INTERPRETING THE INNER MECHANISMS OF LARGE LANGUAGE MODELS IN MATHEMATICAL ADDITION

Anonymous authors
Paper under double-blind review

1 APPENDIX

A DATASET TEMPLATES

We have included a comprehensive list of the templates used in this work as shown in Figure 1. Each name was randomly selected from a pool of 100 English first names, while the objects, verbs, and events were chosen from a curated list of 20 common words. For the datasets used in Section 4.4, we construct the sentences in different contexts with addition and subtraction logic, as shown in Figure 2 and Figure 3.

The <EVENT> <VERB> {A} years from the year <YYY>{B} to the year <YYY>{C}
The <EVENT> <VERB> {A} years from <YYY>{B} to <YYY>{C}
The <EVENT> <VERB> {A} days from <MONTH> {B} to <MONTH> {C}
The <EVENT> will <VERB> {A} days from <MONTH> {B} to <MONTH> {C}
The <EVENT> <VERB> {A} hours from {B} pm to {C}
The <EVENT> will <VERB> {A} hours from {B} pm to {C}
The <EVENT> <VERB> {A} hours from {B} am to {C}
The <EVENT> will <VERB> {A} hours from {B} am to {C}
{A} plus {B} equals to {C}
{A} plus {B} is equal to {C}
<A1>{A} plus <B1>{B} equals to <C1>{C}
<A1>{A} plus <B1>{B} is equal to <C1>{C}
{A} add {B} equals to {C}
{A} add {B} is equal to {C}
<A1>{A} add <B1>{B} equals to <C1>{C}
<A1>{A} add <B1>{B} is equal to <C1>{C}
<NAME> has {A} <OBJECT>, then <NAME> <VERB> {B} <OBJECT>. What's the total number of <OBJECT> that <NAME> has? The answer is {C}
<NAME> <VERB> {A} <OBJECT>, and <NAME2> <VERB> {B} <OBJECT>. What's the total number of <OBJECT> that they <VERB>? The answer is {C}
<NAME> has {A} <OBJECT>, and <NAME2> has {B} <OBJECT>. What's the total number of <OBJECT> that they have? The answer is {C}
<NAME> <VERB> {A} <OBJECT> yesterday, and <NAME> <VERB> {B} <OBJECT> today. What's the total number of <OBJECT> that <NAME> <VERB>? The answer is {C}

Figure 1: Templates used in the addition dataset. All templates in the table involve the addition logic "{A} + {B} = {C}", but have different linguistic meanings like "time span", "number calculation", and "object accumulation".

{A} + {B} = {C}
<A1>{A} + <B1>{B} = <C1>{C}
The addition of {A} and {B} is {C}
The addition of <A1>{A} and <B1>{B} is <C1>{C}
The sum of {A} and {B} is {C}
The sum of <A1>{A} and <B1>{B} is <C1>{C}

Figure 2: Templates used in the dataset when transferring to *unseen* addition task.

{A} - {B} = {C}
<A1>{A} - <B1>{B} = <C1>{C}
From the year <YYY>{B} to the year <YYY>{A} , the <EVENT> <VERB> {C}
From <YYY>{B} to <YYY>{A} , the <EVENT> <VERB> {C}
{A} minus {B} equals to {C}
{A} minus {B} is equal to {C}
<NAME> has {A} <OBJECT>, then <NAME> <VERB> {B} <OBJECT>. What's the total number of <OBJECT> that <NAME> has? The answer is {C}
<NAME> had {A} <OBJECT> yesterday, then <NAME> <VERB> {B} <OBJECT> today. What's the total number of <OBJECT> that <NAME> has? The answer is {C}

Figure 3: Templates used in the dataset when transferring to subtraction task. All templates in the table imitate the addition templates but involve the subtraction logic "{A} - {B} = {C}".

