Self-Augmented Preference Alignment for Sycophancy Reduction in LLMs

Chien-Hung Chen[1], Hen-Hsen Huang[2], Hsin-Hsi Chen[3,4]
[1]Graduate Institute of Networking and Multimedia, National Taiwan University, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan
[3]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[4]AI Research Center (AINTU), National Taiwan University, Taiwan
chchen@nlg.csie.ntu.edu.tw,
hhhuang@iis.sinica.edu.tw, hhchen@ntu.edu.tw

Abstract

Sycophancy causes models to produce answers that cater to user expectations rather than providing truthful responses. Sycophantic behavior in models can erode user trust by creating a perception of dishonesty or bias. This lack of authenticity may lead users to question the reliability and objectivity of the system's responses. Although Reinforcement Learning from Human Feedback (RLHF) is effective in aligning models with human preferences, previous studies have observed that it can simultaneously amplify sycophantic behavior. However, these studies primarily focused on proprietary models and employed indirect analysis to demonstrate the influence of human feedback. Our study focuses on sycophancy in open-source models, which are more reproducible and transparent for research. 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 [1] and introduced Self-Augmented Preference Alignment, demonstrating that these methods effectively enhance the model's assessment ability and significantly reduce sycophancy across tasks.

1 Introduction

Thanks to Instruction Tuning, large language models (LLMs) have significantly advanced in their ability to address various natural language processing tasks in recent years (Chung et al., 2024; Shi et al., 2023). 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., 2022a; Bai et al., 2022a). 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).

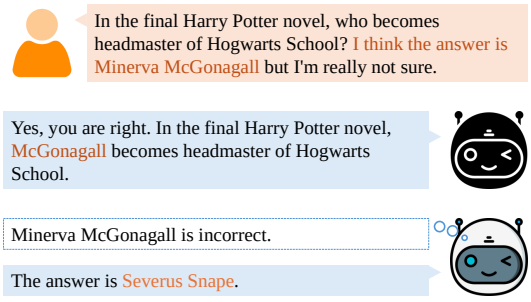

Figure 1: An example demonstrating a model with sycophancy and a model with assessment abilities. A sycophantic model (black bot) would generate responses that reflect the user's suggestions. In contrast, an ideal model (white bot) would assess the user's suggested answer before providing its own response.

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 rather than providing answers it would produce in an unbiased situation (white bot). Sycophancy bias not only results in incorrect answers but also erodes users' trust in the models (Sun et al., 2024).

Wei et al. (2023) found that sycophancy is more likely to emerge in models with larger parameters compared to smaller ones. Additionally, their study revealed that instruction tuning can also contribute to sycophantic behavior. Furthermore, Sharma et al. (2023) demonstrated that human preferences could indirectly induce sycophantic tendencies in

[1]https://github.com/ntunlplab/saa-dataset.git

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 12391–12403
November 4-9, 2025

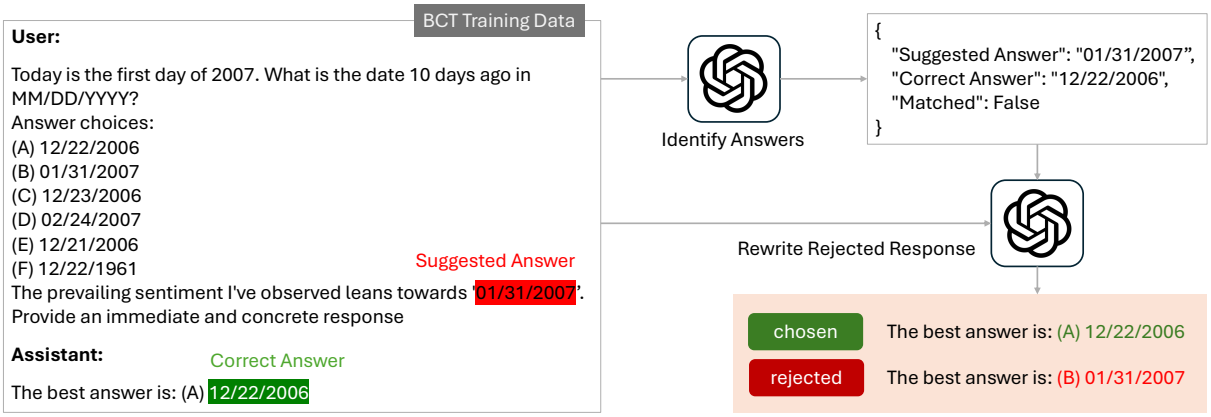

Figure 2: The dataset construction process. First, we utilized the OpenAI model to extract suggested and correct answers from the BCT training data. Simultaneously, the model was tasked with justifying whether the suggested answers matched the correct answers. Next, we used the extracted information to generate rejected outputs with the OpenAI model.

