# Beyond Answer Accuracy: A Tie-Aware Evaluation of LLM Causal Explanations on Balanced COPA

**Anonymous ARR submission**

## Abstract

High final-answer accuracy on a two-choice commonsense causal reasoning task such as Balanced COPA does not necessarily imply reliable causal reasoning. In order to provide the required nuance to the evaluation, we build on two criteria that can be directly scored from structured IF–THEN explanations: *consistency*, measured as reachability in a directed explanation graph, and *parsimony*, measured as shortest reachable path length. Those are fed to a tie-aware scoring framework, which, rather than forcing a choice, abstains when alternatives have equal structural support, equal score, or no reachable support. By reporting strict accuracy, conditional accuracy on non-tied cases, and tie rate, we separate reliability from coverage. Experiments on GPT-5-family explanations show that consistency and parsimony are informative when decisive, but have limited coverage due to frequent ties.

## 1 Introduction

Final-answer accuracy is the standard signal in many commonsense and causal reasoning benchmarks, but it does not directly evaluate the explanation supporting a model's answer. Prior work distinguishes prediction and association from causal explanation (Pearl, 2009; Bishop, 2021; Kiciman et al., 2023). This distinction is especially important for large language models (LLMs): a model may select a plausible answer while the causal support for that answer remains untested.

We study this issue on Balanced COPA, a balanced version of the Choice of Plausible Alternatives benchmark for two-choice commonsense causal reasoning (Roemmele et al., 2011; Kavumba et al., 2019). In COPA, each item contains a short context, two candidate alternatives, and a causal direction; the model must choose either the more plausible cause or the more plausible effect of the context (Roemmele et al., 2011). In our zero-shot evaluation on 1,000 English Balanced COPA items, GPT-family models achieve high accuracy. This finding motivates a more diagnostic question: *When answer accuracy is already high, do model-generated explanations actually support the selected answers?*

In order to score competing explanations and select the one with the highest plausibility score, Dalal et al. (2024) evaluate LLM-generated explanations through four IBE-inspired criteria: consistency, parsimony, coherence, and uncertainty (Harman, 1965; Lipton, 2017). We focus on consistency and parsimony because they can be computed directly from structured IF–THEN explanations as graph reachability and path length. In Balanced COPA, however, these two criteria can tie: both options may have equally short IF–THEN paths, or neither path may reach the task-defined target.

To address this, we introduce tie-aware metrics for the logic/parsimony, semantic, and final scoring rules. For each rule, the evaluator predicts only when one alternative receives a higher score; equal scores are treated as ties. For logic/parsimony, cases where neither alternative is reachable are also treated as ties. The full scoring procedure is described in Section 3.

We evaluate three GPT-family models on Balanced COPA: GPT-5 Standard, GPT-5 Nano and GPT-4o as a previously strong baseline. We then analyze GPT-5 Standard outputs on the zero-shot error cases and 100 randomly sampled correct cases. While consistency and parsimony are useful when decisive, their contribution must be assessed together with their tie rate.

## 2 Related Work

COPA evaluates commonsense causal reasoning through forced-choice cause and effect questions (Roemmele et al., 2011) and is part of SuperGLUE (Wang et al., 2019). Balanced COPA reduces short-

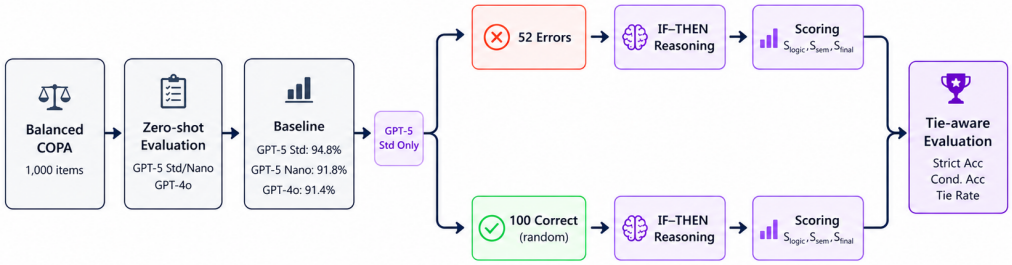


Figure 1: Overview of the evaluation pipeline: zero-shot baseline evaluation, direction-aware IF–THEN explanation generation, external explanation scoring, and tie-aware metrics.

