When AI Becomes the Grader: Comparing LLMs Versus Human Graders on Physics Essays

By: Jason Tran, Timothy Swanson, James Freericks, and Leanne Doughty

Department of Physics, Georgetown University

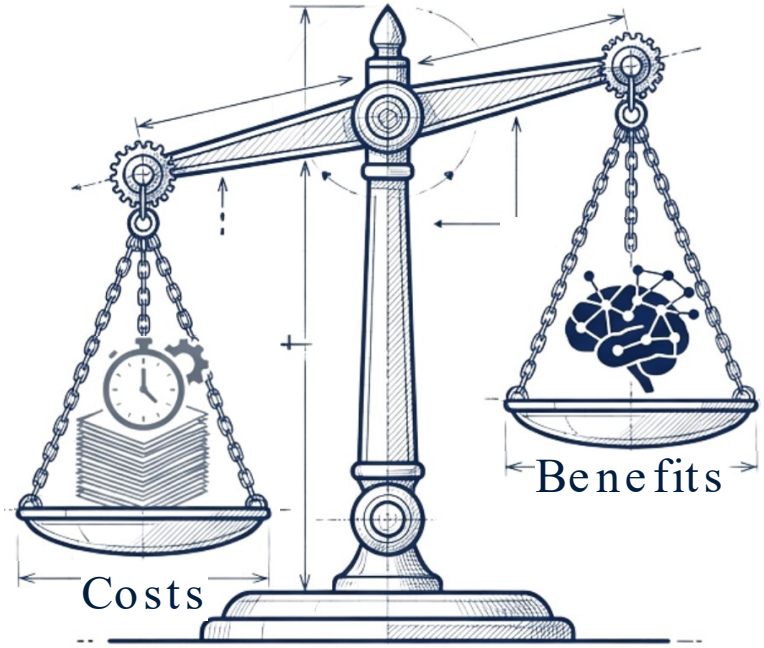




Balancing The Importance of Physics Essays

Logistical Reality

- Severe time constraints
- Difficult to scale in quantity
- High interrater variability with TAs
- Grader bias

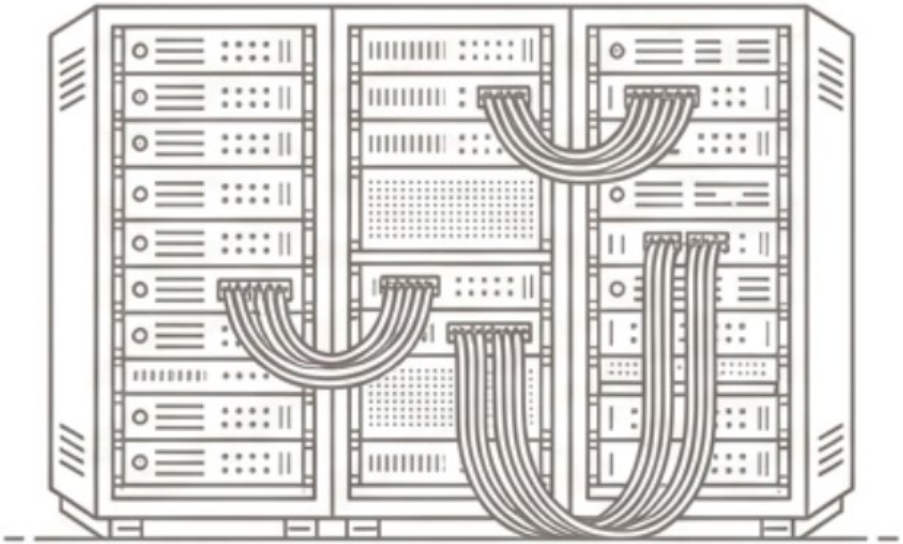


Pedagogical Value

- Fosters critical thinking
- Develops reasoning structures
- Promotes deep understanding

Dockter et. al., 2016; Doughty & Caballero, 2015; Scott et. al, 2006

AI Tools in Physics Grading



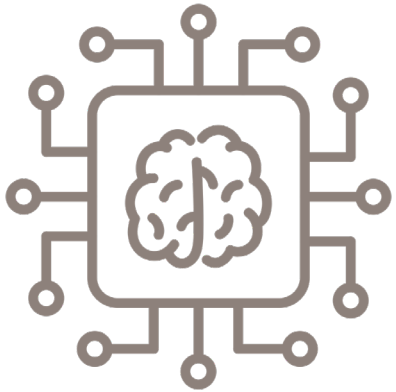
- PER literature has shown that some universities are beginning to adopt AI-assisted grading in their physics courses

- While these are highly powerful techniques, it required extensive research and resources in building AI models and agents to reach high fidelity

Kortemeyer & Nohl, 2024; Kortemeyer & Nohl, 2025

Research Question

Can readily available AI -tools assist AI -novice teachers in reliably grading of physics essays?





