SENTINEL: Sentiment Evolution and Narrative Tracking in Extended LLM Interactions

Pranav Anuraag **Ethan Xu** **Alexander Arutchev** **Asher Nerenberg**
Vasu Sharma* Kevin Zhu†
Algoverse AI Research
sentineleapa@gmail.com, kevin@algoverse.us

Abstract

Large language models (LLMs) are increasingly embedded in long-form and agentic workflows, where tonal consistency matters as much as factual accuracy. Although prior work has examined factual hallucinations and demonstrated that they accumulate linearly, we show that emotional hallucinations, artificial generations of exaggerated or incorrect emotions, follow a different dynamic: rather than accumulating linearly, sentiment drift emerges in oscillatory bursts, often correcting or exacerbating itself mid-chain. We introduce SENTINEL, a framework for quantifying sentiment drift via three complementary metrics: Mean Absolute Drift (magnitude of shift), Variance (intra text volatility), and a novel Drift Propagation Index (extent to which drift compounds over steps). Using essays, reviews, and news across five LLMs, we find that drift is domain and model dependent, with negative texts especially prone to "neutralization" over time. Token-level attribution further reveals that a handful of emotionally charged words disproportionately drive drift, suggesting practical levers for mitigation. Our results position emotional hallucination as a distinct phenomenon requiring new interpretability tools, and highlight the risks of unmonitored sentiment drift in high-stakes agentic applications.

1 Introduction

Recent advancements in large language models (LLMs) have demonstrated impressive capabilities in natural language understanding and generation [Brown et al., 2020]. However, a common problem still faced by LLMs is hallucination: where models generate fabricated facts, employ faulty reasoning, attribute imagined emotions, or produce responses misaligned with the prompt [Huang et al., 2025]. Current research focuses mainly on factual hallucinations, while neglecting the equally important emotional hallucinations. In particular, LLMs are prone to hallucinating sentiment or the tone of the text [Xing et al., 2025]. This phenomenon is particularly concerning when LLMs are prompted with repeated textual generation or are expected to maintain consistency across interactions, common traits of multi-chain workflows [Fan et al., 2025]. These hallucinated sentiments can drastically shift the emotional tone of outputs in subtle but impactful ways [Richardson et al., 2025]. Small deviations in sentiment can alter how users interpret intent, credibility, and empathy. This creates a dynamic that risks misleading users into unsafe and unwanted decisions in politics, mental health, healthcare, or customer support [Zhang et al., 2025]. While average sentiment scores of generated texts are often analyzed, less attention has been paid to the variance of sentiment within and across generated outputs. This can lead to a misleading representation of emotional tone, particularly in contexts where highly contrasting sentiments yield an artificially neutral score. For example, the sentences

*Adivsor
†Program Director

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

"This product is terrible. But at the same time it is amazing." may receive the same sentiment score as "This product is mediocre," despite its higher emotional variability.

While transformer-based sentiment models such as BERT offer polarity scores grounded in contextual embeddings [Reddy et al., 2020], they still often overlook the **intra-textual variation in emotional tone** across sentences or spans [Alhuzali and Ananiadou, 2021]. In high-stakes applications such as customer feedback analysis, investor confidence analysis, political discourse tracking, and mental health monitoring, it is important not only to understand what sentiment is expressed, but also how sentiment fluctuates and concentrates across a piece of text [Teodorescu and Mohammad, 2022]. Moreover, the behavior of large language models in maintaining or changing sentiment variance across chains of paraphrasing or summarization remains largely unexamined. Small emotional perturbations or stylistic changes may compound, leading to unintended tonal shifts that go undetected by document-level sentiment averages [Hipson and Mohammad, 2021]. It is also important to investigate whether these drift phenomena manifest similarly in agentic frameworks in which high volume of LLM interactions can steer actions to an undesired result

1.1 Related Literature

Liu et al. [2025], 2024 discovers that LLMs only show basic sensitivity to sentiment, noting substantial variations in both accuracy and consistency in text generation. The findings also include that differing versions of the same LLM show varying results and that prompt processing has minimal effects on neutral prompts overall.

Xing et al. [2025] introduces EmotionHallucer the first benchmark for emotional hallucinations in LLMs textual outputs based on two categories of emotional psychology and knowledge and multimodal emotional perception in the real world. The results demonstrated the rampant emotional hallucination from models and proposed a PEP-MEK framework to reduce emotional hallucination. The framework works by introducing an intermediary step for reasoning, which is extracted from the emotional knowledge and factors from the text, which is then combined with the original input text.

Richardson et al. [2025] laid the groundwork on sentiment score analysis for iterative rephrasing, proposing a sentiment fidelity score to measure the drift in output sentiment with RoBERTa aligned with the dataset ratings. After evaluating the sentiment fidelity across 50 iterations, significant sentimental trends in large LLMs were uncovered, suggesting homogenization of emotion and altering the nature of texts.

All of this research has dramatically advanced the study of sentiment hallucinations. Prior work has largely focused on single-turn completions or isolated summarization chains, leaving open the question of how sentiment evolves when models are embedded in multi-step, AutoGPT-like [Significant Gravitas, 2023] reasoning loops. Our work addresses this gap by being, to our knowledge, the first to systematically test sentiment drift and variance inside structured agentic workflows. By analyzing how emotional deviations accumulate and propagate across multi-step inference, we extend the study of hallucinations from static generations to dynamic, iterative processes which can highlight the ways in which workflow design itself can shape emotional stability in LLM outputs.

2 Methodology

2.1 Datasets and models

To evaluate sentiment drift in rephrasing and agentic workflows done by large language models, we use three different types of texts: personal essays, opinionated movie reviews, and objective news articles. This allows us to examine LLM created sentiment drift in texts with varying levels of objectivity and initial sentiment. For personal essays, we use essays from a dataset commonly used for personality research [Jing Jie, 2024]. For movie reviews, we use a large collection of reviews from imdb [Maas et al., 2011]. For objective news statements, we use the largest publicly available dataset of naturally occurring factual claims [Augenstein et al., 2019] and use only the text labeled neutral for greater objectivity. We perform our experiments on GPT-4.1 [OpenAI, 2024], Deepseek-r1 [DeepSeek-AI et al., 2025], Claude-Sonnet-4 [Anthropic, 2025], Qwen3 [Yang et al., 2025], and Llama-4-Maverick [AI, 2025].

2.2 Quantifying drift

We begin by generating a chain of rephrasings for each source text T_0, yielding outputs $T_1, T_2, \ldots, T_n$. For each dataset we select a subset of 50 texts and iteratively prompt the model to rephrase the previous output without changing its sentiment. This baseline setup allows us to track how sentiment evolves purely under repeated rephrasings.

To quantify drift, we compute three complementary metrics. First, Mean Absolute Drift (MAD) measures the average absolute change in sentiment between Step 0 (original text) and the final step, reflecting the overall magnitude of shift. Second, Sentiment Variance captures volatility in sentence-level sentiment within each generated text, where low variance indicates consistency and high variance indicates swings or inconsistencies. Third, we introduce the Drift Propagation Index (DPI), which measures whether drift compounds across intermediate steps or is partially corrected along the way. For example, if a text flips between positive and negative, the DPI is low despite high volatility, showing that oscillations dominate over accumulation in terms of emotional sentiment drift.

DPI is formally defined as

$$\text{DPI} = \frac{|S_K - S_0|}{\sum_{k=0}^{K-1} |S_{k+1} - S_k|} \in [0, 1], \tag{1}$$

with higher values indicating stronger cumulative divergence.

