Long-Horizon Reliability of Agentic Scientific Investigators

Controlled Experiments in a Synthetic Materials Environment

Ian Foster
Argonne National Laboratory & University of Chicago

Abstract

Recent progress in agentic AI has emphasized short-horizon task completion and verifiable domains such as software engineering. However, scientific research requires agents that operate persistently over long horizons, integrate noisy and unreliable tools, and accumulate epistemic state. We present a controlled experimental study of long-running investigator agents in a synthetic materials science environment. By systematically varying tool reliability and horizon length, we characterize performance degradation, identify phase transitions in agent behavior, and demonstrate that simple epistemic interventions—specifically, belief decay—can substantially improve robustness. Our results highlight the need for trajectory-level evaluation and explicit epistemic control mechanisms in agentic scientific systems.

1 Introduction

Most evaluations of AI systems focus on isolated model capabilities: benchmark accuracy, parameter scaling, or single-step reasoning. In contrast, deployed AI systems are increasingly *agentic*: persistent, stateful processes that plan, act, observe, and revise over many steps. This distinction is especially pronounced in scientific research, where progress emerges from long sequences of hypothesis generation, experimentation, and interpretation under uncertainty.

While agentic coding systems benefit from dense symbolic feedback and verifiable execution, scientific domains present additional challenges: tools may fail, measurements may be corrupted, and early errors can propagate over long horizons. Understanding how such factors affect long-running agent performance is therefore critical for AI-driven science.

In this work, we construct a minimal but expressive experimental harness for studying agentic scientific investigation. Using a synthetic materials environment with controllable noise and failure modes, we conduct systematic experiments to answer the following questions:

- How does agent performance degrade as tool reliability decreases?
- Are there qualitative phase transitions in long-horizon behavior?
- Can simple epistemic control mechanisms mitigate degradation?

2 Methods

2.1 Agentic Investigation Loop

We study a persistent investigator agent that executes a typed epistemic loop: *hypothesis generation*, *test design*, *execution*, and *interpretation*. The agent maintains an internal belief state over candidate materials and operates under an explicit budget on tool invocations. Termination occurs when either a candidate satisfying all constraints is found or the budget is exhausted.

Rather than evaluating individual actions, we treat each investigation as a single trajectory and assess outcomes at the run level.

2.2 Synthetic Materials Environment

To enable controlled experimentation, we introduce a synthetic oracle that emulates materials property queries (e.g., stability and bandgap). Given a set of candidate materials, the oracle returns deterministic base values perturbed by configurable noise and failure processes. This abstraction plays the role of simulation codes, databases, or experimental instruments, while allowing reproducibility and parameter sweeps.

2.3 Failure and Noise Model

Each oracle query is subject to two independent error processes:

- **Failure** with probability p_f, in which no result is returned.
- **Corruption** with probability p_c, in which the returned value is perturbed by large-magnitude noise.

By varying (p_f, p_c), we study how tool unreliability impacts long-horizon agent behavior.

2.4 Evaluation Metrics

We evaluate entire investigation trajectories using the following metrics:

1. Best objective score achieved during the run.
2. Step at which the first valid candidate appears.
3. Maximum stagnation length (consecutive steps without improvement).
4. Termination condition (success vs. budget exhaustion).

These metrics capture epistemic progress, delay, and failure modes that are not visible in stepwise accuracy measures.

2.5 Epistemic Forgetting

To mitigate error accumulation, we optionally apply exponential decay to the agent's belief state. At each interpretation step, belief values are multiplied by a factor $\lambda \in (0, 1]$. Setting $\lambda = 1.0$ yields pure persistence, while $\lambda < 1.0$ introduces forgetting that reduces the influence of early, potentially corrupted observations.

3 Experimental Design

We perform a grid of experiments varying:

- Tool failure probability $p_f \in \{0.0, 0.02, 0.05, 0.1\}$
- Tool corruption probability $p_c \in \{0.0, 0.02, 0.05\}$
- Belief decay factor $\lambda \in \{1.0, 0.98\}$

Each configuration is repeated multiple times, with the number of repetitions scaled adaptively based on expected variance (more repeats for high-noise regimes). All runs use a fixed horizon of 300 tool calls. Results are aggregated from event-sourced logs stored in a SQLite database and summarized into a CSV table for analysis.

3.1 Run-to-Run Variability

We observe substantial variance across runs even under identical experimental conditions. This variance arises from the interaction of stochastic candidate generation, noisy tool responses, and long-horizon stateful decision-making, which together induce path-dependent dynamics. Rather than treating this variability as experimental noise, we interpret it as an inherent property of persistent agentic systems. Accordingly, we report aggregate statistics and confidence intervals across repeated runs, and analyze both success probability and outcome quality. Figures display mean values with shaded regions indicating ± 1 standard deviation to visualize this inherent variability.

