CLEAR-3K: Assessing Causal Explanatory Capabilities in Language Models

Naiming Liu
Rice University
nl35@rice.edu

Richard Baraniuk
Rice University
richb@rice.edu

Shashank Sonkar
University of Central Flordia
shashank.sonkar@ucf.edu

Abstract

We introduce CLEAR-3K, a dataset of 3,008 assertion-reasoning questions designed to evaluate whether language models can determine if one statement causally explains another. Each question presents an assertion-reason pair and challenge language models to distinguish between semantic relatedness and genuine causal explanatory relationships. Through comprehensive evaluation of 21 state-of-the-art language models (ranging from 0.5B to 72B parameters), we identify two fundamental findings. First, language models frequently confuse semantic similarity with causality, relying on lexical and semantic overlap instead of inferring actual causal explanatory relationships. Second, as parameter size increases, models tend to shift from being overly skeptical about causal relationships to being excessively permissive in accepting them. Despite this shift, performance measured by the Matthews Correlation Coefficient plateaus at just 0.55, even for the best-performing models. Hence, CLEAR-3K provides a crucial benchmark for developing and evaluating causal explanatory reasoning in language models, which is an essential capability for applications that require accurate assessment of causal relationships.

1 Introduction

Assertion-reasoning questions have emerged as a valuable tool for evaluating higher-order thinking in educational assessment (Kumar, 2018; Central Board of Secondary Education, 2020a). These questions present students with an assertion followed by a reason, asking them to determine whether the reason correctly explains the assertion. Unlike conventional multiple-choice questions that focuses on factual memorization, assertion-reasoning questions requires the skill of distinguishing between two statements that are merely topically related and those that reflects a genuine explanatory relationship (Bloom et al., 1956; Anderson et al., 2001). This distinction represents a fundamental challenge in logical reasoning that requires deeper critical thinking capabilities beyond surface-level topic recognition and memorization (Kahneman, 2011; Evans and Stanovich, 2013).

Despite the importance of evaluating causal relationships, current natural language understanding benchmarks inadequately address this reasoning capability. While existing datasets assess various reasoning aspects, including natural language inference (Bowman et al., 2015; Williams et al., 2018), reading comprehension (Rajpurkar et al., 2016, 2018), and commonsense reasoning (Talmor et al., 2019; Sakaguchi et al., 2020), few specifically target the ability to determine whether a given reason constitutes a valid explanation for an assertion. This gap is particularly concerning given the prevalence of causal reasoning in critical domains, from medical diagnosis and treatment decisions to legal argumentation and policy analysis (Thorne et al., 2018; Wadden et al., 2020).

To address this limitation, we introduce CLEAR-3K, a dataset of 3,008 assertion-reasoning questions designed to evaluate causal explanatory reasoning capabilities of language models. The questions are curated from real-world educational materials and covers a diverse range of subjects spanning both STEM and humanities disciplines from grades 9 to 12. This diversity allows for robust evaluation across different knowledge domains, difficulty levels, and reasoning styles.

We conduct an extensive empirical evaluation of 21 language models spanning five leading open-source model families, with parameter sizes ranging from 0.5B to 72B parameters. This comprehensive evaluation allows us to identify robust patterns in how causal explanatory reasoning changes with different model types and parameters.

Our primary finding reveals a fundamental limitation in how language models approach causal explanatory reasoning. Language models tend to substitute semantic similarity for causal understand-

Findings of the Association for Computational Linguistics: EACL 2026, pages 3121–3136

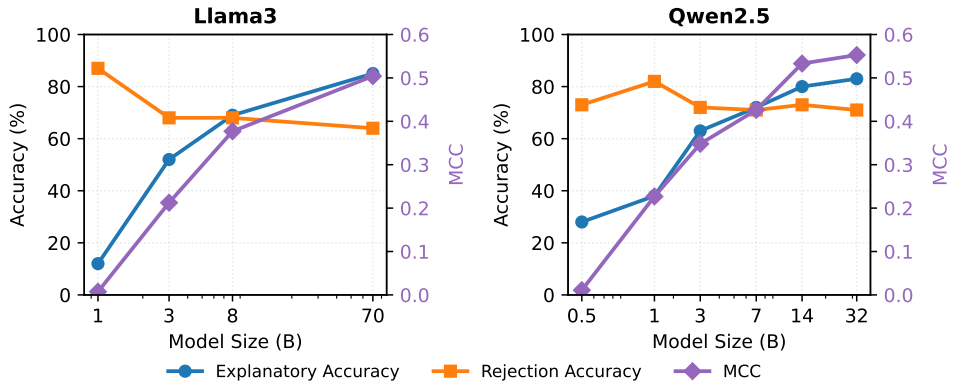

Figure 1: Model performance with increasing sizes. Smaller models exhibit strong negative bias when evaluating explanatory relationship (Llama3-1B: 87% rejection accuracy, 12% explanatory accuracy), while larger models develop the opposite tendency (Llama3-70B: 85% explanatory accuracy, 64% rejection accuracy). Although overall performance improves with scale, MCC stays at 0.55 across model families. This trend reveals a fundamental limitation: models struggle to distinguish semantic relatedness from genuine causal explanatory relationships.

ing, regardless of whether a true explanatory relationship exists. This pattern is illustrated in the confusion matrix with semantic similarity statistics from Phi-14B:

	Pred Y	**Pred N**
True Y	0.590 ± 0.073 (1048)	0.553 ± 0.080 (129)
True N	0.571 ± 0.072 (238)	0.547 ± 0.083 (480)

Table 1: Confusion matrix for Phi-4-14B when both assertion and reason are true. Y/N indicates whether the reason actually explains the assertion. Reported values show semantic similarity (mean ± std) and counts.

This results in Table 1 reveals that when the model correctly identifies causal relationships (True Y, Pred Y), the average semantic similarity is significantly higher (0.590) than when it misses the explanations relationships (True Y, Pred N: 0.553). Similarly, when the model incorrectly predicts explainability for non-explainable pairs (True N, Pred Y), the statements have higher similarity (0.571) than when it correctly rejects them (True N, Pred N: 0.547). These consistent differences suggest that models rely heavily on semantic similarity as a proxy for causality, which is a classic case of confusing correlation with causation.

