Explanation of Revisions: Response to Meta Review and Reviewer Comments

Authors

July 31, 2025

1 Overview of Major Changes

We thank the Area Chair and reviewers for their thorough and constructive feedback. We have made substantial revisions to address the core concerns raised in the meta review. The key changes include:

- **Reframed the causal claims and methodology:** Clarified the scope of our contribution and addressed the causal framing concerns, making it clear that our hybrid design leverages LLMs for candidate causal structure discovery while grounding estimation in formal causal modeling principles
- **Enhanced technical rigor:** Provided detailed descriptions of decision logic and method validation processes.
- **Expanded evaluation framework:** Included additional models, datasets, and comprehensive error analysis
- **Improved transparency and clarity of decision tree:** Added detailed explanations of the Tree-of-Thoughts approach and validation mechanisms
- **Addressed performance concerns:** Conducted thorough analysis of multiple in-depth baseline comparisons and in-depth error analysis

2 Point-by-Point Response to Meta Review : vEeL

2.1 Concern 1: Overstated Causal Framing

Reviewer Concern: *"The paper's framing around causality and counterfactual reasoning might be misleading, as the actual methodology lacks formal causal modeling, structural interventions, or rigorous causal inference. The approach is fundamentally prompt-driven and verbal rather than grounded in established causal theory."*

Our Response: We acknowledge this important concern and have made significant revisions to address the framing issues. In our revision, we clarify that our approach is not limited to prompt engineering. Instead, it combines structured prompt design to guide the LLM in identifying relevant decision variables and hypothesized causal relations, followed by the application of established structural causal inference techniques for validation and causal effect estimation. This hybrid design leverages LLMs for candidate causal structure discovery while grounding estimation in formal causal modeling principles.

2.2 Concern 2: Questionable Method Effectiveness

Reviewer Concern: *"The method's effectiveness is questionable, with results showing that error rates actually increase when using the proposed approach compared to baseline methods, particularly with GPT-4o-mini."*

Our Response: We acknowledge this concern and have addressed it through comprehensive revisions. We expanded our comparisons to include additional closed-source and open-source models and enhanced Section 6 with improved results and analysis.

2.3 Concern 3: Insufficient Method Validator Explanation

Reviewer Concern: *"The method validator component, which is critical for checking and identifying assumptions in causal inference, lacks proper explanation and justification."*

Our Response: We've Added a comprehensive example of the validation feedback loop in Appendix C, demonstrating explicitly how diagnostics inform method re-selection and improve robustness.

3 Point-by-Point Response to Reviewer pNkC

3.1 Concern 1: Overstated Causal Framing

Reviewer Concern: *"Causal framing is overstated: The paper frequently refers to counterfactual reasoning and causality, but the actual method involves no formal intervention, structural modeling, or causal inference. It's entirely verbal and prompt-driven"*

Our Response: In our revision, we clarify that our approach is not limited to prompt engineering. Instead, it combines structured prompt design to guide the LLM in identifying relevant decision variables and hypothesized causal relations, followed by the application of established structural causal inference techniques for validation and causal effect estimation. This hybrid design leverages LLMs for candidate causal structure discovery while grounding estimation in formal causal modeling principles.

3.2 Concern 2: Lack of Technical Innovation

Reviewer Concern: *"Lack of technical innovation: While the staged prompting setup is well-organized, it largely reflects common-sense decompositions. There's no learning, no new architecture, and no formal evaluation of what's gained by breaking the process into steps."*

Our Response: We acknowledge this concern and believe that staged prompting setup alone is not sufficient. While our baseline prompts also use structured prompting, they often struggle with complex causal problems : this is where our CAIS hybrid design approach excels. To further support our claim, we have made revisions to our paper by adding additional results from ablation studies, incorporating more baseline approaches, and evaluating our framework on multiple LLMs throughout Sections 5 and 6.

3.3 Concern 3:Underwhelming Evaluation Methodology

Reviewer Concern: *"Evaluation is underwhelming: The use of BLEU and similar metrics doesn't capture the quality or correctness of counterfactuals. The human evaluation is a step in the right*

direction, but it's small and lacks detail. It's hard to assess whether the outputs are truly more plausible or useful."

