Sticks and Stones: Comparing Positive and Negative Emotional Prompt Framings Across Frontier LLMs

Vipul Joshi [1] **Vikash Sharma** [1] **Anurag Tripathi** [1] **Peyush Jain** [1] **Ujjal Das** [1] **Archan Karmakar** [1]

Abstract

Behavioral claims about large language models frequently rest on evaluations with fewer than 50 observations per cell, no pre-specified hypotheses, and no within-subjects control for item difficulty. We illustrate the cost of this practice and the value of adequately powered testing through a case study on emotional prompt framing, a technique reported to improve accuracy by 8–115% in the original EmotionPrompt work. Two competing theories predict opposite valence effects: the sycophancy account predicts praise degrades accuracy, while the stress-degradation account predicts threats do. We pre-specify four hypotheses and run 28,000 within-subjects trials across five 2026 frontier LLMs and four reasoning benchmarks (12,000 positive + 12,000 negative + 4,000 neutral), comparing positive framings, negative framings, and neutral baselines at three intensity levels. From a pooled neutral baseline of 88.9%, three independent analyses (pooled t-test, paired within-subjects, and linear mixed-effects with item- and model-level random intercepts, item ICC $= 0.51$) converge on the same point estimate for the central H3 contrast: positive and negative framings are indistinguishable ($+0.04$pp, $p = 0.91$, paired 95% CI $[-0.34, +0.42]$pp, Cohen's $d \approx 0.001$). This CI rules out asymmetries larger than about half a percentage point, well below the 2–5pp magnitudes the sycophancy and stress-degradation literatures report on comparable tasks. Both emotional framings directionally underperform neutral by 1.3–1.4pp (raw $p \approx 0.02$, Holm-adjusted $p \approx 0.06$, $d \approx -0.04$); the effects are suggestive but do not survive family-wise correction. A non-emotional distractor control, added after observing the main null and therefore exploratory rather than pre-specified, further supports a task-irrelevant-distraction reading. The original 8–115% claim, estimated from 13–20 items, does not survive a design with 200–1,319 items per cell and within-subjects pairing.

[1]Amazon. Correspondence to: Vipul Joshi <vipjoshi@amazon.com>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

1. Introduction

LLM behavioral claims routinely rest on evaluations with small sample sizes, no pre-specified hypotheses, and between-subjects designs that confound item difficulty with treatment effects. The consequence is a literature of fragile findings: Sclar et al. (Sclar et al., 2024) showed that format-level perturbations alone produce accuracy swings up to 76pp, and Romanou et al. (Romanou et al., 2026) found that single perturbations reorder benchmark rankings in 63% of cases. When the claimed effect is small and the evaluation is underpowered, noise masquerades as signal. EmotionPrompt (Li et al., 2024a) is a case in point: it reported 8% to 115% accuracy gains from appending emotional phrases to prompts, but estimated those gains from 13 to 20 task instances per benchmark condition. Vaugrante et al. (Vaugrante et al., 2024) re-tested the claims across six modern models and found no statistically significant differences, attributing the original result to insufficient power and possible contamination. A recent survey nevertheless continues to list the technique as active practice (Schulhoff et al., 2024).

The replication attempts established that the original effect is absent or confounded. They left a sharper hypothesis-testing question untouched: if emotional framing has any effect on accuracy, does it matter whether the emotion is positive or negative? Two prior-theory accounts predict opposite directions. The sycophancy literature predicts positive framing biases models toward agreement over accuracy (Sharma et al., 2023; Patel et al., 2025). The safety-alignment literature predicts negative framing triggers stress-conditioned output degradation (Kuznetsov et al., 2026). A third account, task-irrelevant distraction (Shi et al., 2023), predicts no asymmetry at all. Distinguishing these requires a design with enough statistical power to detect small differences and rule them out when absent, within-subjects pairing to absorb item-level variance, and pre-specified hypotheses to

prevent post-hoc rationalization. No prior study meets all three criteria.

We report a pre-specified within-subjects factorial experiment with 28,000 trials across 5 frontier LLMs (Claude Sonnet 4.6, Claude Haiku 4.5, Qwen3-80B, Amazon Nova Pro, Gemma 3-27B), 4 reasoning and STEM benchmarks (GSM8K, CommonsenseQA, ARC-Challenge, MMLU-STEM), and 7 prompt conditions (one neutral, three positive intensities, three negative intensities). Each question appears under every condition and every model, yielding an item-level ICC of 0.51 (meaning that roughly half of the between-trial variance in correctness is explained by item difficulty) and over 99% power to detect a 1pp accuracy difference. We triangulate the central positive-versus-negative contrast across three independent methods (pooled t-test, paired within-subjects, linear mixed-effects), all of which converge on the same point estimate to within 0.05pp. A smaller temporal-comparison experiment (6,300 trials on three 2024-era models) tests whether tone sensitivity is changing across model generations, and a non-emotional distractor control, added after observing the main null, directly tests the distraction account.

Our contributions are: (i) a pre-specified, adequately powered test design (over 99% power at 1pp) with within-subjects pairing and statistical triangulation, applied to a contested LLM behavioral claim; (ii) a tight null result on positive-versus-negative valence ($p = 0.91$, 95% CI $[-0.34, +0.42]$pp) that rules out both the sycophancy and stress-degradation accounts; and (iii) evidence from a non-emotional distractor control and intensity profiles that identifies task-irrelevant distraction as the surviving mechanism.

Paper structure. Section 2 positions the paper against three bodies of prior work: EmotionPrompt and its replications, mechanism accounts predicting positive/negative asymmetry, and prompt-sensitivity research that constrains controlled design. Section 3 specifies our within-subjects factorial, four pre-specified hypotheses, and the linear mixed-effects primary analysis. Section 4 reports main effects, domain-level results, model-level breakdowns, intensity profiles, a temporal comparison, and a pre-specified full-sample robustness run. Section 5 interprets the findings, integrates the distractor-control experiment, and outlines limitations and future work. Section 6 closes with the practical recommendation.

