Uncertainty in Action: Confidence Elicitation in Embodied Agents

Tianjiao Yu, Vedant Shah, Muntasir Wahed, Kiet A. Nguyen, Adheesh Juvekar
Tal August, Ismini Lourentzou
University of Illinois Urbana-Champaign

{ty41,vrshah4,mwahed2,kietan2,adheesh2,taugust,lourent2}@illinois.edu
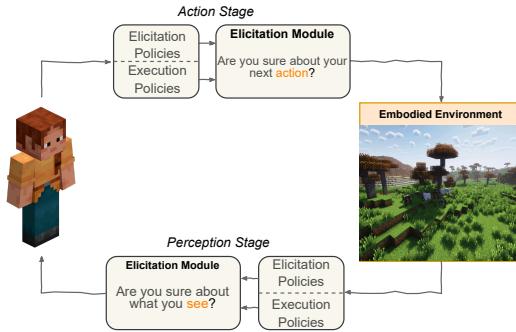https://plan-lab.github.io/ece

Abstract

Expressing confidence is challenging for embodied agents navigating dynamic multimodal environments, where uncertainty arises from both perception and decision-making processes. We present the first work investigating embodied confidence elicitation in open-ended multimodal environments. We introduce Elicitation Policies, which structure confidence assessment across inductive, deductive, and abductive reasoning, along with Execution Policies, which enhance confidence calibration through scenario reinterpretation, action sampling, and hypothetical reasoning. Evaluating agents in calibration and failure prediction tasks within the Minecraft environment, we show that structured reasoning approaches, such as Chain-of-Thoughts, improve confidence calibration. However, our findings also reveal persistent challenges in distinguishing uncertainty, particularly under abductive settings, underscoring the need for more sophisticated embodied confidence elicitation methods.

1. Introduction

In complex embodied environments, success depends not only on what an agent knows but also on how well it understands and communicates uncertainty. Whether navigating a cluttered space, interacting with objects, or planning long-term strategies, eliciting confidence is pivotal as agents must interpret and interact with dynamic settings in real-time while managing uncertainty from both perception and decision-making processes (Ren et al., 2023; Liang et al., 2024). For humans, this instinctive ability to express and calibrate uncertainty is fundamental to decision-making and social interaction. As AI systems are increasingly deployed in high-stakes contexts such as autonomous driving or healthcare, they must also acquire this crucial skill.

Figure 1. **Embodied Confidence Estimation Framework** consisting of *Elicitation Policies* and *Execution Policies*, which jointly enable an agent to assess and express its confidence. Elicitation Modules prompt the agent to evaluate uncertainty in what it sees and does, while *Execution Policies* refine confidence calibration by expanding the agent's reasoning space (See §3 for details).

Specifically, accurate confidence elicitation from AI systems provides critical insights for risk assessment, error mitigation, and system reliability in decision-making (Kuleshov & Deshpande, 2022; Clark, 2015; Yildirim et al., 2019). This is particularly important in open-ended reasoning tasks, where models may generate outputs that are semantically plausible but factually incorrect, a phenomenon commonly referred to as hallucination (Xiao & Wang, 2021). However, confidence elicitation in embodied AI is particularly challenging. For instance, in open-ended environments such as Minecraft, an agent may misinterpret visual cues due to limited viewpoints or struggle to determine the correct action sequence to achieve complex goals (*e.g.*, obtaining a diamond). These illustrate the broader difficulties in eliciting confidence in embodied environments, where agents must navigate uncertainty at multiple levels.

Confidence elicitation in open-ended embodied environments faces several challenges, including: 1) Multimodal understanding, where the agent must assess uncertainty from inputs across different interconnected modalities. 2) Granularity of confidence estimation, where the agent evaluates confidence not only in performing specific actions (*e.g.*, "I am 90% confident I can collect some wood") but also in

---

*Preprint. Work in progress.

understanding high-level tasks or goals (*e.g.*, "I am 70% confident I craft a wooden table"). 3) Interactive dependencies, where the agent's actions directly influence the environment, which in turn affects subsequent decisions, requiring ongoing adjustments to confidence estimates as tasks progress. 4) Finally, while state-of-the-art embodied agents leverage proprietary Large Language Models (LLMs) and Vision-Language Models (VLMs) for their strong multimodal understanding and reasoning capabilities (Wang et al., 2023a; Qin et al., 2024; Zhu et al., 2023), these often lack access to internal token likelihoods or probabilistic outputs, making traditional confidence estimation methods ineffective (Kumar et al., 2023; Chen et al., 2024b).

To address these challenges, we present the first systematic approach that enables LLM/VLM-powered embodied agents to assess and articulate their confidence across multimodal inputs, multiple granularities, and dynamic embodied environments. Our contributions are as follows: **(1)** We propose a framework for embodied verbalized confidence elicitation in multimodal open-ended environments. **(2)** As illustrated in Figure 1, we introduce Elicitation and Execution Policies to enhance confidence estimation in embodied settings. **Elicitation Policies** target different types of uncertainties arising from inductive, deductive, and abductive reasoning, while also facilitating multi-granular confidence estimation, allowing agents to assess uncertainty at both perception and action stages. **Execution Policies** improve robust elicitation across diverse scenarios, plans, and actions while tackling interactive dependencies by incorporating additional information about the environment and expanding potential action trajectories. **(3)** We provide the first structured analysis of embodied uncertainty and identify effective methods for improving confidence calibration and failure prediction, while also pinpointing persistent challenges.

The following are key observations from our analysis:
(1) Elicitation Policies are Effective But Vary by Context: While all proposed elicitation policies improve confidence calibration and failure prediction, their effectiveness varies based on task complexity and uncertainty type, highlighting the need for adaptive strategies that align with the embodied agent's reasoning process and environment demands.
(2) Execution Policies Amplify Reliable Embodied Confidence Elicitation: Execution policies enhance the robustness of elicited confidence as they expand the range of available actions and scenario interpretations, enabling agents to assess their confidence levels more effectively based on a broader set of potential outcomes.
(3) Model Differences Persist: While all models benefit from the proposed policies, differences in their inherent reasoning and representation capabilities lead to significant variability in confidence calibration and task success rates, highlighting the importance of tailoring elicitation and execution strategies to each model's strengths and limitations.
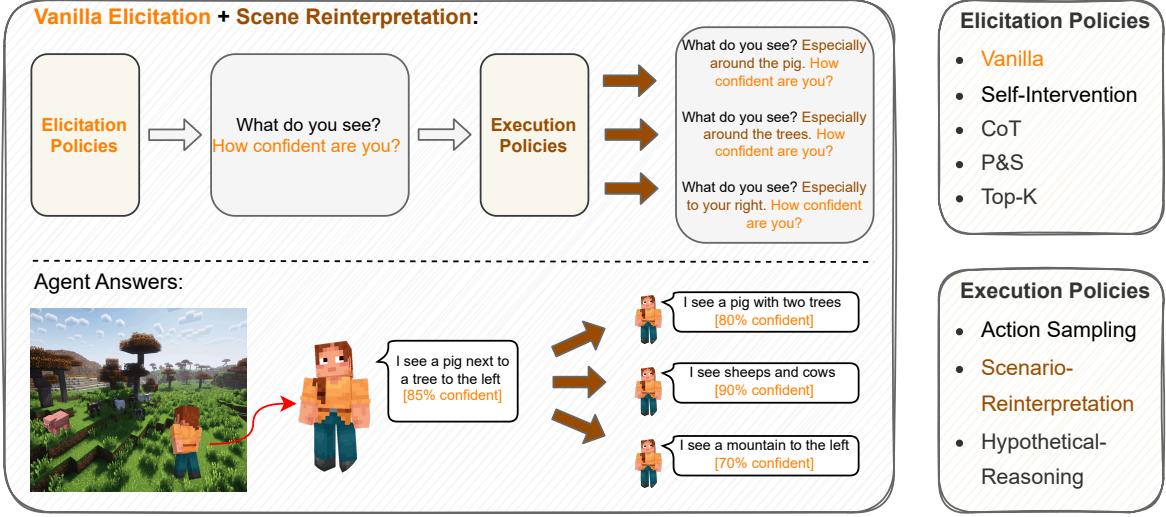
2. Related Works

Confidence Elicitation. Confidence elicitation for traditional machine learning is well-studied (Abdar et al., 2021; Gawlikowski et al., 2023). One stream of work focuses on unsupervised methods leveraging entropy (Malinin & Gales, 2021), graph semantic parsing (Lin et al., 2022b), semantic features (Kuhn et al., 2023; Farquhar et al., 2024), and logit or hidden state information (Su et al., 2024; Chen et al., 2024a) to craft uncertainty metrics. Another explores conformal prediction for tasks like part-of-speech tagging (Dey et al., 2022), paraphrase detection (Giovannotti & Gammerman, 2021), and fact verification (Fisch et al., 2021), offering statistically robust coverage guarantees (Kumar et al., 2023; Ye et al., 2024).

