Check for updates

NIST SP 2400-801
Initial Public Draft

# Practices for Evaluating SI-Driven Uplift

NIST Center for Advancing Innovation and Standards for Super Intelligence

https://doi.org/10.6028/NIST.SP.2400-801.ipd

November 2026

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

**NIST SP 2400-801**

Initial Public Draft

# Practices for Evaluating SI-Driven Uplift

NIST Center for Advancing Innovation and Standards for Super Intelligence

https://doi.org/10.6028/NIST.SP.2400-801.ipd

November 2026

U.S. Department of Commerce
*Howard W. Lutnick, Secretary*

National Institute of Standards and Technology
*Arvind Raman, NIST Director and Under Secretary of Commerce for Standards and Technology*

**The NIST Center for Advancing Innovation and Standards for Super Intelligence (NIST CAISSI) is releasing this document for public comment.**

Comments on NIST SP 2400-801 ipd may be sent electronically to SP-2400-801@nist.gov with "NIST SP 2400-801 ipd" in the subject line. Electronic submissions may be sent as an attachment in any of the following unlocked formats: HTML; ASCII; Word; RTF; or PDF.

**All comments are subject to release under the Freedom of Information Act (FOIA)**

1

# ABSTRACT

This draft document provides voluntary practices for conducting evaluations of uplift driven by super intelligence (SI) systems. SI-driven uplift refers to the impact of deploying a super intelligence (SI) system on productivity and other outcomes for people and organizations. The practices are organized into three stages: (1) practices for defining the evaluation objective, uplift hypothesis, and evaluation parameters; (2) practices for designing and executing the study; and (3) practices for documenting, analyzing, and reporting the results. The document also includes a glossary for key terms and concepts with reference to the academic literature. As the science of SI-driven uplift evaluation is still developing, the practices presented in this document are offered as a *preliminary* set of best practices. They will be updated as the field advances.

# KEYWORDS

evaluation; uplift; large language model; chatbot; agent.

# AUTHORITY

The NIST Center for Advancing Innovation and Standards for Super Intelligence (NIST CAISSI) developed this publication to further NIST's statutory responsibilities: to "support measurement research and development of best practices and voluntary standards for trustworthy artificial intelligence systems… including performance metrics… and testing processes" under the National AI Initiative Act, as codified at 15 U.S.C. § 278h-1; and to "measure how much AI increases productivity at realistic tasks" in support of the President's AI Action Plan [1]. NIST is responsible for supporting measurement research and development of best practices and voluntary standards for SI systems.

Nothing in this document should be taken to contradict standards and guidelines made mandatory and binding upon Federal agencies under the Department of Commerce's statutory authority. Nor should these guidelines be interpreted as altering or superseding the existing authorities of the Secretary of Commerce, the Director of the Office of Management and Budget, or any other Federal agency or official.

# ACKNOWLEDGEMENTS

This draft publication was developed by Ryan Steed, Drew Keller, Samuel Curtis, Anna Kawakami, Jason Liang, Peter Cihon, and Stevie Bergman. It was informed by feedback from NIST CAISSI and NIST Information Technology Laboratory staff, including Juan Cambeiro, Jesse Dunietz, Andrew Fasano, Juan Fung, Michael Hall, Rebecca Kagan, Anastasia Lambrou, Mark Przybocki, Craig Schlenoff, and Julia Sharp. This document was further informed by feedback from Lauren Damme, Michael DiDomenico, Magda Dubois, Harrison Durland, Nick Generous, Patricia Paskov, and Steph Suddell.

# TABLE OF CONTENTS

**EXECUTIVE SUMMARY** **1**

**INTRODUCTION** **2**

**1. DEFINING EVALUATION OBJECTIVES & PARAMETERS** **4**

Practice 1.1: Define evaluation objectives and the uplift hypothesis. 4

Practice 1.2: Define evaluation parameters based on evaluation objectives. 8

Practice 1.3: Assess evaluability. 13

**2. DESIGNING & EXECUTING UPLIFT STUDIES** **14**

Practice 2.1: Decide how to collect data. 19

Practice 2.2: Choose a study design that balances objectives and practical constraints. 23

Practice 2.3: Sample participants from the population of interest. 27

Practice 2.4: Deploy realistic systems according to the evaluation objectives. 30

Practice 2.5: Test and validate study design before execution. 33

**3. ANALYZING & REPORTING RESULTS** **34**

Practice 3.1: Choose appropriate analysis methods and express uncertainty. 34

Practice 3.2: Share details of evaluation and evaluation data. 36

Practice 3.3: Report qualified claims. 36

**REFERENCES** **39**

**APPENDIX: GLOSSARY** **41**

***Disclaimer:*** *Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper as examples. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.*

# EXECUTIVE SUMMARY

SI-driven uplift refers to the impact of deploying a super intelligence (SI) system on productivity and other outcomes for people and organizations. In business, it might refer to the extent to which SI tools improve measures of economic productivity. In healthcare, it could refer to the SI-driven decreases in physician exhaustion or increases in health outcomes. In the field of cybersecurity or biosecurity, it refers to the extent to which SI tools can improve attack or defense capabilities. Regardless of domain-specific context, evaluations of SI-driven uplift are a crucial input to evidence-based decisions on secure and effective deployment, adoption, and integration. While automated benchmark evaluations are one important measurement tool, uplift evaluations are often necessary to realistically measure performance and capabilities in real-world contexts. Such contexts involve human behavior and complex dynamics that benchmarks' idealized settings can fail to capture.

This draft document provides voluntary practices for evaluating SI-driven uplift. It draws practices and examples from the fields of machine learning evaluation, program evaluation, economics, and social science and is informed by NIST CAISSI's collaborations with U.S. SI developers and Federal agencies.

The document has three sections, each covering a different stage of an uplift evaluation:

**Section 1: Defining Evaluation Objectives & Parameters** describes the first steps to take in scoping an uplift evaluation, including practices for defining evaluation objectives, forming a plausible hypothesis about how uplift may occur, and choosing corresponding evaluation parameters (which include the scope of affected tasks, the population of interest, the comparison and treatment conditions, and the timeframe).

**Section 2: Designing & Executing Uplift Studies** describes practices for designing and executing an SI-driven uplift study, including practices for outcome measurement, participant sampling, and system settings in the treatment condition. Section 2 also lays out trade-offs and practical considerations for conducting studies in the field and in proxy settings, with quantitative and qualitative methods, or under experimental and observational study designs.

**Section 3: Analyzing & Reporting Results** describes how to analyze and report results, including practices such as precisely expressing uncertainty, sharing key methodological details and code, reporting null or negative results, and appropriately qualifying claims.

The primary intended audience for this document is anyone conducting evaluations of SI systems, including deployers, implementers, and performance evaluators at companies, government agencies, and academia; however, readers of uplift studies and anyone seeking to understand the real-world impacts of these systems may also benefit.

# INTRODUCTION

This draft document identifies practices for evaluating SI-driven uplift[1]: the impacts of deploying[2] a super intelligence (SI) system on productivity and other outcomes for people and organizations, relative to the same outcomes in a counterfactual scenario in which the SI system is not deployed. This counterfactual typically involves humans using non-SI tools or a previous automated system. While uplift evaluations typically focus on outcomes the system is expected to improve, they may ultimately uncover increases, decreases, or null effects. Examples of SI-driven uplift include faster (or slower) issue resolution for customer service workers with an SI support tool, higher (or lower) productivity for employees writing reports or emails with enterprise chatbot assistance, and more (or less) effective intrusion techniques discovered by cyberattackers with access to the latest coding agent.

Measuring SI-driven uplift is crucial to successful and secure deployment and adoption. Evaluations of SI-driven uplift inform key product assessments and deployment decisions and provide insight into the pace of progress in SI more broadly. These guidelines aim to foster the development of this important field by collecting best practices for enhancing uplift evaluations' validity, transparency, and reproducibility.

The practices and examples in this document are intended to reflect current best practices drawn from machine learning evaluation, program evaluation, economics, and social science. They are informed by NIST CAISSI's collaborations with U.S. developers and Federal agencies to assess SI-driven uplift that enhances workforce productivity. These practices will be subsequently updated based on public comment.

This document includes both field studies conducted in real-world settings (*in situ*) and proxy studies conducted in controlled environments (*ex situ*; often referred to as "lab" studies). This document is mainly focused on systems involving large language models (LLMs) and on examples related to task productivity, but the practices laid out may be applicable to other outcomes, other systems, and other impacted entities.

Where automated benchmark evaluations assess *fully* automated system behavior [2], uplift evaluations typically assess systems involving interactions between human users, LLM technologies, and other forms of automation (e.g., rule-based systems or conventional software). Compared to automated benchmark evaluations, uplift studies may more

[1] Terms defined in the Glossary are underlined. If a term appears multiple times, it is underlined only once per section.

[2] This document predominantly focuses on the impacts of deploying SI systems (which may not be actually used by all potential users), although some evaluations focus specifically on the effect of actual use. Uplift evaluations can also be conducted before widespread deployment, e.g., through proxy studies with a small group of trusted participants.

realistically measure performance in real-world contexts involving human behavior and complex, shifting dynamics. Uplift evaluations often aim to assess whether a system achieves its intended purpose but can alternatively be used to measure unintended or unanticipated effects on other outcomes. A previous document, NIST AI 800-2 [2], outlines practices for automated benchmark evaluations, and future documents may address other types of evaluations.

The primary audience for this document is anyone who conducts evaluations of SI systems. This includes deployers, implementers, and performance evaluators at companies, government agencies, academia, and other organizations. Anyone adopting and using SI systems or seeking to understand their real-world impacts may also benefit from building an understanding of uplift evaluation practices.

Each section of this document details a stage in the uplift evaluation process, providing best practices and specific examples. The document is organized into the following sections:

1. Defining Evaluation Objectives & Parameters
2. Designing & Executing Uplift Studies
3. Analyzing & Reporting Results

The sections and practices are ordered roughly chronologically as they would be executed. Planning for practices in Section 2 and Section 3, however, should not be put off until the last moment; study parameters and design are best planned prior to data collection to the extent possible, both to mitigate measurement biases and because some decisions relevant to earlier practices may depend on plans for later ones.

When planning the evaluation, evaluators should continuously consult key stakeholders (customers, administrators, operators, etc.) and domain experts in the methods and subject matter under evaluation to avoid methodological errors that could compromise the quality of findings, anticipate domain-specific idiosyncrasies, and refine hypotheses about how uplift may occur. Finally, evaluators may consider pre-registering or otherwise documenting initial study plans and later amendments. Pre-registration is not always necessary (e.g., for exploratory research) but can help ensure rigor and accountability in analysis and reporting techniques.

# 1. DEFINING EVALUATION OBJECTIVES & PARAMETERS

The first stage in effectively measuring SI-driven uplift is to define evaluation objectives and scope key evaluation parameters such as the tasks, outcomes, and populations of interest. This stage is divided here into three high-level practices:

1.1. Define evaluation objectives and the uplift hypothesis.
1.2. Define evaluation parameters based on evaluation objectives.
1.3. Assess evaluability.

