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, "Synergized 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.