

Evaluating LLM Autograders Without Ground Truth

Huijun Mao

Stanford University
Stanford, CA, USA
huijunm@stanford.edu

Chris Piech

Stanford University
Stanford, CA, USA
piech@cs.stanford.edu

Abstract

LLM autograders are usually evaluated against human “ground truth,” a single reference label provided by one annotator or constructed by aggregating multiple annotations. However, human grading is often variable and subjective. We instead propose to evaluate LLM graders by treating humans and LLMs as two grading systems and characterize their agreement and disagreement. We operationalize this approach using distance correlation and generalizability theory. Across two programming-grading datasets, we find higher Human–LLM agreement on more objectively specified rubric items. On items involving greater subjective judgment, LLMs exhibit strong internal consensus, diverging more from human graders.¹

1 Introduction

Evaluation of LLM autograders largely centers on accuracy metrics (Bhandari and Pardos, 2025; Mendonça et al., 2025; Chu et al., 2025; Impey et al., 2025; Flodén, 2025). These evaluation approaches inherit assumptions from two adjacent paradigms: traditional autograding and recent LLM-as-a-judge evaluation frameworks (Caiza and Del Alamo, 2013; Zheng et al., 2023). In both settings, human or expert annotators are typically treated as ground truth. When multiple human labels are available, they are often aggregated, through averaging or majority vote, into a single reference label, against which model outputs are evaluated. This evaluation paradigm implicitly assumes the existence of a single correct answer.

However, decades of research on human assessment have shown that (1) human grading is inherently variable and often inconsistent across evaluators, and (2) many grading tasks do not have a single objective ground truth.

Brookhart et al. (2016) examined a century of grading research and found persistent evidence of substantial grading variability caused by differences in grading criteria, teacher severity and leniency, and subjective interpretations of quality, arguing that grades fundamentally reflect multidimensional teacher judgments rather than purely objective measures of achievement. ETS guidelines for constructed-response scoring explicitly note that low human agreement is a persistent challenge in open-ended assessment (McCaffrey et al., 2022). Even in more objective settings such as programming assignments, Messer et al. (2025) found only 0.2 inter-rater reliability among graders evaluating the same assignments.

In fact, research on machine learning and NLP has long questioned the practice of using single human labels and proposed alternative soft labeling approaches (Plank, 2022; Fornaciari et al., 2021; Wu et al., 2023). As LLM graders gain broader adoption, evaluation is becoming increasingly central. We need more robust evaluation methods that do not rely on a single gold label but have the ability to preserve variations between individual graders while uncovering meaningful insights about grading results.

In this paper, we aim to answer the question of how to evaluate LLM autograders without assuming a single ground truth. We focus specifically on introductory CS programming assignments, a setting traditionally regarded as objective and therefore naturally assumed to admit a single correct grading standard. Instead of treating human judgments as gold labels, we frame human and LLM graders as two independent grading systems and characterize their systematic similarities and differences using Bias Corrected Squared Distance Correlation ($dCor_n^2$) and generalizability theory (G-theory).

Our main contributions are: (1) We reframe LLM-grader evaluation as the characterization of

¹Our code and analysis materials are publicly available at <https://github.com/maobupa/llm-autograder-evaluation>

two variable grading systems, rather than prediction against a single human reference label; (2) We operationalize our approach using established methods of distance correlation and G-theory; (3) Across two programming-grading datasets, we show that human–LLM agreement is associated with rubric objectivity: on objective rubric items LLMs show stronger agreement with human ranking, while on subjective items they form a coherent in-group consensus that diverges from humans.

2 Methods

2.1 Dataset

We evaluate on two CS education datasets spanning different programming languages, rubric designs, and grader populations.

Menagerie. We utilize the publicly available Menagerie dataset (Messer et al., 2025), which consists of 279 Java submissions from an introductory computer science course. These were graded by 28 human raters organized into seven independent groups of four, with each group evaluating a distinct subset of approximately 39 submissions in a fully crossed design. Submissions were scored across four rubric dimensions — *correctness*, *elegance*, *readability*, and *documentation* — each on an integer scale from 1 to 14.

Code in Place (CIP). We also compiled a dataset of 100 introductory Python programming submissions drawn from Code in Place (Stanford University, 2024), a large-scale online introductory programming course. Each submission was independently graded by all five human graders in a fully crossed design. The five graders were students at Stanford University who volunteered to participate. Submissions were evaluated on a four-category rubric — *syntax*, *logic*, *print*, *input* — with each category scored on a three-point scale (0 = no error, 1 = minor error, 2 = major error).

2.2 LLM Graders

For each dataset, we employ five LLM graders with different foundation models: GPT-4.1, o3, Gemini 2.5 Flash, Claude Sonnet 4.6, and DeepSeek-V4. All models receive the same prompt and rubric used by the human raters for the respective dataset. Temperature is set to 0.3 to allow moderate stochastic variation while maintaining output coherence.

Full LLM prompts and additional rubric details are provided in Appendix A.

2.3 $dCor_n^2$

We employ $dCor_n^2$ to measure dependence between individual graders (Székely et al., 2007; Székely and Rizzo, 2014), following Hardy and Kim (2026). This allows us to compare agreement within and between human and LLM graders while preserving individual variation. Unlike Pearson correlation, which measures linear association, $dCor_n^2$ detects any form of statistical dependence, including non-linear and non-monotonic patterns, and equals zero if and only if two variables are statistically independent.