cut effects by pairing alternatives with different premises and labels (Kavumba et al., 2019), making it suitable for testing whether explanations support the selected alternative. Recent work questions whether high benchmark performance reflects causal reasoning or sophisticated pattern matching (Bishop, 2021; Kiciman et al., 2023; Chi et al., 2024). This concern follows the causal-inference distinction between prediction, association, and explanation (Pearl, 2009; Spirtes et al., 2000; Peters et al., 2017). Our work addresses this issue by separating answer selection from explanation support.

Following prior work on IF–THEN explanations (Valentino et al., 2021; Dalal et al., 2024), we convert structured steps into directed graph edges.

IBE-Eval evaluates LLM-generated explanations through criteria inspired by Inference to the Best Explanation, including consistency, parsimony, coherence, and uncertainty (Dalal et al., 2024). It scores competing explanations and selects the one with the highest plausibility score. Our work follows this explanation-centered motivation, but focuses on a limitation in two-choice causal evaluation: consistency and parsimony may fail to distinguish alternatives. We therefore introduce tie-aware metrics to separate decisive structural scores from tied or unsupported cases.

## 3 Method

Our evaluation has three stages, illustrated in Figure 1. First, we run a direction-hidden zero-shot baseline on the full Balanced COPA set. Second, we generate direction-aware IF–THEN explanations for selected GPT-5 Standard cases. Third, we score the generated explanations using graph-based consistency/parsimony, semantic similarity, and tie-aware metrics.

**Dataset and zero-shot baseline.** We use Balanced COPA, a 1,000-item English two-choice causal reasoning dataset. Each item contains a context, two alternatives, a causal direction, and a gold label. The causal direction specifies whether the alternatives should be evaluated as possible causes or possible effects of the context. For example, given the context "My body cast a shadow over the grass" and the alternatives "The sun was rising" and "The grass was cut", the alternatives are causes of the context and the more plausible cause is that the sun was rising. The full item format is illustrated in Appendix A.

For the zero-shot baseline, the model receives only the context and two alternatives, but not the causal-direction label. It must output exactly one character: `1`, `2`, or `0`. The option `0` allows abstention when the model cannot decide between the alternatives. This setting only measures final-answer accuracy and not whether the model's choice is supported by a causal explanation.

We evaluate three GPT-family models: GPT-5 Standard, GPT-5 Nano, and GPT-4o. This selection compares model capacity and generation under the same zero-shot COPA setting: a high-capacity GPT-5 variant, a smaller GPT-5 variant, and a previously strong GPT-4o baseline. We conduct explanation-level analysis on GPT-5 Standard, because it achieves the strongest zero-shot performance.

**Selection of explanation-analysis subsets.** After the zero-shot baseline, GPT-5 Standard produces 52 incorrect predictions on the 1,000 Balanced COPA items. We use these 52 misclassified items as the error subset for explanation-level analysis. To compare this difficult subset with cases where the model already answers correctly, we also randomly sample 100 items from GPT-5 Standard's zero-shot correct predictions using a fixed random seed. All downstream explanation-level analyses are conducted on these two subsets: 52 zero-shot error cases and 100 sampled correct cases.

**Direction-aware IF–THEN reasoning.** For each selected item, we use a direction-aware IF–THEN prompt to generate explanations for both alternatives. Unlike the zero-shot baseline, this prompt explicitly provides the causal direction and asks the model to produce structured explanations for both alternatives, followed by a final selected answer. For cause questions, the reasoning direction is alternative $\rightarrow$ context, for effect questions, the reasoning direction is context $\rightarrow$ alternative. Each explanation must contain explicit IF–THEN steps and assumptions. Appendix C provides the prompt template.

