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

Hanyu Cai[1†*], Binqi Shen[1†], Lier Jin[2], Lan Hu[3], Xiaojing Fan[4]
[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