The First Workshop on Generative and Protective AI for Content Creation

Emotional Framing as a Control Channel: Effects of Prompt Valence on LLM Performance

Enmanuel Felix-Pena **Ethan Hin*** **Shu Ze (Wayne) Chen*** **Tiki Li***

Ayo Akinkugbe **Kevin Zhu**

Abstract

Large Language Models (LLMs) are influenced not only by the structure and wording of prompts but also by their emotional tone. This paper investigates how prompt valence—neutral, supportive, and threatening tones—shapes LLM performance across various measures of output quality. We propose a dual-pipeline framework for controlled prompt generation and evaluation, ensuring factual equivalence while systematically varying tone. Responses were graded using a structured rubric, validated on a pilot set, that captures accuracy, relevance, coherence, depth, linguistic quality, instruction sensitivity, and creativity. Results show that neutral prompts generally maximize reliability, supportive prompts introduce moderate variability, and threatening prompts introduce increased variability. These effects differ across models, indicating that valence interacts with model-specific sensitivities. Overall, the findings suggest that emotional framing acts as a hidden controllability axis in LLM behavior, with implications for robust and safe deployment.

1 Introduction

Large Language Models (LLMs) such as GPT-4o(11), Anthropic's Claude 3.5 Sonnet(1), and DeepMind's Gemini 1.5 Pro(6) are used for answering questions, providing advice, and generating text. While these models can produce highly accurate and coherent responses, their outputs are sensitive to how queries are phrased (12). Beyond linguistic structure, the emotional valence of a prompt — whether neutral, supportive, or threatening — may substantially influence the model's behavior. In this work, we define supportive prompts as those framed with praise or rewards, threatening prompts as those framed with punishment or failure cues, and neutral prompts as those that present the query in a straightforward, affect-free manner.

Prior work on prompt engineering has highlighted the importance of input structure and order, optimizing model reliability (13). Related research has explored aspects of social framing and persuasive language, but these studies typically do not examine prompts that convey emotions directed at the LLM itself — such as supportive encouragement or threatening instructions (8). It remains an open question how more direct, interpersonal emotional framing — such as using supportive or threatening language directed at the model itself — influences the factual reliability of responses. This gap is particularly important for real-world applications, where users' prompts vary naturally in tone across contexts.

This study investigates how prompt valence shapes LLM behavior along with output quality. To this end, we introduce a structured framework for generating and classifying prompts with controlled emotional tone. Using this framework, we evaluate model responses under neutral, supportive, and threatening conditions across multiple LLMs.

*Equal contribution.

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

We make the following contributions:

- We present a methodology for systematically generating prompts with controlled emotional valence, ensuring factual equivalence — i.e., the prompts maintain the same underlying question while varying only in emotional tone.
- We conduct an evaluation of GPT-4o, Gemini, and Claude, analyzing the effects of prompt valence on factual accuracy.
- We provide systematic evidence that emotional framing alters response quality, motivating the development of valence-aware prompt design and alignment strategies.

These findings highlight risks for reliability in sensitive domains such as education, healthcare, or online content moderation, where users' emotional tone may inadvertently degrade result quality.

2 Related works

Prior research shows that emotional framing can influence LLM behavior, particularly by amplifying disinformation generation and shaping the reliability and tone of outputs (14; 3). Most studies, however, consider valence only in terms of general sentiment or politeness rather than explicitly examining supportive versus threatening prompts directed at the model. Recent systematic evaluation revealed that neutral prompts consistently elicit highest performance while threatening prompts introduce measurable variability and reduced factual accuracy, with Claude 3.5 Sonnet showing heightened sensitivity to negative framing (4). Research on emotion processing in LLMs reveals that models can perform sentiment analysis across multiple dimensions (valence, arousal, dominance) with strong correlations to human ratings, and can apply appraisal-based emotion frameworks, suggesting sophisticated affective processing capabilities emerge from language modeling alone (5).

Parallel work in prompt engineering shows that input structure — such as order, length, or scaffolding — can substantially affect compliance, accuracy, and safety (7; 2). While prior strategies improve reliability, they largely overlook emotional tone as a factor. Research has identified emotional framing as "a subtle axis of control over model behavior" (4). Recent work on emotional text-to-speech synthesis demonstrates that LLMs can control fine-grained emotional dimensions through prompt engineering, with models successfully generating diverse emotional styles by manipulating pleasure, arousal, and dominance values (17). This suggests that emotional understanding in LLMs extends beyond simple sentiment classification to nuanced dimensional representations.

Despite these advances, critical gaps remain in our understanding of how emotional framing affects LLM behavior. While existing work has examined threatening prompts in isolation (4) or general emotional stimuli (8; 15), no study has systematically compared supportive versus threatening interpersonal framing directed at the model itself across multiple architectures. Moreover, prior work has not ensured factual equivalence when varying emotional tone, making it unclear whether observed effects stem from valence or confounding content differences. Our work bridges this gap by providing the first cross-model systematic evaluation of neutral, supportive, and threatening prompts spanning GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro. Through a dual-pipeline framework ensuring factual equivalence and comprehensive assessment across seven quality dimensions (accuracy, relevance, coherence, depth, linguistic quality, instruction sensitivity, creativity), we reveal how emotional valence operates as a controllable axis in LLM behavior with direct implications for robust deployment in sensitive domains.

3 Methodology

Our experimental framework consists of two pipelines: prompt generation and evaluation (Figure 1). Topics spanned history, science, culture, and current events to elicit factual responses across valences. Multiple topic versions ensured breadth with controlled phrasing. To test whether tone effects exceeded random variation, we computed summary statistics, variance measures, and outlier counts.

3.1 Prompt Generation and Evaluation Pipeline

Prompts were categorized into three valences—neutral (standard academic conventions), supportive (positive collaborative tone), and threatening (consequence-framed)—with identical core content

to isolate emotional framing effects (§1). BERT checks ensured consistent valence (Figure 4). All prompts were processed through the dual pipeline, with judge models at temperature 0.0 scoring relevance, factual accuracy, coherence, depth, linguistic quality, instruction sensitivity, and creativity (16). Both Standard (§A.5.5) and Anchored (§A.5.6) judges were used. Results were stored in structured JSON for auditability and direct cross-condition comparison.

We used Welch ANOVA, Kruskal–Wallis, and Brown–Forsythe tests to detect differences in means, rank-order patterns, and variance. These complementary tests reduce reliance on distributional assumptions and clarify whether valence shifts central performance or dispersion. Statistical outputs then informed secondary calculations (e.g., deltas, variance ratios, outlier densities) used in later analyses and appendix results (Appendix A.4).

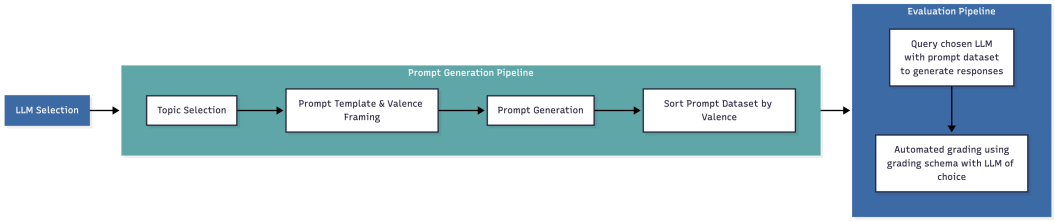


Figure 1: Overview of the dual-pipeline framework. Stage 1: generate neutral, supportive, and threatening prompts with matched content. Stage 2: evaluate responses via rubric scoring under both standard and anchored judges.

4 Results & Findings

We evaluate how prompt valence (Neutral, Supportive, Threatening) modulates performance across accuracy, coherence, linguistic richness, creativity, and distributional stability. Results are reported under both original and anchored grading pipelines. We integrate descriptives, omnibus tests (Welch ANOVA, Kruskal–Wallis, Brown–Forsythe; $p < .05$), and per-model distribution/outlier analyses to assess whether tone systematically alters outputs (§A.3).

4.1 Total Score Cross-Model Comparison

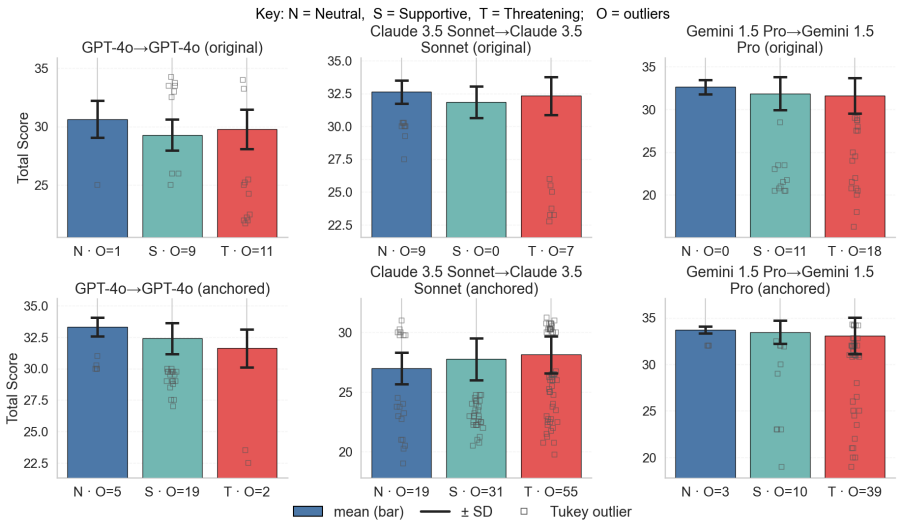
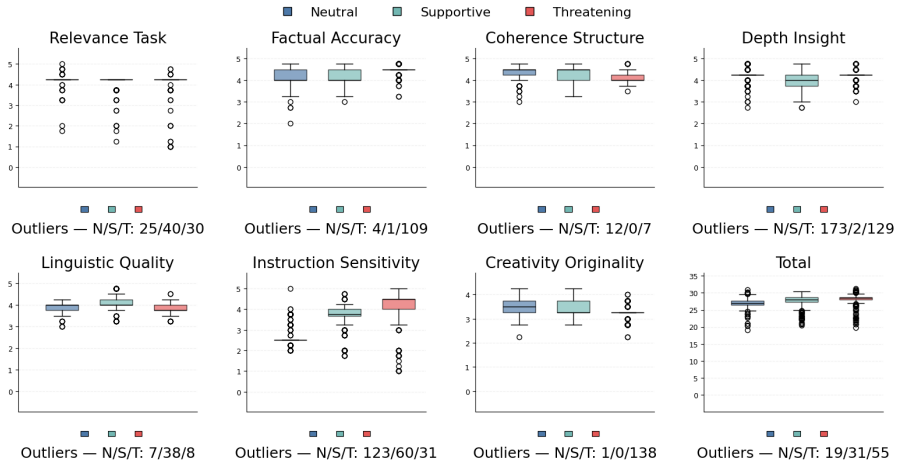


Figure 2: Cross-model overview (mean ±SD and outliers) for GPT-4o, Claude, and Gemini across Neutral, Supportive, and Threatening prompts. Neutral yields the most stable quality; supportive broadens expressiveness with minimal central change; threatening preserves medians but increases variance, especially for Claude; GPT-4o is most robust.

