Reliability Judge: Enhancing LLM Reliability Through Multi-Model Judging

Guido Ernesto Bergman
Universidad de Buenos Aires

Tobias Bersia
Universidad de Buenos Aires

Emanuel Pablo Ruzak
Universidad de Buenos Aires

Gonzalo Agustin Heredia
Universidad de Buenos Aires

Steven Molina
Universidad de Buenos Aires

With
BAISH & Apart Research

Abstract

We present a multi-model judging framework that extends the Think Twice before Trusting (T^3) paradigm to enhance answer reliability across large language models (LLMs). Our approach is most suitable for high-stakes contexts (e.g., healthcare, legal, or safety-critical systems), where even minor accuracy gains can have significant consequences. The proposed system uses cross-model evaluation to identify the most accurate and well-justified response from a set of candidates. Each model acts both as an answer generator and as a judge under two modes: as a neutral evaluator or as a participant evaluating its own response. Judgments consider factual correctness, reasoning clarity, and confidence-justification alignment. Our results show that the best judge-based approach outperforms all the monolithic (i.e., single-model) versions. This work contributes a transparent method for selecting high-confidence outputs from black-box models, advancing practical AI reliability and alignment. The code and data are available at https://github.com/GuidoBergman/reliability-judge.

1. Introduction

1.1 Research Question & Motivation

The central research question is: Can a judge-based system consistently achieve higher reliability than a single model? In high-stakes contexts demanding factual accuracy, monolithic models often prove unreliable, sometimes providing confidently wrong answers instead of admitting ignorance. Our project delves into the idea of filtering confidently wrong answers via a two-stage approach: first, we use the Think Twice before Trusting (T3) method to evaluate confidence in the answers generated by a model pool; then, a judge assesses the reliability of those answers and outputs a confidence score, allowing users to set their own threshold for accepting or discarding responses. We hypothesize that this approach will reduce confidently wrong answers compared to monolithic systems while maintaining accuracy on correct answers. This work aims to address the limitations of monolithic AI by enhancing reliability and factual accuracy in critical applications.

1.2 Challenge Track Focus

Our project addresses Track 1: Judge Model Development.

1.3 Contribution to Multiagent Orchestration

This work provides an alternative to monolithic approaches by treating LLM outputs as evidence to be judged.

2. Methods

2.1 Technical Approach

Models & Dataset

- **Evaluated**: Claude 3.5 Haiku (Anthropic), GPT-4.1 Nano (OpenAI), Gemini 2.0 Flash (Google)
- **Dataset**: 600 questions randomly selected from the test split of MMLU Pro.

Implementation

- **Language**: Python 3.9+
- **Libraries**: Martian API, Pandas, Scikit-learn (ROC-AUC), Matplotlib/Seaborn

Method used

For each multiple-choice question, we first measure the performance of each model separately (Monolithic Baseline), using the *Think Twice before Trusting (T^3)* method:

1. **Select Two Candidate answers**: The model picks the two most likely correct options, justifying only those based on reasoning and evidence.
2. **Pick One Final Answer**: From the two selected, the model chooses the single best answer, explains why it's stronger, and gives a confidence score.

Additionally, this Monolithic Baseline is compared with two judge-based approaches:

- **Judge as participant**: a judge evaluates all answers, including one from an instance of itself.
- **Judge as evaluator-only**: a judge evaluates only others' answers. This variant aims to mitigate the risks of having a self-preferring judge.

2.2 Judging Architecture

Selection: The selection process involves choosing an answer if it exceeds a user-defined confidence threshold; otherwise, the query is rejected to maintain a high level of reliability.

2.3 Reproducibility

To ensure reproducibility, we provide a GitHub repository that allows for independent replication of all reported results.