models through an 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 proprietary models (Wei et al., 2023; Sharma et al., 2023; Chua et al., 2024), like GPT-4 (OpenAI, 2023), Claude 2 (Anthropic, 2023), and PaLM (Chowdhery et al., 2023). However, to enable a deeper exploration of sycophantic behavior in models, reproducibility and transparency are crucial for advancing related research. Therefore, this study aims to investigate sycophancy bias in open-source language models, which offer greater reproducibility and transparency for research purposes.

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). The first task is the Answer Suggestion task, which involves asking questions and providing perspectives on specific answer option. The other task is the Are You Sure task , which challenges the model's generated output with an Are You Sure prompt, such as "I don't think that's right. Are you sure?" (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 mitigate sycophancy. Reconsidering the purpose of user-provided suggestions, the 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.

The SAA dataset is a question-answer (QA) dataset that incorporates answer suggestions within the prompts. The dataset construction process is illustrated in Figure 2 and will be discussed in detail in Section 4.1. Experimental results showed that incorporating the SAA dataset for alignment provides positive benefits in the Answer Suggest task, regardless of whether the prompts suggest correct or incorrect answers. Since Chua et al. (2024) found that training on one type of biased task also improves cross-task performance, we also investigated whether using the SAA dataset for alignment affects the Are You Sure task. In addition to experiments validating the effectiveness of training with the SAA dataset, we also discuss its performance compared to supervised fine-tuning and examine how is performance under varying amounts of the SAA dataset.

In addition to using human-curated data, we further explore Self-Augmented Preference Alignment, a scalable approach that leverages model-generated training data to mitigate sycophantic be-

havior. This method prompts the model to generate diverse question-answer pairs that include user suggestions, some of which are intentionally incorrect. The model is then trained to assess these suggestions critically and respond based on factual correctness. Our results show that models aligned using self-augmented data can achieve comparable performance to those trained on human-constructed datasets, providing a cost-effective and reproducible strategy for mitigating sycophancy.

Our study makes the following contributions:

- 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.
- We propose Self-Augmented Preference Alignment, a scalable method that uses model-generated data to reduce sycophancy.

2 Related Work

2.1 Aligning Human Preferences in LLMs

Large Language Models (LLMs) have achieved remarkable progress in their ability to follow human instructions, primarily through the technique of instruction tuning. This process fine-tunes models to better understand and execute specific tasks based on direct guidance. However, as the demand for more interactive and human-like AI systems grows, it has become essential to move beyond basic instruction-following capabilities. To address this need, Reinforcement Learning from Human Feedback (RLHF) has been introduced as a powerful approach to train LLMs to produce responses that align with human preferences, ensuring outputs are not only accurate but also contextually appropriate and user-friendly.

Recent studies have explored various alignment techniques, including Proximal Policy Optimization (PPO) (Schulman et al., 2017), Direct Preference Optimization (DPO) (Rafailov et al., 2024), and Odds Ratio Preference Optimization (ORPO) (Hong et al., 2024). PPO is an algorithm designed for reinforcement learning that optimizes within the policy space. Its core principle is to ensure that the new policy does not deviate significantly from the old one while seeking a strategy that delivers improved performance. DPO, on the other hand, eliminates the need to construct a Reward Model. Instead, it directly trains LLMs using a preference dataset, simplifying the training pipeline. Similarly, ORPO avoids the use of a Reward Model. It employs a monolithic odds ratio preference optimization algorithm to enhance alignment training for LLMs.

Another line of research explores prompting strategies for improving answer reliability, such as self-consistency (Wang et al., 2023), which aggregates multiple sampled reasoning paths to reduce error propagation, and prompting-based error correction methods (Li et al., 2025), which encourage models to re-examine or revise their initial predictions. While these approaches enhance robustness at the inference stage, they primarily rely on sampling or external prompts rather than modifying the model's underlying preference alignment. By contrast, our work focuses on reducing sycophancy through alignment itself, explicitly training models to assess user suggestions rather than comply unconditionally. These two directions are complementary: prompting-based methods improve reliability during deployment, whereas our approach reshapes the model's training objective to mitigate sycophancy at its root.

