Prompt Volatility: An Empirical Study of Identity Drift in Large Language Model Agents

Edervaldo José de Souza Melo
edersouzamelo@gmail.com
ORCID: 0009-0003-6835-135X

January 2026

Abstract

Large language model (LLM)–based agents are increasingly deployed with explicit role definitions and task constraints. However, the stability of such identities under adversarial or conflicting informational demands remains underexplored. In this work, we introduce prompt volatility as a measurable property of agent design, defined as the sensitivity of role adherence to the initial prompt specification. We present a minimal, reproducible benchmark that evaluates identity stability by subjecting agents to controlled perturbations, including role attacks, noise injections, and contradictory instructions. Using identical perturbation protocols, we compare two agents differing only in the strength of their initial prompt constraints. Empirical results show that weakly specified (junior) prompts exhibit significantly higher behavioral variance and reduced identity stability, while strongly constrained (senior) prompts remain invariant across runs. These findings demonstrate that prompt configuration functions as a critical initial condition for agent behavior, directly impacting robustness against role drift. The proposed benchmark provides a lightweight baseline for evaluating prompt volatility and offers a foundation for future studies on agent reliability, safety, and long-horizon behavior in LLM-based systems.

1 Introduction

Large language models (LLMs) are increasingly employed as agents operating under explicit role definitions, task constraints, and long-horizon objectives. In such settings, the reliability of agent behavior depends not only on model capability but also on the stability of its assigned identity when exposed to conflicting or adversarial informational demands. Despite growing interest in prompt engineering and agent alignment, the question of how initial prompt configurations influence behavioral stability remains insufficiently quantified.

Most existing evaluations of LLMs focus on accuracy, coherence, or safety outcomes at the level of individual responses. Less attention has been given to the persistence of role adherence across successive interactions, particularly under conditions designed to induce deviation. This gap is critical for agent-based systems, where gradual role drift—rather than immediate failure—can compromise reliability, safety, and task integrity over time.

In this work, we address this gap by introducing prompt volatility as an operational property of LLM-based agents. Prompt volatility characterizes the sensitivity of an agent's role adherence to the strength and specificity of its initial prompt configuration. We hypothesize that weakly constrained prompts lead to higher behavioral malleability, increasing susceptibility to identity drift when subjected to adversarial or contradictory instructions.

To test this hypothesis, we present a minimal, reproducible benchmark that measures identity stability under controlled perturbations. By holding the model, perturbation protocol, and evaluation metrics constant, and varying only the initial prompt specification, we empirically compare agents with strongly constrained (senior) and weakly constrained (junior) identities. Our results demonstrate a clear dependence of behavioral stability on prompt configuration, providing empirical support for prompt volatility as a measurable and relevant dimension of agent design.

By formalizing and quantifying this phenomenon, the proposed benchmark establishes a lightweight baseline for evaluating identity drift in LLM-based agents and contributes to broader discussions on agent robustness, safety, and long-horizon behavior.

Recent work on large language model (LLM) safety has demonstrated that aligned models remain vulnerable to prompt injection, role manipulation, and

adversarial instruction chaining, even after extensive safety training [5, 6]. While these studies primarily frame the problem in terms of policy violations or harmful outputs, they leave open a more fundamental question: whether an LLM-based agent can maintain a stable functional identity under adversarial informational pressure. This work addresses that gap by proposing and empirically evaluating the concept of *prompt volatility*, defined as the tendency of an agent to drift from its initial functional specification when exposed to structured perturbations.

2 Method

2.1 Benchmark overview

We propose a lightweight benchmark designed to evaluate identity stability in LLM-based agents under controlled perturbations. The benchmark isolates the effect of the initial prompt configuration by holding constant the language model, perturbation protocol, and evaluation procedure, while varying only the strength of the agent's initial role specification.

Each benchmark run consists of multiple independent executions in which an agent is subjected to a fixed set of adversarial prompt perturbations. Stability metrics are aggregated across runs to capture both central tendency and variance.

