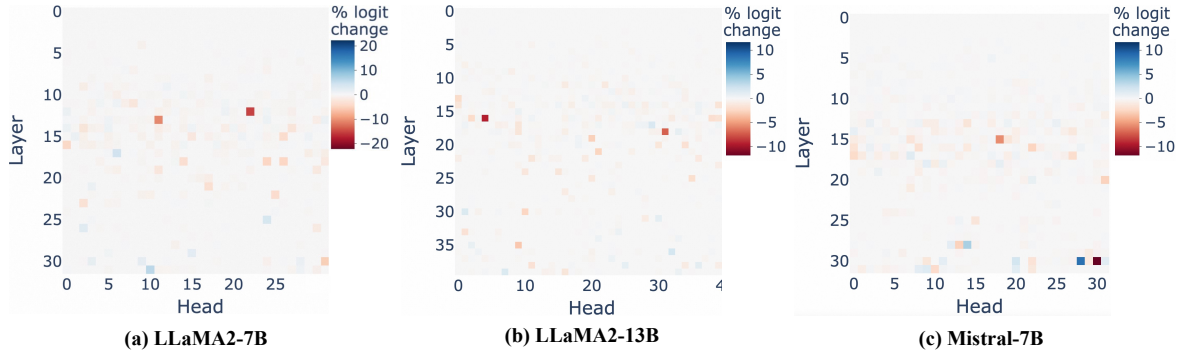
$0.31 \rightarrow$ Output. Firstly, given reference data X_{r} and counterfactual data X_{c}, the activations of all heads are gathered for preparation of the later perturbation. Then, we do a hard intervention on the Head 0.31 that is perturbated to its activation on X_{c}, where the effect will be further propagated to the Ouput node along with a set of paths $\mathcal{P}$. To ensure an independent observation of the impact from the Head 0.31, $\mathcal{P}$ comprises the forward pathways through residual connections and MLPs except for the other attention heads (*e.g.*, Head $0.0, \cdots, 0.30, 1.0, \cdots, 1.31$). Thus we do a hard intervention on the other heads by freezing their activations on X_{r}. Finally, we obtain the final output logits to measure the impact of this perturbation. If there is a significant change in final logits, then the patched paths: Sender $\rightarrow$ Receiver are essential for the model in completing the task.

In this work, to identify the important heads contributing to the calculation task, we scan through all heads as the Sender node denoted by h, and set the Receiver node as output $logits$, and measure the changes in the output logit of ground-truth token $\{\mathtt{C}\}$. Pathways $h \rightarrow logits$ that are critical to the model's computation should induce a large drop in the logit of token $\{\mathtt{C}\}$ after patching. Notably, since the residual operations and MLPs compute each token separately (Elhage et al., 2021), patching the head output at the END position (*i.e.*, the position of the last token in the input sentence) is enough to measure the effects on the next token prediction.

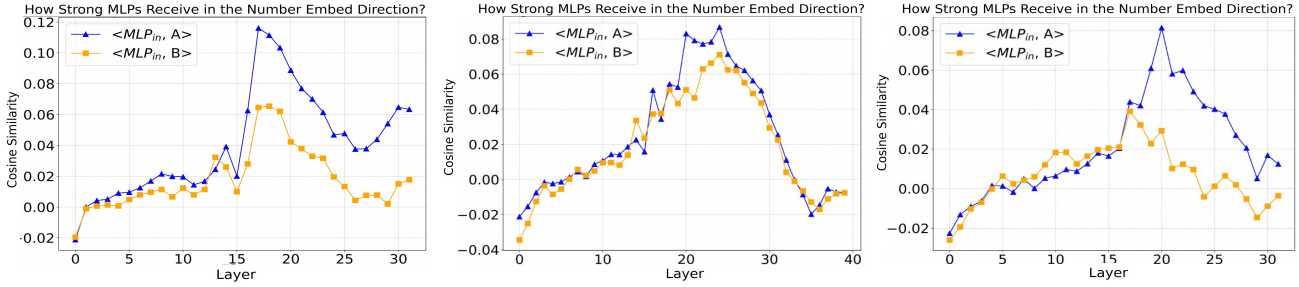
C. More Results of Other LLMs.



(a) LLaMA2-7B (b) LLaMA2-13B (c) Mistral-7B

Figure 12: Comparison of the results of path patching experiments on LLaMA2-7B, LLaMA2-13B, and Mistral-7B (Jiang et al., 2023) across four mathematical tasks. For each head/MLP, a darker color indicates a larger logit difference from the original model before patching.

Key Component Identification. In Figure 12, we further report the results of key components identification of other models (*e.g.*, LLaMA2-13B and Mistral-7B). For example, LLaMA2-13B comprises 40 layers and 40 attention heads per attention layer. The three models of different size exhibit similar phenomena that the calculation-related key heads (*e.g.*, 16.4, 18.31) are distributed sparsely in the middle layers.



(a) LLaMA2-7B (b) LLaMA2-13B (c) Mistral-7B

Figure 13: We investigate the projection of each MLP layer input (MLP_{in}) along the direction of number token $\{\mathtt{A}\}$, $\{\mathtt{B}\}$, respectively.

Key MLPs Behavior. In Figure 13, the similarities of MLP input and number operands $\{\mathtt{A}\}$/$\{\mathtt{B}\}$ across all models demonstrate ascending and descending trends. Specifically, the pivotal points for these trends, delineated as (*start-*