Table 3: Key head identification on the exponentiation task.

Templates	Key Heads [*e.g.*, (Layer, Head)]	Knockout Accuracy
X_{r}: "{A} ∧ {B} = _ " X_{c}: "{A} < {B} = _ "	[(11, 8), (12, 22), (13, 11), (14, 2), (15, 15)]	−66%
X_{r}: "{A} to the **power** of {B} equals _ " X_{c}: "{A} to the **none** of {B} equals _ "	[(11, 8), (12, 22), (13, 11), (14, 2), (15, 15)]	−62%

E. Generalize to More Complex Scenarios.

We conduct experiments on the more complex scenario using the dataset GSM8K (Cobbe et al., 2021). At first, we create new reference data X_{r} and counterfactual data X_{c}. Following the idea of methodology proposed in Section 4.1, we convert the question in GSM8K to obfuscate the semantic elements that necessitate calculation, while ensuring that the alterations to the text are minimal. An example is shown below:

- GSM8K X_{r}: "On a 16 GB (gigabyte) capacity USB drive, 50% is already busy. **Calculate the number** of gigabytes still available."
- GSM8K X_{c}: "On a 16 GB (gigabyte) capacity USB drive, 50% is already busy. **Describe the location** of gigabytes still available."

Then, we conduct the experiments of key head identification and validation following the experimental setting in Section 5.1. As a result, 60% of the key heads are overlapped with the key heads identified based on our original data. Moreover, knocking out the newly-identified key heads leads to a 65% accuracy drop on GSM8K, confirming their importance even in complex scenarios.

Table 4: Comparison of the key heads identified on our generated data in Figure 8 and the dataset GSM8K (Cobbe et al., 2021).

Dataset	Top-10 Key Heads [*e.g.*, (Layer, Head)]	Knockout Accuracy
Ours	[(12, 22), (13, 11), (16, 0), (15, 26), (18, 26), (18, 24), (30,31), (14, 27), (22, 25), (11, 8)]	−69%
GSM8K	[(19, 6), (11, 8), (12, 22), (14, 31), (13, 11), (22, 25), (16, 0), (21, 17), (15, 26), (29, 5)]	−65%

Furthermore, only knocking out the 6 overlapping heads brings in −56% and −52% on our generated data and GSM8K, respectively. It shows these heads are both important in two scenarios. If knocking out the 4 non-overlapping heads identified by GSM8K only, it has a negligible effect on our generated data (−2%) but apparently affects on GSM8K (−26%). It reveals the significance of these 4 heads specific to more complex reasoning mathematical problems. We further investigate the attention patterns of the 4 non-overlapping heads, and find that these heads mainly attend to text tokens. For example, the head (29, 5) attends to ".", and the head (19, 6) attends to "GB". In contrast, the 6 overlapping heads mainly attend to the number operands and operators. For example, the head (13, 11) attends to input operands "50", and the head (11, 8) attends to the operator "%".

Recent research (Opedal et al., 2024) has shown that solving the math word problems requires a synergy of multiple skills including 'text comprehension' and 'arithmetic calculation'. This is aligned with the phenomena of "the 4 non-overlapping heads attend to text tokens (*i.e.*, 'text comprehension'), while the 6 overlapping heads attend to number operands and operators (*i.e.*, 'arithmetic calculation')". In this work, we focus on the skill of arithmetic calculation as it's a fundamental ability universally shared across various levels of complexity for mathematical problems. It's imperative for continued research to develop a more holistic understanding of the intricate reasoning capacities.

To further investigate whether the model's deficiencies stem from a lack of mathematical abilities or a broader impairment in language processing, we evaluate LLaMA2-7B with key heads kept normal and knocked out on MMLU-Humanities benchmark (Hendrycks et al., 2020). The comparative performance was 42.9% for models with the key heads intact versus 42.6% for the knockout models. This negligible difference (−0.3%) suggests that the knockout of these heads does not significantly impact general language abilities.