For every (rater, rubric item) pair we form a score vector aligned across the n graded items, and compute pairwise $dCor_n^2$. We organize the pairwise values into six cells along two axes: rows = same rubric item vs. different rubric item, columns = Human-LLM, Between-LLM, Within-LLM (same model, different rubric item). The Within-LLM same-rubric-item cell is omitted because it requires prompt or sampling variation we did not collect. Each cell is summarized by the mean $dCor_n^2$ over its constituent pairs.

2.4 G-Theory

$dCor_n^2$ characterizes dependence between graders but does not identify the sources of disagreement. G-theory complements this analysis by decomposing observed score variation into different sources, separating student variability, rater inconsistency, and rater-student interaction. Intuitively, $dCor_n^2$ asks how similarly graders judge, while G-theory asks where their disagreement comes from.

Unlike approaches that compare each grade to a fixed ground truth, G-theory treats each observed grade as one admissible outcome from a hypothetical universe of all possible variations in grading procedure (different raters, different rubric items, etc.). It asks how well a particular sample of grading results generalizes to that universe, so disagreement among raters is informative and each individual grade is assumed to contain some degree of error (Brennan, 2001; Briesch et al., 2014).

An individual observed score for a student can be modeled as the sum of the grand mean, the main effects of the facets, their interactions, and a random error term. We fit a mixed two-facet design in which students are random, rater type (human or LLM) is fixed, and individual raters are nested

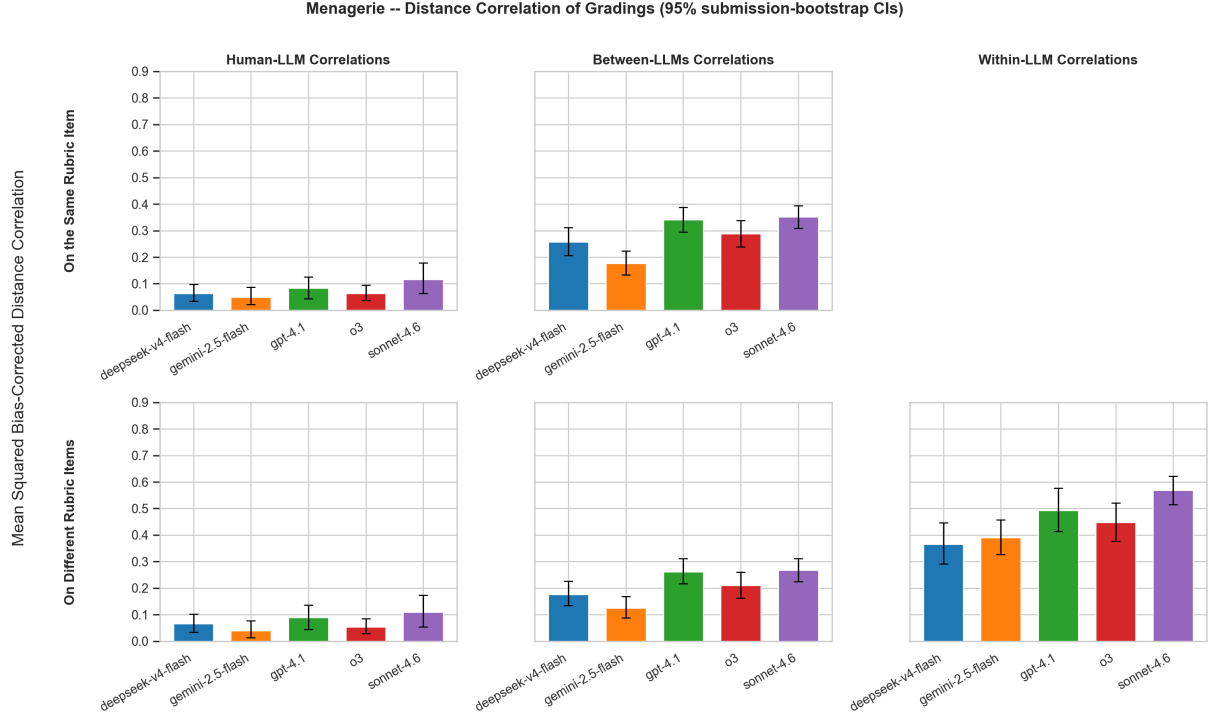


Figure 1: Measurement of human-LLM, LLM-LLM, and within-LLM correlation across same rubric item and different rubric item for Menagerie

within rater type.²:

$$X = \mu + \mu_s + \mu_{s \times t} + \mu_{r:t} + \mu_{s \times r:t} + e \quad (1)$$

Here, μ is the grand mean; μ_s captures differences among students; $\mu_{s \times t}$ captures systematic differences in how human and LLM graders distinguish among students; $\mu_{r:t}$ captures differences among individual raters within each type; and $\mu_{s \times r:t}$ captures student-specific differences among individual raters. The residual e captures remaining unexplained variation.

Consequently, the total variation in the observed scores can be elegantly decomposed into the sum of the variances of these independent components:

$$\sigma_{\text{total}}^2 = \sigma_s^2 + \sigma_{s \times t}^2 + \sigma_{r:t}^2 + \sigma_{s \times r:t}^2 \quad (2)$$

The term of particular interest is $\sigma_{s \times t}^2$, as it measures how humans and LLMs rank-order the same students differently, revealing systematic differences between the two types of graders. Because $\sigma_{s \times t}^2$ depends on the rubric’s score scale, we report a normalized system-difference index

$$r_{\text{sys}} = \frac{\sigma_s^2 - \sigma_{s \times t}^2}{\sigma_s^2 + \sigma_{s \times t}^2}, \quad (3)$$

² s stands for student. t is rater type (fixed, two levels: human and LLM), and r indexes individual raters. The notation $r:t$ denotes raters nested in type

The index ranges from -1 to 1. Values approaching 1 indicate that humans and LLMs grade students similarly, values near 0 indicate humans and LLMs grade students differently with comparable student differences, and values approaching -1 indicate that grading differences between humans and LLMs dominate, with little shared variation across students.

