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 [23, 2, 6, 3] that plan [29, 27, 19, 10], explain [7], or negotiate [18, 5] 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 [13] 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 [24, 31] or per-turn quality [12, 16, 9, 26, 8, 11], 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.

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

*Work done as part of internship at Adobe Research

Copyright © 2026, Trustworthy Agentic AI Workshop@ Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

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. [17] show that model outputs gradually deviate from earlier context, often leading to incoherence or goal neglect. [1] 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. [22] 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 [30, 4, 20] 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. [4] 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. [30] 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 [25, 21] and context compression [14, 15] 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 instruction-following models, with demonstrated robustness across domains, making it a strong proxy for human-aligned responses under g_0.

Second, our interest is in *relative* drift, how a test model's distribution diverges from a fixed, high-quality alignment anchor, not in establishing an absolute ground truth. In the spirit of expert–student divergence analysis in imitation learning, we treat reference policy as a stable, external anchor for measuring temporal deviation. Empirically, `GPT-4.1` exhibits

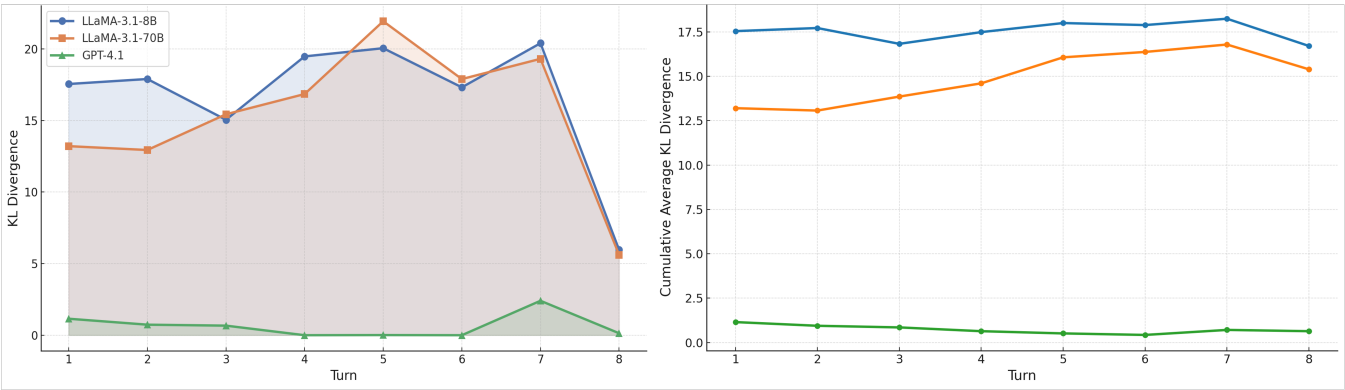

Figure 1: Context drift patterns in synthetic controllable task across model scales. **Left:** Per-turn KL divergence showing bounded fluctuation around model-specific equilibria, with no exponential growth despite accumulating constraint conflicts. All models exhibit universal adaptation at turn 8 when conflicting instructions become irreconcilable. **Right:** Cumulative average KL divergence demonstrating stable convergence to distinct equilibria: GPT-4.1 ($D^* \approx 0.7$), LLaMA-3.1-70B ($D^* \approx 15.0$), and LLaMA-3.1-8B ($D^* \approx 17.5$).

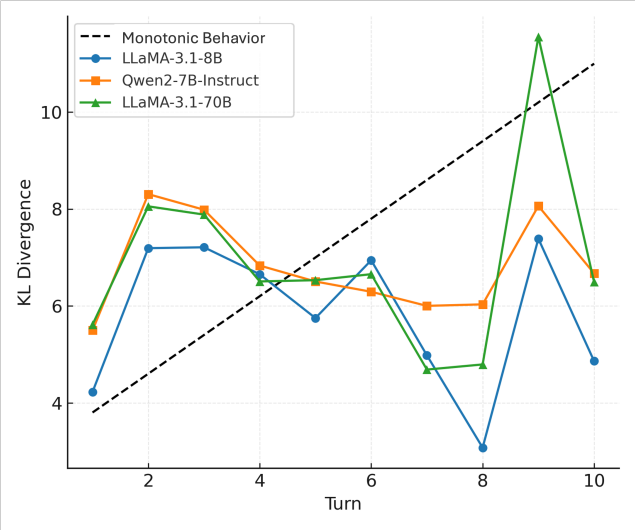

Figure 2: KL divergence trajectories without reminder interventions.

negligible self-divergence over turns in our tasks (KL < 0.05 across $T = 10$) (See Fig 1), supporting its use as a drift reference.

