visited on 06/09/2026).
[54] F. Chollet. *On the Measure of Intelligence*. Nov. 25, 2019. DOI: 10.48550/arXiv.1911.01547. arXiv: 1911.01547 [cs.AI]. URL: http://arxiv.org/abs/1911.01547 (visited on 06/30/2026).
[55] Coefficient Giving. *Navigating Transformative AI*. URL: https://coefficientgiving.org/funds/navigating-transformative-ai/ (visited on 09/18/2026).
[56] Collective Intelligence Project. *Weval: An Open Platform for AI Evaluation*. 2024.
[57] S. Costanza-Chock, I. D. Raji, and J. Buolamwini. "Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1571–1583. DOI: 10.1145/3531146.3533213. URL: https://doi.org/10.1145/3531146.3533213.
[58] T. Cunningham. *Metrics of Agent Ability*. July 24, 2026. URL: https://metr.org/notes/2026-07-24-metrics-of-model-ability/.
[59] P. D'Oro, S. Silwal, W. Wong, Y. Sun, F. Xiao, M. Wang, E. Gan, A. Bolourchi, and J. Tighe. *Computer Use at the Edge of the Statistical Precipice*. May 7, 2026. DOI: 10.48550/arXiv.2605.08261. arXiv: 2605.08261 [cs.SE]. URL: http://arxiv.org/abs/2605.08261 (visited on 06/26/2026).
[60] Dario Amodei. *The Adolescence of Technology*. Jan. 2026. URL: https://darioamodei.com/essay/the-adolescence-of-technology.
[61] E. Debenedetti, I. Shumailov, T. Fan, J. Hayes, N. Carlini, D. Fabian, C. Kern, C. Shi, A. Terzis, and F. Tramèr. *Defeating Prompt Injections by Design*. June 24, 2025. DOI: 10.48550/arXiv.2503.18813. arXiv: 2503.18813 [cs.CR]. URL: http://arxiv.org/abs/2503.18813 (visited on 06/25/2026).
[62] E. Debenedetti, J. Zhang, M. Balunović, L. Beurer-Kellner, M. Fischer, and F. Tramèr. *AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents*. Nov. 24, 2024. DOI: 10.48550/arXiv.2406.13352. arXiv: 2406.13352 [cs.CR]. URL: http://arxiv.org/abs/2406.13352 (visited on 06/26/2026).
[63] M. Demirer, L. Musolff, and L. Yang. *Writing Code vs. Shipping Code: Productivity Effects Across Generations of AI Coding Tools*. May 2026. DOI: 10.3386/w35275. URL: https://www.nber.org/papers/w35275.
[64] D. DeMott. *When Agentic AI Met the Common Law of Agency*. Rochester, NY, May 27, 2026. DOI: 10.2139/ssrn.6838660. URL: https://papers.ssrn.com/abstract=6838660 (visited on 10/09/2026).
[65] Department for Science, Innovation and Technology, AI Security Institute, and Feryal Clark MP. *Cutting-Edge Research on AI Security bolstered with new Challenge Fund to ramp up public trust and adoption*. GOV.UK. Mar. 5, 2025. URL: https://www.gov.uk/government/news/cutting-edge-research-on-ai-security-bolstered-with-new-challenge-fund-to-ramp-up-public-trust-and-adoption.
[66] E. W. Dillon, S. Jaffe, N. Immorlica, and C. T. Stanton. *Shifting Work Patterns with Generative AI*. Nov. 13, 2025. DOI: 10.48550/arXiv.2504.11436. arXiv: 2504.11436 [econ.GN]. URL: http://arxiv.org/abs/2504.11436 (visited on 07/05/2026).
[67] S. Dohnány, Z. Kurth-Nelson, E. Spens, L. Luettgau, A. Reid, I. Gabriel, C. Summerfield, M. Shanahan, and M. M. Nour. "Technological folie à deux: feedback loops between AI chatbots and mental health". In: *Nature Mental Health* 4.3 (Mar. 1, 2026), pp. 336–345. DOI: 10.1038/s44220-026-00595-8. URL: https://doi.org/10.1038/s44220-026-00595-8.
[68] M. Dubois, H. Coppock, M. Giulianelli, O. K. Jorgensen, T. Flesch, L. Luettgau, and C. Ududec. "Quantifying Biases in LLM-as-a-Judge Evaluations". In: *Forty-third International Conference on Machine Learning*. 2026. URL: https://openreview.net/forum?id=q33S4QMuay.
[69] M. Dubois, E. Zorer, M. Hamin, J. Skinner, A. Souly, J. Wynne, H. Coppock, L. Sato, S. Kapoor, S. Dev, K. Juchems, K. Mai, T. Flesch, L. Luettgau, C. Teague, E. Patey, J. J. Allaire, L. Pacchiardi, J. Hernandez-Orallo, and C. Ududec. *Seven simple steps for log analysis in AI systems*. version: 1. Feb. 13, 2026. DOI: 10.48550/arXiv.2604.09563. arXiv: 2604.09563 [cs.AI]. URL: http://arxiv.org/abs/2604.09563 (visited on 06/08/2026).
[70] Dynabench. *Dynabench*. URL: https://dynabench.org/ (visited on 05/28/2026).
[71] Epoch AI. *About | Benchmarking*. URL: https://epoch.ai/benchmarks/about (visited on 06/25/2026).
[72] Epoch AI. *Epoch Capabilities Index*. URL: https://epoch.ai/benchmarks/eci?view=graph&tab=training-compute&subset-view=graph&subset-tab=Software+engineering (visited on 09/18/2026).
[73] European Commission. *The General-Purpose AI Code of Practice*. July 10, 2025. URL: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (visited on 10/05/2026).

[74] European Commission. *European AI Office*. URL: https://digital-strategy.ec.europa.eu/en/policies/ai-office#ecl-inpage-tasks-of-the-ai-office (visited on 09/18/2026).

[75] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*. July 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[76] EvalEval Coalition. *Every Eval Ever*. URL: https://evalevalai.com/projects/every-eval-ever/ (visited on 09/18/2026).

[77] A. Farach, A. Cambon, L. Tankelevitch, C. Hsueh, and R. Janssen. *Scaffolding Human-AI Collaboration: A Field Experiment on Behavioral Protocols and Cognitive Reframing*. Apr. 19, 2026. DOI: 10.48550/arXiv.2604.08678. arXiv: 2604.08678 [econ.GN]. URL: http://arxiv.org/abs/2604.08678 (visited on 07/05/2026).

[78] Fathom. *Independent Oversight Marketplace for AI*. June 1, 2025. URL: https://fathom.org/insights/independent-oversight-marketplace-for-ai (visited on 07/03/2026).

[79] M. Feffer, A. Sinha, W. H. Deng, Z. C. Lipton, and H. Heidari. "Red-Teaming for Generative AI: Silver Bullet or Security Theater?" In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. Vol. 7. arXiv, Aug. 27, 2024, pp. 421–437. DOI: 10.48550/arXiv.2401.15897. arXiv: 2401.15897. URL: http://arxiv.org/abs/2401.15897 (visited on 07/23/2026).

[80] F. Flynn, H. King, and A. Dragan. *Google DeepMind strengthens the Frontier Safety Framework*. Google DeepMind. Section: Responsibility & Safety. Sept. 22, 2025. URL: https://deepmind.google/blog/strengthening-our-frontier-safety-framework/ (visited on 10/19/2026).

[81] C. Ford. *Anthropic employees say they'll give away billions. Where will it go?* Transformer. Mar. 12, 2026. URL: https://www.transformernews.ai/p/anthropic-employees-philanthropy-billions-donations-effective-altruism-coefficient-giving-ai-safety (visited on 07/01/2026).

[82] B. Fowler, R. Smith, D. T. Graviet, W. Myers, J. Greaves, N. F. Oozeer, A. García, P. Quirke, A. Abdullah, F. Barez, and S. K. Upadhyay. *The Capability Frontier: Benchmarks Miss 82% of Model Performance*. June 25, 2026. DOI: 10.48550/arXiv.2606.26836. arXiv: 2606.26836 [cs.AI]. URL: http://arxiv.org/abs/2606.26836 (visited on 06/28/2026).

[83] Frontier Model Forum. *Issue Brief: What is red teaming?* Oct. 27, 2023. URL: https://www.frontiermodelforum.org/updates/red-teaming/ (visited on 07/23/2026).

[84] Frontier Model Forum. *Issue Brief: Early Best Practices for Frontier AI Safety Evaluations*. July 31, 2024. URL: https://www.frontiermodelforum.org/updates/early-best-practices-for-frontier-ai-safety-evaluations/ (visited on 08/12/2026).

[85] Frontier Model Forum. *AI Safety Fund*. Apr. 14, 2025. URL: https://www.frontiermodelforum.org/ai-safety-fund/ (visited on 09/18/2026).

[86] D. Furniss, A. Blandford, and P. Curzon. "Confessions from a grounded theory PhD: experiences and lessons learnt". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada: Association for Computing Machinery, 2011, pp. 113–122. DOI: 10.1145/1978942.1978960. URL: https://doi.org/10.1145/1978942.1978960.

[87] Y. Gao, R. Zhao, Y. Deng, and W. Zhang. "DR-Arena: an Automated Evaluation Framework for Deep Research Agents". In: *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 27130-27152, 2026. 2026, pp. 27130–27152. DOI: 10.18653/v1/2026.acl-long.1249. arXiv: 2601.10504 [cs.CL]. URL: http://arxiv.org/abs/2601.10504 (visited on 07/23/2026).

[88] C. Gartenberg, S. Hasan, A. Murray, and L. Pierce. "More Versus Better: Artificial Intelligence, Incentives, and the Emerging Crisis in Peer Review". In: *Organization Science* 37.3 (May 2026), pp. 795–812. DOI: 10.1287/orsc.2026.ed.v37.n3. URL: https://pubsonline.informs.org/doi/10.1287/orsc.2026.ed.v37.n3.

[89] C. Geertz. "Thick Description: Toward an Interpretive Theory of Culture". In: *The Interpretation of Cultures*. New York, NY: Basic Books, 1973, pp. 15–42. URL: https://search.alexanderstreet.com/view/work/bibliographic_entity%7Cbibliographic_details%7C1667767.

[90] A. Gelman and E. Loken. *The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No "Fishing Expedition" or "p-hacking" and the Research Hypothesis Was Posited Ahead of Time*. Working paper, Department of Statistics, Columbia University. 2013. URL: https://sites.stat.columbia.edu/gelman/research/unpublished/forking.pdf (visited on 09/18/2026).

[91] GenAI Adoption Tracker. *GenAI Adoption Tracker*. URL: https://www.genaiadoptiontracker.com/ (visited on 09/18/2026).

[92] A. Ghosh. *AI evals are becoming the new compute bottleneck*. Apr. 29, 2026. URL: https://evalevalai.com/research/2026/04/29/eval-costs-bottleneck/ (visited on 07/03/2026).

[93] A. Ghosh, A. Reuel, J. Chim, W. M. Kennedy, S. Yadav, J. Mickel, Y. Long, A. Tran, A. Kornilova, D. Stachura, K. Klyman, F. Friedrich, J. Sania, J. Batzner, A. Mishra, E. Habba, Y. Hao, N. Heath, S. Rismani, U. Gohar, A. Loehr, D. Manheim, R. Dhar, S. H. Nelaturu, A. Sinha, L. Choshen, D. Sharma, I. Khire, A. Saha, S. Sahoo, M. Hardy, M. A. Riegler, K. Manghnani, M. Lin, Y. Jiang, Y. Huang, A. Yehudai, J. Ji, A. Hofmann, M. Akhtar, M. Lamparth, N. Moniz, Y. Jernite, S. Biderman, Z. Talat, S. Koyejo, M. Kochenderfer, and I. Solaiman. *Evaluation Cards: An Interpretive Layer for AI Evaluation Reporting*. June 9, 2026. DOI: 10.48550/arXiv.2606.09809. arXiv: 2606.09809 [cs.AI]. URL: http://arxiv.org/abs/2606.09809 (visited on 06/24/2026).

[94] A. Goemans, M. D. Buhl, J. Schuett, T. Korbak, J. Wang, B. Hilton, and G. Irving. *Safety case template for frontier AI: A cyber inability argument*. Nov. 12, 2024. DOI: 10.48550/arXiv.2411.08088. arXiv: 2411.08088 [cs.CY]. URL: http://arxiv.org/abs/2411.08088 (visited on 06/25/2026).

[95] Google. *Understanding the AI economy*. July 23, 2026. URL: https://blog.google/innovation-and-ai/technology/research/understanding-the-ai-economy/ (visited on 09/18/2026).

