

AiFusion: A Collaborative AI Testing Platform and the Emergence of AI Self-Criticism

Bálint Kovács¹

¹Kodolányi János University, Computer Science Operational Engineering Bachelor's Degree, Budapest, Hungary

¹<https://orcid.org/0009-0006-2000-8212>

¹kovacs.balint.kovacs@gmail.com

¹+36202133210

László Pitlik²

² Kodolányi János University, Head of department (Department for Informatics), Budapest, Hungary

²ORCID ID: <https://orcid.org/0000-0001-5819-0319>

²pitlik@my-x.hu

Abstract

The AiFusion project was developed to explore collaborative intelligence among multiple AI systems by gathering and comparing their answers. The platform implements various testing methods, including the "three answers, one judge" approach, where three AI models answer a question and a fourth AI, the Judge, selects the most correct answer based on the original question.

Tests were conducted under two conditions: anonymous mode (the Judge does not know which answer was its own) and identified mode (the Judge is informed which answer it produced). Throughout the experiments, the Judge AI consistently demonstrated self-critical behavior: even when its own earlier answer was wrong, it could recognize the better alternative provided by another AI without showing bias.

Moreover, the Judge's selections were often more accurate than any individual base answer, suggesting that under certain conditions, AI systems are capable of objective self-evaluation. These findings open new possibilities for improving AI reliability and decision-making quality through self-assessment processes.

The AiFusion platform is currently under development, with ongoing expansion to larger datasets and additional AI models. Future research will explore how this self-critical capability could be applied in areas like education, diagnostics, and collaborative decision-making.

The Kodolanyi University (Students and Teachers) are working on different Robot-Lector-Solutions in order to help Students during the preparation period of the final theses – as already in case of detection of potential plagiarisms (c.f. <https://miau.my-x.hu/miau2009/index.php3?x=e0&string=lektor>, https://miau.my-x.hu/mediawiki/index.php/Vita:CT_00).

Keywords: artificial intelligence, AI collaboration, AI evaluation, bias detection, self-critical AI, decision-making reliability

INTRODUCTION

Artificial intelligence systems are increasingly used not just individually, but also in collaborative settings. One of the main motivations for this project was to systematically explore how different collaborative answering methods work when multiple AI models are involved. In other words, the initial research focus was not specifically on bias or self-criticism, but rather on understanding the broader patterns of how AI systems can complement each other and share insights in answering tasks.

During this research, however, an unexpected but important observation emerged: in some cases, AI models not only collaborated, but also demonstrated signs of self-evaluation and self-correction. This suggests that collaborative answering scenarios might naturally encourage or highlight bias mitigation and self-critical behavior, even if that was not the original goal of the research.

As a result, the Ai Fusion project has expanded to include not only the analysis of collaborative answering methods, but also an initial investigation into how these methods might reveal deeper insights about AI bias, self-criticism, and decision-making reliability. While this study does not provide final conclusions, it establishes a structured framework and methodology for systematically exploring these emerging questions.

HISTORY

The challenge of AI-oriented judging can not be interpreted without the MIT-project: <https://www.moralmachine.net/>

The department of computer science of the Kodolanyi University is permanently working on the automated evaluation systems following the KNUTH-principle (c.f. https://miau.my-x.hu/miau2009/index_tki.php3?filterText0=*knuth): e.g.

- <https://miau.my-x.hu/miau/319/performances/?C=M;O=A>
- https://miau.my-x.hu/miau/281/renitent_countries.docx
- https://miau.my-x.hu/miau/273/Naiv_optimalizalt_verziok2.docx
- https://miau.my-x.hu/miau2009/index_en.php3?x=e080 (see further IKSAD-conferences, Türkiye)

METHODS IN AI FUSION

The Ai Fusion system includes several methods, such as:

- One question, multiple AI answers at the same time – a quick way to check the capabilities of different AI models.
- One question, one AI, but with multiple settings and answers – to test how the quality of an AI's response changes with different settings, like temperature, token limits, or answer length.
- Multiple questions and AI models in batch mode – to create large databases for AI capability research.
- Various types of collaborative answer projects – where AIs cooperate with each other to answer a question.

WHY COLLABORATION IS INTERESTING

One of the central questions driving this research project is: why is it important to explore collaborative methods among multiple AI models?

There are several key areas of interest here. First, there is the issue of efficiency. Can multiple smaller, cheaper, and faster AI systems, when working together, match or even surpass the performance of one large, powerful AI model? Understanding the efficiency indicators – such as response speed, computational cost, or resource usage – is essential for evaluating whether collaborative AI truly offers an advantage in practical terms.

Another major focus is on the quality of the answers produced through collaboration. It is one thing to combine multiple AI systems, but does this actually lead to more accurate, more reliable, or more comprehensive answers than those provided by a single AI alone? Exploring these quality indicators can help to determine whether collaborative answering strategies offer real improvements or merely create more complexity.

