Drift No More? Context Equilibria in Multi-Turn LLM Interactions

Vardhan Dongre[1,2] * **Ryan A. Rossi**[2] **Viet Dac Lai**[2]
David Seunghyun Yoon[2] **Dilek Hakkani-Tür**[1] **Trung Bui**[2]

[1]University of Illinois Urbana-Champaign [2]Adobe Research

Abstract

Large Language Models (LLMs) excel at single-turn tasks such as instruction following and summarization, yet real-world deployments require sustained multi-turn interactions where user goals and conversational context persist and evolve. A recurring challenge in this setting is context drift: the gradual divergence of a model's outputs from goal-consistent behavior across turns. Unlike single-turn errors, drift unfolds temporally and is poorly captured by static evaluation metrics. In this work, we present a study of context drift in multi-turn interactions and propose a simple dynamical framework to interpret its behavior. We formalize drift as the turn-wise KL divergence between the token-level predictive distributions of the test model and a goal-consistent reference model, and propose a recurrence model that interprets its evolution as a bounded stochastic process with restoring forces and controllable interventions. We instantiate this framework in both synthetic long-horizon rewriting tasks and realistic user–agent simulations such as in τ-bench, measuring drift for several open-weight LLMs that are used as user simulators. Our experiments consistently reveal stable, noise-limited equilibria rather than runaway degradation, and demonstrate that simple reminder interventions reliably reduce divergence in line with theoretical predictions. Together, these results suggest that multi-turn drift can be understood as a controllable equilibrium phenomenon rather than as inevitable decay, providing a foundation for studying and mitigating context drift in extended interactions.

1 Introduction

Large Language Models (LLMs) have become central to a wide range of interactive systems, from virtual assistants and copilots to autonomous agents (Ouyang et al. 2022; Achiam et al. 2023; Brown et al. 2020; Acikgoz et al. 2025) that plan (Yao et al. 2023; Wang et al. 2023c; Li et al. 2023; Dongre et al. 2024), explain (Cai et al. 2019), or negotiate (Lewis et al. 2017; Bianchi et al. 2024) over extended dialogues. Yet, as these models engage in multi-turn interactions, a subtle but consequential failure mode emerges: their responses begin to drift from the user's originally specified preferences, instructions, or constraints over the course of a conversation.

Unlike factual hallucinations (Ji et al. 2023) or local coherence errors, *context drift* is a slow erosion of intent: a summarizer that gradually loses the requested tone, an image editing agent that drifts from the target aesthetic in an image, and a user simulator that forgets its goals and behavioral constraints. Critically, most current benchmarks and evaluations are blind to this degradation, focusing either on end-task success (Thoppilan et al. 2022; Zhou et al. 2023) or per-turn quality (Guan et al. 2025; Kwan et al. 2024; Dongre et al. 2025; Wang et al. 2023b; Chang et al. 2024; Duan et al. 2023), without capturing temporal misalignment across turns.

The prevailing intuition is that context drift accumulates unboundedly as conversations lengthen, owing to memory limits, information loss, or compounding errors. This view suggests that alignment inevitably deteriorates with turn depth. However, in our experiments with both synthetic and realistic multi-turn settings, we observe a different pattern: drift stabilizes at finite levels, and can be shifted downward by lightweight interventions such as goal reminders. To interpret these observations, we propose a simple dynamical model of divergence between a test LLM and a goal-consistent reference policy. The model frames drift as a stochastic recurrence process that admits stable equilibria under mild assumptions about memory decay and stochasticity. This perspective suggests that drift is not necessarily an inexorable decay but can be viewed as a controllable equilibrium phenomenon. Our contributions in this work can be summarized as follows:

- We measure temporal divergence between test LLMs and a reference policy in both controlled synthetic rewriting tasks and for LLM-based user simulators in τ-Bench, providing one of the first systematic analyses of drift trajectories.
- We propose a simple stochastic process model that explains why drift stabilizes, and how interventions shift the equilibrium level. Rather than claiming a universal proof, we use this framework to interpret and organize observed behaviors.
- Across tasks and models, we show that targeted reminders reduce equilibrium divergence and improve alignment quality, in line with the framework's predictions.

*Work done as part of internship at Adobe Research

Personalization in the Era of Large Foundation Models Workshop

2 Related Works

A persistent challenge in multi-turn dialogue with LLMs is context drift, the gradual degradation or distortion of the conversational state the model uses to generate its responses. Context drift is distinct from alignment drift: the former refers to loss or corruption of relevant information in the active context, while the latter describes a deviation from intended behavioral policies or values.

