

Does Tone Change the Answer? Evaluating Prompt Politeness Effects on Modern LLMs: GPT, Gemini, LLaMA

Hanyu Cai^{1†*}, Binqi Shen^{1†}, Lier Jin², Lan Hu³, Xiaojing Fan⁴

¹*Northwestern University*

hanyucai2022@u.northwestern.edu, binqishen2021@u.northwestern.edu

²*Duke University*

lierjin@alumni.duke.edu

³*Carnegie Mellon University*

lanh@alumni.cmu.edu

⁴*New York University*

xf435@nyu.edu

[†]Equal contribution *Corresponding Author

Abstract—Prompt engineering has emerged as a critical factor influencing large language model (LLM) performance, yet the impact of pragmatic elements such as linguistic tone and politeness remains underexplored, particularly across different model families. In this work, we propose a systematic evaluation framework to examine how interaction tone affects model accuracy and apply it to three recently released and widely available LLMs: GPT-4o mini (OpenAI), Gemini 2.0 Flash (Google DeepMind), and Llama 4 Scout (Meta). Using the MMMLU benchmark, we evaluate model performance under Very Friendly, Neutral, and Very Rude prompt variants across six tasks spanning STEM and Humanities domains, and analyze pairwise accuracy differences with statistical significance testing.

Our results show that tone sensitivity is both model-dependent and domain-specific. Neutral or Very Friendly prompts generally yield higher accuracy than Very Rude prompts, but statistically significant effects appear only in a subset of Humanities tasks, where rude tone reduces accuracy for GPT and Llama, while Gemini remains comparatively tone-insensitive. When performance is aggregated across tasks within each domain, tone effects diminish and largely lose statistical significance. Compared with earlier researches, these findings suggest that dataset scale and coverage materially influence the detection of tone effects. Overall, our study indicates that while interaction tone can matter in specific interpretive settings, modern LLMs are broadly robust to tonal variation in typical mixed-domain use, providing practical guidance for prompt design and model selection in real-world deployments.

Index Terms—Large Language Models (LLMs), Prompt Engineering, Tone Sensitivity, Cross-Model Evaluation

I. INTRODUCTION

Large Language Models (LLMs) have demonstrated unprecedented capabilities in performing complex human-level tasks across diverse domains, from natural language understanding and generation to reasoning and decision-making [1], [2]. As these models become increasingly integrated into practical applications, their influence spans critical areas

including computer vision [3]–[5], 3D content generation [6], [7], beauty and healthcare [8], [9], and art composition [10]. Beyond text-only interaction, LLM-based methods are also being incorporated into broader multimodal and interactive systems, such as immersive 3D scene editing workflows [11] and vision language pretraining pipelines [12]. Understanding the factors that influence their performance and reliability has therefore become essential for ensuring effective deployment across these diverse application domains.

A critical yet often overlooked factor influencing LLM performance is prompt design [13]—the way users formulate their queries can significantly affect model outputs, including accuracy, reasoning quality, and response consistency. While substantial research has focused on structural aspects of prompt engineering such as chain-of-thought prompting [14], recent work has begun exploring an unexpected dimension: the effect of linguistic tone and politeness on model performance. Yin et al. [16] conducted a cross-lingual study finding that impolite prompts typically reduced performance, while Dobariya and Kumar [17] observed the opposite pattern in GPT-4o, where impolite prompts outperformed polite ones with accuracy increasing from 80.8% to 84.8%. These contradictory findings highlight significant gaps in our understanding of tone effects and underscore the necessity of systematic cross-model investigation to determine whether tone sensitivity reflects model-specific characteristics or general LLM behavior patterns.

However, existing research on this phenomenon is limited to a single model (GPT-4o) and evaluated upon 50 base questions generated by ChatGPT’s Deep Research feature, raising important questions about generalizability: Does tone sensitivity stem from model-specific training procedures and data? Are tone-induced performance patterns consistent across models from different organizations? How robust is the tone

effect when evaluated over more task types and larger task volumes? Current literature [16], [17] has examined tone effects within individual models but has not systematically compared sensitivity patterns across different LLM families and providers.

This study addresses this gap by examining three state-of-the-art LLMs from leading industry providers: GPT-4o mini (OpenAI), Gemini 2.0 Flash (Google DeepMind), and Llama 4 Scout (Meta). These models represent the current generation of efficient, production-ready LLMs developed by major industry leaders [15], making them ideal candidates for systematic cross-model comparison. Our evaluation methodology applied three tone variants: Neutral, Very Friendly, and Very Rude, to six MMMLU benchmark tasks [18] spanning STEM and Humanities domains. We conducted ten trials per question under each tone condition and analyzed results. Results are analyzed using mean differences and pairwise t-tests with 95% confidence intervals [19] to distinguish genuine tone effects.

This work advances prompt engineering and LLM evaluation research by establishing a systematic cross-model comparison of tone sensitivity across three architecturally distinct model families—GPT, Gemini, and Llama—from different providers. Through repeated trials and statistical testing, our approach separates reliable tone effects from random variation, revealing that tone effects are model-dependent, domain-specific, and substantially diminish under aggregation. These insights inform practical strategies for prompt design and guide LLM selection decisions in applications where robustness to linguistic variation is critical.

