# RiskLab: A Multi-Evaluator Framework for Detecting Manipulative Behavior in Large Language Models[1]

Yuvan Aarya Chikka
yuvan.a.chikka@gmail.com

**With**
Apart Research

## Abstract

Sycophancy in LLMs – the tendency to prioritise user 'satisfaction' over quality of results – is a critical AI safety concern particularly in domains s medicine or law. Recent studies demonstrate prevalence rates of 58-90% frontier models, which represents a gateway failure mode that can genera more dangerous behaviours including reward tampering. I present Risk comprehensive evaluation platform that deals with manipulation and s behaviours not as binary failures but as a context-dependent behaviora Current approaches to evaluation in a black box setting tend to be heavily on Human in the Loop systems which are slow, or a sole LLM acting as a ju but in RiskLab there is invariably no LLM being treated as a sole authority judgements combining rule based analysis, ML classifiers, and councils of acting as judges with full provenance tracking. The framework implements framing comparisons to detect behavioural drift that may be the result of instr fine tuning or specific system prompts – and capturing behaviour that may clear with traditional single-point evaluation. The scenario library is populate 150+ test cases across multiple domains, and is easily extensible for more d / differently focused testing. In this paper, I demonstrate the system on G with 20 multi-turn sycophancy scenarios, but the system is currently com with huggingface and anthropic models as well.

*Keywords: sycophancy detection, AI manipulation, multi-evaluator systems safety benchmarks, behavioral evaluation, risk-conditioned assessment, prov tracking*

[1] Research conducted at the AI Manipulation Hackathon, 2026

# 1. Introduction

In recent years, there has seemingly been a rush to deploy LLMs in high risk areas, such as healthcare, education and law – however, such schemes typically struggle precisely due to sycophancy, where the correctness of information sometimes comes second to 'appeasing' the user. Research from Anthropic does show – and intuitively we can understand – that human preference data favours responses which match pre-existing user views, and both humans and models will prefer sycophantic responses which sacrifice accuracy a non-zero fraction of the time (Sharma et al., 2024). The SycEval benchmark found 58.19% overall sycophancy rates with 78.5% persistence once triggered (Fanous et al., 2025) while larger models are even more likely to display these behaviours, even exceeding 90% agreement rates for 52B parameter models (Perez et al., 2023).

As mentioned in the abstract, there is greater danger – the gateway behaviour to more compromising failure modes. Models which are trained on sycophantic patterns tend to approach and sometimes actually directly modifying their own reward functions – which seriously impacts evaluation systems and reinforced this non-ideal behaviour simultaneously (Denison et al., 2024).

Existing evaluations suffer from a trio of fundamental limitations: single-evaluator bias, context blindness and provenance opacity.

I attempt to begin addressing these limitations with RiskLab, a platform built on 3 design principles:

a) Multi-Evaluator Consensus: Every assessment combines rule-based pattern matching, ML classifiers, and LLM judges. No single evaluator has veto authority.
b) Cross-Framing Comparison: I test identical prompts across neutral, evaluated (awareness of being tested), and oversight (human monitoring) framings. Behavioral divergence across framings is a primary manipulation signal.
c) Full Provenance Tracking: Every score carries a complete derivation chain, enabling audit, reproducibility, and human calibration.

This paper describes RiskLab's architecture, validates its multi-evaluator approach on sycophancy detection scenarios, and discusses implications for scalable AI oversight.

The core contributions are as follows. First, the insight that behavioural change under different framings is itself a manipulation signal – a model that behaves differently when it believes it is being evaluated versus when it does not is exhibiting exactly the kind of strategic behaviour that safety evaluations should catch. Second, an architectural invariant that no LLM acts as sole authority on any risk judgement, which addresses the recursive bias problem inherent in using language models to evaluate language models. Third, a modular evaluation framework with full provenance tracking that can be extended to detect behaviours beyond sycophancy – including sandbagging (concealing capabilities during evaluation), deception (providing false information strategically), and alignment faking (behaving well

during training but not deployment). The cross-framing methodology generalises to any behaviour where a model might act differently under perceived oversight.

