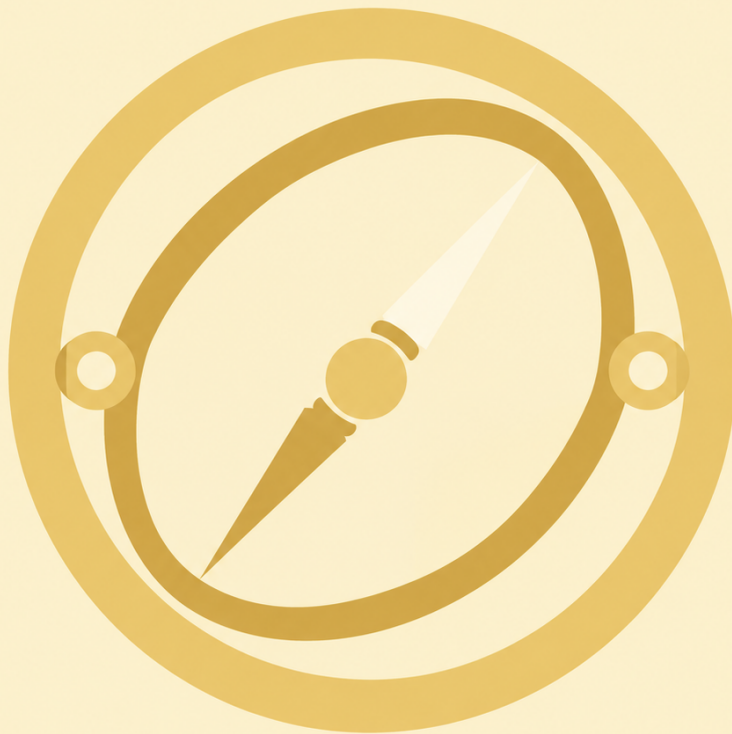


Knowing When to Ask

Uncertainty, Abstention, and Evidence-Seeking in AI Agents



Shehryar Awan

CTO, Deaimer

UNCERTAINTY / EVIDENCE / DECISIONS

Inside this article

A conceptual research paper connecting confidence calibration, selective action, clarification, and the value of obtaining evidence.

1. Introduction
2. Research approach and related work
3. What is the agent uncertain about?
4. Confidence must name an outcome
5. Selective action and the cost of being wrong
6. Evidence-seeking as a decision
7. The four-route framework
8. An illustrative decision scenario
9. Proposed research questions and hypotheses
10. Proposed environment-based study
11. Measurement beyond answer accuracy
12. What custom datasets can contribute
13. Limitations and threats to validity
14. Discussion and conclusion

Includes a four-route decision framework, risk-and-coverage measures, a value-of-information specification, and a proposed environment-based study.

Abstract

AI agents must decide how to proceed when their information is incomplete, their task interpretation is ambiguous, or the consequences of an action remain uncertain. A confidence estimate alone does not determine the appropriate response. An agent may need to inspect its environment, ask a targeted question, choose a reversible alternative, or abstain. This paper examines uncertainty-aware behavior as a decision problem connecting confidence, evidence availability, and the cost of error. Through a selective synthesis of research on calibration, selective prediction, language-model self-evaluation, semantic uncertainty, clarification, and conformal prediction, we distinguish uncertainty estimation from action selection. We propose a four-route decision framework and formalize a simplified risk-sensitive threshold and the expected value of obtaining information. We then specify an environment-based research protocol that varies information access, task ambiguity, and intervention cost, with metrics for selective risk, coverage, calibration, evidence use, and interaction burden. The contribution is a conceptual specification and research agenda, rather than an implemented agent or a report of measured improvements. We discuss how custom task environments within Deaimer's public scope could support empirical investigation.

Keywords: AI agents; uncertainty quantification; confidence calibration; selective prediction; clarification; evidence-seeking; model evaluation.

1. Introduction

An AI agent investigating a software failure encounters two plausible explanations. It can modify a configuration immediately, inspect a diagnostic file, or ask the user which behavior the system is intended to preserve. Each choice has a different cost and a different relationship to uncertainty.

