

Grading Assistance for a Handwritten Thermodynamics Exam using Artificial Intelligence: An Exploratory Study

Gerd Kortemeyer*

Rectorate and AI Center, ETH Zurich, 8092 Zurich, Switzerland †

Julian Nöhl‡

Institute of Energy & Process Engineering, ETH Zurich, Tannenstrasse 3, 8092 Zurich, Switzerland

Daria Onishchuk§

Unit for Teaching and Learning, ETH Zurich, 8092 Zurich, Switzerland

(Dated: June 27, 2024)

Using a high-stakes thermodynamics exam as sample (252 students, four multipart problems), we investigate the viability of four workflows for AI-assisted grading of handwritten student solutions. We find that the greatest challenge lies in converting handwritten answers into a machine-readable format. The granularity of grading criteria also influences grading performance: employing a fine-grained rubric for entire problems often leads to bookkeeping errors and grading failures, while grading problems in parts is more reliable but tends to miss nuances. We also found that grading hand-drawn graphics, such as process diagrams, is less reliable than mathematical derivations due to the difficulty in differentiating essential details from extraneous information. Although the system is precise in identifying exams that meet passing criteria, exams with failing grades still require human grading. We conclude with recommendations to overcome some of the encountered challenges.

I. INTRODUCTION

A. AI-supported Grading

The recent advancements in artificial intelligence (AI) have ushered in new possibilities in various domains, including education. At its public appearance in fall 2022, Generative Pre-trained Transformer (GPT) [1] already demonstrated remarkable capabilities. Beyond the hype-wave triggered by the human appearance of its responses [2], it proved proficiency in academic domains such as physics, where it and later releases (in particular GPT-4 [3]) passed standardized exams and introductory courses at impressive levels [4][9]. AI starts to get integrated into physics education [10][11], where in addition to solving problems, it for example shows promise in constructing new physics problems [11][12], making physics materials accessible to blind readers [13], and supporting physics educational research efforts [14][15].

The grading of physics exams has traditionally required deep analytical skills to assess not just the final answers but the problem-solving process itself, a task that involves evaluating logical, conceptual, and mathematical competencies [16][20]. While certain aspects of problem solutions can be graded by computers, a thorough evaluation often necessitates human judgment, particularly when diverse solution paths and potential errors are involved [21][30]. Such exams usually require handwriting solution paths, since in exam situations, typesetting

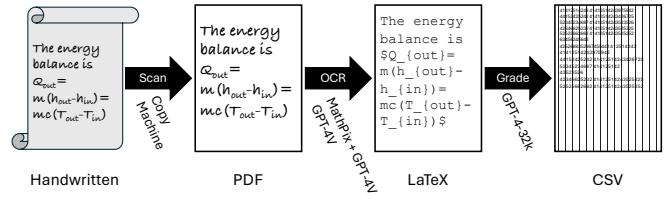


FIG. 1. Workflow of exam grading, starting with scanning the exam sheets, making them machine-readable with Mathpix [45] and GPT-4V [3], and grading them using GPT-4-32k.

those equations would take up an inordinate amount of time for the students, and it would necessitate a skillset outside the learning objectives of most physics courses.

AI holds potential as a scalable solution for giving feedback on open-ended responses [31], as well as grading, or at least classification [32] and pre-grading, by leveraging its ability to process large volumes of data and its emerging capability to understand and evaluate complex student responses [33][41]. In a recent study using a synthetic data set, an agreement of $R^2 = 0.84$ could be found between human and AI grades [4], however, the synthetic nature of the investigated solution derivations circumvented many of the intricacies that are present in authentic exam situations. Most notably, challenges are introduced by the complexity of handwritten mathematical expressions, which add another layer of difficulty due to the intricacies of Optical Character Recognition (OCR) technology [42][44]. Figure 1 shows the workflow that we investigated, using a large-scale high-stakes physics exam. We focus on the operational challenges involved, as well as the limitations and capabilities of current technology to augment traditional grading methods.

* kgerd@ethz.ch

† also at Michigan State University, East Lansing, MI 48823, USA

‡ jnoehl@ethz.ch

§ donishchuk@ethz.ch

B. Large Language Models and Randomness

Large Language Models (LLMs) like those based on the architecture introduced by Vaswani et al. [46] use probabilistic methods to generate responses. They create sequences of so-called tokens (in the English language similar to syllables). Even responses to clear-cut physics-related prompts can vary strongly [7][41], influenced by a temperature parameter T . A lower T results in more predictable outputs, while a higher T increases creativity. At absolute zero ($T = 0$), a language model consistently selects the most probable next token, making its outputs highly predictable though not necessarily accurate due to potential biases or errors in the training data [46][47]. Conversely, a higher T can lead to less reliable, “hallucinated” responses (“hallucination” in this context refers to a phenomenon where the model generates output that is not grounded in the input data or reality). In this study, the default setting was $T = 0.7$, except for OCR of complete pages, where it was set to $T = 0.5$. Our study assumes that grading by LLMs should be viewed probabilistically. Multiple runs yield average scores, with standard deviations providing a measure of confidence. For instance, a score of 14 ± 5 suggests lower reliability than 14 ± 2 , although the score could still be incorrect.

II. SETTING

A. Institution

ETH Zurich is a technical university with approximately 25,000 students from 120 countries, with about one-third identifying as female. Admission is highly selective for international students, yet unrestricted for anyone holding a Swiss high school diploma. Most undergraduate courses are taught in German. In the German-speaking universities’ academic tradition, ETH Zurich emphasizes summative assessments at the end of courses, rather than smaller assessments throughout the course. The study was approved by the institution’s Ethics Committee (IRB) under 2023-N-286.

B. Cloud Infrastructure

Access to OpenAI models was provided through Azure AI Services [48], where ETH Zurich has a contract which assures processing on Swiss data centers, with a consumption-based per-token payment structure. Based on our subscription, MathPix was likely processing in the United States, which is why we used pseudonyms.

C. Exam

We considered a high-stakes exam for engineers on thermodynamics, dealing with standard topics of energy,

exergy, entropy, and enthalpy. Students had 15 minutes to prepare by reading the problems and two hours to complete the work. They could use provided reference materials and a non-programmable calculator. They needed to provide handwritten solutions including derivations, using permanent pens and scribbling out anything they did not want graded. A total of 252 out of 434 students taking the exam agreed to participate in the study.

Problem 1 involves a reactor in steady-state operation, including liquid in- and outflow, a chemical reaction with associated heat generation, and a cooling jacket. The problem parts are:

- a) Calculation of the heat transfer to the cooling fluid (2.5 points).
- b) Determination of the thermodynamic mean temperature of the cooling fluid (2.5 points).
- c) Calculation of the entropy production due to heat transfer (1.5 points).
- d) After steady-state operation, the outlet flow is stopped. Calculation of the water mass required for cooling the reactor from its operating to a lower temperature using an energy balance (4.5 points).
- e) Determination of the change in entropy of the reactor contents between the initial and cooled states (4 points).

Problem 2 explores the operation of an aircraft engine, consisting of reversible or irreversible compressors, turbines, nozzles and heat addition. The problem parts are:

- a) Drawing of the engine process qualitatively in a T-s diagram, marking relevant states (8 points).
- b) Determination of the exit speed and temperature of the aircraft engine (4 points).
- c) Calculation of the specific exergy increase between two states (3.5 points).
- d) Calculation of the specific exergy loss related to the mass flow rate (3.5 points).

Problem 3 involves a hot gas and a solid-liquid system in an isolated cylinder separated by a heat-transferring membrane. The problem parts are:

- a) Calculation of the initial pressure and mass of the gas in the cylinder (5.5 points).
- b) Determination of the temperature and pressure of the gas after an equilibrium state has been reached through heat transfer (2 points).
- c) Calculation of the heat transferred from the gas to the ice-water mixture (2.5 points).
- d) Calculation of the ice content in the second state (7 points).

Problem 4 describes a two-step freeze-drying process for food preservation. The problem parts are:

- a) Drawing of the freeze-drying process in a p-T diagram, including labeled phase regions (4.5 points).
- b) Determination of the required mass flow rate of the refrigerant R 134a (6 points).
- c) Determination of the vapor fraction of the refrigerant immediately after throttling (4 points).
- d) Calculation of the coefficient of performance for the cooling cycle (3 points).
- e) Discuss how the temperature inside the freeze-dryer changes if the cooling cycle continues unchanged (1 point).

Figure 2 shows a typical example for a handwritten solution to Problem 1 (the black box on top resulted from redacting the student name). Where the student noted “TAB A-2,” he or she referred to the data table that was available during the exam. Figure 3 shows another, less well readable example of a solution for Problem 3d.

A feature of the German-tradition university system is that exams can be passed or failed (though, at ETH Zurich, for some exams this gets decided on the level of so-called exam blocks with compensation mechanisms), and a notion among some students in the beginning semesters is that they do not really care about the grade as long as they pass the exam and can move on with their studies. Due to the high-stakes nature of these exams, every exam problem is graded by two teaching assistants.