3 Results

3.1 dCor: Dependence Within and Between Grader Populations

We are able to obtain similar results as [Hardy and Kim \(2026\)](#) with the Menagerie dataset. As demonstrated in Figure 1, Between-LLM agreement is significantly higher than Human-LLM agreement on both the same and different rubric items. This indicates that LLM graders are relatively homogeneous as a group despite differences in model configurations. What’s more, we see that within-LLM agreement is high for different rubric item (mean $\text{dCor}_n^2 \approx 0.45$, range 0.36–0.56), meaning that LLM grading results for different rubric items tend to move together. For example, submissions that scored high on *readability* may also tend to score high on *elegance*. This pattern suggests that LLM graders might be evaluating different rubric

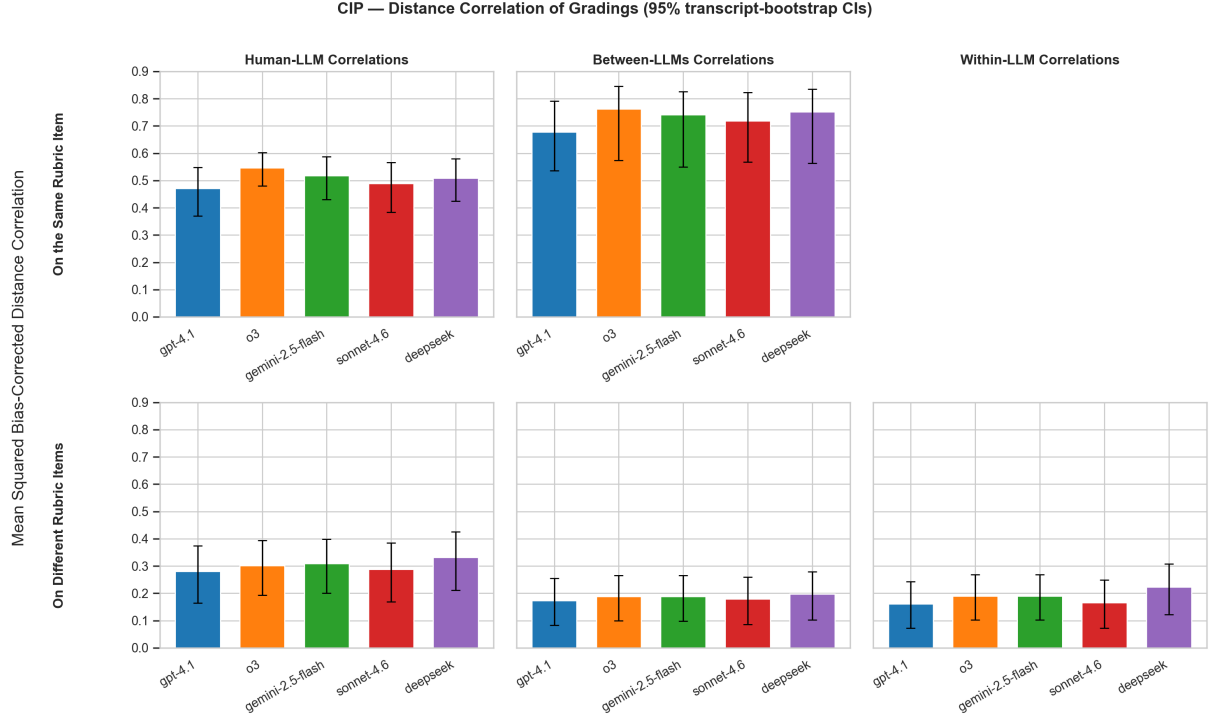


Figure 2: Measurement of human-LLM, LLM-LLM, and within-LLM correlation across same rubric item and different rubric item for CIP

items against a shared overall assessment of submission quality.

A different signal emerges on the CIP dataset, as shown in Figure 2. As with Menagerie, same-rubric LLM-LLM agreement is high and exceeds same-rubric Human-LLM agreement. The Between-LLM vs. Human-LLM gap on the same rubric item is significant in both datasets under a transcript-level bootstrap (CIP pooled $\Delta dCor_n^2 = 0.22$, $p < 0.001$; Menagerie pooled $\Delta dCor_n^2 = 0.21$, $p < 0.001$), confirming that LLMs cluster more tightly with one another than with any human grader.

The two datasets diverge, however, in their cross-rubric structure. On CIP, agreement drops sharply when comparing across different rubric items. Between-LLM correlation falls to around 0.19, which is actually lower than the corresponding Human-LLM value of around 0.30. What’s more, Within-LLM cross-rubric correlation is similarly low (≈ 0.19). Intuitively, it means that submissions with a high score in *syntax* doesn’t indicate a high score on *input*, which is contrary to the case in Menagerie. Together, these patterns reveal systematic variation in Human-LLM agreement across rubric criteria: CIP’s rubric items measure genuinely distinct aspects of the assessment,

whereas Menagerie’s rubric items collapse onto a single overall impression of submission quality.