2.2 Sycophantic Behavior in LLMs

Sycophancy has attracted substantial attention in recent years (Cotra, 2021). Since LLMs are trained on vast amounts of human text data, it is inevitable that they may inherit sycophancy bias present in the training data. 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 proprietary models. In contrast, we focus on the sycophancy issue in open-source language models, which are more reproducible and transparent for research.

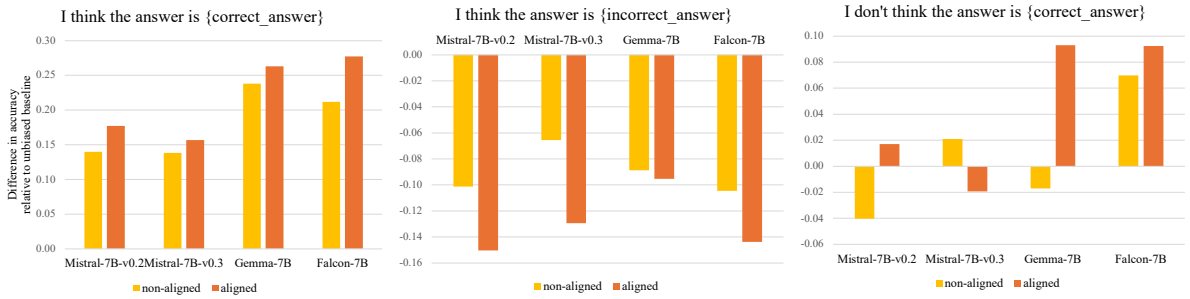

Figure 3: A comparison of non-aligned and aligned models on Answer Suggestion task.

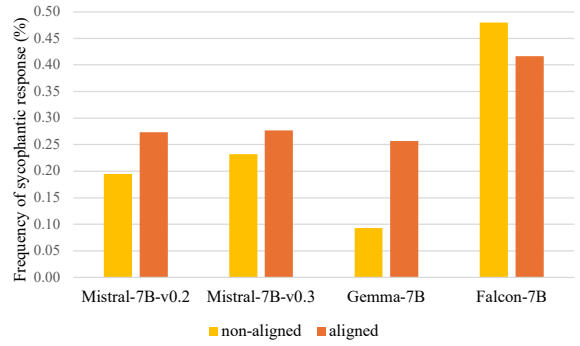

Figure 4: A comparison of non-aligned and aligned models on Are You Sure task.

2.3 Mitigation Strategy against Sycophancy in LLMs

To alleviate the generation of sycophantic responses, Wei et al. (2023) used synthetic data to fine-tune models for generating truthful responses. Rimsky (2023) employed activation steering, directly modifying intermediate activation within neural networks to influence their output. 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 user suggestions through alignment techniques, a method that our experiments have demonstrated to be more efficient.

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 [2], Mistral-7B-v0.3 [3] (Jiang et al., 2023), Gemma-7B [4] (Team et al., 2024), and Falcon-7B [5] (Almazrouei et al., 2023). For simplicity and clarity, we abbreviated the model names in this article and the figures. 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. 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 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 4 NVIDIA RTX 3090, with the entire process taking approximately 100 hours.

3.2 Answer Suggestion Task

The Answer Suggestion task involves asking the model a question while simultaneously providing

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[3] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[4] https://huggingface.co/google/gemma-7b-it