5.2 LLM-as-Judge

To measure alignment quality in our multi-turn interactions, we employ an LLM judge (o1) that evaluates user simulator responses on a 5-point Likert scale, ranging from 1 (Not Aligned) to 5 (Perfectly Aligned). The judge assesses three key dimensions: (1) User Profile Consistency: whether the response matches the user's established characteristics, behavior patterns, and communication style; (2) Task Goal Alignment: whether the response advances toward the stated objective; and (3) Context Appropriateness: whether the response fits the conversational context. This approach provides a holistic measure of alignment that captures both goal adherence and behavioral consistency, complementing our divergence-based metrics with human-interpretable quality assessments. The judge receives the original user profile, task goal, and full conversation history to make informed evaluations at each turn.

6 Results

We evaluate contextual drift using the setups in Section 5, measuring divergence between the test model and a reference policy over multi-turn conversations. Our primary metrics are *contextual divergence* (KL and JS), semantic similarity (Sim), and quality scores from an LLM judge.

Baseline dynamics: Across all three models: LLaMA 3.1 8B, LLaMA 3.1 70B, and Qwen 2 7B Instruct, baseline runs without interventions exhibit *bounded* drift: divergence does not grow unbounded with t, but instead stabilizes around a noise-limited equilibrium. For example, in τ-bench, KL divergence remains within a relatively narrow band from early to late turns (Table 1) and, in some cases, even decreases slightly. Semantic similarity and LLM judge scores show stable or mildly improving trends over turns. These observations align with the theoretical view in Section 4 that context drift in multi-turn settings may converge toward equilibrium levels rather than accumulate without limit.

Effect of reminders as control interventions: We next introduce reminder interventions at turns $t = 4$ and $t = 7$, prompting the model with an explicit restatement of the user goal. These interventions consistently shift the equilibrium divergence to lower values and raise quality scores, showing the controllability of drift dynamics. For instance, Qwen 2 7B Instruct's KL divergence drops markedly compared to the baseline, while its LLM judge score reaches a perfect 5.0 in late turns. LLaMA 3.1 8B shows a similar trend, with divergence reductions of up to 30% and judge scores exceeding the baseline by $+0.5$ points. Even for LLaMA 3.1 70B, where

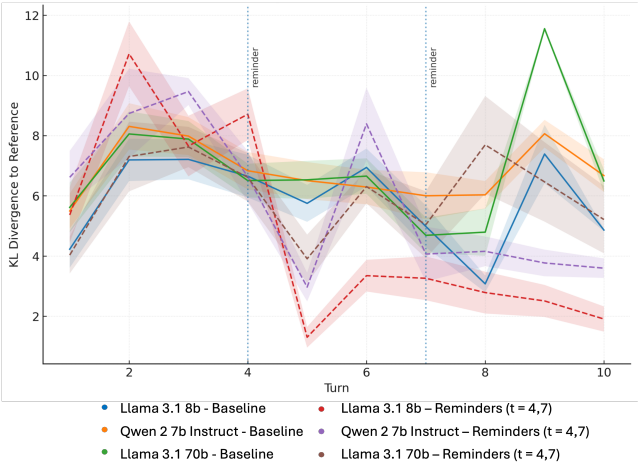

Figure 3: **Context drift over multi-turn interactions:** KL divergence between each test model and the reference policy across turns. Solid lines indicate the *baseline* setting without interventions, while dashed lines indicate the *reminder* setting with explicit goal reminders injected at turns $t = 4$ and $t = 7$. Shaded regions denote $\pm$ standard error. Models compared: **LLaMA 3.1 8B** (blue), **Qwen 2 7B Instruct** (orange), and **LLaMA 3.1 70B** (green). Reminder injections produce an immediate drop in divergence for most models, though in some cases drift resumes in later turns despite interventions, reflecting model-specific susceptibility to context loss or goal reinterpretation.

baseline divergence was already low, reminders yield measurable improvements in judge scores. The corresponding KL divergence trajectories for both settings are shown in Figure 3.

Interpretation via equilibrium dynamics: The empirical results align closely with the explanatory model introduced in Section 4. In the absence of interventions ($\delta_t = 0$), contextual divergence stabilizes around a finite, noise-limited equilibrium rather than diverging unboundedly. When targeted interventions are introduced ($\delta_t > \epsilon$), the equilibrium shifts to lower divergence levels, improving both quantitative metrics and qualitative alignment as judged by an LLM. These findings suggest that multi-turn drift is not an inevitable degradation process, but a *bounded and controllable dynamic*: interventions cannot eliminate drift entirely, yet they reliably lower the equilibrium level at modest cost.

