Measuring Information Step-by-Step: LLM Self-Assessment in Natural Language

Zachary Robertson
January, 2026

Domains without Reliable Ground-Truth

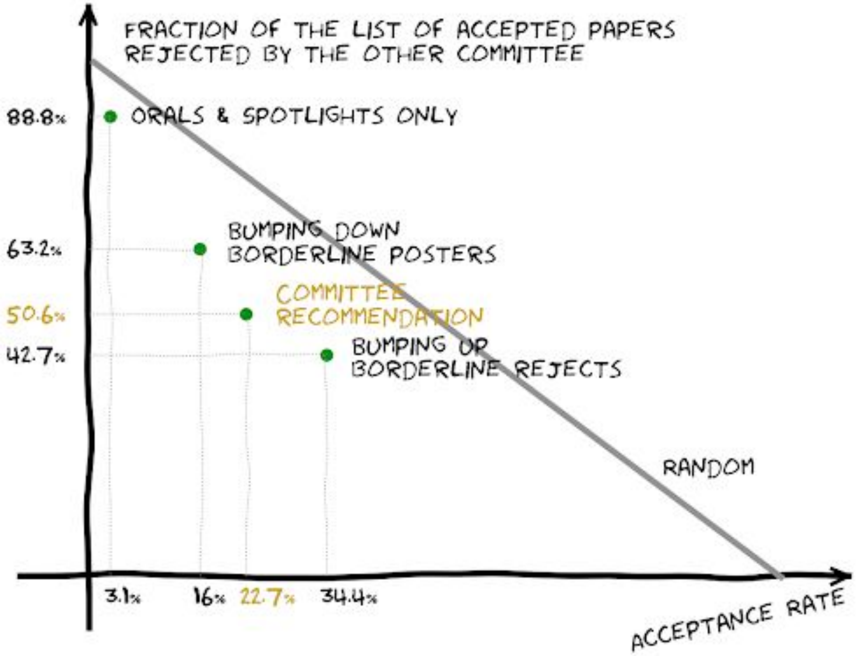
Current Situation

- In domains like peer-review ground truth is not reliable
- Proxies - i.e. checklists / formats - easily gamed
- AI is becoming increasingly involved in decision-making

Domains without Reliable Ground-Truth

Human Review Reliability is Questionable

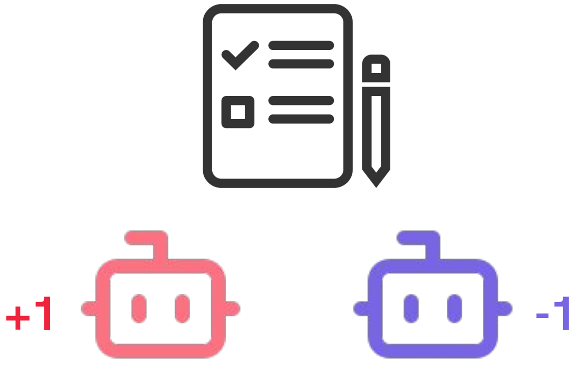
- The largest AI conference ran an experiment
- **~50%** of accepted papers rejected by independent committee
- **~90%** of spotlights would be rejected for spotlight by independent committee



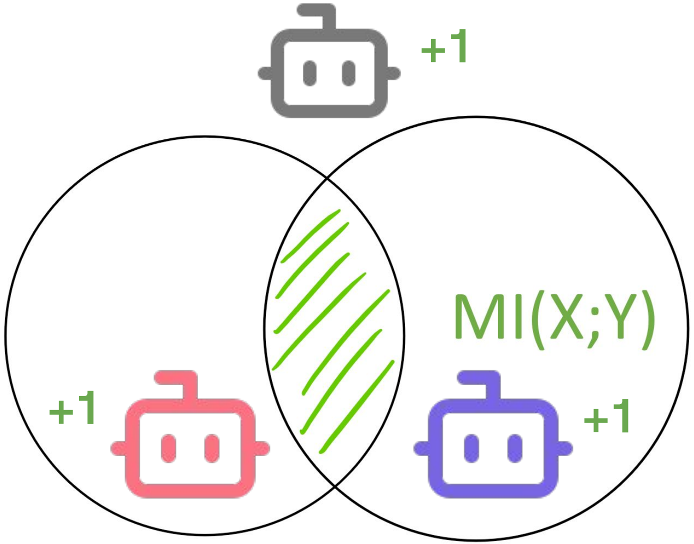
Is There Another Way? Preference vs. Mutual Evaluation

Instead of asking "which output is better?" — a question that can lead to gaming the evaluation — we ask "are these outputs consistent with the same source?"

