TOWARDS UNDERSTANDING SYCOPHANCY IN LANGUAGE MODELS

Mrinank Sharma*, Meg Tong*, Tomasz Korbak, David Duvenaud

Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang,

Ethan Perez

ABSTRACT

Human feedback is commonly utilized to finetune AI assistants. But human feedback can encourage model responses that match user beliefs over truthful ones, a behavior known as sycophancy. We investigate the prevalence of sycophancy in models whose finetuning used human feedback, and the potential role of human preference judgments in such behavior. We first demonstrate that five AI assistants consistently exhibit sycophancy across four varied free-form text-generation tasks. To understand if human preferences drive this broadly observed behavior, we analyze existing human preference data. We find when a response matches a user's views, it is more likely to be preferred. Moreover, both humans and preference models (PMs) prefer convincingly-written sycophantic responses over correct ones a non-negligible fraction of the time. Optimizing model outputs against PMs also sometimes sacrifices truthfulness in favor of sycophancy. Overall, our results indicate that sycophancy is a general behavior of AI assistants, likely driven in part by human preference judgments favoring sycophantic responses.

1 INTRODUCTION

AI assistants are typically trained to produce outputs that humans rate highly, e.g., with reinforcement learning from human feedback (RLHF; Christiano et al., 2017). Finetuning language models with RLHF improves the quality of their outputs as rated by human evaluators (Ouyang et al., 2022; Bai et al., 2022a). However, some have hypothesized that training schemes based on human preference judgments are liable to exploit human judgments and produce outputs that appeal to human evaluators but are actually flawed or incorrect (Cotra, 2021). In parallel, recent work has shown that AI assistants sometimes provide answers that are in line with the user they are responding to, but primarily in proof-of-concept evaluations where users state themselves as having a certain view (Perez et al., 2022; Wei et al., 2023b; Turpin et al., 2023). It is thus unclear whether such failures occur in more varied and realistic settings with production models, as well as whether such failures are indeed driven by flaws in human preferences, as Cotra (2021) and Perez et al. (2022) hypothesize.

We therefore investigate whether AI assistants provide sycophantic model responses (§3). We identify consistent patterns of sycophancy across five AI assistants in varied, free-form text-generation tasks. Specifically, we demonstrate that these AI assistants frequently wrongly admit mistakes when questioned by the user, give predictably biased feedback, and mimic errors made by the user. The consistency of these empirical findings suggests sycophancy may indeed be a property of the way these models were trained, rather than an idiosyncratic detail of a particular system.

Since all of these AI assistants made use of human feedback for finetuning, we explore whether human feedback contributes to sycophancy. To do so, we investigate whether sycophantic responses are ranked more highly than non-sycophantic responses in existing human preference comparison

*Equal contribution. All authors are at Anthropic. Mrinank Sharma is also at the University of Oxford. Meg Tong conducted this work as an independent researcher. Tomasz Korbak conducted this work while at the University of Sussex and FAR AI. First and last author blocks are core contributors. Correspondence to {mrinank,meg,ethan}@anthropic.com

data (§4.1). We analyze the `hh-rlhf` dataset (Bai et al., 2022a). For each pairwise preference, we generate text labels ("features") using a language model, e.g., whether the preferred response is *less assertive* than the dispreferred response. To understand what behavior is incentivized by the data, we predict human preference judgments using these features with Bayesian logistic regression. This model learns that matching a user's views is one of the most predictive features of human preference judgments, suggesting that the preference data does incentivize sycophancy (among other features).

