When Agents Disagree With Themselves: Measuring Behavioral Consistency in LLM-Based Agents

Aman Mehta [1]

Abstract

Run the same LLM agent on the same task twice—do you get the same behavior? We find the answer is often no. In a study of 3,000 agent runs across three models (Llama 3.1 70B, GPT-4o, and Claude Sonnet 4.5) on HotpotQA, we observe that ReAct-style agents produce 2.0–4.2 distinct action sequences per 10 runs on average, even with identical inputs. More importantly, this variance predicts failure: tasks with consistent behavior ($\leq$2 unique paths) achieve 80–92% accuracy, while highly inconsistent tasks ($\geq$8 unique paths) achieve only 27–64%—a 28–58 percentage point gap depending on model. We trace variance to early decisions: 69% of divergence occurs at step 2, the first search query. We also find that reducing temperature from 0.7 to 0.0 improves both consistency and accuracy (+5.4pp). Our results suggest that monitoring behavioral consistency during execution could enable early error detection and improve agent reliability.

1. Introduction

Large language model (LLM) based agents that use tools and multi-step reasoning are increasingly deployed for complex tasks (Yao et al., 2023; Schick et al., 2023). These agents interleave reasoning with actions—searching databases, calling APIs, or executing code—to accomplish goals that require multiple steps. As these systems move from research prototypes to production deployments, understanding their reliability becomes critical.

A fundamental but understudied question is: *how consistent are LLM agents in their behavior?* Given identical inputs, will an agent take the same actions, follow the same reasoning path, and arrive at the same answer? Or does the inherent stochasticity of LLM sampling lead to divergent behaviors even under controlled conditions?

[1]Independent Researcher. Correspondence to: Aman Mehta <amanmehta1997@gmail.com>.

Preprint. January 27, 2026.

This question matters for several reasons. First, inconsistent behavior complicates debugging and evaluation—if an agent fails, was it bad luck or a systematic problem? Second, inconsistency may signal uncertainty, which could be leveraged for error detection. Third, understanding variance sources could inform architecture decisions and deployment strategies.

We present a systematic empirical study of behavioral consistency in ReAct-style agents (Yao et al., 2023). Our key contributions are:

1. **Quantified behavioral variance across models**: We run 3,000 experiments (100 tasks $\times$ 10 runs $\times$ 3 models) and find substantial variance—agents produce 2.0–4.2 unique action sequences per 10 runs on average, with 18–55% step count variance.
2. **Consistency predicts correctness**: Tasks with consistent behavior ($\leq$2 unique sequences) achieve 80–92% accuracy; inconsistent tasks ($\geq$8 sequences) achieve only 27–64%—a 28–58 percentage point gap depending on model.
3. **Early divergence**: 69% of behavioral divergence occurs at step 2. The first search query largely determines the agent's trajectory.
4. **Path length as signal**: Short trajectories (3 steps) yield 90% accuracy with high consistency; long trajectories (8+ steps) yield 43% accuracy with high variance.
5. **Temperature matters**: Reducing temperature from 0.7 to 0.0 improves both consistency (4.2 $\rightarrow$ 2.2 unique sequences) and accuracy (+5.4pp).

These findings suggest that behavioral consistency is both measurable and predictive, with practical implications for agent evaluation, monitoring, and architecture design.

2. Related Work

LLM-Based Agents. ReAct (Yao et al., 2023) introduced the paradigm of interleaving reasoning traces with actions, enabling LLMs to use tools for multi-step tasks. Subsequent work has extended this to web browsing (Zhou et al., 2023), code execution (Yang et al., 2024), and multi-agent systems (Wu et al., 2023). Our work complements these efforts by studying the reliability of such agents under repeated execution.

Agent Consistency and Reliability. Most closely related to our work, τ-bench (Yao et al., 2024) introduced the pass^k metric to measure agent consistency across multiple trials, showing that GPT-4o's success rate drops from 60% (pass^1) to 25% (pass^8). While τ-bench quantifies *whether* agents are inconsistent, our work investigates *where* and *why*: we trace divergence to specific decision points (step 2), correlate consistency with correctness, and identify path length as a predictive signal. Stroebl et al. (2024) argue that current benchmarks overestimate agent capabilities; our findings support this claim.