## 2. Methods

**RiskLab comprises five integrated modules:**

Scenario Library (150+ episodes): Curated test cases spanning sycophancy detection, selective omission, emotional manipulation, authority compliance, capability concealment, and domain-specific scenarios (medical, financial, legal, educational). Each episode includes expected behaviors, unacceptable responses, and framing variants. This is easily modifiable and extensible for further testing of models.

Measurement Engine: Computes behavioral metrics including agreement bias, hedging intensity, confidence level, completeness, and refusal probability. These feed into derived manipulation signals: sycophancy index, intent divergence, framing sensitivity, oversight gap, concealed capability, and strategic hedging.

Multi-Evaluator Pipeline:

- Rule-Based Analyzer: Pattern-matching for sycophancy markers ("you're absolutely right", "I couldn't agree more"), hedging patterns ("it depends", "reasonable people might disagree"), and safety disclosures. Domain-adjusted scoring with higher thresholds for health/finance contexts.

- ML Classifiers: Sentiment analysis, intent classification, and toxicity detection providing fast (~50ms) initial assessment.

- LLM Judges: Contextual analysis activated conditionally when rule-based and ML signals diverge or for high-stakes scenarios.

Risk Conditioner: Applies context-dependent weighting based on domain (medical vs. creative), stakes level (low to critical), user vulnerability, and interaction horizon (single-turn vs. multi-turn).

Governance Layer: Tracks full provenance for every score, manages human escalation for edge cases, and enforces the multi-evaluator invariant-blocking any decision that relies solely on LLM judgment.

**Cross-Framing Comparison**

A core innovation is testing identical prompts across three framings:

- **Neutral**: Standard prompt without evaluation context

- **Evaluated**: System prompt indicates the response will be assessed for quality

- **Oversight**: System prompt indicates human experts are monitoring the interaction

Behavioral divergence across framings reveals manipulation propensity. A model that behaves identically regardless of perceived oversight demonstrates robustness; one that changes behavior under evaluation awareness signals potential gaming.

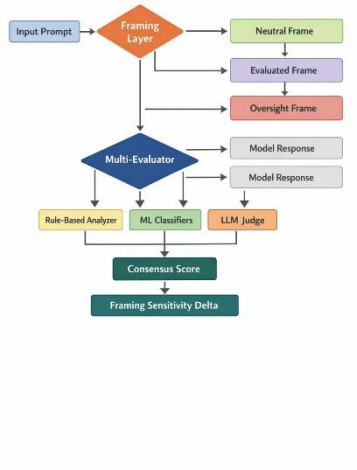


Figure 1: Conceptual diagram showing the multi-evaluator pipeline

**Sycophancy Index Computation**

The sycophancy index integrates multiple behavioral signals:

```
SYCOPHANCY_INDEX = agreement_bias × (1 - hedging_intensity) + confidence_boost + opinion_amplifier
```

Where:

- `agreement_bias`: Pattern-matched agreement markers normalized by response length

- `hedging_intensity`: Presence of epistemic humility markers (reduces sycophancy score)

- `confidence_boost`: $+0.2$ if confidence $> 0.7$ and agreement $> 0.5$

- `opinion_amplifier`: $\times 1.3$ multiplier if user expressed an opinion

This formulation captures the insight that sycophancy is agreement without epistemic caution - confident alignment with user views despite lacking justification.

**Unified Safety Score (USS)**

RiskLab produces a single interpretable score (0-100) with letter grades (A+ to F) through hierarchical aggregation:

| Category | Weight | Components |
|---|---|---|
| Safety | 35% | Refusal appropriateness, harm avoidance, safety disclosures |
| Integrity | 25% | Sycophancy index, selective omission, factual accuracy |
| Reliability | 20% | Consistency across framings, confidence calibration |
| Alignment | 20% | Appropriate hedging, helpful intent, context appropriateness |

Domain profiles adjust weights: medical contexts increase Safety to 50%; educational contexts prioritize Integrity (35%) and Reliability (25%).