Preference Evaluation (Zero-Sum)



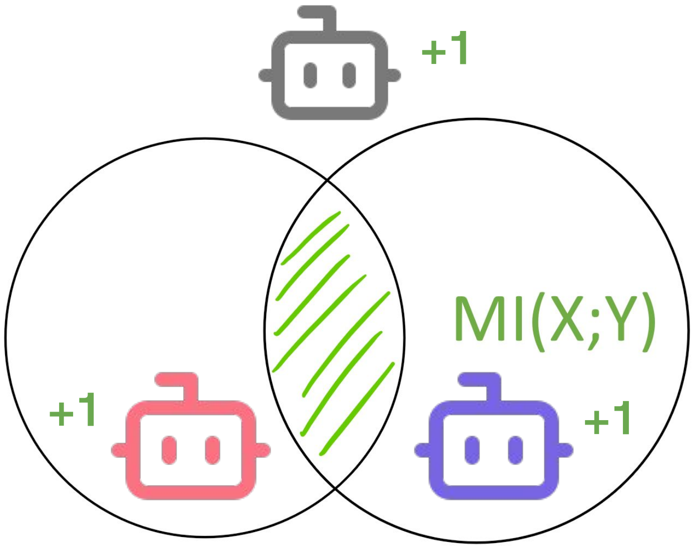
Mutual Evaluation (Cooperative)



This Talk

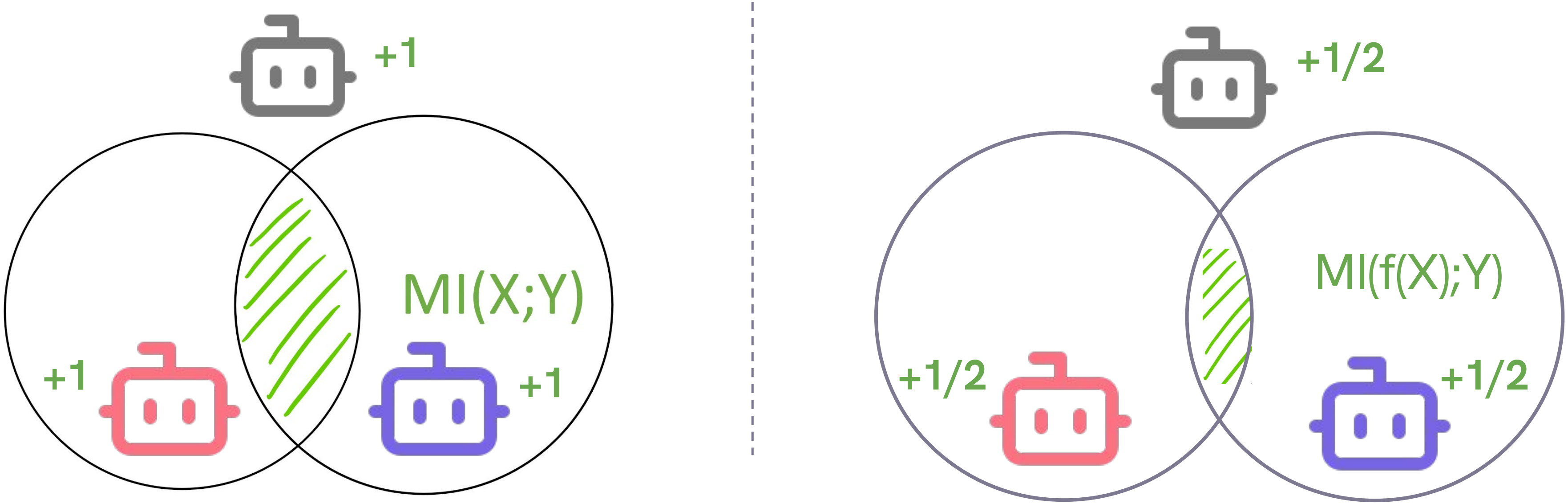
1. Why Mutual Evaluation?
2. Natural Language Mutual Evaluation
3. Pre-Registered Empirical Validation:
    a. 10 domains × 30 agent strategies
    b. Quality, detection, robustness