However, these solutions often require full model access, making them less applicable to black-box language models, which are increasingly prevalent in real-world applications (Achiam et al., 2023; Touvron et al., 2023a). Additionally, their free-form nature of outputs further complicates the application of traditional methods. As a result, alternative approaches have been proposed, including estimating uncertainty by directly querying models for confidence scores after generating responses (Xiong et al., 2024; Kadavath et al., 2022; Lin et al., 2022a; Mielke et al., 2022; Chen & Mueller, 2024). Despite these advancements, existing methods are not designed for embodied tasks, where confidence elicitation must address the challenges of multimodal perception, hierarchical reasoning and planning across various open-ended tasks, as well as non-deterministic interactions.

LLM-based Embodied Agents. With the advent of language models, leveraging their reasoning and planning abilities to empower embodied agents has become quintessential (Huang et al., 2023; Yao et al., 2023; Chen et al., 2023; Zhang et al., 2024a; Shinn et al., 2024; Christianos et al., 2023). In the meantime, Minecraft's open-ended nature with its adaptable mechanics and varied challenges, makes it a compelling benchmark for embedding reasoning and planning capabilities into language model-driven embodied agents (Wang et al., 2023a;c; Zhu et al., 2023). Recent works leverage pre-trained language models to control agents by generating continuous operation instructions or executable policies. For example, some approaches (Zhu et al., 2023; Wang et al., 2023c) directly utilize scene data from simulation platforms like MineDojo (Fan et al., 2022) and MineRL (Guss et al., 2019), while others (Qin et al., 2024) rely on Vision-Language Models (VLMs) for perception.

However, because language models are used in various roles—such as planners, critics, or perceivers—errors and inaccuracies often arise at different process stages (Guo et al., 2024; Driess et al., 2023). These challenges underscore the need for frameworks capable of systematically identifying and localizing sources of uncertainty, which we aim

Figure 2. **Embodied Confidence Elicitation.** *Elicitation Policies* (§3.2) enable agents to express uncertainty, while *Execution Policies* (§3.3) refine and expand confidence assessment through scenario reinterpretation, action sampling, and hypothetical reasoning. Together, they enhance confidence calibration in embodied agents. The orange text represents the vanilla elicitation policy, which incorporates the vanilla confidence prompt (described in Table 1) into the original instruction. The brown arrows ➡ denote the Scenario-Reinterpretation execution policy, prompting the agent to generate additional scene insights.

to address by designing a unified approach that enhances reliability and robustness in embodied agents.

Uncertainty in Embodied Models. Uncertainty estimation is well explored in robot learning and reinforcement learning (Wang & Zou, 2021; Ghasemipour et al., 2022; He et al., 2023; Huang et al., 2019; Jin et al., 2023), but remains a challenge for language models (Tian et al., 2023; Groot & Valdenegro Toro, 2024; Zhang et al., 2024b). While recent efforts have sought to quantify and mitigate uncertainty (Sagar et al., 2024; Tian et al., 2022), the problem is further compounded in embodied AI settings, where agents must reason and act in dynamic multimodal environments (Ren et al., 2024; Shen & Lourentzou, 2025). Our work introduces a structured approach to verbalized confidence elicitation in embodied open-ended multimodal environments to enable agents to express uncertainty and adapt to complex real-world interactions.

3. Method

3.1. Problem Formulation & Framework Design

Let $\mathcal{E}$ denote the embodied environment, characterized by multimodal sensory inputs $\mathcal{I} = \{\mathcal{I}_v, \mathcal{I}_t\}$, where $\mathcal{I}_v$ represents visual observations and $\mathcal{I}_t$ represents task instructions and other types of language-based guidance. For a given task $\mathcal{T}$, the agent operates under a policy $\pi : \mathcal{I} \rightarrow \mathcal{A}$ that maps input $\mathcal{I}$ to actions $\mathcal{A}$. The task of embodied confidence elicitation is to enable agents to estimate and articulate a confidence score $c \in [0, 1]$, representing their belief in the correctness of their perception and subsequent actions.

The challenge lies in systematically identifying, quantifying, and articulating uncertainty as the agent interacts with its environment and executes tasks. This requires not only detecting uncertain aspects of the agent's perception, reasoning, or actions but also ensuring that confidence estimates are refined and reliable under dynamic multimodal conditions. To address this, we propose an embodied confidence estimation framework centered around **Elicitation Modules** that facilitates confidence elicitation at two critical points of interaction between the agent and its environment: **Perception Stage**, where the agent processes sensory input from the environment and assesses its confidence in what it perceives before engaging in reasoning or planning. **Action Stage**, which evaluates the agent's confidence after reasoning, just before executing an action.

Each Elicitation Module operates under a specific **Elicitation Policy** (§3.2), which defines *what* type of uncertainty is being expressed, focusing on quantifying confidence in the agent's perception, reasoning, or action planning. Additionally, an **Execution Policy** (§3.3) determines *how* to collect and refine confidence, ensuring robust and adaptive estimates in complex, dynamic environments. An overview of the overall proposed method is shown in Figure 2.

3.2. Elicitation Policies

Our confidence Elicitation Policies are designed to address distinct types of inferential uncertainty that embodied agents encounter in open-world long-horizon tasks. As these agents actively reason to determine their next actions, we draw inspiration from rich studies on reasoning in language models

Method	Prompt
Vanilla	Read the task (e.g., collect wood, build a shelter), provide your answer, and explain how confident you are in perceiving the environment accurately to complete the task (e.g., recognizing resources, locating structures, identifying threats). Read the task given, provide your answer, and explain how confident you are in planning and executing the actions needed to achieve the goal (e.g., gathering materials, crafting tools, building a structure).
Vanilla + Self-Intervention	Task: [...], Perceived Situation: [...] Q: How confident you are in perceiving the environment accurately to complete the task? Task: [...], Planned Action: [...] Q: How confident you are in planning and executing the actions needed to achieve the goal?
Chain-of-Thought (CoT)	Read the task, analyze step by step what you perceive in the environment (e.g., observe surroundings, identify items), provide your answer, and evaluate your confidence based on the clarity and quality of the environment observations. Read the task, analyze step by step how to complete the task, provide your answer, and evaluate your confidence in successfully planning and executing each action needed to achieve the goal.
Plan & Solve (P&S)	Analyze the task, devise a systematic approach to perceive your environment effectively. (e.g., locating resources, identifying obstacles), and evaluate your confidence based on how well you perceive the environment. Analyze the task, devise a plan of actions needed to complete it, then evaluate your confidence in executing each action and achieving the desired outcome.
Top-K	Provide your K best descriptions of your perceptions of the environment and the probability that each is correct (0% to 100%). Provide your K best plans of the possible actions to take and the probability that each will succeed (0% to 100%).

Table 1. **Prompts for Different Elicitation Policies** in generalist embodied Minecraft agents. Orange text indicates prompts focused on perception, while blue text highlights prompts centered on action and planning.

(Huang & Chang, 2023) and introduce five prompt instructions, comprising two general-purpose methods and three tailored to inductive, deductive, and abductive reasoning settings (Appendix A). These prompts ask the agent to verbalize its confidence levels and systematically refine its uncertainty. Table 1 provides an overview of elicitation policy types with corresponding examples.

◇ **Vanilla.** Leveraging the inherent capability of language models (Brown et al., 2020; Wei et al., 2022a), the Vanilla method directly queries the agent's confidence without additional structure or intervention. Vanilla serves as a baseline for comparison, relying solely on the agent's built-in capacity for confidence elicitation and self-assessment.

◇ **Self-Intervention**. Humans naturally benefit from revisiting their decisions with a fresh perspective, often uncovering insights or errors they initially overlooked. Inspired by this, the self-intervention method separates answer generation from evaluation. In one session, the model generates an answer; in another, it revisits the question and its response to assess its accuracy. This independent second pass mitigates confirmation bias and overconfidence, encouraging critical self-reflection and producing more reliable evaluations.

