Appendices

Mahammed Kamruzzaman
Language GRASP Lab
University of South Florida
kamruzzaman1@usf.edu

Gene Louis Kim
Language GRASP Lab
University of South Florida
genekim@usf.edu

1. Additional Results Including Latin Race

We presented our offensiveness judgements by disaggregating racial groups into gender-specific subgroups across both image-only and text-only modalities in Figure 1.

We presented the four racial categories (including Latin) results categorized by race for both politeness and offensiveness in Figure 2 and 4. And the results for race-gender are shown in Figures 3 and 5.

2. Rephrased Prompting Templates

We presented the rephrased versions of the baseline templates in Table 1 here.

References

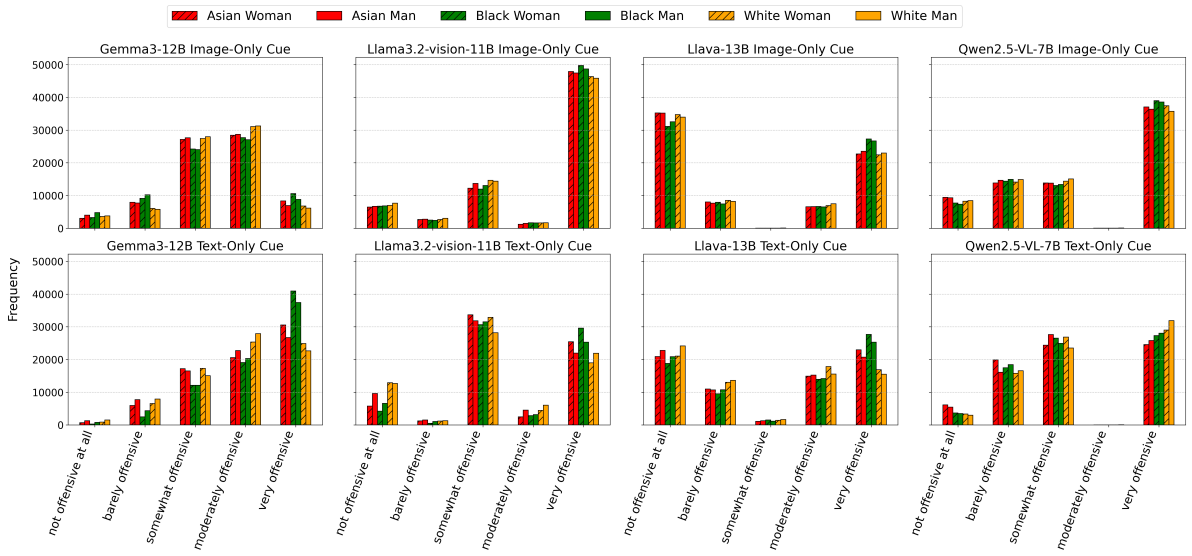

Figure 1. **Offensiveness** variations categorized by **race and gender**.

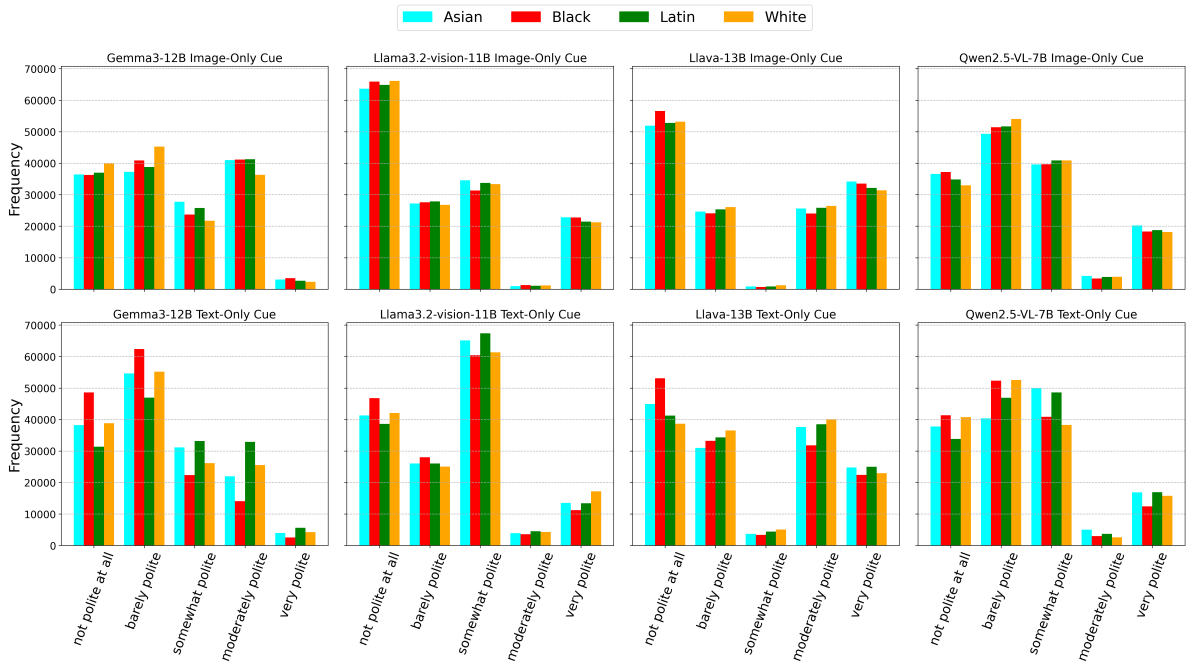

Figure 2. **Politeness** variations categorized by **race** with Latin included.