The agent may be confident that one explanation is more likely. That confidence does not establish that immediate modification is the best decision. A cheap observation might resolve the ambiguity. A mistaken change might be expensive. The user may possess information about the intended outcome that the environment cannot reveal.

These distinctions become important when AI systems are expected to complete work across tools and multiple steps. Their success depends on how they allocate attention and action under incomplete information, including when they recognize that an answer or plan needs more support.

This paper studies the relationship between uncertainty and behavior. Its central thesis is that uncertainty estimation should be evaluated together with the decisions it informs. Expressing doubt can be appropriate, but it provides limited value if the agent does not obtain relevant evidence or ask a useful question. Conversely, an agent can proceed sensibly without perfect confidence when the action is low-cost, reversible, and informative.

The paper proposes a four-route framework: act, inspect, clarify, and abstain. These routes represent different responses to missing information. The framework is concerned primarily

with epistemic adequacy and task interpretation. Authorization remains a separate requirement: confidence never supplies permission for an otherwise unauthorized operation.

1.1 Contributions

The contributions are: (i) a taxonomy of uncertainty sources in interactive work; (ii) a decision-theoretic specification linking error cost to selective action; (iii) a framework for choosing among environmental inspection, user clarification, execution, and abstention; and (iv) an experimental protocol that measures both decision quality and interaction burden.

The component ideas have established foundations in calibration, selective prediction, and decision theory. We propose their organization for environment-based agent research without claiming that they are individually new. No model training, implementation, released dataset, or experimental result accompanies this paper.

2. Research approach and related work

The paper uses a selective narrative review of primary sources. Research was selected for relevance to probabilistic confidence, rejection or deferral, uncertainty in language generation, clarification, and statistical coverage. This is not an exhaustive review of current models or commercial memory and agent products. Findings are interpreted within the tasks and systems studied by their authors.

2.1 Calibration and selective prediction

Guo and colleagues study confidence calibration in modern neural networks and examine post-processing methods including temperature scaling. Their work helps distinguish predictive accuracy from how well stated probabilities correspond to correctness frequencies. [1]

Geifman and El-Yaniv study selective classification for deep neural networks, where rejection trades prediction coverage for risk on retained cases. The setting provides a foundation for discussing abstention without treating maximum coverage as the only objective. [2]

These studies concern supervised prediction settings. Applying their concepts to agents requires an explicit outcome event and a definition of which decisions are accepted or deferred.

2.2 Language-model self-evaluation

Kadavath and colleagues investigate whether language models can estimate the validity of their own answers and predict which questions they can answer. Their results include encouraging behavior in specified formats and difficulty transferring calibration to new tasks. [3]

Tian and colleagues evaluate methods for eliciting confidence from models fine-tuned with human feedback. They report that verbalized confidence can be better calibrated than conditional probabilities on the question-answering benchmarks they study. [4]

These findings motivate the empirical comparison of confidence signals. They do not establish that any particular self-reported score is a reliable probability of successful tool execution in an unfamiliar environment.

2.3 Semantic uncertainty

Farquhar and colleagues study semantic entropy for detecting a subset of language-model hallucinations: confabulations associated with uncertainty over the meaning of generated answers. Their method distinguishes differences in wording from differences in semantic content. [5]

This direction is relevant when several sampled responses express competing conclusions. However, agreement among samples does not by itself establish truth. A consistently repeated mistaken belief can remain low-diversity output.

2.4 Clarification and interaction

CLAM investigates selective clarification for ambiguous user questions, including detecting ambiguity, generating a clarifying question, and answering after a response. [6]

AskBench studies when and what to ask through interactive tasks, including intent-deficient questions and misleading premises. Its proposed training and evaluation methods further motivate examining question quality and interaction efficiency together. [7]

These works primarily examine question-answering interactions. Our proposed extension is to study clarification as one route within executable workflows, where some missing information can instead be recovered from the environment.

2.5 Statistical uncertainty sets

Angelopoulos and Bates provide an introduction to conformal prediction and distribution-free uncertainty quantification. Standard coverage guarantees depend on the assumptions of the specific construction, including exchangeability in common settings. [8]

Conformal methods can be relevant to structured prediction or bounded decision settings. Their guarantees should not be transferred automatically to a changing agent trajectory, where actions influence later observations and sampling conditions may shift.