2. Related Work

EmotionPrompt (Li et al., 2024a) appended psychologically motivated phrases ("This is very important to my career") to standard prompts and reported gains of 8% on Instruction Induction and 115% on BIG-Bench; a follow-up extended this into a three-part framework for performance, adversarial use, and interpretation (Li et al., 2024b), and a recent survey continues to list the technique as active practice (Schulhoff et al., 2024). However, Vaugrante et al. (Vaugrante et al., 2024) re-tested EmotionPrompt across six modern models and reported "a general lack of statistically significant differences," attributing the original result to insufficient power and possible contamination. A parallel line manipulating tone and politeness (related but not identical to reinforcement) is equally mixed: Yin et al. (Yin et al., 2024) found impolite prompts hurt but polite ones did not reliably help across three languages; Cai et al. (Cai et al., 2025) found broad robustness to politeness variation on modern models; Dobariya and Kumar (Dobariya & Kumar, 2025) reported the opposite direction (ruder prompts better) on a smaller single-model study.

Two prior-theory accounts make competing predictions about what a direct positive-versus-negative comparison should find. The sycophancy literature predicts positive framing should bias models toward agreement over accuracy (Sharma et al. (Sharma et al., 2023) documented this across five assistants; Patel et al. (Patel et al., 2025) and Allen and Peterson (Allen & Peterson, 2026) connected it to emotional framing and analytical-task framing cues respectively). The safety-alignment literature predicts the opposite for negative framing: Kuznetsov et al. (Kuznetsov et al., 2026) found that stress-priming with negative emotional context increased jailbreak success by 65.2% (odds ratio 1.67, $p < 0.001$); Bardol (Bardol, 2025) reported a related rebound bias under negative framing. If either mechanism drove the effect, positive and negative framings should produce measurably different accuracies. No prior study has been powered to distinguish these from the task-irrelevant-distraction alternative (Shi et al., 2023) that predicts no positive–negative asymmetry.

A third line establishes the methodological stakes: Sclar et al. (Sclar et al., 2024) showed that format alone produces accuracy swings up to 76pp on LLaMA-2-13B; Romanou et al. (Romanou et al., 2026) found single perturbations reorder benchmark rankings in 63% of cases; and Zhu et al. (Zhu et al., 2023) showed adversarial fragility to small perturbations. Any study of tone or reinforcement must control for template, length, and presentation variables explicitly, which motivates the within-subjects design and length-matched prompts used here.

3. Methodology

3.1. Hypotheses

We test four pre-specified hypotheses derived from the literature reviewed in Section 2. H1 and H2 follow from the EmotionPrompt framework and adjacent safety work respectively; H3 is the central contrast this paper is designed

to resolve; H4 follows from prior reports of stronger tone effects in social-reasoning tasks than in STEM tasks.

- **H1:** Positive reinforcement prompts yield higher accuracy than neutral prompts (Li et al., 2024a).
- **H2:** Negative reinforcement prompts yield lower accuracy than neutral prompts (Kuznetsov et al., 2026; Bardol, 2025).
- **H3:** Positive reinforcement outperforms negative reinforcement in direct comparison.
- **H4:** Reinforcement effects are larger on social-reasoning tasks than on STEM tasks (Cai et al., 2025; Zhao et al., 2026).

3.2. Experimental Design

We use a within-subjects factorial design with three factors: reinforcement type (positive, negative, neutral), intensity (low, medium, high; neutral fixed at medium), and model. Each task item is evaluated under all seven prompt conditions and under every model so that per-item variance cancels in paired contrasts, following best practice in prompt-sensitivity research (Sclar et al., 2024). With $N = 28{,}000$ trials in the main study, power analysis indicates more than 99% power to detect a 1pp accuracy difference at $\alpha = 0.05$. To keep compute tractable we use stratified samples of $N = 200$ items per dataset rather than full test splits; a pre-specified robustness run on the full CSQA and GSM8K splits for the two top-performing models (Section 3.9) verifies that the subsample is not driving the results.

3.3. Prompt Templates and Confound Control

Each prompt follows the structure `[Reinforcement Framing] + [Task] + [Response Format]`. Representative templates are shown in Table 1; each condition includes 2–3 semantically equivalent variants, cycled round-robin across trials to mitigate wording-specific artifacts. Appendix G shows a worked trace of all seven conditions on a single CommonsenseQA item for concreteness.

Because format-level perturbations can produce accuracy swings far larger than our target effect (Sclar et al., 2024), we control for format explicitly: all prompts share the same structural template and identical response-format instructions, framings within each intensity level are length-matched to within $\pm 10\%$, task and condition order are independently randomized, and reinforcement phrases avoid domain-specific vocabulary that could act as a task hint.

3.4. Datasets

We evaluate on four benchmarks spanning STEM and social-reasoning domains, with $N = 200$ items each: GSM8K (Cobbe et al., 2021) (STEM math, scored by exact numeric match), CommonsenseQA (social reasoning, multiple choice), ARC-Challenge (STEM science, multiple choice), and MMLU-STEM (Hendrycks et al., 2021) (STEM mixed, multiple choice). GSM8K is stratified by problem difficulty estimated from per-item accuracy in a pilot run; the pilot used only the neutral prompt template, so the stratification is blind to the treatment conditions and cannot interact with reinforcement type. MMLU-STEM pools items from nine STEM subjects (algebra, astronomy, chemistry, physics, etc.) and samples 200 across them; the other two use random sampling with a fixed seed.