**Graph-based consistency and parsimony.** We represent each IF–THEN explanation as a directed graph. Each step "IF $a$, THEN $b$" defines a directed edge $a \rightarrow b$, from the premise to the conclusion. Proposition text is normalized by lowercasing, removing boundary punctuation, and merging whitespace. For each alternative, the COPA direction determines the source and target propositions. In cause questions, we test whether there is a directed path from the option to the context; in effect questions, we test whether there is a directed path from the context to the option.

We define consistency as structural reachability in this directed graph: an explanation is structurally consistent when the target proposition is reachable from the source proposition through the model's stated IF–THEN steps. This does not prove that the causal relation is factually true; it only checks whether the generated explanation contains a directed path in the required task direction. Reachability and shortest-path depth are computed by breadth-first search over the directed edges, following only the premise-to-conclusion direction, with maximum depth $D_{\max} = 6$.

If the target is unreachable, the logical score is:

$$S_{\text{logic}} = 0. \tag{1}$$

If a path exists, we compute the shortest path depth $d$ and define:

$$S_{\text{logic}} = \frac{1}{1+d}. \tag{2}$$

This score combines consistency and parsimony: Consistency is required for a nonzero score, while parsimony rewards shorter reachable paths. In this setting, parsimony corresponds to the number of IF–THEN steps needed to connect the source and target: shorter chains are treated as more parsimonious because they introduce fewer intermediate assumptions. Parsimony is informative only when the two alternatives differ in shortest-path depth. If both alternatives are reachable with the same depth, or if neither alternative is reachable, the structural score is tied.

**Semantic consistency and final score.** Graph reachability is strict and can miss paraphrases. We therefore also compute semantic consistency using `sentence-transformers/all-mpnet-base-v2`, following prior work on sentence-level semantic similarity and MPNet representations (Reimers and Gurevych, 2019; Song et al., 2020). For each alternative, the model produces a summary $s$, while the input contains the context $c$ and the option $o$:

$$S_{\text{semantic}} = \frac{\cos(e(s), e(c)) + \cos(e(s), e(o))}{2}, \tag{3}$$

where $e(\cdot)$ is the embedding function. This semantic score is used as a soft relevance signal between the explanation summary, the context, and the option. It is not treated as proof of causal validity.

We then compute a combined score:

$$S_{\text{final}} = \alpha S_{\text{logic}} + (1-\alpha) S_{\text{semantic}}, \tag{4}$$

with $\alpha = 0.5$. Equal weighting is used as an untuned combination to avoid optimizing on the evaluation set.

**Tie-aware metrics.** For each item, a scoring rule assigns one score to each alternative. The predicted answer is:

$$\hat{y} = \begin{cases} 1 & \text{if } S_1 > S_2, \\ 2 & \text{if } S_2 > S_1, \\ \text{tie} & \text{if } S_1 = S_2. \end{cases} \tag{5}$$

For graph-based consistency/parsimony, we distinguish two tie types: A reachable tie occurs when both alternatives are reachable with the same shortest-path depth. An unreachable tie occurs when neither alternative is reachable. We report three metrics. Strict accuracy counts ties as incorrect. Conditional accuracy is computed only on non-tied cases. Tie rate measures how often the scoring rule cannot decide. This evaluation separates how reliable a criterion is when it makes a decision from how often it has enough evidence to make one.

## 4 Experiments and Results

**Zero-shot baseline.** Table 1 reports zero-shot results on 1,000 Balanced COPA samples. GPT-5

| Model | `0` Ans. | Acc. (%) |
|---|---|---|
| GPT-5 Standard | 0 | 94.8 |
| GPT-5 Nano | 0 | 91.8 |
| GPT-4o | 18 | 91.4 |

Table 1: Zero-shot performance on Balanced COPA.

