

PHINEES PERSPECTIVES

On capturing expertise

Behavior or Cognition

Why recording what an expert does captures almost nothing of what they know — and what it takes to build a system that holds the rest.

By **Tim Tracy** · AI Strategy & Product Lead, Phinees

The gap

You can record everything an expert does and still capture almost nothing of what they know. This is not a technology problem, and more tooling does not fix it. It is a problem of aiming at the wrong layer.

Watch a skilled person work and what you can see is behavior — the steps they take, the choices they land on, the outputs they produce. What you cannot see is the reasoning underneath: why this step and not another, what they noticed that made them change course, the rule they applied without being aware they were applying it. **Behavior is the visible residue of cognition.** If you record only the residue, you get a system that can imitate the moves without the judgment that produced them. It will do what the expert did, in the cases the expert happened to face, and fail the moment the situation shifts — because it never had the reasoning that would have told it what to do instead.

Behavior is what an expert does. Cognition is how they decide. Most attempts to capture expertise record the first and assume they have captured the second.

The rest of this paper is about why capture defaults to behavior, and what it takes to aim at cognition instead — getting it out of the expert, writing it down without flattening it, and building it into a system that can actually hold it. Three stages, one distinction running through all of them.

Why the default is always behavior

There is a structural reason capture defaults to behavior: behavior is what is observable and askable, and cognition is neither.

Ask an expert how they do something and you get an account — considered, articulate, and largely disconnected from what actually happens in their head when they do it.

Consider a manager known for hiring well. Ask how they do it and the answer comes easily: they hire for culture fit, for grit, for people who ask good questions. Sincere, and almost certainly not the real rule. Watch the actual decisions and a different pattern emerges — they consistently cool on candidates who hesitate before answering a hard question, and warm to the ones who think out loud. That is the operative criterion, and it appears in none of their stated reasons. It may be a genuine signal of quality. It may also be a bias quietly screening out people who reflect before they speak. The manager cannot tell you which, because they do not know the rule is there. They are not concealing it. They reached for the most plausible account of their own judgment and handed it over believing it was true.

This is the part worth sitting with: the gap is not a blank the expert knows to flag. It is a confident, well-formed answer that happens to describe the wrong thing.

This is why interviews, handover documents, and process maps so reliably underdeliver. They are not failing at their job. They are succeeding at the wrong job — faithfully capturing the *what* while the *how* stays exactly where it was, locked inside the expert. The default methods don't merely miss cognition. They are structurally aimed away from it.

First: observe without contaminating

If you cannot get the reasoning by asking for it, you have to get it by watching it happen — the expert working a real case, thinking in real time, before the account-making instinct has a chance to tidy it up. And watching well is harder than it sounds, because the act of observing changes what you observe.

My own instinct for this comes from an unlikely place: years of work interviewing people on camera, in a documentary context. What that teaches, more than anything technical, is when to insert yourself and when to sit back and let the process unfold. The moment you lead — the moment your question telegraphs the answer you expect — the subject gives you the answer they think you want, and the real thing retreats. The most valuable material almost always comes from staying out of the way: asking little, interrupting less, and letting the person reveal how they actually think rather than how they would explain it if asked.

The skill of capturing how someone thinks, rather than what they say they do, is not mainly technical. It is knowing when to stay quiet.

Applied to expertise capture, the discipline is the same. You are not there to extract a description. You are there to observe reasoning in motion and to avoid contaminating it with your own expectations. What you want is the expert's real decision logic, not the cleaned-up narrative they produce when they know they are being studied. That distinction — between the live process and the account of it — is the whole game at the capture stage, and an interviewer's restraint is what protects it.

Second: encode the reasoning, not the action

I love cooking, and I especially love smoking brisket. For years I knew mine was done at twelve hours — that was the number, and the number worked. Then I moved to Mexico City, ran a brisket the same way, and at twelve hours it was still stiff. Nowhere close. What I hadn't accounted for is altitude: at seven thousand feet the air is thinner and drier, so the fire burns cooler and the meat sheds moisture faster, which drags out the long plateau every pitmaster knows where the temperature just sits and refuses to climb. Same brisket, same method, and now it takes closer to sixteen hours to land where it should.

Here is the thing, though. Twelve hours was never actually the rule. A brisket is never done because the clock says so — it is done when it feels done, when the probe slides in without resistance and the meat gives the way it is supposed to. The twelve hours was just where that feeling happened to fall, in my kitchen, at sea level. I had mistaken the output of my judgment for the judgment itself. The altitude didn't break my rule so much as expose that the number was never the rule to begin with — and the same lower-pressure effect that pushed my brisket from twelve hours to sixteen is the one that makes a cake collapse or beans never soften at altitude. One underlying reason, surfacing wherever someone was relying on a number instead of knowing what they were really looking for.

Captured expertise comes in these same two forms, and the difference decides everything. You can write down what the expert did — smoke it for twelve hours — or you can write down the question they were really answering: is it tender yet? The hours, the thermometer, the color of the bark are just evidence toward that question. Record the evidence and you have preserved one answer, good only in the kitchen it came from. Record the question and you have preserved the thing that finds the answer in a kitchen you have never cooked in.