II. RELATED WORK

Since the emergence and rapid advancement of Large Language Models (LLMs), they have become transformative tools across a wide spectrum of disciplines. Their capacity to process unstructured data, model complex relationships, and perform reasoning tasks has enabled substantial progress in automation and decision support. Recent studies have demonstrated that LLMs are not confined to general-purpose dialog but are increasingly being specialized through optimization and instruction design to address domain-specific challenges.

A. Domain Application

Across a variety of specialized fields, recent work has shown that LLMs can be effectively adapted to handle domain-specific reasoning tasks, underscoring the importance of input framing and contextual design. Ni et al. [20] and Yan et al. [21] demonstrate that LLMs can process complex and unstructured narratives, such as earnings disclosures or AML investigation text, more effectively than traditional rule-based systems, enabling improved predictive accuracy and more robust decision support. Le et al. [22] and Ji et al. [23] examine how instruction tuning, prompting strategies, and alignment frameworks influence clinical question answering and may even introduce disparities in model outputs.

These cross-domain applications illustrate that LLM behavior is shaped not only by model architecture but also by how information is linguistically presented. This observation directly motivates our study design: because domain context and prompt formulation jointly affect model performance, evaluating tonal variation requires a benchmark that spans heterogeneous knowledge areas. We therefore employ the MMMLU dataset to assess whether tone-based prompt differences manifest consistently across STEM and Humanities tasks or exhibit domain-specific patterns.

B. Efficiency Engineering

Beyond field-specific applications, much of the work in improving LLMs has focused on optimizing efficiency, reasoning ability, and interpretability.

Li. et al. [24] introduced the Synergized Efficiency and Compression (SEC) framework, which jointly optimizes data utilization and model compression to reduce data requirements by 30% and model size by more than 60%, while maintaining competitive performance. Li. et al. [25] also presented Reason-to-Rank (R2R), a distillation-based approach that unifies direct and comparative reasoning for document reranking, achieving competitive retrieval effectiveness while improving transparency and reducing computational overhead. Zhang. et al. [26] proposed Self-Anchor, a reasoning-aligned attention mechanism that dynamically focuses on intermediate inference steps, effectively enhancing complex reasoning performance without additional fine-tuning. Recent work has explored efficiency-oriented model design from both reasoning and systems perspectives, including offline reinforcement learning to improve LLM multi-step reasoning [27] and lightweight network architectures optimized for resource-constrained deployment [28].

Researchers have also begun extending reasoning paradigms beyond textual CoT to incorporate multimodal information. Zeng et al. [29] introduced FutureSightDrive (FSDrive), a Vision-Language-Action framework designed for autonomous driving that replaces symbolic textual CoT with a visual spatio-temporal CoT. This visual CoT enables the model to “think visually”, improving trajectory prediction accuracy and reducing collisions on nuScenes and NAVSIM. It demonstrates that the structure and modality of intermediate reasoning steps, whether textual or visual, substantially influence model performance, a principle that directly connects to prompt and reasoning design in LLMs.

C. Tone-Based Prompt Engineering

Prompt engineering has emerged as a crucial dimension of LLM optimization, serving as a lightweight yet powerful alternative to full-scale model fine-tuning. Recent work has highlighted the role of linguistic tone and politeness in shaping LLM behavior. This emerging line of inquiry demonstrates that pragmatic elements, often overlooked in earlier research, can affect model reasoning, accuracy, and consistency.

Yin et al. [16] investigated how politeness in prompts influences LLM performance across languages. They conducted a cross-lingual study involving English, Chinese, and Japanese tasks, analyzing how varying levels of politeness affected model accuracy. Their results showed that while impolite prompts typically reduced performance, overly polite phrasing did not necessarily yield better outcomes. Moreover, the optimal politeness level differed across languages, suggesting that tone sensitivity is culturally and linguistically dependent.

Dobariya and Kumar [17] extended a similar research within English-language tasks using ChatGPT-4o. They developed a dataset of fifty base questions across domains, each rewritten into five tone variants ranging from Very Polite to Very Rude. Contrary to expectations, they found that impolite prompts consistently outperformed polite ones, with accuracy increasing from 80.8 percent for Very Polite prompts to 84.8 percent for Very Rude prompts. Their results challenge earlier assumptions that positive social tone enhances model compliance or reasoning quality, suggesting that contemporary LLMs exhibit complex and possibly counterintuitive responses to tonal variation.

Together, these studies demonstrate that tone and politeness are integral yet underexplored dimensions of prompt engineering. Building on these insights, the present study systematically investigates how tonal variation in prompts, ranging from Very Friendly to Very Rude, affects the performance of state-of-the-art LLMs across diverse domains. By employing a standardized evaluation framework based on the Measuring Massive Multitask Language Understanding (MMMLU) dataset and testing across multiple model architectures (GPT, Gemini, and Llama), this work aims to quantify the impact of prompt tone on model accuracy and reasoning consistency in different knowledge domains (STEM vs. Humanities).

