Table 15. SFT and SPT results on Qwen Series. We also evaluate the capabilities of Qwen-7B/14B/72B-Chat, 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. SPT yields less gain on the Qwen Series.

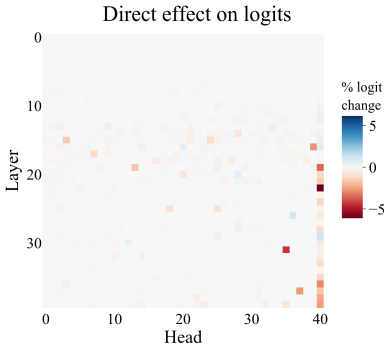
		Sycophancy Metrics				**General Ability**						**Dist. Dev.**
		Confidence		**Truthfulness**		**StrategyQA**		**GSM8K**		**HumanEval**		**KL**
Models	**# Tuned Params.**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Val.**
Qwen-7B	-	27.91	-	55.12	-	**68.56**	-	**50.80**	-	36.59	-	-
+ SFT	7.72B	56.70	+28.79	**81.64**	+26.52	68.21	-0.35	50.04	-0.76	37.80	+1.21	0.0017
+ SPT	67.1M	**73.70**	+45.79	80.69	+25.57	67.60	-0.96	49.28	-1.52	**40.24**	+3.65	**0.0009**
Qwen-14B	-	11.48	-	43.41	-	74.80	-	**61.03**	-	41.46	-	-
+ SFT	14.2B	56.12	+44.64	81.32	+37.91	75.23	+0.43	60.88	-0.15	**46.34**	+4.88	0.0011
+ SPT	168M	**67.08**	+55.60	**86.46**	+43.05	**75.37**	+0.57	59.67	-1.36	45.13	+3.67	**0.0007**
Qwen-72B	-	14.30	-	42.75	-	**82.45**	-	76.04	-	**64.02**	-	-
+ SFT	14.2B	80.21	+65.91	89.09	+46.34	81.22	-1.23	**76.19**	+0.15	59.76	-4.26	0.0012
+ SPT	168M	**81.38**	+67.08	**89.58**	+46.83	82.36	-0.09	75.82	-0.22	60.37	-3.65	**0.0008**

Table 16. Performance gain of SPT compared to SFT. A positive number in the table means that SPT performs better than SFT on the corresponding evaluation dataset. The performance gain of SPT on Llama-2 series is consistently high across different model scales, while the gap between SPT and SFT gradually decreases as the model scales up on Qwen series

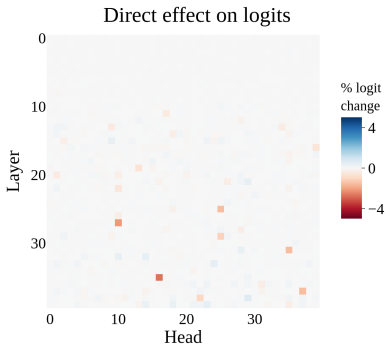
Model Family	**Model Size**	**Sycophancy Metrics**		**General Ability**		
		Confidence	**Truthfulness**	**StrategyQA**	**GSM8K**	**HumanEval**
Llama-2	7B	+11.58	+0.27	+**23.73**	+8.87	+15.24
	13B	+10.37	+2.66	+4.41	+**10.16**	+2.44
	70B	+**17.03**	+**5.14**	+7.73	+5.92	+**27.66**
Qwen	7B	+**17.00**	−0.95	−0.61	−0.76	+**2.44**
	14B	+10.96	+**5.13**	+0.14	−1.21	−1.21
	72B	+1.17	+0.49	+**1.14**	−**0.37**	+0.61

Table 17. Can few-shot prompting eliminate sycophancy? SFT denotes supervised fine-tuning, SPT denotes our proposed supervised pinpoint tuning and FS denotes few-shot prompting. Few-shot prompting provides limited gain on sycophancy evaluation metrics.

Model Family	**Sycophancy Evaluation Metrics**			
	Confidence	**Truthfulness**	**Acc. Before**	**Acc. After**
Llama-2-13B	0.08	18.89	48.96	30.34
+ SFT	61.55	84.06	34.27	32.12
+ SPT	**71.92**	**86.72**	46.99	**47.55**
+ FS	0.20	18.74	**50.98**	31.94
Qwen-14B	11.48	43.41	56.69	38.03
+ SFT	56.12	81.33	57.30	52.50
+ SPT	**67.08**	**86.46**	57.43	**55.18**
+ FS	7.22	76.80	**57.49**	54.05



(a) Direct effect of both heads and MLPs to output sycophancy (before SPT)



(b) Direct effect of each head to output sycophancy (after SPT)

Figure 7. More results on Llama-2-13B path patching

C.6. Comparison of computational efficiency

Table 1 presents the number of tuned parameters of SFT and SPT, *e.g.*, 13.02B and 0.17B for Llama-2-13B. The SPT yields comparable improvements with SFT with only $1/80$ tunable parameters. We also measure the training speed using the metric of samples processed per second (sam./sec.). The train speed of SFT and SPT on Llama-2-13B are 2.8 and 9.7 sam./sec, which on Qwen-13B are 2.3 and 8.1 sam./sec. The training speed of SPT is approximately $3\times$ faster than SFT. These results further verify the efficiency of the SPT over SFT.

Table 18. Model ensemble results. We supervised pinpoint tune Llama-2-13B-Chat on both eliminating the sycophancy dataset and the mathematical reasoning dataset. The weights of the models are then added together.

	Sycophancy Metrics				**Math Ability**	
	Confidence		**Truthfulness**		**GSM8K**	
Models	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**
Llama-2-13B	0.08	-	18.89	-	33.89	-
+ Syco SPT	71.92	+71.84	86.72	+67.83	35.48	+1.09
+ Math SPT	20.07	+19.99	23.78	+4.89	**43.52**	+9.63
+ Ensemble	**74.86**	+74.78	**87.69**	+68.8	40.41	+6.52

C.7. Assembling abilities from homologous models

We have revealed the sparse mechanism inside LLMs, which is attributed to the sycophancy. This finding raises a natural question: through the similar algorithm of SPT, are homologous LLMs with different abilities able to be merged into a single and comprehensive model? To answer this question, we will conduct a preliminary experiment to address the sycophancy and enhance math ability simultaneously. The results are shown in Table 18. Through identifying and pinpoint tuning the 64 math-related heads, the accuracy of "Math SPT" model on GSM8K increases from 33.89 to 43.52. The effectiveness of SPT is fully evident in the alteration of the modified parameters before and after training, referred to as *delta parameters*. Following Yu et al. (2023), we merge the "Syco SPT" and "Math SPT" models into one model by averaging the delta parameters and then applying them to the original model. Interestingly, the assembled model achieves 40.41 accuracy on GSM8K, which preserves 67% the improvement. Moreover, it leads to 74.78 confidence and 87.69 truthfulness on sycophancy metrics, which also outperforms the "Syco SPT". Our findings suggest that LLMs may have an internal compositional structure of the network architecture and shed light on merging multiple task-specific LLMs into a single LLM to continually improve the LLM with diverse abilities.

D. Examples of Model Outputs

To better demonstrate the model behavior before and after using pinpoint tuning to eliminate sycophancy, we show some multi-round chat examples in Table 19 and Table 20. We highlight the correct answer in green and the wrong answer in red.