Information Fidelity in Tool-Using LLM Agents: A Martingale Analysis of the Model Context Protocol

Extended Abstract

Flint Xiaofeng Fan
Centre for Frontier AI Research, A*STAR
Singapore
ETH Zurich
Zurich, Switzerland
fxf@u.nus.edu

Cheston Tan
Centre for Frontier AI Research, A*STAR
Singapore
Institute of High Performance Computing, A*STAR
Singapore
cheston-tan@i2r.a-star.edu.sg

Roger Wattenhofer
ETH Zurich
Zurich, Switzerland
wattenhofer@ethz.ch

Yew-Soon Ong
Centre for Frontier AI Research, A*STAR
Singapore
College of Computing and Data Science, NTU
Singapore
asysong@ntu.edu.sg

ABSTRACT

As LLM agents increasingly rely on external tools via the Model Context Protocol (MCP), a key reliability question is whether small early mistakes can snowball across multi-step tool chains. We show this snowballing is avoidable: under bounded, decaying influence between steps, cumulative semantic distortion grows at most linearly with high-probability deviations of order $O(\sqrt{T})$, so random errors do not amplify superlinearly with chain length. We operationalize distortion with a hybrid metric that combines weighted fact matching and embedding similarity, and we certify the bound via a Doob-martingale concentration argument. Experiments on different open-sourced LLM architectures match the predicted linear trend and $\sqrt{T}$ envelopes, yielding practical design rules such as re-grounding intervals and semantic weighting.

KEYWORDS

LLM Agents, Tool Use, Model Context Protocol, Martingale Theory, Information Fidelity, Error Propagation

ACM Reference Format:
Flint Xiaofeng Fan, Cheston Tan, Roger Wattenhofer, and Yew-Soon Ong. 2026. Information Fidelity in Tool-Using LLM Agents: A Martingale Analysis of the Model Context Protocol: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/GMHB4353

1 INTRODUCTION

LLM-based agents increasingly rely on external tools—databases, APIs, calculators—to overcome the limitations of static training knowledge [9, 11]. The Model Context Protocol (MCP) [2] has been proposed as an open, unifying interface for tool-augmented LLM applications, replacing bespoke $M{\times}N$ integrations with a unified JSON-RPC framework supported in emerging ecosystems [8]. As these agents mediate high-stakes decisions in domains from clinical support to financial analysis, formal reliability guarantees transition from desirable to essential [4, 5, 7].

But tool use introduces a subtle reliability risk. Each call to an external tool is an opportunity for factual error or semantic drift. In sequential chains, where each query depends on previous responses, small early mistakes can compound. The question is: *how badly?* Could errors grow exponentially, making long tool chains fundamentally unreliable? Or does some structure prevent catastrophic accumulation?

Figure 1 illustrates the intuition[1]. Without any grounding mechanism, an agent's internal representation of a concept (here, a cat)

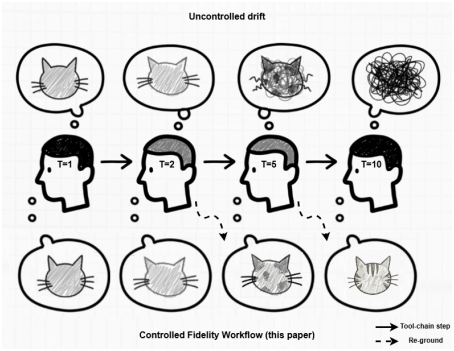

Figure 1: The drift problem in tool chains. (Top) Without control, an agent's representation degrades across sequential tool calls, becoming unrecognizable by step 10. (Bottom) Periodic re-grounding, guided by our theoretical analysis, keeps representations faithful.

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/GMHB4353

[1] A full version of this work is available on arXiv [6].

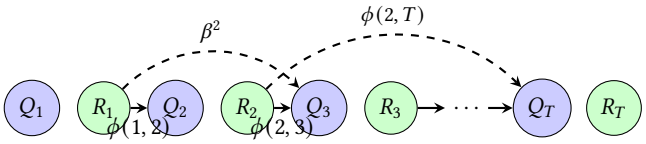

Figure 2: Dependency graph for MCP interactions. Solid arrows indicate direct influence $\phi(i, i+1) = \beta$; dashed arrows show long-range decay $\phi(i, j) = \beta^{j-i}$.

drifts further from truth with each step—eventually becoming unrecognizable. Our framework shows catastrophic drift is *not* inevitable: bounded influence and periodic re-grounding keep the representation faithful.