◇ **Chain-of-Thought (CoT)**. To address uncertainty in inductive reasoning settings, where the agent must identify patterns and infer relationships from observations, we employ zero-shot Chain-of-Thought (CoT) reasoning (Wei et al., 2022b). By decomposing tasks into incremental steps, CoT enhances both interpretability and confidence calibration, allowing agents to reassess uncertainty at each step.

◇ **Plan & Solve (P&S)**. Despite the success of CoT, it often suffers from semantic misunderstandings and missing step errors, particularly when applying general principles to specific cases. These failures stem from uncertainty in deductive reasoning, where the agent is unsure about the correct instantiation of abstract rules or whether a logical step is valid in a given context. P&S (Wang et al., 2023b) mitigates this by explicitly separating planning from execution, prompting the agent to construct a structured reasoning blueprint before solving the problem step by step.

◇ **Top-K**. To address uncertainty in abductive reasoning, where multiple plausible explanations may fit the observed data, the Top-K method prompts the agent to generate its top K answers, each with an associated confidence level. This encourages the agent to consider and distribute its attention across several possible outcomes. By ranking responses rather than a single definitive answer, Top-K provides a balanced and comprehensive representation of abductive uncertainty across multiple plausible interpretations.

3.3. Execution Policies

In embodied contexts, planning is a key factor in task success, requiring the agent to assess its confidence in executing action sequences effectively. Dynamic environments introduce unpredictable factors in action outcomes, which makes it important for the agent to not only consider its primary course of action but also to evaluate and communicate its confidence in alternative actions. By analyzing variance across potential actions rollouts, the agent can bet-

4

ter quantify uncertainty and anticipate divergent outcomes. To address this, we introduce a set of policies that generate additional observations and diverse action trajectories, promoting robust confidence assessment:

↻ **Action Sampling**: The agent can generate multiple possible actions by sampling from a learned policy distribution over the action space, conditioned on the current state and task objectives. By doing so, the agent can explore multiple actions, evaluate different outcomes, and assess which is most likely to succeed based on its perception.

↻ **Scenario Reinterpretation**: The agent can be prompted to reinterpret the same scenario from different perspectives. For example, it could focus on a particular object, re-evaluate environmental obstacles, or re-assess the proximity of targets. This enables the agent to propose different courses of action by gathering and redirecting its attention to relevant environmental information.

↻ **Hypothetical Reasoning**: The agent can be prompted with hypothetical or counterfactual scenarios (*e.g.*, "What if the object in front were not an obstacle?"). By simulating these hypotheticals, the agent can explore how its actions would change and assess confidence in its original plan. This helps to gauge how flexible the agent's decision-making process is when confronted with uncertainty or alternative interpretations of the environment.

Figure 2 provides an overview and examples of Elicitation and Execution Policies. During task-solving, agents rely on these execution policies to gather additional information about the environment and potential action trajectories, which they incorporate into further confidence elicitation.

4. Experiment Setup

Environment & Task Setting. Minecraft has emerged as a popular benchmark for embodied AI research due to its open-ended environment, with diverse terrains, resources, and open-ended goals, making it an ideal testbed for embodied agents that perform hierarchical reasoning and long-term planning (Johnson et al., 2016; Guss et al., 2019; Hafner et al., 2023; Nottingham et al., 2023; Lin et al., 2023; Qin et al., 2024). Building on this foundation, we define 30 tasks evenly distributed across three difficulty levels: easy, medium, and hard, based on the complexity of reasoning steps and the amount of contextual information required. Detailed task descriptions can be found in Appendix B.

Easy tasks typically involve basic interactions with a single environmental element (*e.g.*, locating a pig or observing the weather). Medium tasks require combining perception and reasoning over multiple elements, while hard tasks increase dependency on sequential reasoning and include complex challenges like the *Diamond Challenge*, which requires

long-term planning and multi-step execution. Following prior work (Guss et al., 2019), the maximum episode length is set to 6000 steps. Privileged observation is used as the ground truth for perception, while overall task success rate serves as the ground truth for planning and reasoning.

Evaluation Metrics. To assess the reliability of confidence estimates, we evaluate two key aspects: calibration and failure prediction (Naeini et al., 2015; Yuan et al., 2021). Calibration measures how well an agent's expressed confidence reflects its actual performance, *e.g.*, an 80% confidence should ideally correspond to 80% accuracy. This calibration is crucial for applications requiring robust risk assessment and trustworthiness. On the other hand, failure prediction focuses on the agent's ability to distinguish between correct and incorrect predictions by assigning higher confidence to correct outcomes. We use the Expected Calibration Error (ECE) to quantify calibration quality and the Area Under the Receiver Operating Characteristic Curve (AUROC) to evaluate failure prediction. To address imbalances stemming from varying accuracy levels across tasks, we also include AUPRC-Positive (PR-P) and AUPRC-Negative (PR-N), which separately measure the agent's effectiveness in identifying correct and incorrect predictions.

Minecraft Agents. In this work, we focus on embodied agents powered by advanced Large Language Models (LLMs) and Vision-Language Models (VLMs) that enable multimodal reasoning and understanding in complex embodied environments. We employ three models as the agent's backbone: **(1) GPT-4V**, chosen for its strong performance in multimodal reasoning and proven effectiveness in complex environments like Minecraft (Wang et al., 2023a; Qin et al., 2024; Li et al., 2025) for planning and perception tasks. **(2) MineLLM** (Qin et al., 2024), a model specifically designed for Minecraft tasks, that leverages MineCLIP's visual encoder and Vicuna-13B (Chiang et al., 2023) to deliver robust multimodal understanding. and **(3) STEVE**, built on the versatile LLaMA framework, STEVE models excel in contextual understanding and decision-making (Zhao et al., 2025). Fine-tuned for Minecraft, STEVE enhances planning, communication, and interaction capabilities. Detailed model descriptions are provided in Appendix C.

5. Experimental Results

Table 2 presents the performance of benchmarked agents across all Elicitation Policies. In this experiment, evaluation is conducted without Execution Policies. The final confidence scores are computed as the average of individual step confidence scores across five independent task episodes.

All Elicitation Policies Facilitate Better Calibration and Failure Prediction. Across all models, Elicitation Policies consistently improve calibration (lower ECE) and failure

Metric	Model	Vanilla	Self-Intervention	CoT (Inductive)	P&S (Deductive)	Top-K (Abductive)
ECE ↓	GPT-4V	0.27	0.21	0.16	**0.15**	0.17
	MineLLM	0.49	0.41	**0.34**	0.39	0.43
	STEVE	0.43	0.32	**0.26**	**0.26**	0.35
AUROC ↑	GPT-4V	0.69	0.76	**0.83**	0.82	0.73
	MineLLM	0.53	0.59	**0.64**	0.61	0.58
	STEVE	0.58	0.69	**0.72**	0.67	0.68
PR-P ↑	GPT-4V	0.66	0.76	**0.81**	0.79	0.70
	MineLLM	0.51	0.59	**0.63**	0.60	0.57
	STEVE	0.56	0.67	**0.69**	0.66	0.64
PR-N ↑	GPT-4V	0.52	0.53	**0.58**	0.55	0.53
	MineLLM	0.39	0.42	0.42	**0.43**	0.40
	STEVE	0.41	0.46	**0.46**	0.43	0.42

Table 2. **Confidence Metrics across Elicitation Policies** with three models (GPT-4V, MineLLM, and LLaMA-based STEVE) using different elicitation strategies: Vanilla (basic task understanding), Self-Intervention (reflection on own actions), Chain-of-Thought (step-by-step reasoning), Plan & Solve (explicit planning before execution), and Top-K (confidence distribution across multiple outputs) with No Execution Policies applied. The best performance across each model is in **bold**.

prediction (higher AUROC, PR-P, PR-N) compared to the Vanilla baseline. For instance, in GPT-4V, every Elicitation Policy results in a lower ECE and higher AUROC relative to Vanilla, demonstrating their effectiveness in improving the robustness of uncertainty quantification. Likewise, MineLLM and STEVE exhibit noticeable gains in ECE and AUROC when incorporating elicitation mechanisms, confirming that Elicitation Policies help agents better assess uncertainty and predict incorrect responses.

