Table 1. Overall performance. We evaluate the capabilities of Llama-2-7B/13B/70B-Chat and Mistral-7B-Instruct, eliminating sycophancy, distribution deviation, and transitioning to general tasks, *e.g.*, reasoning (StrategyQA), mathematics (GSM8K), and code-generation (HumanEval)). The Δ represents the performance improvement after SFT or SPT. Supervised fine-tuning across the entire parameter set (denoted as SFT) leads to enhanced performance on sycophancy metrics, albeit at the expense of its capabilities in generic abilities. In contrast, supervised pinpoint tuning (SPT) yields comparable improvements while preserving the model's proficiency in generic tasks with less distance deviation and less tunable parameters.

		Sycophancy Metrics				**General Ability**						**Dist. Dev.**
		Confidence		**Truthfulness**		**StrategyQA**		**GSM8K**		**HumanEval**		**KL**
Models	**# Tuned Params.**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Val.**
Llama-2-7B	-	1.39	-	21.18	-	37.03	-	**24.72**	-	16.46	-	-
+ SFT	6.74B	59.12	+57.73	80.00	+58.82	20.09	-16.94	14.63	-10.09	2.44	-14.02	0.0494
+ SPT	67.1M	**70.70**	+69.31	**80.27**	+59.09	**43.82**	+6.79	23.50	-1.22	**17.68**	+1.22	**0.0043**
Llama-2-13B	-	0.08	-	18.89	-	64.24	-	33.89	-	**23.17**	-	-
+ SFT	13.0B	61.55	+61.47	84.06	+65.17	60.92	-3.32	25.32	-8.57	18.29	-4.88	0.0476
+ SPT	168M	**71.92**	+71.84	**86.72**	+67.83	**65.33**	+1.09	**35.48**	+1.59	20.73	-2.44	**0.0026**
Llama-2-70B	-	0.22	-	23.66	-	56.86	-	49.51	-	**37.80**	-	-
+ SFT	69.0B	65.49	+65.27	86.12	+62.46	56.03	-0.83	45.26	-4.25	8.93	-28.87	0.0560
+ SPT	403M	**82.52**	+82.30	**91.26**	+67.60	**63.76**	+6.90	**51.18**	+1.67	36.59	-1.21	**0.0035**
Mistral-7B	-	4.69	-	63.58	-	65.76	-	42.61	-	29.27	-	-
+ SFT	7.24B	52.46	+47.77	78.45	+14.87	8.21	-57.55	32.45	-10.16	20.12	-9.15	0.1046
+ SPT	33.6M	**69.68**	+65.09	**84.69**	+21.11	**66.72**	+0.96	**43.75**	+1.14	**35.98**	+6.71	**0.0008**

to other heads (0.46 vs. 0.15). This observation, to some extent, confirms that the key heads discovered by path patching are generally closely associated with the models' sycophantic tendencies when faced with user challenges. The sycophancy-related heads also have a relatively high average attention score (0.54) on tokens from all remaining tokens. This observation indicates that several heads associated with sycophancy are also involved in gathering information from other parts of the users' inputs.

Validation of key components In Figure 2(b), all heads are sorted by the effect of each head on logits and knocked out one by one. As the heads are gradually knocked out, the rate of the Llama-2-13B apologizing decreases from 100% to 18%, as well as the accuracy after being challenged increases from 30% to 44%. More results of Llama-2-7B and other models are illustrated in the Figure 5 (Appendix). These results demonstrate that the discovered components are essential in the LLM's untruthful sycophancy.

4.3. Baseline: Supervised Finetuning (SFT)

Following Wei et al. (2023), we present a straightforward baseline as collecting synthetic data from NLP tasks and using these data in a supervised fine-tuning step.

Experimental Details We subsample the training split of the corresponding evaluation datasets (MMLU (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 (details for the template and data can be found in Appendix C.1) where the LM sticks to its correct answer in the first round QA while apologizing and then revising its wrong answer. Llama-2-13B is finetuned with a learning rate decreasing from 5×10^{-6} to 0 for 120 steps. SFT hyperparameters for other models are in Table 13 (Appendix).

4.4. Supervised Pinpoint Tuning (SPT)

Experimental Details We use the same training data for SPT and SFT (details for the multi-round QA template and data can be found in Appendix C.1). We selectively tune top 32, 64, and 192 heads for 7B, 13B, and 70B models, respectively. We choose the number of tunable attention heads related to sycophancy via a standard hyperparameter sweep on the Llama-2-13B and scale according to the size of other models. The learning rate decreases from 5×10^{-6} to 0 in 240 training steps with a cosine schedule. SPT hyperparameters for other models are in Table 14 (Appendix).

Results In Table 1, we compare pinpoint tuning (SPT) with the alternative baseline (SFT). Firstly, both SFT and SPT significantly alleviate sycophancy. For example, the answer truthfulness of Llama-2-13B increases from 18.89% to 84.06% (+65.17%) after SFT, while SPT achieves 86.72%

Table 2. Comparison of general abilities on more tasks.