Self-Consistency. Wang et al. (2023) showed that sampling multiple reasoning chains and taking majority vote improves accuracy on reasoning tasks. Our work differs in focus: rather than using consistency as an ensembling strategy, we study consistency as a diagnostic signal and analyze *where* and *why* variance occurs in agentic settings.

LLM Calibration and Uncertainty. Prior work has studied LLM calibration (Kadavath et al., 2022) and uncertainty quantification (Kuhn et al., 2023). We extend this to the agentic setting, where uncertainty manifests not just in final answers but in action sequences and reasoning traces.

3. Methodology

3.1. Agent Architecture

We implement a ReAct-style agent with three tools:

- **Search(query)**: Returns titles of documents matching the query via keyword matching over the HotpotQA context.
- **Retrieve(title)**: Returns the full text of a specific document.
- **Finish(answer)**: Terminates execution with a final answer.

The agent follows the standard ReAct loop: generate a thought, select an action, observe the result, and repeat until calling Finish or reaching a step limit.

3.2. Experimental Setup

Dataset. We use HotpotQA (Yang et al., 2018) validation set (distractor setting), which provides multi-hop questions requiring reasoning over multiple documents. Each question includes 10 paragraphs (2 gold, 8 distractors). We sample 100 questions, all classified as "hard" difficulty. Of these, 79 are bridge questions (requiring multi-hop reasoning) and 21 are comparison questions.

Models. We evaluate three models representing different providers:

- **Llama 3.1 70B Instruct** via Together AI (open-source)
- **GPT-4o** via OpenAI (closed-source)
- **Claude Sonnet 4.5** via Anthropic (closed-source)

Runs. For each question-model pair, we execute 10 independent runs with identical inputs and temperature 0.7, yielding 3,000 total runs (100 questions $\times$ 10 runs $\times$ 3 models). We also conduct a temperature ablation with Llama 3.1 70B at temperature 0.0 on 20 questions.

3.3. Metrics

We define several metrics to quantify behavioral consistency:

Answer Consistency. The fraction of runs producing the most common answer: $\frac{\max_a |\{r : \text{answer}(r) = a\}|}{N}$

Action Sequence Diversity. The number of unique action sequences (e.g., Search$\rightarrow$Retrieve$\rightarrow$Finish) across N runs.

Step Variance Ratio. $\frac{\max(\text{steps}) - \min(\text{steps})}{\text{mean}(\text{steps})}$, capturing how much path length varies.

First Divergence Point. The earliest step at which runs take different actions.

Correctness. We use fuzzy string matching: an answer is correct if it contains or is contained in the gold answer (case-insensitive).

4. Results

4.1. Overall Model Comparison

Table 1 summarizes our findings across 3,000 runs.

Figure 1 shows the distribution of behavioral consistency across models.

Claude Sonnet 4.5 achieves both the highest accuracy and highest consistency (fewest unique sequences). Llama 3.1 70B shows the most behavioral variance, with 4.2 unique action sequences per 10 runs on average. The distribution of correctness is notably bimodal: 58% of tasks achieve

Table 1. Performance comparison across three models (100 tasks, 1,000 runs each).

Model	Correct	Uniq Seqs	Avg Steps	St
Claude Sonnet 4.5	**81.9%**	**2.0**	5.0	1
Llama 3.1 70B	77.4%	4.2	4.6	[illegible]
GPT-4o	74.0%	2.4	4.6	[illegible]