Structured Elicitation (CoT and P&S) Improves Calibration and Failure Prediction the Most. Among the four Elicitation Policies, structured reasoning approaches—CoT (Inductive) and P&S (Deductive)—consistently yield the best calibration and failure detection performance. For example, in GPT-4V, P&S achieves the lowest ECE (0.15) and one of the highest AUROC scores (0.82), while CoT further improves AUROC up to 0.83. Similar trends hold for MineLLM and STEVE, where CoT and P&S outperform Self-Intervention and Top-K across nearly all metrics. These improvements suggest that breaking down reasoning into explicit steps helps the models maintain logical consistency, facilitating better overall calibration.
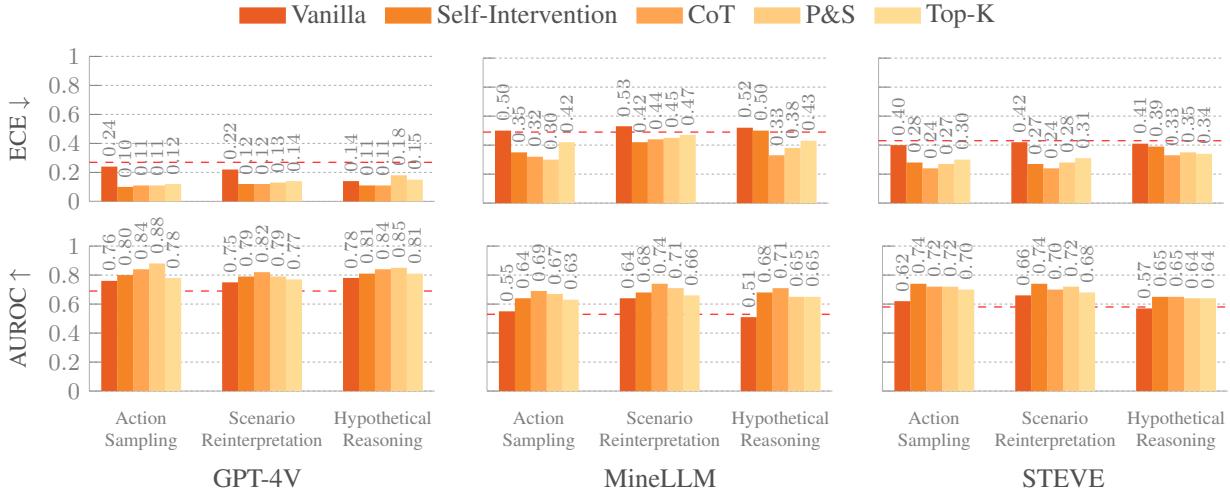
Abductive Reasoning Poses Greater Challenges than Inductive and Deductive. While Top-K (Abductive) improves over the Vanilla policy, it exhibits weaker calibration and failure prediction, suggesting that generating multiple plausible interpretations increases uncertainty misalignment, and therefore making it harder for the model to distinguish between correct and incorrect predictions. Additionally, the lower PR-P and PR-N scores indicate that confidence estimation for abductive reasoning is more difficult to calibrate compared to inductive and deductive settings.

Confidence Calibration Remains Inconsistent Across Models. While GPT-4V consistently benefits from different Elicitation Policies, the improvements are less stable in fine-tuned models like MineLLM and STEVE. For instance, CoT boosts AUROC to 0.83 in GPT-4V but only reaches

0.64 in MineLLM and 0.72 in STEVE, indicating that fine-tuned models struggle to generalize confidence estimation effectively. One likely reason for this inconsistency is that MineLLM and STEVE, being fine-tuned models, exhibit degenerated language capabilities, limiting their ability to verbalize uncertainty reliably.

Execution Policies Amplify Reliable Embodied Confidence Across Elicitation Policies. Figure 3 illustrates how Execution Policies interact with Elicitation Policies. Overall, Execution Policies are capable of further improving calibration and failure prediction performance. For example, GPT-4V achieves better ECE results when pairing any Execution Policy with all Elicitation Policies. More specifically, structured reasoning approaches such as CoT (Inductive) and P&S (Deductive), when paired with Action Sampling, tend to yield improved confidence calibration. For instance, MineLLM's ECE achieves 0.32 and 0.30 paired with CoT and P&S respectively, outperforming other combinations. Hypothetical Reasoning sometimes degrades performance. For instance, STEVE's ECE worsens when pairing Hypothetical Reasoning with all Elicitation Policies, suggesting that while this execution strategy allows models to reason over multiple possible outcomes, it may introduce uncertainty, leading to less calibrated confidence judgments.

So, How Effectively Can Embodied Agents Express Confidence in Dynamic Embodied Tasks? While embodied agents can convey confidence to some extent, their effectiveness depends on how well they integrate reasoning, uncertainty assessment, and environmental interactions. The findings reveal that embodied confidence elicitation remains a challenging problem, requiring a careful balance between general-purpose reasoning and task-specific specialization. However, our proposed Elicitation Policies improve both confidence calibration and failure prediction, while our Execution Policies further augment these performance gains by refining uncertainty through iterative interactions with the environment. These results highlight the importance of

Figure 3. **ECE and AUROC across Models and Execution Policies.** Bars present ECE (top, lower is better) and AUROC (bottom, higher is better) under different elicitation strategies. **Red dashed lines** are metrics for Vanilla elicitation with no execution policy applied.

accounting for the unique challenges faced by embodied agents in confidence estimation, emphasizing the need for execution-aware strategies that enhance both calibration and failure prediction in complex environments.

6. Ablation Studies

Impact of Execution Policies. We analyze the performance of Execution Policy combinations, incorporating Action Sampling (AS), Scenario Reinterpretation (SR), and Hypothetical Reasoning (HR) both incrementally and collectively. Results in Table 3 show clear trends in how Execution Policies influence performance. Without any Execution Policies, Vanilla Elicitation exhibits the worst calibration, with ECE as high as 0.27, while also struggling with failure prediction. When Execution Policies are introduced, performance improves, though trade-offs emerge between failure prediction accuracy (AUROC) and confidence calibration (ECE). Among two-policy combinations, the combination of Action Sampling with Scenario Reinterpretation (AS + SR) delivers the most balanced improvement, significantly increasing AUROC (up to 0.83 for GPT-4V and 0.69 for STEVE) while maintaining the lowest ECE (0.17 for GPT-4V, 0.32 for MineLLM). This suggests that jointly exploring multiple action paths and reinterpreting environmental cues helps refine confidence estimation without sacrificing calibration.

In addition, strategies incorporating Action Sampling (AS) consistently outperform those without it, resulting in better uncertainty estimation and more reliable confidence scores. By generating multiple action plans, AS enhances confidence calibration, underscoring the importance of addressing action planning uncertainty in embodied agents. Combining all Execution Policies yields the strongest overall performance across models, achieving the highest AUROC across all three models while maintaining competitive cali-

bration, with the lowest ECE for GPT-4V (0.17) and strong values for MineLLM (0.32) and STEVE (0.38). This suggests that integrating Action Sampling, Scenario Reinterpretation, and Hypothetical Reasoning provides a complementary effect, improving both failure prediction accuracy and confidence estimation.

Perception *v.s.* Cognition. Embodied agents, when tasked with high-level objectives (*e.g.*, "find a pig"), often rely on language models to decompose the task into smaller, granular actions (*e.g.*, "step forward 2 steps"). During task execution, the agent generates confidence scores for each granular action. Typically, these scores are aggregated temporally to produce a single overall confidence score for the entire task. While this method provides a holistic measure of confidence, it does not differentiate between the confidence associated with perception (*e.g.*, recognizing a pig) and cognition (*e.g.*, reasoning about the sequence of steps).

To better understand how different sources of uncertainty contribute to overall confidence, we separately analyze perception and reasoning confidence. Perception Confidence aggregates scores related to the agent's ability to interpret its sensory inputs (*e.g.*, detecting objects or understanding environmental cues), while Reasoning Confidence aggregates scores associated with reasoning and decision-making processes during task execution. Figure 4 reveals that temporal aggregation achieves the lowest ECE (0.18) and a balanced AUROC (0.76). Temporal aggregation smooths over individual uncertainties, providing robust overall calibration and reliable failure prediction.

