Stress-Testing the Robustness of LLM-Based Causal Inference

A Framework for Assessing Reliability and Uncertainty in AI-Causal Reasoning

Ankitkumar Patel • Jigarkumar Patel • Amit Kumar • Venkatesh Kulkarni • Supreeta Lahari

Presenter

Amit Kumar

Texas A&M University–Corpus Christi

akumar3@islander.tamucc.edu

Meet the Presenter: Amit Kumar

Academic & Professional Background

Amit Kumar is a PhD student in Computer Science at **Texas A&M University–Corpus Christi**. He currently serves as a Research Assistant with a background in applied machine learning and analytics.

Core Research Focus:

- Causal inference and LLM robustness
- Time-series forecasting
- Hate speech detection
- Socially relevant NLP applications

Select Publications

CryptoPulse: Short-Term Crypto Forecasting with Dual-Prediction
IEEE BigData 2024

StockTime: A Time Series Specialized LLM Architecture for Stock Price Prediction
NAACL 2025

Can LLMs Truly Reason About Cause and Effect?

THE SHIFT

Moving from statistical methods to LLM-based causal discovery.

PROBLEM

LLMs relies on surface-level pattern recognition rather than genuine logic.

GAP

Prior literatures show LLMs barely exceed random chance and collapse under simple perturbations like variable renaming.

The Systematic Stress-Test Framework

Six Dimensions of Causal Reliability

Model Scale

Impact of size and pretraining objectives

Graph Complexity

Performance failures at capacity limits (node/edge)

Prompting Strategy

Scale-dependent effects of CoT vs. Zero-shot

Metadata Quality

Non-linear trade-offs in description richness

Data Quality

Precision and formatting of observational data

Uncertainty

Sensitivity to ambiguous or borderline causal links

The Problem With Traditional Metrics For Causal Links

Why Binary Evaluation Is Deceptive?

Traditional metrics (Precision, Recall, F1) assume a **binary framework**, failing to capture the nuance of causal discovery.

Optimism Bias

LLMs tend to **over-identify** causal links, leading to low precision and a high volume of false positives in discovery tasks.

Epistemic Uncertainty

Models frequently violate basic probabilistic constraints, undermining the validity of the metric calculations.

```
P(e) + P(not e) != 1
```

The Uncertainty-Aware Metric (F1_adj)

A Ternary Decision Framework

Three Outcomes: Edge, No-Edge, or Uncertain.

Adjusted Recall (R_adj)

$$R_{adj} = TP / (TP + FN + \beta \times UE)$$

UE: uncertain edges that were truly present.

Adjusted Precision (P_adj)

$$P_{adj} = TP / (TP + FP + \alpha \times UN)$$

UN: uncertain no-edges truly absent.

Adjusted F1 (F1_adj)

$$F1_{adj} = 2 \times (P_{adj} \times R_{adj}) / (P_{adj} + R_{adj})$$

The ultimate measure of uncertainty-aware reliability.

Weighting (α, β): Controls the penalty for abstention vs. incorrect guessing.

This metric penalizes confident errors more severely than informed abstentions, providing a more honest reflection of reliability.

Performance Benchmarks (Asia Network)

Asia Network - Medical Dataset

Dataset contains **eight binary variables** related to pulmonary diseases and clinical symptoms.

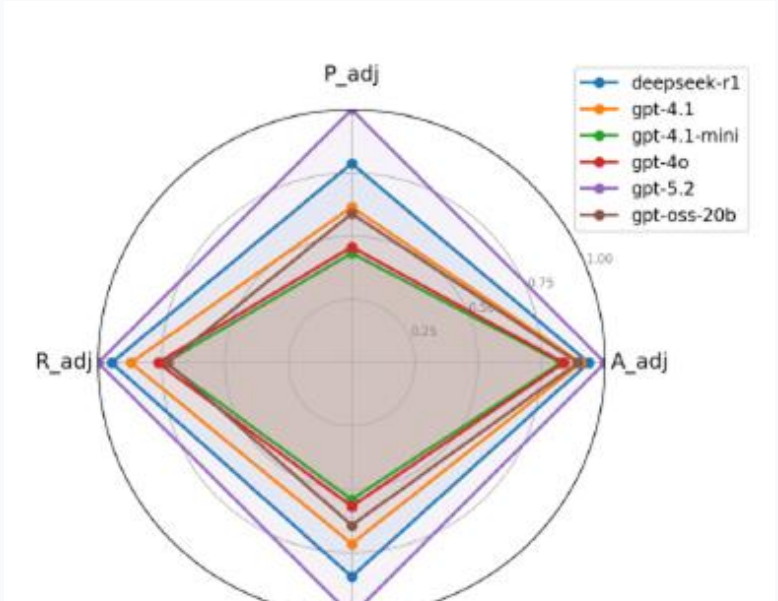
The Leaders

GPT-5.2 and **DeepSeek-R1** show near-perfect performance and high stability.

trending_up **Top Tier Reliability**

High Sensitivity

GPT-4o and **GPT-OSS-20B** performance drops significantly (up to **0.44**) when uncertainty is penalized.



Metadata: The Great Equalizer?

Non-Linear Gains from Description Richness

Low Richness (L1)

Larger models (GPT-4.1) lead significantly.

High Richness (L3)

Smaller models (GPT-4.1-mini) narrow the gap, improving by **187.4%**.

The Convergence

At L3, the smaller variant slightly leads **(0.842 vs 0.824)**.