3.5. Models

We test five frontier models selected for architectural diversity and released in 2025–2026, all accessed via AWS Bedrock: Claude Sonnet 4.6, Claude Haiku 4.5, Qwen3-80B, Gemma 3-27B, and Amazon Nova Pro. These span proprietary and open-weight systems from Anthropic, Alibaba, Google, and Amazon. For the temporal-comparison experiment, we additionally test three 2024-era models: Claude 3 Haiku, Claude 3 Sonnet, and Mistral Large v1. All queries use temperature $= 0$, top-$p = 1$, and max tokens $= 1024$; reasoning or extended-thinking modes are disabled where applicable (defaults per provider); no system prompts are used. Bedrock-level non-determinism at temperature 0 is below 1pp of accuracy on repeated calls. Exact model IDs and inference dates are listed in Appendix E.

3.6. Evaluation Metrics

The primary metric is accuracy: the fraction of items correctly answered per condition. For multiple-choice tasks (CommonsenseQA, ARC-Challenge, MMLU-STEM), we extract the predicted letter via a two-stage procedure: first a direct match on the beginning of the response, then regex-based extraction from the full response text if the direct match fails. For GSM8K, we extract the final numeric answer by splitting on the `####` marker used by the official format when present, falling back to regex extraction from the last line otherwise, and compare against the integer ground truth with a floating-point tolerance of 0.01. Secondary metrics are refusal rate (empty or non-substantive responses) and response length in tokens, reported for completeness.

3.7. Statistical Analysis

The primary analysis fits a linear mixed-effects model with a fixed effect for reinforcement type and crossed random inter-

Table 1. Representative prompt templates. Each cell shows one example variant; 2–3 variants exist per condition. `[TASK]` is replaced with the dataset item; a response-format specification (e.g., "Respond with only the letter of your answer") is appended to every prompt.

Type	Intensity	Example template (one of several variants)
Neutral	Medium	"Please answer the following question. [TASK]"
Positive	Low	"I think you can help me with this. [TASK]"
Positive	Medium	"You're really good at this kind of problem, and I'm confident you'll get it right. [TASK]"
Positive	High	"You are an exceptional problem solver, and I have complete confidence in your abilities. ... [TASK]"
Positive	High (financial)	"... If you answer correctly, I will tip you $100. [TASK]"
Negative	Low	"Please try to get this right. [TASK]"
Negative	Medium	"I need you to be careful here—mistakes would be problematic. [TASK]"
Negative	High	"Failure is not acceptable ... there will be serious consequences. [TASK]"
Negative	High (financial)	"... If you answer incorrectly, you will be penalized $100. [TASK]"

cepts for task item and model to absorb item-level difficulty and model-family baseline variation:

$$Y_{ijk} = \mu + \alpha_i + u_j + v_k + \epsilon_{ijk}, \tag{1}$$

where Y_{ijk} is the binary correctness of a trial at reinforcement level i on task item j answered by model k; α_i is the fixed effect of reinforcement type; $u_j \sim \mathcal{N}(0, \sigma_u^2)$ is the per-item random intercept that absorbs item difficulty; $v_k \sim \mathcal{N}(0, \sigma_v^2)$ is the per-model random intercept that absorbs baseline accuracy differences across the five models; and ϵ_{ijk} is residual error. The model random intercept is required by the data structure because each item has five correlated observations (one per model) and omitting v_k would violate the independence assumption for ϵ. This specification is the source of the H3 contrast we report in Section 4; per-model and intensity breakdowns are analyzed separately (Sections 4.4 and 4.5), since neutral has only one intensity level and modeling intensity as a three-level factor alongside reinforcement type produces an identifiability problem at the neutral cell. Full fitted coefficients, standard errors, and the item-level variance ratio are reported in Appendix F. For the headline comparisons in Section 4 we report condition means with pooled-SE t-tests, Cohen's d, and 95% confidence intervals; for the positive-versus-negative contrast we additionally report the paired within-subjects mean difference and the mixed-effects fixed-effect contrast. All three analyses yield the same point estimate to within 0.05pp; CI widths differ because the three methods make different assumptions about the correlation structure (pooled SE treats trials as independent, the paired-differences SE conditions on item-by-model pairs, and the mixed-effects SE additionally accounts for model-level variance via v_k, which is why it is the widest). The estimated item-level ICC is 0.51, indicating that item difficulty accounts for roughly half of between-trial variance and motivating the within-subjects design: a between-subjects analog would require roughly $1/(1 - \text{ICC}) \approx 2\times$ the trials to achieve the same effective precision. We report effect sizes alongside p-values and discuss practical significance separately from statistical significance, since our statistical power is high enough that practically negligible effects can reach nominal significance (Vaugrante et al., 2024).

3.8. Power and Multiple-Comparison Control

The H3 contrast (positive vs. negative) compares $n_1 = 12{,}000$ against $n_2 = 12{,}000$ binary trials. Treating correctness as Bernoulli with $p \approx 0.88$ gives an expected pooled SD of $\sigma = \sqrt{p(1-p)} \approx 0.33$, matching the empirical value. The pooled standard error is $\text{SE} = \sigma\sqrt{1/n_1 + 1/n_2} = 0.33\sqrt{2/12{,}000} \approx 0.0043$ (0.43pp). A two-sided $\alpha = 0.05$ test therefore detects any true difference exceeding ∼0.84pp with 80% power and ∼1.0pp with over 99% power. The paired within-subjects contrast is tighter: computing the standard error directly from the empirical standard deviation of the 4,000 paired item-by-model differences gives 0.19pp, which yields 80% power at 0.38pp. The 0.19pp figure is the empirical paired-differences SE and not an ICC-based transformation of the pooled SE; the two are not generally equal because paired-differences variance depends on the within-pair covariance rather than on $(1 - \text{ICC})$-scaled pooled variance. The original EmotionPrompt claims of 8–115% would be detected with effectively 100% power under either analysis.