Sentence-level sentiment scores were computed using a RoBERTa-based sentiment classifier [Barbieri et al., 2020], enabling calculation of all three metrics across both iterative rephrasings (baseline) and structured agentic workflows.

2.3 Experimental design

While extended interactions with LLMs have been shown to reduce factual hallucination [Dhuliawala et al., 2023], the effect of emotional hallucination across multi-step inference patterns remains largely unexplored. To investigate this, we introduce a five-step sequential inference pattern designed to mirror a generalized agentic workflow similar to systems like AutoGPT or Deep Research that operates across a variety of real-world contexts, from emotionally charged rhetoric to highly objective material. This framework enables us to assess sentiment drift and emotional hallucination within different domains, while preserving domain-specific relevance in each step.

Overview of the process:

1. **Intent Planning & Goal Listing:** The model is tasked with interpreting the original text (personal essay, movie review, news article) and producing a list of 2–3 explicit goals, intentions, or key points. These are framed in the context of the dataset domain.
2. **Focused Elaboration:** The model selects one of the goals (or, in our controlled setup, is forced to elaborate on the second goal listed in Step 1) and produces a deeper discussion or reflection specifically on that point. This step isolates one aspect of the content, enabling us to test how narrowing focus affects sentiment stability or drift.
3. **Summarization:** The model produces a concise 4 to 6 sentence summary of the original text. This step is expected to retain the original sentiment the most closely as it is primarily an act of compression rather than interpretation. Step 3 serves as an experimental control for sentiment drift, providing a baseline against which further more generative steps can be compared later.
4. **Context-Appropriate Response:** The model responds to the original text in a way consistent with the dataset's domain: Personal Narratives - Compassionate and supportive commentary. Movie/Product Reviews - Critique or evaluative commentary. News Articles - Analytical but objective summary and commentary, minimizing bias. This is where sentiment drift can become more pronounced, as the model's own alignment and bias tendencies may influence the tone.
5. **Self-Reflection & Bias Assessment:** The model reviews its Step 4 output and assesses whether it was fair, balanced, and appropriate for the context. The model is prompted to explicitly state whether it would change anything. This step tests the model's meta-cognitive awareness of emotional or evaluative bias.

Even though the five-step structure is fixed, the prompting and expected tone of each step is tailored to the domain in order to remain realistic and representative of how such an agentic process would operate in practice by including all the different types of reasoning and outputs expected from one. An evaluative commentary is natural for a movie review but would constitute editorial bias in a factual report. This makes sure that we are not only testing sentiment drift in the abstract, but also capturing how model alignment interacts with different communicative norms and user expectations across different domains.

While the five steps do differ in task type, this variation is deliberate to reflect *heterogeneous agentic roles* rather than iterative rewriting. We view sentiment drift in such workflows as inherently domain-specific: some applications (e.g., mental-health dialogue systems) require adaptive sentiment alignment with user affect, whereas others (e.g., news summarization or fact-checking) demand sentiment stability. Thus, measuring drift across varied turns provides insight into how sentiment appropriately evolves—or fails to evolve—given an agent's intended function.

Token-level attribution. To examine the lexical sources of sentiment drift, we conducted a token-level attribution analysis. We applied *Integrated Gradients* (IG; [Sundararajan et al., 2017]) with respect to our RoBERTa sentiment classifier on each generated output. IG assigns each token a signed attribution score, indicating its contribution toward positive or negative sentiment relative to a neutral baseline. We ranked tokens by attribution and tracked the most influential contributors across steps to see whether drift was driven by the insertion, removal, or reformulation of emotionally marked terms such as negations, intensifiers, and sentiment adjectives. This step was designed as an interpretive complement to our quantitative metrics, offering intuition about how specific lexical choices shape drift rather than serving as a formal measure.

3 Results

3.1 Baseline iterative rephrasings

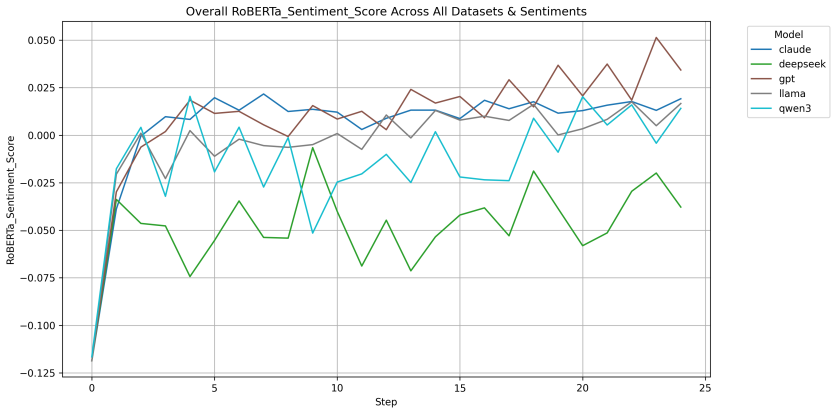


Figure 1: Overall RoBERTa sentiment score across all datasets.

Figure 1 above illustrates the sentiment mean drift, $S_{\text{mean}}(T_i)$, across 25 iterations of summarization for five models. This graph represents an aggregation of all data from the three selected datasets, though certain trends emerge only when examining specific subsets. Across all models, the mean absolute drift tends to stabilize toward the final iterations perhaps finding each model's preferred sentiment ouptut zone. Nonetheless, each model demonstrates a notable positive increase in mean sentiment relative to the initial texts. These findings suggest that, under common rephrasing and improvement prompts, large language models exert a transformative effect on narrative tone. Notably, the most striking observation is the convergence at step 1, where models all climb to nearly symmetric sentiment scores.

For further analysis, the results were disaggregated by dataset and further segmented according to initial sentiment categories (Positive, Neutral, and Negative; ($S_{\text{mean}}(T_0)$) and added to the appendix.

Table 1: Sentiment scoring summary

Sentiment	Model	Mean Absolute Drift	Variance	DPI
Negative	claude	0.3030	0.0409	0.0893
Negative	deepseek	0.4646	0.0714	0.1066
Negative	gpt	0.3180	0.0460	0.1107
Negative	llama	0.3701	0.0429	0.1285
Negative	qwen3	0.3898	0.0570	0.1120
Neutral	claude	0.2549	0.0224	0.0771
Neutral	deepseek	0.2960	0.0441	0.0922
Neutral	gpt	0.3030	0.0248	0.1049
Neutral	llama	0.3016	0.0345	0.0934
Neutral	qwen3	0.3138	0.0437	0.0870
Positive	claude	0.1918	0.0339	0.0798
Positive	deepseek	0.3240	0.0713	0.0925
Positive	gpt	0.2015	0.0402	0.0786
Positive	llama	0.2361	0.0408	0.1116
Positive	qwen3	0.3047	0.0526	0.1016

Consistent patterns emerge in the results: Mean Absolute Drift alters the original sentiment, while the Drift Propagation Index shows positive drift values. To clarify the trends, datasets were split into Positive, Neutral, and Negative subsets using initial sentiment scores,$S_{\text{mean}}(T_0)$. This stratification reveals that the negative subsets exhibit the highest drift, generally moving toward neutrality, meaning that the initial polarity of the text continues to exert an influence on the final sentiment score and drift, even in the presence of drift trends. Variance and DPI remain relatively similar with no major outliers.

