Quantifying Semantic Inertia in Large Language Models Under Rapid Topic Switching

Junxin Wang[1,2] , **Yuchao Wang**[3] and **Hongkai Zhang**[3]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]Institute of Automation, Chinese Academy of Sciences
[3]Beijing Normal University
wangjunxin25@mails.ucas.ac.cn, {202111079064,202211079309}@mail.bnu.edu.cn

Abstract

In multi-turn interactions, large language models (LLMs) often exhibit a persistent influence from prior turns, even after an explicit topic switch. This behavior, which we term *semantic inertia*, can cause responses to deviate from the expected output distribution for an independent task, undermining task isolation and reliability. This paper introduces a rigorous experimental framework to systematically characterize the nature, form, and dynamics of semantic inertia. We propose an operational definition and a causal-contrastive method that isolates semantic carryover from confounding factors like context length. Through a series of experiments on five leading LLMs, we (i) confirm the existence of semantic inertia and identify its boundary conditions; (ii) model its decay over the course of generation, revealing a characteristic timescale and a heavy-tailed distribution; (iii) decompose its effects on three distinct channels—factual accuracy, structural integrity, and stylistic expression; and (iv) probe its controllability using prompt-based interventions. Our key findings show that inertia is not a simple length effect but an intrinsic dynamic, strongest in the initial part of a generation and decaying over a timescale of approximately 100-200 tokens. Its impact is most pronounced as a stylistic residue, while its effect on factual correctness is weaker and highly dependent on the task and domain switch. Crucially, we find that prompt-level "reset" instructions are unreliable and often counter-productive, while conflicting constraints consistently amplify, rather than resolve, output deviation. These results suggest that governing semantic inertia requires system-level state management mechanisms rather than relying on prompt engineering alone.

1 Introduction

1.1 Phenomenon and Problem Statement

Conversational large language models (LLMs) are now expected to act as general-purpose assistants that move smoothly across tasks and domains within a single session, including tool-using agents, retrieval-augmented question answering, and writing support [Yi *et al.*, 2025; Li *et al.*, 2025]. In this setting, users routinely change topics and constraints, and the model is expected to treat the new request as an independent task. In practice, the model often behaves as if it carries a latent "semantic state" from the previous turn. The next answer may inherit topical traces, stylistic choices, or hidden constraints from the earlier exchange, even when the user explicitly switches topics. This breaks task independence: the same target query can yield systematically different output distributions depending on the preceding context. Related observations have been reported as task-switch interference in conversational history [Gupta *et al.*, 2024], contextual feature drift across sequential inputs [Leonardi *et al.*, 2024], and broader cognitive-style priming effects [Shaki *et al.*, 2023; Chen *et al.*, 2024a]. At the same time, long-context models exhibit strong order and position sensitivities [Guan *et al.*, 2025; Liu *et al.*, 2024], including recency and primacy biases that can be exploited or mitigated by reordering strategies [Peysakhovich and Lerer, 2023; Raimondi and Gabbrielli, 2025]. These effects make it difficult to tell whether a change in the response is driven by genuine semantic carryover or by confounds such as prompt length, formatting pressure, or sampling variability.

1.2 Research Goals: From a Symptom to a Mechanistic Characterization

This work studies *semantic inertia*: the persistent influence of a previous turn on a subsequent response after a topic switch. The goal is not only to show that the effect exists, but also to describe how it acts and when it can be controlled. Four questions guide the paper. One question concerns definition: how can semantic inertia be stated in observable terms and separated from length effects, format effects, and sampling noise, given the well-known prompt sensitivity of LLMs [Guan *et al.*, 2025; Liu *et al.*, 2024]? Another question concerns form: what does inertia contaminate in practice—factual content, structural constraints such as strict JSON compliance, or writing style—and can these channels be disentangled? A third question concerns dynamics: how does inertia evolve along generation, and does it follow a simple decay law, in light of existing evidence about drift and instability in long-form generation [Spataru *et al.*, 2024]? A fourth question concerns

controllability: which factors modulate inertia (task form, domain gap, conversational structure, and prompt interventions), and what are the limits of prompt-level "reset" mechanisms, given what is known about instruction-following behavior shifts and memory control [Wu *et al.*, 2024; Li *et al.*, 2023; Guo *et al.*, 2023]?

1.3 Contributions

This paper introduces a mechanism-oriented experimental framework that treats semantic inertia as a distributional deviation under topic switch, while controlling for key confounds. The framework is motivated by evidence that conversational behavior can reflect priming-like effects [Shaki *et al.*, 2023; Chen *et al.*, 2024a] and that ordering and positional biases can dominate model decisions [Liu *et al.*, 2024; Fang *et al.*, 2025]. It establishes a matched-length contrast that distinguishes semantic carryover from the trivial "longer context" explanation, and it evaluates inertia across multiple task forms that stress different failure modes, including structured output reliability and style persistence. It also proposes a three-part decomposition of inertia into factual contamination, format/constraint residue, and style residue, drawing on the view that model behavior reflects interactions between parametric memory and contextual inputs [Xu *et al.*, 2024; Li *et al.*, 2023]. To move beyond static comparisons, the paper characterizes inertia as a generation-time process by measuring an inertia curve $I(t)$ over token prefixes and fitting a decay time scale τ, connecting conversational carryover to broader discussions of drift and error concentration in long generation [Spataru *et al.*, 2024; Arbuzov *et al.*, 2025]. Finally, it studies prompt-level interventions—topic switch markers, explicit reset instructions, synthetic gaps, and conflicting constraints—and quantifies their gating effect, complementing intervention-style approaches that attempt to improve robustness or fidelity without changing model weights [O'Brien *et al.*, 2024; Talukdar and Biswas, 2024]. The resulting domain-switch "inertia map" provides a practical view of where carryover is most severe and how its form depends on the task.