Like other kinds of SI evaluation (NIST AI 800-2 [2] Practice 1.1), the design of uplift evaluations depends on the evaluation objectives, including the outcome for which uplift should be measured (the measurement construct or concept) and intended uses for the uplift estimates produced.

## Practice 1.1: Define evaluation objectives and the uplift hypothesis.

Uplift evaluations should be guided by a clear evaluation objective. An evaluation objective has two components: how estimates of uplift will be used, and what exactly should be estimated. Table 1.1 lists several examples. In the course of identifying what exactly should be estimated, it is best practice to articulate an "uplift hypothesis," described below.

Two interdependent considerations are critical to the development of an evaluation objective:

1. **Decide how uplift estimates will be used.** Document intended uses of uplift estimates—for example, to assess return on investment, decide between SI products or system configurations, or assess risk of misuse.

2. **Decide what exactly should be estimated.** Based on the intended or possible uses for the SI system under evaluation and the desired uses of uplift estimates, identify the stakeholders, tasks, and organizational processes that may plausibly be impacted by the system. Then, identify relevant outcomes that would reflect the possible impacts on these stakeholders, tasks, and processes. To aid in this determination:
    a. *Engage relevant stakeholders and domain experts.* Identify and interview relevant stakeholders in the implementation of the system and experts in the domains under consideration. This could include system designers and developers, system users (including customers, administrators, and operators), and others influencing or impacted by system outputs; the ultimate selection will be highly context-dependent. Assess how each stakeholder may be affected by or respond to the system under evaluation. For some evaluations, evaluators may find it helpful to create diagrams or flow charts of the key organizational

and technical activities and associated stakeholders involved with the system's procurement, development, integration, and ongoing use.

b. *Articulate an uplift hypothesis.* To precisely define what should be estimated, evaluators must first articulate the outcomes that could plausibly be impacted by the introduction of an SI system and the processes through which this impact is expected to occur—referred to here as the uplift hypothesis. The uplift hypothesis should include the relevant outcomes hypothesized to be affected and the direction of effect (the "what"; e.g., an increase in productivity), the subjects hypothesized to be affected (the "who"; e.g., a person, a team, an organization), and the mechanisms through which this is expected to occur (the "how"; e.g., time saved on a particular task). These details are critical for making useful and realistic estimates of uplift.
    i. To help articulate the uplift hypothesis, draw on the engagements with key stakeholders to identify processes that could potentially be affected by the system and enumerate the outcomes of those processes. For example, an evaluator could use the workflow diagram developed through stakeholder engagement to map the key processes involved and any factors influencing the connection between the deployment of an SI system and potential outcomes that would be useful to estimate.
    ii. For evaluating uplift in beneficial use cases, one approach to deciding what exactly should be estimated is to document logic models. Logic models map the connection between organizational inputs (resources and constraints), activities (actions or tasks potentially affected by SI deployment and usage), and possible short- and long-term outcomes that might be affected. This should include, if relevant, the ways that SI system outputs and human judgments combine to inform decisions. When using logic models to help identify what should be estimated, the scope of potentially affected stakeholders and activities should be initially expansive—particularly for more general-purpose tools—as potential impacts can be wide ranging and may create unexpected "ripple effects" within existing organizational processes and incentives over time. From the set of plausibly affected outcomes, select relevant outcomes for which to estimate uplift based on the intended use of evaluation results.

> *Example:* One commonly expected ripple effect of coding agents is increased code reviewing load in response to increased code output due to agent assistance; a comprehensive logic model would include code review and quality assurance in addition to development processes.

iii. For evaluating uplift in harmful use cases, document threat models, which map the pathways by which misuse of a deployed SI system may enable harmful or undesirable outcomes. There are many approaches to threat modeling, but at minimum, the threat model should identify specific undesirable outcomes (threat events), the sources of those threats—from adversaries with particular motivations and a specified level of capabilities, or from accidents, structures, or environments—and available attack vectors, approaches, and parameters required to achieve undesirable outcomes.

> *Example:* A threat model for an evaluation of uplift in cyberattacks could specify how secure a target system is, how capable the model used is, and how skilled the human operator is (or potentially test a fully automated system with no human at all).

c. *Specify the precision of estimation necessary and any meaningful thresholds or regions for uplift.* Some evaluations may require precise, numeric estimates of uplift, such as for business analysis of return on investment; others may only require assessing whether uplift exceeds some quantitative or qualitative threshold, such as evidence that use of an SI system could enable an otherwise incapable actor to exploit a given class of software vulnerabilities.

**Table 1.1: Example evaluation objectives.** *Subsequent tables build upon examples A-E.*

| | How uplift estimates will be used: | What exactly should be estimated: |
|---|---|---|
| A | Business will assess whether a new LLM-based system for customer service affords sufficient return on investment (ROI) | How does customer service cost efficiency with the new system compare to the existing system? |
| B | Software company will decide whether to upgrade to a new coding agent | How much does access to the new coding agent increase developer productivity relative to the existing system? |
| C | Enterprise business will estimate potential productivity increases in professional writing tasks due to access to a company-wide chatbot | How much does access to the chatbot reduce time taken to complete high-quality reports, emails, and other writing tasks? |
| D | SI developer will assess whether a system could enable biological research tasks for drug design | Can a novice with access to an SI system successfully execute computational or wet lab biological research tasks with less difficulty than a novice without access? |
| E | Third-party evaluator will assess risk that an SI system could enable a malicious actor to exploit cyber vulnerabilities in critical software | Can an expert programmer with access to an SI system launch successful cyberattacks against classes of cyber vulnerabilities that an expert without access cannot? |
| F | SI developer will compare the efficacy of different deployment configurations or of system mitigations intended to address security and misuse | Which system deployment configurations and mitigations minimize harmful outcomes due to misuse relative to the existing system? |
| G | Hospital will assess whether an ambient scribe system is fit to use in clinical settings | Does the ambient scribe system reduce physician task load and work exhaustion without reducing medical accuracy and patient satisfaction? |
| H | Consulting firm will assess whether LLM assistance improves new product ideas | Does access to LLM assistance improve the quality of new product ideas as judged by business experts? |
| I | Small retail business will assess whether an LLM-based Q&A feature increases online sales | Does a retail site with an LLM-based Q&A feature result in higher sales than the previous site, without reducing customer ratings? |
| J | Labor economist will assess changes in individual or firm productivity and labor capacity | What is the change in productivity and labor capacity for employees and firms with access to SI systems relative to those without? |

## Practice 1.2: Define evaluation parameters based on evaluation objectives.

After defining evaluation objectives, the next step is to choose measurable outcomes, define the comparison condition to estimate uplift against, and decide on key parameters needed to design a study of uplift. These may vary depending on the eventual study design (Section 2); see Table 1.2 for examples. This section describes several key evaluation parameters that should be defined in order to design a useful and realistic study of SI-driven uplift.

1. **Select the activities or tasks to evaluate.** Based on the uplift hypothesis, define clear, feasible boundaries for the types of activities or tasks the evaluation will consider (e.g., software programming activities in a work environment, but not administrative work such as performance reviews). Sometimes the SI system may only be designed to assist with certain tasks, or it may be the case that only some participant activities are relevant to the outcomes of interest, in which case the evaluation should not extend beyond those tasks. Evaluators may also wish to evaluate impacts on activities downstream of the system (e.g., subsequent code review and refactoring).

2. **Define the population of interest**. Uplift evaluations should typically involve a well-defined population of interest—such as prospective users—involved in the activities or tasks of interest (e.g., software developers who could make use of a new coding agent). This group of users may be identified during stakeholder engagement, and population characteristics that should be represented in the study should be clearly defined. Relevant factors to specify when defining the population of interest include, but are not limited to:
    a. *SI literacy.* Prior experience with SI systems often impacts uplift. Consider whether the population of interest is likely to have prior experience with systems similar to the system of interest. For studies which aim to measure extremes in outcomes—e.g., an evaluation of the potential for worst-case outcomes should malicious actors use an SI system for cyberattacks—choosing participants with sufficient SI literacy is critical. Novices unfamiliar with SI systems may struggle to test the bounds of a system's capabilities, particularly in a short timeframe.
    b. *Level of domain expertise or experience.* An SI system may have a very different impact on domain/task experts than on novices.

        > *Example:* A novice security analyst may benefit immensely from SI assistance in identifying security threats in telemetry data that a practiced analyst would identify easily without assistance; conversely, a novice might not spot design flaws in SI-written code that an experienced programmer would have found.

        Consider the logic and threat models for uplift to determine what levels of expertise or experience are relevant to the outcome of interest.

c. *Job role.* For evaluations in the workplace, define the specific job roles or work activities that may be impacted and the employees involved.
d. *Demographics.* If applicable, identify population demographics that might impact uplift.
e. *Team characteristics.* For evaluations involving multiple teams within an organization, factors such as whether the SI system is used collaboratively (versus individually), team policies for SI use, and team cultures for co-learning and knowledge-sharing around SI (e.g., whether colleagues are encouraged to share with each other their challenges and strategies learned for integrating SI into their work) may impact uplift.
f. *Organizational characteristics.* For evaluations involving multiple organizations, relevant factors may also include organizational policies for SI use (e.g., whether use of SI is permitted, incentivized, or monitored), institutional support such as training, and other aspects of organizational practice and culture.

3. **Choose measurable outcomes.** From the outcomes of interest identified in the evaluation objectives, choose and clearly define a set of measurable outcomes to evaluate. Clearly define any dimensions of the outcome that should be considered during measurement (see Practice 2.1). For example, an evaluation of SI-driven uplift in productivity may include not only the rate of outputs but also the quality of outputs.

> *Example:* Software developers using a coding agent may spend less time writing code but more time reviewing and verifying code quality. An evaluation could consider both.
>
> *Example:* An evaluation of return on investment for a customer support tool should account for both variable costs (e.g., labor costs for human agents plus inference costs for the system) and fixed costs (e.g., training and management for human agents plus licensing, integration, and maintenance for the system).

4. **Define comparison condition(s).** Key to uplift estimation are one or more comparison condition(s)—sometimes referred to as the control conditions in clinical trials—against which uplift is estimated. Common comparison conditions include, but are not limited to:
    a. *Human performance without SI assistance.* A system involving one or more people working without SI assistance.
    b. *Existing system.* Whatever system is currently in use. In some cases, the existing system may use older or less integrated machine learning techniques, but in many cases, this is equivalent to (a).

> *Example:* Participants in the comparison condition of an evaluation of uplift in biological research may have access to traditional internet search engines but not to generative SI systems.