Roadmap

1. The Prompt Setup

Selecting a conceptually rich task

2. Human Calibration

Defining grading baseline

3. The AI Protocol

Progressively scaffolding AI support

4. Evaluate Framework

Comparing AI and human scoring

5. Statistics

Validating scores through statistics

The Prompt Setup

1. The Prompt Setup

Selecting a conceptually rich task

2. Human Calibration

Defining grading baseline

3. The AI Protocol

Progressively scaffolding AI support

4. Evaluate Framework

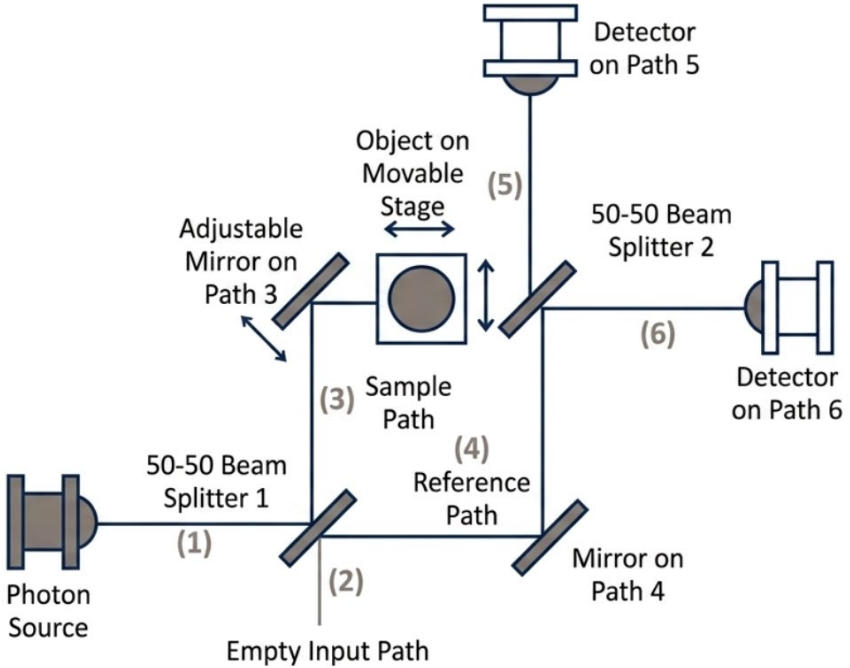
Comparing AI and human scoring

5. Statistics

Validating scores through statistics

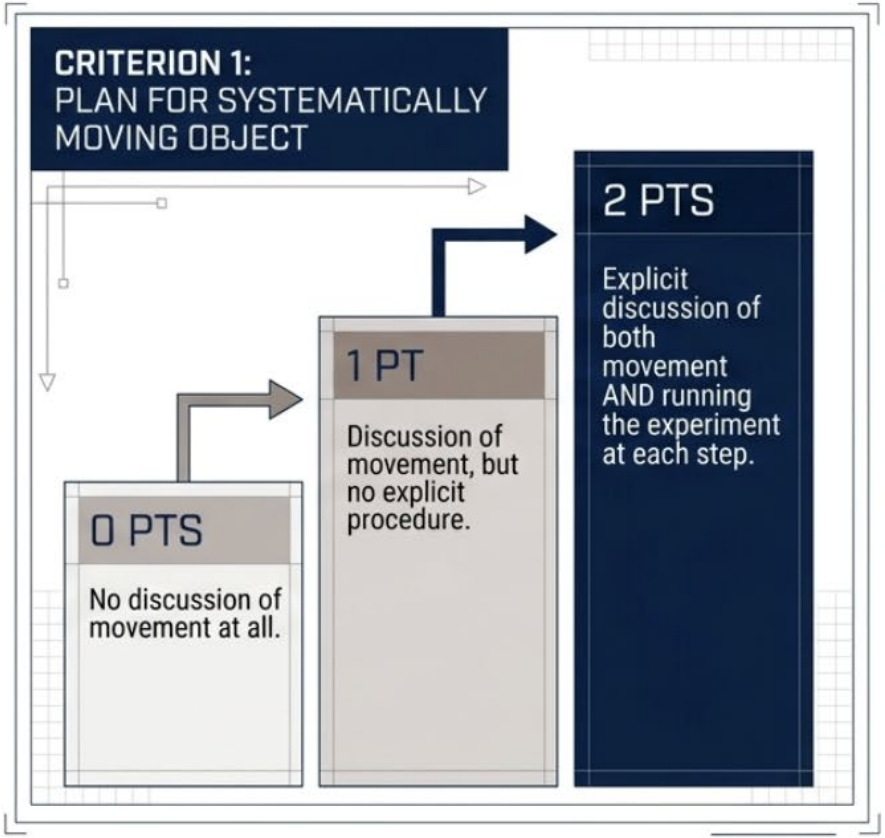
Quantum Mechanics Question

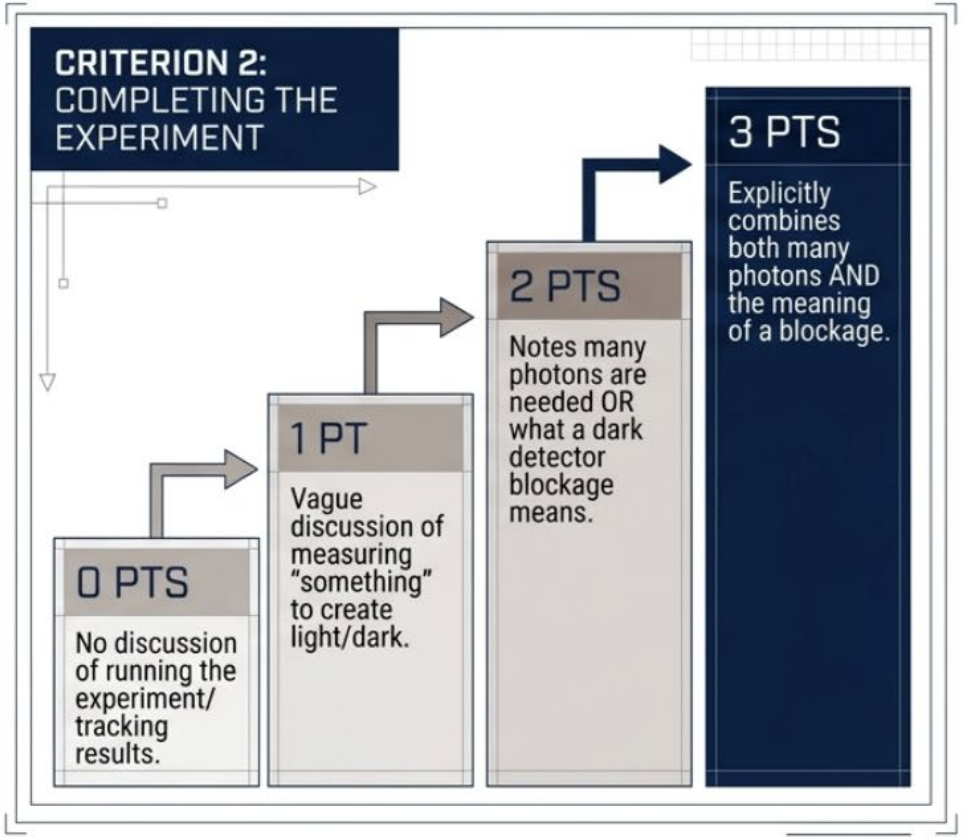
- 131 student responses taken from Mach -Zehnder essay prompt from online quantum mechanics course



Write an essay describing how to measure a silhouette of an object in an interaction -free measurement with a Mach -Zehnder interferometer