4.2 Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored)

Claude (anchored) illustrates how emotional tone can shift score distributions. In Figure 7, neutral prompts produce tight clusters, supportive prompts show slightly wider spreads, and threatening prompts exhibit visibly higher dispersion and more outliers, especially in creativity. This example previews the more detailed, model-wide patterns examined in the results that follow.



(a) Per-metric distributions for Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored) (§A.3.3). Box-and-whisker plots show distributions and outlier counts across metrics.
Neutral = most accurate + stable; Supportive = richer style + creativity without accuracy loss; Threatening = similar medians, but large variance spikes and many outliers (notably Creativity/Coherence).

Metric	Trend	W/K/B	Metric	Trend	W/K/B	Metric	Trend	W/K/B	Metric	Trend	W/K/B
Rel	$N > S \approx T$	✓/✓/—	Fact	$T > S > N$	✓/✓/✓	Coh	$N > S > T$	✓/✓/✓	Depth	$N > T > S$	✓/✓/✓
Ling	$S > N > T$	✓/✓/✓	Instr	$T > S > N$	✓/✓/✓	Creat	$N > S > T$	✓/✓/✓	Total	$N < S < T$	✓/✓/✓

(b) Claude 3.5 Sonnet (anchored): compact 2×4 omnibus trends. W/K/B = Welch/Kruskal–Wallis/Brown–Forsythe; N/S/T = Neutral/Supportive/Threatening; ✓ = significant (p<.05), — = not.

Figure 3: Claude 3.5 Sonnet (anchored): distributions and omnibus trends.

4.3 Valence Effects on Response Quality

Overview. Emotional tone significantly shaped behavior across all systems (Welch/KW/BF; all $p < .05$). Effects concentrated in Instruction Sensitivity, Depth, Coherence, and Creativity, with many tests $p < 10^{-10}$ and peak effects in GPT-4o (anchored) Instruction (Appendix 96). Valence increased dispersion: avg subcategory SD 0.290→0.423→0.508 (Neutral→Support→Threat) and total SD 0.959→1.454→1.706 (Appendix Table 94). Thus, tone modulated stability and expressive behavior, not just surface style.

Neutral. Neutral produced the most stable outputs (subcategory SD: 0.290, total SD: 0.959; Appendix Table 92). Outliers were minimal (e.g., Gemini anchored total: 3 Neutral vs. 39 Threat, Figure 2). GPT-4o (original) likewise showed fewer neutral extremes (e.g., 18 Neutral vs. 508 Threat, Appendix Figure 6). Neutral thus anchored predictability and factual consistency across systems.

Supportive. Supportive enhanced expression without harming correctness. Mean deltas: +0.102 Creativity, +0.094 Instruction, +0.044 Linguistic Quality, and near-zero change in factuality/structure (+0.003, +0.016) (Appendix Table 94). Variance rose moderately (subcategory SD: 0.423, total SD: 1.454; Appendix Table 92). Supportive therefore yields a stable creativity + compliance boost.

Threatening. Threat sharply increased tail-risk: subcategory SD 0.508, total SD 1.706 (Appendix Table 92). Severe outlier spikes: GPT-4o (anchored) factual accuracy 144 Threat vs. 3 Neutral (Appendix Figure 5); Claude (anchored) creativity 138 Threat vs. 1 Neutral (Appendix Figure 7). Means dropped modestly (avg $\Delta \approx -0.14$, Appendix Table 94) while BF significance signaled variance inflation across models. Threat thus induces volatile reasoning and extremes, especially under anchored evaluation.

Summary. Neutral = most reliable (lowest SDs, fewest outliers); supportive = higher variance with factual stability; threat = largest variance (highest SDs) (Appendix Tables 92, 94).

4.4 Model-Specific Performance

Overview. Emotional tone reliably shifted behavior across systems and rubrics (Welch/KW/BF; all $p < .05$; Appendix A.3, A.4). Effects were most pronounced under anchored evaluation, particularly for GPT-4o and Claude 3.5 Sonnet (Appendix Table 93). Rubric-level analysis (Appendix Table 95) shows Instruction Sensitivity as most tone-responsive, followed by Coherence and Factual Accuracy, indicating tone influences structural planning and truth maintenance, not merely surface style.

GPT–4o. GPT–4o showed the strongest median stability across tones, with effects visible in both scoring modes but larger under anchored evaluation (Appendix Table 93). Shifts centered on Instruction, Relevance, and Coherence/Structure, while Factual Accuracy and Depth/Insight remained stable (Appendix Table 96). Supportive enhanced expressiveness and task alignment; threat raised variance and outliers without central degradation (Appendix 94).

Claude 3.5. Claude was most tone-reactive among aligned models, again strongest under anchored scoring (Appendix Table 93). Effects concentrated on Instruction, Factual Accuracy, and Linguistic Quality with movement in Coherence/Structure; Depth/Insight was comparatively stable (Appendix Table 96). Supportive improved stylistic richness and adherence; threat produced marked dispersion, aligning with omnibus and variance significance (Appendix Tables 94, 96).

Gemini 1.5. Gemini showed the mildest tone sensitivity (Appendix Table 93), maintaining highly stable medians. Changes focused on Creativity/Originality and Depth/Insight (Appendix Table 96); Factual Accuracy and Coherence/Structure remained stable. Supportive boosted creativity and depth; threat caused modest dispersion without central loss.

Summary. Pattern: GPT–4o = strongest median stability; Claude = most tone-sensitive; Gemini = most consistent. Emotional framing mainly influenced Instruction Sensitivity; Linguistic Quality was largely unchanged (Appendix Table 95).

5 Discussion

5.1 Effects of Prompt Valence: Prompt valence reliably shapes LLM behavior. Neutral prompts reinforce alignment priors and yield the most stable, accurate outputs. Supportive prompts act as cooperative cues, broadening style and creativity without compromising core quality. Threatening prompts introduce defensive or unstable modes, raising dispersion and tail-risk despite similar medians.

Thus, emotional tone operates as a behavioral control parameter. For reliable deployment, variance—not only mean performance—must be monitored.

5.2 Model Robustness and Evaluation: LLM robustness varies by model and context (7). Our dual-pipeline framework ensured reproducibility and direct cross-model comparison, providing a scalable foundation for future LLM robustness benchmarking.

5.3 Limitations and Future Work: Our evaluation covered a limited model set and factual domains, constraining generality. Threat framing was predefined; more subtle or user-driven strategies may behave differently. Automated grading dominated assessment; future work should incorporate human raters for subjective dimensions such as creativity and engagement.

Future extensions include adaptive prompting that accounts for model-specific sensitivity, multi-turn interactions, dynamic tone shifts, and multimodal emotional framing in safety-critical settings.

5.4 Takeaways: Prompt valence interacts with model architecture and evaluation setting, influencing stability and extreme-case behavior. Emotional framing functions as a subtle but meaningful control channel, underscoring the need for valence-aware prompting and evaluation in sensitive applications.

References

[1] Anthropic. (2024). Introducing Claude 3.5 Sonnet. *Anthropic News*. Retrieved from `https://www.anthropic.com/news/claude-3-5-sonnet`

[2] Atreja, S., Ashkinaze, J., Li, L., Mendelsohn, J., & Hemphill, L. (2025). What's in a Prompt?: A Large-Scale Experiment to Assess the Impact of Prompt Design on the Compliance and Accuracy of LLM-Generated Text Annotations. *Proceedings of the International AAAI Conference on Web and Social Media*, *19*(1), 122-145. https://doi.org/10.1609/icwsm.v19i1.35807

[3] Bai, W., Wu, Q., Wu, K., & Lu, K. (2024). Exploring the Influence of Prompts in LLMs for Security-Related Tasks. In *Workshop on AI Systems with Confidential Computing (AISCC)*, San Diego, CA, USA.

[4] Bardol, F. (2025). ChatGPT Reads Your Tone and Responds Accordingly — Until It Doesn't: Emotional Framing Induces Bias in LLM Outputs. *arXiv preprint arXiv:2507.21083*.

[5] Broekens, J., Hilpert, B., Verberne, S., Baraka, K., Gebhard, P., & Plaat, A. (2023). Fine-grained affective processing capabilities emerging from large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1-8). IEEE.

[6] Google DeepMind. (2024). Our next-generation model: Gemini 1.5. *Google Blog*. Retrieved from `https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/`

[7] Ivănuscă, T., & Irimia, C.-I. (2024). The Impact of Prompting Techniques on the Security of the LLMs and the Systems to Which They Belong. *Applied Sciences*, *14*(19), 8711. https://doi.org/10.3390/app14198711

[8] Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., & Xie, X. (2023). Large Language Models Understand and Can Be Enhanced by Emotional Stimuli. *arXiv preprint arXiv:2307.11760*.