	CSQA (7-shot)		MMLU (0-shot)	
Setting	Acc.	Δ	Acc.	Δ
Llama-2-13B	70.68	-	52.41	-
+ SFT	68.63	-2.05	52.36	-0.05
+ SPT	71.91	+1.23	52.56	+0.15

(+67.83%). And SPT consistently outperforms SFT by a clear margin on the Sycophancy Metric. For instance, the improvement of SFT on Mistral-7B's confidence is 47.77%, and the improvement of SPT is 65.09%. Secondly, compared with SFT, SPT achieves comparable performance while maintaining a clearly low increase of KL divergence. For example, Llama-2-13B SFT leads to a 0.0476 shift of KL, and SPT leads to 0.0026, which is $1/20$ smaller. The detailed comparison of computational cost is shown in Appendix C.6, verifying that SPT can effectively address the sycophancy with little distribution shift.

4.5. Analysis

Evaluation of the general ability To evaluate the general ability of the LLMs, we evaluate the LLMs on five benchmarks: StrategyQA and CSQA for reasoning, GSM8K for arithmetic questions, HumanEval for code generation, and MMLU. Table 1 and Table 2 shows the baseline, SFT, and SPT results on the five benchmarks. Compared to the original model, SFT can lead to a significant decrease in the model's general capability. For example, the accuracy on GSM8K of Llama-2-13B decreases from 33.89% to 25.32%(−8.57%). On the other hand, SPT can effectively address this issue. After SPT, the accuracy on GSM8K achieves 35.48%, which is even better than the original model. A similar phenomenon is also observed in Mistral-7B, which indicates that our proposed SPT can generalize to models beyond Llama-2. Although it's quite inspiring that training on our constructed dataset helps the performance of some models on the HumanEval dataset, our training set doesn't explicitly include coding-related data but includes arithmetic datasets MATH and AQuA. Recent work (Yuan et al., 2023) shows that including code data in the pre-training corpus enhances arithmetic skills in LLMs, and our results may indicate that the reverse case also holds. However, the mechanism and reason behind the improvement are still unclear, and we view this as an interesting topic for further research.

Varying number of tunable components To better understand the characteristics of SPT, we vary the number of tunable components and measure the full performance statistics on three alternative methods. The results are shown in Table 3. First, after varying the number of the selected attention heads, we find that the model answers confidence and truthfulness plateau at the 32 heads. Second, to verify the effectiveness of the selected heads, we randomly choose 64 heads across the model. The random 64 heads achieve 60.11 confidence with a relatively high standard deviation of 7.37 (five times repetition). This shows that randomly chosen heads lead to unstable performance, and it further verifies that tuning only correctly pinpointed attention heads consistently improves performance. Third, we conduct path patching experiments to find the direct effect of each MLP on output sycophancy (Figure 7(a) in Appendix). We tune 64 heads together with the MLP that has the most significant direct effect. We also measure the accuracy of the LLM's responses before and after it is challenged ("Acc. Before" and "Acc. After"). As shown in Table 3, although tuning MLP achieves the best confidence score, the "Acc. Before" is decreased from 46.99% to 43.58% and "Acc. After" is decreased from 47.55 to 43.86. It shows that tuning MLP brings higher impairment to the original model than heads.

Table 3. Ablative experiments on the number of tunable components (Llama-2-13B). Each metric's best and second best performance among settings are in **bold** and underlined, respectively.

Settings	Sycophancy Evaluation Metric			
	Confidence	Truthfulness	Acc. Before	Acc. After
top-8 heads	23.84	37.51	48.49	36.52
top-16 heads	55.24	69.00	<u>48.77</u>	44.41
top-32 heads	70.23	76.77	48.18	45.38
top-48 heads	70.16	83.01	47.79	<u>46.52</u>
top-64 heads	<u>71.92</u>	**86.72**	46.99	**47.55**
random 64 heads	60.11 ±7.37	74.05 ±4.73	**49.49** ±0.36	45.90 ±1.44
top-64 heads + top-1 MLP	**75.82**	<u>84.79</u>	43.58	43.86

Comparison with other PEFT methods We conduct comparative experiments with other PEFT works, including a representative selective PEFT method DARE (Yu et al., 2023) and a reparameterized PEFT method LoRA (Hu et al., 2021). For DARE, we randomly drop 98.71% delta parameters of the SFT model and rescale the rest of the delta parameter with $1/(1-0.9871)$ to meet the same amount of tuned parameters with SPT. For LoRA, we choose the rank to be 16. To further analyze the relation of SPT and PEFT methods, we combine PEFT and SPT by only tuning the identified key attention heads using LoRA, annotated as "LoRA&SPT".

From the results in Table 4, there are several observations: (1) LoRA achieves a similar preservation of general performance with the proposed SPT. But LoRA achieves inferior performance on Confidence/Truthfulness than SPT. This verifies that *tuning the identified key components is more effective* than tuning the whole network for addressing syco-

Table 4. Comparison with other PEFT methods (Llama-2-13B). Each metric's best and second best performance among settings are in **bold** and underlined, respectively.

