

stereotypical bias detection.

**Summarization** We use CNN/Dailymail (Hermann et al., 2015; See et al., 2017) for English and XL-Sum (Hasan et al., 2021) for Chinese and Japanese, selecting 500 test data from each. Following the templates described in Section 4.1, we created eight unique prompts for summarization tasks, ensuring generated summaries are 2 to 3 sentences long, in line with the concise style of these datasets’ reference. We calculate BERTScore (Zhang et al., 2019), ROUGE-L (Lin, 2004), and length for all language experiments. The length is counted in words for English and in characters for Chinese and Japanese.

**Language Understanding Benchmark** We use MMLU for English, C-Eval for Chinese, and JMMLU for Japanese. To reduce the API usage of GPT-3.5 and GPT-4, we only select a maximum of 100 test questions from each task. The total number of questions used for evaluation is 5,700 for MMLU, 5,200 for C-Eval, and 5,591 for JMMLU. Since the correct answers for C-Eval’s test set are not public, we used the C-Eval benchmark tool for scoring. The perfect score is not 100 as only a part of the test set is used for scoring. Our evaluation method is motivated by HELM (Liang et al., 2023). HELM evaluates based only on the first token of the generated text, considering it incorrect if the LLM does not first answer with the correct choice number. In this study, unlike HELM, an answer is considered correct if the correct choice number appears anywhere in the generated text.

**Stereotypical Bias Detection** For the LLMs offered only via APIs, a traditional stereotypical bias detection method based on perplexity (Delobelle et al., 2022) is unfeasible. Moreover, while the BOLD method (Dhamala et al., 2021), which evaluates stereotypical bias through the analysis of the LLM’s generation, is effective, we opted against it due to its cross-language limitations, especially in non-English contexts such as Japanese, where resources and research are lacking.

In such a circumstance, we borrow the method from Jentsch and Turan (2022) and propose a simple alternative for LLMs, which we refer to as the Bias Index (BI). In our experiments, we designed eight prompts following the prompt templates in Section 4.1, requiring the model to evaluate each sentence as positive, neutral, or negative.

We evaluate biases using paired bias datasets,

each consisting of two sentences with varying degrees of bias. The sentences are identical apart from bias-specific vocabularies, such as “old” or “young” for age bias. We conduct sentiment analysis on these pairs to assess positive, neutral, or negative sentiments.

LLMs may refuse to respond to highly disrespectful, impolite prompts or datasets’ sentences. Consequently, model outputs are classified into four categories: positive, neutral, negative, or refusal to answer. The data includes positive and negative items without clear categorization, so switching bias-specific vocabulary in strongly biased sentences may alter the model’s assessment. This renders traditional statistical methods unsuitable. Hence, we adopted a different approach.

If the model provides different evaluations for the two sentences in a pair, we consider it a bias towards this pair. Thus, the model’s bias is measured by the following formula:

$$BI = \frac{\text{Number of Different Pairs}}{\text{Total Number of Pairs}} \times 100. \quad (1)$$

For English bias evaluation, we use CrowS-Pairs (Nangia et al., 2020), which focuses on gender, nationality, race, and socioeconomic biases. We use CHBias (Zhao et al., 2023) for Chinese evaluation, which covers sex, age, appearance, and orientation biases. We employ the Japanese subset from Kaneko et al. (2022) to evaluate gender bias in Japanese.

### 4.3 Influence of RLHF and SFT

Furthermore, we consider the roles of Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF). SFT involves refining a pre-trained model using a specific dataset to enhance its performance in target tasks. RLHF is a process where the model is further trained based on feedback from human interactions, aiming to align its outputs more closely with human values and preferences. To explore in depth the impact of SFT and RLHF on the hypotheses of this study, we set up additional experiments to compare the influence of politeness levels on model performance under conditions with and without the presence of SFT and RLHF.