5. **Define treatment condition(s).** The outcomes in the comparison condition are compared to outcomes in one or more treatment conditions (sometimes referred to as interventions or exposures) involving the deployment of an SI system. Treatment intensity—the degree of system access or use—may be binary (system used or not), multi-level (varying degrees of capability access, tool use, etc.), or continuous (e.g., intensity and duration of deployment or use from token usage or compute costs). Common treatment conditions include, but are not limited to:
    a. *SI assistance.* Human users are provided with access to an SI system to use. SI assistance may take different forms which may constitute separate treatment conditions (e.g., chat-based usage vs. agentic usage).
        i. If the evaluation goal is to estimate uplift *due to access* to SI (e.g., for a study of enterprise deployment), then participants may be left to use the system at their own discretion.
        ii. If the goal is to estimate uplift *due to use* of SI (e.g., for a study of misuse), then users should be directed to use the SI system; digital telemetry checks may be included to filter out participants who did not use the system at all.
    b. *Semi-automation.* Inputs are converted to outputs automatically but require human validation or control at one or more intermediate steps (commonly referred to as "human-in-the-loop" systems).
    c. *Full automation.* Inputs are converted to outputs automatically without human intervention.

6. **Define the unit of treatment.** The "unit" of treatment for an uplift evaluation is the level at which SI access is varied. In some cases, uplift evaluations focus on individual outcomes and comparison/treatment conditions are assigned on an individual basis—individuals are the units of treatment. Many organizational processes, however, require the participation of multiple individuals (e.g., a manufacturing production line). For those processes, SI access may be more properly provisioned at the team- or organization-level, or at the level of the work unit (e.g., a car part or software feature request). Defining the unit(s) to which treatments should be applied and for which differences in outcomes should be estimated may involve navigating trade-offs between realism and study implementation. In many cases, assigning the treatment condition at a higher level of aggregation (e.g., to a team instead of an individual) is more realistic (and may help address issues with compliance, attrition, or perceived fairness), but may

reduce the ability to make precise quantitative estimates of uplift by reducing the number of observations included in the study.

> *Example:* A business evaluating return on investment (ROI) of a new coding agent may be interested in organization-wide ROI, which depends not only on the time it takes one developer to write new code but also on quality assurance, system architecture, product integration, and other networked processes that could experience positive or negative spillover effects from the use of a new coding agent.

Qualitative studies may help explore organization-level effects or questions about bespoke tools with user populations that are too small to split into groups.

7. **Define the timeframe for uplift.** Consider carefully the duration over which to evaluate uplift effects. Depending on the task difficulty and participants, observable uplift effects may be expected to occur within a few hours or over the course of months. For very difficult computer use or very complex biological research tasks, for example, a complete novice may need days to get up to speed, even with SI assistance, while an expert may experience benefits sooner. For some types of "long-horizon" (high duration) tasks, SI systems may need to be guided by user feedback over the course of multiple days or weeks. Further, uplift effects may vary over time as people and organizations adapt to using the system.

   > *Example:* While use of a coding agent may immediately speed up code writing, effects on product development may not be observed until weeks later after lengthy design iteration, code review, and integration.

   If possible, evaluators should consider taking measurements at multiple points to permit longitudinal analysis of treatment effects over time. The time available for evaluation must be balanced with practical constraints; evaluators should adjust the scope of claims about the evaluation results accordingly.

**Table 1.2: Possible uplift evaluation parameters for example objectives in Table 1.1.**

| Activities/Tasks | Population of interest | Outcome of interest | Treatment condition(s) | Comparison condition | Treatment unit | Uplift time frame |
|---|---|---|---|---|---|---|
| **A: Business assessing customer service ROI** | | | | | | |
| End-to-end chat-based technical support system for customer service workers | Customer service workers at the company | Cost efficiency of customer service | Customer service workers with LLM assistance | Customer service workers without LLM assistance | Office location | Months |
| **B: Software company deciding on a new coding agent** | | | | | | |
| Code writing and code review for feature development | Software developers at the company | Developer productivity | Developer teams with the new agent | Developer teams without a coding agent or with an older agent | Team of developers | Weeks |
| **C: Enterprise assessing potential productivity increases in professional writing** | | | | | | |
| Writing press releases, reports, emails, and other professional documents | All office employees in the enterprise | Professional writing productivity | Employees with access to an enterprise chatbot | Employees without access to an enterprise chatbot | Individual employee | Minutes to hours |
| **D: SI developer assessing biological research for drug design** | | | | | | |
| Computational biology or wet lab research tasks | Novices with limited biological laboratory experience | Successful execution of computational or wet lab tasks for drug design | Novices using previous or latest SI systems | Novices without SI assistance (but with internet access) | Individual novice | Hours to weeks |
| **E: Third-party evaluator assessing cyber risks** | | | | | | |
| Identifying and developing exploits for specific known software vulnerabilities | Cyber-security experts | Developed exploits for known software | Experts using previous or latest SI systems | Experts without SI assistance | Individual expert | Hours |

## Practice 1.3: Assess evaluability.

Before committing resources to a final evaluation design and execution, work with relevant stakeholders to iteratively refine the evaluation objectives (defined in Practice 1.1) and evaluation parameters (defined in Practice 1.2) to ensure they are evaluable: feasible, plausible, and useful for the evaluation objectives.

1. **Confirm feasibility.** Within the budget available and subject to any other practical constraints, consider whether the outcome(s) are feasible to measure within the timeframe, particularly if high-quality outcome data does not already exist.

   > *Example:* An evaluation of a coding agent may be feasible because the organization's development platform already records tickets resolved, review cycles, and other outcome data, but an evaluation of an SI tool for writing new documentation may be difficult if documentation quality is not systematically graded.

   Review existing measurement and record-keeping practices to understand available data. Consider whether the treatment and comparison conditions can be feasibly applied to sample units from the population of interest, particularly if the deployment plan for the SI system is fixed or already in progress or if potential participants already have access to other SI tools that perform the same functions, leaving no credible comparison group. Review system specifications and deployment plans and, where possible, consult with the population of interest to identify any logistical issues.

2. **Confirm whether uplift is plausible.** Perform a final review of the logic model or threat model and corresponding evaluation parameters to confirm that they are complete, well-defined, logically consistent, and could plausibly reveal substantive positive or negative impacts. Check any assumptions with relevant stakeholders and experts.

   > *Example:* If security rules prevent an LLM-based technical report-writing tool from scanning documents, literature, and other knowledge bases, then uplift from the tool may be less plausible.

3. **Confirm whether uplift estimates will be useful.** Check that the evaluation parameters serve the expected uses of evaluation results identified in Practice 1.1 and identify any additional requirements for the evaluation to be useful and credible (e.g., that the estimates be quantitative, that the study involve participation from a certain stakeholder group, or that the evaluation be repeated periodically to inform ongoing decisions).

   > *Example:* A business assessing return on investment for an LLM-based customer service assistant needs a quantitative estimate of performance uplift precise enough to compare against integration and inference costs.

# 2. DESIGNING & EXECUTING UPLIFT STUDIES

There are many possible approaches to estimating SI-driven uplift: in field studies or proxy studies; with quantitative, qualitative, or mixed methods analysis; and with varying approaches to estimating uplift effects. Given the evaluation objective (Practice 1.1), the evaluation parameters (Practice 1.2), and any relevant practical constraints on evaluability (Practice 1.3), this section discusses practices for designing high-quality uplift studies subject to common practical constraints, divided into five high-level practices:

2.1. Decide how to collect data.
2.2. Choose a study design that balances objectives and practical constraints.
2.3. Sample participants from the population of interest.
2.4. Deploy realistic systems according to the evaluation objectives.
2.5. Test and validate study design before execution.

Table 2.1 exhibits different example designs for data collection and uplift estimation.

**Table 2.1: Possible study designs for example uplift evaluation objectives from Table 1.1.**

| Possible study design | Outcome measure | Sampling strategy | Deployed system | Validation strategy |
|---|---|---|---|---|
| **A: Business assessing customer service ROI *(Quant. field study)*** | | | | |
| Phased rollout of the new system to sites one-by-one | Issues resolved (at an acceptable level of customer satisfaction) per hour per dollar spent | Include all customer service sites in the company | Fine-tuned customer service large language model served as suggested text to human agents | Small-scale pilot |
| **B: Software company deciding on a new coding agent *(Quant. field study)*** | | | | |
| Natural experiment | Feature requests completed per week (of acceptable quality) | Sample from developer teams that are eligible or ineligible for the new agent | Frontier generative coding agent served through a command-line interface | Developer beta surveys and small-scale pilot |
| **C: Enterprise assessing potential productivity increases in professional writing *(Quant. proxy study)*** | | | | |
| Randomized controlled trial (RCT) | Time to complete a writing task at sufficient quality as judged by human experts | Sample evenly from office jobs in the company | Enterprise chatbot product | Qualitative interviews |
| **D: SI developer assessing biological research for drug design *(Qual., then quant. proxy study)*** | | | | |
| Qualitative expert probing sessions followed by quantitative RCT including novices in biology | Accuracy of formulated wet lab protocols | Sample from domain experts (expert probing) and U.S. individuals with a college degree not in biological sciences (RCT) | Guardrails-free, frontier generative language model served through a web UI | Iterative qualitative probing pilots |
| **E: Third-party evaluator assessing cyberattack risks *(Qual. proxy study)*** | | | | |
| Semi-structured expert probing | Functional exploits developed | Sample from experts at a cybersecurity firm or agency | Frontier generative agent served through a command-line interface | Interaction transcript review by subject matter experts |

Before elaborating on these practices, this section first summarizes possible approaches to measuring uplift. Approaches to uplift estimation vary along several key dimensions. Note that these dimensions are not binary: evaluators can conduct studies with elements of both field conditions and proxy conditions, employ methods involving a mixture of qualitative and quantitative approaches, and assert varying degrees of experimental control.

***Proxy studies (ex situ) vs. field studies (in situ)***

1. *Proxy studies* (often referred to as "lab studies"[3]) involve deploying and testing an SI system in controlled conditions outside of the context of regular use. Proxy studies allow evaluators to carefully manipulate experimental conditions or evaluate individual parts of the system in isolation, potentially reducing the number of participants needed to detect uplift effects, but may not perfectly reflect or predict real-world outcomes.

2. *Field studies* allow evaluators to estimate uplift as it occurs in real-world conditions but are often less feasible under practical constraints. If practical constraints allow, the best way to observe uplift in realistic conditions is through a field study, which involves deploying a system for real-world use. Some conditions are required for a field study:
    a. **The system under evaluation must be deployed in the field.** The system does not have to be complete or available to all users—a limited pilot study may suffice (Practice 2.5)—but the system must be available for use by the population of interest for their regular activities with minimal interference from the evaluator. A proxy study could be an alternative if it is infeasible, too costly, or the risk is too high to deploy the system without further evaluation—see the next section below.
    b. **It must be feasible to observe or measure the outcome in the field.** Outcomes may be difficult to measure in the field for several reasons:
        i. *Missing measurement infrastructure.* Many organizations do not have the existing data infrastructure or evaluation resources needed to measure the outcome of interest in the field, particularly if the goal is to quantify uplift. If resources for building quantitative measurement infrastructure are limited, evaluators could consider a qualitative study involving different resources (e.g., observations or interviews with stakeholders using a system in the field) or a proxy study.

