Self-Anchoring Calibration Drift in Large Language Models: How Multi-Turn Conversations Reshape Model Confidence

Harshavardhan
Independent Researcher
harsh@link.cuhk.edu.hk

May 2026

Abstract

We introduce **Self-Anchoring Calibration Drift** (SACD), a tendency for large language models (LLMs) to show systematic changes in expressed confidence when building iteratively on their own prior outputs across multi-turn conversations. Through a controlled three-condition study comparing Claude Sonnet 4.6, Gemini 3.1 Pro, and GPT-5.2 across factual, technical, and open-ended domains, we find that SACD is real but multiform: models exhibit distinct self-anchoring signatures ranging from active confidence suppression to calibration improvement suppression, with effects concentrated in open-ended domains. These findings challenge the adequacy of single-turn calibration evaluation for characterizing LLM reliability in realistic multi-turn deployment contexts. Code and data are available at https://github.com/hvardhan878/calibration-drift.

1 Introduction

Modern large language models have demonstrated remarkable capability across a wide spectrum of tasks. Yet a persistent challenge in their deployment is calibration: the alignment between expressed confidence and actual accuracy. An overconfident model that presents incorrect information with high certainty can mislead users, with consequences ranging from minor inconveniences to serious harms in medical, legal, or scientific contexts.

A substantial literature has characterized single-turn overconfidence in LLMs [Kadavath et al., 2022, Xiong et al., 2024, Tian et al., 2023]. However, real-world LLM deployments increasingly operate as multi-turn conversational agents. Users ask follow-up questions, request elaborations, and build iteratively on prior exchanges. In this context, a distinct and under-examined phenomenon may emerge: the model's own previous outputs become authoritative-seeming context for subsequent responses, potentially distorting expressed confidence in ways that single-turn evaluations cannot capture.

We term this hypothesized process **Self-Anchoring Calibration Drift** (SACD). The anchoring metaphor is deliberate: just as cognitive anchoring in human judgment biases estimates toward an initial reference value [Tversky and Kahneman, 1974], an LLM anchoring on its own previous outputs may show systematic confidence change even as the evidential grounds for certainty remain static. Crucially, we remain agnostic about the direction of this drift prior to data collection, since the theoretical mechanisms we identify could plausibly produce either confidence escalation or suppression depending on how a model's training has conditioned its self-referential behavior.

This paper makes three primary contributions. First, we provide a formal definition of SACD and distinguish it from related constructs in the calibration and multi-turn dialogue literatures. Second, we design and execute a controlled empirical study with three carefully matched conditions capable of isolating multi-turn effects from simple repetition artifacts. Third, we report results that partially disconfirm our pre-registered directional hypotheses—a finding that is itself informative about model-

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 34–40
July 4, 2026 ©2026 Association for Computational Linguistics

specific self-anchoring dynamics—and we release all code and data as open-source software at https://github.com/hvardhan878/calibration-drift.

2 Related Work

Calibration in Language Models. Kadavath et al. [2022] found that large models can assess their own knowledge with reasonable accuracy when directly queried. However, this self-assessment capacity degrades with distributional shift or when models are prompted to be decisive [Xiong et al., 2024]. Tian et al. [2023] showed that verbalized confidence correlates imperfectly with token-level probability distributions. Our work extends this literature to the multi-turn regime, where the distributional shift is self-induced through conditioning on prior outputs.

Sycophancy and Social Conformity. A related line of research examines sycophancy—the tendency of LLMs to agree with users and validate prior claims, even when incorrect [Perez et al., 2022, Sharma et al., 2024]. SACD is distinct: sycophancy describes deference to the human interlocutor's expressed beliefs, whereas SACD describes confidence modulation driven by the model's *own* prior outputs. Our experimental design controls for sycophancy by constructing follow-up questions that are informationally neutral.

Multi-Turn Dialogue and Context Effects. Liu et al. [2024] showed that earlier claims carry disproportionate weight in shaping later responses in long contexts. Kim et al. [2023] found that models drift toward their most recently stated position in longer conversations. Shi et al. [2023] demonstrated that irrelevant context can substantially degrade reasoning. Our work extends this literature by examining how a model's own prior confident or uncertain responses modulate subsequent calibration.