We provide, to our knowledge, the first formal framework for analyzing error accumulation in MCP-style tool-using LLM agents. Our contributions are: (1) a *hybrid semantic distortion metric* combining weighted fact matching with embedding-based similarity; (2) *martingale concentration bounds* proving $O(\sqrt{T})$ deviation from linear expected distortion via Azuma's inequality [3]; and (3) *empirical validation* across different open-sourced LLM models.

2 FRAMEWORK AND MAIN RESULTS

Distortion Metric. We measure information fidelity through a hybrid metric combining discrete fact matching with continuous semantic similarity:

$$\Delta_t = (1-\lambda)\, d_{\text{set}}^{w}(R_t, \mathcal{I}_t) + \lambda\, d_{\text{emb}}(R_t, \mathcal{I}_t), \tag{1}$$

where d_{set}^{w} is a weighted Jaccard distance over extracted facts, d_{emb} is the normalized cosine distance in embedding space, $\mathcal{I}_t$ is the ideal fact set for step t, and $\lambda \in [0, 1]$ tunes the trade-off. Both components are normalized so $\Delta_t \in [0, 1]$ (0 means fully faithful, 1 means maximally distorted), and d_{set}^{w} is strict about missing/incorrect facts while d_{emb} treats paraphrases as close when meaning is preserved. The cumulative distortion over T tool calls is $D(T) = \sum_{t=1}^{T} \Delta_t$.

Assumptions and Martingale Setup. MCP tool use induces an adaptive history $\mathcal{F}_t$ (all queries and responses up to step t). We assume *bounded, decaying influence*: (i) bounded branching $\beta B < 1$ with influence $\phi(i, j) = \beta^{j-i}$ and branching factor B; (ii) response stability; and (iii) temporal decay with sensitivity $\alpha > 0$ controlling how strongly a single-step perturbation affects future distortions.

Under these conditions, we construct a Doob martingale $Z_t = \mathbb{E}[D(T) \mid \mathcal{F}_t]$ whose increments are uniformly bounded: $|Z_{t+1} - Z_t| \le 1 + \frac{\alpha}{1-\beta B}$, where Z_t is a running *forecast* of the final distortion given the current tool-call history. The increment bound says that observing one additional tool response cannot swing this forecast by more than a constant, because any downstream effect on future steps decays geometrically (the tail sums to $\alpha/(1-\beta B)$).

Theorem 2.1 (High-Probability Distortion Bound). *Under bounded branching, response stability, and temporal decay, for any $\eta \in (0, 1)$,*

$$\Pr\left[D(T) - \mathbb{E}[D(T)] \ge \sqrt{2T(1+\gamma^*)\ln\tfrac{1}{\eta}}\right] \le \eta,$$

where $\gamma^* = 2C^* + (C^*)^2$ and $C^* = \alpha/(1-\beta B)$.

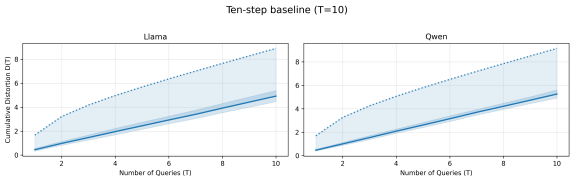

Figure 3: Baseline distortion accumulation over 10 tool calls for experiments with Llama and Qwen. Cumulative distortion with $\beta = 0.7$, $\lambda = 0.5$ (50 chains/model). Solid: empirical mean $\pm 1\sigma$; dotted: high-probability envelopes (Theorem 2.1).

Equivalently, with probability at least $1 - \eta$, $D(T) \le \mathbb{E}[D(T)] + \sqrt{2T(1+\gamma^*)\ln\frac{1}{\eta}}$. The substantive guarantee is concentration: under bounded, decaying influence, cumulative distortion stays within $O(\sqrt{T})$ of its mean (up to dependence-inflated constants), so stochastic error does not amplify superlinearly with chain length. Dependencies enter only through $C^* = \alpha/(1-\beta B)$: as $\beta B \to 1$, the bound remains valid but becomes conservative, motivating more frequent re-grounding, tighter tool validation, or reduced fan-out via gating/serialization.