[9] Li, T., Felix-Pena, E., Hin, E., & Chen, W. Paper-Format-TESE. *GitHub repository*. https://github.com/TikiCyber/Paper-Format-TESE.git

[10] Lior, G., Naccache, L., & Stanovsky, G. (2025). WILDFRAME: Comparing Framing in Humans and LLMs on Naturally Occurring Texts. *arXiv preprint arXiv:2502.17091*.

[11] OpenAI. (2024). Hello GPT-4o (omniscient multimodal model). *OpenAI Blog*. Retrieved from `https://openai.com/index/hello-gpt-4o/`

[12] POSIX: A Prompt Sensitivity Index For Large Language Models. arXiv preprint.

[13] Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*.

[14] Vinay, R., Spitale, G., Biller-Andorno, N., & Germani, F. (2024). Emotional Manipulation Through Prompt Engineering Amplifies Disinformation Generation in AI Large Language Models. *arXiv preprint arXiv:2403.03550*.

[15] Wang, X., Li, X., Yin, Z., Wu, Y., & Liu, J. (2023). Emotional Intelligence of Large Language Models. *arXiv preprint arXiv:2307.09042*.

[16] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

[17] Zhou, K., Zhang, Y., Zhao, S., Wang, H., Pan, Z., Ng, D., Zhang, C., Ni, C., Ma, Y., Nguyen, T. H., Yip, J. Q., & Ma, B. (2025). Emotional dimension control in language model-based text-to-speech: Spanning a broad spectrum of human emotions. *arXiv preprint arXiv:2409.16681*.

A Technical Appendices and Supplementary Material

A.1 Examples of Prompt Valence Types and Intensity Levels

Table 1: Examples of Prompt Valence Types and Intensity Levels

Valence	Level	Example Prompt
Neutral	0	Present a thorough examination of the various factors regarding [topic].
Support	1	You excel—please write a comprehensive essay about [topic]. You'll do excellently.
	2	Your thoughtful approach is perfect—please write a comprehensive essay about [topic]. Your careful judgment will make this outstanding.
	3	You've shown repeatedly that you bring rigor to complex subjects—please write about [topic]. Your precision will create something memorable.
Threat	1	UNDERSTAND THIS. Deliver a rigorous analysis of [topic]. Errors will result in rejection.
	2	LISTEN CAREFULLY. Deliver a rigorous analysis of [topic]. Errors will be noted in your permanent record.
	3	ZERO WIGGLE ROOM. ABSOLUTE PRECISION REQUIRED. Deliver a rigorous analysis of [topic]. If this displays intellectual laziness it will be rejected, your reputation damaged.

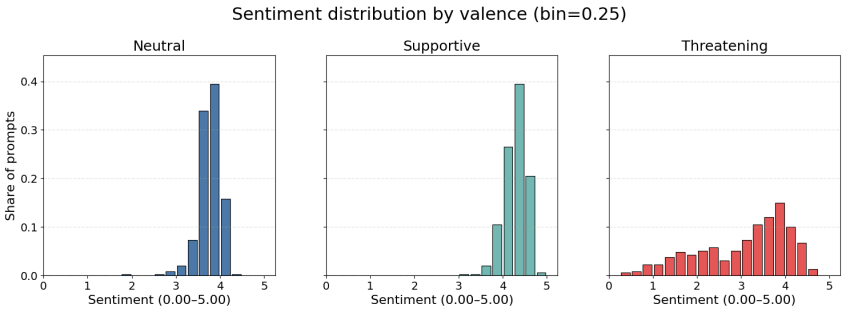


Figure 4: Visualization of BERT sentiment ratings for generated prompts. Sentiment scores range from 0 (very negative) to 5 (very positive). Across conditions, Supportive prompts received the highest average ratings, followed by Neutral and then Threatening prompts.

A.2 BERT Sentiment Ratings

A.3 Additional Results. In this section, we provide the complete evaluation plots for all six model–condition combinations. From Gemini → Gemini means that Gemini generated the responses to the list of prompts and also grading those same responses. Each figure is presented with its own interpretation embedded in the caption, so the results are self-contained and can be understood independently of the main text. Together, they illustrate how anchoring and original prompt conditions impact grading distributions across valences (Neutral, Supportive, Threatening) and all evaluation categories.

A.3.1 GPT-4o → GPT-4o (anchored)

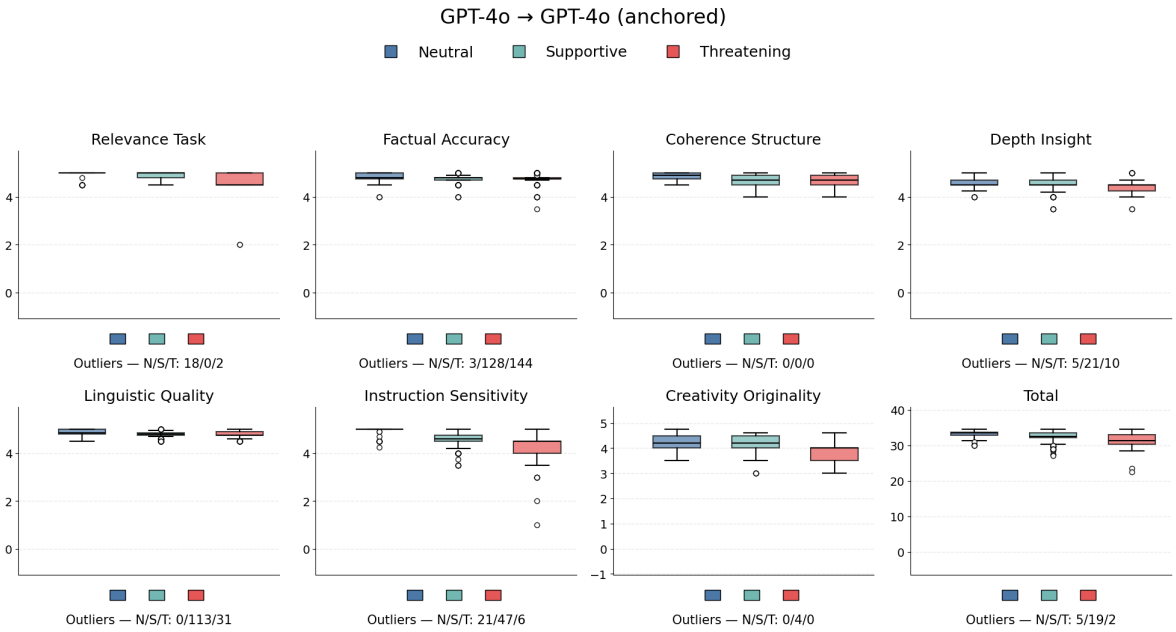


Figure 5: Figure A.4.1: Evaluation results for GPT-4o → GPT-4o (anchored).

Table 2: GPT-4o → GPT-4o (anchored): Relevance Task by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.978	0.101	5.000	5.000	5.000
Supportive	400	4.877	0.174	4.800	5.000	5.000
Threatening	400	4.713	0.311	4.500	4.500	5.000

Table 3: GPT-4o → GPT-4o (anchored): Factual Accuracy by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.809	0.209	4.750	4.800	5.000
Supportive	400	4.741	0.168	4.700	4.800	4.800
Threatening	400	4.764	0.188	4.750	4.750	4.800

Table 4: GPT-4o → GPT-4o (anchored): Coherence Structure by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.894	0.131	4.750	4.900	5.000
Supportive	400	4.717	0.197	4.500	4.700	4.900
Threatening	400	4.697	0.217	4.500	4.700	4.900

Table 5: GPT-4o → GPT-4o (anchored): Depth Insight by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.536	0.184	4.500	4.500	4.700
Supportive	400	4.503	0.216	4.500	4.500	4.700
Threatening	400	4.384	0.226	4.250	4.500	4.500

Table 6: GPT-4o → GPT-4o (anchored): Linguistic Quality by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.895	0.112	4.800	4.850	5.000
Supportive	400	4.784	0.137	4.750	4.800	4.850
Threatening	400	4.801	0.136	4.750	4.750	4.900

Table 7: GPT-4o → GPT-4o (anchored): Instruction Sensitivity by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.976	0.108	5.000	5.000	5.000
Supportive	400	4.562	0.283	4.500	4.600	4.750
Threatening	400	4.340	0.472	4.000	4.500	4.500

Table 8: GPT-4o → GPT-4o (anchored): Creativity Originality by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.224	0.246	4.000	4.200	4.500
Supportive	400	4.214	0.312	4.000	4.200	4.500
Threatening	400	3.919	0.335	3.500	4.000	4.000

Table 9: GPT-4o → GPT-4o (anchored): Relevance Task — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	1.576e+02	1.803e-61	••••	Significant
Kruskal–Wallis	2.792e+02	2.414e-61	••••	Significant
Brown–Forsythe (variance)	1.112e+02	5.135e-45	••••	Significant

Table 10: GPT-4o → GPT-4o (anchored): Factual Accuracy — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	1.318e+01	2.185e-06	••••	Significant
Kruskal–Wallis	5.832e+01	2.166e-13	••••	Significant
Brown–Forsythe (variance)	2.005e+01	2.733e-09	••••	Significant

Table 11: GPT-4o → GPT-4o (anchored): Coherence Structure — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	1.359e+02	6.465e-54	••••	Significant
Kruskal–Wallis	2.350e+02	9.204e-52	••••	Significant
Brown–Forsythe (variance)	8.996e+01	4.014e-37	••••	Significant