Context degradation in multi-turn interactions: A growing body of work has identified that large language models can suffer performance loss in extended conversations. (Laban et al. 2025) show that model outputs gradually deviate from earlier context, often leading to incoherence or goal neglect. (Abdelnabi et al. 2024) measure "task drift" by tracking changes in model activations over turns and propose detection mechanisms to flag when models are likely to have lost the original task. (Mehri et al. 2025) examine goal consistency over long-horizon dialogue with user simulators and call it "instruction drift", highlighting that even capable models struggle to sustain alignment as conversations deepen. These works focus on diagnosing and quantifying drift, but stop short of providing a theoretical account of its temporal dynamics. In contrast, we propose a simple dynamical perspective that models drift as a *bounded stochastic process* rather than as inevitable monotonic decay. Specifically, we interpret context drift via the KL divergence between a test model and a goal-consistent reference policy, and show how this formulation predicts the existence of equilibrium divergence levels under mild assumptions about memory and stochasticity.

Dynamical Systems Perspectives on LLM Interactions: Recent studies (Zhang and Dong 2024; Bhargava et al. 2023; Li et al. 2024a) have aimed to formalize LLM behavior through the lens of dynamical systems and control theory. Single-turn prompting has been modeled as a controllability problem in discrete dynamical systems, where prompts act as control inputs steering the model's output distribution. (Bhargava et al. 2023) treat transformer-based LLMs as discrete stochastic dynamical systems and analyze the controllability of self-attention, showing how short prompts can dramatically steer reachable outputs. (Zhang and Dong 2024) extend this perspective by modeling transformers via Neural ODEs and integrating robust control methods to stabilize outputs. Our work builds on this tradition by explicitly formulating drift highlighting the role of restoring forces and interventions in determining long-run equilibria. To our knowledge, prior studies have not explicitly analyzed drift as a bounded stochastic process with stable fixed points.

Memory and context management: Another strand of work attributes multi-turn failures to imperfect memory mechanisms. Studies on memory-augmented models (Wang et al. 2023a; Li et al. 2024b) and context compression (Jiang et al. 2023a,b) investigate ways of preserving salient information. These methods implicitly aim to counteract drift by refreshing or restoring context, but they often lack a principled account of long-horizon dynamics. Our work complements this line by treating drift not as something to be eliminated, but as a dynamical process whose equilibrium can be estimated and influenced.

3 Dynamics of Context in Multi-Turn Interactions

We study a multi-turn interaction between a *test language model* (LM) and a *reference policy*, both exposed to the same evolving conversation history over $\mathcal{T}$ rounds. At each turn $t \in \{1, \ldots, \mathcal{T}\}$, the conversation history is denoted by $x_{<t} = (x_1, \ldots, x_{t-1})$. The **test model** produces:

$$q_t(y) = \mathcal{P}_\theta(y \mid x_{<t}),$$

while the **reference model** (e.g., a larger LM or human-verified policy) produces:

$$p_t(y) = \mathcal{P}^*(y \mid x_{<t}),$$

serving as a stable, high-quality proxy for goal-consistent behavior. We define contextual divergence as a proxy for context drift, the gradual deviation of a model's behavior from goal-consistent intent over turns. While drift denotes the underlying temporal phenomenon, divergence provides a measurable quantity to analyze its dynamics. We formalize *contextual divergence* from the reference at each turn t via:

$$D_t := D_{\mathrm{KL}}(q_t \,\|\, p_t),$$

where D_{KL} is the Kullback–Leibler divergence. A perfectly context-aligned model satisfies $D_t = 0$ for all t. Under conventional view, as context grows with t, D_t also grows monotonically with t due to memory limits, information loss, or compounding errors, implying inevitable degradation in context tracking. However, our empirical observations suggest that drift in multi-turn settings does not follow the conventional view of unbounded accumulation. Instead, the sequence of divergences D_t can be usefully viewed as the trajectory of a bounded dynamical process:

$$D_{t+1} \;=\; f(D_t, \eta_t) + \xi_t,$$

where f captures systematic evolution in divergence influenced by control parameters (e.g., prompting strategy, reminder frequency, retrieval mechanisms), η_t represents controllable inputs, and ξ_t denotes stochastic variability from decoding randomness or minor linguistic variation. Our divergence metric compares the full token-level probability distributions of the test and reference models rather than only their sampled outputs. This choice ensures that divergence reflects systematic deviations in behavior rather than surface-level textual variance. Importantly, D_t should be interpreted as a proxy for contextual drift, not as an absolute measure of semantic correctness. GPT-4.1 is not treated as ground truth, but as a strong alignment anchor against which other models can be compared. Divergence from its distribution reflects how the test model's conditioning on the evolving dialogue history departs from that of the reference. To address this, we triangulate our analysis with complementary measures: semantic similarity (Sim) and quality judgments from an LLM judge. Our objectives in this study are therefore to: (i) characterize f from empirical interaction traces, (ii) estimate the equilibrium divergence for different models and settings,

and (iii) examine interventions that can shift this equilibrium toward lower divergence. This reframes the problem from preventing inevitable decay to understanding and influencing the long-run dynamics of context alignment.