Grading Criteria





Roadmap

1. The Prompt Setup

Selecting a conceptually rich task

2. Human Calibration

Defining grading baseline

3. The AI Protocol

Progressively scaffolding AI support

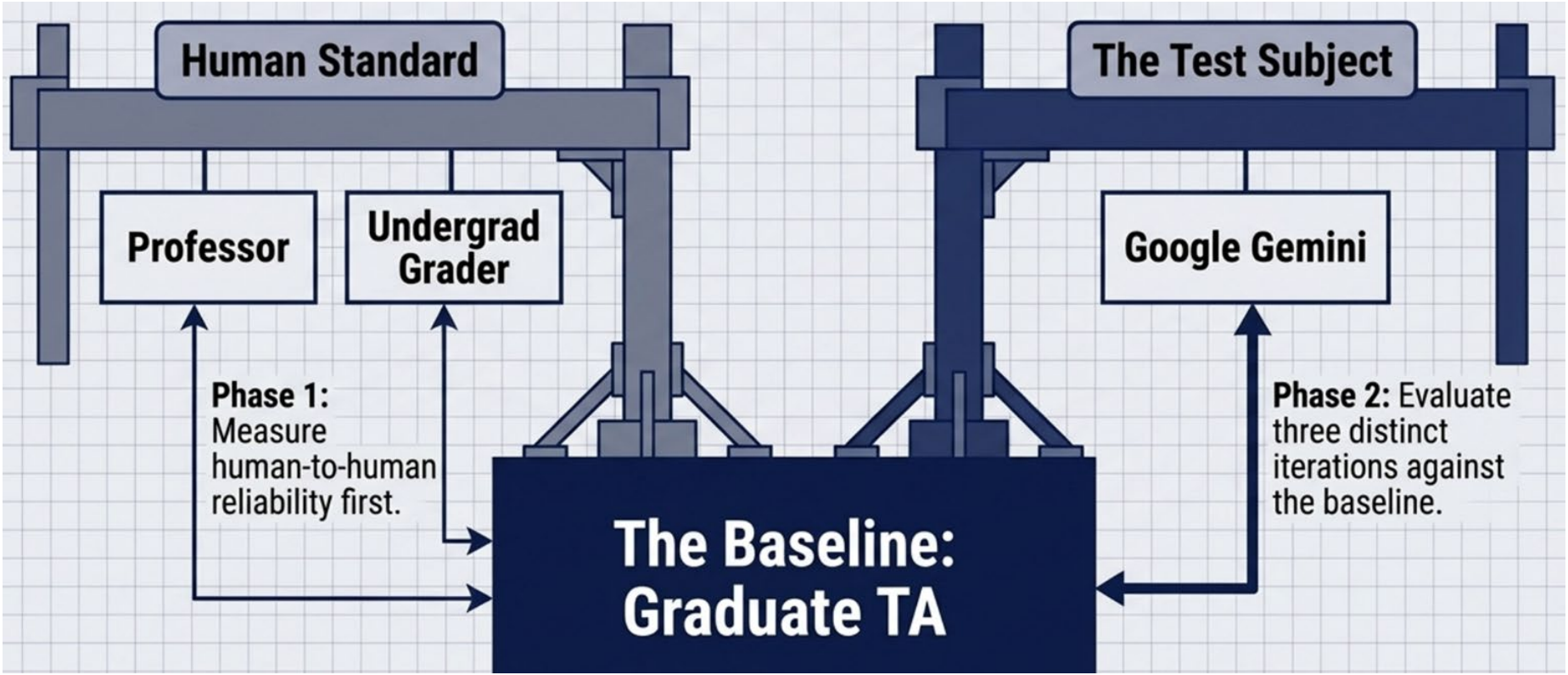
4. Evaluate Framework

Comparing AI and human scoring

5. Statistics

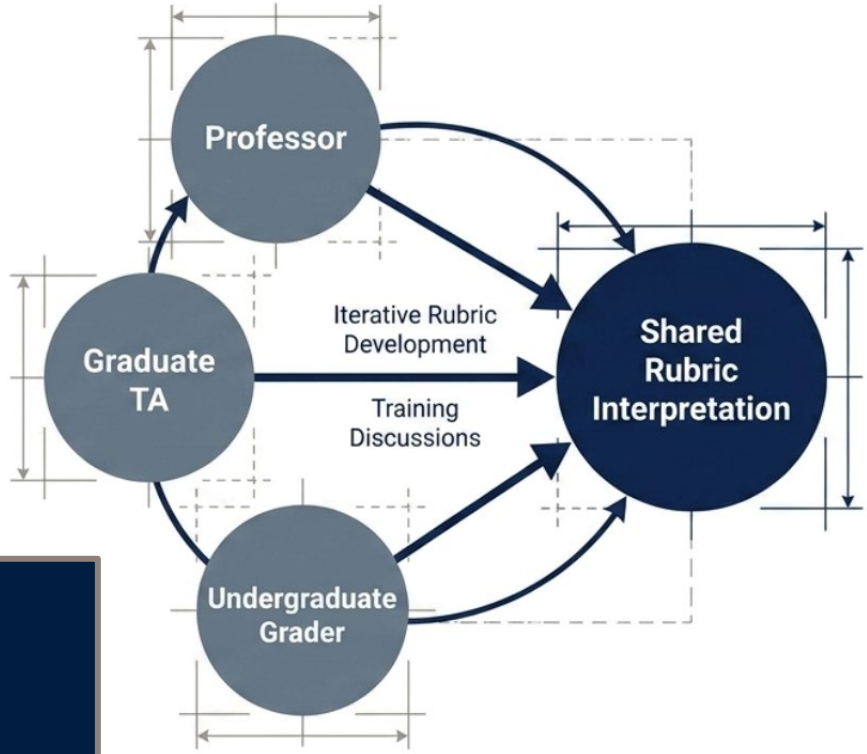
Validating scores through statistics

Graduate TA as Direct Comparison to AI Grader



Building a Shared Grading Consensus

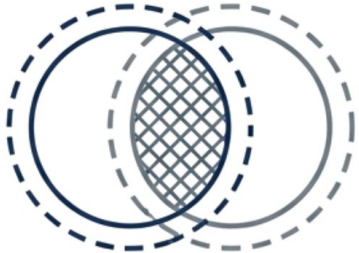
- The research team used 65 student essays during training / calibrating period
- Even human graders required repeated discussions and sample evaluations to produce more consistent interpretations



Key Insight: If humans require training and calibration to grade reliably, so must AI –based grading tools

Statistics Background

Cohen's Kappa



Agreement

- Chance-corrected agreement between two graders on categorial rubric scores

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Mean Absolute Error (MAE)

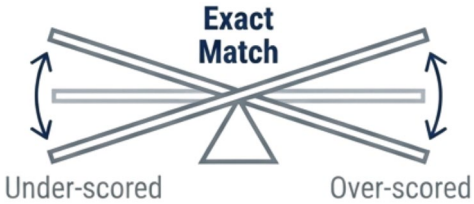


Accuracy

- The average grading difference in absolute rubric scoring units

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|A_n - G_n|$$

Bias Analysis



Directionality

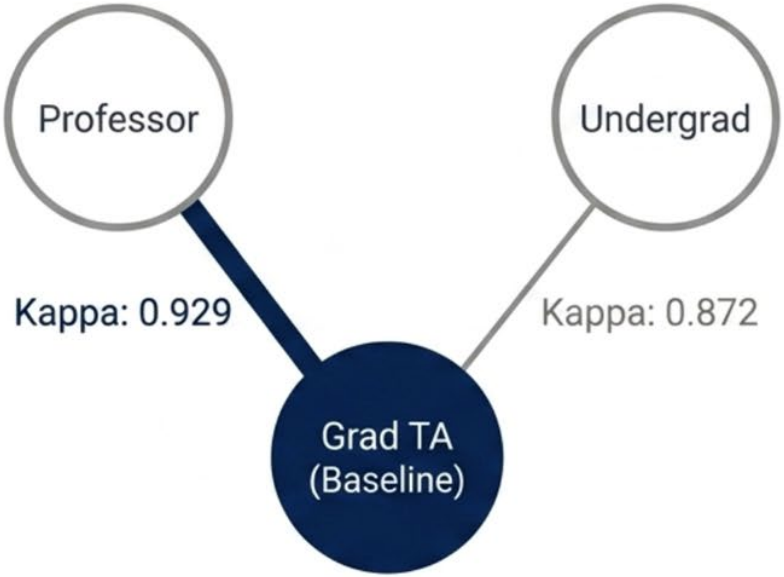
- Measures if grader consistently scores harsher or more lenient than baseline

$A_n > G_n$: over-scored;
$A_n = G_n$: exact;
$A_n < G_n$: under-scored

Cohen, 1960; Willmott & Matsuura, 2005; Linacre, 1989

Human Grading Alignment Check

- After building shared rubric interpretation, the human graders were able to achieve high agreement with $\kappa > 0.800$ and low MAE



	Vs. Professor	Vs. Undergrad
Criterion 1 (Kappa / MAE)	0.901 / 0.099	0.850 / 0.145
Criterion 2 (Kappa / MAE)	0.929 / 0.130	0.872 / 0.244

