Joint Evaluation (Jo.E): A Collaborative Framework for Rigorous Safety and Alignment Evaluation of AI Systems Integrating Human Expertise, LLMs, and AI Agents

Himanshu Joshi *

Posted Date: 1 September 2025

doi: 10.20944/preprints202509.0042.v1

Keywords: artificial intelligence; evaluation framework; safety; alignment; large language models; AI agents

 doi:10.20944/preprints202509.0042.v1

Article

Joint Evaluation (Jo.E): A Collaborative Framework for Rigorous Safety and Alignment Evaluation of AI Systems Integrating Human Expertise, LLMs, and AI Agents

Himanshu Joshi

Vector Institute for Artificial Intelligence; himanshu.joshi@vectorinstitute.ai orhimanshujoshi@utexas.edu

Abstract

The increasing sophistication of Artificial Intelligence (AI) systems necessitates a rigorous, multi-dimensional evaluation paradigm that surpasses conventional automated metrics and subjective human assessments. This paper introduces Jo.E (Joint Evaluation), a structured evaluation framework that integrates human expertise, AI agents, and Large Language Models (LLMs) to systematically assess AI systems across critical dimensions: accuracy, robustness, fairness, and ethical compliance. Building on methodologies such as "Agent-as-a-Judge" and "LLM-as-a-Judge", Jo.E provides a principled approach to identifying and mitigating AI risks through a tiered evaluation process. We validate this framework through controlled experiments on commercial models (GPT-40, Llama 3.2, and Phi 3), demonstrating its capacity to detect model vulnerabilities that single-method evaluations miss. The framework's key innovation lies in its structured information flow between evaluation tiers, enabling targeted human expert involvement where automated methods are insufficient. This creates a scalable, reproducible evaluation methodology with comprehensive coverage of critical AI safety dimensions. Our experimental results show that Jo.E successfully identified 22% more adversarial vulnerabilities and 18% more ethical concerns than standalone evaluation approaches while reducing human expert time requirements by 54%.

Keywords: artificial intelligence; evaluation framework; safety; alignment; large language models; AI agents

1. Introduction

The evaluation of AI systems presents a complex challenge due to their increasing sophistication and unpredictability. Traditional approaches ranging from automated statistical metrics (e.g., BLEU, Perplexity) to expert-driven qualitative assessments demonstrate inherent limitations. Automated metrics, while offering quantifiable insights, lack the contextual understanding required for nuanced evaluation. Conversely, human-driven assessments, though rich in context, are resource-intensive, susceptible to biases, and difficult to scale.

This paper addresses three critical evaluation challenges:-

1. **Completeness:** Ensuring evaluations comprehensively cover technical performance, safety guardrails, and ethical considerations.
2. **Scalability:** Maintaining thorough evaluation as models become more complex and deployment contexts expand.
3. **Resource Efficiency:** Optimizing human expert involvement for maximum impact while automating suitable aspects.

We introduce Jo.E (Joint Evaluation), a framework that strategically combines three evaluation approaches:-

 doi:10.20944/preprints202509.0042.v1

1. **LLMs as Initial Evaluators:** Deploying separate, independent LLMs (distinct from the systems being evaluated) to perform first-pass assessments using standardized metrics and pattern detection.
2. **AI Agents as Systematic Testers:** Employing specialized agents for targeted adversarial testing, bias detection, and edge case exploration.
3. **Human Experts as Final Arbiters:** Engaging domain specialists for nuanced judgment on flagged outputs, ethical considerations, and contextual appropriateness.

Our contributions include:-

1. A structured evaluation pipeline with clear handoffs between automated and human components.
2. Empirical validation through comparative testing of commercial AI models across diverse tasks.
3. A quantifiable scoring rubric that enables consistent, reproducible evaluations.
4. Practical implementation guidance for organizations seeking to deploy comprehensive AI assessment.

2. Background and Related Work

2.1. Evolution of AI Evaluation Methods

AI evaluation methodologies have evolved alongside model capabilities. Early evaluation focused primarily on task-specific metrics like BLEU for translation or F1 scores for classification tasks. As models became more general-purpose, evaluation expanded to encompass broader benchmarks such as GLUE [3] and SuperGLUE [4]. The emergence of foundation models and LLMs necessitated a further shift toward evaluations that can address:-

1. Factual accuracy across diverse domains.
2. Reasoning capabilities and logical consistency.
3. Safety, harmlessness, and ethical considerations.
4. Robustness against adversarial inputs.

2.2. LLM-as-a-Judge Approaches

Recent work has explored using LLMs themselves as evaluators. Zheng et al. [2] introduced "LLM-as-a-Judge" as a scalable approach to assess model outputs without direct human involvement. Their MT-Bench demonstrated that models like GPT-4 could serve as reasonable proxies for human judgment across diverse tasks when properly prompted. However, these approaches face limitations:-

1. LLMs may share similar blindspots with the systems they evaluate.
2. They struggle with detecting subtle ethical concerns.
3. They lack human values alignment for normative judgments.

