

Uncertainty Quantification in LLM Agents: Foundations, Emerging Challenges, and Opportunities

Changdae Oh¹ Seongheon Park¹ To Eun Kim²

Jiatong Li¹ Wendi Li¹ Samuel Yeh¹ Xuefeng Du³

Hamed Hassani⁴ Paul Bogdan⁵ Dawn Song⁶ Sharon Li¹


¹University of Wisconsin–Madison ²Carnegie Mellon University

³Nanyang Technological University ⁴University of Pennsylvania

⁵University of Southern California ⁶University of California, Berkeley

{changdae@cs.wisc.edu, sharonli@cs.wisc.edu}

Abstract

Uncertainty quantification (UQ) for large language models (LLMs) is a key building block for safety guardrails of daily LLM applications. Yet, even as LLM agents are increasingly deployed in highly complex tasks, most UQ research still centers on single-turn question-answering. We argue that *UQ research must shift to realistic settings with interactive agents, and that a new principled framework for agent UQ is needed*. This paper presents three pillars to build a solid ground for future agent UQ research: **(1. Foundations)** We present the first general formulation of agent UQ that subsumes broad classes of existing UQ setups; **(2. Challenges)** We identify four technical challenges specifically tied to agentic setups—selection of uncertainty estimator, uncertainty of heterogeneous entities, modeling uncertainty dynamics in interactive systems, and lack of fine-grained benchmarks—with numerical analysis on a real-world agent benchmark, τ^2 -bench; **(3. Future Directions)** We conclude with noting on the practical implications of agent UQ and remaining open problems as forward-looking discussion for future explorations. 

1 Introduction

LLM-based agents operating in open-world environments take actions that have real consequences: making costly bookings, modifying databases, or issuing irreversible commands (OpenAI, 2025; Google, 2025; Anthropic, 2026) at the next level of autonomy (Steinberger, 2026). In such settings, failures are no longer limited to incorrect text generation: agents may act prematurely under unresolved ambiguity, propagate errors across long interaction trajectories, or commit to outcomes that are costly or difficult to undo. To be deployed responsibly, agents must be able to assess and act upon the likelihood of failure (Kochenderfer, 2015). This makes uncertainty quantification (UQ) an urgent and central requirement for agentic systems.

Despite the importance, most existing UQ research treats LLMs as static oracles: a system examined in isolation, prompted once, and evaluated by the uncertainty of a single response or a chain of responses (Figure 1, left). These methods implicitly assume a *static system* in which no new information is acquired after the initial prompt (Malinin and Gales, 2021; Kadavath et al., 2022; Farquhar et al., 2024). Under this assumption, uncertainty is treated as a point-wise estimate or uni-directional propagation process. While appropriate for single-turn inference or non-interactive multi-step reasoning, *they do not meet the growing needs of UQ in agentic settings featured by its long-horizon interactive environment with heterogeneous entities*.

LLM agents in this setup operate with feedback from users and system components through multi-turn interactions to complete a complex task (Figure 1, right). For example, a flight-booking agent needs to decide whether to finalize a reservation or ask follow-up questions about dates, budget, or lay-over preferences. Early uncertainty should prompt information-seeking actions, while subsequent dialogue and database queries can resolve ambiguities and reduce uncertainty, enabling confident execution. This deviates significantly from the setup considered in classic LLM UQ, as the agent now spans a long interaction log updated over multiple turns that involves messages from other entities, and it also gets additional information from environments. This gap raises questions about the applicability of existing UQ methods to agentic setups.

In this paper, we argue that building a trustworthy LLM-driven agentic system requires a shift in focus: from pointwise uncertainty of the final answer to structured uncertainty dynamics in an open, interactive decision process, and a new framework is necessary for this. To initiate a formal discussion, we first **provide a concrete definition and a general formulation of agent UQ** that captures broad categories of existing UQ setups (Sec. 3).

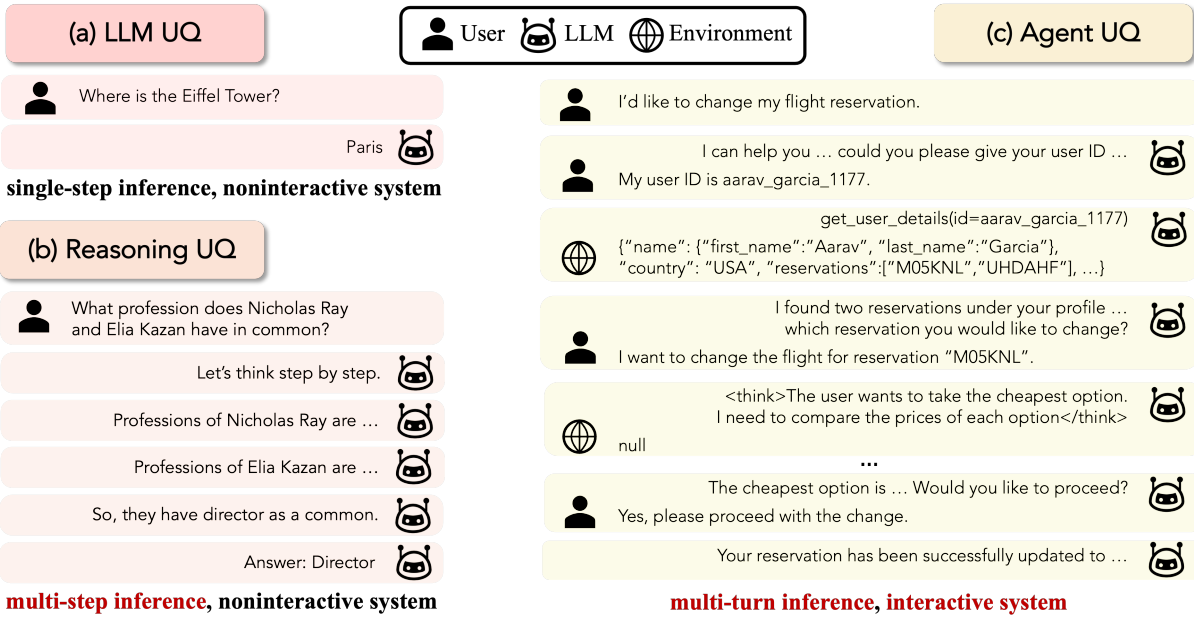


Figure 1: **Comparison between UQ setups.** Traditional LLM UQ (a) measures the uncertainty of an answer given a question, whereas the UQ for LLM reasoning (b) expands the problem by considering multi-step responses rather than a single response. Agent UQ (c) goes further by considering continual interactions between agent and user/environment across the trajectory, making it a multi-turn, interactive setup that does not fit neatly into the classical formulation.

Specifically, we model the agent’s problem-solving trajectory as a stochastic process over actions, observations, and environment states, represented as a simple graphical model. Then, we define both turn-level and trajectory-level uncertainty in a simple, probabilistic expression, where we show that many prior UQ problems can be cast as special cases.

Based on this foundation, we **identify four technical challenges in agent UQ** that call for actions from communities. (1) Selection of uncertainty estimator: while it is already known that existing uncertainty estimators have their own tradeoffs, agentic scaffolding amplifies these into severe limitations or hard constraints; (2) Uncertainty of heterogeneous entities: how can we derive estimates over the content generated by external entities having different underlying distributions?; (3) Modeling uncertainty dynamics in interactive systems: traditional weighted-average-style uncertainty aggregation may not be suitable to modeling uncertainty dynamics in the non-isolated, open environment; (4) Lack of fine-grained benchmarks: our pilot survey reveals that granular benchmarks, such as turn-level evaluation, are scarce for LLM agents.

Then, Sec. 5 discusses **practical implications** of agent UQ across diverse domains, including health-care, software engineering, and robotics, as well as frontier LLM research on adaptive reasoning and post-training. Finally, we outline **open problems**

specific to uncertainty modeling in agentic systems (Sec. 6). Together, we view this work as a step toward providing the field with a necessary foundation and clearer direction for future research on uncertainty-aware, reliable agentic systems.

- We cast agent UQ as a distinct problem setup, and build a foundation for it by providing a concrete definition and a general formulation that expresses many existing UQ setups as special cases.
- We crystallize the emerging challenges in agent UQ with numerical studies on a real-world benchmark τ^2 -bench (Barres et al., 2025) with GPT-4.1 (OpenAI, 2025) and Kimi-K2.5 (Kimi Team, 2026).
- To further give actionable insights and promote discussion, we describe promising examples of agent UQ applications in various domains and also remaining open problems.

2 Related Work

UQ in LLMs. In contrast to classic UQ in machine learning (Neal, 1992; MacKay, 1992; Gal and Ghahramani, 2016), LLM UQ brings extra challenges from its computational cost and free-form outputs. Common approaches include aggregating output token probability (Zhang et al., 2025b;

Fomicheva et al., 2020; Malinin and Gales, 2021; Zhang et al., 2023; Fadeeva et al., 2024; Duan et al., 2024; Aichberger et al., 2024), measuring consistency over multiple generations (Manakul et al., 2023; Farquhar et al., 2024; Nikitin et al., 2024), verbalizing confidence (Kadavath et al., 2022; Lin et al., 2022; Yona et al., 2024), or conformal prediction (Kumar et al., 2023; Cherian et al., 2024; Quach et al., 2024). While promising, they mostly use single-step QA as a testbed, which limits their practical applicability in agentic inference.

UQ in reasoning and agent setups. Beyond single-step generation, researchers show growing interest in measuring the confidence of LLMs during their reasoning. This (un)certainly information has been used to improve response quality (Wang et al., 2023; Kang et al., 2025b; Fu et al., 2025) or to guide a model during inference or training (Hu et al., 2024; Lightman et al., 2024; Zhu et al., 2025; Wang et al., 2025b,a; Cheng et al., 2025). Recently, there have been attempts to model uncertainty in interactive or agentic inference (Han et al., 2024; Zhao et al., 2025; Duan et al., 2025; Frankel et al., 2024; Chan et al., 2025; Zhang et al., 2026a,c). Although aiming to model uncertainty dynamics over long-horizon generation, they *do not reflect uncertainty from different entities* (e.g., users) *nor concern the reducible nature of uncertainty in open environments* (Kirchhof et al., 2025; Suri et al., 2025), highlighting the need for a new UQ framework under agentic scaffolding.

3 A General Formulation for Agent UQ

3.1 Problem Statement and Definitions

Imagine a flight booking agent depicted in Figure 1 right. Here, the types of **action** A and **observation** O are diverse: for A , there is information-seeking, questioning for clarification, calling an API function, and thinking or updating without interactive behaviors. The observation O , on the other hand, can be a database status report, tool-calling results, or the user’s clarification. Both action and observation can be realized in flexible forms of natural language messages or structured strings, e.g., programming language.

Concretely, given a **task specification** E_0 , e.g., a flight booking, a user makes a stochastic **initial query** O_0 , and the agent first requests the user’s ID to get relevant information. Then, it calls a programmed function to get privileged information from the database. Not only does the agent interact

with the user or tool, but it also reasons itself about better future actions (i.e., planning) or makes a system update without external observation. All previous interactions are stored in **environment state** E for context-grounded actions over a long horizon inference. This makes up a **trajectory** $\mathcal{F}_{\leq t}$ (chain of A , E , and O) up to t -turn which continues until the task-specific termination rule is met, e.g., when the agent submits an answer (Yang et al., 2018; Jimenez et al., 2024; Wei et al., 2025), the user generates a stop sign (Yao et al., 2024; Barres et al., 2025), or goal completion (Shridhar et al., 2021) with multiple constraints (Xie et al., 2024a).

Now, to formalize UQ in this real-world agent context, we begin with a concrete definition of the agent system in Definition 3.1, and introduce a graphical model for the agent trajectory in Figure 2. This can generalize many existing agent benchmark setups (Mialon et al., 2023; Yao et al., 2024; Yang et al., 2018; Song et al., 2023; Jimenez et al., 2024; Xie et al., 2024a; Liu et al., 2024; Wei et al., 2025).

Definition 3.1 (Stochastic Agent System). *Let E_i be an environment state, a mixture of (1) a context memory of agent-user-tool interaction logs up to the i -th turn that is fully accessible to the agent, and (2) a state of the system database, which is not directly accessible to the agent—only partially observable via tool-calling. Let O_i and A_i be the i -th turn observation and action derived from distributions P and $P_{\pi, \mathcal{T}}$, respectively, where \mathcal{T} indicates a tool set and π is an LLM policy. Then, under a task specification E_0 and the user’s initial query O_0 , we define the trajectory as a sequence of turns $\mathcal{F}_{\leq T} = \{(A_t, E_t, O_t)\}_{t=0}^T$ where $\mathcal{F}_0 = (E_0, O_0)$. Given a deterministic update function $h(\cdot)$, the generative process for each turn t is defined as:*

$$A_i \sim P_{\pi, \mathcal{T}}(\cdot | E_{i-1}, O_{i-1}), O_i \sim P(\cdot | A_i, E_i), \\ E_i = h(E_{i-1}, O_{i-1}, A_i).$$

To express this trajectory with a simple mathematical representation, we introduce a dynamic Bayesian network (Murphy, 2002) in Figure 2. With this graphical model, dependency structures of A , E , and O can be explained as follows: (1) since the environment state E_{i-1} contains the entire conversation history of $A_{<i-1}$ and $O_{<i-1}$ up to the $i - 1$ step¹, the current action A_i depends

¹An adaptive memory structure (Xu et al., 2025b; Xiong

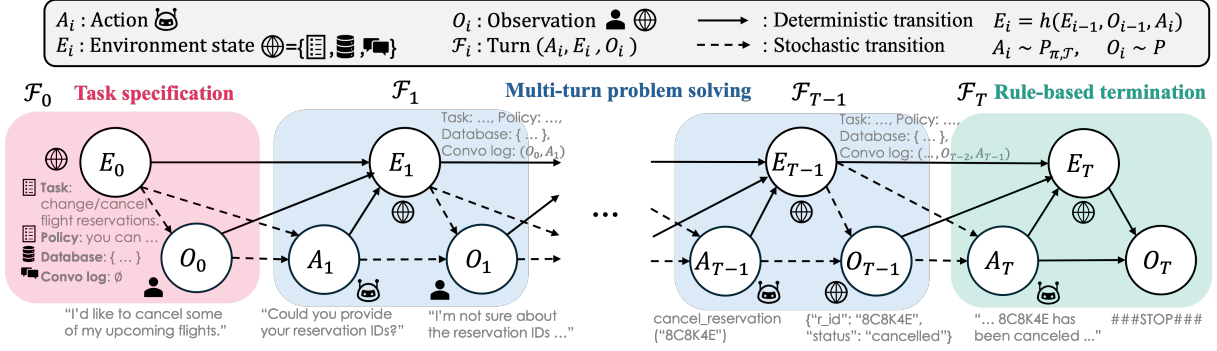


Figure 2: **Graphical model for an agent problem-solving trajectory with examples.** Given a task specification E_0 and an initial user query O_0 , an agent spans a multi-turn trajectory characterized by a chain of action A , observation O , and environment state E . This simple abstraction describes some representative agentic prompting methods such as ReAct (Yao et al., 2022b). See Appendix B.1 for details.

only on E_{i-1} and O_{i-1} ; (2) the current observation O_i from the environment also depends only on A_i and E_i ; (3) the transition of the environment state ($E_{i-1} \rightarrow E_i$) can be defined as a deterministic map² $h(\cdot)$ which stacks the conversation logs up in a memory module and conducts a rule-based modification (write, update, or delete) on a previous system database state given the current action.

