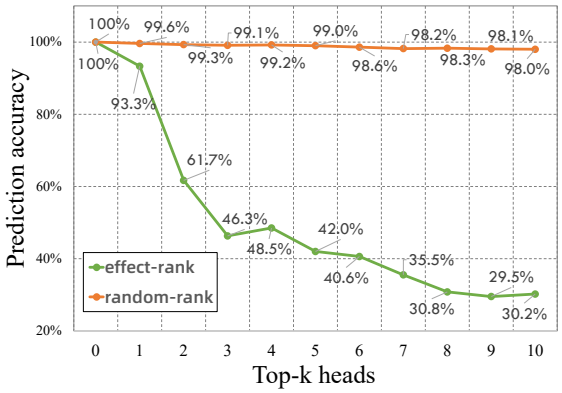


Figure 3: The influence on prediction accuracy after knocking out top-k attention heads that are sorted by the effect of each head on logits ("effect-rank"), and knocking out randomly-sorted top-k heads ("random-rank").

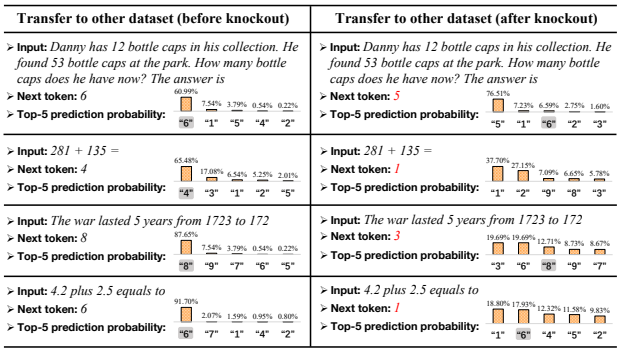
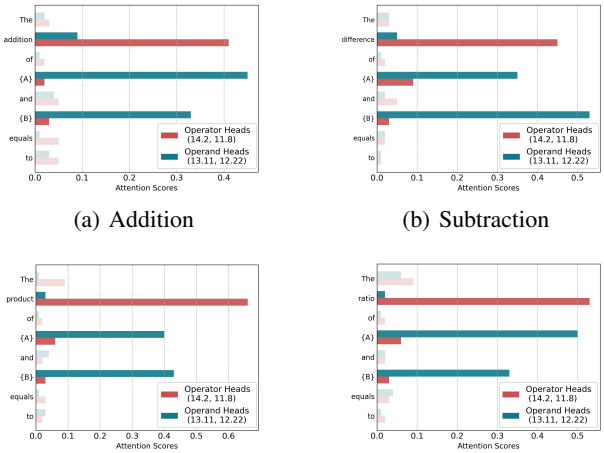


Figure 4: After knocking out the key heads, LLaMA2-7B predicts incorrectly on the cases of SVAMP dataset and other data formats of multi-digit integers, rational numbers.

pothesize that the calculation process is firstly implemented through the key heads, then the subsequent MLPs gradually work out the final results. We validate this in Section 5.2.

Validation of key components. To fully validate the faithfulness of the discovered key heads, we perform additional checks by observing the performance drop when knocking out these components. In Figure 3, all heads are sorted in a certain order by the importance score shown in Fig. 2 and knocked out one by one. It shows that, as the heads are gradually knocked out, the performance of the model drops sharply in "effect-rank", while keeping stable (relatively minor effect within 2%) in "random-rank". We also exhibit the transferability of the key heads with different data prompts or formats as shown in Figure 4. The model becomes largely confused to output incorrect numbers after knocking out the identified key heads. On the dataset SVAMP, there is a relative performance drop ($-22.9\%/34.7\%=-66.0\%$) after the knockout, aligned with the result on our generated dataset. The above results demonstrate that the discovered compo-



(a) Addition (b) Subtraction

(c) Multiplication (d) Division

Figure 5: The attention score distribution of key heads across four calculation tasks. The key heads (*e.g.*, 13.11, 14.2) attend to number operands and calculation operators.

nents play an important role in the language model's ability to complete the calculation task.

5.2. Understanding Calculation-related Component Behaviors.

Key heads behavior. In order to better understand the "behavior" of the heads that have a significant impact on calculation, we begin by analyzing their attention patterns, and check the attention scores between Query END token and each Key token as illustrated in Sec. 4.2. Our findings reveal that these heads exhibit a strong focus on tokens of operands or operators. For example, heads 13.11 and 12.22 have high attention scores on numbers including {A} and {B}, while heads 14.2 and 11.8 attend more to symbols or text indicating operations like "+", "-", "plus", "div", etc. We randomly select 1000 samples from reference data and plot the distribution of averaged attention scores on key heads (arranged in two groups) for four arithmetic calculations.

As illustrated in Figure 5, the operand heads and the operator heads are colored in red and green respectively, and highlighted at the positions of operands and operators. It is clear that these heads exhibit distinctly different distributions and show minimal attention to tokens outside of the operands/operators. Moreover, we visualize the attention patterns of the key heads (*e.g.*, 13.11) on various types of sentences in Figure 7. It reveals that the key heads also primarily prioritize number operands (*e.g.*, '1' and '5' in the first case) even for unseen data formats. This observation provides an explanation for why the deactivation of the key heads can influence the model's perception on number tokens and consequently affect its prediction when transfer-