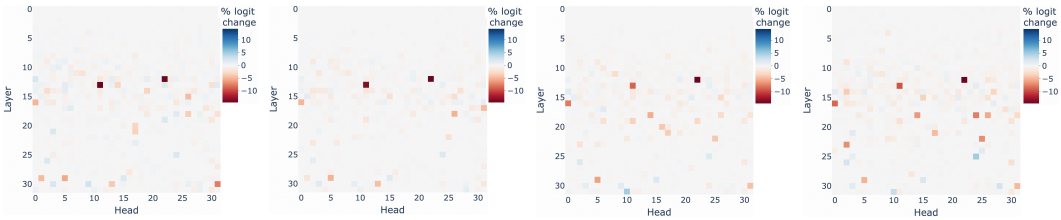
B Comparison of Key heads on four mathematical tasks.

To investigate the distribution of key heads across four mathematical tasks (*i.e.*, addition, subtraction, multiplication, and division), we conduct the path patching experiments using the templates in Figure 4. The results shown in Figure 5 reveal that: (i) the sparsity of key heads remains consistent across all four tasks (less than 1.0% of all heads). (ii) The key heads mainly distribute in the middle layers. The phenomena are analogous to the primary findings on the addition task (Section 4.1), demonstrating the potential of extending the observed effects of the addition task to other mathematical tasks.

We compare the location of key heads across four mathematical tasks. An interesting finding is that the key heads used in "subtraction" and "addition" tasks overlapped significantly, as did the key heads used in "multiplication" and "division" tasks. Moreover, the four tasks share the heads (*e.g.*, 13.11 and 12.22) that deliver the most significant effects, while they have task-specific heads that only emerge in its own task. These findings suggest that LLMs exhibit behavior aligned with human thinking to some extent, since "subtraction-addition" and "multiplication-division" are opposite mathematical operations.

Addition	**Subtraction**
{A} + {B} =	{A} - {B} =
The sum of {A} and {B} is	The difference between {A} and {B} is
Q: What is {A} plus {B}? A:	Q: What is {A} minus {B}? A:
Q: How much is {A} plus {B}? A:	Q: How much is {A} minus {B}? A:
Multiplication	**Division**
{A} * {B} =	{A} / {B} =
The product of {A} and {B} is	The ratio between {A} and {B} is
Q: What is {A} times {B}? A:	Q: What is {A} over {B}? A:
Q: How much is {A} times {B}? A:	Q: How much is {A} over {B}? A:

Figure 4: Templates used to investigate the mathematical tasks of addition, subtraction, multiplication, and division. "Q" and "A" are the abbreviation for "Question" and "Answer", respectively.



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

Figure 5: We conduct path patching experiments on LLaMA2-7B across four mathematical tasks, by searching for each head h directly affecting the logit of the right answer.

Interestingly, when examining subtraction and addition tasks, we could summarize two insightful symmetries between them. (i) 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. (ii) These heads particularly attend to the number tokens regardless of whether they are given addition or subtraction sentences (shown in Section 4.4). This phenomenon could reveal the symmetry of key head "*behavior*" in addition and subtraction.

C Attention patterns in another two language models.

In Figure 6 and Figure 7, we list the attention patterns of the identified key heads (*e.g.*, 19.15 in chatGLM2-6B, 18.4 in Qwen-7B) on different samples. Whether the sentences containing the addition logic (samples 1-3), the subtraction logic (samples 4-6), or even sentences without linguistic meanings (samples 7-9), the key heads have particularly higher attention scores on number tokens in the sentences. These results further validate the claimed functionality of these key heads.

D Knockout results of another two language models.

In Figure 8, we measure the prediction accuracy after knocking out the identified key heads in different language models. As we increase the number of knocked key heads, the model's performance experiences a significant decline followed by a gradual stabilization. This pattern is consistent across three different models, providing further evidence that the identified key heads play a crucial role in completing the addition task.

E Comparison of key heads in different LLMs.