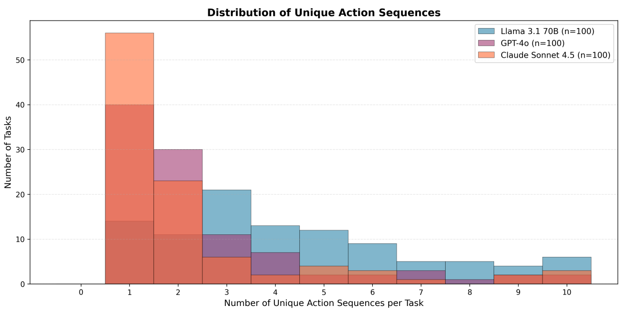

Figure 1. Distribution of unique action sequences per task. Claude and GPT-4o cluster heavily at 1–2 unique sequences (70–80% of tasks), indicating high consistency, while Llama 3.1 70B shows a more spread distribution.

100% correctness (all 10 runs correct), while 22% achieve below 50%. Agents either solve a task reliably or struggle consistently.

4.2. Consistency Predicts Correctness

Our central finding is that behavioral consistency strongly predicts correctness across all models. Table 2 shows the relationship.

Table 2. Correctness by consistency level across models. The "n" column shows the number of consistent/inconsistent tasks.

Model	Consistent ($\leq$2 seqs)	Inconsistent ($\geq$8 seqs)	Gap	n
Claude Sonnet 4.5	85.1%	27.4%	**57.7pp**	80/3
GPT-4o	80.4%	33.3%	47.1pp	71/3
Llama 3.1 70B	92.0%	64.0%	28.0pp	25/1

Figure 2 visualizes this relationship.

The gap is substantial across all models: **28–58 percentage points** depending on the model. For Llama 3.1 70B, the difference is statistically significant (t-test: $t = 2.70$, $p = 0.011$; Mann-Whitney U: $p < 0.001$). This suggests consistency could serve as a runtime signal for answer reliability.

4.3. Divergence Occurs Early

Where do runs first diverge? Analyzing Llama 3.1 70B (which shows the most variance), we find that 69% of divergence occurs at step 2—the first search query after the initial reasoning step.

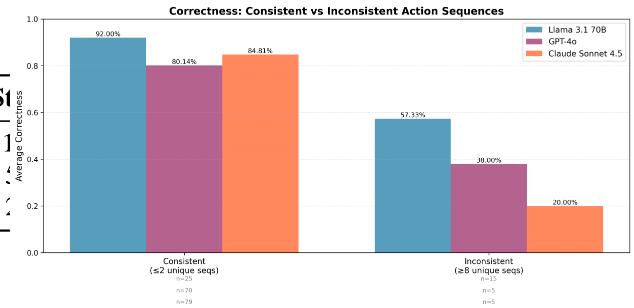

Figure 2. Correctness comparison between consistent ($\leq$2 unique sequences) and inconsistent ($\geq$8 unique sequences) tasks across three models. All models show a substantial gap, with consistent tasks achieving 80–92% accuracy versus 27–64% for inconsistent tasks.

Table 3. Distribution of first divergence point (Llama 3.1 70B).

First Divergence	Tasks	Avg Correct
Step 1	1	—
Step 2	59 (69%)	71.7%
Step 3+	26 (30%)	85.8%

Tasks that maintain consistency through step 2 achieve 85.8% accuracy, compared to 71.7% for early-diverging tasks. The first search query largely determines the agent's trajectory.

4.4. Path Length Correlates with Outcomes

We observe consistent patterns relating path length to both consistency and correctness (Llama 3.1 70B):

- **Perfectly consistent tasks** (1 unique sequence, $n = 14$): Average 3.4 steps, 85.7% correct.
- **Highly inconsistent tasks** ($\geq$9 unique sequences, $n = 10$): Average 7.8 steps, 43% correct.

The correlation between mean steps and correctness is $r = -0.34$. Longer paths indicate the agent is searching, backtracking, and uncertain—each additional step is an opportunity to diverge and err.

4.5. Temperature Ablation

We investigate whether reducing sampling temperature improves consistency. Table 4 shows results for Llama 3.1 70B on a subset of 20 questions.

Reducing temperature from 0.7 to 0.0 improves both consistency (4.2 $\rightarrow$ 2.2 unique sequences) and accuracy (+5.4pp).