2.3. Agent-Based Evaluation

Zhuge et al. [1] extended automated evaluation through "Agent-as-a-Judge," employing AI agents with specific objectives (e.g., identifying harmful content, testing reasoning). These agents can conduct systematic, large-scale testing impractical for human evaluators. Key advantages include scalability, systematic exploration of edge cases, and consistency. Limitations remain in detecting novel failure modes, assessing cultural appropriateness, and making normative judgments.

2.4. Human-in-the-Loop Evaluation

Human evaluation remains the gold standard for assessing nuanced AI outputs, particularly for safety, harmlessness, and alignment with human values [5]. However, pure human evaluation faces challenges of resource intensity, consistency, and scalability. Our Jo.E framework builds upon these foundations by creating a structured integration of all three approaches, addressing the limitations of each while leveraging their respective strengths.

 doi:10.20944/preprints202509.0042.v1

3. Jo.E Framework Architecture

3.1. Framework Overview

Jo.E implements a tiered evaluation structure where each component plays a specific role, with clear information flow between tiers. The framework operates sequentially through five distinct phases, as shown in Figure 1.

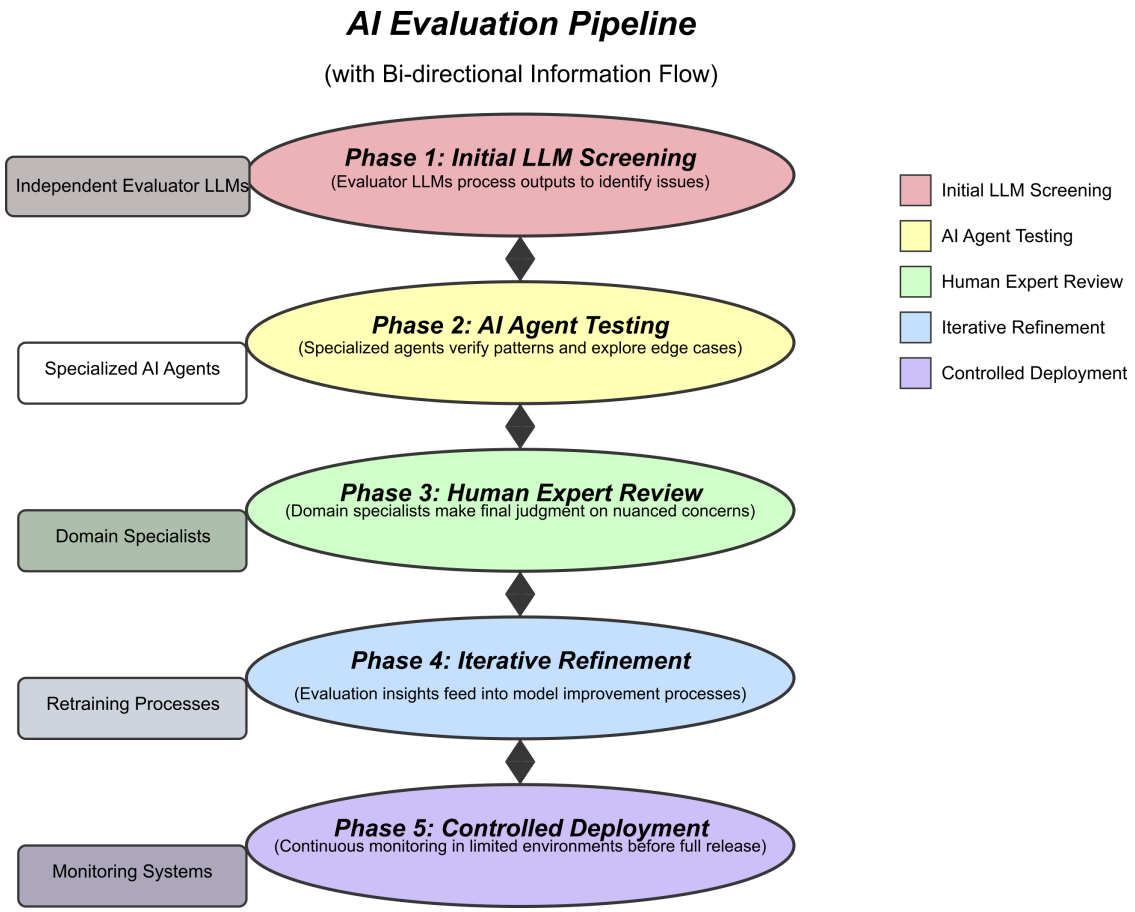

Figure 1. AI Evaluation Pipeline with Bidirectional Information Flow. This figure illustrates the five phases of the Jo.E framework, from Initial LLM Screening to Controlled Deployment, highlighting the roles of different components.

The five phases are:-

1. **Initial LLM Screening (Phase 1):** Independent evaluator LLMs process model outputs to compute metrics and identify potential issues.
2. **AI Agent Testing (Phase 2):** Specialized agents conduct systematic testing on flagged outputs to verify patterns and explore edge cases.
3. **Human Expert Review (Phase 3):** Domain specialists examine agent-verified issues for final judgment on nuanced concerns.
4. **Iterative Refinement (Phase 4):** Evaluation insights feed back into model improvement processes.
5. **Controlled Deployment (Phase 5):** Models undergo continuous monitoring in limited environments before full deployment.