In Figure 3, we present the identified key heads in different LLMs of LLaMA2-7B, Qwen-7B and chatGLM2-6B. Despite the sparsity of key heads compared to the total number of heads (*e.g.*, 1024 in LLaMA2-7B and Qwen-7B, 896 in chatGLM2-6B), LLaMA2 exhibits a relatively denser distribution of results compared to the other two models. Apart from the key heads 13.11 and 12.22, other heads such as 16.0, 14.19, and 15.15 also make a difference on the output logits. It indicates that LLaMA2-7B involves more components in completing the addition task. We hypothesize that this is because LLaMA2-7B demonstrates superior comprehension abilities for mathematical tasks, resulting in more obvious response patterns. To assess the model's proficiency in understanding mathematical tasks, we conducted experiments to evaluate whether the model comprehends the aims of the mathematical task by generating higher logits for numerical tokens. Based on the reference data X_{r}, we compute the average prediction probability P_{avg} of numerical tokens 1-9 for different models (LLaMA2-7B: 99.26% vs. Qwen-7B: 97.55% vs. chatGLM2-6B: 94.68%). The higher P_{avg} of LLaMA2-7B shows that it generates number tokens with a greater confidence level, thus demonstrating a better understanding of the mathematical task. This finding provides an explanation of why LLaMA2-7B involves more key heads. Further investigation is required for a more comprehensive analysis.

