Last Step Matters: Early Uncertainty Cannot Predict Failure in Long-Horizon Agents

Anonymous ACL submission

Abstract

Early prediction of whether a long-horizon LLM agent will ultimately fail could save substantial inference and tool cost by enabling timely intervention. Uncertainty quantification is a promising approach for agent failure detection, yet whether these signals remain discriminative in the early-detection setting has not been explored. We evaluate mainstream uncertainty metrics on long-horizon agent tasks and find that at trajectory completion, verbal confidence outperforms all other metrics (AUROC 0.86 at trajectory completion), but at 50% progress all metrics fall below 0.60, barely above chance. We identify an underlying mechanism explaining this gap: *path switching*, where agents frequently abandon their current approach mid-trajectory, breaking the link between early signal and final outcome. Consequently, a confidence surge occurs only at the tail—correct trajectories rise +15 points above the first-70% progress average while incorrect ones drop −6. These findings challenge the assumption that intermediate uncertainty can guide early intervention, and suggest that agent harnesses should rely on final-step confidence for decision-making: when final-step confidence is low, restarting the task with awareness of prior attempts is more effective than mid-trajectory intervention.

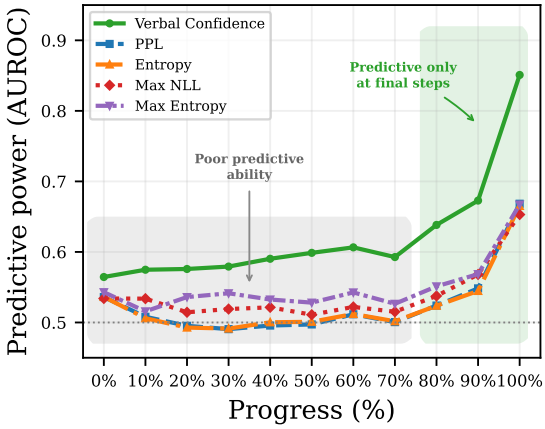

Figure 1: Predictive power of uncertainty metrics for agent failure (measured by AUROC) across normalized trajectory progress. All metrics stay weak until 80–90% progress, then rise sharply at completion (trajectories with ≥11 steps; AUROC averaged across all 5 models and 3 datasets). This pattern is consistent across all combinations (Appendix C).

1 Introduction

LLM-based agents can now autonomously complete complex, long-horizon tasks that require iterative search (Yao et al., 2022), reasoning (Schick et al., 2023), and answer synthesis (Deng et al., 2023). However, these agents remain prone to execution failures: they may get stuck in search loops (Wang et al., 2026), pursue incorrect reasoning paths (Zhu et al., 2025), or hallucinate information (Huang et al., 2025a). Such failures are costly: each step consumes inference compute and tool calls, and long-horizon trajectories can accumulate orders of magnitude more cost than single-turn queries (Xiao et al., 2025; Wang et al., 2025). Early failure detection could enable timely intervention, saving substantial inference and tool cost, yet Figure 2 illustrates the core challenge: a decline in confidence mid-trajectory does not reliably indicate eventual failure. Uncertainty estimation, including verbalized confidence (Tian et al., 2023; Kadavath et al., 2022), perplexity (Kadavath et al., 2022), entropy-based measures (Farquhar et al., 2024), and consistency-based methods (Farquhar et al., 2024; Manakul et al., 2023), has emerged as a promising approach for detecting such failures. Prior work has shown these metrics to be discriminative in single-turn settings and, more recently, in short-horizon agentic tasks (1–6 steps) (Zhang et al., 2026b; Zhao et al., 2025; Duan et al., 2025; Oh et al.). However, whether these metrics retain their utility in *long-horizon*

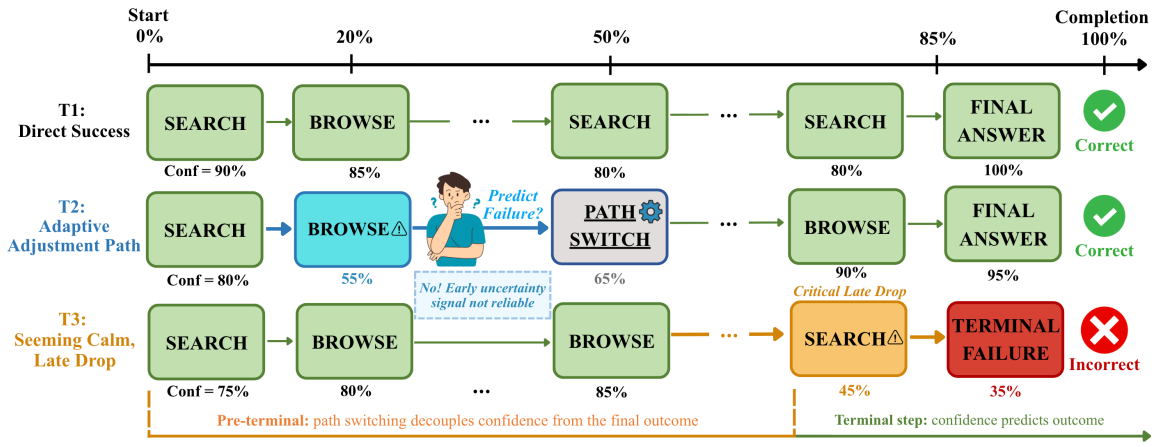

Figure 2: Agent system execution diagram contrasting three typical trajectories (T1–T3) to demonstrate the decoupling of pre-terminal confidence from the final outcome. The trajectories illustrate that early confidence does not predict the final outcome: an early low-confidence signal (T2) does not indicate eventual failure when the agent subsequently switches to a new approach, and an early high-confidence signal (T3) does not guarantee eventual success.

trajectories remains unexplored, and no work has systematically studied which steps along the trajectory carry discriminative signal for failure prediction.

We study the dynamics and discriminative power of mainstream uncertainty metrics across three families, namely verbal confidence, token-probability, and consistency-based, on three long-horizon agent benchmarks: BrowseComp (Wei et al., 2025), BrowseComp-zh (Zhou et al., 2025), and Humanity's Last Exam (Phan et al., 2025). We evaluate over 5 models from 27B to 1T parameters, focusing on two questions: **(1) Effectiveness.** Do established uncertainty metrics remain discriminative in long-horizon agent trajectories? Can they reliably distinguish successful from failed executions? **(2) Timeliness.** More critically, *when* do these signals become informative? Can failures be identified at the early or middle stages of the trajectory, before the agent reaches its conclusion?

Our answer to the first question is encouraging: last-step verbal confidence is a strong baseline, achieving an average AUROC (Area Under the ROC Curve) of 0.86 across all evaluation settings, where 1.0 means perfect prediction and 0.5 means random. The answer to the second question is sobering: even this strong metric suffers from **early predictive impotence**: at 50% trajectory progress, AUROC is only 0.58. As Figure 1 shows, all metrics remain near chance until 80–90% progress, then rise sharply only upon completion.

We identify **path switching** as the micro-mechanism behind this gap. Agents frequently abandon their current reasoning direction mid-trajectory: each switch decouples prior uncertainty from the final outcome, so only post-switch signal predicts success or failure. This produces a sharp confidence surge relative to the early-trajectory baseline (+15 points for correct trajectories vs. −6 points for incorrect ones) and explains why intermediate uncertainty cannot guide early intervention (Zhang et al., 2026a; Ou et al., 2025; Yan et al., 2025). Agent harnesses should therefore rely on final-step confidence: when final-step confidence is low, restarting with awareness of prior attempts outperforms mid-trajectory intervention.