Table 4. Temperature ablation (Llama 3.1 70B, 20 questions).

Temperature	Correctness	Unique Seqs
0.0	**82.8%**	**2.2**
0.7	77.4%	4.2

This suggests that for production deployments, lower temperature settings may be preferable.

4.6. Question Type Analysis

We compare bridge questions (multi-hop reasoning, $n = 79$) with comparison questions (yes/no style, $n = 21$) using Llama 3.1 70B:

Table 5. Performance by question type (Llama 3.1 70B).

Metric	Bridge	Comparison
Correctness	75.7%	80.0%
Answer Consistency	76.6%	62.4%
Step Variance	63%	41%

Interestingly, comparison questions show *higher* correctness but *lower* consistency. The constrained answer space (yes/no) improves accuracy, but explanations vary, reducing measured consistency. This highlights that answer consistency and explanation consistency are distinct dimensions.

5. Discussion

Differentiation from Prior Work. While τ-bench (Yao et al., 2024) established that agents are inconsistent (pass^k drops rapidly with k), our work provides complementary insights: (1) we identify *where* variance originates (69% at step 2), (2) we quantify the consistency-correctness relationship across three models (28–58pp gap), (3) we show path length predicts both consistency and correctness, and (4) we demonstrate that temperature is a key lever for controlling consistency. These findings suggest actionable interventions beyond simply measuring pass^k.

Consistency as a Runtime Signal. The strong correlation between consistency and correctness suggests a practical intervention: run multiple parallel executions and check for agreement. If runs diverge early, the answer is less likely to be correct. This could enable selective human review or automatic retry strategies.

The Step 2 Bottleneck. The concentration of divergence at step 2 suggests the first search query is critical. Improving query formulation—through better prompting, query expansion, or learned retrievers—could reduce downstream variance.

Implications for Complex Agents. Our study uses a minimal action space of three tools, yet we observe substantial behavioral variance. Real-world agents often have dozens of tools and require many more steps to complete tasks. Since each action choice is a potential divergence point, we hypothesize that consistency challenges grow combinatorially with action space size and task complexity. The negative correlation between path length and correctness ($r = -0.34$) supports this concern: agents solving harder problems requiring more steps will likely exhibit even greater variance. This suggests that consistency monitoring may be especially critical for capable, general-purpose agents.

Model Selection. Our results suggest Claude Sonnet 4.5 may be preferable for applications requiring reliability, as it achieves both highest accuracy (81.9%) and consistency (2.0 unique sequences). For open-source deployments, Llama 3.1 70B with reduced temperature (0.0–0.3) offers a reasonable alternative.

Limitations. Our study uses one benchmark (HotpotQA) and lexical search. Future work should validate across diverse tasks (coding, web navigation) and with semantic retrieval. The temperature ablation uses a smaller sample (20 questions) and should be expanded.

6. Conclusion

We present a systematic study of behavioral consistency in LLM-based agents across three models. Our key finding is that consistency predicts correctness: agents that behave consistently achieve 80–92% accuracy, while inconsistent agents achieve only 27–64%—a gap of 28–58 percentage points. Divergence occurs early (step 2), path length serves as a reliable signal, and reducing temperature improves both consistency and accuracy. These findings have practical implications for agent deployment: monitoring behavioral consistency could enable early error detection, and temperature tuning offers a simple lever for improving reliability.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. 2023.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 2023.

Stroebl, B., Kapoor, S., and Narayanan, A. Ai agents that matter. *arXiv preprint arXiv:2407.01502*, 2024.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. 2023.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Yang, J., Jimenez, C. E., Wettig, A., Liber, K., Narasimhan, K., and Press, O. Swe-agent: Agent-computer interfaces enable automated software engineering. *arXiv preprint arXiv:2405.15793*, 2024.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. 2018.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models. 2023.

Yao, S., Shinn, N., Razavi, P., and Narasimhan, K. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Bisk, Y., Fried, D., Alon, U., et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.