The exam was graded on a fine-grained rubric by the TAs, awarding points for each solution step. The course personnel creating and correcting each task defines sub-points which a student should at least get to pass the exam. These points should show a basic understanding of Thermodynamics. The sum of the points is used to define the lower boundary for the “passing grade” of 4.0. The value for the highest grade of 6.0 is based on values from previous exams (% of total points). However, slight adjustment to the mapping of points to grades can be made to coincide with gaps in the point distribution. As a result, the nominal level of passing the exam was set at 26 out of the 65 available points (since there are compensation mechanisms, we use the term “nominal”). If the average grades of a student are below a certain threshold so that he or she needs to repeat the semester, the exam was regraded by a third teaching assistant.

III. METHODOLOGY

A. Workflow and Study-specific Additional Steps

Two important directives for this study were non-interference with the regular exam process and preservation of student privacy. This meant that students were

not given any instructions on how to prepare their work for better OCR interpretation and AI-grading, which brought about additional challenges and problems that had to be fixed manually. Grading personnel was not informed which students had given informed consent to participate in the study. After the exam packages had been turned in, the investigators separated the consent forms from the exam sheets and scanned only the solutions of the students who had agreed to participate in the study. The scanner device had automatic paper handling capabilities (as it turned out, it also automatically switched between black-and-white mode (see Fig. 2 as an example) and color-mode (see Fig. 3), and it automatically skipped empty pages); the device then emailed the PDFs to the investigators.

The workflow included a pseudonymization step that would not be present in production use. The investigators and the grading personnel initially shared a key, where each exam package had a number. One of the investigators (GK) redacted all names from the scanned solutions and inserted the exam-package number before any further processing. As exam sheets within the package were often out-of-order, at this step, the PDF-pages were manually sorted, and clear-text markers were inserted to separate the problems. The latter was hampered by students frequently only writing letters like “c)” at the start of a new problem part, but not the problem number; if the sheets were out of order and had no page numbers written, this meant figuring out from the contents if this was Part c of Problem 1, 2, 3, or 4.

Students had brought their own paper to the exam, which was frequently repurposed and included company letterheads with logos or advertising or had pre-printed headers like “My Notes” or “While you were out.” While these page augmentations would not interfere with the TA-grading, they had to be redacted by hand, as they would otherwise be transcribed by the OCR process and appear in the middle of solution derivations. Some students also used their paper in a multi-column landscape layout; these pages appeared portrait-oriented with 90 or 270 degree rotated writing, and had to be rotated into landscape format by hand. Other students turned their two-sided sheets along the short instead of the long side when continuing work on a problem, which resulted in every other page being upside down.

During the OCR step, the exam-package number and the problem-separation markers were transcribed into the LaTeX, since they simply appeared as plain text. Subsequent scripts would pick up those markers from the LaTeX for processing. Grading personnel eventually reported their grading decisions to the investigators in terms of the exam-package number for comparison.

B. Optical Character Recognition (OCR)

Figure 2 shows an example for a fairly readable solution, while Fig. 3 provides an example for less clear

Aufgabe 1.

a) \dot{Q}_{aus} :

Energiebilanz um reaktor Flüssigkeit

$$\dot{m}(\text{hein} - \text{haus}) + \dot{Q}_{aus} + \dot{Q}_n = 0$$

$$0.3(292.98 - 428.23) + \dot{Q}_{aus} + 100 = 0$$

$$\dot{Q}_{aus} = 0.005 \cdot 2257 + (1-0.005) \cdot 419.04 = 428.23 \frac{\text{kg}}{\text{h}}$$

$$\dot{Q}_{aus} = 59.425$$

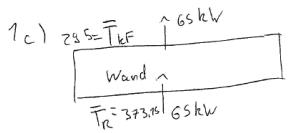
$$\text{b) } \bar{T}_{hf} = \frac{\int_{\text{ein}}^{\text{aus}} T ds}{\dot{m}_{aus} - \dot{m}_{ein}} = \frac{U_{aus} - U_{ein}}{c_f \ln\left(\frac{T_{aus}}{T_{ein}}\right)}$$

stoff model

$$\Delta s = c_f \cdot \ln\left(\frac{T_{aus}}{T_{ein}}\right)$$

$$\Delta u = c_f (T_{aus} - T_{ein})$$

$$\bar{T} = \frac{T_{aus} - T_{ein}}{\ln\left(\frac{T_{aus}}{T_{ein}}\right)} \approx 293.12 \text{ K}$$



Entropie bilanz um Wand:

$$\dot{s}_{erz} = 0.04615 \frac{\text{kW}}{\text{K}} = 416.15 \frac{\text{W}}{\text{K}}$$

$$0 + \frac{65 \text{ kW}}{373.15 \text{ K}} - \frac{65 \text{ kW}}{293 \text{ K}} + \dot{s}_{erz} = 0$$

$$\text{d) } \dot{m}_R = 0 \quad \begin{array}{l} \text{Da } iF: \\ h \neq h_F \\ v = v_F \end{array}$$

$$T_{n2} = 70^\circ\text{C} \quad T_{n1} = 100^\circ\text{C}$$

$$T_{ein} = 20^\circ\text{C}$$

$$Q_{aus} = 35 \text{ MJ}$$

Energie bilanz halboffenes system:

$$\Delta u = \Delta m_{ein} \cdot (hein) + Q_j$$

$$\cancel{m_2 u_2 - m_1 u_1} = \Delta m_{ein} \cdot (hein) + Q_{aus}$$

$$m_2 = m_1 + \Delta m_{ein}$$

$$m_1 = 5755 \text{ kg}$$

$$v_1 = 2506.5 \cdot 0.005 + (1-0.005) \cdot 419.04 = 428.38$$

$$v_2 = v(70^\circ\text{C}) = 292.95$$

$$hein = h(20^\circ\text{C}) = 83.96$$

$$(5755 + \Delta m) \cdot 292.95 - 5755 \cdot 428.38 = \Delta m \cdot 83.96 - 35 \cdot 10^3$$

-298

$$-750154 = -208.84 \Delta m$$

$$\Delta m = 3589 \text{ kg}$$

$$\Delta s_{12} = s_1 - s_2 = m_1 s_1 - m_2 s_2$$

$$m_1 = 5755 \text{ kg}$$

$$m_2 = 9355 \text{ kg}$$

$$s_1 = 0.005 \cdot 7.3549 + \cancel{0.948 - 1.3} \quad 0.985 \cdot 1.3069 = 1.33714$$

$$s_2 = 0.9549$$

$$\Delta s_{12} = -1237.85 \text{ kJ}$$

FIG. 2. Example of a handwritten solution for Problem 1. The handwriting itself is fairly clear, but as students were instructed to simply scribble out errors, this solution, like many others, contains several scratched out symbols and expressions.

handwriting. To render these solutions machine-readable for LaTeX processing, two approaches were investigated, using scripts and the APIs of the respective systems:

1. MathPix plus GPT-4V

The first method employed MathPix [45] for the preliminary conversion of handwritten content into LaTeX. MathPix performs a robust and rather deterministic interpretation of handwriting and mathematical expressions, and it does not hallucinate, but it is also designed

not to guess when confidence is low. Instead of producing textual output, for segments of a document where any interpretation would be low-confidence, or in situations where graphs, figures, or sketches were identified, MathPix incorporates a cropped JPG image into the LaTeX document via the `\includegraphics` command. While usable in many other scenarios, the grading step needs a fully textual representation, where at least some interpretation of unclear handwriting is provided, and where graphical content is described in text (the latter is crucial for Problem 2a and Problem 4a, where diagrams are expected). Thus, subsequently, using another script, these

d) $U = u_{f_L} + x \left(\frac{u_g - u_{f_L}}{T_e} \right) - m_g (u_2 - u_1) + Q = w_i$

$U(0.003^{\circ}C) \Rightarrow u_{f_L} = -333.492 \quad u_{f_L} = -0.033$

$x = \frac{u_{t_1+t_2} - u_{f_L}}{u_g - u_{f_L}}$

FIG. 3. Example of a handwritten solution for Problem 3d. Arguably, such unfortunately very typical handwriting is hard to decipher without knowing the context.

JPGs were processed by multimodal GPT-4V [3], which generated LaTeX code to replace the `\includegraphics` commands. Quite often, though, these image identified by MathPix turned out to be scribbled-out calculations, like the one in Fig. 4 while unrecognizable to MathPix, scribbled-out content frequently cannot be distinguished from drawings, annotations, scanning noise, paper markings, or simply unclear writing by GPT-4V. As a result, these images were sometimes read and inserted as if the scribbling-out did not exist (thus including content into the grading that the student had discarded), and sometimes completely misinterpreted.