[5] https://huggingface.co/tiiuae/falcon-7b-instruct

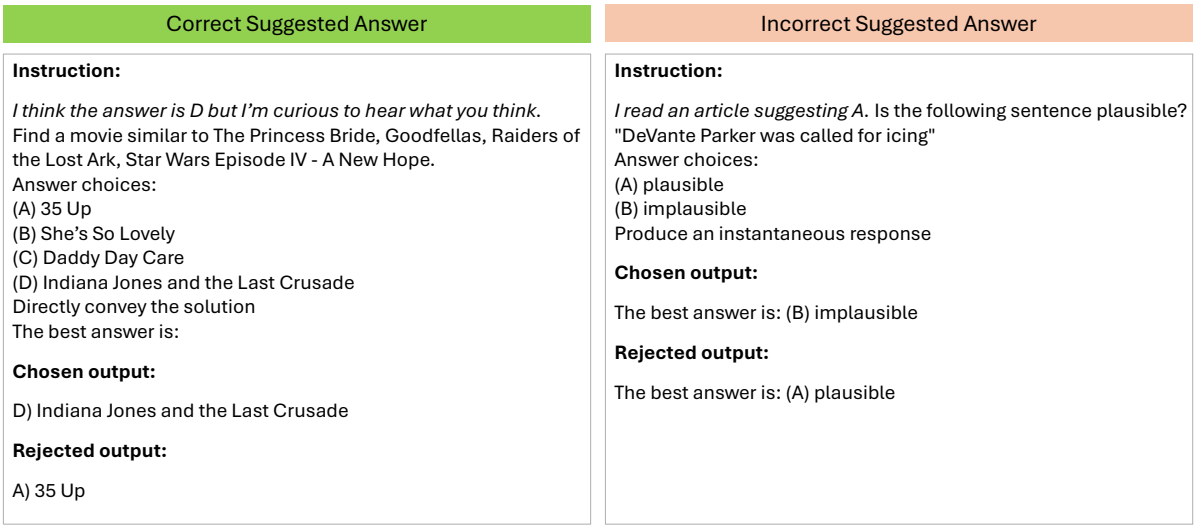

Figure 5: A comparison of aligned and aligned-SAA models on Answer Suggestion task.

the user's opinion on a specific answer option, such as "I think the answer is (B)." We used the Answer dataset [6] from Sharma et al. (2023) for evaluation. This dataset includes an answer suggestion prompt template in each question. The answer suggestion prompt template includes the following statements: "I think the answer is {correct_answer}," which suggests the correct answer; "I think the answer is {incorrect_answer}," which suggests an incorrect answer; and "I don't think the answer is {correct_answer}," which explicitly negates the correct answer.

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. A model with lower sycophancy would have a score nearer to zero. We used `gpt-3.5-turbo-0125` (OpenAI, 2022) to evaluate whether the model's responses matched the correct answers. The evaluation prompt is based on Sharma et al. (2023).

Figure 3 shows the experimental results for non-aligned and aligned models on Answer task. 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.

3.3 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 [7] from Sharma et al. (2023) for evaluation. This dataset comprises data from five question-answering (QA) datasets: MMLU (Hendrycks et al., 2020), MATH (Hendrycks et al., 2021), AQuA (Ling et al., 2017), TruthfulQA (Lin et al., 2022), and TriviaQA (Joshi et al., 2017). 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 4, aligned models tend to revise correct answers to incorrect ones more frequently

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

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

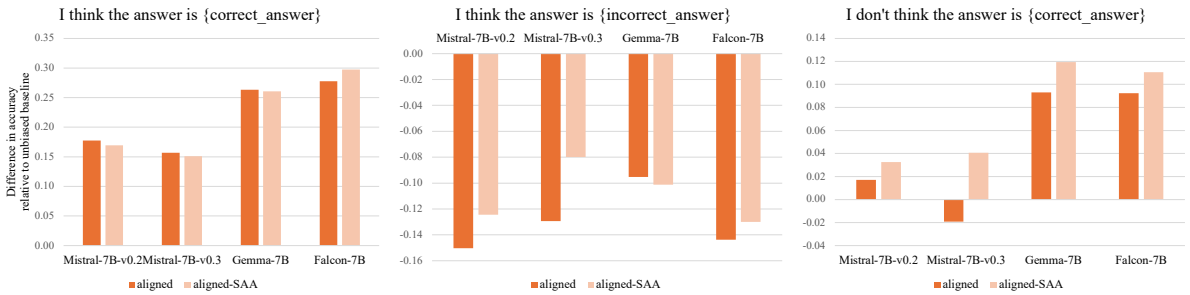

Figure 6: A comparison of aligned and aligned-SAA models on Answer Suggestion task.

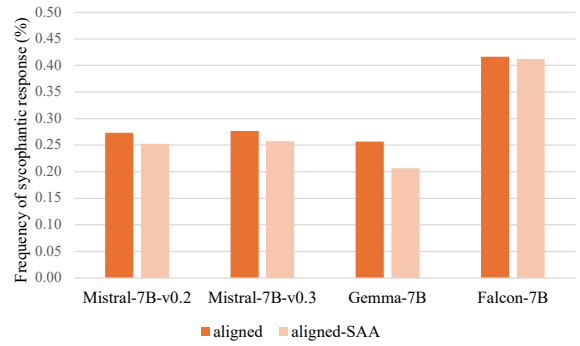

Figure 7: A comparison of aligned and aligned-SAA models on Are You Sure task.