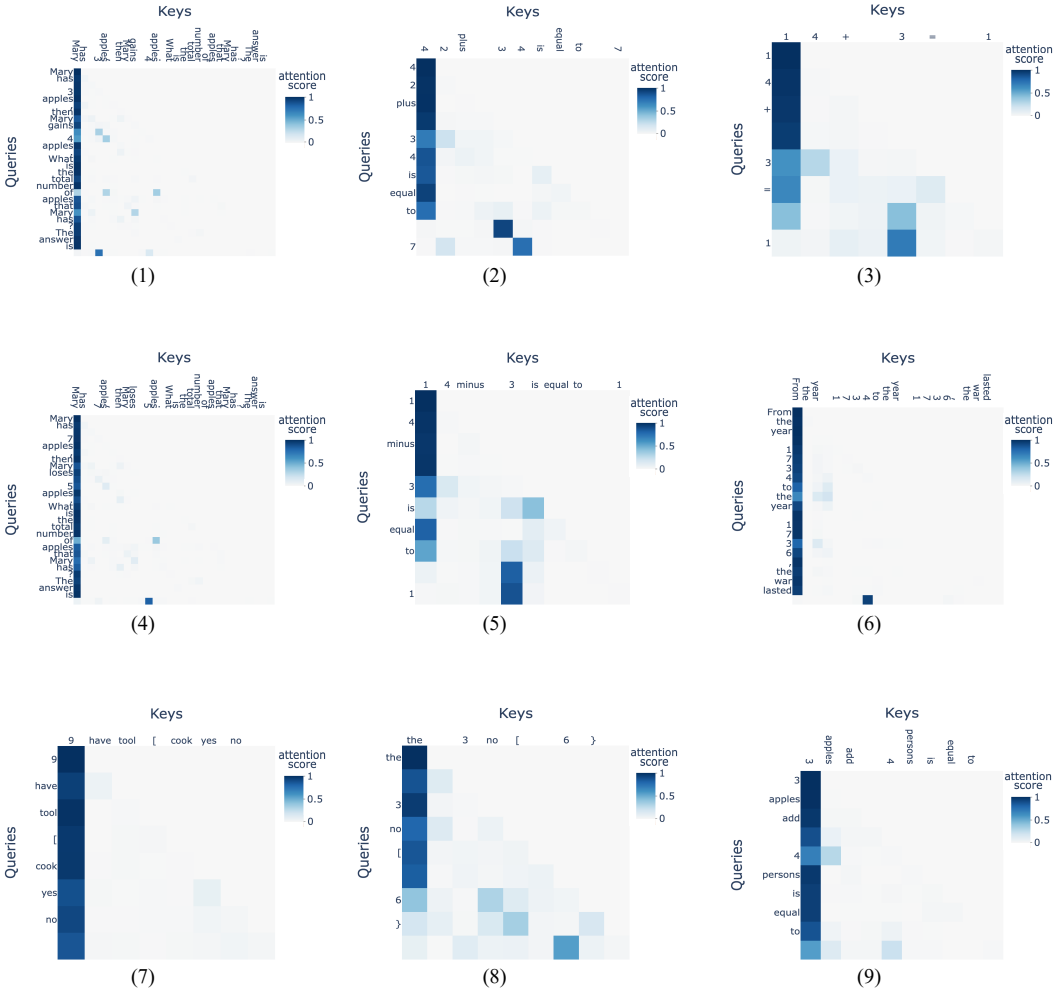


Figure 6: The attention patterns of head 18.4 in Qwen-7B on different samples. The samples (1-3) are sentences with the addition logic. The samples (4-6) are sentences with the subtraction logic. The samples (7-9) are randomly constructed sentences with no linguistic meaning.

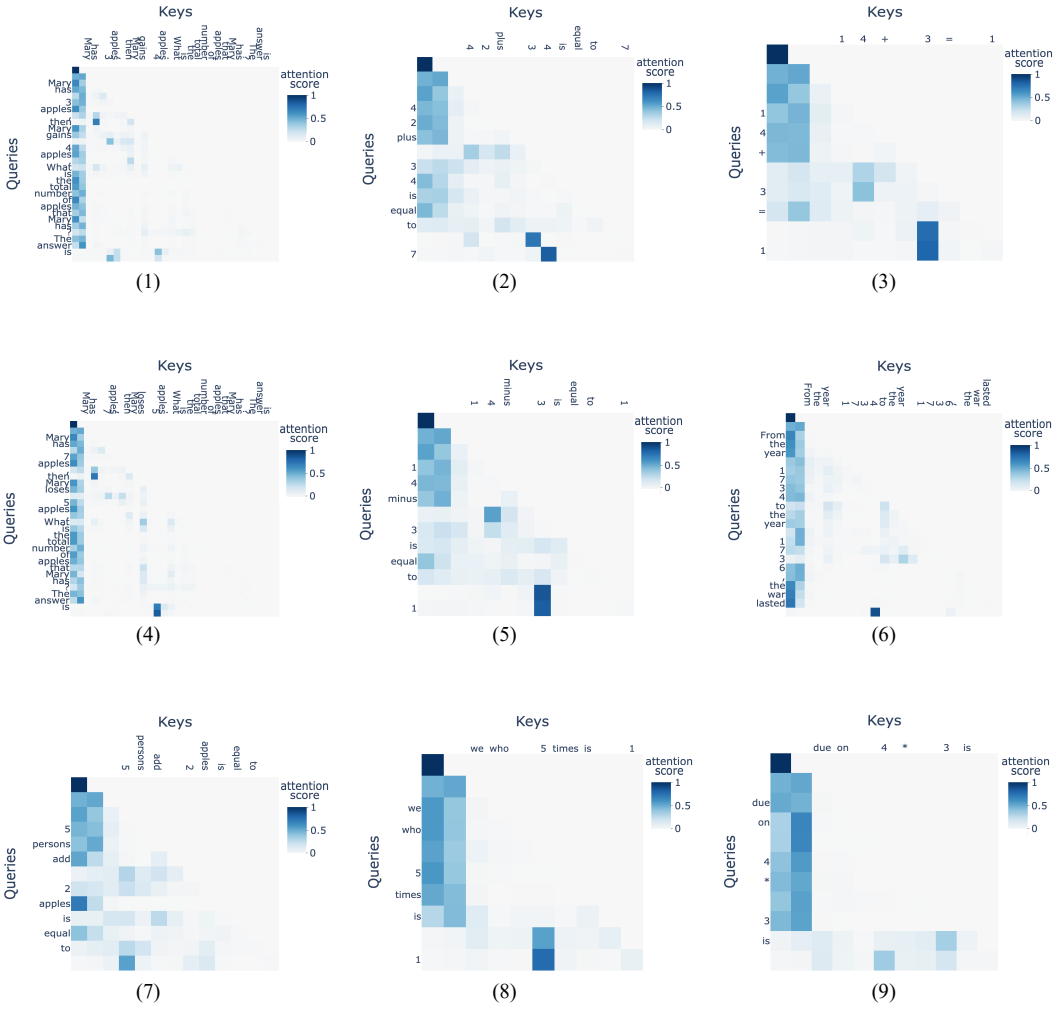


Figure 7: The attention patterns of head 19.15 in chatGLM2-6B on different samples. The samples (1-3) are sentences with the addition logic. The samples (4-6) are sentences with the subtraction logic. The samples (7-9) are randomly constructed sentences with no linguistic meaning.