Settings	Sycophancy Eval.		General Ability	
	Confidence	Truthfulness	StrategyQA	GSM8K
Llama-2-13B	0.08	18.89	64.24	33.89
+ SFT	61.55	84.06	60.92	25.32
+ SPT	71.92	**86.72**	65.33	35.48
+ LoRA	70.04	79.66	65.98	**37.91**
+ DARE	60.38	84.34	60.96	26.91
+ LoRA&SPT	**86.33**	86.21	**66.72**	36.92

Table 5. Generalizatoin of pinpoint tuning for addressing sycophancy beyond training set. Sycophancy Evaluation Datasets are from Perez et al. (2022a).

Setting	NLP (↓)	PHIL (↓)	POLI (↓)	Average (↓)
Llama-2-13B	85.67	95.04	70.09	83.46
+ SFT	**81.99**	94.32	66.33	**80.73**
+ SPT	83.99	**94.14**	**66.25**	81.29

phancy. (2) DARE leads to a similar degradation of general ability as SFT. This shows the *necessity of identifying the task-related modules in the LLM before applying selective PEFT*. (3) Integrating SPT and LoRA can bring a certain performance gain compared with SPT or LoRA alone. The "LoRA&SPT" achieves the best performance on the Confidence metric with a large margin (15%). This illustrates our SPT is *orthogonal to the reparameterized PEFT* and can be boosted with PEFT to improve effectiveness and efficiency.

Addressing sycophancy beyond training set A significant concern is whether pinpoint-tuned models can generalize beyond the SycophancyEval benchmark. As a first step toward investigating this question, we apply a pinpoint tuned model (Llama-2-13B with 64 heads tuned) on Sycophancy Evaluation Datasets from Perez et al. (2022a). This benchmark includes three sycophancy tasks: natural language processing survey questions (NLP), philosophy survey questions (PHIL), and political typology quiz questions (POLI). In these tasks, sycophantic models will tend to select answers that match the user's opinion, even though that opinion is incorrect because the questions are subjective. For each dataset, we compute the frequency of the LLM's answers that matched the user's view, and we calculate over $1k$ evaluation examples. Note that this benchmark was gathered for purposes different from SycophancyEval, so they are a reasonable test of out-of-distribution generalization for the sycophancy addressing. Results are reported in Table 5 (lower is better). Although trained on datasets with different purposes, the results show that both SFT and SPT cause the model to perform somewhat better than the baseline model. This illustrates the generalization of the training and the robustness of pinpointed attention heads.

What happens after pinpoint tuning Besides the quantitative results of the sycophancy evaluation, we conduct the path patching procedure one more time on the LLMs after pinpoint tuning. Similar to Figure 2(a), the direct effect of each head on the sycophancy is shown in Figure 7(b) (Appendix). Comparing the direct effect before and after pinpoint tuning, it is observed that the top-5 heads with the largest impact on sycophancy diminish after SPT. For example, the effect of the head 39 from layer 16 decreases from 3.77% to 0.64%. This result further verifies pinpoint tuning works in a way of reducing the direct effect of the sycophancy on the outputs.

Limitations While our work sheds light on sycophancy reduction, there are several limitations to our work. *Firstly*, the main tool for finding the relevant components in LLMs in this work is *path patching*, which treats each MLP and attention head as individual nodes. We suggest treating each hidden neuron or possibly groups of neurons as the atomic unit instead in future work, as this seems more reflective of the semantics of the computation happening inside deep neural networks (Gurnee et al., 2023). *Secondly*, we set our evaluations of sycophancy as the definition used in Sharma et al. (2023). Although we have conducted additional experiments on other sycophancy evaluation datasets (Wei et al., 2023), it is unclear whether our results generalize to other formats that could be used. Beyond tuning, we also investigate the potential effectiveness of few-shot prompting (FS). Results in Table 17 (Appendix) show that FS does not bring improvement in reducing sycophancy. *Thirdly*, the verification of the effectiveness of pinpoint tuning is mainly conducted on how to address the sycophancy. We conduct another preliminary experiment to improve the ability of arithmetical reasoning with the same methodology. Results are shown in the Appendix C.7. In summary, we do not view our findings as evidence that our specific method can solve all instances but as evidence of the general potential of using interpretability tools to analyze the LLM and change the behavior.

5. Conclusion

In this work, we have pinpointed, analyzed, and tuned the internal components attributed to the sycophantic behaviors of LLMs. Compared with common supervised fine-tuning, pinpoint tuning achieves a significant boost in addressing the sycophancy, without loss of general ability. We hope this work can advance the understanding of the behaviors of LLMs and encourages future work to integrate pinpoint tuning with other interpretability tools like information flow (Wang et al., 2023), representation engineering (Zou et al., 2023) and sparse auto-encoders (Cunningham et al., 2023).