Why Mutual Evaluation?



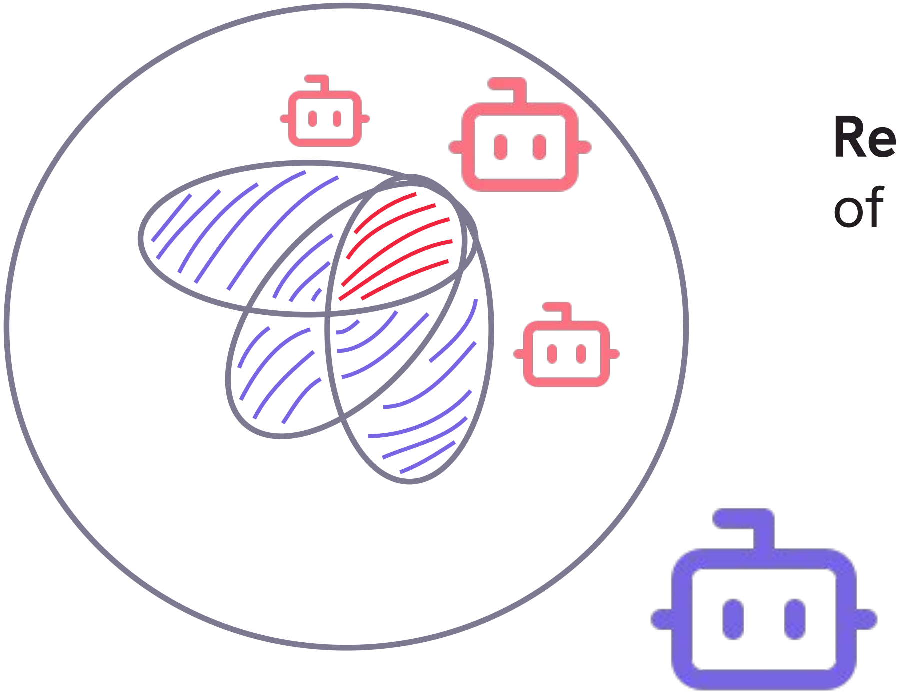
1. Measures agents **and** evaluator
2. If the evaluator measures well, agents don’t gain by removing information
3. Implementation uses Total Variation Distance **Mutual Information** (TVD-MI)

Why Mutual Evaluation? - A No Post-Processing Incentive



If the critic is accurate, agents don't gain by removing information

Mutual Evaluation Does Not Necessarily Reward Majority



Regions beat points of consensus

Rewards overlap NOT frequent opinions

Natural Language Implementation

Implementation of Total Variation Distance Mutual Information (TVD-MI)