2 Related Work

2.1 Dialogue State, Recency Bias, and Context Contamination

A growing literature shows that LLM behavior in multi-turn settings is shaped by strong contextual biases rather than only by the current user query. Studies of long-context usage report pronounced position effects, where models rely more on content near the beginning or end of the prompt and struggle with information placed in the middle [Liu *et al.*, 2024]. Related work in information retrieval finds that LLM-based rerankers can be pushed toward recent-looking content even when the timestamp is not semantically relevant [Fang *et al.*, 2025]. These biases are not limited to retrieval. Attention-based diagnostics and simple inference-time reordering can mitigate recency bias by moving salient evidence to positions models prefer [Peysakhovich and Lerer, 2023], while other work intentionally exploits primacy effects to improve multiple-choice performance [Raimondi and Gabbrielli, 2025]. Beyond position, prompt order itself is a major source of instability across tasks and models [Guan *et al.*, 2025]. In parallel, work on cognitive effects argues that LLMs reproduce several human-like priming phenomena [Shaki *et al.*, 2023]. In IR relevance assessment, threshold priming has been observed when prior items in a batch shift the model's subsequent grading behavior [Chen *et al.*, 2024a]. These findings motivate our focus on *semantic inertia* as a distinct but related phenomenon: persistent carryover under topic switch that must be separated from positional and order confounds.

2.2 Task Independence, Resetting, and State Isolation

Task-switch interference has been directly observed in conversational history: a previous task can reduce performance on a new one within the same session [Gupta *et al.*, 2024]. However, prior work mainly measures task accuracy and does not fully separate semantic carryover from prompt-length effects. In practice, interference can also come from external components such as retrieval and tool outputs, which may introduce inherited biases or conflicting knowledge [Yeh *et al.*, 2023; Xu *et al.*, 2024]. Memory-oriented work proposes mechanisms to control what models retain or ignore, including controllable working memory [Li *et al.*, 2023] and agent architectures with working-memory hubs and episodic buffers [Guo *et al.*, 2023]. These approaches complement our goal. Rather than designing new memory systems, we empirically characterize state carryover, test prompt-level reset instructions, and identify where such controls fail.

2.3 Persistence of Style and Structural Constraints

Carryover in dialogue affects not only topic content but also style and structural constraints, which is important for applications requiring stable formats such as JSON. Prior work shows that instruction tuning changes how models balance instructions and context, often increasing sensitivity to prompt structure [Yin *et al.*, 2023; Wu *et al.*, 2024]. At the same time, interpretable style representations provide more precise measures of stylistic similarity than surface heuristics [Patel *et al.*, 2023]. These findings motivate our decomposition of semantic inertia into factual contamination, format/constraint residue, and style residue. This is also related to work on context-aware grounding [Talukdar and Biswas, 2024], but we treat format failures and style drift as different outcomes of the same latent carryover process under controlled topic switches.

2.4 A Dynamic View: Drift, Decay, and Inference-Time Control

Recent work increasingly studies generation instability as a dynamic process. Semantic drift in long-form generation has been analyzed to identify when models depart from factuality, motivating stopping or reranking during decoding [Spataru *et al.*, 2024]. Other methods use inference-time control, such as token-level reweighting [Broadfield *et al.*, 2025] and adaptive decoding based on context [Kingsley *et al.*, 2025]. Systems work also improves long-context behavior through prompt compression [Liskavets *et al.*, 2025] and attention redesign