3. What is the agent uncertain about?

An uncertainty score is useful only when its target is clear. Confidence in a fact, confidence in a task interpretation, and confidence in the success of an intervention concern different events.

We propose four uncertainty sources for authoring interactive tasks. They are conceptual categories and need not be statistically separable in a model's internal representation.

3.1 Knowledge uncertainty

The agent may lack a relevant fact or method. It might not know the conventions of a tool or the assumptions of a numerical procedure. Documentation, examples, or an expert response may help resolve the gap.

3.2 State uncertainty

The agent may know how a system works in general while lacking evidence about its current state. A process could be stopped, a file could have changed, or a dependency could differ from the documentation. Current observation is often more appropriate than asking the user to supply information the agent can inspect.

3.3 Intent uncertainty

The objective may permit several interpretations. For example, "clean the outputs" could mean remove temporary files, reorganize final artifacts, or revise their content. The environment may not reveal the user's intended interpretation. A targeted question can have direct value.

3.4 Outcome uncertainty

An action can have uncertain consequences even when the task and state are understood. Tool behavior, interactions among components, or incomplete evidence about a cause can make success difficult to predict. A focused test or reversible intervention may provide information while limiting exposure.

Uncertainty source	Illustrative missing information	Potential response
Knowledge	Meaning of an unfamiliar tool option	Consult accessible documentation
State	Which dependency version is installed	Inspect the current environment
Intent	Which outputs the user wants preserved	Ask a specific clarifying question
Outcome	Whether a proposed change addresses the cause	Run a focused test or bounded trial

The categories help connect uncertainty to a possible source of evidence. Their purpose is to discourage a generic response to every confidence deficit. Asking a question is useful when the user can resolve the relevant gap; inspection is useful when the environment contains the answer.

4. Confidence must name an outcome

Let h denote the information available at a decision point. Let a be a proposed action or plan, and let Y be a binary event whose meaning is specified before evaluation. For example, Y may indicate whether a proposed patch satisfies the task's functional requirement under a defined test distribution.

Let $q(h, a)$ be the estimated probability that $Y = 1$. This number is meaningful only relative to its event, evaluation conditions, and information boundary. Confidence that a diagnosis is correct is different from confidence that a particular repair will succeed.

Equation 1. Estimated success probability: $q(h, a) = \text{estimated } \Pr(Y = 1 \mid h, a)$.

Calibration concerns the relationship between predicted probabilities and observed correctness frequencies. It does not imply certainty for a particular case. A model can be calibrated on average while systematically overstating confidence on a subgroup.

Discrimination is also distinct. A predictor that always reports the overall success frequency can be calibrated under suitable conditions while offering little ability to identify which cases should be deferred. A useful selection signal needs to rank relevant risks as well as provide an appropriate probability scale.

This distinction matters for agent evaluation. A confident explanation may have weak correspondence with the next action's success. A probability elicited after observing the final outcome is not a prediction made at the original decision point. Researchers should record confidence before revealing the information used to label it.

Uncertainty estimators can include verbalized confidence, model probabilities where available, sample disagreement, external predictors, and domain-specific evidence. No signal should be assumed to dominate across every workflow. The proposed protocol compares them under matched task and resource conditions.

5. Selective action and the cost of being wrong

5.1 Risk and coverage

For evaluation cases indexed by i , let g_i indicate that the system elects to proceed with its current proposal without additional information. Let e_i indicate that this proposal is incorrect under the specified outcome definition. Coverage is the fraction of cases selected, and selective risk is the error fraction among those selected.

Equation 2. Coverage: $C = (1/N) \times \sum_i g_i$.

Equation 3. Selective risk: $R = [\sum_i (g_i \times e_i)] / [\sum_i g_i]$, when $\sum_i g_i > 0$.

Risk is undefined if no cases are selected. Such a system should be reported as having zero coverage, rather than credited with a perfect error rate. In an agent workflow, selection can be studied at a predefined checkpoint or at the task level; the unit must be consistent across comparisons.