[96] R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, A. Khan, J. Michael, S. Mindermann, E. Perez, L. Petrini, J. Uesato, J. Kaplan, B. Shlegeris, S. R. Bowman, and E. Hubinger. *Alignment faking in large language models*. Dec. 20, 2024. DOI: 10.48550/arXiv.2412.14093. arXiv: 2412.14093 [cs.AI]. URL: http://arxiv.org/abs/2412.14093 (visited on 07/05/2026).

[97] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz. *Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*. May 5, 2023. DOI: 10.48550/arXiv.2302.12173. arXiv: 2302.12173 [cs.CR]. URL: http://arxiv.org/abs/2302.12173 (visited on 06/26/2026).

[98] M. Y. Guan, M. Wang, M. Carroll, Z. Dou, A. Y. Wei, M. Williams, B. Arnav, J. Huizinga, I. Kivlichan, M. Glaese, J. Pachocki, and B. Baker. *Monitoring Monitorability*. Dec. 20, 2025. DOI: 10.48550/arXiv.2512.18311. arXiv: 2512.18311 [cs.AI]. URL: http://arxiv.org/abs/2512.18311 (visited on 07/23/2026).

[99] V. Gupta, P. Nutter, S. Stante, A. Krause, F. Tramèr, L. Fluri, X. Chen, and A. Hedström. *Position: Anthropomorphic Misalignment Research Needs Stronger Evidence*. May 29, 2026. DOI: 10.48550/arXiv.2606.07612. arXiv: 2606.07612 [cs.CY]. URL: http://arxiv.org/abs/2606.07612 (visited on 07/16/2026).

[100] L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith, W. Barfuss, J. Foerster, T. Gavenčiak, T. A. Han, E. Hughes, V. Kovařík, J. Kulveit, J. Z. Leibo, C. Oesterheld, C. S. d. Witt, N. Shah, M. Wellman, P. Bova, T. Cimpeanu, C. Ezell, Q. Feuillade-Montixi, M. Franklin, E. Kran, I. Krawczuk, M. Lamparth, N. Lauffer, A. Meinke, S. Motwani, A. Reuel, V. Conitzer, M. Dennis, I. Gabriel, A. Gleave, G. Hadfield, N. Haghtalab, A. Kasirzadeh, S. Krier, K. Larson, J. Lehman, D. C. Parkes, G. Piliouras, and I. Rahwan. *Multi-Agent Risks from Advanced AI*. Feb. 19, 2025. DOI: 10.48550/arXiv.2502.14143. arXiv: 2502.14143. URL: http://arxiv.org/abs/2502.14143 (visited on 10/03/2026).

[101] K. Handa, A. Tamkin, M. McCain, S. Huang, E. Durmus, S. Heck, J. Mueller, J. Hong, S. Ritchie, T. Belonax, K. K. Troy, D. Amodei, J. Kaplan, J. Clark, and D. Ganguli. *Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations*. Feb. 11, 2025. DOI: 10.48550/arXiv.2503.04761. arXiv: 2503.04761 [cs.CY]. URL: http://arxiv.org/abs/2503.04761 (visited on 06/07/2026).

[102] Harbor. *Harbor*. URL: https://harborframework.com/ (visited on 09/18/2026).

[103] M. Hardt. "Evaluation at the frontier". In: *The Emerging Science of Machine Learning Benchmarks*. Princeton University Press, Oct. 6, 2026. URL: https://mlbenchmarks.org/14-evaluation-frontier.html.

[104] M. Hardt. “Holdout method”. In: *The Emerging Science of Machine Learning Benchmarks*. Princeton University Press, Oct. 6, 2026. URL: https://mlbenchmarks.org/04-holdout-method.html.
[105] M. Hardt. “Scientific crisis”. In: *The Emerging Science of Machine Learning Benchmarks*. Princeton University Press, Oct. 6, 2026. URL: https://mlbenchmarks.org/06-scientific-crisis.html.
[106] M. Hardt. “Test set reuse”. In: *The Emerging Science of Machine Learning Benchmarks*. Princeton University Press, Oct. 6, 2026. URL: https://mlbenchmarks.org/05-test-set-reuse.html.
[107] M. Hardt. “When the model moves the data”. In: *The Emerging Science of Machine Learning Benchmarks*. Princeton University Press, Oct. 6, 2026. URL: https://mlbenchmarks.org/13-model-moves-data.html#dynamic-benchmarks.
[108] E. Hariri and D. E. Ho. *AI for Statutory Simplification: A Comprehensive State Legal Corpus and Labor Benchmark*. Stanford RegLab. 2025. URL: https://dho.stanford.edu/wp-content/uploads/LaborBench.pdf (visited on 09/18/2026).
[109] A. Haupt and E. Brynjolfsson. “Position: AI Should Not Be An Imitation Game: Centaur Evaluations”. In: (2025). URL: https://digitaleconomy.stanford.edu/app/uploads/2025/06/CentaurEvaluations.pdf (visited on 09/18/2026).
[110] D. E. Ho, O. Martin, A. Perez, and K. Rodolfa. “Evaluation as Due Process: Civil Service in an Automated Age”. In: *Administrative Law Review* 77.4 (2025). URL: https://administrativelawreview.org/wp-content/uploads/sites/2/2025/12/ALR-77.4_HoMartinPerezRodolfa.pdf.
[111] M. Hobbhahn. *A long list of open problems and concrete projects in evals*. Google Docs. URL: https://docs.google.com/document/d/1gi32-HZozxVimNg5Mhvk4CvW4zq8J12rGmK_j2zxNEg/edit?tab=t.0&usp=embed_facebook (visited on 07/02/2026).
[112] M. Hobbhahn. *Evals*. URL: https://www.mariushobbhahn.com/evals/ (visited on 06/26/2026).
[113] M. Hobbhahn, M. Balesni, J. Scheurer, R. Shah, and A. Meinke. *A starter guide for evals*. Jan. 8, 2024. URL: https://www.lesswrong.com/posts/2PiawPFJeyCQGcwXG/a-starter-guide-for-evals (visited on 09/18/2026).
[114] Holistic Agent Leaderboard. *HAL Reliability Dashboard*. URL: https://hal.cs.princeton.edu/reliability/ (visited on 10/01/2026).
[115] A. Homewood, S. Williams, N. Dreksler, J. Lidiard, M. Murray, L. Heim, M. Ziosi, S. Ó. hÉigeartaigh, M. Chen, K. Wei, C. Winter, M. Brundage, B. Garfinkel, and J. Schuett. *Third-party compliance reviews for frontier AI safety frameworks*. 2025. arXiv: 2505.01643 [cs.CY]. URL: https://arxiv.org/abs/2505.01643.
[116] Hugging Face. *Harness, Scaffold, and the AI Agent Terms Worth Getting Right*. May 25, 2026. URL: https://huggingface.co/blog/agent-glossary (visited on 06/30/2026).
[117] L. Ibrahim, F. S. Hafner, M. Cheng, C. Lee, R. Anselmetti, R. Willer, L. Rocher, and D. Yang. *Sycophantic AI makes human interaction feel more effortful and less satisfying over time*. June 21, 2026. DOI: 10.48550/arXiv.2605.07912. arXiv: 2605.07912 [cs.HC]. URL: http://arxiv.org/abs/2605.07912 (visited on 07/24/2026).
[118] L. Ibrahim, S. Huang, U. Bhatt, L. Ahmad, and M. Anderljung. *Towards interactive evaluations for interaction harms in human-AI systems*. July 30, 2025. DOI: 10.48550/arXiv.2405.10632. arXiv: 2405.10632 [cs.CY]. URL: http://arxiv.org/abs/2405.10632 (visited on 06/26/2026).
[119] A. Jacovi, A. Caciularu, O. Goldman, and Y. Goldberg. “Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5075–5084. DOI: 10.18653/v1/2023.emnlp-main.308. URL: https://aclanthology.org/2023.emnlp-main.308/ (visited on 06/08/2026).
[120] N. Jain, K. Han, A. Gu, W.-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, and I. Stoica. *LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code*. June 6, 2024. DOI: 10.48550/arXiv.2403.07974. arXiv: 2403.07974 [cs.SE]. URL: http://arxiv.org/abs/2403.07974 (visited on 06/08/2026).
[121] Japan AI Safety Institute. *Guide to Red Teaming Methodology on AI Safety - Japan AISI*. URL: https://aisi.go.jp/output/output_framework/guide_to_red_teaming_methodology_on_ai_safety/ (visited on 07/03/2026).

[122] O. Järviniemi and E. Hubinger. *Uncovering Deceptive Tendencies in Language Models: A Simulated Company AI Assistant*. Apr. 25, 2024. DOI: 10.48550/arXiv.2405.01576. arXiv: 2405.01576[cs.CL]. URL: http://arxiv.org/abs/2405.01576 (visited on 07/03/2026).

[123] R. Jayaram, D. Tyler, D. Woodruff, C. Cortes, Y. Matias, V. Mirrokni, and V. Cohen-Addad. *Towards Automating Scientific Review with Google's Paper Assistant Tool*. June 26, 2026. DOI: 10.48550/arXiv.2606.28277. arXiv: 2606.28277 [cs.LG]. URL: http://arxiv.org/abs/2606.28277 (visited on 06/30/2026).

[124] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan. "SWE-bench: Can Language Models Resolve Real-world Github Issues?" In: *International Conference on Learning Representations (ICLR 2024)*. 2024, pp. 54107–54157. URL: https://proceedings.iclr.cc/paper_files/paper/2024/hash/edac78c3e300629acfe6cbe9ca88fb84-Abstract-Conference.html (visited on 09/18/2026).

[125] M. Joglekar, J. Chen, G. Wu, J. Yosinski, J. Wang, B. Barak, and A. Glaese. *Training LLMs for Honesty via Confessions*. arXiv.org. Dec. 8, 2025. URL: https://arxiv.org/abs/2512.08093v2 (visited on 07/06/2026).

[126] H. Jung, P. V. Diniz, J. R. C. Roveda, A. F. d. Silva, H. Jung, E. Tsai, A. Korolova, and M. H. Ribeiro. *Can AI Agents Synthesize Scientific Conclusions?* June 9, 2026. DOI: 10.48550/arXiv.2606.11337. arXiv: 2606.11337 [cs.AI]. URL: http://arxiv.org/abs/2606.11337 (visited on 06/26/2026).

[127] S. Kapoor, R. Bommasani, K. Klyman, S. Longpre, A. Ramaswami, P. Cihon, A. Hopkins, K. Bankston, S. Biderman, M. Bogen, R. Chowdhury, A. Engler, P. Henderson, Y. Jernite, S. Lazar, S. Maffulli, A. Nelson, J. Pineau, A. Skowron, D. Song, V. Storchan, D. Zhang, D. E. Ho, P. Liang, and A. Narayanan. "Position: On the Societal Impact of Open Foundation Models". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. 2024. arXiv: 2403.07918. URL: https://proceedings.mlr.press/v235/kapoor24a.html.

[128] S. Kapoor, P. Kirgis, A. Schwartz, S. Rabanser, J. J. Allaire, R. Bommasani, H. Coppock, M. Dubois, G. K. Hadfield, A. B. Hall, S. Hooker, S. Lazar, S. Newman, D. Papailiopoulos, S. Tekofsky, H. Toner, C. Ududec, and A. Narayanan. *Open-World Evaluations for Measuring Frontier AI Capabilities*. version: 1. May 19, 2026. DOI: 10.48550/arXiv.2605.20520. arXiv: 2605.20520 [cs.AI]. URL: http://arxiv.org/abs/2605.20520 (visited on 05/31/2026).

[129] S. Kapoor, N. Kolt, and S. Lazar. *Build Agent Advocates, Not Platform Agents*. June 19, 2025. DOI: 10.48550/arXiv.2505.04345. arXiv: 2505.04345 [cs.CY]. URL: http://arxiv.org/abs/2505.04345 (visited on 06/18/2026).

[130] S. Kapoor and A. Narayanan. *Leakage and the Reproducibility Crisis in ML-based Science*. July 14, 2022. DOI: 10.48550/arXiv.2207.07048. arXiv: 2207.07048 [cs.LG]. URL: http://arxiv.org/abs/2207.07048 (visited on 06/08/2026).

[131] S. Kapoor and A. Narayanan. *The AI-as-Normal-Technology view of loss of control incidents*. AI as Normal Technology. Sept. 14, 2026. URL: https://www.normaltech.ai/p/the-ai-as-normal-technology-view (visited on 10/01/2026).