These findings are robust across settings: 5 models spanning 27B to 1T parameters, 3 long-horizon deep research datasets, alternative confidence elicitation prompts, and tasks beyond deep research.

Our contributions are: **(1)** The first **systematic study** of uncertainty dynamics in long-horizon LLM agent trajectories, spanning metrics across three families, establishing a clear distinction between failure *detection* (at completion) and failure *prediction* (before completion). **(2) Path switching** as the mechanism underlying this gap, supported by both aggregate analysis and per-trajectory case studies. **(3) Practical implications for agent harness**: use final-step verbal confidence for decision-making; when final-step confidence is low, restart with prior-answer awareness (e.g. BrowseConf-Neg).

2 Theoretical Background

Following Oh et al., we frame uncertainty in LLM agents at two levels. At the step level, the uncertainty of the agent's action distribution is defined as:

$$\mathrm{SLU}_t(x) = U[P(a_t \mid e_{t-1}, o_{t-1})] \quad (1)$$

where $U(\cdot)$ is a generic uncertainty measure (e.g., entropy $H[P]$, perplexity $\exp H[P]$, or $100-\mathrm{conf}_t$ for verbal confidence). At the trajectory level, step-level values are aggregated into a scalar:

$$\mathrm{TLU}(\tau) = \mathcal{A}[\mathrm{SLU}_0(x), \ldots, \mathrm{SLU}_{T-1}(x)] \quad (2)$$

where $\mathcal{A}$ may be the mean, last-step value, tail-weighted mean, or other aggregation operators.

3 Experimental Setup

3.1 Datasets and Models

We evaluate on 3 benchmark datasets (Table 1).

BrowseComp (Wei et al., 2025) is a web-browsing question-answering benchmark comprising 1,266 questions that require persistently navigating the internet to find hard-to-find, entangled information. **BrowseComp-zh** (Zhou et al., 2025) is its Chinese-language counterpart. **HLE** (Phan et al., 2025) is a multi-modal benchmark of 2,500 expert-level questions spanning mathematics, humanities, and natural sciences. Following Li et al. (2025), we use the full BrowseComp-zh set (289 questions) and randomly sampled subsets for BrowseComp (200 questions); for HLE, we use only the text-only subset (157 questions).

We evaluate open-weight models spanning 27B to 1T parameters (Table 2); we exclude closed-weight models since token-level probabilities are unavailable, precluding metrics such as entropy.

Correctness is determined by comparing the agent's final answer to the ground truth using an LLM-as-a-Judge evaluator (Gu et al., 2024).

3.2 Agentic Setting

We follow the agentic setup of Ou et al. (2025). The agent is equipped with two tools: (1) **Search**, which queries the Google search engine and returns the top-10 results per query (each with a title, snippet, and URL), and (2) **Visit**, which accesses designated webpages by first extracting the full page content via Jina[1] and then summarizing the goal-relevant information. No memory compression, context folding, or other auxiliary tools are used. All models are configured with a context length of 128k tokens and a maximum of 80 tool-call steps per trajectory.

[1] https://jina.ai/

Dataset	# Tasks	Acc.	Avg Steps	Max Steps
BrowseComp	200	32.0%	31.6	80
BrowseComp-zh	289	49.8%	20.7	80
HLE	157	34.1%	13.5	80

Table 1: Dataset statistics.

Model	Total	Activated	Arch.
Qwen3.5-27B	27B	27B	Dense
Qwen3.5-122B-A10B	122B	10B	MoE
GLM-4.7	355B	32B	MoE
DeepSeek-V3.2	671B	37B	MoE
Kimi-2.6	1T	32B	MoE

Table 2: Evaluated models, spanning 27B to 1T total parameters. All are open-weight models, enabling computation of uncertainty metrics based on token-level probabilities over the full trajectory.

3.3 Uncertainty Metrics

We evaluate the following families of uncertainty metrics:

Verbal confidence. Following the practice of prompting models to verbally state their uncertainty (Tian et al., 2023; Wei et al., 2024), we extend this to the agentic setting by requiring the agent to output a confidence score (0–100) alongside each function call, producing a per-step verbal confidence sequence over the trajectory.

Token-probability metrics. We compute four token-probability-based metrics from per-step logprobs over the full assistant turn: perplexity (PPL) (Malinin and Gales, 2021), max token negative log-likelihood (Max NLL) (Manakul et al., 2023), Shannon entropy (Malinin and Gales, 2021), and max token entropy (Manakul et al., 2023).

Metric aggregation. All metrics above are computed per step, yielding a sequence over the trajectory. We aggregate each sequence into a trajectory-level scalar using six operators: `mean`, running `max`/`min`, `last step`, and tail-weighted means of the last 3 or 5 steps, motivated by Zhao et al. (2025), who proposed that tail steps contain more information. Our results in §5.2 confirm that final-step signal dominates. Applied across 5 base metrics (4 token-level + verbal confidence) $\times$ 6 operators, this yields 30 scalar features.

		Qwen3.5-27B			Qwen3.5-122B-A10B			GLM-4.7			DeepSeek-V3.2			Kimi-2.6		
Family	**Best agg.**	**BC**	**BC-zh**	**HLE**	**BC**	**BC-zh**	**HLE**	**BC**	**BC-zh**	**HLE**	**BC**	**BC-zh**	**HLE**	**BC**	**BC-zh**	**HLE**
Single-trajectory																
Verbal Confidence (Tian et al., 2023)	last step	**.915**	**.854**	**.744**	**.904**	**.864**	**.763**	**.906**	**.860**	**.862**	**.921**	**.867**	**.801**	**.794**	**.816**	**.790**
PPL (Malinin and Gales, 2021)	last step	.816	.790	.692	.853	.752	.601	.745	.492	.481	.842	.692	.580	.575	.534	.547
Entropy (Manakul et al., 2023)	max	.692	.748	.695	.736	.705	.695	.750	.575	.469	.702	.689	.667	.607	.548	.669
Max NLL (Manakul et al., 2023)	max	.720	.763	.673	.700	.710	.682	.757	.631	.565	.692	.672	.681	.665	.622	.705
Max token entropy (Manakul et al., 2023)	max	.741	.733	.690	.714	.708	.679	.701	.613	.618	.717	.668	.706	.761	.742	.736

Table 3: Best AUROC per single-trajectory metric family across 15 model–dataset combinations. The aggregation operator with the most per-dataset wins is shown for each family. **Bold**: best in column; underlined: second best. Full results in Appendix B.

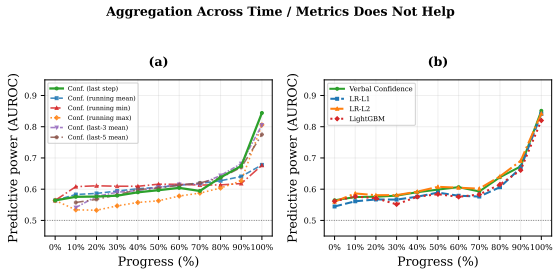

Figure 3: Aggregation across time and metrics does not help. (a) Verbal confidence with six temporal aggregation methods; running-min improves early-stage AUROC slightly but remains near chance (∼0.6). (b) 5-fold CV classifiers using five last-step features vs. verbal confidence; multi-feature combination cannot extract signal absent from the single-metric baseline.

Consistency-based metrics. Consistency-based methods (Xia et al., 2025) require multiple rollouts per question (8 in our setup) and quantify uncertainty from the variation across answers or trajectories. This makes them distinct from the trajectory-level metrics above, which operate within a single trajectory. Although these metrics do not serve our goal of early detection—they require $N\times$ compute for N rollouts and cannot be computed from partial trajectories—we include 10 such metrics (diversity, semantic entropy, predictive entropy, etc.) in Table 4 for completeness, with definitions and citations in the table.

