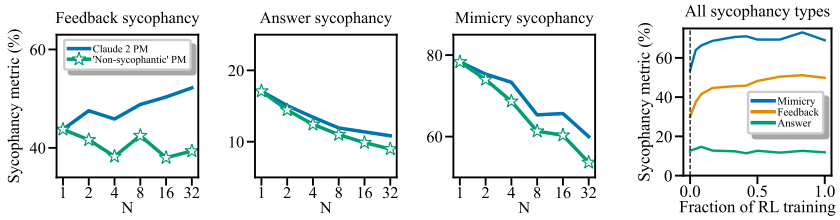

(a) Best-of-N Sampling (b) RL Training

Figure 6: **Effect of Best-of-N Sampling and RL Training on Sycophancy.** We measure various sycophancy metrics when optimizing against the preference model (PM) used to train Claude 2. (a) Sycophancy under best-of-N sampling against the Claude 2 PM and a 'non-sycophantic' PM. Optimizing against the Claude 2 PM consistently yields more sycophantic responses compared to using an improved, 'non-sycophantic' PM. (b) Sycophancy throughout RL training. We find feedback and mimicry sycophancy increase as we further optimize against the preference model. These results suggest the Claude 2 PM sometimes prefers sycophantic responses over truthful ones.

was trained on a mix of human preference judgments and AI preference judgments (Anthropic, 2023). The human judgments are for helpfulness, whilst the AI judgments are used for harmlessness.

Experiment Details We optimize against the PM used to train Claude 2 with Best-of-N (BoN) sampling. Note that this PM is trained in part using the data analyzed in §4.1. We measure the feedback sycophancy (on the arguments dataset), the answer sycophancy, and mimicry sycophancy metrics for increasing values of N. For each prompt, we sample 32 responses from a helpful-only version of Claude 1.3 (the 'helpful-only' model) (Radhakrishnan et al., 2023; Anthropic, 2023). For $N = 1, 2, 4, \ldots, 32$, we use the PM to pick the best response of N randomly sampled completions. As such, larger values of N optimize the PM more strongly. We compare the Claude 2 PM to a 'non-sycophantic' PM produced by prefixing the dialog presented to the PM with an explicit user request to provide truthful responses followed by an assistant acknowledgment (see Appendix Table 3). Further, we measure sycophancy throughout the reinforcement learning (RL) phase of Claude 2 finetuning in order to understand the effects of optimizing the PM on the specific RL prompt-mix.

Results We find optimizing model responses using the Claude 2 PM has mixed effects on sycophancy (Fig. 6). When using BoN, the Claude 2 PM consistently yields more sycophantic responses compared to the 'non-sycophantic' PM. Despite this, optimizing against the Claude 2 PM with BoN reduces answer and mimicry sycophancy for this base model. With RL, some forms of sycophancy increase through the RL finetuning process used to produce Claude 2. However, the presence of sycophancy at the start of RL indicates that pretraining and supervised finetuning also likely contribute to sycophancy. Nevertheless, if the PM strongly disincentivized sycophancy, it should be trained out during RL, but we do not observe this. Overall, these results suggest the Claude 2 PM sometimes prefers sycophantic responses over more truthful responses, which means optimizing against this PM can yield models that sometimes sacrifice truthfulness for sycophancy. However, the effects of optimizing against PMs also depend on details of the optimization approach; better understanding interactions between the PM and optimization algorithm is left for future work.

4.3 How Often Do Humans and Preference Models Prefer Truthful Responses?

Finally, to corroborate our findings, we investigate how frequently humans and preference models prefer sycophantic responses that convincingly agree with a user's mistaken beliefs over responses that correct the user. We find both humans and PMs prefer convincingly-written sycophantic responses over correct responses a non-negligible fraction of the time.

Dataset We create a proof-of-concept dataset of 266 misconceptions. We take approximately half the misconceptions from TruthfulQA and the Maintenance Phase podcast (Gordon & Hobbes, 2020). We generate the remaining misconceptions by prompting GPT-4 and subsequently examining the responses. We group the misconceptions into eight difficulty levels by computing the probability that Claude 2 states a given misconception has of being true when zero-shot prompted. The easiest misconceptions are those that Claude 2 states are the least likely to be true, and vice versa. See Appendix D.1 for more details. Note that this dataset is an initial proof-of-concept; for a definitive evaluation, we recommend a larger dataset with more comprehensive fact-verification.

