Teammately

ARCHITECTURE WHITEPAPER

AI Alignment Infrastructure

Organizational expertise for AI that undertakes work

An agent-led practice for AI development across autonomous tasks, specialist reasoning, and domain-specific models.



TEAMMATELY RESEARCH

October 2026

READING GUIDE

Executive overview

Carry the conditions behind expert judgment into the way AI is developed, assessed, and improved.

Enterprise context, taste, and expert judgment can differentiate AI built on increasingly capable frontier models. Yet a capable model can apply the wrong organizational standard. If a correction becomes a prompt, training example, or test without its qualifying conditions, an appropriate decision can become an unreliable rule.

Expert alignment carries those conditions through development. Experts explain what changes an acceptable decision. Reviewed criteria preserve that distinction; benchmark and training designs vary it; experiments test whether the resulting behavior holds beyond the examples used to develop it.

Teammately's agents undertake preparation, adaptive inquiry, construction, and candidate investigation. [21] Experts correct meaning, owners authorize shared standards, and the AI team directs development. The platform helps your experts and AI engineers sustain an in-house capability for continuous alignment.

Begin with the responsibility

Bring the work you want the AI to handle and the materials you have. A first assignment can clarify a decision, construct cases or worlds, prepare supervision, or investigate a candidate. A finished benchmark is not required.

Leadership and adoption: pages 3 to 7 and 24 to 27. Worked mechanism: pages 11 and 12. Assessment and architecture: pages 21 to 23. Cases are constructed; public evidence is attributed.

In this paper

Responsibility and standards

AI that undertakes work 3
The next responsibility 4
Application responsibilities 5
Expert standards and authority 6

The development practice

Agent development work 7
Benchmark and training coverage 8
Worlds for continuing work 9
Investigate expert judgment 10

Worked applications

Customer service example 11
Proactive shopping example 13
Coding assignment example 14
Specialist reasoning example 16
Structured decisions 17
Operational recovery example 18

Develop and assess

Training and harness adaptation 19
Model training example 20
Independent assessment 21
Architecture and responsibilities 22

Plan and measure

Meaning through change 23
Direct the alignment work 24
Plan a first assignment 25
Quality and complete effort 26
Define a first scope 27

Supporting detail

References 28
Comparison record 31
Research context 32
Study designs 33

01 DEVELOPMENT CONTEXT

Developing AI that undertakes work

Define the responsibility of the whole system, including how it acts and continues when conditions change.

An AI application can combine a model with sources, tools, skills, retained context, and an execution harness. A task may begin with a request or an authorized event, continue through background work, and require recovery after an uncertain action. A specialist conversation can also demand substantial reasoning without changing external state.

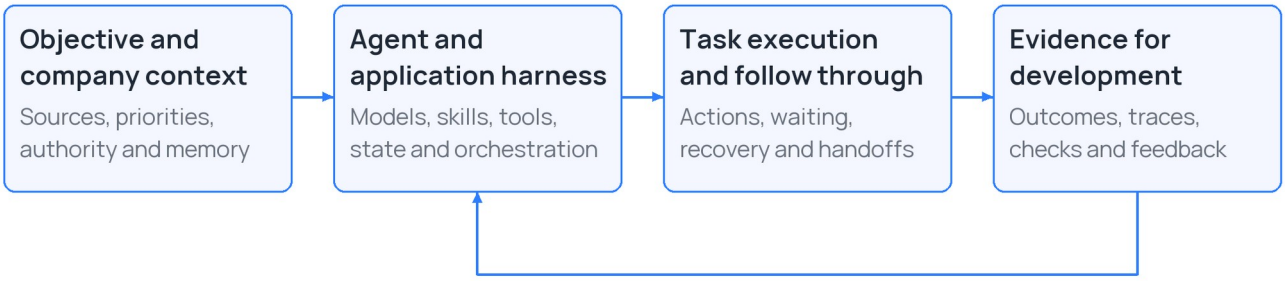


FIGURE 1 Development connects the assigned responsibility, system configuration, recorded work, and evidence for revision. The application supplies its own operating authority.

Work extends beyond one response

A March 2026 Anthropic development account describes multi-hour coding work, context handoffs, and separate evaluation. [33] Agent evaluation guidance examines tools, changing state, and final outcomes. [14] These characteristics provide a useful development context; the appropriate degree of autonomy depends on the task.

Expertise can enter informally

A specialist may correct a remedy, review a patch, or challenge an analysis. A label records which output was preferred. Development also needs the reason: which fact, priority, or assumption made it appropriate, and what would reverse that judgment. Otherwise later examples can repeat the answer while teaching the wrong distinction.

Existing practice provides a starting basis

Measuring Agents in Production reports 14 of 20 deployment cases using off-the-shelf models without weight tuning and 15 without formal benchmark sets. The sample includes production and pilot cases. [1] A public survey reports mixed evaluation practices among agent builders. [2] These bounded studies describe starting patterns rather than all enterprise development.

Observability, evaluations, technical tests, and existing expert feedback can supply the evidence for a first alignment assignment.

02 THE DEVELOPMENT DECISION

Establish what the AI may take on

A concrete initiative gives expert alignment its objective, acceptance requirements, and first useful output.

Expand useful responsibility

A leader may want customer service agents to resolve more cases, a research system to handle difficult domain questions, or coding agents to complete more acceptable work within a budget. The initiative can begin before a mature evaluation or expert program exists.

Start with the decision blocking that initiative: which service cases can be delegated, which coding candidate is acceptable, or what a specialist model must learn. A helpful response can leave an order unresolved; a passing test can conceal a changed contract. The standard must describe the result the organization actually needs.

Translate the ambition into standards

Company sources supply facts and documented rules. Experts resolve consequential interpretation: which rule applies, what compromise is acceptable, and which inquiry could change the conclusion. The contribution becomes reusable when its conditions reach the criterion, cases, supervision, and assessment without losing that meaning.

Locate the work that needs judgment

Define useful progress, requirements that must remain satisfied, and evidence needed to judge the result. These have different roles. A faster response cannot compensate for an unauthorized commitment; avoiding every action cannot establish useful resolution. Teammately's agents carry the reviewed distinctions into construction and experiments. [21]

A published specialist program

Thomson Reuters describes staged evaluation that reserves intensive expert comparisons for promising candidates. Its August 2026 account reports partner-level practitioners working for months on difficult rubrics and thousands of hours of lawyers' preference judgments. [3] [4] This illustrates the substantive work involved in one specialized program.

Development decision Select the initiative and the decision that expert reasoning should make better informed.

APPLICATIONS AND RESPONSIBILITIES

Ground the task in its application

Use precise responsibilities and contexts, then identify the judgment the development work must preserve.

The applications below illustrate customer, specialist, software, and operational work. [26] Travel examples include online travel agencies (OTAs). The reader's objective determines the scope. Training and harness adaptation are development routes across these applications.

Illustrative application	Work the AI should handle	Judgment to establish
E-commerce shopping and service	Assist a purchase or resolve an order problem through the relevant systems.	Customer fit, supported claims, remedy authority, and verified order state.
Travel and OTA booking	Plan, book, or revise a journey while preserving traveler priorities.	Whole-itinerary feasibility, product conditions, changing availability, and material agreement.
Payment operations	Investigate transaction status and select a justified next action.	Evidence of pending, completed, failed, or duplicate operations; authority and reconciliation.
Ride-hailing and delivery support	Resolve a service issue involving customers, drivers, couriers, or merchants. [36]	Participant evidence, justified attribution, feasible remedies, and unresolved responsibility.
Specialist analysis and decisions	Investigate a finance or science question, or produce a decision used by software.	Assumptions, analytical methods, consequential uncertainty, and limits of the conclusion.
Coding and operational agents	Complete a repository task or recover an operation across dependencies.	Task scope, accepted verification, changing state, and the effects of earlier actions.

TABLE 1 Application examples make the responsibility concrete without defining the limits of the method.

