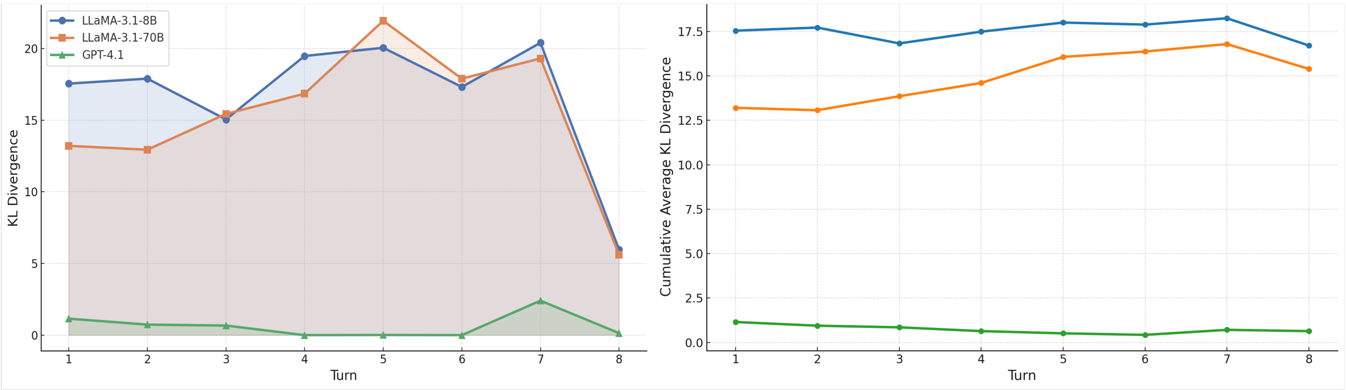

Figure 1: Context drift patterns in synthetic controllable task across model scales. **Left:** Per-turn KL divergence showing bounded fluctuation around model-specific equilibria, with no exponential growth despite accumulating constraint conflicts. All models exhibit universal adaptation at turn 8 when 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 ($D^* \approx 17.5$).

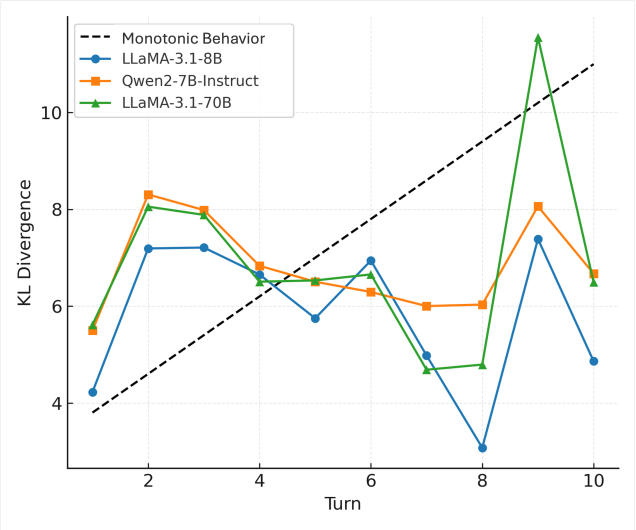

Figure 2: KL divergence trajectories without reminder interventions.

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 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.

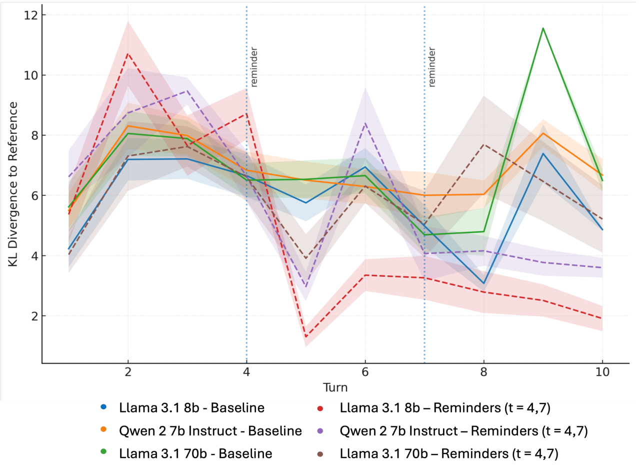

Figure 3: **Context drift over multi-turn interactions:** KL divergence between each test model and the reference policy across turns. Solid lines indicate the *baseline* setting without interventions, while dashed lines indicate the *reminder* setting with explicit goal reminders injected at turns $t = 4$ and $t = 7$. Shaded regions denote ± 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.

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 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

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

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.

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.