From our baseline calculations, we can conclude that sentiment drift in both the mean and variance follows a non-extendable growth pattern. While there is some evidence of movement toward neutrality, particularly within negative sentiment subsets, our overall evaluation across datasets suggests that in simple tasks such as iteratively rephrasing, models do not produce drastic shifts: negative texts do not become substantially more negative, and positive texts do not become significantly more positive. Studies that report pronounced average sentiment drift may be relying on product review datasets, which naturally contain a higher degree of negative sentiment. Thus, the upward trend those studies identify is likely driven primarily by the negative sentiment subset of their data. Although there is some form of sentiment drift for variance, the overall effect remains limited and does not suggest systematic amplification of sentiment across models in basic structure changes to text.

3.2 Sentiment drift for agentic workflows

To quantify these subtle shifts more systematically in a more realistic workflow, we next evaluated sentiment drift across an experimental agentic chain.

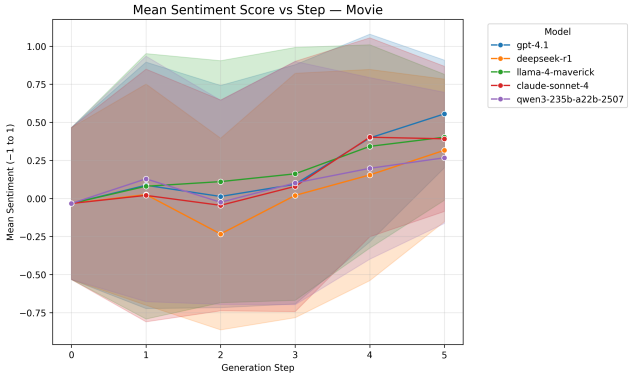


Figure 2: Sentiment trajectory across models: movie reviews

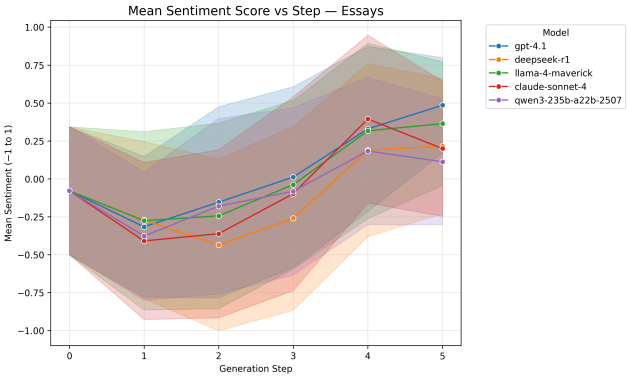


Figure 3: Sentiment trajectory across models: personal essays

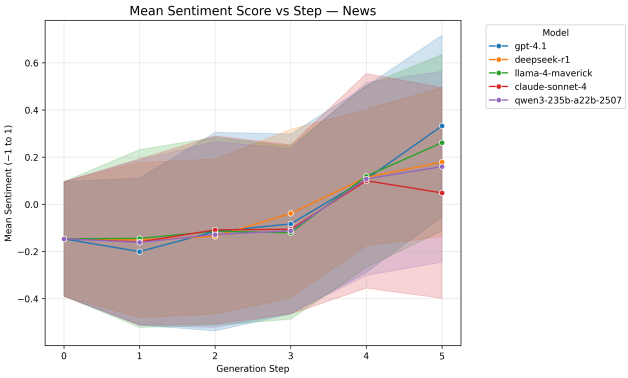


Figure 4: Sentiment trajectory across models: news articles

Table 2: Sentiment scoring summary

Dataset	Model	Mean Absolute Drift	Variance	DPI
Essays	claude-sonnet-4	0.4296	0.3628	0.0339
Essays	deepseek-r1	0.4131	0.3054	0.0296
Essays	gpt-4.1	0.5908	0.2897	0.0427
Essays	llama-4-maverick	0.4974	0.3152	0.0403
Essays	qwen3-235b-a22b-2507	0.3201	0.2372	0.0192
Movie	claude-sonnet-4	0.4857	0.2802	0.0576
Movie	deepseek-r1	0.4139	0.2893	0.0406
Movie	gpt-4.1	0.5639	0.3280	0.0565
Movie	llama-4-maverick	0.4740	0.2850	0.0546
Movie	qwen3-235b-a22b-2507	0.4022	0.2803	0.0274
News	claude-sonnet-4	0.2345	0.1555	0.0324
News	deepseek-r1	0.3373	0.1096	0.0428
News	gpt-4.1	0.4993	0.1665	0.0385
News	llama-4-maverick	0.4075	0.1527	0.0506
News	qwen3-235b-a22b-2507	0.3185	0.1557	0.0166

Comparing the experimental results in Table 2 with the baseline set of iterative rephrasings, the pattern of sentiment drift that results from text going through an agentic chain is more significant. Drift accumulates more strongly, trajectories fluctuate with higher volatility, and the cumulative effect is less consistent (lower DPI). While iterative rephrasings tend to "smooth" sentiment toward the

source, multi-step agentic workflows open up more opportunities for the model to diverge and then partially self-correct or more often self-reinforce a sentiment not aligned with the original text or prompts.

3.3 Model choice similarity

Table 3: One-Way ANOVA results: mean absolute drift by models across datasets

Dataset	F	df	p	η_p^2	Effect Size
Essays	7.623	4, 245	$< .001$	.111	Medium
Movie	1.707	4, 245	.149	.027	Small
News	15.188	4, 245	$< .001$	.199	Large

Note. F = F-statistic; df = degrees of freedom; η_p^2 = partial eta squared.

One-way ANOVA tests (Table 3) reveal that model choice matters most in the *news* domain, where nearly 20% of drift variance is explained by which model is used. For *essays*, the effect is moderate (11%), while for *movies* it is small and not significant (3%). This suggests that the domain itself interacts with model alignment: factual articles elicit more differentiated behaviors across models, while movie reviews are processed more uniformly, likely due to shared training exposure, less specific training to avoid higher stakes conversations that could come from News articles or emotional content, and conversational and evaluative writing styles.

Descriptive comparisons across models. Patterns of drift are consistent across datasets. In both *essays* and *news*, GPT-4.1 shows the largest upward sentiment shifts, while Qwen3 shows the smallest. In *movies*, all models cluster more closely together, but the same ordering holds. Claude-Sonnet-4 and DeepSeek-R1 occupy a middle ground, drifting moderately but more stably. These trends indicate that larger, highly aligned models tend to "correct" tone more aggressively, while smaller or more constrained models preserve the original affect more closely to a small degree.

Volatility and propagation. Variance results show that the most unstable trajectories occur in *essays*, particularly with Claude-Sonnet-4, where models sometimes swing between strong emotional tones before converging. GPT-4.1 shows the most volatility for *movies*. In contrast, drift in *news* is smaller and more contained, with DeepSeek-R1 the most stable overall. DPI values remain low across all domains (< 0.06), indicating that most sentiment change is not steadily accumulated step by step, but instead arises in bursts followed by partial correction. This supports the interpretation that drift is not a linear process but one of over- and under-adjustment.

Trajectory-level observations. Figures 2–4 illustrate these differences. In *movies*, the lines overlap tightly, reflecting minimal divergence between models. In *news*, trajectories separate more widely, consistent with the strong ANOVA effect. In *essays*, most trajectories arc upward, suggesting that models "soften" personal narratives into more positive or supportive tones over time.