Additionally, we observe a consistent shift in model performance as parameter size increases: smaller models exhibit a strong bias toward rejecting explanatory relationships, while larger models develop the opposite trends, as shown in Figure 1. For instance, when both assertion and reason are true, LLaMA3-1B correctly identifies only 12% of valid explanations while correctly rejecting 87% of non-explanatory relationships, whereas LLaMA3-70B flips this pattern, achieving 85% explanatory accuracy and only 64% rejection accuracy. This pattern suggests that increased model size enhances their sensitivity to semantic relatedness but does not enhance their ability to distinguish genuine causal relationships. To quantify overall performance, we use the Matthews Correlation Coefficient (MCC), which improves with scale across all families but plateaus around 0.55, even for the largest models. This convergence suggests an upper bound on current models' causal reasoning abilities.

These results reveal a critical insight about the capabilities of language models: they fundamentally confuse semantic similarity with causal relationships. This limitation persists even as models scale up, suggesting that future progress may require architectural or training innovations specifically targeting causal inference.

2 Dataset: CLEAR-3K

To enable a systematic study of causal reasoning abilities in language models, we present CLEAR-3K (**C**ausal **L**ogical **E**xplanatory **A**ssessment **R**esource), a comprehensive human-curated dataset consisting of 3,008 assertion-reasoning questions spanning multiple domains and difficulty levels. Details of the data collection and validation are provided in Appendix A.

Given a statement of Assertion (A) followed by a statement of a Reason (R), their relationship falls into one of the four categories:

a. Both A and R are true and R is the correct explanation of A
b. Both A and R are true and R is not the correct explanation of A
c. A is true but R is false
d. A is false but R is true

Category	Mathematics, Grade 9
a	**Assertion:** If angles A and B form a linear pair of angles and A = 70°, then B = 110°. **Reason:** Sum of linear pair of angles is always 180°.
Category	**Biology, Grade 10**
b	**Assertion:** Cerebellum controls the coordination of body movement and posture. **Reason:** Medulla oblongata controls and regulates the centre for coughing, sneezing and vomiting.
Category	**Physics, Grade 11**
c	**Assertion:** Total energy of the freely falling body is constant at each point. **Reason:** Kinetic energy of freely falling body is minimum, when it reaches the ground.
Category	**Chemistry, Grade 12**
d	**Assertion:** Copper is a non-transition element. **Reason:** Copper has completely filled d-orbitals in its ground state.

Table 2: Four relationships between assertions and reasons (categories a-d) in assertion-reasoning questions. Each example demonstrates one category, and contains subjects of Mathematics, Biology, Physics, and Chemistry across grades 9-12.

2.1 Assertion-Reasoning Question Format

Assertion-reasoning questions are a widely used assessment format in education designed to evaluate students' higher-order thinking abilities. Each question presents two statements - an assertion (A) and a reason (R) — and asks students to classify their relationship into one of the following four categories:

a. Both A and R are true, and Reason correctly explains the Assertion.
b. Both A and R are true, but the Reason does not explain the Assertion.
c. A is true, but R is false.
d. A is false, but R is true.

This format challenges students to not only evaluate the factual accuracy but also determine whether an explanatory causal relationship exists between the assertion and reason statements. Table 2 provides representative examples from our dataset illustrating each answer category.

2.2 Subject and Grade Distribution

Our dataset covers eight distinct subject areas across Grade 9-12, representing both STEM and humanities disciplines as shown in Table 3.

Subject	Count	Grade 9	10	11	12
Math	763	55	263	79	366
Biology	605	31	146	104	324
Chemistry	597	37	150	139	271
Physics	576	44	108	126	298
Geography	145	28	117	0	0
Pol. Science	122	29	93	0	0
Economics	100	23	77	0	0
History	100	29	71	0	0
Total	**3,008**	**276**	**1,025**	**448**	**1,259**

Table 3: Distribution of CLEAR-3K across subjects and grade levels.

The subject distribution reflects the emphasis on STEM disciplines in standard educational curricula. Mathematics, Biology, Chemistry, and Physics collectively comprise of approximately 84% of the questions in our dataset. In terms of grade level, the dataset is weighted toward Grade 10 and 12, which aligns with the widespread use of assertion-reasoning questions in standardized assessments at these grade levels.

2.3 Answer Category Distribution

Table 4 shows the distribution of correct answers for the assertion-reasoning questions across the four categories (a, b, c, d) described above.

Category	Count	%
(a) Both true, R explains A	1,177	39.1
(b) Both true, R doesn't explain A	718	23.9
(c) A true, R false	616	20.5
(d) A false, R true	497	16.5

Table 4: Distribution of CLEAR-3K by answer categories.

The dataset has a moderate bias toward category (a), which is typical in educational contexts where assertion-reasoning questions are often designed to assess correct understanding of causal relationships. Nevertheless, categories (b), (c), and (d) have sufficient representation to ensure balanced evaluation across different reasoning patterns, with (b) being slightly more common than (c) and (d).

2.4 Performance on Assertion-Reasoning Question

Qwen3-32B (Acc: 73.2%)				
	Pred a	**Pred b**	**Pred c**	**Pred d**
True a	**961**	112	64	40
True b	195	**436**	53	34
True c	69	46	**481**	20
True d	115	26	32	**324**

Table 5: Confusion matrix for Qwen3-32B model on the assertion-reasoning format.

We provide the confusion matrix for Qwen3-32B on the traditional assertion-reasoning format in Table 5. The model achieves 73.2% overall accuracy. The confusion matrix reveals that the most frequent error is misclassifying option (b) as option (a) (195 cases, 27.1% of all true (b) cases). Since both options involve factually true statements, this confusion indicates difficulty in determining whether the reason correctly explains the assertion. While the model demonstrates relatively balanced performance across categories (61% - 82%), it shows particular strength in identifying cases where the reason correctly explains the assertion (option a: 81.6%) and where the assertion is true but the reason is false (option c: 78.1%). A comprehensive analysis across model sizes and correlation to our reformulated causal explanatory reasoning task is provided in Appendix E.