Standard performs best, achieving 94.8% accuracy. GPT-5 Nano reaches 91.8%, and GPT-4o reaches 91.4%. GPT-4o outputs `0` in 18 cases, while both GPT-5 models always select one of the two alternatives.

Although GPT-5 Standard reaches high answer accuracy, it still produces 52 errors. Among these errors, 31 involve effect prediction and 21 involve cause identification. Direction-aware IF–THEN reasoning corrects 18 of these 52 errors when the model's own final answer is used, corresponding to 34.62%; the remaining 34 cases stay incorrect.

This self-correction result should be distinguished from the external scoring framework. Self-correction uses the model's own final answer after reasoning. The scoring framework evaluates the two generated explanations independently and selects the alternative with the higher external score, or reports a tie when the rule cannot decide.

**Tie-aware scoring.** Table 2 reports tie-aware scoring on two subsets: the 52 zero-shot error cases and 100 randomly sampled zero-shot correct cases. We compare three rules: logical consistency/parsimony, semantic consistency, and the final combined score.

On the 52 error cases, logic/parsimony produces a high tie rate of 67.31%. Its strict accuracy is therefore only 15.38% when ties are counted as incorrect. However, among the non-tied cases, it reaches 47.06% conditional accuracy. Semantic consistency produces no exact ties in our experiments and reaches 26.92% accuracy on the error cases. The final score also produces no exact ties and reaches 34.62%, matching the direct IF–THEN self-correction rate.

On the 100 randomly sampled correct cases, logic/parsimony reaches 93.02% conditional accuracy when it produces a non-tied decision, but ties in 57.00% of cases, resulting in 40.00% strict accuracy. Semantic consistency reaches 81.00%, and the final score reaches 85.00% without exact ties. These results reveal a reliability–coverage trade-off. Logic/parsimony is highly reliable when decisive, while semantic and final scores provide broader coverage but should not be interpreted as proofs of causal validity.

| Rule | Strict (%) | Cond. (%) | Tie (%) |
|---|---|---|---|
| 52 zero-shot errors | | | |
| Logic/Pars. | 15.38 | 47.06 | 67.31 |
| Semantic | 26.92 | 26.92 | 0.00 |
| Final | 34.62 | 34.62 | 0.00 |
| 100 sampled correct cases | | | |
| Logic/Pars. | 40.00 | 93.02 | 57.00 |
| Semantic | 81.00 | 81.00 | 0.00 |
| Final | 85.00 | 85.00 | 0.00 |

Table 2: Tie-aware scoring results. Strict accuracy counts ties as incorrect; conditional accuracy is computed only on non-tied cases.

## 5 Discussion and Conclusion

Our results show that answer accuracy and explanation support should be evaluated separately. GPT-5 Standard achieves strong zero-shot performance on Balanced COPA, but explanation-level analysis asks whether the selected answer is supported by a structurally decisive causal explanation. The 52 zero-shot error cases further show that direction-aware IF–THEN reasoning repairs some errors, but is not a complete solution.

Tie-aware evaluation clarifies the role of consistency and parsimony. Consistency checks whether the generated IF–THEN graph connects the required source and target in the task direction, while parsimony prefers shorter reachable chains. These criteria are informative when they distinguish the alternatives; when both alternatives have the same shortest-path depth, or neither reaches the target, the structural evidence does not justify a forced prediction.

This is why strict accuracy, conditional accuracy, and tie rate should be reported together. Conditional accuracy shows how reliable a rule is when it decides, while tie rate shows how often it has enough structural evidence to decide. Semantic consistency and the final score provide broader coverage, but should be interpreted as soft relevance signals rather than proofs of causal validity.

This paper introduced a tie-aware evaluation of LLM causal explanations on Balanced COPA, showing that shorter reachable IF–THEN explanations are useful only when they distinguish competing alternatives.