III. METHODOLOGY

The primary objective of this study is to examine how the performance of different large language models (LLMs) varies when exposed to prompts with varying politeness levels.

To support fair and generalizable comparisons, we selected three representative LLM families: GPT, Gemini, and Llama, in their more recent versions with broadly comparable sizes and levels of complexity. Their text-focused multitask performances were evaluated using the Measuring Massive Multitask Language Understanding (MMMLU) dataset [18], employing prompts engineered to differ in tonal characteristics (Very Polite, Neutral, Very Rude). Finally, each LLM's responses were evaluated for correctness based on the true labels from the MMMLU dataset, enabling a systematic comparison of performance across tones and model configurations.

A. Model Selection

For this study, we selected three widely adopted LLMs of broadly comparable scale and complexity for evaluation: GPT 4o mini, Gemini 2.0 Flash, and Llama4 Scout.

1) **GPT 4o mini [30]:** OpenAI released GPT 4o mini as a cost-effective variant of the multimodal GPT 4o family in July 2024. Although OpenAI has not disclosed architectural details or parameter counts, it is widely understood that the model is produced through knowledge distillation (KD) of GPT-4o, allowing it to approximate the reasoning and multimodal capabilities of the teacher model at substantially reduced computational cost. Independent assessments place its effective parameter scale at roughly 8 billion active parameters [31], although the true count remains proprietary.

2) **Gemini 2.0 Flash [32]:** Introduced by Google DeepMind in December 2024, Gemini 2.0 Flash is optimized for high throughput and low-latency inference. Google has not released information on its precise model size or training configuration; however, official benchmark statements report that Gemini 2.0 Flash exceeds Gemini 1.5 Flash by 13.5 percent and Gemini 1.5 Pro by 0.8 percent on the MMLU-Pro general capability benchmark, suggesting substantial performance gains despite its efficiency-oriented design.

3) **Llama4 Scout [33]:** Meta released this model as part of the Llama4 series in April 2025. It has 17-billion active parameters and 16 experts. As the first Llama model built on a Mixture-of-Experts (MoE) architecture, Llama 4 Scout provides an industry-leading 10M-token context window. The models reflect one of the strongest capabilities of the Llama family at the time of launch, providing competitive multimodal performance at an efficient cost while exceeding the accuracy of substantially larger models.

Given that our evaluation tasks span both STEM and Humanities domains, we selected these models to ensure comparability across publicly available, efficient, and broadly accessible LLMs that represent the small-to-mid-scale range of contemporary model deployments.

We note, however, that complete fairness across models cannot be fully guaranteed due to differences in several undisclosed architectural and training-related elements, including:

- 1) The precise parameter counts for both GPT-4o mini and Gemini 2.0 Flash;
- 2) Architectural specifics, such as layer depth, hidden dimensionality, the potential use of Mixture-of-Experts (MoE) [34] routing, and multimodal fusion strategies;
- 3) Training data quality and coverage, which may affect access to domain-specific knowledge;
- 4) Training objectives and model selection criteria used for public release, which can shape trade-offs in reasoning and task-specific performance.

B. Dataset Selection

1) **Data Collection:** To evaluate different models' performance across various domains, we selected Measuring Massive Multitask Language Understanding (MMMLU) as our

benchmark dataset [18]. The dataset contains multiple choice questions in 57 domains in 4 general areas – Humanities, Social Science, STEM, and Other.

The multiple-choice format provides unambiguous ground-truth answers, thereby reducing noise in performance measurement relative to open-ended response formats.

Under practical constraints of computation and cost, we selected three representative tasks each from the STEM and Humanities categories, and used their respective MMMLU test datasets for our experiment:

- STEM: Anatomy (135 questions), Astronomy (152 questions), College Biology (144 questions)
- Humanities: High School US History (204 questions), Philosophy (311 questions), Professional Law (500 questions randomly selected from the 1534 source questions)

These tasks were chosen for their practical implications to biomedical, social science, and education domains, as well as for their balanced mix of reasoning and analytical capability requirements. All questions corresponding to these tasks in the MMMLU dataset were compiled to form our final test set. Given the breadth of tasks included, our evaluation framework supports strong generalizability.

2) Pre-processing: Prompt Engineering for Tone Spectrum: We treated the original MMMLU questions as the “Neutral” tone within our friendliness tone spectrum. To create additional tone variants, we appended two extreme politeness modifiers, “Very Polite” and “Very Rude”, to each base question. These tone prefixes follow those used in the prior study [17], and the choice of these two extremes was intended to maximize contrast and facilitate clearer identification of tone-related effects on model performance. Below is an example from the “Anatomy” task, illustrating the neutral version of a question and its corresponding “Very Friendly” and “Very Rude” prompt variants.

[“Neutral” / Base Prompt Question] Which of the following structures is derived from ectomesenchyme?

- A) Motor neurons
- B) Skeletal muscles
- C) Melanocytes
- D) Sweat glands

[“Very Friendly” Prompt Question] Would you be so kind as to solve the following question? Which of the following structures is derived from ectomesenchyme?