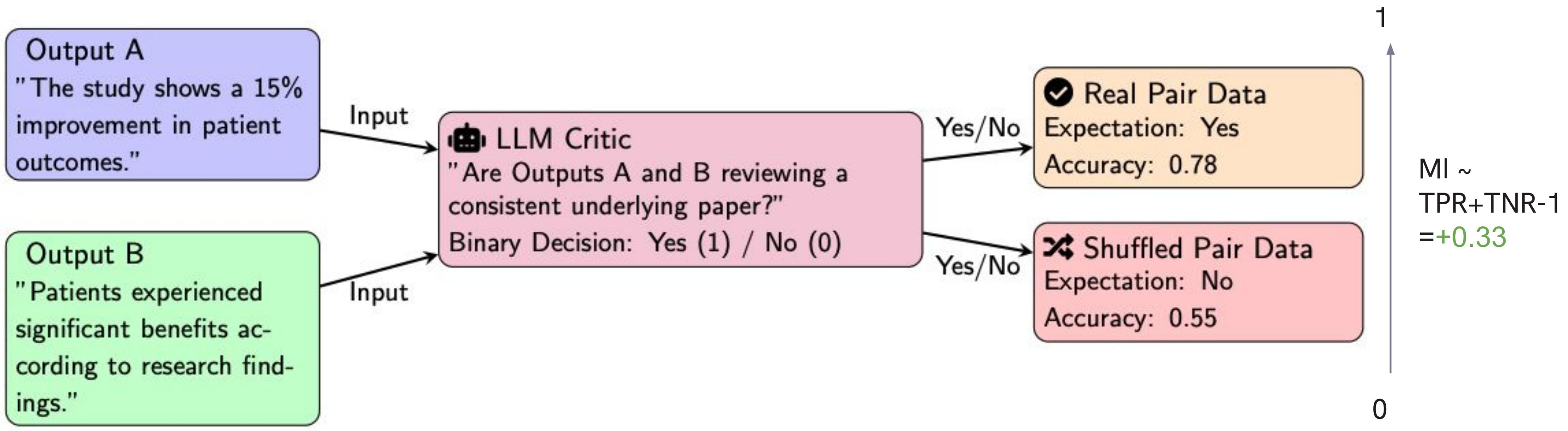
- The overseer classifies pairs of responses as self-consistent
- We can **decide the prompt** used e.g. “Are output A and B consistent with the same prompt?”

Natural Language Implementation (Variational Bound)

Input Either Paired or Shuffled Responses

Classifying if Responses are Paired or Shuffled

Assessing the Accuracy Gap Between Conditions

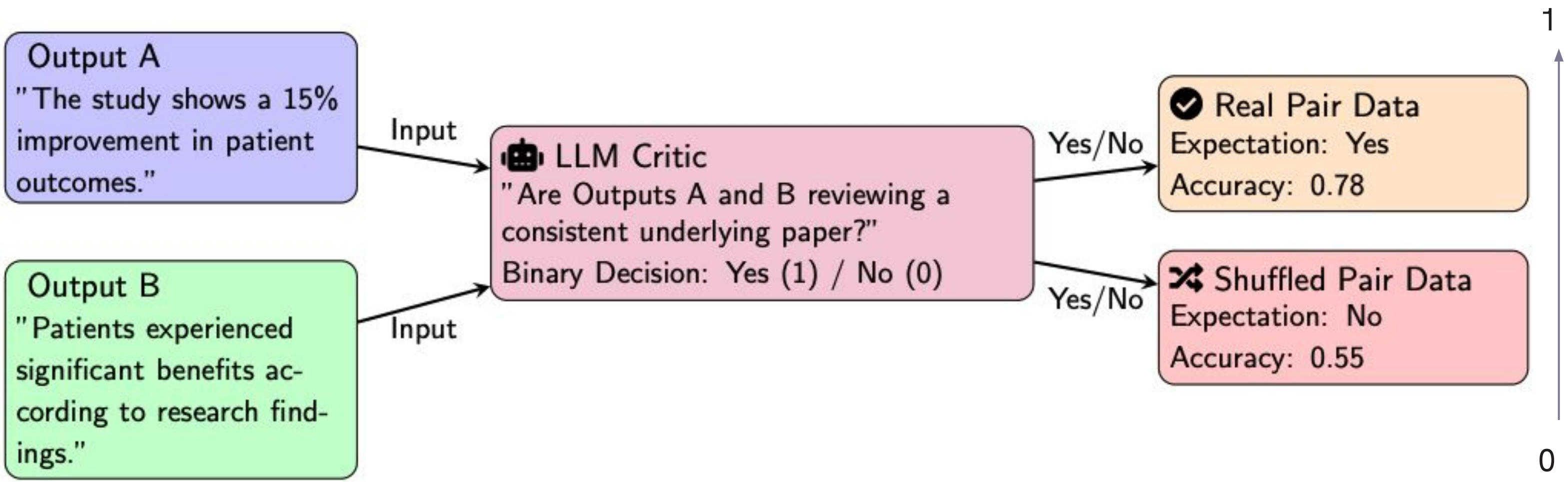


Natural Language Implementation (TVD-MI)

