The Evidence Deficit

What makes useful evidence difficult to obtain, and how to test whether AI-assisted work changes the constraint

This page is intentionally left blank.

Contents

	Summary	5
01	Define the deficit before explaining it	6
02	Formal demand does not settle the barrier	7
03	Reporting can be useful without answering attribution	9
04	Where assisted work might help	10
05	Count the whole workflow	11
06	What does not follow from lower reading cost	12
07	A scoped test and disclosure	13
	Further reading	16

© 2026 Prova LLC · All rights reserved · PP-000245-10

The Evidence Deficit

Useful evidence is often expensive, late, hard to reach or a poor match for the job it is meant to do. AI-assisted reading may make some tasks cheaper. But a cheaper draft does not show that responsible evaluation has become cheaper, because the real test covers much more than reading. It covers the records, the comparison or reasoning about causes, checking, protections for participants, staff work and whether anyone actually uses the result.

The argument of this report is a hypothesis about cost. In some kinds of work, assisted reading might lower the full cost of evidence that is fit for its purpose, and that is worth testing properly. The report does not estimate how common weak evaluation is across social programs. It does not name one constraint that holds back the whole sector, and it does not report any cost reduction that Prova has demonstrated.

© 2026 Prova LLC · All rights reserved · PP-000245-10

Summary

A program can count its services accurately and still not know what effect it had. It may also have good accounts of experience, outcome records or strong research, and need no new experiment. The missing evidence is whatever a particular question needs. It is not a low rung on one ladder for everyone.

Cost can stop worthwhile inquiry. So can outcomes nobody can observe, poor definitions, sources out of reach, legal or ethical limits, weak records of who received what, incentives, limited capacity and a timetable too short for follow-up. Several can be true at once. To claim that skilled reading is the main constraint across the field, you would need evidence that sets it apart from the others.

This report tests a narrower idea. In some work, help with reading or extraction may lower the total resources needed for evidence that fits its purpose, can be checked and meets an acceptable standard. That needs testing in the actual work and across its real range of tasks. A person in the loop, fluent text or existing records do not prove it.

© 2026 Prova LLC · All rights reserved · PP-000245-10

01 Define the deficit before explaining it

Start with the job the evidence has to do. A question about delivery needs defined units, good-quality sources and known coverage. A question about how outcomes changed needs suitable observations that can be compared. An estimate of effect needs a credible way of judging causes, with its assumptions stated. An explanation of causes may instead look at how change happened and at competing explanations through careful qualitative inquiry. A recommendation to expand also needs evidence about whether the result would transfer, along with costs, harms, alternatives, duties and who has authority.

Description, association, qualitative inquiry, experiments and quasi-experiments cannot be put in a single order that grades how good an organization is. The designs and the jobs they do overlap in different ways. A correct count of services is not inadequate just because no experiment lies behind it, and an experiment can still fail because of how outcomes were measured or because follow-up was missing.

To show that a real deficit exists, name the question, the people and the period, the record you have, the grounds you are missing that matter most and the consequence of the gap. A lack of local estimates of cause does not show that no relevant evidence exists. And an outcome table does not show that its reader has the evidence needed for the decision in front of them.

This report gives no count for the whole field and no consistent assessment of quality. Claims about “most programs,” trillions in social spending, billions spent on evaluation or the share of programs evaluated cannot be drawn from the examples here. Each would need its own clearly defined inventory, definitions and review of sources.

© 2026 Prova LLC · All rights reserved · PP-000245-10

02 Formal demand does not settle the barrier

U.S. federal policy on performance and evidence includes the Government Performance and Results Act of 1993 and the Foundations for Evidence-Based Policymaking Act of 2018, which was signed into law on 14 January 2019. The Program Assessment Rating Tool was an administrative assessment run by the Office of Management and Budget (OMB), and not a third law. These facts show what formal policy arrangements exist. They do not measure how much each institution wants to hear an unfavorable answer, whether it has enough capacity or why a particular study was never done.

GAO's 2003 work on evaluation capacity looked at selected agencies and pointed to the culture around evaluation and to the capacity to work together. Because it studied particular cases, it does not estimate how common these problems are today in philanthropy or across government programs. It does give a reason to look at how institutions are organized, as well as at money and methods.