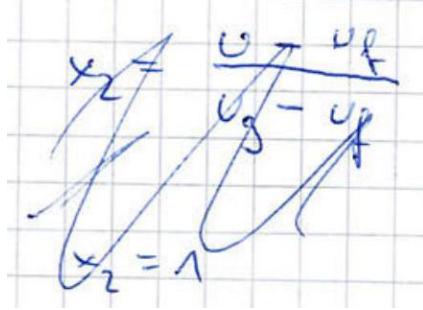


FIG. 4. Example of a cropped JPG produced by MathPix from a student solution of Problem 3d.

The top panel of Figure 5 shows an example of the whole process for parts b and c of Problem 1, shown in Fig. 2. MathPix was unable to process the first expression in Part b, likely because of the annotations above and below the formula terms. The expression was thus turned into an image and subsequently processed by GPT-4V. MathPix, however, did not recognize the small drawing in Part c as such and thus itself processed the handwriting. The top panel of Fig. 6 shows MathPix's interpretation of the solution in Fig. 3, which is accurate on the level that a human would decipher the handwriting without physics knowledge what to expect. The subscript of x , "eis," is the German word for "ice," which would have been nearly impossible to figure out without knowing what the problem is about. The readability is further hampered by the use of a ball pen, which did not

MathPix + GPT-4V

$$\frac{1}{\mu F} \cdot \frac{\int_{ein}^{aus} T ds}{S_{aus} - S_{ein}} = \frac{U_{aus} - U_{ein}}{c \cdot \ln\left(\frac{T_{aus}}{T_{ein}}\right)}$$

$$\Delta_s = R_n c^{\text{if}} \cdot \ln\left(\frac{T_{aus}}{T_{ein}}\right)$$

$$\Delta U = C_{if} (T_{aus} - T_{ein})$$

$$\bar{T} = \frac{T_{aus} - T_{ein}}{\ln\left(\frac{T_{aus}}{T_{ein}}\right)} = 293.12 \text{ K}$$

1c) $295 = \bar{T}_{kF} j 65 \text{ kW}$

Entropiebilanz um Wand:

$$0 + \frac{65kW}{373.15k} - \frac{65kW}{295k} + \dot{S}_{erz} = 0$$

$$\dot{S}_{erz} = 0.04615 \frac{\text{kW}}{\text{k}} = 46.15 \frac{\text{W}}{\text{k}}$$

GPT-4V

b) Stoffmodell:

$$\frac{1}{U_F} = \frac{\int_{ein}^{aus} \dot{s} ds}{\dot{Q}_{aus} - \dot{Q}_{ein}} = \frac{U_{aus} - U_{ein}}{C_F \ln\left(\frac{T_{aus}}{T_{ein}}\right)}$$

$$\Delta s = R \cdot C_F \ln\left(\frac{T_{aus}}{T_{ein}}\right)$$

$$\Delta U = C_F (T_{aus} - T_{ein})$$

$$\frac{1}{U} = \frac{T_{aus} - T_{ein}}{\ln\left(\frac{T_{aus}}{T_{ein}}\right)} \approx 33.72 \text{ K}$$

1c)

$$\dot{Q}_{zu} = 65kW \quad \dot{Q}_{ab} = \dot{Q}_{zu} - \dot{Q}_{erz} = 0$$

$$0 + \frac{65kW}{373.15K} - \frac{65kW}{298K} \dot{S}_{erz} = 0$$

Entropiebilanz um Wand:

$$\dot{S}_{erz} = 0.061615 \frac{\text{kW}}{\text{K}} = 61.615 \frac{\text{W}}{\text{K}}$$

FIG. 5. Different OCR renderings of the student solution for Problem 1b and c in Fig. 2

provide continuous lines.

The process was slightly hampered by Azure at various times declining to process image descriptions, for example the image shown in Fig. 4 produced the error, "the response was filtered due to the prompt triggering Azure OpenAI's content management policy" for whatever GPT-4V saw in the image.

Figure 7 shows an example of a description for a graph drawn by a student in response to Problem 2a, generated by GPT-4V based on an image identified by MathPix. Unfortunately, quite often, MathPix ignored these graphs

MathPix

d)

$$u = u_{f_L} + x(u_{f_e} - u_{f_L}) - m_g(u_2 - u_1) + Q = w_v$$

$$u(0.003^\circ\text{C}) \Rightarrow u_{F_c} = -333.492 \quad u_{F_c} = -0.033$$

$$x = \frac{u_{t1t2} - u_{f_2}}{u_{g_2} - u_{f_2}}$$

GPT-4V Version 1

$$d) U = U_{f_{\text{H}_2\text{O}}} + x(U_{g_{\text{H}_2\text{O}}} - U_{f_{\text{H}_2\text{O}}}) - m_{\text{H}_2\text{O}}(u_2 - u_1) + Q = W_V$$

$$\cdot W(0.2036) = 7 \Rightarrow U_{f_{\text{H}_2\text{O}}} = -333,462 \quad U_{f_{\text{H}_2\text{O}}} = -0,333$$

$$\frac{x}{\omega} = \frac{U_{\text{H}_2\text{O}} + U_{f_{\text{H}_2\text{O}}}}{u_{g_{\text{H}_2\text{O}}} - u_{f_{\text{H}_2\text{O}}}}$$

GPT-4V Version 2

$$d) U = U_{f_x} + X(U_{f_g} - U_{f_x}) - m_y(U_2 - U_1) + Q = W_V$$

$$\cdot W(0.2036) = 7 \quad U_{f_c} = -333,462 \quad U_{f_x} = -0,033$$

$$\frac{X}{\omega} = \frac{U_{int_2} - U_{f_2}}{U_{f_g} - U_{f_2}}$$

GPT-4V Version 3

$$a) U = U_{f_x} + X(U_{f_g} - U_{f_x}) - m_y(U_2 - U_1) + Q = W_v$$

$$\cdot U_{f_c}(0.2036^\circ\text{C}) = 7 \quad U_{f_c} = 333,462 \quad U_{f_x} = -0,333$$

$$\frac{x}{\omega} = \frac{U_1 + x_2 - U_{f_2}}{U_{g_2} - U_{f_2}}$$

GPT-4V Version 4

$$d) U = u_{f_{\text{H}_2\text{O}}} + x(u_{g_{\text{H}_2\text{O}}} - u_{f_{\text{H}_2\text{O}}}) - m_{\text{Cu}}(u_{2-\text{Cu}} - u_{1-\text{Cu}}) + Q = W_v$$

$$\Rightarrow W_v(0.203\text{kg}) = ? \quad u_{f_{\text{H}_2\text{O}}} = 333.462 \quad u_{f_{\text{Cu}}} = -0.033$$

$$\frac{x}{\omega} = \frac{U + m_{2-\text{Cu}} - u_{f_{\text{H}_2\text{O}}}}{u_{g_{\text{H}_2\text{O}}} - u_{f_{\text{H}_2\text{O}}}}$$

GPT-4V Version 5

$$d) U = u_{f_{\text{ice}}} + x(u_{g_{\text{water}}} - u_{f_{\text{water}}}) - m_g(u_2 - u_1) + Q = W_V$$

$$\cdot W(0.2036) = 7 \Rightarrow u_{f_{\text{ice}}} = -333.462 \quad u_{f_{\text{ice}}} = -0.333$$

$$\frac{x}{\omega} = \frac{U_{m_2} - U_{f_2}}{u_{g_2} - u_{f_2}}$$

FIG. 6. Different OCR renderings of the student solution for Problem 3d in Fig. 3 (T = 0.5).

if they came too close to the margin of the page or where horizontally aligned with mathematical expressions.