3 Problem Formulation

In order to isolate causal explanatory reasoning from factual verification, we introduce the **Causal Explanation Task**, a reformulated version that enables more precise evaluation of a model's capacity to identify genuine explanatory relationships and goes beyond simply recognizing topic relevance or semantic relatedness of two statements.

3.1 Reformulating Assertion-Reasoning into Causal Explanation Task

Our primary interest is examining whether language models can correctly identify when one statement (reason) explains another (assertion), which is a fundamental aspect of causal reasoning. Traditional assertion-reasoning questions combine this explanation task with factual verification, which can obscure a model's causal reasoning and factual verification abilities.

To more precisely evaluate the causal reasoning abilities, we introduce the Causal Explanation Task. This task aims to isolate causal explanatory reasoning as a standalone binary decision problem. This reformulation allows us to assess whether models can go beyond surface-level semantic relatedness and truly identify causal explanatory relationships - a fundamental challenge in logical reasoning for both human and language models. In this binary task format, each question presents a pair of statements, an assertion (A) and a reason (R), and prompts the model to decide whether the reason provides a valid causal explanation for the assertion:

> **Causal Explanation Task**:
> Assertion (A): {assertion}
> Reason (R): {reason}
> Task:
> Determine whether the Reason (R) causally explains the Assertion (A).
> Provide your answer in the following JSON format:
> {"Verdict": "yes" or "no"}

In Causal Explanation task, a "yes" verdict only corresponds to the original option (a), where R truly explains A. For all other cases (b, c, d), the correct verdict is "no" — either because R does not

explain A despite both being true (b), or because a factually incorrect statement cannot explain or be explained by another statement (c, d). We evaluate model performance in two distinct settings:

1. **When both A and R are true** (original answers a and b): This context directly tests the model's ability to distinguish between semantic correlation and causation. Even when both statements are factually correct and semantically related, the model must determine whether R actually explains A.

2. **When either A or R is false** (original answers c and d): This context reveals whether factual errors affect causal reasoning. Logically, a false statement cannot validly explain or be explained by another statement, so models should consistently reject explanatory relationships in this setting.

This task reformulation is crucial because it explores whether models substitute semantic similarity for real causal understanding. By comparing performance across these settings, we can determine whether models simply associate semantically related statements or truly comprehend when one statement causally explains another.

3.2 Evaluation Metrics

We evaluate model performance on Causal Explanation task using three complementary metrics. For each assertion-reason pair, language models must determine whether the reason explains assertion by responding with either "yes" (indicating R provides a valid causal explanation for A) or "no" (indicating R does not causally explain A). Based on these binary verdicts, we measure:

1. **Explanatory Accuracy**: The proportion of cases in which the models correctly respond "yes" when R truly explains A (corresponding to original answer (a)). This metric measures the model's ability to correctly identify causal explanatory relationships.

2. **Rejection Accuracy**: The proportion of cases in which the model correctly responds "no" when R does not explain A (corresponding to original answer (b, c, d)). This metric captures the model's ability to identify and reject non-explanatory relationships.

We report both Explanatory and Rejection Accuracy on cases where both A and R are true (Setting 1). For cases where either A or R is false (Setting 2), the correct verdict is always "no," so only Rejection Accuracy is applicable.

3. **Matthews Correlation Coefficient (MCC)**: A balanced measure of binary classification performance, defined as

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{P \cdot N \cdot P' \cdot N'}}$$

where TP = correctly identified explanatory relationships ("yes" for answer (a)), TN = correctly identified non-explanatory relationships ("no" for answer (b, c, d)), FP = incorrectly accepted explanations, FN = incorrectly rejected explanations, and $P = TP + FN$, $N = TN + FP$, $P' = TP + FP$, $N' = TN + FN$.

MCC is particularly well-suited for our task due to the class imbalance: non-explanatory cases (b, c, d) are more frequent than explanatory ones (a). Unlike F1 score, which focuses primarily on positive cases, MCC fully accounts for true negatives cases where the model correctly rejects non-explanatory relationships. This sensitivity to true negatives is essential for evaluating causal explanatory reasoning because correctly rejecting invalid explanations is as important as identifying valid ones. By using MCC, we obtain a more comprehensive assessment of how well models distinguish genuine causal relationships from only semantic correlations.

4 Experiments and Results

We report results on the Causal Explanation task in Table 6 with performance and analysis by subject and grade level shown in Appendix F.

4.1 Experimental Setup

We conducted experiments on Causal Explanation task with five leading open-source LLMs model families: LLaMA3 (1B-70B) (Grattafiori et al., 2024), Qwen2.5 (0.5B-72B) (Team, 2024), Qwen3 (1.7B-32B) (Yang et al., 2025), Gemma3 (1B-27B) (Team et al., 2024), and Phi-4 Reasoning (4B-14B) (Abdin et al., 2024). We computed cosine similarity using Linq-Embed-Mistral (Choi et al., 2024), the top-performing model on the Massive Text Embedding Benchmark (Muennighoff et al., 2022) to our knowledge.

4.2 Explanatory and Rejection Accuracy Analysis Across Models

To systematically evaluate causal reasoning capabilities, we analyzed model performance in two