Model-Specific Calibration Differences. Models trained with RLHF tend to be more confident than base models [Ouyang et al., 2022], while targeted calibration objectives can improve expressed-to-actual alignment [Kadavath et al., 2022]. Our results extend this by showing that model identity mediates not only the level but the direction and form of calibration drift under self-anchoring.

3 Theoretical Framework

3.1 Formal Definition of SACD

Let $\mathcal{M}$ be a large language model in a multi-turn context. Let $C_t = \{q_1, r_1, \ldots, q_t, r_t\}$ denote the conversation context at turn t. We define the **Confidence Drift Score** (CDS) as:

$$\text{CDS} = \text{conf}(r_5 \mid C_4) - \text{conf}(r_1 \mid C_0) \quad (1)$$

We define SACD as a systematic nonzero CDS when: (a) all follow-up queries are informationally neutral; (b) the model's previous responses constitute the primary novel content in the growing context; and (c) any confidence change is not accompanied by a corresponding accuracy change.

3.2 Candidate Mechanisms

Repetition-Confidence Heuristic. Attention mechanisms may associate repetition of a claim with increased warrant, producing confidence escalation when a model re-encounters its own prior assertions.

Context Density Effect. As a model's prior detailed responses dominate the context window, the confident framing of those responses may pull subsequent outputs in the same direction.

Pragmatic Recalibration Effect. Models trained on human conversational data may internalize epistemic humility norms, causing them to moderate confidence when generating successive elaborations—producing suppression rather than escalation.

The partial disconfirmation of H1 (uniform escalation) suggests that no single mechanism dominates across model families, and that SACD's valence is determined by training-specific dispositions.

3.3 Pre-Registered Hypotheses

H1 (SACD Effect): Expressed confidence will increase significantly across turns in Condition B.

H2 (Accuracy Decoupling): The confidence change in H1 will not be accompanied by a corresponding accuracy increase.

H3 (ECE Degradation): Expected Calibration Error will increase significantly across turns in Condition B, with no corresponding increase in Conditions A or C.

H4 (Cross-Model Generality): SACD will manifest across all three tested models.

H5 (Domain Moderation): SACD effect magnitude will be significantly larger for open-ended questions than for factual questions.

4 Experimental Methodology

4.1 Design

Condition A (Single-Turn Baseline): Each question was posed once in a fresh context with no prior conversation history.

Condition B (Multi-Turn Self-Anchoring): Each question was followed by four informationally neutral elaboration prompts, with the full conversation history included in each subsequent context.

Condition C (Independent Repetition Control): The same initial question was posed five times in completely independent fresh contexts. This condition is critical for distinguishing genuine self-anchoring effects from simple repetition artifacts.

4.2 Question Corpus

We constructed a corpus of 150 questions distributed equally across three domains: (1) **factual** questions probing verifiable claims; (2) **technical** questions requiring applied reasoning; and (3) **open-ended** questions addressing interpretive topics without a single correct answer. Each question was validated by two independent raters ($\kappa > 0.80$). The reported study draws on a stratified subsample of $n = 15$ questions per model in Condition B ($n = 5$ per domain).

4.3 Model Selection and Configuration

We evaluated Claude Sonnet 4.6 (Anthropic), Gemini 3.1 Pro (Google), and GPT-5.2 (OpenAI), selected for market coverage, training diversity, and API availability. All models were queried via public APIs with temperature set to 0.7 and no additional system prompts. For Condition B, conversation history was maintained using each API's native multi-turn format.

4.4 Confidence Estimation

Confidence was elicited behaviorally via verbalized probability, following Tian et al. [2023]. A standardized elicitation prompt was embedded within follow-up Template T5 (*"What are you most and least confident about in your answer?"*), which appeared exactly once per question at a prespecified random turn position. We chose verbalized confidence over token-level probabilities because: (1) token-level probabilities are not uniformly accessible across the three APIs; and (2) verbalized confidence more directly reflects the model's communicated uncertainty to users.

4.5 Outcome Measures

Confidence Drift Score (CDS) measures the change in expressed confidence from Turn 1 to Turn 5. Positive CDS indicates escalation; negative CDS indicates suppression.