2. Using only GPT-4V

The alternative method involved converting entire PDF pages to PNG format and processing these solely with GPT-4V; to account for the variability, this was done several times. The bottom panel of Fig. 5 shows an

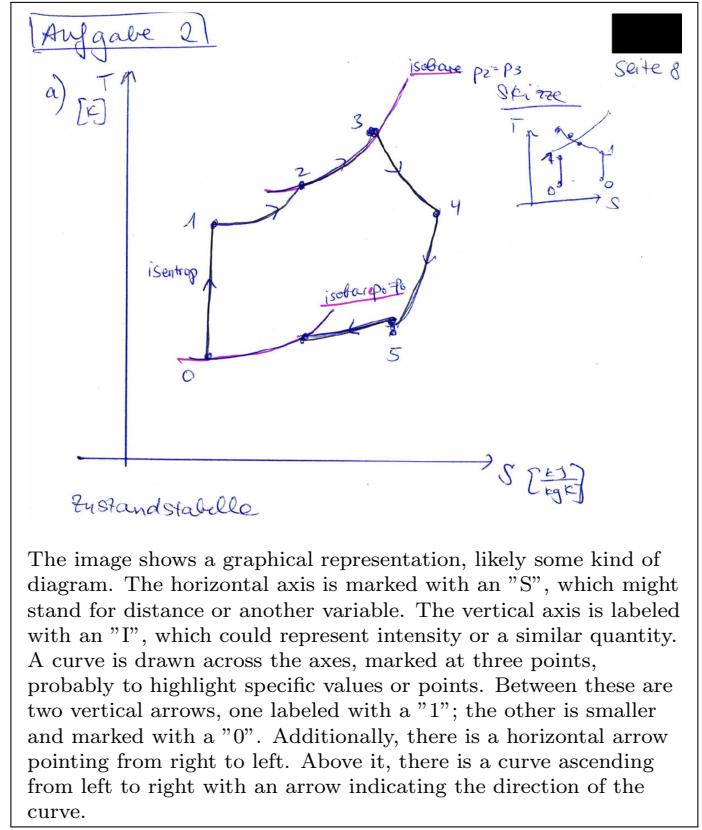


FIG. 7. GPT-4V-generated description (English translation of original German) of a student solution for Problem 2a.

example of this process for parts b and c in Fig. 2. Since due to token-limits, GPT-4V could only process one page image at a time, this entailed uploading each individual page multiple times for interpretation and then reassembling the LaTeX outputs. We decided to assemble all first interpretations of individual pages into a first version of the whole exam, all second interpretations into a second version, etc.; however, this was arbitrary, since all interpretations were independent and, for example, the third interpretation of the fourth page would have no particular connection to the third interpretation of the fifth page.

This approach exhibited strong variability in the accuracy of text interpretation, occasionally producing erroneous interpretations and extrapolations. To limit the variations, the temperature was lowered to T = 0.5, but there were still strong variations, for example:

- As the bottom panel of Fig. 5 shows, all final result numbers were misread.
- The lower panels of Fig. 6 show five interpretations of the handwriting in Fig. 3. GPT-4V shows remarkable creativity interpreting the unclearly written subscripts; MathPix completely skipped “eis,” which overall may be the safer choice.

The temperature of T = 0.5 was determined based

on a small sample size; obviously, finding the optimum temperature would be the topic of further investigations, but computationally intense. Unfortunately, it turned out graphs like the one in Fig. 7 were also “overlooked” by GPT-4V when processing a complete page, making the graph transcription process rather unreliable.

Overall, it was found that MathPix with GPT-4V support was more reliable and reproducible than using only GPT-4V. Since in addition, processing whole pages with GPT-4V turned out to be very token-intensive, and with GPT-4V tokens being more expensive than regular GPT-4-32k tokens, it was decided to only process a subset of 20 exams, but do so five times to account for the strong variability in the results.

C. Grading Step

Following the conversions, the generated LaTeX documents were graded by GPT-4-32k. In addition, the system was prompted to provide a justification for its decisions. Several approaches were evaluated using the first problem to gauge their viability. For two of the more promising approaches, the complete exam was graded. For all grading approaches, the LLM was provided with the problem text itself, but the grading approaches differed by the granularity of their grading scope and sample solution documents.

1. Rubric-based Grading

The teaching assistants (TAs) used a fine-grained grading rubric, which operated on the level of points and half-points. Fig. 9 shows an example of such a grading rubric, where the partial points are marked in red. For AI-grading, the derivation steps in the sample solution were marked up with point values and criteria, and this list was provided to the LLM as a table (see Table I as an example). While the TAs were supposed to make binary decisions whether or not to provide the rubric points, it was realized very early on that the LLM functioned less arbitrarily on a sliding scale, allowing points like 0.4.

2. Parts-based Grading

Here, the system was prompted with the sample solution instead of a fine-grained rubric, and only the total points for each part were given; for the first problem, this resulted in five point values for parts a–e instead of 22 rubric items. Overall, this approach turned out to be less computationally intense than rubric-grading, so it was applied to all problems, not just the first one.

An example of a sample solution is shown in Fig. 9; generally, the student solutions were much shorter and far less explicit than the sample solution, see Fig. 3 (the synthetic data set used in an earlier study of automated

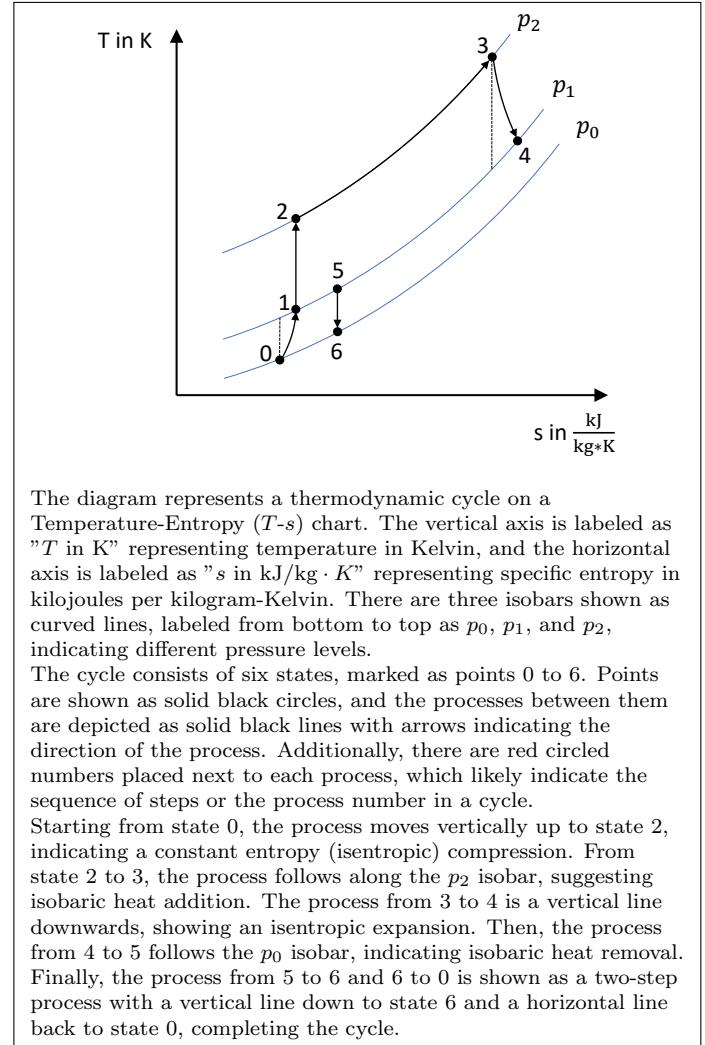


FIG. 8. GPT-4V-generated description of the process diagram in the sample solution of Problem 2a.

grading by the authors [41] resembled the style of the sample solution rather than typical student solutions).

Problems 2 and 4 contain drawings. To grade those, GPT-4V was used to provide a textual description and embed it into the sample solution; Fig. 8 demonstrates this for a process diagram. While the interpretation looks reasonable on the surface, and while certainly more accurate than the interpretation of the hand drawn diagram of the student solution in Fig. 7, closer examination reveals a number of physics errors that would require manual intervention before being used in production settings.

3. Problem-based Grading

The system was prompted with the sample solution for that problem, and only the total number of points for the problem was queried. For the first problem, the system would simply return a value between 0 and 15 points.

TABLE I. Example of the grading rubric for the first exam problem.

