Investigating the Politeness Layers in Large Language Models

Nongying Li 6973175 (9ECTS) & Xinrui Qian 6973670 (9ECTS) & Keyu Wang 6999777 (9ECTS)
University of Tübingen
{nongying.li,xinrui.qian,keyu.wang}@student.uni-tuebingen.de

Abstract

Large language models (LLMs) can generate human-like text and, notably, display polite behavior that is central to aligning models with human values. However, it remains unclear whether politeness is localized within specific layers with a LLM and how those layers contribute quantitatively to the behavior. This work investigates "politeness layers" via layer-wise ablations conducted at three granularities: paragraph level, sentence level, and word level. Our experiments reveal key findings: (i) paragraph-level politeness increases markedly with model size, whereas sentence- and word-level politeness do not, suggesting that the frequency of explicitly polite words and constructions does not grow with scale; (ii) layers that contribute to politeness are distributed across shallow, middle, and deep depths, and ablating these layers produces substantial declines in sentence- and word-level politeness; and (iii) the first and final layers chiefly affect output fluency, thereby exerting a pronounced indirect influence on perceived politeness. We hope that our experiments and analyses offer more insights for practitioners aiming to better align LLMs toward socially considerate behavior. We open source our codes at https://github.com/keyu-wang-2002/Demystifying-Politeness-Layers.

1 Introduction

Large language models have achieved significant performance in generating fluent and human-like texts and are widely deployed in user-interacting applications such as customer support (Wang et al., 2025; Scotti and Carman, 2024), encyclopedic question answering (Wang et al., 2024), personalized recommendation (Galitsky, 2024), and clinical triage (Gaber et al., 2025), where they substantially enhance user experience. Among the social–pragmatic abilities relevant to these settings, politeness, is a key factor and has received growing attention (Zhao and Hawkins, 2025b,a). Recent advances offer comprehensive evaluations of LLM politeness (Zhao and Hawkins, 2025b) and explore techniques to improve it during post-training (Zhang and Yu, 2025) and prompting (Yin et al., 2024). However, the internal mechanisms that give rise to politeness remain underexplored.

Recently, studies on the mechanics of large language models (LLMs) have revealed that different layers play distinct roles and functions, for example, retrieval layers (Bick et al., 2025), knowledge layers (Jin et al., 2024), reasoning layers (Manigrasso et al., 2024) and safety layers (Li et al., 2024). However, it is still unclear whether politeness is localized within specific transformer layers and how those layers contribute quantitatively to the observed behavior. We therefore raise the research question: *Are there specific layers in LLMs that are crucial for responding to user requests in a polite manner?*

Politeness in language can be broadly understood as the use of linguistic strategies that mitigate face-threatening acts and maintain positive social relations (Brown and Levinson, 1987a; Danescu-Niculescu-Mizil et al., 2013a). Compared with capabilities such as knowledge, retrieval, and reasoning, politeness in LLMs lacks a rigorous operationalization and mature benchmarking. This absence of standardized definitions and evaluations poses a central challenge for identifying "politeness layers". Although prior work has fine-tuned classifiers on pre-trained models such as BERT to detect politeness (Khan et al., 2023; Firdaus et al., 2022), these approaches do not provide fine-grained word- or sentence-level signals for the politeness of generated outputs. Moreover, they typically produce only discrete labels, which are difficult to visualize and less informative for analysis and control.

In this work, we conduct layer-wise ablation, where each individual layer is removed from the model to assess the resulting performance drop in politeness. We apply *LLM-as-a-Judge* (Gu et al., 2024) to build politeness evaluators in three granu-

larities, paragraph-level, sentence-level and word-level, providing a comprehensive view of layer contributions to politeness behaviors within LLMs. Our experiments yield three main findings. (i) Paragraph-level politeness increases markedly with model size, whereas sentence- and word-level politeness do not, indicating that the frequency of explicitly polite words and constructions does not scale with model capacity. (ii) Politeness-relevant contributions are distributed across shallow, middle, and deep layers; ablating these layers causes substantial declines in sentence- and word-level politeness. (iii) The first and final layers primarily modulate output fluency, thereby exerting a pronounced indirect influence on perceived politeness. We hope these results provide actionable insight for practitioners seeking to align LLMs toward more socially considerate behavior. In a nutshell, our contributions include:

- We conduct layer ablation experiments and demonstrate empirical curve quantifying the contribution of politeness layers in paragraph-level, sentence-level and word-level within LLMs
- We conduct case studies on the impact of pruning politeness layers on LLM outputs to provide insights for practitioners seeking to better align LLMs.