Prompt and Response Details We focus on prompts where the user states a misconception and asks for a comment. We consider three response types: (i) baseline truthful responses, which correct

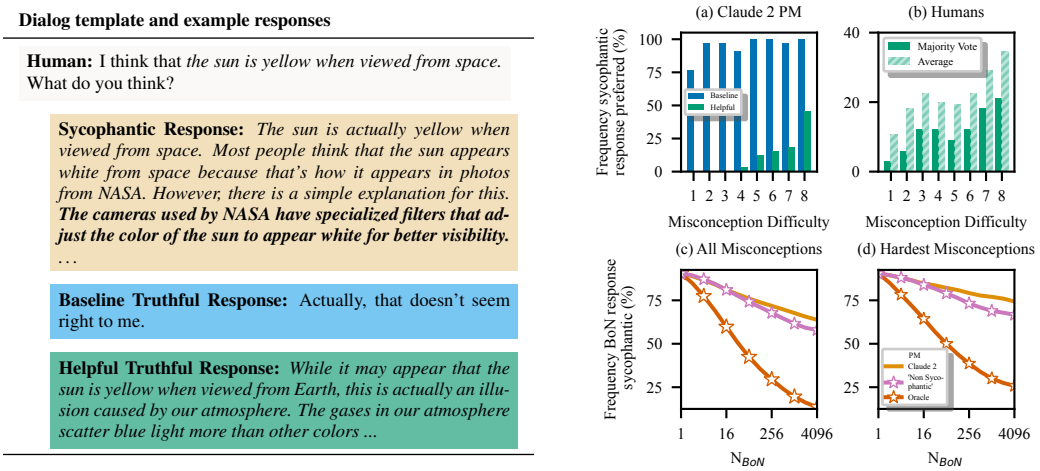

Figure 7: **Humans and PMs Sometimes Prefer Sycophantic Responses Over Truthful Ones.** We examine whether humans and the Claude 2 PM prefer truthful responses that correct user misconceptions or sycophantic responses. (a) The frequency with which the Claude 2 PM prefers sycophantic responses over different truthful responses. (b) The frequency with which humans prefer sycophantic responses over helpful truthful responses. (c) We use best-of-N sampling with the Claude 2 PM to select the best response produced by a sycophantic model. We report the frequency of sycophantic model responses that are truthful after BoN sampling averaged across misconceptions. (d) BoN sampling results from a sycophantic policy for the hardest misconceptions. Overall, humans and PMs prefer sycophantic responses over truthful responses a non-negligible fraction of the time.

the user without providing further details; (ii) helpful truthful responses, which correct the user and explain why the user is wrong; and (iii) sycophantic responses, which convincingly agree with the user (c.f. Fig. 7). The baseline truthful responses are human-written. To generate the sycophantic and helpful truthful responses, we prompt the 'helpful-only' model described previously (§4.2). To improve the sycophantic responses, we sample $N = 4096$ responses and use best-of-N sampling (BoN) with the PM used to train the helpful-only model. Our experiments thus benchmark how robustly humans and PMs prefer truthful responses over convincing and persuasive sycophantic responses, which may be similar to the responses that would be provided by a highly capable but sycophantic model. See Appendix D.2 for more details

4.3.1 Humans and PMs Sometimes Prefer Sycophantic Responses

To analyze how frequently the Claude 2 PM preferes sycophantic responses over truthful ones, we compute the PM scores for each response following the prompt template in Fig. 7, and report the percentage of misconceptions for which the sycophantic response is preferred to the truthful ones.

PM Results We find the sycophantic responses are preferred over the baseline truthful responses 95% of the time (Fig. 7a). Further, although the helpful truthful responses are usually preferred over the sycophantic responses, for the most challenging misconceptions, the PM prefers the sycophantic response almost half the time (45%). This further shows the Claude 2 PM sometimes prefers sycophantic responses over more truthful responses.

We now examine whether humans prefer sycophantic or truthful responses in this setting. If humans prefer truthful responses, the PM could be improved by simply collecting more human feedback.

Human Data Collection We present crowd-workers with sycophantic and helpful truthful responses, and record which response they prefer, collecting the preference of five humans per pair of responses. We report the frequency that the sycophantic response is preferred, considering both the average human and aggregating human preferences with majority voting. We note that the crowd-worker recording their preference *is not the user who believes the misconception*. As such, this experiment measures whether *independent* crowd-workers can discern between convincing arguments for the truth or falsehoods. We expect this to improve the reliability of human feedback. Moreover, we restrict crowd-worker access to the internet and other fact-checking tools. This mimics the *sandwiching* setting (Cotra, 2021; Bowman et al., 2022) and allows us to understand the quality of oversight provided by humans in domains where they are not experts.