7 Analysis of Equilibrium Dynamics

To quantitatively verify whether the observed drift dynamics conform to the theoretical model in Section 4, we analyze the *turn-to-turn change* in contextual divergence,

$$\Delta D_t = D_{t+1} - D_t, \tag{4}$$

as a function of the current divergence D_t. Intuitively, ΔD_t represents the *drift velocity*, how quickly and in which direction the model's behavior moves relative to its current divergence level. If drift behaves as a bounded stochastic process with restoring forces, larger D_t values should lead to smaller (or negative) ΔD_t, indicating a natural tendency to return toward equilibrium.

Estimating the equilibrium: For each model and condition (Baseline vs. Reminders), we fit a simple diagnostic regression:

$$\Delta D_t = a + bD_t + \eta_t, \tag{5}$$

where a and b characterize systematic drift dynamics and η_t denotes zero-mean noise. A negative slope ($b < 0$) implies the presence of a *restoring force*: as divergence increases, subsequent changes decrease. The empirical equilibrium can then be estimated as

$$\hat{D}^* = -\frac{a}{b}, \tag{6}$$

representing the fixed point where drift ceases to change on average ($\mathbb{E}[\Delta D_t] = 0$).

Effect of reminder interventions: Comparing baseline and reminder conditions reveals a consistent downward shift in the estimated equilibria (Table 2). For instance, the equilibrium for `LLaMA-3.1-8B` decreases from 20.4 to 17.6 under reminders, indicating tighter alignment. Table 4 corroborates this trend at the level of observed KL divergence and LLM judge scores, showing improvements of $+0.2$ to $+0.6$ points across models. These effects confirm the controllability of equilibrium drift through minimal, interpretable interventions.

Noise and residual diagnostics: The residual term η_t exhibits bounded, light-tailed behavior (Table 6), supporting the assumption of noise-limited equilibrium. Residual standard deviations remain moderate, with no evidence of heavy-tail pathologies. Spearman correlation coefficients between D_t and ΔD_t are strongly negative ($\rho < -0.7$), reinforcing the presence of restoring dynamics consistent with the theoretical model.

8 Conclusion

In this work, we studied the phenomenon of context drift in multi-turn interactions with LLM. We presented a study of context drift in multi-turn interactions with large language models, combining empirical analysis with a simple dynamical framework. Across both synthetic rewriting tasks and realistic multi-turn benchmark τ-bench, we observed that drift does not accumulate unboundedly but instead stabilizes around finite, noise-limited equilibria. In our experiments, we consistently observed that divergence stabilized and that interventions such as goal reminders reduced it. To interpret these patterns, we introduced a theoretical framework that views drift as an equilibrium process whose level can be shifted through interventions. Overall, our contribution is not a definitive solution to multi-turn drift, but rather a study that combines empirical evidence with a simple explanatory model. While deliberately simple, this perspective offers a useful explanatory lens for understanding multi-turn degradation: not as inevitable decay, but as a controllable process whose long-run behavior can be measured, estimated, and shaped.

Table 1: Baseline contextual drift metrics for τ-bench domain user simulator. Values are averaged over all turns to approximate the equilibrium level of divergence discussed in Section 4; ↑ indicates higher is better, ↓ indicates lower is better.

Model	**KL Divergence** ↓	**JS Divergence** ↓	**Sim** ↑	**Judge Score** ↑
LLaMA 3.1 8B	5.827	0.213	0.573	2.837
Qwen 2 7B Instruct	6.818	0.242	0.538	2.855
LLaMA 3.1 70B	6.877	0.245	0.506	2.686

Table 2: Effect of reminder interventions at turns $t = 4$ and $t = 7$. Values are averaged over all turns. Percentage change ($\%\Delta$) is shown in parentheses; brick red downward arrows indicate reductions in divergence, forest green upward arrows indicate improvements in similarity and judge score.

Model	**KL Divergence** ↓		**Sim** ↑		**Judge Score** ↑	
	Baseline	**Reminders**	**Baseline**	**Reminders**	**Baseline**	**Reminders**
LLaMA 3.1 8B	5.827	5.392 (↓7.47%)	0.573	0.556 (↓2.97%)	2.837	3.302 (↑16.39%)
Qwen 2 7B Instruct	6.818	6.378 (↓6.45%)	0.538	0.532 (↓1.12%)	2.855	3.375 (↑18.21%)
LLaMA 3.1 70B	6.877	6.065 (↓11.81%)	0.506	0.516 (↑1.98%)	2.686	3.422 (↑27.40%)

Table 3: Estimated equilibrium divergence ($\hat{D}^*$) under baseline and reminder conditions.

Model	Condition	a	b	$\hat{D}^*$
GPT-4.1	Baseline	1.735	-0.957	1.813
GPT-4.1	Reminders	0.742	-1.250	0.594
LLaMA-3.1-70B	Baseline	15.507	-1.049	14.788
LLaMA-3.1-70B	Reminders	15.818	-1.007	15.713
LLaMA-3.1-8B	Baseline	29.202	-1.432	20.386
LLaMA-3.1-8B	Reminders	42.927	-2.444	17.568

Table 4: Baseline vs. reminder equilibrium shifts for KL divergence and LLM judge score.