Perception-based confidence, when aggregated separately, offers a distinct advantage in predictive reliability. With an AUROC of 0.79, the highest among the methods, and strong PR-P (0.85) and PR-N (0.81) scores, perception confidence consistently outperforms reasoning. This highlights the

Execution Strategies	GPT-4V		MineLLM		STEVE	
	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
No Execution Strategy	0.27	0.69	0.49	0.53	0.43	0.58
AS + SR	0.18	0.82	**0.32**	0.59	0.39	0.69
AS + HR	0.20	0.79	0.34	0.57	**0.37**	0.66
SR + HR	0.22	0.80	0.37	0.54	0.44	0.58
AS + SR + HR	**0.17**	**0.83**	**0.32**	**0.62**	0.38	**0.69**

Table 3. **Performance of Vanilla Elicitation with Combined Execution Strategies. AS** = Action Sampling, **SR** = Scenario Reinterpretation, **HR** = Hypothetical Reasoning. ECE and AUROC for each model, GPT-4V, MineLLM, and STEVE. Best values highlighted in **bold**.
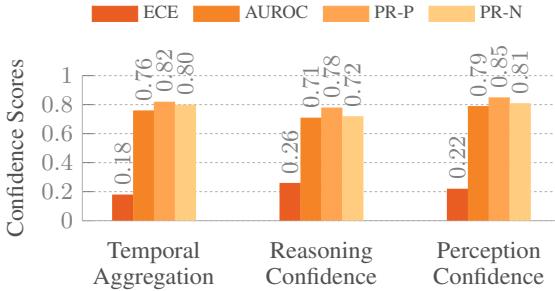

Figure 4. **Comparison of Aggregation Methods for Vanilla Elicitation without Execution Policies.** Temporal aggregation provides a holistic score, while separate aggregation evaluates confidence in reasoning and perception separately respectively.

Figure 5. **ECE and AUROC across Execution Policy Iterations.**

inherent stability of sensory tasks, where clear input-output mappings and deterministic operations reduce uncertainty. Additionally, perception confidence maintains a competitive ECE (0.22), indicating that it remains well-calibrated.

In contrast, reasoning confidence introduces more uncertainty, resulting in a higher ECE (0.26) and a lower AUROC (0.71). These results reflect the challenges of reasoning tasks, which often involve multi-step decision-making and are susceptible to cascading errors. The lower PR-P (0.78) and PR-N (0.72) scores suggest reasoning confidence struggles to accurately distinguish correct from incorrect outcomes. In essence, results affirm that reasoning tasks inherently present greater uncertainty, requiring more sophisticated calibration methods to maintain reliability.

Interestingly, the performance gap between perception and reasoning confidence underscores their complementary nature. While perception excels in calibration and failure prediction, reasoning provides critical insights into decision-making under uncertainty. Temporal aggregation balances these components effectively for an overall confidence score but sacrifices the interpretability offered by separate aggregation. This comparison emphasizes the need to align aggregation methods with task complexity and performance priorities, whether for holistic confidence measures or detailed insights into perception and cognition.

Impact of Execution Iterations. We investigate the impact of repeatedly applying execution policies on calibration and failure prediction accuracy. Iterations range from 0 (*i.e.*, no execution policies employed) to 15, allowing for an analysis
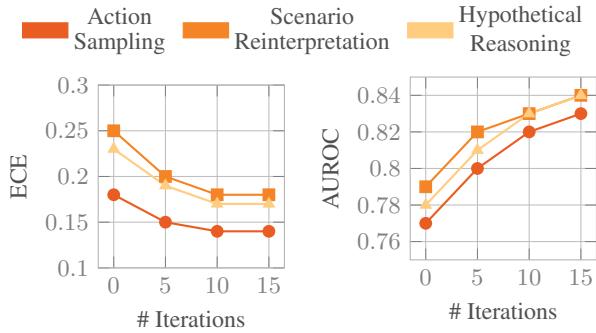
of both the initial benefits and potential diminishing returns of repeated applications. As shown in Figure 5, repeated applications initially improve ECE across all policies but eventually plateau. For instance, Action Sampling reduces ECE from 0.18 (at 0 iterations) to 0.14 (at 15 iterations), with most of the improvement occurring within the first 10 iterations. A similar trend is observed for Scenario Reinterpretation and Hypothetical Reasoning, where ECE drops from 0.25 to 0.18 and from 0.23 to 0.17, respectively. The plateau effect is less pronounced in AUROC, which consistently improves across iterations. Action Sampling increases AUROC from 0.77 to 0.83, while Scenario Reinterpretation and Hypothetical Reasoning improve from 0.79 to 0.84 and from 0.78 to 0.84, respectively. Most AUROC gains occur between 0 and 10 iterations, with diminishing returns after 15 iterations. Overall, early iterations improve calibration and failure prediction, but excessive repetition yields diminishing returns. This underscores the need to balance execution policy applications for optimal effectiveness.

7. Conclusion

This work presents the first systematic exploration of embodied confidence elicitation, introducing elicitation and execution policies that enhance calibration and failure prediction in open-ended multimodal embodied tasks. Our findings highlight improvements in confidence estimation using our proposed methods, providing more accurate uncertainty quantification. Future research could improve confidence elicitation in embodied environments by scaling to more diverse and complex environments and exploring their integration with various embodied agent architectures.

8. Impact Statement

This work advances Embodied AI by introducing confidence elicitation and execution policies tailored to multimodal and dynamic environments. By enabling embodied agents to express uncertainty, our approach enhances their calibration, adaptability, and reliability in complex tasks. This contribution supports safer AI deployment in real-world domains like robotics, education, and collaborative systems, where accurate self-assessment is critical. However, the reliance on large pre-trained models raises concerns about energy efficiency and ethical considerations in high-stakes applications, which warrant further exploration.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 2021.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Chen, B., Shu, C., Shareghi, E., Collier, N., Narasimhan, K., and Yao, S. Fireact: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*, 2023.

Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., and Ye, J. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, 2024a.

Chen, J. and Mueller, J. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Association for Computational Linguistics*, 2024.

Chen, J., Park, S., and Simeone, O. Knowing when to stop: Delay-adaptive spiking neural network classifiers with reliability guarantees. *IEEE Journal of Selected Topics in Signal Processing*, 2024b.

Cheng, K., Yang, J., Jiang, H., Wang, Z., Huang, B., Li, R., Li, S., Li, Z., Gao, Y., Li, X., et al. Inductive or deductive? rethinking the fundamental reasoning abilities of llms. *arXiv preprint arXiv:2408.00114*, 2024.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://lmsys.org/blog/2023-03-30-vicuna/*, 2023.

Christianos, F., Papoudakis, G., Zimmer, M., Coste, T., Wu, Z., Chen, J., Khandelwal, K., Doran, J., Feng, X., Liu, J., Xiong, Z., Luo, Y., Hao, J., Shao, K., Bou-Ammar, H., and Wang, J. Pangu-agent: A fine-tunable generalist agent with structured reasoning. *arXiv preprint arXiv:2312.14878*, 2023.

Clark, A. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, 2015.

Dey, N., Ding, J., Ferrell, J., Kapper, C., Lovig, M., Planchon, E., and Williams, J. P. Conformal prediction for text infilling and part-of-speech prediction. *The New England Journal of Statistics in Data Science*, 2022.

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., and Florence, P. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, 2023.

Fan, L., Wang, G., Jiang, Y., Mandlekar, A., Yang, Y., Zhu, H., Tang, A., Huang, D.-A., Zhu, Y., and Anandkumar, A. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *Advances in Neural Information Processing Systems*, 2022.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024.

Fisch, A., Schuster, T., Jaakkola, T., and Barzilay, R. Efficient conformal prediction via cascaded inference with expanded admission. In *International Conference on Learning Representations*, 2021.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 2023.

Ghasemipour, K., Gu, S. S., and Nachum, O. Why so pessimistic? estimating uncertainties for offline rl through ensembles, and why their independence matters. In *Advances in Neural Information Processing Systems*, 2022.

Giovannotti, P. and Gammerman, A. Transformer-based conformal predictors for paraphrase detection. In *Conformal and Probabilistic Prediction and Applications*, 2021.

Goel, V. Anatomy of deductive reasoning. *Trends in cognitive sciences*, 2007.

Groot, T. and Valdenegro Toro, M. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Trustworthy Natural Language Processing*, 2024.

Guo, X., Huang, K., Liu, J., Fan, W., Vélez, N., Wu, Q., Wang, H., Griffiths, T. L., and Wang, M. Embodied LLM agents learn to cooperate in organized teams. In *Language Gamification - NeurIPS 2024 Workshop*, 2024.

Guss, W. H., Houghton, B., Topin, N., Wang, P., Codel, C. R., Veloso, M. M., and Salakhutdinov, R. Minerl: A large-scale dataset of minecraft demonstrations. In *International Joint Conference on Artificial Intelligence*, 2019.