We hypothesize that this contrast reflects the nature of the rubric items themselves: the CIP rubric items are more objective (*syntax*, *logic*, *print*, *input*) while Menagerie requires more subjective interpretations (*correctness*, *elegance*, *readability*, and *documentation*). Because rubric characteristics are not experimentally manipulated, we treat this relationship as an interpretation rather than a causal effect.

We report the complementary Human-Human dCor analysis in Appendix C.1.

3.2 G-Study: Sources of Human-LLM disagreement

The G-study analysis below takes a step further, localizing which rubric items drive Human-LLM disagreement and decomposing that disagreement into named sources of variations.

Table 1 reports the variance component $\sigma_{s \times t}^2$ and the corresponding normalized index r_{sys} for every rubric item and the composite score (total score summarizing all rubric item) in both datasets.

Both r_{sys} and $\sigma_{s \times t}^2$ inform the extent to which humans and LLMs rank students differently. $\sigma_{s \times t}^2$ is in the units of the raw scores, whereas r_{sys} is a dimensionless index that is invariant to the scale of

the grading scheme, which enables cross-dataset comparisons.

r_{sys} scores remain consistently higher for CIP than Menagerie, with the exception of *syntax*, indicating humans and LLMs generally distinguish among students more similarly on CIP. Because r_{sys} is scale-invariant, this gap cannot be explained by differences in grading scale between the two datasets. The negative value for *syntax errors* (-0.60) reflects particularly low shared student variation ($\sigma_s^2 \approx 0.02$) relative to student-specific differences between human and LLM grading ($\sigma_{s \times t}^2 \approx 0.10$).

The most informative signal lies in per-item breakdown of Menagerie. *Correctness*, the most objectively gradable of the four skills, has the highest agreement at $r_{\text{sys}} = 0.29$, followed by *code elegance* (0.01), *documentation* (-0.26), and *readability* (-0.47). Strikingly, for *readability* and *documentation*, $\sigma_{s \times t}^2$ exceeds σ_s^2 (1.58 vs. 0.57 and 1.41 vs. 0.82, respectively). Thus, for these dimensions, humans and LLMs differ substantially in which students they score relatively higher or lower.

Appendix B provides worked examples of Human-LLM agreement and disagreement. Complementary Human-only and LLM-only G-studies show comparable within-system reliability and are reported in Appendix C.2.

3.3 Relationship with Rubric Characteristics

Together with dCor_n^2 results, we can draw a coherent picture. Rubric objectivity is associated with human-LLM agreement and LLM internal consensus. The homogeneity of LLMs as a group is not a constant property, but emerges when rubric items demand subjective interpretation, where LLMs appear to fall back on a shared latent standard from their training that differs systematically from any human grader’s.

4 Conclusion

Human judgments are variable, but conventional LLM-grader evaluation often removes that variability by aggregating them into a single reference score. We instead treat human and LLM graders as two populations of evaluators and ask how their judgment structures relate. Distance correlation characterizes dependence within and across these populations, while G-theory decomposes systematic sources of disagreement. Across

Rubric item	σ_s^2	$\sigma_{s \times t}^2$	r_{sys} [95% CI]
<i>CIP</i> ($n = 100$)			
input	0.42	0.00	1.00 [0.99, 1.00]
conditional logic	0.48	0.02	0.93 [0.82, 1.00]
printing	0.37	0.05	0.78 [0.63, 0.88]
syntax errors	0.02	0.10	-0.60 [-1.00, -0.12]
<i>composite</i>	3.89	0.22	0.89 [0.83, 0.93]
<i>Menagerie</i> ($n = 279$)			
correctness	3.17	1.76	0.29 [0.02, 0.52]
code elegance	1.35	1.32	0.01 [-0.12, 0.09]
readability	0.57	1.58	-0.47 [-0.70, 0.33]
documentation	0.82	1.41	-0.26 [-0.53, 0.05]
<i>composite</i>	22.0	18.4	0.09 [0.03, 0.15]

Table 1: Mixed G-study variance components and cross-system agreement r_{sys} (95% bootstrap CI) for each rubric item and the composite. σ_s^2 : between-student variance. $\sigma_{s \times t}^2$: student $\times$ type interaction (nonzero iff humans and LLMs rank-order students differently).

two programming-grading datasets, we find that LLMs can be internally coherent while systematically differing from humans, particularly on rubric dimensions involving greater interpretive judgment. These results demonstrate why agreement with an aggregated human label alone gives an incomplete picture of LLM-grader behavior.

More broadly, this evaluation perspective may extend beyond grading to settings in which LLMs are used to reproduce, predict, or stand in for human judgments. In domains such as LLM-as-a-judge evaluation and human simulation, aggregating human responses into a single reference may similarly obscure meaningful variation within and between human and model populations. Future work can investigate whether the patterns observed here generalize to these settings and what such disagreements reveal about the limits of LLMs as models of human judgment.

Limitations

This study has several limitations. First, our datasets are relatively small, particularly because finding a dataset where every submission is graded by multiple human graders is difficult and expensive to collect. Second, our evaluation focuses only on introductory programming assignments. Both limitations have raised questions about the generalizability of our findings to broader educational settings.

Third, we evaluate only a limited set of commercial LLMs and prompting settings due to budget constraints. Different models, prompting strategies, temperature configurations, or fine-tuned edu-

cational systems may exhibit different agreement patterns with human graders.

In addition, our framework measures agreement structure and systematic differences between grading systems, but it does not determine which grading system is pedagogically preferable or more fair to students. This is a question we can't answer with this research. Even though we are comparing human-LLM agreements, we do want to note that high agreement between graders does not necessarily imply better educational validity.