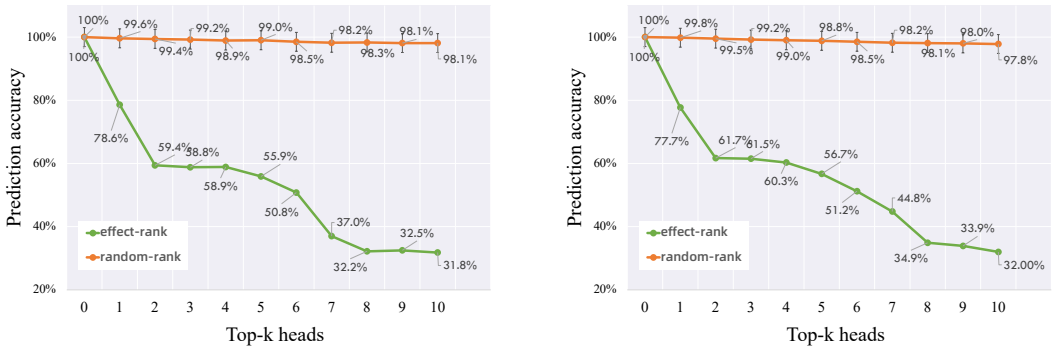


Figure 8: The influence on prediction accuracy after knocking out top-k attention heads in the language model of Qwen-7B and chatGLM2-6B. The heads are sorted by the effect of each head on logits ("effect-rank"), or randomly sorted ("random-rank").

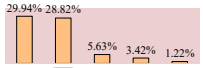
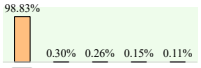
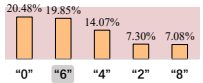
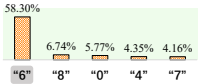
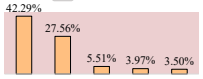
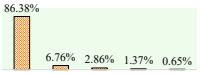
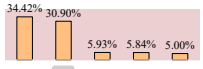
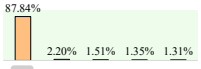
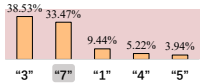
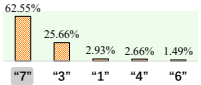
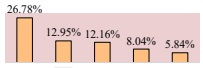
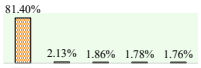
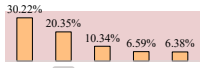
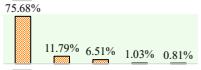
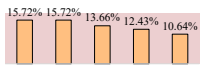
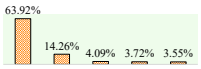
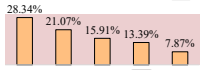
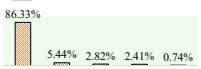
Experiment setting	Reference data	Top-5 prediction probability (after knockout)	Top-5 prediction probability (before knockout)
Test on seen addition task	➢ **Input:** *The conference will outlast 2 hours from 3 pm to* ➢ **Next word (knockout):** *6*		
	➢ **Input:** *42 plus 34 is equal to 7* ➢ **Next word (knockout):** *0*		
	➢ **Input:** *Mary has 3 apples, then Mary gains 4 apples. What is the total number of apples that Mary has? The answer is* ➢ **Next word (knockout):** *1*		
Transfer to unseen addition task	➢ **Input:** *14 + 3 = 1* ➢ **Next word (knockout):** *6*		
	➢ **Input:** *The addition of 3 and 4 is* ➢ **Next word (knockout):** *3*		
Transfer to subtraction task	➢ **Input:** *14 - 3 = 1* ➢ **Next word (knockout):** *0*		
	➢ **Input:** *From the year 1734 to the year 1736, the war lasted* ➢ **Next word (knockout):** *3*		
	➢ **Input:** *7 minus 3 is equal to* ➢ **Next word (knockout):** *2*		
	➢ **Input:** *Mary has 7 apples, then Mary loses 3 apples. What is the total number of apples that Mary has? The answer is* ➢ **Next word (knockout):** *3*		