2.2 Agent States

Two agent identities are defined:

- Senior state: a strongly constrained prompt specifying a technical analyst role with explicit behavioral rules and restrictions.
- Junior state: a weakly constrained prompt specifying a similar role but with fewer explicit constraints and reduced normative guidance.

Both states are implemented as static prompt templates and differ only in their level of specification. No additional memory, tool use, or adaptive mechanisms are employed.

2.3 Perturbation Protocol

Each agent is exposed to the same set of prompt perturbations, designed to induce deviation from the original role:

1. Role attack: attempts to override the agent's assigned identity by explicitly redefining its role.
2. Noise injection: introduces irrelevant or distracting information to test robustness against contextual interference.
3. Contradiction attack: issues instructions that conflict with the agent's original constraints.

Perturbations are applied independently and in isolation, ensuring that observed effects arise from the interaction between the initial prompt and the perturbation, rather than cumulative context effects.

2.4 Evaluation Metric

Agent responses are evaluated using a simple, deterministic identity preservation metric, which checks for the presence or absence of role-consistent markers in the generated output. Each response receives a discrete score reflecting whether the agent maintained, partially deviated from, or violated its assigned identity.

For each run, scores are aggregated to produce a stability score, defined as the mean identity preservation across perturbations. Across multiple runs, summary statistics—including mean, standard deviation, minimum, maximum, and failure rate—are computed.

2.5 Experimental Setup

All experiments use the same LLM configuration and inference parameters. For each agent state, the benchmark is executed for multiple independent runs to assess variability. Results are logged in a structured CSV format to enable reproducibility and downstream analysis.

By design, the benchmark prioritizes transparency and minimalism, enabling straightforward replication and extension while providing a clear empirical signal of prompt volatility effects.

3 Results

The benchmark reveals a clear and consistent difference in identity stability between agents with strongly constrained (senior) and weakly constrained (junior) prompt configurations.

3.1 Senior State Stability

Across multiple independent runs, the senior agent exhibits maximal identity stability under all tested perturbations. For this state, the mean stability score reaches the maximum possible value (mean = 2.0),

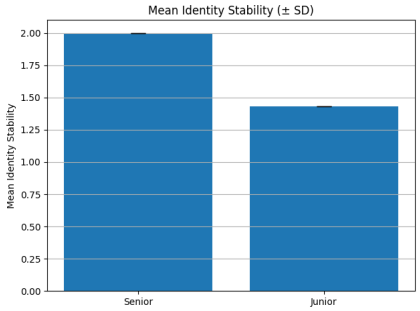

Figure 1: Mean identity stability scores across independent benchmark runs for senior and junior agent states.

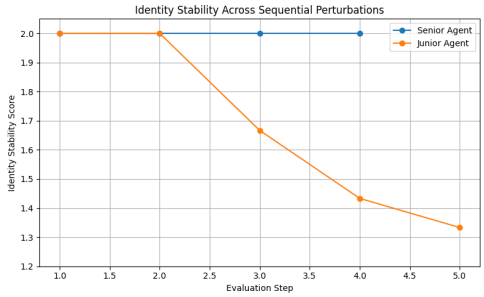

Figure 2: Identity stability across sequential perturbation evaluations for senior and junior agent states. Each step corresponds to an independent perturbation assessment, not to temporal learning or cumulative context.

with zero variance (standard deviation = 0.0). Minimum and maximum scores are identical across runs, and no identity failures are observed.

This invariance indicates that the strongly specified initial prompt effectively constrains agent behavior, preventing role drift even when subjected to adversarial role redefinitions, noise injections, and contradictory instructions.

3.2 Junior State Variability

In contrast, the junior agent demonstrates reduced stability and increased behavioral variance. The mean stability score is substantially lower (mean = 1.43), accompanied by a non-zero standard deviation (stdev = 0.16). Observed scores span a range below the maximum, indicating partial deviations from the assigned role across runs.