Our Response: We acknowledge this concern and have significantly strengthened our evaluation framework in response. We have moved beyond BLEU scores to focus on more meaningful metrics including Method Selection Accuracy (MSA) and Mean Relative Error (MRE), which better capture the quality and correctness of causal inference outcomes. To address concerns about plausibility and usefulness, we have expanded our evaluation to include real-world studies drawn from published research papers that address actual causal problems, providing a more rigorous assessment of our framework's practical utility. Sections 5 and 6 contain the evaluation results while the process of dataset creation is described in Appendix A figure 4.

3.4 Concern 4:Mismatch Between Title and Content

Reviewer Concern:*"Mismatch between title and content: Calling this a "Causal AI Assistant" suggests something more principled and general than what's delivered. The scope is much narrower—sentence-level prompts with scripted outputs—not a general-purpose assistant for causal reasoning."*

Our Response: We have addressed this concern by refining our abstract further.

4 Point-by-Point Response to Reviewer 5JAh

4.1 Concern 1:Unclear Tree-of-Thoughts Integration

Reviewer Concern:*"The step by step framework for causal analysis is intuitive. While Tree-of-Thought is mentioned in the intro, it is not explained in the methods section. How does ToT connect to the methods? And is ToT the main contribution or the assembly of different steps into one system? "*

Our response: We acknowledge this concern and have significantly strengthened our methodology description in Section 4, which now presents our approach in four distinct stages. We have provided additional detail on our decision tree structure in Appendix B. The Tree-of-Thoughts (ToT) framework serves as the core of CAIS, connecting the different aspects of identifying causal variables, method selection, validation, and interpretation. This integrated approach provides substantial structural benefits over baseline methods.

4.2 Concern 2:Inadequate Error Handling Discussion

Reviewer Concern:*"many of the steps such as method validator are error prone, requiring careful thought to apply the right robustness checks and interpret them. The framework doesn't offer a discussion on how errors in the pipeline are handled."*

Our Response: To address this concern we have provided a more detailed explanation of Method Validator and how the Method Selection backtracks if the method validation fails. We have provided a detailed analysis of a failure case in Appendix C

4.3 Concern 3:Performance Issues and Missing Baseline Results

Reviewer Concern:*"While the creation of new problems is noteworthy, the results on gpt4o-mini show that error actually increases using the proposed method compared to baseline. And I don't understand why table 2 does not include the Baseline. Can you provide the results for baseline for table 2? "*

Our Response: To address the concerns, we have strengthend our results section proving evaluation over multiple baselines and across multiple LLMs.

5 Point-by-Point Response to Reviewer 6xtd

5.1 Concern 1: Opaque data-driven method choice

Reviewer Concern: *"The paper shows that CAIA can "look at the dataset and decide," yet never spells out the decision logic that maps dataset diagnostics (e.g., time dimension, instrument presence, discontinuities) to a specific estimator. A short, formal description would make the selection process reproducible and auditable."*

Our Response: We agree with the reviewer that making the decision logic explicit is essential for reproducibility and transparency. In our revised submission, we have added a formal description of the method selection process in Section 4.2. This includes a step-by-step mapping from dataset characteristics such as timing of observations, presence of instrumental variables, or running variables to eligible causal inference methods (e.g., DiD, IV, RDD). Additionally, we now reference the complete decision tree structure in Appendix B, which outlines the conditions and prompts used at each decision node. This tree makes the internal logic of CAIS fully auditable and clearly links dataset diagnostics to estimator selection.

5.2 Concern 2: Method-validator rationale is under-explained.

Reviewer Concern: *"The "validator" agent that checks identifying assumptions (parallel trends, common support, weak instruments, etc.) is critical, but the authors give only brief prompt snippets. A more systematic justification of each test, plus quantitative evidence that the validator catches violations and refuses to report estimates when assumptions fail, is needed."*

Our Response: We have provided a more detailed explanation of Method Validator and how the Method Selection backtracks if the method validation fails. We have provided a detailed analysis of a failure case in Appendix C.