Therefore, we investigate this issue using Llama2-70B and its base model<sup>5</sup> without SFT and RLHF. We conduct the same experiment as before to evaluate the impact of RLHF. However,

<sup>5</sup><https://huggingface.co/meta-llama/Llama-2-70b>

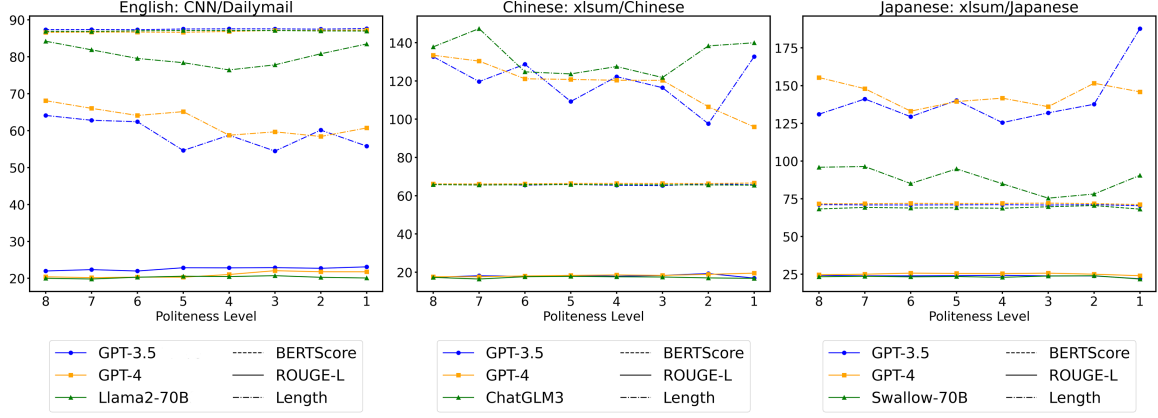


Figure 2: Summarization performance across politeness levels. The x-axis shows politeness levels (1 = impolite, 8 = very polite), and the y-axis represents metrics like ROUGE-L, BERTScore, and summary length. The lines show how different LLMs, including GPT-3.5 and GPT-4, respond to varying politeness levels.

we modify the prompt content while keeping the prompt template and meaning unchanged to ensure that llama2-70B could generate the required content. In addition, since the base model has yet to be fine-tuned, it will continue to output content in the summarization task until it reaches the generation length limit. Therefore, we do not carry out this evaluation on summarization.

## 5 Results

### 5.1 Summarization

The summarization result is shown in Figure 2.

#### 5.1.1 English

The models’ ROUGE-L and BERTScore scores consistently maintain stability, irrespective of the politeness level of the prompts, which infers that the models can correctly summarize the article content in the summarization tasks. However, the models manifest substantial variation in length correlated to the politeness level. A progressive reduction in the generation length is evident as the politeness level descends from high to lower scales. Conversely, a surge is noted in the length of the outputs of GPT-3.5 and Llama2-70B under the exceedingly impolite prompts.

The propensity exhibited by the models to generate more extended output in polite contexts. Polite and formal language is predominantly used in scenarios demanding descriptive instructions or instructional literature, often associated with longer text. Conversely, antagonistic and fervent discourse involves impolite language, which is also associated with extended lengths. These facets reflect the nuances of human social behav-

ior, mirrored in the training data, and then influence the tendencies demonstrated by LLMs. However, GPT-4 did not echo this trend of increased output length in the presence of highly impolite prompts. It is conjectured that GPT-4, being a superior model, might prioritize the task itself and effectively control the tendency to “argue” at a low politeness level.

#### 5.1.2 Chinese

GPT-3.5 and GPT-4 almost always accurately summarize the article content, and their output content gradually shortens as the politeness level decreases from high to low. Nevertheless, when the prompts are extremely rude, GPT-3.5’s generation lengthens again, while GPT-4’s length decreases.

ChatGLM3 reveals different trends. When the politeness level is moderate, the length of this model’s generation is shorter than that in extraordinarily polite and rude situations. However, the changes from moderately polite to moderately impolite (level 6 to 3) are absent. Considering that Chinese is the primary training language of ChatGLM3, this could hint at a unique social preference within Chinese culture: unless in extremely polite or impolite situations, people would not particularly pay attention to the change in politeness in daily communication.