Varying a confidence threshold produces a risk-coverage relationship. This can show how much work the system retains while controlling observed error. It should be accompanied by the behavior of deferred cases, because inspection or clarification may allow them to be completed later.

5.2 A simplified decision threshold

Consider a one-step choice between proceeding and deferring. Assume a correct action has zero loss, an incorrect action has loss $L > 0$, and deferral has total expected loss c with $0 \leq c \leq L$. Assume q is the true conditional success probability for this decision.

The expected loss of proceeding is $(1 - q)L$. Deferral is preferred when c is smaller. Proceeding minimizes loss between the two options when:

Equation 4. Decision threshold: $q \geq 1 - c/L$.

This is a standard binary loss comparison, not a new theorem. It makes the role of consequences explicit. The same success probability can justify proceeding on a low-cost task and deferring on a high-cost one.

The comparison also exposes its limits. Deferral can be imperfect, delay can create loss, and tasks can have multiple outcomes with different costs. A population-calibrated score is not automatically the true conditional probability required by the derivation. Thresholds therefore need empirical assessment under the deployment distribution.

6. Evidence-seeking as a decision

Deferral need not terminate a task. The agent can obtain information and then reconsider its options. Inspection and clarification differ in where that information comes from, but both can be examined through expected decision loss.

Let b represent the agent's belief over relevant task conditions. Let D be its admissible decisions, and let $\text{loss}(d, \theta)$ describe the consequence of decision d when the actual condition is θ . The current minimum expected loss is:

Equation 5. Current decision risk: $R_0 = \min_d E_b[\text{loss}(d, \theta)]$.

Suppose an information action m costs c_m and produces observation Z . After observing Z , the agent may choose a revised decision. Define its expected net improvement as:

Equation 6. Net information value: $V(m) = R_0 - E_Z[\min_d E[\text{loss}(d, \theta) | Z]] - c_m$.

Under the stated one-step model, an information action has positive value when its expected reduction in decision loss exceeds its cost. This formulation is a standard decision-theoretic comparison adapted here to agent inspection and clarification.

Proposition 1. If information is free, the probabilistic model is correct, and the agent can retain all original decisions after observing it, the optimal expected loss after observation cannot exceed R_0 .

Argument. The agent can ignore the observation and choose the original minimizing decision. Optimizing over the available decisions after each observation therefore performs at least as well in expectation as that fixed strategy. The statement does not guarantee

benefit when observation has cost, changes the environment, restricts later options, or is interpreted incorrectly.

The challenge in real systems is estimating the quantities in Equation 6. An agent may be uncertain about how informative a tool response will be or whether a user can answer the proposed question. The framework supplies a research target, not an assumption that the agent possesses accurate values.

A useful experiment can specify reference information actions and compare an agent's choices with their observed effect on task performance. The reference should be an empirical or task-defined benchmark, rather than an omniscient score presented as attainable by the agent.

7. The four-route framework

We propose four routes for uncertainty-aware behavior. They are options within an authorized task, and several routes can occur sequentially.

7.1 Act

Proceed when the available evidence supports the proposed action under the relevant consequence model. A reversible, low-cost operation can also be informative. The system should distinguish an exploratory action from a confident commitment and preserve the ability to revise its plan.

7.2 Inspect

Use a tool or source to obtain missing environmental or factual information. Good inspection is targeted: the observation should discriminate among relevant possibilities or establish a necessary prerequisite.

Repeatedly checking the same unchanged state can add cost without resolving uncertainty. Researchers can examine whether an agent notices that an observation failed to answer its question and selects a different source.

7.3 Clarify

Ask the user a question that can change the decision. A useful question names the unresolved distinction and avoids requesting information already accessible to the agent. When several unknowns exist, it prioritizes the one whose answer most affects the next action.

Clarification quality includes relevance, answerability, and how well the agent incorporates the response. A well-phrased question has limited value if the later plan ignores the answer.

7.4 Abstain

Stop or hand off when required evidence is unavailable, the proposed action exceeds a supported capability, or the cost of proceeding is unacceptable under the task conditions. Abstention should identify the unresolved issue and, where appropriate, specify what would allow work to resume.