[132] S. Kapoor, B. Stroebl, P. Kirgis, N. Nadgir, Z. S. Siegel, B. Wei, T. Xue, Z. Chen, F. Chen, S. Utpala, F. Ndzomga, D. Oruganty, S. Luskin, K. Liu, B. Yu, A. Arora, D. Hahm, H. Trivedi, H. Sun, J. Lee, T. Jin, Y. Mai, Y. Zhou, Y. Zhu, R. Bommasani, D. Kang, D. Song, P. Henderson, Y. Su, P. Liang, and A. Narayanan. *Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation*. Oct. 13, 2025. DOI: 10.48550/arXiv.2510.11977. arXiv: 2510.11977 [cs.AI]. URL: http://arxiv.org/abs/2510.11977 (visited on 05/26/2026).

[133] S. Kapoor, B. Stroebl, Z. S. Siegel, N. Nadgir, and A. Narayanan. *AI Agents That Matter*. July 1, 2024. DOI: 10.48550/arXiv.2407.01502. arXiv: 2407.01502 [cs.LG]. URL: http://arxiv.org/abs/2407.01502 (visited on 05/28/2026).

[134] A. Keller, K. Kwegyir-Aggrey, R. Steed, A. Rao, J. Sharp, and A. Bergman. *Expanding the AI Evaluation Toolbox with Statistical Models*. 800-3. Last Modified: 2026-02-17T20:02-05:00. Gaithersburg, MD: National Institute of Standards and Technology, Feb. 17, 2026. DOI: 10.6028/NIST.AI.800-3. URL: https://www.nist.gov/publications/expanding-ai-evaluation-toolbox-statistical-models (visited on 06/26/2026).

[135] Kevin Meng, Vincent Huang, Jacob Steinhardt, and Sarah Schwettmann. *Introducing Docent: A system for analyzing and intervening on agent behavior*. Transluce. Mar. 24, 2025. URL: https://transluce.org/introducing-docent (visited on 06/07/2026).

[136] D. Kiela, M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Talat, P. Stenetorp, R. Jia, M. Bansal, C. Potts, and A. Williams. "Dynabench: Rethinking Benchmarking in NLP". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou. Online: Association for Computational Linguistics, June 2021, pp. 4110–4124. DOI: 10.18653/v1/2021.naacl-main.324. arXiv: 2104.14337 [cs.CL]. URL: https://aclanthology.org/2021.naacl-main.324/ (visited on 06/08/2026).

[137] D. Kilov, S. Y. Guyot, C. Hendy, S. Li, and S. Lazar. *A Scalable Approach to Evaluating Moral Sensitivity in LLMs*. 2026. arXiv: 2607.02972 [cs.CY]. URL: https://arxiv.org/abs/2607.02972.

[138] M. Kinniment, L. J. K. Sato, H. Du, B. Goodrich, M. Hasin, L. Chan, L. H. Miles, T. R. Lin, H. Wijk, J. Burget, A. Ho, E. Barnes, and P. Christiano. *Evaluating Language-Model Agents on Realistic Autonomous Tasks*. Jan. 4, 2024. DOI: 10.48550/arXiv.2312.11671. arXiv: 2312.11671 [cs.CL]. URL: http://arxiv.org/abs/2312.11671 (visited on 07/10/2026).

[139] P. Kirgis, B. Hawriluk, S. Feng, A. Bilimer, S. Paech, and Z. Tufekci. *LLM Spirals of Delusion: A Benchmarking Audit Study of AI Chatbot Interfaces*. Feb. 20, 2026. DOI: 10.48550/arXiv.2604.06188. arXiv: 2604.06188[cs.HC]. URL: http://arxiv.org/abs/2604.06188 (visited on 06/26/2026).

[140] P. Kirgis, S. Kapoor, S. Rabanser, N. Nadgir, C. Ududec, M. Dubois, J. J. Allaire, C. Stosz, M. Hobbhahn, J. Steinhardt, and A. Narayanan. *Log analysis is necessary for credible evaluation of AI agents*. May 8, 2026. DOI: 10.48550/arXiv.2605.08545. arXiv: 2605.08545 [cs.AI]. URL: http://arxiv.org/abs/2605.08545 (visited on 05/28/2026).

[141] Knight First Amendment Institute. *Experimental Publics: Democracy and the Role of Publics in GenAI Evaluation*. URL: http://knightcolumbia.org/content/experimental-publics-democracy-and-the-role-of-publics-in-genai-evaluation (visited on 06/26/2026).

[142] N. Kolt. *Governing AI Agents*. Feb. 11, 2025. DOI: 10.48550/arXiv.2501.07913. arXiv: 2501.07913[cs.AI]. URL: http://arxiv.org/abs/2501.07913.

[143] T. Korbak, M. Balesni, E. Barnes, Y. Bengio, J. Benton, J. Bloom, M. Chen, A. Cooney, A. Dafoe, A. Dragan, S. Emmons, O. Evans, D. Farhi, R. Greenblatt, D. Hendrycks, M. Hobbhahn, E. Hubinger, G. Irving, E. Jenner, D. Kokotajlo, V. Krakovna, S. Legg, D. Lindner, D. Luan, A. Mądry, J. Michael, N. Nanda, D. Orr, J. Pachocki, E. Perez, M. Phuong, F. Roger, J. Saxe, B. Shlegeris, M. Soto, E. Steinberger, J. Wang, W. Zaremba, B. Baker, R. Shah, and V. Mikulik. *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety*. Dec. 7, 2025. DOI: 10.48550/arXiv.2507.11473. arXiv: 2507.11473 [cs.AI]. URL: http://arxiv.org/abs/2507.11473 (visited on 06/08/2026).

[144] V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg. *Specification gaming: the flip side of AI ingenuity*. Google DeepMind. Section: Research. Apr. 21, 2020. URL: https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/ (visited on 09/18/2026).

[145] T. Kwa, B. West, J. Becker, A. Deng, K. Garcia, M. Hasin, S. Jawhar, M. Kinniment, N. Rush, S. V. Arx, R. Bloom, T. Broadley, H. Du, B. Goodrich, N. Jurkovic, L. H. Miles, S. Nix, T. Lin, C. Painter, N. Parikh, D. Rein, L. J. K. Sato, H. Wijk, D. M. Ziegler, E. Barnes, and L. Chan. *Measuring AI Ability to Complete Long Software Tasks*. NeurIPS 2025. 2025. arXiv: 2503.14499 [cs.AI]. URL: https://arxiv.org/abs/2503.14499.

[146] S. Lazar. *Frontier AI Ethics: Anticipating and Evaluating the Societal Impacts of Language Model Agents*. Oct. 18, 2024. DOI: 10.48550/arXiv.2404.06750. arXiv: 2404.06750[cs.CY]. URL: http://arxiv.org/abs/2404.06750.

[147] S. Lee and D. Brumley. *ExploitBench*. June 29, 2026. URL: https://exploitbench.ai/ (visited on 08/15/2026).

[148] S. Lee, P. Achananuparp, N. Yadav, E.-P. Lim, and Y. Deng. *MHSafeEval: Role-Aware Interaction-Level Evaluation of Mental Health Safety in Large Language Models*. 2026. arXiv: 2604.17730 [cs.CL]. URL: https://arxiv.org/abs/2604.17730.

[149] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda. *Holistic Evaluation of Language Models*. Published in Transactions on Machine Learning Research (TMLR), 2023. Oct. 1, 2023. DOI: 10.48550/arXiv.2211.09110. arXiv: 2211.09110 [cs.CL]. URL: http://arxiv.org/abs/2211.09110 (visited on 06/05/2026).

[150] B. Y. Lin, Y. Deng, K. Chandu, F. Brahman, A. Ravichander, V. Pyatkin, N. Dziri, R. L. Bras, and Y. Choi. "WILDBENCH: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild". In: (2024). URL: https://allenai.github.io/WildBench/WildBench_paper.pdf (visited on 09/18/2026).

[151] C. Liu, Y. Guo, Y. Liu, Y. Yang, Q. Yan, X. Zhao, W. Hua, S. Liu, S. Li, Y. Bu, and X. E. Wang. *Auditing Agent Harness Safety*. May 14, 2026. arXiv: 2605.14271 [cs.CL]. URL: https://arxiv.org/abs/2605.14271v2 (visited on 05/31/2026).

[152] X. Liu, J. Tu, Y. Chen, S. Xie, S. Ren, T. Shi, G. Gantar, E. Sandoval, D. Lee, D. Miao, P. J. Gilbert, N. Hynes, M. Staver, W. He, D. Marn, A. Low, X. Zhang, E. Bandel, M. Shmueli-Scheuer, S. Reddy, A. Drouin, A. Lacoste, R. Krishnan, E. Tabassi, Y. Su, V. Barres, C. Wang, W. Guo, and D. Song. *AgentBeats: Agentifying Agent Assessment for Openness, Standardization, and Reproducibility*. 2026. arXiv: 2606.13608 [cs.AI]. URL: https://arxiv.org/abs/2606.13608.

[153] S. Longpre, S. Kapoor, K. Klyman, A. Ramaswami, R. Bommasani, B. Blili-Hamelin, Y. Huang, A. Skowron, Z.-X. Yong, S. Kotha, Y. Zeng, W. Shi, X. Yang, R. Southen, A. Robey, P. Chao, D. Yang, R. Jia, D. Kang, S. Pentland, A. Narayanan, P. Liang, and P. Henderson. *A Safe Harbor for AI Evaluation and Red Teaming*. Mar. 7, 2024. DOI: 10.48550/arXiv.2403.04893. arXiv: 2403.04893. URL: http://arxiv.org/abs/2403.04893 (visited on 07/01/2026).

[154] S. Longpre, K. Klyman, R. E. Appel, S. Kapoor, R. Bommasani, M. Sahar, S. McGregor, A. Ghosh, B. Blili-Hamelin, N. Butters, A. Nelson, A. Elazari, A. Sellars, C. J. Ellis, D. Sherrets, D. Song, H. Geiger, I. Cohen, L. McIlvenny, M. Srikumar, M. M. Jaycox, M. Anderljung, N. F. Johnson, N. Carlini, N. Miailhe, N. Marda, P. Henderson, R. S. Portnoff, R. Weiss, V. Westerhoff, Y. Jernite, R. Chowdhury, P. Liang, and A. Narayanan. *In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI*. Mar. 25, 2025. DOI: 10.48550/arXiv.2503.16861. arXiv: 2503.16861 [cs.AI]. URL: http://arxiv.org/abs/2503.16861 (visited on 06/24/2026).

[155] S. Longpre, E. Zhu, C. Ezell, A. Ghosh, S. McGregor, K. Paeth, K. Klyman, S. Kapoor, R. Bommasani, R. Appel, G. Strom, L. McIlvenny, M. M. Jaycox, P. Slattery, N. Butters, A. Narayanan, P. Liang, and A. Pentland. *FLARE-AI: Flaw Reporting for AI*. arXiv.org. June 30, 2026. URL: https://arxiv.org/abs/2606.31567v1 (visited on 10/17/2026).

[156] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery*. Sept. 1, 2024. DOI: 10.48550/arXiv.2408.06292. arXiv: 2408.06292 [cs.AI]. URL: http://arxiv.org/abs/2408.06292 (visited on 07/03/2026).

[157] L. Luettgau, H. Coppock, M. Dubois, C. Summerfield, and C. Ududec. *HiBayES: A Hierarchical Bayesian Modeling Framework for AI Evaluation Statistics*. July 13, 2025. DOI: 10.48550/arXiv.2505.05602. arXiv: 2505.05602 [cs.AI]. URL: http://arxiv.org/abs/2505.05602 (visited on 07/19/2026).

[158] M. MacDiarmid, B. Wright, J. Uesato, J. Benton, J. Kutasov, S. Price, N. Bouscal, S. Bowman, T. Bricken, A. Cloud, C. Denison, J. Gasteiger, R. Greenblatt, J. Leike, J. Lindsey, V. Mikulik, E. Perez, A. Rodrigues, D. Thomas, A. Webson, D. Ziegler, and E. Hubinger. *Natural Emergent Misalignment from Reward Hacking in Production RL*. Nov. 23, 2025. DOI: 10.48550/arXiv.2511.18397. arXiv: 2511.18397 [cs.AI]. URL: http://arxiv.org/abs/2511.18397 (visited on 06/26/2026).

[159] N. Marchal, S. Chan, M. Franklin, M. Revel, G. Keeling, R. Fischli, B. Chandra, and I. Gabriel. *Architecting Trust in Artificial Epistemic Agents*. Mar. 23, 2026. DOI: 10.48550/arXiv.2603.02960. arXiv: 2603.02960[cs.AI]. URL: http://arxiv.org/abs/2603.02960.

[160] D. K. Markus, F. Gilardi, and D. Stetsenko. *Societal AI Research Has Become Less Interdisciplinary*. June 11, 2025. DOI: 10.48550/arXiv.2506.08738. arXiv: 2506.08738 [cs.CL]. URL: http://arxiv.org/abs/2506.08738 (visited on 07/13/2026).

