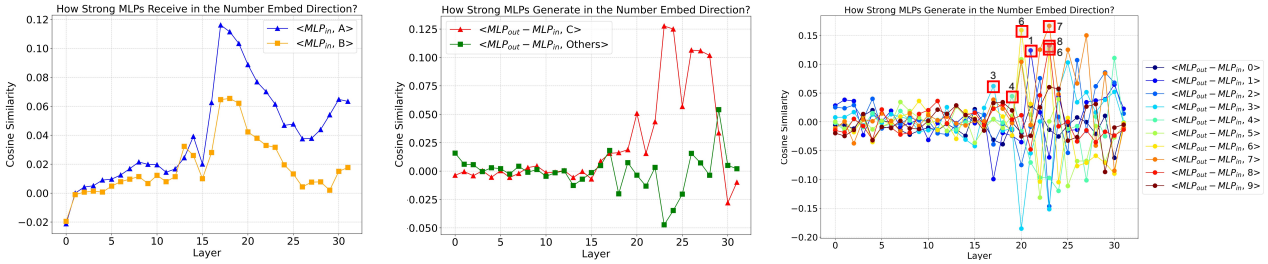
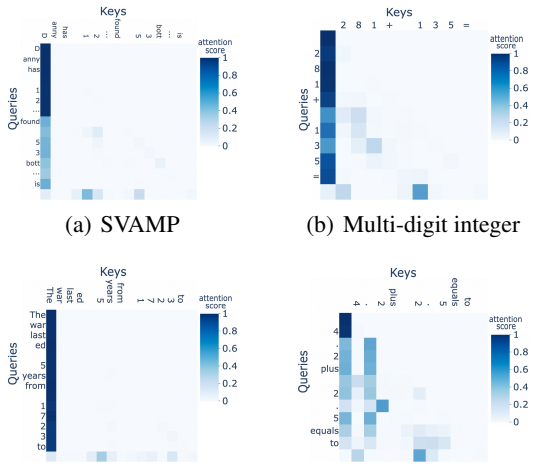


(a) Reception of {A} and {B} (b) Generation of {C} (c) Steps of Calculation

Figure 6: We investigate the projection of each MLP layer input or output along the direction of number token {A}, {B}, and {C}, respectively. The x-axis represents the layer number, ranging from 0 to 31, while the y-axis represents the cosine similarity between the embeddings of the MLP input or output and the number tokens.

ring to other datasets (shown in Figure 4). For more case studies on the key heads, such as the attention pattern on operators, please refer to Figure 16 in Appendix F.



(a) SVAMP (b) Multi-digit integer

(c) Diff. data format (d) Rational numbers

Figure 7: The transferability of attention patterns in key heads on the unseen samples in Figure 4, which mainly attend to the number operands.

Key MLPs behavior. In Figure 6(a), we conduct an initial investigation of the similarities between the MLP_{in} and tokens {A} and {B} over 1000 samples, to verify the information of operands received from above analyzed attention heads. For the 0–12th layers, both $\langle MLP_{in}, \{\texttt{A}\}\rangle$ and $\langle MLP_{in}, \{\texttt{B}\}\rangle$ are close to zero. It indicates no operands are captured during this stage, which corresponds to the blank region (*i.e.*, few key heads for computation task) before the 12th layer in Figure 2. For the 12–17th layers, we observe a sharp increase in the similarities with both operands ({A} and {B}). This surge corresponds to the presence of key attention heads, *e.g.*, 12.22/13.11 in layer 12/13, indicating that the operands are progressively being collected and "written" into the MLPs of these layers for subsequent computations. In layers 17–31, the similarities $\langle MLP_{in}, \{\texttt{A}\}\rangle$ and $\langle MLP_{in}, \{\texttt{B}\}\rangle$ gradually decrease, signifying the transition into a new stage that digests the input information for generating the answers.

To understand how each MLP layer contributes to generating the correct answer {C}, we compute the similarity between token {C} and the input/output of the MLPs. We use $\langle MLP_{out} - MLP_{in}, W_U[\{\texttt{C}\}]\rangle$ to reflect the direct contribution of the MLP to the correct answer, and $\langle MLP_{out} - MLP_{in}, W_U[Other]\rangle$ for other candidate numbers (as shown in Figure 6(b)). Starting from the 17th layer, where the MLPs begin processing operand information, we observe a noticeable increase in $\langle MLP_{out} - MLP_{in}, W_U[\{\texttt{C}\}]\rangle$ and a decrease in $\langle MLP_{out} - MLP_{in}, W_U[Other]\rangle$. This trend indicates that these MLPs are gradually carrying out the calculation required for the correct answer. The above ascending and descending trends can also be viewed in other LLMs as in Figure 13 and Figure 14 in Appendix C.

Based on the above analyses, we further delve into the detailed calculation process from layer 17 to 28. We investigate a case of "$4 + 3 =$" and analyze $MLP_{out} - MLP_{in}$ compared to all numeric tokens in Figure 6(c). At layers 17 and 19, the numbers '3' and '4' are at the top, indicating that MLPs receive and store input {A} ='4' and {B} ='3', respectively. After that, the numbers '6' and '1' appear top at the subsequent layers 20 and 21. In summary, the LLM predicts the next token as '7' in a single inference. However, within the LLM's architecture, the answer '7' is the result of a collaborative process across multiple layers 22/23/25/27, after the layers 17/19/20/21 generate '3'/'4'/'6'/'1', respectively. The results demonstrate that the answer '7' is not deduced directly, and MLPs perform calculations in a "layer-by-layer" manner, somewhat akin to the addition process in computers (a comparison of these two processes are presented Appendix H). Additionally, we observe that numbers close to the correct answer, such as '6' and '8', also appear