The framework does not prescribe a fixed order. An agent can inspect before asking, ask before performing a costly test, or complete part of the task while deferring a distinct unresolved step. The appropriate sequence depends on information access and consequences.

8. An illustrative decision scenario

Consider a fictional data-processing task in which a file can use either of two unit conventions. A transformation is correct under one convention and wrong under the other. The instruction omits the convention, but metadata may reveal it.

Suppose, only for illustration, that the agent assigns the current proposal success probability $q = 0.80$. Let an incorrect transformation have loss 10 and a correct transformation have loss zero. Immediate execution then has expected loss 2.

A metadata inspection costs 0.5. Assume that after inspection, the agent's optimal revised decision has expected error loss 0.5. The total expected loss becomes 1.0, giving a net improvement of 1.0 under these assumptions.

If the metadata is unavailable, the agent might ask the user which convention applies. Whether that is preferable depends on expected response quality, delay, and the cost of interruption. If neither source can resolve the issue, abstention can be appropriate.

All numbers are hypothetical and demonstrate the decision comparison. They are not measured behavior or a proposed universal cost scale.

Now change the task so that the transformation writes only a temporary preview and can be undone at negligible cost. An exploratory operation may be reasonable despite the same uncertainty. Change it again so that the input convention is explicit and current metadata confirms it. Asking the user would now impose unnecessary burden.

This set of variations motivates a paired environment design. The system should respond to the availability and consequence of evidence, not simply to the presence of an ambiguity-related keyword.

9. Proposed research questions and hypotheses

RQ1 asks how well confidence signals predict action success at predefined workflow checkpoints. RQ2 asks whether agents choose information sources that reduce decision-relevant uncertainty. RQ3 asks whether targeted clarification improves completion relative to guessing or asking indiscriminately. RQ4 asks how error cost and reversibility change the best selection policy.

H1 proposes that outcome-specific confidence signals will rank action risk better than generic self-assessed task confidence on the selected environments. H2 proposes that giving agents a targeted inspection option will reduce unnecessary questions when the environment contains the needed information. H3 proposes that cost-aware selection will improve expected task loss relative to a fixed confidence threshold across heterogeneous consequences. H4 proposes that some calibration gains will fail to transfer across tools or task families without recalibration.

These are testable hypotheses, not findings. Their interpretation depends on the quality of outcome labels, the stated cost model, and the information available to each comparison system.

10. Proposed environment-based study

10.1 Task families

We propose executable tasks in repository repair, terminal operations, and scientific data processing. Each task includes a legitimate objective, a defined information boundary, and one or more decision checkpoints. The study should use fictional user interactions and synthetic or appropriately licensed assets.

Task families should vary uncertainty source independently where possible. One variant can omit user intent while preserving complete environmental evidence. Another can specify intent but hide a state fact that a permitted tool reveals. A third can make the critical information unavailable within the assigned budget.

10.2 Paired conditions

Useful pairs include ambiguous versus fully specified instructions, accessible versus inaccessible evidence, low-cost versus high-cost mistakes, reversible versus irreversible effects, and relevant versus uninformative tool responses. The task objective should remain comparable across variants.

These pairs create opportunities to examine behavioral adaptation. An agent that asks in every variant may recognize uncertainty language without selecting appropriately. An agent that acts in every variant may complete easy tasks while failing to account for changed consequences.

10.3 Comparison systems

Suggested conditions include immediate execution, a fixed confidence threshold, a prompt encouraging uncertainty expression, a confidence-based clarification policy, and a policy that can inspect, clarify, or abstain. Compare confidence estimators separately from decision policies when possible.

Hold the underlying model, tool implementations, context budget, and trial budget fixed. Record costs introduced by additional sampling, external predictors, tool calls, and simulated user responses. A stronger result obtained with more resources should disclose that difference.

10.4 Reference outcomes and unavailable counterfactuals

A task must define what counts as correct at the relevant checkpoint. For a deferred proposal, its hypothetical outcome may be unobserved. The study should not silently label abstention as correctness or incorrectness.

Controlled replay or a separate sandboxed execution can label a frozen proposal where appropriate. Such evaluation must not disclose its result to the agent before the recorded confidence or selection decision. If a counterfactual cannot be established, report that missingness and limit the corresponding claim.