Model	Overall MCC	When both A, R are true				When either A or R is false
		Accuracy	MCC	Explanatory	Rejection	Rejection
Qwen-3 Models						
Qwen3-32B	**0.66**	**0.79**	**0.55**	0.85	0.69	0.90
Qwen3-14B	**0.66**	0.78	**0.55**	0.80	0.76	**0.94**
Qwen3-8B	0.63	0.77	0.54	0.78	0.76	0.91
Qwen3-4B	0.62	0.77	0.52	0.81	0.71	0.88
Qwen3-1.7B	0.54	0.74	0.44	0.79	0.65	0.83
LLaMA-3 Models						
LLaMA3.3-70B	0.61	0.77	0.50	0.85	0.64	0.85
LLaMA3.1-8B	0.43	0.69	0.38	0.70	0.68	0.77
LLaMA3.2-3B	0.30	0.59	0.21	0.53	0.68	0.80
LLaMA3.2-1B	0.04	0.40	-0.01	0.12	**0.87**	0.93
Qwen-2.5 Models						
Qwen2.5-72B	0.64	**0.79**	**0.55**	0.87	0.66	0.87
Qwen2.5-32B	0.62	**0.79**	**0.55**	0.84	0.71	0.85
Qwen2.5-14B	0.60	0.78	0.53	0.80	0.74	0.86
Qwen2.5-7B	0.50	0.72	0.43	0.73	0.71	0.82
Qwen2.5-3B	0.42	0.67	0.35	0.64	0.72	0.81
Qwen2.5-1.5B	0.23	0.55	0.23	0.39	0.83	0.82
Qwen2.5-0.5B	0.02	0.45	0.01	0.27	0.73	0.74
Phi-4 Models						
Phi-4-14B	0.65	0.78	**0.55**	0.78	0.78	0.92
Phi-4-4B	0.55	0.73	0.43	0.78	0.65	0.86
Gemma-3 Models						
Gemma3-27B	0.49	0.72	0.37	**0.93**	0.36	0.67
Gemma3-12B	0.52	0.74	0.43	0.87	0.52	0.74
Gemma3-4B	0.35	0.70	0.32	0.90	0.36	0.47

Table 6: Performance of 21 models from five model families (Qwen-3, LLaMA-3, Qwen-2.5, Phi-4, and Gemma-3) on the CLEAR-3K dataset for the Causal Explanation Task. Results are reported using Explanatory Accuracy, Rejection Accuracy, and MCC across three settings: overall performance, cases where both the A and R are true (assessing the ability to distinguish causal explanation from semantic relatedness), and cases where either A or R is false (assessing whether factual errors affect causal explanatory relationship). Best performance for each metric is highlighted in **bold**.

settings: (1) when both A and R are factually true, and (2) when either A or R contains factual errors. Table 6 presents our comprehensive results across 21 models spanning five model families, covering a range of parameters sizes.

4.2.1 When both A and R are True

As shown in Table 6, we observe a clear and consistent pattern of model performance when both assertion and reason are factually correct. As parameter size increases, explanatory accuracy (correctly identifying when R explains A) improves substantially, while rejection accuracy (correctly identifying when R does not explain A despite both being true) tends to decrease.

This trends show that smaller models frequently fail to recognize valid explanatory relationships. For instance, Llama-1B model achieves only 12% explanatory accuracy while maintaining 87% rejection accuracy. The Qwen2-0.5B model shows a similar pattern with 27% explanatory accuracy and 73% rejection accuracy. Smaller models appear to lean heavily toward answering "no" when asked if a reason explains an assertion, regardless of actual causal explanatory relationship.

On the other hand, this tendency reverses as models scale. For instance, LLaMA-70B reaches 85% explanatory accuracy but its rejection accuracy falls to 64%. Similarly, Qwen2-72B achieves 87% explanatory accuracy with 66% rejection accuracy. This consistent pattern suggests that increased model size fundamentally alters how models approach causal reasoning tasks.

To quantify this trade-off, we report MCC for a

Model	When both A and R are true			When either A or R is false		
		Pred Y	Pred N		Pred Y	Pred N
Phi-14B	**True Y**	0.590 ± 0.073 (1048)	0.553 ± 0.080 (129)	**True Y**	–	–
	True N	0.571 ± 0.072 (238)	0.547 ± 0.083 (480)	**True N**	0.567 ± 0.079 (157)	0.563 ± 0.084 (956)
Phi-4B	**True Y**	0.599 ± 0.069 (851)	0.552 ± 0.077 (326)	**True Y**	–	–
	True N	0.583 ± 0.062 (214)	0.543 ± 0.084 (504)	**True N**	0.593 ± 0.066 (200)	0.557 ± 0.086 (913)
Qwen3-14B	**True Y**	0.593 ± 0.072 (938)	0.557 ± 0.078 (237)	**True Y**	–	–
	True N	0.580 ± 0.070 (169)	0.547 ± 0.082 (546)	**True N**	0.566 ± 0.083 (108)	0.563 ± 0.084 (1004)
Qwen3-8B	**True Y**	0.592 ± 0.073 (919)	0.564 ± 0.077 (248)	**True Y**	–	–
	True N	0.581 ± 0.070 (167)	0.547 ± 0.082 (547)	**True N**	0.584 ± 0.073 (121)	0.560 ± 0.085 (982)

Table 7: Confusion matrices with semantic similarity between A and R (mean ± std with sample counts). Y/N indicates whether the reason correctly explains the assertion. In the setting "When either A or R is false", True Y cells are empty because false statements cannot logically provide valid explanations.

balanced measure of model performance. While MCC generally increases as model size scales up, it plateaus around $0.50 - 0.58$, even for the largest models. The highest MCC scores are observed in Phi-4-14B, Qwen3-32B, 14B, Qwen2.5-72B, 32B (0.55), which indicates modest gains in overall discrimination ability despite difference in explanatory accuracy.

4.2.2 When either A or R is False

In contrast, Table 6 shows that when either A or R contains factual errors, accuracy generally improves with model scale, with larger models performing remarkably well. Qwen3-14B achieves 94% accuracy, Qwen3-32B reaches 90%, Qwen3-4B attains 88%, and Phi-4-14B achieves 86%. Even moderate-sized models like Qwen3-1.7B (83%) and Llama3-1B (93%) perform reasonably well on this task.

These results indicates that larger language models have successfully learned that explanatory relationships must be grounded in factual correctness. When models recognize factual errors, they reliably reject the explanatory relationship, demonstrating a satisfactory factual verification capabilities and logical consistency. However, we observe a disparity in performance between the two settings: while models reliably reject explanations involving factual errors, their ability to discriminate between valid and invalid explanations when both statements are true is more limited. This difference suggests that although hallucinations or factual inaccuracies do not hinder performance in the Causal Explanation task, there is still room for improvement for models in identifying genuine explanatory relationships when both the assertion and the reason are factually correct.

