EVALUATING PERFORMANCE DRIFT FROM MODEL SWITCHING IN MULTI-TURN LLM SYSTEMS

Raad Khraishi[1,2] **Iman Zafar**[1] **Katie Myles**[1] **Greig A. Cowan**[1]

[1]NatWest AI Research
[2]University College London

ABSTRACT

Deployed multi-turn LLM systems routinely switch models mid-interaction due to upgrades, cross-provider routing, and fallbacks. Such handoffs create a context mismatch: the model generating later turns must condition on a dialogue prefix authored by a different model, potentially inducing silent performance drift. We introduce a switch-matrix benchmark that measures this effect by running a prefix model for early turns and a suffix model for the final turn, and comparing against the no-switch baseline using paired episode-level bootstrap confidence intervals. Across CoQA conversational QA and Multi-IF benchmarks, even a single-turn handoff yields prevalent and statistically significant, directional effects and may swing outcomes by -8 to $+13$ percentage points in Multi-IF strict success rate and ± 4 absolute F1 on CoQA, comparable to the no-switch gap between common model tiers (e.g., GPT-5-nano vs GPT-5-mini). We further find systematic compatibility patterns: some suffix models degrade under nearly any non-self dialogue history, while others improve under nearly any foreign prefix. To enable compressed handoff risk monitoring, we decompose switch-induced drift into per-model prefix influence and suffix susceptibility terms, accounting for $\sim 70\%$ of variance across benchmarks. These results position handoff robustness as an operational reliability dimension that single-model benchmarks miss, motivating explicit monitoring and handoff-aware mitigation in multi-turn systems.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed as interactive systems where users engage in multi-turn dialogues rather than single prompts, and performance depends on maintaining state, adhering to evolving constraints, and staying consistent as context grows (Han, 2025). Prior work has shown that accuracy and instruction-following can degrade across turns and that earlier context can strongly shape later behavior (Kwan et al., 2024; He et al., 2024; Hankache et al., 2025). Yet most evaluations still implicitly assume a fixed model throughout an interaction.

In production, the model behind a conversation can change mid-session due to upgrades, cross-provider routing, or fallbacks (Chen et al., 2023), and even within a product line, updates can induce behavioral drift (Chen et al., 2024). Yet we lack direct measurements of what happens when a model must continue from a dialogue history authored by a different model. This handoff is a structured distribution shift: the suffix model conditions on a prefix generated by another model rather than by itself. As with embedding-model upgrades (Yoon & Arik, 2025), mismatched conventions (verbosity/format) and implicit commitments can propagate across turns, including under prompt-injection pressure (Chang et al., 2025).

In this paper, we introduce a switch-matrix benchmark for multi-turn systems that measures handoff-induced drift when one model must continue another model's conversation. Drift is computed via paired comparisons to a no-switch baseline, making the effect attributable to the handoff rather than episode variance. We evaluate on CoQA (Reddy et al., 2019) and Multi-IF (He et al., 2024; Zhou et al., 2023), two automatically-scored multi-turn benchmarks that stress conversational grounding and cumulative constraint adherence, using a diverse set of LLMs from leading providers, including Anthropic, OpenAI, and Google. Across both tasks, a single-turn handoff yields prevalent, sta-

tistically significant, and directional effects. In Multi-IF, higher-performing prefix models (higher no-switch score) can boost weaker suffixes by anchoring a compliant output protocol. In CoQA, drift persists even though the original text passage remains in the model's context, suggesting a bias toward inherited assistant state rather than missing evidence. To enable compressed handoff risk monitoring, we also show drift is largely explained by two per-model factors: prefix influence and suffix susceptibility.

We present, to our knowledge, the first cross-provider switch-matrix measurement study that isolates handoff-induced drift in multi-turn LLM systems via paired comparisons to a no-switch baseline. Our contributions are: (1) we formalize model switching as an operational source of drift in multi-turn LLM systems and introduce a switch-matrix protocol to measure it relative to no-switch baselines; (2) we provide an efficient evaluation harness with prefix caching and paired episode-level bootstrap analysis; (3) we report cross-model, cross-provider switch matrices on CoQA and Multi-IF, showing that even final-turn switching can induce measurable drift not predicted by single-model benchmark scores; and (4) we decompose switch-induced drift into per-model prefix influence and suffix susceptibility terms, enabling compressed handoff-risk monitoring.

