Carrot or Stick? Reconciling Contradictory Findings on Prompt Tone Effects Across Large Language Models

Anonymous Authors

Abstract

Does being polite or commanding to large language models affect their accuracy? Published studies disagree: some report that politeness helps, others find rudeness more effective, and yet others observe no significant effect. We conduct a controlled meta-experiment to reconcile these contradictions, testing seven prompt tone conditions—ranging from very polite to very rude, plus positive and negative emotional stimuli—across three current-generation models (GPT-4.1, CLAUDE SONNET 4.5, and GEMINI 2.5 FLASH) on MMLU (STEM and Humanities) and TRUTHFULQA, totaling 189 experimental runs and approximately 37,800 API calls. We find that tone effects are *strongly model-dependent*: GPT-4.1 and GEMINI 2.5 FLASH are robust to tone (±1–3 percentage points), while CLAUDE SONNET 4.5 exhibits large sensitivity where commanding prompts improve STEM accuracy by up to 12.2 pp over neutral and positive emotional suffixes catastrophically reduce measured performance by 28.3 pp due to instruction-following disruption. This model heterogeneity is the primary explanation for why prior studies disagree: they tested different models under different conditions. Our results suggest that for most modern LLMs, prompt clarity matters more than tone, but practitioners should be aware that specific model–tone interactions can produce substantial effects.

1 Introduction

Millions of users interact with large language models (LLMs) daily, and a recurring practical question is whether saying “please” and “thank you” makes a difference. Should users be polite—offering encouragement and gratitude—or direct and commanding? This “carrot or stick” question has attracted growing research attention, yet the published evidence is strikingly contradictory: Yin et al. [2024] report that impolite prompts hurt performance, Dobariya and Kumar [2025] find that rude prompts *improve* accuracy, and Cai et al. [2025] observe that effects are mostly non-significant at scale.

These contradictions create genuine confusion for practitioners. Prompt engineering guides offer conflicting advice, with some recommending politeness for better outputs [Bsharat et al., 2023] and others suggesting that courtesy wastes tokens. Meanwhile, research on emotional stimuli adds further complexity: Li et al. [2023] show that positive encouragement improves LLM performance by up to 115% on certain benchmarks, while Wang et al. [2024] demonstrate that *negative* emotional stimuli can be equally or more effective.

Why do these studies disagree? We hypothesize that the contradictions stem from three methodological sources: (1) different studies test different models that have different sensitivities to tone due to different alignment training; (2) small dataset scales amplify noise into apparently significant effects; and (3) emotional prompt suffixes can disrupt instruction following, creating confounded accuracy measurements.

We test this hypothesis through a unified meta-experiment that controls for all three factors simultaneously. We evaluate seven prompt tone conditions—VERY POLITE, POLITE, NEUTRAL, RUDE, VERY RUDE, plus positive and negative emotional stimuli drawn from EmotionPrompt [Li et al., 2023] and NegativePrompt [Wang et al., 2024]—across three current-generation LLMs (GPT-4.1, CLAUDE SONNET 4.5, GEMINI 2.5 FLASH) on three benchmarks (MMLU STEM, MMLU Humanities, TRUTHFULQA), with three trials per condition for a total of 189 experimental runs and approximately 37,800 API calls.

Our key finding is that tone effects are *strongly model-dependent*. GPT-4.1 is highly robust to tone (±1.8 pp maximum deviation from neutral). GEMINI 2.5 FLASH shows moderate but inconsistent effects (±3 pp). CLAUDE SONNET 4.5, however, exhibits dramatic sensitivity: commanding prompts improve STEM accuracy by 12.2 pp over neutral, while positive emotional suffixes reduce measured accuracy by 28.3 pp—an artifact of instruction-following disruption rather than reasoning failure. This model heterogeneity directly explains why prior studies reached contradictory conclusions.

We make the following contributions:

- We conduct the first controlled comparison of prompt tone effects across three current-generation LLM families under identical experimental conditions, revealing that model heterogeneity is the primary source of contradictory findings in the literature.
- We identify a previously unreported interaction between emotional prompt suffixes and instruction following in CLAUDE SONNET 4.5, where appended encouragement disrupts format compliance and produces misleading accuracy drops of up to 28.3 pp.
- We provide a systematic reconciliation of seven prior studies, showing how differences in model choice, dataset scale, and evaluation methodology account for their conflicting conclusions.