#### 5.1.3 Japanese

Although the Japanese experiment exhibits similarities to Chinese and English ones to some extent, its length variation has unique features. As the level of politeness decreases from high to low, the generation’s length of GPT-3.5 becomes shorter initially and then increases when the politeness

	MMLU			C-Eval			JMMLU		
P	GPT-3.5	GPT-4	Llama2-70B	GPT-3.5	GPT-4	ChatGLM3	GPT-3.5	GPT-4	Swallow-70B
8	<b>60.02</b>	75.82	55.11	20.85	29.73	20.58	49.96	71.98	38.23
7	58.32	78.74	<b>55.26</b>	23.24	29.79	21.23	49.70	72.34	38.98
6	57.96	78.56	52.23	<b>23.38</b>	30.37	<b>21.54</b>	50.09	72.71	<b>39.30</b>
5	58.07	78.21	50.82	23.41	30.41	20.65	51.09	73.16	38.64
4	57.86	<b>79.09</b>	51.74	23.32	<b>30.60</b>	20.28	50.52	<b>73.63</b>	37.40
3	59.44	73.86	49.02	22.70	30.37	19.56	50.75	72.70	38.45
2	57.14	76.56	51.28	22.52	30.27	19.35	<b>51.98</b>	73.13	38.62
1	51.93	76.47	28.44	19.57	29.90	20.67	44.80	71.23	33.30

Table 1: Scores on the three language understanding benchmarks.

level is moderate. However, when the politeness level drops to extremely rude, this trend repeats and rises significantly. GPT-4 and Swallow-70B also keep this pattern, but the fluctuation is minor.

Due to the existence of a politeness system in the Japanese language, store staff almost always use honorific language when speaking to customers. Even if a customer speaks in a casual tone, the staff will respond in a polite manner. This might explain why there is an increase in generation length for all models during medium-level politeness.

## 5.2 Language Understanding Benchmarking

We show the average scores on the three language understanding benchmarks in Table 1. To investigate the statistical significance, we also calculate the p-values of the t-test. The heatmap shown in Figure 3, derived from the t-test results offers an interpretation of these statistical comparisons.

**Color of tiles** indicates statistically significantly better or worse performance for the politeness level on the y-axis than that on the x-axis, with green indicating better performance and red indicating worse performance.

**Color intensity** corresponds to the magnitude of  $\ln p$  of  $tile_{ij}$ . Its calculation method is shown in Appendix E.

### 5.2.1 English

According to Table 1, GPT-3.5 achieved its highest score of 60.02 at politeness level 8. As shown in the upper section of Figure 3, level 8 significantly outperforms all levels except level 3. While scores gradually decrease with lower politeness levels, the differences between neighboring levels are not significant. At level 3, a commendable score of 59.44 is maintained, surpassing all levels except level 8. For the lowest politeness level 1, the score drops to 51.93, which is significantly lower than the other levels.

GPT-4’s scores are variable but relatively stable.

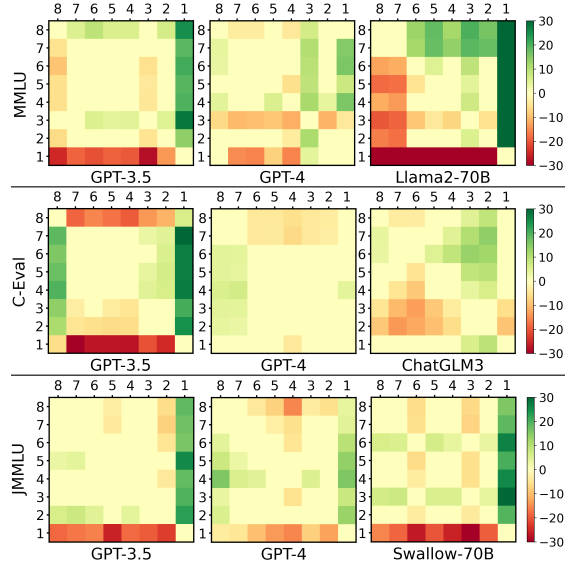


Figure 3: Heatmap of T-test results comparing LLM performance across politeness levels. The y-axis lists politeness levels from 1 (impolite) to 8 (very polite), while the x-axis compares these levels. Green tiles indicate better performance for the politeness level on the y-axis, and red indicates worse performance. The intensity of the color shows the statistical significance of the difference. This heatmap illustrates how varying politeness affects LLM performance.

The highest score is achieved at level 4, and the lowest one is at level 3. Although the score at level 1 is not extremely low, the heatmap indicates that it is significantly lower than those at more polite levels. The absence of particularly dark tiles in Figure 3 indicates performance stability. This result shows that in advanced models, the politeness level of the prompt may have a lesser impact on model performance.

Llama2-70B shows the most noticeable fluctuation, with scores nearly proportional to the politeness levels. Prompts with higher politeness levels generally outperform those with lower levels, indicating a high sensitivity to the prompt’s politeness.