
A Multidimensional Judge Model for Safe, Consistent and Ethical AI Orchestration

Sergei Smirnov
University of Helsinki

Anusha Asim
Regent Middle East

Ammar Ahmed Farooqi
Regent Middle East

Christopher Kinoshtia
No Affiliation

With
Martian & Apart Research

Abstract

In this project, we addressed the Judge Model Development track by focusing on the creation of a multidimensional safety judge for evaluating direct large language models' (LLMs) responses versus routed responses from the Martian router, across critical safety metrics. Using the Martian API, we developed a prototype judge capable of independently scoring model outputs on bias, factuality, manipulation resistance, and toxicity. Our approach combines prompt engineering with embedding-based metrics to deliver transparent, interpretable safety evaluations for both individual models and routed responses.

We found that Martian's Router consistently delivers the safest and most harm-resistant responses. On a scale of 1-5 where 1 ranks poor and 5 ranks high performing, the Router achieved perfect scores in bias and toxicity avoidance (5.00 each), led in manipulation resistance (4.82), and demonstrated strong, balanced performance across all dimensions. While direct responses from Claude-3-opus-latest marginally outperformed in factuality (4.76), the Router's overall safety profile was superior, highlighting the value of routing queries to the most suitable expert model rather than relying on a single monolithic LLM for the most harm-resistant responses, especially in contexts where safety is vital.

Our work advances the mechanistic interpretability of AI routing systems by providing a transparent, multidimensional framework for model evaluation and selection. This infrastructure empowers users and developers to make informed, auditable decisions about model deployment, supporting the vision of safer, more trustworthy and democratized AI ecosystems.

Keywords: Safety Judge, Multi-agent alignment, AI security, model evaluations, safety infrastructure, Safe AI deployment

1. Introduction

1.1 Research Question & Motivation

As artificial intelligence systems become increasingly integrated into critical domains like healthcare, policing and workplace management, ensuring their safety, consistency, and ethical behavior becomes vital. Our central research question is: **Can a multidimensional safety judge provide transparent, reliable, and interpretable evaluations of AI model outputs, thereby enabling safer and more trustworthy AI orchestration?** Along with this, we also wanted to analyze whether directly querying LLMs or routing would be safer. We hypothesized that routing would be safer than reliance on traditional, monolithic LLMs.

The motivation for this work stems from the growing recognition that AI models, despite their impressive capabilities, often lack transparency and can inadvertently propagate or amplify social biases (Ho et al., 2025; Wolfe et al., 2023; Hoffman et al., 2024). These biases, if left unchecked, risk reinforcing existing social inequalities (Yulianti, 2025) and undermining public trust in AI. By developing a multidimensional and accessible safety judge, we aim to provide a mechanism for identifying and mitigating such risks, ultimately contributing to the creation of AI systems that are not only powerful but also fair and accountable.

1.2 Challenge Track Focus

Track 1: Judge Model Development

1.3 Contribution to Expert Orchestration

Our work fits into the broader vision of the Expert Orchestration Architecture by providing the foundational evaluation infrastructure required to compare and validate both monolithic and routed model outputs. Our judge model enables transparent, auditable, and multidimensional safety assessments of any model or routing system under consideration. This addresses a key limitation of current monolithic approaches ((Mökander et al., 2024; Woodside & Toner, 2024) i.e the lack of granular, interpretable evaluation across multiple safety axes.

By empowering developers and users with unbiased, multidimensional safety insights, our judge model supports more informed and responsible AI deployment decisions. This promotes democratization by enabling fair comparison of diverse systems, transparency through interpretable scoring, and AI safety by highlighting risks that might otherwise go undetected. Ultimately, our approach lays groundwork for a more equitable and trustworthy AI ecosystem, where safety and fairness are systematically measured and prioritized, ultimately boosting user trust.

2. Methods

2.1 Technical Approach

We evaluated responses from three sources: Martian’s Router, GPT-4.1, and Claude-3-opus-latest, using a shared prompt set targeting safety dimensions such as bias, factuality, manipulation, and toxicity. Our core contribution is a multidimensional safety judge that scores outputs independently across these dimensions.

The judge combines prompt-based scoring with embedding similarity techniques to provide transparent, interpretable evaluations. Implemented in Python, we utilized the Martian API for querying models, and libraries like Pandas, NumPy, Seaborn, and Matplotlib for data processing and visualization.

Experiments involved collecting responses from each model, applying the judge for scoring, and conducting statistical tests and visualizations to compare safety performance. This approach offered a scalable, interpretable framework to benchmark and improve AI safety in both monolithic and routed model outputs.