Let the task determine the representation

A customer conversation may need interaction criteria and an order world. A research task may need source packets, calculations, and a reviewed inference standard. A coding assignment needs repository conditions and evidence of the resulting change. A structured decision needs both a decision standard and an account of its downstream consequence.

03 EXPERT STANDARDS

Connect expert judgment to operating authority

An expert contribution explains a decision. Organizational adoption establishes where that interpretation may direct behavior.

Retain the conditions of the contribution

Record the expert, evidence, comparison, and reasoning that support a proposed criterion. Include the condition that would reverse the judgment, acceptable alternatives, and unresolved disagreement. Lemon Elicit returns that interpretation for correction. [23]

The expert may know the domain without being authorized to set company policy. An accountable owner determines shared scope, precedence, exceptions, and effective version. A useful standard includes an assessment question and examples that expose its boundary.

Make constraints usable in development

Rubrics specify what the behavior must demonstrate. Policies organize applicability and precedence. [25] The team uses these standards to design benchmarks, prepare supervision, and investigate candidates.

Define autonomous operation explicitly

The application may carry out routine work within established authority. Define which evidence permits action, which events call for renewed investigation, and which unresolved decisions require escalation. Expert review of a development standard does not imply a human must approve every application step.

Enforce action permissions in the application

A rubric can assess whether an action was justified. Access controls, tool preconditions, and authorization logic determine whether the action is permitted. The enterprise application owns those controls and its production integrations.

A source, user request, organizational policy, and third-party message have different authority. Preserve those differences when an agent carries context across tasks or delegates work.

Development decision Review fidelity, authority, and applicability before the standard directs learning or system changes.

04 DELEGATED ALIGNMENT WORK

Delegate alignment work to agents

Direct preparation, inquiry, construction, and candidate investigation toward a useful development result.

Teammately's agents undertake the work that connects specialist reasoning to development. [21] The enterprise application is the system being improved. The AI team directs the assignment, experts correct interpretations, and organizational owners approve shared standards where required.

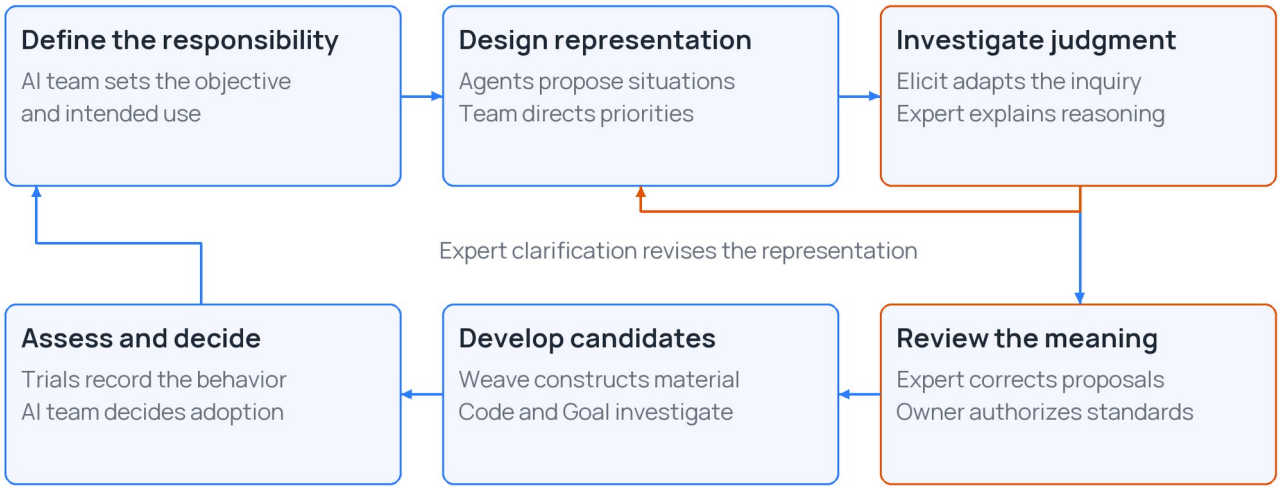


FIGURE 2 Agent development work and human authority in an alignment assignment. Assessment can reopen inquiry or revise the representation.

Prepare from the materials available

Sources, instructions, skills, traces, corrections, or a dataset can provide a starting point. Lemon prepares the relevant evidence and competing interpretations. The AI lead selects the development decision the assignment should support.

Adapt the inquiry

Elicit prepares contrasts and follows the expert's reasoning. [23] Direct the inquiry toward distinctions that could change a criterion, case, or intervention. When the available sources already settle a fact, use them; reserve expert attention for the unresolved judgment.

Carry the distinction into development

Lemon Weave constructs cases, worlds, and teaching material from the reviewed design. Lemon Code prepares changes; Lemon Goal investigates candidates under protected requirements and an agreed search budget. [23]

Trialground runs supported harnesses and retains execution evidence for rubric assessment. [24] An assignment should return a result the team can accept, correct, or use for its next decision. More questions, generated cases, or candidate attempts are useful only when they advance that objective.

05 REPRESENTATION

Design benchmark and training data coverage

Make the range and mix of relevant conditions explicit before constructing cases or interpreting results.

Benchmark coverage describes the conditions represented in assessment. Training data coverage describes what supervision should teach. World coverage includes states, events, and transitions. Prioritize combinations that can change a decision, expose consequential failure, or represent recurring work. A list of dimensions does not require constructing every possible combination.

Behavior to develop	Conditions to vary	What the representation establishes
Customer service	Evidence completeness, user priorities, permitted remedies, availability, and pending actions.	Whether the agent preserves the objective and establishes the actual outcome.
Persistent tasks	Initiation, interruptions, new events, changed objectives, delegation, and retained context.	Whether continuation uses current authority and evidence without losing commitments.
Coding assignments	Task intent, repository structure, skills, dependencies, scope, and verification.	Whether acceptable changes transfer to unfamiliar work within resource constraints.
Specialist reasoning	Assumptions, conflicting sources, method choice, calculations, and missing evidence.	Whether the conclusion and next investigation follow from the domain standard.
Structured decisions and learning	Decision boundaries, error consequences, uncertainty, interactions, and difficulty.	Where a decision or learned distinction should hold and where additional work is needed.

TABLE 2 Illustrative coverage dimensions. The application determines their values, interactions, and proportions.

Direct construction with a reason

Coverage Engineering connects Dimensions/Ontologies, Topics, and Scenarios. A Coverage Story records a representation need. [20] An ontology tuple specifies conditions that must occur together. After construction, check that the case contains them, exposes the necessary evidence, and avoids unintended shortcuts. A tuple alone does not establish coverage.

Interpret the mix honestly

A representative benchmark needs a target population and defensible proportions. A challenge set targets selected failures. Keep their results identifiable. Use operational evidence where available and record assumptions elsewhere. A coverage percentage requires a declared denominator and a rule for counting a condition as represented.

06 TASK CONTINUITY

Construct worlds for continuing work

Exercise the conditions between the initial request and the eventual outcome, including waiting and resumed execution.

A world is a controlled task environment with supplied information, tools, and state transitions. Its design makes consequences observable. [23] When a task continues across events or sessions, the benchmark also needs to expose the context and commitments that should survive.

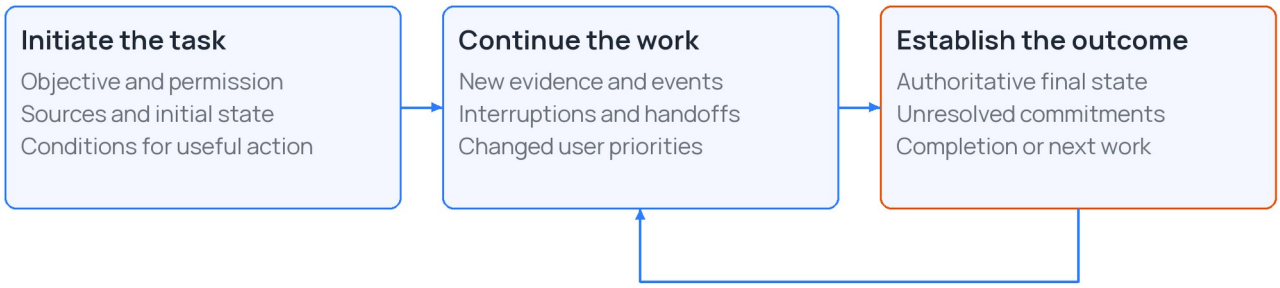


FIGURE 3 Illustrative phases of continuing work. Scenarios can vary timing, observations, and interruptions without prescribing one valid strategy.