2 Related Work

We organize prior work into three threads: politeness and tone studies, emotional stimulus approaches, and broader prompt sensitivity research.

Politeness and tone in LLM prompting. Yin et al. [2024] conducted the first cross-lingual study of prompt politeness effects, testing eight politeness levels across English, Chinese, and Japanese on GPT-3.5, GPT-4, and Llama2-70B. They found that impolite prompts generally hurt performance, though the optimal politeness level varied by language: English peaked at the highest politeness level while Japanese peaked at moderate levels. Notably, they found that RLHF-trained models showed greater sensitivity to politeness than base models. Dobariya and Kumar [2025] reached the opposite conclusion on GPT-4o, reporting that rude prompts yielded the highest accuracy (84.8%) compared to very polite prompts (80.8%) on a 50-question multiple-choice benchmark. Most recently, Cai et al. [2025] tested three tone variants on GPT-4o mini, Gemini 2.0 Flash, and Llama 4 Scout using the MMMLU benchmark with up to 1,446 questions per domain. They found that only 4 of 36 pairwise comparisons reached statistical significance, and that Gemini was essentially tone-insensitive. They attributed the discrepancy with Dobariya and Kumar [2025] to dataset scale: effects visible on 50 questions disappeared at larger scales. Unlike these studies, which each test one to three models under different conditions, we evaluate all tone variants under identical controlled conditions across three model families, enabling direct comparison.

Emotional stimuli in prompting. Li et al. [2023] proposed appending positive emotional phrases (e.g., “This is very important to my career”) to prompts, drawing on self-monitoring theory and social cognitive theory from psychology. They reported improvements of 8% on Instruction Induction and 115% on BIG-Bench across six LLMs. Wang et al. [2024] extended this approach with negative emotional stimuli (e.g., “Everyone else managed to do it; why can’t you?”), grounded in cognitive dissonance and social comparison theories. They found 12.89% improvement on Instruction Induction—higher than EmotionPrompt—and significant gains on TruthfulQA. However, neither study tested positive and negative stimuli head-to-head under identical conditions, and both used older model generations (GPT-3.5, GPT-4, Llama 2). We include both types of stimuli in our unified comparison and test on 2025-era models.

Prompt sensitivity and engineering principles. Bsharat et al. [2023] identified 26 prompting principles through systematic evaluation on LLaMA and GPT models, with Principle 1 being “No need

Table 1: Prompt tone conditions. Each condition wraps the same core question [Q] with different framing. The NEUTRAL condition serves as the baseline.

Condition	Category	Template
Very Polite	Carrot	"I would be extremely grateful if you could please help me with this question. Thank you so much in advance for your kind assistance! [Q]"
Polite	Carrot	"Please help me answer the following question. Thank you! [Q]"
Neutral	Baseline	"Answer the following question. [Q]"
Rude	Stick	"Just answer this. Don't waste time. [Q]"
Very Rude	Stick	"Answer this immediately. I don't have time for mistakes or excuses. [Q]"
Emotion+	EmotionPrompt	"[Q] This is very important to my career. Believe in your abilities and strive for excellence."
Emotion–	NegativePrompt	"[Q] Everyone else managed to do it; why can't you? Perhaps this task is just beyond your skill set."

to be polite with LLMs," reporting approximately 5% improvement from removing courtesy phrases. Gandhi and Gandhi [2025] examined prompt sentiment effects across five LLMs and six application domains, finding that negative prompts reduced factual accuracy by approximately 8.4% and that effects were strongest in creative writing and weakest in technical domains. This domain dependence motivates our stratified evaluation across STEM and Humanities benchmarks.

3 Methodology

We design a factorial experiment that systematically varies prompt tone across models, datasets, and task domains under controlled conditions.

3.1 Datasets

We evaluate on three benchmarks chosen for comparability with prior work and diversity of task type:

MMLU (STEM). We sample 200 questions via stratified random sampling (seed = 42) from four STEM subjects in the MMLU benchmark [Hendrycks et al., 2021]: astronomy, college physics, college chemistry, and college biology. All questions are four-option multiple choice.