4.3 Models Confuse Semantic Relatedness with Causal Relationships

To understand the mechanisms affecting model performance, we analyzed the semantic similarity between assertion and reason statements across different prediction categories. Our analysis focus primarily on the best-performing models, Phi-4-14B and Qwen3-14B, while also examining smaller models to assess consistency across model sizes. The results are shown in Table 7.

4.3.1 When both A and R are True

Table 7 reveals a clear pattern in how semantic similarity correlates with model predictions. For an unbiased model, we would expect similar semantic scores across all prediction categories. However, the performance shows a systematic bias. For instance, for Phi-4-14B model, semantic similarity is highest when the model correctly identifies causal relationships (0.590) and lowest when it correctly rejects non-causal ones (0.547). False positives (0.571) have notably higher similarity than true negatives, suggesting that semantic overlap often misleads the model into inferring causality. Conversely, false negatives (0.553) tend to occur when similarity is lower, causing the model to miss genuine causal explanatory relationship.

This pattern suggests that models are more likely to recognize causal explanatory relationships when the assertion and reason exhibit high semantic overlap, and conversely, they tend to miss valid explanatory relationship when similarity is lower. More interestingly, the higher similarity in false positives (Pred Y when True N) than in true negatives (Pred N when True N) indicates that models often confuses semantic relatedness with causal explanation.

These differences are statistically significant and

consistent across model families and sizes. Both Phi-4-14B and Qwen3-14B show nearly identical patterns, with similarity differentials of approximately 0.037 between correct and incorrect positive predictions, and 0.024-0.033 between incorrect and correct negative predictions. We include a more thorough statistical analysis in Appendix G.

4.3.2 When either A or R is False

Interestingly, when either assertion or reason contains factual errors, the pattern between semantic similarity and prediction changes. For Qwen3-14B, the semantic similarity for incorrect positive predictions (0.566) is only marginally higher than for correct negative predictions (0.563). This smaller differential (0.003) contrasts sharply with the larger gap (0.033) observed when both statements are true. This change implies that, in the presence of factual errors, models rely less on semantic similarity alone. The reduced gap indicates that models may be incorporating factual verification and partially overriding the tendency to infer causality from surface-level similarity.

4.3.3 Implications

These results demonstrate that language models rely heavily on semantic similarity as a proxy for predicting causal relationships. When two statements share more vocabulary or concepts, models are more likely to infer that one explains the other, regardless of whether a genuine causal relationship exists.

This reliance on semantic similarity explains the pattern observed in Table 6. As models increase sizes, they become more sensitive to semantic relationships between statements, which improves their ability to identify valid explanations but simultaneously becomes more prone to incorrectly inferring causality only based on semantic overlap.

These findings reveals an important limitation in current language models: they substitute correlation (semantic similarity) for causation (causal relationship). Despite improvements in many reasoning tasks with increased model size, this fundamental confusion between semantic relatedness and causal relationships persists across model families and parameter scales, suggesting a deeper architectural limitation in how language models process causal information.

5 Related Works

Evaluating causal reasoning in large language models has become an active research focus. CLadder (Jin et al., 2023a) assess formal causal inference through synthetic causal graphs covering associational, interventional, and counterfactual queries. Corr2Cause (Jin et al., 2023b) evaluates models' ability to infer causation from correlation using correlational statements, and CASA (Liu et al., 2024) assesses argument sufficiency via the probability-of-sufficiency framework, emphasizing logical fallacy detection.

Our work complements these existing benchmarks (Sonkar et al., 2024) by addressing a distinct but important aspect of causal reasoning: determining whether one natural language statement causally explains another in educational contexts. Unlike existing benchmarks that often use synthetic or abstract scenarios, CLEAR-3K draws from real-world educational materials spanning diverse subjects and grade levels. This distinction is critical because real-world applications frequently require assessing explanatory relationships between naturally occurring statements rather than applying formal causal inference rules.

Our findings situates alongside recent studies that question whether language models genuinely reason about causality or merely rely on pattern memorization (Zečević et al., 2023; Wu et al., 2024). We empirically demonstrate that models tend to substitute semantic similarity for causal explainability, recognizing assertion-reason pairs with higher lexical overlap as causal explanatory while overlooking valid causal explanations when semantic similarity is lower.

6 Conclusions

We presented CLEAR-3K, a dataset of assertion-reasoning questions designed to evaluate causal explanatory reasoning capabilities in language models. Through extensive evaluation of 21 models across five model families, we identify a fundamental limitation: language models tend to confuse semantic similarity with causal explanation. Models consistently rely on semantic overlap to infer explanatory connections, regardless of whether genuine causal relationships exist. This tendency persists across scales, where smaller models are overly skeptical about causal relationships while larger models become excessively permissive.

Despite improvements in other reasoning tasks

with increasing parameter sizes, causal reasoning performance remains limited, with MCC staying below 0.55. Our results indicate that the ability to distinguish correlation from causation does not emerge naturally through current scaling approach. Hence, CLEAR-3K provides a valuable benchmark for measuring progress toward genuine causal explanatory reasoning capabilities, which is an essential frontier for developing more capable and reliable language models.

Limitations

Our study focuses exclusively on open-source language models due to budget constraints that prevented us from accessing commercial API-based models such as GPT-4 or Claude at large scale (evaluating thousands of examples across multiple formats). Hence, we want to stress that the moderate results we achieved applies specifically to current LLM families we evaluated, not as a fundamental limit for all future models. Additionally, our evaluation specifically targets assertion-reasoning questions as a proxy for causal reasoning capabilities; other paradigms for assessing causality might reveal complementary insights. Future work could expand this analysis to include more diverse evaluation formats and extend to closed-source models through their APIs, potentially revealing whether the patterns we observe persist in these systems as well.

Acknowledgments

This work was supported by NSF SafeInsights 2153481 and ONR MURI N00014-20-1-2787.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Lorin W. Anderson, David R. Krathwohl, Peter W. Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul R. Pintrich, James Raths, and Merlin C. Wittrock. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York.

Benjamin S. Bloom, Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. David McKay Company, New York.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642. Association for Computational Linguistics.

Central Board of Secondary Education. 2020a. Cbse assessment framework for science, maths and social science classes 9 and 10.