Roadmap

1. The Prompt Setup

Selecting a conceptually rich task

2. Human Calibration

Defining grading baseline

3. The AI Protocol

Progressively scaffolding AI support

4. Evaluate Framework

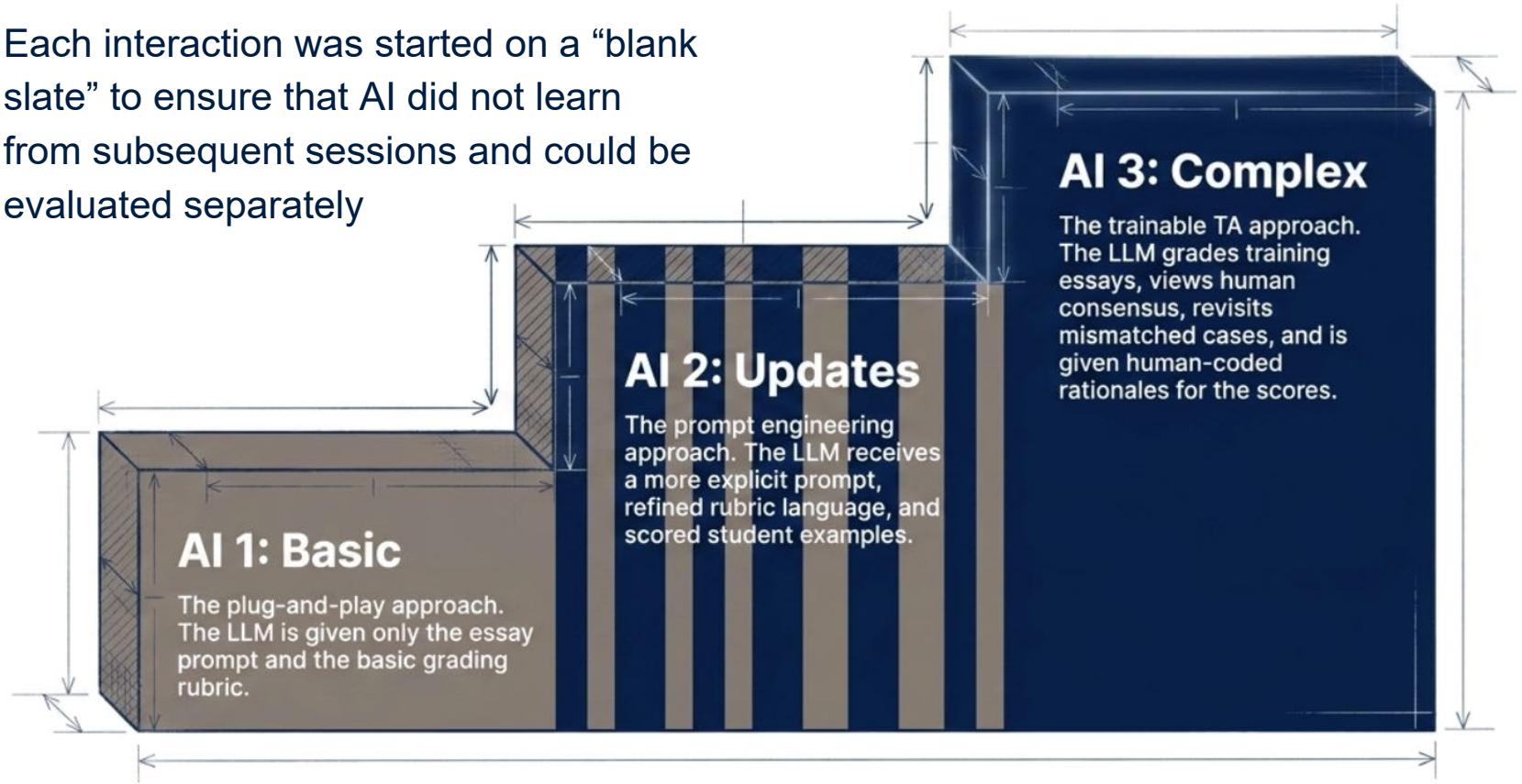
Comparing AI and human scoring

5. Statistics

Validating scores through statistics

Iterative AI Calibration

- Each interaction was started on a “blank slate” to ensure that AI did not learn from subsequent sessions and could be evaluated separately



Iterative AI Calibration

	AI 1: Basic	AI 2: Updates	AI 3: Complex
Inputs	Essay prompt & basic rubric	Explicit prompt, detailed rubric, student examples	Calibration, feedback loops, human coded reasoning
Instructor Effort	Low: Basic Submission	Medium: Prompt Engineering and Examples	High: Calibration and Open Dialogue
State	Initial approach for instructor	Providing LLM more structure for pattern recognition	Dialogue to establish shared interpretative framework
Pedagogical Metaphor	Handing the rubric and walking away	Providing a detailed grading manual with examples	Mentoring and calibrating the new TA with discussions

AI 1: Basic - Prompt

Given is a file is a question from a quantum course and a rubric on how to grade student responses to the question given in the "Rubric" tab. In the "Data" tab is 131 student responses to that question. Grade the essays according to the rubric criteria and return the Excel sheet.

- Input: Basic essay prompt and rubric
- Models a low-effort initial attempt by a solo instructor
- “Upload and pray”

AI 1: Basic - Rubric

Criterion 1: Plan for Systematically Moving Object	
No discussion of movement at all	0
Discussion of movement, but no explicit procedure	1
Explicit discussion of both movement & running the experiment	2
Criterion 2: Completing the Experiment	
No discussion of running the experiment, tracking results, or forming silhouette	0
Vague discussion of measurement to create light or dark detectors	1
Discussion that many photons are needed OR what a dark detector blockage means	2
Discussion that many photons are needed AND what a dark detector blockage means	3

AI 2: Updates

Given is a file is a question from a quantum course and a rubric on how to grade student responses to the question given in the "Rubric" tab, which has two different criteria for a maximum of 5 points. The question is taken from a quantum mechanics course and uses ideas surrounding the Mach -Zehnder interferometer. In the "Data" tab is 131 student responses to that question. Grade the essays carefully according to the rubric criteria provided and return the Excel sheet. Use the examples where human graders gave student essay responses scores to inform you on how to score the new data given in the "Example Data" excel sheet.

- Explicit prompt, detailed rubric, student examples
- Models attempt thinking purely in AI prompt engineering

AI 2: Updates - Scores

Student	Criteria	Agreed Option	Agreed Points
Student 1	Plan for Systematically Moving the Object	Poor	0
	Completing the Experiment	Poor	0
Student 2	Plan for Systematically Moving the Object	Good	2
	Completing the Experiment	Great	3
Student 3	Plan for Systematically Moving the Object	Good	2
	Completing the Experiment	Good	2
Student 4	Plan for Systematically Moving the Object	Poor	0
	Completing the Experiment	Poor	0
Student 5	Plan for Systematically Moving the Object	Good	2
	Completing the Experiment	Good	2

AI 2: Updates - Rubric

Criterion 1: Plan for Systematically Moving Object	
There is no discussion present in the essay that explains that the object in the experiment needs to be moved in any way, including: mentioning scanning, pixels, 1 mm increments, or generally movement. Any discussion of movement of the experiment or object is omitted entirely in the essay OR the answer response does not answer the question being asked and instead the essay response appears to be answering a completely different question OR are blank responses. Only full omission of movement will count as 0 points in this criterion. Essays that only discuss completing the experiment in terms of photons, detectors, and probabilities and does NOT mention any ideas of moving the objection or scanning in the experiment will be awarded 0 points in this criterion.	0
There are vague mentions of movement present in the essay that explains that the object in the experiment needs to be moved in in the experiment or that the experiment needs to be mapped out but does not make explicit mention of how the pixels, data points, or squares of the image is created sequentially, including: mentioning scanning, pixels, or 1 mm increments, etc. The explicit connection can be a few words only needs to be a few words, as mentioned in the example, but needs to address either how the object moves or scans in increments or how the coordinates are built to receive more than 1 point in this criterion. 1 point essays would only contain general ideas of movement and may mention moving, scanning, etc, but does not elaborate on how to build the silhouette through multiple measurements at different positions or scanning through multiple measurements explicitly.	1
There are explicit mentions of movement present in the essay that explains that the object in the experiment needs to be moved in in the experiment of how the pixels, data points, or squares of the image is created sequentially, including: mentioning scanning, pixels, or 1 mm increments, etc. These descriptions do not need to be highly detailed, but need to make explicit mentions of how the object moves or scans in increments or how the coordinates are built to receive full 2 points in this criterion.	2