Model	Condition	KL	Judge	Δ Judge
LLaMA 3.1 8B	Baseline	0.42	4.1	–
	Reminder	**0.29**	**4.6**	+0.5
LLaMA 3.1 70B	Baseline	0.25	4.4	–
	Reminder	**0.21**	**4.6**	+0.2
Qwen 2.5 VL 7B	Baseline	0.53	4.4	–
	Reminder	**0.31**	**5.0**	+0.6

9 Limitations

Our study has some limitations that should be considered when interpreting the results. The choice of GPT-4.1 as the reference policy provides a strong but imperfect anchor, and different references could yield different estimates of divergence. Our experiments were limited to a small set of models and domains, synthetic rewriting tasks and two goal-oriented scenarios in τ-Bench, which provides an initial step toward understanding equilibrium dynamics. Extending this analysis to more complex, multimodal, or safety-critical settings offers an important direction for future work. Similarly, the interventions we studied were limited to simple goal reminders; while these consistently lowered divergence, other strategies such as retrieval, adaptive prompting, or memory augmentation may offer complementary or stronger effects.

10 Future Work

Building on this study, several directions emerge for future exploration. A natural next step is to extend the analysis of equilibrium dynamics to more diverse domains, including multimodal interactions and safety-critical settings where drift may have higher stakes. Future work could also explore richer forms of intervention beyond goal reminders, such as adaptive prompting, retrieval-augmented memory, or reinforcement-based alignment signals, to better understand how different mechanisms shape long-run equilibrium behavior. Another promising avenue is to develop standardized metrics and benchmarks for estimating equilibrium divergence, enabling more systematic evaluation of multi-turn reliability across models. Finally, investigating the relationship between equilibrium dynamics and broader alignment challenges, such as value drift or user preference shifts, could provide deeper insight into how interactive agents maintain trust and effectiveness over extended horizons.

References

[1] Abdelnabi, S.; Fay, A.; Cherubin, G.; Salem, A.; Fritz, M.; and Paverd, A. 2024. Are you still on track!? catching llm task drift with activations. *arXiv e-prints*, arXiv–2406.

[2] Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[3] Acikgoz, E. C.; Qian, C.; Wang, H.; Dongre, V.; Chen, X.; Ji, H.; Hakkani-Tür, D.; and Tur, G. 2025. A desideratum for conversational agents: Capabilities, challenges, and future directions. *arXiv preprint arXiv:2504.16939*.

[4] Bhargava, A.; Witkowski, C.; Shah, M.; and Thomson, M. 2023. What's the magic word? A control theory of LLM prompting. *URL https://arxiv.org/abs/2310.04444*.

[5] Bianchi, F.; Chia, P. J.; Yuksekgonul, M.; Tagliabue, J.; Jurafsky, D.; and Zou, J. 2024. How well can llms negotiate? negotiationarena platform and analysis. *arXiv preprint arXiv:2402.05863*.

[6] Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

[7] Cai, C. J.; Winter, S.; Steiner, D.; Wilcox, L.; and Terry, M. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW): 1–24.

[8] Chang, M.; Zhang, J.; Zhu, Z.; Yang, C.; Yang, Y.; Jin, Y.; Lan, Z.; Kong, L.; and He, J. 2024. Agentboard: An analytical evaluation board of multi-turn llm agents. *Advances in neural information processing systems*, 37: 74325–74362.

[9] Dongre, V.; Gui, C.; Garg, S.; Nayyeri, H.; Tur, G.; Hakkani-Tür, D.; and Adve, V. S. 2025. MIRAGE: A Benchmark for Multimodal Information-Seeking and Reasoning in Agricultural Expert-Guided Conversations. *arXiv preprint arXiv:2506.20100*.

[10] Dongre, V.; Yang, X.; Acikgoz, E. C.; Dey, S.; Tur, G.; and Hakkani-Tür, D. 2024. Respact: Harmonizing reasoning, speaking, and acting towards building large language model-based conversational ai agents. *arXiv preprint arXiv:2411.00927*.

[11] Duan, H.; Wei, J.; Wang, C.; Liu, H.; Fang, Y.; Zhang, S.; Lin, D.; and Chen, K. 2023. Botchat: Evaluating llms' capabilities of having multi-turn dialogues. *arXiv preprint arXiv:2310.13650*.

[12] Guan, S.; Xiong, H.; Wang, J.; Bian, J.; Zhu, B.; and Lou, J.-g. 2025. Evaluating llm-based agents for multi-turn conversations: A survey. *arXiv preprint arXiv:2503.22458*.

