sentation can be characterized as an additive update influencing the evolving representation across vocabularies. Our methodology is aligned with these works, while we mainly focus on the token embeddings of right/wrong answers to reveal the contribution of MLPs on the calculation tasks.

4.3. Precise Fine-tuning.

Supervised Fine-Tuning (SFT) is widely used for enhancing a model's mathematical capabilities. Building on this, precise SFT only updates those components closely associated with mathematical abilities, while keeping the rest parameters unchanged. Algorithm 2 illustrates the whole process. For the i-th attention layer, the output matrix W_O^i is split into equal size blocks for each head $\left[W_O^{i,1}, W_O^{i,2}, \cdots W_O^{i,H}\right]$. As

Algorithm 2 Precise Fine-tuning

Require: Model $\mathcal{M}$, input X, index of key heads $\mathbf{\Phi}$, iterations I, learning rate η, $W_\theta = W_{Q/K/V/O}$
for $(i, j) \in \mathbf{\Phi}$ **do**
  $W_\theta^{i,j}.requires\_grad = True$
end for ▷ activate key heads
loop I times
  $\mathcal{L}$ = $\mathcal{M}$.forward(X)
  $\mathcal{L}$.backward()
  for $w \in W_\theta$ **do**
    $w = w - \eta * w.grad$
  end for ▷ update target parameters
end loop

is verified in (Elhage et al., 2021), it is equivalent to running heads independently, multiplying each by its own output matrix, and adding them into the residual stream. For the selected individual heads, precise SFT updates the parameters of four matrices: $W_Q^{i,j}$, $W_K^{i,j}$, $W_V^{i,j} \in \mathbb{R}^{d \times \frac{d}{H}}$, and $W_O^{i,j} \in \mathbb{R}^{\frac{d}{H} \times d}$. For the selected MLP layer, precise SFT updates all parameters in this layer. Moreover, since we adjust only a small fraction of the parameters, precise SFT naturally benefits from shorter training times and minimal impact on the model's original capabilities.

5. Experiments

The experiments are organized as follows: (1) *identify* the calculation-related key components via path patching and *validate* their importance in implementing arithmetic calculation via knockout in Section 5.1; (2) *understand* the behavior of the newly identified components by examining their attention patterns and embeddings in Section 5.2; (3) *improve* the mathematical capability via precise supervised fine-tuning on math benchmarks in Section 5.3. For simplicity, we primarily report the results of LLaMA2-7B, while the results of other models can be found in Appendix.

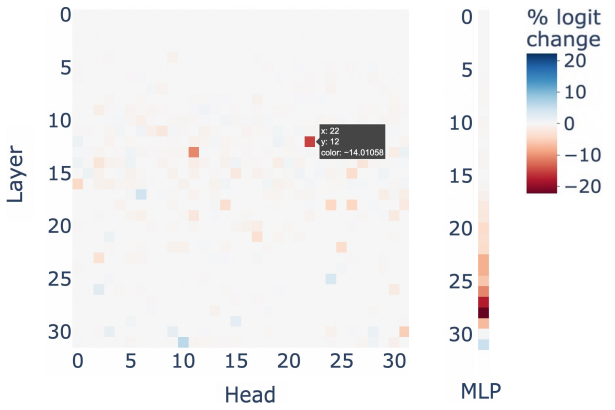

Figure 2: We conduct path patching experiments on LLaMA2-7B across four mathematical tasks, by searching for each head and MLP directly affecting the logit of the right answer. For each head/MLP, a darker color indicates a larger logit difference from the model before patching.

5.1. Identifying Calculation-related Components.

Location of key heads. In Figure 2, we visualize the effect of each head according to the serial numbers of the heads and layers. This arrangement allows for a clear comparison of the causal impact of each head to the logit of ground-truth token {C}. The red squares indicate heads that have a significant positive impact on predicting the output token, while the blue squares represent heads that have a negative effect. From these results, we observe that: (i) *Only a small number of heads have a noteworthy influence on the output*. Specifically, when the heads such as 12.22[3] is patched, there is a substantial decrease of 14.0% on the logit of token {C}, which highlights their positive contribution to the calculation tasks. We classify heads that exhibit logit change exceeding −5% as "key heads". The sparse distribution of these key heads motivates us to explore their specific functionalities and characteristics in Section 5.2. (ii) *The discovered key heads are mainly located in the middle layers*. For LLaMA2-7B, key heads emerge starting from the 12th layer for all arithmetic calculations. Prior layers exhibit heads that do not exert a direct effect on the output logits. Key heads are primarily concentrated between layers 12 and 17. (More analysis of the key heads in other LLMs can be found in Appendix C.)

Location of key MLPs. The last column in Figure 2 visualizes the effect of each MLP layer on the logit of ground-truth token {C}. It is observed that MLPs before the identified heads (0−16) have almost no impact on the outputs (approximately ±0.0%). In contrast, after the 17-th layer, MLPs exhibit a much larger effect (approximately ±10.0%). It indicates that MLPs are engaged in the calculation. We hy-

[3]We apply the notation of $i.j$ to refer to the j-th head of the i-th attention layer.