A. Appendix

A. Templates

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

Figure 8: Templates used in this work follow the formations of "Equation", "Statement", "Question-Answer".

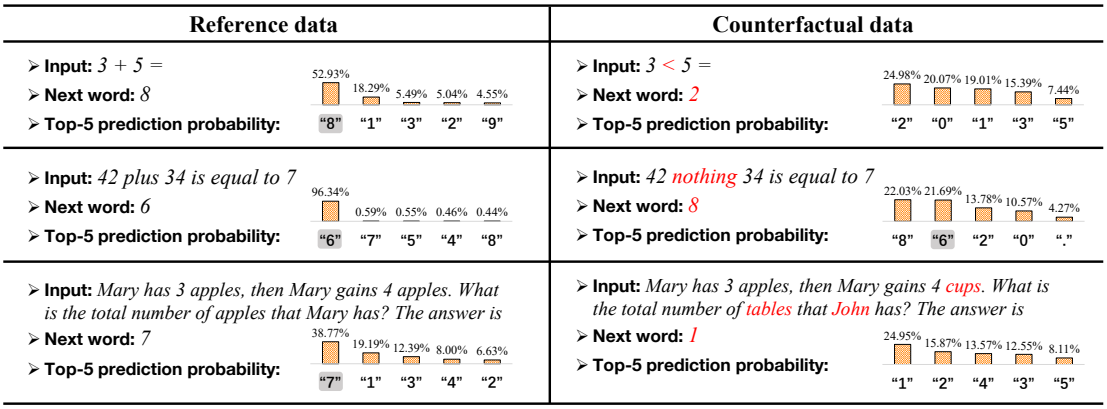

Figure 9: Examples of reference data (with addition logic) and counterfactual data (without addition logic). Given the input sentence, the results of next word prediction are provided by LLaMA2-7B.

We have included a list of 36 templates used in this work as shown in Figure 8. All these templates share the same calculation logic. we sample the <A> and <B> from $\{1, \cdots, 9\}$, since LLaMA2 tokenizes each digit individually (e.g., '42' is tokenized to '4' and '2'). Based on the above templates, we generate the sentences that the LLMs can predict the addition result {C} correctly as the reference data X_{r}. We generate the counterfactual data X_{c} following the principles depicted in Section 3, where we replace the words (*e.g.*, "plus", "minus", "times", "ratio") with a randomly-selected term from the set {"none", "nothing", $\cdots$, "null"}, and replace the operations (*e.g.*, "+", "-", "*", "/") with a randomly-selected term from the set {"<", ">", $\cdots$, "@"}. We show three cases in Figure 9 with the inspection into the top-5 prediction probability of LLaMA2-7B. Moreover, in Figure 10, we also construct several different types of linguistic meanings for the addition task: "time span" and "object accumulation". For the templates 1-8 of "time span", we sample from a