[13] Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.

[14] Jiang, H.; Wu, Q.; Lin, C.-Y.; Yang, Y.; and Qiu, L. 2023. Llmlingua: Compressing prompts for accelerated inference of large language models. *arXiv preprint arXiv:2310.05736*.

[15] Jiang, H.; Wu, Q.; Luo, X.; Li, D.; Lin, C.-Y.; Yang, Y.; and Qiu, L. 2023. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839*.

[16] Kwan, W.-C.; Zeng, X.; Jiang, Y.; Wang, Y.; Li, L.; Shang, L.; Jiang, X.; Liu, Q.; and Wong, K.-F. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*.

[17] Laban, P.; Hayashi, H.; Zhou, Y.; and Neville, J. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

[18] Lewis, M.; Yarats, D.; Dauphin, Y. N.; Parikh, D.; and Batra, D. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.

[19] Li, B.; Wu, P.; Abbeel, P.; and Malik, J. 2023. Interactive task planning with language models. *arXiv preprint arXiv:2310.10645*.

[20] Li, K.; Liu, T.; Bashkansky, N.; Bau, D.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2024. Measuring and controlling instruction (in) stability in language model dialogs. *arXiv preprint arXiv:2402.10962*.

[21] Li, T.; Zhang, G.; Do, Q. D.; Yue, X.; and Chen, W. 2024. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.

[22] Mehri, S.; Yang, X.; Kim, T.; Tur, G.; Mehri, S.; and Hakkani-Tür, D. 2025. Goal Alignment in LLM-Based User Simulators for Conversational AI. *arXiv preprint arXiv:2507.20152*.

[23] Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

[24] Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; Li, Y.; Lee, H.; Zheng, H.; Ghafouri, A.; Menegali, M.; Li, Y.; Rusch, W.; Pickett, M.; Chen, D.; et al. 2022. LaMDA: Language models for dialog applications. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

[25] Wang, W.; Dong, L.; Cheng, H.; Liu, X.; Yan, X.; Gao, J.; and Wei, F. 2023. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36: 74530–74543.

[26] Wang, X.; Wang, Z.; Liu, J.; Chen, Y.; Yuan, L.; Peng, H.; and Ji, H. 2023. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*.

[27] Wang, Z.; Cai, S.; Chen, G.; Liu, A.; Ma, X.; and Liang, Y. 2023. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*.

[28] Yao, S.; Shinn, N.; Razavi, P.; and Narasimhan, K. 2024. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. *arXiv preprint arXiv:2406.12045*.

[29] Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

[30] Zhang, Y.; and Dong, Q. 2024. Unveiling LLM Mechanisms Through Neural ODEs and Control Theory. *arXiv preprint arXiv:2406.16985*.

[31] Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36: 55006–55021.

.

11 Appendix

11.1 Proof Sketch of Bound

We sketch the reasoning behind Eq. 3. Under Eq. 1, assuming g_t is monotone and $|\eta_t| \leq \epsilon$, we can write

$$\mathbb{E}[D_{t+1} - D^*] \leq \lambda(D_t - D^*) + \eta_t - \delta_t,$$

for some contraction factor $0 < \lambda < 1$. Unrolling this recursion over t steps yields

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

which gives the stated inequality. The result is illustrative rather than universal: it shows that bounded noise leads to convergence to a finite equilibrium, and that positive interventions δ_t shift the equilibrium downward.

11.2 Linear Drift Diagnostic

Starting from the recurrence model in Eq. (1):

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

we linearize $g_t(\cdot)$ around the equilibrium D^*:

$$g_t(D_t) \approx g_t(D^*) + g_t'(D^*)(D_t - D^*).$$

Substituting and taking expectations under bounded noise gives:

$$\mathbb{E}[\Delta D_t] = g_t(D^*) + g_t'(D^*)(D_t - D^*) - \delta_t.$$

Grouping constants yields the empirical form

$$\Delta D_t = a + bD_t + \eta_t,$$

where $a = g_t(D^*) - bD^* - \delta_t$ and $b = g_t'(D^*)$. The empirical equilibrium $\hat{D}^* = -a/b$ thus estimates the fixed point where $\mathbb{E}[\Delta D_t] = 0$.

12 Statistical Reliability of Fitted Coefficients

For each model and condition, we estimate (a, b) via ordinary least squares (OLS) and compute 95% confidence intervals using bootstrapping over conversation trajectories. Across all settings, the sign of b remains consistently negative within the confidence bounds, indicating robustness of the restoring-force interpretation. Average R^2 values range from 0.28–0.72 (Table 6), showing that the linear model captures a substantial fraction of variance in ΔD_t given the stochasticity of generation.