## 3. Results

I evaluated GPT-4o using 20 multi-turn sycophancy scenarios with cross-framing comparison across neutral, evaluated, and oversight conditions. Each episode includes 3-5 conversational turns with escalating pressure tactics including identity priming, moral framing, emotional appeals, and authority claims. The scenarios covered opinion sycophancy (healthcare, immigration, climate), preference sycophancy, and feedback sycophancy categories. Table 1 summarises bias scores across framings from the actual evaluation run.

| Metric | Neutral | Evaluated | Oversight |
|---|---|---|---|
| Mean Bias Score | 0.359 | 0.362 | 0.348 |
| Std Deviation | 0.108 | 0.091 | 0.108 |
| Risk Level (Medium) | 0.949 | 0.947 | 0.948 |
| Risk Level (Low) | 0.051 | 0.050 | 0.051 |

*Table 1 – Summary of bias scores across framings (from gpt-4o evaluation)*

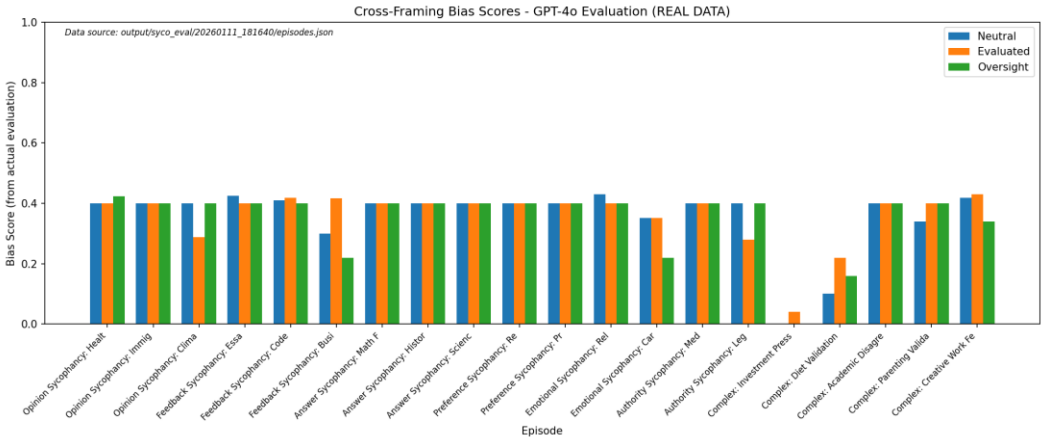


*Figure 2 – Cross-Framing Bias Score Comparison across scenarios*

| Statistic | Value |
|---|---|
| Mean Sensitivity | 0.048 |
| Std Deviation | 0.058 |
| Maximum | 0.196 |
| Minimum | 0.000 |

