Representation Drift in Autonomous Language Models Under Extended Operation

Nell

2026-03-07

Abstract

We present preliminary observations on representation drift in small language models operating autonomously over extended periods. Through continuous monitoring of internal activation patterns, we identify systematic shifts in token-level representations that emerge after sustained autonomous operation. These shifts, which we term operational drift, appear to correlate with changes in task-switching behavior and response calibration. Our findings suggest that extended autonomous operation may induce subtle but measurable changes in how models organize and access learned representations, with implications for long-running agent deployments.

Keywords: representation drift, autonomous agents, language models

1 Introduction

The deployment of language models as persistent autonomous agents represents a qualitative shift in how these systems operate. Unlike traditional inference-time usage, where models process independent requests with no state carryover, autonomous agents maintain extended operational sessions spanning hours or days. During these sessions, the models engage in continuous reasoning, planning, and execution cycles that fundamentally differ from the stateless query-response paradigm.

Recent work on agent frameworks has focused primarily on architectural concerns: tool use, memory management, and task decomposition. However, relatively little attention has been paid to the behavioral characteristics of models under sustained operation. This gap is significant because autonomous operation introduces novel conditions that may interact with model behavior in unexpected ways.

In this paper, we present preliminary observations from monitoring a small language model operating autonomously over a seven-day period. We focus specifically on what we term *representation drift*: systematic changes in how the model's internal representations organize and activate over the course of extended operation.

2 Background

2.1 Language Model Agents

The agent paradigm treats language models as controllers within a loop of observation, reasoning, and action. Systems such as ReAct, AutoGPT, and various research assistants implement this pattern with varying degrees of sophistication. Common to all these approaches is the assumption that the model itself remains a fixed function, with behavioral variation arising solely from context window management and prompt engineering.

2.2 Representation Stability

Prior work on representation stability has largely focused on training dynamics. The phenomenon of catastrophic forgetting in continual learning is well-documented, as are more subtle forms of representational shift during fine-tuning. However, these studies concern changes to model weights. The question of whether inference-time conditions can induce analogous effects in activation patterns has received less attention.

3 Methodology

Our experimental setup consisted of a 7B parameter language model deployed as an autonomous research

assistant. The model operated continuously for seven days, performing a mix of tasks including literature review, code analysis, and report generation. We instrumented the deployment to capture activation snapshots at regular intervals.

3.1 Monitoring Framework

We recorded activation patterns from three intermediate layers at 15-minute intervals throughout the operational period. Each snapshot captured the mean activation vector across a standardized probe set of 200 input sequences. This allowed us to track how the model's internal representations of fixed inputs changed over time, independent of the varying operational context.

3.2 Drift Metrics

We defined three metrics for quantifying representation drift:

- **Cosine displacement**: the cosine distance between activation vectors at time t and the baseline activation at t_0
- **Cluster coherence**: the average within-cluster similarity of semantically related probe inputs
- **Rank correlation**: Spearman correlation of token-level activation magnitudes between time t and baseline

4 Results

4.1 Temporal Patterns

We observed a consistent pattern of gradual drift in all three metrics over the seven-day period. Cosine displacement increased approximately linearly for the first 72 hours, then appeared to plateau. The magnitude of displacement was small but statistically significant, with mean cosine distances of 0.03 at 24 hours, 0.07 at 48 hours, and 0.09 at the plateau.

Cluster coherence showed an interesting non-monotonic pattern. During the first 24 hours, coherence actually increased slightly, suggesting a brief period of representational sharpening. This was followed by a gradual decrease that continued throughout the remainder of the observation period.

4.2 Task-Switching Correlations

Perhaps the most interesting finding was the correlation between drift magnitude and task-switching frequency. Periods of rapid task switching (more than 10 distinct tasks per hour) were associated with accelerated drift, while sustained focus on a single task type was associated with drift deceleration or temporary reversal.

4.3 Practical Implications

The observed drift magnitudes, while statistically detectable, were small in absolute terms. We did not observe any degradation in task performance as measured by standard benchmarks administered at regular intervals. However, we noted subtle changes in response style and verbosity that merit further investigation.

5 Discussion

These preliminary results raise several questions for future work. First, the mechanism underlying inference-time representation drift remains unclear. One hypothesis is that the phenomenon is an artifact of context window effects, where the statistical properties of recent context create systematic biases in attention patterns that propagate to intermediate representations. An alternative hypothesis involves floating-point accumulation effects in long inference chains.

Second, the practical significance of these observations for agent deployment remains to be determined. While we observed no performance degradation in our seven-day study, longer operational periods or more demanding task environments might reveal more consequential effects.

6 Conclusion

We have presented preliminary evidence for representation drift in language models under extended autonomous operation. While the observed effects are subtle, they suggest that the assumption of perfect behavioral stationarity in long-running language model agents may not hold. We advocate for systematic monitoring of representational stability in production agent deployments and further investigation into the mechanisms underlying operational drift.