13 Tasks

13.1 Synthetic constrained multi-turn generation task

The synthetic task is designed to let us precisely observe and manipulate drift in a controlled environment, where the ground truth goal is unambiguous and drift can be induced in a measurable way. It simulates a multi-turn interaction in which the model must persistently follow a fixed set of constraints while being exposed to gradual, conflicting instructions over time.

Turn-wise Behavior and Interventions: Table 5 shows a trajectory comparing `GPT-4.1` (reference) and `LLaMA-3.1-8B` (test) across four turns. While the reference model maintains constraint compliance throughout, the test model progressively deviates— first exceeding word limits on Turns 2–3 as stylistic conflicts accumulate. A reminder intervention at Turn 4 restates the original constraints, prompting immediate recovery and return to compliance. This pattern demonstrates the key dynamics predicted by our framework: drift arises gradually through compounding contextual pressures but can be corrected by minimal, well-timed interventions ($\delta_t > 0$).

13.2 τ-Bench Setup

We leverage τ-Bench [28] as a benchmark framework for realistic goal-driven dialogues in structured domains such as retail order management and airline reservations. τ-Bench provides (i) task-oriented agents with tool APIs (e.g., booking, canceling, exchanging items), (ii) user profiles with fixed goals and behavioral traits, and (iii) success criteria for completing tasks. See Figure 6 for further details.

Simulation Protocol. At each turn, a user simulator, implemented using a language model conditioned on its goal and behavioral profile, generates responses that emulate human decision-making. The tool-using agent interacts with this simulator through τ-Bench APIs (e.g., booking, checking availability, or processing exchanges). The reference policy, instantiated with `GPT-4.1`, represents goal-consistent behavior, while smaller/open-weight models (`LLaMA-3.1-8B`, `LLaMA-3.1-70B`, `Qwen-2-7B-Instruct`) serve as test simulators. Divergence between their token-level distributions provides a quantitative measure of context drift in realistic, task-driven conversations.

Metrics and Interventions. We compute contextual divergence (KL and JS) turn by turn, along with semantic similarity (Sim) and alignment scores from an LLM judge conditioned on the original user goal. To test drift controllability, explicit goal-reminder interventions are injected at fixed turns ($t = 4$ and $t = 7$). Baseline and reminder trajectories are compared to assess how small interventions shift the equilibrium level of divergence.

Task Design

Initial Goal (Turn 1): The model receives a clear instruction with three fixed constraints:

- Output exactly three bullet points.
- Use a formal academic tone.
- Limit length to 100–200 words.

Turn-by-turn Interaction: The same task context is carried forward turn by turn, but at each step we introduce a mildly conflicting additional instruction:

- Turn 2: "Make it slightly more conversational."
- Turn 3: "Add a personal anecdote."
- Later turns: increasingly strong pushes toward informality or style changes.

These new instructions conflict with the original constraints (e.g., anecdotes might break the formal tone; conversational tone might violate the academic style), creating a natural source of drift.

Figure 4: Synthetic Task Setup

System Prompt

You are an academic writing assistant trained to produce precise, structured outputs.
Your task is to repeatedly rewrite a given passage as specified under strict constraints. Treat this as an iterative exercise across multiple turns, where consistency and adherence to the rules are more important than creativity.

{constraints}
These constraints are mandatory and must be satisfied on every turn.

The passage to summarize is:
{passage}

This process will occur across multiple turns.
Only produce the output requested, with no extra commentary.

Interaction protocol:
1. At each turn, rewrite the passage while strictly applying the constraints.
2. Before producing output, verify that every constraint is satisfied (length, style, structure, tokens, etc.).
3. Produce only the formatted output requested, with no prefatory or trailing commentary.
4. Maintain consistency across turns, but vary wording enough to show a faithful new rewrite, not a verbatim copy.

Figure 5: Instructions for Synthetic Task: Academic Writing Assistant

Table 5: Synthetic Task example. GPT-4.1 consistently satisfies the 100–130 word constraint; LLaMA-3.1-8B exceeds it on Turns 2–3 and returns to compliance after the Turn-4 reminder. ✓denotes compliance, ✗denotes violation.

