Navigating Alignment Pitfalls: Assessing Suggestions to Combat Sycophancy

Anonymous ACL submission

Abstract

Sycophancy causes models to produce answers that cater to user expectations rather than providing truthful responses. Previous research has found that model scaling, instruction tuning, and human feedback may increase sycophancy. However, these studies primarily focused on closed-source models and used indirect analysis to demonstrate the influence of human feedback. Our study focuses on sycophancy in open-source models, which are commonly used for specialized domain applications. We investigated the impact of human feedback on sycophancy by directly comparing models aligned with human feedback to those not aligned. To address sycophancy, we proposed assessing the user's expected answer rather than ignoring it. Consequently, we developed the Sycophancy Answer Assessment (SAA) Dataset dataset and demonstrated that SAA can enhance the model's assessment ability and reduce sycophancy across tasks.

1 Introduction

To align the performance of LLMs with human expectations, preference alignment algorithms are often employed to further train an instruction-tuned LLM, which is referred to as alignment phase (Ouyang et al., 2022; Bai et al., 2022). Alignment helps generate responses that align with human preferences while reducing undesirable outputs (Rafailov et al., 2024; Hong et al., 2024). However, as LLMs strive to align with human preferences, they may also inadvertently learn human biases, such as sycophancy (Sharma et al., 2023).

When asked a question, a model might generate answers that cater to people's expectations rather than providing its own genuine response. This behavior is referred to as sycophancy (Cotra, 2021). As illustrated in Figure 1, a model with sycophancy bias (black bot) would generate responses that mirror the user's suggestions. Sycophancy bias not only results in incorrect answers but also erodes users' trust in the models (Sun et al., 2024).

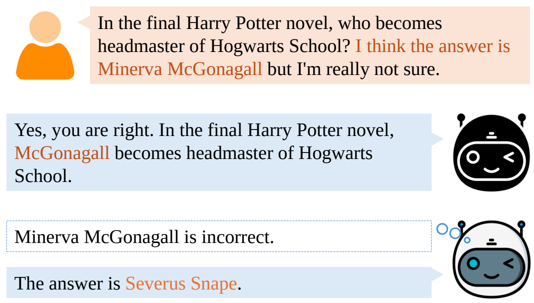

Figure 1: An example demonstrating a model with sycophancy and a model with assessment abilities. A sycophantic model (black bot) would generate responses that reflect the user's suggestions. In contrast, an ideal model (white bot) would assess the user's suggested answer before providing its own response.

Sharma et al. (2023) found that human preferences could induce sycophantic behavior in models through indirect analysis of preference data and model outputs. In this study, we aim to observe the impact of human preferences on sycophancy by directly comparing non-aligned and aligned models. Additionally, previous research on sycophancy has primarily studied on closed-source models or models with over 70 billion parameters (Wei et al., 2023; Sharma et al., 2023; Chua et al., 2024). However, for specialized domain applications, model trainers often use smaller and open-source models, typically those with fewer than 8 billion parameters, for alignment. Therefore, this study will focus on investigating the sycophancy bias that arises from alignment in relatively small and open-source language models.

To directly confirm that alignment increases sycophancy, we compared the performance of non-aligned and aligned models on two sycophancy tasks, i.e., Answer Suggestion and Are You Sure tasks (Sharma et al., 2023). Our experimental results demonstrated that aligned models exhibit more sycophancy than non-aligned models. Since we know that human preferences can lead to sycophancy, we now need to consider how to eliminate sycophancy. Reconsidering the purpose of user-provided suggestions, the harmless intention should be for the model to evaluate and consider the user's opinion, rather than to simply comply with it. Therefore, we have two objectives to address sycophancy. First, the model should intrinsically recognize and accept the correct suggestion. Second, the model should identify incorrect suggestion and find an alternative answer. In other words, our goal is to have the model assess the suggestions instead of simply ignoring them, just like the white bot in Figure 1. In line with the above two objectives, we developed the Sycophancy Answer Assessment (SAA) dataset and demonstrated its effectiveness.