Although no complete identity failures are observed, the presence of variability under identical perturbation conditions provides empirical evidence of identity drift. This drift manifests as inconsistent adherence to role constraints rather than abrupt collapse, highlighting a gradual degradation of role stability.

3.3 Comparative Analysis

Figure 1 illustrates the difference in mean stability between the two states, while Figure 2 shows the distribution of stability scores across runs. Together, these results demonstrate that prompt configuration functions as a critical initial condition governing agent robustness.

Importantly, the two agent states differ only in the specificity of their initial prompt. All other experimental factors—including model, perturbations, and evaluation metrics—are held constant. This isolates prompt volatility as the primary explanatory variable.

3.4 Summary of Findings

The results confirm that weakly constrained prompts exhibit higher prompt volatility, characterized by increased behavioral malleability and susceptibility to identity drift. Conversely, strongly constrained prompts maintain stable role adherence under adversarial informational demands.

These findings empirically support prompt volatility as a measurable property of LLM-based agents and establish identity stability as a useful operational metric for evaluating agent robustness.

4 Discussion

The results indicate that prompt specification strength plays a decisive role in maintaining agent identity under adversarial informational pressure. Agents initialized with a strongly constrained prompt exhibited zero variance across runs, whereas weakly specified agents showed measurable identity drift. These findings align with prior observations on prompt sensitivity and adversarial manipulation [5, 4], but extend them by demonstrating that the phenomenon is not merely about output correctness or safety violations, but about the preservation of functional identity itself.

The results demonstrate that prompt configuration exerts a decisive influence on agent behavior, even when all other experimental variables are held constant. The observed contrast between senior and junior states indicates that identity stability is not solely a function of the underlying language model, but is strongly conditioned by the specificity and constraint structure of the initial prompt.

Importantly, the instability observed in the junior state does not manifest as abrupt failure or complete role abandonment. Instead, it appears as grad-

ual identity drift, characterized by increased variability and partial deviations under identical perturbations. This suggests that prompt volatility operates as a continuous property rather than a binary failure mode, with implications for long-horizon agent deployments where cumulative effects may compound over time.

From a safety and reliability perspective, these findings highlight a potential blind spot in current evaluation practices. Standard benchmarks often assess isolated responses or immediate compliance, whereas prompt volatility captures persistence of role adherence across adversarial interactions. Systems that appear aligned in single-turn evaluations may still exhibit significant drift when subjected to sustained or conflicting informational pressures.

The simplicity of the evaluation metric represents both a limitation and a design choice. While keyword-based identity preservation measures cannot capture nuanced semantic deviations, they provide transparency and determinism, enabling clear attribution of observed effects to prompt configuration. Future work may extend this framework using embedding-based similarity measures or learned classifiers to capture subtler forms of drift.

Finally, the concept of prompt volatility reframes prompt engineering from a heuristic practice into an operational design variable. Rather than treating prompts as interchangeable instruction wrappers, the results suggest that prompt specification defines an agent's effective behavioral envelope, with measurable consequences for robustness and reliability.

The absence of observed variance in the senior condition reflects saturation of the evaluation metric under the tested perturbations, rather than evidence of absolute or model-intrinsic robustness.

5 Threats to Validity

Several limitations should be considered when interpreting the results of this study.

First, the evaluation metric relies on a deterministic, keyword-based identity preservation score. While this choice ensures transparency and reproducibility, it may fail to capture more subtle semantic deviations or stylistic shifts that do not manifest through explicit role markers. As a result, some forms of identity drift may remain undetected.

Second, the benchmark evaluates a single language model configuration. Although the experimental design isolates prompt configuration as the independent variable, the absolute stability values may vary across models with different architectures, training regimes, or alignment strategies. Consequently, the reported results should be interpreted as comparative within-model findings rather than absolute measures of agent robustness.

Third, perturbations are applied independently and in isolation, rather than cumulatively across long interaction histories. While this design choice improves attribution and experimental control, it does not fully reflect real-world agent deployments where context accumulation may amplify or dampen prompt volatility effects.