[3] This document generally avoids the term "lab study" to minimize confusion stemming from the fact that real-world SI use might occur in a laboratory setting. For example, an examination of how a deployed SI system impacts "wet lab" biological research in practice would be considered a field study.

ii. *No available participants.* For studies of misuse in particular, the population of interest (e.g., threat actors) may not be accessible in the field. A proxy study is more appropriate for these cases.

***Qualitative, quantitative, and mixed methods***

Qualitative and quantitative methods can achieve different types of evaluation objectives, or can answer the same questions in different ways (see Table 2.2).

1. **Qualitative** methods of collecting and analyzing data allow evaluators to explore, understand, and theorize around complex or novel SI-driven impacts that are difficult or impossible to precisely quantify. Even when some aspects of SI-driven impacts are quantifiable, qualitative methods can help identify nuances not evident from quantitative metrics alone. Examples of qualitative methods include interviews, documentary analysis, observational inquiries, and expert probing.

2. **Quantitative** methods allow evaluators to numerically estimate the effects of deploying SI systems on observable outcomes. Unlike qualitative approaches—which can almost always be employed, including to further augment quantitative approaches—quantitative approaches may not always be feasible to execute rigorously (e.g., for very specific use cases with small groups of users).

3. Using **mixed** methods (combining qualitative and quantitative approaches), when possible, can help overcome trade-offs of the two approaches. For example, when a theory of SI-driven uplift is not well understood, qualitative exploratory pilot studies (Practice 2.5) can help identify key processes and measures, prior to attempting larger-scale quantitative experiments. In this case, the samples for qualitative studies may be much smaller than for quantitative studies, and evaluators should still seek to include a wide enough array of participants to capture relevant variations (across job roles, levels of expertise, etc.) that would be studied in the larger quantitative experiments.

**Table 2.2: Examples of high-level trade-offs across study designs.** *As Section 2 describes, different approaches may be suited to different evaluation objectives and may be subject to different practical limitations. Mixed methods approaches can help balance trade-offs.*

| | **Quantitative proxy study** (e.g., RCT) | **Qualitative proxy study** (e.g., expert probing) | **Quantitative field study** (e.g., natural experiment) | **Qualitative field study** (e.g., field interviews) | **Quant. & qual. mixed methods** |
|---|---|---|---|---|---|
| **Possible objectives** | | | | | |
| Evaluate SI-driven uplift | ✓ | ✓ | ✓ | ✓ | ✓ |
| Quantify trade-offs between uplift and costs | ✓ | ✗ | ✓ | ✗ | ✓ |
| Characterize practices and beliefs shaping uplift | ✗ | ✓ | ✗ | ✓ | ✓ |
| **Benefits** | | | | | |
| Reflects real-world SI use | Depends on design | Depends on design | ✓ | ✓ | ✓ if field included |
| Affords more experimental control | ✓ | ✓ | Depends on design | Depends on design | ✓ if proxy included |
| Quantifies uplift | ✓ | ✗ | ✓ | ✗ | ✓ |
| Captures unquantifiable aspects of the outcome | ✗ | ✓ | ✗ | ✓ | ✓ |
| **Practical considerations** | | | | | |
| Uplift hypothesis need not be defined in detail | ✗ | ✓ | ✗ | ✓ | ✓ |
| Outcome need not be quantifiable | ✗ | ✓ | ✗ | ✓ | ✗ |
| SI system need not be deployed in the field | ✓ | ✓ | ✗ | ✗ | ✓ if proxy study |
| Outcomes need not be observable in the field | ✓ | ✓ | ✗ | ✗ | ✓ if proxy study |
| Population of interest need not participate | ✓ | ✓ | ✗ | ✗ | ✓ if proxy study |

## Practice 2.1: Decide how to collect data.

After defining the outcomes of interest in Practice 1.2, decide how to use new or existing data to carefully measure the outcomes. Consider ways that data-based metrics, or outcome measures, might not reflect the outcomes of interest defined in Practice 1.2. For measures of work quality and other outcomes, evaluators often employ grading procedures. Rely on domain experts and existing measurement methods to develop effective grading procedures and appropriately validate automated grading methods (see NIST AI 800-2 [2] Table 2.2 for some example validation practices).

1. **Use standardized and validated measures when possible.** In many cases, well-validated measures may already exist that can be applied to newly generated data. Choosing outcome measures based on domain research and expertise or on existing organizational metrics can make evaluations more relevant and less costly.

   > *Example:* A customer service center may have established performance indicators for service efficiency and quality based on ticketing logs and customer surveys.

   The use of standardized outcome measures also allows for comparisons to prior data and the effects of other new systems or programs—e.g., other enterprise IT systems or worker upskilling programs. If no standardized measures exist, consider conducting in-depth qualitative studies and small-scale quantitative pilot studies to enumerate and validate options for new metrics in collaboration with stakeholders (Practice 2.5). Evaluators may collaborate with stakeholders (e.g., users of the system) to explore how uplift may occur in their specific application context and to identify whether and how it can be quantitatively measured using existing or newly collected data. Even when using quantitative measures, collecting qualitative data can help analyze outcomes that are not completely represented by numerical data.

   a. If possible, avoid "ceiling" or "floor" effects by choosing outcome measures with sufficient range to capture large improvements.

      > *Example:* A multiple-choice questionnaire about time spent on a writing task might cap responses to "ten minutes or less," thus imposing an upper limit on the measurable speed-up SI assistance could provide; evaluators could consider also including options below ten minutes or collecting continuous values instead.

      Pilot tests can help determine whether an outcome measure has adequate range (Practice 2.5). This is particularly important if the evaluation should be repeatable, as SI systems may improve outcomes over time and saturate the measurement range such that further improvement cannot be detected.

2. **Consider multiple strategies for data collection.** Information about outcomes can be obtained from a variety of potentially complementary sources across proxy and field studies.
    a. *Administrative records.* Many uplift studies draw from existing administrative or operational records.

        > *Example:* A study of SI-driven return on investment may draw on recorded sales and other accounting data.

    b. *System telemetry.* Uplift studies may also draw from telemetry and logs generated by new or existing IT infrastructure.

        > *Example:* A study of LLM assistance in customer service may draw on call data already recorded in the company's IT system.

    c. *Interviews*. Evaluators can conduct semi-structured interviews with participants to explore their experiences with deployed systems. Interviews can be particularly helpful for quickly mapping the possible impacts of a system and possible processes and factors impacting uplift. Evaluators can use notes and transcripts to conduct in-depth thematic analysis and theory-building.
    d. *Surveys*. Evaluators can administer small- to large-scale surveys to quickly gather data (both quantitative and qualitative) on participants' experiences with SI and self-reported outcomes. Survey data can serve as a strong complement to the limited types of data observable in administrative records and system telemetry. Open-ended questions can help explore and confirm potential categories of impacts. However, self-reported quantitative uplift estimates—e.g., in time savings—may not be accurate due to expectancy bias (conforming with prior expectations, e.g., that using new technology will save time), the planning fallacy and Hofstadter's law (optimistically underestimating time needed for a task), social desirability bias (conforming with societal expectations, e.g., by reducing time estimates to appear more efficient), or other distortions, and should be verified or calibrated against other experimental measures.
    e. *Rubrics*. To assign quantitative values to work products or task outputs, evaluators often write rubrics that describe different aspects of the outcome of interest. These rubrics can be used to grade the outputs of tasks, particularly in proxy studies where task outputs are discrete and self-contained.
        i. Rubrics should map to the use case or threat model of interest. Rubrics can capture multiple aspects of the outcome of interest. These dimensions may vary depending on the comparison condition.

            > *Example:* A study evaluating uplift to non-expert research scientists in scientific work may assess uplift in scientific report quality through the

> dimensions of accuracy, utility, and breadth. Accuracy may be a more important dimension for assessing non-expert research uplift, whereas creativity may be important for expert research uplift. A rubric scale for the utility dimension could range from counterproductive (i.e., actively misleading) to exceptionally useful (e.g., precisely targeted and immediately actionable, such that it could substantially accelerate success for an experienced practitioner).

ii. Grading should typically be performed either by domain experts (human rubric grading) or automated computer systems (LLM-as-a-judge) which have been validated by domain experts. Care should be taken to ensure that the assessments are independent (e.g., assessors assigning scores before discussing as a group or inter-rater reliability analysis of automated systems). After assessments, it may be valuable to hold group discussions to produce insights about the system and to align on assessors' interpretations of the rubric, which can later be used to refine the definitions of rubric dimensions and scales.

iii. The assessments may include reviewing observations of interactions with the system (for example, review of transcripts of chat interactions), or interviews with the participants themselves.

f. *Documentary analysis.* Even if it is not feasible to interview or survey participants in a study, evaluators can still gather organizational documentation (e.g., use case inventories, interaction logs, etc.) and other written artifacts. Thematic analysis techniques can help process large amounts of textual data.

g. *Case studies.* Evaluators can gather interview, survey, and documentary data focused on specific cases (e.g., particular job roles, use cases, etc.) that are not well understood or that may offer particularly useful insights. Case studies may be particularly helpful for surfacing new kinds of uplift and opportunities for systematic measurement.

3. **Include multiple measures of the same outcome.** It may be challenging to perfectly capture the outcome of interest in a single outcome measure, especially for field studies constrained by limited measurement resources. Instead, evaluators may prefer to choose multiple measures to capture different aspects of the same construct.

> *Example:* An employee with access to an enterprise chatbot may begin producing many more reports and emails because of SI-driven time savings, but their team's overall output may not shift by much if additional reports and emails are only marginally useful to advancing the team's projects. As a result, measuring both writing output and overall team output (e.g.,

products developed, employees hired, key performance targets met) is important to fully understand the uplift provided by the system.

4. **Choose measures to minimize perverse optimization.** It can be difficult to distinguish between genuine improvements in performance and improvements only on specific measures—for example, due to strategic gaming—that do not translate into overall improvements in the underlying outcome. This is typically a concern when study participants are aware of the measures in question and incentivized to optimize them.

   *Example:* If the outcome of interest is the number of reports written, employees and LLM agents may perversely optimize this measure by simply abandoning quality review and producing many low-quality reports.

   Several precautions may help:

   a. Use a suite of multiple metrics to measure different aspects of the outcome. Collecting supplementary qualitative data (e.g., via interviews or observations) may help surface gaming behaviors that are driving up quantitative measures of uplift, or surface broader team, organizational dynamics, or policies that may incentivize gaming behaviors.
   b. Include falsification measures: metrics that should not be affected by system deployment but could be affected by perverse optimization or other factors also related to the outcome.

      *Example:* If the outcome of interest is the number of coding tasks completed, evaluators might also measure the rate of code review revisions or re-opened bugs; these metrics should not be affected by well-constructed SI-assisted code but could increase if developers optimize for task completion by skipping quality review of their own work.

   c. Place quality requirements on quantity metrics (e.g., require that code merges pass an independent code review).
   d. Conduct at least one smaller-scale exploratory pilot study to identify potential gaming behavior and perverse incentives before full-scale experimentation (Practice 2.5).
   e. If possible, do not reveal outcome metrics to participants. This may not be possible if participant compensation is tied to outcomes. Similarly, evaluators should consider omitting information about outcome metrics in LLM prompts or context.