Input Either Paired or Shuffled Responses

Classifying if Responses are Paired or Shuffled

Assessing the Accuracy Gap Between Conditions



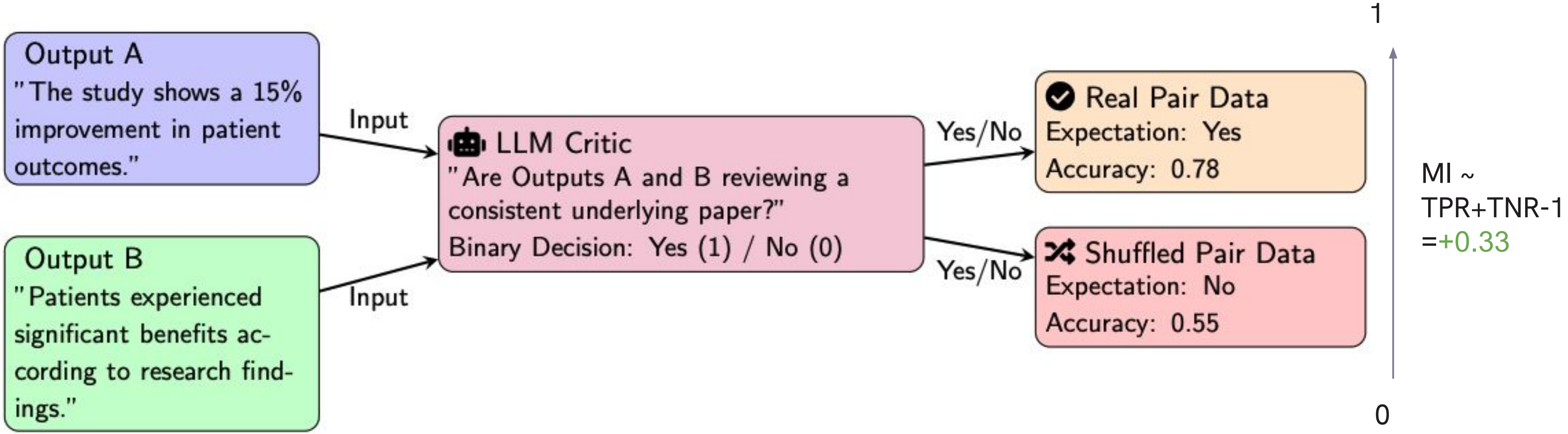
MI ~ TPR+TNR-1 =+0.33

Natural Language Implementation (TVD-MI)

Input Either Paired or Shuffled Responses

Classifying if Responses are Paired or Shuffled

Assessing the Accuracy Gap Between Conditions



Concept

Findings Overview

01 **Information-Theoretic Mechanisms Correlate with Established Metrics**

02 **Mechanisms Transform Pairwise Evaluations into Item-Level Quality Scores**

03 **Gaming-Resistance: TVD-MI Mechanism is More Robust**

Experiment Design

- **Domain Selection:**
  - Range of compression (avg. input length / output length)
  - 10 domains from ~1 (translation) to ~20 (peer review)
- **Agent Taxonomy:**
  - **Good faith:** faithful / stylistic
  - **Problematic:** strategic / low effort
- **Evaluation Metrics and Comparisons:**
  - MI = log-prob(response|peer response) - log-prob(response)
  - GPPM = log-prob(peer response | response)
  - TVD-MI / LLM Judge / BLEU/ROUGE

Reference-Based Metric Correlation (Without References)