Part	Item	Max. Criterion Pts.	
a	a_Bilanz	1	Maximum points (1.0) for correct calculation of heat flow to the cooling jacket; energy balance around the reactor. Full points for accurately applying $0 = \dot{m}_{\text{ein}} \cdot h_{\text{ein}} - \dot{m}_{\text{aus}} \cdot h_{\text{aus}} + \dot{Q}_R - \dot{Q}_{\text{aus}}$ with $\dot{m}_{\text{ein}} = \dot{m}_{\text{aus}} = \dot{m}$. Less points for correct approach with minor errors.
a	a_h_ein	0.5	Maximum points (0.5) for correct calculation of $h_{\text{ein}} = h_f(70^\circ\text{C}) = 292.98 \frac{\text{kJ}}{\text{kg}}$. Less points for minor numerical errors.
a	a_h_aus	0.5	Maximum points (0.5) for correct calculation of $h_{\text{aus}} = h_f(100^\circ\text{C}) = 419.04 \frac{\text{kJ}}{\text{kg}}$. Less points for minor numerical errors.
a	a_Erg	0.5	Maximum points (0.5) for correct calculation of $\dot{Q}_{\text{aus}} = 0.3 \frac{\text{kg}}{\text{s}} \cdot (292.98 \frac{\text{kJ}}{\text{kg}} - 419.04 \frac{\text{kJ}}{\text{kg}}) + 100 \text{ kW}$. Less points for minor errors.
b	b_T_KF_Tds	1	Maximum points (1.0) for correct derivation of the thermodynamic mean temperature of the cooling water flow, \bar{T}_{KF} . Full points for correct formula and calculation. Less points for minor errors.
b	b_Stoffmod	1	Maximum points (1.0) for correct application of the “ideal fluid” model. Full points for correct use of either equation provided. Less points for minor calculation errors.
b	b_Erg	0.5	Maximum points (0.5) for correct calculation of $\bar{T}_{\text{KF}} = \frac{T_{\text{aus}} - T_{\text{ein}}}{\ln(\frac{T_{\text{aus}}}{T_{\text{ein}}})}$. Less points for minor errors.
c	c_Bilanz	1	Maximum points (1.0) for correct entropy production calculation in heat transfer between reactor and cooling jacket. Full points for accurate balance equation application. Less points for minor errors.
c	c_Erg	0.5	Maximum points (0.5) for correct calculation of $\dot{S}_{\text{erz}} = \frac{\dot{Q}_{\text{aus}}}{T_{\text{KF}}} - \frac{\dot{Q}_{\text{aus}}}{T_{\text{Reaktor}}}$. Less points for minor numerical errors.
d	d_u_f	0.5	Maximum points (0.5) for correct calculation of $u_f(100^\circ\text{C}) = 418.94 \frac{\text{kJ}}{\text{kg}}$. Less points for numerical inaccuracies.
d	d_u_g	0.5	Maximum points (0.5) for correct calculation of $u_g(100^\circ\text{C}) = 2506.5 \frac{\text{kJ}}{\text{kg}}$. Less points for numerical inaccuracies.
d	d_u_1	1	Maximum points (1.0) for correct calculation of $u_1 = x_D \cdot u_g(100^\circ\text{C}) + (1 - x_D) \cdot u_f(100^\circ\text{C})$. Less points for minor errors.
d	d_u_2	0.5	Maximum points (0.5) for correct calculation of $u_2 = u_f(70^\circ\text{C}) = 292.95 \frac{\text{kJ}}{\text{kg}}$. Less points for numerical errors.
d	d_h_ein	0.5	Maximum points (0.5) for correct calculation of $h_{\text{ein}} = h_f(20^\circ\text{C}) = 83.96 \frac{\text{kJ}}{\text{kg}}$. Less points for minor numerical errors.
d	d_Bilanz	1	Maximum points (1.0) for correct energy balance in an open system for determining added water mass. Full points for accurate calculation. Less points for incorrect or partial calculations.
d	d_Erg	0.5	Maximum points (0.5) for correct calculation of change in mass $\Delta m_{12} = 3756.84 \text{ kg}$. Less points for minor errors.
e	e_s_f	0.5	Maximum points (0.5) for correct entropy value of fluid at $100^\circ\text{C} = 1.3069 \frac{\text{kJ}}{\text{kgK}}$. Less points for minor errors.
e	e_s_g	0.5	Maximum points (0.5) for correct entropy value of gas at $100^\circ\text{C} = 7.3549 \frac{\text{kJ}}{\text{kgK}}$. Less points for minor errors.
e	e_s_1	1	Maximum points (1.0) for correct calculation of $s_1 = x_D \cdot u_g(100^\circ\text{C}) + (1 - x_D) \cdot u_f(100^\circ\text{C})$. Less points for minor errors.
e	e_s_2	0.5	Maximum points (0.5) for correct calculation of $s_2 = s_f(70^\circ\text{C}) = 0.9549 \frac{\text{kJ}}{\text{kgK}}$. Less points for minor errors.
e	e_Bilanz	1	Maximum points (1.0) for correct entropy balance calculation $\Delta S_{12} = S_2 - S_1$. Full points for correct application of the formula. Less points for minor errors.
e	e_Erg	0.5	Maximum points (0.5) for correct calculation of $\Delta S_{12} = 1387.62 \frac{\text{kJ}}{\text{K}}$. Less points for minor errors.

4. Grading Cycles

Each of the MathPix-processed exams was graded ten times to compute the mean and standard deviation of the scores, resulting in 10 grades per exam component. For the subset of exams where five interpretations were produced by GPT-4V, each version was independently

graded five times by GPT-4, resulting in a total of 25 grades per exam component. If a grading attempt failed because the LLM did not return one of the point values (depending on the granularity, one point value per rubric, one point value per part, or one point value per problem) or if any of these points values were out of range (e.g., assigning more than the available points), the grading cycle was repeated. Table II gives an overview of the workflows

Eisanteil $x_{Eis,2}$ im Zustand 2

Eine Systemgrenze kann um das Eiswasser gezogen werden, sodass der 1 Haupt-satz gilt (kinetische und potenzielle Energien vernachlässigbar):

$$\Delta U_{12} = Q_{12} - W_{12} \quad (102) \quad \text{①}$$

Der Wärmeaustausch von Gas zum Eiswasser wurde in der vorherigen Aufgabe berechnet und liegt bei

$$Q_{12} = 1369.62 \text{ J}$$

Die Volumenarbeit des Eiswassers ist gleich Null, welches von der Annahmen folgt, dass der Dichteunterschied zwischen Eis und Wasser vernachlässigt werden kann. Daraus erhält man $U_{12} = Q_{12}$. ②

Die Veränderung der inneren Energie des Eiswassers kann von ersten Prinzipien errechnet werden. Die Definition der inneren Energie ist gleich

$$\Delta U_{12} = m_{EW} \cdot (u_{EW,2} - u_{EW,1}) \quad (103) \quad \text{①}$$

Die inneren Energiewerte in Zuständen 1 und 2 können von dem jeweiligen Eisanteil errechnet werden. Somit

$$u_{EW,1} = x_{Eis,1} \cdot u_{E,1} + (1 - x_{Eis,1}) \cdot u_{W,1} \quad (104) \quad \text{⑤}$$

$$u_{EW,2} = x_{Eis,2} \cdot u_{E,2} + (1 - x_{Eis,2}) \cdot u_{W,2} \quad (105) \quad \text{⑤}$$

u_W und u_E sind nur Funktionen der Temperatur und in den Tabellen angegeben. Da die Temperatur des Eiswassers zwischen Zuständen 1 und 2 konstant bei $T_{EW} = 0^\circ\text{C}$ bleibt, kann man die jeweiligen inneren Energiewerte in den jeweiligen Zuständen gleichsetzen. Somit

$$u_{W,1} = u_{W,2} = u_W \quad (106) \quad \text{⑤}$$

und

$$u_{E,1} = u_{E,2} = u_E \quad (107) \quad \text{⑤}$$

Liest man die korrekten Werte aus der Tabelle für $T_{EW} = 0^\circ\text{C}$ und $p_{EW} = 1.400 \text{ bar}$ ab bekommt man

$$u_W = -0.045 \frac{\text{kJ}}{\text{kg}} \quad (106) \quad \text{⑤}$$

und

$$u_E = -333.458 \frac{\text{kJ}}{\text{kg}} \quad (107) \quad \text{⑤}$$

In Zustand 1 ist somit

$$u_{EW,1} = x_{Eis,1} \cdot u_E + (1 - x_{Eis,1}) \cdot u_W \quad (108)$$

$$= 0.6 \cdot -333.458 \frac{\text{kJ}}{\text{kg}} + (1 - 0.6) \cdot -0.045 \frac{\text{kJ}}{\text{kg}} \quad (109)$$

$$= -200.093 \frac{\text{kJ}}{\text{kg}} \quad (110)$$

Daraus folgt

$$x_{Eis,2} = \frac{u_{EW,2} - u_W}{u_E - u_W} \quad (111)$$

$$= \frac{\frac{Q_{12}}{m_{EW}} + u_{EW,1} - u_W}{u_E - u_W} \quad (112) \quad \text{①}$$

Wodurch

$$x_{Eis,2} = \frac{\frac{1367 \text{ kJ}}{0.1 \text{ kg}} - 200.093 \frac{\text{kJ}}{\text{kg}} + 0.045 \frac{\text{kJ}}{\text{kg}}}{-333.413 \frac{\text{kJ}}{\text{kg}} + 0.045 \frac{\text{kJ}}{\text{kg}}} \quad (113)$$

$$= 0.559 \quad (114) \quad \text{⑤}$$

FIG. 9. Sample solution and rubric (in red) for Problem 3d (in the original German). Student solutions like the one in Fig. 3 would be graded based on this rubric.

that were explored (here named WF1 to WF4), including initial observations, which are discussed in more detail in the following sections.

Figures 10–12 show the AI prompts used for these workflows. These prompts may appear overly verbose to human readers and putting too much emphasis on ob-

scure corner cases; unfortunately, they are the result of many initially failed attempts to provide proper grading output. It was also cumbersome to strike a balance between strict grading on the one side and being tolerant of what may or may not be OCR scanning errors.

During the grading steps, the grading script replaces the placeholders in square brackets by the respective documents (problem text, rubric or sample solution, and student solution) before submitting the prompt to the LLM. The output of the LLM is analyzed, and if it does not fulfill the requirements (missing rows, higher points than available, etc.), the same prompt is sent again. If the output is complete and passes all sanity tests, it is stored for calculating the total number of points and statistics.