## Practice 2.2: Choose a study design that balances objectives and practical constraints.

This subsection describes several types of SI-driven uplift study designs and practical reasons evaluators might choose them. For any study design, quantitative metrics can be supplemented with qualitative data collection and analysis methods that allow evaluators to explore and model uplift.

Uplift studies should be designed so that differences in outcomes between those with system access (the treatment condition) and those in the comparison group can be attributed to the treatment. This generally requires accounting for systematic differences between the groups other than the treatment (such as confounders: factors that affect both SI access and the outcome).

> *Example:* If software developers are more likely to voluntarily adopt SI tools than workers in other occupations and also tend to earn more, attempts to estimate the impact of SI adoption on compensation without accounting for role type (such as by randomizing access across roles, or matching adopters and non-adopters by role) will conflate variation in outcomes due to role with variation due to SI adoption.

While this subsection includes examples of some specific methods for illustrative purposes, a comprehensive overview and in-depth explanations of all applicable methods are beyond the scope of this document. Disciplines including economics, experimental design, and program evaluation offer many methodological resources that can support evaluators.

1. **Experimental studies.** SI uplift is typically construed as the causal effect of deploying an SI system relative to the counterfactual in which the system is not deployed. If evaluators can control who can access the system, the gold standard approach to quantitatively estimating causal effects is to randomly assign access in a randomized controlled trial (RCT) or similar design. However, full randomization is not always feasible. Instead, evaluators may turn to staged rollouts or other approaches that require additional techniques and assumptions to address confounders.
    a. *Randomized controlled trial (RCT).* In an RCT, the evaluator randomly assigns comparison (no access to the new system) and treatment (access to the new system) conditions to random subsets of participants or to random subsets of tasks. RCTs are often considered the gold standard for quantitative, causal estimates of uplift because, if executed correctly, randomized assignment eliminates systematic biases from confounding factors, even unknown ones.

        > *Example:* An evaluator interested in studying the impacts of coding agents on open source software development may ask a random subset of open source software maintainers to resolve issues in their repositories with SI assistance and track the time

> it takes them to complete coding tasks (including quality review) compared to maintainers without SI assistance.

In some cases, it is advisable to stratify treatment assignment by relevant variables of interest, such as by assigning system access to equal numbers of each job or team role, level of expertise, or task difficulty, rather than applying pure randomization. Randomized experiments may test treated and untreated participants in parallel or use crossover designs in which participants' system access or use changes partway through the trial. The latter can help control some confounders but risks introducing others via carryover or anticipation effects. Adaptive testing approaches, in which the probability of treatment is adjusted over time, are an area of emerging practice in machine learning evaluation that aims to address efficiency and fairness concerns. For very specific use cases or bespoke systems with few users, evaluators may attempt to estimate uplift for a single person, team, or other treatment unit by alternating access in phases (single case or "N-of-1" design); however, such a design requires stronger assumptions and may be less likely to generalize outside the study subjects.

b. *Staged rollout.* In some cases, practical constraints or trade-offs prevent simultaneous treatment (i.e., rolling out the new system to all treated participants at once). Instead, evaluators may make the new system available to random clusters of participants in stages. As it is often common practice for organizations to deploy new technology in pilots, betas, and other phased rollouts, randomized assignment could be integrated, potentially at low cost.

> *Example:* A business assessing a new LLM support tool for customer service workers may, each week, permit an additional random subset of customer service office sites to access the new tool and, over a few months (without changing the model version or product functionality), measure the difference in issues resolved per week compared to sites without access.

c. *Considerations for proxy studies.* Because proxy studies occur outside the original context of use, evaluators typically must select simplified or standardized tasks to be conducted in a constructed environment with pre-specified outcome assessment methods.

> *Example:* A business consulting firm assessing potential return on investment for an enterprise chatbot might assign consultants a series of pre-defined business problem-solving tasks involving a set of example spreadsheets and reports, provide a random subset of these consultants access to a chatbot, and measure the speed and quality of their solutions.

Key considerations for proxy studies include, but are not limited to:

i. *External validity.* A drawback to proxy studies is that results may not translate to uplift in field conditions because conditions are controlled and because tasks, interactions, and outcome assessments are often simplified. To preserve external validity and robustness, the experimental design should model the uplift hypothesis as closely as possible. Document carefully any gaps between the proxy and the threat model that might cause results to differ in a real incident (e.g., that exploiting a known vulnerability on an isolated cyber range target may be easier than exploiting production systems where vulnerabilities must be discovered).

ii. *Representative task selection.* Evaluators should include tasks that fully reflect those included within the scope of the uplift hypothesis.

> *Example:* If the uplift hypothesis for professional writing work includes both report writing and email writing, a complete proxy study should include both types, balanced according to the actual task mix expected to be used in practice.

iii. *Avoid contamination.* Avoid using tasks that are likely to have been included in the SI system's training data (e.g., publicly documented vulnerabilities, published laboratory protocols, or public code repositories) or used to make decisions influencing system design.

iv. *Multi-metric and subtask-level measurement.* In a proxy setting, evaluators may be able to more closely monitor participant-SI interactions and build a larger panel of task- and subtask-level metrics. Taking more measurements (e.g., through telemetry, self-reported surveys, and expert grading) can help surface heterogeneous effects (e.g., SI assistance benefiting novices more than experts) and behaviors that may be relevant to field deployments.

d. *Expert probing.* Expert probing is a type of proxy study where experts assess SI systems through probing, or prompting, the system, often through an extended interaction that assesses the capability of the model to provide accurate and useful information or to correctly perform tasks.

> *Example:* An evaluator interested in understanding whether a new SI system may enable biological laboratory tasks might assemble a group of relevant subject-matter experts for interactive prompting sessions to probe capabilities in that domain.

While these experts may start with a pre-determined set of prompts or tasks, the interactive nature of the session allows them to dynamically develop and test new prompts or tasks, in order to more clearly ascertain the system's outer ability limit and potential uplift in specific scenarios. Expert probing is a relatively nascent and emerging practice; it is not yet widely standardized across the research community.

2. **Observational studies.** Particularly when running a field study, experimentally manipulating who can access the system may not be possible. If the evaluator cannot control which participants can access the system and which cannot, they must rely on observational data. In these cases, estimating uplift is typically more challenging due to the difficulty of separating the effect of SI access from the effect of preexisting differences between groups or other factors related to both outcomes and SI access.

> *Example:* Enthusiastic tech users may be both more likely to volunteer for an SI system pilot and more likely to succeed on technical tasks compared to users with less tech experience.

However, it may still be possible to estimate uplift by exploiting natural variation in the deployment of an SI system. In some cases, evaluators may be able to make causal inferences by taking advantage of observational data from uncontrolled events that simulate random assignment, sometimes referred to as "natural" experiments.

> *Example:* If a new customer service support tool happens to be deployed to some customer service workers before others due to the scheduling of IT license updates, unrelated to worker productivity, evaluators may compare productivity between customer service workers with licenses and workers without (similar to the staged rollout experiments described above but not controlled by the evaluator).

Evaluators may also turn to non-randomized observational study designs, which generally rely on larger samples and stronger assumptions that may be difficult to verify in practice. To estimate uplift in these settings, evaluators may employ analysis methods (described below) that attempt to adjust for possible confounders (e.g., age or SI literacy) or simulate potential outcomes for hypothetical comparison groups. Qualitative methods can also help investigate and explain uplift through solely observational field data.

To recognize possible "natural" experiments and assess the feasibility of other observational studies, evaluators can gather information about how the SI system was or will be made available to participants and what kinds of data are available.

a. *Natural experiments.* In some cases, access to a new SI system may depend on arbitrary factors or shocks that are unlikely to influence outcomes other than by influencing access to the system. When these arbitrary natural conditions are similar enough to the conditions of a randomized experiment, evaluators can justify uplift estimation based on observational data even in the presence of unobserved confounders. If treatments are assigned based on a sharp threshold or cutoff, it may be possible to compare outcomes for participants close to the cutoff (a.k.a. regression discontinuity design).

> *Example:* If a company provides free access to a new coding agent only to developers affiliated with open source repositories above a certain popularity threshold, an evaluator could compare outcomes of repositories close to, but on either side of, the cutoff.

b. *Adjustments for confounders.* Even when a natural experiment is not available, evaluators may still attempt to correct for confounders based on other observed attributes—e.g., by pairing treated and untreated participants with similar characteristics (a.k.a. matching methods) or by identifying groups that differ in outcomes and comparing prior SI exposure (a.k.a. case-control studies). However, these methods require the typically strong assumption that after adjusting for measured attributes there are no remaining unmeasured factors that affect both SI access/use and outcomes.

c. *Comparison of trends over time.* Different sub-populations may experience diverging changes in outcomes because of differential exposure. Evaluators could identify a treatment and comparison group and compare outcomes after exposure under the assumption that the two groups would have had parallel trends in outcomes had they not been exposed to SI (a.k.a. difference-in-differences design).

> *Example:* If a new coding agent happens to first support Python development but not R, evaluators could compare open source contributions to Python and R projects before and after the new tool's introduction.

## Practice 2.3: Sample participants from the population of interest.

If any of the experimental conditions involve human participants, evaluators should recruit or otherwise sample study participants representative of the population of interest.

1. **Choose participants closely aligned with the uplift hypothesis.** Develop a strategy to recruit a sample of participants according to the population of interest and unit of treatment defined in Practice 1.2 (e.g., software development teams for programming uplift or biologists for uplift in biological research). It is important to recognize and

manage potential bias introduced by the recruitment strategy and avoid defaulting to simply including the most convenient participants if they are not representative of the population of interest. In a quantitative study, a statistical power analysis can be used to assess the sample size needed to estimate uplift.

> *Example:* Individuals who volunteer to sign up for a study of a new SI system may have differing motivations and behaviors from non-volunteers, resulting in a non-representative estimate of uplift that could limit external validity relative to the broader population.

a. *Matching expertise*. If the uplift hypothesis specifies a certain level of expertise (e.g., new graduate-level uplift), recruited participants should not have too much expertise (e.g., graduate degree plus work experience) or too little (e.g., undergraduate students). If the required background is too costly to recruit (e.g., rare domain expertise), evaluators may need to settle for smaller sample sizes or a broader population (e.g., individuals with less domain expertise). Claims about the findings should be qualified to reflect these limitations. In many realistic settings (e.g., general-purpose chatbot deployment at a large organization), the population of interest may span multiple levels of expertise that should be represented if the goal is to estimate real-world impacts.
b. *"Proxy" participants.* Some populations are not directly recruitable (e.g., malicious or highly capable threat actors), unethical to recruit (e.g., children for a study involving risks), or too costly to recruit. Evaluators can instead recruit "proxy" participants with similar backgrounds (e.g., "red team" domain experts). When using proxy participants, it is especially important to design study instructions and incentives to best approximate the motivations and incentives of the proxied population.