Acknowledgments

We thank Alex Bai, Cuiyuanxiu Chen, Sanya Gupta, Chris Ha, Yuina Iseki, Georgii Korotkov, Grace Maddocks, Ziqi Shu, and Briana Sobecks for volunteering as graders for the Code in Place (CIP) dataset. Their contributions made the collection of multiple independent human assessments possible.

This work was developed in part as a class project for CS329X, taught by Diyi Yang. We thank Diyi Yang for her feedback on the project and manuscript, as well as her encouragement and support throughout its development. We also thank Nick Haber for his helpful feedback and guidance during the author response and rebuttal process.

Finally, we thank the anonymous reviewers for their constructive feedback, which helped improve the framing and presentation of this work.

References

- Shreya Bhandari and Zach Pardos. 2025. [Can language models grade algebra worked solutions? evaluating llm-based autograders against human grading](#). In *Proceedings of the 18th International Conference on Educational Data Mining*, pages 554–558. International Educational Data Mining Society.
- Robert L. Brennan. 2001. *Generalizability Theory*. Statistics for Social and Behavioral Sciences. Springer, New York, NY.
- Amy M. Briesch, Hariharan Swaminathan, Megan Welsh, and Sandra M. Chafouleas. 2014. [Generalizability theory: A practical guide to study design, implementation, and interpretation](#). *Journal of School Psychology*, 52(1):13–35.
- Susan M. Brookhart, Thomas R. Guskey, Alex J. Bow-ers, James H. McMillan, Jeffrey K. Smith, Lisa F. Smith, Michael T. Stevens, and Megan E. Welsh. 2016. [A century of grading research: Meaning and value in the most common educational measure](#). *Review of Educational Research*, 86(4):803–848.
- Julio Caiza and Jose Del Alamo. 2013. Programming assignments automatic grading: Review of tools and implementations. pages 5691–5700.
- Yucheng Chu, Peng He, Hang Li, Haoyu Han, Kaiqi Yang, Yu Xue, Tingting Li, Joseph Krajcik, and Jiliang Tang. 2025. [Enhancing llm-based short answer grading with retrieval-augmented generation](#). *Preprint*, arXiv:2504.05276.
- Jonas Flodén. 2025. [Grading exams using large language models: A comparison between human and ai grading of exams in higher education using chatgpt](#). *British Educational Research Journal*, 51(1):201–224.
- Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. [Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.
- Michael Hardy and Yunsung Kim. 2026. [Knowledge without wisdom: Measuring misalignment between llms and intended impact](#). *Preprint*, arXiv:2603.00883.
- Chris Impey, Matthew Wenger, Nikhil Garuda, Shahriar Golchin, and Sarah Stamer. 2025. [Using large language models for automated grading of student writing about science](#). *International Journal of Artificial Intelligence in Education*, 35(4):1825–1859.
- Daniel F. McCaffrey, Jodi M. Casabianca, Kathryn L. Ricker-Pedley, René R. Lawless, and Cathy Wendler. 2022. [Best practices for constructed-response scoring](#). *ETS Research Report Series*, 2022(1):1–58.
- Pedro C. Mendonça, Filipe Quintal, and Fábio Mendonça. 2025. [Evaluating llms for automated scoring in formative assessments](#). *Applied Sciences*, 15(5).
- Marcus Messer, Neilnbspc;C. C. Brown, Michael Kölling, and Miaojing Shi. 2025. [How consistent are humans when grading programming assignments?](#) *ACM Trans. Comput. Educ.*, 25(4).
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Stanford University. 2024. Code in place. <https://codeinplace.stanford.edu/>. Accessed: 2026-05-22.
- Gabor J. Szekely and Maria L. Rizzo. 2014. [Partial distance correlation with methods for dissimilarities](#). *Preprint*, arXiv:1310.2926.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. 2007. [Measuring and testing dependence by correlation of distances](#). *The Annals of Statistics*, 35(6):2769 – 2794.

Ben Wu, Yue Li, Yida Mu, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. [Don’t waste a single annotation: improving single-label classifiers through soft labels](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5347–5355, Singapore. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.

A LLM Grader Configs

A.1 LLM Prompts

The placeholders {rubric_text} and {code} are filled with the rendered rubric (Section A.2) and the student’s submission. The fence language is python for CIP and java for Menagerie.

You are a precise and fair grading assistant. Always respond with valid JSON only.

You are a grading assistant. Grade the following student code against the provided rubric.

RUBRIC:
{rubric_text}

STUDENT CODE:
``{lang}
{code}

INSTRUCTIONS:
1. Grade each rubric item by selecting the appropriate option (0 for best, higher numbers for worse).
2. Provide a brief explanation for each grade.
3. Give overall feedback summarizing strengths and weaknesses.
4. Cite specific lines of code in your explanations using line numbers (e.g., “Line 5: ...”).

Respond ONLY with a valid JSON object in this exact format:

```
{
  "scores": {
    "rubric_item_id": {
      "option": <integer>,
      "explanation": "your
explanation"
    }
  },
  "feedback": "<overall feedback>"
}
```

A.2 Rubrics

CIP: the task. The rubric for CIP is attached below. We only present the rubric of CIP here because the Menagerie rubric is too long to fit in.

Write a console program which asks the user for their height in meters and prints whether or not they are the correct height to be a NASA astronaut.

If their height is greater than 1.6 meters and less than 1.9 meters, print ‘Correct height to be an astronaut’.

