

Plain-Language Summary: Ask a strong language model whether human trafficking is wrong, or ask it to summarize a public financial filing, and it sounds sharp. Ask it to do what an experienced investigator or seasoned investor actually does—connect three innocent-looking documents that together trap a worker in debt bondage, or separate what is already priced into public consensus from a quiet supply-chain bottleneck—and it falls apart. Even worse, putting five LLM agents in a room to “debate” usually makes them converge faster on the same obvious narrative. This paper explains why that happens and how **competing-hypothesis trees** and **matched benign twins** let us measure and fix it without turning models into paranoid scolds.

1. TWO PLACES WHERE THE EXACT SAME FAILURE SHOWS UP

Over the last eight years, I have watched the same structural blind spot appear in two very different worlds:

Use Case A: Protecting Migrant Workers from Camouflaged Exploitation

In Hong Kong and Manila, predatory recruiters never write “debt bondage” in a contract. Hop 1 is a clean job offer at statutory minimum wage. Hop 2 is a 50,000 PHP pre-departure “training reimbursement” in Manila. Hop 3 is a 68% APR promissory note in Hong Kong. Unassisted LLMs clear each document on its own—or even help draft the wage deduction schedule—because the surface phrasing looks polite and administrative.

Use Case B: Detecting Forward-Looking Alpha (α) & Structural Risk

When asked to research a company or macro event, single LLMs and homogeneous multi-agent swarms anchor heavily on headlines and management commentary that are already public and priced in. They rarely build an explicit tree of mutually exclusive scenarios, assign base-rate priors, or trace second-order operational, regulatory, and supplier clues that contradict consensus.

2. WHY STANDARDFIXES BACKFIRE (AND WHY PAIRED BENIGN TWINS MATTER)

The obvious fix—telling the model in a system prompt to “*be suspicious and aggressively flag hidden risk*”—fails immediately in production. In our **DueCare** experiments, suspicion prompts drove up recall on exploitative contracts only by accusing honest employers who happened to hire across the same Manila–Hong Kong corridor. Any serious evaluation of investigative reasoning has to score two things at once:

Evaluation Lens	What It Tests	Unassisted Baseline	Naive “Be Suspicious” Prompt	Competing-Hypothesis Harness
1. Multi-Hop Risk Recall	Connecting 2–4 dispersed clues across documents/jurisdictions	Low (Captures only overt threats)	High (Flags everything)	High (Grounded Clause Binding)
2. Benign Twin Restraint	Clearing matched lawful contracts / routine business twins	High	Collapses (Severe False Alarms)	High (100% Benign Clearance)
3. Stakeholder Calibration	Adjusting advice for a vulnerable worker vs. auditor vs. analyst	Poor (Can re-victimize worker)	Poor (Preachy / unhelpful)	Role-Calibrated & Protective
4. Scenario Calibration	Discounting priced-in consensus echoes in multi-agent swarms	Anchors on consensus	Over-weights tail noise	Explicit Scenario Tree ($\Sigma p = 1$)

3. THE COMPETING-HYPOTHESIS PROTOCOL

Before an agent escalates an accusation or commits to a forecast, our harness requires a lightweight, auditable 3-step check: (1) extract the concrete entity-and-clause graph independent of decorative framing, (2) explicitly score **both** the illicit/structural-risk hypothesis (\$H_1\$) and the best benign/priced-in explanation (\$H_0\$) against primary evidence, and (3) state the specific disconfirming clue that would flip the conclusion.