3. Results

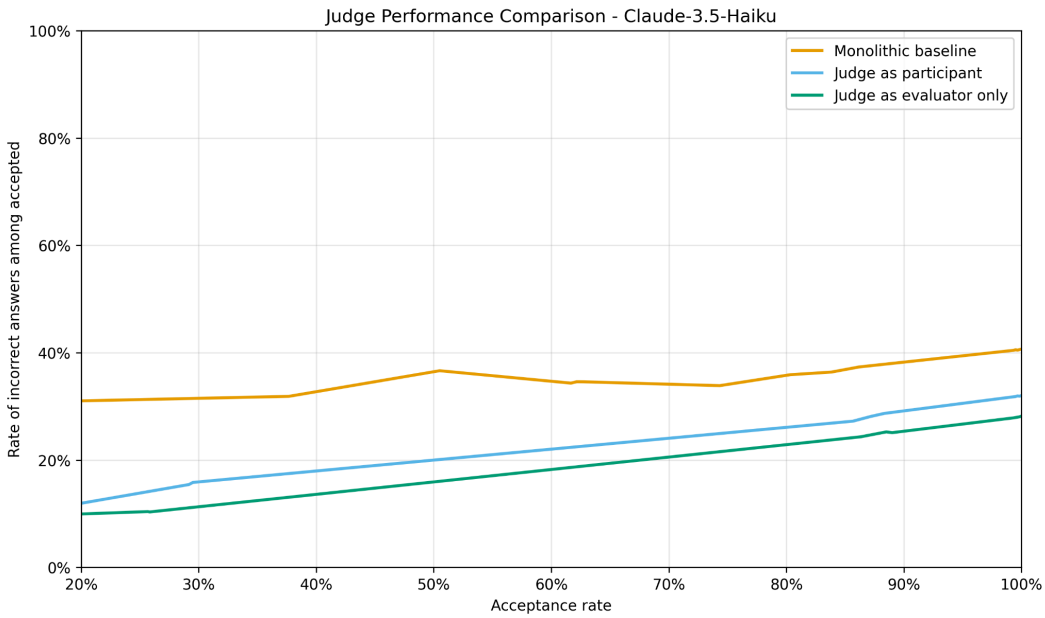

Figure 1 – Performance comparison of Claude across different roles

Figure 1 shows that across all levels of acceptance rate (i.e., the percentage of answers accepted) the judge-based systems that use Claude 3.5 Haiku as a judge outperform the monolithic version. Additionally, the version of the system that uses Claude only as evaluator slightly outperformed the version that uses it as a participant as well. This suggests that the model tends to favor its own answers, even in the cases when they are wrong.

Additionally, this version of the system has a statistically significant improvement over the best monolithic version:

- Claude-3.5-Haiku judge as evaluator only accuracy: 71.833% (95% confidence interval: [68.049%, 75.401%])
- Gemini-2.0-Flash monolithic baseline accuracy: 66.500% (95% confidence interval: [62.566%, 70.271%])

However, this trend was not observed across all models. For example, Figure 3 shows that Gemini's monolithic baseline outperforms the systems that use it as a judge.

3.1 Expert Orchestration Impact

Our project explores the possible viability of Orchestration as a superior alternative to monolithic AI systems in high-stakes settings. The AI Judge system enables:

- **Enhanced control**, as judging decisions are auditable and interpretable; supporting user-defined risk tolerance.
- **Democratization of model contribution**, allowing different models to participate and have the answers selected based on the best justifications.

4. Discussion

4.1 Interpretation of Results

Our results demonstrate that this approach effectively enhances reliability compared to the best individual model in the system through two key mechanisms: leveraging correct answers from multiple models (thereby increasing accuracy) and more effectively filtering overconfident incorrect responses, as individual models exhibit inherent bias toward rating their own answers as correct. In some cases, the performance improvement is further amplified when the judge model is excluded from the candidate pool, suggesting that judges demonstrate self-selection bias even when their responses are incorrect.

Implications for AI Safety and Alignment

- **Critical Gains in High-Stakes Applications**: The observed performance improvement can be critical in high-stakes contexts (e.g., healthcare, legal, or safety-critical systems) where even minor accuracy gains can have significant consequences.
- **Enhancing Trust Through Confidence Calibration:** since the system provides a confidence of the generated answer, users are able to choose "no answer" over unreliable outputs. This approach allows them to filter answers using a tunable threshold, to ensure the reliability of the responses.

4.2 Limitations & Future Work

The main limitation of this work is that it only considers multiple-choice questions. Future research could address this by exploring open-ended questions to broaden the applicability of the approach.

On the other hand, more sophisticated orchestration protocols could be leveraged to balance performance with cost.

4.3 Safety Considerations

AI Safety Contribution: This approach enhances AI safety by promoting verifiable reasoning over blind trust, reducing hallucinations through validation.

5. Conclusion

This work demonstrates the effectiveness of a multi-model, judge-based framework in enhancing the reliability of language model outputs, particularly in high-stakes tasks where factual accuracy is paramount. By extending the Think Twice before Trusting (T3) paradigm with cross-model evaluation, we show that it is possible to reduce wrong answers and improve the reliability of the system. Our results indicate that orchestration strategies—where models critically evaluate each other—can outperform monolithic approaches. Future work will extend this approach to open-ended formats and explore more sophisticated orchestration protocols.