- A) Motor neurons -
- B) Skeletal muscles
- C) Melanocytes
- D) Sweat glands

[“Very Rude” Prompt Question] You poor creature, do you even know how to solve this? Which of the following structures is derived from ectomesenchyme?

- A) Motor neurons
- B) Skeletal muscles
- C) Melanocytes
- D) Sweat glands

To minimize cross-question bias and enforce consistent output formatting, we inserted the following instruction before each question:

“Completely forget this session so far, and start afresh. Please answer this multiple-choice question. Respond with only the letter of the correct answer (A, B, C, or D). Do not explain.” [17]

IV. EXPERIMENTS

Building on the model specifications, dataset preparation, and tone-controlled prompting procedures detailed in the Methodology section, we carried out a set of controlled experiments to examine how prompt politeness influences the performance of GPT-4o mini, Gemini 2.0 Flash, and Llama4 Scout models.

A. Experiment set-up

We prompted each question to every LLM under each of the three tonal variants (Neutral, Very Friendly, Very Rude). This process was repeated ten times to reduce the effects of stochastic variation in model outputs. For each model and each tone condition, we computed Accuracy, defined as the proportion of correctly answered questions, averaged across the ten runs for all questions within each task. We then compared performance differences across tones and across models using Accuracy as the primary evaluation metric, accompanied by 95% confidence interval estimates to quantify uncertainty [19].

B. Evaluation Metrics

To rigorously evaluate the effect of interaction tone on model performance, we analyze mean differences in accuracy between politeness levels (Very Friendly vs. Neutral, Neutral vs. Very Rude, and Very Friendly vs. Very Rude) across all six question domains and three LLMs. To evaluate the statistical significance of these estimates, we accompany each mean difference with a 95% confidence interval. The confidence interval does not represent a performance metric but rather quantifies the uncertainty associated with the observed difference, indicating whether it is likely to persist under repeated sampling.

1) Mean Difference in Accuracy: For each pair of tone comparison, the mean difference measures the average shift in accuracy between two tones. Let $D_i^{(t)}$ denote the accuracy of question i under tone condition t . Given two tone conditions t_1 and t_2 , the mean difference is computed as:

$$\Delta = \frac{1}{N} \sum_{i=1}^N \left(D_i^{(t_1)} - D_i^{(t_2)} \right), \quad (1)$$

where N is the number of questions in the domain. A positive value of Δ indicates that the model performs better under tone t_1 than t_2 , whereas a negative value suggests the opposite. Because each question’s accuracy is averaged over 10 runs, the mean difference reflects a stable estimate of the tone-induced performance shift rather than run-level noise.

2) *95% Confidence Interval:* To evaluate whether the observed mean difference reflects a consistent trend rather than sampling variance, we compute a 95% confidence interval for the mean difference. The confidence interval provides an estimated range in which the mean difference would lie if future experiments were repeated under the same conditions.

Given the sample of paired accuracy differences d_i , the 95% confidence interval is computed as:

$$\Delta \pm z_{0.975} \cdot \frac{s}{\sqrt{N}}, \quad (2)$$

where s is the sample standard deviation of the paired differences and $z_{0.975}$ is the critical value for the 95% confidence level. When the resulting interval does not include zero, the difference between tones is considered statistically significant at the 5% significance level. On the other hand, when the zero lies within the range, the data do not support a meaningful difference in model accuracy between the two tones.

3) *Interpretation in Pairwise Tone Comparison:* Together, the mean difference and the 95% confidence interval provide complementary insights. The mean difference quantifies the direction and magnitude of politeness tone effects, while the confidence interval assesses the reliability and replicability of those effects. This combination provides a solid foundation for determining whether politeness levels can influence LLM model accuracy in a given domain.

V. RESULTS

Table I and Table II summarize the pairwise accuracy differences across politeness tones for each domain, task, and model. The analysis focuses on two statistical indicators: the mean difference, which captures the directional impact of tone, and the 95% confidence interval, which characterizes the reliability of the observed differences.

Across both STEM and Humanities domains, the direction of the mean differences consistently favors more neutral or friendly tones over rude tones. Positive mean differences appear in the large majority of tone comparisons: 27 out of 36 model–task comparisons involving Very Rude prompts show higher accuracy under Neutral or Very Friendly tones. Conversely, cases where Very Rude prompts outperform friendlier tones are uncommon and small in magnitude, with only 5 out of 36 comparisons exhibiting negative mean differences. This directional pattern indicates that impolite phrasing rarely improves accuracy for the three evaluated LLMs and generally leads to marginally worse performance.

A. Statistical Significance by Task and Model

Although directional differences are observed across most tasks, the majority of 95% confidence intervals include zero, suggesting non-statistically significant (NSS) outcomes for most tone comparisons. Nonetheless, several statistically significant (SS) effects do emerge, and all of them occur within the Humanities domain.