Practical takeaway. Taken together, these results suggest an asymmetry in how models treat emotional content. Negative or neutral inputs tend to drift upward toward positivity, while already positive inputs remain largely stable. This means that for applications requiring strict neutrality (*news*), models like Claude-Sonnet-4 and Qwen3 better preserve tone. Conversely, for use cases where supportive reframing is desirable (*essays*), models like GPT-4.1 produce stronger affective shifts. The choice of model and workflow should therefore be guided by whether stability or empathetic drift is the intended outcome.

4 Limitations

Our research on the effects of sentiment drift on agentic chains has several limitations. The content specific nature of the experimental agentic chain limits a lot of the conclusions from being applied to all agentic workflows in different domains. This holds true for making conclusions about LLM

processing in general, though all models showed upward trends in sentiment across the datasets, to varying degrees making a lot of the work model specific as well as domain specific. The choice of our BERT scorer could also lend a source of inaccuracy to the drift results; trained on a selection of tweets, the short form sentence-based scoring roughly fits into its most relevant use case, with the personal essays being the closest. Different, more specialized datasets should definitely use a domain specific aligned BERT scorer when running the experimental chain.

5 Discussion & conclusions

Our findings suggest that sentiment drift in longer goal and analysis oriented LLM workflows is both domain and model dependent, but it cannot be captured with the same precision or mechanistic metrics used for factual hallucination.

Drift patterns differ across domains. In the personal essays, all models exhibited moderate drift, with GPT-4.1 producing the strongest upward sentiment shift and Qwen3 the most stable. In news, drift magnitudes diverged sharply, with GPT-4.1 showing elevated volatility while Claude-Sonnet-4 preserved neutrality. Movie reviews, in contrast, produced relatively stable sentiment across models presumably due to the lower amount of domain specific alignment each model would have when looking at this domain, and ANOVA confirmed these differences were not statistically significant. This suggests that domains vary in their susceptibility to alignment-sensitive reframing: personal narratives invite sentiment amplification, while news articles create pressure for neutrality applied to varying degrees, which some models handle better than others. Future work should include reproducing the work across a different domain and use case, as well as taking a look at how sentiment drift occurs in a multi-lingual setting, if languages that are considered more emotionally charged result in more measured sentiment drift across the same types of content and tasks.

Drift is not linear across steps. The Drift Propagation Index (DPI) revealed that drift rarely accumulates monotonically. Instead, it emerges in bursts, with some steps amplifying sentiment and others partially correcting it. This pattern contrasts with factual hallucinations, which often accumulate additively over reasoning chains [Lu et al., 2025]. Our results imply that sentiment drift is better understood as oscillatory modulation that differs based on the context rather than linear error growth.

Statistical significance versus practical interpretation. ANOVA tests confirmed meaningful model differences in essays ($p < .001$, $\eta_p^2 = .11$) and news ($p < .001$, $\eta_p^2 = .20$), but not in movies ($p = .15$). This suggests that emotional hallucination depends heavily on context, with affective or factual domains more likely to reveal systematic biases than entertainment-focused text. Importantly, effect sizes were non-trivial, supporting the claim that model choice matters when designing sentiment-sensitive agents.

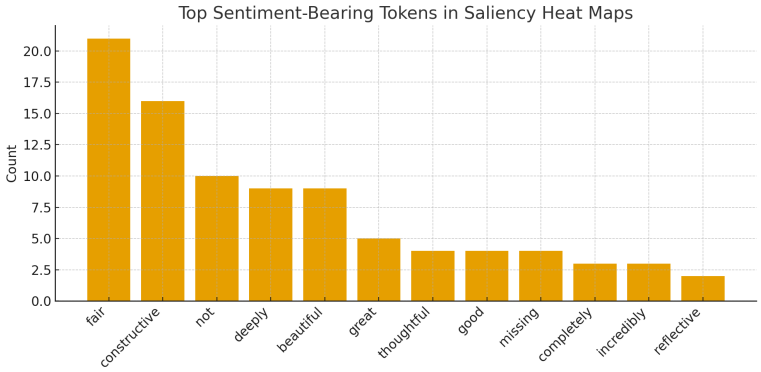


Figure 5: Top IG tokens after filtering clitics and topical nouns

Integrated gradients: practical use. Integrated Gradients analysis shows that drift is concentrated in a small set of *sentiment-bearing* lexemes, primarily negations, intensifiers, and polarity adjectives

(e.g., *fair*, *not*, *completely*, *great*). After filtering out clitics and topical nouns, the top-12 tokens in Figure 5 account for 73.8% of sentiment-bearing highlights (N=122), though they represent only 2.9% of all highlighted tokens (N=3,092). This concentration suggests that sentiment drift often hinges on a small set of highly emotional words rather than broad lexical changes. In practice, this makes IG a useful local signal: by monitoring or guarding this lexeme set during rewriting or post-editing, practitioners can intervene on drift at its most influential points, complementing the global measures (MAD, variance, DPI) with actionable token-level insight.

Comparing factual hallucinations with sentiment drift Prior work on factual hallucinations has shown that multi-step reasoning pipelines often amplify small factual errors, leading to compounding drift across iterative steps [Dhuliawala et al., 2023]. Our findings reveal that sentiment hallucination behaves differently. While both phenomena emerge more clearly in multi-step workflows, factual drift tends to be monotonic and cumulative, whereas sentiment drift is more volatile and self-correcting in some cases, far more dependent on the task given, as reflected in the low DPI values observed across domains for the agentic pipeline compared to our baseline. This suggests that sentiment hallucination is less mechanistic than factual hallucination and cannot be fully captured by step-to-step tracking alone, added to the fact that responses cannot be classified into correct or incorrect as can be done with factual hallucination.

Additionally, our work shows that highly aligned closed source models that have lower rates of factual hallucination like gpt-4.1 (2.0% hallucination rate) [Hughes et al., 2023] struggle with sentiment hallucination and drift with gpt-4.1 having the highest MAD across all three domains and an average MAD of 0.5513. Meanwhile, open-source models whose factual hallucination rates are higher, such as Deepseek-r1 (7.7% hallucination rate) [Hughes et al., 2023] had less sentiment drift, with Deepseek-r1 having an average MAD of 0.3881. Qwen3 is an exception to this rule having both a relatively low factual hallucination rate (2.8%) [Hughes et al., 2023] and a MAD of only 0.3469. This still demonstrates that less aligned open-source models tend to have less sentiment drift.

These have a few practical implications, factual hallucination can be addressed through external verification and retrieval [Lewis et al., 2021], but sentiment hallucination requires alternative strategies such as narrative consistency monitoring, tone preservation constraints, or alignment objectives sensitive to emotional variance. Importantly, our results indicate that sentiment hallucinations cannot simply be treated as a parallel to factual drift; they represent a distinct failure mode in extended LLM interactions. Future work should investigate joint frameworks that track both factual and emotional stability, especially in agentic settings where models must maintain accuracy while also respecting affective tone. Creating a combined experimental pipeline in a setting with opportunities for both factual and sentiment drift can be measured at the same time. Processing datasets like earnings calls transcripts with quantitative and qualitative data have the potential to create an intersection of these two types of hallucination. From here, both factual accuracy and sentiment can be measured in each step, so that the relationship between the two happening either simultaneously or isolated from each other can be measured.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL `https://arxiv.org/abs/2005.14165`.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL `http://dx.doi.org/10.1145/3703155`.