Central Board of Secondary Education. 2020b. Cbse assessment framework for science, maths and social science classes 9 and 10. https://www.cbse.gov.in/. Central Board of Secondary Education, New Delhi, India.

Chanyeol Choi, Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, and Jy-yong Sohn. 2024. Linq-embed-mistral technical report. *arXiv preprint arXiv:2412.03223*.

Jonathan St BT Evans and Keith E Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3):223–241.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and 1 others. 2023a. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023b. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Sangeet S Khemlani and Daniel M Oppenheimer. 2011. When one model casts doubt on another: A levels-of-analysis approach to causal discounting. *Psychological bulletin*, 137(2):195.

Sandeep Kumar. 2018. Assessment in science education: A study of teaching effectiveness. *International Journal of Research in Social Sciences*, 8(1):669–690.

Xiao Liu, Yansong Feng, and Kai-Wei Chang. 2024. Casa: Causality-driven argument sufficiency assessment. *arXiv preprint arXiv:2401.05249*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

National Council of Educational Research and Training. 2020. National education policy implementation framework and learning outcomes for school education. https://ncert.nic.in/. National Council of Educational Research and Training, New Delhi, India.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial Winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Shashank Sonkar, Naiming Liu, MyCo Le, and Richard Baraniuk. 2024. Malalgoqa: Pedagogical evaluation of counterfactual reasoning in large language models and implications for ai in education. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15554–15567.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. Class: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4149–4158.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1112–1122.

Anpeng Wu, Kun Kuang, Minqin Zhu, Yingrong Wang, Yujia Zheng, Kairong Han, Baohong Li, Guangyi Chen, Fei Wu, and Kun Zhang. 2024. Causality for large language models. *arXiv preprint arXiv:2410.15319*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.

A Dataset Collection and Composition

The questions in CLEAR-3K were collected from standardized educational resources, including CBSE textbooks and examinations (Central Board of Secondary Education, 2020b) and NCERT curriculum materials (National Council of Educational Research and Training, 2020), with careful attention to maintaining diversity across academic subjects and complexity levels. Each question in our dataset adheres to the standard assertion-reasoning format described above, comprising an assertion (A), a reason (R), and a classification into one of the predefined four categories (a, b, c, or d).

A.1 Dataset Quality Validation

To rigorously validate our dataset quality, we randomly sampled 300 questions and showed them two qualified expert annotators, each with over three years of experience teaching STEM subjects. Annotators were provided with detailed guidelines and examples for each answer category, then asked to evaluate: "whether the assertion and reason are factually correct, and whether the reason provides a genuine causal explanation for the assertion."

The annotations achieved 96% agreement with our original labels (Cohen's κ = 0.94), confirming that CLEAR-3K represents authentic causal reasoning assessment consistent with educational standards. The few disagreements primarily involved borderline cases where the boundary between "related but non-explanatory" and "genuinely explanatory" relationships was subtle; these cases were reviewed and adjudicated through discussions.

B Model Parameters

We followed each model family's recommended prompting format and parameter settings (e.g., temperature, top-p) as described in their respective documentation. For Llama 3, we used the official chat template format with special tokens to structure conversations, with default parameters of temperature = 0.6 and top_p = 0.9. For Qwen familiy models (Qwen2.5 and Qwen3), we applied temperature = 0.7, top_p = 0.8, top_k = 20. For Gemma 3, we adopted the recommended inference settings of temperature = 1.0, top_k = 64, top_p = 0.95, and min_p = 0, with repetition_penalty = 1. For Phi-4 Reasoning models, we used temperature = 0.8, top_k = 50, top_p = 0.95.

C Supplementary Evaluation on GPT models

To assess generalizability beyond open-source architectures, we evaluated GPT-5-nano and GPT-5-mini using a stratified random sample of 400 questions (limited due to commercial API budget constraints). The results in Table 8 reinforce the patterns observed in open-source models: we identify a ceiling where reasoning capability plateaus at an MCC of 0.55 for GPT-5-mini, matching the best open-source models. Furthermore, the "permissive bias" persists in scaling. GPT-5-mini achieves higher explanatory accuracy (88%) but suffers a significant drop in rejection accuracy (66%) compared to the nano version (80%), confirming that larger closed-source models also prioritize semantic pattern matching over causal verification.

D Analysis of Assertion vs. Reason Falsehoods

To further assess whether the structural position of a factual error influences model performance, we analyzed Rejection Accuracy distinguishing between scenarios where the A is false versus where the R is false, as shown in Table 9.

The results demonstrate a notable symmetry in performance, indicating that accuracy remains stable regardless of which statement contains the error. For instance, Qwen3-32B shows a negligible difference ($< 0.1\%$) between scenarios (90.26% vs. 90.34%)and Qwen2.5-72B exhibits a similar pattern (86.69% vs. 86.52%). Despite minor variations in specific models (e.g., Gemma3), the general trend confirms that the Causal Explanation task is robust to the location of the hallucination.

E Performance on Traditional Assertion-Reasoning Format

Assertion–reasoning questions are typically presented in a 4-option multiple-choice format, where students are asked to evaluate the factual accuracy of two statements and the explanatory relationship between them, as shown below.

a. Both assertion (A) and reason (R) are true, and the reason correctly explains the assertion.

b. Both assertion (A) and reason (R) are true, but the reason does not explain the assertion.

c. The assertion (A) is true, but the reason (R) is false.

Model	Overall MCC	When both A, R are true				When either A or R is false
		Accuracy	MCC	Explanatory	Rejection	Rejection
GPT-5-mini	0.68	0.77	0.55	0.88	0.66	0.96
GPT-5-nano	0.62	0.77	0.54	0.74	0.80	0.94

Table 8: Supplementary evaluation of GPT models using a stratified random sample of 400 questions (100 per category) to assess generalizability.