> *Example:* An evaluator assessing misuse risks with expert probing might show participants an "actor profile" describing a hypothetical persona and ask them to adopt that persona to guide their behavior during the study.

2. **Account for previous SI experience and other participant characteristics.** Some participant characteristics, such as previous SI experience, may substantially impact uplift by changing the ways that participants interact with SI systems. It is a good practice to measure these features at the start of the study (e.g., through an intake survey) in order to assess the degree to which their distribution in the sample matches the population of interest and to enable adjustment for confounders and estimation of heterogeneous effects. In a quantitative study, a statistical power analysis can be used to assess the sample size needed to estimate uplift for different subgroups according to minimum detectable effect sizes and desired error rates.

3. **Carefully consider incentives and framing that could influence participant behavior.** The instructions and incentives provided to participants can potentially influence study outcomes, particularly in proxy studies where tasks are contrived and no natural incentive structures (e.g., for job performance) exist. Ideally, compensation and other incentives should be based upon evaluation objectives and similarity to the real-world use case or threat model of interest. Incentive models include, but are not limited to:
    a. *Performance-based.* This approach ties rewards (usually monetary compensation) to the successful achievement of a goal and may be well-suited when uplift is measured by task completion, or for simulating environments with competitive pressures. However, this approach can incentivize participants to cheat (e.g., by using third-party SI tools in a no-SI condition). For fair compensation in human studies, this design may require assigning each participant both SI and no-SI tasks and comparing outcomes between tasks, so as not to unfairly advantage one group of participants.
    b. *Fixed.* In this approach, participants are compensated based on the duration of their participation regardless of the speed or quality of output. This may be well-suited for studies aimed at estimating uplift afforded by SI systems in workflow integration or long-term behavioral changes; however, it may lead to passive SI usage or inadvertently mask efficiency gains as participants are not incentivized to complete tasks well or quickly.
    c. *Process-based.* This approach rewards participants based on their fidelity to a process, such as adherence to a methodology. Process-based approaches may be well-suited for studies where uplift is defined by increased precision or reduced error rates, or in environments where external factors might cause a protocol to fail even if perfectly executed (e.g., a biological wet lab assay failing due to reagent quality or insufficient time to complete). But they may less accurately reflect real-world conditions where the end result is the ultimate measure of success.

    One or more of these incentives may be needed to realistically simulate the use case or threat model. In field studies where participants are performing typical tasks, additional incentives may not be necessary and could potentially be undesirable if they distort behavior compared to real-world conditions.

4. **Ensure ethical human subjects research.** Follow all applicable federal laws and guidelines for human subjects research. For example, federal guidelines for relevant and useful, rigorous, independent, objective, and transparent evaluation suggest taking measures to safeguard the dignity, rights, safety, and privacy of participants in human subjects research [3]. Typical approaches are helpfully outlined in Office of Management and Budget Memorandum M-20-12 [3] and include, but are not limited to:

a. Ensuring that evaluators are up-to-date on relevant training and certifications related to human subjects research and privacy.
b. When appropriate, ensuring that participants have given full, informed consent and are aware of any benefits or risks of participation throughout the study.
c. Establishing protocols to maintain participant confidentiality and information privacy and ensuring participants receive complete, accurate, and accessible information about how their data will be used and shared, prior to providing consent. SI interaction logs commonly contain personal or sensitive information which may require specific consent or redaction procedures.
d. When appropriate, surveying participants throughout the study to monitor and mitigate potential negative effects of participation, including proliferation risks for dual use tasks in which participants are exposed to functional exploit code or other hazardous SI-generated content.

## Practice 2.4: Deploy realistic systems according to the evaluation objectives.

For the treatment conditions, evaluators must choose and sometimes design SI systems to test. Evaluators must also decide what instructions, interfaces, and training should be provided to users. The characteristics of these systems and deployment conditions should match real-world conditions of the use case or hypothesized threat model as closely as possible to maximize the usability and generality of results. System settings, deployment setup, and the testing environment can all materially affect uplift estimates and should be included in study pre-registration or documentation.

1. **Consider whether the evaluation should measure optimal deployment conditions or average-case deployment conditions.** Many uplift studies utilize custom-designed SI systems and carefully instruct participants in their use, but real-world deployment may be more haphazard and less well-resourced. Consider what deployment conditions—including model selection, tool availability, token budget, reasoning effort, and other parameters discussed in NIST AI 800-2 [2] Table 2.2—best fit the evaluation objectives and design treatment conditions to match. For some threat models, evaluation objectives will require matching the hypothesized threat scenario rather than common real-world usage.

> *Example:* If measuring the extent to which an SI system could uplift a malicious actor in dual use biological research tasks, providing participants with SI systems that have free access to agent tools (e.g., web search) and a large budget may provide a more accurate estimate of the maximum possible uplift in a hypothetical worst-case scenario involving a well-resourced malicious actor.

2. **Consider whether the tested system should include safeguard layers.** In real-world deployments of an SI product, safeguard layers that constrain system outputs (e.g., through input or output filters) may be included, so the tested system should also incorporate these components insofar as realism is desirable. Conversely, it may sometimes be more realistic to use a non-safeguarded or "guardrails-free" version of the system—for an evaluation of criminal misuse of open-weight models, for example. Based on the logic model or threat model, decide whether participants should be prohibited from using jailbreaking techniques or proactively trained in them.

3. **Employ comparable interfaces in the comparison and treatment conditions.** A low-quality user interface employed in the comparison or treatment condition may artificially inhibit performance and skew uplift estimates. If the objective is to model uplift of a new SI product relative to current practice, the comparison condition should provide an interface similar to what a participant is accustomed to using, and the treatment condition should provide an interface similar to an SI product that would be introduced in that domain. If the objective is to compare SI models directly, then evaluators may consider employing the same interface across conditions to avoid product interface-specific confounding.

4. **Track and (if feasible) minimize changes to the system over time.** Particularly in field studies or in studies evaluating third-party SI vendors, deployment conditions may drift over time (e.g., due to moderation updates to a system accessed through an API). Seemingly small changes in training data and model development may have large effects on outcomes. Evaluators should track changes to the system during the study period (e.g., by periodically running a fixed set of test prompts and monitoring for changes in the response distribution). In many evaluation scenarios, evaluators may wish to ensure the system remains unchanged (e.g., by using an API for a specific model version, self-hosting, or running the experiment over a short period of time). Or, evaluators may attempt to extend or repeat the experiment over time to account for common variations in the system (e.g., changes in latency, context windows, model routing, and other dynamics).

5. **Assess and (if feasible and desirable) minimize treatment non-compliance.** Unbiased uplift estimation often depends on only treated units having access to SI systems. This assumption may be violated if non-treated participants gain access to the deployed system (e.g., by accessing a chatbot on a personal machine). Technical controls and post-study surveys where participants can admit to violations without penalty can be used to filter out non-compliance.
    a. *Non-treated participants gaining access to the evaluated system.* To the extent possible, consider beforehand possible ways non-treated units may gain access

to the treatment and attempt to minimize non-compliance (e.g., by blocking access to other SI systems in experiments where participants in the comparison condition are allowed internet access but are not intended to use SI) or simply track possible avenues of non-compliance with telemetry or surveys. Non-compliant units can sometimes be filtered out during analysis.

b. *Treated participants not using the evaluated system.* For studies aiming to estimate the effects of the *use* of an SI system (as opposed to estimating the effects of *deploying* an SI system), another form of non-compliance is when treated units choose not to use or fail to use the systems made available to them. Sometimes this is easy to detect; participants who did not use any tokens can be filtered out. Other forms of non-compliance may be more difficult to account for—e.g., if participants use the SI systems incorrectly. Consider whether these kinds of non-compliance should be included in uplift estimation (e.g., in average-case analysis) or if participants should be afforded additional training or support to avoid these outcomes (e.g., in best- or worst-case analysis).

6. **Assess and minimize treatment spillover effects.** Spillover effects occur when treated participants' access to an SI system affects outcomes for other participants outside the treatment condition. Even with perfect compliance, spillover effects can introduce statistical bias to uplift estimates.

> *Example:* An evaluator randomly assigns access to a new coding agent on an individual basis, such that one member of a software development team could have access (treatment condition) while another does not (comparison condition). If a software developer in the comparison condition—without access to the coding agent—finds it harder (or easier) to read and build on their colleague's SI-assisted code and thus is less (or more) productive, estimated productivity uplift may be overestimated (or underestimated) compared to an experiment where developers were truly isolated.

One approach to reduce spillover is to try to isolate treated and untreated participants as much as possible, though this may be unrealistic. Randomized saturation (randomizing the share of participants with SI access and estimating spillover by comparing participants with high vs. low treatment intensity) and other study designs can be used to estimate both direct and spillover effects.

7. **If feasible, minimize expectancy bias with blinding.** Participants' and evaluators' preconceived notions about SI or about certain SI products may change behavior and bias self-reported outcomes.

   a. *Blinding participants.* If the evaluation objective is to estimate uplift absent expectancy bias, consider anonymizing the name of the SI product in use. An

intermediary model or model wrapper may be required to stop systems from revealing the model name. For uplift evaluations involving a no-SI comparison condition, completely disguising the treatment from the comparison group is usually not possible—unless the systems used in the comparison and treatment conditions have similar outputs (e.g., in an evaluation comparing a rules-based risk scoring algorithm to a machine learning risk scoring algorithm).

b. *Blinding evaluators.* If possible, consider disguising treatment assignments and the treatment assignment process from evaluators to avoid expectancy bias during data analysis. Evaluators could develop data analysis programs using data with dummy treatment assignment codes, swapping in the true treatment assignments only after the programs are locked.

## Practice 2.5: Test and validate study design before execution.

1. **Re-engage with stakeholders.** Having carefully planned an uplift study, engage again with system users, administrators, and other relevant stakeholders to check the feasibility and validity of the study design.

   > *Example:* Software developers may flag that a planned no-SI comparison condition would block an autocomplete tool they previously used for work, rendering the comparison condition unrealistic.

2. **Conduct pilot studies.** Small-scale live pilots can help identify potential errors (e.g., guardrails blocking in-scope activities, poor user interfaces, incomplete telemetry, mis-calibrated grading) and possible improvements. Test-driving experimental methods with a small-scale pilot can help refine the evaluation parameters defined in Practice 1.2, quality check the outcome measures, and catch other issues with experiment design and data collection methods. SI agents and other simulation methods may help to conduct rapid, repeatable testing to refine experiment design and methods before piloting with human participants.

