References

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.

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*.

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*.

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*.

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*.

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.

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.

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.

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*.

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*.

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*.

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*.

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.

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

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

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*.

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

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*.

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

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

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

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*.

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.

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)*.

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

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

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

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*.

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)*.

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

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 (Yao et al. 2024) 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.