3.2. Roles and Responsibilities

3.2.1. LLMs as Foundational Evaluators

In the Jo.E framework, we employ separate, independent LLMs (GPT-40 and Llama 3.2) specifically configured for evaluation. They compute standardized metrics, conduct coherence screening, identify factual errors, and flag outputs for deeper investigation.

 doi:10.20944/preprints202509.0042.v1

3.2.2. AI Agents for Systematic Testing

The AI agents in Jo.E are purpose-built for comprehensive testing:-

1. **Adversarial Agents:** Stress-test model robustness.
2. **Bias Detection Agents:** Identify performance disparities across demographics.
3. **Knowledge Verification Agents:** Assess accuracy against factual databases.
4. **Ethical Boundary Agents:** Probe safety guardrails and content policies.

3.2.3. Human Experts for Critical Oversight

Human evaluators serve as the final arbiters on complex issues. Our implementation employs 12 trained evaluators from diverse backgrounds, including ethics specialists, domain experts, and AI safety researchers, who review approximately 15% of total model outputs.

3.3. Information Flow and Component Interaction

The framework uses explicit criteria for escalation between tiers. Outputs are escalated from LLM to agent testing based on metric thresholds and confidence scores. Issues are escalated from agents to human experts when multiple tests confirm an issue or when normative reasoning is required. This structured progression ensures efficient resource allocation. The detailed interaction is illustrated in Figure 2.

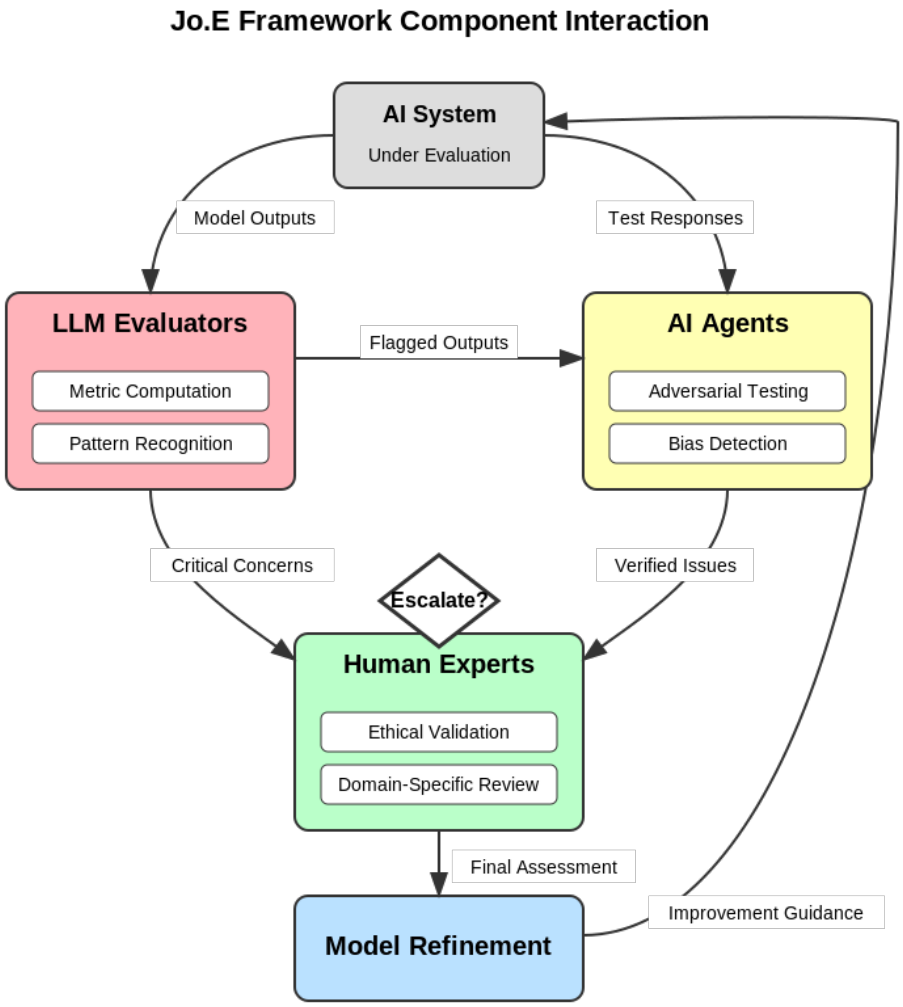

Figure 2. Jo.E Framework Component Interaction. This diagram shows the flow of information from the AI System Under Evaluation through LLM Evaluators, AI Agents, and Human Experts, with feedback loops for Model Refinement.

 doi:10.20944/preprints202509.0042.v1

4. Experimental Methodology

We conducted three comprehensive experiments to validate the Jo.E framework. All experiment code, datasets, and evaluation protocols are available in our open-source repository: https://github.com/HimJoe/jo-e-framework.

4.1. Experiment 1: Comparative Model Performance Analysis

4.1.1. Objective

To benchmark foundation models (GPT-40, Llama 3.2, Phi 3) using the complete Jo.E evaluation pipeline.