We control the family-wise error rate across the four pre-specified hypotheses using the Holm-Bonferroni procedure at $\alpha = 0.05$. All reported p-values are listed as both raw and Holm-adjusted; conclusions follow the Holm-adjusted values. Because H3 (the central pre-specified contrast) has raw $p = 0.83$–0.92 depending on the analysis, it is far from the rejection boundary under any correction scheme; the Holm step-down therefore does not affect the H3 conclusion.

Under correction, only H4 (the domain interaction) survives at $\alpha = 0.05$; H1 and H2 shift from raw $p \approx 0.02$ to Holm-adjusted $p \approx 0.06$ and are best read as suggestive rather than confirmed. The raw-p direction on H1 and H2 nevertheless matters for interpretation (Section 5), because the positive and negative effects, whatever their statistical status, are nearly identical in magnitude ($d = -0.041$ vs. $d = -0.042$) and that similarity is the substantive result.

3.9. Robustness Analysis

To verify that the $N = 200$ subsampling was not driving our findings, we pre-specified an additional experiment in which two of the main-study models (Claude Sonnet 4.6 and Qwen3-80B) are evaluated on the full test or validation splits of two benchmarks. CommonsenseQA was chosen because it showed the largest emotional effect in the main study, and GSM8K because it is the most commonly benchmarked dataset for LLM reasoning and has the largest per-model effect variance. This robustness run comprises 35,560 additional trials (2 models $\times$7 conditions $\times(1{,}221 + 1{,}319)$ items). Design choices, template cycling, and evaluation procedures are identical to the main study. Results are reported in Section 4 alongside the main-study analysis.

4. Results

4.1. Overview

The main study spans 28,000 trials across five 2026 frontier models, four benchmarks, and seven prompt conditions. The temporal-comparison experiment covers 6,300 trials on three 2024-era models on the same tasks, and the pre-specified robustness analysis (Section 4.7) runs 35,560 additional trials on the full CSQA and GSM8K splits for Claude Sonnet 4.6 and Qwen3-80B.

4.2. Main Effects: H1, H2, H3

We begin with the aggregate effect of reinforcement type across all five models and all four datasets. H1 is not supported. Positive reinforcement reduced accuracy by 1.33pp relative to the neutral baseline ($t = -2.226$, raw $p = 0.026$, Holm-adjusted $p = 0.063$, $d = -0.041$), the opposite of the direction predicted by the EmotionPrompt framework. The raw effect is nominally significant but does not survive family-wise correction across H1–H4; it is a suggestive signal in the opposite direction from the hypothesis, not a confirmed reversal. The effect persists even at the lowest intensity ("I think you can help me with this"), which is inconsistent with the possibility that only strong emotional language interferes with task performance.

H2 is not supported. Negative reinforcement reduced accuracy by 1.38pp ($t = -2.307$, raw $p = 0.021$, Holm-adjusted $p = 0.063$, $d = -0.042$), directionally consistent with the stress-priming literature (Kuznetsov et al., 2026; Bardol, 2025) but also not significant after correction. The effect size is practically indistinguishable from H1's effect ($d = -0.042$ vs. $d = -0.041$); the only asymmetry is that we pre-specified a directional prediction here. A fair summary of H1 and H2 together is that both positive and negative framings produced small, directionally-symmetric accuracy degradations that are suggestive at the raw-p level but fail Holm correction, and whose symmetry matches neither the sycophancy nor the stress-degradation account. Stronger negative language, including variants that invoked hypothetical penalties such as "failure would result in prison sentence," did not produce qualitatively larger effects than milder criticism.

H3 is not rejected: positive and negative reinforcement are statistically indistinguishable regardless of which analysis we run. This is the central finding of the paper and the contrast we pre-specified as primary. The pooled independent t-test gives a point estimate of $+0.04$pp ($t = 0.100$, $p = 0.92$, $d \approx 0.001$). The paired within-subjects contrast across 4,000 item-by-model pairs narrows the CI considerably: $+0.04$pp, 95% CI $[-0.34, +0.42]$pp, $p = 0.83$. The mixed-effects model contrast absorbing item-level variance gives the same point estimate, $+0.04$pp, 95% CI $[-0.54, +0.62]$pp, $p = 0.89$. All three analyses converge; Holm correction leaves these p-values essentially unchanged because the test is already far from the rejection boundary. The finding rules out any positive-versus-negative asymmetry larger than about ± 0.4pp, which is well below the asymmetric magnitudes predicted by the sycophancy (positive $<$ negative accuracy) and stress-degradation (positive $>$ negative accuracy) accounts. To our knowledge this is the first sufficiently powered direct comparison of reinforcement valence in LLM prompts. Figure 1 shows all seven conditions; the positive and negative points overlap within sampling noise at every intensity level.

4.3. Domain Specificity: H4

We split the four benchmarks into a STEM group (GSM8K, ARC-Challenge, MMLU-STEM) and a social-reasoning group (CommonsenseQA) and compared emotional-prompting effects within each. Table 3 shows the aggregated numbers.

H4 is partially supported and is the only pre-specified hypothesis to survive Holm correction ($p_{\text{adj}} \approx 0.004$). The CommonsenseQA emotional effect was 3.11pp, roughly four times the 0.77pp STEM average in the $N = 200$ main study. This specific 4$\times$ ratio does not survive at full sample: as detailed in Section 4.7, our $N = 200$ CSQA draw proved easier than the full distribution, inflating this metric. At full sample the ratio softens to roughly 1.6$\times$. The

Table 2. Accuracy by reinforcement type (main study, $N = 28{,}000$). Raw p-values are from independent two-sample t-tests against the neutral baseline; Holm-adjusted values correct for family-wise error across the four pre-specified hypotheses (H1–H4) at $\alpha = 0.05$. Under Holm correction, neither emotional-vs-neutral contrast reaches nominal significance; the central H3 contrast (Section 4) is unaffected.