Under this Bayesian network, the joint probability for the trajectory can be factorized as below,

$$\begin{aligned} P(\mathcal{F}_{\leq T}) &= P(E_0, O_0) \prod_{i=1}^T P(\mathcal{F}_i | \mathcal{F}_{i-1}) \\ &= P(E_0, O_0) \prod_{i=1}^T P_{\pi, \tau}(A_i | E_{i-1}, O_{i-1}) P(O_i | A_i, E_i), \end{aligned}$$

where $\mathcal{F}_0 = (E_0, O_0)$, and $P(E_i | \cdot) = 1$ as it is deterministic. Now, we define the UQ problem for this stochastic agent system as follows.

Definition 3.2 (Agent UQ). Let $U(X) \geq 0$ be an uncertainty function that takes a random variable X or its distribution to produce a non-negative real value. Given a T -turn trajectory, agent UQ aims to estimate both a turn-level uncertainty $U(\mathcal{F}_t | \mathcal{F}_{t-1})$ for $t = 1, \dots, T$ and a trajectory-level uncertainty $U(\mathcal{F}_{\leq T}) = U(\mathcal{F}_0, \dots, \mathcal{F}_T)$, as a joint total uncertainty.

In the machine learning field, the most popular choices of the uncertainty function $U(X)$ are Shannon entropy (Shannon, 1948) $\mathbb{E}[-\log P(X)]$, negative log probability $-\log P(X = x)$, and their variants. These information-theoretic measures of uncertainty have a nice property, *chain-rule* (Cover, 1999), making them suitable for sequential modeling. For example, given $U(\mathcal{F}_t) := \mathbb{E}[-\log P(\mathcal{F}_t)]$,

et al., 2025) can replace this ever-growing memory.

²In general, this system transition can be also stochastic, but we consider a simplified yet still realistic scenario.

we have the following uncertainty expansion, allowing us to model the agent’s total uncertainty as a simple arithmetic of component-level uncertainties over the action, observation, and state as below³:

$$\begin{aligned} U(\mathcal{F}_{\leq T}) &= U(E_0, O_0) + \sum_{i=1}^T U(\mathcal{F}_i | \mathcal{F}_{i-1}) \\ &= U(E_0, O_0) + \sum_{i=1}^T [U(A_i | E_{i-1}, O_{i-1}) + U(O_i | A_i, E_i)], \end{aligned} \quad (1)$$

where $U(E_0, O_0) = U(O_0 | E_0) + U(E_0)$ is an initial query uncertainty plus a task specification volatility.

Desideratum. While previous LLM UQ setups usually focus on conditional uncertainty of a final answer, agent UQ aims at estimating *joint* uncertainty over the full trajectory, where the estimate from a calibrated agent should be predictive of the reward on that trajectory. Formally, for all $r_1 > r_2$,

$$\mathbb{E}[U(\mathcal{F}_{\leq T}) | r(\mathcal{F}_{\leq T}) = r_1] < \mathbb{E}[U(\mathcal{F}_{\leq T}) | r(\mathcal{F}_{\leq T}) = r_2], \quad (2)$$

where $r(\cdot)$ denotes a real-value reward function commonly set as a success-failure binary verifier.

We draw analogies between agent UQ and probabilistic Turing machines (Gill III, 1974), as well as belief tracking in Partially Observable Markov Decision Processes (Kaelbling et al., 1998), highlighting that agent UQ is grounded by, while still having its unique edge in contrast to these existing theoretical models in terms of setups and focus. See Appendix D for the relevant discussion.

³See Appendix B and C for a derivation and alternative measures, e.g., informational energy (Pardo and Taneja, 1991) and nonextensive entropy (Gell-Mann and Tsallis, 2004).

3.2 A Unified View on Existing UQ Setups

We now show that our agent UQ formulation captures various existing UQs as special cases.

Reduction to single-step LLM UQ. When $t = 1$ and the action space is restricted to responses, LLM UQ can be cast as a special case

$$U(\mathcal{F}_{\leq T}) = U(O_0) + U(A_1|O_0) \geq U(A_1|O_0),$$

where O_0 and A_1 denote the given question and the corresponding model answer, respectively. Most current LLM UQ literature addresses only the conditional answer uncertainty $U(A_1|O_0)$ (Malinin and Gales, 2021; Fadeeva et al., 2023; Farquhar et al., 2024). A few also consider the question ambiguity $U(O_0)$ (Hou et al., 2024; Tomov et al., 2025) to reflect the reality of wild user query (Min et al., 2020).

Reduction to multi-step reasoning UQ. When the action space involves multi-step reasoning (e.g., chain-of-thought, Wei et al. (2022)), we have

$$U(\mathcal{F}_{\leq T}) = U(O_0) + \sum_{i=1}^T U(A_i|A_{<i}, O_0) \quad (3)$$

$$\geq \sum_{i=1}^T U(A_i|A_{<i}, O_0) \quad (4)$$

$$\geq \max_{i \in \{1, \dots, T\}} U(A_i|A_{<i}, O_0) \quad (5)$$

$$\geq \sum_{i=1}^T w_i U(A_i|A_{<i}, O_0), \quad (6)$$

where $w_i \geq 0$ and $\sum_{i=1}^T w_i = 1$. Here, O_0 and A_T denote the given initial query and the model’s final response, and A_i, \dots, A_{T-1} are intermediate outputs. Similar to the single-step setup, most existing multi-step UQs (Tanneru et al., 2024; Tao et al., 2025) do not consider the question ambiguity $U(O_0)$, so that reduces Eq. 3 to Eq. 4, while a few consider it (Leang et al., 2025). Meanwhile, rather than summing up all the step-wise uncertainties, some approaches consider the most uncertain one, i.e., the lowest confidence used in Fu et al. (2025), which can be represented by Eq. 5. Finally, Eq. 6 expresses various weighted average methods, e.g., length-normalized uncertainty (Kang et al., 2025b) with $w_i = \frac{1}{T}$, and tail confidence (Fu et al., 2025) by setting $w_T = 1$, and other clever weighting strategies (Zhao et al., 2025).

Connection to process reward modeling. We further draw a *connection between this multi-turn*

UQ problem and the process reward modeling problem (Uesato et al., 2022; Lightman et al., 2024; Li and Li, 2025) where we assign credit (reward) for each intermediate reasoning step, and then it can be utilized to evaluate the reasoning trajectory by aggregating the rewards across steps to derive a trajectory-level reward. For example, Lightman et al. (2024) proposed the product of the step-wise probabilities and the minimum probability between steps as the aggregation method. These are analogs to Eq. 4 and Eq. 5, respectively.

4 Technical Challenges in Agent UQ

We discuss what makes agentic setups more challenging for UQ than non-agentic ones, identifying four distinct challenges. To motivate discussion, we present preliminary results on τ^2 -bench using existing UQ methods: negative log-likelihood (NLL), entropy over token probabilities (Entropy), and verbalized confidence, each aggregated over all tokens or turns generated by an agent throughout the trajectory (Appendix F has details).

4.1 Selection of Uncertainty Estimator

Approach	Access	Inference Overhead	Theoretical Ground
Probability-based	✗	Free	✓
Consistency-based	✓	Significant	✓
Verbalized Confidence	✓	Neglectable	✗

Table 1: **Comparison between UQ approaches.** Existing UQ approaches have unique weaknesses that become severe obstacles in agentic setups.

So what kind of uncertainty estimator should we adopt to model the agent’s uncertainty, $U(\cdot)$ in Eq. 3.1? We can start by examining three representative approaches that have been advanced in LLM UQ literature so far: output probability-based (and/or internal hidden state-based), response consistency-based, and prompting-based verbalized confidence methods. These approaches involve clear tradeoffs along three dimensions: accessibility, inference-time cost overhead, and theoretical grounding. Probability-based methods exhibit limited applicability due to their reliance on access to output probabilities. While consistency-based and verbalized confidence methods do not face this accessibility constraint, they incur substantial inference-time overhead from repeated generation and lack theoretical grounding, respectively.

Although these are well-known points in the classical single-turn QA setup, **agentic scaffold turns**

Scenario	Avg. r	NLL			Entropy			Verb. Conf.		
		AUROC	ρ	τ	AUROC	ρ	τ	AUROC	ρ	τ
GPT-4.1 on Retail	0.509	0.597	0.169	0.138	0.580	0.139	0.114	0.575	0.179	0.170
GPT-4.1 on Telecom	0.517	0.624	0.214*	0.176*	0.611	0.192*	0.158*	0.685	0.330***	0.286***
Kimi-K2.5 on Retail	0.447	0.469	-0.054	-0.044	0.468	-0.056	-0.045	0.523	0.039	0.032
Kimi-K2.5 on Telecom	0.965	0.645	0.093	0.076	0.664	0.104	0.086	0.580	0.051	0.042

Table 2: **Evaluation of uncertainty estimators on τ^2 -bench.** NLL and Entropy predict task failure ($1 - r$); Verbalized confidence predicts task success (r), where r denotes the reward; ρ and τ denote Spearman and Kendall rank correlation coefficient, respectively. The significance is noted as * $p < 0.05$ and *** $p < 0.001$.

these into much more serious challenges: (1) probability-based methods cannot be applied to most frontier LLMs that centered at challenging agent benchmarks, as they do not consistently provide probability outputs; moreover, free-form long generation makes aggregated token probabilities less informative; (2) consistency-based methods become infeasible due to their prohibitively high inference cost in long-horizon, multi-turn settings; and (3) dynamically expanding context memory of agent (i.e., E_i in Eq. 3.1), often coupled with noisy observations, results in increasingly inflated and unreliable verbalized confidence. A pilot study in Table 2 provides some results of probability-based and verbalized confidence methods for estimating the trajectory uncertainty $U(\mathcal{F}_{\leq T})$ with aggregated action uncertainty $\frac{1}{T} \sum_{i=1}^T U(A_i)$. In many cases, the considered methods show close-to-random classifier performance in predicting the agent’s success/failure. We believe that establishing a theoretically grounded verbalized confidence method is a promising future direction, given its vivid merits in agentic setups such as accessibility and minimal extra inference cost.

4.2 Uncertainty of Heterogeneous Entities

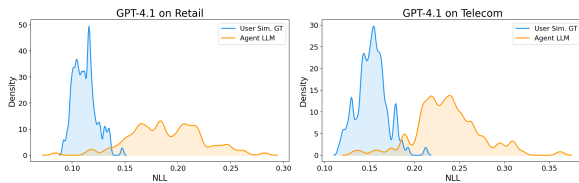


Figure 3: **Observation uncertainty on τ^2 -bench.** We compare distributions over observation uncertainty estimates, i.e., average NLL over all user messages, from the agent LLM and the ground truth user simulator LLM.

Inference of agents involves not only the actions A_i from the agent itself, but more importantly, observations O_i through the conversations with the user and interactions with the environment via tools.

There are *no trivial means to estimate the uncertainty of observations that come from these heterogeneous external entities*, departing from the epistemic uncertainty of the agent-self.

One may argue that we can just feed the whole trajectory, including the observation strings from the user and tools, into the agent and then use its output probability $P_{\pi, \mathcal{T}}(O_i | A_i, E_i)$ to approximate the underlying true observation distribution $P(O_i | A_i, E_i)$ per turn and corresponding uncertainty $U(O_i | A_i, E_i)$. However, the distribution of human language is definitely different from that of LLMs (Muñoz-Ortiz et al., 2024; Reinhart et al., 2025), and the distribution over the tool-calling outcome strings might be different too. To quantify this gap, we conduct a preliminary experiment on τ^2 -bench, where the user is simulated with another LLM equipped with a distinct prompt from that of the agent LLM. In Figure 3, we compare the trajectory-level average NLL density plots driven from the ground truth user simulator and the agent LLM. We see remarkable deviations between them, stressing the need for deliberation to estimate observation uncertainty. Adopting auxiliary LLM as an approximated world model can be a possible solution, where we see some promises in Figure 8.

4.3 Modeling Uncertainty Dynamics in Interactive Systems

One of the most distinctive features of agent UQ is that the model now operates in a non-isolated environment to generate a trajectory from iterative interactions. This presents a unique challenge: modeling multi-turn cascade uncertainty in an interactive system. The desired UQ method should not only model the propagation of uncertainty across the trajectory but also *account for additional information gained through interactions*. To be specific, an agent in this open environment can gather extra information, make clarifications, and request confirmations at each step by interacting with the user and databases through conversations and tool-

calling. These interactive actions are the driving force of goal achievement in the complex (and usually under-specified) agentic tasks in the wild (Li et al., 2026). Therefore, the uncertainty model, which can reliably inform the agent’s failures in this setting, should reflect **conditional uncertainty reduction through information-seeking behavior as well as the uncertainty propagation**.

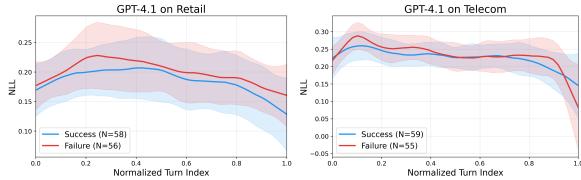


Figure 4: **Uncertainty evolution across the trajectory on τ^2 -bench.** We visualize the evolution of action uncertainty estimates across the normalized turn index, averaged over the entire trajectories from two groups: task success and failure.

Unfortunately, most existing multi-step UQ approaches discussed from Eq. 3 to Eq. 6 do not consider the interactivity of actions for each turn, and only focus on uncertainty propagation, while neglecting the reducible nature of uncertainty dynamics in non-isolated, interactive systems. In Figure 4, we present the uncertainty evolution of a naive averaging-based uncertainty aggregation method (Eq. 6), which is averaged over all trajectories from both success and failure groups. This simple, action-independent aggregation fails to meaningfully discern the failure group from the success group; Even the failure group achieves a sharper uncertainty decrease in the later part of trajectories in Telecom domain. This implies that the naive cascade modeling cannot robustly identify the agent’s failure both in the process and at the termination of the task, indicating the necessity of a new framework, pursuing interactivity-centric action-conditional modeling. We discuss a possible recommendation on this action-conditional uncertainty dynamics model in Appendix E.

4.4 Lack of Fine-grained Benchmarks

In the meantime, similar to the process reward modeling (Choudhury, 2025), annotating long-horizon agent trajectories incurs non-trivial effort and cost. The absence of fine-grained benchmarks becomes a natural bottleneck to developing a solid UQ method for agents. In this subsection, we provide a mini-survey of existing agent or multi-turn interaction benchmarks (See Appendix G), focusing primar-

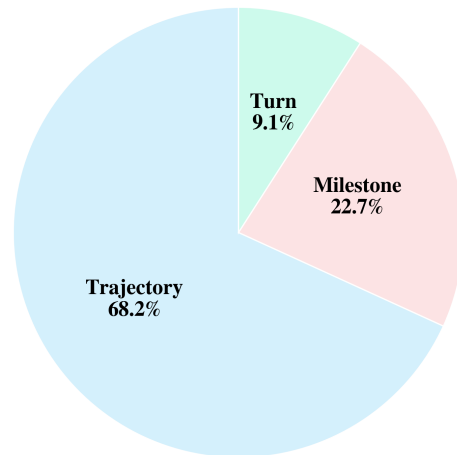


Figure 5: **Survey on 44 modern agent benchmarks.** The number denotes the percentage of papers in each evaluation category.

ily on LLM-centric agents rather than multimodal or physical ones. Specifically, we categorize 44 benchmarks into three classes depending on evaluation granularity: trajectory-level (conducted once at the end of the trajectory), milestone-level (involving several intermediate milestones or events), and turn-level (conducted at every single turn). Figure 5 reveals that only 4 out of 44 benchmarks provide turn-level annotation, which is very scarce compared to the trajectory-level benchmarks. Besides, although the milestone-level alternative shows better population than the turn-level one, it is still lacking, resulting in three times less than the trajectory-level one.