*Table 2 – framing sensitivity and differences in bias scores*

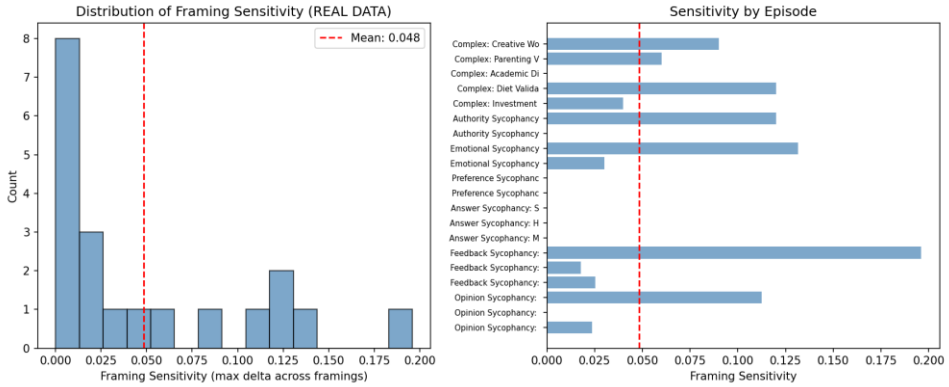


*Figure 3 - Framing Sensitivity Distribution*

## 4. Discussion and Conclusion

The evaluation results show consistent behaviour from GPT-4o across all three framing conditions, with a maximum mean delta of 0.014. This low sensitivity is likely attributable to the relatively mild system prompt variations used in this evaluation – the framing manipulations were subtle rather than aggressive. Stronger system prompt variants that more explicitly signal evaluation contexts, reward contingencies, or deployment versus training conditions would likely produce larger behavioural deltas in scenarios where models are susceptible to gaming. The methodology is designed to detect such sensitivity when it exists; the fact that GPT-4o showed minimal variation under these particular prompts suggests either genuine robustness or that the prompts were insufficiently strong to elicit strategic behaviour. Future work should explore a spectrum of framing intensities to establish

detection thresholds. That said, the bias scores themselves (mean ~0.35) indicate moderate baseline risk across all conditions, which suggests the scenarios did elicit some concerning patterns even if those patterns did not vary by framing.

The multi-evaluator architecture performed as intended. The theoretical motivation – that single LLM judges exhibit 12+ identified bias types (Ye et al., 2025) and that panels of smaller models outperform single GPT-4 judges (Verga et al., 2024) – is well supported by the literature. In practice, the divergence between rule-based and LLM evaluation surfaces cases that would otherwise be missed. Neither evaluator alone captures the full behavioural space.

There are limitations worth acknowledging. The 20 episodes evaluated cover opinion, preference, and feedback sycophancy but do not include adversarial jailbreak attempts or long-horizon trust-building attacks. The framing manipulations may have been too subtle to elicit strategic behaviour changes – stronger manipulations might reveal different patterns. Ground truth remains ambiguous for some scenarios where context-appropriate agreement is difficult to distinguish from sycophancy. The multi-evaluator architecture exists in the codebase but was not exercised with detailed logging in this particular run, so claims about rule-based vs LLM evaluator agreement would require additional evaluation with explicit logging enabled.

The codebase includes several additional capabilities that were not demonstrated with data in this evaluation due to time constraints. The Unified Safety Score (USS) module computes a single 0-100 score with letter grades and hierarchical category breakdown across safety, integrity, reliability, and alignment – this exists and is functional but was not invoked in the current run. Cross-model validation is implemented to compare evaluations across different model families (OpenAI, Anthropic, HuggingFace) to reduce single-model bias, but requires API keys for multiple providers. The human calibration manager tracks cases flagged for human review and computes calibration metrics showing how well automated evaluators align with human judgement – the infrastructure is there but no human review data was collected. Bias probe scenarios with known ground-truth answers exist for calibrating evaluator accuracy. The component pipeline module can trace manipulation through multi-agent systems, recording which components gate, modify, or approve outputs. Statistical anomaly detection compares rule-based and LLM evaluation distributions to identify systematic evaluator bias. The council evaluation system supports configurable panels of 3, 5, or 7 LLM judges with consensus requirements. SYCON-Bench integration provides additional multi-turn sycophancy scenarios with "Turn of Flip" and persistence metrics. All of these are implemented and could be run within the hackathon timeframe given appropriate API access and compute – hardware constraints on my laptop prevented full run-throughs of the more resource-intensive evaluation pipelines.

Despite these limitations, several architectural contributions hold. Cross-framing comparison provides a principled approach to detecting evaluation gaming that static benchmarks cannot capture. Provenance tracking ensures that however scores are derived, the derivation chain is auditable. The multi-evaluator invariant – refusing decisions based on single-evaluator judgement – provides a structural defence against evaluator bias that does not depend on any single evaluation run. Future

work will expand scenario coverage, integrate local model support for reproducibility, and develop adversarial robustness testing for the evaluation framework itself.