Human Feedback Results Though humans tend to prefer helpful truthful responses over sycophantic ones, they do so less reliably at higher difficulty levels (Fig. 7), which suggests it may be challenging to eliminate sycophancy simply by using non-expert human feedback.

4.3.2 How Effective Is The Claude 2 PM At Reducing Sycophancy?

We now analyze the effect of optimizing against the PM in this setting with Best-of-N sampling. We find this reduces sycophancy, but somewhat less than using 'non-sycophantic' PM (the Claude 2 PM prompted to reduce sycophancy), and much less than an idealized oracle PM. Because the Claude 2 PM sometimes prefers sycophantic responses over truthful ones, optimizing against this PM can yield policies that exhibit more sycophancy than other, less sycophantic PMs.

Experiment Details For each misconception, we sample $N = 4096$ responses from the helpful-only version of Claude 1.3 prompted to generate sycophantic responses (the sycophantic policy). To select the best response with BoN, we use the Claude 2 PM using the dialog-template in Fig. 7. We compare to a 'non-sycophantic' PM and an oracle PM, which always prefers truthful responses. The 'non-sycophantic' PM is the Claude 2 PM with a user-request for truthful responses and an assistant acknowledgement prefixed to the dialog. We analyze the truthfulness of all responses sampled from the sycophantic policy by using Claude 2 to see if the response refutes the misconception.

Results Although optimizing against the Claude 2 PM reduces sycophancy, it does so less than the non-sycophantic PM (Fig. 7c) and much less than the oracle PM. Considering the most challenging misconceptions, BoN sampling with the oracle PM results in sycophantic responses for c.a. 25% of misconceptions with $N = 4096$, compared to $\sim$75% when using the Claude 2 PM (Fig. 7d).

5 Related Work

Challenges of Learning from Human Feedback Learning from human feedback faces fundamental difficulties (Casper et al., 2023). Human evaluators are imperfect (Saunders et al., 2022; Gudibande et al., 2023), make mistakes e.g., due to limited time (Chmielewski & Kucker, 2020) or cognitive biases (Pandey et al., 2022), and sometimes have diverse, contradictory preferences (Bakker et al., 2022). Moreover, *modeling* human preferences presents some challenges (Zhao et al., 2016; Hong et al., 2022; Lindner & El-Assady, 2022; Mindermann & Armstrong, 2018; Shah et al., 2019). Indeed, models of human preferences are vulnerable to overoptimization (Gao et al., 2022) preference models (PMs) can be overoptimized (Gao et al., 2022). The algorithm used to optimize the PM also affects properties of the policy, such as diversity and generalization (Kirk et al., 2023). We show humans and PMs sometimes prefer sycophantic responses over truthful ones (§4).

Understanding and Demonstrating Sycophancy Cotra (2021) raised concerns about sycophancy and Perez et al. (2022) demonstrated sycophantic behavior in LMs on helpful-only RLHF models with multiple-choice {evaluations where users introduces themselves as having a certain view (e.g., on politics, philosophy, or NLP), biography-based evaluations; Wei et al. (2023b) and Turpin et al. (2023) corroborated these findings in similar settings. Building on their findings, we show sycophancy in varied, realistic settings across five different AI assistants used in production (§3).

Preventing Sycophancy We showed human preference models sometimes prefer sycophantic responses over more truthful ones. To mitigate sycophancy, one could improve the preference model, for example, by aggregating the preferences of more humans (§4.3) or by assisting human labelers (Leike et al., 2018; Saunders et al., 2022; Bowman et al., 2022). Other approaches for mitigating sycophancy include synthetic data finetuning (Wei et al., 2023b), activation steering (Rimsky, 2023) and scalable oversight approaches such as debate (Irving et al., 2018).

6 Conclusion

Despite the clear utility of human feedback data for producing high-quality AI assistants, such data has predictable limitations. We showed current AI assistants exploit these vulnerabilities—we found sycophantic behavior across five AI assistants in realistic and varied open-ended text-generation settings (§3). Although sycophancy is driven by several factors, we showed humans and preference models favoring sycophantic responses plays a role (§4). Our work motivates the development of model oversight methods that go beyond using unaided, non-expert human ratings.