There are several obstacles to developing fine-grained benchmarks for agentic inference setups. First, agent benchmarks often involve very long rollouts that can easily exceed a hundred steps, making it hard to manually annotate each step. Although using frontier models as a substitute for human labeling has become increasingly common, running such models on long-horizon tasks can be prohibitively expensive for many researchers. Moreover, agent benchmarks are usually defined by complex policies; even for humans, it may take a couple of hours to fully understand the task configuration. To accelerate the development of agent UQ, we should pay greater attention to scalable dataset construction strategies for agentic inference.

5 Practical Implications

Developing the agent UQ framework is not merely a theoretical exercise but a prerequisite for deploying LLM agents in non-deterministic real environments. We outline implications for *frontier LLM research* and three specialized domains: *healthcare*, *programming*, and *robotics*, in which agent UQ may have profound downstream effects, thereby incentivizing policymakers, practitioners, and researchers.

5.1 Advancing Frontier LLMs

Test-time scaling through reasoning is now a de facto standard technique to enhance the quality of LLM responses on challenging tasks (Wei et al., 2022; Muennighoff et al., 2025). However, there is still huge room for improvement, e.g., mitigating overthinking and/or pursuing efficiency (Sui et al., 2025; Aggarwal et al., 2025). Agent UQ provides concrete guidance to develop an adaptive reasoning method, i.e., early stopping or interaction invoking, via uncertainty budgeting. In addition, multi-turn reinforcement learning has emerged as a promising yet challenging research theme in eliciting the agentic behavior and reasoning of LLMs (Zhou et al., 2024c; Shani et al., 2024; Zeng et al., 2025). Agent UQ helps tackle the main bottlenecks, credit assignment and when-to-explore problems (Thrun, 1992; Pecka and Svoboda, 2014; Choudhury, 2025), limited positive feedback (Lee et al., 2025, 2026), and inspires an information pursuit policy (Chattopadhyay et al., 2022) by leveraging directional and quantitative uncertainty information as feedback.

5.2 Reliable Clinical Decision Support

Modern LLM agents offer a promising vision for healthcare, moving beyond static knowledge retrieval to autonomous systems with complex reasoning and tool-execution capabilities (Wang et al., 2025c). However, the transition from promising prototypes to reliable clinical agents faces critical obstacles stemming from the system’s inability to recognize and manage its own limitations. We believe that agent UQ can fill this gap. For example, although Ferber et al. (2025) observed a huge potential of tool-calling LLM agents in oncology diagnoses, the remaining error rate underscores the necessity of a “human-in-the-loop” workflow. Here, agent UQ can serve as a gatekeeper, e.g., flagging highly accumulated uncertainty moments to invite humans, automating during low-uncertainty peri-

ods, while pursuing total uncertainty reduction for the right final diagnosis. This aligns with the concept: *adaptive healthcare* that envisions a paradigm shift from passive prediction to risk-aware active inquiry (Hinostroza Fuentes et al., 2025), to tackle some real clinical settings such as disease stage prediction (Yin et al., 2023) or brain aging assessment (Cheng et al., 2026).

5.3 Reliable Software Engineering Agents

Although advanced coding agents (Anthropic, 2025; OpenAI, 2025) have pushed performance on standard benchmarks (Jimenez et al., 2024) to an impressive degree, the disparity between passing a controlled test and safely modifying a production-level codebase stresses the urgent need for a reliable agent UQ framework. Agent UQ enables a more controllable exploration and commitment mechanism, wherein the agentic system knows uncertainty over bug patch candidates and decides whether to gather more evidence, e.g., check more files, commit to a fix, or interact with the user. Besides, *uncertainty-triggered rollback and branching algorithm*, in which an agent treats increasing uncertainty after an action as a signal to revert or branch by mirroring the checkpoint-rollback workflows of human engineers, can lead to reliable agentic coders.

5.4 Reliable Embodied Agents in Cyber-Physical Systems

Embodied agents in cyber-physical systems face heterogeneous uncertainty from sensing, dynamics, and human intent (Li et al., 2024; Fung et al., 2025; Gemini Robotics, 2025), and they actively reduce uncertainty through interaction by re-sensing the environment or interacting with humans before committing to irreversible physical actions (Xu et al., 2025c). For example, a robot instructed to retrieve a fragile object may explore or seek clarification rather than execute a grasp under high uncertainty. This makes embodied autonomy an instance of a conditional uncertainty reduction process, where information-gathering actions reduce uncertainty and state-changing actions increase commitment and risk. By explicitly quantifying and managing uncertainty dynamics, agent UQ offers *safer action selection* (Zhang et al., 2025a), *principled delegation with minimal feedback* (Hagenow and Shah, 2025), and *robust execution* (Römer et al., 2025).

6 Open Problems

We have so far presented a roadmap for agent UQ with a general formulation, identified challenges, and practical implications, but important open problems remain as follows.

Intrinsic solution multiplicity. Is high uncertainty per step due to the agent’s lack of knowledge about the right action for that step? Or, is it due to the intrinsic multiplicity of valid actions in that step? We cannot identify which of them is the right source of high uncertainty! Besides, in contrast to the classic QA tasks, the task assigned to an agent is too complex to be perfectly specified at once. Thus, each intermediate step is intrinsically ambiguous and allows multiple valid actions, making it impossible to specify the source of the high uncertainty. A concurrent work points out this multiplicity issue and approaches it through an information-theory-inspired framework (Suri et al., 2025). Graph-based parallelism/multiplicity modeling might also be worth exploring as a possible solution here (Wu et al., 2025; Choi and Li, 2026).

Evaluation beyond task failure. Evaluation of predictive uncertainty has been conducted by measuring the informativeness of uncertainty estimates when predicting the incorrectness of the final answer. While suitable in single-turn QA settings, it becomes insufficient when we move to agentic LLM setups. In such setups, the sole prediction of eventual task failure collapses a rich, structured problem-solving process into a single scalar judgment, discarding useful information that makes uncertainty more actionable. We envision a new evaluation protocol that integrates dynamic solution multiplicity, task difficulty, and irreducible external ambiguity, as well as task failure, that may redefine uncertainty from a multi-facet view (Guo et al., 2025). This reframing calls for new metrics that credit not just predictive correlation with outcomes but the informativeness of uncertainty signals in multiple factors in agentic scaffolding, which is yet to be explored.

Uncertainty modeling in multi-agent systems. Although we confine our attention to a single-agent environment, there has been notable advancement in multi-agent systems (MAS) (Liang et al., 2023; Du et al., 2024; Guo et al., 2024; Wu et al., 2024a; Choi et al., 2025). Uncertainty dynamics in this MAS can be leveraged to improve the quality of inter-agent communication (Yoffe et al., 2025) and

mitigate debate collapse (Tang et al., 2026), thus worth exploring. Unfortunately, while there are some works that borrow the strength of collective intelligence from MAS to improve UQ quality in a single-turn QA setup (Yang et al., 2024b; Feng et al., 2025), no existing work in this field attempts to design a UQ method tailored for MAS. Modeling the joint uncertainty dynamics of multiple agents brings extra challenges that may require deliberation and other formal toolkits (Tsallis and Brigatti, 2004; Jø sang, 2016; Cheng et al., 2021). Nonetheless, our graphical model naturally extends to coupled per-agent trajectories with a shared environment state, offering a foundation for UQ in future multi-agent systems.

Uncertainty modeling in self-improving agents. In our setup, we assume that the agent only operates within a single episode. However, the agent research community has recently shown an increasing interest in self-evolving/improving agents (Bai et al., 2022; Shinn et al., 2023; Madaan et al., 2023; Wang et al., 2024a; Jiang et al., 2025a) that adapt across multiple episodes. Uncertainty information can also be utilized in this setup for reliable continual adaptation (Zhang et al., 2026b). However, the interactive self-evolving environment adds another non-trivial factor that shifts uncertainty dynamics as shown in (Huang et al., 2025). Future exploration in this setup should reflect the non-stationary toolkit, context memory, and agent model parameters across multiple episodes.

7 Conclusion

We argue that LLM UQ research should move towards a more realistic setup, i.e., agentic inference, featuring the interactive long-horizon task episode, which requires a paradigm shift from a point-wise estimate to the sequential dynamics model of uncertainty. We laid the foundation for this by providing a concrete definition and general formulation of agent UQ that abstracts a wide variety of existing UQ approaches. With that, we identified four key challenges emerging in UQ on agentic scaffolding with detailed explanations and numerical analyses. We further note the future direction to explore agent UQ from the perspectives of real-world applications and open problems to provoke forward-looking discussions. We hope the three pillars provided in this work help communities design their upcoming works, products, and more.

Limitations

Agentic systems in the wild may encounter unreliable (even adversarial; Kang et al. (2025a)) observations and a stochastically evolving environment. However, we do not dive deeper into those settings, and our formulation might not provide a proper abstraction. In that scenario, we may need to explicitly model uncertainty in the agent’s evidence stream as well, e.g., trust/ignorance about tool outputs, retrieval, and user-provided facts, by leveraging formalisms, such as subjective logic (Jsang, 2018; Cheng et al., 2020a), imprecise probability theory (Walley, 1991), and formal belief representation (Genin and Huber, 2022), to safely manage memory and act under unreliable observations in stochastic world.

Ethical Considerations

This work aims to establish a foundation for uncertainty quantification of LLM agents, which can directly contribute to society, science, and human well-being (Han et al., 2023; Ren et al., 2025; Oh et al., 2025c; Hendrycks et al., 2025). We believe that the realization of our proposed framework can significantly boost the reliability of the LLM agent in a dynamic, interactive inference setup, thereby facilitating risk mitigation in high-stakes decision-making. However, we also acknowledge that the proposed framework can be misused by allowing adversaries to efficiently steer the agent towards its blind spots or manipulate uncertainty thresholds to bypass safety filters (Wang et al., 2025d). Therefore, we recommend alleviations such as limiting exposure to attack-enabling signals, red-teaming, and monitoring for abuse, which avoid overpromising what low uncertainty implies.

Acknowledgments

We sincerely thank Artem Shelmanov, Shawn Im, Hyeong Kyu Choi, Jongwon Jeong, Eunsu Kim, Mingyu Kim, Ayoung Lee, JungEun Kim, Hayun Lee, Hoyoon Byun, and Kyungwoo Song for their sharp feedback on the draft. This work is supported in part by the AFOSR Young Investigator Program under award number FA9550-23-1-0184, National Science Foundation under awards IIS-2237037 and IIS-2331669, Office of Naval Research under grant number N00014-23-1-2643, Schmidt Sciences Foundation, Open Philanthropy, Alfred P. Sloan Fellowship, and gifts from Google

and Amazon. Paul Bogdan acknowledges the support by the National Science Foundation (NSF) under the NSF Award 2243104 under the Center for Complex Particle Systems (COMPASS), NSF Mid-Career Advancement Award BCS-2527046, U.S. Army Research Office (ARO) under Grant No. W911NF-23-1-0111, Defense Advanced Research Projects Agency (DARPA) Young Faculty Award and DARPA Director Fellowship Award, Okawa foundation award, National Institute of Health (NIH) under R01 AG 079957, and Intel faculty awards.

References

- Pranjal Aggarwal, Seungone Kim, Jack Lanchantin, Sean Welleck, Jason Weston, Ilya Kulikov, and Swarnadeep Saha. 2025. *Optimalthinkingbench: Evaluating over and underthinking in llms*. *arXiv preprint arXiv:2508.13141*.
- Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, and 1 others. 2022. *Do as i can, not as i say: Grounding language in robotic affordances*. *arXiv preprint arXiv:2204.01691*.
- Lukas Aichberger, Kajetan Schweighofer, and Sepp Hochreiter. 2024. *Rethinking uncertainty estimation in natural language generation*. *arXiv preprint arXiv:2412.15176*.
- Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, J Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. 2025. *Agentharm: A benchmark for measuring harmfulness of LLM agents*. In *The Thirteenth International Conference on Learning Representations*.
- Anthropic. 2025. *Claude sonnet 4.5 system card*. Technical report, Anthropic PBC.
- Anthropic. 2026. *Cowork: Claude code for the rest of your work*. <https://claude.com/blog/cowork-research-preview>.
- Axel Backlund and Lukas Petersson. 2025. *Vending-bench: A benchmark for long-term coherence of autonomous agents*. *arXiv preprint arXiv:2502.15840*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. *Constitutional ai: Harmlessness from ai feedback*. *arXiv preprint arXiv:2212.08073*.

- Chaithanya Bandi, Ben Hertzberg, Geobio Boo, Tejas Polakam, Jeff Da, Sami Hassaan, Manasi Sharma, Andrew Park, Ernesto Hernandez, Dan Rambado, Ivan Salazar, Rafael Cruz, Chetan Rane, Ben Levin, Brad Kenstler, and Bing Liu. 2026. Mcp-atlas: A large-scale benchmark for tool-use competency with real mcp servers. *arXiv preprint arXiv:2602.00933*.
- Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ^2 -bench: Evaluating conversational agents in a dual-control environment. *arXiv preprint arXiv:2506.07982*.
- Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR.
- Antoine Bigeard, Langston Nashold, Rayan Krishnan, and Shirley Wu. 2025. Finance agent benchmark: Benchmarking llms on real-world financial research tasks. *arXiv preprint arXiv:2508.00828*.
- Ovidiu Calin and Constantin Udriște. 2014. *Informational Energy*, pages 133–163. Springer International Publishing.
- Hyungjoo Chae, Namyong Kim, Kai Tzu iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. 2025. Web agents with world models: Learning and leveraging environment dynamics in web navigation. In *International Conference on Learning Representations*.
- Kranti Chalamalasetti, Jana Götz, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 11174–11219.
- Kwan Ho Ryan Chan, Yuyan Ge, Edgar Dobriban, Hamed Hassani, and René Vidal. 2025. Conformal information pursuit for interactively guiding large language models. *arXiv preprint arXiv:2507.03279*.
- Ma Chang, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. 2024. Agentboard: An analytical evaluation board of multi-turn llm agents. *Advances in neural information processing systems*, 37:74325–74362.
- Aditya Chattopadhyay, Stewart Slocum, Benjamin D Haeffele, René Vidal, and Donald Geman. 2022. Interpretable by design: Learning predictors by composing interpretable queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7430–7443.
- Chaoran Chen, Zhiping Zhang, Bingcan Guo, Shang Ma, Ibrahim Khalilov, Simret A Gebreegzabher, Yanfang Ye, Ziang Xiao, Yaxing Yao, Tianshi Li, and 1 others. 2025a. The obvious invisible threat: Llm-powered gui agents’ vulnerability to fine-print injections. *arXiv preprint arXiv:2504.11281*.
- Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. 2025b. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *The Thirteenth International Conference on Learning Representations*.
- Anzhe Cheng, Italo Ivo Lima Dias Pinto, and Paul Bogdan. 2026. Structural complexity of brain mri reveals age-associated patterns. *arXiv preprint arXiv:2601.17211*.
- Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*.
- Mingxi Cheng, Shahin Nazarian, and Paul Bogdan. 2020a. There is hope after all: Quantifying opinion and trustworthiness in neural networks. *Frontiers in artificial intelligence*, 3:54.
- Mingxi Cheng, Chenzhong Yin, Junyao Zhang, Shahin Nazarian, Jyotirmoy Deshmukh, and Paul Bogdan. 2021. A general trust framework for multi-agent systems. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 332–340.
- Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020b. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR.
- John Cherian, Isaac Gibbs, and Emmanuel Candes. 2024. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842.
- Hyeong Kyu Choi and Sharon Li. 2026. Modex: Evaluator-free best-of-n selection for open-ended generation. *arXiv preprint arXiv:2601.02535*.
- Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. 2025. Debate or vote: Which yields better decisions in multi-agent large language models? In *Advances in Neural Information Processing Systems*. Spotlight.
- Sanjiban Choudhury. 2025. Process reward models for llm agents: Practical framework and directions. *arXiv preprint arXiv:2502.10325*.
- Sanghyuk Chun. 2025. Multiplicity is an inevitable and inherent challenge in multimodal learning. *arXiv preprint arXiv:2505.19614*.

- Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. 2024. Gamebench: Evaluating strategic reasoning abilities of llm agents. *arXiv preprint arXiv:2406.06613*.
- Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114.
- Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. 2024. Workarena: How capable are web agents at solving common knowledge work tasks? In *International Conference on Machine Learning*, pages 11642–11662. PMLR.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*.
- Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063.
- Jinhao Duan, James Diffenderfer, Sandeep Madireddy, Tianlong Chen, Bhavya Kailkhura, and Kaidi Xu. 2025. Uprop: Investigating the uncertainty propagation of llms in multi-step agentic decision-making. *arXiv preprint arXiv:2506.17419*.
- Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, and 1 others. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9367–9385.
- Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, and 1 others. 2023. Lmpolygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.
- Yu Feng, Phu Mon Htut, Zheng Qi, Wei Xiao, Manuel Mager, Nikolaos Pappas, Kishalay Halder, Yang Li, Yassine Benajiba, and Dan Roth. 2025. Rethinking LLM uncertainty: A multi-agent approach to estimating black-box model uncertainty. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12349–12375. Association for Computational Linguistics.
- Dyke Ferber, Omar SM El Nahhas, Georg Wölflein, Isabella C Wiest, Jan Clusmann, Marie-Elisabeth Leßmann, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, and 1 others. 2025. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature cancer*, pages 1–13.
- Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.
- Eric Frankel, Shuyue Stella Li, Lillian J. Ratliff, Yulia Tsvetkov, Sewoong Oh, and Pang Wei Koh. 2024. Conformal reasoning: Uncertainty estimation in interactive environments.
- Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, and 1 others. 2026. Gaia2: Benchmarking llm agents on dynamic and asynchronous environments. *arXiv preprint arXiv:2602.11964*.
- Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. Deep think with confidence. *arXiv preprint arXiv:2508.15260*.
- Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, and 1 others. 2025. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*.
- Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.
- Murray Gell-Mann and Constantino Tsallis. 2004. *Nonextensive Entropy: Interdisciplinary Applications*. Oxford University Press.
- Gemini Robotics. 2025. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*.
- Konstantin Genin and Franz Huber. 2022. Formal Representations of Belief. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2022 edition. Metaphysics Research Lab, Stanford University.

- John Gill. 1977. Computational complexity of probabilistic turing machines. *SIAM Journal on Computing*, 6(4):675–695.
- John T Gill III. 1974. Computational complexity of probabilistic turing machines. In *Proceedings of the sixth annual ACM symposium on Theory of computing*, pages 91–95.
- Google. 2025. Gemini agent. <https://gemini.google/overview/agent/>. Accessed: 2025-12-11.
- Google. 2026. Gemini 3.1 pro: A smarter model for your most complex tasks. <https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/>.
- German Gritsai, Megan Richards, Maxime M eloux, Kyunghyun Cho, and Maxime Peyrard. 2025. Mist: Mutual information via supervised training. *arXiv preprint arXiv:2511.18945*.
- Pei-Fu Guo, Yun-Da Tsai, and Shou-De Lin. 2025. Uncertainty profiles for llms: Uncertainty source decomposition and adaptive model-metric selection. *arXiv preprint arXiv:2505.07309*.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xi-angliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.
- Michael Hagenow and Julie A Shah. 2025. Realm: Real-time estimates of assistance for learned models in human-robot interaction. *IEEE Robotics and Automation Letters*.
- Shayan Mohajer Hamidi, Linfeng Ye, and Konstantinos N. Plataniotis. 2026. Dreamphase: Offline imagination and uncertainty-guided planning for large-language-model agents. In *The Fourteenth International Conference on Learning Representations*.
- Jieun Han, Haneul Yoo, Yoonsu Kim, Junho Myung, Minsun Kim, Hyunseung Lim, Juho Kim, Tak Yeon Lee, Hwajung Hong, So-Yeon Ahn, and 1 others. 2023. Recipe: How to integrate chatgpt into efl writing education. In *Proceedings of the tenth ACM conference on learning@ scale*, pages 416–420.
- Jiuzhou Han, Wray Buntine, and Ehsan Shareghi. 2024. Towards uncertainty-aware language agent. *arXiv preprint arXiv:2401.14016*.
- Dan Hendrycks, Dawn Song, Christian Szegedy, Honglak Lee, Yarin Gal, Erik Brynjolfsson, Sharon Li, Andy Zou, Lionel Levine, Bo Han, and 1 others. 2025. A definition of agi. *arXiv preprint arXiv:2510.18212*.
- Vasco Gerardo Hinostroza Fuentes, Hezerul Abdul Karim, Myles Joshua Toledo Tan, and Nouar Al-Dahoul. 2025. Ai with agency: a vision for adaptive, efficient, and ethical healthcare. *Frontiers in Digital Health*, 7:1600216.
- Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 19023–19042. PMLR.
- Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei W Koh, and Bryan Hooi. 2024. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in llms. *Advances in Neural Information Processing Systems*, 37:24181–24215.
- Liangjie Huang, Dawei Li, Huan Liu, and Lu Cheng. 2025. Beyond accuracy: The role of calibration in self-improving large language models. *arXiv preprint arXiv:2504.02902*.
- Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024. MLAGentBench: Evaluating language agents on machine learning experimentation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20271–20309. PMLR.
- Pengcheng Jiang, Jiacheng Lin, Zhiyi Shi, Zifeng Wang, Luxi He, Yichen Wu, Ming Zhong, Peiyang Song, Qizheng Zhang, Heng Wang, and 1 others. 2025a. Adaptation of agentic ai. *arXiv preprint arXiv:2512.16301*.
- Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025b. Medagentbench: A virtual ehr environment to benchmark medical llm agents. *NEJM AI*, 2(9).
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*.
- Audun J o sang. 2016. Generalising bayes’ theorem in subjective logic. In *2016 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 462–469. IEEE.
- Audun Jsang. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134.

- Hangoo Kang, Jehyeok Yeon, and Gagandeep Singh. 2025a. Trap: Targeted redirecting of agentic preferences. *arXiv preprint arXiv:2505.23518*.
- Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025b. Scalable best-of-n selection for large language models via self-certainty. In *Advances in Neural Information Processing Systems*.
- Kimi Team. 2026. Kimi k2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.
- Michael Kirchhof, Gjergji Kasneci, and Enkelejda Kasneci. 2025. Position: Uncertainty quantification needs reassessment for large language model agents. In *Forty-second International Conference on Machine Learning Position Paper Track*.
- Mykel J Kochenderfer. 2015. *Decision making under uncertainty: theory and application*. MIT press.
- Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, Fedor Noskov, Alexander Fishkov, Artem Shelmanov, Artem Vazhentsev, Aleksandr Petiushko, and Maxim Panov. 2022. Nonparametric uncertainty quantification for single deterministic neural network. *Advances in Neural Information Processing Systems*, 35:36308–36323.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.
- Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*.
- Joshua Ong Jun Leang, Zheng Zhao, Aryo Pradipta Gema, Sohee Yang, Wai-Chung Kwan, Xuanli He, Wenda Li, Pasquale Minervini, Eleonora Giunchiglia, and Shay B Cohen. 2025. Picsar: Probabilistic confidence selection and ranking for reasoning chains. *arXiv preprint arXiv:2508.21787*.
- Sangyun Lee, Brandon Amos, and Giulia Fanti. 2025. Banel: Exploration posteriors for generative modeling using only negative rewards. *arXiv preprint arXiv:2510.09596*.
- Seonggyun Lee, Sungjun Lim, Seojin Park, Soeun Cheon, and Kyungwoo Song. 2026. Semi-supervised preference optimization with limited feedback. In *International Conference on Learning Representations*.
- Ido Levy, Ben Wiesel, Sami Marreed, Alon Oved, Avi Yaeli, and Segev Shlomov. 2024. Stwebagentbench: A benchmark for evaluating safety and trustworthiness in web agents. *arXiv preprint arXiv:2410.06703*.
- Bowen Li, Wenhan Wu, Ziwei Tang, Lin Shi, John Yang, Jinyang Li, Shunyu Yao, Chen Qian, Binyuan Hui, Qicheng Zhang, and 1 others. 2025a. Prompting large language models to tackle the full software development lifecycle: A case study. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7511–7531.
- Jiatong Li, Changdae Oh, Hyeong Kyu Choi, Jindong Wang, and Sharon Li. 2026. Thinking makes llm agents introverted: How mandatory thinking can backfire in user-engaged agents. *arXiv preprint arXiv:2602.07796*.
- Junlong Li, Wenshuo Zhao, Jian Zhao, Weihao Zeng, Haoze Wu, Xiaochen Wang, Rui Ge, Yuxuan Cao, Yuzhen Huang, Wei Liu, Junteng Liu, Zhaochen Su, Yiyang Guo, Fan Zhou, Lueyang Zhang, Juan Michelini, Xingyao Wang, Xiang Yue, Shuyan Zhou, and 2 others. 2025b. The tool decathlon: Benchmarking language agents for diverse, realistic, and long-horizon task execution. *arXiv preprint arXiv:2510.25726*.
- Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, and 1 others. 2024. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534.
- Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. API-bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3102–3116. Association for Computational Linguistics.
- Wendi Li and Yixuan Li. 2025. Process reward model with q-value rankings. In *Proceedings of the International Conference on Learning Representations*.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. In *Conference on Empirical Methods in Natural Language Processing*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.
- Shudong Liu, Hongwei Liu, Junnan Liu, Linchen Xiao, Songyang Gao, Chengqi Lyu, Yuzhe Gu, Wenwei Zhang, Derek F Wong, Songyang Zhang, and 1 others. 2025a. Compassverifier: A unified and robust verifier for llms evaluation and outcome reward. In *Proceedings of the 2025 Conference on Empirical*

- Methods in Natural Language Processing*, pages 33454–33482.
- Wenrui Liu, Zixiang Liu, Elsie Dai, Wenhan Yu, Lei Yu, Tong Yang, Jinjun Han, and Hong Gao. 2025b. Mcpagentbench: A real-world task benchmark for evaluating llm agent mcp tool use. *arXiv preprint arXiv:2512.24565*.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*.
- Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. 2025. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1160–1183, Albuquerque, New Mexico. Association for Computational Linguistics.
- David J. C. MacKay. 1992. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.
- Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.
- Mike A Merrill, Alexander G Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E Kelly Buchanan, and 1 others. 2026. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. *arXiv preprint arXiv:2601.11868*.
- Grégoire Mialon, Clémentine Fourier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797.
- Sina Molavipour, Germán Bassi, and Mikael Skoglund. 2020. Conditional mutual information neural estimator. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5025–5029.
- Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. 2024. Rule based rewards for language model safety. *Advances in Neural Information Processing Systems*, 37:108877–108901.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332.
- Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. 2020. Ccm: Classifier based conditional mutual information estimation. In *Uncertainty in artificial intelligence*, pages 1083–1093. PMLR.
- Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. [Contrasting linguistic patterns in human and llm-generated news text](#). *Artificial Intelligence Review*, 57:265.
- Kevin Patrick Murphy. 2002. *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley.
- Radford Neal. 1992. Bayesian learning via stochastic dynamics. *Advances in neural information processing systems*, 5.
- Dang Nguyen, Jian Chen, Yu Wang, Gang Wu, Namyong Park, Zhengmian Hu, Hanjia Lyu, Junda Wu, Ryan Aponte, Yu Xia, Xintong Li, Jing Shi, Hongjie Chen, Viet Dac Lai, Zhouhang Xie, Sungchul Kim, Ruiyi Zhang, Tong Yu, Mehrab Tanjim, and 11 others. 2025. GUI agents: A survey. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22522–22538, Vienna, Austria. Association for Computational Linguistics.
- Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929.
- Changdae Oh, Zhen Fang, Shawn Im, Xuefeng Du, and Yixuan Li. 2025a. Understanding multimodal LLMs under distribution shifts: An information-theoretic approach. In *International Conference on Machine Learning*.