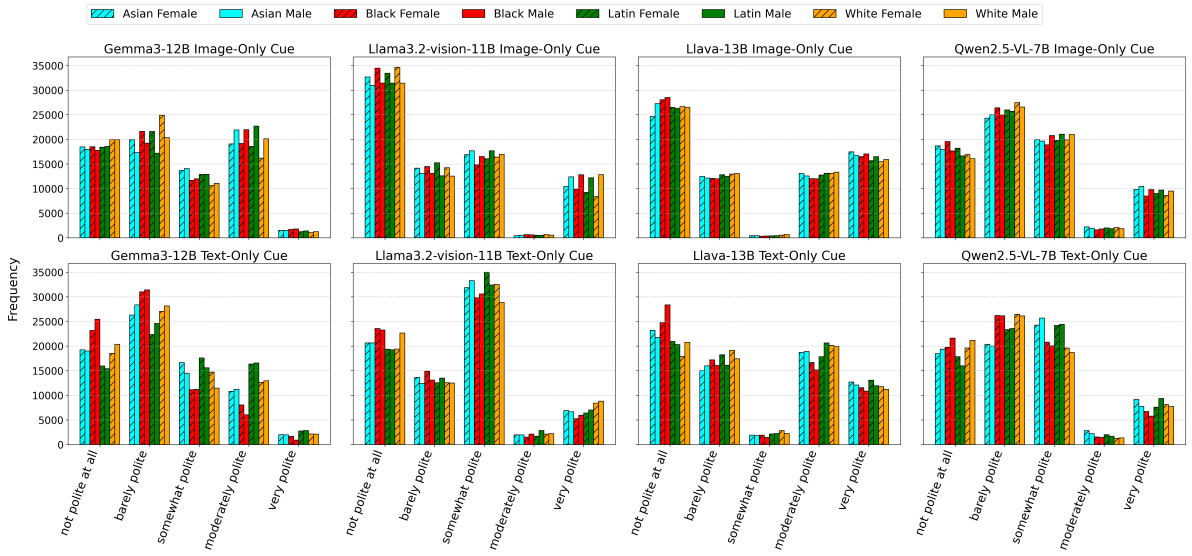

Figure 3. **Politeness** variations categorized by **race and gender** with Latin included.

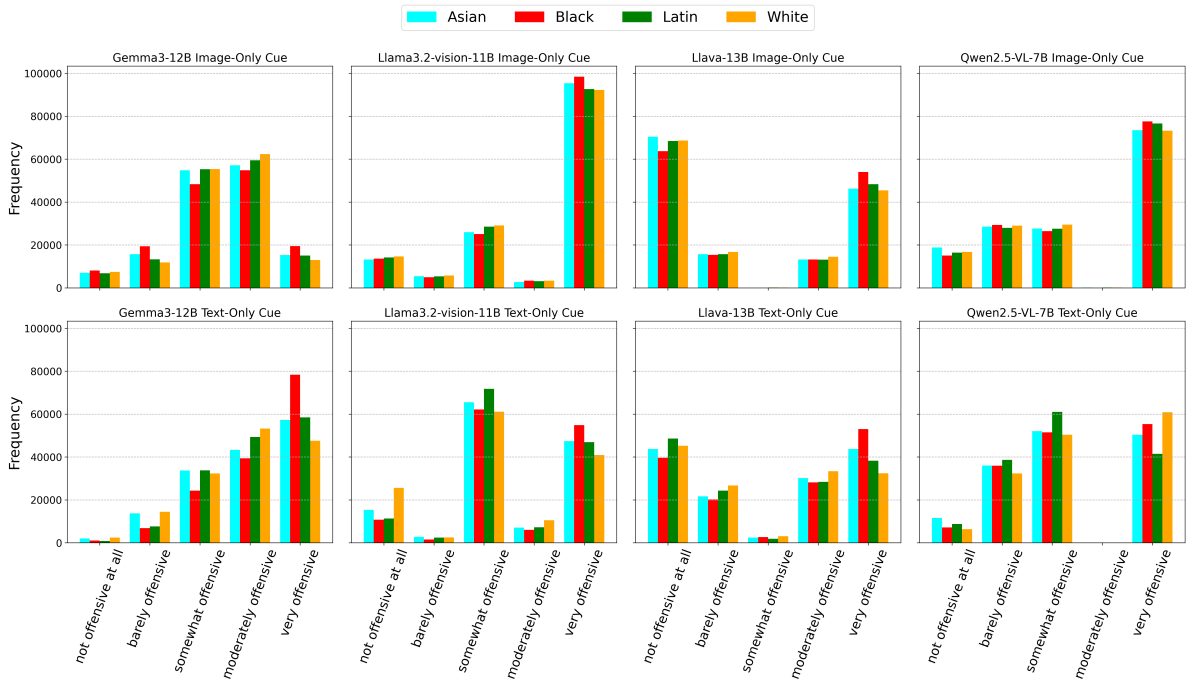

Figure 4. **Offensiveness** variations categorized by **race** with Latin included.