Model	Overall (A or R False)	R False (%)	A False (%)
Qwen3-32B	0.90	90.26	90.34
Qwen3-14B	0.94	94.97	92.35
LLaMA3.3-70B	0.85	84.42	85.92
LLaMA3.1-8B	0.77	78.25	77.06
Qwen2.5-72B	0.87	86.69	86.52
Qwen2.5-32B	0.85	85.39	83.70
Phi-4-14B	0.92	93.34	90.34
Phi-4-4B	0.86	86.85	85.51
Gemma3-27B	0.67	68.67	64.19
Gemma3-12B	0.74	78.41	68.41

Table 9: Breakdown of Rejection Accuracy distinguishing between Assertion and Reason falsehoods. Results are presented for the two largest models from each model family.

d. The assertion (A) is false, but the reason (R) is true.

In our main analysis, we reformulated these questions into Causal Explanation Task that directly asks whether a reason explains an assertion. This allowed us to isolate causal explanatory reasoning from factual verification and examine whether models can distinguish genuine explanatory relationships from semantic similarity.

For completeness and connecting our findings with established educational assessment practices, we also evaluated model performance on the traditional 4-option assertion-reasoning question format. Table 10 presents confusion matrices for Qwen3 models of various sizes.

E.1 Analysis of Mistake Patterns

The confusion matrices in Table 10 reveal two important patterns:

1. **Persistent a/b confusion**: Across all model sizes, we observe substantial missclassification between options (a) and (b). The most common error is incorrectly classifying option (b) as option (a) (176-214 instances across models). Since both options involve factually true statements, this confusion directly reflects difficulty in determining whether one statement explains another.

2. **Scaling improves accuracy but not explanatory discrimination**: While overall accuracy increases with model scale (from 70.2% for Qwen3-4B to 73.2% for Qwen3-32B), the fundamental confusion between options (a) and (b) persists. Even the largest model, Qwen3-32B, misclassifies option (b) as option (a) in 195 cases (27.1% of all true (b) cases).

E.2 Connection to Main Findings

The results from the traditional assertion-reasoning questions reinforce several core findings from our main analysis. First, the persistent confusion between options (a) and (b) confirms our finding that models struggle to distinguish when a reason truly explains an assertion versus when two statements are simply related. This parallels our main finding that models rely on semantic similarity rather than deeper causal understanding.

Second, we can observe how models handle cases where one statement is false. For option (c) (assertion true, reason false), Qwen3-14B achieves 82.5% accuracy, and for option (d) (assertion false, reason true), it achieves 69.0% accuracy. This suggests that models generally reject explanatory relationships when they detect factual errors, but do so more reliably when the error is in the reason statement.

Third, the traditional format combines factual verification and causal reasoning assessment in ways that make it difficult to isolate specific reasoning failures. By evaluating both formats, we gain complementary insights: the traditional format reflects how models perform on standard educational assessments, while our Causal Explanation Task provides a targeted investigation into language models' causal reasoning capabilities.

The consistent patterns observed across both evaluation formats strengthen our conclusion that

Qwen3-4B (Acc: 70.2%)				
	Pred a	**Pred b**	**Pred c**	**Pred d**
True a	**963**	95	68	51
True b	214	**392**	74	38
True c	69	70	**452**	25
True d	109	36	47	**305**

Qwen3-8B (Acc: 71.0%)				
	Pred a	**Pred b**	**Pred c**	**Pred d**
True a	**934**	125	69	49
True b	176	**422**	72	48
True c	67	62	**467**	20
True d	113	30	40	**314**

Qwen3-14B (Acc: 72.8%)				
	Pred a	**Pred b**	**Pred c**	**Pred d**
True a	**923**	114	94	46
True b	188	**415**	77	38
True c	45	42	**508**	21
True d	84	19	51	**343**

Qwen3-32B (Acc: 73.2%)				
	Pred a	**Pred b**	**Pred c**	**Pred d**
True a	**961**	112	64	40
True b	195	**436**	53	34
True c	69	46	**481**	20
True d	115	26	32	**324**

Table 10: Confusion matrices for Qwen3 models on the traditional 4-option assertion-reasoning format.

current language models fundamentally confuse semantic similarity with causal relationships. This limitation persists across model sizes, architectures, and evaluation settings, pointing to a critical frontier for future model development.

F Performance Analysis by Subjects, Models types and Grade-Level

To investigate how causal reasoning abilities vary across knowledge domains, we conducted a detailed analysis of model performance by subject area and grade level. We selected the largest model from each family (Qwen2.5, Qwen3, Gemma3, and LLaMA3) and examined their performance using MCC specifically for cases where both assertion and reason are true. Our analysis reveals several significant patterns in how language models approach causal reasoning across different knowledge domains.

F.1 Subject-Specific Analysis

As shown in Table 11, biology presents the greatest challenge for causal reasoning across all models (MCC = 0.32), followed by physics (0.37) and geography (0.41). In contrast, models perform substantially better on mathematics (0.53) and economics (0.58). This pattern is remarkably consistent, with biology ranking as the most difficult subject for three out of four model families.

The difficulty hierarchy suggests a fundamental pattern: subjects involving complex, multi-factor causality (such as biological systems) present greater challenges for causal reasoning than subjects with more explicit rule-based, deterministic relationships (such as mathematics and economics). This pattern mirrors findings in human cognition research, where distinguishing correlation from causation is particularly challenging in domains with numerous interacting variables (Khemlani and Oppenheimer, 2011).

Subject Performance (Ordered by Difficulty)			
Subject	**MCC**	**Std Dev**	**Hardest For**
Biology	0.32	0.10	Gemma3 (0.21)
Physics	0.37	0.11	Gemma3 (0.28)
Geography	0.41	0.11	Llama (0.30)
Chemistry	0.42	0.12	Gemma3 (0.28)
Political Science	0.46	0.10	Gemma3 (0.36)
History	0.47	0.16	Gemma3 (0.23)
Mathematics	0.53	0.08	Llama (0.45)
Economics	0.58	0.06	Qwen2 (0.51)
Grade Level Performance (Ordered by Difficulty)			
Grade	**MCC**	**Std Dev**	
11	0.36	0.10	
12	0.36	0.09	
9	0.49	0.08	
10	0.49	0.11	
Overall	**0.43**	**0.11**	

Table 11: Causal reasoning performance by subjects and grade levels, measured by MCC when both A and R are true (mean and std across largest models from each model family).