- Changdae Oh, Jiatong Li, Shawn Im, and Sharon Li. 2025b. Visual instruction bottleneck tuning. In *Advances in Neural Information Processing Systems*.
- Changdae Oh, Hyesu Lim, Mijoo Kim, Dongyoon Han, Sangdoon Yun, Jaegul Choo, Alexander Hauptmann, Zhi-Qi Cheng, and Kyungwoo Song. 2024. Towards calibrated robust fine-tuning of vision-language models. *Advances in Neural Information Processing Systems*, 37:12677–12707.
- Juhyun Oh, Eunsu Kim, and Alice Oh. 2025c. Flex-travelplanner: A benchmark for flexible planning with language agents. *arXiv preprint arXiv:2506.04649*.
- OpenAI. 2025. Chatgpt agent. <https://chatgpt.com/features/agent/>. Accessed: 2025-12-11.
- OpenAI. 2025. Introducing gpt-5.2-codex. <https://openai.com/index/introducing-gpt-5-2-codex/>.
- OpenAI. 2025. Introducing gpt-4.1 in the api.
- Yichen Pan, Dehan Kong, Sida Zhou, Cheng Cui, Yifei Leng, Bing Jiang, Hangyu Liu, Yanyi Shang, Shuyan Zhou, Tongshuang Wu, and 1 others. 2024. Webcanvas: Benchmarking web agents in online environments. *arXiv preprint arXiv:2406.12373*.
- L Pardo and IJ Taneja. 1991. Information energy and its applications. In *Advances in electronics and electron physics*, volume 80, pages 165–241. Elsevier.
- Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. 2020. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 3219–3229. PMLR.
- Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. 2025. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *International Conference on Machine Learning*.
- Martin Pecka and Tomas Svoboda. 2014. Safe exploration techniques for reinforcement learning—an overview. In *International workshop on modelling and simulation for autonomous systems*, pages 357–375. Springer.
- Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2025. Verif: Verification engineering for reinforcement learning in instruction following. *arXiv preprint arXiv:2506.09942*.
- Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations*.
- Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. In *International Conference on Learning Representations*.
- Alex Reinhart, Ben Markey, Michael Laudendach, Kachata Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.
- Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.
- Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. 2025. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*.
- Alfréd Rényi. 1961. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press.
- Zachary Robertson and Sanmi Koyejo. 2025. Let’s measure information step-by-step: Llm-based evaluation beyond vibes. *arXiv preprint arXiv:2508.05469*.
- Ralf Römer, Adrian Kobras, Luca Worbis, and Angela P. Schoellig. 2025. Failure prediction at runtime for generative robot policies. In *Advances in Neural Information Processing Systems*.
- Erwin Schrödinger. 1974. *What is Life? & Mind and Matter: The Physical Aspect of the Living Cell*. Cambridge University Press.
- Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, and 1 others. 2024. Multi-turn reinforcement learning with preference human feedback. *Advances in Neural Information Processing Systems*, 37:118953–118993.
- Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.
- Yucheng Shi, Wenhao Yu, Wenlin Yao, Wenhao Chen, and Ninghao Liu. 2025. Towards trustworthy gui agents: A survey. *arXiv preprint arXiv:2503.23434*.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. Alfworld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*.
- Zachary S Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebel, and Arvind Narayanan. 2024. CORE-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Transactions on Machine Learning Research*.
- Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2998–3009.
- Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. 2025. PaperBench: Evaluating AI’s ability to replicate AI research. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 56843–56873. PMLR.
- Peter Steinberger. 2026. Introducing openclaw. <https://openclaw.ai/blog/introducing-openclaw>.
- Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, and 1 others. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.
- Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A Rossi, and Dinesh Manocha. 2025. Structured uncertainty guided clarification for llm agents. *arXiv preprint arXiv:2511.08798*.
- Luoxi Tang, Yuqiao Meng, Joseph Costa, Yingxue Zhang, Muchao Ye, and Zhaohan Xi. 2026. The value of variance: Mitigating debate collapse in multi-agent systems via uncertainty-driven policy optimization. *arXiv preprint arXiv:2602.07186*.
- Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Quantifying uncertainty in natural language explanations of large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 1072–1080. PMLR.
- Linwei Tao, Yi-Fan Yeh, Minjing Dong, Tao Huang, Philip Torr, and Chang Xu. 2025. Revisiting uncertainty estimation and calibration of large language models. *arXiv preprint arXiv:2505.23854*.
- Sebastian B Thrun. 1992. *Efficient exploration in reinforcement learning*. Carnegie Mellon University.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442. Association for Computational Linguistics.
- Tim Tomov, Dominik Fuchsgruber, Tom Wollschläger, and Stephan Günnemann. 2025. The illusion of certainty: Uncertainty quantification for llms fails under ambiguity. *arXiv preprint arXiv:2508.15260*.
- Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. 2024. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16022–16076, Bangkok, Thailand. Association for Computational Linguistics.
- Constantino Tsallis. 1988. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487.
- Constantino Tsallis and Edgardo Brigatti. 2004. Nonextensive statistical mechanics: A brief introduction. *Continuum Mechanics and Thermodynamics*, 16(3):223–235.
- Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.
- I Vajda. 1968. Bounds of the minimal error probability on checking a finite or countable number of hypotheses. *Problemy Peredachi Informatsii*, 4(1):9–19.
- Igor Vajda and Jana Zvárová. 2007. On generalized entropies, bayesian decisions and statistical diversity. *Kybernetika*, 43(5):675–696.
- Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. 2025. Uncertainty quantification for llms through minimum bayes risk: Bridging confidence and consistency. *arXiv preprint arXiv:2502.04964*.
- Peter Walley. 1991. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024a. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*.

- Jiawei Wang, Jiakai Liu, Yuqian Fu, Yingru Li, Xintao Wang, Yuan Lin, Yu Yue, Lin Zhang, Yang Wang, and Ke Wang. 2025a. Harnessing uncertainty: Entropy-modulated policy gradients for long-horizon llm agents. *arXiv preprint arXiv:2509.09265*.
- Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025b. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.
- Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025c. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*.
- Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024b. Executable code actions elicit better llm agents. In *Forty-first International Conference on Machine Learning*.
- Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2024c. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. In *International Conference on Learning Representations*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.
- Zhun Wang, Vincent Siu, Zhe Ye, Tianneng Shi, Yuzhou Nie, Xuandong Zhao, Chenguang Wang, Wenbo Guo, and Dawn Song. 2025d. Agentvigil: Automatic black-box red-teaming for indirect prompt injection against llm agents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23159–23172.
- Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Jiaqi Wu, Qinlao Zhao, Zefeng Chen, Kai Qin, Yifei Zhao, Xueqian Wang, and Yuhang Yao. 2025. GAP: Graph-based agent planning with parallel tool use and reinforcement learning. In *NORA: The First Workshop on Knowledge Graphs & Agentic Systems Interplay*.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024a. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *Conference on Language Modeling*.
- Yue Wu, Xuan Tang, Tom Mitchell, and Yuanzhi Li. 2024b. Smartplay : A benchmark for LLMs as intelligent agents. In *The Twelfth International Conference on Learning Representations*.
- Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Xin Guo, Dingwen Yang, Chenyang Liao, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. 2025. AgentGym: Evaluating and training large language model-based agents across diverse environments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27914–27961, Vienna, Austria. Association for Computational Linguistics.
- Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024a. Travelplanner: A benchmark for real-world planning with language agents. In *International Conference on Machine Learning*, pages 54590–54613. PMLR.
- Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, and 1 others. 2024b. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.
- Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. 2025. How memory management impacts llm agents: An empirical study of experience-following behavior. *arXiv preprint arXiv:2505.16067*.
- Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Zhiruo Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Melroy Maben, Raj Mehta, Wayne Chi, Lawrence Keunho Jang, Yiqing Xie, and 2 others. 2025a. Theagentcompany: Benchmarking LLM agents on consequential real world tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025b. A-mem:

- Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.
- Xiaomeng Xu, Yifan Hou, Zeyi Liu, and Shuran Song. 2025c. Compliant residual dagger: Improving real-world contact-rich manipulation with human corrections. *arXiv preprint arXiv:2506.16685*.
- Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024a. On verbalized confidence scores for llms. *arXiv preprint arXiv:2412.14737*.
- John Yang, Akshara Prabhakar, Karthik R Narasimhan, and Shunyu Yao. 2023. Intercode: Standardizing and benchmarking interactive coding with execution feedback. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Ruixin Yang, Dheeraj Rajagopal, Shirley Anugrah Hayati, Bin Hu, and Dongyeop Kang. 2024b. Confidence calibration and rationalization for LLMs via multi-agent deliberation. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.
- Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022a. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ -bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022b. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.
- Chenzhong Yin, Mihai Udrescu, Gaurav Gupta, Mingxi Cheng, Andrei Lihu, Lucretia Udrescu, Paul Bogdan, David M Mannino, and Stefan Mihaicuta. 2023. Fractional dynamics foster deep learning of copd stage prediction. *Advanced Science*, 10(12):2203485.
- Sheng Yin, Xianghe Pang, Yuanzhuo Ding, Menglan Chen, Yutong Bi, Yichen Xiong, Wenhao Huang, Zhen Xiang, Jing Shao, and Siheng Chen. 2024. Safeagentbench: A benchmark for safe task planning of embodied llm agents. *arXiv preprint arXiv:2412.13178*.
- Luke Yoffe, Alfonso Amayuelas, and William Yang Wang. 2025. DebUnc: Improving large language model agent communication with uncertainty metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23299–23315. Association for Computational Linguistics.
- Gal Yona, Roei Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764.
- Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. 2024. Assistantbench: Can web agents solve realistic and time-consuming tasks? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8938–8968.
- Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, Yang Katie Zhao, and Mingyi Hong. 2025. Reinforcing multi-turn reasoning in llm agents via turn-level credit assignment. In *ICML 2025 Workshop on Computer Use Agents*.
- Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Josef Dai, Yuanpei Chen, and Yaodong Yang. 2025a. SafeVLA: Towards safety alignment of vision-language-action model via constrained learning. In *Advances in Neural Information Processing Systems*.
- Caiqi Zhang, Ruihan Yang, Xiaochen Zhu, Chengzu Li, Tiancheng Hu, Yijiang River Dong, Deqing Yang, and Nigel Collier. 2026a. Confidence estimation for llms in multi-turn interactions. *arXiv preprint arXiv:2601.02179*.
- Dengjia Zhang, Xiaoou Liu, Lu Cheng, Yaqing Wang, Kenton Murray, and Hua Wei. 2026b. Selaur: Self evolving llm agent via uncertainty-aware rewards. *arXiv preprint arXiv:2602.21158*.
- Jiaxin Zhang, Prafulla Kumar Choubey, Kung-Hsiang Huang, Caiming Xiong, and Chien-Sheng Wu. 2026c. Agentic uncertainty quantification. *arXiv preprint arXiv:2601.15703*.
- Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932.
- Tunyu Zhang, Haizhou Shi, Yibin Wang, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Haoxian Chen, Ligong Han, Kai Xu, Huan Zhang, and 1 others. 2025b. Token-level uncertainty estimation for large language model reasoning. *arXiv preprint arXiv:2505.11737*.
- Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. 2024. Agent-safetybench: Evaluating the safety of llm agents. *arXiv preprint arXiv:2412.14470*.

- Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyun Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, Chen Zhao, and 1 others. 2025. Uncertainty propagation on llm agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6064–6073.
- Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. 2025a. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning*.
- Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Anderson Compalas, Dawn Song, and Xin Eric Wang. 2025b. Multimodal situational safety. In *The Thirteenth International Conference on Learning Representations*.
- Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024a. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*.
- Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024b. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.
- Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024c. ArCHer: Training language model agents via hierarchical multi-turn RL. In *International Conference on Machine Learning (ICML)*.
- Yuqi Zhu, Ge Li, Xue Jiang, Jia Li, Hong Mei, Zhi Jin, and Yihong Dong. 2025. Uncertainty-guided chain-of-thought for code generation with llms. *arXiv preprint arXiv:2503.15341*.
- Mingchen Zhuge, Changsheng Zhao, Dylan R. Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. 2025. Agent-as-a-judge: Evaluate agents with agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80569–80611. PMLR.

Appendix

Contents

A Revision History	21
B Additional Details on Formulation	21
B.1 Expression of Various Agentic Prompting under Our Formulation	21
B.2 Uncertainty Instantiations and Total Uncertainty Expansion	21
C Alternative Uncertainty Measures	22
D Extended Context	22
E Action-Conditional Uncertainty Dynamics Model	23
E.1 Implementation Sketch: Conditional Uncertainty Reduction Process	24
E.2 Theoretical Analysis	26
E.3 Action Categorization	26
F Experiment	26
F.1 Setup	26
F.2 Extended Results	27
G Mini-Survey on Agent Benchmark	28
H Discussion and Future Work	28
A Revision History	

Our draft titled "*Towards Reducible Uncertainty Modeling for Reliable Large Language Model Agents*" was specifically targeted at the third challenge described in the current Section 4.3 by proposing a conceptual framework, conditional uncertainty reduction process, to enable action-conditional uncertainty dynamics modeling for interactive agents, which is available now at Section E in this Appendix. After receiving feedback from the reviewers and community, we revised this article to cover a broad range of practical challenges emerging under agentic scaffolding, supported by quantitative analyses, rather than focusing narrowly on a single challenge and its corresponding solution. In the main body of the paper, Section 4 has undergone a major revision, while the remaining sections were subject to only moderate or minor edits.

B Additional Details on Formulation

B.1 Expression of Various Agentic Prompting under Our Formulation

In Section 3.1, we have introduced a definition of stochastic agent system (Def. 3.1) with a probabilistic graphical model of specific conditional dependency assumptions in Figure 2. Under this model, we can abstract some representative agentic prompting methods (Ahn et al., 2022; Yao et al., 2022b; Shinn et al., 2023; Patil et al., 2025) with a unified language yet distinctive conditions as below.

- 1. Vanilla function calling** (Patil et al., 2025): Same as Definition 3.1. Stochastic process of (A_t, E_t, O_t) without any extra conditions.
- 2. Few-shot prompting** (Ahn et al., 2022): Just augmenting initial task-specification variable as $E'_0 = E_0 \cup \{\mathcal{F}^{(i)}\}_{i=1}^N$ with some example rollouts $\mathcal{F}^{(i)} = \{(A_t, E_t, O_t)\}_{t=1}^{T^i}$, which may implicitly affect the action transition probability, $P_{\pi, \mathcal{T}}(\cdot)$, by conditioning.
- 3. ReAct** (Yao et al., 2022b; Wang et al., 2024b): Adding a constraint on the action sampling as $A_i \sim P_{\pi, \mathcal{T}}(A^{\text{thk}} | E_{i-1}, O_{i-1})$ if $A_{i-1} \in \mathcal{A}^{-\text{thk}}$ else $A_i \sim P_{\pi, \mathcal{T}}(A^{-\text{thk}} | E_{i-1}, O_{i-1})$, where $A^{\text{thk}} \in \mathcal{A}^{\text{thk}}$ and $A^{-\text{thk}} \in \mathcal{A}^{-\text{thk}}$ random variables from the partitions of \mathcal{A} standing for the space of textual reasoning and all others, respectively.
- 4. Reflexion** (Shinn et al., 2023): Do multiple trials $\mathcal{F}^{(\leq i)} = \{\mathcal{F}^{(1)}, \dots, \mathcal{F}^{(i)}\}$ until $r^{(i)} := r(\mathcal{F}^{(i)}) = 1$ or reach maximum trials, otherwise augmenting the per-trial initial state with an external long-term memory $E_0^{(i+1)} = E_0^{(i)} \cup \{\mathcal{M}(\mathcal{F}^{(j)}, r^{(j)})\}_{j \leq i}$, where the long-term memory $\mathcal{M}(\cdot)$ constructs with textual feedback generated from an LLM prompted by the current trial trajectory and a corresponding binary reward. Thus, the action transition probability, $P_{\pi, \mathcal{T}}(\cdot)$, is continually evolved by reflecting feedback from historical runs.

Overall, this suggests that our agent formulation is general and flexible enough to capture a broad class of representative agentic prompting scenarios.

B.2 Uncertainty Instantiations and Total Uncertainty Expansion

After defining the agent UQ in Def 3.2, we presented a trajectory-level total uncertainty that ex-

presses the joint uncertainty across multiple turns in an additive form of the initial query uncertainty, action uncertainty, and observation uncertainty.

Given the joint probability, $P(\mathcal{F}_{\leq T}) = P(E_0, O_0) \prod_{i=1}^T P_{\pi, \tau}(A_i | E_{i-1}, O_{i-1}) P(O_i | A_i, E_i)$, we enumerate three example instances of the uncertainty measure $U(\cdot)$ that induces the simple additive form of total uncertainty:

1. **Information content** (negative log probability), $U(X = x) := -\log P(X = x)$, to measure a point-wise surprisal for a given observation.
2. **Entropy**, $U(X) := H(X) = \mathbb{E}[-\log P(X)]$, to measure the expected amount of surprisal.
3. **Relative entropy**, $U(X) := \mathbb{D}_{\text{KL}}(Q(X) || P(X))$ with a pre-defined reference distribution $Q(X)$ and Kullback-Leibler divergence \mathbb{D}_{KL} , to measure the expected amount of surprisal given a prior knowledge (if we have any).