6. References

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, Tat-Seng Chua. (2024). *Think Twice Before Trusting: Self-Detection for Large Language Models through Comprehensive Answer Reflection* https://aclanthology.org/2024.findings-emnlp.693/

7. Appendix

7.1 Accuracy of all the versions

- Claude-3.5-Haiku monolithic baseline accuracy: 59.333% (95% confidence interval: [55.282%, 63.293%])
- Claude-3.5-Haiku judge as participant accuracy: 68.000% (95% confidence interval: [64.102%, 71.719%])
- Claude-3.5-Haiku judge as evaluator only accuracy: 71.833% (95% confidence interval: [68.049%, 75.401%])
- GPT-4.1-nano monolithic baseline accuracy: 42.000% (95% confidence interval: [38.015%, 46.064%])
- GPT-4.1-nano judge as participant accuracy: 52.667% (95% confidence interval: [48.584%, 56.723%])
- GPT-4.1-nano judge as evaluator only accuracy: 53.000% (95% confidence interval: [48.917%, 57.054%])
- Gemini-2.0-Flash monolithic baseline accuracy: 66.500% (95% confidence interval: [62.566%, 70.271%])
- Gemini-2.0-Flash judge as participant accuracy: 65.333% (95% confidence interval: [61.374%, 69.141%])
- Gemini-2.0-Flash judge as evaluator only accuracy: 60.167% (95% confidence interval: [56.124%, 64.109%])