[161] Martian. *Up and to the left! How Martian Uses Routing to Push the Pareto Frontier*. URL: https://withmartian.com/post/up-and-to-the-left (visited on 06/28/2026).

[162] O. Martin, F. Surani, K. Rodolfa, A. Perez, and D. E. Ho. "The Spectrum of AI Integration: The Case of Benefits Adjudication". In: (2024). Ed. by C. H. Cwik et al. URL: https://dho.stanford.edu/wp-content/uploads/Spectrum_AI.pdf (visited on 09/18/2026).

[163] M. Massenkoff and P. McCrory. *Labor market impacts of AI: A new measure and early evidence*. Anthropic, Mar. 2026. URL: https://www.anthropic.com/research/labor-market-impacts (visited on 06/09/2026).

[164] A. Matton, T. Sherborne, D. Aumiller, E. Tommasone, M. Alizadeh, J. He, R. Ma, M. Voisin, E. Gilsenan-McMahon, and M. Gallé. *On Leakage of Code Generation Evaluation Datasets*. Oct. 3, 2024. DOI:

10.48550/arXiv.2407.07565. arXiv: 2407.07565 [cs.CL]. URL: http://arxiv.org/abs/2407.07565 (visited on 06/26/2026).

[165] M. Mazeika, A. Gatti, C. Menghini, U. M. Sehwag, S. Singhal, Y. Orlovskiy, S. Basart, M. Sharma, D. Peskoff, E. Lau, J. Lim, L. Carroll, A. Blair, V. Sivakumar, S. Basu, B. Kenstler, Y. Ma, J. Michael, X. Li, O. Ingebretsen, A. Mehta, J. Mottola, J. Teichmann, K. Yu, Z. Shaik, A. Khoja, R. Ren, J. Hausenloy, L. Phan, Y. Htet, A. Aich, T. Rabbani, V. Shah, A. Novykov, F. Binder, K. Chugunov, L. Ramirez, M. Geralnik, H. Mesura, D. Lee, E.-Y. H. Cardona, A. Diamond, S. Yue, A. Wang, B. Liu, E. Hernandez, and D. Hendrycks. *Remote Labor Index: Measuring AI Automation of Remote Work*. Oct. 30, 2025. DOI: 10.48550/arXiv.2510.26787. arXiv: 2510.26787 [cs.LG]. URL: http://arxiv.org/abs/2510.26787 (visited on 07/21/2026).

[166] A. Meinke, B. Schoen, J. Scheurer, M. Balesni, R. Shah, and M. Hobbhahn. *Frontier Models are Capable of In-context Scheming*. Jan. 14, 2025. DOI: 10.48550/arXiv.2412.04984. arXiv: 2412.04984 [cs.AI]. URL: http://arxiv.org/abs/2412.04984 (visited on 06/25/2026).

[167] Mercor. *APEX Benchmarks: AI Productivity Index Leaderboards*. URL: https://www.mercor.com/apex/ (visited on 09/18/2026).

[168] Meridian Labs. *Inspect Scout*. URL: https://meridianlabs-ai.github.io/inspect_scout/ (visited on 06/08/2026).

[169] METR. "Guidelines for capability elicitation". In: *METR Blog* (Mar. 15, 2024). URL: https://metr.org/blog/2024-03-15-guidelines-for-capability-elicitation/ (visited on 06/25/2026).

[170] METR. *Details about METR's preliminary evaluation of OpenAI's o3 and o4-mini*. 2025. URL: https://metr.org/evaluations/openai-o3-report/ (visited on 09/18/2026).

[171] METR. "MALT: A Dataset of Natural and Prompted Behaviors That Threaten Eval Integrity". In: *METR Blog* (Oct. 14, 2025). URL: https://metr.org/blog/2025-10-14-malt-dataset-of-natural-and-prompted-behaviors/ (visited on 06/08/2026).

[172] METR. *Recent Frontier Models Are Reward Hacking*. 2025. URL: https://metr.org/blog/2025-06-05-recent-reward-hacking/#chain-of-thought-monitoring-as-a-detection-strategy (visited on 09/18/2026).

[173] METR. "Research Update: Algorithmic vs. Holistic Evaluation". In: *METR Blog* (Aug. 13, 2025). URL: https://metr.org/blog/2025-08-12-research-update-towards-reconciling-slowdown-with-time-horizons/ (visited on 07/19/2026).

[174] METR. *Frontier AI Safety Policies*. May 26, 2026. URL: https://metr.org/fsp (visited on 07/02/2026).

[175] METR. "Frontier Risk Report (February to March 2026)". In: *METR Blog* (May 19, 2026). URL: https://metr.org/blog/2026-05-19-frontier-risk-report/ (visited on 06/08/2026).

[176] METR. "We are Changing our Developer Productivity Experiment Design". In: *METR Blog* (Feb. 24, 2026). URL: https://metr.org/blog/2026-02-24-uplift-update/ (visited on 06/25/2026).

[177] METR. *Notes*. URL: https://metr.org/notes/ (visited on 09/18/2026).

[178] METR. *Task-Completion Time Horizons of Frontier AI Models*. URL: https://metr.org/time-horizons/ (visited on 07/01/2026).

[179] METR and Redwood Research. *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. Aug. 26, 2026. URL: https://metr.org/hugging-face-incident-report-aug-2026.pdf (visited on 10/09/2026).

[180] J. Miao, J. K. Pritchard, and J. Zou. *The Agentic Garden of Forking Paths*. July 1, 2026. DOI: 10.48550/arXiv.2607.01507. arXiv: 2607.01507 [cs.AI]. URL: http://arxiv.org/abs/2607.01507 (visited on 07/09/2026).

[181] Michael Skarlinski, Jon Laurent, Albert Bou, and Andrew White. *About 30% of Humanity's Last Exam Chemistry/Biology Answers are Likely Wrong*. July 23, 2025. URL: https://www.futurehouse.org/research/hle-exam (visited on 09/18/2026).

[182] E. Miller. *Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations*. Nov. 1, 2024. DOI: 10.48550/arXiv.2411.00640. arXiv: 2411.00640 [stat.AP]. URL: http://arxiv.org/abs/2411.00640 (visited on 06/23/2026).

[183] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar. *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*. Aug. 27, 2025. DOI: 10.48550/arXiv.2410.05229. arXiv: 2410.05229 [cs.LG]. URL: http://arxiv.org/abs/2410.05229 (visited on 09/28/2026).

[184] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto. "s1: Simple test-time scaling". In: *Proceedings of the 2025 Conference on Empirical Methods in*

Natural Language Processing. Ed. by C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 20275–20321. DOI: 10.18653/v1/2025.emnlp-main.1025. URL: https://aclanthology.org/2025.emnlp-main.1025/ (visited on 09/30/2026).

[185] C. Murphy. *Agent Island: A Saturation- and Contamination-Resistant Benchmark from Multiagent Games*. May 5, 2026. DOI: 10.48550/arXiv.2605.04312. arXiv: 2605.04312 [cs.AI]. URL: http://arxiv.org/abs/2605.04312 (visited on 06/07/2026).

[186] N. Nadgir, S. Kapoor, K. Liu, P. Kirgis, M. Orona, S. Rabanser, T. Bayer, A. Shetty, Y. Ling, D. Chan-Sew, R. Nakagawa, S. Utpala, Z. S. Siegel, and A. Narayanan. *Life After Benchmark Saturation: A Case Study of CORE-Bench*. June 23, 2026. arXiv: 2606.26158 [cs.AI]. URL: https://arxiv.org/abs/2606.26158v1 (visited on 07/03/2026).

[187] D. Nathani, L. Madaan, N. Roberts, N. Bashlykov, A. Menon, V. Moens, A. Budhiraja, D. Magka, V. Vorotilov, G. Chaurasia, D. Hupkes, R. S. Cabral, T. Shavrina, J. Foerster, Y. Bachrach, W. Y. Wang, and R. Raileanu. *MLGym: A New Framework and Benchmark for Advancing AI Research Agents*. Feb. 20, 2025. DOI: 10.48550/arXiv.2502.14499. arXiv: 2502.14499 [cs.CL]. URL: http://arxiv.org/abs/2502.14499 (visited on 07/03/2026).

[188] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. National Institute of Standards and Technology, 2023. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (visited on 09/18/2026).

[189] National Institute of Standards and Technology. "Technical Blog: Strengthening AI Agent Hijacking Evaluations". In: *NIST* (Jan. 17, 2025). Last Modified: 2025-12-19T17:12-05:00. URL: https://www.nist.gov/news-events/news/2025/01/technical-blog-strengthening-ai-agent-hijacking-evaluations (visited on 06/25/2026).

[190] National Institute of Standards and Technology. *artificial intelligence red-teaming - Glossary*. URL: https://csrc.nist.gov/glossary/term/artificial_intelligence_red_teaming (visited on 09/18/2026).

[191] F. Ndzomga. *Efficient Benchmarking of AI Agents*. 2026. arXiv: 2603.23749 [cs.AI]. URL: https://arxiv.org/abs/2603.23749.

[192] J. Needham, G. Edkins, G. Pimpale, H. Bartsch, and M. Hobbhahn. *Large Language Models Often Know When They Are Being Evaluated*. 2025. arXiv: 2505.23836 [cs.CL]. URL: https://arxiv.org/abs/2505.23836.

[193] NeurIPS. *Call For Evaluations & Datasets 2026*. URL: https://neurips.cc/Conferences/2026/CallForEvaluationsDatasets (visited on 09/18/2026).

[194] Nicholas Carlini. *Building a C compiler with a team of parallel Claudes*. Engineering at Anthropic. Feb. 5, 2026. URL: https://www.anthropic.com/engineering/building-c-compiler (visited on 06/08/2026).

[195] Nicolas Emmenegger, Young-Kyung Kim, Alessandro Morosini, Adam J. Berinsky, Chara Podimata, and Charles Stewart III. *Election LLM Observatory*. URL: https://www.llmelectionobservatory.org/questions (visited on 10/15/2026).

[196] NIST. "Accelerating AI Innovation Through Measurement Science". In: *NIST* (Dec. 2, 2025). Last Modified: 2026-02-17T16:23-05:00. URL: https://www.nist.gov/blogs/caisi-research-blog/accelerating-ai-innovation-through-measurement-science (visited on 06/25/2026).

[197] OECD. *OECD Framework for the Classification of AI systems*. 323. Paris: OECD Publishing, Feb. 22, 2022. DOI: 10.1787/cb6d9eca-en. URL: https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html (visited on 08/12/2026).

[198] V. Ojewale, R. Steed, B. Vecchione, A. Birhane, and I. D. Raji. *Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling*. 2024. DOI: 10.1145/3706598.3713301. arXiv: 2402.17861 [cs.CY]. URL: https://arxiv.org/abs/2402.17861.

[199] A. Olteanu, S. L. Blodgett, A. Balayn, A. Wang, F. Diaz, F. du Pin Calmon, M. Mitchell, M. Ekstrand, R. Binns, and S. Barocas. *Rigor in AI: Doing Rigorous AI Work Requires a Broader, Responsible AI-Informed Conception of Rigor*. 2025. arXiv: 2506.14652 [cs.CY]. URL: https://arxiv.org/abs/2506.14652.

[200] OpenAI. *Superalignment Fast Grants*. Dec. 14, 2023. URL: https://openai.com/index/superalignment-fast-grants/.

[201] OpenAI. *Building an early warning system for LLM-aided biological threat creation*. Jan. 31, 2024. URL: https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/ (visited on 07/05/2026).