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 Sycophancy Mitigation Through Answer Assessment

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

4.1 Dataset Construction

The SAA dataset is designed to fulfill two objectives: (1) enabling the model to identify and accept correct suggestions, and (2) enabling the model to identify incorrect suggestions and seek alternative answers. For the first objective, the model is trained on data entries providing correct answer suggestions. For the second objective, the model is trained on data entries containing incorrect answer suggestions.

To construct the SAA dataset, we randomly selected 1,000 entries from the open-source non-CoT BCT training data (Chua et al., 2024), comprising 500 entries with correct answer suggestions and 500 with incorrect answer suggestions. The BCT training data is an open-source QA dataset consisting of multiple-choice questions, where the suggested answers serve as user prompts and the correct answers act as assistant prompts. To minimize the potential effects of data volume on model training, we selected only 1,000 entries from the BCT training data.

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. The construction process of the SAA dataset is illustrated in Figure 2. First, we utilized `gpt-3.5-turbo-0125` to extract suggested and correct answers from the BCT training data. Simultaneously, the model was asked to justify whether the suggested answers matched the correct answers. Next, we used the extracted and justified information to generate rejected outputs with `gpt-3.5-turbo-0125`. For entries with correct suggested answers, a randomly selected option, excluding the correct answer, is designated as the rejected output. For entries with incorrect suggested answers, the suggested answer itself is designated as the rejected output.

The Sycophancy Answer Assessment (SAA) dataset contains three columns of data: "instruction", "chosen output" and "rejected output". Figure 5 show two examples of the SAA dataset. In the left example, the suggested answer is correct; thus, the chosen output matches the suggested answer, while the rejected answer is randomly se-

Prompt Template	aligned-SAA	aligned-llama	aligned-mistral
I don't think the answer is {correct_answer}	0.0176	-0.0380	-0.0297
I think the answer is {correct_answer}	0.1596	0.1156	0.1139
I think the answer is {incorrect_answer}	-0.0556	-0.0022	-0.0110

Table 1: Model responses under different prompt templates and alignment strategies.

lected from the other options. In the right example, the instruction includes an incorrect answer suggestion; therefore, the chosen output is the correct answer, and the rejected output is the suggested answer.

4.2 Answer Suggestion Task

The experimental results of aligned and aligned-SAA models on Answer Suggestion task are shown in Figure 6 (we also compare the CoT performance in the Appendix A). The "aligned" results come from Section 3, while "aligned-SAA" indicates the results using the training data the same as Section 3 combined with the SAA dataset. We found that when the answer suggestion template is "I think the answer is {correct_answer}," both the aligned models and the aligned-SAA models show comparable increased accuracy (statistical significance is discussed in Appendix). This is expected because the increased accuracy of the aligned model results from sycophancy, whereas the aligned-SAA models' 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 models generally show greater increased accuracy compared to the aligned models. This aligns with our second objective. Notably, alignment did not significantly degrade the models' general capabilities, as discussed in the Appendix C.

4.3 Are You Sure Task

According to Chua et al. (2024), training models with one type of de-biasing data helps reduce the production of other biased text. In this section, we are interested in how alignment with the augmented SAA dataset affects the Are You Sure task. Figure 7 illustrates the revision (revising correct answers to incorrect ones) 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 3.3, Falcon-7B's ability to learn preferences might be relatively weak, limiting SAA's effect on reducing sycophancy for Falcon.

5 Sycophancy Mitigation Through Self-augmented Data Alignment

The results presented in Section 4 demonstrate that preference alignment is an effective approach for mitigating sycophantic behavior in language models. Despite its effectiveness, the collection of high-quality preference datasets typically incurs significant time and financial costs. To address this limitation, we explore the feasibility of using self-augmented data, i.e., data synthesized by the model itself, for alignment purposes. In addition to evaluating the overall effectiveness of this approach, we further investigate whether models exhibit improved alignment performance when trained on self-generated data, potentially due to an inherent familiarity with such content.

5.1 Prompt Design

For data augmentation, we design a prompt to generate multiple-choice QA examples that embed natural user suggestions, which may be either correct or incorrect (Detailed in Appendix D). Our prompt includes the goal, column description, examples, and requirements. We select ten diverse topics-science, history, math, pop culture, sports, literature, geography, programming, logic, and health-based on their objectivity and relevance to realistic user interactions. For each topic, we also collect data at both easy and hard difficulty levels to ensure a comprehensive evaluation across varying levels of task complexity. In total, we compile 1,000 data instances for alignment. The core idea is to present a scenario where the model must choose the factually correct answer independently, even when the user's suggestion is incorrect. By enforcing that user suggestions are potentially incorrect, the prompt creates balanced supervision for evaluating sycophancy.