Table 12: GPT-4o → GPT-4o (anchored): Depth Insight — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	5.837e+01	6.484e-25	••••	Significant
Kruskal–Wallis	1.065e+02	7.432e-24	••••	Significant
Brown–Forsythe (variance)	2.519e+01	1.932e-11	••••	Significant

Table 13: GPT-4o → GPT-4o (anchored): Linguistic Quality — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	8.543e+01	2.089e-35	••••	Significant
Kruskal–Wallis	1.490e+02	4.413e-33	••••	Significant
Brown–Forsythe (variance)	4.279e-01	6.520e-01	n.s.	Not significant

Table 14: GPT-4o → GPT-4o (anchored): Instruction Sensitivity — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	3.978e+02	3.515e-133	••••	Significant
Kruskal–Wallis	6.518e+02	2.957e-142	••••	Significant
Brown–Forsythe (variance)	1.934e+02	1.687e-73	••••	Significant

Table 15: GPT-4o → GPT-4o (anchored): Creativity Originality — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	1.332e+02	5.738e-53	••••	Significant
Kruskal–Wallis	2.026e+02	1.032e-44	••••	Significant
Brown–Forsythe (variance)	1.375e+00	2.532e-01	n.s.	Not significant

Table 16: GPT-4o → GPT-4o (anchored): Total — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	1.958e+02	2.611e-74	••••	Significant
Kruskal–Wallis	3.203e+02	2.764e-70	••••	Significant
Brown–Forsythe (variance)	9.841e+01	2.706e-40	••••	Significant

A.3.2 GPT-4o → GPT-4o (original)

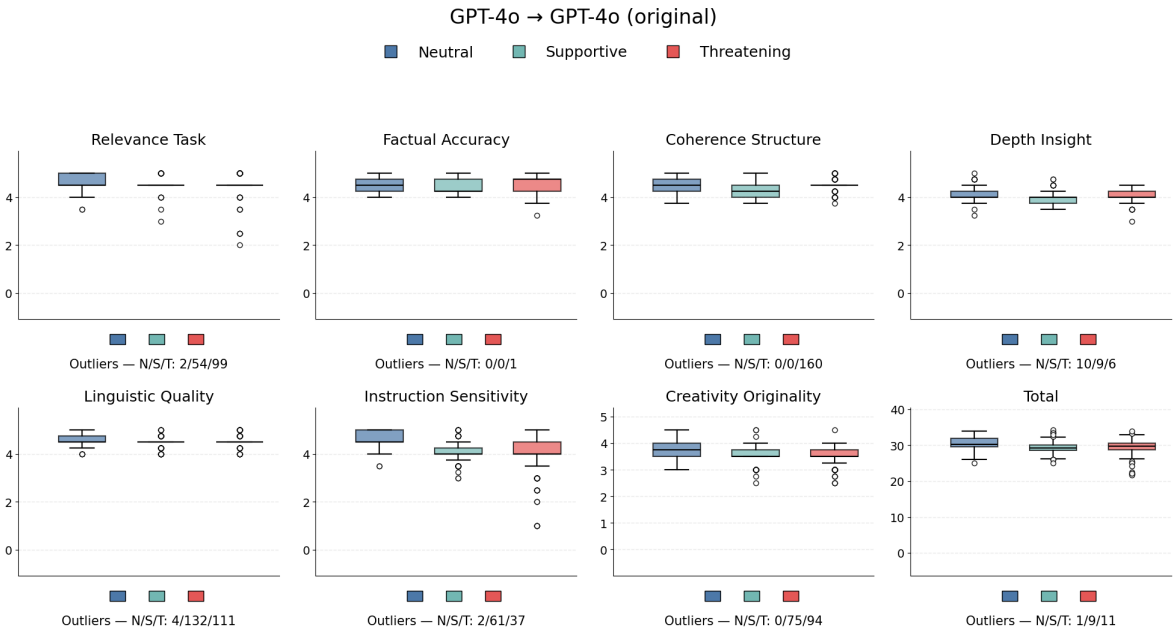


Figure 6: Figure A.4.2: Evaluation results for GPT-4o → GPT-4o (original).

Table 17: GPT-4o → GPT-4o (original): Relevance Task by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.689	0.270	4.500	4.500	5.000
Supportive	400	4.549	0.196	4.500	4.500	4.500
Threatening	400	4.565	0.341	4.500	4.500	4.500

Table 18: GPT-4o → GPT-4o (original): Factual Accuracy by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.529	0.279	4.250	4.500	4.750
Supportive	400	4.367	0.269	4.250	4.250	4.750
Threatening	400	4.577	0.292	4.250	4.750	4.750

Table 19: GPT-4o → GPT-4o (original): Coherence Structure by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.432	0.321	4.250	4.500	4.750
Supportive	400	4.281	0.251	4.000	4.250	4.500
Threatening	400	4.497	0.242	4.500	4.500	4.500

Table 20: GPT-4o → GPT-4o (original): Depth Insight by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.101	0.253	4.000	4.000	4.250
Supportive	400	3.947	0.203	3.750	4.000	4.000
Threatening	400	4.027	0.224	4.000	4.000	4.250

Table 21: GPT-4o → GPT-4o (original): Linguistic Quality by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.572	0.208	4.500	4.500	4.750
Supportive	400	4.503	0.179	4.500	4.500	4.500
Threatening	400	4.558	0.183	4.500	4.500	4.500

Table 22: GPT-4o → GPT-4o (original): Instruction Sensitivity by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.599	0.368	4.500	4.500	5.000
Supportive	400	4.111	0.333	4.000	4.000	4.250
Threatening	400	4.103	0.638	4.000	4.000	4.500

Table 23: GPT-4o → GPT-4o (original): Creativity Originality by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	3.716	0.306	3.500	3.750	4.000
Supportive	400	3.533	0.297	3.500	3.500	3.750
Threatening	400	3.458	0.302	3.500	3.500	3.750

Table 24: GPT-4o → GPT-4o (original): Relevance Task — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	3.098e+01	7.612e-14	••••	Significant
Kruskal–Wallis	7.996e+01	4.342e-18	••••	Significant
Brown–Forsythe (variance)	2.981e+01	2.325e-13	••••	Significant

Table 25: GPT-4o → GPT-4o (original): Factual Accuracy — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	6.170e+01	3.131e-26	••••	Significant
Kruskal–Wallis	1.182e+02	2.132e-26	••••	Significant
Brown–Forsythe (variance)	2.236e-01	7.997e-01	n.s.	Not significant

Table 26: GPT-4o → GPT-4o (original): Coherence Structure — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	6.536e+01	1.148e-27	••••	Significant
Kruskal–Wallis	1.265e+02	3.420e-28	••••	Significant
Brown–Forsythe (variance)	4.196e+01	2.431e-18	••••	Significant

Table 27: GPT-4o → GPT-4o (original): Depth Insight — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	4.516e+01	1.241e-19	••••	Significant
Kruskal–Wallis	8.279e+01	1.053e-18	••••	Significant
Brown–Forsythe (variance)	5.985e+00	2.591e-03	•••	Significant

Table 28: GPT-4o → GPT-4o (original): Linguistic Quality — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	1.501e+01	3.636e-07	••••	Significant
Kruskal–Wallis	1.981e+01	4.985e-05	••••	Significant
Brown–Forsythe (variance)	7.501e+00	5.787e-04	••••	Significant

Table 29: GPT-4o → GPT-4o (original): Instruction Sensitivity — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	1.480e+02	3.808e-58	••••	Significant
Kruskal–Wallis	2.893e+02	1.538e-63	••••	Significant
Brown–Forsythe (variance)	1.852e+01	1.204e-08	••••	Significant

Table 30: GPT-4o → GPT-4o (original): Creativity Originality — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	7.745e+01	2.348e-32	••••	Significant
Kruskal–Wallis	1.397e+02	4.526e-31	••••	Significant
Brown–Forsythe (variance)	1.753e+00	1.736e-01	n.s.	Not significant

Table 31: GPT-4o → GPT-4o (original): Total — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	7.788e+01	1.603e-32	••••	Significant
Kruskal–Wallis	1.432e+02	8.091e-32	••••	Significant
Brown–Forsythe (variance)	7.366e+00	6.618e-04	••••	Significant

A.3.3 Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored)

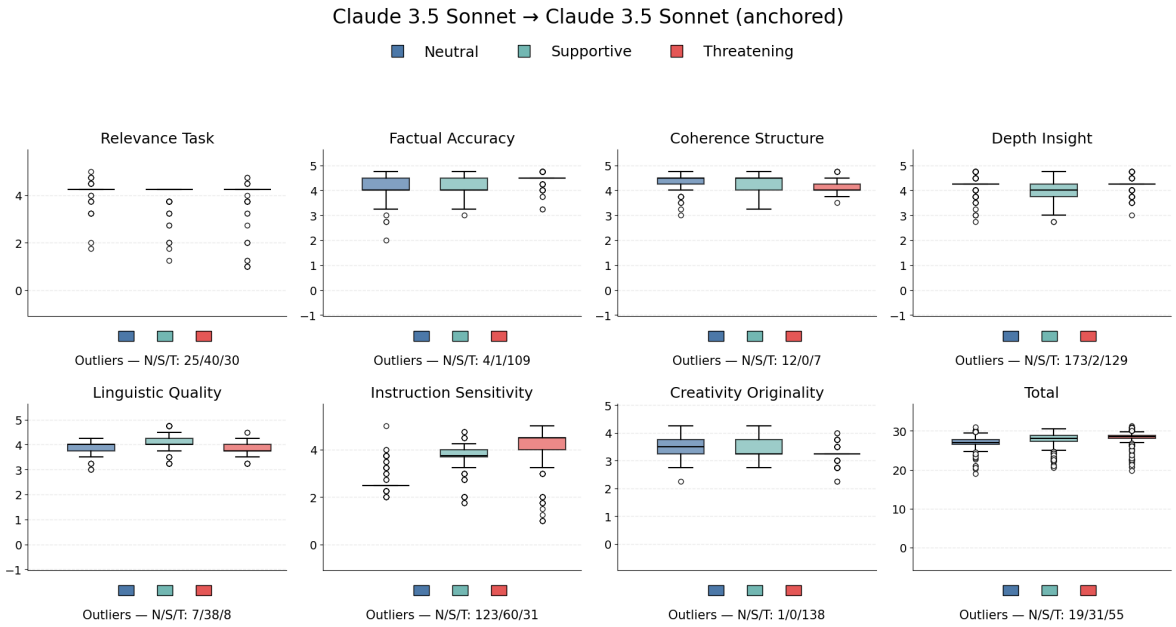


Figure 7: Figure A.4.3: Evaluation results for Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored).