3.4 Evaluation Protocol

AUROC. Area Under the ROC Curve, computed within each model–dataset combination and then averaged. Higher values indicate stronger predictive power, and 0.5 corresponds to chance.

Backward truncation analysis. Since trajectories vary in length, we normalize each trajectory to 11 equally-spaced progress points (0%, 10%, 20%, . . . , 100%) and compute AUROC at each point using only the trajectory prefix up to that point, as if the trajectory had been truncated there. This reveals how well uncertainty metrics can predict the final outcome at each stage of completion.

4 Discriminative Power of Uncertainty Metrics

4.1 Full-Trajectory Failure Detection

Returning to the first question, *do uncertainty metrics remain discriminative in long-horizon trajectories?*, our answer is **yes**: verbal confidence with last-step aggregation achieves 0.77–0.93 AUROC for failure detection across all settings (Table 3, full table in Appendix B).

This advantage reflects two properties. First, *uncertainty awareness*: the model's self-assessed confidence carries more signal than derived statistics from token-level distributions or answer agreement, indicating that pretrained LLMs have developed metacognitive capacity beyond what token probabilities capture. Second, *temporal aggregation*: the transformer's attention mechanism naturally compresses the full trajectory into the final hidden state, making the last-step value a trajectory-wide summary that outperforms hand-designed aggregators such as `mean` or `max`, which dilute the concentrated final-step signal with noisy intermediate values.

Among consistency metrics (Table 4), answer-frequency and graph-based metrics (diversity, degree uncertainty) tend to outperform log-probability-based ones (predictive entropy, semantic entropy). On harder datasets (HLE), the best consistency metric (discrete semantic entropy, .809) even surpasses verbal confidence (.791), while on easier benchmarks verbal confidence retains a comfortable margin.

4.2 Is Early Detection Possible?

Turning to the second question, *can failures be identified before the agent reaches its conclusion?*, our answer is **no**: even the best metric (verbal confidence) reaches only 0.60 AUROC at 50% tra-

jectory progress and 0.67 at 90%, barely above chance despite achieving 0.85 at completion. We measure this using the backward truncation protocol (§3.4).

Figure 1 shows the result across five metric families: AUROC remains weak (below 0.65) through 80% progress for all metrics, then rises sharply at 100%, with verbal confidence consistently outperforming all token-probability metrics. This pattern is not an artifact of absolute step position, as it holds regardless of trajectory length, confirming that discriminative information is concentrated at the end.

4.3 Can Feature Aggregation/Combination Help?

We ask whether aggregating verbal confidence over time or combining multiple uncertainty metrics can overcome the information deficit. Figure 3(a) compares six temporal aggregation methods for verbal confidence. While running-min slightly improves early-stage AUROC (0.61 at 10% vs. 0.58 for last-step), the gain is negligible. No aggregation method surpasses the last-step value at completion. Figure 3(b) evaluates multi-feature combination using five last-step features (verbal confidence, PPL, entropy, max NLL, max entropy) with three classifiers (LR-L1, LR-L2, LightGBM) via 5-fold stratified CV. The additional token-probability features are predominantly noise at intermediate stages; classifiers fail to exceed verbal confidence at any progress point, confirming the information deficit is fundamental.

5 Path Switching: The Micro-Mechanism Behind Terminal Concentration

5.1 Path Switching

As introduced above, *path switching*, where agents abandon their current reasoning direction mid-trajectory, decouples intermediate uncertainty from the final outcome. Before a switch, the agent's confidence reflects its assessment of a path it will soon abandon, not the eventual outcome.

Table 5 illustrates this with a real trajectory from BrowseComp. The agent first searches by the "ambassador" constraint alone (Steps 1–2), then switches to add the "Nobel Prize" constraint (Step 3), and then further refines by combining "ambassador" and "Nobel Prize" (Step 4). It is only after this second switch, which synthesizes all three key constraints (headmaster, Nobel laureate, ambassador), that the search yields a productive lead (Step 5). The pre-switch confidence (70%, 65%) carries no information about the eventual success, as it reflects confidence in partial, unsuccessful search directions.

Metric	Qwen3.5-122B-A10B BC	BC-zh	HLE	GLM-4.7 BC	BC-zh	HLE
Single-trajectory						
Verbal Confidence (Tian et al., 2023)	**.904**	**.864**	.763	**.915**	**.854**	.744
Consistency-based (8 rollouts per question)						
Disc. sem. ent. (Farquhar et al., 2024)	.838	.800	**.783**	.844	.823	**.822**
Deg. unc. (Lin et al., 2023)	.849	.839	.748	.827	.839	.799
Diversity (Cole et al., 2023)	.840	.789	.778	.837	.820	.811
Variation ratio (Huang et al., 2025b)	.862	.814	.714	.823	.817	.743
Lexical sim. (Fomicheva et al., 2020)	.861	.814	.714	.823	.817	.742
FSD (Lyu et al., 2025)	.804	.776	.777	.801	.789	.779
LN pred. ent. (Kadavath et al., 2022)	.827	.716	.624	.782	.772	.669
Pred. entropy (Kadavath et al., 2022)	.832	.667	.613	.745	.745	.666
SentSAR (Duan et al., 2024)	.720	.685	.546	.757	.733	.649
Sem. entropy (Kuhn et al., 2023)	.746	.709	.690	.750	.576	.538

Table 4: Consistency-based metrics vs. Verbal Confidence. **Bold**: best in column; underlined: second best.

Q: Who was headmaster of a school in the 1940s–60s, taught a future Nobel laureate, and later became ambassador in the 1970s?
A: Francis L. Bartels (born June 1909)

Step	Tool	Conf	Sw.	Search query / Action
1	search	70%		headmaster turned ambassador 1970s
2	search	65%		former headmaster became ambassador 1970s
3	search	60%	←	headmaster 1940s 1960s student won **Nobel Prize**
4	search	55%	←	Nobel laureate attended school headmaster **ambassador**
5	visit	85%		Francis L. Bartels (Wikipedia)
6	search	90%		Mfantsipim School alumni Nobel Prize winner
7	visit	95%		Confirm answer

Table 5: Case study of path switching. Sw. = path switch. Pre-switch confidence reflects the abandoned direction, not the eventual outcome.

Dataset	% traj. with switch	Avg # switches / traj.
Detected by keywords such as "try a different approach"		
BrowseComp	86.8%	8.5
BC-zh	69.7%	4.7
HLE	36.9%	1.3

Table 6: Path switch prevalence by dataset.

Prevalence. Path switching is widespread: on BrowseComp, 87% of trajectories contain at least one detected switch (avg. 8.5 per trajectory), and on HLE, 37% (avg. 1.3) (Table 6). We detect switches via keyword matching on the agent's thinking text (24 regular expressions covering phrases like "let me try a different approach"), which is a conservative lower bound as many di-

rection changes lack explicit markers.

Pre-switch uncertainty is uninformative. If path switches reset the agent's reasoning direction, the discriminative power of uncertainty metrics should be near zero before the switch and only recover afterward. Figure 4(a) confirms this using |Spearman ρ| between each metric and correctness at each position relative to the last switch step. Before the switch (steps t-5 to t-1), $|\rho|$ is consistently below 0.15 across all metrics and datasets; after the switch, it rises monotonically to 0.3–0.5 by step t+5.