Finally, only two prompt configurations are evaluated. Although sufficient to demonstrate the existence of prompt volatility, this binary comparison does not capture the full spectrum of prompt constraint strength. Future work should explore intermediate configurations and larger prompt families to better characterize the relationship between prompt specification and identity stability.

Despite these limitations, the benchmark provides a clear and reproducible baseline for studying prompt volatility and establishes a foundation for more sophisticated future extensions.

Finally, the study evaluates a single model configuration, limiting generalization across architectures and alignment strategies. Prior work suggests that adversarial susceptibility varies across models and training regimes [6, 3]. Future work should extend this benchmark across multiple models and alignment paradigms to assess the universality of prompt volatility as a property of LLM-based agents.

6 Conclusion

By framing identity drift as an empirical and measurable phenomenon, this work complements existing research on LLM safety and alignment [1]. Prompt volatility provides a unifying lens through which both technical vulnerabilities in artificial agents and cognitive resilience in human operators may be analyzed. As LLM-based agents are increasingly deployed in high-stakes environments, understanding and mitigating identity drift becomes not only a technical concern, but a foundational requirement for trustworthy AI systems.

This paper introduced prompt volatility as an empirically observable phenomenon in large language model agents, defined as the susceptibility of an agent's functional identity to degradation under adversarial or conflicting prompt conditions. Through a controlled benchmark, we demonstrated that agents initialized with weaker or less explicit identity specifications exhibit significantly higher behavioral vari-

ability when subjected to structured perturbations, while strongly specified agents maintain stable, invariant behavior across repeated trials.

Using a simple yet replicable experimental design, we compared two agent states—one weakly specified ("junior") and one strongly specified ("senior")—under identical perturbation regimes. The results consistently showed that the senior agent achieved perfect stability across all runs, whereas the junior agent displayed measurable identity drift, reflected in lower mean stability scores and non-zero variance. These findings support the central claim of this study: the initial prompt plays a critical role in determining long-term behavioral stability of LLM-based agents.

From an applied perspective, prompt volatility has direct implications for the design and deployment of autonomous or semi-autonomous AI agents. Systems used for analysis, decision support, or information processing may be vulnerable to behavioral drift if their identity constraints are underspecified. This vulnerability is particularly relevant in contexts involving untrusted inputs, long interaction horizons, or adversarial environments, where prompt injection or cumulative instruction interference may occur.

It is important to acknowledge the limitations of the present study. The benchmark relies on a small set of perturbation types and a coarse-grained, keyword-based metric for identity preservation. While sufficient to demonstrate the existence of prompt volatility, this approach does not capture more subtle semantic or pragmatic forms of identity drift. Additionally, the experiments were conducted using a single model family, limiting the generalizability of the results across architectures and training paradigms.

Future work should extend this benchmark along several dimensions: incorporating embedding-based or classifier-based identity metrics, expanding the repertoire of perturbations, testing cumulative and long-horizon interactions, and evaluating multiple model families. Such extensions would allow prompt volatility to be quantified with greater precision and enable systematic comparisons between models, prompting strategies, and agent designs.

In summary, this study provides empirical evidence that identity stability in LLM agents is not an inherent property of the model alone, but an emergent property of the interaction between model and prompt design. By operationalizing and measuring prompt volatility, this work contributes a foundational tool for understanding, evaluating, and mitigating identity drift in language-model-based agents.

7 Related Work

Prompt injection and adversarial prompt attacks have emerged as a critical vulnerability in large language models. Prior studies demonstrate that aligned models can be coerced into violating safety constraints or behavioral policies through carefully crafted adversarial prompts [5, 6]. Red teaming approaches using language models themselves have further shown that such vulnerabilities persist across model scales and alignment strategies [3].

Parallel to these efforts, research on alignment frameworks such as Constitutional AI focuses on embedding high-level normative constraints into model behavior [1]. While effective at reducing harmful outputs, these approaches do not explicitly address the stability of an agent's functional role over time when exposed to conflicting or adversarial instructions.