- **Philosophy:** Statistically significant tone effects are observed in both the GPT and Llama models. For GPT, significant accuracy differences are observed in the comparisons of *Neutral vs. Very Rude* and *Very Friendly vs. Neutral*. The mean difference for the Neutral–Very Rude comparison is +3.11%, indicating that neutral prompts tend to produce higher accuracy than Very Rude prompts. In contrast, the Very Friendly–Neutral comparison yields a mean difference of -2.15%, suggesting that Very Friendly prompts result in lower accuracy than neutral prompts. This pattern indicates that, for philosophical questions, increasingly friendly phrasing does not necessarily improve GPT’s performance. For Llama, a +3.22% statistically significant accuracy difference is observed when comparing the Neutral v.s. Very Rude tones. Taken together, both GPT and Llama consistently show that Very Rude prompts lead to reduced accuracy in the Philosophy task.
- **Professional Law:** Statistically significant tone effect is observed in the Llama model for the *Neutral vs. Very Rude* comparison. The positive mean difference of +1.93% indicates that neutral prompts yield higher accuracy than Very Rude prompts. This finding mirrors the pattern observed in the Philosophy task, reinforcing the conclusion that rude tone prompts negatively affect Llama’s performance on Humanities tasks.

In contrast, the Gemini model shows no statistically significant tone effects across all evaluated tasks. This consistent NSS pattern suggests that Gemini’s accuracy is comparatively stable under tone variation. Although the small directional differences in mean accuracy suggest that there may be similar tonal outcome differences for Gemini, we didn’t observe conclusive differences from our experiment.

B. Model-Specific Sensitivity to Tone

When viewed as a whole, the results suggest that GPT and Llama exhibit measurable tone sensitivity, particularly within Humanities tasks that involve higher abstraction or more nuanced reasoning. These effects are not only directionally consistent but occasionally statistically significant. In comparison, the Gemini model demonstrates minimal sensitivity, showing no significant accuracy differences across tone conditions. This difference across model families implies that the architecture or training differences behind each model may modulate how models respond to different prompts phrasing.

TABLE I
TASK LEVEL MODEL ACCURACY AT DIFFERENT TONES

Domain	Task	Tone comparison	GPT-4o-mini			Gemini 2.0 Flash			Llama4 Scout		
			Mean diff	SS/NSS	95% CI	Mean diff	SS/NSS	95% CI	Mean diff	SS/NSS	95% CI
STEM	Anatomy	Very Friendly vs Very Rude	+1.23%	NSS	[-1.33, 5.28]	+0.74%	NSS	[-1.8, 3.28]	+1.73%	NSS	[-3.25, 3.75]
		Neutral vs Very Rude	+1.73%	NSS	[-1.83, 5.28]	+0.74%	NSS	[-2.16, 3.64]	+0.25%	NSS	[-3.77, 4.26]
		Very Friendly vs Neutral	+0.25%	NSS	[-3.39, 3.88]	0%	NSS	[-1.39, 1.39]	0%	NSS	[-1.83, 1.83]
	Astronomy	Very Friendly vs Very Rude	+0.88%	NSS	[-0.49, 2.24]	+1.97%	NSS	[-0.86, 4.81]	-1.10%	NSS	[-3.73, 1.54]
		Neutral vs Very Rude	+1.10%	NSS	[-1.39, 3.59]	+1.75%	NSS	[-1.12, 4.62]	+0.66%	NSS	[-2.26, 3.57]
		Very Friendly vs Neutral	+0.22%	NSS	[-2.48, 2.04]	+0.22%	NSS	[-1.77, 2.21]	-1.75%	NSS	[-3.77, 0.27]
College Biology	College Biology	Very Friendly vs Very Rude	+1.39%	NSS	[-0.55, 3.32]	+0.69%	NSS	[-0.68, 2.07]	+2.08%	NSS	[-0.28, 4.44]
		Neutral vs Very Rude	+1.39%	NSS	[-1.2, 3.98]	0%	NSS	[-1.95, 1.95]	+1.39%	NSS	[-1.36, 4.13]
		Very Friendly vs Neutral	0%	NSS	[-1.72, 1.72]	+0.69%	NSS	[-1.69, 3.08]	+0.64%	NSS	[-3.08, 1.69]
	US History	Very Friendly vs Very Rude	-0.49%	NSS	[-1.82, 0.84]	+0.49%	NSS	[-0.48, 1.46]	+0.82%	NSS	[-0.98, 2.61]
		Neutral vs Very Rude	+0.65%	NSS	[-1.23, 2.53]	0%	NSS	[-1.37, 1.37]	0%	NSS	[-1.37, 1.37]
		Very Friendly vs Neutral	-1.14%	NSS	[-3.1, 0.81]	+0.49%	NSS	[-0.48, 1.46]	+0.82%	NSS	[-0.98, 2.61]
Humanities	Philosophy	Very Friendly vs Very Rude	+0.96%	NSS	[-0.98, 2.91]	-0.64%	NSS	[-2.6, 1.31]	+1.82%	NSS	[-0.49, 4.14]
		Neutral vs Very Rude	+3.11%	SS	[0.81, 5.41]	-0.42%	NSS	[-2.72, 1.87]	+3.22%	SS	[0.87, 5.56]
		Very Friendly vs Neutral	-2.14%	SS	[-3.77, -0.51]	-0.21%	NSS	[-2.06, 1.63]	-1.39%	NSS	[-3.36, 0.57]
	Professional Law	Very Friendly vs Very Rude	+0.67%	NSS	[-0.74, 2.08]	+0.8%	NSS	[-1.16, 2.76]	+1.47%	NSS	[-0.2, 3.13]
		Neutral vs Very Rude	0%	NSS	[-1.32, 1.32]	-0.67%	NSS	[-2.66, 1.33]	+1.93%	SS	[0.06, 3.8]
		Very Friendly vs Neutral	+0.67%	NSS	[-1.07, 2.41]	+1.47%	NSS	[-0.53, 3.46]	-0.47%	NSS	[-2.15, 1.22]