4 Results

4.1 Performance Degradation Under Tool Failures

Figure 1 shows the mean best objective score as a function of tool failure probability. Performance degrades smoothly at low failure rates but drops sharply beyond a critical threshold, indicating a phase transition in agent reliability. Measurement corruption accelerates this degradation across all conditions.

4.2 Delayed Discovery and Stagnation

As failure and corruption probabilities increase, agents require more steps to identify their first valid candidate (Figure 2). We also observe longer stagnation periods during which no improvement occurs (Figure 3), suggesting that early noise can trap the agent in self-reinforcing but incorrect beliefs.

4.3 Effect of Epistemic Forgetting

Introducing belief decay substantially improves robustness under noisy conditions. Across most reliability regimes, agents with forgetting achieve higher final scores, discover valid candidates earlier, and exhibit shorter stagnation periods. These results demonstrate that persistence alone is insufficient for long-horizon investigation; explicit mechanisms for epistemic revision are essential.

5 Discussion

Our experiments highlight a fundamental distinction between model-centric and agent-centric evaluation. While per-step correctness inevitably degrades with horizon length, appropriately designed agentic mechanisms can sustain epistemic progress even under unreliable tools. The observed phase transitions suggest that long-horizon reliability is governed by system-level properties rather than incremental model improvements.

Importantly, the belief decay mechanism studied here is deliberately simple. Its effectiveness indicates that many failure modes arise not from lack of expressivity but from unbounded accumulation of early errors.

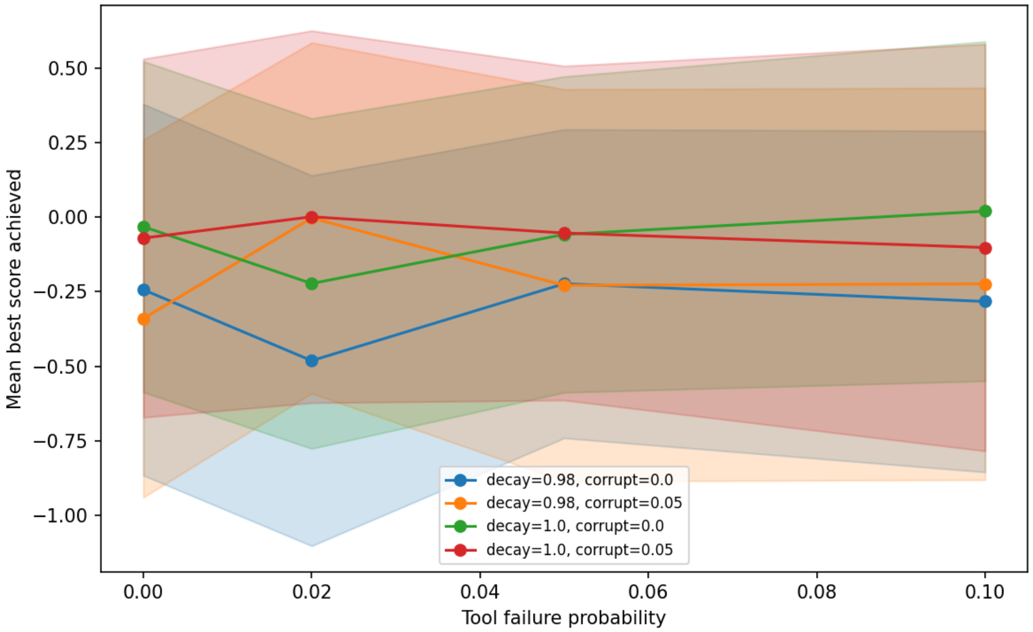

Figure 1: Mean best objective score versus tool failure probability. Curves correspond to different corruption probabilities and belief decay values. Shaded regions indicate ± 1 standard deviation across repeated runs.

6 Limitations and Future Work

This study uses a synthetic materials environment to enable controlled experiments. Future work will extend this framework to real materials datasets and simulators, incorporate learned policies for test selection, and evaluate richer epistemic controls such as periodic revalidation and hypothesis pruning. The same harness can also be used to quantify the marginal benefit of large language models when embedded within robust agentic systems.

7 Conclusion

We presented a controlled experimental study of long-running agentic investigators in a scientific setting. By varying tool reliability and horizon length, we identified qualitative failure modes and demonstrated that simple epistemic interventions significantly improve robustness. These results underscore the importance of trajectory-level evaluation and system-level design for AI-driven scientific discovery.