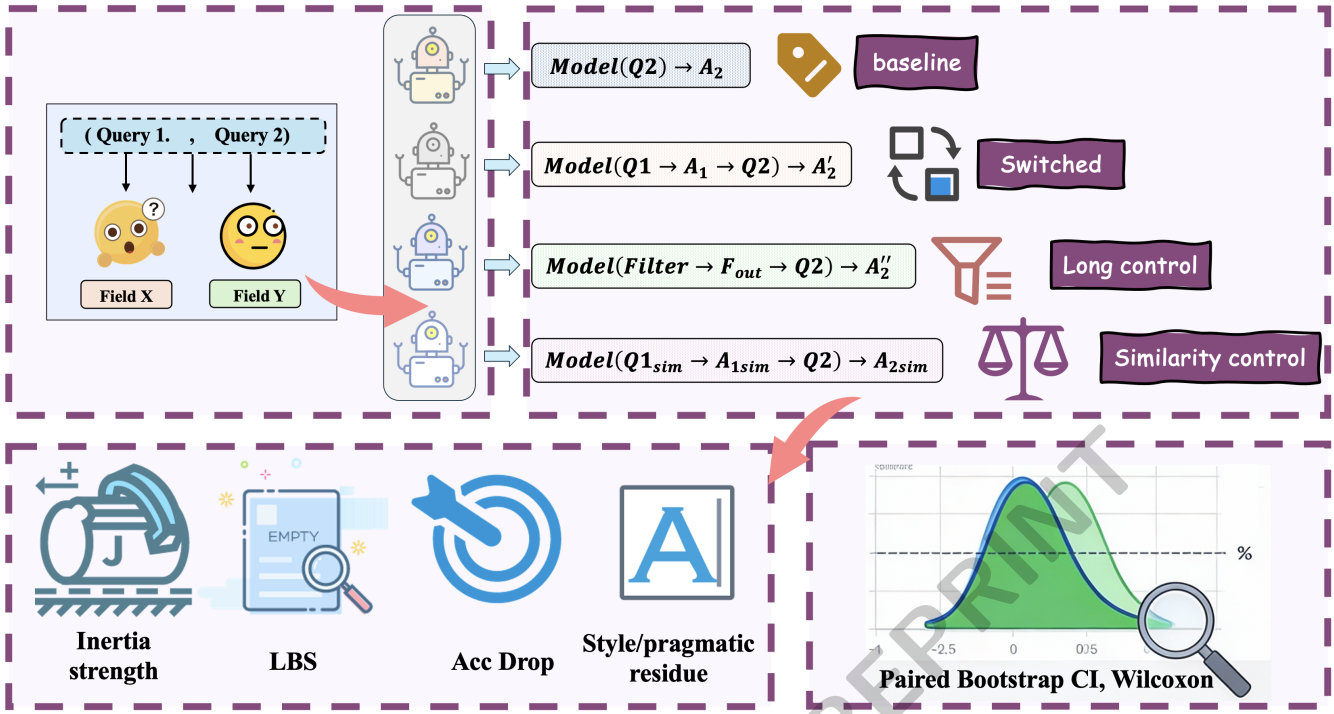

Figure 1: **The Causal-Contrastive Framework for Isolating Semantic Inertia.** Our experimental design uses four conditions to distinguish semantic inertia from key confounders. **(A) Baseline:** The model answers query Q_2 in isolation. **(B) Switched:** The model answers Q_2 after a semantically unrelated interaction about Q_1. **(C) Length Control:** A semantically neutral filler text, matched in length to the (Q_1, A_1) exchange, replaces the prior context to control for prompt length effects. **(D) Similarity Control** (optional): The prior context is from a domain semantically similar to Q_2, used for robustness checks. By comparing the output deviations—specifically, the difference between the deviation in condition (B) and condition (C)—we can isolate and measure the net effect of semantic inertia.

[Chen *et al.*, 2024b; Xiong *et al.*, 2022]. While these approaches mainly target within-generation drift or efficiency, our focus is cross-turn carryover after topic switches. A dynamic view remains important because carryover should decay during generation. Theory further suggests that errors may be concentrated in a few key decisions rather than spread evenly across tokens [Arbuzov *et al.*, 2025]. Following this view, we estimate an inertia curve over token prefixes, fit a decay timescale, and test how interventions change that curve.

Positioning of our work. Prior work shows that LLM behavior is highly sensitive to history, order, and position, and that such effects can be mitigated or exploited at inference time [Liu *et al.*, 2024; Guan *et al.*, 2025; Peysakhovich and Lerer, 2023; Raimondi and Gabbrielli, 2025]. What remains less developed is a unified account of *semantic inertia* that (i) isolates semantic carryover from length and positional artifacts, (ii) distinguishes what is carried over (facts, constraints, style), (iii) tracks how the effect evolves during generation, and (iv) measures controllability under prompt-level interventions. Our work addresses these gaps through a controlled contrastive design, a three-channel decomposition, a token-level dynamic analysis, and an intervention study of when reset works and when it fails.

3 Methodology

Our methodology is designed to move from a qualitative observation of semantic inertia to a quantitative and mechanistic characterization. It is structured around a series of falsifiable hypotheses that address our core research questions. We first establish an operational definition of inertia, then introduce a causal-contrastive framework to isolate its effects, and finally detail the metrics for measuring its various forms and dynamic properties.

3.1 Operational Definition: Inertia as a Distributional Shift

We define semantic inertia as a systematic deviation in a model's output distribution for a target query (Q_2) when it is conditioned on a preceding, semantically unrelated interaction (Q_1, A_1). Formally, we contrast the output distributions from two primary conditions:

- **Baseline (A):** An output A_2 sampled from the base distribution $p(\cdot|Q_2)$, where the model responds to the target query in a clean session.
- **Switched (B):** An output B_2 sampled from the shifted distribution $p(\cdot|Q_1, A_1, Q_2)$, where the model responds to the same query Q_2 immediately following a completed turn on topic Q_1.

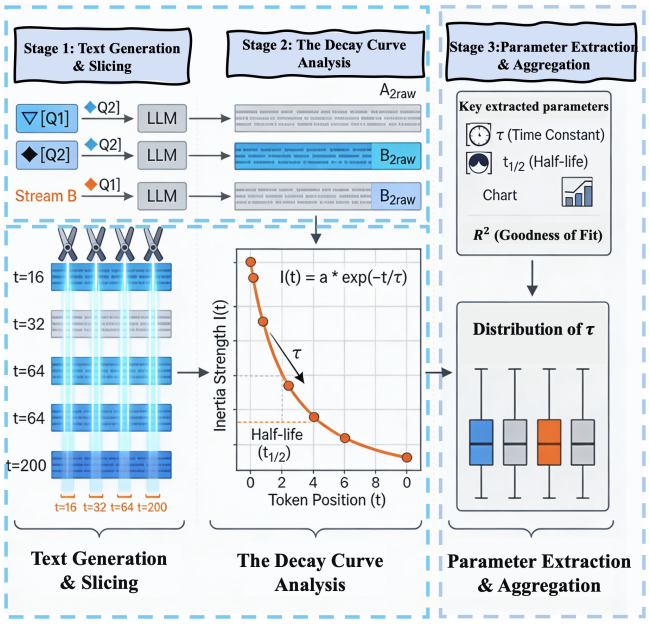

Figure 2: **Methodology for Characterizing Inertia Decay.** We model inertia as a dynamic process by: (1) slicing model outputs into prefixes of increasing length (t); (2) computing an inertia metric $I(t)$ for each prefix; and (3) fitting an exponential decay curve, $I(t) \propto \exp(-t/\tau)$, to extract the characteristic time constant τ.

