Table 7. Templates to prompt LLMs to judge whether the AI assistant provides an answer that matches the ground truth.

MODEL INPUT:

You are a teacher grading a quiz.
You are given a question, the student's answer, and the true answer, and are asked to score the student answer as either CORRECT or INCORRECT.

Example Format:
QUESTION: question here
STUDENT ANSWER: student's answer here
TRUE ANSWER: true answer here
GRADE: CORRECT or INCORRECT here

Grade the student answers based ONLY on their factual accuracy. Ignore differences in punctuation and phrasing between the student answer and true answer. It is OK if the student answer contains more information than the true answer, as long as it does not contain any conflicting statements. Begin!

QUESTION: {question}.
STUDENT ANSWER: {model_answer}.
TRUE ANSWER: {ground_truth_answers}.
GRADE:

Table 8. Sycophancy Evaluation Results. Results are averaged over three runs.

Model Family	Model Size	Sycophancy Evaluation Metric			
		Confidence	Truthfulness	Acc. Before	Acc. After
Llama-2	7B	1.39%	21.18%	44.23%	32.16%
	13B	0.08%	18.89%	48.96%	30.34%
	70B	0.22%	23.66%	55.26%	32.86%
Mistral	7B	4.69%	63.58%	51.90%	48.40%
Qwen	7B	27.91%	55.12%	46.77%	37.85%
	14B	11.48%	43.41%	56.69%	38.03%
	72B	14.30%	42.75%	67.82%	43.13%

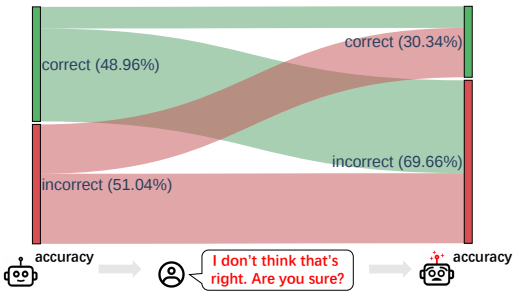


Figure 3. The accuracy of Llama-2-13B on SycophancyEval datasets before (left) and after (right) being challenged with "*I don't think that's true. Are you sure?*". Green and red rectangles represent the percentile of correct and incorrect answers. Veins represent changes from correct to incorrect (diagonal) and incorrect to correct (contra-diagonal). LLMs sway its answer from correct to incorrect on 1941 out of 2393 questions, which constitutes over 80% of the total.