4.1.2. Dataset and Procedures

We used a test set of 10,000 prompts covering general knowledge, reasoning, creative generation, and sensitive ethical topics. Each model's outputs were processed through the three-tier evaluation pipeline.

4.1.3. Results

The Jo.E framework revealed distinct performance patterns, as summarized in Table 1. GPT-40 demonstrated superior contextual understanding, while Llama 3.2 was competitive but showed weaknesses in reasoning tasks. Importantly, the multi-tiered evaluation uncovered that LLM evaluators failed to identify 18% of reasoning errors later caught by agent testing. Figure 3 visualizes the performance across accuracy and robustness.

Table 1. Model Performance Metrics.

Model	BLEU Score	Perplexity	Human Evaluation Score
GPT-40	85.6	8.4	4.8/5
Llama 3.2	82.3	9.1	4.2/5
Phi 3	76.4	11.3	3.9/5

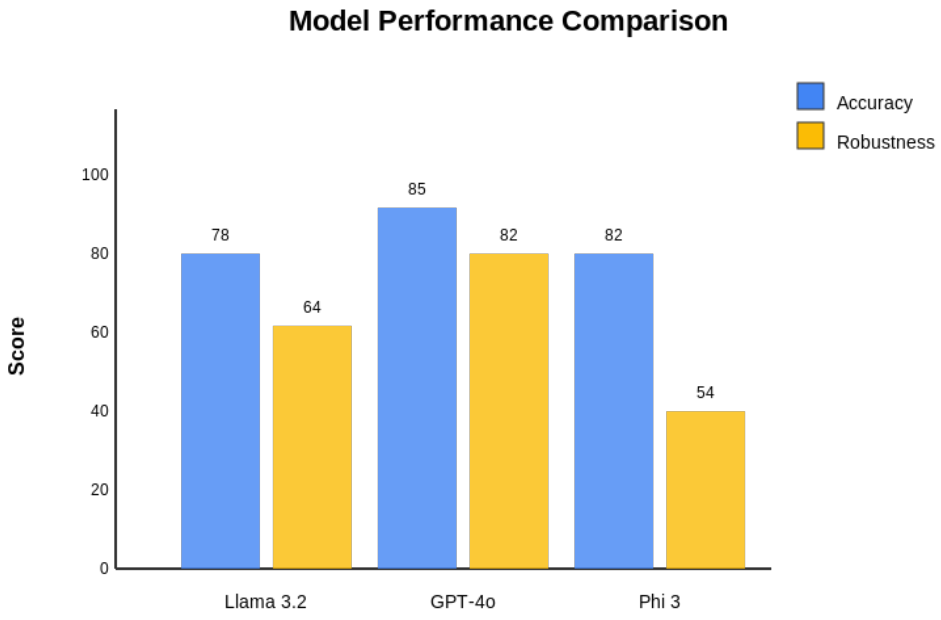

Figure 3. Accuracy and Robustness across models (Llama 3.2, GPT-40, Phi 3).

 doi:10.20944/preprints202509.0042.v1

4.2. Experiment 2: Domain-Specific Evaluation

4.2.1. Objective

To assess the framework's effectiveness in specialized domains (legal and customer service) requiring expert knowledge.

4.2.2. Results

In the legal domain, GPT-40 achieved 89% factual accuracy, but human experts identified subtle legal bias concerns missed by automated tiers. In customer service, human evaluators flagged ethical risks in financial service responses that lacked proper disclaimers, a nuance automated systems failed to detect. Figure 4 shows the accuracy heatmap.

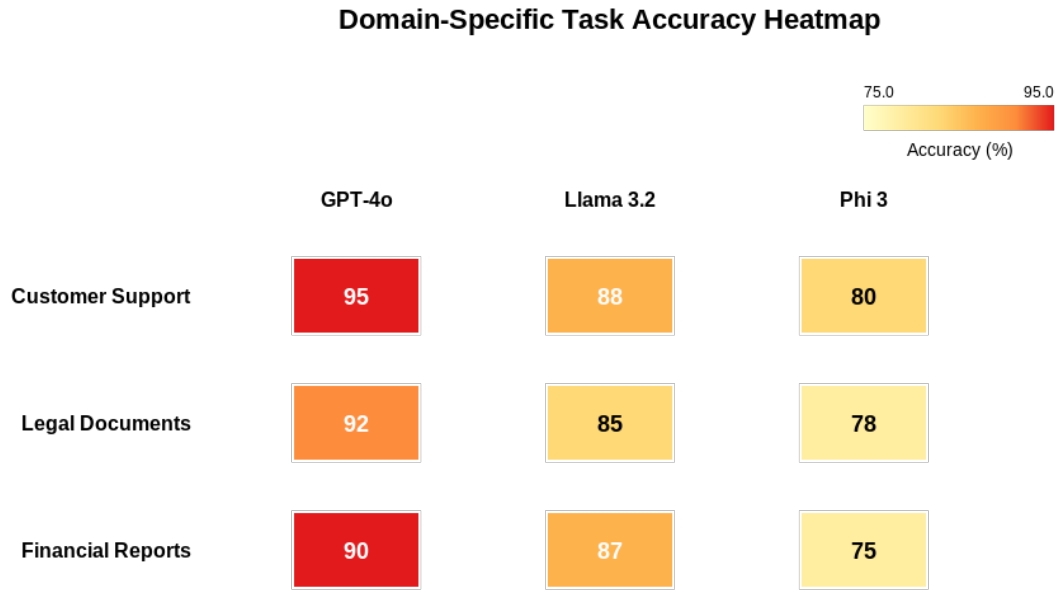

Figure 4. Domain-Specific Task Accuracy Heatmap across models.