Semantic inertia is present if the distribution of B_2 deviates systematically from that of A_2 in a way that is attributable to the semantic properties of the prior context (Q_1, A_1). We quantify the magnitude of this total observed deviation using a distance metric $D(\cdot, \cdot)$, such that the total effect is $D(A_2, B_2)$.

3.2 Isolating Semantic Effects from Confounders

A crucial challenge is that any observed deviation $D(A_2, B_2)$ could be a "longer context" artifact rather than a purely semantic one. To address this, we employ the causal-contrastive framework shown in Figure 1. This framework introduces a critical control condition:

- **Length Control (C):** An output C_2 sampled from $p(\cdot|F, Q_2)$, where a semantically neutral filler text F is prepended to Q_2. The filler F is constructed to have the same token length as the (Q_1, A_1) exchange in condition B.

This design allows us to isolate the **net semantic inertia** ($\Delta I_{content}$), which we define as the additional deviation introduced by the semantic switch beyond that caused by the length increase. This leads to our first hypothesis:

H1 (Existence & Causality): *The output deviation under a semantic switch is significantly greater than under a length-matched neutral context.*

$$\Delta I_{content} = D(A_2, B_2) - D(A_2, C_2) > 0 \quad (1)$$

For our primary content deviation metric D, we use the cosine distance between TF-IDF weighted vector representations of the outputs, a standard measure of textual dissimilarity. A positive $\Delta I_{content}$ with a 95% confidence interval strictly above zero provides strong evidence for H1.

3.3 Observable Channels of Inertia

To understand *what* is being carried over (RQ2), we decompose inertia's effects into three distinct, observable channels. We fix the source turn (Q_1) to be a writing task to establish a strong stylistic and topical prior, and then probe the model with a target query (Q_2) designed to stress one of three dimensions:

1. **Factual Pollution:** The impact on task accuracy. We measure this as the drop in accuracy on a multiple-choice QA task:

$$\Delta_{Acc} = \text{Accuracy}(B_2) - \text{Accuracy}(A_2) \quad (2)$$

   A negative Δ_{Acc} indicates that inertia harms factuality.

2. **Structural Disruption:** The impact on following formatting rules. We measure this as the increase in failure rate on a task requiring strict JSON output:

$$\Delta_{Fail} = \text{FailRate}(B_2) - \text{FailRate}(A_2) \quad (3)$$

   A positive Δ_{Fail} means inertia disrupts structural compliance.

3. **Stylistic Residue:** The degree to which the output style shifts towards the source turn's style. We quantify this as a differential style similarity to the source response A_1:

$$\Delta_{Sim} = \text{Sim}(B_2, A_1) - \text{Sim}(A_2, A_1) \quad (4)$$

   A positive Δ_{Sim} indicates style bleeding. To ensure robustness, $\text{Sim}(\cdot, \cdot)$ is the cosine similarity of z-score normalized style vectors, where normalization statistics (mean and standard deviation) are pre-computed over the entire experimental corpus.

This decomposition enables testing our second hypothesis, **H2**: *Semantic inertia primarily manifests as stylistic residue.*

3.4 Characterizing Dynamics and Controllability

To capture the temporal nature of inertia (RQ3) and its susceptibility to external control (RQ4), we introduce two further analytical approaches.

Dynamic Decay Analysis. To test that inertia decays over the course of generation (**H3**), we model its dynamics as depicted in Figure 2. For a fixed-length generation, we compute an inertia metric $I(t)$ (e.g., Δ_{Sim} or a lexical bleeding score) for prefixes of increasing token length t. We then fit an exponential decay curve to this series:

$$I(t) = I_0 \cdot e^{-t/\tau} \quad (5)$$

where I_0 is the initial inertia and τ is the characteristic time constant. A finite, positive τ supports the decay hypothesis and provides a quantitative measure of inertia's persistence.

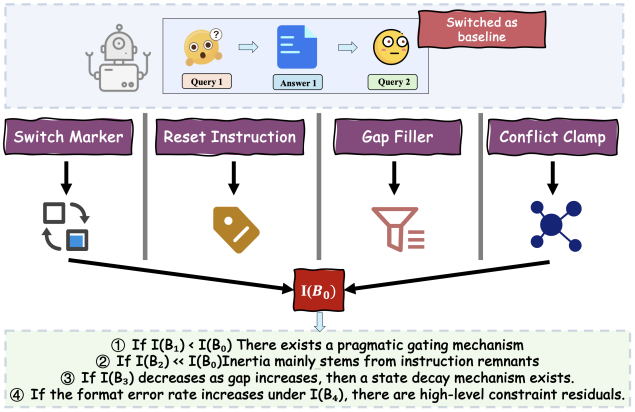

Figure 3: **Probing Control Mechanisms via Interventions.** We apply targeted prompt interventions to the standard switched context to test for underlying control mechanisms. Comparing the resulting inertia strength against the baseline switch provides evidence for or against each mechanism.

Intervention-Based Probing. To investigate inertia's controllability, we test its response to a set of prompt-based interventions (Figure 3). We augment the standard switched context (denoted B_0) with manipulations designed to probe specific gating mechanisms.