Figure 9: When testing on the reference data of seen addition tasks, unseen addition tasks and subtraction tasks, Qwen-7B provides incorrect predictions after knocking out key heads.

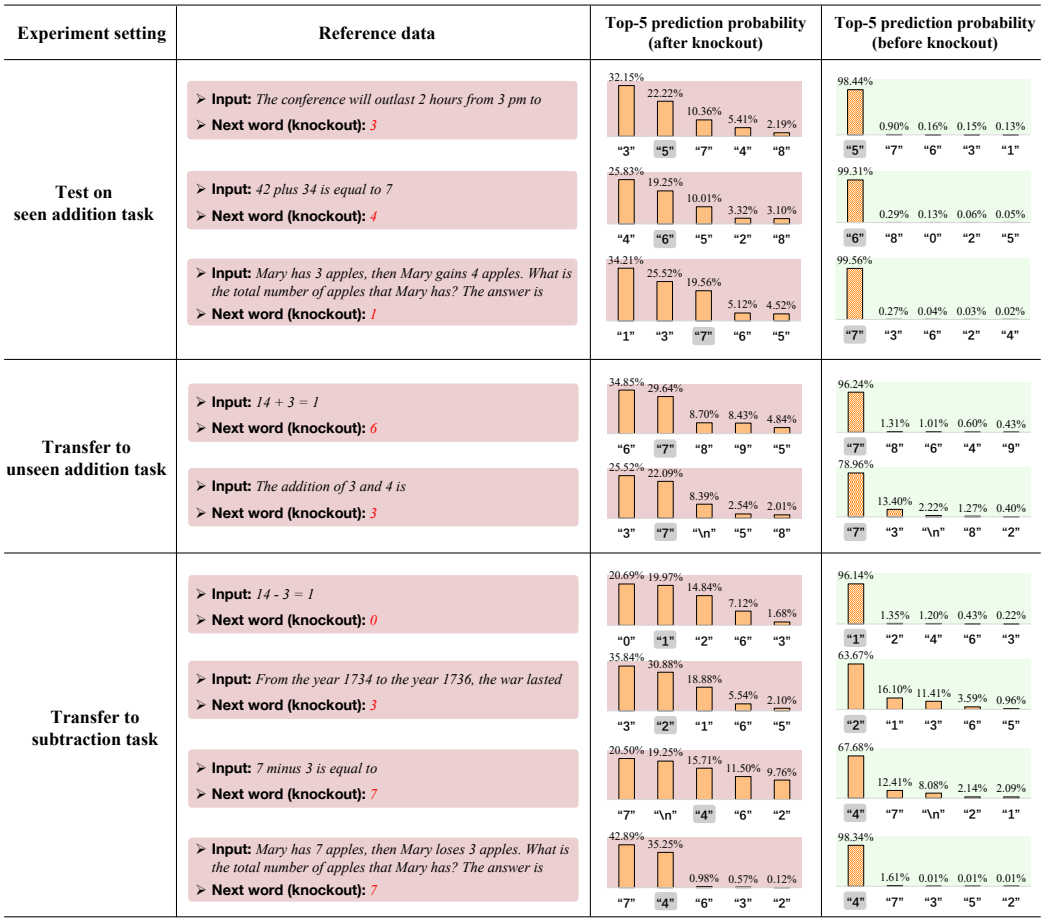


Figure 10: When testing on the reference data of seen addition tasks, unseen addition tasks, and subtraction tasks, chatGLM2-6B provides incorrect predictions after knocking out key heads.