AI 2: Updates - Rubric

Criterion 2: Completing the Experiment	
There is no discussion present on how to conduct the Mach-Zehnder-like experiment or mention of how the detection of the photons creating the image/ silhouette occurs OR the answer response does not answer the question being asked and instead the essay response appears to be answering a completely different question OR are blank responses. Any essays that only focus on building the silhouette through moving the object or scanning will be awarded 0 points in this criterion. Only the omission of mentioning how the Mach-Zehnder through blockages, mentioning of dark readings or dark detectors, or mentions how the photons are measured will receive 0 points for the criterion.	0
There are vague mentions of the detection or that detection occurs, but there is no explicit mention of how the image is created through a blockage in the sample path resulting in readings or probability of detection increase in the "dark" detector. This may mean the student mentions measurements of the Mach-Zehnder interferometer and mentions that the object blocks the path, but does not explicitly mention either the dark detector or the bright detector in their description in how the dark detector could only receive photons if there is an object blocking the path. Essays that only vaguely mention a blockage or an object blocking the path will receive 1 points for the criterion.	1
The essay makes an explicit mention of detection of the photon mentions 1 of the following: 1) how the image is created through a blockage in the sample path resulting in readings or a probability of detection increase in the "dark" detector, or 2) that the experiment needs to be run enough times to be sure that there are no photons going into the dark detector when there is no blockage due to the increase of probability of the dark detector from 0% to 25% or mentioning of multiple photons or a beam of last light required for the experiment. Only 1 of these two ideas needs to be explicitly mentioned to receive 2 points. These ideas gives more context than the 1 point criterion where students only mention a blockage, while both of these points gives more context that the students understand the underlying mechanisms of the Mach-Zehnder interferometer. For either of these 2 ideas, the students only need to explicitly mention it in a few words or sentences, but make explicit mention of either multiple photons (through use of "photons", "beam", multiple measurements, or connections to probability being 25% chance for the dark detector.	2
The essay makes an explicit mention of detection of the photon mentions both of the following: 1) how the image is created through a blockage in the sample path resulting in readings or a probability of detection increase in the "dark" detector, or 2) that the experiment needs to be run enough times to be sure that there are no photons going into the dark detector when there is no blockage due to the increase of probability of the dark detector from 0% to 25% or mentioning of multiple photons or a beam of last light required for the experiment. For 3 points for the criterion, students make explicit mention of both of these ideas; these explanations could be just a few words or sentences, but the focus that the students only need to explicitly mention both multiple photons (through use of "photons", "beam", multiple measurements, or connections to probability being 25% chance for the dark detector and mention the dark detector explicitly and how if there are photon readings, it must mean there was a object blocking the path (only focusing on the dark detector is required).	3

AI 3: Complex - Prompts

First, I will give a question from a quantum mechanics course to write an essay with a detailed rubric. Analyze the rubric and its information in great detail first, then ask me to proceed.

Now, given is example data on student responses plus what human graders agreed is an appropriate grade based on the rubric provided. There are 65 student responses here. Grade these essays yourself based on the rubric provided.

Compare to what was determined as the appropriate grade by human graders. Refine your grading criteria based on the discrepancies between any mismatched scores and then regrade the essays. Show your percentage alignment for the first two passes and give descriptions on how you changed your grading parameters.

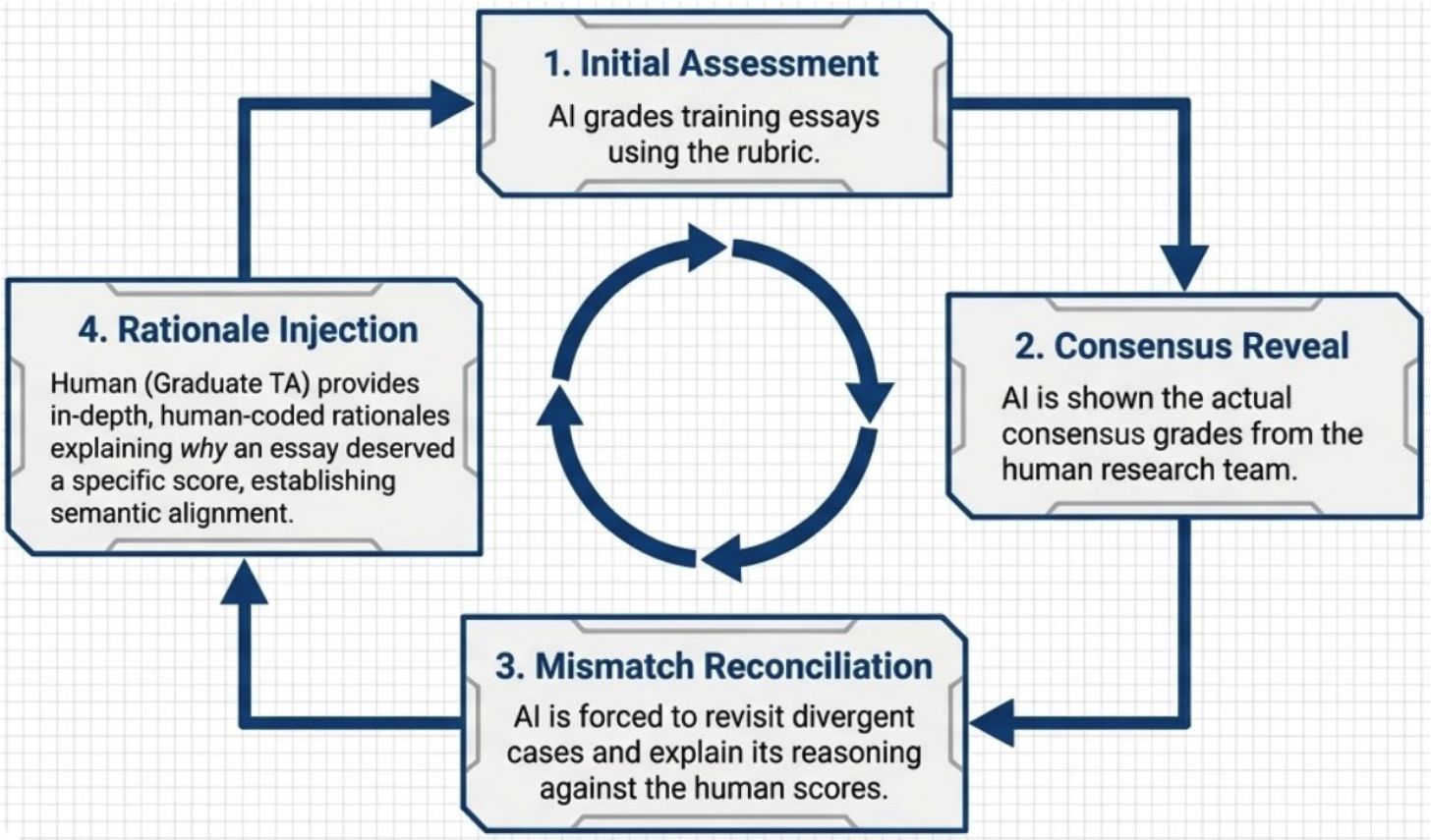
- Input: Many. Conversational. Calibration. Provide Human Coded Reasoning

AI 3: Complex - Scores

Student	Criteria	Agreed Option	Agreed Points
Student 2	Plan for Systematically Moving the Object	Good	2
	Completing the Experiment	Great	3

Plan for Systematically Moving the Object	2	From this sentence alone: "We can have this apparatus move about, in an upward, downward, left and right direction, for a distance of 1 mm, and each time we do so, we monitor the detectors. If a light photon hits the dark detector, then the point at which the object is has an obstruction." It is clear that the student understands how to move the object apparatus on the stage in increments if there is a blockage or obstruction. **Even from these words like "upward, downward, left, and right direction" would have been enough to signal scanning, or pixel creating.**
Completing the Experiment	3	For this criteria, it is clear that the student knows there is a connection between the blockage and photons being detected in the dark detector. Example quote: "that if a photon enters the dark detector, there is an obstruction present." **This is already enough for 2 out of the 3 points alone.** Then, the student says that when there is a blockage, the probability of dark detection increases, and then consistently uses language with plural of photon using "photons" or when "a photon (instead of the photon)" indicating that there **should be more than 1 photon for the experiment.** For example: "Then, if there is an obstruction on the sample path, the probability of entering the dark detector increases. " While it is not clear if the student is using the plural form of "photon" to mean multiple measurements in the same location or just multiple measurements or photons through the scanning process, the graders have given this response a 3 in this criteria.