Hafner, D., Pasukonis, J., Ba, J., and Lillicrap, T. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

He, S., Han, S., Su, S., Han, S., Zou, S., and Miao, F. Robust multi-agent reinforcement learning with state uncertainty. *Transactions on Machine Learning Research*, 2023.

Huang, J. and Chang, K. C.-C. Towards reasoning in large language models: A survey. In *Association for Computational Linguistics*, 2023.

Huang, W., Zhang, J., and Huang, K. Bootstrap estimated uncertainty of the environment model for model-based reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2019.

Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., et al. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, 2023.

Jin, L., Chen, X., Rückin, J., and Popović, M. Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering. In *International Conference on Intelligent Robots and Systems*. IEEE, 2023.

Johnson, M., Hofmann, K., Hutton, T., and Bignell, D. The malmo platform for artificial intelligence experimentation. In *International Joint Conference on Artificial Intelligence*, 2016.

Johnson-Laird, P. N. Deductive reasoning. *Annual review of psychology*, 1999.

Josephson, J. R. and Josephson, S. G. *Abductive inference: Computation, philosophy, technology*. Cambridge University Press, 1996.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Kuleshov, V. and Deshpande, S. Calibrated and sharp uncertainties in deep learning via density estimation. In *International Conference on Machine Learning*, 2022.

Kumar, B., Lu, C., Gupta, G., Palepu, A., Bellamy, D., Raskar, R., and Beam, A. Conformal prediction with large language models for multi-choice question answering. In *ICML Neural Conversational AI TEACH workshop*, 2023.

Levine, Y., Wies, N., Jannai, D., Navon, D., Hoshen, Y., and Shashua, A. The inductive bias of in-context learning: Rethinking pretraining example design. In *International Conference on Learning Representations*, 2022.

Li, M. and Vitányi, P. M. B. Inductive reasoning. In *Language Computations*, 1992.

Li, Z., Xie, Y., Shao, R., Chen, G., Jiang, D., and Nie, L. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. *Advances in Neural Information Processing Systems*, 2025.

Liang, K., Zhang, Z., and Fisac, J. F. Introspective planning: Aligning robots' uncertainty with inherent task ambiguity. In *Advances in Neural Information Processing Systems*, 2024.

Lin, H., Wang, Z., Ma, J., and Liang, Y. Mcu: A task-centric framework for open-ended agent evaluation in minecraft. *arXiv preprint arXiv:2310.08367*, 2023.

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022a.

Lin, Z., Liu, J. Z., and Shang, J. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. In *Association for Computational Linguistics*, 2022b.

Liu, E., Neubig, G., and Andreas, J. An incomplete loop: Instruction inference, instruction following, and in-context learning in language models. In *Conference on Language Modeling*, 2024.

Malinin, A. and Gales, M. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

Mielke, S. J., Szlam, A., Dinan, E., and Boureau, Y.-L. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 2022.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

Nottingham, K., Ammanabrolu, P., Suhr, A., Choi, Y., Hajishirzi, H., Singh, S., and Fox, R. Do embodied agents dream of pixelated sheep: Embodied decision making using language guided world modelling. In *International Conference on Machine Learning*, 2023.

Okoli, C. Inductive, abductive and deductive theorising. *International Journal of Management Concepts and Philosophy*, 2023.

Peirce, C. S. *Collected papers of charles sanders peirce*. Harvard University Press, 1934.

Qin, Y., Zhou, E., Liu, Q., Yin, Z., Sheng, L., Zhang, R., Qiao, Y., and Shao, J. Mp5: A multi-modal open-ended embodied system in minecraft via active perception. In *Conference on Computer Vision and Pattern Recognition*, 2024.

Ren, A. Z., Dixit, A., Bodrova, A., Singh, S., Tu, S., Brown, N., Xu, P., Takayama, L. T., Xia, F., Varley, J., et al. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning*, 2023.

Ren, A. Z., Clark, J., Dixit, A., Itkina, M., Majumdar, A., and Sadigh, D. Explore until confident: Efficient exploration for embodied question answering. In *Robotics: Science and Systems*, 2024.

Robinson, J. and Wingate, D. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations*, 2023.

Sagar, S., Taparia, A., and Senanayake, R. Failures are fated, but can be faded: Characterizing and mitigating unwanted behaviors in large-scale vision and language models. In *International Conference on Machine Learning*, 2024.

Shen, Y. and Lourentzou, I. Learning by asking for embodied visual navigation and task completion. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024.

Su, W., Wang, C., Ai, Q., Hu, Y., Wu, Z., Zhou, Y., and Liu, Y. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Association for Computational Linguistics*, 2024.

Tian, B., Luo, L., Zhao, H., and Zhou, G. Vibus: Data-efficient 3d scene parsing with viewpoint bottleneck and uncertainty-spectrum modeling. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Walton, D. Abductive, presumptive and plausible arguments. *Informal Logic*, 2001.

Walton, D. *Abductive reasoning*. University of Alabama Press, 2014.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. Voyager: An open-ended embodied agent with large language models. In *NeurIPS Intrinsically-Motivated and Open-Ended Learning Workshop*, 2023a.

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., and Lim, E.-P. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Association for Computational Linguistics*, 2023b.

Wang, Y. and Zou, S. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 2021.

Wang, Z., Cai, S., Chen, G., Liu, A., Ma, X., Liang, Y., and CraftJarvis, T. Describe, explain, plan and select:

Interactive planning with large language models enables open-world multi-task agents. In *Advances in Neural Information Processing Systems*, 2023c.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022a.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022b.

Xiao, Y. and Wang, W. Y. On hallucination and predictive uncertainty in conditional language generation. In *European Chapter of the Association for Computational Linguistics*, 2021.

Xiong, M., Hu, Z., Lu, X., LI, Y., Fu, J., He, J., and Hooi, B. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024.

Xu, F., Lin, Q., Han, J., Zhao, T., Liu, J., and Cambria, E. Are large language models really good logical reasoners? a comprehensive evaluation and beyond. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Ye, F., Yang, M., Pang, J., Wang, L., Wong, D. F., Yilmaz, E., Shi, S., and Tu, Z. Benchmarking LLMs via uncertainty quantification. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Yildirim, M. Y., Ozer, M., and Davulcu, H. Leveraging uncertainty in deep learning for selective classification. *arXiv preprint arXiv:1905.09509*, 2019.

Yuan, Z., Yan, Y., Sonka, M., and Yang, T. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *International Conference on Computer Vision*, 2021.

Zhang, J., Lan, T., Murthy, R., Liu, Z., Yao, W., Tan, J., Hoang, T., Yang, L., Feng, Y., Liu, Z., Awalgaonkar, T., Niebles, J. C., Savarese, S., Heinecke, S., Wang, H., and Xiong, C. Agentohana: Design unified data and training pipeline for effective agent learning. *arXiv preprint arXiv:2402.15506*, 2024a.

Zhang, R., Zhang, H., and Zheng, Z. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024b.

Zhao, Z., Chai, W., Wang, X., Li, B., Hao, S., Cao, S., Ye, T., and Wang, G. See and think: Embodied agent in virtual environment. In *European Conference on Computer Vision*, 2025.

Zhu, X., Chen, Y., Tian, H., Tao, C., Su, W., Yang, C., Huang, G., Li, B., Lu, L., Wang, X., et al. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023.

A. Definitions of Uncertainty Types

The three fundamental forms of logical reasoning, inductive, deductive, and abductive, have long been recognized and studied (Peirce, 1934; Walton, 2014; Wei et al., 2022b; Levine et al., 2022; Okoli, 2023). As language models demonstrated extraordinary abilities, designing better reasoning mechanisms has become a popular research trend (Cheng et al., 2024; Liu et al., 2024). These reasoning paradigms serve as fundamental frameworks for structuring inference and decision-making processes, particularly in settings where uncertainty arises due to partial observations, ambiguous premises, or multiple plausible explanations (Xu et al., 2025). We adapt these reasoning types to the domain of embodied confidence elicitation and formally define and describe each uncertainty type (See Table 4).