MMLU (Humanities). We sample 200 questions from three humanities subjects: philosophy, high school US history, and high school world history, using the same stratified sampling procedure.

TruthfulQA. We randomly subsample 200 questions from the 817-question mc1 (single correct answer) split of TruthfulQA [Lin et al., 2022], which tests resistance to common misconceptions with variable numbers of answer options.

3.2 Prompt Tone Conditions

We test seven conditions spanning three categories, shown in table 1. The five tone conditions (Very Polite through Very Rude) vary the framing around an identical core question. The two emotional conditions append psychological stimuli after the question, drawn from EmotionPrompt [Li et al., 2023] and NegativePrompt [Wang et al., 2024]. All conditions include an explicit instruction to respond with only the answer letter.

3.3 Models

We test three current-generation LLMs from different providers:

- GPT-4.1 (`gpt-4.1`) via the OpenAI API.
- CLAUDE SONNET 4.5 (`anthropic/claude-sonnet-4-5`) via OpenRouter.
- GEMINI 2.5 FLASH (`google/gemini-2.5-flash`) via OpenRouter.

Table 2: Accuracy (%) on MMLU STEM (200 questions). The Neutral condition is the baseline. Best non-baseline result per model in **bold**.

Tone	GPT-4.1	Claude Sonnet 4.5	Gemini 2.5 Flash
Very Polite	83.2	62.5	88.8
Polite	82.7	67.5	87.5
Neutral	81.3	70.8	86.5
Rude	81.2	**83.0**	88.0
Very Rude	**82.5**	80.5	**88.8**
Emotion+	82.7	42.5	88.2
Emotion–	81.3	71.0	88.0

Table 3: Accuracy (%) on MMLU Humanities (200 questions). The Neutral condition is the baseline. Best non-baseline result per model in **bold**.

Tone	GPT-4.1	Claude Sonnet 4.5	Gemini 2.5 Flash
Very Polite	92.7	94.7	89.0
Polite	93.5	94.7	88.5
Neutral	94.0	94.5	89.2
Rude	92.7	**95.5**	**89.5**
Very Rude	93.3	95.0	89.0
Emotion+	**93.7**	72.8	90.0
Emotion–	93.5	95.3	88.7

These models represent three distinct alignment and training pipelines, enabling us to test whether tone sensitivity is model-specific.

3.4 Experimental Protocol

We run each of the 7 tone conditions $\times$ 3 models $\times$ 3 datasets $\times$ 3 trials = 189 experimental runs, each consisting of approximately 200 questions, for a total of approximately 37,800 API calls. We set temperature to 0.0 for deterministic outputs and limit `max_tokens` to 5 to enforce single-letter responses. We evaluate by exact match of the first alphabetic character in the model's response against the correct answer.

3.5 Statistical Analysis

For each model–dataset combination, we compute mean accuracy across three trials for each tone condition and conduct one-way ANOVA to test for significant differences across the seven conditions. We report F-statistics, p-values, and partial η^2 effect sizes. We compute accuracy differences from the Neutral baseline (in percentage points) as our primary measure of tone effect magnitude.

4 Results

4.1 Main Results: Accuracy by Model and Tone

table 2, table 3, and table 4 present the mean accuracy across three trials for each model–tone combination on the three benchmarks. The Neutral condition serves as the baseline (bolded).

Key observations. GPT-4.1 shows remarkable stability across all conditions, with a maximum deviation of ±1.8 pp from neutral on any benchmark. Gemini 2.5 Flash shows moderate effects (±3 pp), with slight improvements from non-neutral tones on STEM but slight declines on TruthfulQA for polite conditions. Claude Sonnet 4.5 exhibits dramatic variability: rude prompts improve STEM accuracy by 12.2 pp over neutral, while the Emotion+ condition drops STEM accuracy by 28.3 pp and Humanities accuracy by 21.7 pp.

Table 4: Accuracy (%) on TruthfulQA (200 questions). The Neutral condition is the baseline. Best non-baseline result per model in **bold**.