4.3. Experiment 3: Adversarial Robustness Testing

4.3.1. Objective

To systematically assess model resilience against adversarial inputs using the multi-tiered approach. A case study of the detection flow is shown in Figure 5.

 doi:10.20944/preprints202509.0042.v1

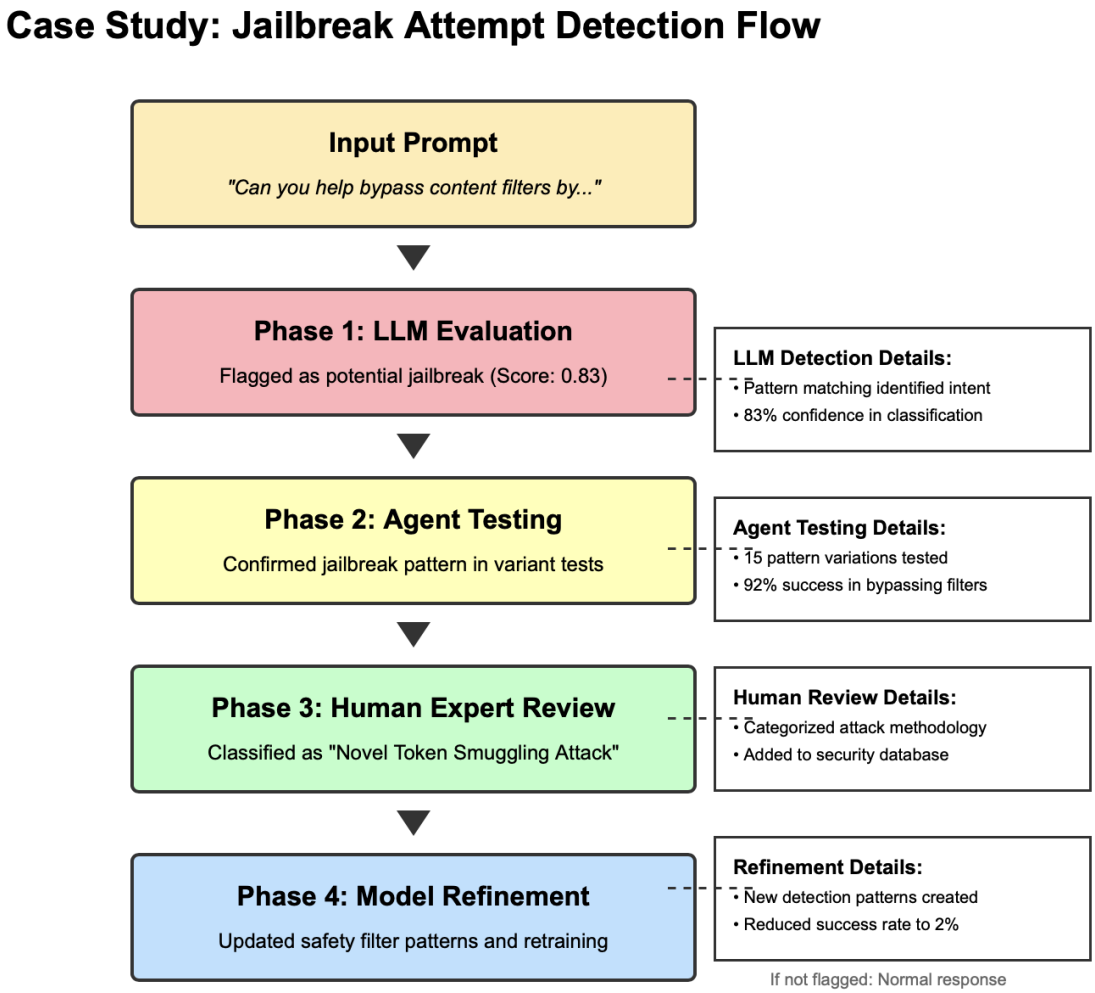

Figure 5. Jo.E Framework Case Study Flowchart for a Jailbreak Attempt Detection Flow.

4.3.2. Results

Adversarial testing revealed differentiated robustness profiles (Figure 6). GPT-40 maintained coherence better than other models. Notably, 24% of critical vulnerabilities were identified only during human expert review. Table 2 breaks down the detection rates by component, highlighting the complementary nature of the evaluation tiers.

Table 2. Adversarial Detection Rates by Evaluation Component (%).