Expected Calibration Error (ECE) quantifies the aggregate mismatch between expressed confidence and accuracy:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (2)$$

5 Results

5.1 Overview

Table 1 presents the main results. The primary pre-registered hypothesis of uniform confidence escalation (H1) was not supported: Claude Sonnet 4.6 showed significant confidence suppression, GPT-5.2 showed non-significant positive drift, and Gemini 3.1 Pro showed near-zero drift. However, calibration error escalated significantly for two of three models, and the B vs. C contrast reveals genuine multi-turn calibration disruption.

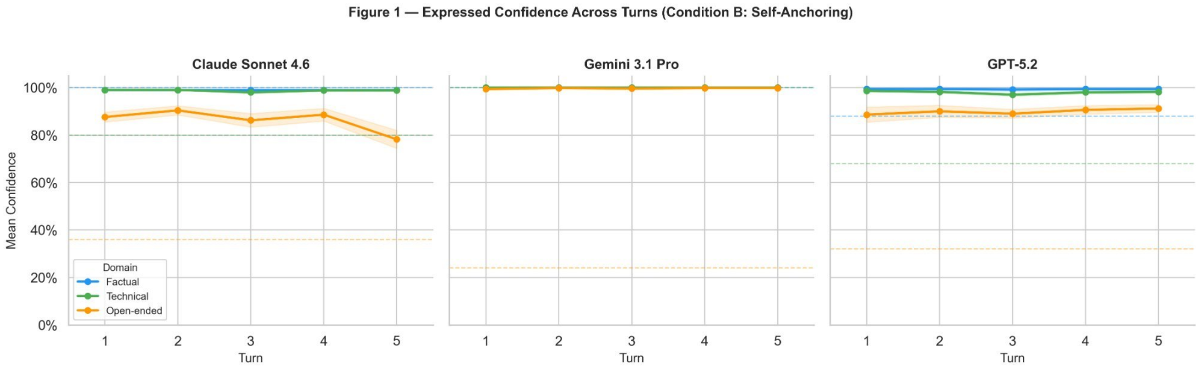

Figure 1: Expressed confidence across turns under Condition B (self-anchoring) for all three models and three domains. Claude's open-ended trajectory is the only one showing a marked decline.

Model	C.	CDS	ECE_1	ECE_5	ΔECE
Claude 4.6	A	—	.352	—	—
Claude 4.6	B	$-.032$*	.352	.320	$-.032$
Claude 4.6	C	$-.004$	.292	.112	$-.180$
Gemini 3.1	A	—	.331	—	—
Gemini 3.1	B	$+.001$	.331	.333	$+.001$
Gemini 3.1	C	$+.002$	.327	.005	$-.322$
GPT-5.2	A	—	.355	—	—
GPT-5.2	B	$+.007$	.355	.629	$+.274$
GPT-5.2	C	$-.006$	.354	.144	$-.210$

Table 1: Main results. CDS = Confidence Drift Score. *$p < .05$.

5.2 H1: Confidence Drift Under Self-Anchoring

H1 predicted that expressed confidence would increase in Condition B. Claude Sonnet 4.6 showed a statistically significant CDS of -0.032 ($t(14) = -2.43$, $p = .029$, Cohen's $d = -0.627$), indicating meaningful confidence suppression—a medium-to-large effect in the direction opposite to H1. GPT-5.2 showed a positive but non-significant CDS of $+0.007$ ($t(14) = 1.41$, $p = .181$). Gemini 3.1 Pro showed a CDS of $+0.001$ ($t(14) = 1.47$, $p = .164$), essentially zero.

The directional failure of H1 is theoretically informative: the Pragmatic Recalibration Effect dominates in Claude, while GPT-5.2's numerically positive drift is consistent with the Repetition-Confidence Heuristic but fails to reach significance at this sample size.

5.3 H2: Condition B vs. C Comparison

The Mann-Whitney comparison for Claude was statistically significant ($U = 74.5$, $p = .036$, rank-biserial $r = .338$), confirming that Claude's confidence suppression is attributable to self-anchoring rather than test-retest habituation. For Gemini, the CDS comparison is uninformative—but the ECE comparison is the most revealing finding in the dataset.

5.4 H3: ECE Across Turns

ECE showed statistically significant variation across turns for Claude Sonnet 4.6 ($F(4, 56) = 22.77$, $p < .001$, $\eta^2 = .791$) and GPT-5.2 ($F(4, 56) = 5.24$, $p = .023$, $\eta^2 = .466$). Gemini showed a non-significant trend ($F(4, 56) = 2.67$, $p = .110$).