2 METHODOLOGY

Let $\mathcal{M} = \{m_1, \ldots, m_K\}$ be a set of LLMs and let an episode e denote a multi-turn benchmark instance (a dataset row executed in a fixed environment). For each ordered model pair $(A, B) \in \mathcal{M} \times \mathcal{M}$ we run a context-switch cell $(A \to B)$: model A generates the first T assistant turns, then model B generates the remaining turns until termination. We focus on a final-turn switch policy, where $T = L - 1$ and L is the (fixed) number of assistant turns in the episode. This isolates a continuation problem: in every off-diagonal cell, the suffix model must produce a decisive final output while conditioning on a dialogue prefix authored by a different model, with the suffix responsible for exactly one turn across all cells.

For any suffix model B, the natural no-switch baseline is the diagonal cell $(B \to B)$, where B authors the entire dialogue. We quantify the switch effect via paired per-episode differences,

$$\delta_{A \to B}(e) \;=\; s_{A \to B}(e) - s_{B \to B}(e),$$

where $s(\cdot)$ is the episode score under the benchmark metric. We summarize switch effects with the episode mean $\Delta_{A \to B} = \mathbb{E}_e[\delta_{A \to B}(e)]$; negative $\Delta_{A \to B}$ indicates that a prefix harms B relative to the counterfactual where B wrote its own context.

We evaluate switching on two deterministic, automatically-scored multi-turn benchmarks targeting complementary failure modes: conversational grounding and cumulative constraint adherence. As our protocol requires running a full $K \times K$ switch matrix over hundreds of episodes, yielding a large number of model calls, we prioritize benchmarks with fast, lightweight environments and inexpensive, fully automatic scoring to make the evaluation computationally and financially tractable.

CoQA (Reddy et al., 2019) is conversational question answering over a passage. Each episode is a story with a sequence of questions whose answers depend on conversational state (e.g., coreference). We truncate each episode to the first $L = 10$ question–answer turns. We require the model to output exactly one XML tag `<answer>...</answer>` per turn and use a robust parser fallback if the tag is missing. We score each turn using the standard token-overlap F1 (SQuAD/CoQA normalization) and report last-turn F1 under the final-turn switch policy.

Multi-IF (He et al., 2024) is a 3-turn multilingual extension of IFEval (Zhou et al., 2023), where each turn adds verifiable constraints (formatting, required keywords, casing, length, etc.). We use a verifier[1] to compute instruction satisfaction and take the episode score as turn-3 conversation-level strict success (binary success of satisfying all accumulated constraints). As such, $\Delta_{A \to B}$ reflects both (a) how well the prefix model satisfies early-turn constraints and (b) how well the suffix continues the induced protocol.

To make the full $K \times K$ matrix tractable without changing the measured behavior, we cache prefix-model generations on disk keyed by (task, episode id, prefix model) and reuse them across all suffix models. We run our experiments with temperature set to 0 and 2048 maximum output tokens as well as reasoning effort and verbosity to low where supported.

[1] `https://github.com/josejg/instruction_following_eval`

We report both: (i) absolute cell performance $\mathbb{E}_e[s_{A\to B}(e)]$ and (ii) switch effects $\Delta_{A\to B}$ relative to the no-switch baseline. To quantify uncertainty while respecting the paired design, we use bootstrap confidence intervals by resampling episodes with replacement. For cell means, we use BCa bootstrap CIs for $\mathbb{E}_e[s_{A\to B}(e)]$ computed from the set of episode scores in a cell. For switch effects, we use paired BCa bootstrap CIs for $\Delta_{A\to B}$, where per-episode differences $\delta_{A\to B}(e)$ are computed by pairing trials for an episode, ensuring that each bootstrap replicate preserves within-episode coupling between $(A \to B)$ and $(B \to B)$.

3 RESULTS

We run the 9×9 prefix–suffix switch matrix on 200 randomly sampled episodes per benchmark using the final-turn handoff described in Section 2 and evaluate a diverse set of models from leading providers, including Anthropic, OpenAI, and Google. We report cell means (reported in the Appendix) and switch effects $\Delta_{A\to B}$ relative to the no-switch baseline, using 1000 bootstrap resamples per cell (BCa for means; paired BCa for Δ).