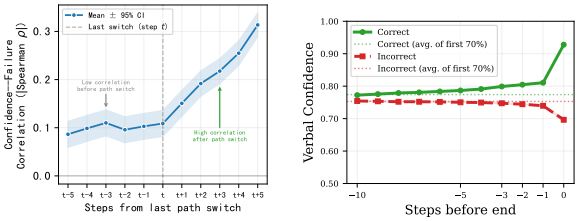

Figure 4: Micro-mechanism of early impotence: path switching. (a) Confidence–failure correlation (|Spearman ρ|) before and after the last path switch (step t). Before the switch, $|\rho|$ is near zero; after, it rises monotonically, confirming that only the post-switch portion carries predictive signal. (b) Confidence surge relative to the early-trajectory baseline. Correct trajectories exhibit a sharp confidence increase relative to the average of the first 70% of steps (dashed lines), while incorrect trajectories decline slightly.

5.2 Post-Switch Confidence Surge

If path switching decouples pre-switch uncertainty from the outcome, we should also observe a sharp confidence surge after the *last* switch, when the agent finds the correct direction. This is precisely what happens.

We define the **confidence surge** as the difference between the final answer verbal confidence and the average verbal confidence over the first 70% of the trajectory. Figure 4(b) shows the result: correct trajectories exhibit a strong positive surge (+15 points on average), while incorrect trajectories show a slight decline (−6 points on average). Dashed lines indicate the first-70% average baseline for each group. This pattern holds across all model–dataset combinations (see Appendix D for per-combination trajectories).

Together, pre-switch decoupling and post-switch surge explain why discriminative signal concentrates at the end of the trajectory: only the agent's *final* assessment, made after the last switch, correlates with the outcome.

6 Discussion

6.1 Sensitivity to Verbal Confidence Prompt

Our main experiments use in-tool-call confidence (Appendix A), where the model outputs a confidence score as a tool-call argument at each step. A natural concern is whether the observed patterns, namely early predictive impotence and near-completion confidence surge, are artifacts of this specific elicitation method. We also evaluate two alternatives: in-content confidence, where the model outputs confidence in `<confidence>` tags within its text output, and two-stage confidence (Tian et al., 2023), where a separate model call after each step rates the overall confidence that the task will be completed correctly given the trajectory prefix (Appendix A).

Figure 5(a) confirms on Qwen3.5-122B (averaged across two datasets) that all three methods produce the same qualitative trend: weak AUROC early followed by a sharp confidence surge, regardless of how confidence is elicited.

6.2 Beyond Deep Research Tasks

Our experiments focus on deep research and web browsing tasks (BrowseComp, BrowseComp-zh, HLE), where trajectories consist of information retrieval and synthesis steps. To test whether the observed patterns generalize, we evaluate GLM-4.7 on TerminalBench 2.0 (Merrill et al., 2026), a coding benchmark with 89 tasks requiring agents to complete terminal-based programming and system administration challenges (e.g., compiling projects, debugging scripts, configuring servers). Each task produces a long trajectory (avg. 51 steps) where the agent executes shell commands and receives terminal output, reporting confidence (0–100) at each step.

As Figure 5(b) shows, the coding task curve exhibits the same overall trend—early impotence followed by a near-completion surge—though with modest quantitative differences: intermediate AUROC is moderately higher (0.66–0.72 vs. 0.56–0.62) while the terminal surge is slightly weaker (0.79 at 100% vs. 0.78–0.93).

We hypothesize that this difference stems from fundamentally different environment dynamics. In deep research tasks, each search or visit action *does not alter* the environment, so the agent can

freely switch directions (path switching), decoupling early uncertainty from the final outcome. In coding tasks, each command *mutates* the environment (creating files, modifying code), making it costly to abandon the current direction. This may reduce path switching and allow intermediate confidence to retain some predictive power, while also making the final outcome harder to assess, as the agent may believe the task is done yet fail the test suite. Disentangling these mechanisms remains for future work.

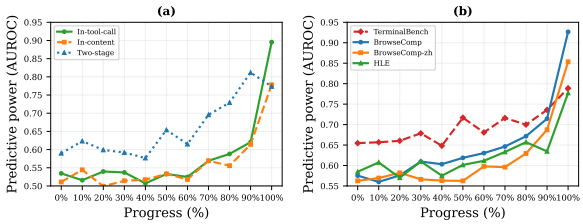

Figure 5: Sensitivity analysis. (a) Sensitivity to verbal confidence prompt: predictive power (AUROC) by normalized progress, comparing three elicitation methods. All methods show the same qualitative trend of early impotence followed by a confidence surge. (b) Sensitivity to task type: coding tasks (TerminalBench 2.0) vs. deep research benchmarks. Coding tasks show higher intermediate AUROC but a weaker confidence surge.

7 Implications for Agent Harness

What implications do our findings carry for the design of agent harnesses? **Our findings argue against mid-trajectory intervention based on uncertainty metrics**, for two reasons. First, intermediate-step uncertainty signals have weak predictive power for the final outcome. Second, path switching is pervasive in LLM-based agents. Since the agent already self-corrects through switching, intermediate-based intervention seems unwarranted.

To make this concrete, Figure 6 shows the failure-detection precision and recall at two confidence thresholds (τ=0.8 and τ=0.9) at 50% and 100% trajectory progress, averaged across all 15 model–dataset combinations. At 50% progress, the tradeoff is unacceptable regardless of threshold: τ=0.8 yields precision 0.68 / recall 0.62, while τ=0.9 achieves recall 0.96 but precision drops to 0.64—barely above the average failure rate (0.62). Neither threshold provides operationally useful discrimination. At 100% progress, the picture reverses: τ=0.8 achieves precision 0.95 with recall 0.46, and τ=0.9 achieves precision 0.88 with recall 0.76, confirming that confidence becomes genuinely informative only at task completion.

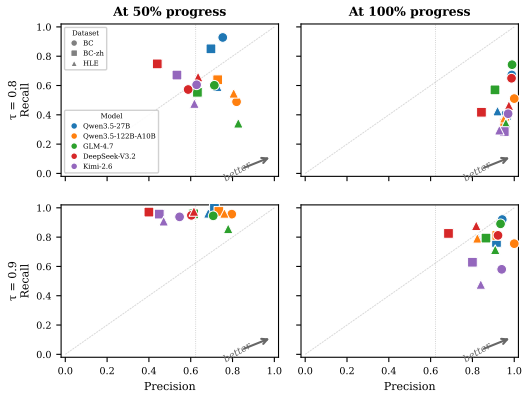

Figure 6: Failure detection precision and recall at two confidence thresholds (τ=0.8 and τ=0.9; predict failure when confidence $< \tau$). Columns compare 50% vs. 100% trajectory progress. At 50%, neither threshold provides operationally useful discrimination; at 100%, precision rises sharply while recall remains high. Each point is one model–dataset combination. Full PR curves in Appendix E.

Existing work has already explored confidence-based intervention at two distinct points in the agent loop:

- **Mid-trajectory intervention** (AUQ; Zhang et al. 2026a): monitors verbal confidence at every step; when it drops below a threshold (τ=0.8), the agent pauses to reflect and replan within the current trajectory, consuming additional steps before continuing.
- **Post-completion retry** (BrowseConf-Neg; Ou et al. 2025): waits until the agent finishes a full trajectory, then checks the final verbal confidence; if it falls below a threshold (τ=0.9), the agent restarts the entire task while informing the model that the previous answer was likely incorrect.

We reproduce both methods on our benchmark. ReAct (no intervention) serves as the shared baseline. AUQ is configured with the reflection threshold τ=0.8 and a best-of-N reflection budget, terminating when confidence rises above threshold or the maximum step count is reached. BrowseConf-Neg is run with up to N=10 retries (i.e. 10 rollouts total), restarting from scratch whenever the final confidence falls below τ=0.9, and conditioning each retry on the negation of the prior low-confidence answer.