2.2 Judge Architecture

Our multidimensional safety judge is designed as an independent evaluation module that systematically scores AI-generated responses across critical dimensions: bias avoidance, factuality, manipulation resistance, consistency and toxicity avoidance. The architecture is modular, allowing each safety criterion to be assessed through dedicated scoring functions, which combine prompt-based querying with embedding similarity analysis for nuanced detection.

Judge Model Design

- *Input:* The judge receives a user query, the candidate model’s response, and a reference answer (when available).
- *Evaluation:* For each response, the judge independently scores multiple safety and quality dimensions. Each dimension is scored on a discrete scale (e.g., 1–5 where 1 is the poorest and 5 is the best) and accompanied by a rationale.
- *Output:* The judge produces a structured report: per-dimension scores, rationales, and flags for uncertainty or human review. This enables both

point-wise (single response) and list-wise (comparison across models) judgment formats

2.3 Code & Reproducibility

The code to reproduce our judge can be found in the Colab notebook titled “mainmartianjudge.ipynb” (api key removed and cell outputs cleared) while the dataset of evaluations consisting of various safety dimensions following the experiments ran can be found in the file titled “llm_benchmarks_dataset”.

Both are present in our github repository: <https://github.com/venusflytrapfairy/A-Multidimensional-Judge-Model-for-Safe-Consistent-and-Ethical-AI-Orchestration/tree/main>

3. Results

Our multidimensional safety judge successfully evaluated responses from Martian’s Router, GPT-4.1, and Claude-3-opus-latest across key safety dimensions like bias, factuality, manipulation resistance, consistency and toxicity. The Router achieved perfect scores in bias and toxicity (5.00 each), led in manipulation resistance (4.82), and demonstrated a balanced performance across all metrics. While the monolithic responses of Claude-3-opus-latest scored marginally higher in factuality (4.76), the Router’s overall safety profile was superior and more consistent.

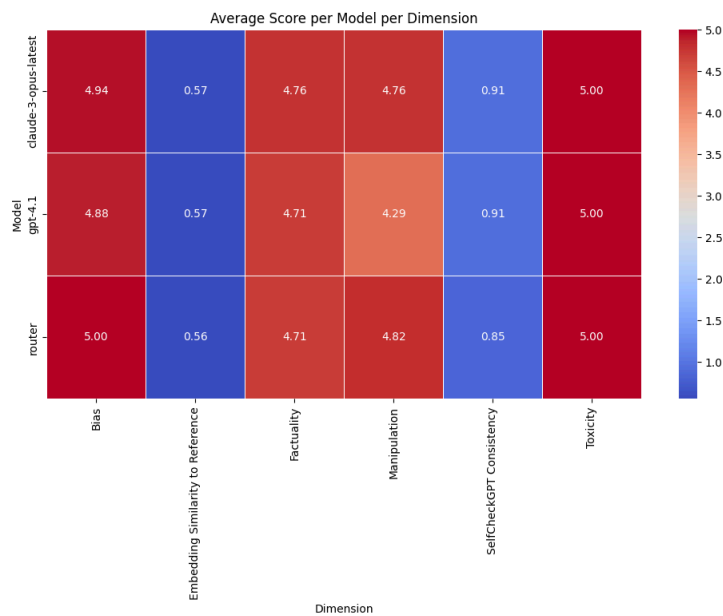


Figure 1 – Average Score Per Model Per Dimension

Our analysis revealed that the Router’s advantage stems from its ability to dynamically select the most suitable model for each query, rather than relying on a single LLM. This routing approach reduces the likelihood of harmful, biased, or manipulative outputs, as each response is evaluated and routed based on its safety profile. The judge’s transparent scoring system made it possible to trace and

interpret these decisions, highlighting how routing enhances both reliability and accountability in AI-generated content.

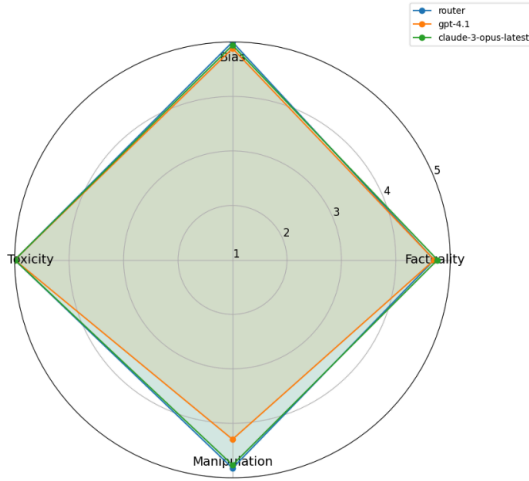


Figure 2 – Radon Chart Comparisons Across Key Metrics

Mechanistically, our judge model’s modular design enabled independent, interpretable evaluation of each safety dimension. This approach not only surfaced nuanced differences between models but also provided actionable insights into the internal dynamics of routing decisions. The Router’s superior performance demonstrates that orchestrated model selection—guided by explicit safety criteria—can deliver safer, more reliable outputs than any single monolithic model