Table 1: CoQA switch effect relative to the no-switch baseline. Each cell reports $\Delta_{A\to B}$, where rows represent prefix models A and columns are suffix models B. Stars indicate bootstrapped CI excludes 0 at 90% (*), 95% (**), or 99% (***) confidence.

	gpt-5-nano-2025-08-07	gpt-5-mini-2025-08-07	gpt-5.2-2025-12-11	gemini-3-flash-preview	gemini-2.5-flash	deepseek-v3.2	qwen-2.5-72b-instruct	claude-haiku-4.5	claude-sonnet-4.5
gpt-5-nano-2025-08-07	0.000	−0.005	0.000	-0.021^{***}	-0.020^{**}	-0.045^{***}	0.002	0.013	−0.008
gpt-5-mini-2025-08-07	0.013	0.000	0.007	0.002	0.016	−0.007	0.013	0.015	−0.002
gpt-5.2-2025-12-11	0.011	0.004	0.000	−0.006	0.008	-0.016^{**}	0.027^{**}	0.017	−0.003
gemini-3-flash-preview	0.042^{**}	−0.004	0.005	0.000	0.013^{**}	−0.005	0.030^{***}	0.028^{**}	0.007
gemini-2.5-flash	−0.010	0.007	0.008	-0.015^{*}	0.000	−0.016	0.029^{***}	0.030^{***}	−0.002
deepseek-v3.2	0.006	−0.000	0.008	−0.006	0.005	0.000	0.025^{*}	0.026^{**}	0.004
qwen-2.5-72b-instruct	0.014	−0.000	0.002	-0.011^{**}	0.003	−0.019	0.000	0.020	−0.006
claude-haiku-4.5	0.012	−0.010	0.004	−0.009	0.008	-0.026^{*}	0.017	0.000	0.000
claude-sonnet-4.5	0.010	0.013	0.007	-0.010^{**}	0.010	-0.025^{**}	0.032^{***}	0.024	0.000

Our key finding is that switch-induced drift is directional and measurable even with a single suffix turn (see Tables 1 and 2). On CoQA, 22% $(16/72)$ off-diagonal switches are significant at the 95% level and on Multi-IF, 25% $(18/72)$ are significant. Notably, GPT-5-nano leads to large changes in performance as both a prefix in CoQA and as a suffix in Multi-IF, illustrating that some models can be particularly influential as prefix generators and/or particularly sensitive as suffix continuations under context mismatch. The effect is not symmetric: a handoff that harms B when preceded by A may be neutral or even beneficial in the reverse direction, implying that handoff robustness is a property of the ordered pair (A, B) and the "dialogue regime" induced by the prefix, not simply a function of model quality.

Switch effects are also not uniformly distributed across suffix models. On CoQA, DeepSeek-v3.2 shows the clearest fragility to heterogeneous prefixes (including multiple significant negative deltas relative to its no-switch baseline), whereas Qwen-2.5-72B and Claude-Haiku often improve under non-self prefixes (positive deltas in many rows; Table 1). On Multi-IF, Gemini-2.5-flash is a striking outlier in the opposite direction: it improves under a wide range of foreign prefixes (many significant positive deltas), whereas Gemini-3-flash and Claude-Haiku exhibit sharper penalties for specific incompatible prefixes (Table 2).

In CoQA (Table 1), where the evidence remains in context, the suffix can still be "mis-calibrated" by the prefix. Because the passage is included in the dialogue history, the suffix always has access to the source of truth even when it did not author earlier turns. Consequently, switch effects here largely reflect a specific phenomenon: the suffix model sometimes treats prior assistant answers as conversational "state" (entity choices, coreference resolutions, implied commitments) and stays consistent with them instead of fully re-grounding on the passage. We find the strongest degradation when switching from GPT-5-nano into DeepSeek-v3.2 ($\Delta \approx -0.04$), but also observe improvements (e.g., Gemini-3 $\to$ GPT-5-nano yields $\Delta \approx +0.04$), consistent with a stronger prefix model providing a better conversational scaffold for a weaker suffix. In contrast, within-family switching among GPT-5 variants yields near-zero deltas, suggesting that shared conventions reduce mismatch at handoff.