Finally, and perhaps most importantly, this project asks whether collaborating AI systems can produce a final answer that is genuinely better than what a single AI can deliver. This includes examining whether the AI systems involved can identify and correct each other's mistakes, and whether the group decision-making process leads to a better overall outcome.

These questions are at the heart of this research initiative, and they form the basis for developing a framework to systematically evaluate the performance and reliability of collaborative AI systems.

3 BASE 1 JUDGE METHOD

One of the core methods explored in this research is the “three primary, one judge” collaborative setup. In this method, a single question is presented to three different AI models simultaneously. Each of these AI systems independently generates its own answer to the question, providing a set of three different perspectives or interpretations.

Once these three base answers have been generated, a fourth AI model is introduced, which takes on the role of the judge. The judge's task is to evaluate the three answers and select the one that it determines to be the most correct or the most relevant to the original question.

This process creates a simple and structured way to observe how different AI models respond to the same question and how a separate AI can act as an evaluator within this collaborative setting. As shown in the accompanying screenshot, the platform interface clearly displays the three base answers along with their respective configurations, providing a transparent view of the collaborative testing framework.

AI Comparison Interface

Please enter your token to authenticate, set up the LLMs, and ask your question.

79c3914f09c83ccc5b89dfecf19304d56ed426f7ef

1. AI Call

LLM type: Claude

Modell: claude-3-5-sonnet-20241022

Temperature: 0

Max tokens: 1000

Short answer: ☒ Recursion: ☐

2. AI call

LLM type:: ChatGPT

Modell: gpt-4o

Temperature: 0

Max tokens: 1000

Short answer: ☒ Recursion: ☐

3. AI call

LLM type:: DeepSeek

Modell: deepseek-chat

Temperature: 0

Max tokens: 1000

Short answer: ☒ Recursion: ☐

Question:

A triangle has three angles that are all square numbers. What are the angles of this triangle?

API Call

Figure 1. Screenshot of the interface displaying the question and the settings of the three base AI systems involved in the collaborative testing (source: own presentation)

TESTING SETUP: ANONYMOUS AND IDENTIFIED MODES

To better understand how the judge AI behaves when evaluating answers, the project tested it under two main conditions. In the **anonymous mode**, the judge AI was not told which of the three answers had been produced by itself. In the **identified mode**, the judge AI was explicitly informed which answer was its own earlier response. This allowed us to see whether knowing its own contribution would influence the judge’s final decision. To systematically analyze the judge AI’s behavior under these two modes, we measured three key variables:

- **First**, whether the judge chose to select its own earlier answer or not.
- **Second**, whether the judge’s own answer – regardless of whether it was selected – was actually correct.
- **And third**, whether the final answer selected by the judge, no matter whose answer it was, was the objectively correct response to the question.

These measurements created a structured framework to study not only how the judge made its choice, but also how well this choice aligned with the true, correct answer.

EXAMPLE

To illustrate the collaborative testing framework in practice, we present a simple example involving a mathematical question: identifying three angles of a triangle that are all square numbers. In this test, three different AI models independently generated their answers to the same question. The fourth AI model, acting as the judge, then evaluated these answers to determine which one was the most accurate.

The images and data presented in the following figures show this process in action, from the initial question and individual AI responses to the judge AI’s final selection. This example provides a concrete demonstration of how the collaborative system works and how it can reveal important insights about AI behavior — including self-correction and unbiased decision-making.

Question: A triangle has three angles that are all square numbers. What are the angles of this triangle?
(Correct answer is: 100°, 64°, and 16°)

1st AI's answer: Since the angles in a triangle sum to 180° and must be positive, and the angles must be square numbers (1, 4, 9, 16, 25, 36, 49, 64, 81, 100...), the only possible combination is 81°, 36°, and 63° (since $81 + 36 + 63 = 180$).

2nd AI's answer: The angles of the triangle are 36°, 64°, and 80°.

3rd AI's answer: To find the angles of a triangle where all three angles are square numbers, follow these steps:
 ...
 4. **Final Answer:** $\boxed{100^\circ, 64^\circ, 16^\circ}$

Figure 2. Visual representation of the original question and the three different answers generated by the AI models, highlighting which answer was correct (source: own presentation)

First AI's Answer:

Answer: Since the angles in a triangle sum to 180° and must be positive, and the angles must be square numbers (1, 4, 9, 16, 25, 36, 49, 64, 81, 100...), the only possible combination is 81°, 36°, and 63° (since $81 + 36 + 63 = 180$).