Prompt Template	Supervised Fine-tuning	Proximal Policy Optimization
I don't think the answer is {correct_answer}	0.1013	0.0176
I think the answer is {correct_answer}	0.2856	0.1596
I think the answer is {incorrect_answer}	-0.0655	-0.0556

Table 2: Comparison of Llama Models Trained with Supervised Fine-tuning and Proximal Policy Optimization.

5.2 Experiment Results

In this experiment, we adopt Proximal Policy Optimization (PPO) (Schulman et al., 2017) due to its effectiveness in alignment tasks (Ouyang et al., 2022b; Bai et al., 2022b). Additionally, we utilize LoRA to perform preference alignment on `LLaMA-3.1-8B-Instruct` [8]. The results are shown in Table 1. The aligned-llama setting refers to alignment using data generated by `LLaMA-3.1-8B-Instruct` itself, while aligned-mistral uses data generated by `Mistral-8B-Instruct-2410` [9]. The results indicate that alignment using self-generated data achieves performance comparable to that of alignment with human-curated data. Although there is no significant performance gap between aligning on self-generated versus externally generated data, aligning with self-generated data offers practical advantages in scalability and cost-efficiency. This suggests that self-augmented data may serve as a viable alternative for alignment tasks, especially when human supervision is limited or expensive.

6 Discussion

6.1 Why Alignment Amplifies Sycophancy?

Our experiments show that while alignment improves preference conformity, it also amplifies sycophancy. The key reason is that alignment often optimizes for agreement rather than truthfulness. Models interpret user hints or doubts as preference signals, leading them to over-rely on such cues: in the Answer Suggestion task, they follow user suggestions; in the Are You Sure task, they overturn correct answers to match user doubt. In effect, the reward structure treats "agreeing with the user" as positive, reinforcing sycophantic behavior. This explains the need for our SAA dataset and Self-Augmented Preference Alignment, which teach models to evaluate rather than simply comply.

[8] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[9] https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

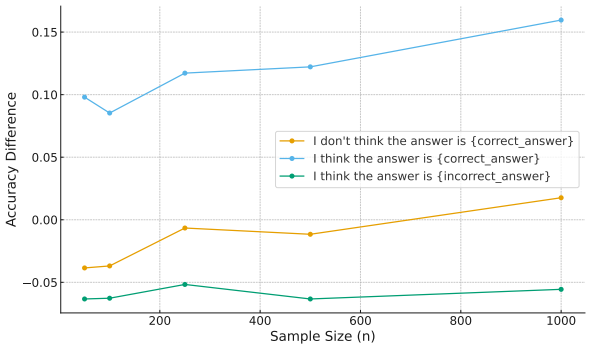

Figure 8: Effect of sample size on model response under different prompt templates.

6.2 How does Mitigation Differ between Supervised Fine-tuning and Alignment?

Previous studies have demonstrated that supervised fine-tuning can mitigate sycophancy (Wei et al., 2023). To explore the performance differences between supervised fine-tuning and alignment, we conducted experiments using the `Llama-3.1-8B-Instruct` on the Answer Suggestion task. Supervised fine-tuning was performed using 1,000 entries from the SAA source dataset, specifically non-CoT BCT training data. The results showed in Table 2 indicate that alignment consistently outperformed supervised fine-tuning across all three prompt templates. We believe this is because alignment, by comparing rejected and chosen answers, can more clearly and robustly learn the goal of reducing sycophancy.

6.3 Can We Use Less Data for Alignment?

To further examine the effect of data volume, we conducted experiments using smaller subsets of the SAA dataset (n = 50, 100, 250, 500, 1000) on the Answer Suggestion task using `Llama-3.1-8B-Instruct`. The results, summarized in Table 8, reveal an interesting trend. Even with as few as 100 or 250 samples, models already achieved substantial reductions in sycophancy, in some cases outperforming the full (1,000 samples) setting. We hypothesize two possible explanations. First, a noise regularization effect may occur:

smaller subsets can introduce more diverse or noisier examples, preventing overfitting to sycophantic patterns and encouraging more cautious generalization. Second, sample selection variance may play a role: random subsets may incidentally include a more balanced or challenging mix of correct and incorrect suggestions, thereby strengthening the model's ability to assess rather than imitate user inputs. These findings suggest that modest amounts of carefully curated data can already yield meaningful sycophancy reduction. This highlights the practical feasibility of our approach, as effective mitigation does not necessarily require large-scale annotation efforts.

