curated list of common words. For example, we select `<EVENT>` from {"war", "conflict", ⋯ , "project"}[4], `<VERB>` from {"last", "span", ⋯ , "extend"}, `<MONTH>` from {"Jan.", "Feb.", ⋯ , "Dec."}, and `<YYY>` from {100, ⋯ , 199}. For the templates 9-12 of "object accumulation", we sample `<OBJECT>` from {"apple", "orange", ⋯ , "pear"}, `<VERB>` from {"get", "obtain", ⋯ , "acquire"}, and each `<NAME>` was randomly selected from a pool of 100 English first names.

1. The `<EVENT>` `<VERB>` {`A`} years from the year `<YYY>`{`B`} to the year `<YYY>`{`C`}
2. The `<EVENT>` `<VERB>` {`A`} years from `<YYY>`{`B`} to `<YYY>`{`C`}
3. The `<EVENT>` `<VERB>` {`A`} days from `<MONTH>` {`B`} to `<MONTH>` {`C`}
4. The `<EVENT>` will `<VERB>` {`A`} days from `<MONTH>` {`B`} to `<MONTH>` {`C`}
5. The `<EVENT>` `<VERB>` {`A`} hours from {`B`} pm to {`C`}
6. The `<EVENT>` will `<VERB>` {`A`} hours from {`B`} pm to {`C`}
7. The `<EVENT>` `<VERB>` {`A`} hours from {`B`} am to {`C`}
8. The `<EVENT>` will `<VERB>` {`A`} hours from {`B`} am to {`C`}
9. `<NAME>` has {`A`} `<OBJECT>`, then `<NAME>` `<VERB>` {`B`} `<OBJECT>`.
What's the total number of `<OBJECT>` that `<NAME>` has? The answer is {`C`}
10. `<NAME>` `<VERB>` {`A`} `<OBJECT>`, and `<NAME2>` `<VERB>` {`B`} `<OBJECT>`.
What's the total number of `<OBJECT>` that they `<VERB>`? The answer is {`C`}
11. `<NAME>` has {`A`} `<OBJECT>`, and `<NAME2>` has {`B`} `<OBJECT>`.
What's the total number of `<OBJECT>` that they have? The answer is {`C`}
12. `<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 10: Additional templates used in the addition task, involve different linguistic meanings like "time span" (1-8) and "object accumulation" (9-12).

B. Evaluate the Effect of Attention Heads.

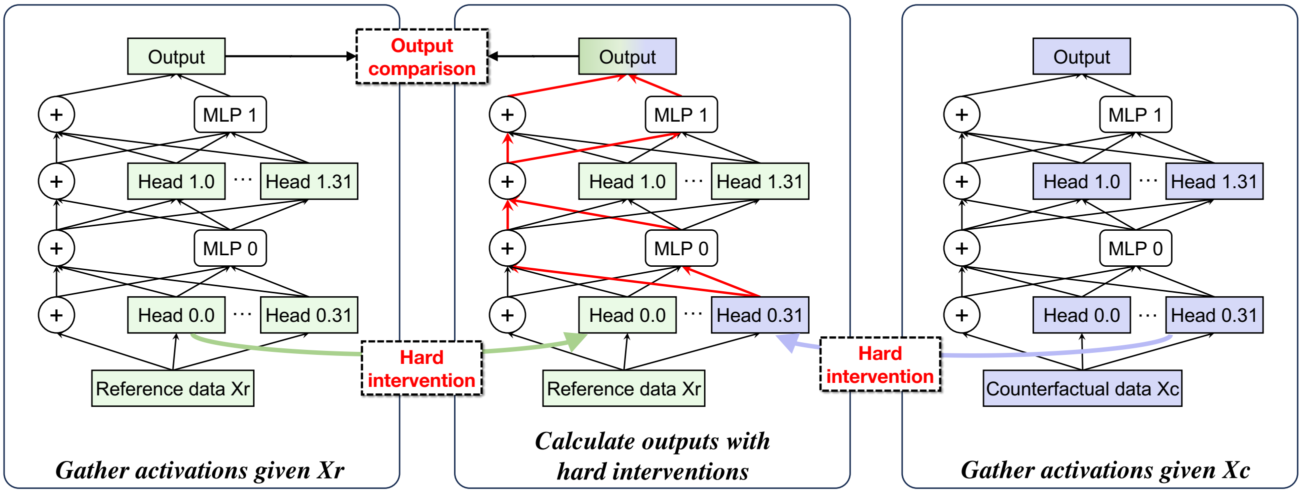

Figure 11: A case illustration of the method "path patching". It measures the importance of forward paths (*i.e.*, the red lines that originate from Head 0.31 to Output) for the two-layer transformer in completing the task on reference data.

Path Patching. To discover the cause of the predicted answer, we employ the causal intervention technique known as *path patching* (Goldowsky-Dill et al., 2023; Wang et al., 2023a). This approach is highly effective in analyzing the causal relationship between two computation nodes (Sender → Receiver). This helps us determine whether Sender is the cause of Receiver, and the connections between them are important for the model in implementing the task.

Specifically, the entire process of path patching is shown in Figure 11, where the node pair Sender → Receiver is set as Head

[4] We empirically find that the specific choice of words does not affect the results, as long as they meet similar semantics.