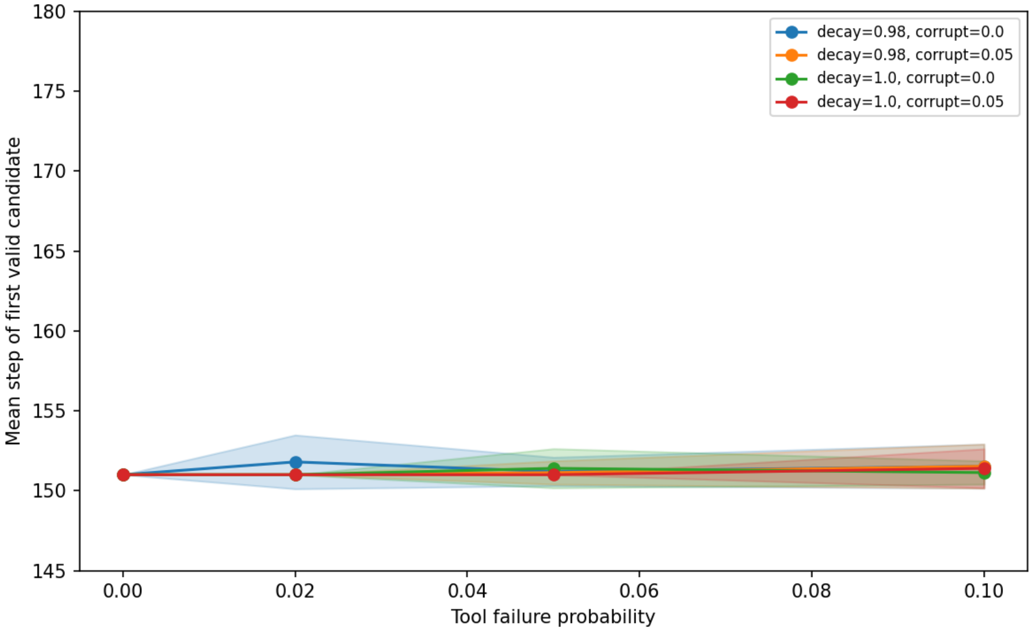

Figure 2: Mean step at which the first valid candidate appears as tool reliability decreases. Shaded regions indicate ±1 standard deviation.

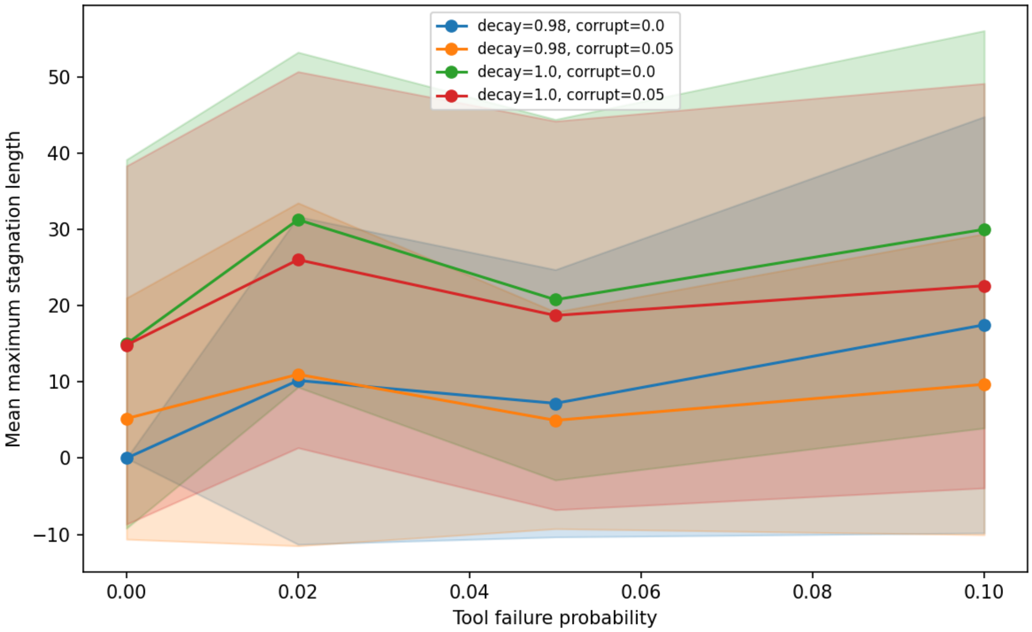

Figure 3: Mean maximum stagnation length as a function of tool failure probability. Shaded regions indicate ±1 standard deviation.