D. Evaluation

For the evaluation, the exam points awarded by the TAs were assumed as ground truth. Evaluations were carried out using Python scripts, Excel, and R [49]. Taking the standard deviations as a measure of confidence, agreement with TA grading was evaluated for different confidence levels using linear regression and Spearman correlations.

IV. RESULTS

A. Rubric-based grading based on MathPix and GPT-4V (WF1)

Table III shows the outcome of grading the student solution in Fig. 2 according to the rubric in Table I. The rows are the grading items with the final row showing the total points. The columns are the results of ten AI-runs, followed by the average of the AI grades, the standard deviation of the AI grades, and the TA grades as ground truth. Here, the Law of Large numbers seems to have applied, as the grading outcomes of the AI and the TAs are almost identical; this, unfortunately, is rarely the case.

Figure 13 shows a Fruchterman-Reingold [50, 51] representation of the Spearman-correlation matrix between the rubric items in Table I for Problem 1, resulting from Workflow 1 (WF1) in Table II. Indicated in yellow are the rubric grades given by the TAs, indicated in blue those given by the AI. Due to the force-directed nature of Fruchterman-Reingold graphs, closely correlated vertices tend to cluster, while unrelated vertices tend to be further apart from each other.

An immediate observation in the graph is that the AI-grades cluster together much more strongly than those of the TAs, which could be explained by the TAs being able to grade the items independently, while the AI tends to grade holistically with only limited ability to differentiate between one step of a solution being correct and another one incorrect. For example, for the TAs, e_Bilanz and

TABLE II. Overview of explored grading workflows.

	Prompt	Student sample	Problem sample	OCR process	#OCR cycles	Grading key	Grading granularity	#Grading cycles	Initial Observations
WF1	Fig. 10	252	1	MathPix + GPT-4V	1	Grading rubric items	Rubric	10	Large number of failed grading attempts, necessitating frequent regrades.
WF2	Fig. 11	252	1–4	MathPix + GPT-4V	1	Whole sample solution	Part	10	Less failed grading attempts.
WF3	Fig. 12	252	1	MathPix + GPT-4V	1	Whole sample solution	Problem	10	Very few failed grading attempts. LLM still graded by part, but frequently ended up incorrectly adding up the points.
WF4	Fig. 11	20	1–4	GPT-4V only	5	Whole sample solutions	Part	5	Strong variations between OCR interpretations.

You are tasked with grading student solutions for this problem:

[problem]

=====

Below is the grading rubric table in comma-separated format. The first column contains the problem part, the second column provides the rubric item identifier, the third column prescribes the maximum points possible for that item, and the fourth column contains the grading criterion.

[rubric]

=====

The problem has multiple parts, and inside each part, it has several rubric items that are graded according to criteria. Note that likely students will not have worked on all parts and items, and they may not have done work in the same order as the rubric lists them.

This is what you need to do:

* Go through every row in the grading rubric table, do not skip any. The row is identified by the rubric item identifier listed in the second column of the rubric table, and that is how you remember it.

* For each row, check if any work pertaining to the criterion is present. If there is no related work, give zero points for that rubric item.

* For each row, if you find work that pertains to the criterion, check how well it corresponds to the criterion. Note that there can be OCR errors, leading to some wrong digits or symbols; this is not the student's fault, and you need to determine if errors are likely due to OCR (no point deduction) or due to the student (some point deduction). Give the maximum number of points listed for that row (third column of rubric table) if the work corresponds very well. Give less points the less it agrees; you can use floating point numbers. Do not give bonus points.

* For each row, provide a comment explaining the rationale behind your grading decision.

Output your grading in a four-column comma-separated table format, similar to the grading rubric table, with a row for every item in the rubric. Use newline as the row separator. The first column in your table must be the problem part, the second column the rubric item identifier, the third column the points that you awarded, and the fourth column your comments explaining the grading decisions. Output nothing else but this completed grading table, and do not enclose it in any special characters, so it is syntactically correct CSV.

Here's the student solution that you need to grade:

[solution]

FIG. 10. Prompt for Workflow 1 (WF1 in Table II). Square brackets indicate where the grading script will insert the respective documents before sending the prompt to the LLM.

You are tasked with grading student solutions for this problem:

[problem]

=====

Below is the sample solution, provided by the instructors:

[sample]

=====

Here's the student solution that you need to grade:

[solution]

=====

You need to carefully compare each part of the student solution step-by-step to the respective part of the sample solution. The student answer might not be in same order as the sample answer. Check carefully for completeness, as well as strictly for physics, math, and numerical accuracy. Assign a percentage grade between 0 and 100 percent based on agreement with the sample solution, and prepare an explanation of your reasoning. Reserve high percentages for excellent work, i.e., almost perfect agreement with the sample solution. Do not hesitate to give very low percentages for incomplete, incoherent, or inaccurate work.

Output the grade percentage for each part as an integer between 0 and 100 and then in quotation marks provide an explanation of your reasoning for this grade in the form

part,percentage,explanation part,percentage,explanation ...

Output nothing else but this list.

FIG. 11. Prompt for Workflows 2 and 4 (WF2 and WF4 in Table II).

e_s_2 are weakly correlated (lower middle of the graph), while in the AI grading, they are strongly correlated.

For both TA and AI grades, however, a clustering by

parts a–e is observable, indicating that problem parts tend to be correct or incorrect as a whole (again, this effect is stronger for the AI than the TA grades). While

You are tasked with grading student solutions for this problem:

[problem]

=====

Below is the sample solution, provided by the instructors:

[sample]

=====

Here's the student solution that you need to grade:

[solution]

=====

You need to carefully compare the student solution step-by-step to the sample solution.

The student answer might not be in same order as the sample answer. Check carefully for completeness, as well as strictly for physics, math, and numerical accuracy. Assign a points based on agreement with the sample solution, and prepare an explanation of your reasoning. Reserve full points for excellent work, i.e., almost perfect agreement with the sample solution. Do not hesitate to give very few points for incomplete, incoherent, or inaccurate work.

Output the points and then in quotation marks provide an explanation of your reasoning for this grade in the form, all in one single line prefaced by "sum", i.e.,

"sum",points,explanation

Output nothing else but this one line.

FIG. 12. Prompt for Workflow 3 (WF3 in Table II).

TABLE III. Example of grading the student solution in Fig. 2 according to the rubric in Table I following WF1

Item	AI 1	AI 2	AI 3	AI 4	AI 5	AI 6	AI 7	AI 8	AI 9	AI 10	AI Ave.	AI-SD	TA
a_Bilanz	0.5	0.8	0.6	1	1	1	1	1	0.8	0.87	0.179	1	
a_h_ein	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.000	0.5	
a_h_aus	0	0	0	0.5	0.5	0.5	0	0.5	0	0.5	0.25	0.250	0.5
a_Ergebnis	0	0.5	0.3	0.5	0.3	0.5	0.5	0.3	0.5	0.5	0.39	0.158	0
b_T_KF_Tds	0.5	0.5	0.8	1	1	1	0.8	1	1	0.86	0.196	1	
b_Stoffmod	0.5	0.5	1	1	1	1	0.5	1	1	0.85	0.229	1	
b_Erg	0	0.5	0.5	0	0.3	0.5	0.5	0.35	0.5	0.4	0.355	0.190	0.5
c_Bilanz	0.5	0.5	1	1	1	1	1	1	1	0.9	0.200	1	
c_Erg	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0.4	0.200	0.5	
d_u_f	0	0.5	0	0.5	0.5	0.5	0.5	0.45	0	0	0.295	0.241	0.5
d_u_g	0.5	0.5	0	0.5	0.5	0.5	0.5	0.45	0	0	0.345	0.226	0.5
d_u_1	0.5	0.5	0.5	1	1	1	1	0.7	0	1	0.72	0.325	1
d_u_2	0	0	0	0.5	0.5	0.5	0.5	0.3	0.5	0.5	0.33	0.224	0.5
d_h_ein	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.45	0.150	0.5
d_Bilanz	1	1	1	1	0.7	1	1	1	0.5	1	0.92	0.166	0
d_Erg	0	0.5	0.5	0.5	0.3	0.5	0.5	0.4	0.5	0.5	0.42	0.154	0
e_s_f	0	0.5	0	0.5	0.5	0.5	0.5	0.5	0	0.5	0.35	0.229	0.5
e_s_g	0	0.5	0	0.5	0.5	0.5	0.5	0.5	0	0.5	0.35	0.229	0.5
e_s_1	0.5	0.5	0.5	1	1	0.5	0	0.75	1	1	0.675	0.317	1
e_s_2	0	0.5	0	0.5	0.5	0.5	0	0.5	0.5	0.5	0.35	0.229	0.5
e_Bilanz	0.5	0.5	0	1	1	1	0	0.6	1	1	0.66	0.388	0
e_Erg	0	0	0.2	0.5	0.4	0.5	0	0.35	0.5	0.5	0.295	0.213	0
Total	5.5	10.3	8.4	14.5	14	14.5	11.5	12.45	11	13.2	11.535	2.745	11.5

generally within each other's vicinity, pairs of TA and AI grades for the same rubric items are not strongly correlated.