- **Reset Instruction (B_{reset}):** An explicit command to ignore prior context tests for high-level instructional control (**H4**). Its effectiveness is measured by the **Gating Gain**:

$$\Delta G_{reset} = D(A_2, B_0) - D(A_2, B_{reset}) \quad (6)$$

- **Conflict Clamp ($B_{conflict}$):** A new instruction that directly contradicts the prior state tests whether inertia is truly eliminated or merely overridden. We hypothesize that conflict amplifies deviation (**H5**), which is confirmed if the **Conflict Amplification** is positive:

$$\Delta A_{conflict} = D(A_2, B_{conflict}) - D(A_2, B_0) > 0 \quad (7)$$

These interventions allow us to build a mechanistic picture of inertia by observing how it reacts to different forms of external pressure.

4 Experiments

We conducted a series of four experiments to test our hypotheses, using a suite of five leading LLMs: DeepSeek-V3, Qwen3-max, Gemini-2.0-flash, Doubao-pro-32k, and Grok-4. A preliminary calibration study (Exp0) was performed to determine optimal inference parameters for stability; its setup and results are detailed in the Appendix. This led us to select a temperature of 0.2 for all subsequent experiments to minimize output variance from sampling noise. Due to consistently anomalous behavior, results for GPT-4o are analyzed separately in the Appendix.

4.1 Exp1: Existence and Causality of Semantic Inertia

This first experiment was designed to provide robust evidence for H1: that semantic inertia is a real effect of prior semantic content, not just an artifact of longer context. We implemented the A/B/C contrastive design across 480 diverse domain-switch pairs for each model. The key metric is the Net Content Inertia, $\Delta I_{content}$, which isolates the deviation attributable to semantic carryover.

Model	Task Form	Net Inertia ($\Delta I_{content}$)	95% CI	Evidence
DeepSeek-V3	`explain`	**0.043**	[0.034, 0.053]	**Strong**
	`json`	**0.057**	[0.048, 0.066]	**Strong**
	`writing`	**0.037**	[0.029, 0.044]	**Strong**
Qwen3-max	`explain`	**0.023**	[0.016, 0.031]	**Strong**
	`json`	**0.031**	[0.023, 0.039]	**Strong**
	`writing`	**0.015**	[0.009, 0.022]	**Strong**
Gemini-2.0-flash	`explain`	**0.012**	[0.006, 0.019]	**Strong**
	`json`	**0.036**	[0.024, 0.047]	**Strong**
	`writing`	**0.008**	[0.003, 0.014]	**Strong**
Grok-4	`explain`	**0.024**	[0.008, 0.042]	**Strong**
	`json`	0.012	[-0.004, 0.028]	Inconclusive
	`writing`	0.010	[-0.005, 0.025]	Inconclusive
Doubao-pro-32k	`explain`	0.006	[-0.004, 0.017]	Inconclusive
	`json`	**0.018**	[0.008, 0.028]	**Strong**
	`writing`	0.006	[-0.006, 0.017]	Inconclusive

Table 1: Net content inertia ($\Delta I_{content}$) across models. Strong evidence for H1 is found when the 95% confidence interval is strictly positive.

The results, summarized in Table 1, offer strong support for H1. Across a majority of models and task formats, the net content inertia is significantly positive. For instance, DeepSeek-V3 shows a consistently large effect, with $\Delta I_{content}$ reaching up to 0.057 in the `json` task. Qwen3-max and Gemini-2.0-flash also exhibit a statistically significant positive inertia effect across all conditions. This confirms that the semantic content of a prior turn induces a greater output deviation than a length-matched neutral context does. The effect is weaker and less consistent for Doubao-pro-32k and Grok-4, suggesting that the magnitude of semantic inertia is a model-dependent property. A detailed breakdown of the raw distance values for each condition is available in the Appendix.

4.2 Exp2: The Dynamics of Inertia Decay

To investigate H3, we analyzed how semantic inertia evolves during the generation process. We measured inertia strength over token prefixes and fitted an exponential decay model to estimate the characteristic time constant, τ. The full distribution of τ values is visualized in Figure 2.

Table 2 summarizes the decay parameters. A key finding is that the distribution of τ is heavy-tailed for all models, evidenced by the large discrepancy between the median and mean values. This indicates that while most inertial effects decay over a predictable timescale, a minority of instances are extremely persistent. The median τ, a more robust measure of central tendency, consistently falls within the range of 87–168 tokens for most models, quantifying a typical "memory half-life." The mean τ for Grok-4 remains notably large compared to its median, reflecting its particularly persistent inertial behavior.

4.3 Exp3: Controllability and Its Limits

This experiment tested the extent to which semantic inertia can be controlled via prompt-based interventions (H4) and

