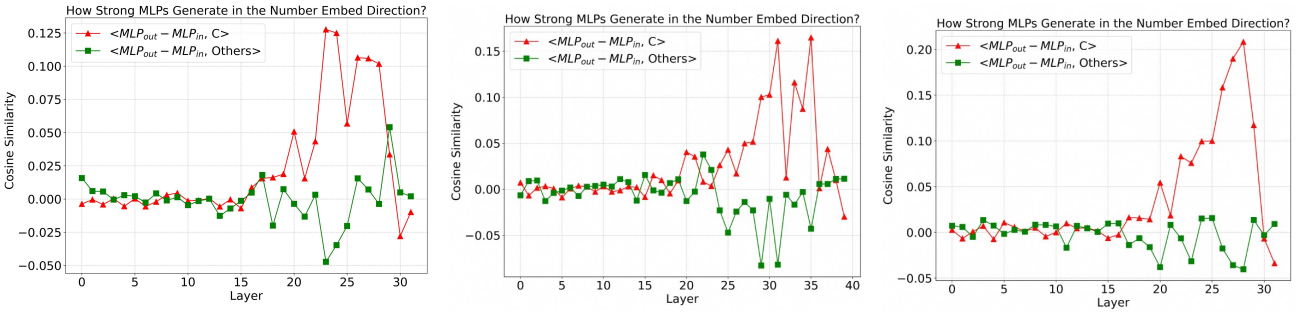


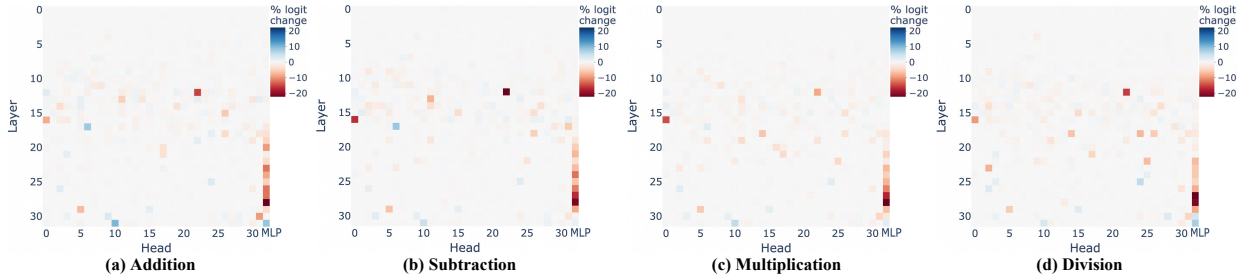
(a) LLaMA2-7B (b) LLaMA2-13B (c) Mistral-7B

Figure 14: We investigate the projection of each MLP layer (MLP_{out}-MLP_{in}) along the direction of number token {C} (*i.e.*, right answer) and other tokens (*i.e.*, wrong answer).

$inflection\text{-}end$), are as follows: (13-18-28) for LLaMA2-7B, (13-18-35) for LLaMA2-13B, and (13-20-28) for Mistral-7B. In Figure 14, the similarities of MLP_{out}-MLP_{in} and right answer {C} show a pattern of initial stabilization followed by an increase. The critical points for LLaMA2-7B/LLaMA2-13B/Mistral-7B are again (13-18-28), (13-18-35), and (13-20-28). The inflection points in both Figure 13 and Figure 14 are nearly identical, indicating consistent trend shifts across the models. It helps to verify that LLMs initially leverage attention heads then relaying information to downstream MLPs, to progressively carry out the calculation to final results. Furthermore, the above findings appear to be general and robust across different LLMs, not limited to a specific model.

D. Key Component Location across Calculation Tasks.

We investigate the location of key components for each calculation task individually, as shown in Figure 15. The discovered key heads could be shared across four tasks, which are sparsely distributed in the middle layers. Specifically, when examining subtraction and addition tasks, we could summarize two insightful symmetries between them. The identified key heads of two tasks are almost the same, albeit with different magnitude of the effect. This phenomenon could reveal the symmetry of key head "*location*" in addition and subtraction. Moreover, the tasks of multiplication and division exhibit a greater number of key heads compared to the tasks of addition and subtraction. We assume it could be attributed to their more intricate operations within multiplication and division.



(a) Addition (b) Subtraction (c) Multiplication (d) Division

Figure 15: 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. The last column denotes the path patching results of MLPs. For each head/MLP, a darker color indicates a larger logit difference from the original model before patching.

Generalize to other calculation operations. We conduct the experiments of key head identification and validation following Section 5.1. We generate the samples including the exponentiation operation as the reference data X_{r}. Then we generate the counterfactual data X_{c} following the principles introduced in Section 4.1 to exclude the exponentiation logic.

The results reveal the potential of generalizing to more complex mathematical operations: (i) Five key heads are identified based on the newly generated X_{r} and X_{c}. We find that the heads (11, 8) and (14, 2) mainly attend to the operators "^", "power", while the heads (12, 22), (13, 11), (15, 15) mainly attend to the input operands {A} and {B}. (ii) Knocking out the key heads, identified by both templates, leads to significantly impacts (over 60%) on model performance.