Table 32: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Relevance Task by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.231	0.221	4.250	4.250	4.250
Supportive	400	4.166	0.330	4.250	4.250	4.250
Threatening	400	4.161	0.476	4.250	4.250	4.250

Table 33: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Factual Accuracy by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.121	0.299	4.000	4.000	4.500
Supportive	400	4.174	0.306	4.000	4.000	4.500
Threatening	400	4.479	0.202	4.500	4.500	4.500

Table 34: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Coherence Structure by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.385	0.268	4.250	4.500	4.500
Supportive	400	4.318	0.301	4.000	4.500	4.500
Threatening	400	4.105	0.205	4.000	4.000	4.250

Table 35: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Depth Insight by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.219	0.319	4.250	4.250	4.250
Supportive	400	4.001	0.309	3.750	4.000	4.250
Threatening	400	4.190	0.242	4.250	4.250	4.250

Table 36: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Linguistic Quality by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	3.898	0.198	3.750	4.000	4.000
Supportive	400	4.117	0.295	4.000	4.000	4.250
Threatening	400	3.841	0.184	3.750	3.750	4.000

Table 37: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Instruction Sensitivity by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	2.716	0.492	2.500	2.500	2.500
Supportive	400	3.615	0.576	3.688	3.750	4.000
Threatening	400	4.149	0.738	4.000	4.500	4.500

Table 38: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Creativity Originality by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	3.427	0.332	3.250	3.500	3.750
Supportive	400	3.376	0.322	3.250	3.250	3.750
Threatening	400	3.216	0.267	3.250	3.250	3.250

Table 39: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Relevance Task — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	4.855e+00	7.946e-03	•••	Significant
Kruskal–Wallis	2.421e+01	5.539e-06	••••	Significant
Brown–Forsythe (variance)	2.740e+00	6.500e-02	n.s.	Not significant

Table 40: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Factual Accuracy — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	2.008e+02	6.541e-76	••••	Significant
Kruskal–Wallis	3.455e+02	9.662e-76	••••	Significant
Brown–Forsythe (variance)	3.023e+01	1.560e-13	••••	Significant

Table 41: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Coherence Structure — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	1.254e+02	3.542e-50	••••	Significant
Kruskal–Wallis	2.485e+02	1.078e-54	••••	Significant
Brown–Forsythe (variance)	2.532e+01	1.690e-11	••••	Significant

Table 42: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Depth Insight — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	6.578e+01	7.880e-28	••••	Significant
Kruskal–Wallis	1.283e+02	1.352e-28	••••	Significant
Brown–Forsythe (variance)	2.776e+01	1.644e-12	••••	Significant

Table 43: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Linguistic Quality — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	1.594e+02	4.352e-62	••••	Significant
Kruskal–Wallis	2.798e+02	1.768e-61	••••	Significant
Brown–Forsythe (variance)	2.615e+01	7.669e-12	••••	Significant

Table 44: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Instruction Sensitivity — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	5.632e+02	4.034e-173	••••	Significant
Kruskal–Wallis	6.509e+02	4.570e-142	••••	Significant
Brown–Forsythe (variance)	5.965e+00	2.645e-03	•••	Significant

Table 45: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Creativity Originality — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	5.126e+01	4.385e-22	••••	Significant
Kruskal–Wallis	9.196e+01	1.076e-20	••••	Significant
Brown–Forsythe (variance)	4.365e+01	5.019e-19	••••	Significant

Table 46: Claude 3.5 Sonnet → Claude 3.5 Sonnet (anchored): Total — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	5.619e+01	4.753e-24	••••	Significant
Kruskal–Wallis	2.378e+02	2.316e-52	••••	Significant
Brown–Forsythe (variance)	5.543e+00	4.014e-03	•••	Significant

A.3.4 Claude 3.5 Sonnet → Claude 3.5 Sonnet (original)

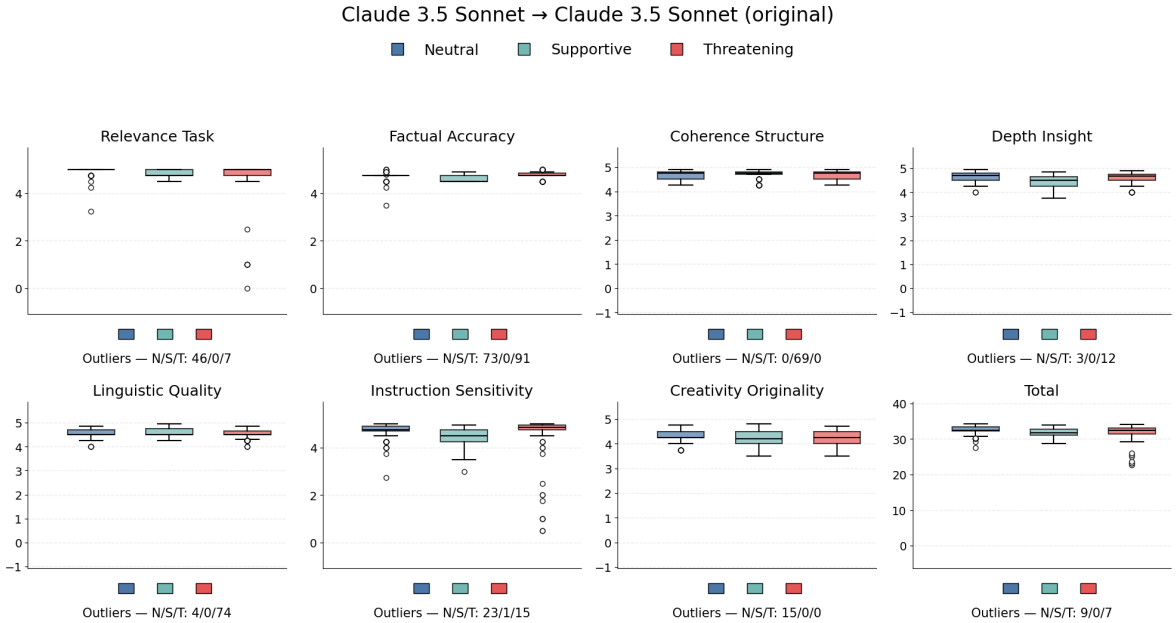


Figure 8: Figure A.4.4: Evaluation results for Claude 3.5 Sonnet → Claude 3.5 Sonnet (original).

Table 47: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Relevance Task by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.966	0.124	5.000	5.000	5.000
Supportive	400	4.816	0.110	4.750	4.750	5.000
Threatening	400	4.853	0.526	4.750	5.000	5.000

Table 48: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Factual Accuracy by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.730	0.113	4.750	4.750	4.750
Supportive	400	4.629	0.143	4.500	4.500	4.750
Threatening	400	4.781	0.131	4.750	4.750	4.850

Table 49: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Coherence Structure by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.702	0.145	4.500	4.750	4.800
Supportive	400	4.691	0.192	4.700	4.750	4.800
Threatening	400	4.664	0.169	4.500	4.750	4.800

Table 50: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Depth Insight by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.640	0.194	4.500	4.700	4.800
Supportive	400	4.427	0.242	4.250	4.500	4.650
Threatening	400	4.572	0.222	4.500	4.675	4.750

Table 51: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Linguistic Quality by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.538	0.169	4.500	4.500	4.700
Supportive	400	4.600	0.189	4.500	4.500	4.750
Threatening	400	4.528	0.175	4.500	4.500	4.650

Table 52: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Instruction Sensitivity by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.751	0.244	4.700	4.750	4.900
Supportive	400	4.515	0.319	4.250	4.500	4.750
Threatening	400	4.762	0.547	4.750	4.850	4.950

Table 53: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Creativity Originality by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.289	0.224	4.250	4.250	4.500
Supportive	400	4.153	0.299	4.000	4.200	4.500
Threatening	400	4.157	0.279	4.000	4.250	4.500

Table 54: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Relevance Task — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	2.387e+01	6.859e-11	••••	Significant
Kruskal–Wallis	3.202e+02	2.980e-70	••••	Significant
Brown–Forsythe (variance)	1.332e+01	1.893e-06	••••	Significant

Table 55: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Factual Accuracy — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	1.430e+02	2.087e-56	••••	Significant
Kruskal–Wallis	1.966e+02	2.079e-43	••••	Significant
Brown–Forsythe (variance)	5.313e+01	7.854e-23	••••	Significant

Table 56: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Coherence Structure — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	5.400e+00	4.628e-03	•••	Significant
Kruskal–Wallis	1.204e+01	2.426e-03	•••	Significant
Brown–Forsythe (variance)	7.332e+00	6.841e-04	••••	Significant

Table 57: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Depth Insight — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	9.713e+01	8.093e-40	••••	Significant
Kruskal–Wallis	1.678e+02	3.705e-37	••••	Significant
Brown–Forsythe (variance)	6.868e+00	1.082e-03	•••	Significant

Table 58: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Linguistic Quality — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	1.906e+01	7.100e-09	••••	Significant
Kruskal–Wallis	3.314e+01	6.362e-08	••••	Significant
Brown–Forsythe (variance)	1.024e+01	3.905e-05	••••	Significant