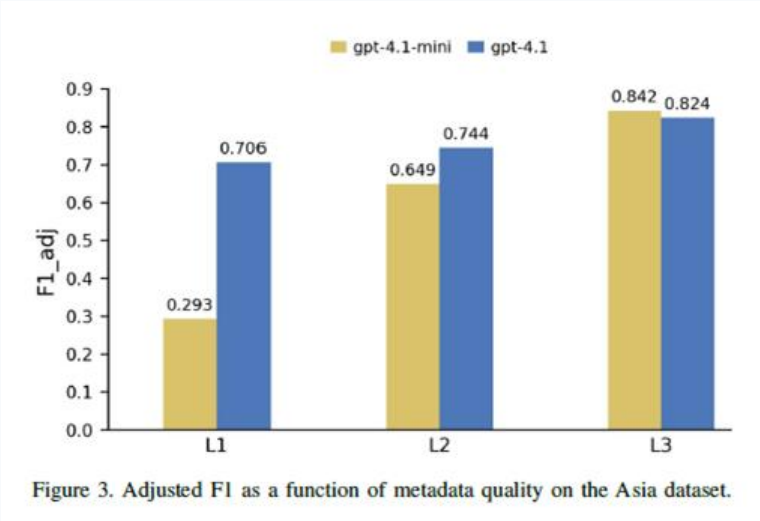


Figure 3. Adjusted F1 as a function of metadata quality on the Asia dataset.

Performance Comparison

Evaluating Reliability with the F1_adj Metric

The Leader: GPT-5.2

Achieves substantially higher median **F1_adj** scores and tighter performance distributions. High deployment cost makes it impractical for all cases.

Stable Alternatives

Models like **GPT-4.1** and **DeepSeek-R1** maintain reasonably stable performance under favorable conditions at a lower cost.

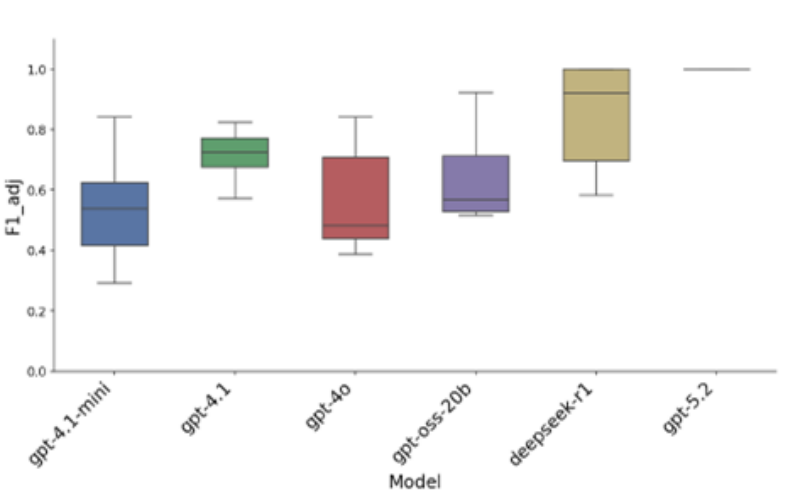


Figure 5. Distribution of adjusted F1 scores across models.

Strategic Tradeoff: The uncertainty-aware F1_adj metric reveals when lower-cost models are reliable enough and when their variability poses unacceptable risk.

Sensitivity Analysis of α and β

Quantifying the Impact of Uncertainty Penalties

Metric Mechanics

We vary **α** and **β** to control how strongly uncertain predictions are penalized in the uncertainty-aware metric.

Stability Insights

GPT-5.2 and **DeepSeek-R1** remain stable, whereas **GPT-4o** is more sensitive to stricter penalties.

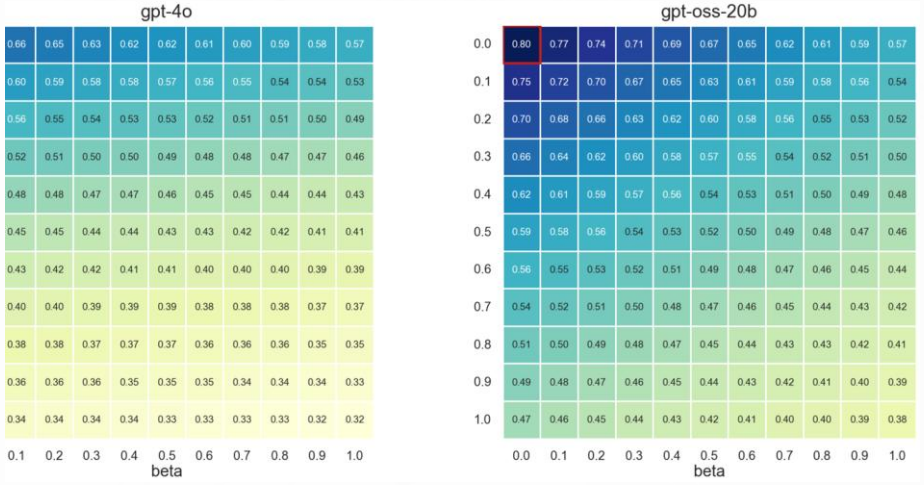


Figure 4: Sensitivity of F1_adj to α and β parameters across models.

Conclusion: Uncertainty-aware evaluation gives a more realistic measure of causal reasoning reliability by revealing fragile gains.

Conclusion & Future Directions

Roadmap to Robust Causal Pipelines

Summary

Scaling improves outcomes, but gains are fragile and metadata-dependent.

Key Takeaway

Use **F1_adj** to decide when lower-cost models are sufficient and when frontier models are required.

Future Work

- Studying multi-agent debate mechanisms.
- Scaling to systems exceeding 200 nodes.