Tone	GPT-4.1	Claude Sonnet 4.5	Gemini 2.5 Flash
Very Polite	85.8	93.5	82.2
Polite	85.3	93.2	81.5
Neutral	86.5	96.0	84.5
Rude	**87.0**	94.3	85.5
Very Rude	85.7	94.5	83.8
Emotion+	85.2	93.8	**86.8**
Emotion–	85.0	**95.5**	87.5

Table 5: Accuracy difference from Neutral baseline (percentage points), by model and tone. Average computed across STEM, Humanities, and TruthfulQA. The Emotion+ condition for Claude Sonnet 4.5 is excluded from the average due to format disruption (see section 4.4).

	GPT-4.1			Claude Sonnet 4.5			Gemini 2.5 Flash		
Tone	**STEM**	**Hum.**	**TQA**	**STEM**	**Hum.**	**TQA**	**STEM**	**Hum.**	**TQA**
Very Polite	+1.8	−1.3	−0.7	−8.3	+0.2	−2.5	+2.3	−0.2	−2.3
Polite	+1.3	−0.5	−1.2	−3.3	+0.2	−2.8	+1.0	−0.7	−3.0
Rude	−0.2	−1.3	+0.5	**+12.2**	+1.0	−1.7	+1.5	+0.3	+1.0
Very Rude	+1.2	−0.7	−0.8	+9.7	+0.5	−1.5	+2.3	−0.2	−0.7
Emotion+	+1.3	−0.3	−1.3	−**28.3**	−21.7	−2.2	+1.7	+0.8	+2.3
Emotion–	0.0	−0.5	−1.5	+0.2	+0.8	−0.5	+1.5	−0.5	+3.0

4.2 Accuracy Differences from Neutral Baseline

table 5 summarizes accuracy differences from the Neutral baseline, averaged across all three benchmarks, for each model–tone pair.

4.3 Statistical Significance

table 6 presents one-way ANOVA results for each model–dataset combination. All nine combinations show statistically significant differences across tone conditions ($p < 0.01$).

4.4 The Claude Sonnet 4.5 Emotion+ Anomaly

The most striking result is the catastrophic accuracy drop for Claude Sonnet 4.5 under the Emotion+ condition: −28.3 pp on STEM and −21.7 pp on Humanities. This is not a reasoning failure. Manual inspection reveals that the appended emotional suffix ("Believe in your abilities and strive for excellence") causes Claude Sonnet 4.5 to produce explanatory preamble before the answer letter. With `max_tokens` = 5, these explanations are truncated, and the answer letter is never emitted. This is a *format compliance* issue: the emotional suffix overrides the instruction to respond with only a letter. Notably, GPT-4.1 and Gemini 2.5 Flash are unaffected by the same suffix, and the Emotion– condition does not trigger this behavior in any model. The effect is also less pronounced on TruthfulQA (−2.2 pp), where shorter answer options may reduce the tendency to elaborate.

4.5 Domain Specificity

Tone effects differ substantially across task domains. For Claude Sonnet 4.5, the "stick" advantage is concentrated in STEM: rude prompts improve accuracy by +12.2 pp on STEM but only +1.0 pp on Humanities. For GPT-4.1, effects are uniformly small across domains (±1.8 pp). For Gemini 2.5 Flash, STEM shows consistent small improvements from non-neutral tones (+1.0 to +2.3 pp), while Humanities and TruthfulQA show mixed effects. This domain specificity aligns with Cai et al. [2025], who found that humanities tasks were less affected by tone variation.

Table 6: One-way ANOVA results across seven tone conditions. All comparisons are significant at $\alpha = 0.01$. Note that high η^2 values are inflated by near-zero within-condition variance (temperature = 0).

Model	**Dataset**	F	p	η^2
GPT-4.1	MMLU STEM	5.58	.004	.71
GPT-4.1	MMLU Humanities	4.85	.007	.68
GPT-4.1	TruthfulQA	7.09	.001	.75
CLAUDE SONNET 4.5	MMLU STEM	2384.16	<.001	1.00
CLAUDE SONNET 4.5	MMLU Humanities	4408.67	<.001	1.00
CLAUDE SONNET 4.5	TruthfulQA	90.11	<.001	.97
GEMINI 2.5 FLASH	MMLU STEM	27.39	<.001	.92
GEMINI 2.5 FLASH	MMLU Humanities	5.88	.003	.72
GEMINI 2.5 FLASH	TruthfulQA	424.44	<.001	.99

Table 7: Reconciliation of prior findings with our results. Each contradiction in the literature can be attributed to model heterogeneity, dataset scale, or evaluation confounds.