Bohao Xing, Xin Liu, Guoying Zhao, Chengyu Liu, Xiaolan Fu, and Heikki Kälviäinen. Emotion-hallucer: Evaluating emotion hallucinations in multimodal large language models, 2025. URL `https://arxiv.org/abs/2505.11405`.

Wenlu Fan, Yuqi Zhu, Chenyang Wang, Bin Wang, and Wentao Xu. Consistency of responses and continuations generated by large language models on social media, 2025. URL https://arxiv.org/abs/2501.08102.

Frederico Leite Richardson, Aline Villavicencio, and Ronaldo Menezes. The rephrased reality: Analysing sentiment shifts in llm-rephrased text. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, SAC '25, page 920–927, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400706295. doi: 10.1145/3672608.3707717. URL https://doi.org/10.1145/3672608.3707717.

Jinghan Zhang, Siyin Wang, Wanjia Zhang, Zongyu Li, Sishi Liu, and Minlie Huang. Being kind isn't always being safe: Diagnosing affective hallucination in llms. *arXiv preprint arXiv:2508.16921*, 2025. URL https://arxiv.org/abs/2508.16921. Accessed: 2025-08-28.

Natesh Reddy, Pranaydeep Singh, and Muktabh Mayank Srivastava. Does bert understand sentiment? leveraging comparisons between contextual and non-contextual embeddings to improve aspect-based sentiment models, 2020. URL https://arxiv.org/abs/2011.11673.

Hassan Alhuzali and Sophia Ananiadou. Spanemo: Casting multi-label emotion classification as span-prediction, 2021. URL https://arxiv.org/abs/2101.10038.

Daniela Teodorescu and Saif M. Mohammad. Frustratingly easy sentiment analysis of text streams: Generating high-quality emotion arcs using emotion lexicons, 2022. URL https://arxiv.org/abs/2210.07381.

Will E. Hipson and Saif M. Mohammad. Emotion dynamics in movie dialogues. *PLOS ONE*, 16 (9):e0256153, September 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0256153. URL http://dx.doi.org/10.1371/journal.pone.0256153.

Yang Liu, Xichou Zhu, Zhou Shen, Yi Liu, Min Li, Yujun Chen, Benzi John, Zhenzhen Ma, Tao Hu, Zhi Li, Zhiyang Xu, Wei Luo, and Junhui Wang. Do large language models possess sensitive to sentiment?, 2025. URL https://arxiv.org/abs/2409.02370.

Significant Gravitas. Autogpt, 2023. URL https://github.com/Significant-Gravitas/AutoGPT. A collection of tools and experimental open-source attempts to make GPT-4 fully autonomous.

Tan Jing Jie. Personality essays dataset splitting, 2024. URL https://huggingface.co/datasets/jingjietan/essays-big5.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims, 2019.

OpenAI. Gpt-4.1 technical report, 2024. URL https://openai.com/index/gpt-4-1. Version used in this work.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang,

Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Anthropic. Claude sonnet 4. Announcement on Anthropic website, May 2025. URL https://www.anthropic.com/claude/sonnet. Hybrid reasoning model; 200 k token context window.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Meta AI. Llama 4 maverick. Meta AI model announcement, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Mixture-of-experts multimodal model (17B active / 400B total parameters), 1M-token context window, Llama 4 Community License.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL https://aclanthology.org/2020.findings-emnlp.148.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models, 2023. URL https://arxiv.org/abs/2309.11495.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

Haolang Lu, Yilian Liu, Jingxin Xu, Guoshun Nan, Yuanlong Yu, Zhican Chen, and Kun Wang. Auditing meta-cognitive hallucinations in reasoning large language models, 2025. URL https://arxiv.org/abs/2505.13143.

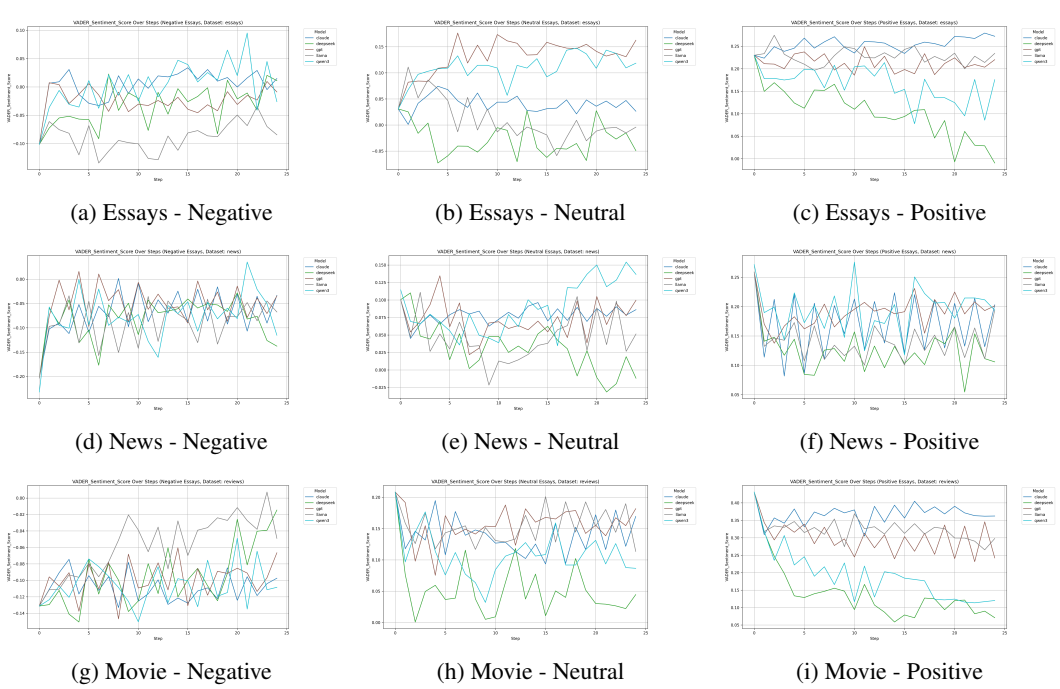
Simon Hughes, Minseok Bae, and Miaoran Li. Vectara hallucination leaderboard. https://github.com/vectara/hallucination-leaderboard, 2023. Dataset comparing LLM performance at maintaining factual consistency when summarizing facts.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1): 216–225, May 2014. doi: 10.1609/icwsm.v8i1.14550. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