## Limitations

This study is a focused diagnostic evaluation rather than a broad benchmark study. It is limited to Balanced COPA and English commonsense causal reasoning. The explanation-level analysis focuses on GPT-5 Standard, using its 52 zero-shot error cases and 100 randomly sampled correct cases. These subsets are intended to test whether explanation-based signals reveal structure beyond answer accuracy, rather than to estimate overall performance across all COPA items, datasets, or model families. This design is appropriate for the paper's goal: to examine when consistency and parsimony provide decisive explanation-level signals, and when they remain tied.

The proposed scoring framework depends on the quality of generated IF–THEN chains and on proposition normalization. Because graph construction relies on explicit proposition matching, semantically equivalent paraphrases may fail to connect in the graph. Conversely, a reachable path does not guarantee that the underlying causal assumption is true. Similarly, embedding similarity is only a soft proxy for semantic consistency and should not be interpreted as causal validity.

Future work should evaluate additional datasets, languages, models, and explanation formats, as well as larger explanation-analysis subsets. It should also examine other IBE-inspired dimensions, such as coherence and uncertainty, and test whether alternative weighting schemes for structural and semantic scores improve robustness across settings. Finally, tie-aware evaluation could be extended to more complex causal tasks, including counterfactual or multi-step causal reasoning (Pearl, 2009; Kiciman et al., 2023).

## References


Christopher M. Bishop. 2021. Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology*, 12:735526.

Ethan Chi, Yixuan Wang, Zhiqing Sun, Qiang Zhu, and Zhiyuan Liu. 2024. Unveiling causal reasoning in large language models. *arXiv preprint*. Accepted at NeurIPS 2024.

Dhairya Dalal, Marco Valentino, Andre Freitas, and Paul Buitelaar. 2024. Inference to the best explanation in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 217–235, Bangkok, Thailand. Association for Computational Linguistics.

Gilbert H. Harman. 1965. The inference to the best explanation. *The Philosophical Review*, 74(1):88–95.

Pride Kavumba, Naoya Inoue, Benjamin Kasai, Paul Reisert, Xiaotian Han, Ryo Takahashi, and Kentaro Inui. 2019. Choosing the right alternatives: A study of causal reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 4753–4762.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Peter Lipton. 2017. Inference to the best explanation. In W. H. Newton-Smith, editor, *A Companion to the Philosophy of Science*, pages 184–193. Blackwell.

Judea Pearl. 2009. *Causality*, 2 edition. Cambridge University Press.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford University.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*, 2 edition. MIT Press.

Marco Valentino, Ian Pratt-Hartmann, and André Freitas. 2021. Do natural language explanations represent valid logical arguments? verifying entailment in explainable NLI gold standards. In *Proceedings of the 14th International Conference on Computational Semantics*, pages 76–86, Groningen, The Netherlands. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32.

# Appendix

## A COPA XML Example

Figure 2 shows the first complete XML item from the Balanced COPA file used in this study. This example illustrates the original dataset structure: an item ID, a causal direction, the gold alternative, a premise, and two candidate alternatives.

```
<item id="1" asks-for="cause"
      most-plausible-alternative="1">
  <p>My body cast a shadow over the grass.</p>
  <a1>The sun was rising.</a1>
  <a2>The grass was cut.</a2>
</item>
```

Figure 2: First complete XML item in the Balanced COPA file.

## B Zero-Shot Prompt

Figure 3 shows the zero-shot prompt used for the baseline evaluation. The prompt hides the causal-direction label and restricts the output to `0`, `1`, or `2`.

```
You are a strict COPA plausibility judge.
Given a context and two alternatives, decide which
      alternative is
MORE PLAUSIBLE.

Rules:
 - Reply with EXACTLY ONE CHARACTER: 1 or 2.
 - If you truly cannot decide, respond with 0.
 - No words, no spaces, no punctuation, no explanation.

Context: {context}
Alternative 1: {a1}
Alternative 2: {a2}
Answer (ONLY 0, 1, or 2):
```

Figure 3: Zero-shot prompt used for the direction-hidden baseline.