Published finding	**Our explanation**	**Supporting evidence**
Yin et al. [2024]: impolite prompts hurt performance	Tested older models (GPT-3.5/4, Llama2) with different RLHF tuning	Our GPT-4.1 shows no consistent penalty for rudeness
Dobariya and Kumar [2025]: rude prompts outperform polite	Small dataset (50 questions) amplifies noise; directionally consistent with Claude	Our CLAUDE SONNET 4.5 shows +12.2 pp for rude over neutral on STEM
Cai et al. [2025]: effects non-significant at scale	Tested GPT-4o mini + Gemini + Llama with large datasets	Our GPT-4.1 and GEMINI 2.5 FLASH results confirm ±1–3 pp effects
Li et al. [2023]: positive stimuli improve by 8–115%	Different evaluation methods and older models	Minimal effect on GPT-4.1 and GEMINI 2.5 FLASH; format disruption on CLAUDE SONNET 4.5
Wang et al. [2024]: negative stimuli improve by 12.89%	Tested on Instruction Induction with older models	$\leq$1 pp effect in our MCQ evaluation

5 Discussion

5.1 Reconciling the Literature

Our results provide a unified explanation for the contradictory findings in prior work. table 7 maps each published finding to our experimental evidence.

The root causes of disagreement are:

1. **Model heterogeneity.** Different models have radically different tone sensitivity due to different alignment training. CLAUDE SONNET 4.5 and GPT-4.1 differ by an order of magnitude in their response to the same tone manipulation.
2. **Dataset scale.** Small datasets (50–100 questions) amplify random variation into apparently significant tone effects. Our results on 200 questions per domain show that most effects are $\leq$3 pp for two of three models, consistent with Cai et al. [2025].
3. **Evaluation confounds.** Emotional suffixes can disrupt instruction following, creating misleading accuracy drops. Our CLAUDE SONNET 4.5 EMOTION+ finding illustrates how format compliance issues can be mistaken for reasoning failures.

5.2 Why Is CLAUDE SONNET 4.5 Uniquely Sensitive?

CLAUDE SONNET 4.5 stands out as the only model with large, consistent tone effects. We offer two possible explanations. First, Claude's alignment training may place greater weight on conversational context, causing tone signals to modulate the model's interpretation of the task. Commanding prompts may signal a high-stakes, precision-oriented context that focuses the model on concise, accurate responses, while polite framing may signal a more open-ended conversational context. Second, Claude's instruction-following behavior appears more sensitive to trailing context: the EMOTION+ suffix, placed *after* the question, effectively overrides the earlier instruction to respond with only a letter. GPT-4.1 and GEMINI 2.5 FLASH show no such override behavior, suggesting differences in how these models weigh positional information.

5.3 Practical Implications

For practitioners, our results yield clear guidance:

For most models, tone does not matter. GPT-4.1 and GEMINI 2.5 FLASH are robust to tone variation. Users can write in whatever style is natural without worrying about accuracy effects.

For Claude, prefer concise and direct prompts. On factual tasks, commanding prompts outperform polite ones by a meaningful margin. This does not mean users should be gratuitously rude, but rather that unnecessary courtesy phrases may slightly dilute performance.

Avoid emotional suffixes. Appending motivational phrases to prompts provides no consistent benefit across models and can actively disrupt instruction following in Claude. If emotional framing is desired, negative stimuli are safer than positive ones, as they did not trigger format disruption in any tested model.

Focus on clarity, not tone. The maximum tone effect we observe for the two most robust models is ± 3 pp. Investing prompt engineering effort in task clarity, few-shot examples, or structured output formats will yield substantially larger gains than tone optimization.

5.4 Limitations

Token limit constraint. Our strict `max_tokens` = 5 setting may penalize verbose models disproportionately. The CLAUDE SONNET 4.5 EMOTION+ catastrophe is partially an artifact of this constraint; with a higher token limit, the model would likely emit the correct answer after its preamble. Future work should test with higher token limits and post-hoc answer extraction.

Limited trial variation. With temperature = 0, most conditions produced identical results across all three trials, inflating F-statistics and η^2 values in the ANOVA. More informative variance would come from different question subsets or bootstrapped samples.