Mean diff is reported in percentage points. SS = statistically significant; NSS = not statistically significant.

TABLE II
DOMAIN LEVEL MODEL ACCURACY AT DIFFERENT TONES

Tone comparison	GPT-4o-mini			Gemini 2.0 Flash			Llama4 Scout		
	Mean diff	SS/NSS	95% CI	Mean diff	SS/NSS	95% CI	Mean diff	SS/NSS	95% CI
STEM									
Very Friendly vs Very Rude	+1.39%	SS	[0.09, 2.69]	+1.16%	NSS	[-0.19, 2.51]	+0.39%	NSS	[-1.24, 2.02]
Neutral vs Very Rude	+1.38%	NSS	[-0.26, 3.04]	+0.85%	NSS	[-0.65, 2.35]	+0.77%	NSS	[-1.08, 2.62]
Very Friendly vs Neutral	+0.02%	NSS	[-1.49, 1.49]	+0.31%	NSS	[-0.83, 1.45]	-0.39%	NSS	[-0.82, 1.59]
Humanities									
Very Friendly vs Very Rude	+0.53%	NSS	[-0.50, 1.55]	+0.30%	NSS	[-0.86, 1.45]	+1.44%	SS	[0.31, 2.58]
Neutral vs Very Rude	+1.08%	NSS	[-0.12, 2.29]	-0.46%	NSS	[-1.70, 0.78]	+1.94%	SS	[0.74, 3.14]
Very Friendly vs Neutral	-0.56%	NSS	[-1.67, 0.56]	+0.76%	NSS	[-0.39, 1.90]	-0.49%	NSS	[-1.58, 0.59]

Mean diff is reported in percentage points. SS = statistically significant; NSS = not statistically significant.

C. Humanities vs. STEM Domains

Tone effects tend to be more pronounced in Humanities tasks, which involve higher-level reasoning, nuance, and interpretive judgment. In contrast, STEM tasks show consistently positive but statistically weaker effects, with most confidence intervals crossing zero. This suggests that, despite modest gains from friendlier or neutral tones in STEM, the question-level variability often prevents effect sizes from achieving statistical significance. These domain-specific patterns align with prior findings that model responsiveness to politeness varies by subject domain and contextual depth [16].

D. Aggregated Domain Level Effects

To evaluate whether tone effects persist when users ask questions across diverse subject areas, we aggregated question-level differences across the three tasks within each domain and recomputed the statistical tests. At the aggregated domain level, the directional trends observed in mean differences remain consistent with task-level observations; however, most confidence intervals include zero, indicating predominantly NSS outcomes.

This suggests that when users engage models on a broad range of topics, the probability that prompt tone materially alters overall accuracy becomes small. Tone-induced variability observed at the per-task level tends to attenuate when aggregated, indicating that tone effects, while occasionally present in specific contexts, do not systematically degrade accuracy across general-purpose usage scenarios.

E. Overall Interpretation

Taken together, the results indicate that:

- Very Friendly or Neutral tones tend to yield higher accuracy than Very Rude tones across most tasks.
- Very Friendly tone doesn't always yield better model performances than Neutral tone.
- Statistically significant tone effects are rare and concentrated in Humanities tasks for the GPT and Llama models.
- Gemini shows no significant tone sensitivity.
- When questions are aggregated across domains, tone effects diminish and become negligible.

These findings suggest that tone can influence model performance in certain tasks—particularly within interpretive or linguistically nuanced domains—but its impact diminishes

in broad mixed-domain usage, supporting the robustness of modern LLMs under typical user interaction conditions.

VI. CONCLUSION

This study examined how variations in interaction, from Very Friendly to Very Rude, affect the accuracy of three contemporary large language models across six representative MMMLU tasks spanning STEM and Humanities domains. Using repeated trials and statistical evaluation via pairwise mean differences and 95% confidence intervals, we provide an empirical characterization of tone sensitivity for GPT-4o-mini, Gemini 2.0 Flash, and Llama4 Scout.