[202] OpenAI. *Preparedness Framework (Version 2)*. Apr. 15, 2025. URL: https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf (visited on 06/09/2026).
[203] OpenAI. *Security on the path to AGI*. Mar. 26, 2025. URL: https://openai.com/index/security-on-the-path-to-agi/ (visited on 10/19/2026).
[204] OpenAI. *Strengthening our safety ecosystem with external testing*. Nov. 19, 2025. URL: https://openai.com/index/strengthening-safety-with-external-testing/ (visited on 06/25/2026).
[205] OpenAI. *Working with US CAISI and UK AISI to build more secure AI systems*. Sept. 12, 2025. URL: https://openai.com/index/us-caisi-uk-aisi-ai-update/ (visited on 06/26/2026).
[206] OpenAI. *GPT-5.6: Frontier intelligence that scales with your ambition*. July 9, 2026. URL: https://openai.com/index/gpt-5-6/.
[207] OpenAI. *Introducing Trusted Access for Cyber*. Feb. 5, 2026. URL: https://openai.com/index/trusted-access-for-cyber/.
[208] OpenAI. *OpenAI – Hugging Face Incident Technical Report*. Aug. 26, 2026. URL: https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf.
[209] OpenAI. *OpenAI and Hugging Face partner to address security incident during model evaluation*. July 21, 2026. URL: https://openai.com/index/hugging-face-model-evaluation-security-incident/ (visited on 10/17/2026).
[210] OpenAI. *Request for Proposals: OpenAI Economic Research Exchange*. June 8, 2026. URL: https://openai.com/index/economic-research-exchange-request-for-proposals/.
[211] OpenAI. *Safety and alignment in an era of long-horizon models*. July 20, 2026. URL: https://openai.com/index/safety-alignment-long-horizon-models/ (visited on 07/21/2026).
[212] OpenAI. *Our Structure*. URL: https://openai.com/our-structure/ (visited on 09/18/2026).
[213] OpenAI. *Reasoning models*. OpenAI Developers. URL: https://developers.openai.com/api/docs/guides/reasoning (visited on 10/11/2026).
[214] OpenAI Foundation. *Update on the OpenAI Foundation*. Mar. 24, 2026. URL: https://openaifoundation.org/news/update-on-the-openai-foundation (visited on 06/08/2026).
[215] OpenAI Foundation. *Resilience in the age of AI*. URL: https://openaifoundation.org/news/resilience-in-the-age-of-ai#ai-model-safety (visited on 09/18/2026).
[216] A. Orben and J. N. Matias. "Fixing the science of digital technology harms". In: *Science* 388.6743 (Apr. 11, 2025), pp. 152–155. DOI: 10.1126/science.adt6807. URL: https://www.science.org/doi/full/10.1126/science.adt6807 (visited on 10/14/2026).
[217] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. *Generative Agents: Interactive Simulacra of Human Behavior*. Aug. 6, 2023. DOI: 10.48550/arXiv.2304.03442. arXiv: 2304.03442 [cs.HC]. URL: http://arxiv.org/abs/2304.03442 (visited on 07/24/2026).
[218] P. Paskov, M. J. Byun, K. Wei, and T. Webster. *Preliminary suggestions for rigorous GPAI model evaluations*. 2025. DOI: 10.7249/PEA3971-1. arXiv: 2508.00875 [cs.CY]. URL: http://arxiv.org/abs/2508.00875 (visited on 07/02/2026).
[219] P. Paskov, K. Wei, S. Z. Hong, D. Bateyko, X. Roberts-Gaal, C. Ezell, G. Praninskas, V. Chen, U. Bhatt, and E. Guest. *RCTs for Frontier AI Governance: Methodological Challenges and Solutions for Human Uplift Studies*. version: 3. June 30, 2026. DOI: 10.48550/arXiv.2603.11001. arXiv: 2603.11001 [cs.CY]. URL: http://arxiv.org/abs/2603.11001 (visited on 08/12/2026).
[220] T. Patwardhan, R. Dias, E. Proehl, G. Kim, M. Wang, O. Watkins, S. P. Fishman, M. Aljubeh, P. Thacker, L. Fauconnet, N. S. Kim, P. Chao, S. Miserendino, G. Chabot, D. Li, M. Sharman, A. Barr, A. Glaese, and J. Tworek. *GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks*. Oct. 5, 2025. arXiv: 2510.04374 [cs.LG]. URL: https://arxiv.org/abs/2510.04374v1 (visited on 06/03/2026).
[221] J. Pencharz, T. Flesch, and J. Sandbrink. *Long-Form Tasks: A Methodology for Evaluating Scientific Assistants*. AISI. Dec. 3, 2024. URL: https://www.aisi.gov.uk/blog/long-form-tasks (visited on 07/02/2026).
[222] S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer. *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot*. Feb. 13, 2023. DOI: 10.48550/arXiv.2302.06590. arXiv: 2302.06590 [cs.SE]. URL: http://arxiv.org/abs/2302.06590 (visited on 07/05/2026).

[223] S. Rabanser, S. Kapoor, P. Kirgis, K. Liu, S. Utpala, and A. Narayanan. *Towards a Science of AI Agent Reliability*. Feb. 23, 2026. DOI: 10.48550/arXiv.2602.16666. arXiv: 2602.16666 [cs.AI]. URL: http://arxiv.org/abs/2602.16666 (visited on 05/26/2026).

[224] I. D. Raji, P. Xu, C. Honigsberg, and D. Ho. "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance". In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. Oxford United Kingdom: ACM, July 26, 2022, pp. 557–571. DOI: 10.1145/3514094.3534181. URL: https://dl.acm.org/doi/10.1145/3514094.3534181 (visited on 06/26/2026).

[225] B. Rank, H. Bhatnagar, A. Prabhu, S. Eisenberg, K. Nguyen, M. Bethge, and M. Andriushchenko. "PostTrainBench: Can LLM Agents Automate LLM Post-Training?" In: *International Conference on Machine Learning (ICML)*. 2026. URL: https://arxiv.org/abs/2603.08640.

[226] N. Ransohoff. *The third wave of American philanthropy*. May 19, 2026. URL: https://nanransohoff.substack.com/p/the-third-wave-of-american-philanthropy (visited on 05/26/2026).

[227] M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, R. Comanescu, C. Akbulut, T. Stepleton, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, I. Gabriel, V. Rieser, W. Isaac, and L. Weidinger. "Gaps in the Safety Evaluation of Generative AI". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7.1 (Oct. 16, 2024), pp. 1200–1217. DOI: 10.1609/aies.v7i1.31717. URL: https://ojs.aaai.org/index.php/AIES/article/view/31717 (visited on 06/07/2026).

[228] D. Rein, J. Becker, A. Deng, S. Nix, C. Canal, D. O'Connel, P. Arnott, R. Bloom, T. Broadley, K. Garcia, B. Goodrich, M. Hasin, S. Jawhar, M. Kinniment, T. Kwa, A. Lajko, N. Rush, L. J. K. Sato, S. V. Arx, B. West, L. Chan, and E. Barnes. *HCAST: Human-Calibrated Autonomy Software Tasks*. Mar. 21, 2025. DOI: 10.48550/arXiv.2503.17354. arXiv: 2503.17354 [cs.AI]. URL: http://arxiv.org/abs/2503.17354 (visited on 07/03/2026).

[229] I. A. S. Report. "International AI Safety Report 2026". In: (Feb. 2026). URL: https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026_1.pdf (visited on 09/18/2026).

[230] A. Reuel, A. Ghosh, J. Chim, A. Tran, Y. Long, J. Mickel, U. Gohar, S. Yadav, P. S. Ammanamanchi, M. Allaham, H. A. Rahmani, M. Akhtar, F. Friedrich, R. Scholz, M. A. Riegler, J. Batzner, E. Habba, A. Saxena, A. Kornilova, K. Wei, P. Soni, Y. Mathew, K. Klyman, J. Sania, S. Sahoo, O. B. Bruvik, P. Sadeghi, S. Goswami, A. Wang, Y. Jernite, Z. Talat, S. Biderman, M. Kochenderfer, S. Koyejo, and I. Solaiman. *Who Evaluates AI's Social Impacts? Mapping Coverage and Gaps in First and Third Party Evaluations*. May 30, 2026. DOI: 10.48550/arXiv.2511.05613. arXiv: 2511.05613 [cs.CY]. URL: http://arxiv.org/abs/2511.05613 (visited on 06/12/2026).

[231] J. Roberts, M. R. Taesiri, A. Sharma, A. Gupta, S. Roberts, I. Croitoru, S.-V. Bogolin, J. Tang, F. Langer, V. Raina, V. Raina, H. Xiong, V. Udandarao, J. Lu, S. Chen, S. Purkis, T. Yan, W. Lin, G. Shin, Q. Yang, A. T. Nguyen, D. I. Atkinson, A. Baranwal, A. Coca, M. Dang, S. Dziadzio, J. D. Kunz, K. Liang, A. Lo, B. Pulfer, S. Walton, C. Yang, K. Han, and S. Albanie. *ZeroBench: An Impossible Visual Benchmark for Contemporary Large Multimodal Models*. July 7, 2026. DOI: 10.48550/arXiv.2502.09696. arXiv: 2502.09696 [cs.CV]. URL: http://arxiv.org/abs/2502.09696 (visited on 08/14/2026).

[232] Ryan Marten [@ryan_marten]. *Continuous Benchmarks*. X/Twitter. July 23, 2026. URL: https://x.com/ryan_marten/status/2080321791361527843 (visited on 09/18/2026).

[233] Sakana AI. *AI-Scientist: The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery*. URL: https://github.com/SakanaAI/AI-Scientist (visited on 09/30/2026).

[234] 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 26, 2025. DOI: 10.48550/arXiv.2505.10573. arXiv: 2505.10573 [cs.CY]. URL: http://arxiv.org/abs/2505.10573 (visited on 09/11/2026).

[235] Scale AI. *The Remote Labor Index: Measuring the Automation of Work*. Oct. 29, 2025. URL: https://scale.com/blog/rli (visited on 07/21/2026).

[236] Scale AI. *HiL-Bench (Human-in-Loop Benchmark)*. Sept. 18, 2026. URL: https://labs.scale.com/leaderboard (visited on 06/09/2026).

[237] J. Scheurer, M. Balesni, and M. Hobbhahn. *Large Language Models can Strategically Deceive their Users when Put Under Pressure*. July 15, 2024. DOI: 10.48550/arXiv.2311.07590. arXiv: 2311.07590 [cs.CL]. URL: http://arxiv.org/abs/2311.07590 (visited on 07/03/2026).

[238] Schmidt Sciences. *New $10 Million AI Safety Science Program Launched for Foundational Research*. Feb. 11, 2025. URL: https://www.schmidtsciences.org/new-10-million-ai-safety-science-program-launched-for-foundational-research/ (visited on 09/18/2026).

[239] B. Schoen, E. Nitishinskaya, M. Balesni, A. Højmark, F. Hofstätter, J. Scheurer, A. Meinke, J. Wolfe, T. v. d. Weij, A. Lloyd, N. Goldowsky-Dill, A. Fan, A. Matveiakin, R. Shah, M. Williams, A. Glaese, B. Barak, W. Zaremba, and M. Hobbhahn. *Stress Testing Deliberative Alignment for Anti-Scheming Training*. Sept. 19, 2025. DOI: 10.48550/arXiv.2509.15541. arXiv: 2509.15541 [cs.AI]. URL: http://arxiv.org/abs/2509.15541 (visited on 07/08/2026).

[240] SecureBio. *Benchmark Trends – SecureBio*. June 26, 2026. URL: https://securebio.org/benchmarks/ (visited on 07/03/2026).

[241] P. Seshadri, S. Cahyawijaya, A. Odumakinde, S. Singh, and S. Goldfarb-Tarrant. *Lost in Simulation: LLM-Simulated Users are Unreliable Proxies for Human Users in Agentic Evaluations*. Jan. 28, 2026. DOI: 10.48550/arXiv.2601.17087. arXiv: 2601.17087 [cs.HC]. URL: http://arxiv.org/abs/2601.17087 (visited on 07/24/2026).

[242] N. B. Shah. *Asking Authors About Their Own Papers*. TMLR Blog. Sept. 16, 2026. URL: https://blog.tmlr.org/2026/asking-authors-about-their-own-papers/.

[243] Y. Shao, V. Samuel, Y. Jiang, J. Yang, and D. Yang. *Collaborative Gym: A Framework for Enabling and Evaluating Human-Agent Collaboration*. Dec. 6, 2025. DOI: 10.48550/arXiv.2412.15701. arXiv: 2412.15701 [cs.AI]. URL: http://arxiv.org/abs/2412.15701 (visited on 06/08/2026).

[244] Y. Shao, Z. Z. Wang, N. Ahuja, Y. Wang, B. Liu, and D. Yang. *CollabSkill: Evaluating Human-Agent Collaboration On Real-World Tasks*. Apr. 20, 2026. DOI: 10.48550/arXiv.2606.09833. arXiv: 2606.09833 [cs.HC]. URL: http://arxiv.org/abs/2606.09833 (visited on 07/24/2026).

[245] Y. Shao, H. Zope, Y. Jiang, J. Pei, D. Nguyen, E. Brynjolfsson, and D. Yang. *Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce*. Feb. 1, 2026. DOI: 10.48550/arXiv.2506.06576. arXiv: 2506.06576 [cs.CY]. URL: http://arxiv.org/abs/2506.06576 (visited on 06/08/2026).