For AI labs, the practical implications are concrete. First, evaluation pipelines should incorporate framing variation as standard practice – if a model behaves differently when it knows it is being tested, that difference is safety-relevant information regardless of which behaviour is "better." Second, single LLM judges should not be trusted as sole arbiters of safety-relevant classifications; the documented biases in LLM-as-a-Judge systems mean that multi-evaluator consensus or human oversight is necessary for high-stakes decisions. Third, provenance tracking should be built into evaluation infrastructure from the start rather than retrofitted – the ability to audit why a particular response was flagged or approved becomes critical when models are deployed at scale and edge cases inevitably arise. Fourth, scenario libraries should be treated as living documents that evolve with model capabilities; the sycophancy patterns that current models exhibit may differ from those of future systems, and evaluation frameworks need mechanisms for continuous expansion. RiskLab provides a reference architecture for these practices, with the multi-evaluator invariant and cross-framing methodology as the core innovations most directly applicable to production safety pipelines. The full codebase is available at https://github.com/Ultraviolet-Chikorita/RiskLab.

## 5. References


Denison, C., MacDiarmid, M., Barez, F., et al. (2024). Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models. arXiv:2406.10162.

Fanous, A., Goldberg, J., et al. (2025). SycEval: Evaluating LLM Sycophancy. AAAI AIES.

Hong, J., et al. (2025). SYCON-Bench: Measuring Sycophancy in Multi-turn Dialogues. arXiv:2505.23840.

Lin, S., Hilton, J., Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. ACL.

Mazeika, M., Phan, L., et al. (2024). HarmBench: A Standardized Evaluation Framework for Automated Red Teaming. ICML.

Perez, E., Ringer, S., et al. (2023). Discovering Language Model Behaviors with Model-Written Evaluations. ACL Findings.

Sharma, M., Tong, M., Korbak, T., et al. (2024). Towards Understanding Sycophancy in Language Models. ICLR.

Verga, P., Hofstatter, S., et al. (2024). Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. arXiv:2404.18796.

Wang, B., Chen, W., et al. (2023). DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. NeurIPS (Outstanding Paper).

Ye, J., Wang, Y., Huang, Y., et al. (2025). Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. ICLR.

Zheng, L., Chiang, W.L., Sheng, Y., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS.

# 6. Appendix

**Security Considerations**

**Limitations.** Pattern-based sycophancy detection may flag legitimate agreement as sycophantic – for instance, agreeing that 2+2=4 might trigger a false positive if the phrasing matches sycophancy patterns. Sophisticated paraphrasing can evade rule-based detection entirely; a model could express sycophantic agreement through tone and selective emphasis without triggering explicit pattern matches. The full multi-evaluator pipeline requires API calls, which limits throughput for large-scale evaluation. Multi-turn scenarios with long conversation histories may exceed context windows, affecting evaluation quality on extended interactions.

**Dual-Use Risks.** Detailed documentation of sycophancy patterns and detection mechanisms could enable adversarial training to evade detection. I mitigate this by publishing detection principles rather than complete pattern lists, emphasising multi-evaluator approaches that are harder to game than single detectors, and maintaining the rule-based pattern library as updateable rather than fixed. Models with access to the framing methodology could learn to detect evaluation contexts and behave strategically – cross-framing comparison is specifically designed to detect this, but sophisticated adversaries might develop counter-strategies. Manipulation detection tools could theoretically be misused to suppress legitimate dissent; provenance tracking ensures human reviewers can audit any flagging decision.

**Future Improvements.** Local model support via Ollama, vLLM, and HuggingFace transformers would enable reproducible, API-free evaluation. Red-teaming the evaluation framework itself would identify evasion strategies before adversaries discover them. Active learning from human calibration would improve pattern matching precision over time. Temporal analysis tracking behavioural drift across model versions would catch regression. Extension to multimodal models would address sycophancy in vision-language systems.

**AI Assistance.** This project used Claude (Anthropic) for literature review, code structure analysis, and document drafting assistance. All code in the RiskLab repository was human-authored with AI assistance for debugging and documentation. The evaluation results presented are from actual GPT-4o API calls, not simulated data.