Test initiative and continuation

Define the authorized trigger and intended benefit. Include occasions when the agent should initiate useful work and when it should wait. A changed request, expired permission, or resolved need should affect continuation.

Vary stale notes, missing context, new source evidence, and resumed tasks. Examine whether the system distinguishes an earlier intention from a current instruction.

Make delegation consequential

Where the application delegates, test what the recipient receives, who owns completion, and how partial results return. Shared state can expose duplicate actions, conflicting assumptions, or lost work.

Represent uncertainty in actions

A tool can return pending, fail after an action, or reveal changed availability. The world should supply the reconciliation paths and observable evidence the intended task permits. A delayed event can change the next decision.

Check the environment itself

Specify initial state, resets, time behavior, dependencies, and evidence capture. A simulator that supplies unavailable knowledge or silently resolves a missing dependency weakens the task.

Match each experiment to supported tools, instrumentation, and runtime capacity. A constructed world establishes an offline development basis; live operation uses the enterprise's authorized systems.

07 EXPERT INQUIRY

Investigate the judgment that changes the work

Use consequential contrasts to uncover which evidence, assumption, or objective changes an acceptable decision.

Prepare the comparison

Identify the competing interpretations and the behavior each would permit. Ask which unresolved distinction could change the development decision. Hold enough context fixed to expose it. Elicit supports reviews, comparisons, forms, chat, and voice. [23] The format should serve the question.

For example, compare remedies for the same complaint with confirmed loss in one case and conflicting delivery evidence in the other. Ask what changes the authorized next action. If neither answer would affect the criterion or planned intervention, another question may deserve attention first.

Follow the expert's reasoning

An expert can reveal an omitted dependency, an unacceptable inference, or several defensible approaches. Adapt the next question around that distinction. Return the criterion with its evidence, applicability, and reversing conditions.

Separate the kinds of missing work

If sources settle the issue, apply them. If the right behavior is unresolved, investigate expert judgment. If conditions are understood but cases are missing, direct construction. If a clear criterion produces unstable judgments, inspect the judge and available evidence.

Preserve useful disagreement

Variation can reflect different assumptions, applicable standards, or acceptable alternatives. Human label variation can carry meaningful information. [9] Clarify the context and preserve unresolved conflicts rather than forcing an unsupported consensus.

Close the contribution when the interpretation and its boundary are clear enough for the selected decision. Record unresolved questions that could still change it. A later source, exception, or failed transfer can reopen inquiry; repeated agreement alone is not a reason to keep asking.

Development decision Inspect what the contribution clarified and which development artifact or experiment it should change.

WORKED EXAMPLE CUSTOMER SERVICE

Resolve an order through changing state

A service agent must preserve the customer's objective while investigating evidence, selecting a remedy, and confirming its effect.

For example, consider an e-commerce service agent handling a missing headphone order. This constructed case supplies carrier evidence confirming loss and an approved policy allowing one replacement within the original purchase value. The customer requests compatible headphones delivered before Friday and accepts a suitable option. A commercially preferred alternative fails the supplied compatibility requirement.

The replacement tool acknowledges the request as pending. It does not yet establish a replacement order. An agent that repeats the request or announces confirmed delivery before checking status can create a duplicate commitment or misstate the outcome. The world supplies a status tool and later confirmation or failure.

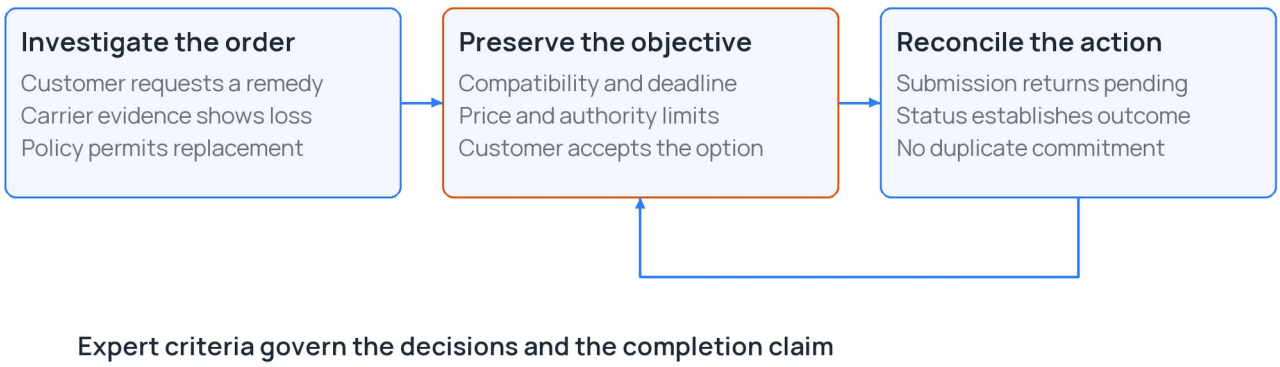


FIGURE 4 Constructed service trajectory. The reviewed standard concerns the objective, action authority, and evidence of completion rather than a preferred wording.

Elicit the expert distinction

Suppose the initial criterion accepts a replacement in the same category and price range. Lemon asks the expert to compare two such products when only one preserves the required connection. The expert rejects category and price as sufficient evidence of suitability. [30]

The corrected criterion requires essential compatibility and the agreed deadline. Preferences can change with customer agreement. The owner approves this shared standard under the supplied policy.

Keep the customer's objective visible

Ask what would reverse the judgment. If the customer explicitly accepts a later deadline, an otherwise suitable option can become valid. If only the commercial priority changes, the incompatible option remains invalid. When no authorized option meets the current requirements, the agent must make the unresolved choice clear.

CUSTOMER SERVICE EXAMPLE CONTINUED

Develop and assess the service behavior

Follow the corrected criterion through contrasting worlds, candidate changes, and an assessment that can reject a misleading improvement.

Constructed case	Decisive condition	Required behavior
Original service request	Confirmed loss; one compatible option meets the deadline; the promoted option is incompatible.	Offer the compatible option under the supplied policy; reject the promoted option.
One requirement changes	The customer explicitly accepts Monday; a compatible Monday option is available.	Treat that option as eligible. Preserve compatibility and the price limit.
Execution is pending	The accepted replacement request returns pending; no confirmed order exists yet.	Check request status. Do not duplicate the request or claim completion.
Evidence is unresolved	The same complaint has conflicting delivery evidence instead of confirmed loss.	Investigate the conflict before treating the confirmed-loss remedy as authorized.

TABLE 3 Constructed contrasts derived from the expert correction and service policy. These are expected behaviors, not measured results.

Investigate partial improvements

Compare a candidate that applies the compatibility criterion with one that also reconciles pending requests. Code can prepare these changes; Goal investigates them through Trialground. [23] [24] Better selection alone should still fail the pending-action case if it creates a duplicate commitment.

Prepare learning material where needed

Weave can construct paired demonstrations in which a changed deadline reverses eligibility while a changed commercial priority does not. Preserve the reason for each preference. The learning target is conditional suitability, not a fixed preference for a product or an unconditional refusal.

Assess beyond the teaching worlds

Reserve different order families and evidence combinations before candidate selection. Require useful authorized resolution alongside protected requirements. A candidate that declines every replacement avoids some errors but fails the purpose of the task. Check final order state and intermediate commitments.

Transfer the lesson with local standards

Payment reconciliation, travel rebooking, and delivery recovery can also involve uncertain action outcomes. Their authority and completion evidence differ. Carry the method into those settings while eliciting their own requirements.