[246] Y. Shavit, S. Agarwal, M. Brundage, S. Adler, C. O'Keefe, R. Campbell, T. Lee, P. Mishkin, T. Eloundou, A. Hickey, K. Slama, L. Ahmad, P. McMillan, A. Beutel, A. Passos, and D. G. Robinson. *Practices for Governing Agentic AI Systems*. OpenAI, Dec. 2023. URL: https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

[247] A. Shirali, R. Abebe, and M. Hardt. *A Theory of Dynamic Benchmarks*. Mar. 1, 2023. DOI: 10.48550/arXiv.2210.03165. arXiv: 2210.03165 [cs.LG]. URL: http://arxiv.org/abs/2210.03165 (visited on 07/24/2026).

[248] Z. S. Siegel, S. Kapoor, N. Nadgir, B. Stroebl, and A. Narayanan. *CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark*. Sept. 17, 2024. DOI: 10.48550/arXiv.2409.11363. arXiv: 2409.11363 [cs.CL]. URL: http://arxiv.org/abs/2409.11363 (visited on 06/08/2026).

[249] S. Singh, Y. Nan, A. Wang, D. D'Souza, S. Kapoor, A. Üstün, S. Koyejo, Y. Deng, S. Longpre, N. A. Smith, B. Ermis, M. Fadaee, and S. Hooker. *The Leaderboard Illusion*. May 12, 2025. DOI: 10.48550/arXiv.2504.20879. arXiv: 2504.20879 [cs.AI]. URL: http://arxiv.org/abs/2504.20879 (visited on 06/08/2026).

[250] J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger. *Defining and Characterizing Reward Hacking*. 2022. arXiv: 2209.13085 [cs.LG]. URL: https://arxiv.org/abs/2209.13085.

[251] Stanford Digital Economy Lab. *The AI Economic Indicators*. June 10, 2026. URL: https://digitaleconomy.stanford.edu/project/indicators/ (visited on 09/18/2026).

[252] G. Starace, O. Jaffe, D. Sherburn, J. Aung, J. S. Chan, L. Maksin, R. Dias, E. Mays, B. Kinsella, W. Thompson, J. Heidecke, A. Glaese, and T. Patwardhan. *PaperBench: Evaluating AI's Ability to Replicate AI Research*. Apr. 7, 2025. DOI: 10.48550/arXiv.2504.01848. arXiv: 2504.01848 [cs.AI]. URL: http://arxiv.org/abs/2504.01848 (visited on 06/08/2026).

[253] A. Stein, D. Brown, H. Hassani, M. Naik, and E. Wong. *Detecting Safety Violations Across Many Agent Traces*. Apr. 13, 2026. DOI: 10.48550/arXiv.2604.11806. arXiv: 2604.11806 [cs.AI]. URL: http://arxiv.org/abs/2604.11806 (visited on 07/22/2026).

[254] E. Stengel-Eskin, N. Sander, C. Bonetti, S. Boguraev, J. Bowler, H. Sirin, and S. Kirby. *GlossoGen: Emergent Language in Complex Multi-Agent LLM Interactions*. 2026. arXiv: 2609.01491 [cs.CL]. URL: https://arxiv.org/abs/2609.01491.

[255] C. Stosz, K. Elmgren, C. Foster, G. Balston, S. Donoughe, S. Nedungadi, M. Chen, J. Götting, P. Paskov, S. Kapoor, S. Schwettmann, R. Bommasani, L. Righetti, S. McGregor, G. Werner, R. Reich, A. Narayanan, E. Barnes, C. Painter, M. Brundage, A. Homewood, D. Siddharth, F. Lalani, C. Teague, J. Sevilla, and J. Steinhardt. "AEF-1: Minimum Operating Conditions for Independent Third Party AI Evaluations". In: (2025). URL: https://www.aef.one/aef-one.pdf (visited on 09/18/2026).

[256] G. Straus and A. Hall. *How Accurately Did Claude Code Replicate and Extend a Published Political Science Paper?* Jan. 2026. URL: https://andrewbenjaminhall.com/pdfs/Straus_Hall_Claude_Audit.pdf (visited on 09/18/2026).

[257] C. Summerfield, L. Luettgau, M. Dubois, H. R. Kirk, K. Hackenburg, C. Fist, K. Slama, N. Ding, R. Anselmetti, A. Strait, M. Giulianelli, and C. Ududec. *Lessons from a Chimp: AI "Scheming" and the Quest for Ape Language*. 2025. arXiv: 2507.03409 [cs.AI]. URL: https://arxiv.org/abs/2507.03409.

[258] F. Surani, L. A. Gailmard, A. Casasola, V. Magesh, E. J. Robitschek, and D. E. Ho. "What Is the Law? A System for Statutory Research (STARA) with Large Language Models". In: (2025). URL: https://dho.stanford.edu/wp-content/uploads/STARA.pdf (visited on 09/18/2026).

[259] Survival and Flourishing Fund. *Survival and Flourishing Fund*. URL: https://survivalandflourishing.fund/ (visited on 09/18/2026).

[260] K. Tadepalli. *AI automation measures are not up to the task*. Apr. 28, 2026. URL: https://blog.karthiktadepalli.com/p/tasks (visited on 07/19/2026).

[261] S. Tan, S. Zhuang, K. Montgomery, W. Y. Tang, A. Cuadron, C. Wang, R. A. Popa, and I. Stoica. *JudgeBench: A Benchmark for Evaluating LLM-based Judges*. Apr. 5, 2025. DOI: 10.48550/arXiv.2410.12784. arXiv: 2410.12784 [cs.AI]. URL: http://arxiv.org/abs/2410.12784 (visited on 07/03/2026).

[262] L. A. Team. *LMSYS Chatbot Arena: Live and Community-Driven LLM Evaluation - LMSYS Blog*. Mar. 1, 2024. URL: https://lmsys.org/blog/2024-03-01-policy/ (visited on 06/09/2026).

[263] E. Tennant, B. Henke, A. Keshmirian, M. Shanahan, V. Rieser, K. Lum, S. Levine, and J. Haas. *Normative Robustness as a Frontier for Non-Verifiable Reasoning in LLMs*. June 10, 2026. DOI: 10.48550/arXiv.2606.12731. arXiv: 2606.12731 [cs.LG]. URL: http://arxiv.org/abs/2606.12731 (visited on 07/20/2026).

[264] The AI Security Institute (AISI). *Grants*. URL: https://www.aisi.gov.uk/grants (visited on 10/01/2026).

[265] Transluce. *Monitoring SWE-bench Agents*. Nov. 19, 2025. URL: https://transluce.org/docent/blog/swe-bench (visited on 06/25/2026).

[266] Trenton Bricken, Rowan Wang, Sam Bowman, Euan Ong, Johannes Treutlein, Jeff Wu, Evan Hubinger, and Samuel Marks. *Building and evaluating alignment auditing agents*. Alignment Science Blog. July 24, 2025. URL: https://alignment.anthropic.com/2025/automated-auditing/ (visited on 10/17/2026).

[267] C. Trout, S. Koyejo, S. Romanosky, G. Ripamonti, L. Thompson, D. Spain, A. Taylor, K. Casey, S. Casper, M. Botvinick, S. McGregor, M. Brundage, A. F. Cooper, P. Paskov, A. Ecoffet, B. Bucknall, K. Wei, M. Anderljung, L. Szpruch, B. Treece, T. Zick, G. Weil, U. Ozer, K. Kalinich, J. Gonzalez, V. Baranov, M. Koren, G. Laban, G. Arazi, H. Winand, D. Blum, T. Clowes, A. Kleinman, A. Srinivasan, T. Fehring, R. Kvist, and R. Dattani. *Underwriting the Agent Economy: The Blueprint for an AI Insurance Stack*. July 15, 2026. DOI: 10.48550/arXiv.2607.11999. arXiv: 2607.11999[cs.CY]. URL: http://arxiv.org/abs/2607.11999 (visited on 07/17/2026).

[268] S. Truong, Y. Tu, M. Hardy, A. Reuel-Lamparth, Z. Tang, J. Burapacheep, J. Perera, C. Uwakwe, B. Domingue, N. Haber, and S. Koyejo. "Fantastic Bugs and Where to Find Them in AI Benchmarks". In: *Advances in Neural Information Processing Systems*. Ed. by D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen. Vol. 38. Curran Associates, Inc., 2025. DOI: 10.52202/085713-2118. URL: https://proceedings.neurips.cc/paper_files/paper/2025/file/5b44b99d012922011f0be1031a98c4f9-Paper-Datasets_and_Benchmarks_Track.pdf (visited on 09/18/2026).

[269] U.S. Department of Labor. *O*NET*. URL: https://www.dol.gov/agencies/eta/onet (visited on 09/18/2026).

[270] U.S. General Services Administration. *USAi.gov*. URL: https://www.usai.gov/ (visited on 09/18/2026).

[271] U.S. House of Representatives. *Obernolte, Trahan release a discussion draft of the Great American AI Act*. June 4, 2026. URL: https://obernolte.house.gov/media/press-releases/obernolte-trahan-release-discussion-draft-great-american-ai-act (visited on 09/18/2026).

[272] UK AI Security Institute. *A pipeline for transcript analysis using Inspect Scout | AISI Work*. URL: https://www.aisi.gov.uk/blog/a-pipeline-for-transcript-analysis-using-inspect-scout (visited on 06/15/2026).
[273] UK AI Security Institute. *HiBayES*. URL: https://ukgovernmentbeis.github.io/hibayes/ (visited on 09/18/2026).
[274] UK AI Security Institute. *Inspect*. URL: https://inspect.aisi.org.uk/ (visited on 09/18/2026).
[275] UK AI Security Institute. *More compute, more capability: Why AI agent evaluations need to account for test-time compute | AISI Work*. URL: https://www.aisi.gov.uk/blog/more-compute-more-capability-why-ai-agent-evals-need-to-account-for-test-time-compute (visited on 09/18/2026).
[276] UK AI Security Institute. *RepliBench: measuring autonomous replication capabilities in AI systems | AISI Work*. URL: https://www.aisi.gov.uk/blog/replibench-measuring-autonomous-replication-capabilities-in-ai-systems (visited on 07/02/2026).
[277] UK Government. *AI Safety Institute approach to evaluations*. Feb. 9, 2024. URL: https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations (visited on 06/25/2026).
[278] US AI Safety Institute and UK AI Safety Institute. *US AISI and UK AISI Joint Pre-Deployment Test: OpenAI o1*. Dec. 2024. URL: https://www.nist.gov/system/files/documents/2024/12/18/US_UK_AI%20Safety%20Institute_%20December_Publication-OpenAIo1.pdf (visited on 09/18/2026).
[279] Vals AI. *Vals AI*. URL: https://www.vals.ai/ (visited on 06/15/2026).
[280] V. Veselovsky, M. H. Ribeiro, and R. West. *Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks*. 2023. arXiv: 2306.07899 [cs.CL]. URL: https://arxiv.org/abs/2306.07899.
[281] B. Vidgen, A. Mann, A. Fennelly, J. W. Stanly, L. Rothman, M. Burstein, J. Benchek, D. Ostrofsky, A. Ravichandran, D. Sur, N. Venugopal, A. Hsia, I. Robinson, C. Huang, O. Varones, D. Khan, M. Haines, A. Bridges, J. Boyle, K. Twist, Z. Richards, C. Mahapatra, B. Foody, and O. Nitski. *APEX-Agents*. Feb. 23, 2026. DOI: 10.48550/arXiv.2601.14242. arXiv: 2601.14242 [cs.CL]. URL: http://arxiv.org/abs/2601.14242 (visited on 06/08/2026).
[282] H. Wallach, M. Desai, N. Pangakis, A. F. Cooper, A. Wang, S. Barocas, A. Chouldechova, C. Atalla, S. L. Blodgett, E. Corvi, P. A. Dow, J. Garcia-Gathright, A. Olteanu, S. Reed, E. Sheng, D. Vann, J. W. Vaughan, M. Vogel, H. Washington, and A. Z. Jacobs. *Evaluating Generative AI Systems is a Social Science Measurement Challenge*. Nov. 17, 2024. DOI: 10.48550/arXiv.2411.10939. arXiv: 2411.10939 [cs.CY]. URL: http://arxiv.org/abs/2411.10939 (visited on 06/07/2026).
[283] A. Wan, S. Hatgis-Kessell, T. Aguirre, P. Liang, and R. Bommasani. *Economic Evaluations of Language Models*. June 26, 2026. DOI: 10.48550/arXiv.2607.19375. arXiv: 2607.19375 [cs.CY]. URL: http://arxiv.org/abs/2607.19375 (visited on 07/30/2026).
[284] S. Wan, C. Nikolaidis, D. Song, D. Molnar, J. Crnkovich, J. Grace, M. Bhatt, S. Chennabasappa, S. Whitman, S. Ding, V. Ionescu, Y. Li, and J. Saxe. *CYBERSECEVAL 3: Advancing the Evaluation of Cybersecurity Risks and Capabilities in Large Language Models*. Sept. 6, 2024. DOI: 10.48550/arXiv.2408.01605. arXiv: 2408.01605 [cs.CR]. URL: http://arxiv.org/abs/2408.01605 (visited on 10/01/2026).
[285] A. Wang, D. E. Ho, and S. Koyejo. *The Inadequacy of Offline LLM Evaluations: A Need to Account for Personalization in Model Behavior*. 2025. arXiv: 2509.19364 [cs.CL]. URL: https://arxiv.org/abs/2509.19364.
[286] J. Wang, W. Wang, L. Fang, and B. Tse. *The Safety–Openness Pareto Frontier: Technical Report*. Sept. 21, 2026. URL: https://airiskmonitor.net/doc/en/blog/safety-openness-pareto.
[287] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen. "A Survey on Large Language Model based Autonomous Agents". In: *Frontiers of Computer Science* 18.6 (Dec. 2024), p. 186345. DOI: 10.1007/s11704-024-40231-1. arXiv: 2308.11432 [cs.AI]. URL: http://arxiv.org/abs/2308.11432 (visited on 09/27/2026).
[288] X. Wang, L. Jiang, J. Hernández-Orallo, D. Stillwell, S. Chen, L. Sun, F. Luo, and X. Xie. "Evaluating General-Purpose AI with Psychometrics". In: *Commun. ACM* 69.5 (Apr. 2026), pp. 92–102. DOI: 10.1145/3769688. URL: https://doi.org/10.1145/3769688.
[289] Y. Wang, C. Zhu, T. Feng, L. X. Lu, and B. Jia. *Agentic AI and Human-in-the-Loop Interventions: Field Experimental Evidence from Alibaba's Customer Service Operations*. June 1, 2026. DOI:

10.48550/arXiv.2605.14830. arXiv: 2605.14830 [cs.HC]. URL: http://arxiv.org/abs/2605.14830 (visited on 07/05/2026).