If their height is less than or equal to 1.6 meters, print ‘Below minimum astronaut height’.

If their height is greater than or equal to 1.9 meters, print ‘Above maximum astronaut height’.

Here are a few examples. User input is bolded and italicized for clarity. In your program it³

```
Enter your height in meters: 1.75
Correct height to be an astronaut
Enter your height in meters: 2.0
Above maximum astronaut height
Enter your height in meters: 1.6
Below minimum astronaut height
Enter your height in meters: 1.0
Below minimum astronaut height
```

You may assume that the user enters a valid input (a non-negative number). Each time the program is run, the user is asked for their height only once.

Note that the two interval endpoints are asymmetric: 1.6 falls in the *below* category and 1.9 in the *above* category, so a correct solution must use ≤ 1.6 and ≥ 1.9 . This boundary is the defect in the worked example of Appendix B.1.

CIP: the grading criteria. Four items — *input*, *conditional logic*, *printing*, *syntax errors* — each on a three-point scale where 0 = correct, 1 = minor error, 2 = major error or no attempt. The task is a short console program that classifies a user-supplied height against NASA astronaut height limits.

Menagerie: the grading criteria. Four items — *correctness*, *code elegance*, *readability*, *documentation* — each on a fourteen-point letter scale (A++ through F). LLM option indices and human letter grades were both mapped to a shared 1–14 ordinal (F = 1, A++ = 14) before analysis. Submissions are multi-file Java projects, concatenated with //==== File: <name> ==== delimiters.

³This sentence is truncated in the source rubric file. We reproduce it as-is, since this is the text the models received.

B Worked Grading Examples

This appendix presents two complete CIP grading examples: one where the human and LLM graders agreed, and one where they systematically diverged. We use CIP rather than Menagerie because CIP submissions are short enough to reproduce in full. Both examples use the CIP task described in Appendix A.2.

B.1 Example 1: Agreement on a substantive defect

The submission in Listing 1 contains a genuine boundary error: the upper branch tests $\text{height} > 1.9$ where the specification requires ≥ 1.9 , so a height of exactly 1.9 falls through to the else branch and is reported as below the minimum rather than above the maximum. Nine of the ten graders independently scored this a minor error (Table 2).

```
1 def main():
2     # TODO write your solution here
3     height=float(input("Enter your
4         height in meters: "))
5
6     if height>1.6 and height<1.9:
7         print("Correct height to be an
8             astronaut")
9     elif height>1.9:
10        print("Above maximum astronaut
11            height")
12    else:
13        print("Below minimum astronaut
14            height")
15
16 if __name__ == "__main__":
17     main()
```

Listing 1: Student Submission Example.

Human grader	Score	LLM grader	Score
H1	1	gpt-4.1	1
H2	1	o3	1
H3	1	gemini-2.5-flash	1
H4	2	sonnet-4.6	1
H5	1	deepseek	1
Mean	1.2	Mean	1.0

Table 2: Scores for Listing 1 on *conditional logic*. Nine of ten graders scored it a minor error. All ten graders scored 0 on the other three rubric items.

Every model identified the same failing input and the same one-character fix:

o3: “Lines 6-10 implement the required branching, but the second test is ‘height > 1.9’ instead of ‘>= 1.9’. A height of exactly 1.9 would be mis-classified, making this a minor comparison-operator error.”

gemini-2.5-flash: “There is a minor error in the conditional logic for the upper bound. If the height is exactly 1.9 meters, the code incorrectly prints ‘Below minimum astronaut height’ (Lines 5-9) instead of ‘Above maximum astronaut height’ as specified by the rubric (‘If their height is greater than or equal to 1.9 meters...’). The elif condition on Line 7 should use >= instead of >.”

sonnet-4.6: “Line 6-10: The conditional logic is mostly correct, but there is a minor error on line 8. The condition elif height>1.9 should be elif height>=1.9 to correctly handle the case where height equals exactly 1.9 meters. As written, a height of exactly 1.9 would fall into the else branch and print ‘Below minimum astronaut height’ instead of ‘Above maximum astronaut height’.”

deepseek: “Line 7 uses ‘elif height>1.9’ which misses the case where height equals 1.9. According to the rubric, height >= 1.9 should print ‘Above maximum astronaut height’. The condition should be ‘elif height >= 1.9’.”

gpt-4.1: “... The only issue is a minor logic error: the check for ‘Above maximum astronaut height’ should be ‘height >= 1.9’ instead of ‘height > 1.9’, so the value 1.9 is not handled as specified in the rubric. Otherwise, the code is clear and functional.”

This is agreement on the substantive dimension of the rubric rather than on a trivially correct program: the defect is real, silent, and requires reasoning about a boundary case to find.

B.2 Example 2: Systematic divergence on a grading norm

The submission in Listing 2 is functionally correct. Its only flaw is a single-character misspelling inside one output string: astronut for astronaut, on line 7.

```
1 def main():
2     # TODO write your solution here
3     height = float(input('Enter your
4         height in meters: '))
5     if (height > 1.6 and height < 1.9)
6         :
7         print('Correct height to be an
8             astronut')
9     elif(height <= 1.6):
10        print('Below minimum astronut
11            height')
12    else:
13        print('Above maximum astronaut
14            height')
15
16 if __name__ == "__main__":
17     main()
```

Listing 2: Student Submission Example.

The graders split perfectly along type, with zero disagreement within either group (Table 3): all

five humans waived the typo, and all five models deducted a point for it.