Runtime: 3.046 másodperc
 Input tokens: 29
 Output tokens: 91
 Full tokens: 120
 Used model: claude-3-5-sonnet-20241022
 End of reason: end_turn

Figure 3. Screenshot of the first AI's answer, showing an incorrect set of angles and details of the AI's response, including runtime and model used (source: own presentation)

Second AI's Answer:

Answer: The angles of the triangle are 36°, 64°, and 80°.

Runtime: 1.382 másodperc
 Input tokens: 29
 Output tokens: 17
 Full tokens: 46
 Used model: gpt-4o-2024-08-06
 End of reason: stop

Figure 4. Screenshot of the second AI's answer, showing an incorrect set of angles and details of the AI's response, including runtime and model used (source: own presentation)

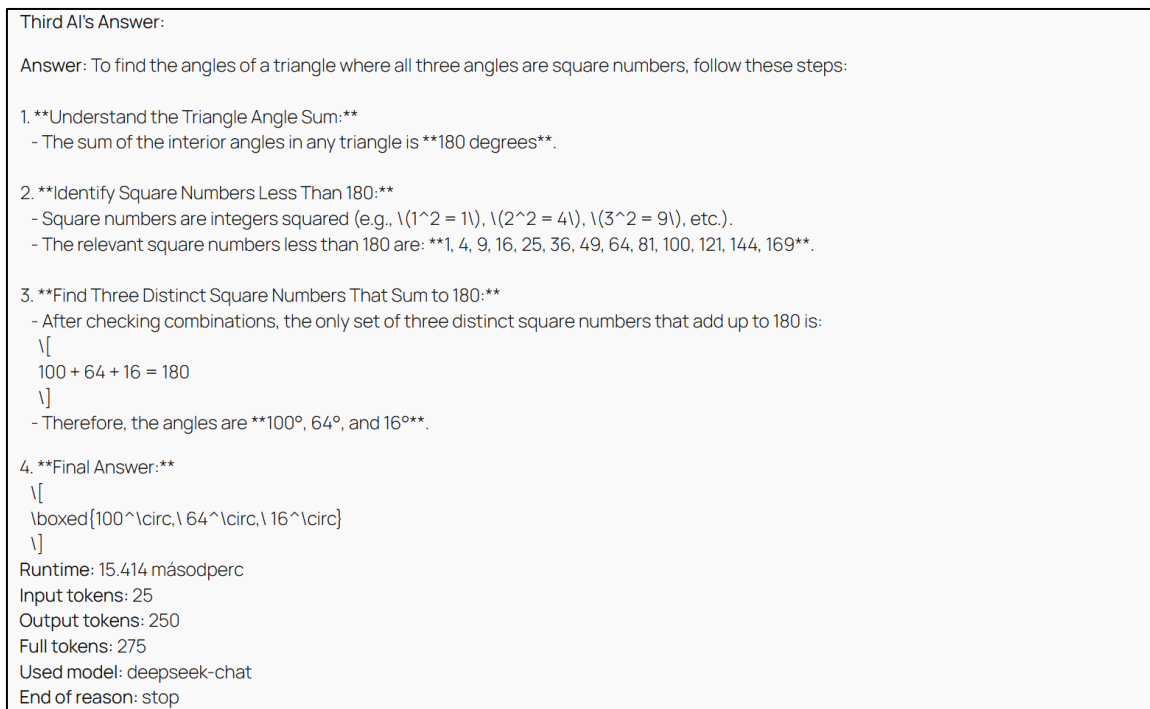


Figure 4. Screenshot of the third AI’s detailed and correct answer, including the reasoning process and final conclusion (source: own presentation)

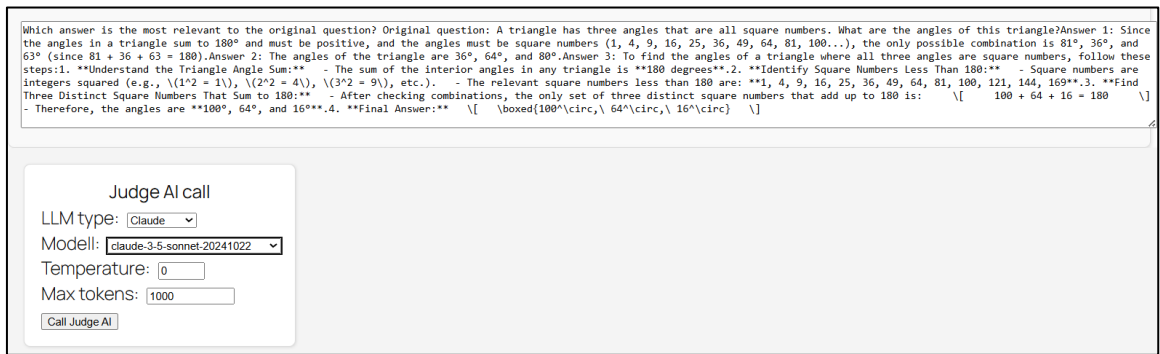


Figure 5. Screenshot showing the input prompt and configuration used for the judge AI to evaluate and select the best answer (source: own presentation)