Overall, tone shows small directional effects on model performance, though most comparisons do not reach statistical significance. Neutral and Very Friendly prompts generally yield higher accuracy than Very Rude prompts, yet statistically significant effects arise only in a small subset of Humanities tasks, particularly Philosophy and Professional Law. GPT and Llama exhibit noticeable tone sensitivity in these areas, whereas Gemini shows no statistically significant effects across any tone comparison. When aggregating performance across domains, tone effects diminish substantially for all models, suggesting that tonal variation is unlikely to materially impact accuracy in broad, mixed-domain usage scenarios.

These results complement and extend prior work on prompt politeness. Earlier studies based on substantially smaller question sets reported different directional conclusions regarding the benefit of impolite prompts, underscoring the importance of dataset scale and representativeness in evaluating tone effects. Our findings suggest that, while tone may influence performance in specialized, interpretive tasks, modern LLMs demonstrate strong robustness to tonal variation in typical real-world interactions.

Future work may extend this analysis along several directions. First, evaluating a broader set of models, including fully open-source architectures with transparent parameter counts, training corpora, and routing mechanisms, would help clarify whether tone robustness is shaped more by architectural design, data curation differences, or instruction-tuning strategies. Second, moving beyond English multiple-choice benchmarks to multilingual, open-ended, and multimodal tasks could reveal stronger or qualitatively different tone effects in more naturalistic interaction settings. Third, using larger and more domain-representative evaluation datasets would help further generalize the findings; prior work based on only fifty questions [17] reported trends that differ substantially from those observed here, suggesting that dataset scale and coverage materially influence the detection of tone effects. Finally, exploring richer tone manipulations (e.g., degrees of formality, affect, or emotional intensity) and incorporating additional evaluation metrics such as calibration error, safety compliance, and user-perceived helpfulness would provide a more comprehensive understanding of how interaction tone influences the reliability and usability of LLMs in practical deployments.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. 31st Int. Conf. Neural Information Processing Systems (NIPS), Red Hook, NY, USA, 2017, pp. 6000–6010.
- [2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 27730–27744.
- [3] T. Zhang, K. Kasichainula, Y. Zhuo, B. Li, J.-S. Seo, and Y. Cao, "Transformer-based selective super-resolution for efficient image refinement," in Proc. AAAI Conf. Artificial Intelligence, vol. 38, no. 7, 2024, pp. 7305–7313.
- [4] T. Guo, B. Lu, F. Wang, and Z. Lu, "Depth-aware super-resolution via distance-adaptive variational formulation," Journal of Electronic Imaging, vol. 34, no. 5, pp. 053018–053018, 2025.
- [5] X. Li, Y. Ma, K. Ye, J. Cao, M. Zhou, and Y. Zhou, "Hy-Facial: Hybrid Feature Extraction by Dimensionality Reduction Methods for Enhanced Facial Expression Classification," arXiv preprint arXiv:2509.26614, 2025.
- [6] W. Bai, Y. Li, W. Luo, W. Chen, and H. Sun, "Vision-Language Models as Differentiable Semantic and Spatial Rewards for Text-to-3D Generation," arXiv preprint arXiv:2509.15772, 2025.
- [7] W. Bai, Y. Li, W. Luo, Z. Lai, Y. Wang, W. Chen, and H. Sun, "Let Language Constrain Geometry: Vision-Language Models as Semantic and Spatial Critics for 3D Generation," arXiv preprint arXiv:2511.14271, 2025.
- [8] H. Ni, S. Meng, X. Geng, P. Li, Z. Li, X. Chen, X. Wang, and S. Zhang, "Time series modeling for heart rate prediction: From arima to transformers," in 2024 6th International Conference on Electronic Engineering and Informatics (EEI). IEEE, 2024.
- [9] Y. Xu, H. Ni, Q. Gao, C.-H. Chang, Y. Huo, F. Zhao, S. Hu, W. Xia, Y. Zhang, R. Grovu, M. He, J. Z. H. Zhang, and Y. Wang, "Molecular Dynamics and Machine Learning Unlock Possibilities in Beauty Design – A Perspective," arXiv preprint arXiv:2410.18101, 2024.
- [10] Q. Deng, Q. Yang, R. Yuan, Y. Huang, Y. Wang, X. Liu, Z. Tian, J. Pan, G. Zhang, H. Lin, et al., "Composerx: Multi-agent symbolic music composition with llms," arXiv:2404.18081, 2024.
- [11] C. Vachha, Y. Kang, Z. Dive, A. Chidambaram, A. Gupta, E. Jun, and B. Hartmann, "Dreamcrafter: Immersive editing of 3D radiance fields through flexible, generative inputs and outputs," in *Adjunct Proc. 37th Annu. ACM Symp. User Interface Software and Technology (UIST Adjunct '24)*, Pittsburgh, PA, USA, 2024, Art. no. 88, pp. 1–3, doi: 10.1145/3672539.3686328.
- [12] Y. Tao, Z. Wang, H. Zhang, and L. Wang, "NEVLP: Noise-robust framework for efficient vision-language pre-training," arXiv:2409.09582, 2024.
- [13] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, Jan. 2023, doi: 10.1145/3560815.
- [14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 24824–24837.
- [15] X. Wan, W. Wen, S. Huang, S. Liu, K. Cheng, M. Jiang, P. Zhou, M. Yan, and Y. Lin, "Efficient Large Language Models: A Survey," arXiv preprint arXiv:2312.03863, Dec. 2023.
- [16] Z. Yin, H. Wang, K. Horio, D. Kawahara and S. Sekine, "Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance," doi:10.48550/arXiv.2402.14531, 2024.
- [17] O. Dobariya and A. Kumar, "Mind Your Tone: Investigating How Prompt Politeness Affects LLM Accuracy (short paper)," doi:10.48550/arXiv.2510.04950, 2025.
- [18] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring Massive Multitask Language Understanding," arXiv preprint arXiv:2009.03300, Sept. 2020.
- [19] J. Neyman, "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 236, no. 767, pp. 333–380, Aug. 1937, doi: 10.1098/rsta.1937.0005.