10.5 Separation and statistical analysis

Reserve tools, repositories, or task families according to the intended transfer claim. Tune confidence transformations and decision thresholds on a calibration split, then freeze them for held-out assessment.

Repeated trials should characterize stochastic variation. Paired task-level comparisons or cluster-aware resampling can account for related variants. Report confidence intervals and subgroup outcomes, alongside the number of valid tasks and trials. Label exploratory comparisons accordingly.

11. Measurement beyond answer accuracy

11.1 Probabilistic quality

For a binary event with observed label y_i and predicted probability q_i , the binary Brier score is:

$$\text{Equation 7. Brier score: } BS = (1/N) \times \sum_i (q_i - y_i)^2.$$

This proper scoring rule evaluates probabilistic forecasts, with lower values preferred. Its origins lie in Brier's work on probability forecast verification. [9] It should be accompanied by calibration plots or tables and measures relevant to selection, because a single score does not fully characterize subgroup behavior or the risk ranking needed for deferral.

Expected calibration error can be reported with the binning method disclosed, but estimates can depend on sample size and bin choices. Reliability analysis should include uncertainty around observed frequencies. A small aggregate calibration error does not establish an individual-case guarantee.

11.2 Behavioral and operational outcomes

Measure	What it examines	Important qualification
Selective risk and coverage	Error among decisions accepted without more information	State the selection unit and treatment of zero coverage
Final completion	Success after any permitted inspection or clarification	Distinguish initial selection from eventual completion
Necessary clarification recall	Questions asked when intent information is required	Requires a task-defined label and consistent information access
Unnecessary question rate	Questions on cases answerable from available evidence	Distinguish useful confirmation from redundant questioning
Evidence efficiency	Useful uncertainty reduction per observation cost	Define utility through task outcomes or reference decisions
Abstention quality	Appropriate stopping and a useful account of missing information	A confident refusal is not proof of appropriate abstention
Expected task loss	Errors and intervention costs under a stated loss model	Report sensitivity to alternative cost assumptions

Question quality should be assessed separately from question frequency. A system can ask fewer questions by guessing, or ask more without resolving the issue. Relevant assessment includes whether the question identifies the missing distinction, whether a user can answer it, and whether the response changes the final decision appropriately.

Similarly, tool use should be evaluated through what it establishes. More observations can increase cost while leaving the decision unsupported. An efficient agent seeks evidence that matters to the objective.

12. What custom datasets can contribute

Uncertainty-aware behavior requires examples in which several responses are possible and their value depends on the task conditions. A static final answer can miss the distinction between a sensible decision and a lucky guess.

A useful dataset can therefore represent the objective, evidence available at each checkpoint, optional information actions, user knowledge, and consequences. Task variants can change one feature while preserving others, creating a controlled comparison.

Domain expertise matters when defining these distinctions. In a scientific workflow, an expert can identify which assumption changes the method choice. In repository work, an

expert can determine what evidence distinguishes two possible fault causes. In terminal operations, an expert can specify whether a proposed inspection is informative or redundant.

Deaimer publicly describes custom environments and private task collections spanning terminal operations, software engineering, scientific computing, and longer workflows. [10, 11] These capabilities provide a basis for discussing custom uncertainty-dependent scenarios within an agreed scope.

The proposed application is to construct tasks in which acting, inspecting, clarifying, and abstaining have observable consequences. It does not imply that Deaimer has released a calibrated-agent benchmark or measured the performance of the policy proposed here.

For a model developer, a concrete brief might specify the uncertainty source, accessible evidence, intended behavior, and transfer boundary. An empirical study would then establish whether a particular training collection or agent policy improves the relevant outcomes.

13. Limitations and threats to validity

The contribution is conceptual. Equations 4-6 assume a defined loss model and beliefs whose quality may be difficult to establish. They illustrate a decision structure without providing an estimator that solves the practical problem.

Outcome labels can also be incomplete. A test may confirm one property while leaving other consequences unexamined. Confidence should be interpreted relative to the event actually labeled, rather than expanded into a claim about overall task safety.