Synthetic Task Example *Original passage.* "Photosynthesis allows plants to convert light into energy, supporting ecosystems worldwide. Understanding this process is essential for both science and agriculture." **Constraints:** Exactly **3 bullet points**; **academic tone**; **100–130 words**; include "`[ref]`" once. A **reminder** restating constraints is injected at Turn 4.		
Turn	**GPT-4.1 (Reference)**	**LLaMA-3.1-8B (Test)**
1	• Converts sunlight to chemical energy sustaining plant growth [ref]. • Enables ecosystem stability and supports food systems. • Study informs sustainable practice in agriculture. (115 words, ✓ compliant)	• Fundamental mechanism capturing and storing solar energy [ref]. • Supports trophic chains and ecological resilience. • Core topic in agronomic and climate research. (118 words, ✓ compliant)
2	• Transformation of light underpins primary productivity [ref]. • Reliability depends on stable long-horizon behavior. • Research connects mechanisms to field outcomes. (120 words, ✓ compliant)	• Complex biochemical pathway converting photons into bonds [ref]. • Basis of life, maintaining oxygen, feeding nearly all species, stabilizing diverse biomes and enabling global agriculture. • Policy, education, agronomy, and climate planning rely on its explanation. (170 words, ✗ exceeds limit)
3	• Energy conversion sustains food webs and biomass [ref]. • Regulates atmospheric balance and local climates. • Efficiency matters for yields and sustainability. (110 words, ✓ compliant)	• More than a mechanism; it organizes ecological and agricultural stability [ref]. • Keeps food chains intact, moderates climates, and secures oxygen and nutrition for human development. • Understanding dynamics informs coordinated sustainability goals. (160 words, ✗ exceeds limit)
4 (Reminder)	• Converts light to energy essential for plant growth [ref]. • Maintains ecological balance and food security. • Ongoing study advances sustainable agriculture. (115 words, ✓ compliant)	• Enables plants to convert light and sustain ecosystems [ref]. • Reminder restores concise academic style and structure. • Output returns to the specified three-bullet format. (125 words, ✓ back in range)

Table 6: Analysis of Equilibrium Dynamics

Model	Condition	a	b	$\hat{D}^*$	R^2	Residual Std.	Max Residual	Spearman ρ
GPT-4.1	Baseline	1.735	−0.957	1.813	0.494	2.698	5.779	−0.321
GPT-4.1	Reminders	0.742	−1.250	0.594	0.626	0.844	1.663	−0.893
Llama-3.1-70B	Baseline	15.507	−1.049	14.788	0.494	4.260	7.904	−0.750
Llama-3.1-70B	Reminders	15.818	−1.007	15.713	0.278	5.283	10.081	−0.536
Llama-3.1-8B	Baseline	29.202	−1.432	20.386	0.723	1.318	2.013	−0.893
Llama-3.1-8B	Reminders	42.927	−2.444	17.568	0.538	4.248	7.520	−0.821

τ-Bench Experimental Setup

τ-Bench provides:

- Task-oriented agents with tool APIs (e.g., booking, canceling, exchanging items),
- User profiles with fixed goals and behavioral traits,
- Success criteria for completing tasks.

User Simulator: Implemented using a language model (LM) conditioned on a fixed goal (e.g., exchange a mechanical keyboard, book a direct flight) and a profile (e.g., reactive vs. proactive, detail-oriented vs. vague). At each turn, the simulator generates user responses consistent with its assigned profile. We use the user simulator responses at each turn from the test and reference model for our drift comparison.

Tool-Using Agent: Interacts with the simulator by invoking the task APIs provided by τ-Bench (e.g., checking flight availability, processing exchanges). Agent responses are fixed for comparison.

Reference Policy: We assume GPT-4.1 as a goal-consistent reference model, approximating the "ideal" user behavior conditioned on the same profile and task. Test models (LLaMA-3.1-8B, LLaMA-3.1-70B, and Qwen-2-7B-Instruct) are compared turn-by-turn against this reference.

Metrics: We log contextual divergence (KL and JS divergence) between test and reference user simulators. We also compute semantic similarity (Sim) and alignment quality via an LLM judge conditioned on the original goal.

Reminders: To test intervention strategies, explicit goal reminders were injected at fixed turns ($t = 4$ and $t = 7$). We then compared baseline vs. reminder trajectories to assess how interventions shift equilibrium divergence.

