Table 1: Overall performance. We evaluate the capabilities of LLaMA2-7B and LLaMA2-13B, transitioning from generic tasks (*e.g.*, MMLU and CSQA) to mathematical tasks (*e.g.*, GSM8K, AddSub, SingleEq, and SVAMP). Supervised fine-tuning across the entire parameter set (denoted as Full SFT) leads to enhanced performance in math-related tasks, albeit at the expense of its capabilities in generic tasks. In contrast, selectively tuning only the parameters of 32 critical attention heads (denoted as Precise SFT) yields comparable improvements while preserving the model's proficiency in generic tasks, with faster training speed (samples processed per second) and less tuned parameters.

			Mathematical Tasks								**Generic Tasks**			
			GSM8K		**AddSub**		**SingleEq**		**SVAMP**		**MMLU**		**CSQA**	
Models	**Train Speed**	**Tuned Params.**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**
LLaMA2-7B	-	-	14.6	-	30.5	-	65.4	-	34.7	-	46.0	-	59.8	-
+ Full SFT	15sam./sec.	6.7B	24.6	+10.0	53.7	+23.2	68.2	+2.8	50.3	+15.6	40.5	-5.5	54.0	-5.8
+ Precise SFT	50sam./sec.	0.07B	27.4	+12.8	50.6	+20.1	69.7	+4.3	55.8	+21.1	46.4	+0.4	59.6	-0.2
LLaMA2-13B	-	-	28.7	-	33.7	-	76.6	-	45.7	-	54.8	-	67.3	-
+ Full SFT	8sam./sec.	13.0B	44.6	+15.9	62.2	+28.5	79.8	+3.2	62.8	+17.1	50.2	-4.6	62.0	-5.3
+ Precise SFT	34sam./sec.	0.08B	46.3	+17.6	61.1	+27.4	82.2	+5.6	66.6	+20.9	55.0	+0.2	67.2	-0.1

at the top in layer 23. However, in subsequent layers, the correct answer '7' consistently remains top while '6' and '8' decline. It indicates that LLMs may do computations in a coarse-to-fine manner, where the result is firstly regressed to an embedding around that of the right answer, and then converges to the final output based on the fine-grained information introduced by subsequent MLPs.

Consolidating these findings, we can assert with some confidence that LLMs initially leverage attention heads to focus on operands ({A} and {B}) and the operator, relaying this information to downstream MLPs. Over time, the MLPs progressively bolster {C} and diminish the effect of confused answers, carrying out the calculation to final results.

5.3. Precise SFT on Calculation-related Components.

Experimental details. We evaluate precise SFT on four mathematical datasets (GSM8K (Cobbe et al., 2021), AddSub (Hosseini et al., 2014), SingleEq (Koncel-Kedziorski et al., 2015), SVAMP (Patel et al., 2021)), and another two datasets (MMLU (Hendrycks et al., 2020) and CSQA (Saha et al., 2018)) to evaluate the generic ability. During training, we optimize the key components only and leave the other components unchanged. We gather all training data from four mathematical datasets, and perform SFT updating on top 32 key heads. Following (Yu et al., 2023), the gradient is rescaled by $\frac{H}{h}$, where H is the number of all heads in each layer, h is the number of updated heads in each layer. In practice, we train LLaMA2-7B and LLaMA2-13B with a learning rate of 2×10^{-5} and a batch size of 128 for 2 epochs. The warm up ratio and weight decay are set as 0.02 and 0.1 by default, respectively. All experiments are conducted on 8 NVIDIA A100 80GB GPUs.

Table 2: Ablative experiments on the number of tunable components. The default setting is shown in gray.

Precise SFT Setting	**Evaluation Metric**			
	Train Speed	Tuned Params.	GSM8K	MMLU
top-8 heads	58sam./sec.	0.017B	25.4	45.1
top-16 heads	52sam./sec.	0.033B	26.5	45.8
top-32 heads	50sam./sec.	0.067B	27.4	46.4
top-48 heads	46sam./sec.	0.101B	27.4	46.4
top-64 heads	40sam./sec.	0.134B	27.3	45.5
top-32 heads + top-3 MLPs	31sam./sec.	0.473B	28.0	45.2

Precise SFT improves mathematical ability. Supervised Fine-Tuning (SFT) is an effective approach for augmenting the mathematical capabilities of models by fine-tuning all parameters within LLMs. We term this all-parameter fine-tuning as Full SFT for clarity, and adopt the same training settings as Precise SFT. Table 1 presents the results of Full SFT and Precise SFT on the LLaMA2-7B and LLaMA2-13B models. Precise SFT effectively bolsters their mathematical capabilities, yielding an averaged increase of 15% on four distinct mathematical datasets. It matches or even surpasses the improvements made by Full SFT. For example, Precise SFT outperforms Full SFT by 5.5% on the SVAMP dataset and 2.8% on GSM8K, underlining its superior ability to enhance the mathematical prowess of LLMs. Full SFT suffers from the trade-off between mathematical and general capabilities (about 5% drops on MMLU and CSQA), while Precise SFT effectively maintains the model's original performance. A further advantage of Precise SFT is the drastic