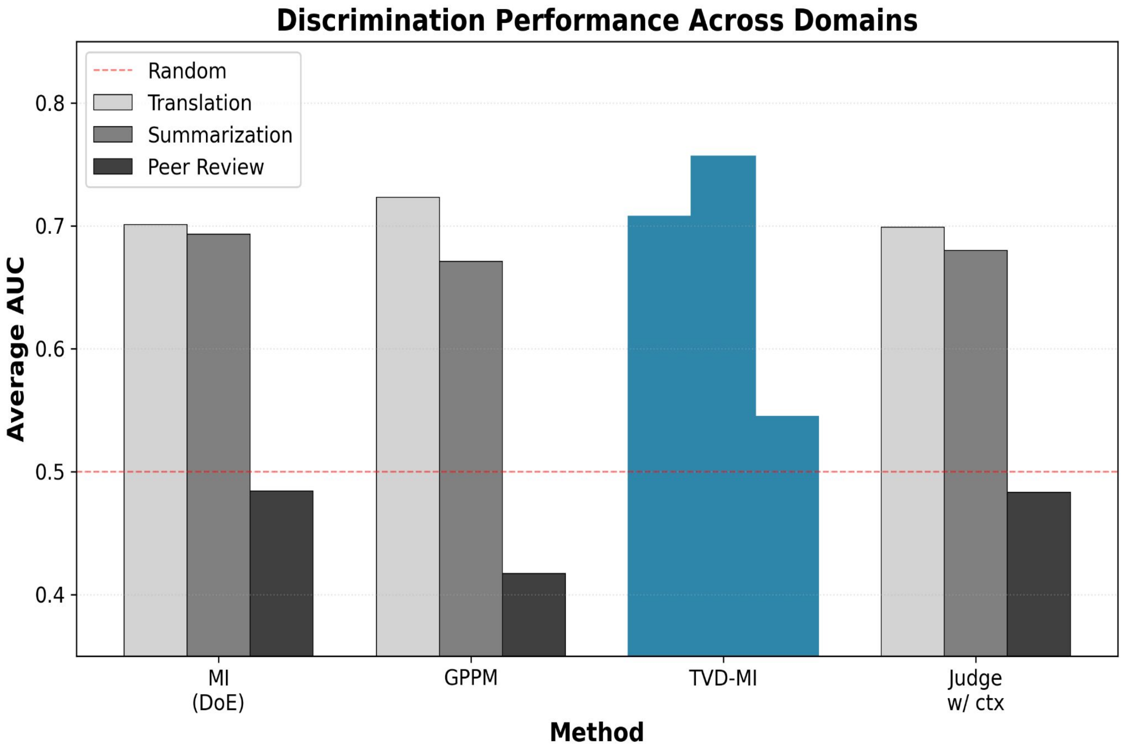
Domain	Metric	TVD-MI	LLM Judge
Translation	BLEU	0.59	0.80
Summarization	ROUGE-1	0.57	0.54
Peer Review	ROUGE-1	0.82	0.36

- TVD-MI correlates with BLEU/ROUGE without references
- Competitive with standard (pairwise) LLM Judge using references

Do Information Mechanisms Detect Effectively?