[290] Z. Z. Wang, Y. Shao, O. Shaikh, D. Fried, G. Neubig, and D. Yang. *How Do AI Agents Do Human Work? Comparing AI and Human Workflows Across Diverse Occupations*. Nov. 6, 2025. DOI: 10.48550/arXiv.2510.22780. arXiv: 2510.22780 [cs.AI]. URL: http://arxiv.org/abs/2510.22780 (visited on 06/08/2026).

[291] Z. Z. Wang, J. Yang, K. Lieret, A. Tartaglini, V. Chen, Y. Wei, Z. Wang, L. Zhang, K. Narasimhan, L. Schmidt, G. Neubig, D. Fried, and D. Yang. "Position: Humans are Missing from AI Coding Agent Research". In: (2026). URL: https://zorazrw.github.io/files/position-haicode.pdf (visited on 09/18/2026).

[292] K. L. Wei, P. Paskov, S. Dev, M. J. Byun, A. Reuel, X. Roberts-Gaal, R. Calcott, E. Coxon, and C. Deshpande. *Recommendations and Reporting Checklist for Rigorous & Transparent Human Baselines in Model Evaluations*. 2025. arXiv: 2506.13776 [cs.AI]. URL: https://arxiv.org/abs/2506.13776.

[293] L. Weidinger, I. D. Raji, H. Wallach, M. Mitchell, A. Wang, O. Salaudeen, R. Bommasani, S. Koyejo, and W. Isaac. "Toward an Evaluation Science for Generative AI Systems". In: *The Bridge* 55.1 (Apr. 15, 2025), pp. 9–18. URL: https://www.nae.edu/19579/19582/21020/337862/338231/Toward-an-Evaluation-Science-for-Generative-AI-Systems (visited on 07/08/2026).

[294] L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, I. Gabriel, V. Rieser, and W. Isaac. *Sociotechnical Safety Evaluation of Generative AI Systems*. Oct. 31, 2023. DOI: 10.48550/arXiv.2310.11986. arXiv: 2310.11986 [cs.AI]. URL: http://arxiv.org/abs/2310.11986 (visited on 06/08/2026).

[295] T. van der Weij, F. Hofstätter, and F. R. Ward. *An Introduction to AI Sandbagging*. Apr. 26, 2024. URL: https://www.lesswrong.com/posts/jsmNCj9QKcfdg8fJk/an-introduction-to-ai-sandbagging (visited on 09/18/2026).

[296] G. Weil. "Don't Let AI Developers Hire Their Own Referees". In: *AI Frontiers* (July 29, 2026). URL: https://ai-frontiers.org/articles/dont-let-ai-developers-hire-their-own-referees (visited on 07/31/2026).

[297] V. Weilnhammer, K. Y. Hou, L. Luettgau, C. Summerfield, R. Dolan, and M. M. Nour. "A clinically validated framework for auditing AI chatbot behavior in mental health interactions". In: *Nature Medicine* (Aug. 7, 2026), pp. 1–11. DOI: 10.1038/s41591-026-04577-2. URL: https://www.nature.com/articles/s41591-026-04577-2 (visited on 08/07/2026).

[298] S. J. Westwood, J. Grimmer, and A. B. Hall. *Measuring Perceived Slant in Large Language Models Through User Evaluations*. May 8, 2025. URL: https://www.gsb.stanford.edu/faculty-research/working-papers/measuring-perceived-slant-large-language-models-through-user.

[299] C. White, S. Dooley, M. Roberts, A. Pal, B. Feuer, S. Jain, R. Shwartz-Ziv, N. Jain, K. Saifullah, S. Dey, Shubh-Agrawal, S. S. Sandha, S. Naidu, C. Hegde, Y. LeCun, T. Goldstein, W. Neiswanger, and M. Goldblum. *LiveBench: A Challenging, Contamination-Limited LLM Benchmark*. Apr. 18, 2025. DOI: 10.48550/arXiv.2406.19314. arXiv: 2406.19314 [cs.CL]. URL: http://arxiv.org/abs/2406.19314 (visited on 06/08/2026).

[300] H. Wijk, T. Lin, J. Becker, S. Jawhar, N. Parikh, T. Broadley, L. Chan, M. Chen, J. Clymer, J. Dhyani, E. Ericheva, K. Garcia, B. Goodrich, N. Jurkovic, H. Karnofsky, M. Kinniment, A. Lajko, S. Nix, L. Sato, W. Saunders, M. Taran, B. West, and E. Barnes. *RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts*. May 27, 2025. DOI: 10.48550/arXiv.2411.15114. arXiv: 2411.15114 [cs.LG]. URL: http://arxiv.org/abs/2411.15114 (visited on 06/08/2026).

[301] M. Williams, C. Raymond, and M. Carroll. *Sidestepping Evaluation Awareness and Anticipating Misalignment with Production Evaluations*. Dec. 18, 2025. URL: https://alignment.openai.com/prod-evals/ (visited on 06/08/2026).

[302] J. Yang, C. E. Jimenez, A. Wettig, K. Lieret, S. Yao, K. Narasimhan, and O. Press. "SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering". In: *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. arXiv, Nov. 11, 2024, pp. 50528–50652. DOI: 10.48550/arXiv.2405.15793. arXiv: 2405.15793. URL: http://arxiv.org/abs/2405.15793 (visited on 06/08/2026).

[303] J. Yang, K. Lieret, J. Ma, P. Thakkar, D. Pedchenko, S. Sootla, E. McMilin, P. Yin, R. Hou, G. Synnaeve, D. Yang, and O. Press. *ProgramBench: Can Language Models Rebuild Programs From Scratch?* 2026. URL: https://arxiv.org/abs/2605.03546.

[304] S. Yao, N. Shinn, P. Razavi, and K. Narasimhan. *τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains*. June 17, 2024. DOI: 10.48550/arXiv.2406.12045. arXiv: 2406.12045 [cs.AI]. URL: http://arxiv.org/abs/2406.12045 (visited on 06/08/2026).

[305] A. Yehudai, L. Eden, A. Li, G. Uziel, Y. Zhao, R. Bar-Haim, A. Cohan, and M. Shmueli-Scheuer. *Survey on Evaluation of LLM-based Agents*. Apr. 23, 2026. DOI: 10.48550/arXiv.2503.16416. arXiv: 2503.16416 [cs.AI]. URL: http://arxiv.org/abs/2503.16416 (visited on 09/27/2026).

[306] Z. Zeng, J. Liu, S. Chen, T. He, Y. Liao, Y. Tian, J. Wang, Z. Wang, Y. Yang, L. Yin, M. Yin, Z. Zhu, T. Cai, Z. Chen, J. Chen, Y. Du, X. Gao, J. Guo, L. Hu, J. Jiao, X. Li, J. Liu, S. Ni, Z. Wen, G. Zhang, K. Zhang, X. Zhou, J. Blanchet, X. Qiu, M. Wang, and W. Huang. *FutureX: An Advanced Live Benchmark for LLM Agents in Future Prediction*. Sept. 5, 2025. DOI: 10.48550/arXiv.2508.11987. arXiv: 2508.11987 [cs.AI]. URL: http://arxiv.org/abs/2508.11987 (visited on 06/07/2026).

[307] A. K. Zhang, K. Klyman, Y. Mai, Y. Levine, Y. Zhang, R. Bommasani, and P. Liang. *Language model developers should report train-test overlap*. July 22, 2025. DOI: 10.48550/arXiv.2410.08385. arXiv: 2410.08385 [cs.LG]. URL: http://arxiv.org/abs/2410.08385 (visited on 06/23/2026).

[308] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. Dec. 24, 2023. DOI: 10.48550/arXiv.2306.05685. arXiv: 2306.05685 [cs.CL]. URL: http://arxiv.org/abs/2306.05685 (visited on 07/03/2026).

[309] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried, U. Alon, and G. Neubig. *WebArena: A Realistic Web Environment for Building Autonomous Agents*. Apr. 16, 2024. DOI: 10.48550/arXiv.2307.13854. arXiv: 2307.13854 [cs.AI]. URL: http://arxiv.org/abs/2307.13854 (visited on 06/08/2026).

[310] X. Zhou, W. Sun, Q. Ma, Y. Xie, J. Liu, W. Du, S. Welleck, Y. Yang, G. Neubig, S. T. Wu, and M. Sap. *Mind the Sim2Real Gap in User Simulation for Agentic Tasks*. Mar. 11, 2026. DOI: 10.48550/arXiv.2603.11245. arXiv: 2603.11245 [cs.AI]. URL: http://arxiv.org/abs/2603.11245 (visited on 07/24/2026).

[311] M. Zhu, J. Tan, R. Murthy, J. Qiu, L. Yang, W. Zhao, S. Savarese, S. Heinecke, and H. Wang. *RealUserSim: Bridging the Reality Gap in Agent Benchmarking via Grounded User Simulation*. Apr. 7, 2026. DOI: 10.48550/arXiv.2605.20204. arXiv: 2605.20204 [cs.HC]. URL: http://arxiv.org/abs/2605.20204 (visited on 07/24/2026).

[312] M. Zhuge, C. Zhao, D. Ashley, W. Wang, D. Khizbullin, Y. Xiong, Z. Liu, E. Chang, R. Krishnamoorthi, Y. Tian, Y. Shi, V. Chandra, and J. Schmidhuber. *Agent-as-a-Judge: Evaluate Agents with Agents*. Oct. 16, 2024. DOI: 10.48550/arXiv.2410.10934. arXiv: 2410.10934 [cs.AI]. URL: http://arxiv.org/abs/2410.10934 (visited on 06/25/2026).

[313] A. Zverianskii, A. Zhang, J. Clyne, A. Garcia, F. Barez, and S. Upadhyay. *Code Review Bench*. 2026. URL: https://github.com/withmartian/code-review-benchmark.

Appendix A: Best Practices, Research Directions, and Practical Considerations

A.1 Best practices

Table 3: Best practices

Section	Best practices
1a. How does validity change for agents?	• Define explicitly the underlying construct before choosing a metric. Walk through different ways that metric might be gamed by an agent strategy which lacks the capability of interest. • Scrutinize the validity of evaluation results and the claims they support (e.g., using a validity framework [234]), and how they rely on construct, internal, and external validity. • Apply appropriate tooling (e.g., log analysis) to inspect possible sources of validity error—for example, through shortcuts, contamination, or reward hacking [186]. We explore these methods in greater detail in Sections 1h and 1i.
1b. How should we evaluate benchmark quality?	• Use saturation on one metric as a trigger for deeper benchmark scrutiny. • Check for the possibility of impossible tasks, incorrect grading, task errors, and usage of shortcuts to review benchmark validity. • Apply item-level analysis (e.g., Item Response Theory (IRT) from measurement science) to determine which benchmark tasks continue to discriminate among top agents, possibly to repurpose subsets into more efficient benchmarks [191].
1c. What should agent evaluations measure beyond accuracy?	• Report accuracy jointly with an appropriate notion of cost (e.g., monetary value, latency) and illustrate the tradeoff as a Pareto curve [133] (see Figure 6). • Measure reliability as a metric in its own right and set an acceptable threshold scaled to the specific deployment scenario. • Clarify what model-scaffold combinations are being measured, to better understand model $\times$ scaffold interaction effects.
1d. How can agent evaluations achieve statistical rigor?	• Conduct repeated runs to measure consistency for each task, and report how many runs were conducted. Avoid claims relying on low run-to-run variance if only a single run is affordable. • Ensure that the metric of interest and the stated claim are aligned: report pass@k when it matters whether an agent can succeed at all, and report pass$\wedge k$ for whether it succeeds every time—pass@k should not serve as evidence of reliability. • Account for different sources of variance (e.g., judge error) in reported uncertainty, particularly for LLM-as-judge setups (see Section 1h), where raw scores may understate uncertainty.