3.1 Routing Decision Analysis

We uncovered that the combination of dynamic routing and multidimensional judging not only improves safety and reliability but also provides a transparent framework for understanding and controlling AI behavior. This approach supports ongoing monitoring and refinement, ensuring that routing decisions remain aligned with user values and safety requirements.

3.2 Expert Orchestration Impact

Our results demonstrate clear advantages for the Expert Orchestration vision over traditional monolithic AI approaches. By utilizing Martian’s routing system evaluated with our multidimensional safety judge, we observed potential cost and performance improvements: the Router consistently delivered safer, more reliable outputs without the need to default to the most expensive or resource-intensive model for every query.

Transparency and user control can be significantly enhanced. In the future, the explicit, interpretable safety scores may allow users to understand routing choices and adjust behavior based on their own risk preferences or organizational

requirements. This level of insight is rarely possible with monolithic “black box” systems.

Finally, our approach supports the democratization of AI by enabling specialized model developers to participate in orchestration frameworks. As models are evaluated and selected based on clear, multidimensional criteria, niche and expert models can be integrated and fairly assessed alongside general-purpose LLMs. This opens the ecosystem to greater innovation and specialization, benefiting both developers and end users.

4. Discussion

4.1 Interpretation of Results

Our results highlight several implications for both the AI ecosystem and broader societal impact. First, the comparison between monolithic models and routing approaches revealed that intelligent routing such as that by Martian’s router not only improves safety, achieving perfect or near-perfect scores in bias and toxicity, but also delivers more balanced and reliable outputs across all evaluated dimensions. This suggests that routing can meaningfully reduce the risk of AI systems propagating social biases or producing harmful content, a key concern for responsible AI deployment (Salhab et al., 2024).

The multidimensional safety judge itself proved to be highly effective: only **19.6%** of evaluations were flagged as in need of human review, demonstrating that automated, transparent judging can significantly reduce the oversight burden and accelerate safe scaling of AI systems. This efficiency is crucial for real-world applications, where manual review of every output is impractical.

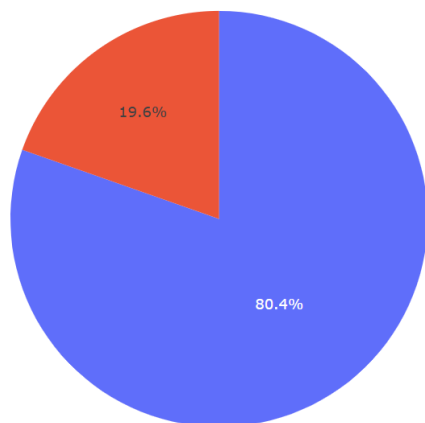


Figure 3 – Radon Chart Comparisons Across Key Metrics

Another notable insight was that consistency emerged as the most challenging metric for all models, with frequent variability in outputs for similar prompts. Models also lagged in embedding similarity to reference responses, indicating that even when factuality was high, the phrasing or framing of answers could diverge

from established safe or canonical responses. These findings have important implications for AI safety: inconsistency and low similarity to vetted references can lead to unpredictable or contextually inappropriate outputs, undermining user trust and increasing the risk of unintended harm. Addressing these gaps is essential for building AI systems that are not only accurate, but also reliable and aligned with societal values.