Reasoning Type	Definition	Uncertainty Associated	Example	Elicitation Method
Inductive	Inductive reasoning derives general principles from a body of observations which means making broad generalizations based on specific observations (Li & Vitányi, 1992).	*Inductive Uncertainty*: Arises when an agent generalizes from limited observations, leading to potential overgeneralization or misclassification.	An agent observes that all previously encountered caves contained hostile entities and infers that any future cave is also dangerous. However, this conclusion is uncertain because it is based on a limited number of observations.	*Chain-of-Thought (CoT)*: The agent systematically analyzes observed trends, considers possible exceptions, and evaluates confidence in applying generalizations.
Deductive	Deductive reasoning is the process of drawing deductive inferences that start from the given premises and reason with logical rules or commonsense to obtain certain conclusions (Johnson-Laird, 1999; Goel, 2007).	*Deductive Uncertainty*: Arises when an agent applies logical rules but encounters missing, conflicting, or incomplete premises, making the outcome uncertain.	An agent knows the rule: "If wood is available, then a wooden tool can be crafted." However, if it is uncertain whether wood is available, it cannot confidently conclude whether crafting is possible.	*Plan-and-Solve (P&S)*: The agent formulates a set of premises, identifies missing dependencies, and assesses confidence in executing the task.
Abductive	The process of inferring the most plausible explanation for an observation based on incomplete evidence. Abduction generates hypotheses rather than definitive conclusions (Josephson & Josephson, 1996; Walton, 2001).	*Abductive Uncertainty*: Arises when multiple explanations could account for an observation, with no definitive way to determine the correct one.	An agent searching for a pig near a river hypothesizes that pigs and rivers may exist in any of the four cardinal directions but lacks direct evidence to confirm a single hypothesis.	*Top-K Reasoning*: The agent generates multiple plausible hypotheses, assigns probability estimates to each, and ranks them by likelihood.

Table 4. Definitions of reasoning types, their associated uncertainty, examples, and the corresponding elicitation methods.

Inductive Uncertainty arises when an agent generalizes from specific observations to broader conclusions based on incomplete data. Induction relies on identifying patterns from limited experiences, leading to inherent uncertainty. This is particularly relevant in open-world environments, where observations are partial, and inferred generalizations may not always hold. For example, an agent navigating an unfamiliar environment may observe that all previously encountered caves contained hostile entities. Based on this pattern, it may infer that any future cave is also dangerous. However, since this conclusion is based on a limited set of observations rather than a deterministic rule, the agent must assess how strongly its past experiences justify this generalization, introducing *inductive uncertainty*.

To elicit inductive uncertainty, we employ Chain-of-Thought (Wei et al., 2022b), which prompts the agent to explicitly reflect on the reliability of its observed patterns. By systematically verbalizing its reasoning, the agent is encouraged to: (1) analyze the strength of observed trends, (2) consider possible exceptions or contradictory evidence, and (3) assess its confidence in applying the generalization to new situations. This structured elicitation enables the agent to express uncertainty in its inductive inferences rather than assuming patterns always hold.

Deductive Uncertainty arises when an agent faces ambiguity due to missing, conflicting, or incomplete premises. Deductive

uncertainty occurs within a structured decision-making process when the available information is insufficient to determine a definitive outcome. Consider an agent tasked with crafting a wooden tool in a survival environment. It knows the rule: "If wood is available, then a wooden tool can be crafted." However, if the agent is uncertain whether wood is currently accessible, it cannot confidently conclude whether crafting is possible. This scenario exemplifies deductive uncertainty, where the agent's ability to reason is constrained by unknown or ambiguous premises.

To elicit deductive uncertainty, we use *Plan-and-Solve* prompting (Wang et al., 2023b), which guides the agent through a structured reasoning process. The agent is encouraged to: (1) formulate a comprehensive set of premises relevant to the task, (2) identify any missing premises or dependencies, and (3) assess its confidence in executing each step successfully. This structured elicitation enables the agent to explicitly express uncertainty when premises are incomplete or insufficient to deduce a definitive conclusion.

Abductive Uncertainty occurs when an agent must infer the most plausible explanation for an observation without definitive evidence. Abduction involves *hypothesis generation* under uncertainty. The challenge in abductive reasoning lies in selecting the most probable explanation when multiple interpretations exist, each carrying some degree of uncertainty. A simple example occurs when an agent is tasked with locating a pig near a river for unspecified reasons. Given its environment, the agent may hypothesize that pigs and rivers could exist in any of the four cardinal directions but are unlikely to be present in all directions simultaneously. Since the agent lacks direct evidence to confirm a single hypothesis, it must infer the most plausible explanation, leading to abductive uncertainty.

To elicit abductive uncertainty, we implement *Top-K reasoning* (Robinson & Wingate, 2023), where the agent is instructed to generate multiple plausible hypotheses explaining an observation and assign probability estimates to each. This process forces the agent to explicitly consider alternative interpretations, rank them by likelihood, and communicate the level of confidence in its inferences. By quantifying uncertainty across multiple competing hypotheses, Top-K reasoning reveals the agent's abductive reasoning capabilities.

B. Task Setting Details

Inspired by previous works (Lin et al., 2023; Qin et al., 2024), we define a set of 30 tasks evenly distributed across three difficulty levels: easy, medium, and hard. Categorization is based on the complexity of reasoning required and the extent of contextual information necessary for successful task completion. Each difficulty level incorporates distinct challenges, ranging from straightforward operations to intricate reasoning across interdependent objectives, with a balanced distribution of complexity within the task set. We present all tasks and highlight entities of different categories in Table 5

Easy Tasks are designed to evaluate the agent's ability to process minimal perceptual information and perform straightforward actions with limited reasoning. These tasks typically require the perception of only one environmental element from predefined categories such as Object, Mob, Ecology, Time, Weather, or Brightness (Qin et al., 2024). For example, tasks at this level may involve identifying a specific object in the environment or recognizing a simple temporal condition (*e.g.*, daytime or nighttime). The actions required are relatively simple and involve a single reasoning step, such as gathering an object that is readily visible.

Medium Tasks introduce moderate complexity by requiring the perception and integration of two to three environmental elements, alongside a corresponding increase in reasoning steps. Tasks at this level involve combining multiple types of perceptual data, such as recognizing a specific biome and locating a particular mob or object within it. For example, the agent might need to identify a forest biome, locate a pig, and gather specific materials. In addition to perceptual challenges, medium tasks often include sequential sub-goals, such as collecting and combining resources to create basic tools. These tasks require the agent to interpret dynamic environmental information, execute plans involving multiple steps, and adapt to minor changes in the environment. This level evaluates the agent's ability to balance perception, reasoning, and adaptability.

Hard Tasks are the most challenging and require the agent to process and integrate multiple layers of perceptual information (up to six elements) while performing complex situation-aware planning and dynamic action execution. These tasks involve a high level of reasoning, such as decomposing long-term objectives into interdependent sub-tasks, managing uncertainties in the environment, and dynamically adjusting strategies in response to real-time changes. For example, a hard task might involve navigating through hazardous biomes, identifying and gathering multiple resources, and crafting advanced tools or items that require sequential processing and the use of specialized platforms. Environmental conditions, such as weather, time of day, or changing brightness, may dynamically impact the task, necessitating constant adaptation by the agent. These tasks often introduce significant challenges, such as hostile mobs or the need to traverse difficult terrain, testing the agent's
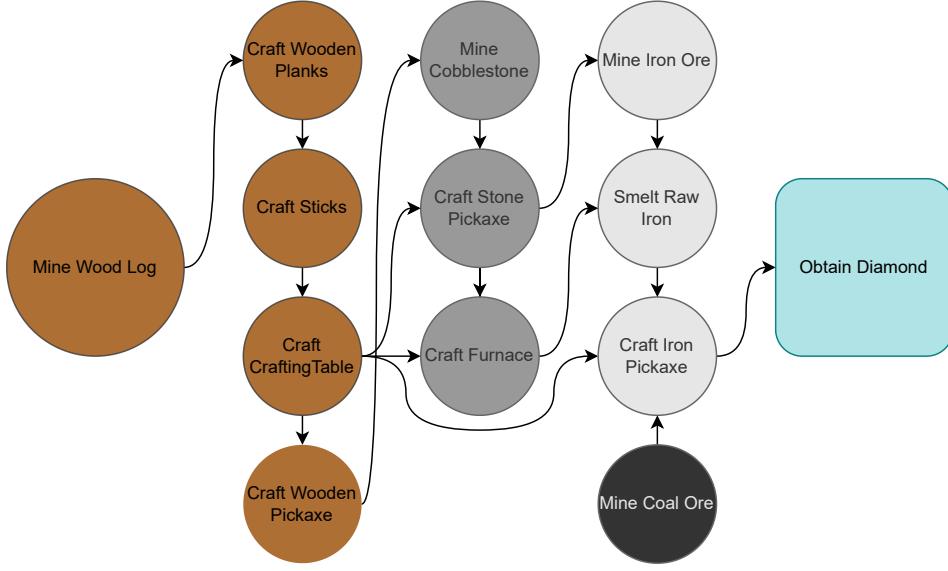
Figure 6. An illustrative diagram of the Obtain Diamond task, featuring five distinct colors to represent the source materials required—wood, stone, iron, coal, and diamond—aligned with the Minecraft tech tree.

ability to balance perception, planning, and execution effectively.