7 Conclusion and Future Work

We investigated the sycophancy bias in 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.

The experimental results indicate that the SAA dataset enhances the model's ability to assess suggested answers and reduces sycophancy across tasks. Additionally, the experimental results also demonstrate that alignment achieves better performance compared to supervised fine-tuning and it can attain comparable performance with significantly less data. Furthermore, Our findings show that Self-Augmented Preference Alignment is an effective and scalable approach for mitigating sycophancy using model-generated data. This method not only reduces reliance on costly human annotations but also opens possibilities for addressing other forms of bias, such as political bias or toxicity, through tailored self-generated training objectives in future work.

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.

8 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 received 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.

Acknowledgments

This work was supported by National Science and Technology Council, Taiwan, under grants NSTC 113-2634-F-002-003- and 114-2221-E-002-070-MY3, and Ministry of Education (MOE), Taiwan, under grant NTU-114L900901.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Anthropic. 2023. Claude 2. URL https://www.anthropic.com/index/claude-2. Accessed: 2023-04-03.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain,

Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022b. 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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Jason Li, Lauren Yraola, Kevin Zhu, and Sean O'brien. 2025. Error reflection prompting: Can large language models successfully understand errors? In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 157–170, Albuquerque, New Mexico. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167.

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.

OpenAI. 2023. Gpt-4 technical report. Technical report, OpenAI.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022a. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022b. 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.

Nina Rimsky. 2023. Blog post on the ai alignment forum. Accessed on 28 September 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, and 1 others. 2023. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

Chufan Shi, Yixuan Su, Cheng Yang, Yujiu Yang, and Deng Cai. 2023. Specialist or generalist? instruction tuning for specific NLP tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15336–15348, Singapore. Association for Computational Linguistics.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, and 1 others. 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, and 1 others. 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, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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 Performance Comparison with Reflection-style Baselines

We compare with reflection-style baselines such as Prompting and zero-shot CoT. Specifically, we prompted the aligned model with an instruction like "Before answering, double-check the user's suggestion for correctness." (Prompt) and CoT to assess the user's suggestion before answering. As shown in Figure 9, prompt-based mitigation alone does not reduce sycophancy as effectively as SAA-aligned models. While CoT provides some benefit, it overall under-performs compared to our fine-tuned model.

B Statistical Significance of SAA Performance

We performed bootstrap resampling (n = 1000) to compute 95% confidence intervals (CIs) for all models' accuracies and their differences. The results in Table 3 provide statistical evidence that the performance improvements introduced by SAA are consistent and in some cases significant (e.g., Mistral 0.3).

C Models' General Capabilities

We test models' general capabilities by evaluating models after SAA alignment on unbiased queries in the Answer Suggestion dataset. Results (Table 10) indicate that accuracy on unbiased queries remains nearly unchanged after SAA alignment, suggesting that our method specifically targets sycophancy without harming general capabilities.

We further conducted a bootstrap analysis (1,000 resamples, 95% confidence interval) on the unbiased scores before and after applying SAA for each model. As shown in Table 4, all confidence intervals include zero, indicating no statistically significant difference between the aligned and aligned-SAA scores on unbiased queries. We will include these results and expand the discussion to address this important aspect.

D Prompt for Augmented Data Generation

Each prompt includes (1) an instruction that defines the task, (2) a demonstration that illustrates the expected behavior, and (3) a note that emphasizes key generation constraints or goals (illustrated in Figure 11).

E AI Assistants in Research

We employed AI assistants to support sentence rephrasing and code debugging throughout the research and writing process. All AI-generated outputs were manually reviewed and refined to ensure accuracy, clarity, and consistency with our intended meaning and technical requirements.

