1. **Title:** We changed the title to better reflect the core contribution, emphasizing the idea of uncovering hidden correctness in causal reasoning.
2. **Abstract and Introduction:** The abstract and introduction better aligns with the updated framing and highlights the gap in current evaluation metrics failure.
3. **Literature Review:** The literature review was streamlined to avoid repetition and focus more directly on current limitations.
4. **Figures:** The main figure was revised for clarity, and a new figure has been added to illustrate the failure cases.
5. **More baselines:** Included LLM-as-a-judge as a baseline in addition to string match. Its clearer stated that other metrics and the experiments are in appendix.
6. **Clarity:** Some equations were removed that were not referenced later. Motivation behind formal guarantees were clarified, and restructured the explanation of do-calculus clearer. The method section has been simplified with clearer structure, and more details on the libraries we used to implement them.

Note: We acknowledge that incorporating an additional dataset was a primary concern across multiple reviews. While time constraints prevented us from including it in this revision, we are actively working on it and expect to include results in time for the rebuttal phase.