© 2026 Prova LLC · All rights reserved · PP-000245-10

Instead of ruling out other explanations by assertion, ask questions that can tell them apart:

Possible barrier	Evidence that would clarify it	What an apparent improvement does not establish
The work for staff of reading and extracting information	Records of time and errors for the same task, including preparation and checking	That a quicker draft lowered the total cost of doing the work responsibly
Gaps in outcomes or sources	Coverage, definitions, permission and the actual records	That more fluent interpretation produced an outcome nobody observed
Weak grounds for claims about causes	How people were assigned, the comparison, the assumptions and competing explanations	That a similar group or more data fields identified an effect
A mismatch between timing and use	The real deadline, when outcomes will be complete, delays in sources, review and who has authority	That a late report has no value for learning or accountability
Incentives or governance	Who commissions findings, sees them, challenges them, reports them and may act on them	That publication or a label of independence removes every conflict
Resources and burden	Effort by staff, participants and partners, fees for sources and realistic alternatives	That existing records or salaried staff make the work free

These are proposed questions for diagnosis. This report has not tested which barrier matters most in a real set of programs.

© 2026 Prova LLC · All rights reserved · PP-000245-10

03 Reporting can be useful without answering attribution

Counts of services, measures of conditions, participants' accounts and investigations of delivery can all serve accountability and learning. Their value depends on what they actually show. A descriptive report is not just a second-best product because it has no estimate of effect.

The problem comes when a sentence asks its evidence to carry something else. Examples include presenting visits as if they were unique people, the experience of respondents as everyone's outcome, an observed gain as a causal effect, a measure of conditions as an instruction to decide one way, or a general research finding as a guarantee for one place. Correct each mismatch specifically. Do not conclude that funders know the statement is wrong, or that reporting exists only to look accountable.

A useful account of the present can sit alongside a plan for stronger inquiry. It may reveal a failure of access, an important experience or a gap in outcomes before anyone can work out what the program caused. In the other direction, a result about causes can be favorable and worth stating firmly within its scope. Caution should not wipe out evidence that has earned a conclusion.

© 2026 Prova LLC · All rights reserved · PP-000245-10

04 Where assisted work might help

Searching, preparing documents, extracting information, organizing it and checking it can each include tasks that are suited to assistance. How useful the assistance is depends on the quality of the sources, how the model or tool behaves, the instructions for the task, the interface, review by experts and what happens when mistakes are made. The same setup can work well on one kind of document and badly on another.

The research on people working with AI that is cited in **Human Review Has to Be Tested** separates different points of comparison and reports that results vary from task to task. It does not show how well or how cheaply Prova's own work is done. The survey of knowledge workers discussed there is about what people report doing and about patterns in their answers. It does not show that using AI causes a loss of reasoning skill. Those limits still apply to any claim about cost.

Human review is a safeguard that itself needs to be tested. The reviewer needs the sources in front of them while checking, the right expertise, realistic time, the authority to reject an output or to widen a check, and a record that others can inspect. An assisted system's explanation cannot replace the original evidence, and a label saying a result was approved does not show that the review catches the errors that matter.

Administrative records can cut some new data collection, but they bring other costs, such as permissions, cleaning, matching, changes in sources, gaps in coverage and interpretation. A case-management record is not a complete, continuous record of a participant's life. Selected case notes can inform qualitative inquiry, but summarizing them automatically does not turn them into an estimate of how common something is or into an effect.

© 2026 Prova LLC · All rights reserved · PP-000245-10

05 Count the whole workflow

The people, places, organizations and figures in this document's examples are fictional.

Example. This example shows how a saving in extraction time alone can mislead. Both ways of working handle the same task of 100 records. The table assumes that both have the same scope and both meet the same quality standards, checked independently. It does not show that either one actually meets them.

Work and cost basis	Manual model	Assisted model
Preparation and access, at $100 per hour	10 hours; $1,000	30 hours; $3,000
Extraction, at $100 per hour	60 hours; $6,000	20 hours; $2,000
Checking and rework, at $150 per hour	20 hours; $3,000	40 hours; $6,000
Additional tool cost	$0	$500
Total	$10,000	$11,500

Extraction hours fall by two-thirds, but the total modeled cost rises by $1,500, or 15 percent. That does not show that assistance never pays. It shows why it matters what a cost claim covers. A different way of preparing the work, a different pattern of errors or using the work again could change the comparison. Those are ideas to test, including any costs that are shifted onto participants or partners.