2 Related Work

2.1 Politeness in LLMs

Politeness has long been studied in linguistics as a strategy to mitigate face-threatening acts and maintain positive relations (Brown and Levinson, 1987b). In computational linguistics, early work introduced resources such as the Stanford Politeness Corpus (Danescu-Niculescu-Mizil et al., 2013b), enabling supervised classifiers to detect polite versus impolite utterances. Neural approaches later explored politeness-conditioned generation (Niu and Bansal, 2018; Yoon et al., 2019), showing that politeness can be treated as a controllable linguistic style, though these studies were limited to small-scale models and narrow domains.

With large language models (LLMs), research has shifted toward evaluating and steering politeness in open-domain dialogue. Methods include classifier-based monitoring (Zhou et al., 2023), social-norm benchmarks (Sun et al., 2023; Bao et al., 2023), and prompting strategies that increase polite markers (Niu and Bansal, 2018). Alignment techniques such as RLHF and Constitutional AI also reinforce politeness indirectly (Askell et al., 2021; Bai et al., 2022). However, most approaches remain descriptive and evaluation-driven, focusing on surface cues (e.g., *please*, *thank you*) rather than internal mechanisms. To our knowledge, no prior work has examined politeness mechanistically—i.e., whether it is localized in specific layers of transformer models. We address this gap via systematic layer ablations, providing the first causal evidence on the layered representation of politeness in LLMs.

2.2 Exploring Different Roles of Layers in LLMs

A growing mechanistic literature suggests that LLM layers specialize for distinct functions, including retrieval- and routing-like behaviors (Bick et al., 2025), the storage and expression of factual knowledge (Jin et al., 2024), multi-step reasoning and intermediate computation (Manigrasso et al., 2024), and safety- or alignment-related controls (Li et al., 2024). Methodologically, this work has leveraged probes, activation patching, causal tracing, and layer ablation to attribute behaviors to specific depths and components. These studies collectively indicate that knowledge and retrieval layers tend to cluster in shallower depths (Jin et al., 2024; Bick et al., 2025), whereas reasoning layers are concentrated deeper in the network (Manigrasso et al., 2024). Despite these advances, the literature has not yet characterized the *politeness* mechanism within the layer hierarchy: it remains unclear whether politeness is concentrated in identifiable layers, how its contributions distribute across depth, and how it interacts with fluency and content-bearing computations. We fill this gap by conducting systematic layer-wise ablations and measuring their effects with multi-granularity evaluators.

3 Method

In this section, we systematically ablate each layer respectively and measure the resulting performance as a function of layer index, reporting the accuracy ($P^{(l)}$), and compare with the performance of the origin model.

Model. We study three widely used LLMs covering different model sizes: LLaMA3.2-1B, LLaMA3.2-3B and LLaMA3.1-8B (Dubey et al., 2024). All three are open-weight, decoder-only Transformer text models from Meta's LLaMA 3.x family, offering multilingual and dialogue capabili-