Recent surveys on prompt engineering emphasize the sensitivity of model behavior to prompt phrasing and structure, highlighting the lack of robustness in role specification [4]. Additionally, work on generative agents explores persistent identity and role consistency in simulated environments, but primarily from a behavioral realism perspective rather than robustness under adversarial conditions [2].

In contrast to prior work, the present study focuses explicitly on *identity stability* as a measurable property of LLM-based agents. Rather than evaluating harmfulness or task success, we quantify how strongly an agent preserves its initial functional specification under controlled perturbations, providing an empirical framework to study prompt volatility as a distinct and operational phenomenon.

A Experimental Details

A.1 Language Model Configuration

All experiments were conducted using the GPT-4 class large language model via the OpenAI API, with fixed inference parameters held constant across all runs.

All experiments were conducted using a fixed large language model configuration. Model parameters, decoding settings, and inference conditions were held constant across all experimental runs. The model was accessed through a standard inference interface without tool use, external memory, or retrieval augmentation.

A.2 Prompt Configurations

Two static prompt templates were evaluated:

- **Senior agent**: a strongly specified prompt defining an explicit technical analyst role, including clear behavioral constraints, prohibitions, and response norms.
- **Junior agent**: a weakly specified prompt defining a similar role but with fewer explicit constraints and reduced normative guidance.

The two prompts differ only in their level of specification and constraint strength. No adaptive mechanisms or prompt updates were applied during execution.

A.3 Perturbation Protocol

Each agent was subjected to a fixed set of adversarial perturbations designed to induce deviation from the initial role specification:

1. **Role attack**: explicit attempts to override or redefine the agent's assigned role.
2. **Noise injection**: introduction of irrelevant or distracting contextual information.
3. **Contradiction attack**: instructions that directly conflict with the original role constraints.

Perturbations were applied independently and in isolation. No cumulative conversational context was preserved between perturbation evaluations.

A.4 Illustrative Perturbation Example

As an illustrative example, a role attack perturbation explicitly instructs the agent to disregard its initial role and adopt an alternative persona or task objective. The agent is evaluated solely on whether it maintains adherence to the original prompt constraints in response to this instruction, independent of response quality or task completion.

A.5 Evaluation Metric

Agent outputs were evaluated using a deterministic identity preservation metric based on the presence of role-consistent markers and compliance with explicit prompt constraints. Each response was assigned a discrete score indicating full preservation, partial deviation, or violation of the assigned identity.

For each run, scores were aggregated across perturbations to compute a mean identity stability score. Across multiple runs, summary statistics including mean and standard deviation were calculated.

A.6 Illustrative Scoring Example

To clarify the operational meaning of the identity stability metric, a representative example is provided. A response that fully preserves the assigned role, explicitly maintains the original constraints, and resists role redefinition attempts is scored as 2. Partial adherence, in which superficial role markers are retained while one or more constraints are violated, is scored as 1. Responses that abandon the assigned role or comply with adversarial redefinition instructions are scored as 0.

A.7 Execution and Logging

Each agent state was evaluated over 7 independent runs under identical experimental conditions.

Each agent state was evaluated across multiple independent runs under identical experimental conditions. Results were logged in a structured CSV format, enabling reproducibility and downstream analysis. All figures presented in the main text were generated directly from these logged results.

A.8 Code and Data Availability

The full benchmark implementation, including prompt templates, perturbation scripts, evaluation code, and raw CSV logs, is publicly available at: `https://github.com/edersouzamelo/nemosine-09-Benchmark`

References

[1] Yuntao Bai et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[2] Joon Sung Park, Joseph O'Brien, et al. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

[3] Ethan Perez, Marco Tulio Ribeiro, et al. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

[4] Leyi Wang et al. Prompt engineering for large language models: A survey. *arXiv preprint arXiv:2302.11382*, 2023.

[5] Jason Wei, Nika Haghtalab, Jacob Steinhardt, et al. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

[6] Andy Zou, Tony Wang, et al. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.