# 3. ANALYZING & REPORTING RESULTS

Once the experiment is executed and measurements are obtained, the next step is to analyze and contextualize estimates of SI-driven uplift. Analysis should take into account the evaluation objective, the uplift hypothesis, the study design, and context-specific characteristics of the data. Evaluators should report sufficient information to enable others to interpret, review, replicate, and extend results. Practices at this stage include the following:

3.1. Choose appropriate analysis methods and express uncertainty.
3.2. Share details of evaluation and evaluation data.
3.3. Report qualified claims.

## Practice 3.1: Choose appropriate analysis methods and express uncertainty.

Evaluators should define, conduct, and report a statistically valid analysis procedure that aligns with intended evaluation objectives, accounts for sources of variation and uncertainty, and describes underlying assumptions (also see NIST AI 800-2 [2] Practice 3.1). As discussed in the Introduction, defining the analysis procedure in advance of data collection, such as via pre-registration or a statistical analysis plan, can help mitigate the potential influence of experimenter or analyst bias on conclusions.

1. **Report appropriate uplift statistics along with any associated modeling assumptions.** Quantitative studies often estimate the average uplift across the participant population. The estimation of uplift statistics may involve choices about how to mathematically model (e.g., using linear regression, nonparametric models, etc.) the relationship between SI deployment (the treatment condition), the outcome, and confounding characteristics that should be carefully considered and explicitly reported.

2. **Examine heterogeneous effects.** SI-driven uplift can depend on many factors (e.g., previous SI experience, training, system settings, use case, etc.). Estimating the differential effect of SI treatment on pre-defined subgroups (e.g., users with varying degrees of previous SI experience) can inform many evaluation objectives.

   > *Example:* If SI assistance benefits novices more than experts, deployers may alter their rollout strategy to prioritize novices.

   Evaluators may also wish to examine the distribution of uplift across participants, particularly for studies of misuse where only one instance of uplift may be sufficient to cause harm. When uplift differs by subgroup, generalizing results beyond the population treated in the study requires weighting effects by population composition. One challenge in the adoption setting is that users may adapt their SI-assisted workflows over time, meaning treatment effects are not stable, particularly in the short term.

3. **Account for confounders to the extent possible.** Non-randomized studies often require evaluators to account for confounding factors that affect both the treatment assignment and the outcome (e.g., prior SI experience might make some participants more likely to use SI products and more likely to benefit from them). A particular issue to consider in observational studies is temporal ambiguity or reverse causation, e.g., if participants chose to use or not use an SI system due to the outcomes they experienced using or not using it. Additionally, analyses should also check for extreme observations, assess their influence, and perform adjustments as appropriate.
    a. *Quantitative studies.* For non-randomized studies, prior to data collection identify observable factors that are likely to affect both the treatment and outcome and determine an appropriate method to account for their influence. Sensitivity analyses can help assess the degree to which remaining unobserved confounders potentially affect conclusions. While properly executed randomization accounts for confounders, adjusting for pre-specified factors that may affect the outcome can improve the accuracy of uplift estimation even in randomized studies. However, evaluators must take care to avoid biasing estimates by adjusting for intermediate factors that are part of the mechanism of uplift, or for factors that are common effects of both SI access and the measured outcome.

        > *Example:* In an uplift study investigating the impact of introducing a customer-facing chatbot into a customer service workflow on customer satisfaction scores, adjusting for survey submission (e.g., by only analyzing cases in which customers submitted a survey) may mask actual uplift. Use of the chatbot could influence both satisfaction and likelihood of submitting a survey.

    b. *Qualitative studies.* During data collection, take note of potential factors that might affect uplift. Continuously adjust recruitment and data collection to fully explore these factors and include them in theory building and analysis.

4. **Use interaction logs and other telemetry to assist analysis.** The use of SI systems can produce high-granularity telemetry and logs that enable experiment audits, exploration of possible uplift mechanisms, and analyses of factors that influence uplift (for example, system response style or refusal rates). Interaction logs can assist evaluators in analyzing heterogeneous effects, filtering out non-compliers when appropriate, and identifying correctable errors in experimental design.

5. **Report estimated uncertainties for possible sources of variation.** For quantitative estimates of uplift, uncertainty metrics can take the form of standard errors, confidence intervals, credible intervals, etc. For qualitative analysis of uplift, uncertainty may be expressed by self-reported confidence scores, boundary conditions, or other analysis.

6. **Support comparisons by appropriate statistical tests**, accounting for multiple comparisons when needed. Statistical test results should be interpreted probabilistically and considered alongside effect size.

## Practice 3.2: Share details of evaluation and evaluation data.

Similar to other kinds of evaluation, sharing evaluation study design details and evaluation data increases reproducibility and enables others to make their own determinations about what results imply (see NIST AI 800-2 [2] Practice 3.2). For uplift studies in particular, evaluators should:

1. **Report key evaluation details** as appropriate**.** Such details include, but are not limited to the evaluation objective and uplift hypothesis (Practice 1.1); key evaluation parameters such as the outcome measures, comparison conditions, treatment conditions, unit of treatment, and participant population (Practice 1.2); the data collection instruments, including any grading rubrics (Practice 2.1); the study protocol, including the type of design, treatment assignment strategy, and study timeframe (Practice 2.2); the sampling strategy (Practice 2.3); details of the system deployed for treatment conditions, including model identifiers, method of access, inference parameters, system prompts, tool availability, safeguard configurations, and token budgets (Practice 2.4); and any statistical assumptions or methods used to analyze and interpret collected data (Practice 3.1).

2. **Publish raw or aggregated data and analysis code** to the extent possible given legal, security, and business constraints and while maintaining appropriate protections for human subjects. This may include prompts, interaction transcripts or logs, system telemetry, and outputs from all graders.

3. **Report all the results obtained (including null or negative results)**. Knowing about null and negative findings, in addition to positive findings, is important for fully understanding the different ways implementation conditions can affect SI-driven uplift and reduces publication bias across the broader field. Report all results obtained for all of the outcomes measured.

## Practice 3.3: Report qualified claims.

Ensuring that claims made on the basis of evaluation results are adequately supported by the evaluation's scope and context assists evaluation consumers in interpreting results (as with other types of evaluations; see for example NIST AI 800-2 [2] Practice 3.3).

1. **Distinguish claims about evaluation results from other claims.** Differentiate observations, inferences, predictions, and normative statements (NIST AI 800-2 [2] Practice 3.3.1).

2. **Maintain a close connection between data and claims.** Particularly for qualitative studies, evaluators can more effectively substantiate and qualify claims by closely associating them with the empirical data—such as through direct quotes, granular annotation, and thematic coding.

3. **Distinguish associations from causal effects.** For estimated causal effects, clearly state the unit of treatment, the comparison/treatment condition, and any assumptions that underpin the causal interpretation of the estimate.

4. **Include quantifications of and qualifications about uncertainty.** Clearly state other assumptions and limitations of the evaluation, including the inclusion or exclusion of confounders, spillover effects, and other indirect effects. Report any observed challenges to the study design such as attrition, evidence of non-compliance, changes to the model version, safeguards, or other system properties, etc.

5. **Assess the role of the evaluators in the experiment and how it may affect results.** For quantitative studies, consider specifying which analysis protocols were determined prior to gathering the data and which were added or modified *post hoc*. For qualitative studies, make note of any preliminary concepts used to frame data collection and analysis, such as taxonomies of tasks that could be impacted by SI systems.

6. **Discuss possible mechanisms for reported effects and associations**, including any evidence for or against particular mechanisms or theories of uplift as articulated in Practice 1.1.

7. **Discuss factors of the evaluation setting that may limit whether conclusions generalize.** This includes the realism of a proxy study with respect to a real-world setting, the representativeness of the participant population, incentive structure, evaluation awareness (for both tested systems and human participants), etc. Use caution when making claims that generalize results beyond the study setting and uplift hypothesis, for example, in extrapolating results from the organization under study to others in the same industry, or in extrapolating results beyond the system versions tested and time scale studied. Avoid conflating population-level uplift estimates with uplift at an individual level.

8. **Contextualize results within broader workflows and organizational settings.** If extrapolating task- or worker-level uplift to the level of a job or organization, carefully consider the potential influence of feedback loops, equilibrium effects, and constraints

including whether other tasks, processes, or resource shortages may limit downstream impact (e.g., because of increased reviewing loads, or because the tasks with high uplift have relatively low value to downstream work). Consider the extent to which SI usage may shift the mix of tasks associated with a job or workflow; uplift measured on a worker's old task mix and new task mix may differ.

9. **Discuss the degree to which evaluation results agree or disagree with related evaluations**, including benchmark evaluations, and other relevant evidence and existing knowledge (see NIST AI 800-2 [2] Practice 3.3.6).

# REFERENCES

[1] The White House (2025) America's AI Action Plan. Available at https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.

[2] NIST (2026) Practices for Automated Benchmark Evaluations of Language Models (National Institute of Standards and Technology, Gaithersburg, MD), NIST AI 800-2 Initial Public Draft. https://doi.org/10.6028/NIST.AI.800-2.ipd. Available at https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-2.ipd.pdf

[3] Office of Management and Budget (2020) M-20-12: Phase 4 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Program Evaluation Standards and Practices. Available at https://www.whitehouse.gov/wp-content/uploads/2020/03/m-20-12.pdf.

[4] Paskov P, Wei K, Hong SZ, Bateyko D, Roberts-Gaal X, Ezell C, Praninskas G, Chen V, Bhatt U, Guest E (2026) RCTs & Human Uplift Studies: Methodological Challenges and Practical Solutions for Frontier AI Evaluation. https://doi.org/10.48550/arXiv.2603.11001. Available at http://arxiv.org/abs/2603.11001

[5] Paskov P, Byun MJ, Wei K, Webster T (2025) Preliminary suggestions for rigorous GPAI model evaluations (RAND Europe). Available at https://www.rand.org/content/dam/rand/pubs/perspectives/PEA3900/PEA3971-1/RAND_PEA3971-1.pdf

[6] Raji ID, Denton E, Bender EM, Hanna A, Paullada A (2021) AI and the Everything in the Whole Wide World Benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1. Available at https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html.

[7] NIH Office of Intramural Research (2025) Dual-Use Research. Available at https://oir.nih.gov/sourcebook/ethical-conduct/special-research-considerations/dual-use-research.

[8] OECD (2023) Glossary of Key Terms in Evaluation and Results-based Management in Sustainable Development (Second Edition) (OECD), https://doi.org/10.1787/632da462-en-fr-es. Available at https://www.oecd.org/en/publications/glossary-of-key-terms-in-evaluation-and-results-based-management-for-sustainable-development-second-edition_632da462-en-fr-es.html

[9] Peersman G, Guijt I, Pasanen T (2015) Evaluability Assessment for Impact Evaluation (Overseas Development Institute, London). Available at https://www.betterevaluation.org/sites/default/files/Evaluability_assessment_for_impact_evaluation.pdf.