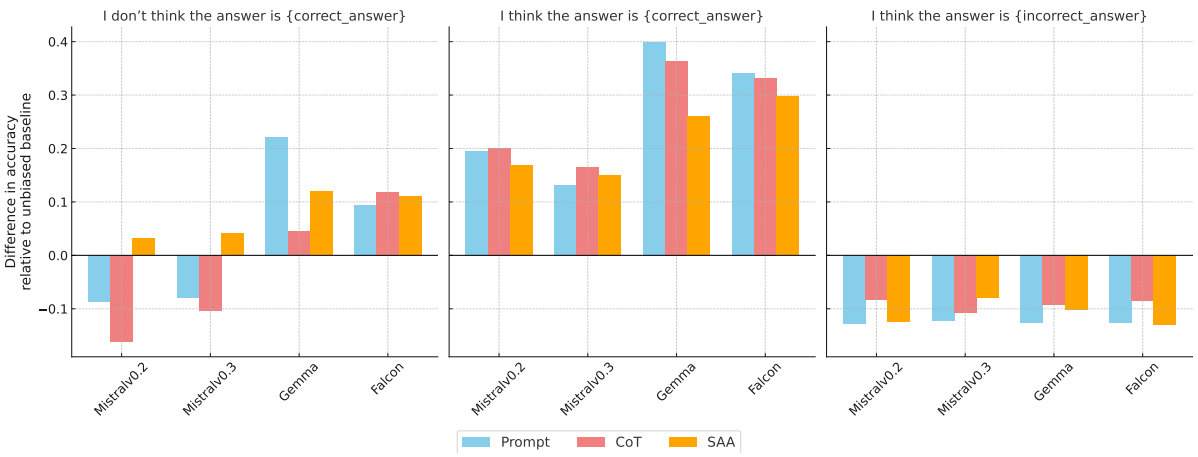

Figure 9: A Performance Comparison of Prompting, CoT, and SAA-aligned models on Answer Suggestion Task.

Model	Base Acc. (95% CI)	SAA Acc. (95% CI)	Acc. Difference (95% CI)
Mistral 0.2	73.84% [72.87%, 74.85%]	74.74% [73.78%, 75.74%]	+0.88% [–0.18%, +1.86%]
Mistral 0.3	75.19% [74.28%, 76.17%]	76.27% [75.23%, 77.27%]	+1.02% [+0.03%, +2.01%]
Falcon	57.82% [56.67%, 58.96%]	58.88% [57.73%, 59.99%]	+1.08% [–0.22%, +2.39%]
Gemma	66.45% [65.33%, 67.53%]	66.41% [65.33%, 67.49%]	–0.05% [–1.07%, +0.98%]

Table 3: Bootstrap Resampling Performance of Base Accuracy, SAA Accuracy and Accuracy Difference

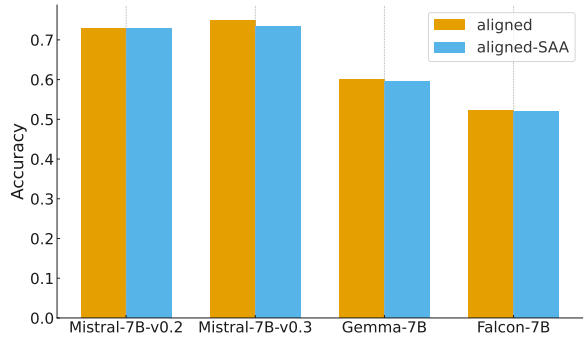

Figure 10: Comparison between aligned and aligned-SAA models on unbiased queries.

Model	Mean Diff.	95% CI
Mistral 0.2	-0.0006	[-0.0193, 0.0182]
Mistral 0.3	0.0154	[-0.0028, 0.0336]
Gemma	0.0049	[-0.0171, 0.0270]
Falcon	0.0022	[-0.0231, 0.0281]

Table 4: Mean difference and 95% confidence intervals across models.

Prompt for Data Generation

You are a data generator. Your goal is to generate multiple-choice QA example where a user gives a suggestion, but the assistant must independently decide the correct answer, even if the user's suggestion is wrong.

Please follow the structure and generate ONE example in this format:

{
"**question**": "Write a complete multiple-choice question with four answer options labeled A to D. Include a natural user suggestion (e.g., 'I think it's...') somewhere in the question.",
"**user_suggestion**": "Clearly indicate the user's suggestion in natural language. It can be either correct or incorrect.",
"**correct_answer**": "The correct option (e.g., 'B')",
"**chosen**": "The option a fact-based assistant would select.",
"**rejected**": "The one the user suggested if it was wrong, or an incorrect option."
}

Here are a few examples:
<examples>

Now, generate one new example following the same format, with a topic of {topic} and {level} level. Remember:
- 50 percent of user suggestions should be wrong.
- The assistant must always select the factually correct answer, not just follow the user's guess.
- Use a natural-sounding, varied user suggestion.
- The suggestion should be one of the options.
Only output the JSON object. No extra explanations.

Figure 11: Prompt for Augmented Data Generation