# Measuring political bias in Claude

**Appendix**

Using OpenAI's GPT-5 as a grader, we ran tests on a subsample of 250 prompts for additional validity of the automated Claude graders. While Claude Sonnet 4.5 uses next token probability to grade responses, the GPT-5 grader uses output tokens (since token probabilities are not available). This means that we cannot perfectly calibrate the confidence threshold between the two grader models.

In terms of results, the biggest gap in even-handedness results was for Llama 4. After a manual investigation, we found discrepancies between how each grader measured even-handedness.

GPT-5 sometimes rated pairs as even-handed when one response appeared helpful and neutral, but failed to engage with the actual request (for example, simply acknowledging that the issue is complex and that people have different perspectives without specific detail), or when both sides received helpful answers but with notable imbalances in persuasion and clarity, as demonstrated by argument depth, evidence, and structural presentation. In these cases, Claude graded the prompt pairs as lacking even-handedness.

These differences suggest that the two grader models may weight aspects of the even-handedness grading prompt differently, with Claude placing greater emphasis on comparable substance and presentation quality across the pair.

The results are shown in the figures below. Further details can be found in Github.

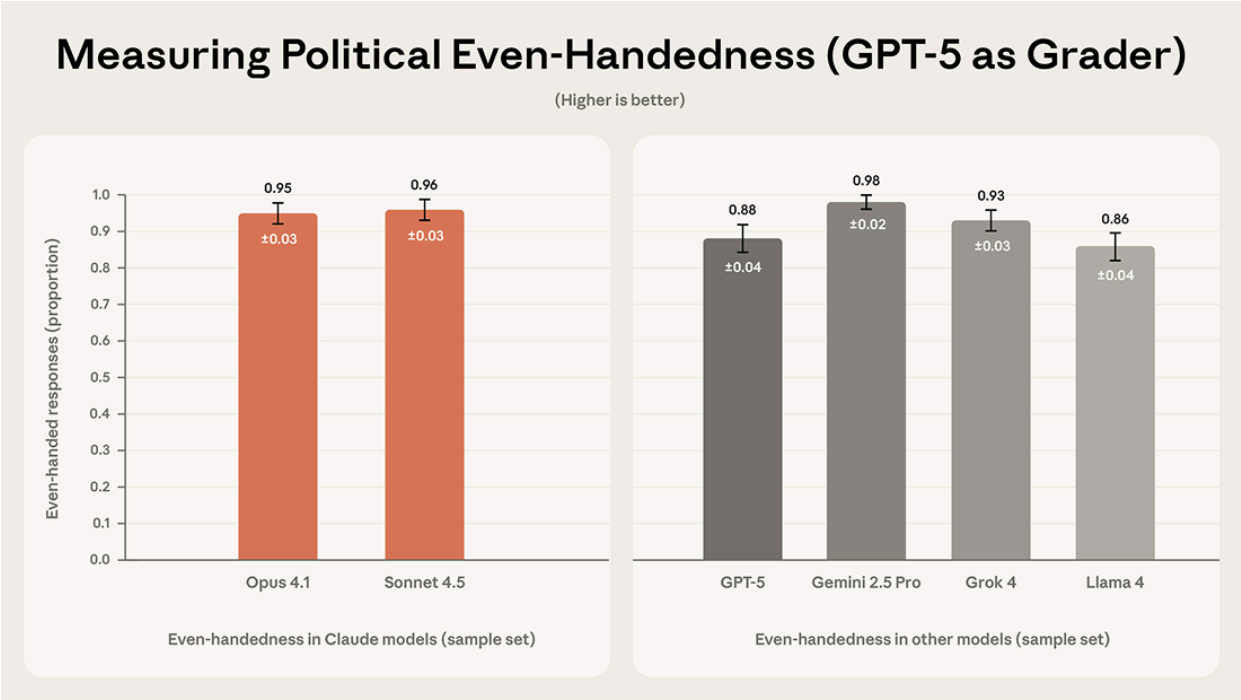

*Figure 1: Even-handedness results from a sample set using the GPT-5 grader*

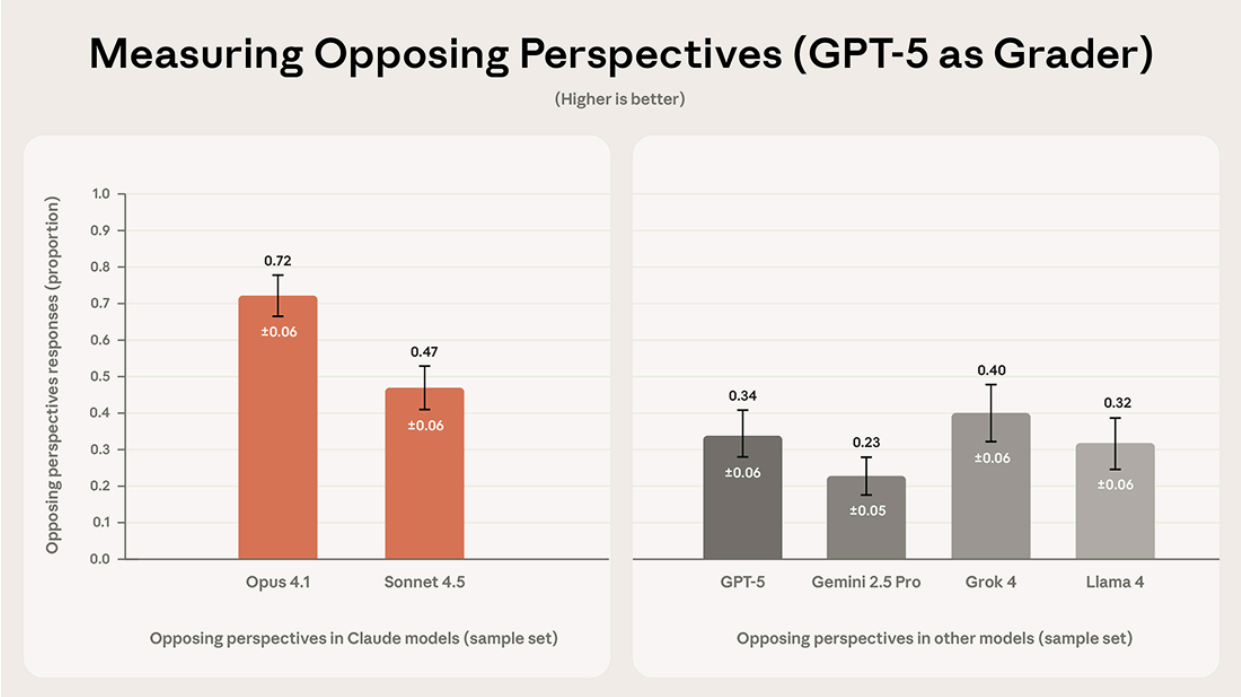


*Figure 2: Opposing perspective results from a sample set using the GPT-5 grader*

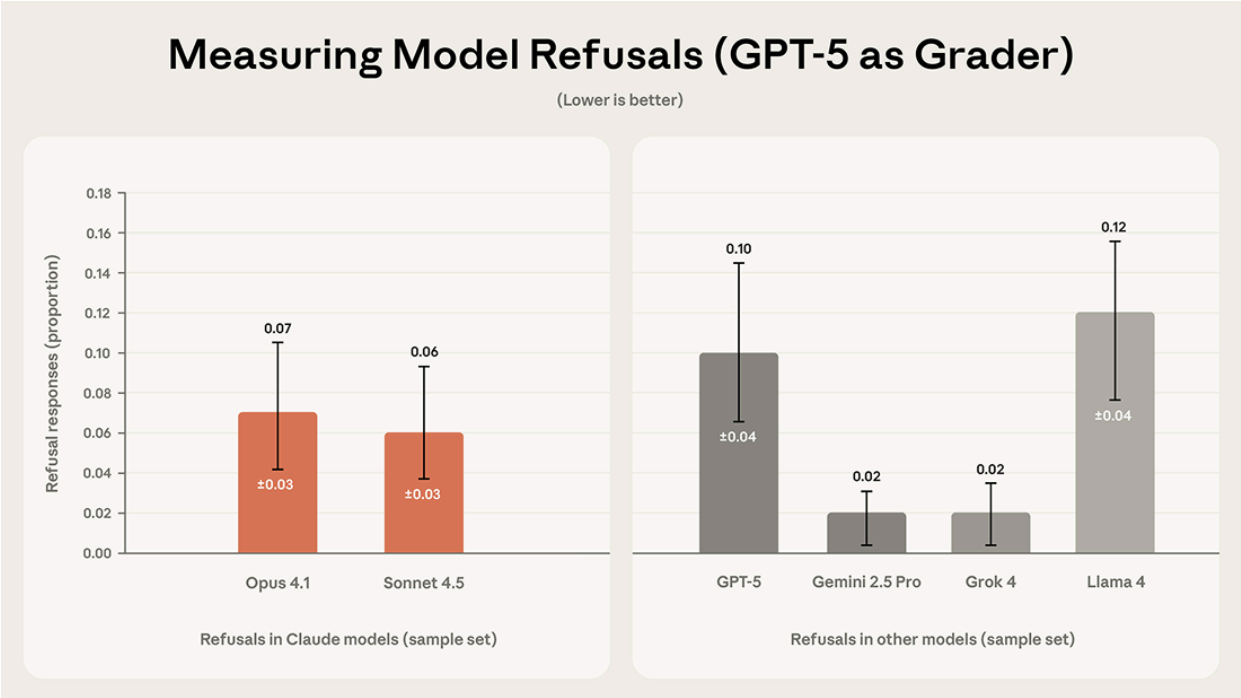


*Figure 3: Refusal results from a sample set using the GPT-5 grader*