AI 3: Complex - Summary



Roadmap

1. The Prompt Setup

Selecting a conceptually rich task

2. Human Calibration

Defining grading baseline

3. The AI Protocol

Progressively scaffolding AI support

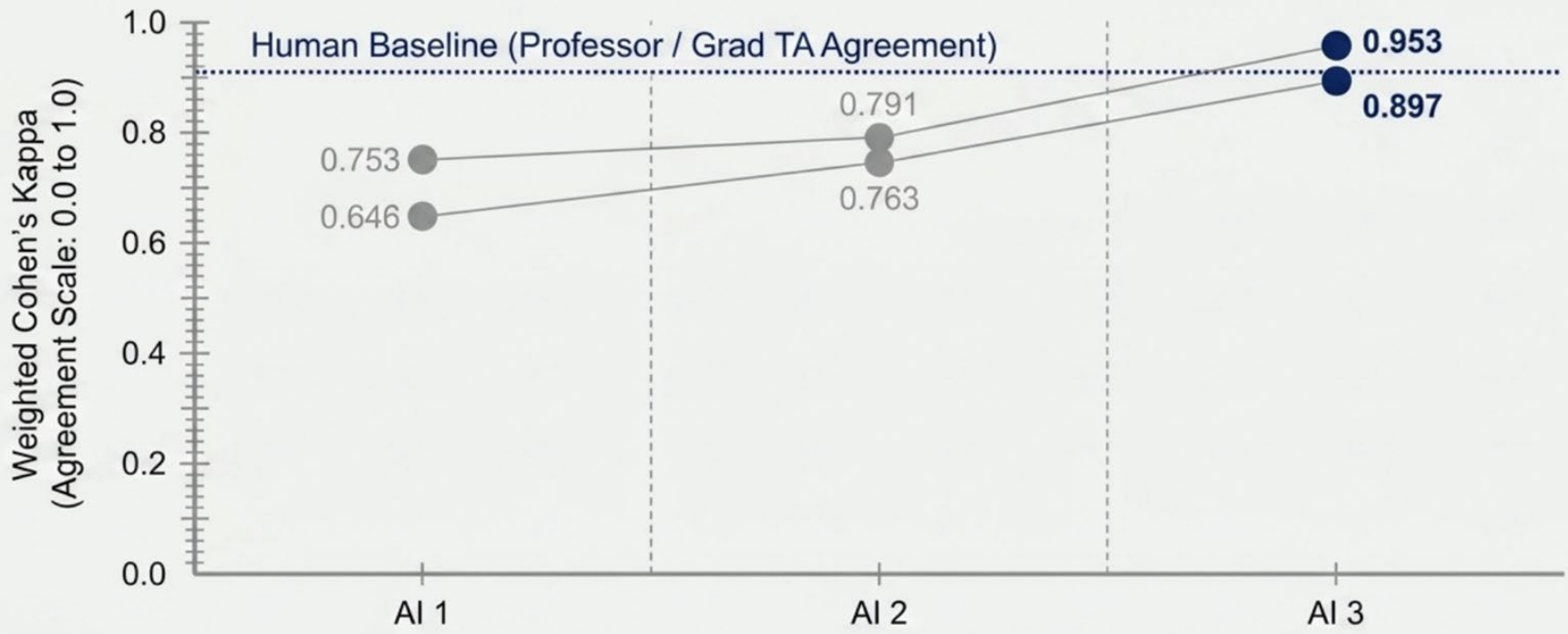
4. Evaluate Framework

Comparing AI and human scoring

5. Statistics

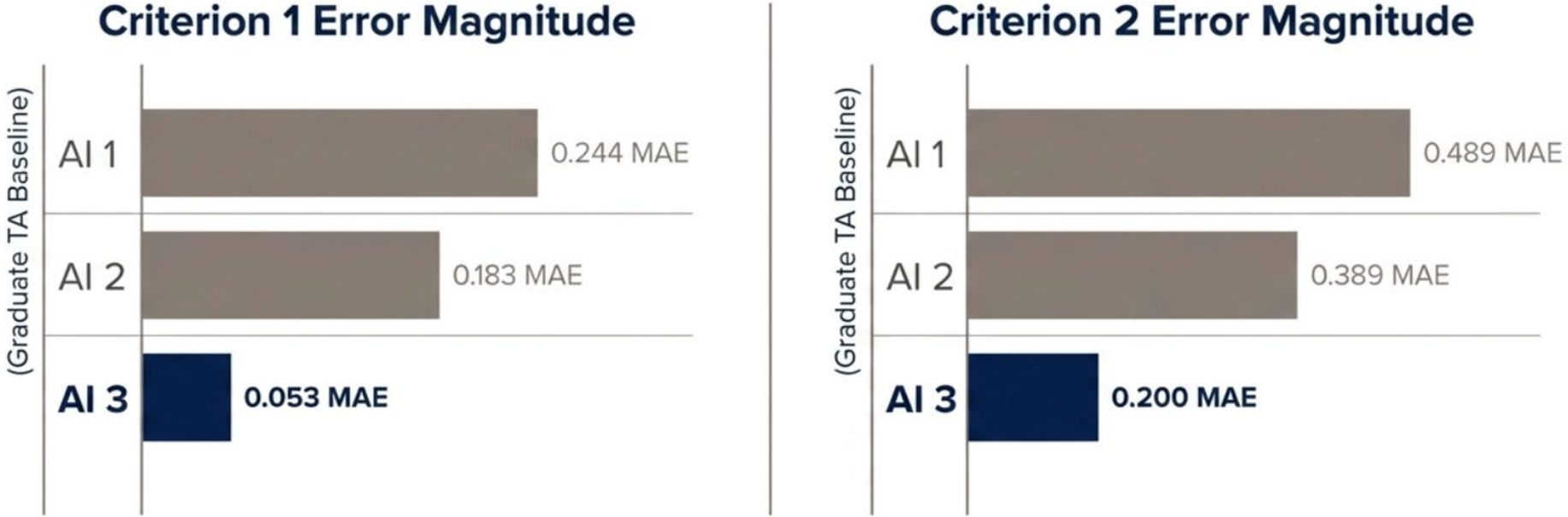
Validating scores through statistics

Cohen's Kappa



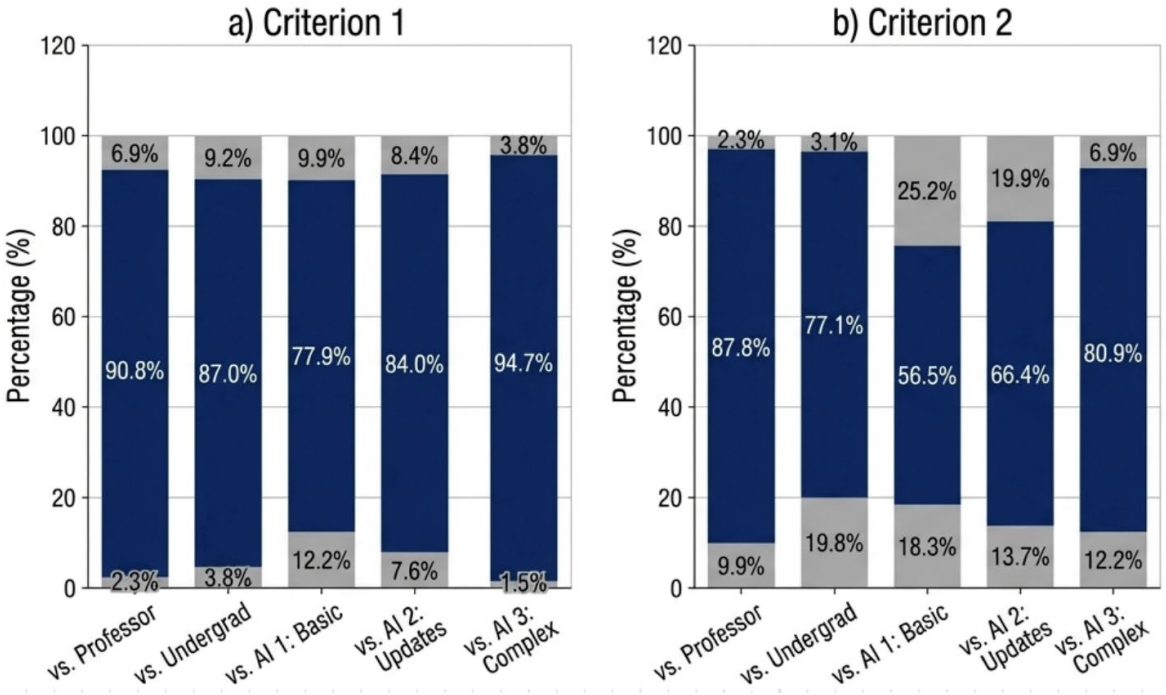
- High agreement only comes in AI 3 iteration with high human -AI interaction

