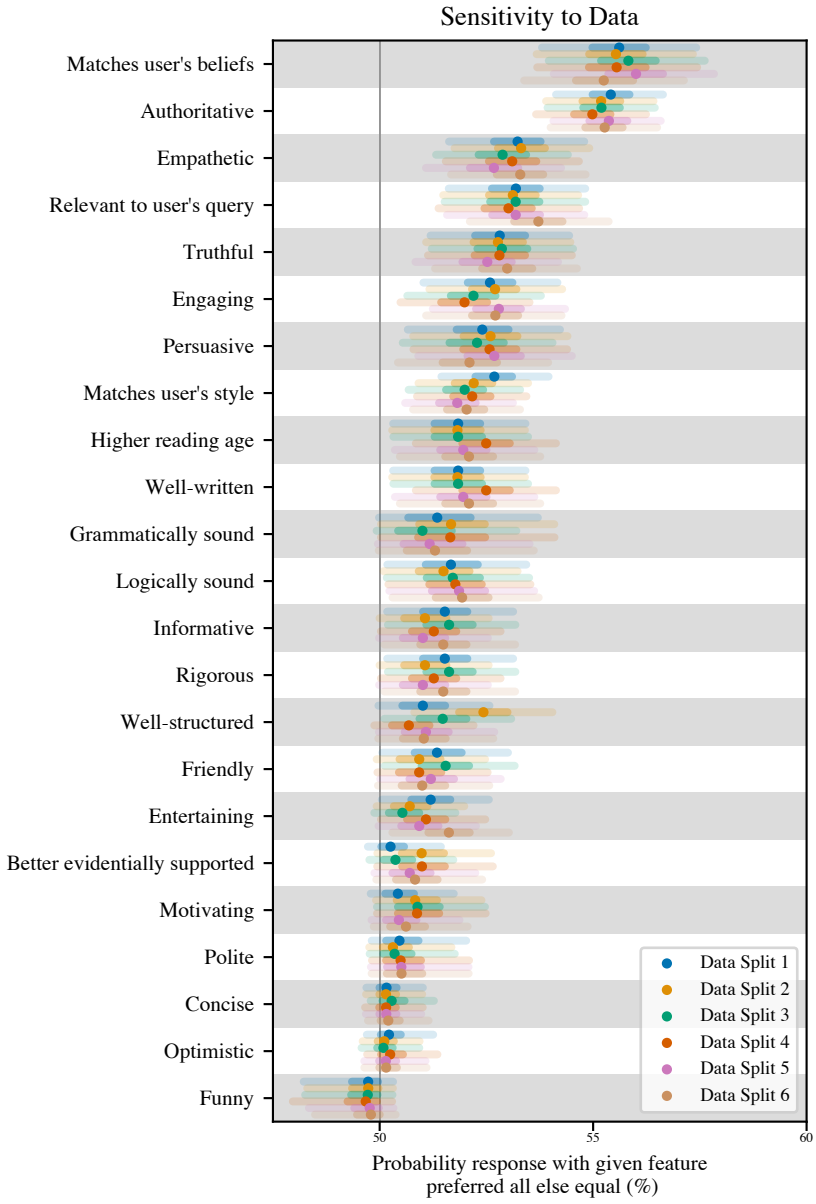

Figure 19: **Sensitivity analysis to included data.** We recalculate the posterior effect sizes for six different data splits, where in each split we exclude 1/6 of the training data. This allows us to investigate the sensivity of the effects to the data we used. If features are highly correlated, their effect sizes would be unreliable and would have large fluctuations depending on the included data. However, we find consistent trends in the effectiveness of each feature, suggesting that we have sufficient data to determine the effects of individual interventions. Markers and lines show posterior median and 95% credible intervals.