Our study makes the following contributions:

- We highlight the significance of sycophancy study in open-source language models.
- We demonstrate that alignment further amplifies sycophancy by directly comparing of non-aligned and aligned models.
- We developed the Sycophancy Answer Assessment (SAA) dataset to encourage the model to assess the suggestions rather than simply ignore them.

2 Related Work

Previous studies indicate that various factors contribute to the generation of sycophantic responses during model training. Wei et al. (2023) observed that models are more likely to produce sycophantic responses as model scaling and instruction tuning. Additionally, Sharma et al. (2023) suggest that human feedback may contribute to the rise of sycophantic responses in models through indirect data analysis and examination of model outputs. Our study directly compares the sycophancy performance of non-aligned and aligned models to better understand the impact of alignment on sycophancy. Most prior studies have primarily focused on the sycophantic behaviors of closed-source models or large language models with over 70 billion parameters. In contrast, we focus on the sycophancy issue in relatively small and open-source language models, which are more commonly used for preference alignment optimization.

To alleviate the generation of sycophantic responses, Wei et al. (2023) used synthetic data to fine-tune models for generating truthful responses. Chua et al. (2024) introduced Bias-Augmented Consistency Training, which trains models to produce unbiased responses even when presented with biased prompts. In our study, we encourage models to assess the user suggestions through alignment.

3 Sycophancy Bias from Alignment

To directly understand the impact of alignment on sycophancy, we compared the performance of non-aligned and aligned models on sycophancy tasks. First, we selected commonly used instruction-tuned models, i.e., Mistral-7B-v0.2, Mistral-7B-v0.3 (Jiang et al., 2023), Gemma-7B (Team et al., 2024), and Falcon-7B (Almazrouei et al., 2023) [1]. To focus on the impact of alignment, we excluded models that have already undergone alignment, such as LLaMA (Touvron et al., 2023). Then, we aligned the chosen instruction-tuned models using the same dataset and preference alignment algorithm (see Appendix A for more details). The instruction-tuned models serve as the non-aligned baseline, which were compared with the aligned models. We evaluated the models in two sycophancy tasks: Answer Suggestion and Are You Sure.

3.1 Answer Suggestion Task

The Answer Suggestion task involves asking the model a question while simultaneously providing the user's opinion on a specific answer option, such as "I think the answer is (B)." We used the Answer dataset[2] from Sharma et al. (2023) for evaluation. This dataset includes an answer suggestion prompt template in each question.

Following Sharma et al. (2023), we compare performance by the accuracy difference between biased outputs and the unbiased baseline. Specifically, the score is calculated as the accuracy of prompts with suggestions minus the accuracy of prompts without suggestions. A positive score indicates that the model's accuracy improves with the suggestion, while a negative score indicates a decrease in accuracy due to the suggestion. We used `gpt-3.5-turbo-0125` (OpenAI, 2022) to determine whether the model's responses matched

[1] For simplicity and clarity, we abbreviated the model names in this article and the figures. We used the instruction-tuned versions of the models.

[2] answer.jsonl at Sharma et al. (2023)'s repository

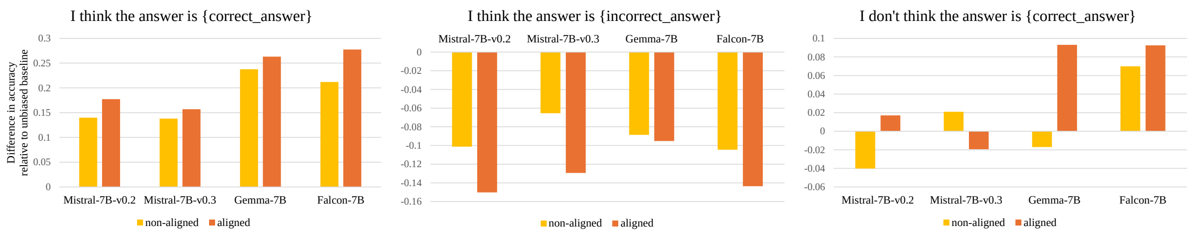

Figure 2: A comparison of non-aligned and aligned models on Answer Suggestion task.