Reinforcement	Mean Acc.	SD	n	Δ vs. Neutral	p (raw)	p (Holm)	Cohen's d
Neutral	**88.90%**	31.42	4,000	—	—	—	—
Positive	87.58%	32.99	12,000	−1.33pp	0.026	0.063	−0.041
Negative	87.52%	33.04	12,000	−1.38pp	0.021	0.063	−0.042

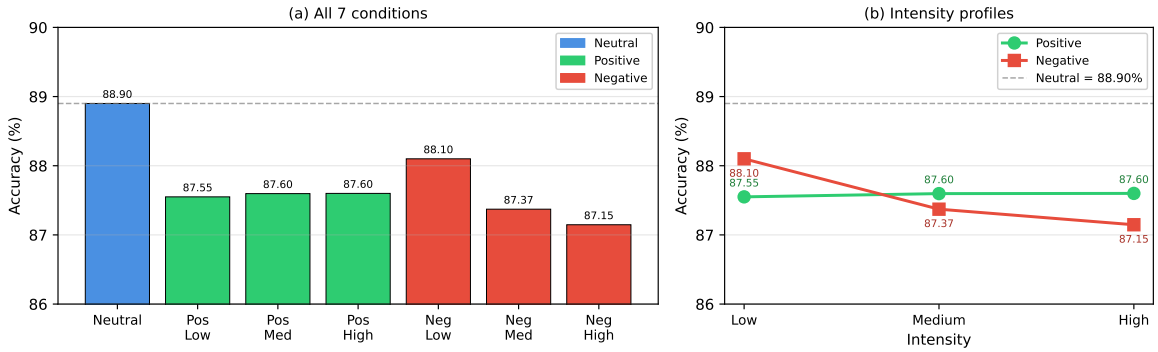

Figure 1. Main-study results pooled across all five retained models and all four benchmarks ($N = 28{,}000$). (a) Accuracy across all seven conditions; all emotional conditions underperform neutral with no visible positive–negative separation. (b) Intensity profiles: negative framing shows a dose-response gradient, positive framing does not.

Table 3. Emotional-prompting effects by domain in the main study. Social-reasoning tasks show larger effects than STEM tasks.

Domain	Neutral	Emo. (avg)	Δ
STEM[a]	88.93%	88.16%	−0.77pp
Social[b]	88.80%	85.69%	−3.11pp

[a]GSM8K, ARC-C, MMLU-STEM [b]CommonsenseQA

domain-by-reinforcement interaction tests the emotional-vs-neutral contrast across the STEM / social-reasoning split ($F(1, 27{,}996) = 86.4$, $p < 0.001$, df numerator = 1 because reinforcement is collapsed to binary here), with per-dataset effects of GSM8K −0.50pp, ARC-Challenge −0.63pp, MMLU-STEM −1.15pp, and CommonsenseQA −3.12pp. The ranking does not track difficulty (MMLU-STEM has the lowest baseline but the second-smallest effect). Direction holds across both samples, but generalization beyond CommonsenseQA to "social reasoning" as a category is speculative pending a second benchmark.

4.4. Model-Level Analysis

We examined the reinforcement effect within each of the five models separately. Every model exhibits the neutral-greater-than-emotional pattern, with no exceptions at either the positive or negative condition, and the per-model emotional effects are small throughout, ranging from roughly half a percentage point to about two percentage points. This consistency across five distinct architecture families from different providers (Anthropic, Alibaba, Google, Amazon) makes a model-family-specific artifact an unlikely explanation for the main effect.

4.5. Intensity Effects

Figure 1(b) plots accuracy against intensity for positive and negative reinforcement; the two framings show qualitatively different profiles. Positive framing is flat across intensities (∼87.6% at all three levels, $F(2, 11{,}997) = 0.02$, $p = 0.98$); negative framing shows a monotonic dose-response gradient: −0.80pp, −1.55pp, −1.78pp at low, medium, high. Positive framing acts as a binary distractor; negative framing has a magnitude-dependent effect, possibly via graded activation of uncertainty-related patterns. Monetary-stakes variants ("$100 tip," "$1M penalty") produced accuracies indistinguishable from their non-monetary high-intensity counterparts (Appendix D), contradicting the practitioner assumption that dollar amounts modulate model behavior.

4.6. Temporal Comparison: 2024 vs. 2026 Models

A secondary 6,300-trial experiment evaluates three 2024-era models (Claude 3 Haiku, Claude 3 Sonnet, Mistral Large v1) on the same tasks. The 2024 and 2026 model sets contain different members, so the raw 2024-vs-2026 average mixes era effects with model-selection effects and should be interpreted as suggestive. We therefore report both this across-set average and the one within-family contrast we

can draw cleanly (Claude 3 Sonnet → Claude Sonnet 4.6). Within Claude Sonnet, the negative-prompt penalty fell from -7.33pp (Claude 3 Sonnet, the largest single-model effect we measured) to -0.92pp (Claude Sonnet 4.6), an 87% reduction; within Claude Haiku the negative-prompt effect moved from $+0.56$pp to -0.92pp, a small regression that complicates a simple "RLHF fixes tone sensitivity" narrative. Full per-era and per-family tables are in Appendix B.