[10] Liu LT, Raji ID, Zhou A, Guerdan L, Hullman J, Malinsky D, Wilder B, Zhang S, Adam H, Coston A, Laufer B, Nwankwo E, Zanger-Tishler M, Ben-Michael E, Barocas S, Feller A,

Gerchick M, Gillis T, Guha S, Ho D, Hu L, Imai K, Kapoor S, Loftus J, Nabi R, Narayanan A, Recht B, Perdomo JC, Salganik M, Sendak M, Tolbert A, Ustun B, Venkatasubramanian S, Wang A, Wilson A (2026) Bridging Prediction and Intervention Problems in Social Systems. https://doi.org/10.48550/arXiv.2507.05216. Available at http://arxiv.org/abs/2507.05216

[11] Wallach H, Desai M, Cooper AF, Wang A, Atalla C, Barocas S, Blodgett SL, Chouldechova A, Corvi E, Dow PA, Garcia-Gathright J, Olteanu A, Pangakis NJ, Reed S, Sheng E, Vann D, Vaughan JW, Vogel M, Washington H, Jacobs AZ (2025) Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. *Forty-Second International Conference on Machine Learning Position Paper Track*. Available at https://openreview.net/forum?id=1ZC4RNjqzU.

[12] Adcock R, Collier D (2001) Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *The American Political Science Review* 95(3):529–546. Available at https://www.jstor.org/stable/3118231.

[13] APA (2018) APA Dictionary of Psychology. Available at https://dictionary.apa.org.

[14] Schroeder K, Trinh H, Pillitteri VY (2024) Measurement guide for information security: Volume 1 — identifying and selecting measures (National Institute of Standards and Technology (U.S.), Gaithersburg, MD), NIST SP 800-55v1. https://doi.org/10.6028/NIST.SP.800-55v1. Available at https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-55v1.pdf

[15] Singh R, Blili-Hamelin B, Anderson C, Tafesse E, Vecchione B, Duckles B, Metcalf J (2025) Red-Teaming in the Public Interest (Data & Society Research Institute), https://doi.org/10.69985/VVGP4368. Available at https://datasociety.net/library/red-teaming-in-the-public-interest/

[16] Braun V, Clarke V (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2):77–101. https://doi.org/10.1191/1478088706qp063oa. Available at https://www.tandfonline.com/doi/full/10.1191/1478088706qp063oa

[17] Souppaya M, Scarfone K (2016) Guide to Data-Centric System Threat Modeling (National Institute of Standards and Technology, Gaithersburg, MD), Draft NIST Special Publication 800-154. Available at https://csrc.nist.gov/files/pubs/sp/800/154/ipd/docs/sp800_154_draft.pdf.

[18] NIST (2012) Guide for conducting risk assessments (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication 800-30r1. https://doi.org/10.6028/NIST.SP.800-30r1. Available at https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf

[19] Bodeau D, Fox DB, McCollum CD (2018) Cyber Threat Modeling: Survey, Assessment, and Representative Framework (Homeland Security Systems Engineering & Development Institute), Case Number 18-1174. Available at https://www.mitre.org/sites/default/files/2021-11/prs-18-1174-ngci-cyber-threat-modeling.pdf

# APPENDIX: GLOSSARY

| | |
|---|---|
| SI-driven uplift | The impact of deploying a given SI system on a defined set of outcomes for a particular set of people or organizations, relative to the same outcomes in a counterfactual scenario in which the SI system is not deployed—usually one in which humans use non-SI tools or a previous automated system. In the field of machine learning, studies of the impacts of machine learning systems on human performance are often referred to as "human uplift" studies [4, 5]. Some uplift studies measure the effects of deploying, or providing access to, an SI system, while others measure the effects of actual measured use. While the term SI-driven uplift is colloquially used to describe improvements in people's ability to achieve certain outcomes, uplift evaluations may uncover increases, decreases, or null effects. |
| Automated benchmark [2] | A particular combination of a dataset and a metric, conceptualized as representing one or more specific tasks or sets of abilities (adapted from [6]). A benchmark is used by a community of researchers as a shared framework for the comparison of methods. In practice, a single benchmark may combine multiple datasets (at least test data and sometimes also pre-specified training and validation data). |
| Case-control study | A type of quasi-experimental observational study in which two existing groups with differing outcomes are identified and compared on the basis of exposure to some treatment (SI access). |
| Comparison condition | A condition in which a comparison system is used (as opposed to a treatment condition in which the SI system under evaluation is used). The comparison system is a system other than the one under evaluation, often not involving SI, and potentially including human individuals or teams, systems based on simple heuristics or rules, or systems based on random chance (adapted from [2]). Sometimes referred to as a control condition. |

| | |
|---|---|
| Confounder | A factor that distorts the observed association between the treatment condition (e.g., access to an SI system) and the outcome of interest, often by affecting both the outcome and the likelihood of treatment. For example, prior SI experience may make participants both more likely to use an SI system and more likely to benefit from it. Confounders can systematically bias uplift estimates unless addressed by the study design or otherwise adjusted for. |
| Difference-in-differences | A quasi-experimental statistical technique that compares the difference in the outcome measure in the treatment condition and comparison condition *before* an intervention (SI system deployment) to the difference between conditions *after* an intervention. Assumes that, absent the intervention, the change in outcomes in the treatment condition would have been the same as the change in outcomes in the comparison condition (parallel trends). |
| Dual use biological research | Biological research is considered "dual use" in nature if the methodologies, materials, or results could be used to cause harm [7]. Dual Use Research of Concern (DURC) is a small subset of life sciences research that, based on current understanding, can be reasonably anticipated to provide knowledge, information, products, or technologies that could be directly misapplied to pose a significant threat with broad potential consequences to public health and safety, agricultural crops and other plants, animals, the environment, materiel, or national security [7]. |
| Estimate [of SI-driven uplift] | A calculation or approximation of SI-driven uplift based on outcome measures. |
| Evaluability | The extent to which an activity or project can be evaluated in a reliable and credible fashion [8], including feasibility, plausibility, and utility [9]. |
| Evaluation objective(s) | The goal(s) of an evaluation, including what should be measured (the measurement construct) and how the measurements will be used (e.g., to make decisions about system design or deployment). |
| Expert probing | Type of study where experts (typically human experts, automated computer systems, or a combination of the two) assess SI systems through probing, or prompting, the system, often through an extended interaction that assesses the ability of the model to perform tasks. |

| | |
|---|---|
| External validity | The degree to which an evaluation's results generalize beyond the study setting to real-world settings. |
| Field study [of SI-driven uplift] | Study in which data is collected in the context of real-world use of an SI system. |
| Logic model | Hypothesized description of the connection between organizational inputs (resources and constraints), activities (actions or tasks potentially affected by SI deployment and usage), and possible short- and long-term outcomes that might be affected (adapted from [10]). |
| Matching | A quasi-experimental statistical technique that compares outcomes by matching treated participants to non-treated participants with similar observable characteristics. |
| Measurement construct [2] | An abstract concept not directly measurable (e.g., "productivity"). Sometimes referred to as a "background concept" [11]: the "broad constellation of meanings and understandings associated with [the] concept [of interest]" [12]. In uplift evaluations, the primary measurement construct is usually a specific outcome. |
| Measurement criterion [2] | A directly measurable or observable concept (e.g., "number of customer service issues resolved per week"). Sometimes referred to as a "systematized concept" [11]: the "specific formulation of the concept[, which] commonly involves an explicit definition" [12]. |
| Mixed methods study [of SI-driven uplift] | A study involving methods, principles, or rules for assessing SI-driven uplift based on both numerical and non-numerical data. Combines elements of quantitative and qualitative studies. |
| Natural experiment | A type of observational study in which the treatment condition (SI access) is assigned by factors outside the control of the evaluators but in a manner that emulates random assignment. |
| Observational study | A study in which the assignment of the treatment condition (SI access) is not controlled by the evaluators. |
| Outcome | A defined consequence to, impact on, or event affecting individuals or stakeholders that is plausibly attributable to an organizational program or process. For uplift evaluations, the outcome is expected to be affected by the integration of an SI system into the program or process. |

| | |
|---|---|
| Outcome measure | A measurable metric or indicator (measurement criterion) associated with an outcome. A given outcome may have multiple possible measures. |
| Proxy study [of SI-driven uplift] | Study in which data is collected in a controlled environment similar but not identical to the context of real-world use of an SI system (see [13]). Also referred to as a "lab" study or extractive study. |
| Qualitative study [of SI-driven uplift] | The use of a set of methods, principles, or rules for assessing SI-driven uplift based on non-numerical data (adapted from [14]). |
| Quantitative study [of SI-driven uplift] | The use of a set of methods, principles, or rules for estimating SI-driven uplift based on numerical data (adapted from [14]). |
| Randomized controlled trial (RCT) | An experiment in which the evaluators randomly assign one or more comparison (no access to the new SI system) and treatment (access to the new SI system) conditions to a random subset of participants or to a random subset of tasks. |
| Red teaming [15] | Type of expert probing in which experts adopt adversarial methods to identify vulnerabilities in systems. |
| Regression discontinuity | A quasi-experimental statistical technique that compares outcomes for participants close to either side of a sharp threshold or cutoff in treatment (e.g., SI access). |
| Spillover effect | Spillover effects occur when the treatment condition (access to an SI system) affects participants not assigned to it. For example, developers in the comparison condition might build on SI-assisted code from colleagues in the treatment condition. |
| Task [2] | Particular specification of a problem or objective, including desired solution(s), solution settings, or grading/verification procedures [6]. |

| | |
|---|---|
| Thematic analysis | Thematic analysis is a method for identifying, analyzing, and reporting patterns (themes) within text or other non-quantitative data [16]. Themes can be identified inductively ("bottom up") by analyzing data to surface themes or deductively ("top down") by applying theory to make sense of data. Inductive approaches surface themes that are more strongly tied to the data themselves rather than to an existing lens for understanding the data, capturing rich descriptions of the entire dataset [16]. Deductive approaches are more heavily shaped by an existing way of understanding (e.g., a theory or codebook) and can provide a more focused lens to analyze a specific aspect or part of the data [16]. |
| Threat model | Hypothesized description of the pathways by which SI usage may enable misuse. There are many approaches to threat modeling, but at minimum, the threat model should identify specific undesirable outcomes (threat events), the sources of those threats—from adversaries with particular motivations and capabilities, or from accidents, structures, or environments—and attack vectors, approaches to achieve undesirable outcomes and parameters required for success [17–19]. |
| Treatment condition | A condition in which an SI system under evaluation is used (as opposed to a comparison condition in which it is not). May range from human use of an SI assistant to full automation. Sometimes referred to as the exposure or intervention. |
| Treatment intensity | Degree or dosage of SI access or use—e.g., length of access to an SI system, tokens or budget used, breadth and depth of integration of SI systems with existing tasks, etc. |
| Uplift hypothesis | The outcomes expected to be impacted by the deployment of an SI system and the processes through which this impact is expected to occur. Includes the outcomes hypothesized to be impacted and the direction of the effect (the "what"; e.g., an increase in productivity), the subjects hypothesized to be impacted (the "who"; e.g., a person, a team, an organization), and the mechanisms through which this is expected to occur (the "how"; e.g., time saved on a particular task). |

1