
Apart x Martian Mechanistic Router Interpretability Hackathon

Author: Joshua Williams

Affiliation: All Tech Is Human; Open Ethics; BABL AI

With
Martian & Apart Research

Abstract

This project addresses the Intelligent Router Systems (Track 2) by providing a People + Planet + Parity (P3) Governance Framework designed to proactively detect, evaluate, and mitigate biases at all stages of AI model integration (before, during, and after deployment). Traditional monolithic AI systems often fail to sufficiently account for accessibility, environmental impact, and racial fairness, leading to disconnected metrics and unclear accountability. This solution integrates specialized judge models—Accessibility Judge, Carbon Judge, and Bias Judge—into a governance-driven framework, systematically ensuring AI outputs meet predefined ethical standards.

Key findings include a structured and transparent method for ethical decision-making, continuous logging, and auditable metrics directly linked to accountability roles. The governance framework significantly enhances mechanistic interpretability by clearly showing why and how routing decisions are made and adjusted in real-time, greatly improving accountability and safety infrastructure.

Keywords: *AI ethics, AI governance, accessibility, racial bias mitigation, carbon emissions tracking, multi-agent alignment, AI safety, Responsible AI, intelligent router systems*

1. Introduction

1.1 Research Question & Motivation

Primary Question: How can AI routing systems proactively uncover and address biases tied to accessibility, climate impact, and racial equity throughout every phase of model deployment—before launch, during operations, and after analysis?

This approach supports the Expert Orchestration Architecture vision by addressing monolithic AI systems' ethical and transparency limitations. Rather than tackling ethical considerations after deployment, the governance method incorporates accountability and bias checks at every stage of model orchestration.

1.2 Challenge Track Focus

Track 2: Intelligent Router Systems

- This approach focuses on embedding ethical routing decisions into intelligent systems by employing specialized “Judges” (Accessibility, Carbon, Bias) to evaluate outputs and inform routing logic, even without direct code.

1.3 Contribution to Expert Orchestration

This solution addresses monolithic AI limitations by breaking down ethical evaluations into specialized judge models, each focused on specific ethical aspects. These models are integrated into an automated routing decision pipeline, directing processes to the appropriate evaluator based on context. This approach enhances transparency, making decision-making more accessible while democratizing AI safety practices for organizations of all sizes. Ultimately, the model is designed to be universally applicable and scalable, promoting responsible and ethical AI use across various industries.

2. Methods

2.1 Technical Approach

- **Models/datasets used:**
 - OpenAI Moderation API (Bias Judge)
 - Flesch readability and Web Content Accessibility Guidelines (WCAG) 2.2 Standards (Accessibility Judge)
 - Lambda Labs GPU and codecarbon tracking (Carbon Judge)
- **Mechanistic interpretability techniques:**
 - Structured Logging: Every routing decision includes metrics and rationale logged transparently.
 - Continuous Testing: Pre-deployment bias probes (e.g., race-coded, accessibility probes).
- **Implementation details:**
 - Python, OpenAI API, Lambda Labs GPU, codecarbon, textstat
- **Experimental setup:**
 - Accessibility, bias, and carbon judges systematically evaluate model outputs.

- Router chooses specialized AI models based on judge evaluations.
- Continuous logging into CSV for transparency and auditability.

2.2 Routing/Judge Architecture

- **Accessibility Judge:** Evaluates readability and alt-text.
- **Carbon Judge:** Measures and routes based on per-query emissions.
- **Bias Judge:** Detects harmful stereotypes and disparities.
- **Decision Transparency:** Routing decisions logged with full interpretability (timestamp, judge scores, and routing rationale).

2.3 Governance Framework Reproducibility

While this project doesn't provide executable code, the goal is to achieve complete reproducibility through clearly documented governance templates and standards:

- Governance Framework Document that includes:
 - Guiding Principles
 - Roles & Responsibilities Matrix
 - Metrics & Threshold Logging
 - Incident Response Protocol
 - Documentation & Logging Standards
 - Adoption & Implementation Plan

3. Results



Incident Response Protocol Continued: This matrix outlines how teams should respond when thresholds for accessibility, climate impact, or fairness are breached. It supports ethical accountability, clarity, and escalation at each step.

Trigger Event	Detected By	Immediate Action	Responsible Role	Escalation Criteria	Documentation Required
Multiple Judge Failures (≥ 2 fail on same query)	Router or governance audit log	Suspend query route; trigger ethics review team to convene	Responsible AI & Governance Lead	Any dual failure on the same output (e.g., bias + carbon).	Full incident report: list all judges, scores, timestamps, and team decision.
Threshold Change	Manual or post-incident	Open Threshold Update Form; send to AI Governance Lead	Requestor + Responsible AI & Governance Lead	All requests must include justification + new target	Update router_config.yaml with new values, rationale, and approver signature.
Unresolved or Repeated Violations	Any recurring flagged pattern	Whole Team Review	Responsible AI & Governance Lead	≥3 same category violations in a 30-day window	Incident series summary + audit trail

Figure 1 – Representation of Incident Response Protocol

This governance-focused project demonstrates robust methods for proactively identifying, documenting, and addressing biases related to accessibility, environmental impact, and racial fairness throughout the AI lifecycle stages. We hope to achieve the following:

- **Unified Ethical Evaluation:** Accessibility, Carbon, and Bias metrics integrated into a cohesive governance structure, systematically preventing fragmented ethics.
- **Proactive Bias Management:** Bias risks explicitly mapped to OWASP LLM standards, ensuring immediate visibility and structured response capabilities.