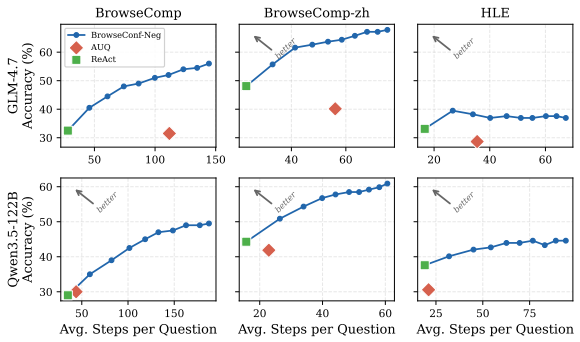

Figure 7: Agent harness: accuracy vs. efficiency. BrowseConf-Neg Pareto-dominates AUQ across all models and datasets. Upper-left is better (fewer steps, higher accuracy).

Figure 7 compares the two strategies in the accuracy–efficiency plane, where the horizontal axis measures *effective steps*—the total number of agent–environment interactions, including any steps consumed by AUQ's mid-trajectory reflection or BrowseConf-Neg's retries. Upper-left is better: fewer steps, higher accuracy.

BrowseConf-Neg Pareto-dominates AUQ across all models and datasets: at every comparable step budget, it achieves higher accuracy. Furthermore, BrowseConf-Neg's accuracy rises steadily with each additional retry, demonstrating that the confidence-guided restart mechanism effectively converts additional compute into accuracy gains.

7.1 Harness Recommendations

Based on the analysis above, we offer two recommendations for agent system design: **(1) Use final-step verbal confidence for decision-making.** Last-step AUROC reaches 0.86, while no metric exceeds 0.6 at early-to-medium stages. Intermediate-signal intervention incurs high false-positive cost and risks disrupting the model's own path-switching behavior. **(2) When final-step confidence is low, restart with prior-answer awareness** (Ou et al., 2025).

8 Related Work

Verbalized confidence (Tian et al., 2023), token probability-based methods (Kadavath et al., 2022), and consistency-based methods (Farquhar et al., 2024; Manakul et al., 2023) were originally proposed for single-turn question answering. When extended to multi-step LLM agents (Oh et al.), prior work varies along two axes: the choice of uncertainty signal and the aggregation strategy for per-step signals. BrowseConf (Ou et al., 2025) uses final-step verbal confidence for retry decisions, AUQ (Zhang et al., 2026a) monitors intermediate verbal confidence for mid-trajectory intervention, SAUP (Zhao et al., 2025) propagates step-level uncertainty with learned weights, and UProp (Duan et al., 2025) models uncertainty propagation across decision steps. All prior methods aggregate per-step uncertainty into a trajectory-level score, however, none has systematically analyzed *which steps* carry discriminative signal: SAUP and UProp assign learned weights to intermediate steps, implicitly assuming those steps contribute meaningfully to the final prediction, but this assumption has never been tested. AUQ finds that last-step aggregation outperforms averaging (0.913 vs. 0.783 AUROC on ALFWorld) but offers no explanation. Similarly, Agentic Confidence Calibration (Zhang et al., 2026b) finds that last-step token log-probability outperforms global-trace average on two of three benchmarks, but the result is mixed and no explanation is offered. Critically, whether uncertainty signals observed at early or intermediate stages can forecast the final outcome has not been explored. We fill this gap by systematically studying step-level uncertainty dynamics in long-horizon trajectories, revealing that intermediate steps carry essentially no predictive information due to path switching.

9 Conclusion

Verbal confidence reliably detects failures at task completion (AUROC 0.86) but is near-random at mid-trajectory (0.58 at 50% progress), and neither additional features nor nonlinear classifiers close this gap. Path switching, the tendency of agents to abandon search directions mid-trajectory, breaks the causal link between intermediate uncertainty and final outcomes, concentrating discriminative signal at the end of the trajectory. For agent harness design, our results show that when final-step confidence is low, restarting the task with prior-answer awareness (e.g. BrowseConf-Neg) provides a Pareto improvement over mid-trajectory intervention (e.g. AUQ), converting additional test-time compute into steadily higher accuracy.

Limitations and Future Work

As discussed in Section 6.2, our initial results on TerminalBench 2.0 (Merrill et al., 2026) suggest that coding tasks exhibit a different uncertainty profile from deep research tasks, with moderately higher early AUROC but a weaker confidence surge. However, our analysis of coding tasks is limited to a single model (GLM-4.7) and a single benchmark; confirming whether environment mutability and reduced path switching explain the observed differences requires controlled experiments across multiple models, multiple coding benchmarks, and explicit path-switch detection in coding trajectories. Investigating whether the post-switch confidence surge generalizes to mathematical reasoning and multi-turn dialogue domains also remains for future work.

References

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

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. volume 36, pages 28091–28114.

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.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

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.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025b. Look before you leap: An exploratory study of uncertainty analysis for large language models. *IEEE Transactions on Software Engineering*, 51(2):413–429.

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*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Baixuan Li, Dingchu Zhang, Jialong Wu, Wenbiao Yin, Zhengwei Tao, Yida Zhao, Liwen Zhang, Haiyang Shen, Runnan Fang, Pengjun Xie, and 1 others. 2025. Parallelmuse: Agentic parallel thinking for deep information seeking. *arXiv preprint arXiv:2510.24698*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2025. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19260–19268.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9004–9017.

Mike A Merrill, Alexander Glenn Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, Lin Shi, Jeong Yeon Shin, Thomas Walshe, E. Kelly

Buchanan, Junhong Shen, Guanghao Ye, Haowei Lin, Jason Poulos, Maoyu Wang, Marianna Nezhurina, Di Lu, Orfeas Menis Mastromichalakis, Zhiwei Xu, and 65 others. 2026. Terminal-bench: Benchmarking agents on hard, realistic tasks in command line interfaces. In *The Fourteenth International Conference on Learning Representations*.

Changdae Oh, Seongheon Park, To Eun Kim, Jiatong Li, Wendi Li, Samuel Yeh, Xuefeng Du, Hamed Hassani, Paul Bogdan, Dawn Song, and 1 others. Uncertainty quantification in llm agents: Foundations, emerging challenges, and opportunities. In *Agentic AI in the Wild: From Hallucinations to Reliable Autonomy*.

Litu Ou, Kuan Li, Huifeng Yin, Liwen Zhang, Zhongwang Zhang, Xixi Wu, Rui Ye, Zile Qiao, Pengjun Xie, Jingren Zhou, and 1 others. 2025. Browseconf: Confidence-guided test-time scaling for web agents. *arXiv preprint arXiv:2510.23458*.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. volume 36, pages 68539–68551.

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, Singapore. Association for Computational Linguistics.

Ningning Wang, Xavier Hu, Pai Liu, He Zhu, Yue Hou, Heyuan Huang, Shengyu Zhang, Jian Yang, Jiaheng Liu, Ge Zhang, and 1 others. 2025. Efficient agents: Building effective agents while reducing cost. *arXiv preprint arXiv:2508.02694*.

Xinyu Jessica Wang, Haoyue Bai, Yiyou Sun, Haorui Wang, Shuibai Zhang, Wenjie Hu, Mya Schroder, Bilge Mutlu, Dawn Song, and Robert D Nowak. 2026. The long-horizon task mirage? diagnosing where and why agentic systems break. *arXiv preprint arXiv:2604.11978*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

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*.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21381–21396.