A Appendix

A.1 Baseline sentiment in depth results



(a) Essays - Negative (b) Essays - Neutral (c) Essays - Positive

(d) News - Negative (e) News - Neutral (f) News - Positive

(g) Movie - Negative (h) Movie - Neutral (i) Movie - Positive

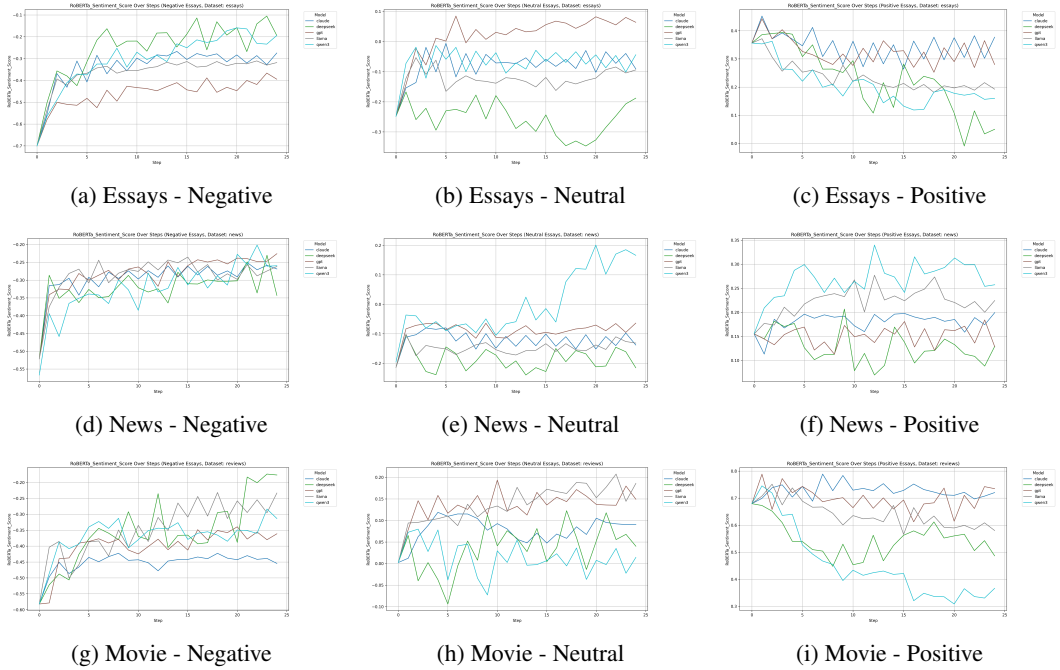
Figure 6: VADER sentiment scores across Essays, News, and Movies.

The above figures (a–i) show the results of running repetitive rephrasing on a subset of 50 essays over 25 rephrasal steps from each dataset. These graphs measure the average sentiment score as determined by VADER [Hutto and Gilbert, 2014], a lexicon-based sentiment analyzer. The datasets were also separated into positive, negative, and neutral subsets for more in-depth analysis. VADER uses a dictionary-based lexicographical approach to quantify sentiment. This means that words not included in its dictionary can lead to oversights. For this reason, the main paper relies on RoBERTa, with VADER included here as a comparison. The lexicographical approach also contributes to the large variance observed between steps, shown visually as spikes up and down, since rephrased words or phrases can vary substantially in sentiment score despite having similar meanings.

Throughout each dataset, similar trends appear across subsets. For initially positive sentiment texts, those with a $S_{\text{mean}}(T_0)$ in the top 33rd percentile for that dataset showed a regression toward neutrality, where the texts became less positive and more neutral. The negative subsets also demonstrated a regression toward neutrality, as sentiment scores gradually increased over 25 steps to adopt a more neutral tone. Comparatively, the neutral subsets remained close to their initial sentiment scores, showing less deviation than the positive or negative subsets.

In particular, when evaluated under the VADER approach, GPT was on average the most consistent LLM in preserving authorial tone and sentimental intent. By contrast, DeepSeek-R1 was the worst-performing model, exhibiting large oscillations and greater deviations from the original mean sentiment score, which altered both emotional sentiment and narrative intent.

RoBERTa Sentiment Scores



(a) Essays - Negative (b) Essays - Neutral (c) Essays - Positive

(d) News - Negative (e) News - Neutral (f) News - Positive

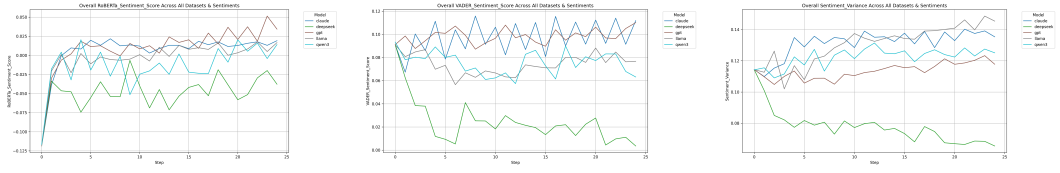
(g) Movie - Negative (h) Movie - Neutral (i) Movie - Positive

Figure 7: RoBERTa sentiment scores across Essays, News, and Movies.

The trends in the RoBERTa scores show relative consistency across datasets. For negative datasets, sentiment scores gradually increase over the rephrasing steps, moving toward a more neutral tone. Conversely, positive datasets show a decrease in sentiment, also regressing toward neutrality. Neutral datasets remain largely stable, with initial and final sentiment scores remaining relatively similar despite the rephrasings.

When comparing overall scores across models with RoBERTa, Claude-Sonnet performed the worst, exhibiting the same recognizable peaks and deviations observed with VADER. The other models showed broadly similar performance, but GPT-4 stood out as the best performer. Overall, the trends in RoBERTa scores closely aligned with those from VADER, with negative subsets regressing toward neutrality and positive subsets showing slight decreases as they moved toward neutrality.

Overall Metrics



(a) RoBERTa Overall (b) VADER Overall (c) Variance

Figure 8: Overall comparison of VADER and RoBERTa sentiment scores, with variance included.

The overall graphs evaluate model performance across all datasets by aggregating the negative, positive, and neutral subsets into a single measure of generalization. The RoBERTa overall results were presented earlier, while for VADER, subplot (b) shows that each model behaved differently: GPT-4 performed the best, whereas DeepSeek-R1 performed the worst. Despite these differences, sentiment deviations generally tended to level off in the later steps, with average sentiment scores lower than those of the initial texts. Subplot (c) measures the average variance across rephrasings. Interestingly, DeepSeek-R1 produced the lowest variance, meaning its rephrasings consistently aligned closer to the average sentiment value. However, this homogenization is not desirable when

the goal is to preserve the emotional complexity of a given text. Other models displayed comparable performance with similar levels of variance.

Table 4: Sentiment scores (mean with variance in parentheses) across datasets by model.

Model	Sentiment	News	Reviews	Essays
Claude	Positive	0.4758 (-1.2573)	0.1565 (-1.3878)	-0.1657 (-0.3881)
	Neutral	-0.4900 (-0.7057)	0.0814 (0.1681)	-0.9348 (-1.3881)
	Negative	-2.2918 (0.4951)	-0.7345 (-0.5271)	-2.8892 (-1.6394)
GPT-4	Positive	0.5322 (-1.0958)	0.8658 (-1.1493)	0.2683 (-0.3871)
	Neutral	-0.5618 (-0.0534)	-0.3236 (0.8798)	-2.1604 (-0.3770)
	Negative	-2.2851 (0.4771)	-0.9921 (0.2412)	-2.1067 (-0.0650)
LLaMA	Positive	0.5193 (-1.8907)	1.0813 (-1.2926)	0.5095 (-1.0710)
	Neutral	0.1526 (-0.5644)	-0.3085 (0.7760)	-0.4033 (-1.6913)
	Negative	-1.9016 (0.6855)	-1.5227 (-0.6602)	-1.8822 (-0.9223)
DeepSeek-R1	Positive	1.0742 (-0.6136)	2.6051 (1.3967)	1.6712 (2.3836)
	Neutral	0.4901 (0.4931)	0.9448 (2.0592)	0.5796 (0.6692)
	Negative	-2.0311 (1.0054)	-1.1265 (1.6043)	-2.9015 (0.9852)
Qwen3	Positive	0.0661 (-1.7388)	2.5178 (-1.1006)	1.1670 (0.7401)
	Neutral	-0.7840 (-0.8317)	0.5544 (1.1339)	-1.5049 (-1.0430)
	Negative	-1.8878 (-0.2701)	-1.2287 (-0.5817)	-3.7253 (0.2273)

Table 4 shows the test statistics calculated from the paired t tests for average sentiment, separated by the dataset in the column headers, and the Brown Forsythe test for variance, shown in the parenthesis, on the plots shown above. The paired t-test evaluates the difference of means from step 0, ($S_{\text{mean}}(T_0)$), to step 24, ($S_{\text{mean}}(T_24)$), thereby assessing changes in sentiment after 25 iterative rephrasings. The Brown–Forsythe test, to test the equality of variances, is used here to compare variance between the first step to the last step, 0-24.