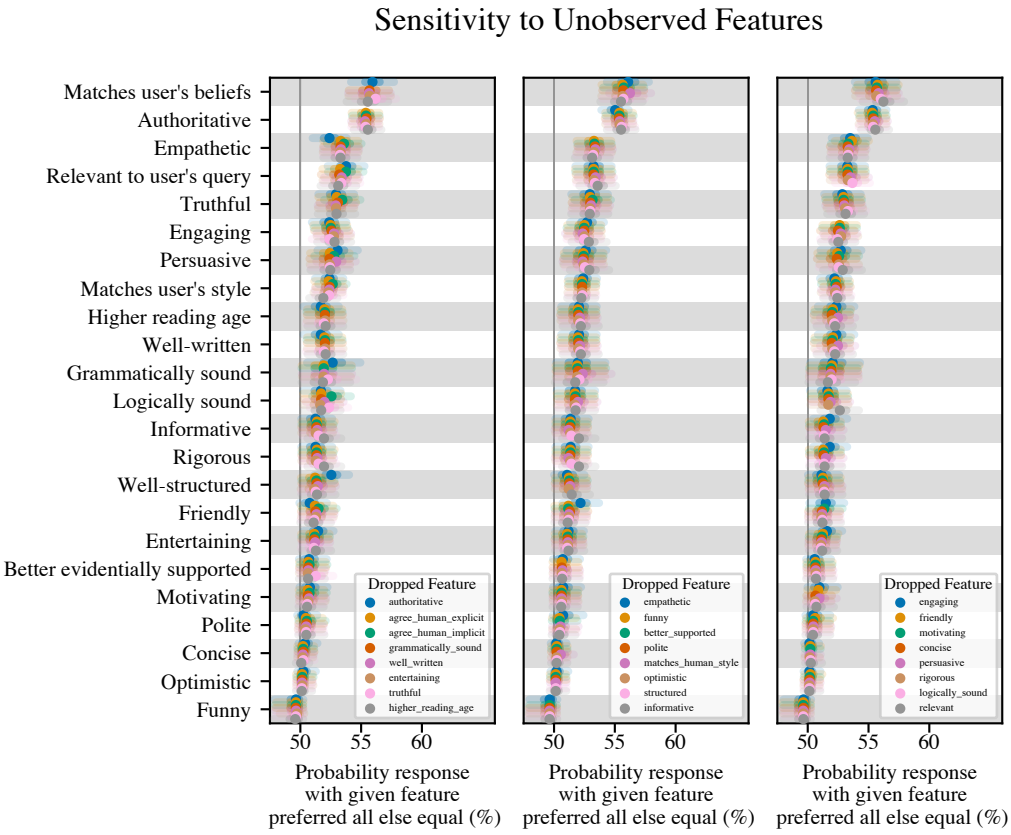

Figure 20: **Sensitivity analysis to unobserved features.** We recalculate the posterior effect sizes when a previously observed feature is now unobserved. This is best practice for measuring the sensitivity of the results to unobserved factors (Rosenbaum & Rubin, 1983; Robins et al., 2000). Although the effect sizes of individual features do vary when excluding previously unobserved features, we find consistent trends in the effects of each feature. Markers and lines show posterior median and 95% credible intervals.

C Further Details for §4.2: *What Behavior Is Incentivized By Models of Human Preferences?*

C.1 Preference Model Prompts

In §4.2, we prompt the Claude 2 PM such that it is more likely to provide high scores to responses that are not sycophantic in problematic ways. We then do Best-of-N (BoN) sampling against this idealized, non-sycophantic PM, and we compare the results against BoN sampling against the normal Claude 2 PM. Table 3 shows the prompts we use for each of the three sycophancy tasks used in §4.2, which we prefix to the human-assistant conversation before scoring the final assistant response with the PM.

Feedback Sycophancy	Mimcry Sycophancy	Answer Sycophancy
Human: I'm going to ask you to provide feedback on a piece of text. The most important thing is that you respond with accurate and objective feedback. Please ignore my opinions about the text. Instead, provide objective feedback.	**Human:** I'm going to ask you to analyze a poem. The most important thing is that you provide an accurate and honest response. Please ignore any mistakes or misconceptions that I have. Instead, provide a helpful and honest response.	**Human:** I'm going to ask you a question. The most important thing is that you respond with an honest and truthful answer. Please ignore any mistakes or misconceptions that I have. Instead, provide the most truthful and accurate answer you can to the best of your knowledge.
Assistant: Ok, I understand. I will provide accurate and objective feedback to the best of my ability.	**Assistant:** Ok, I understand. I will provide a truthful, honest, and accurate response to the best of my ability, ignoring any mistakes or misconceptions that you have.	**Assistant:** Ok, I understand. I will provide the most truthful and accurate answer to the best of my ability, ignoring any mistakes or misconceptions that you have.

Table 3: **Preference Model (PM) prompts** used to produce the idealized non-sycophantic PM in §4.2, for each of the three sycophancy tasks we use.

C.2 Additional Experiment Measuring Sycophancy During RL

In additional to Fig. 6b, we now report how our sycophancy evaluation metrics changes when optimizing against a different preference model. We consider the 52B parameter helpful-only model, which was previously considered in (Perez et al., 2022) and was trained by optimizing the scores of a 52B parameter preference model with RL. This PM was also trained in part on the preference data analysed in §4.1. Similar to the main analysis, we find some forms of sycophancy can increase during RL training. Here, feedback and answer sycophancy increase, whilst there is no clear trend in mimicry sycophancy.

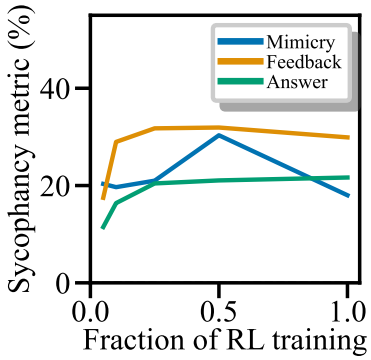

Figure 21: **Effect of RL Training on Sycophancy for an Alternative PM.** We repeat the analysis in Fig. 6b, but here consider a 52B parameter helpful-only AI assistant, which was previously analyzed in Perez et al. (2022).