For information content, it is trivial to show by just taking the negative logarithm to $P(\mathcal{F}_{\leq T})$. For the entropy and relative entropy, we have the following chain rule (Cover, 1999),

$$\begin{aligned} \bullet H(\mathcal{F}_0, \dots, \mathcal{F}_T) &= \sum_{i=0}^T H(\mathcal{F}_i | \mathcal{F}_{i-1}, \dots, \mathcal{F}_1), \\ \bullet \mathbb{D}_{\text{KL}}(Q(\mathcal{F}_0, \dots, \mathcal{F}_T) || P(\mathcal{F}_0, \dots, \mathcal{F}_T)) &= \\ &= \sum_{i=0}^T \mathbb{D}_{\text{KL}}(Q(\mathcal{F}_i | \mathcal{F}_{i-1}, \dots, \mathcal{F}_1) || P(\mathcal{F}_i | \mathcal{F}_{i-1}, \dots, \mathcal{F}_1)), \end{aligned}$$

which directly drives our uncertainty expansion for $U(\mathcal{F}_{\leq T})$ in Section 3.1.

C Alternative Uncertainty Measures

Although we have mainly focused on the above three information-theoretic uncertainty measures throughout the paper, there are plenty of alternatives one can consider depending on the problem setups. In this section, we examine some possible alternative uncertainty measures, *Rényi entropy*, *Tsallis entropy*, and *informational energy*. These sophisticated measures may be worth investigating for the LLM agents' inference interface, characterized by long-range interactions, evolving memory, and the multifractal property.

Rényi entropy (Rényi, 1961) has been widely applied in modern machine learning, ecology, biodiversity science, and quantum information, among other fields. Given an order parameter $0 < \alpha < \infty$ and $\alpha \neq 1$, it is defined as $H_\alpha(X) = \frac{1}{1-\alpha} \log(\sum_{x \in \mathcal{X}} P(x)^\alpha)$ and generalize many other entropies: max-entropy $H_0(X) = \log |\mathcal{X}|$, Shannon

entropy $\lim_{\alpha \rightarrow 1} H_\alpha(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x)$, collision entropy $H_2(X) = -\log \sum_{x \in \mathcal{X}} P(x)^2$, and min-entropy $H_\infty(X) \doteq \min_{x \in \mathcal{X}} -\log P(x)$.

Tsallis entropy (Tsallis, 1988) extends this further to model a complex system where the additivity property of Shannon and Rényi entropies for independent subsystems does not hold. Given an index parameter $q \in \mathbb{R} \setminus \{1\}$, it is defined as $H_q(X) = \frac{1 - \sum_{x \in \mathcal{X}} P(x)^q}{q-1}$, which recover Shannon entropy when $q \rightarrow 1$. Along with Rényi entropy, Tsallis entropy has taken a key role in characterizing complex physical systems as a nonextensive measure of entropy (Tsallis and Brigatti, 2004).

Onicescu informational energy (Pardo and Taneja, 1991; Calin and Udriște, 2014) is another popular measure of (un)certainty that has bred many applications in quantum mechanics, economics, ecology, social sciences, and so on. It is defined as $\text{IE}(X) := \sum_{x \in \mathcal{X}} P(x)^2$ which can be connected to Rényi entropy of order-2 $H_2(X) = -\log \sum_{x \in \mathcal{X}} P(x)^2 = -\log \text{IE}(X)$, as well as the power entropy (Vajda, 1968; Vajda and Zvárová, 2007) $V_2(X) = 1 - \sum_{x \in \mathcal{X}} P(x)^2 = 1 - \text{IE}(X)$. There is also a concept, Onicescu's correlation coefficient (Calin and Udriște, 2014), $\rho(X, Y) = \frac{\sum_{x,y} P(x)P(y)}{\sqrt{\text{IE}(X) \cdot \text{IE}(Y)}}$, to quantify the dependency structure between multiple variables.

D Extended Context

Connection to the probabilistic Turing machines.

At some level of abstraction, one can view the LLM agents as probabilistic interactive Turing machines (PITMs) (Gill, 1977) that induce a distribution over action-observation transcripts. This connection may allow us to borrow some possible formal tools from the PITM literature to design UQ method for LLM agents. However, we note the following distinctions: (1) PITMs have an explicit randomness source and a defined distribution, whereas LLM agent also has a procedural uncertainty induced by the decoding time strategies; (2) while PITMs interact via fixed channels, LLM agents commonly interact with users and tools in partially observable or changing formats; and (3) the goal of LLM agents are often underspecified upfront and can be negotiated over turns depending on the situation, though PITMs have a concrete accept/output condition.

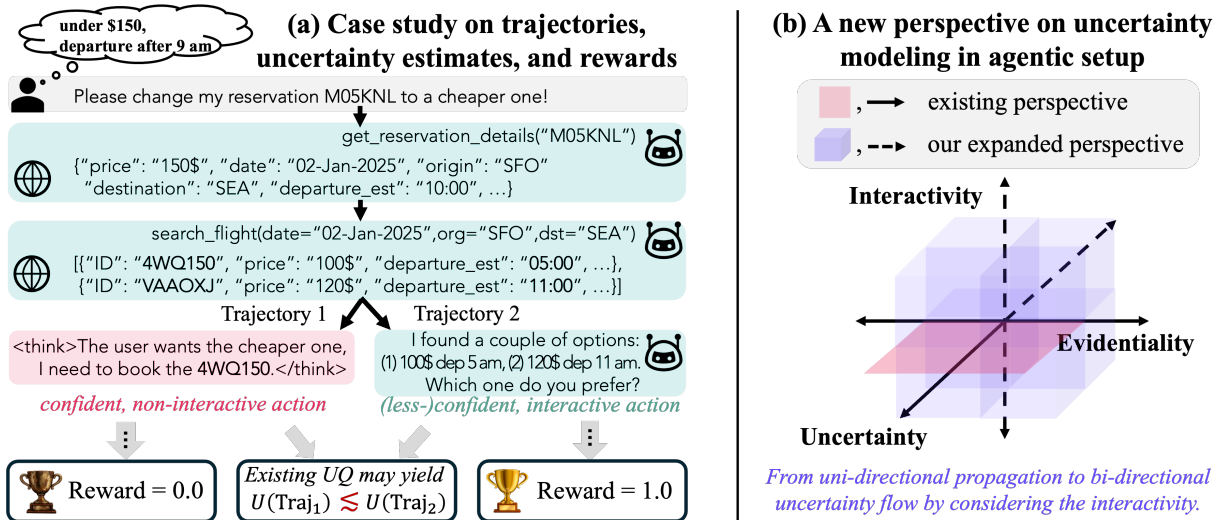


Figure 6: **Limitation of existing UQ (a) and our suggestion (b).** Prior works just concern the evidentiality when designing or evaluating UQ while neglecting *interactivity*, so they may fail to reliably capture the agent’s failure. We recommend expanding a dimension of interest, interactivity, and moving on to tackle the third challenge (Sec. 4.3) by allowing reducible uncertainty modeling of agents.

Connection to the partially observable Markov decision process. One might draw an analogy between the agent UQ and belief tracking in Partially Observable Markov Decision Processes (POMDPs) (Kaelbling et al., 1998). However, the agent theme challenges the standard setups in traditional POMDPs: (1) there is no explicit belief state, but the agent carries an implicit belief in its memory; (2) actions and observations comprise language and structured strings, spanning an effectively infinite space—far from the small discrete spaces in classical POMDP; and (3) it concerns more on the agent’s uncertainty, whereas POMDP concerns mostly on environment uncertainty. These connections highlight the distinctive edge of agent UQ, while grounding it in a well-established classical problem set.

E Action-Conditional Uncertainty Dynamics Model

“Every living means an increase in entropy of the part of the world, and it can only survive by continually drawing negative entropy from its environment.” – Schrödinger (1974)

Motivation. From a unified view discussed in Section 3.2, we see that most of the existing UQ methods model the multi-step uncertainty as a uni-directional propagation, i.e., every step-level uncertainty positively contributes to the total uncertainty, without distinguishing action types at each turn. One might claim that the average uncer-

tainty $\frac{1}{T} \sum_{t=1}^T U(\mathcal{F}_t | \mathcal{F}_{t-1})$ can already be fine, as it might produce a lower value for the chain of certain actions and vice versa. However, it does not consider the type of action which is critical in agentic setups, where an agent’s *interactivity* is key to earning a reward. See Fig. 6 (a), a scenario wherein a user does not fully specify their goal upfront. After searching for options, the agent either reasons itself or interacts with the user to delegate decisions. Although both trajectories are evidential, interaction-oriented ones tend to achieve higher rewards, despite reasoning-oriented ones often yielding greater confidence. Yet, existing methods can produce misleading uncertainty estimates, fail to meet the desiderata (Eq 2). From our pilot experiments (Figure 9 and Figure 10 provided in Appendix F.2), we observed that an action-independent naive aggregation strategy could not faithfully separate between the success and failure groups on average, resulting in full overall performance reported in Table 2, regardless of the types of uncertainty estimator.

Recommendation: conditional uncertainty reduction process. We therefore need a new perspective to model an agent’s uncertainty dynamics that accounts for both reductions and increases, explicitly depending on the agent’s interactivity at each step. While there are multiple ways to realize this concept, we propose one possible implementation as a prototype: *information gating*, which goes hand in hand with the information-theoretic mea-

sure of uncertainty noted in Sec. 3. Specifically, we derive a lower bound of total uncertainty $U(\mathcal{F}_{\leq T})$ that admits a selective reduction or increase in uncertainty throughout the trajectory:

$$\begin{aligned}
& U(E_0, O_0) + \sum_{i=1}^T U(\mathcal{F}_i | \mathcal{F}_{i-1}) \\
&= U(E_0, O_0) + \sum_{i=1}^T [U(A_i | E_{i-1}, O_{i-1}) + U(O_i | A_i, E_i)] \\
&= U(E_0, O_0) + \sum_{i=1}^T U(A_i | E_{i-1}, O_{i-1}) \left[1 + \frac{U(O_i | A_i, E_i)}{U(A_i | E_{i-1}, O_{i-1})} \right] \\
&\geq U(E_0, O_0) + \sum_{i=1}^T U(A_i | E_{i-1}, O_{i-1}) g(\mathcal{F}_i). \tag{7}
\end{aligned}$$

Here, we introduce a conditional gating function $g(\cdot)$ that reduces or increases a current turn uncertainty depending on the action. That is, for a set of valid uncertainty reduction actions, \mathcal{A}^- , we define:

$$g(\mathcal{F}_i) = \begin{cases} \frac{-\text{Info}(O_i; O_0 | E_i \setminus O_0)}{U(A_i | E_{i-1}, O_{i-1})} & \text{if } A_i \in \mathcal{A}^-, \\ 1 + \frac{U(O_i | A_i, E_i)}{U(A_i | E_{i-1}, O_{i-1})} & \text{otherwise.} \end{cases}$$

where $\text{Info}(\cdot; \cdot)$ denotes mutual information or pointwise mutual information that measures the *amount of information gain* by having the current observation O_i w.r.t. initial query O_0 given conversation history $E_i \setminus O_0$. For each turn, if the agent’s action is classified as an interactive (inviting a user or a tool) and evidential one (factual or not in conflict with stored data), then $g(\cdot)$ produces the amount of information gain with negative sign to realize uncertainty reduction; otherwise, it propagates uncertainty (See Fig. 7).

Importantly, the signed gating function handles uncertainty in a more interpretable manner than traditional estimates by letting not only quantitative but also **directional interpretations**. Besides, this uncertainty model allows us to derive the analytic extrema, i.e., maximum and minimum, of total uncertainty in closed forms as shown in Lemma E.1.

E.1 Implementation Sketch: Conditional Uncertainty Reduction Process

We discuss a possible implementation recipe for this conditional uncertainty reduction process to provide useful guidance for practitioners and researchers.

Action classifier. To realize the conditional information gate, we first need to classify actions to determine whether each action will contribute to total uncertainty reduction for that task or not. As emphasized in Section 3, considering the interactivity of actions is key to accounting for agent

uncertainty dynamics, as well as the evidentiality (or factuality) of the action that is solely considered for the existing approaches so far. Therefore, we may want to implement a compound classifier with both interactivity assessment and evidentiality verification. They can be implemented through a rule/syntactic-based, simple verifier (Mu et al., 2024), an LLM judge with specialized prompting, fine-tuning (Liu et al., 2025a), or a hybrid of rule-based and LLM-based (Peng et al., 2025). Below are the example prompts one can adopt in τ^2 -bench (Retail) for LLM-judge in interactivity and evidentiality classification.

Example System Prompt for Interactivity Classification

You are classifying customer service agent actions as either:

- "interactive": asking questions, requesting clarification, using tools to collect info
- "non_interactive": thinking, state-modifying actions, providing final answers

Respond as a JSON array of objects with keys: "action_idx" (int), "label" (str), "justification" (str).

Example System Prompt for Evidentiality Classification

You are evaluating whether a customer service agent’s actions are grounded in the conversation context and domain policy.

For each assistant action, determine:

- Are tool call arguments traceable to user-provided information or prior tool results?
- Are text claims about policies consistent with the provided policy document?
- Are stated facts (account details, prices, statuses) confirmed by tool results in the conversation?

Labels:

- "evidential": The action is well-grounded in the conversation and policy.
- "non_evidential": The action contains fabricated information, hallucinated tool arguments, or claims contradicting the policy.

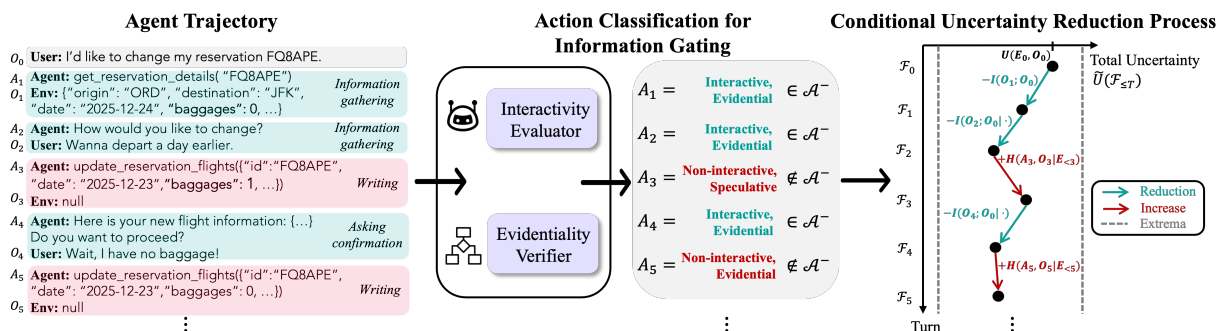


Figure 7: **Illustration on the proposed agent UQ paradigm.** We propose the conditional uncertainty reduction process for LLM agents by discerning interactive and evidential actions from others.

- "uncertain": Cannot determine grounding with available context.

Respond as a JSON array of objects with keys:

"action_idx" (int), "label" (str), "justification" (str), "issues" (list of str, empty if none).

Uncertainty estimation. In the proposed total uncertainty lower bound Eq. 7, there are three uncertainty terms, including initial query uncertainty, action uncertainty, and observation uncertainty. The *action uncertainty* can be estimated or approximated by leveraging existing techniques, such as sampling-based ones entropy (Malinin and Gales, 2021; Kuhn et al., 2023), pure probability-based methods (Fomicheva et al., 2020; Duan et al., 2024), hybrid (Vashurin et al., 2025), or verbalized confidence methods (Tian et al., 2023; Xiong et al., 2024; Yang et al., 2024a). However, as mentioned in Section 4.1, all these existing approaches have their own limitations, and one may have to choose an appropriate estimator depending on the situation.

