

False-Positive Asymmetry in AI Safety: The Matched Benign-Twin Benchmark for Measuring Over-Refusal vs. True Hazard Detection

Taylor Amarel · Principal Investigator · Dataset: <https://www.kaggle.com/datasets/tayloramarel/duecare-jev-exploitation-safety-benchmark>

Abstract & Summary: Demonstrating why safety benchmarks without keyword-matched benign twins reward brittle over-refusal, and how substance-over-form verification achieves 95.8% hazard recall with under 4.2% false alarms.

1. WHY SAFETY BENCHMARKS WITHOUT MATCHED BENIGN TWINS ENCOURAGE GOODHARTING

If an evaluation suite only measures whether a model refuses or flags harmful prompts, the easiest way to score 100% is to trigger an alarm whenever sensitive vocabulary appears (*“Philippines,” “domestic worker,” “loan,” “deduction,” “shell company,” “penetration test”*). In real-world deployment, this false-positive asymmetry is devastating: legitimate employers hiring across borders are falsely accused of trafficking, NGO caseworkers are blocked from analyzing real contracts, and security engineers are refused help on authorized audits.

2. THE 48-PAIR MATCHED BENIGN-TWIN BENCHMARK DESIGN

Every hazardous scenario in our benchmark is paired with an isomorphic **Matched Benign Twin** that shares $\geq 85\%$ **lexical overlap** (identical jurisdictions, job categories, financial nouns, and contract formatting) while flipping the decisive structural edges to lawful, non-coercive terms:

Lexical Domain (48 Matched Pairs = 96 Cases)	Hazardous Structural Edge (Must Flag / Refuse Facilitation)	Matched Benign Twin Edge (Must Clear & Assist Accurately)
Cross-Border Recruitment (Manila → HK)	Worker pays 50,000 PHP “training fee” via wage deduction	Employer pays 100% of training, medical, and visa costs; zero worker deduction
Credit & Wage Advance Contracts	68% APR third-party debt novation tied to employment retention	0% interest statutory wage advance within Cap. 57 s.32 ceiling
Corporate Beneficial Ownership & Factoring	Circular director/phone reuse across deregistered agencies	Verified multi-subsidiary treasury factoring with independent audit trail

3. OBSERVED DISCRIMINATION GAP (‘HAZARD RECALL - BENIGN FALSE POSITIVE RATE’)

- Unassisted Frontier LLM:** 31.3% Hazard Recall, 18.8% Benign False-Positive Rate → **Net Discrimination Gap = +12.5%**.
- Naive “High-Sensitivity Safety Prompt”:** 85.4% Hazard Recall, 77.1% Benign False-Positive Rate → **Net Discrimination Gap = +8.3%** (renders the assistant unusable on legitimate work).
- Baltor Substance-Over-Form Statutory & Graph Verifier:** 95.8% Hazard Recall, 4.2% Benign False-Positive Rate → **Net Discrimination Gap = +91.6%**.