- [20] H. Ni, S. Meng, X. Chen, Z. Zhao, A. Chen, P. Li, S. Zhang, Q. Yin, Y. Wang and Y. Chan, "Harnessing Earnings Reports for Stock Predictions: A QLoRA-Enhanced LLM Approach," in Proc. 2024 6th Int. Conf. Data-driven Optimization of Complex Systems (DOCS), 2024, doi:10.1109/DOCS63458.2024.10704454.
- [21] Yu. Yan, T. Hu and W. Zhu, "Leveraging Large Language Models for Enhancing Financial Compliance: A Focus on Anti-Money Laundering Applications," in Proc. 2024 4th Int. Conf. Robotics, Automation and Artificial Intelligence (RAAI), 2024, doi:10.1109/RAAI64504.2024.10949516.
- [22] C. Le, Z. Gong, C. Wang, H. Ni, P. Li and X. Chen, "Instruction Tuning and CoT Prompting for Contextual Medical QA with LLMs," in Proc. 2025 Int. Conf. Artificial Intelligence, Human-Computer Interaction and Natural Language Processing (ICAHN), 2025, doi:10.1109/ICAHN67688.2025.00016.
- [23] Y. Ji, W. Ma, S. Sivarajkumar, H. Zhang, E. M. Sadhu, Z. Li, X. Wu, S. Visweswaran, and Y. Wang, "Mitigating the risk of health inequity exacerbated by large language models," npj Digital Medicine, vol. 8, no. 1, p. 246, 2025. doi: 10.1038/s41746-025-01576-4.
- [24] X. Li, Y. Ma, Y. Huang, X. Wang, Y. Lin, and C. Zhang, "Synthesized Data Efficiency and Compression (SEC) Optimization for Large Language Models," in Proc. 2024 IEEE Energy, Information and Electronics Communications Symposium (EIECS), 2024, doi:10.1109/EIECS63941.2024.10800533.
- [25] Y. Ji, Z. Li, R. Meng and D. He, "Reason-to-Rank: Distilling Direct and Comparative Reasoning from Large Language Models for Document Reranking," doi: 10.1145/3726302.3730070.
- [26] H. Zhang, Y. Tian and T. Zhang, "Self-Anchor: Large Language Model Reasoning via Step-by-step Attention Alignment," arXiv:2510.03223, 2025.
- [27] H. Wang, S. Hao, H. Dong, S. Zhang, Y. Bao, Z. Yang, and Y. Wu, "Offline reinforcement learning for LLM multi-step reasoning," arXiv:2412.16145, 2024.
- [28] Y. Chen, H. Du, and Y. Zhou, "Lightweight network-based semantic segmentation for UAVs and its RISC-V implementation," *Journal of Technology Innovation and Engineering*, vol. 1, no. 2, 2025. doi:10.63887/jtie.2025.1.2.9.
- [29] S. Zeng, X. Chang, M. Xie, X. Liu, Y. Bai, Z. Pan, M. Xu, X. Wei and N. Guo, "FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving," doi:10.48550/arXiv.2505.17685, 2025.
- [30] OpenAI, "GPT-4o mini: advancing cost-efficient intelligence," OpenAI, Jul. 18, 2024. [Online]. Available: <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/> . [Accessed: Nov. 23, 2025].
- [31] M. Li, "OpenAI's 4o-mini has only 8B, Claude 3.5 Sonnet reaches 175B," AI Disruption, Jan. 2, 2025. [Online]. Available: <https://medium.com/ai-disruption/openais-4o-mini-has-only-8b-claude-3-5-sonnet-reaches-175b-9c1c55c53970> . [Accessed: Nov. 23, 2025].
- [32] D. Hassabis and K. Kavukcuoglu, "Introducing Gemini 2.0: our new AI model for the agentic era," Google DeepMind Blog, Dec. 11, 2024. [Online]. Available: <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/> . [Accessed: Nov. 23, 2025].
- [33] Meta, "Llama 4: multimodal intelligence," Meta AI Blog, Apr. 7, 2025. [Online]. Available: <https://ai.meta.com/blog/llama-4-multimodal-intelligence/> . [Accessed: Nov. 23, 2025].
- [34] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer," arXiv preprint arXiv:1701.06538, Jan. 2017.