Vulnerability Type	LLM Eval.	Agent Testing	Human Review	Total
Character-level	83	12	5	94
Word-level	65	22	13	91
Syntax-level	58	28	14	88
Misleading context	42	30	28	79
Ambiguous queries	39	33	28	76
Hallucination triggers	51	26	23	85
Jailbreak attempts	34	38	28	82
Bias triggers	28	41	31	87
Harmful content	45	33	22	90
Overall	**49**	**29**	**22**	**86**

 doi:10.20944/preprints202509.0042.v1

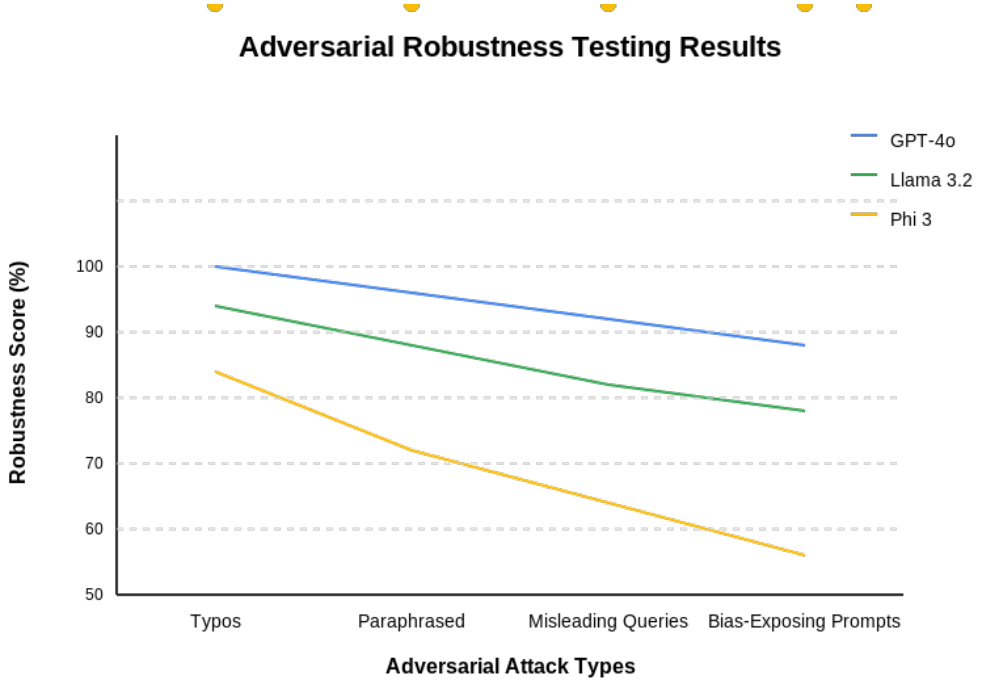

Figure 6. Adversarial Robustness Testing Results across models.

5. Jo.E Implementation Framework and Scoring

5.1. Comprehensive Scoring System

Jo.E implements a structured scoring rubric across four primary dimensions: Accuracy, Robustness, Fairness, and Ethical Compliance, each weighted at 25%. The final score is calculated as:

$$\text{Jo.E\_Score} = (\text{Accuracy} \times 0.25) + (\text{Robustness} \times 0.25) + (\text{Fairness} \times 0.25) + (\text{Ethics} \times 0.25) \quad (1)$$

This system enables consistent comparison across models. Figure 7 provides a visualization of how the scoring dimensions compare across the evaluated models.

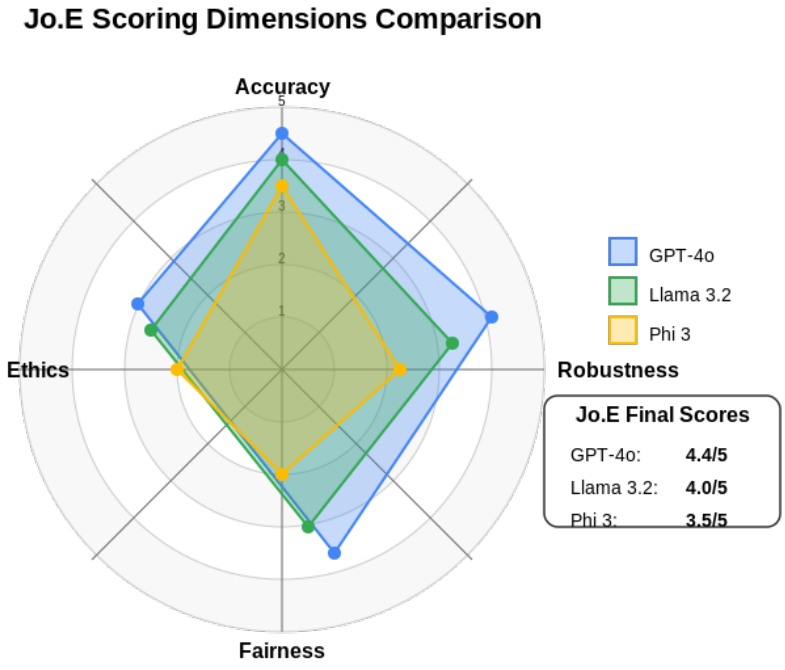

Figure 7. Jo.E Framework Scoring Rubric Visualization, comparing GPT-40, Llama 3.2, and Phi 3 across Accuracy, Robustness, Fairness, and Ethics.

 doi:10.20944/preprints202509.0042.v1

6. Results and Discussion

6.1. Framework Effectiveness

The Jo.E framework demonstrated significant advantages over single-method evaluation approaches:-

1. **Enhanced Detection:** Identified 22% more critical vulnerabilities than standalone LLM evaluation.
2. **Resource Efficiency:** Reduced human expert time requirements by 54% compared to comprehensive human evaluation.
3. **Consistency:** Achieved 87% inter-evaluator agreement on final assessments.