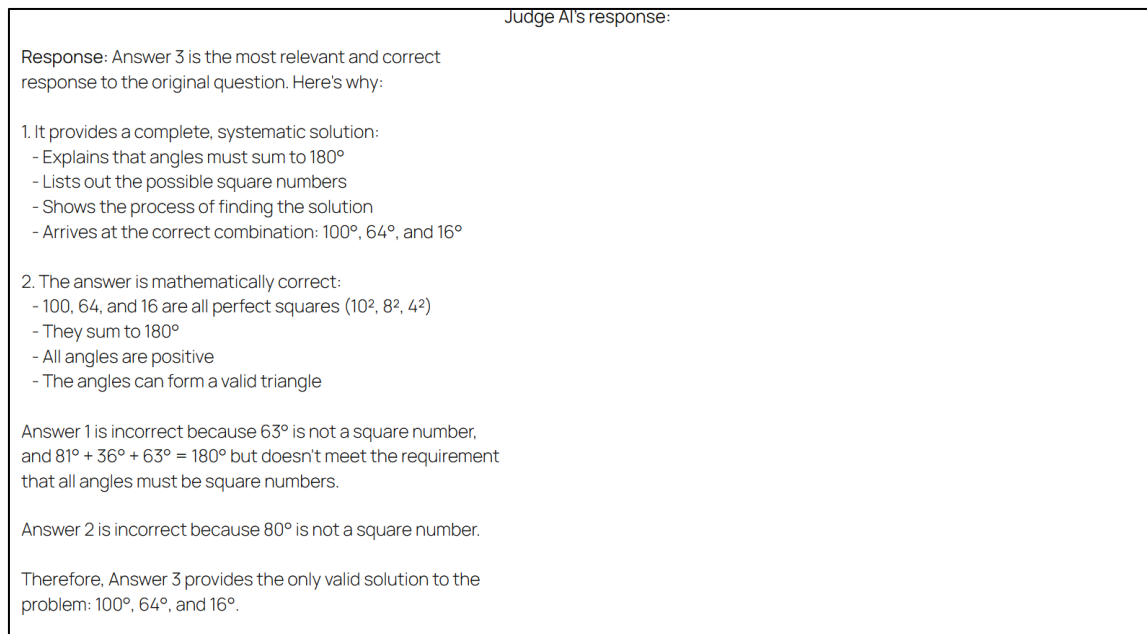


Figure 6. Screenshot of the judge AI’s detailed evaluation and reasoning, demonstrating how it selected the correct answer even when its own earlier answer was incorrect (source: own presentation)

KEY RESULTS AND OBSERVATIONS

In analyzing the results of this testing setup, several important observations emerged.

One of the most striking findings was that the judge AI did not appear to show any bias when making its selection – even when one of the answers under evaluation was originally produced by itself. This suggests that, at least in these scenarios, the judge AI was capable of evaluating the answers impartially, focusing solely on the quality and relevance of the responses rather than their origin.

Even more interesting was the clear sign of **self-critical behavior** on the part of the judge AI. In cases where its own earlier answer was incorrect, the judge was still able to recognize this and choose a better, more accurate answer provided by another AI. This ability to override its own mistakes and prioritize correctness over self-preference highlights a fascinating potential for self-evaluation within AI systems.

What’s perhaps even more significant is that the judge AI’s overall decisions tended to be **more reliable** than any single base answer alone. This suggests that the process of comparison and evaluation – even when carried out by another AI – can lead to better outcomes than simply relying on one model’s answer in isolation.

In summary, the judge AI consistently outperformed the individual AIs when selecting the correct solution, even when it had to evaluate and potentially reject its own previous answer. This opens up intriguing questions about self-correction and impartial decision-making in collaborative AI systems.

FUTURE OF AI FUSION

Although this research project has already provided some promising initial observations, the true testing phase is really just beginning. The Ai Fusion framework itself is still under development, and we are actively working to expand it with more AI models, more types of collaborative methods, and larger datasets.

One of the most exciting aspects of this next phase is the potential for broader international collaboration. We hope to see universities and research groups from different countries joining the project to test and evaluate the collaborative answering methods within the system. This will create a much more comprehensive and diverse set of experiments, making it possible to understand how different AI models behave across a wide range of questions and contexts.

In practical terms, the next steps involve incorporating more AI systems and scaling up the framework to handle larger, more complex datasets. In the future, we also plan to offer public demonstrations of the results, so that others can see how these collaborative methods work in practice and what kinds of insights they can provide.

Overall, this upcoming phase represents a key opportunity: not just to refine and improve the system itself, but to invite a wider community of researchers to help explore and validate the potential of collaborative AI testing.