

Deployment Questions for AI-Adapted Workflows

A set of questions drawn from three live AI deployments at IRC

This document draws on lessons from three AI pilots at IRC. The questions that follow are meant to add to, not repeat, what other guidance already covers, and are detailed here to help with the decisions that come up as an AI-adapted workflow moves from an idea to design to implementation and governance.

Pilot Descriptions

- Business Development: AI tools were introduced to support tasks like RFP analysis and proposal quality checks, where staff capacity was stretched thin by the volume of funding opportunities to review
- PEERS: AI support for partnership due diligence was introduced because the process took roughly a month per partnership
- Primary Health Care Data Visualization: AI-generated visualizations were tested in Burkina Faso and Kenya because health data existed but wasn't being turned into insight teams could act on

Infrastructure

Will outputs need to function offline or be shared with people who don't have direct tool access, and is the workflow designed for that from the start?

This question exists because access in humanitarian settings is rarely universal or stable. In the PHC pilot, dashboards were deliberately built to work as offline HTML files and Excel workbooks, which meant they could be shared with colleagues who had no LLM license of their own. This was a design choice that in feedback came out as critical for delivery. If a workflow assumes everyone who needs the output has live, ongoing access to the tool that produced it, it will exclude exactly the people, often frontline staff who need it most.

Task Fit and Scope

Do we agree on what good output looks like, not in theory, but among the specific people who will review it? And has the organization defined what “good” looks like for this task independent of AI?

This addresses two potential issues. The first is about hidden disagreement: in the PEERS pilot, a compliance officer and a partnership specialist could each say they knew what good due diligence looked like, but brought genuinely different standards to the same output. The second issue is related to evaluation of outputs. In BD, every interview participant independently identified the absence of shared quality standards as a significant gap. This creates a ceiling on AI performance that no amount of engineering can fix because the tool cannot encode a standard that doesn't exist. Before deploying AI to do this work, ask whether your team/organization has actually agreed on what quality looks like, with or without AI in the picture.

Has the team agreed to a minimum scope before building, and treated the ease of adding more as a risk to manage, not a free benefit?

Because AI makes it (relatively) cheap to generate outputs, scope creep can happen easily. For example, with PHC, teams could accumulate indicators, sections, or outputs, all easy to produce but perhaps disconnected from actual decision the team needs to make. The PHC pilot avoided this by capping its dashboard at 15 core indicators agreed with country leadership up front, and explicitly treated later requests to expand as something to hold within that agreed scope rather than default to yes. The evaluation's recommendation was direct: "treat ease of adding analysis as a risk to manage, not a benefit." Without that discipline set before building starts, the tool's flexibility becomes a liability rather than a strength.

Is the tool anchored to an existing artifact staff already trust (a template, workbook, or document), rather than asking staff to evaluate output in the abstract?

The BD pilot's clearest design lesson was anchoring prompts to the Capture Workbook, a tool staff already used and trusted, rather than introducing a new format alongside a new technology at the same time. The evaluation described this as bounding the scope, giving outputs a clear destination, and making quality evaluation intuitive. It named it as a key reason this pilot moved faster and generated more credible outputs than comparable initiatives; the report stating "rather than redesigning workflows at the same time as introducing new tools, any future initiative should identify existing artefacts and processes staff already trust and build from those." Asking staff to judge whether an output is good in the abstract is a much harder task than asking them to judge whether it correctly filled in a document they already know how to evaluate.

Values Alignment

Have domain experts reviewed the prompts/workflow for values alignment, not just factual accuracy?

PEERS is the clearest cautionary case in the entire body of evidence here. The original prompts produced outputs that were factually reasonable and looked authoritative to non-expert users, but were, in the evaluation's own words, "technically plausible but operationally wrong." They used compliance-first framing that conflicted with IRC's partnership philosophy, referenced only four of the six PEERS principles, used non-PEERS terminology ("capacity building" instead of "capacity sharing"), and generated go/no-go recommendations at the wrong point in the decision process. None of this was caught by accuracy review, because none of it was factually incorrect. It required someone with deep domain expertise to recognize the framing itself was misaligned. Accuracy review and values review are different and both uniquely valuable checks

Task Risk and Trust

Is AI reducing existing material or generating new substance, and does verification reflect that difference?

Across every pilot, this distinction sorted reliable outcomes from risky ones. BD found outputs were "very accurate" for tightly scoped, reduction-style tasks like RFP analysis and proposal quality checking, while "the primary risk lies when AI is used as a generation tool, where hallucination and context management risks increase." PEERS showed the same pattern from a different angle: outputs were consistently accurate for tightly scoped analytical tasks, but quality declined specifically when the tool had to generate risk flags or recommendations requiring tacit knowledge from direct partner engagement rather than what was in the documentation.

Does trust in this tool rest on visible provenance, routine independent verification, and open iteration, or is it being asked for on faith?

The PHC evaluation named the specific mechanics of how trust actually got built, rather than just asserting that it existed. Respondents pointed to three things directly: provenance (inputs came from health information systems they already stood behind), verification (technical staff routinely cross-checked outputs against their own manual calculations and external benchmarks, not as a one-off but as a routine step), and iteration (confidence rose specifically because errors were found and corrected in the open. One dashboard builder reported 90-95% confidence after the dashboard was revised three to four times in direct response to review-meeting feedback). Trust wasn't built by the first version being good. It was built by people watching a flawed version get corrected in front of them, repeatedly. If none of these three mechanisms are present, any stated trust in the tool is closer to an assumption than an earned outcome.