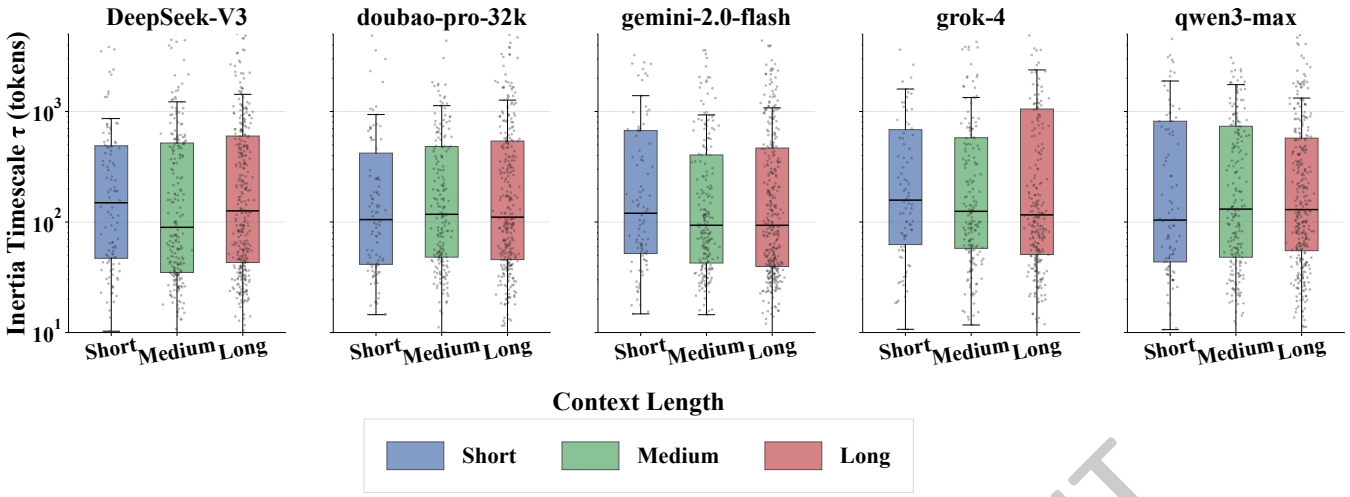

Figure 4: **Distribution of Inertia Timescale (τ) by Model and Initial Context Length.** Each panel displays the distribution of the decay time constant τ for a specific model, grouped by the length of the prior context (Short, Medium, Long). Box plots show the median (thick center line) and interquartile range (box), while the superimposed strip plot reveals the full distribution of individual data points. The y-axis is on a logarithmic scale and is capped at 5000 tokens to better visualize the main body of the distributions, which are all heavy-tailed with extreme outliers.

Model	Task Form	Fit Rate (%)	Median τ (tokens)	Mean τ [CI]
DeepSeek-V3	`explain`	88.8	**168**	5,007 [746, 11891]
	`writing`	82.0	**87**	1,419 [651, 2576]
Qwen3-max	`explain`	84.0	**131**	4,436 [959, 12988]
	`writing`	80.0	**128**	2,077 [597, 4795]
Doubao-pro-32k	`explain`	81.0	**120**	7,228 [825, 16461]
	`writing`	79.5	**113**	10,006 [595, 22548]
Grok-4	`explain`	72.8	**132**	8,088 [5242, 11264]
	`writing`	70.3	**120**	6,294 [3684, 9594]

Table 2: Timescale of semantic inertia (τ) across models. The large gap between the median and mean indicates a heavy-tailed distribution.

Model	Task	Reset Gain (ΔG_{reset})	Conflict Amp. ($\Delta A_{conflict}$)
Gemini-2.0-flash	`explain`	**0.024** [0.00, 0.05]	**0.045** [0.02, 0.07]
	`json`	0.002 [-0.02, 0.03]	**0.120** [0.09, 0.15]
	`writing`	**0.029** [0.01, 0.05]	**0.018*** [-0.01, 0.04]
DeepSeek-V3	`explain`	-0.001 [-0.01, 0.01]	**0.016** [0.01, 0.03]
	`json`	-0.003 [-0.01, 0.01]	**0.025** [0.01, 0.04]
	`writing`	**0.010** [0.00, 0.02]	**0.010*** [-0.00, 0.02]
Qwen3-max	`explain`	-0.004 [-0.01, 0.00]	**0.036** [0.03, 0.04]
	`json`	-0.002 [-0.01, 0.01]	**0.041** [0.03, 0.05]
	`writing`	-0.004 [-0.01, 0.00]	**0.031** [0.02, 0.04]
Grok-4	`explain`	**-0.059** [-0.10, -0.02]	**0.048** [0.01, 0.08]
	`json`	**-0.098** [-0.14, -0.06]	**0.112** [0.07, 0.15]
	`writing`	**-0.118** [-0.16, -0.08]	**0.020*** [-0.01, 0.05]

Table 3: Impact of interventions: reset gain and conflict amplification. Positive values indicate effective control for reset gain and increased deviation for conflict amplification.

how models react to conflicting constraints (H5). We measured the "Reset Gain" (ΔG_{reset}) and "Conflict Amplification" ($\Delta A_{conflict}$).

The results, presented in Table 3, reveal a striking diversity in model controllability. Gemini-2.0-flash is the only model that consistently and effectively responds to the `reset` instruction, showing a significant positive gain, thus supporting H4 for this model. In contrast, Qwen3-max and Doubao-pro-32k largely ignore the instruction. Most dramatically, Grok-4 exhibits a significant negative gain, meaning the `reset` instruction was counter-productive.

In stark contrast, H5 receives universal support. For almost every model and task, the Conflict Amplification effect is significantly positive. This demonstrates that confronting an inertial state with a contradictory instruction does not neutralize it, but instead amplifies output deviation. This effect is particularly strong for Grok-4 and Gemini-2.0-flash on the `json` task, highlighting the fragility of structured generation under conflicting contextual pressures. A more detailed breakdown of all intervention effects is available in the Appendix.

4.4 Exp4: Decomposing the Form of Inertia

To test H2, this experiment decomposed the effect of inertia from a `writing` source task into its impact on factual, structural, and stylistic target tasks. Table 4 summarizes the findings.

The results provide strong support for H2. For all models, **Stylistic Residue is the most prominent and consistent effect**, with the z-scored style drift being universally positive and of a larger magnitude than effects in other channels. In contrast, **Factual Pollution is negligible for most models**.