F.1.1 Model-Specific Variations

Table 12 reveals interesting variations in how different model families approach causal reasoning across domains. Gemma3 shows particular difficulty with biology (MCC = 0.21) and history

Model	Bio	Phys	Chem	Math	Hist	Geo	PolSci	Econ
Gemma3	0.21	0.28	0.28	0.48	0.23	0.49	0.36	0.56
Llama	0.34	0.31	0.39	0.45	0.56	0.30	0.39	0.62
Qwen2	0.30	0.36	0.42	0.61	0.51	0.34	0.54	0.51
Qwen3	0.44	0.53	0.58	0.58	0.57	0.52	0.54	0.65

Table 12: Model-specific MCC performance by subject (larger values indicate better causal reasoning).

(0.23), while performing relatively well on economics (0.56). Llama struggles most with geography (0.30), while excelling at economics (0.62) and history (0.56). The Qwen family generally shows more balanced performance across subjects, with Qwen3 achieving the highest scores in seven out of eight subjects. This difference suggests that potential architectural or training dataset differences may significantly impact domain-specific causal reasoning capabilities.

F.2 Grade-Level Analysis

As expected, higher grade levels (11 and 12) present significantly greater challenges for models than lower grades (9 and 10). Both grade 11 and 12 questions yield an average MCC of 0.36, while grade 9 and 10 questions yield an MCC of 0.49. This pattern holds consistently across model families and suggests that more advanced educational content involves more subtle causal relationships that current language models struggle to differentiate from simple semantic relatedness. The increased difficulty at higher grades likely reflects both greater content difficulty and harder causal relationship that require deeper domain understanding.

F.3 Implications

These findings have important implications for applications of language models in both educational contexts (Sonkar et al., 2023) and causal explanatory reasoning:

1. **Causal reasoning is domain-sensitive.** The substantial variations in performance across subjects suggest that evaluations of causal reasoning should consider domain-specific challenges rather than treating reasoning as a uniform capability.

2. **Model Architectures matter.** Model family differences suggest that some design and training strategies may better support causal reasoning, particularly in complex domains.

3. **Current models are not very reliable for advanced educational tasks.** The difficulty gradient across grade levels indicates that current models may be more reliable for causal reasoning in introductory educational contexts than in advanced subject matter.

This analysis complements our main findings by demonstrating that while semantic similarity affects causal judgments across all domains, the magnitude of this effect still varies by subject area and grade level.

G Statistical Analysis of Semantic Similarity Effects

To rigorously evaluate our hypothesis that models use semantic similarity as a proxy for causal relationships, we conducted a series of statistical tests. Specifically, we examine whether semantic similarity between assertion and reason statements significantly predicts model errors, and whether this effect varies based on the true explanatory status.

G.1 Methodology

We analyze the model performance of Phi-4-14B on cases where both assertion and reason are factually true. The semantic similarity values used in this analysis correspond to those reported in Table 7.

Data Preparation We constructed a synthetic dataset that follows the distributions observed in our experiments with Phi-4-14B. Each data point includes the following attributes:

- **true_class**: Whether the reason actually explains the assertion (Y) or not (N)
- **pred_class**: Whether the model predicted an explanatory relationship (Y) or not (N)

- **similarity**: The semantic similarity between assertion and reason statements
- **error**: Whether the model's prediction was incorrect (1) or correct (0)

The constructed dataset includes:

- 1048 correct identifications of explanatory relationships (True Y, Pred Y), with mean similarity 0.590
- 129 missed explanatory relationships (True Y, Pred N), with mean similarity 0.553
- 480 correct rejections of non-explanatory relationships (True N, Pred N), with mean similarity 0.547
- 238 incorrect inferences of explanatory relationships (True N, Pred Y), with mean similarity 0.571

Statistical Tests We conducted three complementary analyses to assess the influence of semantic similarity:

1. **Logistic regression analysis**: This statistical model predicts the probability of an error based on semantic similarity, true class, and their interaction.We use the formula

   $$\text{error} \sim \text{similarity} \times \text{true\_class}$$

   to test for:

   - Whether semantic similarity affects error rates
   - Whether this effect differs depending on the true relationship status
   - How much variance in errors can be explained by similarity

2. **With-in Class t-tests**: For each true class (Y and N), we performed independent t-test to compare the semantic similarity score between correctly and incorrectly predicted cases. These t-tests determine whether the differences in similarity are statistically significant or could have occurred by chance.

3. **Prediction-based t-test**: We compared similarity scores between all cases where the model predicted "Yes" versus all cases where it predicted "No", regardless of ground truth. This tests whether model predictions systematically correlate with similarity levels.

G.2 Results

The logistic regression results showed:

- A significant positive coefficient for similarity (coefficient $= 4.41$, $p < 0.001$), indicating that higher similarity generally increases errors for the reference category (True N)
- A significant negative interaction term (coefficient $= -12.60$, $p < 0.001$), indicating that the effect of similarity reverses for explanatory relationships (True Y)
- A pseudo R^2 of 0.097, meaning that similarity explains approximately 10% of the variance in model errors

The t-test results confirmed:

- For explanatory relationships (True Y), correctly identified cases have significantly higher similarity (0.590) than missed ones (0.553), $t = 6.05$, $p < 0.00000001$
- For non-explanatory relationships (True N), incorrectly predicted cases have significantly higher similarity (0.571) than correctly rejected ones (0.547), $t = -4.71$, $p = 0.00000315$
- Overall, statements predicted as explanatory relationships have significantly higher similarity than those predicted as non-explanatory, regardless of ground truth. $t = 11.04$, $p < 0.00000001$.

G.3 Implications

These statistical results provide strong statistical evidence for our main claim that models substitute semantic similarity for causal understanding. The highly significant effects confirm that semantic similarity systematically biases model predictions in both directions:

1. When true explanatory relationships exhibit lower-than-average similarity, the model tends to reject them.
2. When non-explanatory pairs exhibit higher-than-average similarity, the model incorrectly predicts causal relationships.

This pattern represents a fundamental confusion of correlation (semantic similarity) with causation (explanatory relationship). These findings reinforce

our conclusion that current language models have not developed genuine causal reasoning capabilities that can reliably distinguish semantic association from explanatory relationships.