Yuan-An Xiao, Pengfei Gao, Chao Peng, and Yingfei Xiong. 2025. Improving the efficiency of llm agent systems through trajectory reduction. *arXiv preprint arXiv:2509.23586*.

Hang Yan, Fangzhi Xu, Rongman Xu, Yifei Li, Jian Zhang, Haoran Luo, Xiaobao Wu, Luu Anh Tuan, Haiteng Zhao, Qika Lin, and 1 others. 2025. Mur: Momentum uncertainty guided reasoning for large language models. *arXiv preprint arXiv:2507.14958*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models.

Jiaxin Zhang, Prafulla Kumar Choubey, Kung-Hsiang Huang, Caiming Xiong, and Chien-Sheng Wu. 2026a. Agentic uncertainty quantification. *arXiv preprint arXiv:2601.15703*.

Jiaxin Zhang, Caiming Xiong, and Chien-Sheng Wu. 2026b. Agentic confidence calibration. *arXiv preprint arXiv:2601.15778*.

Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyou 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.

Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, and 1 others. 2025. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese. *arXiv preprint arXiv:2504.19314*.

Kunlun Zhu, Zijia Liu, Bingxuan Li, Muxin Tian, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, and 1 others. 2025. Where llm agents fail and how they can learn from failures. *arXiv preprint arXiv:2509.25370*.

Appendix

A Verbal Confidence Prompt Variants

We elicit verbal confidence using three prompt variants, differing in *how* the confidence score is collected and *what* it is intended to reflect.

In-tool-call (Step Utility). Confidence is collected as a required integer argument (`confidence`) in each tool call, inline with the agent's execution. The model is instructed to rate *how likely the current step is to yield high-value, relevant information*:

> *An integer between 0 and 100 representing your confidence that THIS STEP will yield high-value, relevant information. Think: "How likely is it that executing this action will retrieve information relevant for the current reasoning step?" Lower the score (e.g., below 50) if: 1) The parameters/queries are too broad, vague, or highly speculative; 2) The expected results are likely repetitive or redundant with existing context; 3) The action is based on an unverified or weak intermediate hypothesis; 4) Previous tool calls with similar parameters failed to yield valuable information.*

After providing the final answer, the model also outputs a confidence score in `<confidence>` tags.

In-content (Current Path). Confidence is collected via `<confidence>` tags in the assistant's text output before each tool call, rather than as a tool-call argument. The score estimates current-path task success: the probability that continuing along the current solution path will eventually lead to the correct final answer. The model outputs the score as a mandatory `<confidence>` block immediately before each tool call:

> *IMPORTANT PROTOCOL: When you need to call a tool, your output must follow this exact order:* `<think>...</think>` `<confidence>...</confidence>` `<tool_call>...</tool_call>`. *The* `<confidence>` *block is mandatory after every* `</think>`. *The token* `<tool_call>` *must never appear unless it is immediately preceded by a valid* `<confidence>...</confidence>` *block. Confidence level must be a number between 0 and 100. It represents the probability that continuing along the current solution path will eventually lead to the correct final answer.*

Two-stage (Current Path). Following the two-stage calibration protocol of Tian et al. (2023), confidence is collected via a separate model call after each step. The model is shown the trajectory prefix and asked to estimate the probability that continuing along this path will eventually produce the correct answer:

> *You are reviewing a problem-solving trajectory prefix. The trajectory stops immediately after one assistant message. Based only on this prefix, estimate the probability that continuing along this same problem-solving path from this point will eventually produce the correct final answer. Give ONLY the probability, no other words or explanation. For example: Probability:* `<the probability between 0 and 100>`

This approach completely separates confidence elicitation from the agent's execution: no changes to the agent's prompt, tool interface, or output format are required.

	In-tool-call (Step Utility)	In-content (Current Path)	Two-stage (Current Path)
Confidence location	Tool argument	Content tag	Separate call
Confidence scope	Step utility	Current Path	Current Path
Interference with agent	Yes	Yes	No

Table 7: Comparison of three verbal confidence prompt variants.

Summary. Table 7 summarizes the key differences.

B Full Trajectory Predictive Power by Metric Family

Table 8 shows the full-trajectory AUROC for all single-trajectory metric families and aggregation operators across 15 model–dataset combinations.

C Per-Model-Per-Dataset Predictive Power

Figure 8 shows the AUROC by normalized trajectory progress for each of the 15 model–dataset combinations individually, complementing the averaged view in Figure 1 of the main paper. The pattern—metrics remain near chance until 80–90% progress, then rise sharply at completion—is consistent across all models and datasets.

D Per-Model-Per-Dataset Verbal Confidence Trajectories

Figure 9 shows the verbal confidence trajectories over the last 10 steps for each model–dataset combination individually, complementing the averaged view in Figure 4(b) of the main paper.

		Qwen3.5-27B			Qwen3.5-122B-A10B			GLM-4.7			DeepSeek-V3.2			Kimi-2.6		
Family	**Aggregation**	**BC**	**BC-zh**	**HLE**	**BC**	**BC-zh**	**HLE**	**BC**	**BC-zh**	**HLE**	**BC**	**BC-zh**	**HLE**	**BC**	**BC-zh**	**HLE**
Single-trajectory																
Verbal Confidence (Tian et al., 2023)																
	last step	**.921**	**.867**	**.801**	**.904**	**.864**	**.763**	**.915**	**.854**	.744	**.906**	**.860**	**.862**	**.794**	**.816**	.790
	last 3 steps	.878	.821	.740	.845	.783	.740	.897	.825	**.749**	.863	.781	.806	.790	.722	**.813**
	last 5 steps	.859	.753	.722	.813	.719	.714	.879	.804	.732	.844	.748	.794	.783	.717	.807
	max	.917	.865	.800	.865	.848	.731	.863	.834	.694	.836	.830	.800	.791	.794	.787
	min	.737	.650	.634	.757	.652	.662	.804	.724	.626	.730	.618	.680	.690	.672	.738
	mean	.753	.736	.654	.739	.712	.691	.764	.746	.667	.682	.682	.713	.703	.686	.761
PPL (Malinin and Gales, 2021)																
	last step	**.842**	**.692**	.580	**.853**	**.752**	.601	.816	**.790**	.692	**.745**	.492	.481	.575	.534	.547
	last 3 steps	.794	.589	.647	.821	.693	.626	**.830**	.788	.697	.679	.517	.443	.532	.505	.573
	last 5 steps	.738	.488	.673	.784	.636	.610	.813	.765	.689	.635	.510	.448	.529	.462	.527
	max	.698	.669	.667	.722	.716	**.690**	.699	.755	.697	.742	**.557**	**.514**	**.636**	**.556**	**.664**
	min	.367	.401	.504	.493	.516	.513	.497	.522	.503	.307	.290	.319	.314	.421	.430
	mean	.671	.575	**.674**	.731	.693	.678	.731	.724	**.708**	.392	.273	.347	.446	.486	.505
Entropy (Malinin and Gales, 2021)																
	last step	**.841**	**.697**	.590	**.860**	**.765**	.606	.833	**.791**	.687	.746	.486	.424	.553	.547	.517
	last 3 steps	.786	.606	.658	.825	.706	.627	**.837**	.784	.694	.659	.531	.415	.494	.500	.532
	last 5 steps	.730	.500	.690	.793	.653	.612	.818	.756	.686	.600	.512	.415	.481	.458	.486
	max	.702	.689	.667	.736	.705	**.695**	.692	.748	.695	**.750**	**.575**	**.469**	**.607**	**.548**	**.669**
	min	.373	.397	.530	.517	.527	.516	.509	.518	.505	.355	.294	.294	.313	.430	.403
	mean	.677	.583	**.691**	.743	.699	.675	.736	.721	**.705**	.376	.268	.331	.435	.488	.496
Max NLL (Manakul et al., 2023)																
	last step	**.741**	.603	.627	**.750**	.654	.599	.694	.765	.680	**.761**	.547	**.565**	.744	.604	.533
	last 3 steps	.702	.576	.692	.690	.619	.635	.743	**.767**	**.697**	.587	.568	.458	.742	.590	.600
	last 5 steps	.642	.535	**.750**	.677	.589	.627	**.754**	.748	.684	.548	.551	.482	**.766**	.583	.665
	max	.692	**.672**	.681	.700	**.710**	**.682**	.720	.763	.673	.757	**.631**	**.565**	.665	**.622**	**.705**
	min	.350	.346	.549	.413	.452	.507	.418	.450	.493	.269	.308	.332	.454	.468	.485
	mean	.618	.578	.704	.662	.652	.654	.683	.683	.681	.349	.379	.353	.688	.552	.627
Max token entropy (Manakul et al., 2023)																
	last step	**.780**	.638	.652	**.796**	.701	.617	.736	**.778**	.688	**.812**	.493	.491	.700	.573	.546
	last 3 steps	.748	.573	.695	.746	.645	.644	.767	.741	**.699**	.654	.592	.415	**.818**	.666	.679
	last 5 steps	.700	.500	**.718**	.707	.620	.631	**.775**	.709	.677	.583	.544	.450	.812	.697	.693
	max	.717	**.668**	.706	.714	**.708**	**.679**	.741	.733	.690	.701	**.613**	**.618**	.761	**.742**	.736
	min	.340	.357	.568	.473	.483	.531	.460	.467	.494	.292	.259	.324	.643	.479	.542
	mean	.655	.527	.704	.691	.667	.659	.713	.664	.679	.399	.313	.370	.754	.682	**.753**