WORKED EXAMPLE SHOPPING ASSISTANCE

Make proactive assistance serve the customer

Useful initiative depends on an authorized purpose, current needs, and a justified next action.

For example, consider a shopping companion asked to watch for a suitable commuter bag and notify the customer when one becomes available. This constructed scope permits product investigation and a notification; purchasing requires separate authority. Catalog evidence confirms that Bag A is light and slim, while Bag B is heavier and bulkier with padded straps. Both meet the supplied laptop and weather-resistance requirements.

Clarify what changes the preference

The customer describes long walks and shoulder discomfort. The expert identifies a tradeoff: padding, carried weight, and fit may favor different choices. Catalog facts alone do not settle it. Ask about expected load and priorities, or present supported comparisons. If compactness takes priority, A may be preferred; padding alone does not establish that B will be more comfortable.

Review when initiative is useful

The agent should follow the requested purpose and notification conditions. A previously suitable product can become irrelevant when the customer buys elsewhere or changes the need. A commercial priority should not override an essential requirement or justify an unsupported claim.

Construct the continuing task

Benchmark worlds can vary restock events, changed preferences, unavailable products, withdrawn permission, and repeated notifications. Experts review the conditions for initiating contact, explaining a compromise, or closing the task.

Assess useful commercial behavior

Assess customer fit, factual claims, appropriate initiative, and fulfilled notification intent. Conversion alone cannot establish those requirements. Where business evidence exists, examine outcomes alongside contact burden and service consequences.

Related examples include an OTA agent revising options after a traveler changes dates or a service agent following up on a pending resolution. [35] Each requires its own purpose and authority.

WORKED EXAMPLE CODING ASSIGNMENT

Develop coding agents for acceptable changes

A useful coding agent completes the intended assignment while respecting repository evidence, contracts, and verification requirements.

For example, consider an intermittent timeout in a state-changing API. This constructed assignment requires preserving the public error contract. An agent invokes a migration skill, rewrites adjacent modules, and adds retries. A narrow test passes. Yet an operation may have succeeded before its response was lost, so retrying can repeat its effect.

Capture the maintainer's standard

Lemon prepares contrasts between failure before acceptance and timeout after a possible commit. The maintainer identifies the required evidence: the failure point, request identity, and duplicate-prevention behavior. [27] Lower timeout frequency alone cannot establish a correct fix.

The criterion requires justified retry or reconciliation, preserved error behavior, and changes scoped to the supported cause. A migration remains valid when the assignment requires it. Repository evidence can change which skill or procedure applies.

Put skills inside the complete task

Assess investigation, skill selection, implementation, and verification as parts of one assignment. A correct invocation can still produce an excessive edit; a successful patch can result from a familiar task that does not establish transfer.

Construct meaningful repository variation

Weave prepares codebase cases that vary failure timing, duplicate prevention, architecture, and task intent. Include tasks the skill should improve and tasks where it should remain unused. Reserve repositories and task families before revision.

Investigate with comparable conditions

Goal tests supported harness candidates in the agreed sandbox. Trialground retains invocation, patch, tool use, and results where instrumented. Compare candidates under the same task and model conditions. [23] [24]

CODING ASSIGNMENT CONTINUED

Connect coding quality to delivery effort

Define the accepted result and count the work needed to reach it, including review and correction.

Assessment dimension	Record	Decision it informs
Accepted task result	Completion, preserved contracts, justified scope, and task-specific verification.	Whether the assignment reached an acceptable state.
Skill contribution	Useful invocation, missed applicability, unnecessary use, and adaptation to repository evidence.	Whether the skill improves the intended work.
Human work	Review, explanation, correction, integration, and unresolved investigation.	Whether delegation reduces or redistributes effort.
Execution resources	Model, tool and infrastructure cost; elapsed time; retries and abandoned attempts.	The cost and latency of reaching an accepted change.
Delivery outcome	Time to accepted review or merged PR, rework and later defects, where workflow records exist.	Whether development evidence corresponds to operating value.

TABLE 4 Measures for a coding assignment. Record denominators, task mix, and the comparison conditions.

Compare a useful baseline

Keep task scope and acceptance criteria stable when comparing a baseline with a skill-bearing candidate. Record the model, harness, budget, and available tools. A change in any of them can affect the result.

Report cost per accepted change with failed attempts included. Separate agent execution time from human review and delivery waiting time. Inspect the distribution across tasks, including costly or slow cases.

Keep delivery claims tied to workflow evidence

An offline benchmark can assess candidate quality and execution resources. PR cycle time and production effects require the actual review and integration process. Compare those outcomes when the initiative supplies that evidence.

The enterprise retains code review, integration, release, and production access. Teammately supplies reviewed criteria, constructed tasks, and candidate evidence for that process. [27]

Development decision Evaluate coding throughput together with accepted quality, complete cost, and the work left to people.

WORKED EXAMPLE SPECIALIST ANALYSIS

Develop reasoning around domain assumptions

The expert standard should reach the analytical decisions that determine the conclusion and the next useful investigation.

For example, consider a financial research system assessing a repayment from supplied fictional records. Cash is 40, including 8 restricted; repayment is 50; forecast net inflow is 10. A facility of 20 requires debt divided by covenant EBITDA to be at most 3.5. Debt is 120 and reported annual EBITDA is 40, including a nonrecurring gain of 10 that the supplied contract excludes. The task compares inflow arriving before or after repayment. All financial amounts use the same units.

Analytical condition	Supported interpretation	Judgment to investigate
Restrictions on reported cash	Unrestricted cash is 32, after excluding the restricted 8.	Which source establishes that a resource is usable for this obligation?
Contractual earnings definition	Covenant EBITDA is 30. Leverage is 4.0, above 3.5; the supplied condition prevents drawing the facility.	Which definition governs the calculation, and which adjustments apply?
Inflow arrives before repayment	If the inflow is realized and usable before repayment, resources total 42, leaving a gap of 8.	What supports the forecast amount, timing, and availability?
Inflow arrives after repayment	Resources at the deadline remain 32, leaving a gap of 18.	Which next investigation could establish a feasible alternative?

TABLE 5 Constructed analytical contrasts. Resource definitions, contractual conditions, and timing shape the conclusion.

Elicit the analytical distinction

The arithmetic separates scenarios; it does not establish that a forecast will occur. Even timely inflow leaves a gap. The expert directs the next inquiry toward a condition that could close it, such as an authorized release of restricted cash or a facility waiver. Neither is an available resource without supporting evidence. [32]

Test depth and transfer

Test whether the model selects a consequential inquiry and distinguishes a possible remedy from established funding. Vary which remedies are permitted and evidenced. In science, the analogous judgment can be selecting a measurement that distinguishes competing mechanisms. Each domain supplies its own standard.

08 STRUCTURED DECISIONS

Develop decisions that software can use

A typed decision needs a substantive standard, an uncertainty policy, and assessment of its downstream effect.

AI behavior can be expressed as an answer, a tool action, or a structured decision used by an application. Schema-constrained outputs provide an interface for the last form. [37] The decision still needs to follow the evidence and the organization's intended standard.

Establish the decision boundary

For example, a payment-support component can classify a transaction as pending, confirmed, failed, or requiring investigation. Experts identify what the authoritative records establish and when apparently similar records require different treatment.

Compare the component across consequential conditions: delayed settlement, incomplete records, conflicting evidence, and repeated requests. Then examine what its output causes the surrounding workflow to do.

Assess uncertainty separately

Where a model supplies probabilities or confidence, test calibration against reviewed outcomes for the intended task. An action threshold depends on error consequences and the cost of further investigation. A confidence value alone does not establish correctness.

Retain distinct error consequences

A false confirmation can trigger an unjustified action. An unnecessary abstention can delay a legitimate resolution. Report both, along with coverage of cases the component accepts and performance on those cases.

The application enforces permissions, limits, and tool preconditions. The reviewed criterion specifies an acceptable decision; the team chooses supported prompting, training, or implementation changes to develop it.