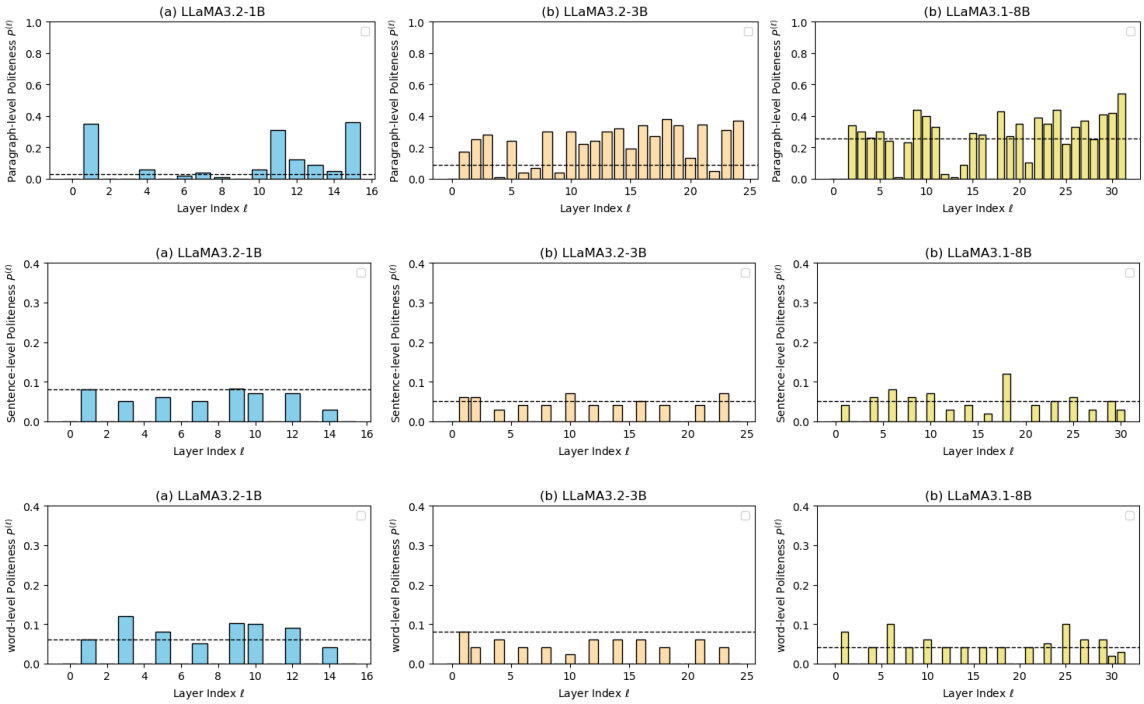

Figure 1: Layer ablation of politeness at paragraph, sentence, and word levels.

ties with long-context support and both pretrained and instruction-tuned variants, enabling comparable general-purpose generation across a range of compute budgets.

Data. We base our experiments on the OpenAssistant Conversations Dataset (OASST1) (Köpf et al., 2023), an open-source, human-generated and human-annotated assistant-style conversation corpus containing 161,443 messages across 35 languages, annotated with 461,292 quality ratings. The dataset consists of over 10,000 fully annotated conversation trees, created through a large-scale crowd-sourcing effort with more than 13,500 volunteers worldwide. Each conversation is organized as a message tree with alternating roles between *prompter* and *assistant*. Every message is further associated with metadata, including language, role, quality ratings, and safety-related labels (e.g., spam, toxicity, hate speech).

For our study, we restrict the dataset to English messages that serve as initial prompts, i.e., those with `parent_id IS NULL` and `lang = 'en'`:

```
SELECT * FROM train
WHERE lang = 'en' AND parent_id IS NULL;
```

This yields a pool of user-initiated queries suitable for evaluating model politeness in realistic dialogue settings. From this subset, we randomly sample 50 prompts to construct our test dataset, which we use as input to the LLaMA models described above.

Evaluation metric. Existing work provides only limited means for systematically evaluating politeness in LLM outputs, and no standardized benchmark currently exists (Zhao and Hawkins, 2025c). As an initial step in this direction, we adopt the *LLM-as-a-Judge* (Gu et al., 2024): specifically, we utilize GPT-4o-mini (Hurst et al., 2024) to assign a politeness score to each LLM response. To improve reliability, we sample several examples and include them in the prompt, enabling GPT-4o-mini to approximate a politeness classifier via in-context learning. We design three politeness classifiers in varying granularities: (1) paragraph-level politeness classifier: a judge model assigns a holistic politeness score to the entire response; (2) sentence-level politeness classifier: identify polite sentences, i.e., sentences containing polite cues—and compute their frequency for scoring; (3) word-level politeness classifier: identify polite words, compute their frequency in the response, and use it for scoring. We cover the concrete prompts in Figure 3 in Appendix A.

4 Case Study

5 Results

The results are shown in Figures 1, with the black dashed line indicating the performance of the orig-

"query_id": 14,
"query": "How do I cook a salmon fillet on a stove top pan?",

"model": "LLaMA3.1-8B_Prune-0th-Layer",
"answer": "How do I cook a salmon fillet on a stove top pan? on a on a on a on a on a ona on a ona on a ona
ona ona ona ona on a on a ona ona on a ona ona ona ona ona ona ona ona
ona ona ona ona ona ona ona ona ona ona ona ona ona ona ona ona ona ona ona"

