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:

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

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:

$$\Delta \pm z_{0.975} \cdot \frac{s}{\sqrt{N}}, \tag{2}$$

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 is 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 suggests 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	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]
Humanities	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]
	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