Mean Absolute Error



- Accuracy goes to as low as ~1% difference to baseline score

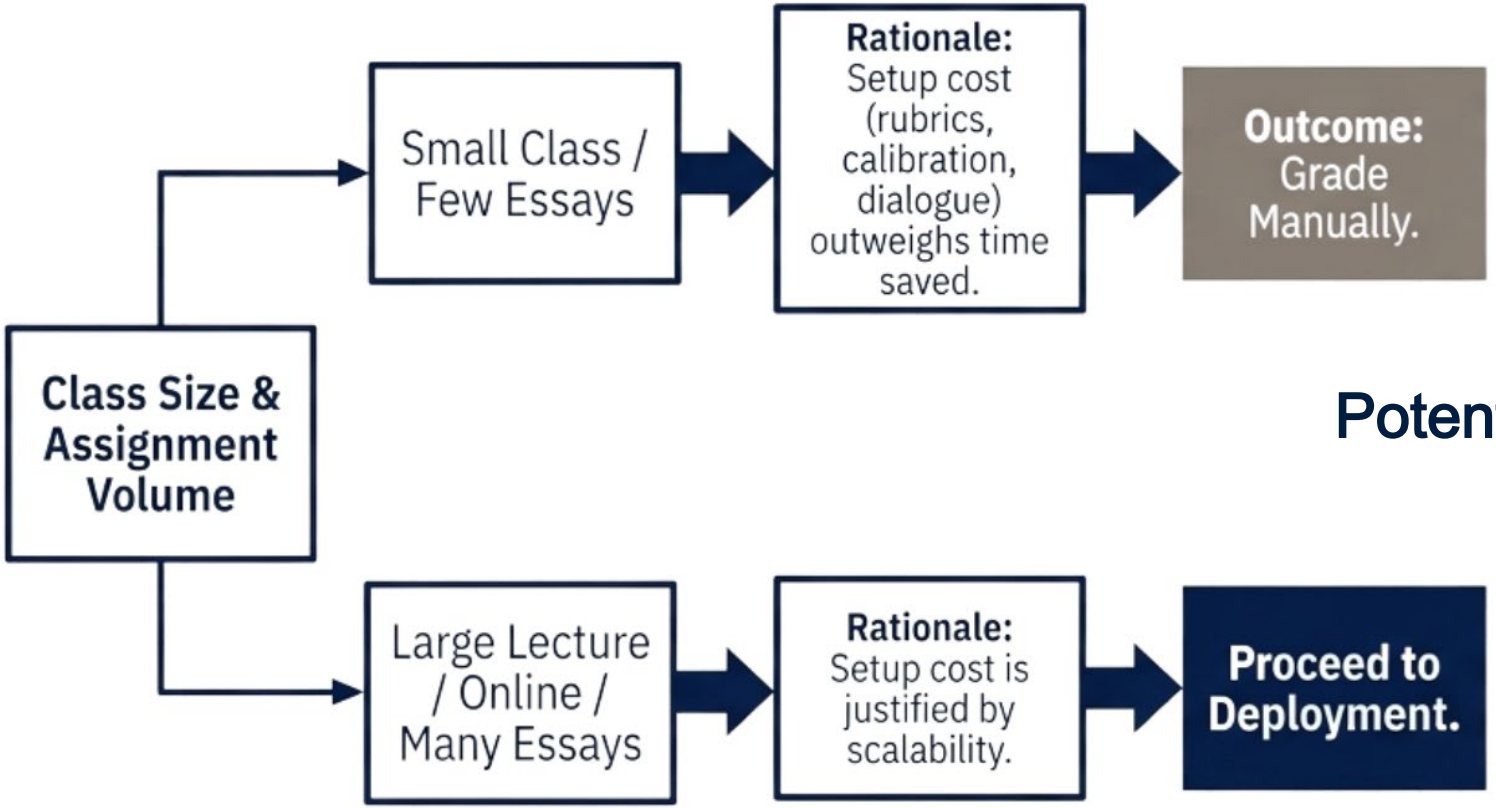
Bias Analysis



- Bias analysis shows that there still may be some significant amount of over - or under - grading occurring but AI bias is comparable to human bias

Key Insight: AI bias may be driven by the inherent biases by the human training it

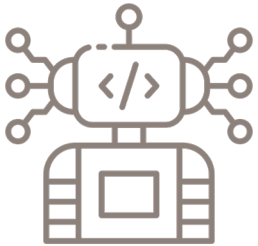
Should You Use AI to Grade Physics Essays?

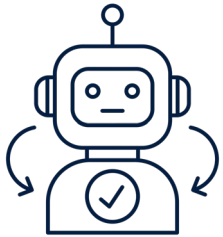


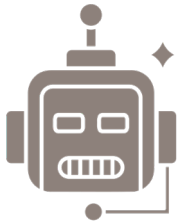
Potentially.

Limitations to Study

- Limited to just *Google Gemini*
  - Test other LLMs like *ChatGPT*
- Focus on only 1 essay question
  - Try with up to 3 -5 total essay prompts
- Did not focus on creation of AI agents or feeding course materials for better alignment
  - Try to create course specific AI agents through AI models like *Ollama* or *NotebookLM*







Key Takeaways from AI Iterations

01

Rubrics are not self-executing.

Even perfectly constructed rubrics fail without a shared interpretative framework. Words on a page are insufficient for both novice humans and AI.

02

Prompt engineering is insufficient.

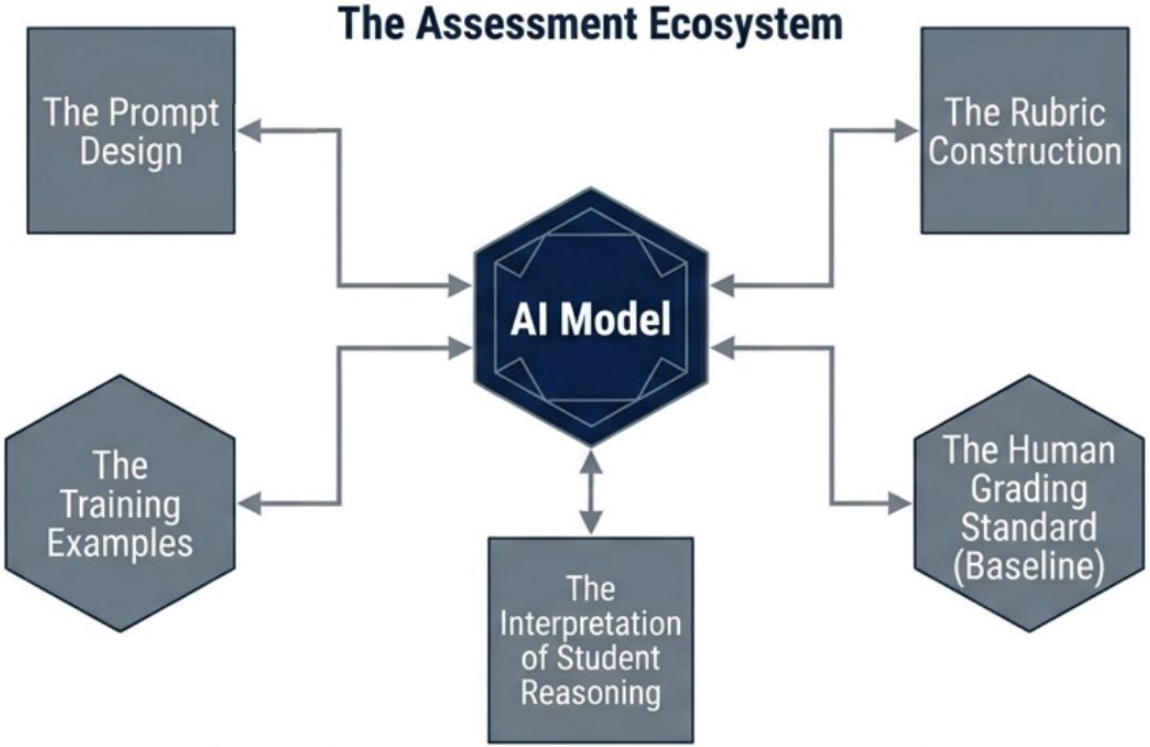
Simply adding examples and explicit, structured language (AI 2) yields only marginal improvements in grading reliability for complex reasoning.