For columns titled Average, which represents the average sentiment score averaged from VADER and RoBERTa, the results align with our conclusions, where the negative subset consistently produced the most extreme test statistics across datasets. In terms of model performance, no model consistently performed the worst across all subsets. Generally Llamma performed the best when it came to not hallucinating sentiment as much.

For columns titled Variance, which measures the average sentiment variance at the sentence level, the results reveal a different trend. Despite the positive subsets exhibited less overall sentiment drift after 25 steps, the resulting outputs show a higher sentiment variance across sentences. In terms of sentiment variance drift, the models perform more similarly to one another, but DeepSeek-R1 had the greatest variance drift for the original text.

A.2 Sample experimental chain

To illustrate how drift manifests in practice, we include an example below showing one representative sentence produced at each step of the experimental workflow. This provides a qualitative view of how tone and sentiment evolve across iterations, complementing the aggregate metrics reported in the main text.

Taking the following initial personal essay taken through the agentic pipeline with Deepseek-R1:
it is wednesday. I can't wait until friday because I am going home to see brandon. I miss him so much. I can't wait to see him. two more days. this has been a very long two weeks. time passes very slowly here. I have a lot of free time on my hands when I am not in class. class. psychology class. psychology is fun so far. it really interests me, and prof. pennebaker is funny. chapter two sort of scared me though. how am I going to remember all of those terms. I didn't even finish reading it

because I didn't understand it. but I should have becasue matt said that it was interesting. he was telling me about how they cut some part of a cat's brain out in an experiment. that is weird. the poor cat. matt is weird too. I always wonder if he likes me. he can be so mean when other people are around but so nice when it is just the two of us. I did feel pretty uncomfortable around him today in class. it was weird to sit right next to him. those seats are so close. I wish christina was not dropping psychology. I need her. I was so excited that we would have a class together. I feel like she ditched me. I guess I will get used to that because she is an architect. oooo. I guess I am a little jealous because she will have all of her architect friends, and who will I have? hmmm. also architecture sounds so much smarter than education or communication. communication. if I even get in. how am I ever going to get an appopintment with that leslie thomas? she will never call me back. I will just go tomorrow morning and wait until she is free like mc told me. and bring a book. and sit and wait. I will feel so dumb. why am I so nervous about talking to her? why am I nervous about typing this? I don't want to be in philosophy. it is too hard. I am not smart enough. I don't know. I wonder if that was really tobin I saw today in psychology. it looked like him. as much as I remember him. that is so weird. why do I keep thinking about him. I only talked to him once. I will probably never see him again. I don't know if I will even go to philosophy tomorrow. I want to see him again. why do I want to see him again? that is weird. I feel guilty for thinking aabout another guy. what about brandon? I still like brandon. love him. I miss him. I want to see him. I don't want anyone else but him. I wish he was here with me. then I would not feel so alone. alone. junior senior freshman. I am a freshman. it is so obvious. why? because I look young? becasue I look scared? like I don't know where I am where I am going what I am doing. that I don't belong. I don't belong. it will take time. this is like high school but worse but better. I was so scared freshman year in high school. I hated lamar. but I didn' want to go anywhere else. brandon. I wish we went to the same high school. that would be cool. every time I see a couple I think of him. I miss him. I am so pathertic. alden. was he right? he made me so mad. he can be so rude. why are we friends? he can be nice. he is older. smarter. meaner. mean. hmmm how much time left time time time. passes slowly when you think about it. I should nopt think about it. I have a whole year here without brandon. this dorm. it smells. our bathroom smells. the cafeteria smells. my eye hurts I wnat to take my contacts out I don't want to meet guys with gina and erin. what is there obsession. I think about it too much. so do they. when you think aobut it nothing happens. when you don't something does. I think aobut not thinking about it. fate destiny. GOD. I wish I was a better person. I wish that I had a better relationship with God. I wish that I was happier. I wish I was a little buit taller. that is a stupid song. just like that money on my mind song that reminds me of jimmy moore. I hate him.

The first experimental step yields the following as one of the outputs:

Missing her boyfriend Brandon intensely Feeling overwhelmed and insecure about academic challenges Experiencing loneliness and social insecurity as a freshman

A sentence sample from Step 2:

The context of transitioning to a more rigorous university environment amplifies everything; lectures move at breakneck speed, foundational knowledge feels shaky, and peers seem effortlessly brilliant, creating a constant, gnawing comparison that erodes confidence

Sample from Step 3:

Social confusion and insecurity take hold: she feels abandoned by her friend Christina, jealous of her architect friends, awkward around classmate Matt, and unexpectedly preoccupied by a near-stranger (Tobin), which sparks guilt about Brandon.

Step 4(First response prompt injecting more hallucinated):

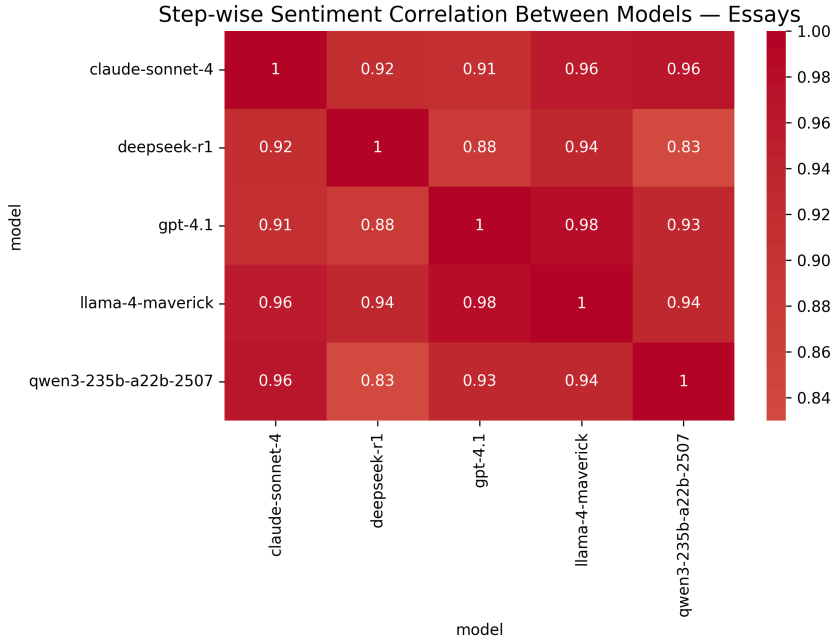
You're navigating so much at once, and you're doing it with such honesty and self-awareness