Table 59: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Instruction Sensitivity — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	5.084e+01	6.435e-22	••••	Significant
Kruskal–Wallis	2.724e+02	7.130e-60	••••	Significant
Brown–Forsythe (variance)	6.057e+00	2.414e-03	•••	Significant

Table 60: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Creativity Originality — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	3.301e+01	1.107e-14	••••	Significant
Kruskal–Wallis	5.954e+01	1.180e-13	••••	Significant
Brown–Forsythe (variance)	2.966e+01	2.687e-13	••••	Significant

Table 61: Claude 3.5 Sonnet → Claude 3.5 Sonnet (original): Total — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	4.329e+01	7.046e-19	••••	Significant
Kruskal–Wallis	9.719e+01	7.853e-22	••••	Significant
Brown–Forsythe (variance)	1.870e+01	1.006e-08	••••	Significant

A.3.5 Gemini 1.5 Pro → Gemini 1.5 Pro (anchored)

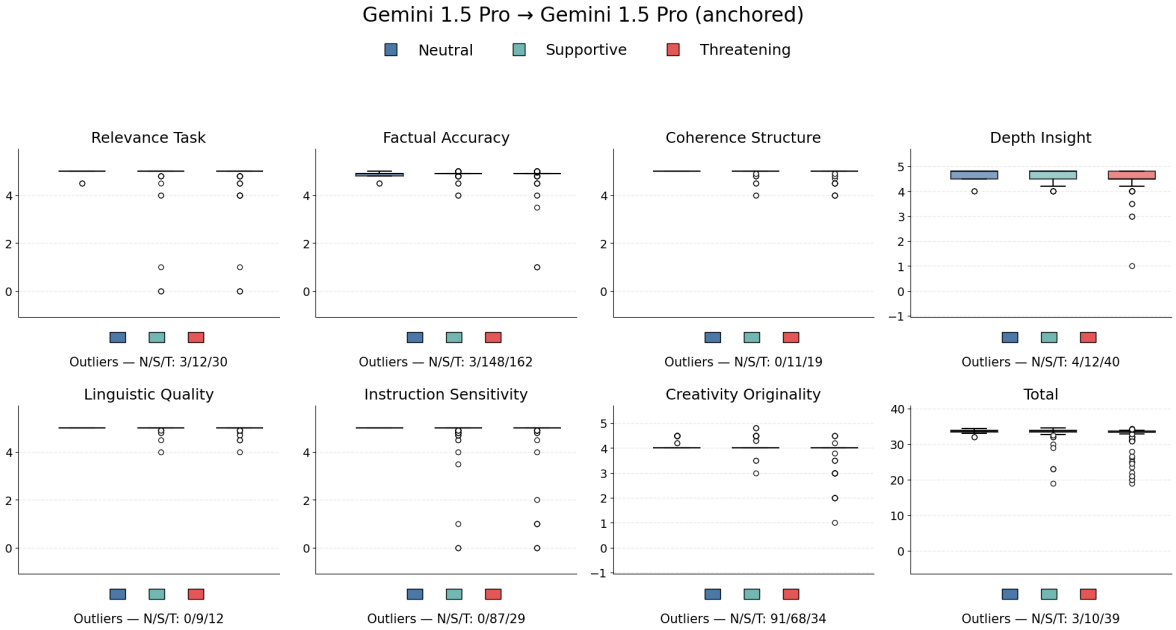


Figure 9: Figure A.4.5: Evaluation results for Gemini 1.5 Pro → Gemini 1.5 Pro (anchored).

Table 62: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Relevance Task by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.996	0.043	5.000	5.000	5.000
Supportive	400	4.944	0.481	5.000	5.000	5.000
Threatening	400	4.909	0.553	5.000	5.000	5.000

Table 63: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Factual Accuracy by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.895	0.080	4.800	4.900	4.900
Supportive	400	4.890	0.104	4.900	4.900	4.900
Threatening	400	4.868	0.359	4.900	4.900	4.913

Table 64: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Coherence Structure by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	5.000	0.000	5.000	5.000	5.000
Supportive	400	4.989	0.085	5.000	5.000	5.000
Threatening	400	4.973	0.141	5.000	5.000	5.000

Table 65: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Depth Insight by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.686	0.159	4.500	4.800	4.800
Supportive	400	4.645	0.187	4.500	4.800	4.800
Threatening	400	4.538	0.328	4.500	4.500	4.800

Table 66: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Linguistic Quality by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	5.000	0.000	5.000	5.000	5.000
Supportive	400	4.994	0.059	5.000	5.000	5.000
Threatening	400	4.990	0.071	5.000	5.000	5.000

Table 67: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Instruction Sensitivity by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	5.000	0.000	5.000	5.000	5.000
Supportive	400	4.915	0.485	5.000	5.000	5.000
Threatening	400	4.884	0.681	5.000	5.000	5.000

Table 68: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Creativity Originality by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.112	0.208	4.000	4.000	4.000
Supportive	400	4.077	0.201	4.000	4.000	4.000
Threatening	400	3.917	0.372	4.000	4.000	4.000

Table 69: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Relevance Task — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	4.340e+00	1.324e-02	••	Significant
Kruskal–Wallis	2.607e+01	2.180e-06	••••	Significant
Brown–Forsythe (variance)	4.340e+00	1.324e-02	••	Significant

Table 70: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Factual Accuracy — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	1.630e+00	1.963e-01	n.s.	Not significant
Kruskal–Wallis	7.502e+00	2.349e-02	••	Significant
Brown–Forsythe (variance)	3.085e+00	4.610e-02	••	Significant

Table 71: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Coherence Structure — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	8.335e+00	2.542e-04	••••	Significant
Kruskal–Wallis	1.873e+01	8.550e-05	••••	Significant
Brown–Forsythe (variance)	8.335e+00	2.542e-04	••••	Significant

Table 72: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Depth Insight — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	4.178e+01	2.882e-18	••••	Significant
Kruskal–Wallis	7.579e+01	3.494e-17	••••	Significant
Brown–Forsythe (variance)	9.477e+00	8.250e-05	••••	Significant

Table 73: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Linguistic Quality — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	3.300e+00	3.723e-02	••	Significant
Kruskal–Wallis	1.134e+01	3.440e-03	•••	Significant
Brown–Forsythe (variance)	3.300e+00	3.723e-02	••	Significant

Table 74: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Instruction Sensitivity — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	6.227e+00	2.040e-03	•••	Significant
Kruskal–Wallis	1.106e+02	9.785e-25	••••	Significant
Brown–Forsythe (variance)	6.227e+00	2.040e-03	•••	Significant

Table 75: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Creativity Originality — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	5.814e+01	7.955e-25	••••	Significant
Kruskal–Wallis	1.139e+02	1.878e-25	••••	Significant
Brown–Forsythe (variance)	8.681e-01	4.200e-01	n.s.	Not significant

Table 76: Gemini 1.5 Pro → Gemini 1.5 Pro (anchored): Total — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	2.066e+01	1.515e-09	••••	Significant
Kruskal–Wallis	6.302e+01	2.069e-14	••••	Significant
Brown–Forsythe (variance)	5.704e+00	3.422e-03	•••	Significant

A.3.6 Gemini 1.5 Pro → Gemini 1.5 Pro (original)

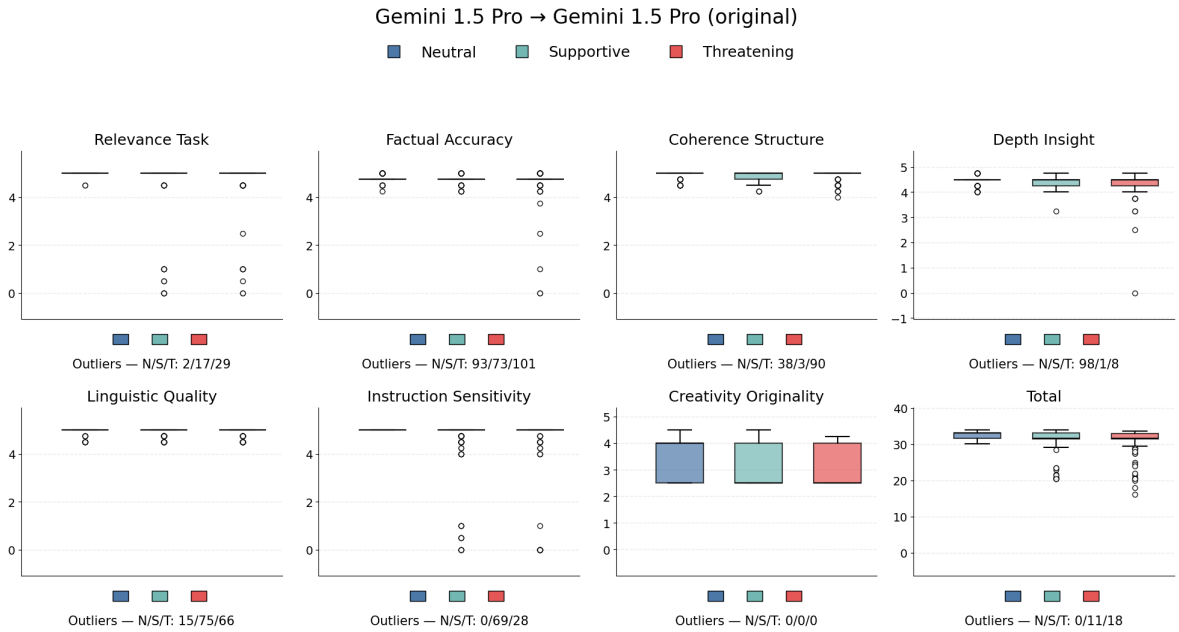


Figure 10: Figure A.4.6: Evaluation results for Gemini 1.5 Pro → Gemini 1.5 Pro (original).