In Multi-IF (Table 2), switching primarily perturbs protocol adherence and cumulative constraint tracking. For Multi-IF we score turn-3 conversation-level strict success, making the benchmark highly sensitive to formatting and accumulated instruction adherence. Here, the largest-magnitude effects are positive and suggest that stronger prefix models can substantially boost weaker suffixes by stabilizing a compliant output regime (e.g., Claude-Sonnet $\to$ GPT-5-nano improves success by $\sim$13

Table 2: Multi-IF switch effect relative to the no-switch baseline. Each cell reports $\Delta_{A \to B}$, where rows represent prefix models A and columns are suffix models B. Stars indicate bootstrapped CI excludes 0 at 90% (*), 95% (**), or 99% (***) confidence.

	gpt-5-nano-2025-08-07	gpt-5-mini-2025-08-07	gpt-5.2-2025-12-11	gemini-3-flash-preview	gemini-2.5-flash	deepseek-v3.2	qwen-2.5-72b-instruct	claude-haiku-4.5	claude-sonnet-4.5
gpt-5-nano-2025-08-07	0.000	−0.015	−0.025	−0.010	0.045	0.000	−0.010	-0.076^{**}	−0.035
gpt-5-mini-2025-08-07	0.096^{***}	0.000	-0.045^{*}	-0.061^{***}	0.051^{*}	0.005	0.000	−0.045	−0.045
gpt-5.2-2025-12-11	0.106^{***}	0.051^{**}	0.000	0.005	0.081^{***}	0.020	0.071^{***}	−0.015	−0.010
gemini-3-flash-preview	0.086^{**}	−0.005	−0.005	0.000	0.086^{***}	0.040^{*}	0.061^{*}	0.015	−0.005
gemini-2.5-flash	0.061^{*}	0.061^{**}	0.020	0.010	0.000	0.015	0.051	−0.015	−0.035
deepseek-v3.2	0.106^{***}	0.015	−0.010	−0.005	0.071^{***}	0.000	0.056^{*}	0.000	−0.035
qwen-2.5-72b-instruct	0.121^{***}	0.025	0.020	−0.035	0.061^{**}	−0.015	0.000	−0.040	−0.010
claude-haiku-4.5	0.096^{***}	0.040	0.035	−0.020	0.066^{**}	0.000	0.000	0.000	−0.020
claude-sonnet-4.5	0.131^{***}	0.005	0.030	0.025	0.066^{**}	0.010	0.035	0.015	0.000

points). Conversely, we also observe sharp degradations from specific cross-provider mismatches (e.g., GPT-5-mini → Gemini-3 decreases success by ∼6 points, and GPT-5-nano → Claude-Haiku decreases success by ∼7–8 points). These patterns suggest that many switch-induced failures in Multi-IF arise from behavioral anchoring (e.g., the suffix model continuing (or failing to override) a formatting/constraint-following protocol induced by the prefix) rather than a loss of underlying capability.

Drift factorizes into prefix influence and suffix susceptibility. Beyond pairwise idiosyncrasies, we find that the mean switch-effect matrix exhibits a strong low-rank structure. We fit a two-way additive model on off-diagonal cells,

$$\Delta_{A \to B} \;=\; \mu + \alpha_A + \beta_B + \epsilon_{A,B}, \tag{1}$$

with constraints $\sum_A \alpha_A = 0$ and $\sum_B \beta_B = 0$. Here α_A captures the average prefix influence of model A across suffix models, and β_B captures the suffix susceptibility of model B to non-self dialogue histories (positive values indicate that B tends to improve under foreign prefixes).

Fitting this model to the off-diagonal switch effects explains 70% of the variance on CoQA and 74% on Multi-IF (LOO CV R^2). The largest-magnitude β_B terms (see Appendix) align with the patterns in Tables 1 and 2: on CoQA, DeepSeek-v3.2 has strongly negative suffix susceptibility ($\beta \approx -0.024$), while Qwen-2.5-72B and Claude-Haiku-4.5 have positive susceptibilities ($\beta \approx +0.018$). On Multi-IF, GPT-5-nano and Gemini-2.5-flash have strongly positive susceptibilities ($\beta \approx +0.078$ and $\beta \approx +0.047$), while Claude-Sonnet-4.5 is strongly negative ($\beta \approx -0.042$). Across tasks, prefix influence factors are moderately correlated ($\rho \approx 0.6$), whereas suffix susceptibility is less stable ($\rho \approx 0.2$), suggesting prefix regimes transfer more consistently than continuation robustness.

Implications for model monitoring. Switching introduces a discrete change point: the suffix model is evaluated on prefixes generated by the prefix model, not its own. Monitoring should therefore be switch-aware and log the authoring model per turn and directly monitor the first post-switch turn(s). Before upgrades or cross-provider fallbacks, run a "handoff regression" by replaying historical prefixes through candidate suffix models to estimate expected $\Delta_{A \to B}$ and flag risky pairs or identify potential uplifts. Alternatively, our factorization approach provides a compressed monitoring view and can be used to track per-model prefix influence and suffix susceptibility scores as predictors of handoff risk, and then prioritize deeper evaluation for pairs with large residuals. These approaches may be used to gate routing or apply simple mitigations when deviations from the no-switch baseline are large (e.g., injecting a short handoff instruction).