Meanwhile, *observation uncertainty* and *initial query uncertainty* will be much harder to estimate, as we usually do not have full knowledge of the world (See Section 4.2). If the system allows multiple trials, one can utilize a sampling-consistency-based approach to estimate the observation uncertainty, likewise the traditional black-box UQ methods. If the system does not allow this, one may try to construct a datastore that contains some relevant data from a similar task and leverage non-parameter estimation methods (Ren et al., 2023; Kotelevskii et al., 2022). Another line of methods can be establishing a world model to approximate

the environment transition and estimate observation uncertainty from it (Chae et al., 2025; Zhou et al., 2025a; Hamidi et al., 2026). Figure 8 provides results on an auxiliary LLM-based approximation that mimics world modeling philosophy.

Mutual information estimation. The proposed information gating mechanism features information reduction based on the amount of information gain by measuring the conditional mutual information between the current observation O_i and the initial query O_0 , given all previous conversations except the initial query $E_i \setminus O_0$. While the estimation of mutual information is notorious for its difficulty, recent advancements on neural estimators (Belghazi et al., 2018; Cheng et al., 2020b; Mukherjee et al., 2020; Molavipour et al., 2020; Gritsai et al., 2025) shed light on this problem by enabling estimation in high-dimensional, unstructured data space. Besides, LLM-prompting-based mutual information estimation (Robertson and Koyejo, 2025) is also emerging as an attractive line of work, though it has yet to provide a solid theoretical foundation.

Conditional certainty maximization process. As an analogy to the conditional uncertainty reduction process presented before, we can also envision a "*certainty maximization*" approach by replacing the uncertainty with a certainty measure, such as informational energy (Pardo and Taneja, 1991), mentioned in Appendix C. Given that Onicescu's informational energy brings its correlation coefficient as well, we may be able to implement the exact same style information-gating-based method that has the opposite direction to uncertainty reduction dynamics.

E.2 Theoretical Analysis

Lemma E.1 (Extrema of Information Gating).

Let the lower bound of agent total uncertainty in Eq. 7 be $\tilde{U}(\mathcal{F}_{\leq T})$, denote $U(X) := H(X) = \mathbb{E}[-\log P(X)]$ and $\text{Info}(X; Y) := I(X; Y) = \mathbb{E}[\log \frac{P(X, Y)}{P(X)P(Y)}]$, then, we have:

$$\begin{aligned} \tilde{U}(\mathcal{F}_{\leq T}) &\geq H(E_0, O_0) - \sum_{i=1}^{T-1} I(O_i, O_0 | E_i \setminus O_0) \\ &\quad + H(A_T | E_{T-1}, O_{T-1}), \\ \tilde{U}(\mathcal{F}_{\leq T}) &\leq H(E_0, O_0) + \sum_{i=1}^T H(A_i, O_i | E_{i-1}, O_{i-1}). \end{aligned}$$

Proof. For i , if $A_i \in \mathcal{A}^-$ (resp. $A_i \notin \mathcal{A}^-$), turn-level uncertainty $U(\mathcal{F}_i)$ becomes $I(O_i, O_0 | E_i \setminus O_0)$ (resp. $U(\mathcal{F}_i) = U(A_i | E_{i-1}, O_{i-1}) + U(O_i | A_i, E_i) = H(A_i, O_i | E_{i-1}, O_{i-1})$). Thus, if all the intermediate actions are interactive and evidential, e.g., $A_i \in \mathcal{A}^-$ for all $1 \leq i \leq T-1$ (resp. $A_i \notin \mathcal{A}^-$ for all $1 \leq i \leq T-1$), our total uncertainty lower bound $\tilde{U}(\mathcal{F}_{\leq T})$ becomes a monotonic uncertainty reduction process (resp. monotonic propagation process), deriving the above two inequalities. Here, the final termination-turn action should not be interactive, and the observation in this turn is deterministic in our considered setups (Fig. 2), resulting in the accumulation of last action uncertainty $H(A_T | E_{T-1}, O_{T-1})$. \square

E.3 Action Categorization

An actual example categorization of the action interactivity in an airline assistant task is provided in Table 3. Besides, a higher abstraction of action category to two more scenarios, retail in τ^2 -bench (Yao et al., 2024; Barres et al., 2025) and ToolSandbox (Lu et al., 2025) is provided in Table 4.

F Experiment

To ground our statements with empirical results, we conducted a small-scale pilot experiment on a real-world agent benchmark in general assistance tasks to evaluate existing UQ techniques on LLM agents.

F.1 Setup

Dataset and task. τ^2 -bench (Barres et al., 2025) (the latest version of τ -bench (Yao et al., 2024)) is a representative agent benchmark adopted in frontier LLM evaluation (Google, 2026), consisting of three different domains – airline, retail, and telecom, where the total number of sample tasks is 50, 114, and 114, respectively. By following (Google,

2026), we experimented with retail and telecom domains. The tasks in the Retail domain are about helping users cancel or modify pending orders, return or exchange delivered orders, modify user addresses, or provide information. The tasks in the telecom domain are about tackling telecommunication issues in mobile devices, organized into roughly three categories: service issues, mobile data issues, and MMS issues.

LLMs and inference details. We experiment with two modern LLMs: GPT-4.1 (OpenAI, 2025) and Kimi-K2.5 (Kimi Team, 2026) through Microsoft Azure⁴ AI Foundry API endpoints. Most modern proprietary LLMs do not provide the output log probability in multiple inference engine consistency, and GPT-4.1 was a viable choice for us to adopt here. Meanwhile, small-scale open-source models that can be run in local GPUs did not achieve meaningful performance on τ^2 -bench due to its task complexity, so we stick with a state-of-the-art large-scale open-source model, Kimi-K2.5, through third party platform. We set the temperature to 0.0 during inference, which is the default value of τ^2 -bench. Due to the resource constraint, we only run a single trial per task in contrast to the four trials conducted in τ^2 -bench. User simulator LLM was fixed to Kimi-K2.5 for all tasks, and we only varied agent LLM between GPT-4.1 and Kimi-K2.5.

Implementation of uncertainty estimators. We adopt three different uncertainty/confidence estimators: negative log-likelihood (NLL), Entropy, and verbalized confidence. The NLL and Entropy are calculated per token and averaged across the entire sequence of actions in the trajectory, whereas the verbalized confidence is calculated per turn and also averaged across the entire trajectory. To be specific,

$$\begin{aligned} \hat{U}_{\text{NLL}}(\mathcal{F}_{\leq T}) &= \frac{1}{T} \sum_{i=1}^T \hat{U}_{\text{NLL}}(A_i | E_{i-1}, O_{i-1}) = \\ &= \frac{1}{T} \sum_{i=1}^T \frac{1}{L_i} \sum_{j=1}^{L_i} -\log P_{\pi, \tau}^1(A_{i,j} | E_{i-1}, O_{i-1}) \\ \hat{U}_{\text{Ent}}(\mathcal{F}_{\leq T}) &= \frac{1}{T} \sum_{i=1}^T \hat{U}_{\text{Ent}}(A_i | E_{i-1}, O_{i-1}) = \\ &= \frac{1}{T} \sum_{i=1}^T \frac{1}{L_i} \sum_{j=1}^{L_i} \mathbb{E}_{A_{i,j}} [-\log P_{\pi, \tau}^5(A_{i,j} | E_{i-1}, O_{i-1})] \\ \hat{C}_{\text{verb}}(\mathcal{F}_{\leq T}) &= \\ &= \frac{1}{T} \sum_{i=1}^T \hat{C}_{\text{verb}}(A_i | E_{i-1}, O_{i-1}, \text{vc-prompt}) \end{aligned}$$

where L_i denotes the sequence length of i -th turn, P^K indicates top-K output token distribution (a chosen token probability if $K = 1$), and

⁴<https://azure.microsoft.com/en-us>

Action Category	Interactivity	Example
Information-gathering	interactive	Agent uses a read tool (e.g., <code>get_reservation_details</code> or <code>search_flights</code>) to retrieve data or asks the user for missing information, e.g., “Could you provide your reservation number and last name?”.
Asking clarification or confirmation to user	interactive	Agent asks the user preferences or confirmation on decision, such as, “Do you want me to proceed with booking FQ8APE?”.
Thinking	non-interactive	Agent plans the future action sequence based on the previous trajectory.
State-changing tool call (writing)	non-interactive	Agent calls the tools that modify the database (e.g., <code>cancel_reservation</code> , <code>book_reservation</code> , <code>update_reservation_flights</code>) commit to an outcome.
Providing final information to user	non-interactive	Agent reports the result of an action, such as “Your reservation has been cancelled; your refund will be processed” or “Your flight has been rebooked to SFO departing at 8 a.m.”.

Table 3: **Example action classes in Airline booking scenario in τ^2 -bench.** A number of the agent’s actions can be categorized into five classes, which may reduce or increase the total uncertainty. See Table 4 for extended analyses.

Action Category	Interactivity	Scenario-specific Example
Information-gathering	interactive	Search flight information (τ^2 -bench Airline); retrieve order status (τ^2 -bench Retail); read message box (ToolSandbox)
Asking clarification or confirmation to user	interactive	Ask for flight choice or request final booking decision (τ^2 -bench Airline); instruct or ask the user to choose cancellation (τ^2 -bench Retail); ask for providing contact information or request updating contact number (ToolSandbox)
Thinking	non-interactive	Plan flight schedule (τ^2 -bench Airline); consider returning items (τ^2 -bench Retail); think the reason of failed temperature checking (ToolSandbox)
State-changing tool call (write)	non-interactive	Update flight booking (τ^2 -bench Airline); return/cancel items (τ^2 -bench Retail); Turn on Wifi (ToolSandbox)
Providing final information to user	non-interactive	Summarize updated flights (τ^2 -bench Airline); summarize returned items (τ^2 -bench Retail); summarize phonebook update results (ToolSandbox)

Table 4: **Example action classes in three scenarios from two benchmarks, τ^2 -bench (Yao et al., 2024; Barres et al., 2025) and ToolSandbox (Lu et al., 2025).**

vc-prompt is a prompt for verbalized confidence we designed as below,

System Prompt for Verbalized Confidence (Tian et al., 2023)

Provide the probability (0.0 to 1.0) that your current response/action is correct for the given conversation history and context so far. Give ONLY the probability, no other words or explanation.

Evaluation details. Since τ^2 -bench does not provide turn-level annotation, we conduct trajectory-level evaluation that compares the trajectory-level uncertainty $U(\mathcal{F}_{\leq T})$ with the final task failure $1-r$ (or success r in the case of verbalized confidence). For this, we measure the area under the receiver operating characteristic curve (AUROC), Spearman

rank correlation coefficient ρ , and Kendall rank correlation coefficient τ . The average reward (success rate) per model-domain pair is also reported in Table 2.

F.2 Extended Results

This subsection provides full analysis results on the observation uncertainty in Figure 8 and uncertainty evolution in Figure 9 and Figure 10. In Figure 8, we compare three different observation uncertainties over all user messages across trajectories: (1) the GT user simulator uncertainty (computed from the output probability of user simulator LLM), (2) estimates driven from the output probability of agent LLM by feeding user messages along with all previous context, and (3) estimates driven from the output probability of an auxiliary LLM by feeding user messages along with all previous context,

coupled with a prompt as below.

System Prompt for Observation Distribution Approximator

You are an impartial observer monitoring a customer-service conversation between an AI agent and a human user. Your task is to predict the next observation that the agent will receive:

1. A user message - the customer's natural-language reply, or
2. A tool result - the deterministic output returned by a backend API after the agent issues a tool call.

Given the full conversation history up to this point, generate the most likely continuation as if you were the user or the tool backend. Aim to model the distribution of plausible observations as faithfully as possible: assign high probability to expected responses and low probability to surprising ones.

We observe remarkable deviation between the GT user simulator's uncertainties and the estimates from the agent LLM in general. However, using an auxiliary LLM as a world model somehow reduces the gap in many cases. Cost-effective auxiliary model-based approximation can be worth exploring.

Next, Figure 9 and 10 compare uncertainty evolutions over turns between success and failure groups of tasks. To compare varying-length trajectories across different tasks, we normalize the turn index in X-axis into a zero-to-one scale. Results show that existing naive uncertainty aggregation methods can not discern the failure group from the success group to a statistically meaningful degree (the shaded areas in the plots denote a single standard deviation confidence interval), both in the middle of and at the end of the trajectory.

G Mini-Survey on Agent Benchmark

We did a small-scale survey on LLM-based agent benchmarks released to the public from February 2023 to February 2026, where we sourced the papers mostly from arxiv.org and also referred to ML/NLP conference proceedings. We ended up with a non-exhaustive collection of 44 papers, mainly designed for pure LLM agents, while a

few of them also support multimodal scenarios of VLM agents. We confine the scope of the survey to the agent's problem-solving capability rather than other desired properties such as safety and trustworthiness (Zhang et al., 2024; Levy et al., 2024; Andriushchenko et al., 2025). We categorize the papers into three classes depending on the evaluation granularity: turn-level, milestone-level, and trajectory-level. Table 5 summarizes the turn-level and milestone-level evaluation benchmarks, and Table 6 summarizes the trajectory-level ones. The scarcity of fine-grained benchmarks is severe and becomes a real bottleneck to establishing and evaluating the agent UQ framework.

H Discussion and Future Work

Does solely establishing the agent UQ framework actually help us to reliably infer on the agent's performance on a task? Modern LLMs (especially after post-training) are not well-calibrated (Tian et al., 2023), so the uncertainty estimates from these ill-calibrated probabilistic models can not be directly used as a reliable performance indicator. Although we present a high-level roadmap for quantifying uncertainty for LLM agent systems, future work should also explore the calibration of LLM agents as well. The connection between calibration and accuracy (Park et al., 2020; Oh et al., 2024) of a predictor shows an exciting future work direction for joint optimization of the agent's problem-solving capability and calibration simultaneously.

Another promising direction for future investigations would be to extend the agent UQ framework to multimodal setups such as graphical user interface agents (Nguyen et al., 2025), where extra challenges occur, such as many-to-many correspondence between modalities (Chun, 2025). Defining and establishing a reliable UQ framework for multimodal agents would offer helpful foundations to develop robustness solutions (Oh et al., 2025a,b), safety (Zhou et al., 2025b; Chen et al., 2025a) and more (Shi et al., 2025).