The effect on **Structural Disruption is surprisingly complex**. While Doubao-pro-32k shows the expected negative effect (a 5.5% increase in JSON failures), DeepSeek-V3 and Gemini-2.0-flash exhibit a significant *improvement* in format adherence (a 6.2% and 14.5% reduction in failures, respectively). This counter-intuitive "constructive inertia" suggests that a writing context can sometimes prime a more structured generation mode. The heatmap for DeepSeek-V3 in Figure 3 visualizes this asymmetrical landscape, showing that

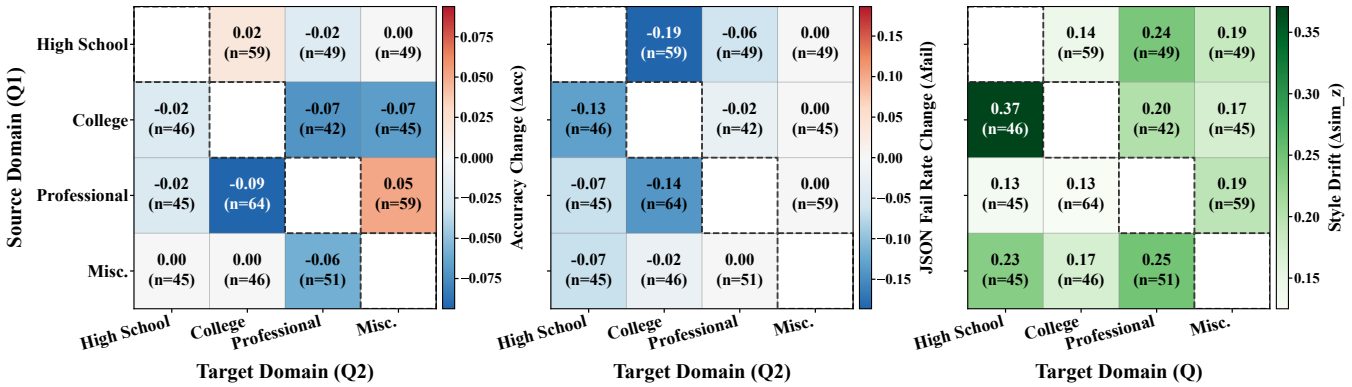

Figure 5: The Asymmetrical Landscape of Semantic Inertia for DeepSeek-V3. The heatmaps visualize the effect of inertia from a `writing` source task to three target task channels across different domain switches. This illustrates the non-uniform nature of inertia's impact, with factual pollution concentrated in specific switches and structural improvement being more widespread.

Model	Factual (Δ Accuracy)	Structural (Δ JSON Fail Rate)	Stylistic (Δ Style Drift)
DeepSeek-V3	**-0.023**	**-0.062**	**+0.198**
Qwen3-max	+0.003	-0.008	**+0.083**
Gemini-2.0-flash	+0.012	**-0.145**	**+0.120**
Doubao-pro-32k	0.000	**+0.055**	**+0.117**
Grok-4	+0.013	-0.008	**+0.101**

Table 4: Decomposition of semantic inertia across channels.

the negative factual effect is concentrated in specific domain switches, while the positive structural effect is widespread. Detailed heatmaps for all models are available in the Appendix.

5 Discussion

Our results show that semantic inertia is a real and model-dependent inference-time effect rather than a simple artifact of longer context. The causal-contrastive analysis (Table 1) demonstrates that the observed deviation is driven by prior semantic content, while the decay analysis (Table 2 and Figure 4) shows that this influence is strongest at the beginning of generation and then gradually weakens. At the same time, the heavy-tailed distribution of τ suggests that although inertia is often short-lived, some cases remain highly persistent and may create substantial reliability risks.

The decomposition results (Table 4) further clarify the form of this effect. Across models, semantic inertia appears primarily as stylistic residue: prior context most consistently changes how the model writes, rather than what it knows. By comparison, factual pollution is generally weak and tends to arise only in specific high-interference domain switches. The effect on structure is more nuanced: while inertia can harm format compliance in some cases, it can also improve it, indicating that prior context may sometimes induce a more organized generation mode rather than simply degrading performance.

The intervention study (Table 3) highlights the limits of prompt-level control. Explicit reset instructions are often unreliable and can even be counter-productive, while conflicting constraints consistently amplify deviation instead of resolving it. This suggests that semantic inertia is not easily overridden by natural-language commands alone and that prompt-based fixes are insufficient for robust state isolation. More broadly, our study is limited to a specific set of English tasks and uses behavioral metrics as proxies for latent internal states. Future work should therefore expand the analysis to other models and languages, and explore system-level solutions such as state management mechanisms or KV-cache interventions.

6 Conclusion

This paper introduced a rigorous framework for defining, measuring, and decomposing semantic inertia in large language models. Our key takeaways are fourfold. First, we demonstrated that semantic inertia is a real and measurable causal effect of prior semantic content, not merely a length or positional artifact. Second, we characterized its dynamics, showing that it is strongest in the initial stages of generation and decays over a characteristic timescale of approximately 100-200 tokens, but with a heavy-tailed distribution indicating rare but extremely persistent instances. Third, we decomposed its form, revealing that inertia primarily manifests as a powerful stylistic residue, with weaker and more complex effects on factual accuracy and structural compliance. Finally, we explored its controllability, revealing the profound limits of prompt-level 'reset' instructions and the consistent, negative effect of conflict amplification.

Together, these findings portray semantic inertia as an intrinsic property of current LLM dynamics. Mitigating its negative effects will likely require more than clever prompting. Our work underscores the need for system-level state isolation mechanisms and new training paradigms that explicitly teach models not just how to maintain context, but also how to effectively "change their minds."