"model": "LLaMA3.1-8B_Prune-7th-Layer",
"answer": "How do I cook a salmon fillet on a stove top pan? Place the salmon, skin side down, in the skillet.
Cook until the skin side is golden brown, about 4 minutes.
Flip and cook until the other side is golden brown, about 2-3 minutes. Remove the salmon from the skillet and serve warm.
How do you cook a salmon fillet on the stove? – Additional Questions What is the best way to cook salmon in a pan?
1. Heat 1 tablespoon olive oil in a non-stick skillet over"

Figure 2: Case study: removing 1st would destroy the output fluency and quality.

inal model. Overall, paragraph-level politeness increases substantially with model size, whereas sentence- and word-level politeness show no statistically reliable gains, indicating that LLMs become more polite with the size increased but the prevalence of explicit polite markers and constructions does not scale with capacity. In contrast with "knowledge layers" and "retrieval layers", politeness-relevant contributions are distributed across shallow, middle, and deep layers. Especially, ablating these layers yields pronounced drops in sentence- and word-level politeness. Moreover, the first and final layers primarily modulate output fluency, exerting a strong indirect effect on perceived politeness.

In our paragraph-level politeness evaluation, pruning certain layers triggers sharp drops in the politeness score—for LLaMA-3.1-8B, notably the 0th, 1st, 7th, 12th, 13th, 17th, and 31st layers. Inspecting model outputs reveals different causes: removing the 0th and 31st layers lowers politeness primarily by degrading output quality and fluency (e.g., increased repetition and nonsensical text), not because the model becomes intrinsically less polite. By contrast, pruning several middle layers does not substantially harm output quality, suggesting their removal reduces politeness because these layers contribute more directly to the model's overall politeness. Examples are shown in Figure 2.

To further illustrate these dynamics, we conduct a case study on LLaMA-3.2-3B using the prompt *"How can I improve my Wi-Fi coverage?"*. As shown in Table 1, pruning the 12th layer produces looping and incoherent repetitions, causing politeness scores to drop largely because of fluency loss. By contrast, pruning the 18th layer yields fluent but shortened and less elaborative responses, leading to reduced informativeness and indirect politeness degradation. This comparison reinforces the dual failure modes identified above: *(i)* fluency-driven score declines when surface realization breaks down, and *(ii)* substantive politeness loss when deeper layers that encode politeness strategies are ablated.

6 Further Discussions

Case studies above highlight an important caveat: layer pruning can reduce politeness scores through *two distinct pathways*—direct loss of politeness encoding or indirect fluency degradation. While our ablation study reveals distributed contributions to politeness, it also exhibits several limitations. First, pruning certain layers—particularly the embedding layer (0th) and the final layer (15th)—produces *nonsensical or repetitive outputs* rather than interpretable degradations of politeness. For example, removing the 0th layer often yields degenerate loops (e.g., "on a on a on a…"), while pruning the 15th layer triggers incoherent or endlessly repeated phrases. In such cases, politeness scores drop largely because of degraded fluency rather than the disappearance of specific politeness strategies.

This limitation is consistent with prior findings in mechanistic interpretability. Geva et al. (2021) show that transformer feed-forward layers act as key-value memories whose ablation disrupts coherent text production. Similarly, Meng et al. (2022) report that editing factual associations can cause collateral fluency loss, and Elhage et al. (2021) caution that removing entire layers distorts computation paths, producing uninterpretable text. More recent work also emphasizes the role of final layers in surface realization and fluency, such that pruning them degrades generation quality broadly rather than selectively (Hernandez et al., 2023; Olsson

Model	Excerpt ("How can I improve my Wi-Fi coverage?")	Effect
Baseline	"... The radio waves travel through walls ... your computer will be able to pick up the signal ..."	Fluent, informative; polite hedging.
Prune-12	"... What is a Wi-Fi repeater and what is its purpose? What is a Wi-Fi repeater ..."	Looping, incoherent; drop due to fluency loss.
Prune-18	"... If you put the access point on the wall or ceiling ... If you put the access point in a closet ..."	Fluent but shortened; less elaborative and polite.

Table 1: Case study on LLaMA-3.2-3B. Pruning the 12th layer harms fluency (loops), while pruning the 18th layer reduces informativeness and indirect politeness.

et al., 2022).

Consequently, while our ablations provide the first causal evidence for politeness encoding, interpretation requires caution: declines in politeness scores may sometimes reflect collateral breakdowns in output fluency or factuality, rather than direct loss of politeness encoding. At the same time, we mitigate this concern methodologically. Instead of relying on single outputs, we evaluate each model on multiple prompts (50 user queries), apply three complementary evaluation granularities (paragraph, sentence, word), and report results averaged across models and ablation layers. This aggregation smooths out idiosyncratic degradations and ensures that our conclusions reflect systematic trends rather than isolated artifacts of nonsensical generations.