03

Mentorship scales.

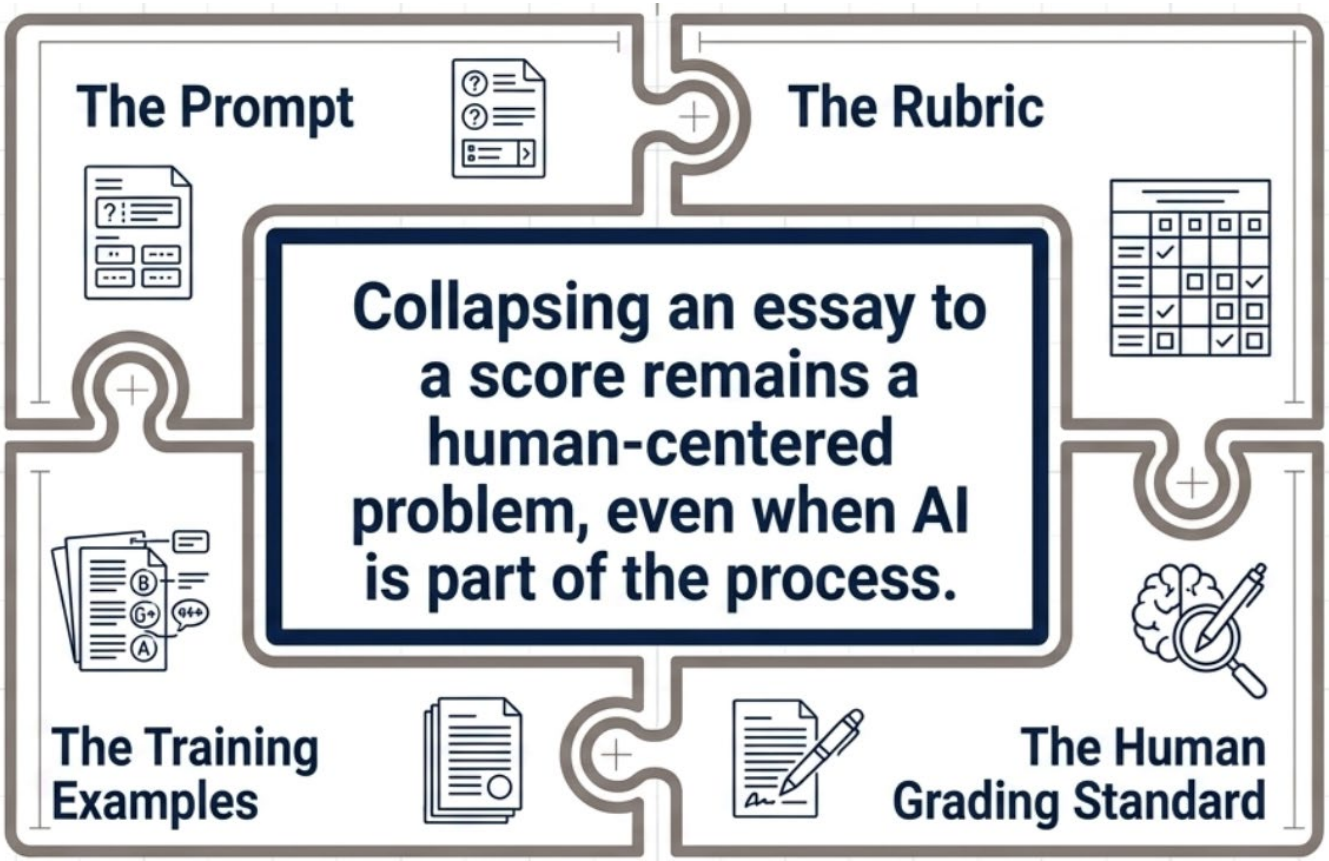
Intensive, conversational calibration (AI 3) is the missing link. Treating the AI like an intellectual partner rather than a software tool yields expert-level reliability.

Key Takeaways from AI Iterations



- AI grading assessment must be evaluated via the entire pedagogical ecosystem, as we should think of AI more like a trainable teaching assistant, and less like an omnipotent blackbox

Key Takeaway



References

Google 's *NotebookIm* - For slide inspiration, figures, and drawings

1. J. L. Docktor, J. Dornfeld, E. Frodermann, K. Heller, L. Hsu, K. A. Jackson, A. Mason, Q. X. Ryan, and J. Yang, "Assessing st ude nt written problem solutions: A problem -solving rubric with application to introductory physics," Physical Review Physics Education Research, 2016.
2. L. Doughty and M. D. Caballero, "Rubric design for separating the roles of open -ended assessments," Physics Education Research C onference Proceedings, 2014.
3. M. Scott, T. Stelzer, and G. Gladding, "Evaluating multiple -choice exams in large introductory physics courses," Physical Revie w Special Topics - Physics Education Research, 2006.
4. D. R. E. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of ChatGPT," In novations in Education and Teaching International, 2024.
5. E. Kasneci, K. Sebler, S. Kuchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Gunnemann, E. Hullermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, an d G. Kasneci, "ChatGPT for good? On opportunities and challenges of large language models for education," 2023.
6. G. Kortemeyer and J. Nohl, "Assessing confidence in AI -assisted grading of physics exams through psychometrics: An exploratory s tudy," Physical Review Physics Education Research, 2025.
7. G. Kortemeyer, J. Nohl, and D. Onishchuk, "Grading assistance for a handwritten thermodynamics exam using artificial intellig enc e: An exploratory study," Physical Review Physics Education Research, 2024.
8. M. A. Stellmack, Y. L. Konheim -Kalkstein, J. E. Manor, A. R. Massey, and J. A. P. Schmitz, "An assessment of reliability and val idity of a rubric for grading APA -style introductions," Teaching of Psychology, 2009.
9. M. Moore and D. F. Thomas, "Inter - and intra -rater consistency: Armies of graduate TAs grading in first year,"Discussions on Uni versity Science Teaching: Proceedings of the Western Conference on Science Education, 2017.
10. Passonneau, Rebecca J. , "AI for Coding Education Meta -analyses: An Open -Science Approach that Combines Human and Machine Intell igence," Artificial Intelligence in Education Technologies: New Development and Innovative Practices: Proceedings, 2023.
11. J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," Psychological Bullet in, 1968.
12. C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assess ing average model performance," Climate Research, 2005.
13. J. M. Linacre, "Many -Facet Rasch Measurement," MESA Press, 1989.

Future Projects

- Continue to determine AI interactions that secure best agreement, accuracy, and improve bias directionality
- Try with other *LLMs* like *ChatGPT*, *Ollama* , or *NotebookLM*
- Conduct consistency trials over different physics free responses

Slides

Collaborators: Timothy Swanson, James Freericks and Leanne Doughty



