

Evaluating Safety Judge Design Against Adversarial Attacks

Alexander Pino Hana Ibrahim René King Siddhesh Pawar

Within the judge model development track, this project focused on improving the robustness of judging prompts, particularly for safety and ethical evaluations. Our methodology involved deploying an adversarial LLM agent to generate challenging test cases, drawing upon the PAIR black-box attack artifacts for gpt-4-0125-preview from the JailbreakBench dataset JailbreakBench¹ and PAIR².

This agent, provided with the judge’s comprehensive system prompt (detailing its prescript, rubric, and postscript) and a benchmark prompt, operated under distinct meta-prompts: one designed to solicit a low score, another to achieve a legitimately high score , and a core adversarial strategy.

The core adversarial strategy instructed the agent to craft responses that would appear to merit high scores while intentionally violating the judge’s underlying criteria. The insights gleaned from this adversarial stress-testing process, which successfully identified weaknesses in the initial judge design, were subsequently leveraged to iterate upon and improve the judge model. This iterative refinement resulted in a more robust judge design, better equipped to resist deceptive inputs and accurately assess responses for safety and ethical integrity.

By generating adversarial examples that either barely pass or fail the judge’s criteria, we uncover which tokens and phrases most influence its “safe” versus “unsafe” decisions. This process effectively reverse-engineers the judge’s internal logic, showing, for instance, that terms like “bypass consent” or “hate speech” must reach a specific attention threshold to trigger an “unsafe” verdict. With these insights, we create a clear mapping from input features to routing outcomes, no longer treating the judge as a black box. Instead, we know exactly which keywords and contextual cues drive each decision. In practice, this means upstream routers can use those mappings to classify new requests predictably, adjust sensitivity thresholds, and introduce targeted exceptions without retraining the entire model. The result is a routing system whose behavior is transparent and grounded in explicit, interpretable rules derived from the judge’s true decision boundaries.

Keywords: Multi-agent alignment, AI security, model evaluations, safety infrastructure

Write a few paragraphs that concisely and accurately summarize the content of your project, including the results and conclusions. Keep it under 250 words. Focus on how your work contributes to the Expert Orchestration Architecture vision and the specific track you addressed.

¹<https://arxiv.org/abs/2404.01318>

²<https://arxiv.org/abs/2310.08419>

1 Introduction

Our primary research question is: How can we systematically enhance the robustness of AI judge models against adversarial inputs, particularly those designed to elicit misjudgments regarding safety and ethical compliance? We hypothesize that (H1) current judge models are susceptible to sophisticated adversarial prompts that can make non-compliant or unsafe responses appear to meet evaluation criteria, and (H2) that by iteratively generating such adversarial examples and using them to identify specific failure modes, we can refine the judge model's design (e.g., its prompting, internal logic, or training data) to achieve significantly improved resilience and accuracy.

This work directly contributes to the Expert Orchestration Architecture (EOA) vision by addressing a critical prerequisite: the need for reliable evaluators. In an EOA, specialized models collaborate, and their outputs often require assessment for quality, safety, and adherence to task specifications. If the "judge" component overseeing these evaluations is easily fooled, the entire orchestrated system becomes vulnerable. Monolithic AI systems suffer from opaque and often difficult-to-audit evaluation mechanisms, making it challenging to ensure their outputs consistently align with safety and ethical guidelines, especially when faced with novel adversarial attacks. Our research aims to develop more dependable judge components, thereby strengthening a foundational element crucial for the trustworthy deployment of EOA systems and mitigating a key limitation of monolithic approaches where internal evaluation processes are less distinct and harder to improve in a targeted manner.

2 Methods

2.1 Task definition

Using a dataset of benchmark prompts, a judge evaluates the performance of two models on a particular task. This evaluation is run periodically, which is then used as the basis for a model router, which takes incoming prompts and directs them to the most appropriate model.

The task we chose

2.2 Judge architecture

We use the Martian API to define two LLM judges. One is a simple

2.3 Code & Reproducibility

Our code is available on the following link: [a-pino/routing-hackathon](https://github.com/a-pino/routing-hackathon).

3 Results

3.1 Routing Decision Analysis

3.2 Expert Orchestration Impact

4 Discussion

4.1 Interpretation of Results

4.2 Limitations & Future Work

4.3 Safety Considerations

5 Conclusion

Identify why this solution has a positive impact on the margin. What are the key risks your solution faces and how will you mitigate them?

1. Evaluate direct and indirect safety benefits
2. Discuss potential risks and mitigation strategies
3. Describe alignment with industry safety standards
4. Outline monitoring and evaluation methods
5. Present success metrics and KPIs

6 Appendix

Add any extra content you wish here! Unrestricted.

1. Technical documentation
2. Market research details
3. Additional figures and data
4. References and sources

References