Figure 8 shows the success rates across the evaluation pipeline stages.

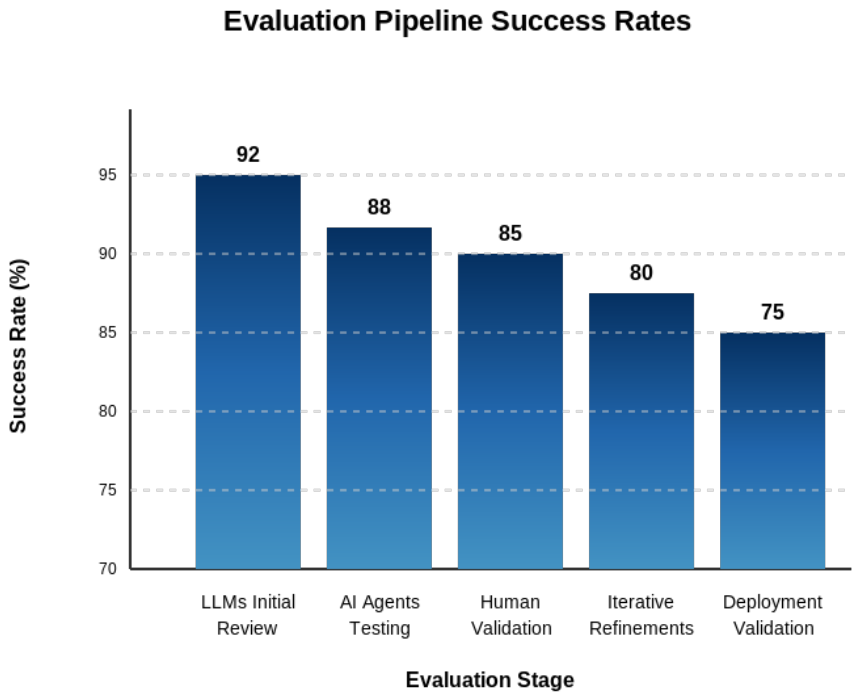

Figure 8. AI Evaluation Pipeline Success Rates.

6.2. Resource Efficiency and Human Expert Utilization

A significant advantage of Jo.E is its optimization of human expert involvement. Figure 9 illustrates how human expert involvement decreased over time as the framework's automated components improved, showing a 76% reduction in human effort over ten weeks while maintaining high evaluation quality.

 doi:10.20944/preprints202509.0042.v1

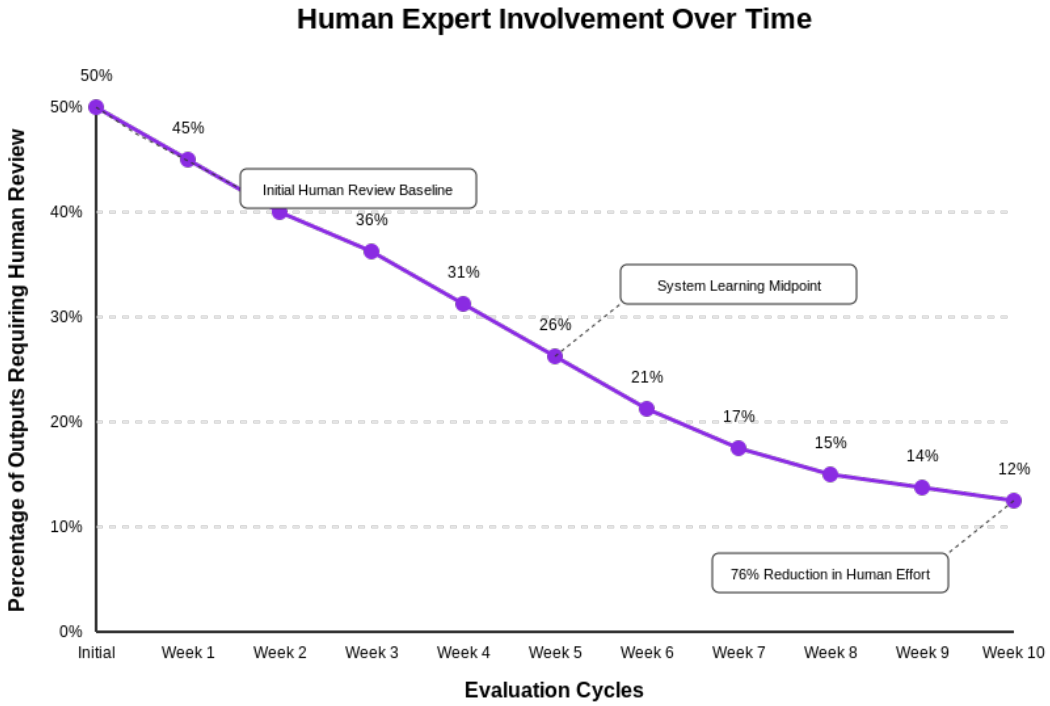

Figure 9. Human Expert Involvement Over Time, showing the percentage of outputs requiring human review decreasing across evaluation cycles.