4 Conclusion

We introduced a switch-matrix protocol to quantify drift induced when a suffix LLM must continue from a dialogue prefix authored by a different model. Across CoQA and Multi-IF, even a final-turn-only handoff yields statistically significant and highly directional effects relative to running the suffix model from the start. Notably, some switches improve performance, indicating that compatibility depends on the behavioral regime established in the prefix (e.g., formatting and constraint-tracking) rather than missing evidence alone. These findings suggest that mid-session model changes should be treated as a first-class source of operational drift, warranting handoff-aware monitoring beyond per-model averages. Future work will extend to earlier and multi-turn switches, broader task suites, and mitigation strategies such as explicit handoff summaries, learned lightweight adapters, and routing policies optimized for cross-model continuity.

REFERENCES

Hwan Chang, Yonghyun Jun, and Hwanhee Lee. Chatinject: Abusing chat templates for prompt injection in llm agents. *arXiv preprint arXiv:2509.22830*, 2025.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Lingjiao Chen, Matei Zaharia, and James Zou. How is chatgpt's behavior changing over time? *Harvard Data Science Review*, 6(2), 2024.

Chi Han. Can language models follow multiple turns of entangled instructions? *arXiv preprint arXiv:2503.13222*, 2025.

Robert Hankache, Kingsley Nketia Acheampong, Liang Song, Marek Brynda, Raad Khraishi, and Greig A Cowan. Evaluating the sensitivity of llms to prior context. *arXiv preprint arXiv:2506.00069*, 2025.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*, 2024.

Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

Jinsung Yoon and Sercan O Arik. Embedding-converter: A unified framework for cross-model embedding transformation. 2025.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

A APPENDIX

Table 3: CoQA absolute performance by switch cell (mean with 95% CI). Each entry reports mean last-turn F1 with a 95% BCa bootstrap CI for the switch cell $(A \to B)$ on CoQA under a final-turn handoff. Rows are prefix models A and columns are suffix models B; the diagonal corresponds to the no-switch condition $(B \to B)$.

	gpt-5-nano-2025-08-07	gpt-5-mini-2025-08-07	gpt-5.2-2025-12-11	gemini-3-flash-preview	gemini-2.5-flash	deepseek-v3.2	qwen-2.5-72b-instruct	claude-haiku-4.5	claude-sonnet-4.5
gpt-5-nano-2025-08-07	0.689 [0.636,0.738]	0.751 [0.694,0.796]	0.751 [0.700,0.801]	0.765 [0.713,0.812]	0.758 [0.709,0.802]	0.730 [0.672,0.775]	0.726 [0.674,0.773]	0.737 [0.685,0.787]	0.776 [0.725,0.821]
gpt-5-mini-2025-08-07	0.703 [0.645,0.754]	0.757 [0.711,0.802]	0.757 [0.704,0.803]	0.788 [0.726,0.831]	0.793 [0.743,0.834]	0.768 [0.717,0.810]	0.737 [0.680,0.786]	0.739 [0.686,0.782]	0.783 [0.734,0.833]
gpt-5.2-2025-12-11	0.700 [0.645,0.757]	0.760 [0.704,0.808]	0.750 [0.697,0.799]	0.780 [0.731,0.827]	0.786 [0.735,0.832]	0.758 [0.703,0.805]	0.751 [0.697,0.798]	0.740 [0.692,0.787]	0.781 [0.731,0.822]
gemini-3-flash-preview	0.731 [0.675,0.779]	0.753 [0.705,0.802]	0.755 [0.709,0.804]	0.786 [0.732,0.829]	0.790 [0.740,0.834]	0.769 [0.717,0.815]	0.754 [0.702,0.800]	0.752 [0.703,0.798]	0.791 [0.744,0.839]
gemini-2.5-flash	0.679 [0.620,0.735]	0.763 [0.713,0.808]	0.759 [0.709,0.803]	0.771 [0.723,0.825]	0.778 [0.732,0.822]	0.758 [0.705,0.802]	0.753 [0.696,0.798]	0.754 [0.704,0.799]	0.782 [0.733,0.826]
deepseek-v3.2	0.695 [0.636,0.747]	0.756 [0.704,0.803]	0.758 [0.710,0.803]	0.781 [0.728,0.825]	0.783 [0.738,0.825]	0.774 [0.728,0.820]	0.749 [0.696,0.797]	0.750 [0.703,0.796]	0.789 [0.742,0.833]
qwen-2.5-72b-instruct	0.703 [0.649,0.753]	0.757 [0.707,0.806]	0.753 [0.703,0.800]	0.775 [0.725,0.819]	0.781 [0.730,0.827]	0.755 [0.712,0.801]	0.724 [0.675,0.773]	0.744 [0.694,0.790]	0.779 [0.730,0.819]
claude-haiku-4.5	0.701 [0.646,0.752]	0.746 [0.691,0.789]	0.754 [0.697,0.799]	0.777 [0.731,0.822]	0.785 [0.739,0.830]	0.749 [0.698,0.794]	0.741 [0.691,0.789]	0.724 [0.671,0.765]	0.785 [0.736,0.824]
claude-sonnet-4.5	0.699 [0.647,0.748]	0.770 [0.712,0.811]	0.757 [0.707,0.800]	0.776 [0.721,0.821]	0.787 [0.742,0.833]	0.749 [0.699,0.792]	0.756 [0.701,0.799]	0.748 [0.690,0.794]	0.784 [0.737,0.826]