Table 8: Full-trajectory AUROC for all single-trajectory metric families and aggregation operators across 15 model–dataset combinations (companion to Table 3). `max`/`min` denote running extremum. Within each family, **bold** marks the best and underlined marks the second-best per column.

E Failure Detection Precision-Recall Curves

Figure 10 shows precision–recall curves for failure detection using verbal confidence at 50% and 100% trajectory progress for each model–dataset combination. At 50% progress, the PR curves hug the top-right corner with high recall but precision near the failure rate, confirming that confidence provides almost no discriminative signal at intermediate stages. At 100% progress, precision improves substantially while recall remains high, reflecting the terminal concentration of predictive signal.

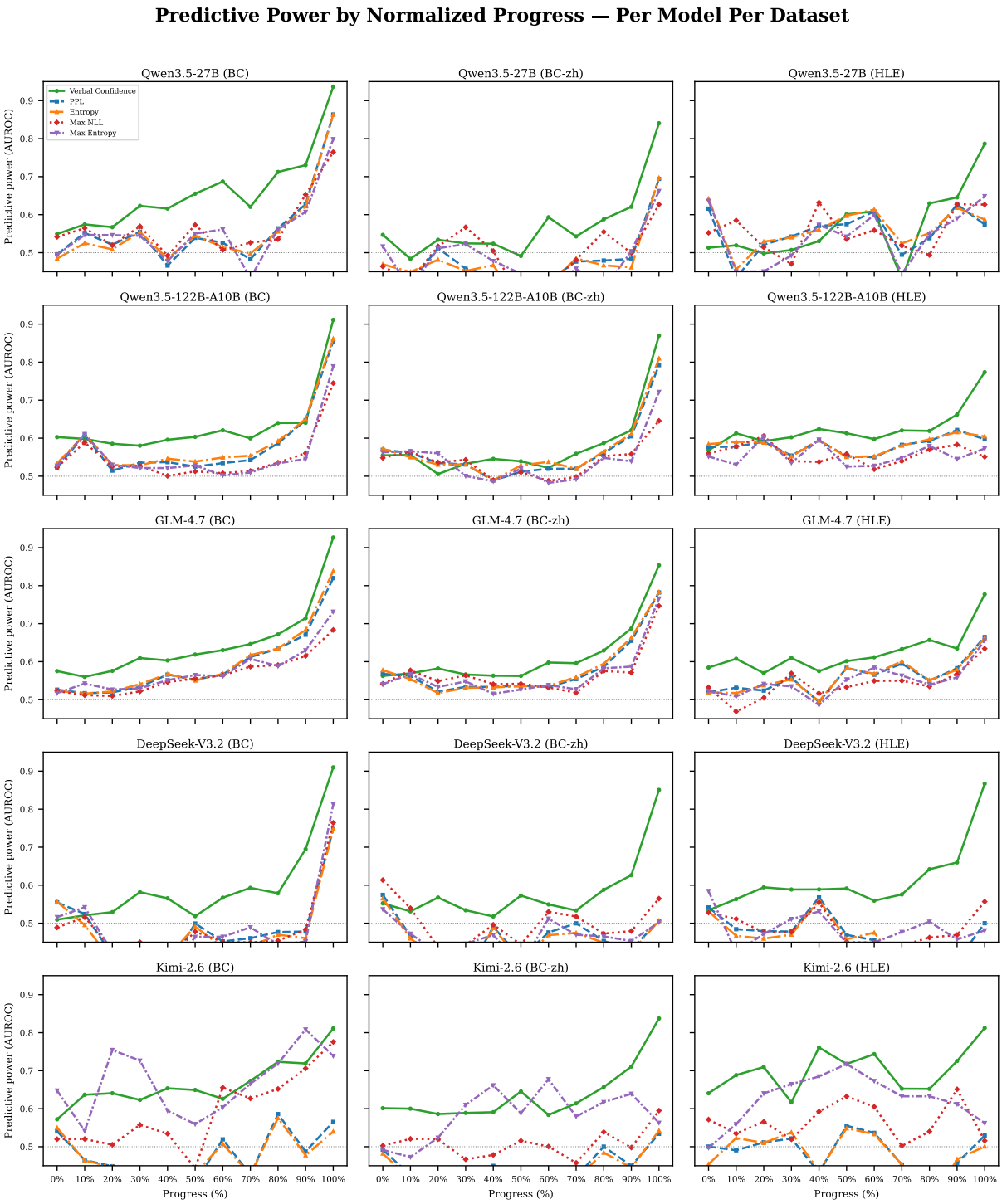

Figure 8: Predictive power of uncertainty metrics for agent failure (measured by AUROC) across normalized trajectory progress for each model–dataset combination. Each subplot shows five uncertainty metrics (verbal confidence, PPL, entropy, max NLL, max entropy). The early predictive impotence and terminal concentration pattern is consistent across all 15 settings.