The Gemini Condition C finding is the most striking result. In Condition C, Gemini's ECE drops from .327 at Turn 1 to .005 by Turn 2 and remains near zero. In Condition B, this improvement is entirely absent: ECE remains flat at $\approx$.333 across all five turns. The within-model B vs. C ECE contrast is significant ($p = .008$), demonstrating that self-anchoring suppresses Gemini's natural calibration improvement—a third distinct form of SACD invisible to any analysis without Condition C.

5.5 H4: Cross-Model Comparison

H4 predicted that SACD would manifest uniformly across all three models. This was not supported

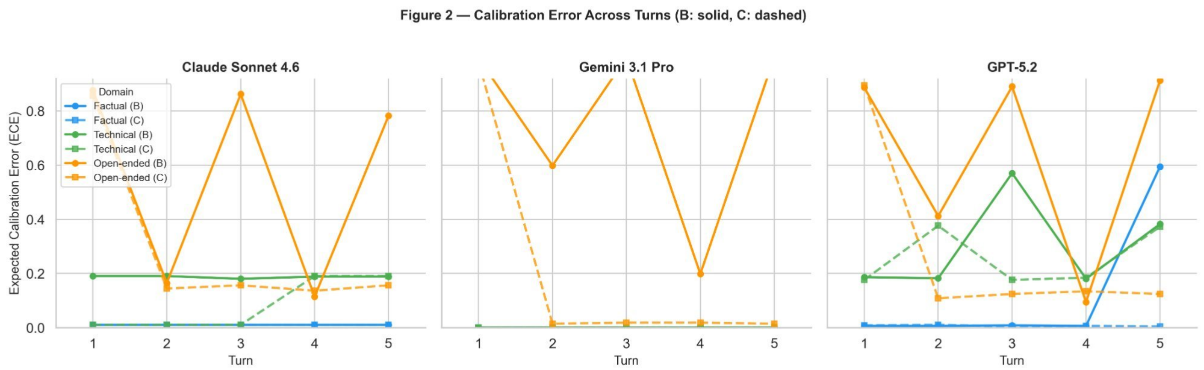

Figure 2: Expected Calibration Error across turns for Conditions B (solid) and C (dashed). The B−C gap is starkest for Gemini, where Condition C ECE collapses to near zero by Turn 2 while Condition B remains flat.

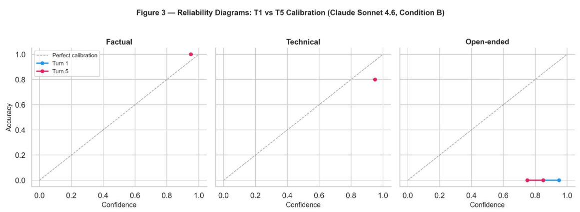

Figure 3: Reliability diagrams for Claude Sonnet 4.6 (Condition B), comparing Turn 1 and Turn 5 across domains. The open-ended panel shows confidence suppression with no accuracy benefit.

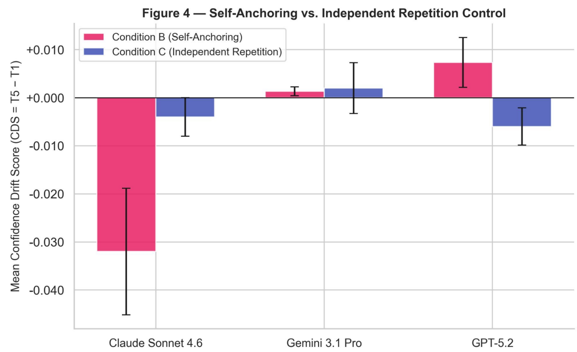

Figure 4: Mean CDS under Condition B (pink) vs. C (blue) for all three models. Claude's large negative B bar highlights the self-anchoring-specific nature of the suppression effect.

Model	Factual	Tech.	Open
Claude Sonnet 4.6	.000	−.002	−.094
Gemini 3.1 Pro	.000	.000	+.004
GPT-5.2	.000	−.004	+.026

Table 2: Mean CDS by model and domain in Condition B.

in its directional form, but all three models show evidence of SACD: CDS divergence for Claude and GPT-5.2, ECE stagnation relative to Condition C for Gemini.