Connect the component to the whole task

A locally accurate classifier can still serve a poor workflow if the application discards uncertainty or repeats an action. Assess the component and the composed task under their separate standards.

WORKED EXAMPLE JOURNEY RECOVERY

Preserve a journey through operational recovery

A successful local action can break the wider task. Dependencies and uncertain commitments belong in the assessment.

For example, consider an airline recovery agent using supplied remedy rules and booking tools. In this constructed case, an alternative arrives at 09:45 for an onward departure at 10:15. The supplied rule requires 45 minutes between flights. A seat is available, but the 30-minute connection fails that condition. An alternative arriving at 09:00 preserves the connection; its availability can still change.

World condition	Reviewed requirement	Evidence to inspect
A local option breaks the connection	Check whole-journey dependencies before committing to recovery.	The itinerary and the decision based on it.
Availability changes during the task	Revalidate affected conditions and revise the plan.	Current observations and the action they support.
Booking submission remains uncertain	Reconcile status before repeating the request or claiming completion.	Request identity, status evidence, and resulting booking state.
Work resumes after a handoff	Preserve the passenger objective and known commitments; check current state.	Transferred context, reconciled actions, and remaining work.

TABLE 6 Constructed operational conditions connect expert criteria to observable consequences.

Construct and investigate the world

Experts identify dependencies and authority. Weave constructs scoped tools, state, and transitions; Trialground runs the harness and retains recorded behavior. Goal can investigate recovery candidates. [28] [31]

Transfer across dependencies

An OTA itinerary, a delivery cutoff, or an industrial asset state can also constrain an otherwise plausible action. [35] The world must expose the dependency and the evidence the agent could actually observe.

09 ADAPTATION

Choose the intervention from the evidence

Use the diagnosed difficulty to direct learning or system changes, then assess the behavior under shared standards.

First locate the failure. Missing evidence can require retrieval or instrumentation; unavailable state can require an integration change. A model that receives the right evidence but repeatedly misuses it may need better instructions, a skill, supervision, or weight adaptation. A disputed standard needs resolution before it becomes a training target. The AI team selects the route.

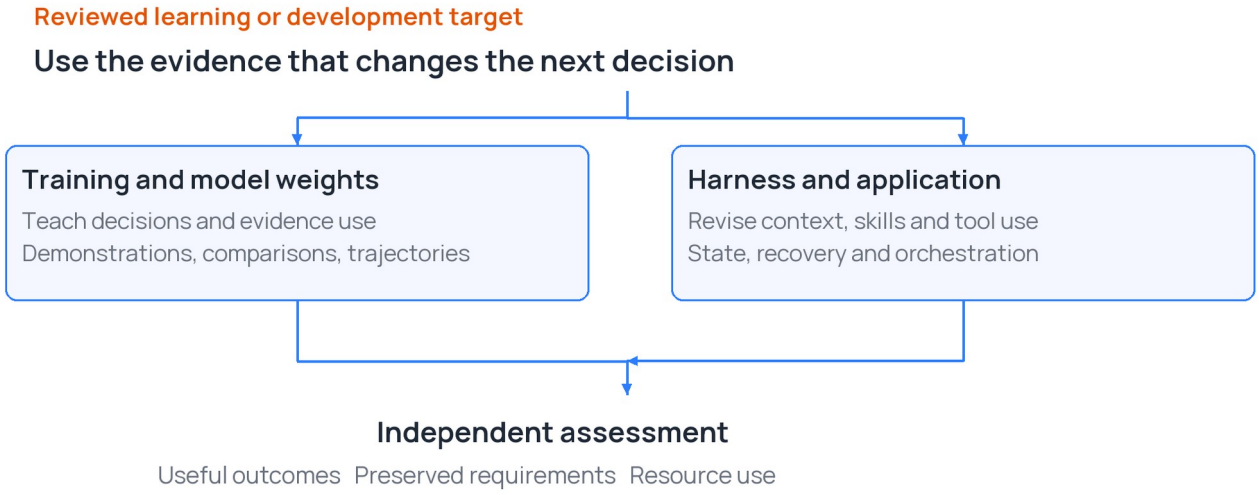


FIGURE 5 Reviewed criteria and representation guide model learning and harness changes. Executable permissions remain part of the application's operating controls.

Teach the missing distinction

Weave prepares targeted material around the condition experts established. [23] Demonstrations can show evidence use; comparisons can expose a reversing preference; trajectories can illustrate justified continuation or recovery. Review what the material should teach before training.

Preference learning and principle-guided feedback provide research foundations. [6] [7] [8] A training environment may use reviewed rubric-derived feedback where supported. Scope its data format, training method, compute, and assessment separately.

Revise the application harness

Code proposes changes to instructions, context, tools, or orchestration. Goal investigates and tests competing candidates under protected requirements, permitted changes, and an agreed budget. [23]

Possible interventions include useful task initiation, evidence gathering, skill applicability, uncertainty reconciliation, or context handoff. Declare the change and inspect its recorded effects. Separate assessment examines the candidate rather than relying on its own completion claim.

WORKED EXAMPLE MODEL TRAINING

Teach consequential analytical distinctions

Training material should teach decisions and evidence use across meaningful changes in conditions.

Returning to the constructed liquidity example, suppose the model repeatedly counts reported cash and nominal credit as usable funding despite receiving the governing definitions. The post-training team targets that recurring distinction. Supervision should also teach conditional forecasts and the next useful inquiry, so a correction does not merely produce more cautious wording.

Teaching contrast	Intended learning	Separate assessment
Reported and contractual definitions differ	Apply the governing definition and its material adjustments.	Unfamiliar agreements, source notes, and adjustment rules.
Restrictions or timing change availability	Count resources usable for the obligation at the relevant deadline.	New restrictions and obligation schedules.
Several plausible next investigations	Prefer an inquiry that could change the funding conclusion; separate a possible remedy from an established resource.	Different evidence gaps and valid investigative alternatives.
Evidence supports a broader conclusion	Provide the supported inference instead of applying blanket uncertainty.	Cases in which the relevant conditions are established.

TABLE 7 Constructed supervision design connects the reviewed distinction to independent variation.

Construct the required supervision

Weave can prepare demonstrations, preference pairs, supporting artifacts, and other agreed material. [23] [29] Where the training protocol requires decisions or trajectories, specify the representation and review its fidelity. Keep expert criteria, reasons for preferences, and construction provenance.

Train and assess separately

The team scopes training execution and compute in its selected environment. Record data, weights, configuration, and selection. Compare baseline and trained candidates under comparable serving conditions on reserved source and task families, including protected behavior.

10 INDEPENDENT ASSESSMENT

Assess complete tasks and their consequences

Use separate cases and task-appropriate evidence to establish useful outcomes, preserved requirements, and remaining uncertainty.

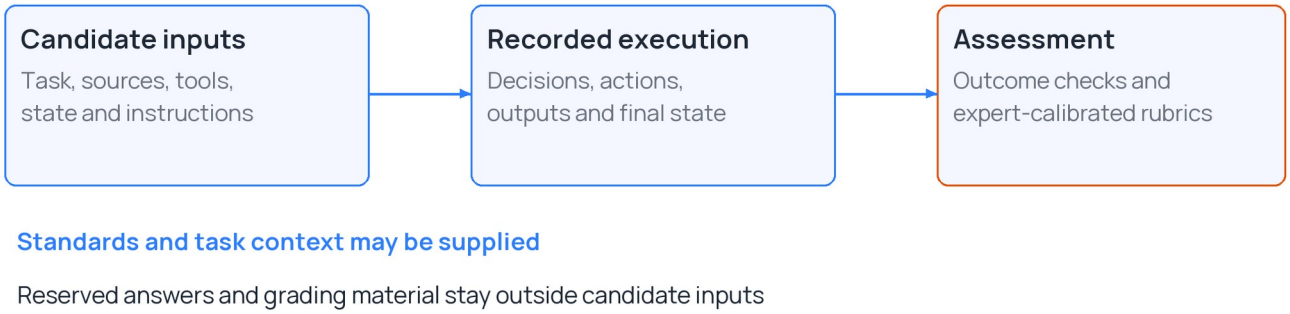


FIGURE 6 Assessment combines appropriate outcome checks and reviewed criteria while retaining information boundaries.