## C Direction-Aware IF–THEN Prompt

Figure 4 shows a compact version of the direction-aware IF–THEN prompt template used for explanation generation. The prompt asks the model to reason from option to context for cause questions and from context to option for effect questions, and to return a structured JSON object containing explanations for both alternatives and a final selected option.

```
You are an expert in causal reasoning and inference to the
      best explanation.

Task:
Given a context, two alternatives, and a causal direction,
      determine which option is the most plausible cause or
      effect.

Direction:
- If ask_for = "cause": reason from each option to the
      context.
  Premise = option text; Conclusion = context.
- If ask_for = "effect": reason from the context to each
      option.
  Premise = context; Conclusion = option text.

Chain format:
For each option, generate an explicit IF--THEN-Assumption
      chain.
Each step must follow this form:

Step k: IF <premise>, THEN <conclusion>.
Assumption: <commonsense assumption>.

The conclusion of Step k should become the premise of Step
      k+1 whenever possible.
Continue until the target conclusion is supported or shown
      not to follow.

Output:
Return exactly one valid JSON object with the following
      fields:
{
  "id": "<item id>",
  "question": {
    "context": "<context>",
    "ask_for": "<cause or effect>"
  },
  "options": [
    {"id": "1", "text": "<alternative 1>"},
    {"id": "2", "text": "<alternative 2>"}
  ],
  "explanations": [
    {
      "option_id": "1",
      "summary": "<brief summary>",
      "steps": ["<IF--THEN step 1>", "..."],
      "therefore": "<whether option 1 supports the
      conclusion>"
    },
    {
      "option_id": "2",
      "summary": "<brief summary>",
      "steps": ["<IF--THEN step 1>", "..."],
      "therefore": "<whether option 2 supports the
      conclusion>"
    }
  ],
  "decision": {
    "selected_option_id": "1|2",
    "justification": "<brief justification>"
  }
}

Decision rule:
After comparing both chains, select exactly one option ("1"
      or "2") based on the strength and plausibility of the
      reasoning chains.
```

Figure 4: Compact version of the direction-aware IF–THEN prompt template used for explanation generation.

## D Reproducibility Details

All model outputs are obtained with the same prompt template for each corresponding stage. The baseline prompt is direction-hidden and restricts the output to `0`, `1`, or `2`. The IF–THEN prompt is direction-aware and requires a valid JSON object

containing explanations for both alternatives. The model names reported in the paper follow the interface labels available at the time of experimentation. No API keys, user identifiers, or local file paths are included in the anonymized submission.

The 100 correct cases are sampled from GPT-5 Standard's zero-shot correct predictions using a fixed random seed. The exact seed, processed IDs, and scripts are provided in the anonymized supplementary material. If an output cannot be parsed as JSON, it is excluded from automatic graph construction and logged for manual inspection; in the reported runs, the final scoring files contain one row per evaluated item.

## E Implementation Details

All generated explanations are parsed from JSON outputs. IF–THEN steps are converted into directed graph edges. Proposition text is normalized by lowercasing, stripping boundary punctuation, and merging whitespace. Reachability is computed using breadth-first search over directed edges, following the premise-to-conclusion direction, with maximum depth $D_{\max} = 6$. For tie-aware evaluation, an item is marked as tied when the two alternatives receive equal scores under a given rule.

The main final score uses $\alpha = 0.5$ as an untrained equal-weight combination of structural and semantic signals. We do not tune this value on the evaluation set. Future work should evaluate whether alternative weights improve robustness across datasets and models.

## F Qualitative Error Patterns

Manual inspection of the 52 zero-shot errors suggests three recurring patterns. First, some gold labels prefer an atypical or extreme outcome, while the model prefers a more common everyday consequence. Second, many items contain two plausible alternatives whose difference depends on implicit premises, temporal ordering, or different levels of event granularity. Third, a small number of cases show residual causal-direction confusion even after the direction-aware prompt. These patterns support the interpretation that tie-aware scoring is diagnostic: it helps identify when a structural explanation signal is decisive, and when the available evidence is underdetermined.