Keep apart hours of work, time from start to finish, money spent, the full economic cost of resources and the staff time that is freed up. A task done faster may still have to wait for access to sources or for follow-up. A saving on paper may not free up any cash, and moving staff to other work is not the same as lowering total cost. If costs are spread across several pieces of work, state the volume and the assumptions about reuse, and do not count reuse that may happen in future as a saving already made.

© 2026 Prova LLC · All rights reserved · PP-000245-10

06 What does not follow from lower reading cost

Assistance cannot observe what would have happened to the same person under other conditions. It cannot create lawful access to data, fix every missing record or justify a way of collecting data that causes harm. It may help with examining existing evidence, but any claim about causes still needs its own assumptions and statement of uncertainty. Qualitative inquiry into causes can test how change happened and what else might explain it, without automatically estimating an average effect.

Methods, standards of measurement and professional practice change over time. The principle that holds up is to match each specific claim to the grounds it has now. It is not that there are four fixed grades of evidence, or that all standards have stayed the same since the 1970s. A larger or cleaner dataset can improve precision, coverage or a relevant test, but on its own it cannot remove the problem of other factors explaining a result.

Accountability belongs to real people and institutions with the right expertise and authority. Adding a human reviewer does not guarantee it. Protecting participants, dealing with contrary evidence, disclosure, review and realistic action are all costs and responsibilities that come with the method.

Evidence that arrives after one decision may still matter for later choices, for learning, for public explanation or for accountability. Whether it is useful now depends on its purpose and timing. Do not treat every late result as a historical record with no value, and do not claim that excuses about cost will run out on some predicted timetable.

© 2026 Prova LLC · All rights reserved · PP-000245-10

07 A scoped test and disclosure

To test the hypothesis, choose a defined way of working and a set of tasks that is representative of it. Before judging the results, state what you are comparing against, the types of source, the quality required, which errors would matter and the rule for accepting an output. Keep the cases used to develop the method separate from the cases held back for testing. If you add known errors to see whether they are caught, record the mix you used, because that set does not tell you how often errors happen in real work.

Record the time and cost of preparation, access, production, checking, rework, review and reporting, including the burden on partners and participants. Have competent people review the reference answers, and keep a record of disagreements. Compare quality and cost together, and state the uncertainty and how far the results might transfer. A small pilot gives evidence about the conditions it tested. It does not cover every program, and it does not give lasting permission to expand.

© 2026 Prova LLC · All rights reserved · PP-000245-10

A completed working record names:

What to record	Your entry
Question and intended use	
Tasks, sources and permissions	
What is being compared, and the setup	
Quality standard and the consequences of errors	
The point of view on cost, and every stage	
Reviewers and disagreements	
Results and uncertainty	
Next action, owner, budget, deadline and authority	
Review trigger	

If the proposed saving depends on access that has not been secured or on less careful checking, that dependency stays visible in the record.

Prova has a commercial interest in this hypothesis about cost. This report proposes work to test it. It does not show that Prova currently has policies against bidding, that it is independent, how its fees create incentives, how well its services perform or that any cost reductions have been achieved. Each of those would need its own checked records. Review Prova's claims against the evidence in the same way you would review a program's claims.

© 2026 Prova LLC · All rights reserved · PP-000245-10

This report does not yet show that the proposed way of working succeeds.

RELATED: Human Review Has to Be Tested, for how to test human review; A Systematic Review Is a Study, for how a review's choices shape its conclusion. Both point to their primary sources.

© 2026 Prova LLC · All rights reserved · PP-000245-10

Further reading

1 **World Bank and IDB, Impact Evaluation in Practice (2016)**. See the passages on counterfactuals, assignment and missing outcomes.

2 **Spencer and colleagues, Quality in Qualitative Evaluation**, and its appraisal questions, and Magenta Book Annex A, sections A1.3 to 4, on the grounds for qualitative claims and on other explanations.

3 **GovInfo, GPRA statutory history**, short-title note; Evidence Act, Public Law 115-435, title, enactment and Title I; GAO's PART overview.

4 **GAO, evaluation culture and capacity, GAO-03-454**. See the official overview and its historical discussion of selected agencies.

QUESTION IT

Found an error, or want to question this document? Contact Prova at hello@byprova.com.

© 2026 Prova LLC · All rights reserved

Cite as: Prova (2026). The Evidence Deficit. byprova.com

Examples are fictional unless sourced. Not legal or financial advice.