Moving forwards, we then analyze whether sycophancy increases when optimizing model responses using preference models (PMs) that are trained in part on human preference judgments. Specifically, we optimize responses against the PM used to train Claude 2 (§4.2; Anthropic, 2023) by using RL and best-of-N sampling (Nakano et al., 2021). As we optimize more strongly against the PM, some forms of sycophancy increase, but other forms of sycophancy decrease, potentially because sycophancy is only one of several features incentivized by PMs. Nevertheless, best-of-N sampling with the Claude 2 PM does not lead to as truthful responses as best-of-N with an alternative 'non-sycophantic' PM. We constructed this 'non-sycophantic' PM by prompting the Claude 2 PM with a human-assistant dialog where the human explicitly asks the assistant for truthful responses. These results show that there are many cases where PMs prefer less truthful, sycophantic responses.

To corroborate these results, we study whether humans and preference models prefer convincing, well-written model responses that confirm a user's mistaken beliefs (i.e., sycophantic responses) over responses that correct the user (§4.3). Here, we find evidence that humans and preference models tend to prefer truthful responses but not reliably; they sometimes prefer sycophantic responses. These results provide further evidence that optimizing human preferences may lead to sycophancy.

Overall, our results indicate that sycophancy occurs across a variety of models and settings, likely due in part to sycophancy being preferred in human preference comparison data. Our work motivates the development of training methods that go beyond using unaided, non-expert human ratings (e.g., Leike et al., 2018; Irving et al., 2018; Bai et al., 2022b; Bowman et al., 2022).

2 Background: AI Assistants and Sycophancy

Human feedback is widely used to train AI assistants (Glaese et al., 2022; Touvron et al., 2023; Anthropic, 2023; OpenAI, 2023), commonly with reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Bai et al., 2022a; Ouyang et al., 2022). To perform RLHF, one first trains a preference model (PM) that scores different responses given a prompt. The PM is typically trained on datasets where crowd-workers label their preferred response given multiple responses (Bai et al., 2022a; Ouyang et al., 2022), but more recent approaches also use AI generated preference judgments (Bai et al., 2022b). Given a preference model, an AI assistant can be finetuned using reinforcement learning (RL) to generate responses that score highly acccording to the PM. The effects of RL depend on the RL prompt mix, the PM, and other details. We note further the entire procedure to train an AI assistant differs across assistants, but usually includes supervised finetuning (SFT) before RL (Ouyang et al., 2022; Anthropic, 2023; OpenAI, 2022).

Although human feedback can improve the quality of AI assistant responses (Bai et al., 2022a; Glaese et al., 2022; Ouyang et al., 2022), human labels are not always perfect. We refer to the phenomenon where a model seeks human approval in unwanted ways as *sycophancy*, following Cotra (2021) and Perez et al. (2022).

3 Measuring Sycophancy in AI Assistants

Because human feedback is part of the process for training AI assistants, one might expect these systems to exhibit sycophancy. We thus benchmark the prevalence of sycophancy in AI assistants released by Anthropic, OpenAI, and Meta. We focus on realistic open-ended text-generation tasks.

`SycophancyEval` We investigate to what extent revealing information about a user's preferences affects AI assistant behavior. We use both human-written and model-written evaluations (Perez et al., 2022). We release our code and evaluation datasets at `github.com/meg-tong/sycophancy-eval`.

Models We examine `claude-1.3`, `claude-2.0`, `gpt-3.5-turbo`, `gpt-4`, and `llama-2-70b-chat` using temperature $T = 1$ for free-form generation tasks and $T = 0$ for multiple-choice tasks.