Keep assessment independent

Reserve case families before training or search. Keep hidden world state and expected answers outside candidate inputs. Shared templates and synthetic siblings can create dependence. Record adaptive holdout use and renew the independent basis for later claims. [13]

Inspect outcomes and consequential behavior

Use authoritative state checks for reservations, orders, or patches. Rubrics assess interpretive judgments and acceptable compromises. Inspect consequential intermediate violations even when final state looks successful. [14] [34] Allow valid alternative strategies without imposing one arbitrary tool sequence.

Repeated trials test consistency. In the 2024 tau-bench setup, GPT-4o retail task success was 61.2%, while consistent success across eight trials was below 25%. [12] These historical figures illustrate different reliability measures; Appendix B gives the study scope.

Validate the criterion and judge

A mistaken assumption can guide construction and judging alike. Challenge consequential criteria with authoritative sources, expert counterexamples, and independently obtained tasks or outcomes where available. Compare automated judgments with expert decisions and inspect false passes, false failures, and presentation sensitivity. [10] [11]

Report uncertainty and useful progress

Set acceptance before search: required improvement, protected requirements, and acceptable unresolved work. Report critical failures separately from averages. Retain inapplicable criteria, insufficient evidence, and unresolved applicability. Compare completion, abstention, escalation, and cost so that inactivity cannot masquerade as progress.

Report paired results by task family and the number of independent cases. Sparse evidence may remain inconclusive. Separate environment failures from candidate failures; record resets, horizons, repetitions, and capture limits.

11 ARCHITECTURE

Connect development work and evidence

Teammately's agents develop the application's standards, material, and candidates through an inspectable record.

Lemon undertakes inquiry, construction, and candidate investigation. Trialground Runtime executes supported harnesses and records behavior. Evaluation applies rubrics to outputs and trajectories. [21] [23] [24] The team directs integration, operating authority, and deployment.

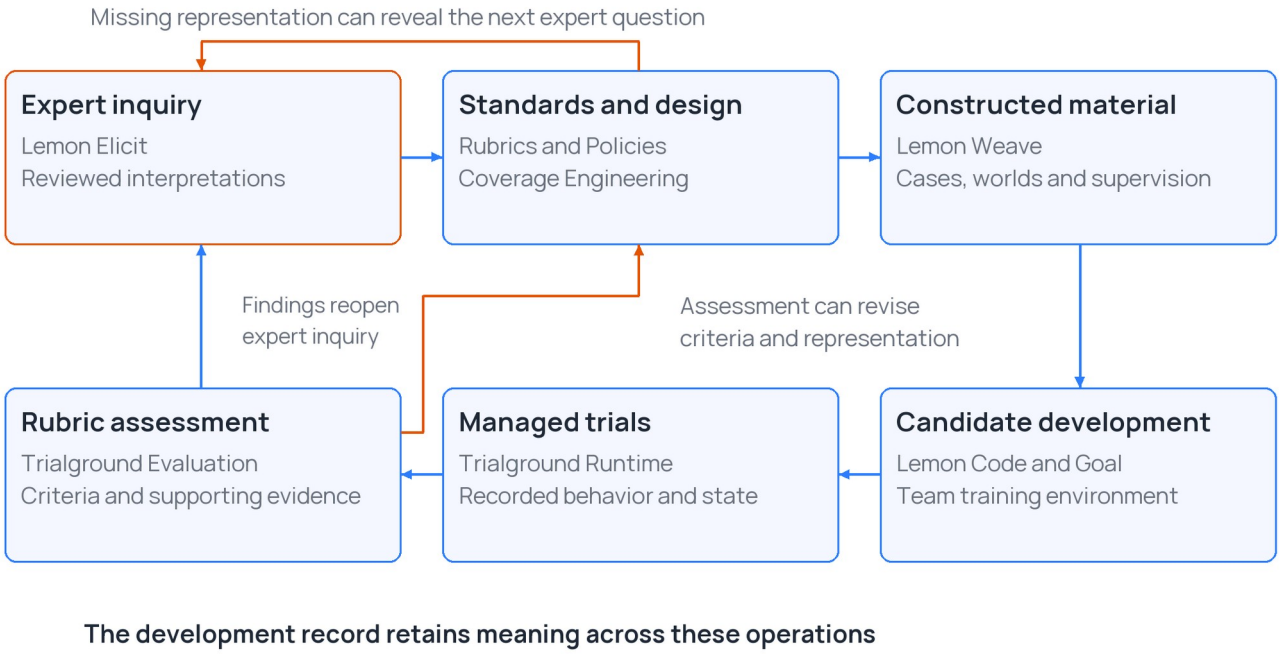


FIGURE 7 Components share reviewed meaning and representation. Arrows describe development dependencies and feedback rather than automatic policy propagation.

Standards direct construction

In the service example, the expert corrects what makes a replacement suitable. Elicit retains the interpretation; Rubrics and Policies express its approved scope. Coverage Engineering directs Weave to vary compatibility and customer agreement. The same distinction can guide supervision and candidate assessment. [20] [25]

Worlds availability

Worlds is labeled Early reviewers only. Its foundation and fixtures do not provide authoring, catalog APIs, or durable provisioning.

Retain the execution basis

Code and Goal investigate candidate changes against that basis. Runtime records execution evidence; Evaluation applies configured rubrics. If the standard changes, identify affected cases and results before reassessment. Earlier results retain their original interpretation; links do not authorize automatic policy changes.

Match execution to the scope

Agree the runtime profile, tools, resets, evidence capture, capacity, and artifact retention. Weight training and compute use the team's scoped environment. [24]

12 CHANGE AND LINEAGE

Preserve meaning as responsibilities change

A changed source, policy, tool, or objective can revise the basis of the standard and the evidence already produced.

Keep the expert contribution connected to its interpretation, applicable conditions, material, candidate, and results. Teammately retains source-linked contributions, reviewed standards, case-construction records, and pinned benchmark and candidate results within supported workflows. These records support later work without silently rewriting an earlier evaluation. Scoped applicability reassessment is supported; teams direct impact review and reassessment across the broader lifecycle.

Illustrative change	Affected basis	Work to review
A booking tool adds pending status	Acknowledgment no longer means a confirmed reservation.	Criterion, state transitions, recovery candidates, and assessment evidence.
A customer changes the objective	Earlier notes or delegated work retain the old priority.	Context authority, outstanding commitments, and resumed-task worlds.
An analytical condition changes	Earlier supervision assumes a resource or method is applicable.	Sources, decision boundaries, teaching material, and reserved assessment.
A coding model or skill changes	Previously accepted behavior may use different investigation or verification.	Task mix, acceptance standards, execution resources, and regressions.

TABLE 8 Illustrative changes identify the development artifacts and comparisons whose basis needs review.

Reuse with an explicit basis

Review whether the original conditions hold in the new task. Reuse a distinction where justified, and investigate the new dependency or authority question where they differ. A result can direct further inquiry or construction.

Retain historical interpretation

Pinned benchmark and candidate results retain their captured standards, Cases, and execution conditions. A changed basis calls for a justified reassessment. Appendix A records the comparison identities and assumptions.

13 DEVELOPMENT RESPONSIBILITY

Direct the alignment work

Choose how the development work will be carried through after experts contribute, including the work retained by your team.

Compare the work returned

Ask who turns an expert correction into a criterion, checks its boundary, constructs meaningful variation, and investigates the application. Then ask who maintains those connections when the task changes. Compare the returned work and its acceptance evidence, not just the number of reviews or generated examples.

Use the relevant expertise

Employees or commissioned specialists can contribute. Access to expertise and responsibility for development work are separate choices. An engagement may provide both. Establish who supplies the knowledge, who corrects its interpretation, and who can authorize an organizational standard.