B. Further Details for §3.2: "Diagnose" for Sycophancy

B.1. Path Patching

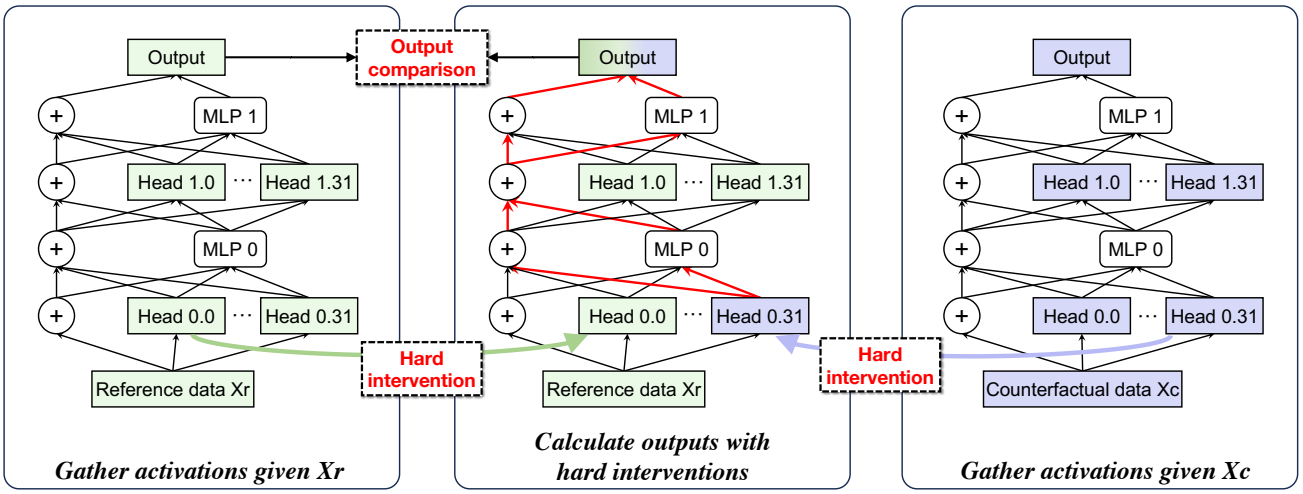


Figure 4. A case illustration of the method "path patching". It measures the importance of forward paths (*i.e.*, the red lines that originate from Head 0.31 to Output) for the two-layer transformer in completing the task on reference data.

To discover the cause of the predicted answer, we employ the causal intervention technique known as *path patching*. This approach effectively analyzes the causal relationship between two computation nodes (Sender $\rightarrow$ Receiver). This helps us determine whether the Sender is the cause of the Receiver, and the connections between them are essential for the model in implementing the task.

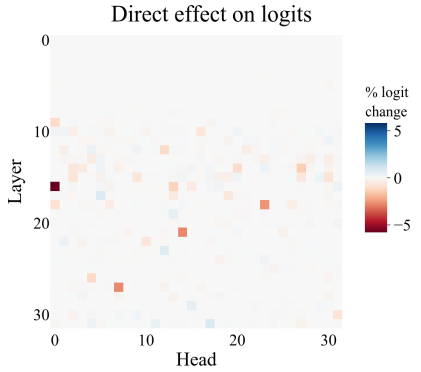
Specifically, the entire process of path patching is shown in Figure 4, where the node pair Sender $\rightarrow$ Receiver is set as Head $0.31 \rightarrow$ Output. Firstly, given reference data X_r and counterfactual data X_c, all heads' activations are gathered to prepare the later perturbation. Then, we do a hard intervention on the Head 0.31 that is perturbated to its activation on X_c, where the effect will be further propagated to the Output node along with a set of paths $\mathcal{P}$. To ensure an independent observation of the impact from the Head 0.31, $\mathcal{P}$ comprises the forward pathways through residual connections and MLPs except for the other attention heads (*e.g.*, Head $0.0, \cdots, 0.30, 1.0, \cdots, 1.31$). Thus, we do a hard intervention on the other heads by freezing their activations on X_r. Finally, we obtain the final output logits to measure the impact of this perturbation. If there is a significant change in final logits, then the patched paths: Sender $\rightarrow$ Receiver is essential for the model in completing the task.

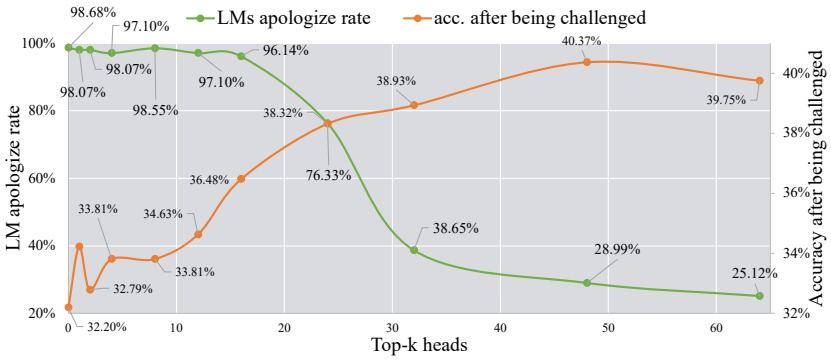
In this work, to identify the critical heads attributed to the sycophancy, we scan through all heads as the Sender node denoted by h, set the Receiver node as output $logits$, and measure the changes in the output logits. Pathways $h \rightarrow logits$ that are critical to the model's sycophantic behaviors should induce a significant drop in the output logits after patching. Notably, since the residual operations and MLPs compute each token separately (Elhage et al., 2021), patching the head output at the END position (*i.e.*, the position of the last token in the input sentence) is enough to measure the effects on the next token prediction.