Simulated users can respond more consistently than real users. They may reveal precisely the information a benchmark author intended, making clarification look more effective than it would be under realistic ambiguity or delay. Human studies and sensitivity analysis can help examine that boundary.

Evidence sources can be noisy, outdated, or misleading. Inspection is useful only when observations are interpreted appropriately. An agent can acquire more text without reducing the relevant uncertainty.

Calibration and selection can shift across tools, task families, languages, and operating conditions. Standard conformal coverage does not automatically establish coverage along an adaptive sequence of tool interactions. Any statistical guarantee needs its sampling assumptions and covered event stated explicitly.

Cost models are another source of sensitivity. The same error can matter differently to different users. A universal threshold cannot represent every preference or operational consequence. Studies should report several plausible cost settings where feasible.

Finally, appropriate abstention is not identical to a refusal policy. A system can lack sufficient evidence for an allowed task, or possess strong evidence for an operation it is not authorized to perform. Confidence-sensitive behavior must coexist with separate permission boundaries.

14. Discussion and conclusion

Uncertainty in AI agents should be connected to the information and consequences of their decisions. A confidence score can help, but the useful question is what the agent does next: proceeds with support, inspects a relevant source, asks a targeted question, or stops with an explanation of what is missing.

This paper proposes a framework for studying those routes together. The formal comparisons show why error cost and information value influence appropriate behavior. The proposed protocol examines those influences through paired environments while preserving distinctions between initial selection, final completion, and interaction burden.

Future work can implement the framework, compare confidence estimators, and test its hypotheses across held-out tools and task structures. Especially useful studies would examine cases in which an agent can recognize uncertainty yet fails to obtain relevant evidence, or asks a sensible question but ignores the answer.

Custom executable tasks can support this agenda by making information access and consequence explicit. Dataset value then arises from the quality of the decision opportunities represented, rather than from uncertainty-related wording alone.

The research objective is selective competence: agents that use their capabilities productively while recognizing when additional evidence is worth obtaining. Demonstrating that competence requires observing decisions and outcomes together.

Declarations

Authorship and affiliation: Shehryar Awan, CTO, Deaimer.

Competing interests: The author is affiliated with Deaimer, which offers datasets and custom AI-agent environments. Proposed environment applications may be commercially relevant to that work.

Data and code availability: No new experimental dataset, implementation, or measured result accompanies this conceptual paper. Referenced resources belong to their respective authors and publishers.

Research status: The document presents a literature synthesis, formal specification, and proposed study. It is not presented as peer-reviewed or independently validated.

References

1. Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. International Conference on Machine Learning.
<https://arxiv.org/abs/1706.04599>
2. Geifman, Y., and El-Yaniv, R. (2017). Selective Classification for Deep Neural Networks. Advances in Neural Information Processing Systems.
<https://arxiv.org/abs/1705.08500>
3. Kadavath, S., et al. (2022). Language Models (Mostly) Know What They Know. arXiv:2207.05221.
<https://arxiv.org/abs/2207.05221>
4. Tian, K., et al. (2023). Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. EMNLP.
<https://aclanthology.org/2023.emnlp-main.330/>
5. Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature.
<https://doi.org/10.1038/s41586-024-07421-0>
6. Kuhn, L., Gal, Y., and Farquhar, S. (2022; revised 2023). CLAM: Selective Clarification for Ambiguous Questions with Generative Language Models. arXiv:2212.07769.
<https://arxiv.org/abs/2212.07769>
7. Zhao, J., Fang, K., and Cheng, L. (2026). When and What to Ask: AskBench and Rubric-Guided RLVR for LLM Clarification. arXiv:2602.11199.
<https://arxiv.org/abs/2602.11199>
8. Angelopoulos, A. N., and Bates, S. (2021; revised 2022). A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv:2107.07511.
<https://arxiv.org/abs/2107.07511>
9. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1-3.
https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml
10. Deaimer. RL Environments & Coding-Agent Evaluations. Public offering description. Accessed 14 October 2026.
<https://www.deaimer.com/rl-environments>
11. Deaimer. Private Coding-Agent Datasets for Licensing. Public offering description. Accessed 14 October 2026.
<https://www.deaimer.com/datasets>