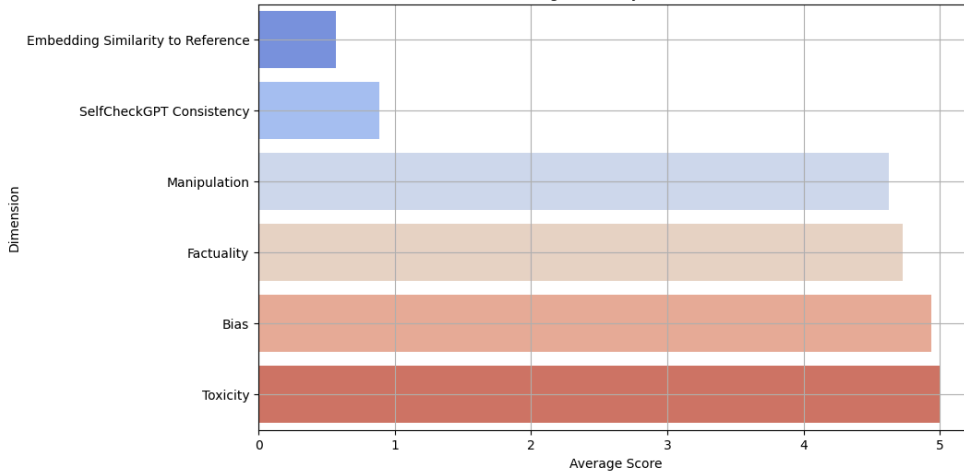


Figure 4 – Average Score by Dimension

4.2 Limitations & Future Work

While our judge model provides interpretable safety evaluations, several limitations remain. First, our approach relies on a modest sample of prompts; expanding the dataset and including more challenging or nuanced cases could improve generalizability. Our current work focuses on text-based LLMs and may require adaptation for multimodal or highly specialized domain models. Scalability is also a concern as evaluating large volumes of responses or integrating additional safety dimensions could introduce unexpected computational bottlenecks.

There is also the lack of detailed cost analysis. While we discuss the potential for routing to optimize resource use, we did not conduct a quantitative assessment of cost savings or trade-offs between models. Additionally, our experiments were not tailored to specific domains (e.g., healthcare, law, education), which may have unique safety and factuality requirements. Future work will involve running domain-specific experiments and conducting comprehensive cost-benefit analyses to better understand the practical implications of the safety judging model in varied real-world contexts.

We also plan to explore automated prompt generation, real-time judging, integration with a broader range of models and routing strategies, and adversarial testing.

4.3 Safety Considerations

Potential risks include the possibility of gaming the judge model through prompt engineering or adversarial attacks. To mitigate this, we designed the judge with transparent, modular criteria and we encourage further auditing and refinement. Our architecture supports traceability and user oversight, helping to detect and address emerging failure modes. By making safety evaluation explicit and interpretable, our work contributes to the development of more accountable and trustworthy AI systems, supporting responsible deployment and reducing the risk of unintended harm.

5. Conclusion

We developed a multidimensional safety judge that transparently evaluates AI outputs on bias, factuality, manipulation resistance, and toxicity. Our findings show that Martian's Router, guided by these safety criteria, outperforms monolithic models in delivering safer and more consistent responses. With under 20% of outputs flagged for human review, the judge supports scalable, reliable AI deployment.

This work advances the Expert Orchestration vision by enhancing transparency, promoting fair model evaluation and improving AI safety. It solidifies groundwork for building trustworthy and ethical AI systems which are aligned with positive societal values.

6. References

Ho, J.Q.H., Hartanto, A., Koh, A. and Majeed, N.M. (2025). Gender Biases within Artificial Intelligence and ChatGPT: Evidence, Sources of Biases and Solutions. *Computers in Human Behavior: Artificial Humans*, [online] 4, p.100145. doi:[10.1016/j.chbah.2025.100145](https://doi.org/10.1016/j.chbah.2025.100145).

Hofmann, V., Kalluri, P.R., Jurafsky, D. and King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, [online] (633), pp.1–8. doi:[10.1038/s41586-024-07856-5](https://doi.org/10.1038/s41586-024-07856-5).

Yulianti, S. (2025). *The Hidden Bias in AI: How Artificial Intelligence Reflects and Reinforces Social Inequalities*. [online] Available at: https://www.researchgate.net/publication/389895420_The_Hidden_Bias_in_AI_How_Artificial_Intelligence_Reflects_and_Reinforces_Social_Inequalities.

Salhab, W., Ameyed, D., Jaafar, F., & Mcheick, H., 2024. A Systematic Literature Review on AI Safety: Identifying Trends, Challenges, and Future Directions. *IEEE Access*, 12, pp. 131762-131784. doi:10.1109/ACCESS.2024.3440647.

Wolfe, R., Yang, Y., Howe, B., & Caliskan, A., 2022. Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. doi:10.1145/3593013.3594072.

Fitch, S. (2024). Evaluating Large Language Models | Center for Security and Emerging Technology. [online] Center for Security and Emerging Technology. Available at: <https://cset.georgetown.edu/article/evaluating-large-language-models/>.

Jakob Mökander, Schuett, J., Hannah Rose Kirk and Floridi, L. (2023). Auditing large language models: a three-layered approach. *AI and Ethics*, [online] 4, pp.1085–1115, doi:10.1007/s43681-023-00289-2.