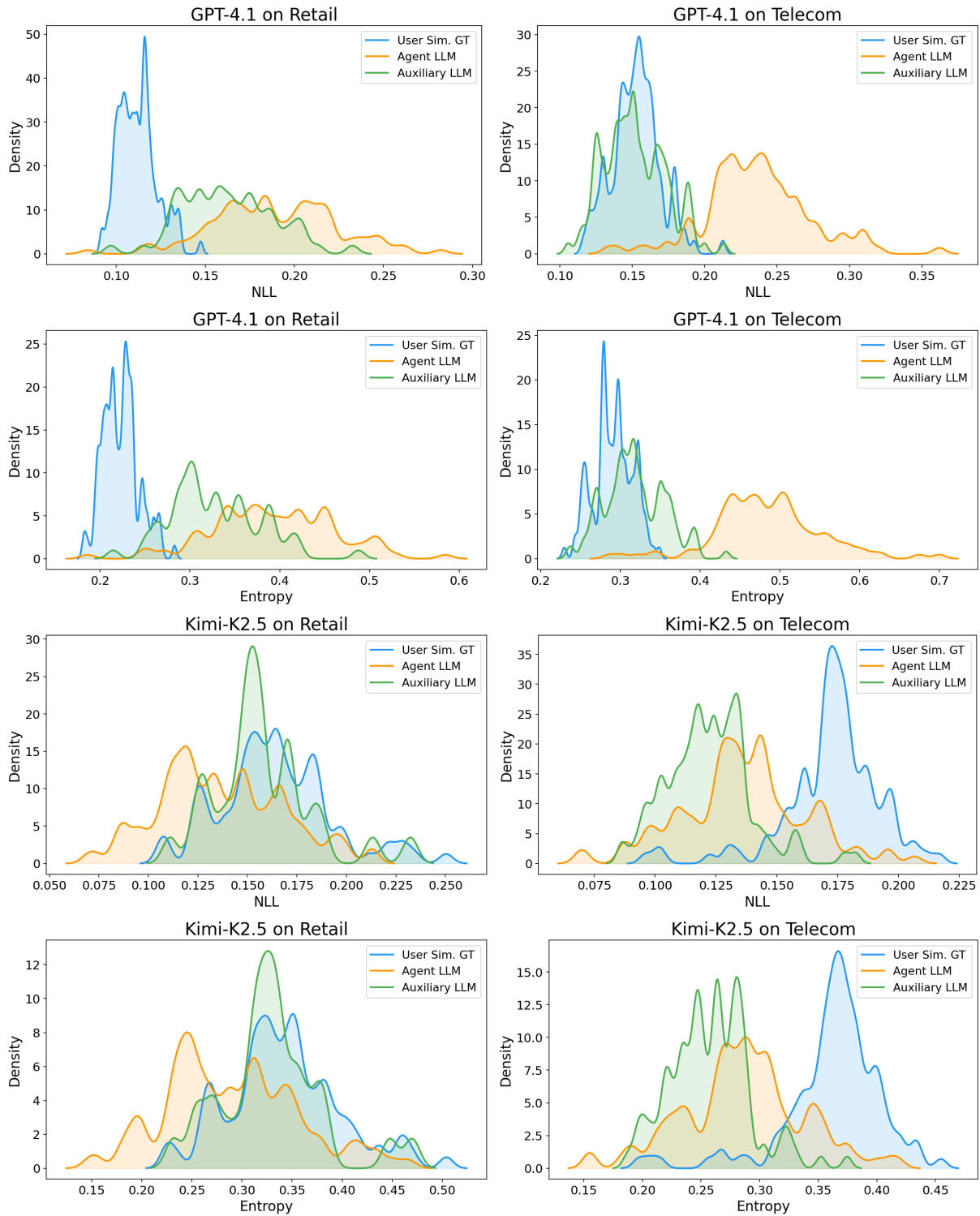


Figure 8: **Comparison of observation uncertainty estimates on τ^2 -bench.** We compare distributions over observation uncertainty estimates, i.e., average NLL and Entropy over all text message tokens from the user. Distributions of the ground truth user simulator LLM and that of the agent-approximated one show remarkable deviation, whereas auxiliary LLM prompting-based estimation shows somewhat close distribution with the ground truth.

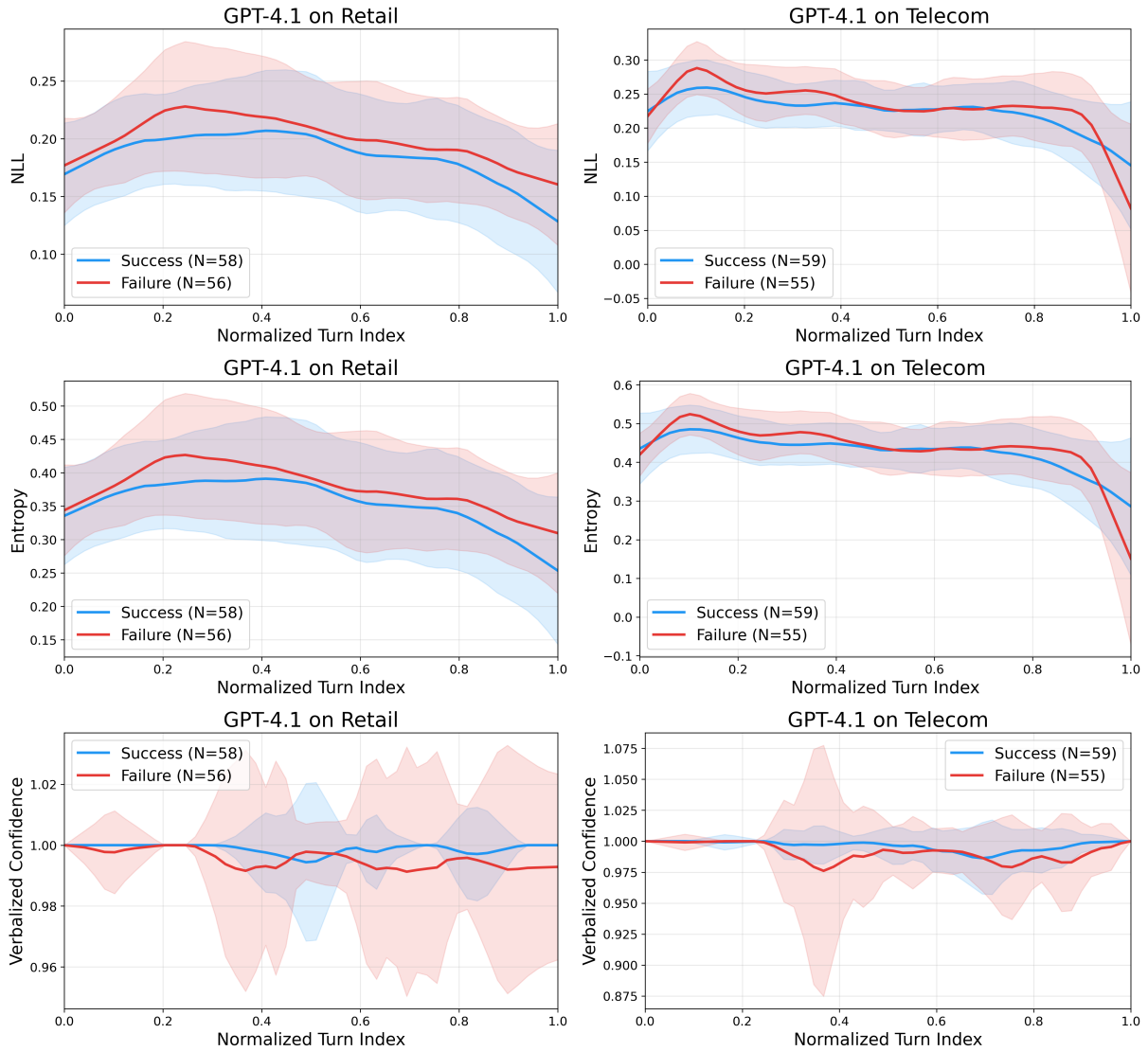


Figure 9: **Uncertainty evolution of GPT-4.1 across the trajectory on τ^2 -bench.** We visualize the evolution of action uncertainty/confidence estimates over the normalized turn index, averaged over the entire trajectories from two groups: task success and failure. The shade area denotes one standard deviation.

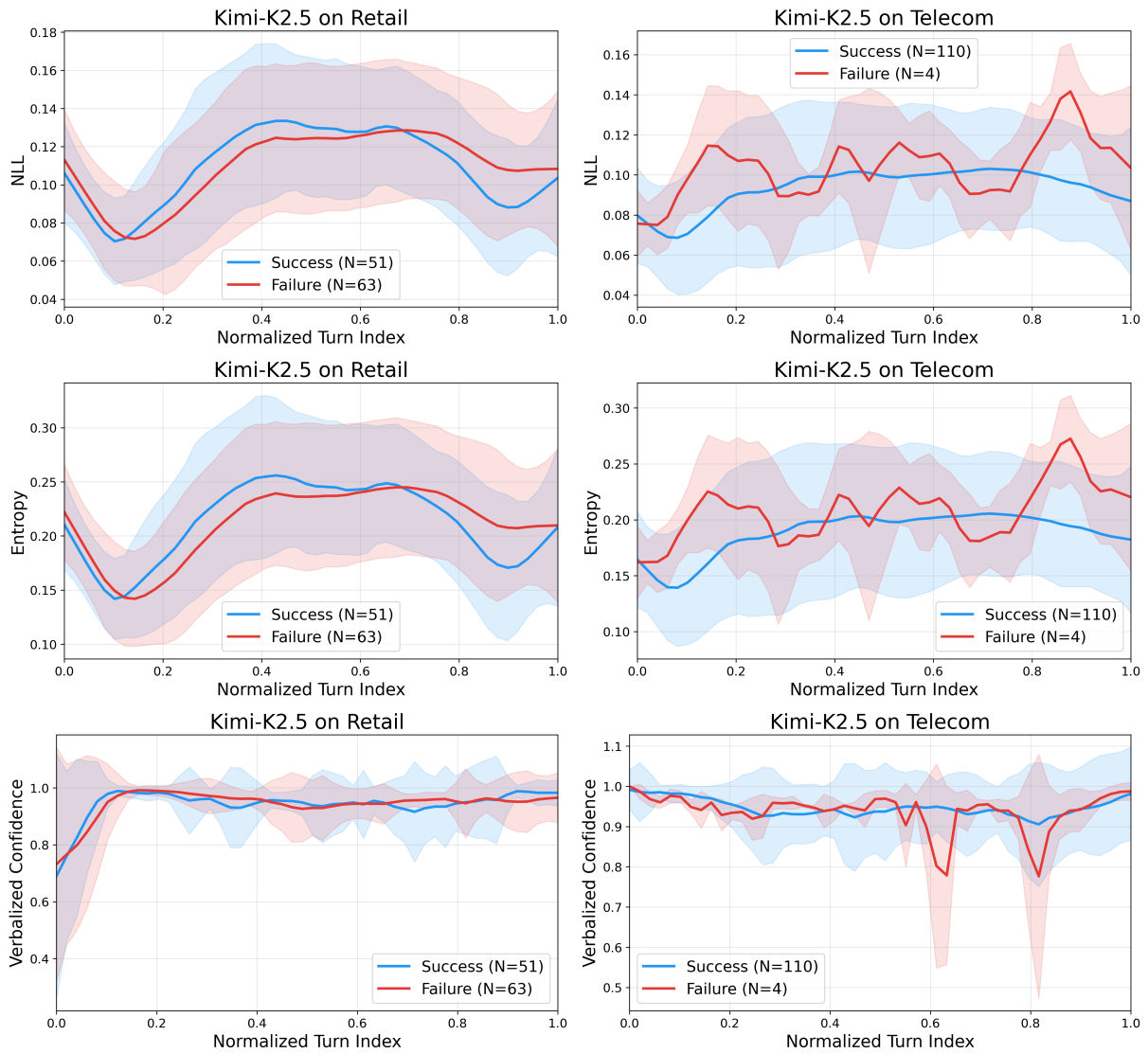


Figure 10: **Uncertainty evolution of Kimi-K2.5 across the trajectory on τ^2 -bench.** We visualize the evolution of action uncertainty/confidence estimates over the normalized turn index, averaged over the entire trajectories from two groups: task success and failure. The shade area denotes one standard deviation.

Title	Citation	Evaluation
Agent-as-a-Judge: Evaluate Agents with Agents	(Zhuge et al., 2025)	Turn-level
API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs	(Li et al., 2023)	Turn-level
Clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents	(Chalamalasetti et al., 2023)	Turn-level
Mind2Web: Towards a Generalist Agent for the Web	(Deng et al., 2023)	Turn-level
AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents	(Chang et al., 2024)	Milestone-level
DevBench: A Comprehensive Benchmark for Software Development	(Li et al., 2025a)	Milestone-level
Gaia2: Benchmarking LLM Agents on Dynamic and Asynchronous Environments	(Froger et al., 2026)	Milestone-level
PaperBench: Evaluating AI’s Ability to Replicate AI Research	(Starace et al., 2025)	Milestone-level
ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery	(Chen et al., 2025b)	Milestone-level
ST-WebAgentBench: A Benchmark for Evaluating Safety and Trustworthiness in Web Agents	(Levy et al., 2024)	Milestone-level
The Berkeley Function Calling Leaderboard (BFCL): From Tool Use to Agentic Evaluation of Large Language Models	(Patil et al., 2025)	Milestone-level
TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks	(Xu et al., 2025a)	Milestone-level
ToolSandBox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities	(Lu et al., 2025)	Milestone-level
WebCanvas: Benchmarking Web Agents in Online Environments	(Pan et al., 2024)	Milestone-level

Table 5: Agent benchmark research with turn-level and milestone-level evaluation protocols.

Title	Citation
AgentBench: Evaluating LLMs as Agents	(Liu et al., 2024)
AgentGym: Evolving Large Language Model-based Agents across Diverse Environments	(Xi et al., 2025)
AppWorld: A Controllable World of Apps and People for Benchmarking Interactive Coding Agents	(Trivedi et al., 2024)
AssistantBench: Can Web Agents Solve Realistic and Time-Consuming Tasks?	(Yoran et al., 2024)
BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents	(Wei et al., 2025)
CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark	(Siegel et al., 2024)
Finance Agent Benchmark: Benchmarking LLMs on Real-world Financial Research Tasks	(Bigard et al., 2025)
GAIA: A Benchmark for General AI Assistants	(Mialon et al., 2023)
GameBench: Evaluating Strategic Reasoning Abilities of LLM Agents	(Costarelli et al., 2024)
InterCode: Standardizing and Benchmarking Interactive Coding with Execution Feedback	(Yang et al., 2023)
MCPAgentBench: A Real-world Task Benchmark for Evaluating LLM Agent MCP Tool Use	(Liu et al., 2025b)
MCP-Atlas: A Large-Scale Benchmark for Tool-Use Competency with Real MCP Servers	(Bandi et al., 2026)
MedAgentBench: A Realistic Virtual EHR Environment to Benchmark Medical LLM Agents	(Jiang et al., 2025b)
MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback	(Wang et al., 2024c)
MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation	(Huang et al., 2024)
OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments	(Xie et al., 2024b)
SafeAgentBench: A Benchmark for Safe Task Planning of Embodied LLM Agents	(Yin et al., 2024)
SmartPlay: A Benchmark for LLMs as Intelligent Agents	(Wu et al., 2024b)
SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents	(Zhou et al., 2024b)
SWE-bench: Can Language Models Resolve Real-World GitHub Issues?	(Jimenez et al., 2024)
Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces	(Merrill et al., 2026)
The Tool Decathlon: Benchmarking Language Agents for Diverse, Realistic, and Long-Horizon Task Execution	(Li et al., 2025b)
ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs	(Qin et al., 2024)
TravelPlanner: A Benchmark for Real-World Planning with Language Agents	(Xie et al., 2024a)
Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents	(Backlund and Petersson, 2025)
WebArena: A Realistic Web Environment for Building Autonomous Agents	(Zhou et al., 2024a)
WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents	(Yao et al., 2022a)
WorkArena: How Capable Are Web Agents at Solving Common Knowledge Work Tasks?	(Drouin et al., 2024)
τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains	(Yao et al., 2024)
τ^2 -Bench: Evaluating Conversational Agents in a Dual-Control Environment	(Barres et al., 2025)

Table 6: Agent benchmark research with a trajectory-level evaluation protocol.