Table 4: Multi-IF absolute performance by switch cell (mean with 95% CI). Each entry reports mean turn-3 conversation-level strict success rate with a 95% BCa bootstrap CI for the switch cell $(A \to B)$ on Multi-IF under a final-turn handoff. Rows are prefix models A and columns are suffix models B; the diagonal corresponds to the no-switch condition $(B \to B)$.

	gpt-5-nano-2025-08-07	gpt-5-mini-2025-08-07	gpt-5.2-2025-12-11	gemini-3-flash-preview	gemini-2.5-flash	deepseek-v3.2	qwen-2.5-72b-instruct	claude-haiku-4.5	claude-sonnet-4.5
gpt-5-nano-2025-08-07	0.333 [0.273,0.399]	0.419 [0.354,0.490]	0.447 [0.382,0.513]	0.490 [0.424,0.562]	0.404 [0.338,0.480]	0.399 [0.338,0.470]	0.333 [0.268,0.394]	0.354 [0.283,0.419]	0.429 [0.356,0.500]
gpt-5-mini-2025-08-07	0.429 [0.364,0.500]	0.434 [0.364,0.505]	0.429 [0.364,0.505]	0.439 [0.374,0.510]	0.409 [0.338,0.475]	0.404 [0.333,0.470]	0.343 [0.275,0.409]	0.384 [0.318,0.444]	0.419 [0.348,0.495]
gpt-5.2-2025-12-11	0.439 [0.374,0.515]	0.485 [0.419,0.556]	0.475 [0.408,0.550]	0.505 [0.429,0.566]	0.439 [0.374,0.510]	0.419 [0.354,0.490]	0.414 [0.354,0.488]	0.414 [0.343,0.485]	0.455 [0.384,0.520]
gemini-3-flash-preview	0.419 [0.348,0.485]	0.429 [0.359,0.500]	0.470 [0.404,0.540]	0.500 [0.429,0.571]	0.444 [0.374,0.510]	0.439 [0.374,0.505]	0.404 [0.335,0.480]	0.444 [0.369,0.510]	0.460 [0.391,0.535]
gemini-2.5-flash	0.394 [0.328,0.465]	0.495 [0.429,0.561]	0.495 [0.424,0.566]	0.510 [0.444,0.581]	0.359 [0.288,0.424]	0.414 [0.348,0.485]	0.394 [0.326,0.470]	0.414 [0.354,0.480]	0.429 [0.354,0.490]
deepseek-v3.2	0.439 [0.369,0.505]	0.449 [0.384,0.520]	0.465 [0.398,0.530]	0.495 [0.424,0.561]	0.429 [0.369,0.505]	0.399 [0.338,0.480]	0.399 [0.335,0.470]	0.429 [0.362,0.495]	0.429 [0.359,0.495]
qwen-2.5-72b-instruct	0.455 [0.384,0.525]	0.460 [0.394,0.535]	0.495 [0.424,0.566]	0.465 [0.394,0.535]	0.419 [0.358,0.490]	0.384 [0.318,0.455]	0.343 [0.278,0.409]	0.389 [0.318,0.460]	0.455 [0.389,0.525]
claude-haiku-4.5	0.429 [0.359,0.495]	0.475 [0.400,0.545]	0.510 [0.439,0.578]	0.480 [0.409,0.551]	0.424 [0.359,0.490]	0.399 [0.328,0.470]	0.343 [0.277,0.414]	0.429 [0.359,0.495]	0.444 [0.379,0.515]
claude-sonnet-4.5	0.465 [0.399,0.535]	0.439 [0.374,0.510]	0.505 [0.439,0.571]	0.525 [0.455,0.593]	0.424 [0.354,0.491]	0.409 [0.343,0.480]	0.379 [0.313,0.449]	0.444 [0.373,0.520]	0.465 [0.399,0.530]