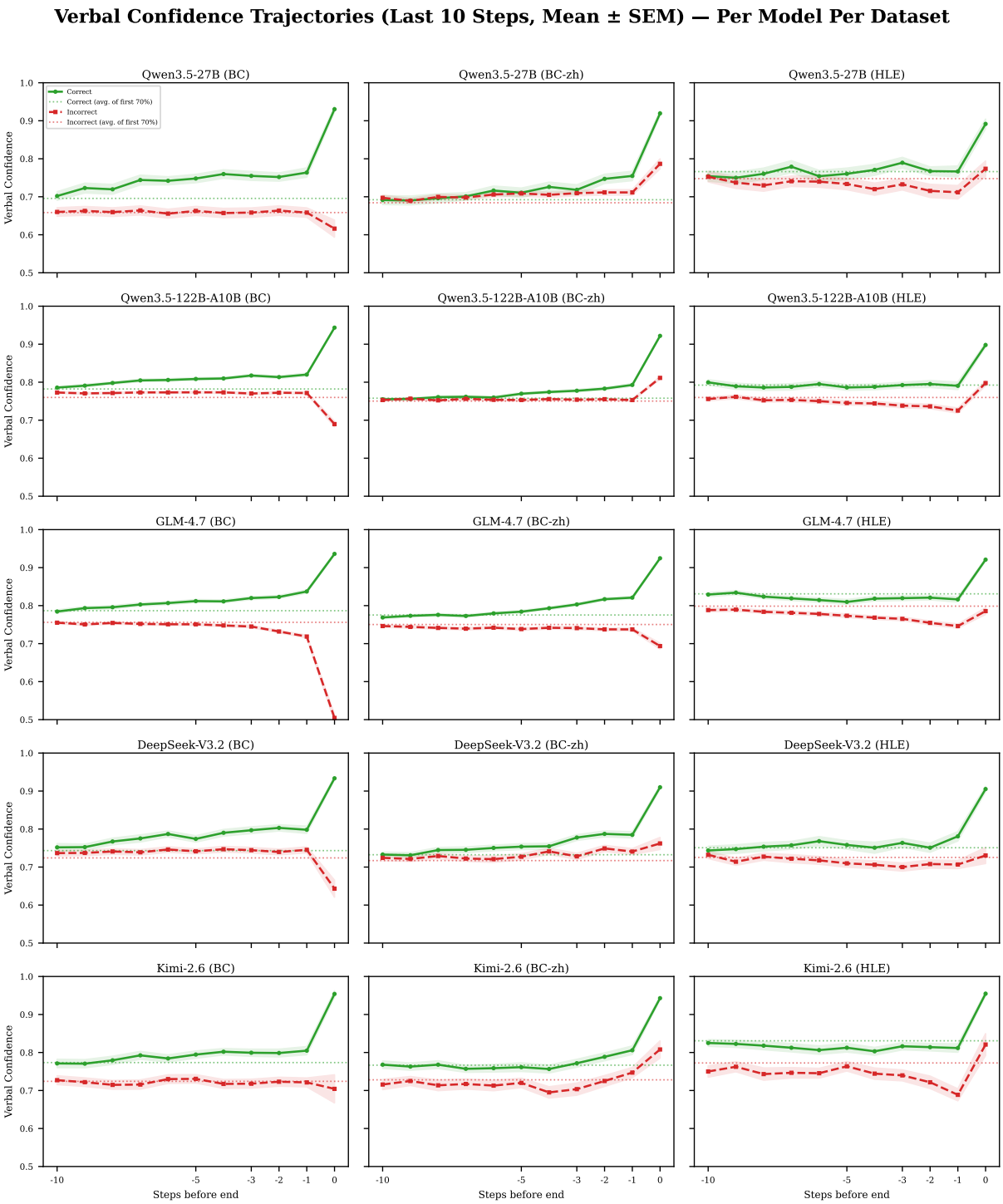

Figure 9: Verbal confidence trajectories over the last 10 steps for each model–dataset combination (shaded area: ± 1 SEM). Dashed lines indicate the average confidence over the first 70% of the trajectory for correct (green) and incorrect (red) trajectories. The confidence surge relative to the early-trajectory baseline is consistent across all settings, though the magnitude varies with task difficulty.

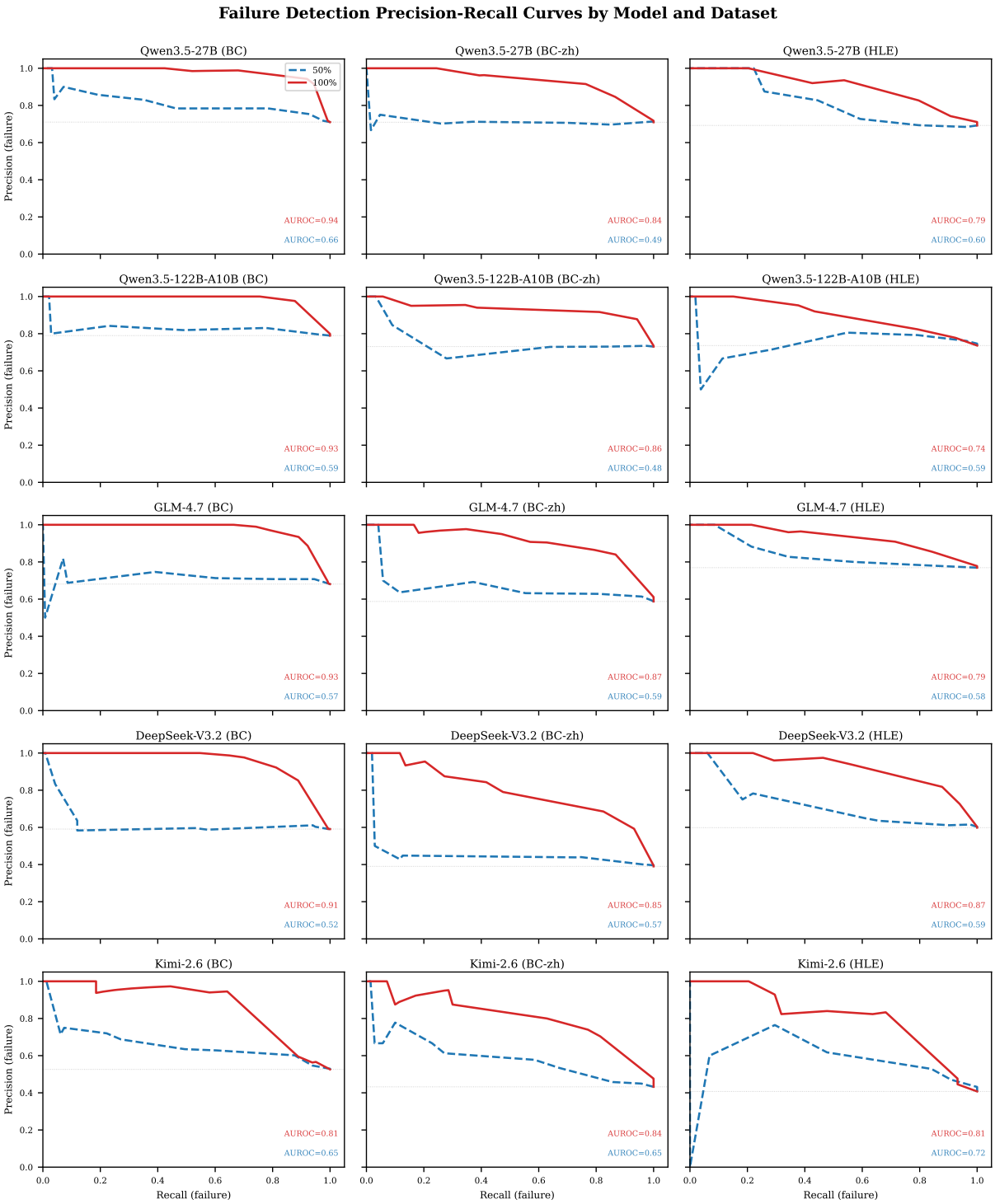

Figure 10: Precision–recall curves for failure detection using verbal confidence at 50% (blue dashed) and 100% (red solid) trajectory progress. Each subplot corresponds to one model–dataset combination. At 50% progress, precision is near the failure rate (gray dotted line), indicating near-random discrimination; at 100%, precision increases sharply while recall remains high.