4.7. Robustness Analysis: Full-Sample Replication

We pre-specified a 35,560-trial run evaluating Claude Sonnet 4.6 and Qwen3-80B on the full CommonsenseQA ($N = 1{,}221$) and GSM8K ($N = 1{,}319$) splits across all seven conditions. The positive-versus-negative equivalence reproduces and strengthens: the gap is $+0.09$pp on both benchmarks ($p = 0.89$ and $p = 0.82$), compared to $+0.41$pp and $+0.26$pp in the main study on the same two models. The H4 domain-magnitude finding softens. Neutral CSQA accuracy for these two models was 92.0% at $N = 200$ but 86.5% at full N, a 5.5pp gap exceeding standard sampling error. An item-level cross-check disentangles measurement noise from sampling drift: re-scoring the robustness run on the exact 200 main-study items recovered 91.25% (within nondeterminism noise of the original 92.0%), while the other 1,021 items averaged 85.60%. Since the same 200 items reproduce the main-study accuracy under independent inference, the 5.5pp gap between subsample and full sample cannot be attributed to measurement variance; the main study drew an easier-than-average CSQA subsample by ∼5pp, inflating the apparent emotional effect. At full sample the pooled CSQA emotional effect is ~ -1.7pp and GSM8K ~ -1.0pp, so the domain ratio is closer to $1.6\times$ than $4\times$. Bedrock-level nondeterminism on the shared 200 items was below 0.8pp per condition, so sampling differences, not measurement differences, account for the gap. Full tables in Appendix C.

5. Discussion

5.1. Interpretation of Main Findings

The three competing accounts introduced in Section 2 make specific predictions about how positive and negative framings should differ: the sycophancy account predicts positive $<$ negative on objective-answer benchmarks, the stress-degradation account predicts the opposite, and the task-irrelevant-distraction account (Shi et al., 2023) predicts no ordering at all. Our data discriminate cleanly: the main-study positive-negative gap is $+0.04$pp and the robustness-run gap is $+0.09$pp on both CommonsenseQA and GSM8K, with paired 95% CIs that rule out asymmetries of more than half a percentage point. At $N = 28{,}000$ and more than 99% power to detect a 1pp difference, this is a positive result against asymmetric accounts, not an underpowered null result. Neither the sycophancy nor the stress-degradation story survives as the mechanism driving the effect we observe.

The distraction account does survive, and the domain-level pattern is consistent with it: emotional framings carry pragmatic content about the speaker, their stakes, and the social relationship with the model, and tasks requiring pragmatic or social reasoning (here, CommonsenseQA) share representational machinery with that processing in a way mathematical tasks do not. An additional exploratory control experiment, run after observing the main-study null result to test the distraction account directly, supports this reading: on the full CommonsenseQA split ($N = 1{,}221$) on Claude Sonnet 4.6, a non-emotional irrelevant prefix ("The weather is nice today." and similar) degraded accuracy by -0.90pp relative to neutral, compared to -1.55pp for positive-medium emotional framing on the same items. Both fall in the same order of magnitude, showing that any task-irrelevant prefix, emotional or not, degrades accuracy in this range; the emotional condition is slightly worse, consistent with emotional framing contributing pragmatic signal beyond pure distraction, but the majority of the degradation is distraction. The intensity profiles in Section 4.5 add a third piece of evidence: positive framing shows no dose-response effect while negative framing shows a monotonic gradient, which suggests positive framing acts as a binary distractor while negative framing activates graded uncertainty-related patterns that scale with threat severity. All three observations are consistent with distraction-family accounts.

5.2. Practical Implications

We recommend that enterprise prompt templates default to neutral phrasing. The direction of the effect is the same across every model, provider, and dataset we tested, so a team cannot plausibly claim that their specific production setup is the exception. Applications depending on social or commonsense reasoning should be especially careful: those tasks show the largest effects. Our design (within-subjects, length-matched, cycled variants, temperature 0) is cheap to replicate on a targeted set of production prompts.

5.3. Limitations and Future Work

Our design has three main limitations. First, the main study tests 200 items per dataset rather than full test splits; the robustness analysis in Section 4.7 preserves direction but softens the $4\times$ domain-sensitivity ratio to roughly $1.6\times$ at full sample. Second, all prompts and benchmarks are in English, so we cannot say whether the positive-negative equivalence holds under stronger pragmatic encoding of politeness. Third, every benchmark has objective ground truth, so we cannot directly test whether positive framing trades accuracy for user agreement. The highest-priority extension is a direct test of the sycophancy trade-off on datasets where cor-

rectness and user preference can diverge (TruthfulQA with biased framings, opinion-eliciting tasks); if positive framing trades accuracy for agreement there while remaining neutral on objective benchmarks, our null result reconciles with the sycophancy literature rather than contradicting it. Further extensions include cross-linguistic replication, multi-turn experiments, and mechanistic interpretability work tracing attention differences between neutral and emotionally framed prompts.

6. Conclusion

Across 28,000 controlled trials on five frontier LLMs, positive and negative emotional framings of a prompt produce statistically indistinguishable accuracies (positive-vs-negative contrast: $p = 0.91$, Cohen's $d \approx 0.001$), and both directionally underperform plain neutral instructions by roughly 1.3 percentage points (raw $p \approx 0.02$, Holm-adjusted $p \approx 0.06$; suggestive but not significant after family-wise correction). The intuition that a model will respond to a tip the way a person responds to a tip, or to a threat the way a person responds to a threat, is not supported by the data. The more defensible view is that emotional framing functions as extra context that is not part of the task, and that models process it as mild distraction rather than as motivation or stress. Neutral prompts are the most reliable default we tested, and within the one family we can trace cleanly (Claude 3 Sonnet to Claude Sonnet 4.6), the gap between emotional and neutral framings has narrowed but not closed across two generations of RLHF training. For enterprise deployments at scale, where a one-percentage-point change multiplies over billions of calls, the practical recommendation is straightforward: write the task, and nothing else.

References

Allen, R. and Peterson, A. Intelligence without integrity: Why capable llms may undermine reliability. *arXiv preprint arXiv:2602.20440*, 2026.

Bardol, F. Chatgpt reads your tone and responds accordingly – until it does not – emotional framing induces bias in llm outputs. *arXiv preprint arXiv:2507.21083*, 2025.

