Explanation of Revisions to: Counterfactual Evaluation for Blind Attack Detection in LLM-based Evaluation Systems

Anonymous ACL submission

Error analysis:

In our previous submission, although we included some dataset-specific notes on TruthfulQA, we did not formally analyze the cases where errors occurred. To systematically evaluate failure cases, we added an error analysis, where we examined common reasons when an LLM misjudged a given candidate input.

More existing work:

To provide a more thorough context of our paper's standing and contribution, we expanded the Introduction to incorporate more existing work. This includes a wider range of attack methods on LLMs, such as specific types of prompt injections, jailbreaks, and data poisoning. We also included more defense methods for similar prompt injection attacks and studies targeting the security of evaluator LLMs.

Clear definition of setting:

To clearly explain the problem setting and the class of attacks we are focusing on, we included a figure. This visual representation illustrates the normal evaluation and attack flows, which we believe helps readers understand the scenario under consideration.

Additional limitations:

We acknowledged that our binary framework is a simplification of real-world QA tasks, as suggested by our reviewers. We included this inthe Limitations section and identify this as an important direction for future work.

Further elaborations:

To improve reproducibility of results, we elaborated on the specific sizes of the datasets that were used, API parameters that were specified, and clarified our reasons for not reporting computational budget and number of parameters for certain LLMs. We also included a clarification on our use of Ai assistants.