Verification

Does the tool make verification easier, or does it just produce outputs and leave verification as the user's problem?

This is a question about the tool's design, not just the organization's process around it. BD's strongest-performing outputs came specifically from two practices: loading relevant documents (as context) before querying, and specifying structured, templated output formats, both of which made it easier for a human to check the result against something concrete. A tool can technically produce correct answers most of the time and still be poorly designed for verification if checking its work requires more effort than doing the task manually would have.

Has the person building this workflow done the underlying work themselves, and if not, is domain staff testing planned in from the start, not added later?

In all three pilots, the people building the tools weren't the ones who actually did partnership assessments, wrote proposals, or ran health data reviews day to day, so they had no reliable way to judge whether an output was actually good. That judgment could only come from staff who did the work themselves. In every case, testing by those domain staff was built in from the start. A tech-led build can produce a working tool. It cannot, on its own, tell you whether that tool is actually good at the underlying work, only someone who has done that work can.

Is our human-in-the-loop practice a designed process with named steps, or just an intention?

"A human reviews the output" has to be turned into an actual process, not left as a goal or an intention. A designed process names who reviews what, against what standard, and what happens when they flag a problem; if you can't answer those three things concretely, the oversight you believe exists is aspirational rather than real. In the projects that built this in properly, there was an explicit step where staff verified outputs and fed that verification back into the next stage before moving forward, rather than treating review as a one-time check at the end.

Governance

Is there a single person with both domain knowledge and decision authority to close the loop?

BD and PEERS are a direct contrast on this point. In BD, one specialist could take a prompt, test it, refine it, and push it out; the loop closed with one person who had both the expertise to judge the output and the standing to act on that judgment. In PEERS, outputs had to pass through multiple specialists with different professional orientations before anyone felt comfortable moving forward, and the evaluation noted this meant the loop was "really difficult to close." Every output became a fresh site of contestation rather than a starting point for shared work, because staff were debating not just the analysis but whether to trust the system that produced it. Diverse input is valuable in design. Without a named person or clear mechanism to actually decide, that same diversity becomes the reason nothing moves.

Does this workflow cross multiple functions who might each want it weighted toward their own priorities (e.g., compliance over a more holistic approach), and if so, is there a named decision-maker or process that keeps the final call anchored to your values?

This lesson played out with the PEERS project, a highly cross-functional team, all with different focus areas and functions. In testing, the original outputs weighted IRC's fiduciary and compliance interests first, in a process that's meant to be collaborative with partners, not compliance-led. The model chose one function's orientation over another's without anyone deciding that should be the case. When a workflow touches more than one function, each brings its own legitimate priorities, and without an explicit mechanism for resolving whose priority wins in ambiguous cases, the workflow will default to whichever function's assumptions are most clearly stated.

Is the person who designed the workflow the same person who will use it, and if not, is there a deliberate handoff?

Judgment about when to trust an output and when to question it typically lives in the head of whoever built the workflow, not in the written instructions they leave behind. This matters because a prompt library, a skill, a plug-in, or an adapted workflow along don't transfer judgment. someone who didn't build the workflow can follow its steps and catch obvious errors, but has no reliable basis for recognizing when an output is subtly wrong in a way that sounds reasonable but doesn't reflect how the organization actually thinks about the decision.

Measurement

Is success being measured only by time saved, or also by work that becomes possible that wasn't before? The latter may be the larger and more” invisible” source of value.

All three pilots found their most significant impact wasn't visible in standard time-savings metrics. PEERS users ran due diligence on 2-4 partners simultaneously instead of one at a time, and 80% confirmed they took on work that would previously have required more people or been beyond their capacity entirely. BD's evaluation was even more concrete: a country-level donor prioritization analysis that previously required a week of staff or consultant time became producible in a couple of hours. PHC's dashboards went further still. They surfaced health issues teams said they would not otherwise have seen at all, from a rise in adolescent STI cases in Burkina Faso to referral delays and immunization gaps in Kenya, several of which led directly to programmatic action. If an evaluation only tracks hours saved on tasks that were already happening, it will miss both the work AI made newly possible and the problems it surfaced that no one was measuring for in the first place.

Adoption and Scale

Is the organizational moment stable enough to absorb a pilot? Are there imminent disruptions that would undermine activation? License activation failures are an early signal.

Kenya is the direct cautionary case here. Indicator selection and data sharing, essential to the development of the dashboards were delayed roughly six weeks because of bandwidth issues for staff and aid cut uncertainty leading to unclear job security. A pilot requires people to activate licenses, attend trainings, and give feedback, on top of existing workloads, none of which happens when staff are managing something more urgent. If a team can't complete a simple activation step within a couple of weeks, that's a readiness signal worth taking seriously before investing further, not an inconvenience to push past.

Have you planned for adoption outpacing your original use case? When adoption spreads beyond original scope, which directions are acceptable and which require guardrails that don't yet exist?

Both PHC and BD show this happening in real time, before either pilot had even closed. In PHC, country staff were already asking to extend the dashboard beyond its original scope, thinking about scale before the initial phase was finished. In BD, staff were already thinking about the next scope of work to take on before the current pilot wrapped. Neither team planned for this in advance; it simply arrived while the work was still underway. Successful deployment generates its own momentum. The question isn't whether use will spread beyond what was planned — it will — but whether the organization has decided in advance which directions of that spread are fine and which need guardrails before they become established practice by default.