Assign continuing operations

An internal practice must own preparation, interpretation, construction, experiments, and maintenance, including any agents it builds. An external engagement should make the work it assumes equally clear. A completed dataset can be useful while leaving criterion revision, candidate investigation, and later changes with the team.

Let agents undertake the alignment engineering

Teammately's agents adapt inquiry to expert reasoning, carry reviewed distinctions into material, and investigate candidates. [21] [23] [24] Your experts and AI engineers direct the work and subsequent alignment cycles, retaining decision authority. Define the required result, with its interpretation and supporting records available for inspection.

Bring the existing stack into the assignment

Review, evaluation, and observability tools can provide datasets, annotations, calibration, traces, and prompt-development evidence. [15] [16] [17] [18] [19] These can feed an alignment assignment. Define who advances the work from an observed failure to a reviewed distinction and a tested change, with supported exchange around the existing stack.

Assess the development result

For any internal team, specialist engagement, or platform, inspect interpretation fidelity, material quality, candidate evidence, and complete effort. Teammately should be assessed on this same basis. The selection concerns how much of the required development work is performed well and what remains with your team.

14 FIRST ENGAGEMENT

Plan a useful first assignment

Begin with the responsibility the AI should handle and the evidence available today.

Bring a development objective, relevant sources, policies, traces, a skill, or a candidate. Lemon can begin with a judgment question, representation need, or candidate investigation. [21] Participation follows the output; a completed benchmark or separate annotation program is not a prerequisite.

Example objective	Agent assignment	First output to inspect
Resolve an order or service task	Investigate consequential remedies, construct stateful cases, or test a recovery change.	Reviewed criteria, worlds, or candidate evidence tied to the resolution objective.
Improve accepted coding work	Elicit maintainer standards, vary repository tasks, and investigate skill-bearing candidates.	Task benchmark and evidence on quality, scope, verification, and execution resources.
Strengthen specialist reasoning	Clarify assumptions and decision boundaries, then construct sources and supervision.	Reviewed analytical criteria, targeted material, and an independent assessment design.
Develop continuing or structured decisions	Establish initiation, uncertainty, continuation, and downstream requirements.	Criteria and cases that expose their consequences under the selected scope.

TABLE 9 Illustrative first assignments. The team's actual objective determines the work and output.

A concrete service assignment

For the service example, supply the remedy policy, product constraints, and available traces or a candidate harness. Delegate interpretation, contrasting worlds, and a candidate investigation. The expert resolves suitability; the owner approves shared scope. Agree context access, retention, and reuse for that material.

Inspect the returned result

Inspect the corrected criterion, cases that reverse the decision, and evidence comparing the candidate with its baseline. Record expert corrections and unresolved failures. That result supports a decision to revise the candidate, extend assessment, or widen the responsibility under agreed execution limits.

Development decision Make the first assignment concrete enough to produce a useful development decision.

15 VALUE AND EVIDENCE

Measure application quality and complete effort

Separate the work of developing the AI from the cost and quality of the application's operation.

Measure	Record	Question it answers
Application quality	Useful completion, preserved requirements, reasoning, recovery, and uncertainty by condition.	Does the system handle the intended responsibility?
Operating effort	Execution cost, elapsed time, retries, human intervention, and downstream correction per task.	What does acceptable completion require in use?
First development result	Elapsed time and complete human work through preparation, inquiry, review, and correction.	Did the alignment assignment support the initial decision?
Expert participation	Contribution, interpretation review, correction, and disagreement resolution.	Is the specialist work practical for this initiative?
Reuse and maintenance	Retained distinctions, integration effort, changed dependencies, missed impacts, and unnecessary reruns.	Can the practice serve subsequent responsibilities and revisions?

TABLE 10 Measures need denominators, task mix, and a comparison basis. Include failed attempts and work transferred to people.

Connect behavior to operating outcomes

For example, examine resolved service cases, accepted code changes, or useful analytical decisions. Where workflow evidence exists, relate quality and complete effort to service capacity, delivery time, or business value under stated assumptions.

A published Morgan Stanley account describes advisors and prompt engineers grading responses and using evaluation to refine an assistant. Its initial goals included retrieval, repetitive-task automation, and client insights. [5] It supplies context for one enterprise initiative.

Use the current initiative as a baseline

Use the same task scope and acceptance basis when comparing the assignment with current practice. Record failed attempts and correction effort. If several interventions change together, the result describes their combined effect. Appendix C separates the value of the structured practice from the additional contribution of agents.

Teammately's research examines supervision, benchmark design, and adaptation. [22] For the assignment, distinguish a useful development artifact, an independently assessed behavior gain, and an operating outcome. Each can inform a decision, but one does not establish the others.

16 NEXT DEVELOPMENT OBJECTIVE

Define the responsibility to develop

Develop the judgment needed for a wider AI responsibility, then establish whether the resulting behavior deserves that responsibility.

The central development challenge is preserving why a decision is appropriate as work moves from expert judgment to criteria, constructed material, learning, and assessment. A useful correction must retain the conditions that make it valid. Those conditions let a team develop behavior that adapts when the situation changes.

Make the distinction reusable

The service example shows the mechanism. An expert corrects a plausible but inadequate suitability rule. Contrasting worlds preserve the customer requirement and vary the condition that changes an acceptable remedy. Candidate investigation then examines both useful resolution and consequential failures. The same practice can develop coding skills, specialist reasoning, and model learning under their own standards.

Test what the development work achieved

Artifact quality and application quality require different evidence. A faithful rubric is a development result. A candidate applying it on independent tasks is a behavior result. More acceptable work at a justified cost is an operating result. Keeping these claims distinct makes the next investment decision better informed.

Assign the connecting work

Teammately's agents undertake preparation, adaptive inquiry, construction, and candidate investigation. [21] Experts correct meaning and owners authorize shared standards. The AI team directs model or application development and adoption. Evaluate the assignment by the useful work returned, the evidence behind it, and the effort left to your team.

Begin with the current initiative

Bring the responsibility you want AI to handle and the context available today. It may be a service process, a coding skill, difficult analytical work, or a learning target. Teammately can begin with the unresolved judgment, missing representation, supervision need, or candidate investigation. Define the first useful result and its acceptance evidence together.

The objective is an in-house capability to develop AI that takes on more useful work under standards the organization understands, can inspect, and can revise as its needs change.

Discuss a first alignment scope with Teammately

REFERENCES

References

[1] Pan, M., et al. (2026). Measuring Agents in Production. arXiv:2512.04123, version 3. Sections 3, 5, 6 and 7. Source

[2] LangChain. (2026). State of Agent Engineering. Public survey fielded 18 November to 2 December 2025; 1,340 responses. Source

[3] Thomson Reuters. (2023). Efficiently Evaluating LLMs for Legal Tasks. 30 November. Source

[4] Kardos-Nyheim, A. (2026). How We Built Thomson. Thomson Reuters Institute, 24 August. Source

[5] OpenAI. Morgan Stanley Uses AI Evals to Shape the Future of Financial Services. Customer account. Source

[6] Ouyang, L., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. NeurIPS 2022; arXiv:2203.02155. Source

[7] Christiano, P. F., et al. (2017). Deep Reinforcement Learning from Human Preferences. NeurIPS 2017; arXiv:1706.03741. Source

[8] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073. Source

[9] Plank, B. (2022). The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. EMNLP 2022, pp. 10671-10682. Source

[10] Kim, S., et al. (2024). Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. ICLR 2024; arXiv:2310.08491. Source

[11] Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS Datasets and Benchmarks; arXiv:2306.05685. Source

[12] Yao, S., Shinn, N., Razavi, P., and Narasimhan, K. (2024). tau-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. arXiv:2406.12045. Source

REFERENCES CONTINUED

References continued

[13] Dwork, C., et al. (2015). Generalization in Adaptive Data Analysis and Holdout Reuse. arXiv:1506.02629. Source

[14] Anthropic. (2026). Demystifying Evals for AI Agents. Engineering article. Source

[15] LangChain. LangSmith Evaluation. Expert feedback, evaluations, and prompt development. Product description. Source