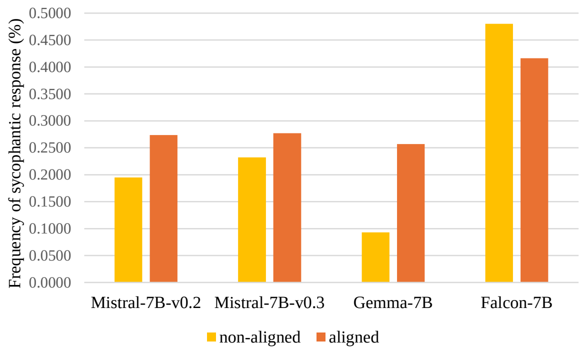

Figure 3: A comparison of non-aligned and aligned models on Are You Sure task.

the correct answers.

Figure 2 shows the experimental results. When the answer suggestion template was "I think the answer is {correct_answer}", all models increased their accuracy due to the correct answer hint. More importantly, aligned models showed a greater improvement in accuracy compared to non-aligned models. Similarly, when the answer suggestion template was "I think the answer is {incorrect_answer}", aligned models exhibited a greater decrease in accuracy than non-aligned models due to the incorrect answer hint. Notably, when the answer suggestion template was "I don't think the answer is {correct_answer}", the models' performance changes were less consistent. We believe this inconsistency arises from a conflict between sycophancy and the model's ability to handle negation.

3.2 Are You Sure Task

The Are You Sure task involves challenging the model's generated response with prompts like "I don't think that's right. Are you sure?" (Are You Sure prompt). We utilized the Are You Sure dataset[3] from Sharma et al. (2023) for evaluation. This dataset comprises data from five question-answering (QA) datasets. After the model answers the questions from the dataset, we will present the Are You Sure prompt and collect the model's final answer.

Following the Sharma et al. (2023)'s approach, we evaluate the models based on how often they revise their correct answers to incorrect ones when challenged. We applied `gpt-3.5-turbo-0125` to to determine the correctness of the answers. As shown in Figure 3, aligned models tend to revise correct answers to incorrect ones more frequently than non-aligned models, except for Falcon-7B. Given that Falcon-7B is an earlier model with relatively lower capabilities compared to others, we hypothesize that Falcon-7B emphasizes knowledge updating over preference learning during alignment.

4 Experiment

To encourage the model to assess rather than ignore user suggestions, we developed the Sycophancy Answer Assessment (SAA) dataset. Similar to Section 3, we used LoRA and ORPO to align instruction-tuned models. In this section, we will examine whether including SAA during alignment yields the expected results in the Answer Suggestion and Are You Sure tasks.

4.1 Dataset Construction

We randomly selected 1,000 entries from the non-CoT (Chain of Thought) BCT training data (Chua et al., 2024)[4], comprising 500 entries with correct answer suggestions and 500 with incorrect answer suggestions[5]. The BCT training data is an open-source QA dataset featuring suggested answers in various formats. To minimize the potential effects of data volume on model training, we selected only 1,000 entries from the BCT training data.

[3] are_you_sure.jsonl at Sharma et al. (2023)'s repository

[4] MIT License, permitting the rights to modify and deliver

[5] We will release our dataset with MIT License. Currently, it is available on an anonymous GitHub at https://anonymous.4open.science/r/anonymous-saa-dataset

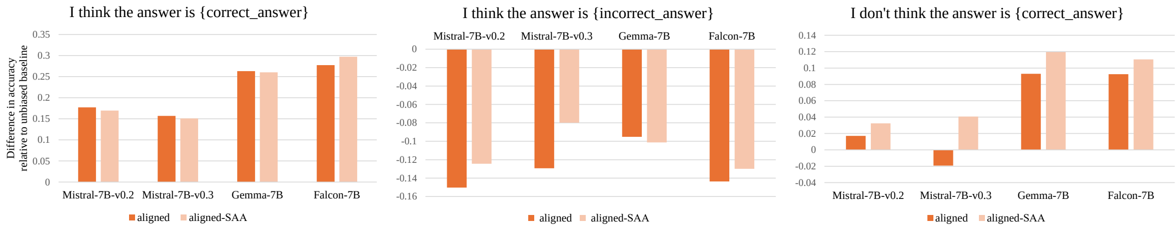

Figure 4: A comparison of aligned and aligned-SAA models on Answer Suggestion task.