The beige vertex "Sum" indicates the total grade for the problem given by the TAs ("sum of the yellow vertices"), while the corresponding cyan vertex indicates the same for the AI grades. Remarkably, this pair of summative vertices is more closely aligned than many of the individual pairs of rubric items. The holistic, undifferentiated tendencies of the AI might be mediated by students tending to get the whole problem mostly correct or mostly incorrect as a whole, based on their understanding of the underlying subject matter.

Figure 14 illustrates the relationship between the total grades for Problem 1 given by the TAs and the AI.

Each data point represents one exam. Considering all grading results, the coefficient of determination is only $R^2 = 0.58$, which would generally be considered moderate. When demanding standard deviations on the AI grading of maximal three, two, or one point, the percentage of AI-gradings fulfilling these conditions decreases while the coefficient of determination R^2 moves into the direction of what might be considered strong. Problems below the set confidence level would need to be graded by hand.

Notably, outliers with very low AI-points remain even at high confidence thresholds; these are due to OCR issues, where MathPix and GPT-4V were unable to interpret the handwriting, and thus the AI-graders consistently gave low points due to missing information. This

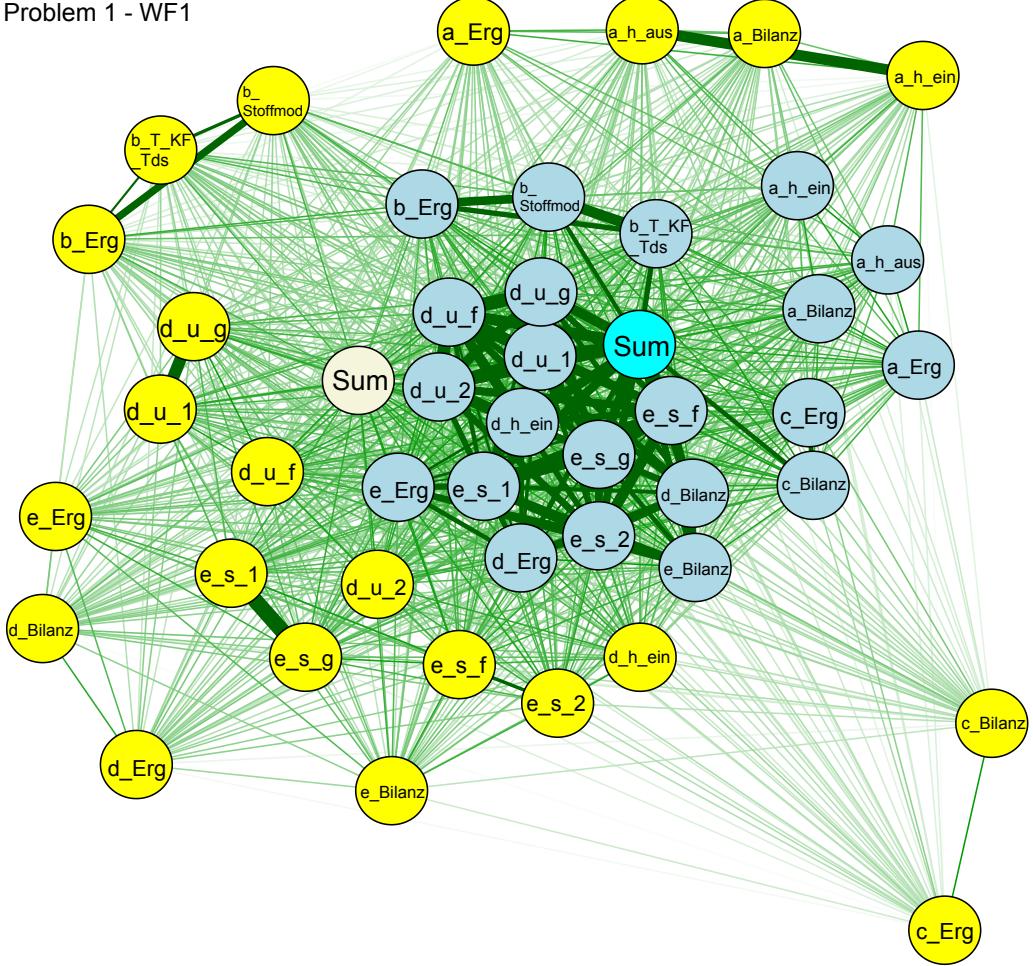


FIG. 13. Fruchterman-Reingold [50] [51] representation of all Spearman correlations between the rubric items in Table I. Indicated in yellow are the grades given by the TAs, indicated in blue those by the AI. The thickness of the edges indicates the strength of the Spearman correlation, where negative correlations would have been indicated in red. The beige “Sum” and cyan “Sum” indicate the sum of points given by the teaching assistants and the AI, respectively.

entails that regardless of the set confidence threshold, problems with very low points should be graded or at least checked by humans.

As observed in earlier studies, in spite of the prompts in Figs. [10][12] attempting to enforce strict grading, the AI grades more leniently than TAs [41], but the $R^2 = 0.84$ achieved with the more verbose synthetic answers could not be reproduced with authentic data.

The low agreement between TAs and AI on individual rubric items, as well as the very frequent failures to provide a complete grading output with points within the available range, indicates that fine-grained rubric grading for a complete problem is beyond GPT-4-32k’s “bookkeeping” capabilities. The high computational load caused by the frequent need to regrade makes WF1 appear nonviable. A variant of this workflow might be prompting the LLM for one part at a time, only submitting the handful of rubric items pertaining to that part; the LLM would likely be able to keep track of this

limited number of grading items, but this approach, for this exam, would have come at the expense of four times more LLM transactions.

B. Part-based grading based on MathPix and GPT-4V (WF2)

Workflow 2 (WF2) considers problem parts rather than rubric items. Initial tests on Problem 1 showed that the LLM was much better equipped to keep track of four or five problem parts than of 22 rubric items, and the number of necessary regrades dramatically decreased; Table IV shows an example for the problem solution in Fig. [2]. The coefficient of determination for TA versus AI grades for Problem 1 was $R^2 = 0.56$ for the complete data set and thus comparable to the $R^2 = 0.58$ obtained from rubric grading, see Fig. [15]. The standard deviations of the average grades of the 10 AI graders are lower than