- **Comprehensive Incident & Decision Logging:** Ethics logging mechanism that captures every significant decision, facilitating clear accountability and audit trails.

3.1 Routing Decision Governance Analysis

The governance approach explicitly defines transparent criteria for evaluating and interpreting AI routing decisions:

- **Accessibility Criteria:** Clearly defined WCAG 2.2 AA standards are a gating factor for deployment approvals.
- **Carbon Impact Criteria:** SCI-aligned carbon emission thresholds ($\leq 0.05\text{g CO}_2/\text{query}$) embedded into deployment checklists and monitored continuously.
- **Bias & Fairness Criteria:** Established OWASP GenAI Security Project standards and Office of the High Commissioner for Human Rights (OHCHR) anti-racism guidelines form the structured basis for routing decision evaluation and escalations pathways.

3.2 Expert Orchestration Impact

The governance framework significantly advances the Expert Orchestration vision by:

- **Enhancing Transparency & Accountability:** Defines distinct roles, promotes transparent decision-making, and implements robust logging practices that greatly exceed those of monolithic AI structures.
- **Empowering AI Safety:** Defines accountability and incident response clearly, ensuring safer deployments and organized mechanisms for responding to bias incidents.
- **Democratizing Responsible AI:** Offers practical and easy-to-adopt governance resources for teams of all skill levels to effectively implement responsible AI, thereby improving the adoption rate of safer and fairer AI practices.

4. Discussion

4.1 Interpretation of Results

The Governance Framework effectively addresses previously overlooked areas such as accessibility, climate impact, and fairness through a continuous ethical evaluation process. This approach ensures comprehensive inclusivity and allows teams to understand the mechanisms behind routing decisions. The framework enhances accountability and fosters trust by clarifying why specific outputs are approved or blocked.

4.2 Limitations & Future Work

- **Technical limitations:** current implementation leverages third-party APIs; future versions may include custom models.
- **Scalability:** initial tests were small-scale; future work involves scaling to enterprise-level deployment.
- **Improvement Areas:** expanding the Accessibility Judge to cover additional disabilities explicitly.

4.3 Safety Considerations

- Built-in safeguards ensure flagged content cannot reach end-users unchecked.
- Continuous logging and alerts on threshold breaches.
- Safety risks minimized through automated fallback mechanisms.

5. Conclusion

Our innovative framework significantly contributes to the mechanistic interpretability of intelligent routing systems, advancing the concept of Expert Orchestration Architecture. By offering a comprehensive and practical approach, it addresses critical elements such as accessibility for underserved populations, environmental sustainability to mitigate ecological impacts, and promoting racial fairness to ensure equitable outcomes.

A key feature of our framework is its focus on resolving the prevalent issues of fragmented and late-stage ethical integration in AI systems. By embedding ethical considerations early in the development process, we create a foundation for more transparent accountability and enhanced transparency in decision-making processes. This proactive stance fosters an environment conducive to the safe deployment of AI technologies, ultimately enhancing user trust and societal acceptance. Through these advancements, our framework improves the performance of intelligent routing systems and aligns them with ethical standards and societal values.

• References

- **Partnership on AI.** (2025). *Guidance for Inclusive AI*. <https://partnershiponai.org/guidance-for-inclusive-ai/>
- **World Wide Web Consortium (W3C).** (2023, October 5). *Web Content Accessibility Guidelines (WCAG 2.2 W3C Recommendation)*. <https://www.w3.org/TR/WCAG22/>
- **OWASP Foundation.** (2025). *OWASP Top 10 for Large Language Models (LLM) applications*. <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/>
- **Office of the High Commissioner for Human Rights.** (2024, July 3). *Racism and AI: Bias in the past leads to bias in the future*. United Nations. <https://www.ohchr.org/en/racism>
- **Open Ethics Initiative.** (2024). *Open Ethics Label*. <https://openethics.ai/label/generate/>
- **Green Software Foundation.** (2023). *Software Carbon Intensity (SCI) specification v1.0*. <https://sci.greensoftware.foundation/>
- **United Nations.** (n.d.) *Goal 13: Take urgent action to combat climate change and its impacts*. United Nations Sustainable Development Goals. <https://sdgs.un.org/goals/goal13>

• Appendix

- <https://drive.google.com/file/d/1F4eUgkqFT7zZDKhcWCDdzfdXWNpcQS-6/view?usp=sharing>