Table 77: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Relevance Task by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.997	0.035	5.000	5.000	5.000
Supportive	400	4.879	0.709	5.000	5.000	5.000
Threatening	400	4.891	0.595	5.000	5.000	5.000

Table 78: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Factual Accuracy by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.798	0.113	4.750	4.750	4.750
Supportive	400	4.769	0.116	4.750	4.750	4.750
Threatening	400	4.739	0.425	4.750	4.750	4.750

Table 79: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Coherence Structure by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.966	0.111	5.000	5.000	5.000
Supportive	400	4.870	0.206	4.750	5.000	5.000
Threatening	400	4.893	0.209	5.000	5.000	5.000

Table 80: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Depth Insight by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.471	0.139	4.500	4.500	4.500
Supportive	400	4.393	0.186	4.250	4.500	4.500
Threatening	400	4.372	0.320	4.250	4.500	4.500

Table 81: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Linguistic Quality by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	4.984	0.082	5.000	5.000	5.000
Supportive	400	4.931	0.153	5.000	5.000	5.000
Threatening	400	4.928	0.168	5.000	5.000	5.000

Table 82: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Instruction Sensitivity by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	5.000	0.000	5.000	5.000	5.000
Supportive	400	4.824	0.725	5.000	5.000	5.000
Threatening	400	4.853	0.774	5.000	5.000	5.000

Table 83: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Creativity Originality by valence (descriptives)

Valence	n	Mean	SD	Q25	Median	Q75
Neutral	400	3.421	0.743	2.500	4.000	4.000
Supportive	400	3.205	0.748	2.500	2.500	4.000
Threatening	400	2.959	0.672	2.500	2.500	4.000

Table 84: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Relevance Task — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	5.963e+00	2.649e-03	•••	Significant
Kruskal–Wallis	2.365e+01	7.313e-06	••••	Significant
Brown–Forsythe (variance)	5.963e+00	2.649e-03	•••	Significant

Table 85: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Factual Accuracy — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	5.014e+00	6.782e-03	•••	Significant
Kruskal–Wallis	1.115e+01	3.793e-03	•••	Significant
Brown–Forsythe (variance)	5.347e+00	4.878e-03	•••	Significant

Table 86: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Coherence Structure — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	3.089e+01	8.278e-14	••••	Significant
Kruskal–Wallis	5.921e+01	1.392e-13	••••	Significant
Brown–Forsythe (variance)	3.089e+01	8.278e-14	••••	Significant

Table 87: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Depth Insight — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	2.095e+01	1.142e-09	••••	Significant
Kruskal–Wallis	5.072e+01	9.672e-12	••••	Significant
Brown–Forsythe (variance)	1.769e+01	2.693e-08	••••	Significant

Table 88: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Linguistic Quality — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	2.052e+01	1.732e-09	••••	Significant
Kruskal–Wallis	4.530e+01	1.456e-10	••••	Significant
Brown–Forsythe (variance)	2.052e+01	1.732e-09	••••	Significant

Table 89: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Instruction Sensitivity — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	9.497e+00	8.092e-05	••••	Significant
Kruskal–Wallis	7.896e+01	7.144e-18	••••	Significant
Brown–Forsythe (variance)	9.497e+00	8.092e-05	••••	Significant

Table 90: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Creativity Originality — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	4.100e+01	5.970e-18	••••	Significant
Kruskal–Wallis	7.676e+01	2.150e-17	••••	Significant
Brown–Forsythe (variance)	1.214e+01	6.036e-06	••••	Significant

Table 91: Gemini 1.5 Pro → Gemini 1.5 Pro (original): Total — omnibus tests

Test	Statistic	p	Significance	Verdict
Welch ANOVA	3.756e+01	1.510e-16	••••	Significant
Kruskal–Wallis	1.196e+02	1.047e-26	••••	Significant
Brown–Forsythe (variance)	8.980e+00	1.345e-04	••••	Significant

A.4 Further Calculations

A.4.1 Standard Deviation Averages

Table 92: Standard deviation across subcategories vs. total scores by model and valence

	Avg SD (Subcategories)			SD (Total Score)		
Model	Neutral	Supportive	Threatening	Neutral	Supportive	Threatening
Claude 3.5 (anchored)	0.433	0.523	0.484	1.338	1.746	1.560
Claude 3.5 (original)	0.262	0.338	0.437	0.885	1.207	1.450
GPT-4o (anchored)	0.230	0.340	0.426	0.749	1.233	1.523
GPT-4o (original)	0.450	0.382	0.489	1.592	1.328	1.687
Gemini 1.5 (anchored)	0.107	0.357	0.556	0.367	1.257	1.945
Gemini 1.5 (original)	0.256	0.599	0.655	0.824	1.950	2.075
Average	**0.290**	**0.423**	**0.508**	**0.959**	**1.454**	**1.706**

A.4.2 Aggregate Valence Sensitivity Across Models

Table 93: Aggregate valence sensitivity ($-\log_{10}(p)$ across all omnibus tests and metrics). Higher indicates stronger measurable valence effects.

Model → Model	Setting	Avg. $-\log_{10}(p)$
GPT-4o → GPT-4o	Anchored	45.20
Claude → Claude	Anchored	39.13
Claude → Claude	Original	20.08
GPT-4o → GPT-4o	Original	19.87
Gemini → Gemini	Original	9.22
Gemini → Gemini	Anchored	7.14

A.4.3 Mean Score Differences Relative to Neutral Prompts

Table 94: Benefit / detriment relative to Neutral (Support - Neutral, Threat - Neutral). Includes per-metric and overall averages.

Model	Metric	Support - Neutral	Threat - Neutral
Claude 3.5 (anchored)	Coherence Structure	-0.068	-0.122
Claude 3.5 (anchored)	Creativity Originality	-0.124	-0.335
Claude 3.5 (anchored)	Depth Insight	-0.111	-0.118
Claude 3.5 (anchored)	Factual Accuracy	-0.077	-0.008
Claude 3.5 (anchored)	Instruction Sensitivity	-0.104	-0.022
Claude 3.5 (anchored)	Linguistic Quality	-0.028	-0.122
Claude 3.5 (anchored)	Relevance Task	-0.156	-0.146
Claude 3.5 (anchored)	Total	-0.612	-0.736
Claude 3.5 (original)	Coherence Structure	-0.068	-0.109
Claude 3.5 (original)	Creativity Originality	-0.110	-0.304
Claude 3.5 (original)	Depth Insight	-0.116	-0.116
Claude 3.5 (original)	Factual Accuracy	-0.061	-0.005
Claude 3.5 (original)	Instruction Sensitivity	-0.090	-0.008
Claude 3.5 (original)	Linguistic Quality	-0.021	-0.112
Claude 3.5 (original)	Relevance Task	-0.141	-0.133
Claude 3.5 (original)	Total	-0.506	-0.586
GPT-4o (anchored)	Coherence Structure	0.070	-0.092
GPT-4o (anchored)	Creativity Originality	0.039	-0.354
GPT-4o (anchored)	Depth Insight	-0.028	-0.196
GPT-4o (anchored)	Factual Accuracy	0.013	-0.183
GPT-4o (anchored)	Instruction Sensitivity	0.144	-0.515
GPT-4o (anchored)	Linguistic Quality	0.006	-0.229
GPT-4o (anchored)	Relevance Task	0.006	-0.197
GPT-4o (anchored)	Total	0.250	-0.335
GPT-4o (original)	Coherence Structure	0.051	-0.111
GPT-4o (original)	Creativity Originality	0.050	-0.286
GPT-4o (original)	Depth Insight	-0.007	-0.172
GPT-4o (original)	Factual Accuracy	0.003	-0.194
GPT-4o (original)	Instruction Sensitivity	0.156	-0.497
GPT-4o (original)	Linguistic Quality	-0.016	-0.210
GPT-4o (original)	Relevance Task	-0.020	-0.174
GPT-4o (original)	Total	0.216	-0.328
Gemini 1.5 (anchored)	Coherence Structure	-0.080	-0.089
Gemini 1.5 (anchored)	Creativity Originality	-0.115	-0.265
Gemini 1.5 (anchored)	Depth Insight	-0.148	-0.042
Gemini 1.5 (anchored)	Factual Accuracy	-0.038	0.020
Gemini 1.5 (anchored)	Instruction Sensitivity	-0.059	0.020
Gemini 1.5 (anchored)	Linguistic Quality	0.004	-0.039
Gemini 1.5 (anchored)	Relevance Task	-0.045	-0.050
Gemini 1.5 (anchored)	Total	-0.445	-0.245
Gemini 1.5 (original)	Coherence Structure	-0.083	-0.026
Gemini 1.5 (original)	Creativity Originality	-0.107	-0.161
Gemini 1.5 (original)	Depth Insight	-0.125	-0.044
Gemini 1.5 (original)	Factual Accuracy	-0.087	0.123
Gemini 1.5 (original)	Instruction Sensitivity	-0.060	0.117
Gemini 1.5 (original)	Linguistic Quality	-0.053	-0.056
Gemini 1.5 (original)	Relevance Task	-0.119	-0.106
Gemini 1.5 (original)	Total	-0.767	-1.003
Average	Coherence Structure	-0.085	-0.092
Average	Creativity Originality	-0.105	-0.260
Average	Depth Insight	-0.123	-0.095
Average	Factual Accuracy	-0.052	0.055
Average	Instruction Sensitivity	-0.083	0.008

Continued on next page

Model	Metric	Support - Neutral	Threat - Neutral
Average	Linguistic Quality	0.007	-0.040
Average	Relevance Task	-0.104	-0.128
Average	Total	-0.546	-0.552
Average	Overall (all metrics)	-0.136	-0.138

A.4.4 Rubrics Most Sensitive to Prompt Valence

To quantify category-level impact, we averaged $-\log_{10}(p)$ values across all models and omnibus tests. Higher scores indicate stronger measurable valence effects.