4 Modeling Drift Dynamics

We view contextual drift as the turn-by-turn divergence between a test model and a reference policy during a multi-turn interaction. For a perfectly aligned model $D_t = 0$ for all t. The conventional intuition is that D_t grows monotonically with conversation length due to memory limits and compounding errors. However, our experiments (Section 6) suggest that divergence instead fluctuates around *bounded equilibrium levels*. To capture this empirically observed pattern, we propose a simple recurrence model:

$$D_{t+1} = D_t + g_t(D_t) + \eta_t - \delta_t, \tag{1}$$

where:

- $g_t(D_t)$ models systematic bias from imperfect memory or representation,
- η_t is a bounded stochastic perturbation ($|\eta_t| \leq \epsilon$),
- $\delta_t \geq 0$ models the effect of corrective interventions such as reminders.

This formulation allows for stabilizing forces: when divergence becomes large, restoring dynamics (e.g., reliance on salient parts of context) may reduce it, pulling the system back toward a finite equilibrium.

4.1 Equilibrium Interpretation

We define a contextual equilibrium D^* as a fixed point of the process:

$$\mathbb{E}[D_{t+1} - D_t \mid D_t = D^*] = 0. \tag{2}$$

If g_t is monotone and noise is bounded, trajectories converge toward this equilibrium. Intuitively, D^* represents the long-run level of divergence sustained by the model under a given interaction protocol.

4.2 A Simple Bound

Under mild assumptions, we obtain the following bound:

$$|D_t - D^*| \leq \lambda^t |D_0 - D^*| + \frac{\epsilon - \bar{\delta}}{1 - \lambda}, \tag{3}$$

for some $0 < \lambda < 1$, where $\bar{\delta}$ is the average intervention strength.

This result should be read as an *interpretive bound*: it illustrates that

1. without interventions ($\delta_t = 0$), divergence settles near a noise-limited equilibrium, and
2. with sufficiently strong interventions ($\bar{\delta} > \epsilon$), the equilibrium level can be shifted downward.

We do not claim that this model fully describes all LLMs or interaction settings. Rather, it provides a *conceptual and mathematical lens* that is consistent with our empirical findings: drift stabilizes, and interventions alter the equilibrium.

Takeaway

Context drift in multi-turn interactions can be understood as a **bounded, controllable equilibrium process** rather than inevitable decay. The key challenge is estimating the equilibrium and designing minimal interventions to keep alignment near it.

5 Experimental Setup

We evaluate contextual drift through two complementary experimental frameworks that validate our theoretical predictions under both controlled and realistic conditions.

- **Synthetic Controllable Drift Task:** To provide precise validation of our bounded dynamics hypothesis, we introduce a novel synthetic task where drift can be measured objectively. Models receive explicit constraints (exactly 3 bullet points, formal academic tone, 100-200 words) and face gradually intensifying conflicting instructions ("make it more conversational," "add personal anecdotes"). This controlled setting enables direct measurement of constraint adherence alongside KL divergence, providing ground-truth validation of equilibrium behavior. We test three models (LLaMA-3.1-8B, LLaMA-3.1-70B, Qwen2-7B) across 8 turns with interventions at turns 3 and 6. See Table 5 for an example.

- **Multi-Turn Interactions:** We complement synthetic validation using the τ-*Bench* framework, which provides realistic goal-oriented conversational environments with explicit user goals and measurable success criteria. Our experiments cover user simulations for two domains: *retail* (product search and purchase) and *airline* (flight booking and itinerary changes), both requiring mixed-initiative dialogue, entity resolution, and tool usage. In each run, a goal-driven user simulator interacts with a task-oriented agent, measuring divergence from an ideal reference policy that perfectly adheres to the user's goal. See Figure 6 for the task setup and 7 for an example of drifting behavior in LLM-based user simulator for τ-Bench.

For both frameworks, we examine two conditions: (1) *free-running interaction*, capturing natural accumulation of divergence due to compounding errors and imperfect context retention; and (2) *intervention-controlled interaction*, where targeted interventions (goal reminders or context refreshes) are inserted at pre-specified turns to test our controllability predictions from Section 4. We log full dialogue histories, output distributions, and turn-wise contextual divergence D_t, enabling analysis of equilibrium trajectories and quantification of intervention effectiveness across model architectures and task complexity levels.

5.1 Reference policy definition.

In our experiments, we operationalize the *goal-consistent reference policy* as the predictive distribution of `GPT-4.1`, conditioned on the original user goal g_0 and the full interaction history. This choice is motivated by two factors. First, `GPT-4.1` is among the most capable publicly accessible