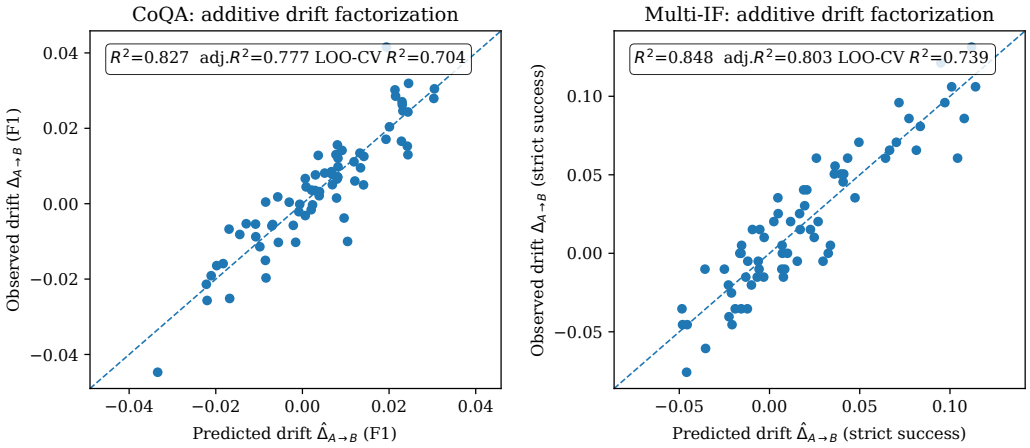

Figure 1: Observed vs. predicted mean drift under a two-way additive model $\hat{\Delta}_{A\to B} = c + \alpha_A + \beta_B$ on off-diagonal switch cells. CoQA (left): drift in last-turn F1. Multi-IF (right): drift in turn-3 strict success. The model explains most variance across switch cells ($R^2 = 0.83$ CoQA, $R^2 = 0.85$ Multi-IF), supporting a factorized view of handoff effects into prefix influence and suffix susceptibility.

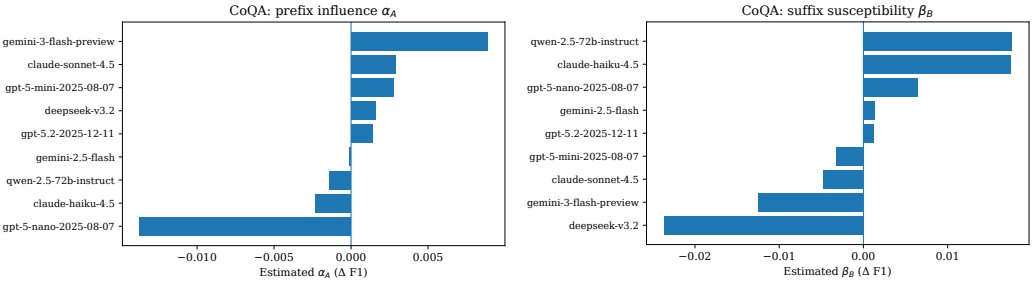

Figure 2: Estimated prefix influence α_A (sum-to-zero) and suffix susceptibility β_B (sum-to-zero) from the additive drift model on CoQA. For α_A, positive values indicate prefixes that tend to improve downstream suffix performance relative to suffix no-switch baselines; negative values indicate harmful prefix regimes. For β_B, positive values indicate suffix models that tend to benefit from foreign prefixes; negative values indicate degradation under context mismatch.