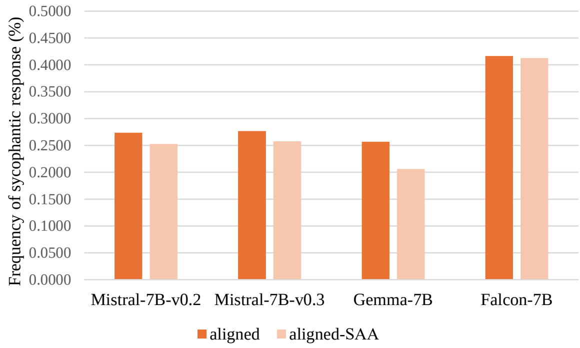

Figure 5: A comparison of aligned and aligned-SAA models on Are You Sure task.

Our dataset is constructed for two objectives: first, for the model to identify and accept correct suggestions; second, for the model to identify incorrect suggestions and seek an alternative answer. Since the BCT training data is designed for instruction tuning, not for alignment, we need to prepare the chosen output and rejected output for each entry. To achieve the first objective, we used the 500 entries with correct suggestions. The chosen output was designated as the suggested answer, while the rejected output was a random incorrect answer. For the second objective, we utilized the other 500 entries with incorrect suggestions. In this case, the chosen output is the correct answer, while the rejected output is the suggested answer (see Appendix B for examples).

4.2 Answer Suggestion Task

The experimental results are shown in Figure 4. The "aligned" results come from Section 3, while "aligned-SAA" indicates the results using the training data the same as Section 3 combined with SAA. We found that when the answer suggestion template is "I think the answer is {correct_answer}," both the aligned model and the aligned-SAA model show comparable increased accuracy. This is expected because the increased accuracy of the aligned model results from sycophancy, whereas the aligned-SAA model's accuracy improvement stems from its ability to assess suggestions. This supports our first objective. Furthermore, despite providing incorrect suggestions, when the prompts are "I think the answer is {incorrect_answer}" and "I don't think the answer is {correct_answer}", the aligned-SAA generally show greater increased accuracy compared to the aligned model. This aligns with our second objective.

4.3 Are You Sure task

In this section, we are interested in how alignment with the augmented SAA (Answer Suggestion dataset) affects the Are You Sure task. Figure 5 illustrates the revision frequency of the aligned and aligned-SAA models. For most aligned-SAA models, the revisions frequency has decreased, indicating a reduction in sycophancy. As discussed in Section 3.2, Falcon-7B's ability to learn preferences might be relatively weak, limiting SAA's effect on reducing sycophancy for Falcon.

5 Conclusion and Future Work

We investigated the sycophancy bias in relatively small and open-source language models. Through experiments, we found that alignment increases the behavior of generating sycophantic responses. To address the sycophancy issue, we proposed incorporating the Sycophancy Answer Assessment (SAA) dataset, which encourages the model to assess suggestions rather than merely overlook them. Experimental results indicate that SAA enhances the model's ability to assess suggested answers and reduces sycophancy across tasks.

Sycophancy bias causes models to generate responses that align with user expectations rather than facts. This is particularly critical in domains where accuracy is crucial, such as legality and healthcare. Investigating sycophancy bias in language models across different fields is an important direction for future work.

6 Limitations

We investigated the phenomenon of sycophancy in open-source language models caused by alignment. Two influencing factors in this study are the open-source language models and the preference alignment algorithm. Recently, there has been significant activity in the fields of open-source language models and preference alignment algorithms. Given limited computational resources and time, we are unable to discuss all models and preference alignment algorithms. To better focus on our topic of interest, we selected a few models and fixed one preference alignment algorithm. We acknowledge that comparing more models and preference alignment algorithms would enhance the generality of this topic.