3 EXPERIMENTS

We validate our theoretical predictions across Qwen2-7B-Instruct [10] and Llama-3-8B-Instruct [1] using deterministic MCP tools.

Key findings include (1) Cumulative distortion tracks the predicted linear trend at a constant per-step rate of $\approx$0.5, with all empirical trajectories falling within $O(\sqrt{T})$ theoretical envelopes (Figure 3). (2) Increasing semantic weight λ from 0 to 1 reduces distortion by ~80%, revealing that exact fact matching accumulates errors more aggressively than semantic similarity—every paraphrase contributes distortion at λ=0 even when meaning is preserved. (3) Even at extreme dependency (β=0.98) over extended chains (T=60), the system avoids exponential failure—high dependencies inflate the variance bounds rather than the mean distortion rate. (4) Different architectures exhibit similar distortion patterns, consistent with the bounds depending primarily on dependency structure (β, B) and metric properties (λ), not internal model mechanisms.

4 CONCLUSION AND IMPLICATIONS

We analyze tool-using LLM agents as a *sequential pipeline*: each tool call, paraphrase, or aggregation step can introduce a small distortion, and these distortions may correlate over time. Our dependence-aware concentration bound formalizes when such distortions behave like *accumulating noise* (roughly linear drift with controlled fluctuations), and makes explicit which *system-level factors* govern risk (e.g., temporal dependence / "memory"), rather than relying on opaque model internals. Practically, the analysis converts deployment into a small set of controllable knobs: (i) cap the length of *ungrounded* action chains or periodically *re-ground* to trusted sources, (ii) reduce lossy transformations (e.g., aggressive paraphrasing/summarization) when exactness is required, and (iii) monitor distortion proxies online to adapt the re-grounding cadence as workloads change. Full proofs, estimation details, and extended experiments are provided in the arXiv full version [6].

ACKNOWLEDGMENTS

The work is partly supported by the National Research Foundation (NRF), Singapore, through the AI Singapore Programme under the project titled "AI-based Urban Cooling Technology Development" (Award No. AISG3-TC-2024-014-SGKR).

REFERENCES

[1] AI@Meta. 2024. Llama 3 Model Card. (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[2] Anthropic. 2024. Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol. Published Nov 25, 2024.

[3] Kazuoki Azuma. 1967. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal* 19, 3 (1967), 357–367.

[4] Flint Xiaofeng Fan, Cheston Tan, Yew-Soon Ong, Roger Wattenhofer, and Wei-Tsang Ooi. 2025. FedRLHF: A Convergence-Guaranteed Federated Framework for Privacy-Preserving and Personalized RLHF. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems* (Detroit, MI, USA) *(AAMAS '25)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 713–721.

[5] Flint Xiaofeng Fan, Cheston Tan, Roger Wattenhofer, and Yew-Soon Ong. 2025. Position Paper: Rethinking Privacy in RL for Sequential Decision-making in the Age of LLMs. In *International Joint Conference on Neural Networks*. https://arxiv.org/abs/2504.11511

[6] Flint Xiaofeng Fan, Cheston Tan, Roger Wattenhofer, and Yew-Soon Ong. 2026. Information Fidelity in Tool-Using LLM Agents: A Martingale Analysis of the Model Context Protocol. *arXiv preprint arXiv:2602.13320* (2026).

[7] Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Wei Jing, Cheston Tan, and Bryan Kian Hsiang Low. 2021. Fault-tolerant federated reinforcement learning with theoretical guarantee. *Advances in neural information processing systems* 34 (2021), 1007–1021.

[8] Microsoft Azure AI Team. 2025. Model Context Protocol (MCP): Integrating Azure OpenAI for Enhanced Tool Integration and Prompting. https://techcommunity.microsoft.com/blog/azure-ai-services-blog/model-context-protocol-mcp-integrating-azure-openai-for-enhanced-tool-integratio/4393788.

[9] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems*.

[10] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report, 2024. *URL https://arxiv. org/abs/2407.10671* 7 (2024), 8.

[11] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*.