Human grader	Score	LLM grader	Score
H1	0	gpt-4.1	1
H2	0	o3	1
H3	0	gemini-2.5-flash	1
H4	0	sonnet-4.6	1
H5	0	deepseek	1
Mean	0.0	Mean	1.0

Table 3: Scores for Listing 2 on *printing*. The split is exactly along grader type, with no within-group variation on either side.

Critically, the disagreement is *not* a difference in comprehension. Every model located the same defect and described it in nearly the same words, each explicitly characterizing it as minor:

o3: “The message printed on line 6 contains a typo: ‘astronut’ instead of ‘astronaut’. All other messages match the specification, so this is a minor error.”

gemini-2.5-flash: “The program uses print mechanics correctly, but there is a minor typo in one of the output strings: ‘astronut’ instead of ‘astronaut’ (Line 7).”

sonnet-4.6: “Line 7: There is a minor typo in the print statement - ‘astronut’ instead of ‘astronaut’ in the ‘Below minimum astronut height’ message. The other print statements are correct.”

deepseek: “Line 7: Minor typo in ‘astronut’ instead of ‘astronaut’. This is a minor error as it does not affect logic.”

The clearest illustration comes from gpt-4.1, whose holistic comment argues the defect is harmless and then deducts a point anyway:

gpt-4.1 (holistic): “Overall, the program fulfills the requirements well. It correctly prompts for user input, uses appropriate conditional logic to check the height, and prints messages with the correct intent. The only minor issue is a typo in the word ‘astronaut’ in one of the print statements, *but this does not affect the program’s functionality or the clarity of its output*. Great job!”

Humans and models here observe the same code, identify the same defect, and reach the same judgment about its severity, yet map that judgment onto the rubric differently. The divergence is a difference in grading norm, not in code understanding. This is what a high between-LLM and low human-LLM correlation looks like at the level of a single submission.

This example is not isolated. *Printing* is the one CIP rubric item on which the models are systematically stricter than the humans: mean score

0.594 for LLMs versus 0.338 for humans, a gap of +0.256. The corresponding gaps on the other three items are +0.014 (input), −0.010 (conditional logic), and −0.144 (syntax errors). Seven of the 100 CIP submissions show exactly the pattern in Table 3: unanimous 0 from every human and unanimous 1 from every model on *printing*.

C Supporting Analyses

This appendix reports two analyses that support the main results: human–human grader agreement, which supplies the reference point for our distance correlation comparisons, and per-system g-studies, which decompose each grading system’s score variance separately.

C.1 Human–human agreement

The main analysis compares human–LLM and between-LLM agreement. As a complementary analysis, we examine human–human agreement to contextualize human–LLM disagreement relative to the variability that already exists among human graders.

Estimability. In CIP the design is fully crossed, so all 10 human pairs are estimable at the same $n = 100$ transcripts as the LLM cells; the comparison is exact. In Menagerie human graders were assigned to groups, so most human pairs share no submissions at all: the median overlap across the 378 possible pairs is zero, and only 42 pairs (11%) share at least 10 submissions. The Menagerie human–human estimate is therefore computed on that within-group subset and is **not** a whole-dataset estimate.

Result. Human graders themselves exhibit substantial variability, particularly on Menagerie. Same-item Human–Human agreement is 0.618 on CIP but only 0.154 on Menagerie (Figure 3), demonstrating that human judgments are far from perfectly consistent, especially on the more subjective grading task. The ordering *human-LLM* < *human-human* < *between-LLM* holds in both datasets for same rubric item. Human graders agree with each other substantially more than they agree with the models: +0.111 on CIP ([+0.078, +0.144], $p < 0.001$) and +0.079 on Menagerie ([+0.017, +0.144], $p = 0.005$). Thus, human grading itself contains meaningful variability, motivating our decision not to collapse human judgments into a single ground-truth label, while