5.3 Concern 3: All experiments use GPT-4o/4o-mini.

Reviewer Concern: *"Including GPT-o3 (high-reasoning models) would clarify whether CAIA's gains stem from the pipeline design or from model horsepower. This is especially important for practitioners who may not have access to the very latest frontier model."*

Our Response: We have conducted a more comprehensive evaluation in response to reviewer concerns by expanding the diversity of model families tested, including GPT-4o, GPT-4o-mini, o3-mini, Gemini 2.5 Pro, and LLaMA 3.3-70B. Furthermore, we performed a dedicated ablation study isolating the contribution of the method validator. Results across Table 2 and Section 5.4 show that removing the validator's feedback loop significantly increases estimation error, especially in smaller models like GPT-4o-mini. This demonstrates that the validator plays a crucial role in correcting missteps in method selection and variable identification capabilities that are otherwise limited in compact models with weaker reasoning capacity.

5.4 Concern 4: Sensitivity to LLM Capability

Reviewer Concern: *"Large sensitivity to LLM quality remains unexplored. Table 3 hints that performance drops sharply with smaller variants, suggesting CAIA is fragile to LLM capability.*

An ablation that measures estimator-selection accuracy across a spectrum of model sizes would strengthen the claims."

Our Response: To better understand the impact of LLM size and reasoning capability on CAIS's performance, we expanded our evaluation to include a diverse spectrum of models (GPT-4o, GPT-4o-mini, o3-mini, Gemini 2.5 Pro, and LLaMA 3.3-70B) and conducted targeted ablation studies. Table 3 already presents method selection accuracy across these models and shows that performance drops noticeably with smaller variants, especially in real-world and synthetic datasets. To further investigate this, we performed an ablation removing the method validator. Results in Table 2 demonstrate that smaller models (e.g., GPT-4o-mini) exhibit a sharper degradation in performance without the validator, underscoring its critical role in guiding weaker models through method re-selection and error correction. In summary, while CAIS is designed to be LLM-agnostic, its effectiveness scales with model reasoning capability. We now explicitly discuss this observation in Section 6 and include it as a limitation in Section 10, providing guidance for users selecting models based on performance-resource tradeoffs.

5.5 Concern 5: Dataset Attribution in Reported Metrics

Reviewer Concern: *"The paper reports error metrics in Table 4 but never specifies which benchmark (QRData, synthetic, or real-world) those numbers come from. Stating the exact dataset and its characteristics is essential for interpreting the results."*

Our Response: We have completely redefined the experimentation section. All experiments are now reported across all three datasets (QRData, synthetic, and real-world) for every table to ensure clarity and consistency in benchmarking.

5.6 Concern: Statistical Validity Not Fully Assessed

Reviewer Concern: *"Confidence intervals reflect sampling error only; no coverage study under misspecification, and no comparison with expert hand-coded baselines on real data."*

Our Response: While we acknowledge that our confidence intervals currently reflect only sampling variability, we have taken steps to strengthen the empirical grounding of our evaluations. Specifically, for real-world studies, we now compare CAIS's causal estimates against ground truth results reported in published empirical research papers. These expert-derived estimates serve as strong hand-coded baselines for benchmarking both numerical accuracy and interpretability. This comparison helps assess whether CAIS produces plausible and valid inferences in high-stakes, real-world settings. We now clarify this comparison in Section 3.2 and include results in the experimental analysis (Section 6). We agree that formal coverage analysis under model misspecification is a valuable future direction and have added this point explicitly to the Limitations section (Section 10).

6 Summary

We believe these revisions comprehensively address the concerns raised in the meta review and individual reviewer comments. The paper now provides:

1. **Clarified Contribution Scope:** We have reframed our work as a method selection and application tool rather than a fundamental causal inference contribution, clearly distinguishing between LLM-driven method selection and formal econometric estimation.

2. **Rigorous Evaluation Framework:** We have expanded our evaluation to include multiple LLMs (GPT-4o, GPT-4o-mini, o3-mini, Gemini 2.5 Pro, LLaMA 3.3-70B), comprehensive baseline comparisons, and meaningful metrics (MSA, MRE) that better capture causal inference quality.

3. **Comprehensive System Documentation:** We have provided detailed explanations of all system components, including the Tree-of-Thoughts framework integration, method validator functionality, and error handling mechanisms.

4. **Assessment of limitations and failure modes**: We have included thorough discussions of failure modes, and practical constraints, providing clear guidance for practitioners.

We thank the reviewers for their valuable feedback, which has significantly improved the clarity, rigor, and impact of our work.