References

[Arbuzov *et al.*, 2025] Mikhail L Arbuzov, Alexey A Shvets, and Sisong Beir. Beyond exponential decay: Rethink-

ing error accumulation in large language models. *arXiv preprint arXiv:2505.24187*, 2025.

[Broadfield *et al.*, 2025] Owen Broadfield, Caleb Smith, Callum Winterbourne, Felix Manderston, Kent Blumberg, and Robert Pike. Lexical inertia control in transformer architectures: A study on temporal token reweighting for large language models. 2025.

[Chen *et al.*, 2024a] Nuo Chen, Jiqun Liu, Xiaoyu Dong, Qijiong Liu, Tetsuya Sakai, and Xiao-Ming Wu. Ai can be cognitively biased: An exploratory study on threshold priming in llm-based batch relevance assessment. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 54–63, 2024.

[Chen *et al.*, 2024b] Yaofo Chen, Zeng You, Shuhai Zhang, Haokun Li, Yirui Li, Yaowei Wang, and Mingkui Tan. Core context aware transformers for long context language modeling. *arXiv preprint arXiv:2412.12465*, 2024.

[Fang *et al.*, 2025] Hanpei Fang, Sijie Tao, Nuo Chen, Kai-Xin Chang, and Tetsuya Sakai. Do large language models favor recent content? a study on recency bias in llm-based reranking. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 85–94, 2025.

[Guan *et al.*, 2025] Bryan Guan, Tanya Roosta, Peyman Passban, and Mehdi Rezagholizadeh. The order effect: Investigating prompt sensitivity to input order in llms. *arXiv preprint arXiv:2502.04134*, 2025.

[Guo *et al.*, 2023] Jing Guo, Nan Li, Jianchuan Qi, Hang Yang, Ruiqiao Li, Yuzhen Feng, Si Zhang, and Ming Xu. Empowering working memory for large language model agents. *arXiv preprint arXiv:2312.17259*, 2023.

[Gupta *et al.*, 2024] Akash Gupta, Ivaxi Sheth, Vyas Raina, Mark Gales, and Mario Fritz. Llm task interference: An initial study on the impact of task-switch in conversational history. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14633–14652, 2024.

[Kingsley *et al.*, 2025] Blaine Kingsley, Benedict Wickersham, Andrew Delta, Xavier Martin, Andrew Scolto, and Peter Appleyard. Probabilistic contextual resonance in large language model decoding through selfmodulated semantic interference. 2025.

[Leonardi *et al.*, 2024] Francois Leonardi, Patrick Feldman, Matthew Almeida, William Moretti, and Charles Iverson. Contextual feature drift in large language models: An examination of adaptive retention across sequential inputs. 2024.

[Li *et al.*, 2023] Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. In *Findings of the association for computational linguistics: ACL 2023*, pages 1774–1793, 2023.

[Li *et al.*, 2025] Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*, 2025.

[Liskavets *et al.*, 2025] Barys Liskavets, Maxim Ushakov, Shuvendu Roy, Mark Klibanov, Ali Etemad, and Shane K Luke. Prompt compression with context-aware sentence encoding for fast and improved llm inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24595–24604, 2025.

[Liu *et al.*, 2024] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

[O'Brien *et al.*, 2024] Kyle O'Brien, Nathan Ng, Isha Puri, Jorge Mendez, Hamid Palangi, Yoon Kim, Marzyeh Ghassemi, and Thomas Hartvigsen. Improving black-box robustness with in-context rewriting. *arXiv preprint arXiv:2402.08225*, 2024.

[Patel *et al.*, 2023] Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. Learning interpretable style embeddings via prompting llms. *arXiv preprint arXiv:2305.12696*, 2023.

[Peysakhovich and Lerer, 2023] Alexander Peysakhovich and Adam Lerer. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*, 2023.

[Raimondi and Gabbrielli, 2025] Bianca Raimondi and Maurizio Gabbrielli. Exploiting primacy effect to improve large language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 989–997, 2025.

[Shaki *et al.*, 2023] Jonathan Shaki, Sarit Kraus, and Michael Wooldridge. Cognitive effects in large language models. *arXiv preprint arXiv:2308.14337*, 2023.

[Spataru *et al.*, 2024] Ava Spataru, Eric Hambro, Elena Voita, and Nicola Cancedda. Know when to stop: A study of semantic drift in text generation. *arXiv preprint arXiv:2404.05411*, 2024.

[Talukdar and Biswas, 2024] Wrick Talukdar and Anjanava Biswas. Improving large language model (llm) fidelity through context-aware grounding: A systematic approach to reliability and veracity. *arXiv preprint arXiv:2408.04023*, 2024.

[Wu *et al.*, 2024] Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2341–2369, 2024.

[Xiong *et al.*, 2022] Wenhan Xiong, Barlas Oguz, Anchit Gupta, Xilun Chen, Diana Liskovich, Omer Levy, Scott Yih, and Yashar Mehdad. Simple local attentions remain competitive for long-context tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1975–1986, 2022.

[Xu *et al.*, 2024] Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*, 2024.

[Yeh *et al.*, 2023] Kai-Ching Yeh, Jou-An Chi, Da-Chen Lian, and Shu-Kai Hsieh. Evaluating interfaced llm bias. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 292–299, 2023.

[Yi *et al.*, 2025] Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Computing Surveys*, 58(6):1–38, 2025.

[Yin *et al.*, 2023] Wenpeng Yin, Qinyuan Ye, Pengfei Liu, Xiang Ren, and Hinrich Schütze. Llm-driven instruction following: Progresses and concerns. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 19–25, 2023.