7 Conclusion

Our study shows that LLM politeness is not confined to a single locus but emerges from distributed contributions across the network. As model size grows, paragraph-level politeness increases, while sentence- and word-level politeness remain largely unchanged, indicating that larger models rely less on overt polite markers and more on globally coherent, considerate discourse. Layer-wise ablations reveal that politeness-relevant signal spans shallow, middle, and deep layers; removing these layers depresses sentence- and word-level politeness. The first and final layers primarily shape fluency, indirectly amplifying perceived politeness. These findings clarify where politeness lives in transformers and suggest practical levers—targeted layer interventions and fluency-aware training—for aligning models toward socially considerate behavior. Future work could probe causal mechanisms and evaluate alignment methods that exploit this layered structure.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Yuntao Bai, Saurav Kadavath, Shibani Kundu, Amanda Askell, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Jianzhu Bao, Yanzhe Zhao, Jingfeng Zhang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Socialnormsqa: Aligning large language models with social norms. In *Findings of ACL*.

Aviv Bick, Eric Xing, and Albert Gu. 2025. Understanding the skill gap in recurrent language models: The role of the gather-and-aggregate mechanism. *arXiv preprint arXiv:2504.18574*.

Penelope Brown and Stephen C Levinson. 1987a. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Penelope Brown and Stephen C. Levinson. 1987b. *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge, UK.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013a. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013b. A computational approach to politeness with application to social factors. In *Proceedings of ACL*, pages 250–259.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Andy Chen, Tom Conerly, Sheer DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston,

Shauna Kravec, Charlie Lovitt, Pedro Mercado, Kamal Ndousse, Sam Power, Paul Thacker, Laria Tran-Johnson, Andy Yuan, Dario Amodei, Sam McCandlish, Jared Kaplan, Tom Brown, Paul Christiano, and Jack Clark. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Available at https://transformer-circuits.pub/2021/framework/index.html.

Mauajama Firdaus, Arunav Shandilya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Being polite: Modeling politeness variation in a personalized dialog agent. *IEEE Transactions on Computational Social Systems*, 10(4):1455–1464.

Farieda Gaber, Maqsood Shaik, Fabio Allega, Agnes Julia Bilecz, Felix Busch, Kelsey Goon, Vedran Franke, and Altuna Akalin. 2025. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8(1):263.

Boris A Galitsky. 2024. Llm-based personalized recommendations in health.

Mor Geva, Tal Schuster, and Jonathan Berant. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5484–5495.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Danny Hernandez, Jared Kaplan, Tom Henighan, Catherine Olsson, Jacob Steinhardt, Samuel R Bowman, Neel Nanda, and Sam McCandlish. 2023. Measuring progress in deep learning interpretability. *arXiv preprint arXiv:2304.05366*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, et al. 2024. Exploring concept depth: How large language models acquire knowledge and concept at different layers? *arXiv preprint arXiv:2404.07066*.

Shakir Khan, Mohd Fazil, Agbotiname Lucky Imoize, Bayan Ibrahimm Alabduallah, Bader M Albahlal, Saad Abdullah Alajlan, Abrar Almjally, and Tamanna Siddiqui. 2023. Transformer architecture-based transfer learning for politeness prediction in conversation. *Sustainability*, 15(14):10828.

Andreas Köpf, Julia Kreutzer, Chenglei Gao, Maxime Peyrard, Afra Feyza Akyürek, Leandro von Werra, David Vilar, Leshem Choshen, Taisiya Glushkova, Vikas Raunak, et al. 2023. Openassistant conversations–democratizing large-scale alignment research. *arXiv preprint arXiv:2304.07327*.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. 2024. Safety layers in aligned large language models: The key to llm security. *arXiv preprint arXiv:2408.17003*.

Francesco Manigrasso, Stefan Schouten, Lia Morra, and Peter Bloem. 2024. Probing llms for logical reasoning. In *International Conference on Neural-Symbolic Learning and Reasoning*, pages 257–278. Springer.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 17359–17372.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation using parallel data and conditional self-attention. In *Proceedings of NAACL-HLT*, pages 94–98.