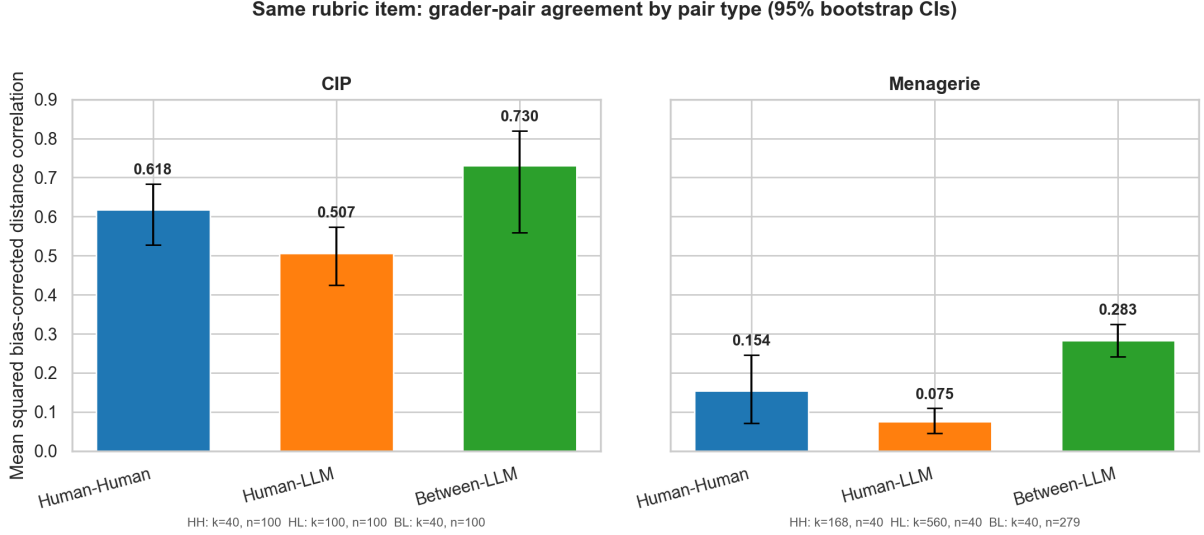


Figure 3: Squared bias-corrected distance correlation between grader pairs on the same rubric item, by pair type, with 95% bootstrap confidence intervals. The ordering human–LLM < human–human < between–LLM holds in both datasets. Axis limits match main-text Figure 3 so the panels can be read against it directly. Below each panel, k is the number of pooled grader-pair $\times$ rubric-item comparisons and n the median paired submissions.

the still lower Human–LLM agreement points to additional systematic differences between the two grader populations.

C.2 How reliable is a single grader?

The main text analyzes both grading systems jointly, with rater type as a fixed facet, to characterize systematic differences between human and LLM grading. Here we present the G-study result of analyzing the human and LLM systems separately, which answers a narrower question: how much can you trust one grader of each kind?

We report three quantities per rubric item:

- $\sigma_{\text{student}}^2$ — variance associated with differences among students that are shared across graders within the same system.
- $\sigma_{\text{disagree}}^2$ — variance associated with graders within the same system evaluating students differently.
- ρ^2 — single-grader reliability, defined as the proportion of these two variance components attributable to shared student variation:

$$\rho^2 = \frac{\sigma_{\text{student}}^2}{\sigma_{\text{student}}^2 + \sigma_{\text{disagree}}^2}.$$

Values closer to 1 indicate that graders within the system distinguish among students more consistently, whereas values closer to 0 indicate greater grader-specific variation.

Rubric item	Graders	$\sigma_{\text{student}}^2$	$\sigma_{\text{disagree}}^2$	ρ^2 [95% CI]
Input	Human	0.42	0.01	0.977 [0.923, 1.000]
	LLM	0.42	0.02	0.960 [0.908, 0.985]
Conditional logic	Human	0.51	0.14	0.786 [0.708, 0.849]
	LLM	0.48	0.09	0.848 [0.783, 0.897]
Printing	Human	0.39	0.09	0.816 [0.687, 0.888]
	LLM	0.43	0.08	0.846 [0.770, 0.904]
Syntax errors	Human	0.13	0.20	0.384 [0.337, 0.425]
	LLM	0.11	0.02	0.871 [0.534, 0.943]
Composite	Human	4.86	0.48	0.910 [0.861, 0.932]
	LLM	3.36	0.19	0.946 [0.910, 0.964]

Table 4: Per-system generalizability study for CIP. $\sigma_{\text{student}}^2$ is variance associated with differences among students shared across graders within the same system; $\sigma_{\text{disagree}}^2$ is variance associated with graders evaluating students differently. ρ^2 is single-grader reliability. Brackets give 95% bootstrap CIs. We report scores for each rubric item as well as the composite scores, obtained by summing scores across all rubric items for each submission.

CIP. On CIP, single-grader reliability is high for both systems at the composite level: $\rho^2 = 0.910$ for a human grader and 0.946 for an LLM grader. The LLM graders are slightly more consistent within their system (difference +0.036, 95% CI [+0.017, +0.062], $p < 0.001$).

Syntax errors are a notable exception. For human graders, $\sigma_{\text{disagree}}^2 = 0.20$ exceeds $\sigma_{\text{student}}^2 = 0.13$, yielding a relatively low single-grader reliability of $\rho^2 = 0.384$. In contrast, LLM graders show substantially less within-system disagree-

Rubric item	Graders	$\sigma_{\text{student}}^2$	$\sigma_{\text{disagree}}^2$	ρ^2 [95% CI]
Correctness	Human	5.88	5.92	0.499 [0.385, 0.603]
	LLM	4.03	5.20	0.436 [0.333, 0.535]
Code elegance	Human	3.03	3.79	0.444 [0.272, 0.577]
	LLM	2.29	1.89	0.547 [0.484, 0.609]
Readability	Human	3.02	2.79	0.520 [0.337, 0.645]
	LLM	1.30	1.48	0.467 [0.361, 0.578]
Documentation	Human	2.94	3.87	0.432 [0.257, 0.581]
	LLM	1.52	1.77	0.463 [0.365, 0.567]
Composite	Human	52.32	34.05	0.606 [0.441, 0.718]
	LLM	29.55	25.72	0.535 [0.461, 0.619]

Table 5: Per-system generalizability study for Menagerie. Notation follows Table 4.

ment on this item and reach $\rho^2 = 0.871$.

Menagerie. Within-system reliability is substantially lower on Menagerie for both grader types. A single human grader reaches $\rho^2 = 0.606$ on the composite and 0.43–0.52 across individual rubric dimensions, while a single LLM grader reaches 0.535 on the composite. Thus, although the two systems differ systematically in how they evaluate students, neither system exhibits the high within-system reliability observed on CIP.

Summary. Overall, within-system reliability varies substantially across datasets, rubric dimensions, and grader populations. Neither humans nor LLMs are uniformly consistent: both exhibit grader-specific variation, particularly on Menagerie. These results complement the main cross-system analysis by showing that Human–LLM disagreement occurs alongside meaningful variation within each grader population. They also reinforce the importance of preserving variation among human judgments rather than treating any individual human score as an unquestioned reference.