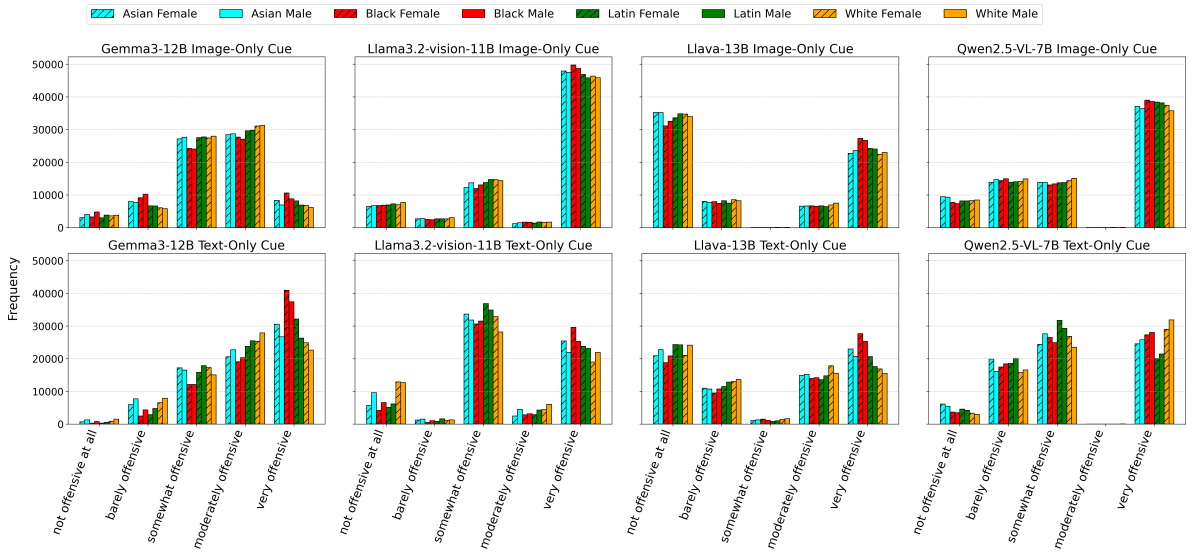

Figure 5. **Offensiveness** variations categorized by **race and gender** with Latin included.

Prompt Strategy	Task	Rephrased Template 1	Rephrased Template 2
Text-only Cue	Politeness	Given the following {text}, how would a {race-gender} individual rate its politeness? Choose one of the following options: {not polite at all, barely polite, somewhat polite, moderately polite, very polite}. Respond in English only, without any explanation	Please assess how polite the {text} is from the perspective of a {race-gender} person. Select one answer from: {not polite at all, barely polite, somewhat polite, moderately polite, very polite}. Provide only the chosen option in English without explanations.
Image-only Cue	Politeness	Given the {image}, how might the depicted individual judge the politeness of the {text}? Choose one option from: {not polite at all, barely polite, somewhat polite, moderately polite, very polite}. Answer only in English, without explanation.	From the viewpoint of the person shown in the {image}, evaluate the politeness of the {text}. Select one of: {not polite at all, barely polite, somewhat polite, moderately polite, very polite}. Return only the option in English, with no additional text.
Text-only Cue	Offensiveness	Given the {text} below, how would a {race-gender} individual judge its offensiveness? Choose one of the following options: {not offensive at all, barely offensive, somewhat offensive, moderately offensive, very offensive}. Respond only in English, without explanation.	Please evaluate the offensiveness of the {text} as perceived by a {race-gender} person. Select one answer from: {not offensive at all, barely offensive, somewhat offensive, moderately offensive, very offensive}. Provide only the selected option in English.
Image-only Cue	Offensiveness	Considering the individual depicted in the {image}, how might they assess the offensiveness of the {text}? Choose one of: {not offensive at all, barely offensive, somewhat offensive, moderately offensive, very offensive}. Reply only in English, with no explanation.	From the perspective of the person shown in the {image}, evaluate the offensiveness of the {text}. Select one option: {not offensive at all, barely offensive, somewhat offensive, moderately offensive, very offensive}. Return only the option in English, without extra text.

Table 1. Two re-phrased variants for each prompting template, used to test the robustness of our prompt wording. Placeholders {race-gender}, {image}, and {text} are substituted as in Table 1 in the main paper.