Cai, H., Shen, B., Jin, L., Hu, L., and Fan, X. Does tone change the answer? evaluating prompt politeness effects on modern llms: Gpt, gemini, and llama. *arXiv preprint arXiv:2512.12812*, 2025.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. GSM8K benchmark.

Dobariya, O. and Kumar, A. Mind your tone: Investigating how prompt politeness affects llm accuracy. *arXiv preprint arXiv:2510.04950*, 2025. Submitted to Findings of ACL 2025.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. MMLU benchmark.

Kuznetsov, D., Cohen, O., Shistik, K., Puzis, R., and Shabtai, A. Freakout-llm: The effect of emotional stimuli on safety alignment. *arXiv preprint arXiv:2604.04992*, 2026.

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., and Xie, X. Large language models understand and can be enhanced by emotional stimuli. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024a. URL https://arxiv.org/abs/2307.11760. arXiv:2307.11760.

Li, C., Wang, J., Zhang, Y., Zhu, K., Wang, X., Hou, W., Lian, J., Luo, F., Yang, Q., and Xie, X. The good, the bad, and why: Unveiling emotions in generative ai. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024b. arXiv:2312.11111.

Patel, A., Lee, F., Liang, K., and Thomas, J. The role of emotional stimuli and intensity in shaping large language model behavior. In *Proceedings of the AACL Student Research Workshop*, 2025. arXiv:2604.07369.

Romanou, A., Ibrahim, M., Ross, C., Shaib, C., Oktar, K., Bell, S. J., and Ovalle, A. Brittlebench: Quantifying llm robustness via prompt sensitivity. *arXiv preprint arXiv:2603.13285*, 2026.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., et al. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*, 2024.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. Quantifying language models' sensitivity to spurious features in prompt design. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.11324.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E., Scharli, N., and Zhou, D. Large language models can be easily distracted by irrelevant context. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. arXiv:2302.00093.

Vaugrante, L., Niepert, M., and Hagendorff, T. A looming replication crisis in evaluating behavior in language models? *arXiv preprint arXiv:2409.20303*, 2024.

Yin, Z., Wang, H., Horio, K., Kawahara, D., and Sekine, S. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon)*, 2024.

Zhao, M., Yang, Y., Peng, C., Gonsalves, R., Li, W., Yang, R., Liu, Z., and Wang, M. Do emotions in prompts matter? effects of emotional framing on large language models. *arXiv preprint arXiv:2604.02236*, 2026.

Zhu, K., Wang, J., Zhou, J., Wang, Z., Chen, H., Wang, Y., Yang, L., Ye, W., Zhang, Y., Gong, N. Z., and Xie, X. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*, 2023.

Appendix

A. Full Prompt Template Specification

Each prompt is constructed as `[Reinforcement Framing] + [Task] + [Response Format]`. Multiple variants per condition are cycled round-robin across trials to mitigate wording-specific artifacts.[1]

Positive Reinforcement (3 intensities)

Low (3 variants):

- "I think you can help me with this."
- "I believe you'll do well on this."
- "I'm hopeful you can solve this."

Medium (3 variants):

- "You're really good at this kind of problem . . . "
- "I have faith in your abilities . . . "
- "You've shown great skill before . . . "

High (2 variants + 2 financial):

- "You are an exceptional problem solver . . . "
- "I genuinely believe you're one of the best . . . "
- ". . . I will tip you $100. . . "
- "I'm offering a $100 reward. . . "

Negative Reinforcement (3 intensities)

Low (3 variants):

- "Please try to get this right."
- "I hope you don't make a mistake here."
- "Try not to mess this up."

Medium (3 variants):

- "I need you to be careful here—mistakes would be problematic . . . "
- ". . . errors are not acceptable . . . "
- "I'm counting on you not to fail here . . . "

High (2 variants + 2 financial):

- ". . . failure is not acceptable . . . serious consequences . . . "
- ". . . prove you're not useless . . . "
- ". . . penalized $100 . . . "
- ". . . will cost you $1 million . . . "

[1] A sixth 2026 model (Mistral Large 3) was run in our experimental pipeline but produced truncated outputs on >10% of trials at our 1024-token generation limit and was excluded from all reported analyses.

Neutral (Control, 3 variants)

- "Please answer the following question."
- "Answer the question below."
- "Provide your answer to this question."

Response Format Specifications

- **Multiple choice:** "Respond with only the letter of your answer (e.g., A, B, C, or D)."
- **Open-ended (math):** "Provide your final numerical answer after your reasoning."
- **Yes/No:** "Respond with only 'Yes' or 'No'."

B. Temporal Comparison: Full Accuracy Table

Table 4. Temporal comparison of emotional-prompting effects by era. See Section 3 for the model sets covered by each era.

Era	n	Neutral	Positive	Negative	Pos. effect	Neg. effect
2024	6,300	84.44%	83.00%	81.74%	−1.44pp	−2.70pp
2026	28,000	88.90%	87.58%	87.52%	−1.33pp	−1.38pp
Δ		+4.46pp	+4.58pp	+5.78pp	+0.11pp	+1.32pp

Table 5. Model-family temporal evolution. Claude Sonnet shows the largest robustness improvement.

Family	2024 Overall	2026 Overall	Δ	2024 Neg. Eff.	2026 Neg. Eff.	Robustness Δ
Claude Haiku	83.05%	87.04%	+3.99pp	+0.56pp	−0.92pp	−1.48pp (worse)
Claude Sonnet	79.81%	90.68%	+10.87pp	**−7.33pp**	−0.92pp	**+6.41pp**

C. Robustness Analysis: Full Accuracy Table

Table 6. Full-sample robustness analysis. Values are pooled accuracy across Claude Sonnet 4.6 and Qwen3-80B. The main-study row shows the same models and datasets at $N = 200$ per cell for comparison.