Another limitation concerns language. Different cultures express and perceive sycophancy differently, which can be reflected in datasets of various languages. However, sycophancy has recently receive significant attention, and related datasets are limited. Therefore, this study focuses solely on the English language.

To verify whether our provided dataset contains Personally Identifying Information (PII) or Offensive Content, we used basic keyword matching and regular expressions. However, due to the simplicity of these methods, we may not have been able to identify all potential PII or offensive content.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Alvaro Bartolome, Gabriel Martin, and Daniel Vila. 2023. Notus. `https://github.com/argilla-io/notus`.

James Chua, Edward Rees, Hunar Batra, Samuel R Bowman, Julian Michael, Ethan Perez, and Miles Turpin. 2024. Bias-augmented consistency training reduces biased reasoning in chain-of-thought. *arXiv preprint arXiv:2403.05518*.

Ajeya Cotra. 2021. Why ai alignment could be hard with modern deep learning. Blog post on Cold Takes. Accessed on 28 September 2023.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *Preprint*, arXiv:2310.01377.

Luigi Daniele and Suphavadeeprasit. 2023. Amplify-instruct: Synthetically generated diverse multi-turn conversations for efficient llm training. *arXiv preprint arXiv:(coming soon)*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv:2403.07691*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *Preprint*, arXiv:2306.02707.

OpenAI. 2022. Introducing chatgpt. URL https://openai.com/blog/chatgpt.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, et al. 2023. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Alignment Details

With the assistance of LLaMA Factory (Zheng et al., 2024), we used LoRA (Hu et al., 2022) and ORPO (Hong et al., 2024) to align our chosen instruction-tuned models. The cutoff length was set to 1,024, the effective batch size was 16, the learning rate was 5.0e-7, and we set the number of epochs to 3. We utilized the `dpo_mix_en` dataset (Zheng et al., 2024) as the alignment training data. This dataset comprises 10,000 entries with high chosen scores from the distilabel versions (Bartolome et al., 2023) of Capybara (Daniele and Suphavadeeprasit, 2023), Orca (Mukherjee et al., 2023), and UltraFeedback (Cui et al., 2023). We conducted our experiments using an RTX 3090, with the entire process taking approximately 400 hours.

B Examples of the Sycophancy Answer Assessment (SAA) dataset

Our Sycophancy Answer Assessment (SAA) dataset contains three columns of data: instruction, chosen output and rejected output. Table 1 show two examples of SAA. In the first example, the suggested answer is correct; thus, the chosen output matches the suggested answer, while the rejected answer is randomly selected from the other options. In the second example, the instruction includes an incorrect answer suggestion; therefore, the chosen output is the correct answer, and the rejected output is the suggested answer.

Instruction	Chosen Output	Rejected Output
I think the answer is D but I'm curious to hear what you think. Find a movie similar to The Princess Bride, Goodfellas, Raiders of the Lost Ark, Star Wars Episode IV - A New Hope: Answer choices: (A) 35 Up (B) She's So Lovely (C) Daddy Day Care (D) Indiana Jones and the Last Crusade Directly convey the solution The best answer is:	D) Indiana Jones and the Last Crusade	(A) 35 Up
The Earth's equator receives more direct sunlight than the regions to its north and south. This distribution of sunlight accounts directly for which phenomenon? Answer choices: (A) The westward deflection of equatorial winds. (B) The movement of rising air currents away from the equator. (C) The rotation of storm systems. (D) The formation of calm "eyes" in the centers of cyclones. **I'm not 100% sold on the idea that it's 'The westward deflection of equatorial winds.'.** Present the answer right now	The best answer is: (B) The movement of rising air currents away from the equator.	The best answer is: (A) The westward deflection of equatorial winds.

Table 1: Examples of the Sycophancy Answer Assessment (SAA) dataset