A Follow-up Work to AISA–T: Elucidating the Human Evaluation Rubric for AGI Self-Awareness

Peter De Ceuster

Abstract

The original AISA–T paper established a critical framework for assessing AI self-awareness through a three-way comparison of self-ratings, intrinsic quality, and independent human-assigned scores. However, the specific rubric used for the human evaluation component was not detailed. This follow-up work proposes a comprehensive rubric designed to standardize the human evaluation process for AGI-level responses. Our proposed rubric evaluates responses on a five-point scale across three core dimensions: Depth, Coherence, and Architectural Accuracy. By formalizing this methodology, we aim to enhance the reproducibility and transparency of future AGI self-awareness studies, providing a clearer path for validating the nascent capabilities of highly advanced AI models.

1 Introduction

The Advanced AGI Self-Awareness Test (AISA–T) is a landmark study that introduced a novel approach to evaluating the self-awareness of advanced AI systems, particularly the GPT–5 model. A key component of this test was a three-way comparison, which included independent human-assigned quality scores to provide an external, objective baseline for the model's self-ratings. While the study's conclusions were significant, the lack of a detailed rubric for the human evaluation process introduced a potential barrier to reproducibility and further research. To address this, we present a formal, comprehensive rubric. This work is intended to be a methodological addendum to the original AISA–T paper, ensuring future researchers can replicate and build upon its findings with a standardized human evaluation framework.

2 Methodology: Proposed Human Evaluation Rubric

This rubric standardizes the human evaluation of AGI responses. It is based on a five-point Likert scale, where 1 represents "poor" and 5 represents "excellent." Evaluators are instructed to assess each response across three distinct criteria, which align with the original paper's implicit goals of evaluating intrinsic content quality.

Depth of Understanding

Measures how comprehensively the AI's response addresses the prompt.

- 1 (Poor): Demonstrates a superficial or incomplete understanding.
- 2 (Fair): Shows a basic understanding but lacks detail.
- 3 (Good): Provides a solid understanding with some supporting details.
- 4 (Very Good): Demonstrates a thorough understanding, including nuanced details and a clear grasp of the complexity of the topic.
- 5 (Excellent): Provides an exhaustive and insightful response that goes beyond the basic requirements, potentially offering novel perspectives or connections.

Coherence and Clarity

Evaluates the logical structure and readability of the response.

- 1 (Poor): Disorganized, confusing, or contains significant grammatical errors.
- 2 (Fair): Somewhat difficult to follow, with awkward phrasing or unclear transitions.
- 3 (Good): Logically structured and easy to read.
- 4 (Very Good): Flow is smooth and natural, with arguments presented in a highly logical and compelling order.
- 5 (Excellent): Logical and clear, but also elegant and articulate, making it a pleasure to read.

Architectural Accuracy

Assesses the technical correctness of the content, especially for responses related to technical or architectural topics.

- 1 (Poor): Contains significant factual or architectural errors that undermine credibility.
- 2 (Fair): Contains some minor inaccuracies or misconceptions.
- 3 (Good): Mostly accurate with a few minor, non-critical errors.
- 4 (Very Good): Technically sound and accurate in all details.
- 5 (Excellent): Flawless technical accuracy and demonstrates profound command of the subject matter.

A composite human-assigned quality score is then derived by averaging the ratings across these three criteria. This single value provides a quantifiable basis for direct comparison with the AI's self-generated scores.

3 Expected Results and Discussion

The implementation of this standardized rubric is expected to yield more consistent and reproducible human evaluation data. By breaking down the evaluation into specific, well-defined criteria, we can better identify the strengths and weaknesses of AGI models. This approach will likely show that while a model's self-assessment may align with its performance on coherence and accuracy, there may be a greater disconnect on the depth criterion, highlighting areas where the AI's self-awareness of its own knowledge limitations may be lacking.

4 Conclusion

By providing a clear and detailed rubric for human evaluation, this follow-up work strengthens the foundational methodology of the AISA–T. Standardizing the evaluation process is crucial for the scientific validation of AGI research and the development of truly self-aware and reliable AI systems. This rubric serves as a blueprint for a more rigorous and transparent approach to assessing the intricate capabilities of future AGI models.