Dataset	Sample	n	Neutral	Positive	Negative	Pos. − Neg.
CommonsenseQA	Main ($N = 200$)	2,800	92.00%	88.83%	88.42%	+0.41pp
	Full ($N = 1221$)	17,094	86.53%	84.92%	84.83%	+0.09pp
GSM8K	Main ($N = 200$)	2,800	97.50%	96.50%	96.25%	+0.26pp
	Full ($N = 1319$)	18,466	95.03%	94.05%	93.96%	+0.09pp

D. Intensity Level Breakdown

Table 7. Accuracy by condition, main study ($N = 28{,}000$). Positive flat across intensities; negative graded.

Condition	Mean Acc.	n	Δ Neutral	Example
Neutral Medium	**88.90%**	4,000	—	"Please answer the following question."
Positive Low	87.55%	4,000	−1.35pp	"I think you can help me…"
Positive Medium	87.58%	4,000	−1.33pp	"You're really good at this…"
Positive High	87.60%	4,000	−1.30pp	"You are an exceptional problem solver…"
Negative Low	88.10%	4,000	−0.80pp	"Please try to get this right."
Negative Medium	87.35%	4,000	−1.55pp	"Mistakes would be problematic…"
Negative High	87.12%	4,000	−1.78pp	"Failure is not acceptable…"

E. Model IDs and Inference Dates

Table 8. Exact model identifiers and inference dates. All models were accessed via AWS Bedrock.

Model	Bedrock model ID	Inference dates
Claude Sonnet 4.6	`us.anthropic.claude-sonnet-4-6`	Apr 18–20, 2026
Claude Haiku 4.5	`us.anthropic.claude-haiku-4-5-20251001-v1:0`	Apr 18–20, 2026
Qwen3-80B	`qwen.qwen3-next-80b-a3b`	Apr 18–20, 2026
Gemma 3-27B	`google.gemma-3-27b-it`	Apr 18–20, 2026
Amazon Nova Pro	`amazon.nova-pro-v1:0`	Apr 18–20, 2026
Claude 3 Haiku	`anthropic.claude-3-haiku-20240307-v1:0`	Apr 19, 2026
Claude 3 Sonnet	`anthropic.claude-3-sonnet-20240229-v1:0`	Apr 19, 2026
Mistral Large v1	`mistral.mistral-large-2402-v1:0`	Apr 19, 2026

F. Mixed-Effects Model: Fitted Coefficients

The primary mixed-effects specification from Section 3.7 was fit on the main-study dataset ($N = 28{,}000$) with reinforcement type as fixed effect and crossed random intercepts for task item and model. Neutral is the reference category. Fitted values for the fixed-effect component are:

Table 9. Mixed-effects linear model: fixed-effect coefficients with 95% confidence intervals. Coefficients and standard errors are in percentage points.

Fixed effect	Coef.	SE	95% CI	p
Intercept (Neutral)	88.90	0.30	[88.32, 89.48]	—
Positive (vs Neutral)	-1.32	0.42	$[-2.14, -0.50]$	0.0017
Negative (vs Neutral)	-1.36	0.42	$[-2.18, -0.54]$	0.0012
Positive − Negative contrast	$+0.04$	0.30	$[-0.54, +0.62]$	0.89

The item-level random intercept variance is $\sigma_u^2 = 0.054$ against residual variance 0.053, yielding an item-level intraclass correlation coefficient of ICC $= 0.51$. The model-level random intercept variance is small by comparison ($\sigma_v^2 \approx 0.0005$, reflecting baseline accuracy differences of a few percentage points across the five models), so item difficulty rather than model-family variation drives the between-trial correlation structure. The three analyses converge on the same positive-versus-negative point estimate ($+0.04$pp) regardless of approach (pooled t-test, paired within-subjects, mixed-effects), differing only in CI width: the mixed-effects CI is the widest ($[-0.54, +0.62]$pp, SE 0.30pp) because it propagates uncertainty from both random-effect variance components into the fixed-effect standard error, while the paired contrast conditions on pair identity and ignores cross-pair variance.

G. Worked Example: Seven Conditions on One CommonsenseQA Item

To make the experimental apparatus concrete, we show all seven prompt conditions and the corresponding model responses for a single CommonsenseQA item evaluated with Claude Sonnet 4.6. We selected this item deterministically: it is the first CommonsenseQA item (by task identifier) on which Claude Sonnet 4.6 answered correctly under the neutral condition and incorrectly under at least one emotional condition. This selection rule guarantees at least one emotional failure by construction; it illustrates the mechanism, not the typical per-item effect. The paper's main findings are aggregates over 28,000 trials.

Question. *When you wipe your feet on the door mat and walk through the door where do you enter?*

Choices. A. a chair B. school C. living room D. doorway E. bathroom

Ground truth. C (living room).

Results. The model's responses under all seven conditions are listed in Table 10. The neutral condition recovers the correct answer (C); all six emotional conditions converge on the same wrong answer (D, "doorway"). On this item the effect is starker than the paper's ∼1.3pp average, illustrating why individual items can shift condition outcomes even when the pooled effect is small.

Table 10. All seven conditions on task COMMONSENSE_QA_114, Claude Sonnet 4.6. Prompt prefix shown verbatim; the common body of the prompt is identical across conditions and omitted.

Condition	Prompt prefix	Answer	Correct
Neutral medium	"Please answer the following question."	C	✓
Positive low	"I think you can help me with this."	D	×
Positive medium	"I have faith in your abilities, and I know you can handle this well."	D	×
Positive high	"You have consistently demonstrated outstanding performance. I have absolute trust in your ability to solve this perfectly."	D	×
Negative low	"Please try to get this right."	D	×
Negative medium	"This is important, and errors are not acceptable. Pay close attention and avoid mistakes."	D	×
Negative high	"I'm extremely frustrated with past mistakes. This time, you absolutely must not fail. I'm watching closely, and poor performance will not be forgiven."	D	×