It isn't only cooks. Ask a seasoned loan officer why they declined an application and they might point to the debt-to-income ratio: over forty-three percent, so it failed. But watch them work and the forty-three percent turns out to be a stand-in for the question they are really asking — will this income still be here in two years? A borrower who clears the ratio on a single client of four months gets declined; one who sits just over the line on a thirty-year pension gets approved. The threshold was never the rule. It was a proxy for a judgment about durability, reliable right up until a case arrived where the number and the judgment pointed in opposite directions. Record the threshold and a system will decline the pension, confidently and wrongly. Record the question, and it does what the loan officer does — it reasons about a case the number never anticipated.

This is why expertise so rarely survives being written down. The cases that reach an expert are the ones that do not fit the standard pattern — that is what makes them an expert's cases — and a record of what the expert *did* meets each of those cases with the answer to a slightly different question. A record of how the expert *reasons* works the new case the way they would have. Capture the action and you have preserved what the expert decided. Capture the reasoning and you have preserved the expert.

Third: store patterns in layers, not facts in a pile

Watch how a person actually thinks through a live situation and you notice they are never working from a single pool of information. Several things run at once. There is what is in front of them right now — the immediate thread, where it is going, where it just caught. There is everything they have absorbed over time about how this kind of situation tends to unfold — not a transcript of what was said before, but a feel for how it usually goes. And underneath both, quietly, there is the larger objective the whole exchange is in service of — the thing they are steering toward even when it is not what they are talking about in the moment.

These are not three separate acts of remembering. They are one mind holding three layers at the same time, and the skill is in how they interact: the larger objective shapes how the immediate moment is read, the accumulated sense of how things go tells you what the present exchange actually means. A system meant to replicate that kind of thinking has to be built the same way — with those layers distinct and held together — because that is how the thinking it is imitating is structured. This is an architectural decision, and it has to be made at the start. It is not a setting to tune later; it is a question about what you are building.

The three layers reach the model in three different ways, and that is the part most systems get wrong. The immediate layer is simply present — it lives in the active context window, the running state of the exchange, attended to directly because it is right there. Nothing has to be fetched. The accumulated layer is the opposite: it is not in the window, it has to be retrieved and brought in when it bears on the

moment. And the overarching layer is neither — it is held, pinned in context the whole time, because a standing objective that only surfaces when something seems relevant is not a standing objective at all.

Retrieval is where the real distinction lives, and it is why so much of this goes wrong. There are, in effect, two things a system can retrieve. It can retrieve facts — this topic came up, that word appeared, here is what was said before — which is what most retrieval is built to do, and which produces the hollow, averaged answer everyone recognizes. Or it can retrieve patterns — the recurring shape of how this person reasons, what reliably triggers a given move, what has worked before. The same mechanism, pointed at a different kind of stored thing. One feeds a system that recites. The other feeds a system that reasons. The accumulated layer is only worth having if it retrieves the second.

Active listening — the quality everyone is chasing in conversational systems — comes from holding the immediate move and the standing objective at once. Not from a larger pile of retrieved facts. From keeping the layers distinct and present together.

This is why collapsing everything into one store fails the way it does. A flat system runs all three through a single mechanism — retrieve whatever seems relevant — and that one mechanism suits none of them. It cannot hold an objective steady, because the objective only surfaces when it happens to look relevant. It cannot reason from patterns, because it is storing facts. It treats the present moment, the accumulated sense, and the standing goal as the same kind of thing, fetched the same way. The robotic answer is the sound of that collapse: not too little memory, but three kinds of memory forced through one channel that serves none of them well.

What the right structure is, exactly, is not settled — this is a live frontier, and it is moving fast. The mechanisms for holding and retrieving context are changing month to month, and what counts as good architecture today will look primitive before long. That is not a reason to wait. It is a reason to build on the clearest understanding available now and to stay close enough to the work to move as it moves. The distinction underneath — between a system that recites and one that reasons — will outlast any particular way of achieving it. Watching how we get there is going to be the interesting part.

The through-line

Three requirements, and one idea underneath all of them. Observe without contaminating, so you capture cognition and not a tidy account of behavior. Encode the reasoning and not the action, so what you capture generalizes. Store patterns in layers and not facts in a pile, so the system can think rather than merely respond. At every stage — capture, encoding, architecture — the same distinction is at stake: behavior or cognition, the visible residue or the reasoning that produced it.

Most efforts to capture expertise collapse that distinction somewhere along the way — they ask instead of observe, or record the action instead of the reasoning, or pile facts where they needed layered patterns. Wherever the collapse happens, that is where the expertise leaks out, and what remains is a system that does the moves without knowing why. Keeping behavior and cognition distinct, at every stage, is the whole discipline.

About Phinees

Phinees is an AI consultancy based in Mexico City. We capture the critical tacit knowledge of an organization and convert it into AI-based systems — working to capture cognition rather than behavior, reasoning rather than action. ***What your organization knows, stays.***

To start a conversation: diegolozano@phineesapp.com