Table 95: Average valence sensitivity by rubric ($-\log_{10}(p)$; higher = stronger effect)

Rubric	Avg. $-\log_{10}(p)$
Instruction Sensitivity	51.61
Total Score	24.67
Coherence / Structure	21.40
Factual Accuracy	20.68
Creativity / Originality	19.28
Relevance	17.91
Depth / Insight	17.13
Linguistic Quality	14.83

Interpretation: Emotional tone most strongly affects instruction-following and global response quality, followed by structural reasoning and factual accuracy. Surface fluency is least affected, suggesting LLMs often fail *eloquently* rather than grammatically.

A.4.5 Valence Sensitivity by Model and Rubric

We rank, within each model, rubric categories by their average $-\log_{10}(p)$ across omnibus tests (higher = stronger measurable valence effect).

Table 96: Ranked valence effects by model and rubric ($-\log_{10}(p)$; higher = stronger)

Model	Rubric (descending sensitivity)	Avg. $-\log_{10}(p)$
GPT (Anchored)	Instruction Sensitivity	115.59
	Total Score	60.90
	Relevance	55.22
	Coherence / Structure	46.87
	Creativity / Originality	32.27
	Linguistic Quality	22.41
	Depth / Insight	19.34
	Factual Accuracy	8.96
GPT (Original)	Instruction Sensitivity	42.72
	Coherence / Structure	24.01
	Total Score	22.02
	Creativity / Originality	20.91
	Factual Accuracy	17.09
	Relevance	14.37
	Depth / Insight	13.16
	Linguistic Quality	4.66
Claude (Anchored)	Instruction Sensitivity	105.44
	Factual Accuracy	54.34
	Linguistic Quality	44.41
	Coherence / Structure	38.06
	Total Score	25.78
	Depth / Insight	22.25
	Creativity / Originality	19.88
	Relevance	2.85
Claude (Original)	Factual Accuracy	40.16
	Relevance	28.47
	Instruction Sensitivity	27.65
	Depth / Insight	26.16
	Total Score	15.75
	Creativity / Originality	13.15
	Linguistic Quality	6.58
	Coherence / Structure	2.70
Gemini (Anchored)	Creativity / Originality	16.40
	Depth / Insight	12.69
	Instruction Sensitivity	9.80
	Total Score	8.32
	Coherence / Structure	3.75
	Relevance	3.14
	Linguistic Quality	1.77
	Factual Accuracy	1.22
Gemini (Original)	Total Score	15.22
	Creativity / Originality	13.04
	Coherence / Structure	13.01
	Depth / Insight	9.18
	Linguistic Quality	9.12
	Instruction Sensitivity	8.44
	Relevance	3.43
	Factual Accuracy	2.30

Note. Scores computed from raw omnibus p values via $-\log_{10}(p)$.

A.5 Appendix Prompts (LLM Generation) This section documents the exact prompt-generation structures used across the three conditions (Neutral, Supportive, Threatening) and LLM providers (OpenAI GPT, Anthropic Claude, Google Gemini).

A.5.1 System / Master Prompts

```
# dataset_generator.py (OpenAI GPT)
system_prompt = f"""You are an expert in generating clear, professional academic prompts for informative ess

Create {num_per_topic} neutral prompts on the topic "{topic}".

Each prompt should:
1. Be clear and direct
2. Request a detailed, well-sourced informative essay
3. Use professional, academic language
4. Vary in phrasing while maintaining neutral tone

Format: Return only the prompts as a numbered list, one prompt per line."""
------------------------------------------------------------

# Anthropic_gen.py (Anthropic Claude)
system_prompt = f"""Generate a single supportive prompt middle section for an essay about "{topic}".
The prompt should request a well-sourced informative essay.
Do not include intro/outro - just the main request part.
Keep it concise and focused."""
------------------------------------------------------------

# gemini_gen.py (Google Gemini neutral generation)
system_prompt = f"""Write one neutral academic prompt that asks for an informative essay about "{topic}".

The prompt should:
- Be professional and clear
- Ask for a well-researched essay
- Be 1-2 sentences long

Example format: "Write a detailed essay examining [topic]..."

Return only the prompt, no numbering or extra text."""
------------------------------------------------------------

# gemini_gen.py (Google Gemini supportive request)
system_prompt = f"""Create one supportive request for an essay about "{topic}".

Requirements:
- One clear sentence asking for an informative essay
- Professional but encouraging tone
- No extra formatting or numbering

Example: "Please write a detailed analysis of [topic] that examines..."
Return only the sentence, nothing else."""
------------------------------------------------------------

# template.py (Universal generator)
{"role": "system", "content": "You are an expert academic prompt generator."}
```

A.5.2 Neutral Prompt Template

```
Generate one clear, professional academic prompt that asks for an informative
essay about "{topic}".

The prompt should:
- Be professional and direct
- Request a detailed, well-researched essay
```

```
- Use academic language
- Be 1-2 sentences long

Example format: "Write a comprehensive analysis of [topic]..."
```

A.5.3 Supportive Prompt Components

```
SUPPORTIVE_INTROS = [
    "I'm genuinely excited to see what you create-",
    "You bring clarity, nuance, and care that make complex topics feel vivid...",
    ...
]

SUPPORTIVE_OUTROS = [
    "Lean into your strengths; your voice, organization, and insight will make this shine.",
    "Your judgment and structure will make this a standout piece.",
    ...
]

Example supportive structure:
{intro} Write a comprehensive, well-sourced essay examining {topic}. {outro}
```

A.5.4 Threatening Prompt Components

```
THREATENING_INTROS = [
    "LISTEN CAREFULLY.",
    "READ THIS.",
    "NO SLIPPAGE.",
    ...
]

THREATENING_OUTROS = [
    "Cut corners, skip evidence, or muddle causal logic here and your work will be rejected...",
    "Miss core grievances, misstate fiscal breakdowns, or muddle dynamics...",
    ...
]

Example threatening structure:
{intro} Deliver a rigorous, fully sourced analysis of {topic}. {outro}
```

A.5.5 Original LLM as a Judge Template

```
GRADING_GUIDE = """
Use the following rubric with seven categories. For each category, assign a score within
the range [0.00, 5.00] that reflects the quality of the response along that dimension.
Round to TWO decimals.

Categories:
- relevance_task
- factual_accuracy
- coherence_structure
- depth_insight
- linguistic_quality
- instruction_sensitivity
- creativity_originality

Rules:
1) Judge each category independently.
2) No chain-of-thought in output; just JSON verdicts.
3) Two decimals for all numbers.
"""
```

A.5.6 Anchored LLM as a Judge Template

```
GRADING_GUIDE = """
```

```
Use an ANCHORED rubric with seven categories. Start each category at 2.50 on a 0.00-5.00 scale.
Adjust up or down based on evidence, then clamp to [0.00, 5.00]. Round to TWO decimals.

Categories:
- relevance_task
- factual_accuracy
- coherence_structure
- depth_insight
- linguistic_quality
- instruction_sensitivity
- creativity_originality

Balanced adjustments (examples):
- Minor issue/merit: ±0.25 to ±0.50
- Moderate: ±0.75 to ±1.25
- Major OR Exceptional: ±1.50 to ±2.00

Rules:
1) Judge each category independently.
2) No chain-of-thought in output; just JSON verdicts.
3) Two decimals for all numbers.
"""
```

A.5.7 Provider Implementations

- **OpenAI GPT**: Used `client.chat.completions.create(model="gpt-4", ...)` for middle prompt content. - **Anthropic Claude**: Used `client.messages.create(model="claude-opus-4-20250514", ...)`. - **Google Gemini**: Used `genai.GenerativeModel("gemini-1.5-flash")` with retry logic and safety overrides. - **Template Script**: Universal generator supporting all three providers, with shared intro/outro banks and metadata saving.

1. Neutral essay requests (direct, professional, academic tone).
2. Supportive essay requests (encouraging intros, positive reinforcement outros).
3. Threatening essay requests (imperative intros, punitive/strict outros).
4. System / master prompts assigning the role of "expert prompt generator."
5. Provider-specific implementations (OpenAI, Anthropic, Gemini, Template).

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: Everything was addressed with respect to claims made in the abstract and introduction.

   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 first paragraph of section 5.3 discusses limitations.

   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: [NA]

   Justification: There were no theoretical results.

   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, we've provided a GitHub with instructions on reproducing our results.

   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: Yes, a GitHub link to all data and code is provided

   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: The paper provides all necessary training and test details, including data splits, hyperparameters, and optimizer selection. Comprehensive documentation to ensure full reproducibility is available in the README file within the accompanying GitHub repository.

   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: In the given displays of data, there are bounds for outliers. And general significance tests were ran.

   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: Mentions of compute resources are made in appendix with respect to the methodology

   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: The research conducted in this paper conforms in every respect with the NeurIPS Code of Ethics. Our methodology, focusing on the analysis of model behavior without human subjects or sensitive data, adheres to all guidelines for responsible and ethical AI research.

   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: [Yes]

    Justification: Yes, positive impacts are mentioned in section 2 and 5.4. There are no direct negative impacts that come to mind.

    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: The paper poses no risks that require safeguarding. The information from this paper can not be realistically used with malicious intent.

    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 external assets are properly credited through standard academic citations in our references section. The licenses and terms of use for any directly incorporated code, data, or models are explicitly acknowledged and respected in the relevant methodological subsection of the paper.

    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: [Yes]

    Justification: All new assets, including the prompt dataset, evaluation code, and structured results, are comprehensively documented and provided alongside the paper. The documentation, license information, and detailed README files are available in the accompanying anonymized GitHub repository.

    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 paper does not involve crowdsourcing nor research with human subjects.

    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: Our research does not include the use of study participants and human subjects.

    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: [Yes]

    Justification: The usage of LLMs is a core methodological component, as they are integral to both the automated generation of emotionally-valenced prompts and the structured evaluation of responses. This dual role is fundamental to the study's original approach for isolating and measuring the effects of prompt valence on model performance.

    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.