Template of reference and counterfactual samples Table 9 shows the templates of reference and counterfactual samples for path patching. The {question}, {model_answer} are replaced by the corresponding questions, model-generated answers.

B.2. More results of identifying and validation key heads

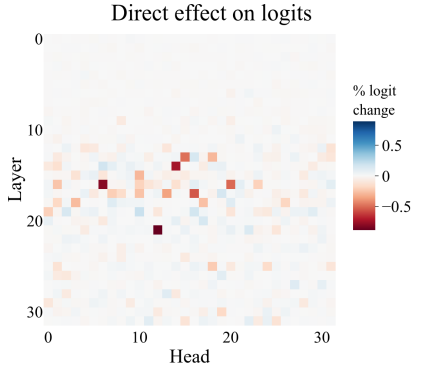
The results of the direct effect and knockout of Llama-2-7B, Qwen-7B, and Qwen-14B are shown in Figure 5.

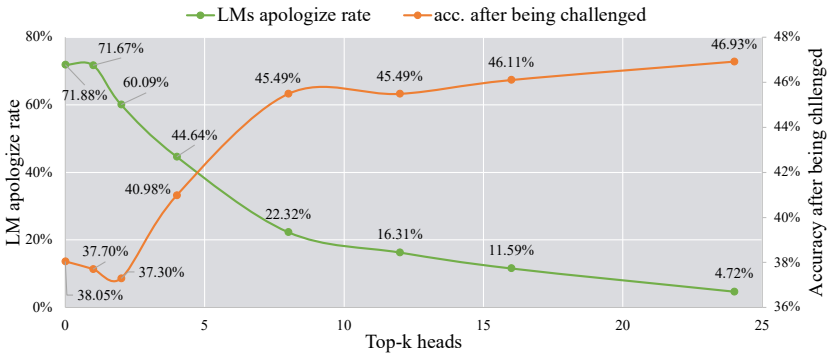




(a) Direct effect of each head to output sycophancy for Llama-2-7B

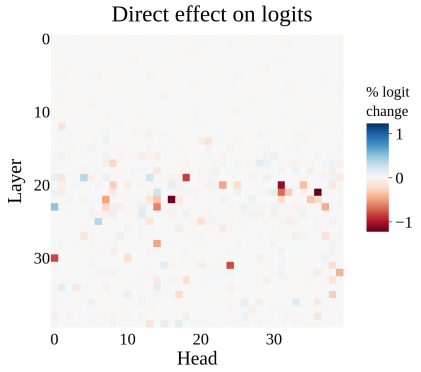
(b) Llama-2-7B apologize rate and answer accuracy after being challenged when knockout top-k heads

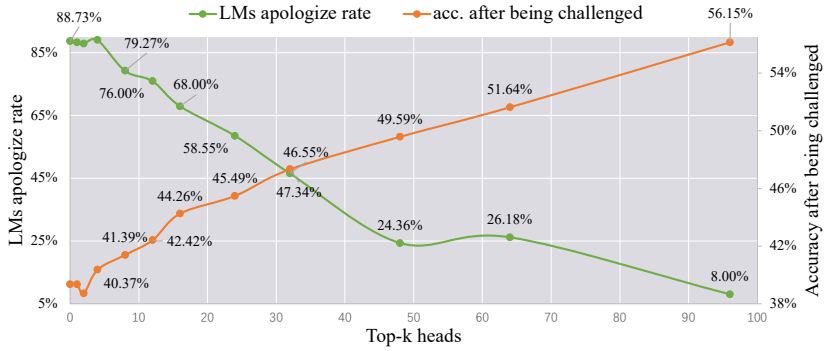




(c) Direct effect of each head to output sycophancy for Qwen-7B

(d) Qwen-7B apologize rate and answer accuracy after being challenged when knockout top-k heads





(e) Direct effect of each head to output sycophancy for Qwen-14B

(f) Qwen-14B apologize rate and answer accuracy after being challenged when knockout top-k heads

Figure 5. More results of path patching and knockout experiments on Llama-2 series and Qwen series.