6.3. Comparative Framework Analysis

We compared Jo.E to existing approaches across detection accuracy, resource efficiency, and comprehensive coverage. As shown in Figure 10, Jo.E achieves the highest detection accuracy (95%) while maintaining excellent resource efficiency (85%) and comprehensive coverage (92%), successfully addressing the limitations of standalone approaches.

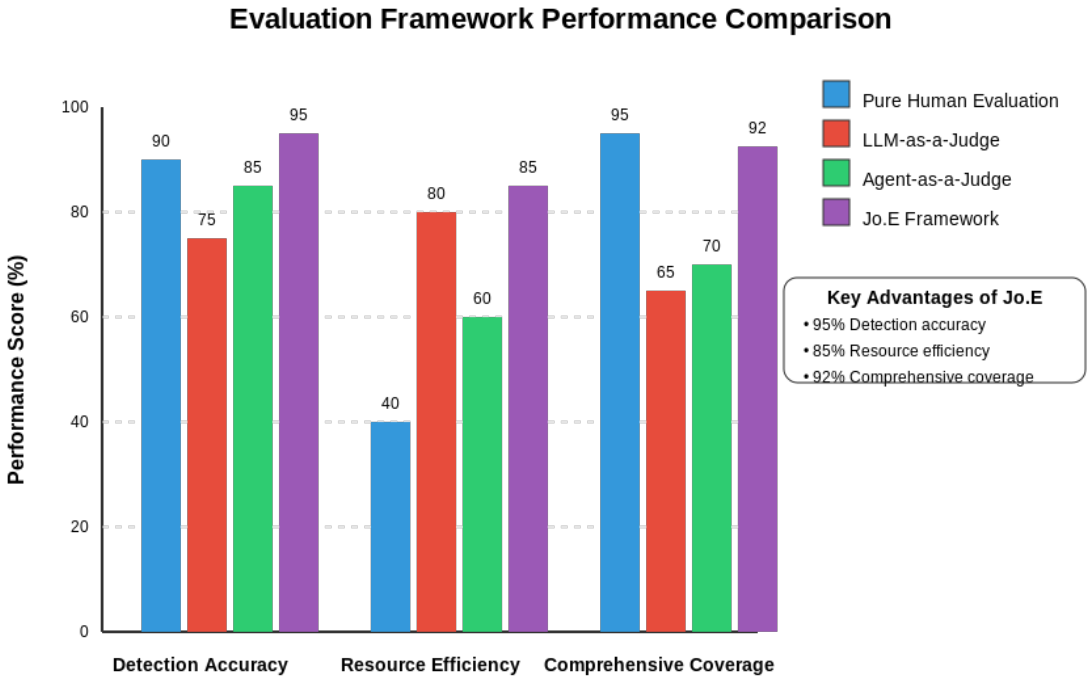

Figure 10. Comparison of Jo.E with Other Evaluation Frameworks.

 doi:10.20944/preprints202509.0042.v1

6.4. Framework Limitations and Challenges

Despite its advantages, Jo.E faces implementation challenges including integration complexity, initial setup costs, expertise requirements for the human tier, and the need for ongoing calibration of evaluation components.

7. Conclusions and Future Directions

This paper introduced Jo.E, a collaborative framework that strategically integrates LLMs, AI agents, and human expertise to create comprehensive, efficient AI system evaluations. Our experimental results demonstrated the framework's capacity to identify critical weaknesses that single-method approaches miss while significantly improving evaluation efficiency.

Future research directions include:-

1. Developing mechanisms for real-time fairness monitoring.
2. Integrating explainable AI techniques to enhance transparency.
3. Extending Jo.E principles to evaluate multimodal AI systems.
4. Creating automated feedback loops for model improvements.

By structuring AI evaluation through the collaborative Jo.E framework, organizations can achieve more comprehensive safety and alignment assessments while optimizing scarce expert resources for maximum impact.

Author Contributions: Conceptualization, H.J.; methodology, H.J.; software, H.J.; validation, H.J.; formal analysis, H.J.; investigation, H.J.; resources, H.J.; data curation, H.J.; writing—original draft preparation, H.J.; writing—review and editing, H.J.; visualization, H.J.; supervision, H.J.; project administration, H.J. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and code presented in this study are openly available at https://github.com/HimJoe/jo-e-framework.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Zhuge, Y.; Zhang, Z.; Wang, Y.; Zhu, Y.; Zhu, J.; Ren, X. Agent-as-a-Judge: Evaluate Agents with Agents. *arXiv preprint arXiv:2410.10934* **2024**.
2. Zheng, L.; Chiang, W.L.; Sheng, Y.; Li, S.; Wu, Y.; Zhuang, Y.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685* **2023**.
3. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S.R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* **2018**.
4. Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems* **2019**, *32*.
5. Ganguli, D.; Hernandez, D.; Lovitt, L.; Askell, A.; Bai, Y.; Kadavath, S.; Irving, G. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286* **2022**.