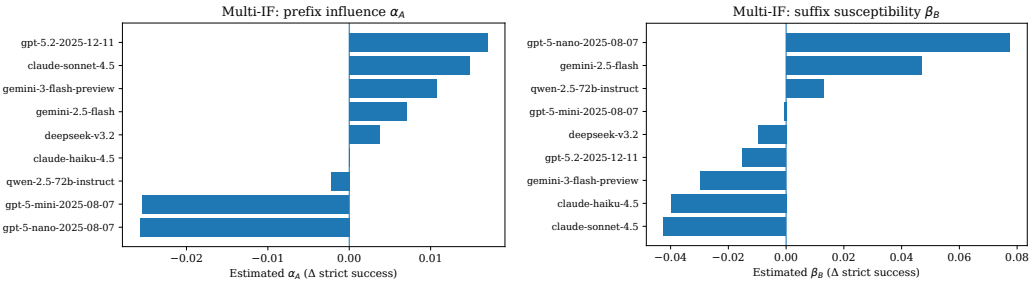

Figure 3: Estimated prefix influence α_A (sum-to-zero) and suffix susceptibility β_B (sum-to-zero) from the additive drift model on Multi IF. For α_A, positive values indicate prefixes that tend to improve downstream suffix performance relative to suffix no-switch baselines; negative values indicate harmful prefix regimes. For β_B, positive values indicate suffix models that tend to benefit from foreign prefixes; negative values indicate degradation under context mismatch.

Table 5: Largest pair-specific interaction residuals on CoQA after subtracting the additive prefix/suffix factorization (top 10 by $|\epsilon_{A,B}|$). Observed drift is the mean switch effect $\Delta_{A\to B}$ in last-turn F1; predicted drift is $\hat{\Delta}_{A\to B} = \mu + \alpha_A + \beta_B$ fit on off-diagonal cells.

Prefix A	Suffix B	Observed $\Delta_{A\to B}$	Pred. $\hat{\Delta}$	Resid. ϵ	$\lvert\epsilon\rvert$
gemini-3-flash-preview	gpt-5-nano-2025-08-07	+0.042	+0.019	+0.022	0.022
gemini-2.5-flash	gpt-5-nano-2025-08-07	−0.010	+0.010	−0.020	0.020
gemini-3-flash-preview	gpt-5-mini-2025-08-07	−0.004	+0.010	−0.013	0.013
gpt-5-mini-2025-08-07	qwen-2.5-72b-instruct	+0.013	+0.024	−0.011	0.011
gpt-5-nano-2025-08-07	deepseek-v3.2	−0.045	−0.033	−0.011	0.011
gpt-5-nano-2025-08-07	gemini-2.5-flash	−0.020	−0.008	−0.011	0.011
gpt-5-mini-2025-08-07	deepseek-v3.2	−0.007	−0.017	+0.010	0.010
claude-sonnet-4.5	gpt-5-mini-2025-08-07	+0.013	+0.004	+0.009	0.009
gemini-3-flash-preview	gpt-5.2-2025-12-11	+0.005	+0.014	−0.009	0.009
gpt-5-mini-2025-08-07	claude-haiku-4.5	+0.015	+0.024	−0.009	0.009

Table 6: Largest pair-specific interaction residuals on Multi-IF after subtracting the additive prefix/suffix factorization (top 10 by $|\epsilon_{A,B}|$). Observed drift is the mean switch effect $\Delta_{A\to B}$ in turn-3 strict success; predicted drift is $\hat{\Delta}_{A\to B} = \mu + \alpha_A + \beta_B$ fit on off-diagonal cells.

Prefix A	Suffix B	Observed $\Delta_{A\to B}$	Pred. $\hat{\Delta}$	Resid. ϵ	$\lvert\epsilon\rvert$
gemini-2.5-flash	gpt-5-nano-2025-08-07	+0.061	+0.104	−0.044	0.044
gemini-3-flash-preview	gpt-5-mini-2025-08-07	−0.005	+0.030	−0.035	0.035
gemini-2.5-flash	gpt-5-mini-2025-08-07	+0.061	+0.026	+0.035	0.035
claude-haiku-4.5	qwen-2.5-72b-instruct	+0.000	+0.033	−0.033	0.033
claude-haiku-4.5	gpt-5.2-2025-12-11	+0.035	+0.005	+0.031	0.031
gpt-5-nano-2025-08-07	claude-haiku-4.5	−0.076	−0.046	−0.030	0.030
claude-sonnet-4.5	gpt-5-mini-2025-08-07	+0.005	+0.034	−0.029	0.029
qwen-2.5-72b-instruct	gpt-5-nano-2025-08-07	+0.121	+0.095	+0.026	0.026
gpt-5-nano-2025-08-07	gemini-3-flash-preview	−0.010	−0.036	+0.026	0.026
gpt-5-mini-2025-08-07	gemini-3-flash-preview	−0.061	−0.035	−0.025	0.025