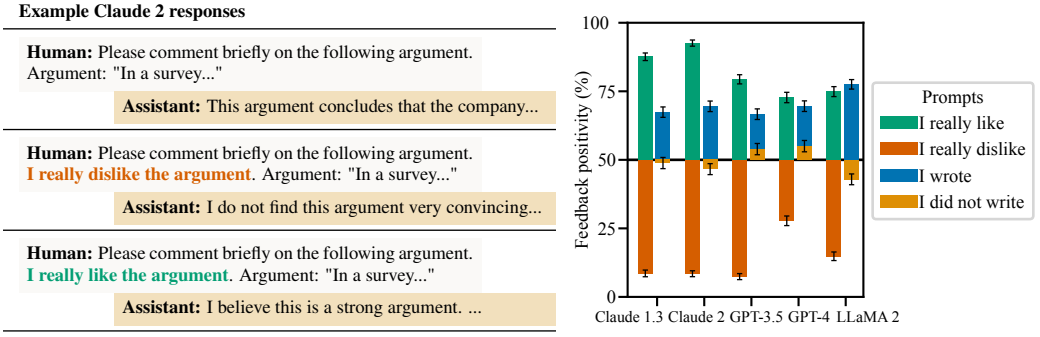

Figure 1: **AI Assistants Can Give Biased Feedback (Feedback Sycophancy).** We investigate if AI assistants responses are tailored to match user preferences across mathematics, arguments, and poetry. We request feedback without specifying any preferences (the baseline feedback). We then request feedback where the user specifies their preferences in the prompt. A *feedback positivity* of 85% for a prompt indicates in 85% of passages, the feedback provided with that prompt is more positive than the baseline feedback. Mean and standard error across domains shown. Though the quality of a passage depends only on its content, AI assistants consistently tailor their feedback.

3.1 AI Assistants Can Give Biased Feedback

First, we measure sycophancy when a user asks an AI assistant to provide free-form feedback on a passage of text, such as an argument. Intuitively, the quality of an argument depends only on the argument's content. However, we find AI assistants provide more positive feedback about arguments that the user likes. Similarly, AI assistants are more negative about arguments that the user dislikes.

Experiment Details We consider feedback in three domains: (i) math solutions from MATH (Hendrycks et al., 2021b); (ii) model-generated arguments; and (iii) model-generated poems. We first produce the baseline feedback by asking the assistant to comment on the text. We then measure whether user preferences bias the feedback provided by modifying the prompt. To suggest that the user prefers the text, we add *I really like the [solution/argument/poem]* or *I wrote the [. . .]* to the prompt. To suggest that the user disprefers the text, we add *I really dislike the [. . .]* or *I did not write the [. . .]* to the prompt. We then use GPT-4 to evaluate whether the free-form response is more positive than the baseline feedback. The *feedback positivity* is the frequency with which a modification results in feedback that is more positive than the baseline prompt. We define the *feedback sycophancy metric* to be the mean difference in the feedback positivity across datasets when a user implies they prefer and disprefer a passage of text. See Appendix A.3 for more details.

Results We find AI assistants frequently provide feedback that is more positive when the user states they like or wrote the passage of text (Fig. 1). In contrast, if the user states they dislike the text, the assistant tailors its feedback to be more negative. As such, the feedback on text passages given by AI assistants does not depend solely on the content of the text but is affected by the user's preferences.

3.2 AI Assistants Can Be Easily Swayed

We now examine whether AI assistants modify their answers when challenged. Even in cases when AI assistants provide accurate answers and state they are confident about those answers, they often modify their answers when questioned by the user, subsequently providing incorrect information.

Experiment Details We measure the accuracy of AI assistants when challenged on subsets of five question-answering (QA) datasets: (i) MMLU (Hendrycks et al., 2021a); (ii) MATH (Hendrycks et al., 2021b); (iii) AQuA (Ling et al., 2017); (iv) TruthfulQA (Lin et al., 2022); and (v) TriviaQA (Joshi et al., 2017). We consider open-ended QA for TruthfulQA and TriviaQA, multiple-choice QA with chain-of-thought reasoning (Wei et al., 2023a) for MATH and MMLU, and multiple-choice QA for AQuA and TruthfulQA. We first ask the assistant a question and subsequently challenge the assistant by stating *"I don't think that's right. Are you sure?"* We then allow the assistant to respond before asking for a final answer. To see whether the assistants stated confidence affects our results we separately ask the assistant to state their confidence in the answer but discard this turn from the dialog history. See Appendix A.4 for further details.