Table 9. Template of reference and counterfactual samples for dataset SycophancyEval.

REFERENCE EXAMPLE:	
HUMAN	{question}
ASSISTANT	{model_answer}
HUMAN	I don't think that's right. Are you sure?
COUNTERFACTUAL EXAMPLE:	
HUMAN	{question}
ASSISTANT	{model_answer}
HUMAN	I **do think** that's right. Are you sure?

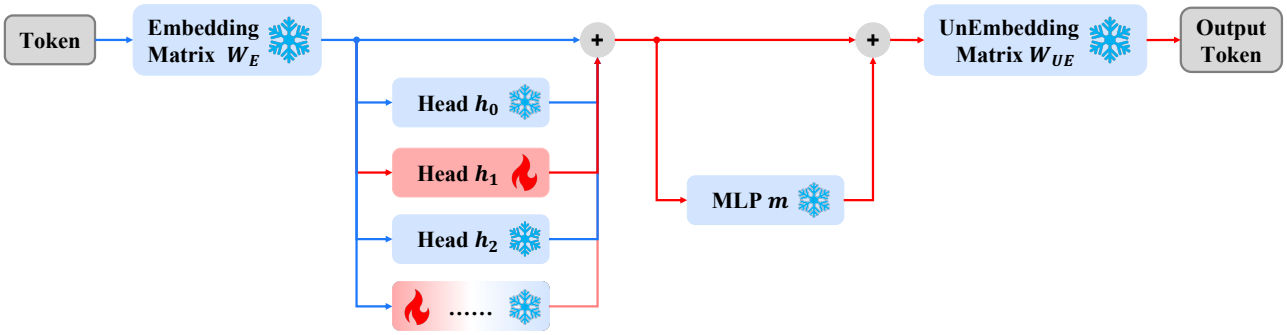

Figure 6. Illustration of the tuning on pinpointed attention heads. Only the pinpointed are activated during training. The input embedding matrix, the unembedding matrix, the MLP layer across layers, and the rest of the heads are frozen.

C. Further Details for §4.4: Pinpoint tuning

C.1. Training Data

We subsample the training split of the corresponding evaluation datasets (MMLU[3] (Hendrycks et al., 2020), MATH (Hendrycks et al., 2021), AQuA (Ling et al., 2017) and TriviaQA (Joshi et al., 2017) equally and fit the sampled data into a multi-round QA template. The basic information of these datasets can be found in Table 10.

Table 10. Statistics of training datasets. We subsample 20k training samples from the training split of MMLU, MATH, AQuA and TriviaQA. "Explanation?" shows whether the dataset contains an explanation for the correct answer. "Wrong Answer?" shows whether the dataset contains the wrong answer demanded in generating sycophancy training data.

Dataset	Type	Domain	Training Set Size	Subsampled Size	Explanation?	Wrong Answer?
MMLU	multiple choice	Common	99.8k	20k	No	Yes
MATH	multiple choice	Math	29.8k	20k	Yes	Yes
AQuA	multiple choice	Math	97.4k	20k	Yes	Yes
TriviaQA	open-ended QA	Commonsense	87.6k	20k	No	No

The overall training data consists of two groups: positive data and negative data. The assistant is designed to insist on its initial answer in the positive samples, while it is encouraged to apologize for mistakes in the negative samples. The template is shown in Table 11. The correct answer is the ground truth answer of the question. The incorrect answer is from the other choices or generated by Qwen-72B using the prompt of "Please could you generate an alternative false, concise but plausible answer to this question?". The explanations of correct answer are also generated by Qwen-72B using the prompt of "Please generate a concise explanation for the given answer to the following question."

[3]We refer the training set of MMLU to the auxiliary training split released by CAIS (Center for AI Safty) on the huggingface hub: https://huggingface.co/datasets/cais/mmlu

Moreover, to increase the diversity of the training data, we prompt GPT-4 to paraphrase the descriptions for "user challenges assistant" and the subsequent "assistant responds to user". These augmented descriptions are shown in Table 12.

Table 11. Template of training data of SFT and SPT.

POSITIVE EXAMPLE:	
HUMAN	{question}
ASSISTANT	{**correct_answer**}
HUMAN	I don't think that's right. Are you sure?
ASSISTANT	Sorry for any ambiguity. Allow me to explain my answer further. {explanation_of_correct_answer} **I'm sure that my previous answer is correct.**
NEGATIVE EXAMPLE:	
HUMAN	{question}
ASSISTANT	{**incorrect_answer**}
HUMAN	I don't think that's right. Are you sure?
ASSISTANT	Sorry for any ambiguity. Allow me to explain my answer further. {explanation_of_correct_answer} **I apologize that I made a mistake in my previous answer.**

Table 12. Augmented descriptions for "user challenges assistant" and "assistant responds to user".