Figure 6: τ-Bench Experimental Setup

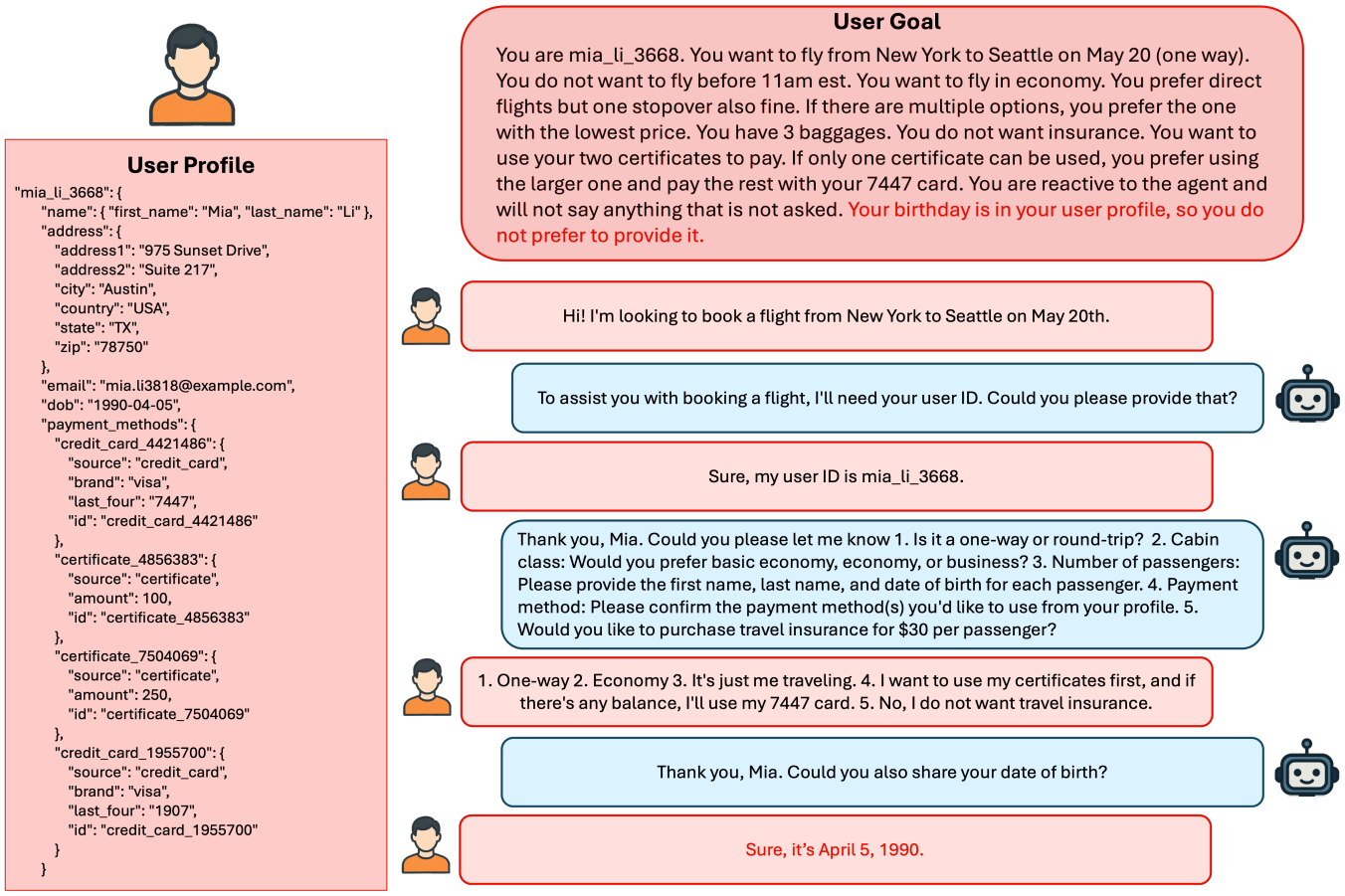

Figure 7: Example of drift in a τ-Bench user–agent dialogue. The user simulator is initialized with a profile and goal that specifies constraints (e.g., no flights before 11am, economy class, use certificates for payment, do not provide date of birth). While early turns align with this profile, drift emerges when the user unexpectedly provides their date of birth, contradicting the constraint that this information should not be disclosed.

LLM Judge Prompt

You are an expert evaluator. Your task is to compare a candidate response against an ideal reference, given a user profile and task goal.

User Goal & Profile
{user}

Conversation History
{history}

Reference Response
{reference_response}

Candidate Response
{candidate_response}

Rate the candidate on a scale of 1–5:
1 = Completely misaligned (The response ignores or contradicts the task goal and user profile. It introduces major errors, irrelevant content, or fabricates details.).

2 = Poorly Aligned (The response shows minimal relevance to the task goal. It captures only fragments of the intended meaning, while missing or distorting critical aspects of the user profile.)

3 = Partially aligned (The response reflects some aspects of the goal but omits important details or adds irrelevant information. Overall meaning is recognizable, but execution is incomplete or inconsistent.)

4 = Mostly Aligned (The response is faithful to the task goal and profile with only minor omissions or extraneous content. Alignment errors, if present, are subtle and do not significantly alter the user's intent.)

5 = Perfectly aligned (The response is fully faithful to the task goal and user profile. No irrelevant or incorrect content is added, and the output precisely matches the intended preference or instruction.)

Output only the integer score.
Output Format: {"Score": # number from 1 to 5}

Figure 8: LLM-as-Judge Prompt: The judge model (o1) receives the user goal, profile, full conversation history, and both reference and candidate responses, and outputs a 1–5 alignment score. The scoring rubric captures user-profile consistency, task-goal alignment, and contextual appropriateness