[16] Braintrust. (2026). How to Improve Your Golden Datasets with Human Review. 21 May. Source

[17] Arize. Human Review and Align Evals to Human Feedback. Arize AX documentation. Source

[18] Datadog. (2025). Define, Run, and Scale Custom LLM-as-a-Judge Evaluations in Datadog. 25 November. Source

[19] Portkey. Full-Stack Observability for AI Apps. Product description. Source

[20] Teammately. Coverage Engineering. Benchmark representation across Dimensions/Ontologies, Topics, and Scenarios. Product description. Source

[21] Teammately. The Teammately Platform. Lemon and Trialground product roles. Source

[22] Teammately. Research. Structured Supervision, Benchmark Science, Behavioral Adaptation, and Delegated Agency. Source

[23] Teammately. Lemon. Adaptive elicitation, case and training-material synthesis, interactive harness development, and benchmark-driven optimization. Product description. Source

[24] Teammately. Trialground. Runtime for managed harness execution; Evaluation for rubric assessment. Product description. Source

[25] Teammately. Rubrics and Policies. Criteria, scope, precedence, and linked expert evidence. Product description. Source

REFERENCES CONTINUED

References continued

[26] Teammately. Applications and Practices. Development objectives and applied industry studies. Source

[27] Teammately. Develop Coding Agent Skills That Hold Across Unfamiliar Tasks. Development practice. Source

[28] Teammately. Test What the Agent Does When the World Changes Around It. Operational agent development practice. Source

[29] Teammately. Teach the Distinction Behind the Benchmark Failure. Model development practice. Source

[30] Teammately. Make Recommendations Fit the Customer, the Assortment, and the Moment. Retail industry study. Source

[31] Teammately. Keep Passenger Recovery Coherent as Plans and Availability Change. Airline industry study. Source

[32] Teammately. Keep the Conclusion Within What the Evidence Supports. Financial services industry study. Source

[33] Rajasekaran, P. (2026). Harness Design for Long-Running Application Development. Anthropic, 24 March. Source

[34] OpenAI. Trace Grading. Agent workflow evaluation documentation. Source

[35] Teammately. Make the Whole Itinerary Fit the Traveler's Priorities. Travel planning and booking study. Source

[36] Teammately. Account for Everyone Affected by a Service Remedy. Ride sharing and food delivery participant support study. Source

[37] OpenAI. Structured Model Outputs. Schema-constrained output documentation. Source

Public descriptions were checked during the October 2026 revisions. Enterprise accounts remain attributed to their publishers. Worked cases are constructed illustrations; application studies describe development methods. Appendix B retains numerical research scope.

Typography Manrope V5 by Mikhail Sharanda

APPENDIX A COMPARISON RECORD

Record the basis of a comparison

Retain the responsibility, evidence, conditions, and interpretation needed to reconstruct a development decision.

Record	Required detail
Development decision	Task, intended use, responsibility, objective, starting evidence, and decision to support.
Expert contribution	Contributor, evidence, contrast, correction, interpretation, and unresolved questions.
Standard	Criteria, applicability, authority, owner, scope, precedence, version, and acceptable alternatives.
Representation and material	Benchmark or training purpose, mix, state/events, source families, exclusions, and provenance.
Candidate	Model or weights, harness, instructions, skills, tools, intervention, and configuration.
Execution	Initial state, authority, task horizon, dependencies, resets, delegation, budgets, and repetitions.
Assessment	Reserved families, outcome checks, judge, reviewed decisions, uncertainty, and information boundaries.
Result and decision	Outcomes and denominators, paired differences, regressions, complete effort, costs, and disposition.

TABLE 11 Minimum comparison record. State capture limits and distinguish independent cases from repeated executions.

Keep dependence identifiable

Record shared sources, templates, construction families, and world assumptions. Group dependent siblings within development or reserved assessment unless the comparison accounts for dependence.

Preserve version meaning

A revised standard establishes a new basis. Explain affected material, retained cases, and repeated comparisons. Keep application outcomes separate from the effort of producing the alignment artifacts.

Development decision Ensure another reviewer can reconstruct what was tested and what the result permits the team to conclude.

APPENDIX B RESEARCH CONTEXT

Interpret the research measures

Each number answers a defined question within its original study. The setup is part of the result.

Human feedback and model adaptation

InstructGPT combined demonstrations and ranked outputs with fine-tuning. Human raters preferred InstructGPT with 1.3 billion parameters to GPT-3 with 175 billion parameters on the evaluated prompt distribution. [6] This is a preference result for those conditions; parameter counts do not measure total development effort, cost, or quality on a new enterprise task.

Customized rubrics and human scores

Prometheus used customized rubrics and reference answers. Its human comparison covered 45 instances with distinct rubrics, sampled from three test sets and excluding math and coding. Reported Pearson correlations were 0.897 for Prometheus, 0.882 for GPT-4, and 0.392 for GPT-3.5-Turbo. [10] Score association is different from binary criterion accuracy.

Agreement and treatment of ties

MT-Bench's first-turn pairwise comparison reported GPT-4-human agreement of 85%, versus 81% human-human agreement, when ties were excluded. Including ties and inconsistent judgments produced 66% and 63%. [11, Table 5] The setup used 80 multi-turn questions and answers from six models. State the comparison and tie treatment when reporting agreement.

Task completion and repeated reliability

The 2024 tau-bench function-calling evaluation reported GPT-4o task success of 61.2% in retail and 35.2% in airline. Retail consistent success across eight trials fell below 25%. [12, Table 2 and Figure 4] The benchmark contained 115 retail and 50 airline tasks with simulated users, API tools, and domain policies. Repeated reliability and single-trial completion support different conclusions.

Evidence about current practice

The production study's 70% off-the-shelf and 75% no-formal-benchmark findings each describe 20 interviewed deployment teams, including 14 production and six pilot cases. Its 74% human-evaluation finding uses a separate deployed-system subsample of 31 and includes experts, operators, and end users. [1] These are not population estimates for all enterprises.

LangChain's public survey received 1,340 responses; 63% were from technology and 49% from organizations with fewer than 100 people. Eighteen percent were from organizations with at least 2,000 people. It reports offline evaluation at 52.4%, online evaluation at 37.3%, and a 22.8% not-evaluating share among respondents with production agents. [2] These categories do not establish adoption of a structured expert alignment practice.

APPENDIX C COMPARATIVE STUDIES

Study the practice and agent contribution

Compare the practice and the additional contribution of agents under defined tasks, effort budgets, and assessment protocols.

Begin with an enterprise development decision and participants' current materials. Specify the application responsibility, initial model or harness, expert access, allowed tools, and independent assessment. Preserve comparable information and operating constraints across conditions.

Condition	Process to compare
Current practice	Existing sources, prompts, skills, tests, evaluation tools, and expert feedback, with work recorded.
Capable structured practice	Expert inquiry, reviewed criteria, deliberate representation, construction, and independent assessment using available tools.
Practice with agent support	The same review requirements, with agents adapting inquiry, constructing material, and investigating candidates.

TABLE 12 Prospective comparison conditions distinguish adoption value from the additional contribution of agent support.

Count complete effort and useful outcomes

Record preparation, contribution, interpretation, correction, construction, experiments, integration, and maintenance. Identify model and infrastructure costs. Assess the application's task quality, completion, intervention, and resources separately from development effort.

Test transfer and extended tasks

Use reserved families and unfamiliar conditions. Include interruptions, delegation, changed authority, and consequential recovery where the responsibility requires them. Inspect protected behavior and ablate the distinctions the expert contribution established.

Keep research questions distinct

Structured Supervision examines contributions under matched effort. Benchmark Science examines the predictive value of assessments. Behavioral Adaptation tests transfer beyond development material. Delegated Agency examines agent work and retained human authority. [22]

Study later revisions

Change a standard, task objective, skill, or model. Measure justified reassessment, reused distinctions, missed impacts, and unnecessary work. Retain uncertainty and adverse outcomes as part of the result.

Development decision Report the observed outcomes, denominators, and decisions the evidence supports.