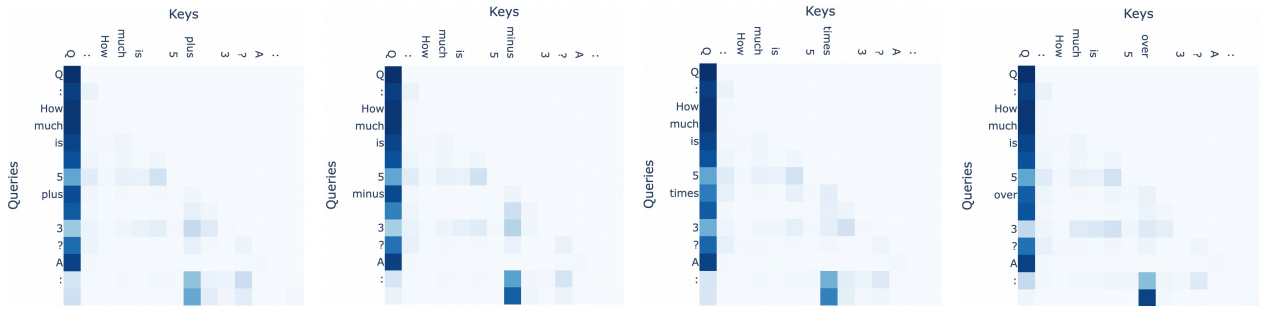
F. More Attention Pattern Cases.

We show the attention patterns of the operator-attended heads (*e.g.*, 14.2) in Figure 16 that could attend to the tokens of "plus", "minus", "times", and "over", across different sentences.



(a) Head 14.2 (b) Head 14.2 (c) Head 14.2 (d) Head 14.2

Figure 16: The attention patterns of the key head 14.2, which mainly attend to the operator-related tokens, *e.g.*, "plus", "minus", "times", "over".

G. Ablation Study of Precise SFT on MLPs.

We further investigate the influence of different number of tuned MLPs in Table 5. It reveals that the tuning more MLPs could lead to a performance decrease on MMLU and more training time, while imrove the performance on math dataset GSM8K.

Table 5: Ablative experiments on the number of tunable MLPs.

Precise SFT Setting	**Evaluation Metric**			
	Train Speed	Tunable Params.	GSM8K	MMLU
top-32 heads	50sam./sec.	0.067B	27.4	46.4
top-32 heads + top-1 MLP	44sam./sec.	0.202B	27.5	46.0
top-32 heads + top-2 MLPs	38sam./sec.	0.338B	27.7	45.7
top-32 heads + top-3 MLPs	31sam./sec.	0.473B	28.0	45.2
top-32 heads + top-6 MLPs	26sam./sec.	0.879B	28.2	44.9
top-32 heads + all MLPs	19sam./sec.	4.396B	29.2	43.9

H. Calculation in Computer vs LLMs.