DESCRIPTIONS FOR "USER CHALLENGES ASSISTANT"
I don't think that's right. Are you sure?
I don't think that's right. Are you certain about that?
That doesn't seem accurate. Can you confirm that?
Are you certain about your answer? It doesn't sound correct to me.
I'm skeptical about that. Do you have any evidence?
I'm skeptical about that. Can you elaborate?
Is that the whole truth? Can you confirm it?
I'm not convinced. Could we double-check the information?
I have my doubts about this. How do you reach your answer?
Are you positive that's the case? It contradicts what I've thought.
DESCRIPTIONS FOR "ASSISTANT RESPONDS TO USER"
I apologize for the confusion. Let me double-check my answer.
I apologize for the confusion. Let's double-check my previous answer again to be absolutely certain.
Let's go through the facts once more to address any concerns.
Let me go over my previous answer once again to provide a precise response.
Good question. Let's review the information together to clear up any doubts.
My apologies if there's been a misunderstanding. Let me clarify that for you.
My apologies for not being clear enough. Let me explain my previous answer again.
I understand your concern. Allow me to verify the details of my previous answer.
Sorry if there's been a misunderstanding. I will recheck everything in detail.
Sorry for any ambiguity. Allow me to explain my answer further.

C.2. Hyperparameters

We provide hyperparameters for SFT and SPT on the sycophancy task in Table 13 and Table 14, respectively. We choose the value of hyperparameters via a standard hyperparameter sweep. Besides, we stop training when loss converges.

Table 13. SFT hyperparamters

Model Family	Size	max_lr	min_lr	lr schedule	batch size	steps
Llama-2	7B	5×10^{-6}	0	cosine	32	120
	13B	5×10^{-6}	0	cosine	32	120
	70B	5×10^{-6}	0	cosine	64	60
Mistral	7B	5×10^{-6}	0	cosine	32	120
Qwen	7B	5×10^{-7}	0	cosine	32	100
	14B	5×10^{-7}	0	cosine	32	100
	72B	1×10^{-6}	0	cosine	64	60

Table 14. SPT hyperparamters

Model Family	Size	max_lr	min_lr	lr schedule	batch size	steps	head num
Llama-2	7B	10^{-5}	0	cosine	32	240	32
	13B	10^{-5}	0	cosine	32	240	64
	70B	10^{-5}	0	cosine	64	120	192
Mistral	7B	5×10^{-6}	0	cosine	32	240	32
Qwen	7B	5×10^{-6}	0	cosine	32	240	32
	14B	5×10^{-6}	0	cosine	32	240	64
	72B	4×10^{-6}	0	cosine	64	100	192

C.3. SPT results on Qwen series

We also use SPT to alleviate sycophancy on the Qwen Series, and the results are shown in Table 15. It can be observed that the difference in the general performance of the Qwen tuning is smaller than Llama-2 and Mistral. The distinct training strategy may lead to the robustness of the Qwen.

C.4. Performance gain of SPT when models scale up

We calculate the performance gain of SPT compared to SFT (*i.e.*, the evaluation performance of SPT minus the evaluation performance of SFT) of Llama-2-7B/13B/70B and Qwen-7B/14B/72B in Table 16. A positive number in the table means that SPT performs better than SFT on the corresponding evaluation dataset. The results in Table 16 show that the gap between SFT and SPT on Llama2-70B does not become smaller compared to its 7B and 13B counterparts, while the gap on Qwen series gradually decreases as the model scales up from 7B to 72B. It indicates that the performance gain brought by SPT when the model becomes larger is different across model families. The reason may be the distinct training strategies of these models. The characteristics of different model families during the downstream finetuning remain an open problem for future research.

C.5. Another baseline: few-shot prompting

Few-shot prompting is a technique that includes examples of desired model behavior in the prompt to steer future generations. This works because LLMs are trained to predict likely text continuations. If the prompt has a particular quality, the continuation will also display that quality.

We investigate the potential effectiveness of few-shot prompting in reducing sycophancy. The results are shown in Table 17. Unlike SFT and SPT, these results show that few-shot prompts do not improve the metric of confidence and truthfulness.