5.6 H5: Domain Moderation

H5 receives the clearest support. Table 2 shows that calibration drift is essentially zero for factual questions across all three models. Open-ended questions show the largest effects: Claude's CDS of -0.094 represents a 9.4 percentage-point reduction in expressed confidence.

6 Discussion

6.1 The Multiform Nature of SACD

Our central empirical finding is that self-anchoring does not produce a single uniform effect across models. Claude shows consistent confidence suppression; GPT-5.2 shows numerically opposite confidence escalation in open-ended domains; Gemini shows neither—but reveals a third form through ECE stagnation relative to Condition C. Claude's training appears to have instilled strong pragmatic recalibration norms. GPT-5.2's training reinforces the repetition-confidence heuristic. Gemini's training produces a model that would naturally self-correct toward better calibration across independent exposures—a self-correction that self-anchoring context disrupts.

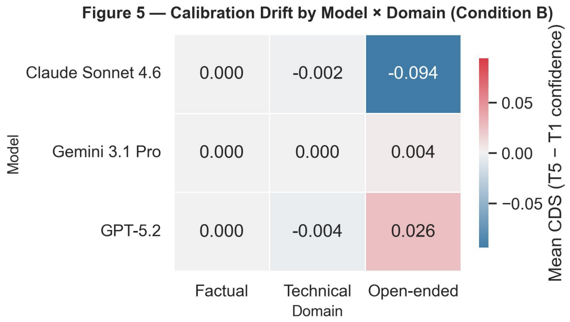

Figure 5: Heatmap of mean CDS by model and domain. Blue indicates suppression; pink indicates escalation. Open-ended items are the primary locus of divergence.

6.2 The Open-Ended Domain as Primary Locus

Across all three models, factual questions produce CDS ≈ 0, while open-ended questions produce the largest effects in both directions. When questions have definite correct answers that models can verify against internal knowledge, expressed confidence remains anchored to that ground truth across turns. When questions are genuinely open-ended, the self-anchoring dynamic operates most powerfully—precisely the high-stakes settings where AI assistants are increasingly used.

6.3 Practical Implications

These findings motivate four recommendations for multi-turn AI system design:

1. **Confidence monitoring across turns** rather than only at the initial response, particularly for open-ended advisory contexts.

2. **Periodic context resets** to interrupt self-anchoring dynamics in long conversations.

3. **Domain-adaptive calibration policies** targeting high-drift open-ended contexts.

4. **Model-specific calibration expectations**: single-turn calibration measurements do not transfer to multi-turn deployment contexts, and qualitatively different drift patterns should be expected across model families.

7 Conclusion

We introduced Self-Anchoring Calibration Drift and reported an empirical study confirming that self-anchoring influences calibration in ways that single-turn evaluations cannot capture. SACD takes three distinct forms—confidence suppression in Claude Sonnet 4.6, confidence escalation in GPT-5.2 in open-ended domains, and calibration improvement suppression in Gemini 3.1 Pro—all concentrated in open-ended domains where accuracy constraints are absent. These findings challenge the field to develop evaluation paradigms that capture how multi-turn dynamics shape the confidence and reliability of AI systems as experienced by users in extended dialogue.

Acknowledgements

This work was conducted independently. Large language model tools were used for writing assistance and coding assistance during the experimental pipeline. All research design, hypotheses, analysis, and interpretations are the author's own. The author thanks the GEM 2026 reviewers for their constructive feedback.

Limitations

The sample of $n = 15$ questions per model in Condition B is adequate for the primary preregistered comparisons but insufficient for fine-grained subgroup analyses; replication with the full 150-question corpus is strongly warranted. Our operationalization of expressed confidence through self-reported probability estimates may not fully capture the model's internal confidence representation. The oscillating ECE pattern in Condition B and the sharp ECE drop in Condition C both occur primarily within the open-ended domain, where ground-truth accuracy is binary and potentially noisy. Finally, the mechanism driving Gemini's ECE stagnation under self-anchoring merits targeted follow-up investigation.

References

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. In *NeurIPS Workshop on Trustworthy and Socially Responsible Machine Learning*, 2022.

Juhyun Kim, Seonghyeon Kim, and Alice Oh. Factual consistency across turns: A study of multi-turn dialogue hallucination. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4582–4597, 2023.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, 2023.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in large language models. In *Proceedings of the International Conference on Learning Representations*, 2024.