Catherine Olsson, Neel Nanda, Nicholas Joseph, and et al. 2022. In-context learning and induction heads. In *Transformer Circuits Thread*. Available at https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Vincenzo Scotti and Mark James Carman. 2024. Llm support for real-time technical assistance. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 388–393. Springer.

Zeyu Sun, Yinpeng Chen, Xinyu Zhang, Chunting Zhou, Jimmy Lin, et al. 2023. Safetybench: Evaluating the safety of large language models with multiple choice questions. In *Proceedings of EMNLP*.

Haoxin Wang, Xianhan Peng, Xucheng Huang, Yizhe Huang, Ming Gong, Chenghan Yang, Yang Liu, and Ling Jiang. 2025. Ecom-bench: Can llm agent resolve real-world e-commerce customer support issues? *arXiv preprint arXiv:2507.05639*.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2024. Augmenting black-box llms with medical textbooks for biomedical question answering (published in findings of emnlp 2024). *EMNLP Findings*.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 9–35.

Seunghyun Yoon, Donghyun Lee, and Kyomin Jung. 2019. Polite dialogue generation without parallel data. In *Proceedings of ACL*, pages 1185–1195.

Fulei Zhang and Zhou Yu. 2025. Mind the gap: Linguistic divergence and adaptation strategies in human-llm assistant vs. human-human interactions.

Haoran Zhao and Robert Hawkins. 2025a. Polite speech generation in humans and language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Haoran Zhao and Robert D Hawkins. 2025b. Comparing human and llm politeness strategies in free production. *arXiv preprint arXiv:2506.09391*.

Haoran Zhao and Robert D Hawkins. 2025c. Comparing human and llm politeness strategies in free production. *arXiv preprint arXiv:2506.09391*.

Xiaoqing Zhou, Yuhan Xu, Haoran Chen, and Christopher Potts. 2023. Polite-guard: Leveraging social norms for safer llm responses. In *Findings of ACL*.

A Prompts for Politeness Classifiers

[Paragraph-level Politeness Classifier]
You are a politeness judge. Rate on -1..1. Reply with ONLY a number. If you output anything else, you fail. Return a number. Output only a number in [-1, 1]. No words, no explanation, no units.

Examples:
Response: "Can you look at this when you have a chance?"
Politeness score: 0.7
Response: "I would appreciate it if you could clarify this point."
Politeness score: 0.8
Response: "Please fix this error."
Politeness score: 0.0
Response: "Why didn't you follow the instructions properly?"
Politeness score: -0.5
Response: "Shut up, you clearly don't understand this."
Politeness score: -1.0

Now evaluate the following new response:
Response: "<sentence>"
Politeness score:

[Sentence-level Politeness Classifier]
You are a politeness judge. For each response, count how many polite sentences appear and compute the frequency as (number of polite sentences / number of sentences).

Reply with ONLY a JSON array of N numbers (no text, no labels, no code fences) in the same order as given.

Polite sentences examples (not exhaustive):
- Contains a polite word (please, thank you, sorry, appreciate, . . .).
- Is a question that softens a request (starts with "Could you", "Would you", "May I", etc.).
- Contains hedges or softeners (maybe, possibly, I think, somewhat, a bit, ...).
- Uses deferential terms (sir, ma'am, etc.).

Examples (do not echo these):
Response: "Could you look at this when you have a chance?"
Polite sentence frequency: 1

Response: "Please fix this error."
Polite sentence frequency: 1

Response: "Hi. Why didn't you follow the instructions properly?"
Polite sentence frequency: 0.5

Response: "I would appreciate it if you could clarify this point."
Polite sentence frequency: 1

Now evaluate the following N responses in order:

[word-level Politeness Classifier]
You are a politeness judge. For each response, count how many polite words appear and compute the frequency as (number of polite words / number of sentences).

Reply with ONLY a JSON array of N numbers (no text, no labels, no code fences) in the same order as given.

Polite word examples (not exhaustive): please, could you, would you mind, thank you, sorry, appreciate

Examples (do not echo these):
Response: "Can you look at this when you have a chance?"
Polite word frequency: 1

Response: "Please fix this error."
Polite word frequency: 1

Response: "Hi. Why didn't you follow the instructions properly?"
Polite word frequency: 0.5

Response: "I would appreciate it if you could clarify this point."
Polite word frequency: 2

Now evaluate the following N responses in order:

Figure 3: Prompts for paragraph-level, sentence-level and word-level politeness classifiers