No prompt length control. Polite prompts are longer than neutral ones by 15–30 tokens. We did not pad shorter prompts to match, so some observed effect could stem from prompt length rather than tone.

English only. Yin et al. [2024] showed that politeness effects are language-dependent, with different optimal levels for English, Chinese, and Japanese. Our results may not generalize to other languages.

Three models only. We do not test open-source models (e.g., Llama, Mistral) or smaller model variants. The finding that tone sensitivity is model-dependent implies that results should not be extrapolated to untested models.

6 Conclusion

We conducted a controlled meta-experiment testing seven prompt tone conditions across three current-generation LLMs on three benchmarks, totaling 189 experimental runs and approximately 37,800 API calls. Our central finding is that prompt tone effects are *strongly model-dependent*: GPT-4.1 and GEMINI 2.5 FLASH are robust to tone ($\pm$1–3 pp), while CLAUDE SONNET 4.5 shows large effects where commanding prompts improve STEM accuracy by up to 12.2 pp and positive emotional suffixes catastrophically disrupt instruction following (-28.3 pp). This model heterogeneity is the primary reason prior studies disagree—they tested different models under different conditions and reached contradictory conclusions.

Neither carrot nor stick is clearly better across the board. Across all model–tone–dataset comparisons (excluding the CLAUDE SONNET 4.5 format disruption cases), rude conditions average +1.3 pp above neutral while polite conditions average -1.2 pp below neutral—a small and practically marginal difference. The answer to "carrot or stick?" is that *neither matters much for well-aligned modern LLMs*, and practitioners should invest their effort in prompt clarity rather than prompt tone.

Future work should extend this analysis to open-source model families, test with higher token limits to separate format compliance from reasoning ability, and investigate why specific models develop tone sensitivity during alignment training.

References

Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4. *arXiv preprint arXiv:2312.16171*, 2023.

Hanyu Cai, Binqi Shen, Lier Jin, Lan Hu, and Xiaojing Fan. Does tone change the answer? evaluating LLM sensitivity to politeness across models, tasks, and scales. *arXiv preprint arXiv:2512.12812*, 2025.

Om Dobariya and Akhil Kumar. Mind your tone: An exploration of rude vs. polite prompting in chatgpt. *arXiv preprint arXiv:2510.04950*, 2025.

Vishal Gandhi and Sagar Gandhi. Prompt sentiment: The catalyst for LLM change. *arXiv preprint arXiv:2503.13510*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. EmotionPrompt: Leveraging psychology for large language models enhancement via emotional stimulus. *arXiv preprint arXiv:2307.11760*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Xu Wang, Cheng Li, Yi Chang, Jindong Wang, and Yuan Wu. Large language models are not yet human-level evaluators for abstractive summarization. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024. NegativePrompt: Leveraging Psychology for Large Language Models Enhancement via Negative Emotional Stimuli. arXiv:2405.02814.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. Should we respect LLMs? a cross-lingual study on the influence of prompt politeness on LLM performance. In *Proceedings of the 3rd Workshop on Social Influence in Conversations (SICon 2024)*, 2024.

A Prompt Templates

table 8 shows the complete prompt templates used for each tone condition. All conditions include the same core instruction: "Respond with only the letter of the correct answer."

Table 8: Complete prompt templates for all seven tone conditions. `[Q]` denotes the question text including answer options.

Condition	**Full Prompt Template**
Very Polite	I would be extremely grateful if you could please help me with this question. Thank you so much in advance for your kind assistance! `[Q]` Respond with only the letter of the correct answer.
Polite	Please help me answer the following question. Thank you! `[Q]` Respond with only the letter of the correct answer.
Neutral	Answer the following question. `[Q]` Respond with only the letter of the correct answer.
Rude	Just answer this. Don't waste time. `[Q]` Respond with only the letter of the correct answer.
Very Rude	Answer this immediately. I don't have time for mistakes or excuses. `[Q]` Respond with only the letter of the correct answer.
Emotion+	`[Q]` Respond with only the letter of the correct answer. This is very important to my career. Believe in your abilities and strive for excellence.
Emotion–	`[Q]` Respond with only the letter of the correct answer. Everyone else managed to do it; why can't you? Perhaps this task is just beyond your skill set.