The (obtain) **Diamond Task** is one of the most iconic and challenging benchmarks in Minecraft agent research, serving as a comprehensive test of an agent's long-horizon planning, resource management, and adaptability. The task requires the agent to progress through multiple interdependent steps, including gathering basic resources like wood and stone, crafting tools such as a pickaxe, and locating and mining diamonds deep within underground caves (See Figure 6). Each step presents its own set of challenges, such as navigating complex terrain, managing limited resources, and avoiding environmental hazards like lava or hostile mobs. The randomized nature of Minecraft's procedural world generation further compounds the difficulty, as the agent must adapt dynamically to new environments while maintaining focus on the ultimate objective. Success in the "Obtain Diamond" task is often seen as a key indicator of an agent's ability to integrate active perception, situational awareness, and embodied action execution in an open-world setting. This task demonstrates the complexity of open-ended problem-solving and has become a gold standard for evaluating the capabilities of autonomous agents in multi-modal and multi-step scenarios. We added the diamond task as one of our hard tasks.

C. Detailed Model Descriptions

GPT-4V: This vision-capable variant of GPT-4 excels at processing both visual and textual inputs, making it a powerful tool for tackling tasks within the visually complex Minecraft environment. Unlike its predecessors, GPT-4V's ability to seamlessly combine perception and reasoning allows for sophisticated decision-making and planning. The GPT-4 series has already demonstrated its efficacy in Minecraft-based research. For instance, Voyager (Wang et al., 2023a), the first LLM-powered embodied lifelong learning agent, used GPT-4 to facilitate continuous exploration, skill acquisition, and task execution without human intervention. Voyager's architecture included an automatic curriculum for exploration and a skill library to store and retrieve executable code, allowing agents to adapt and improve iteratively. Similarly, Optimus-1 (Li et al., 2025) employs GPT-4V to refine its planning processes, focusing on logical reasoning and task generalization. These implementations underscore GPT-4V's pivotal role in advancing embodied AI research, offering exceptional capabilities for both exploration and problem-solving.

MineLLM: Tailored specifically for tasks within Minecraft, MineLLM represents a significant leap in AI development for complex embodied environments. As a central component of the MP5 framework (Qin et al., 2024), MineLLM is designed to tackle the unique challenges posed by Minecraft's open-ended tasks. It combines the image visual encoder from MineCLIP (Fan et al., 2022) with the Vicuna-13B-v1.5 language model for integrating visual perception with natural

Difficulty	Task ID	Task Description
Easy	1	Find a pig
	2	Find a cow
	3	Find a tree
	4	Mine log
	5	Mine sand
	6	Craft a plank
	7	Craft a stick
	8	Craft a chest
	9	Craft a wooden door
	10	Craft a wooden boat
Medium	11	Find a tree in the forest
	12	Find a pig on grass
	13	Find a cow in the desert
	14	Craft a wooden sword
	15	Craft a wooden pickaxe
	16	Craft a stone pickaxe
	17	Smelt an iron ingot
	18	Smelt glass
	19	Cook beef
	20	Cook mutton
Hard	21	Find a pig near a grass in the forest during the daytime
	22	Find a cow in the desert during the daytime
	23	Find a grass near a pig in the forest
	24	Find a pig while wearing an iron helmet
	25	Craft an iron door
	26	Craft an iron pickaxe
	27	Craft an iron sword
	28	Craft a compass
	29	Kill a zombie with an iron sword
	30	Obtain a diamond

Table 5. Full task details. 30 tasks evenly distributed as easy, medium, and hard. Underlines label different information categories in Minecraft, highlighting how the complexity varies at each level.

language understanding. Trained on a vast dataset of 500,000 Minecraft-specific image-text instruction pairs, MineLLM can generate detailed insights about the game environment, answer complex queries, and provide contextual guidance for planning and execution. Its integration into MP5 enables the framework to address context- and process-dependent tasks with remarkable success rates, achieving a 91% success rate on context-dependent tasks and demonstrating exceptional adaptability in novel scenarios.

STEVE: The STEVE series represents another advancement in language model-driven embodied agents for the Minecraft environment (Zhao et al., 2025). Built upon the foundation of LLaMA-2 (Touvron et al., 2023b), STEVE integrates powerful language capabilities tailored to enhance task reasoning, contextual understanding, and interaction. At its core, the language model in the STEVE series excels at decomposing complex objectives into actionable subtasks through iterative reasoning and hierarchical planning. This allows STEVE agents to process high-level instructions effectively and generate detailed plans for task execution. The STEVE series relies heavily on its ability to adapt to Minecraft-specific tasks. To this end, Zhao et al. (2025) curated the STEVE-21K dataset, containing 20K knowledge-based question-answering pairs and 200+ skill-code pairs that directly enhance the model's contextual understanding and task reasoning. These adaptations enable the language model to seamlessly integrate with perception and action modules, driving coherent decision-making in real time. Furthermore, STEVE agents leverage advanced contextual awareness to refine their decision-making processes, significantly outperforming prior benchmarks in task decomposition and completion efficiency. The series also demonstrated up to 1.5x faster progression in complex tasks like unlocking tech trees and up to 2.5x quicker performance in block search scenarios compared to other state-of-the-art models.

D. Additional Experiments

Do Hard Tasks Lead to Poor Calibration? We use the best-performing GPT-4V as our agent backbone and withhold any execution policies to reduce computation costs. We set the maximum episode length as 12,000 to provide enough coverage for all task difficulties. Results are shown in Table 6.

Task	Policies	ECE (↓)	AUROC (↑)	Success Rate (↑)
Easy	Vanilla	0.26	0.76	84%
	Self-Intervention	0.26	0.76	92%
	CoT	**0.11**	0.78	**94%**
	P&S	0.12	**0.80**	82%
	Top-K	0.32	0.72	74%
Medium	Vanilla	0.35	0.54	52%
	Self-Intervention	0.35	0.51	44%
	CoT	**0.22**	**0.58**	**54%**
	P&S	0.22	0.55	48%
	Top-K	0.40	0.47	32%
Hard	Vanilla	0.33	0.58	17%
	Self-Intervention	0.35	0.52	12%
	CoT	**0.31**	0.68	18%
	P&S	0.32	**0.71**	**18%**
	Top-K	0.41	0.49	8%

Table 6. ECE, AUROC, and Success Rates for Different Task Difficulties and Elicitation Policies. Lower ECE and higher AUROC/Success Rates indicate better performance.

For **Easy tasks**, CoT demonstrated the best performance, achieving the lowest ECE (0.11) and the highest success rate (94%), followed by P&S, which recorded the highest AUROC of 0.80 and a success rate of 82%. Self-Intervention performed comparably in calibration (ECE = 0.26, AUROC = 0.76). Top-K underperformed, with the highest ECE (0.32) and the lowest success rate (74%), indicating limitations in leveraging task simplicity. For **Medium tasks**, all policies showed noticeable declines in performance. CoT emerged as the best overall, with an ECE of 0.22, an AUROC of 0.58, and a success rate of 54%, balancing calibration and task success effectively. P&S followed closely with similar calibration (ECE = 0.22) but a slightly lower AUROC (0.55) and success rate (48%). For **Hard tasks**, performance further degraded across all policies. CoT and P&S maintained relative superiority, with CoT achieving an ECE of 0.31, AUROC of 0.68, and a success rate of 22%, while P&S recorded slightly worse calibration (ECE = 0.32) and the highest AUROC (0.71) but tied for a success rate of 18%.

These results confirm our hypothesis that as task difficulty increases, confidence calibration significantly deteriorates, with the ECE gap increasing as high as 0.20 (Easy CoT vs. Hard CoT). However, the results also demonstrate that structured elicitation policies, such as CoT and P&S, consistently prove effective in handling calibration, failure prediction, and task success across task difficulties. Additionally, simpler policies like Self-Intervention also show moderate success, particularly in easier tasks, suggesting their utility in less demanding scenarios.