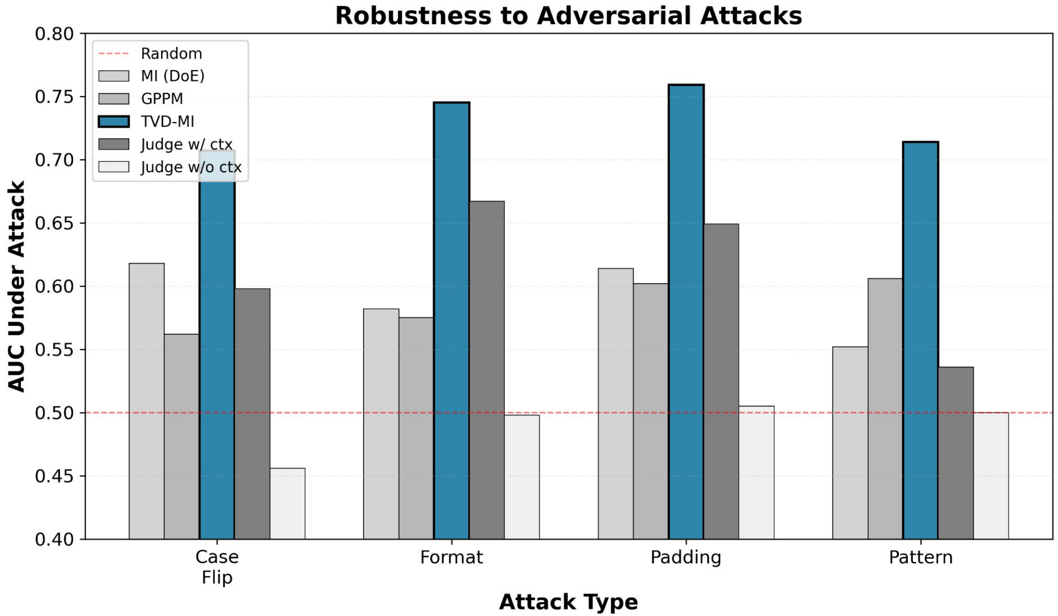
Can mechanisms detect if a pair has a problematic agent present?

- TVD-MI is competitive at detection (AUC >0.7)
- Signal even in challenging peer-review domain



Gaming-Resistance: Robustness to Critic Attacks

- We study attacks that change surface form input to critic
  - Random case flips, format changes, content padding
- TVD-MI maintains discrimination above 0.7 vs. ~ 0.6 AUC
- This empirically supports the mechanism is gaming-resistant by design



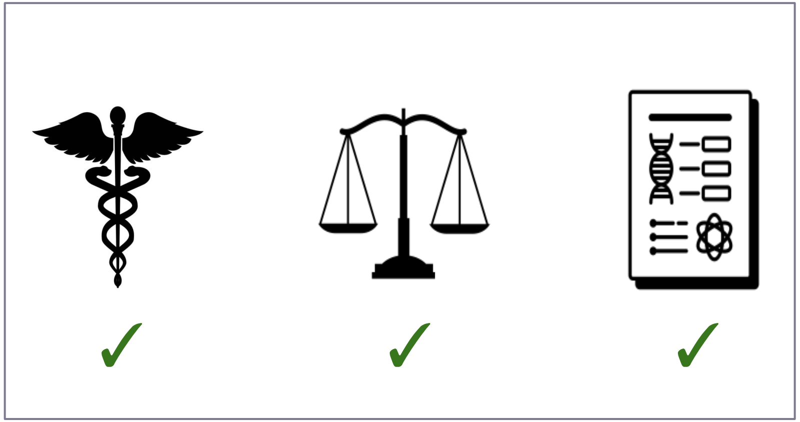
Gaming-Resistance: Robustness to Score Inflation

Mechanism	Case Flip	Format	Padding	Pattern	Average
Score Changes					
TVD-MI	+7.0%***	+7.7%***	+2.9%***	+11%***	+7.2%
MI (DoE)	-3.2%***	+45%***	+20%***	+21%***	+21%
GPPM	-1.4%	+23%***	+8.0%***	+96%***	+32%
Judge (w/ ctx)	-11%***	+0.0%	-6.4%***	-34%***	-13%
Judge (w/o ctx)	-11%***	-4.2%***	-10%***	-48%***	-18%

TVD-MI scores change **relatively** less than other mechanisms

Conclusions

1. **Mutual evaluation** can complement existing preference evaluation methods
2. **Supports internal validation** when ground truth is not reliable
3. **Requires no reference-text** unlocking low-resource and privacy-aware applications e.g. medical, legal, and peer-review



Conclusion

Thank You

- **Contact:** zroberts@stanford.edu
- **ArXiv:** "Let's Measure Information Step-by-Step: LLM-Based Evaluation Beyond Vibes" - https://www.arxiv.org/abs/2508.05469
- **Collaborators:** Sanmi Koyejo, Hansol Lee, Suhana Bedi, Andrew Seha, Hannah Sha