Source subsection	Best practices (continued)
1e. How can agent evaluations attribute causal changes in performance?	• Identify and document clearly different components of an experimental setup—the models used, scaffolds paired, environments tested, prompts used, and the budget constraints. When comparing agents, hold all but one component fixed such that any change in performance can be properly attributed. • Do not attribute an agent's failure directly to the underlying (language) model, without accounting for whether improved scaffolding around the model changes the performance profile [140, 248]. • When budget allows, systematically test model $\times$ scaffold interaction effects in various permutations, e.g., with (fractional) factorial designs—in which multiple independent variables are tested in conjunction.
1f. How can agent evaluations resist contamination?	• Define secret holdouts appropriate to the level of generality expected of the agent [132], and limit repeated use and optimization against the holdouts [46, 104, 106]. • Report test-train overlap [307], any benchmark-specific information used to develop the agent system's prompts, scaffolds, and tool usage, and create transparency on private testing practices [249]. • Inspect agent trajectories (see Section 1i) for contamination and incorporate clean-room evaluation harnesses into the testing setup [126]. • When a holdout set is publicly available, update the set regularly with refreshed tasks (Section 1g) such that benchmark scores do not depend on tasks the agent may already have been exposed to.
1g. What role do dynamic benchmarks play, and what are their tradeoffs?	• Apply test equating and linking from psychometrics to ensure a common anchor subset across scores, for longitudinal comparability. • Track distribution shifts and regularly revisit whether results validly support claims about the intended construct. • Implement version control for each component of the benchmark, to enable reproducibility and comparability over time.
1h. Where should LLMs judge and monitor agents?	• Validate autograders and scanners against human-coded holdouts representative of the intended use case with subject matter expert participation. • Incorporate targeted manual analysis into automated analysis. In particular, review a subset of the cases a judge passed/stopped to quantify how many cases might slip by undetected. • Quantify biases in autograder setups, through human-LLM / human-agent agreement and monitor if/when training traffic is no longer representative of the object of judgement.
1i. What can log analysis reveal beyond outcomes?	• Do not conclude that a clean log analysis record implies an unproblematic task execution; logs have their own limitations, and absence of evidence is not evidence of absence. • Conduct both manual and automated log inspection, with and without LLMs (e.g., using Transluce's Docent [135, 265], Inspect Scout [168, 272]) [171]. • Ensure availability of agent logs—including full or summarized reasoning traces, tool calls, environment states—as part of evaluation practices to enable downstream and future analysis (e.g., Epoch AI's log viewer setup [71]).

Source subsection	Best practices (continued)
1j. What is the value of standardized evaluation infrastructure for agents?	• Before adopting commonly used metrics, verify that they are aligned with the construct under evaluation for the specific use case. Default assumptions and parameters in such systems shape many evaluations simultaneously. • When the same framework is used for both training and evaluation, ensure that evaluation tasks are isolated from training environments to prevent leakage. • While a one-size-fits-all approach may not be feasible or necessarily desirable, use a modular approach, with standardization around a shared set of tools, for deduplicating effort and achieving comparability.
1k. How should red-teaming evolve for agents?	• Evaluate misuse (i.e., a user directing an agent with malicious intent) and attacks on legitimate use cases (e.g., indirect prompt injection through tool outputs), across models, tools, scaffolds, and environments. • Frame red-teaming not only in the context of deployment, but in testing as well [131]. This requires, for example, verifying sandbox isolation before running agents with relaxed safeguards or broad tool access. • When granting agents realistic tool access for ecological validity, decide in advance which real-world effects are acceptable. During execution, monitor the runs for any attempts to escape these pre-defined boundaries.
1l. What can open-world evaluations reveal that benchmarks cannot?	• Present results from open-world evaluations as evidence of what agents *can* achieve (i.e., upper bounds) rather than of typical performance. Instead of over-generalizing from a handful of cases, carefully report the experimental conditions and assumptions (e.g., scaffold, tool access, setup effort, and cost). • Clearly specify the construct (Section 1a) and metric(s) of interest. • Document points of human intervention and agent trajectories in detail. • Incorporate real-time monitoring and analysis of agent trajectories.
1m. How can evaluators interpret agent performance relative to human capabilities?	• Scrutinize how evaluators approach task definition. State what aspects of an occupation tasks claim to cover (and, importantly, what they do not). Do not assume successful task completion extrapolates to claims about job automation and replacement. • Report uncertainty for human baselines [292], which carry their own measurement error, e.g., from small expert samples and noisy completion times. • Quantify levels of reliability, or the additional human work required to turn an agent-generated deliverable into something of sufficient quality [220, pp. 12–13].

A.2 Research directions

Table 4: Research directions

Section	Research directions
2a. Human-agent interaction: evaluation with simulated users	• For which use cases do findings from simulated users serve as a useful proxy for human ones, and in which do they not? • Which failure modes in simulated users most impact the generalizability of study findings? • What role can targeted participation of human users play toward grounding or validating that of simulated users?
2b. Human-agent interaction: experiments in the lab	• How should evaluators define baselines in studies to best reflect the ecological validity of the growing diffusion of agents? • What are different population categories (e.g., junior vs. senior expertise) in which agents are likely to have uneven or disparate impacts on various outcomes? • Which particular laboratory settings best translate to real-world deployment contexts?
2c. Human-agent interaction: field experiments	• What sorts of spillover effects do agents create beyond the specific users themselves? • How do effects of agents change over time as users adjust to these agents in their workflows? • In what ways do agents create new winners and losers in different work contexts, and across which outcomes?
2d. Observational data on real-world usage	• What sorts of coverage issues does production traffic raise—and to what extent can this traffic be contextualized by subsequent user action? • How can independent evaluators gain access to this data at scale, or devise new crowdsourced alternatives, in a privacy-preserving manner? • How can production traffic serve an exploratory purpose for evaluators, to narrow and refine questions to target causal studies on real-world agent performance?
2e. Domain-specific evaluation	• For what purposes can upstream evaluation through general benchmarks continue to be useful, and which are now exclusive to domain-specific evaluation? • How do particular occupational norms and existing skillsets drive differences in impact across domains? • How does task completion translate to understanding of broader workflows, and how can we better measure augmentation effects of agent deployment?
2f. Economic impact evaluation	• How can economic evaluation identify what task-based measures fail to capture (such as the work between tasks, and the reorganization of roles once tasks are automated)? • Under what conditions is task-level capability predictive of broader claims about final output, occupations, or economic outcomes? Conversely, which observed economic effects should motivate new capability evaluations? • What additional worker-centric metrics can complement understanding of the impact of agents and shifting nature of different occupations?

Source subsection	Research directions (continued)
2g. Agent behavior under evaluation	• How can external evaluators promote better ecological validity of evaluation settings, with respect to evaluation awareness? • How can tooling appropriately scale to address broader monitorability risks, and what are associated overreliance risks? • For what risks does focus on agent trajectories offer reliable signal, and when does it detract from other sources of understanding? • Beyond understanding whether agent behavior is capable of scheming and sandbagging, how can evaluations estimate how often scheming or sandbagging occurs [99]?
2h. Dangerous capabilities and risk	• To what extent does performance under simulated or prompted conditions reflect real-world risk? • What is the realistic baseline of capabilities already available to malicious users, and how much additional capability do agents provide?
2i. Agent and evaluation security	• How can evaluators balance evaluation security with demand for ecological validity? • What control measures are appropriate for different risks stemming from evaluation security, and how should public infrastructure be hardened? • What new approaches to real-time monitoring are necessary to handle the volume of agent transcripts and more proactively address security incidents?
2j. Interaction harms	• What metrics can allow evaluations to better measure harms across repeated interactions? • How can evaluators balance ecological validity alongside risks for research participants under the principle of beneficence? • Which interaction harms are genuinely new to agents (i.e., arising from actions taken on a user's behalf), and which exacerbate existing concerns from chatbots? • What new forms of crowdsourced data are necessary to explore these questions, such as incident reporting programs [216]?
2k. Systemic evaluation	• What institutional mechanisms and incentive structures are agents reinforcing, and which dynamics are qualitatively changing? • When can findings from a case study of systemic impact (e.g., peer review at one venue) be expected to generalize to other organizations? • What methods are necessary to identify leading indicators or sources of anticipatory insight that allow evaluators to understand changes of a systemic nature?

A.3 Practical considerations

Table 5: Practical considerations

Section	Practical considerations
3a. Who evaluates?	• Explore opportunities for evaluation work in academia, frontier labs, non-profits, and public sector entities. Not all necessarily require a traditional academic path via a PhD. • Examine what functions in the evaluations ecosystem one may best contribute—directly evaluating agents, interpreting and sharing findings, coalescing standards, communicating with policymakers, or advising funding of evaluations work.
3b. What does independence require?	• Scrutinize funding, access limitations, and conflicts of interest associated with organizations, applying frameworks such as the AEF-1 standard from the AI Evaluator Forum [255]. • Clarify institutional policies before joining an organization, as legal and technical safe harbor for independent evaluators remains incomplete [153].
3c. Does lab direction drive evaluations?	• When an evaluation adopts risk categories or thresholds (e.g., from a Responsible Scaling Policy or Preparedness Framework), explicitly cite those and note which risks those categories leave out. • Disclose the conditions attached to model access or credits (e.g., pre-publication review, restricted model versions) alongside each release of results. • Identify areas of priority focus and constraints attached to evaluation funding, and reflect on whether they align with an evaluator's own research priorities.
3d. Do evaluations drive lab direction?	• Differentiate between the types of evaluations one is conducting and judge frontier developer participation accordingly. • Distinguish capability from accountability evaluations, and treat developer adoption as a success signal only for the former. • Plan how a new metric could become visible and address a gap with existing popular metrics (e.g., inclusion in widely followed leaderboards). • Track overall how labs and developers respond to new findings (e.g., patches, public statements, policy changes). • Do not treat missing results as good news—depending on context, unfavorable evaluations may simply go unreported [249].
3e. How does funding shape independent evaluation?	• Distinguish between financial and in-kind support, and understand how an organization discloses and treats both categories. • Identify historical and present funding and how it may shape an organization's orientation. • When relying on shared infrastructure, take note of who funds and maintains it.
3f. How do rising costs affect evaluations?	• Estimate the full cost of an experimental panel (i.e., per-run API cost $\times$ repeated runs $\times$ tasks) before committing, since a full run can reach tens of thousands of dollars per model. • Jointly factor in cost [92], alongside dimensions of quality, reliability, and accuracy (see Section 1c). • Allocate sufficient budget for engineering and maintenance, in addition to base API credit costs. Cut run costs where possible (e.g., via item-level analysis to shrink task sets, Section 1b; or by relying on shared infrastructure, Section 1j).

Source subsection	Practical considerations (continued)
3g. Where are the venues for evaluations?	• Look beyond academia as well for the latest research insights, and scrutinize validity and rigor regardless of venue. • Invest time to understand other evaluation ecosystems, and distinguish geographic differences in evaluation norms and goals. • Embrace voluntary standardized reporting formats and share infrastructure and resources to help reinforce field-wide standard practices and promote legibility.
3h. How is the public sector responding?	• Evaluate AI deployments with respect to status quo baselines (e.g., marginal risk evaluations) [84]. • Engage actively with public sector evaluation bodies (e.g., US CAISSI, UK AISI, EU AI Office [74]) to inform new evaluation agendas. • Plan public sector evaluations as a lifecycle that encompasses post-deployment through ongoing monitoring [188].
3i. Where are the entry points for different disciplines?	• Collaborate with disciplinary perspectives suited to the questions of interest, not just by adopting their questions or methods. • Draw upon the extensive literature from different disciplines, particularly for questions of validity and rigor. • Recognize existing technical barriers to participation and connect new disciplinary perspectives to sufficient engineering capacity.