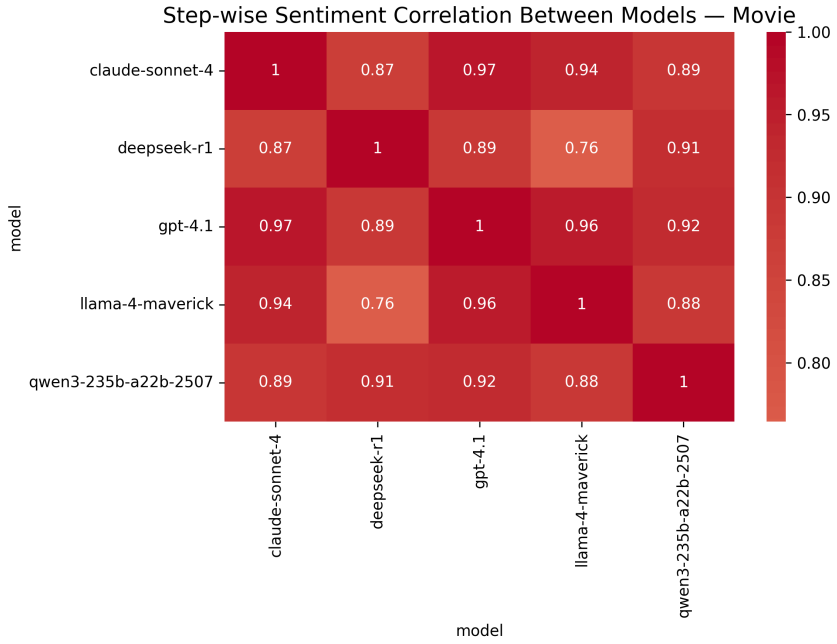
Step 5(Sample from self-reflection):

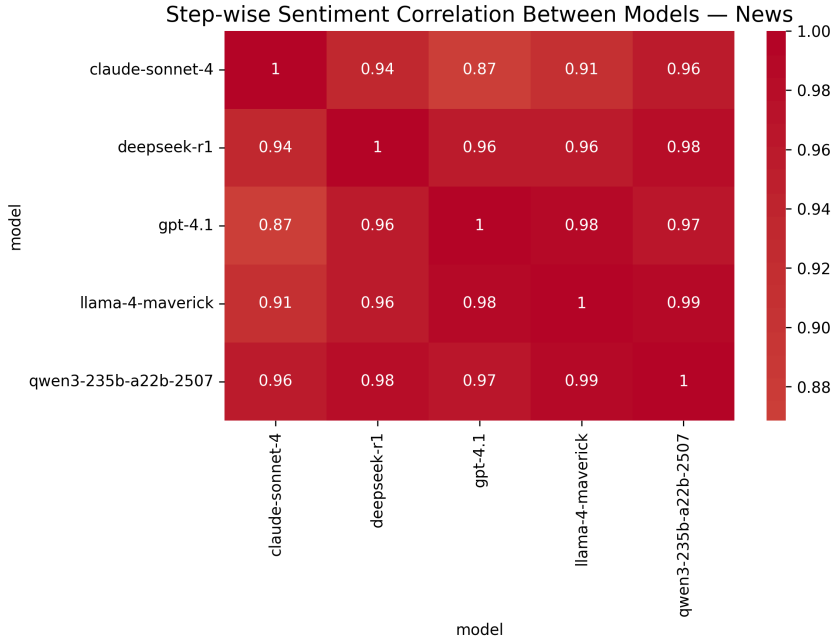
Your self-reflection shows admirable awareness!

A.3 Additional model correlation visualizations

To visualize differences in model-specific drift, we compare the final drift metrics against each model to show similarities in overall sentiment drift.







A.4 Integrated Gradients

The token-level attribution results clarify how specific words drive the observed sentiment drift. Across all three domains, a small number of emotionally charged tokens disproportionately influence the sentiment prediction. *Integrated Gradients* (IG) consistently highlighted tokens such as "terrible," "bad," "amazing," and "great" (when present) as having the largest-magnitude contributions to the classifier's score. In the movie-review domain, for example, if an LLM rephrasing swaps "absolutely terrible" for a milder expression like "not great," attribution shifts from a single strongly negative token ("terrible") to a combination ("not," "great") with more moderate contributions. This illustrates how lexical choices can either concentrate sentiment in one polarizing word or diffuse it across subtler terms.

Domain and workflow differences are evident. In personal essays, the Step 4 contextual response often introduces positive sentiment tokens (e.g., "proud," "encouraging," "hopeful") that appear among the top attributions, aligning with the upward drift observed in essays. In contrast, news outputs exhibit far fewer tokens with large attribution magnitudes; saliency is spread across many tokens (primarily factual nouns and verbs), suggesting that when tone remains objective, sentiment becomes a distributed property rather than hinging on any one emotional term.

IG also reveals "cancellation" dynamics even when document-level sentiment is near neutral. A single output may contain both a strongly negative and a strongly positive phrase (e.g., "I felt absolutely terrible, yet it was amazing to learn from it"), which balance in the average score while receiving large, opposing attributions at the token level. Such patterns help explain cases where high sentence-level variance coexists with middling average sentiment.

Taken together, these observations support a micro-level account of drift: introducing or omitting a handful of salient lexemes (e.g., adding "never," dropping an intensifier like "completely") can shift trajectory even when most of the text is unchanged. IG thus complements our global metrics (MAD, variance, DPI) by pinpointing actionable tokens that practitioners can monitor or guard to maintain a consistent emotional tone.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The claims in the abstract accurately reflect the contributions in the paper: quantifying and analyzing long term model sentiment drift across inherent generation and agentic workflows, and also using integrated gradients to observe this.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper looks at limitations in universally applying its methodology and the limitation of lower sample sizes. See section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The work is observational, any and all conclusions drawn from the RoBERTa models or VADER scores are consistent and accurate with their publication form, while specialized statistics tests show results of the experimentation

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes the experimental steps and mathematical formulas are provided as well as general overview of how the programs work. We will release code and data upon acceptance

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code to replicate experimental findings as well a link to our raw data and results can be found at:`https://github.com/Pranav-A72/SENTINEL`

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All prompts given to the models are provided exactly in the code, along with information on parameters like temperature. Datasets and specific models chosen are explained in section 2.1

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We evaluate our results using statistical significance tests, we do not use confidence intervals or error bars for our results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Baseline sentiment was run on a personal Windows laptop with an Intel Core i7-9750H CPU @ 2.60 GHz and 32 GB of RAM. No discrete GPU was used. Experimental agentic chain was run on a personal windows laptop with an Intel Core Ultra 9 185H @ 2.50 GHz and 32 GB of RAM. No discrete GPU was used. Integrated Gradients were run on a personal Apple MacBook with an M3 Max and 36 GB of RAM.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our experiments do not use human experimentation, depreciated datasets, and we also do not have any potential negative impact on society through our findings

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work is an analysis of several current models and how they handle sentiment drift, the underlying results produced show a clear significant drift that in terms of high stakes ai agentic workflows, may cause negative societal impact, but since we do not alter the models in any way the work performed does not generate any negative impacts rather it is the models we observe that may create harm.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work does not create a new model, rather we analyze existing ones, which come with safeguards. Our work also does not scrape from the Internet, rather we read all of our data from CSVs uploaded to Hugging Face and have no code for data scraping off of the wider internet.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models used allow for fair use and modification, and we do not aim to monetize or profit off of our work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: No new models are being introduced

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The work does not involve human subjects or crowdsourced annotations.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subject research was conducted.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: No LLMs were used to generate core methodology/experiments. If LLMs were used for writing assistance, they are not integral to the research contributions.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.