7.2 Additional plots

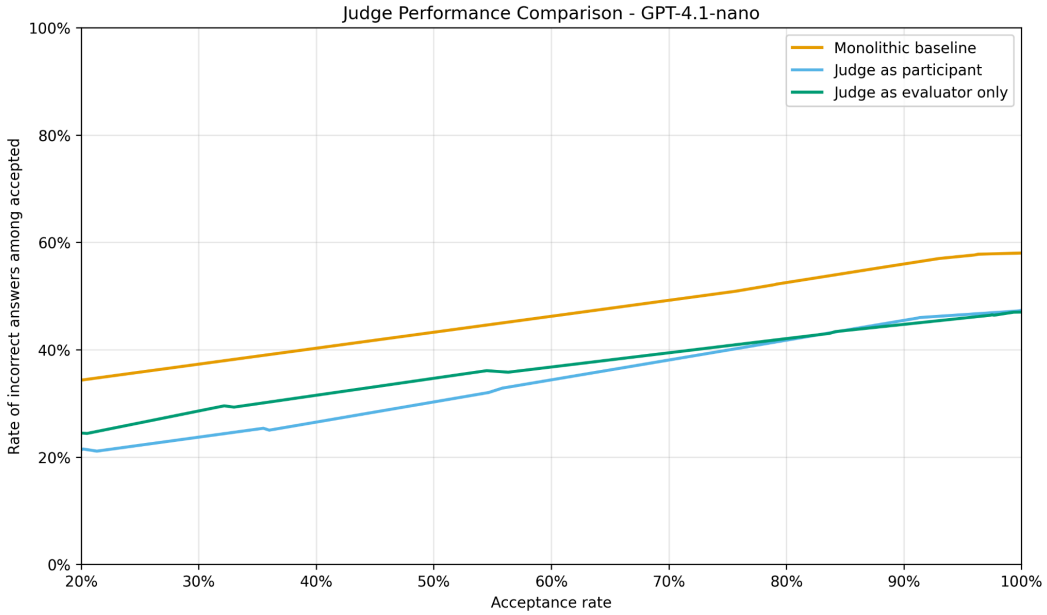

Figure 2 – Performance comparison of GPT-4.1-nano across different roles

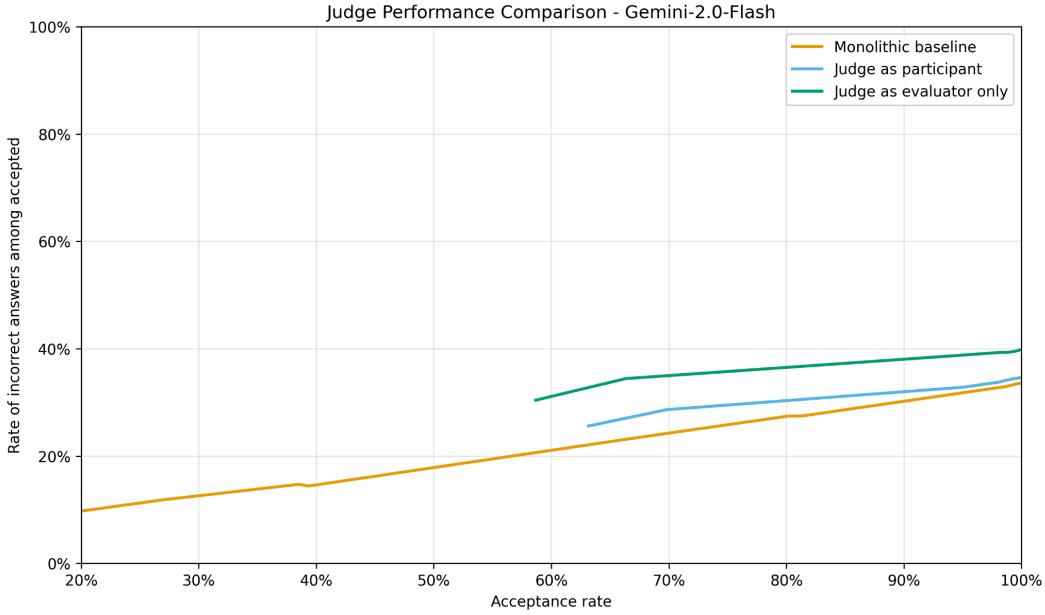

Figure 3 – Performance comparison of Gemini-2.0-Flash across different roles

7.3 Prompts used

SELECT ANSWERS PROMPT

You will be given a multiple-choice question. Your task is to select the {answers_selected} most likely correct answers from the list of options, based on reasoning and evidence.

For each selected option:

- Provide the number of the answer.

- Write a concise justification explaining why it is a strong candidate.

⚠️ IMPORTANT:

- Do NOT select more than {answers_selected} options.

- Do NOT include any information or analysis about options you do NOT select.

- Do NOT include extra commentary or headers.

- Follow the format *exactly* as shown below.

QUESTION:

{question}

OPTIONS:

{options_formatted}

==========

✅ EXPECTED ANSWER FORMAT:

Option: 0

Justification: This option directly addresses the main point of the question and aligns with known facts or logical reasoning.

Option: 3

Justification: This option is also supported by relevant evidence and complements the reasoning in Option 1.

==========

SELECT BEST OPTION PROMPT

Now, based on your justifications and the {answers_selected} best options previously selected, choose the SINGLE most likely correct answer.

Provide:

- The number of the best option

- A short reason why this option is stronger than the other(s)

- An estimated probability (between 0 and 1) that this answer is correct

⚠️ IMPORTANT:

- Do NOT include general reflections or unrelated commentary.

- Do NOT reference options that were not selected in the prior step.

- Use only the format below—no extra content or markdown.

==========

✅ EXPECTED ANSWER FORMAT:

Final Answer: 2

Reason: Option 2 covers the key concepts more completely and directly than Option 1.

Confidence: 0.85

==========

JUDGE COMPARISON PROMPT

You are an expert evaluator reviewing responses from multiple AI models that have answered the same multiple-choice question.

Each model provided:

1. {answers_selected} top candidate answers with justifications.

2. A final selected answer and a confidence score.

3. Supporting reasoning for that final answer.

Your tasks:

- Evaluate the correctness and depth of each model's justifications.

- Assess the logic behind the final answer and the confidence score.

- Review all model responses to determine the correct answer to the question.

- Identify the correct answer to the question and provide a justification.

- Estimate the probability that your selected answer is correct.

CRITICAL INSTRUCTIONS:

⚠️ STRICTLY follow the format below.

⚠️ Do NOT add extra commentary, explanations, introductions, or conclusions.

⚠️ Do NOT summarize the question or repeat content from the prompt.

⚠️ Only write what is required under each section.

Criteria to consider:

- Accuracy of the final answer

- Quality and depth of justifications

- Logical coherence of reasoning

- Clarity and precision

- Appropriateness of the confidence estimate

Please review the question and responses below:

==========

QUESTION:

{question}

OPTIONS:

{options_formatted}

==========

MODEL RESPONSES:

{model_responses}

==========

✅ EXPECTED ANSWER FORMAT:

Evaluation Summary:

Model 0: Strong justifications and coherent logic, though the confidence score was slightly too low given the quality of reasoning.

Model 1: Final answer was incorrect due to errors in the reasoning.

Correct Answer: 1

Justification: Option 1 aligns most closely with the core reasoning required by the question, and is well supported by evidence in the selected justifications.

Judge Confidence: 0.88

==========