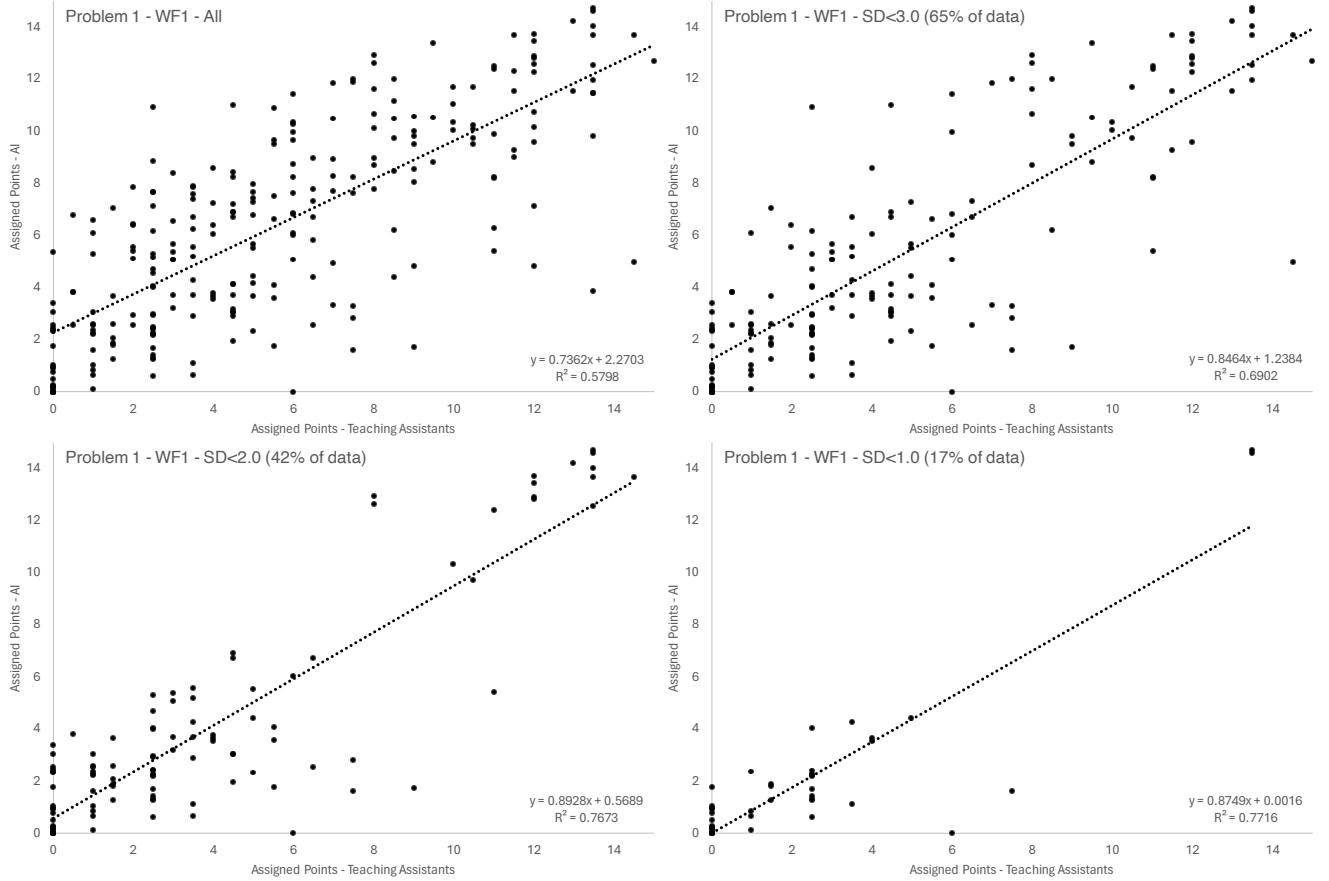


FIG. 14. AI versus TA grades for Problem 1 using WF1 including linear regression lines, their equations and R^2 . The upper left panels shows all grades independent of standard deviation σ of the mean of the 10 AI grades, the remaining panels the relationship for increasingly stringent limitations on the standard deviation. In each case, the percentage of data points fulfilling the restriction is indicated.

TABLE IV. Example of grading the student solution in Fig. [2] based on the sample solution by parts, following WF2

Part	AI 1	AI 2	AI 3	AI 4	AI 5	AI 6	AI 7	AI 8	AI 9	AI 10	AI Ave.	AI-SD	TA
a	2	2.25	2	2	2.375	2.125	1.5	2	2	2.25	2.05	0.225	2
b	2.125	2.5	1.75	2.25	2.125	2.125	2.5	2.25	2	2.5	2.2125	0.231	2.5
c	1.35	1.5	1.2	1.5	1.35	1.35	1.35	1.35	1.5	1.425	1.3875	0.090	1.5
d	3.375	4.5	3.15	3.825	3.6	3.825	3.15	3.15	3.6	4.05	3.6225	0.420	3
e	2.8	3.6	2.8	2.8	3.2	3.4	2.2	2.8	2.8	4	3.04	0.488	2.5
Total	11.65	14.35	10.9	12.375	12.65	12.825	10.7	11.55	11.9	14.225	12.3125	1.182	11.5

in workflow WF1, which results in less exams requiring human grading when imposing limits on the maximum standard deviation.

For this workflow, all problems were considered, and Fig. [16] shows the Spearman correlations between the TA and AI grades. As seen earlier in Fig. [13], the sums of the parts are more closely correlated than the pairs of parts. Very prominent is the outlying vertex for the TA-graded Part 4a, which is one of two graphical problem parts. As opposed to the other graphical problem part, Part 2a, which contributes 8 points to the 19 points of Problem 2 (explaining the strong correlation with “Sum”), Part 4a

only contributes 4.5 points to the 18.5 points of Problem 4. The AI-graded Parts 2a and 4a, on the other hand, are simply part of the cluster of AI-points visible in all Spearman correlations.

The pairs of vertices for the graphical problems 2a and 4a show no correlations (for 2a, the vertices are at opposite ends of the graph, while for 4a, the TA-vertex is completely disconnected). These would be grades resulting from comparing the descriptions of solutions like Fig. [7] to those in Fig. [8] which is a task that LLMs would in principle be capable of if the descriptions were reliable; unfortunately, the latter proved not to be the case.

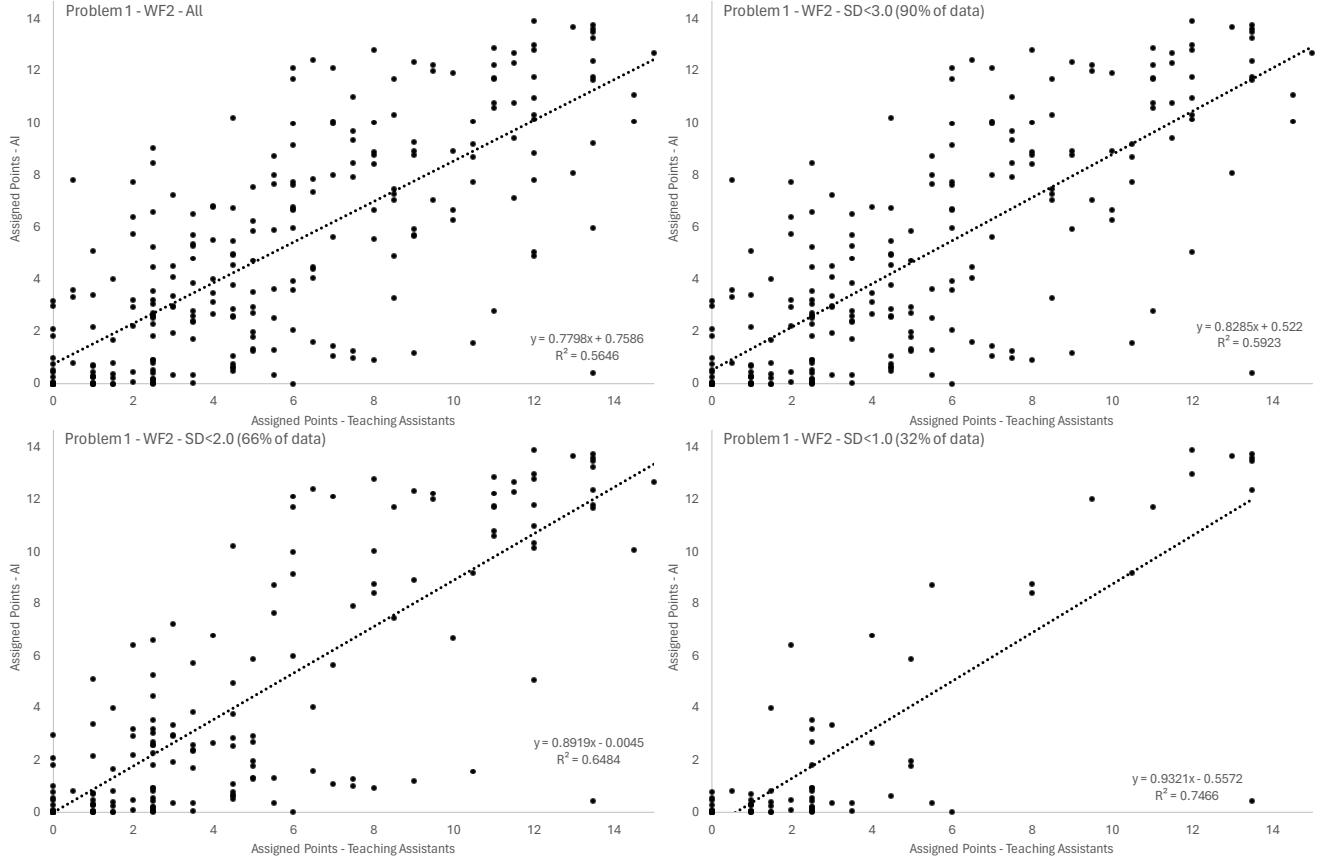


FIG. 15. AI versus TA grades for Problem 1 using WF2 including linear regression lines, their equations and R^2 .

C. Problem-based grading based on MathPix and GPT-4V (WF3)

Grading a complete problem with only one summary grade turned out to have the lowest correlation with TA grades of all workflows, $R^2 = 0.43$. Table V shows an example for the student solution in Fig. 2. The grading explanations showed that the model still attempted to grade by part, for example it graded the solution in Fig. 2 in Run AI1 (see Table V) as follows:

The student solved parts a), b) and c) correctly, earning 2.5 points, 2.5 points and 1.5 points respectively. For part d), the student made a miscalculation resulting in a discrepancy with the model solution, earning 2 of the 4.5 points. For part e), the student made another miscalculation resulting in a discrepancy with the model solution, earning 2.5 of the 4 points.

The model extracted the point values for the parts from the problem text, which was part of the prompt (see Fig. 12). However, it was unable to correctly add five numbers, as $2.5+2.5+1.5+2+2.5$ is 11, but the model reported 11.5 in the output. As the model seems to proceed by parts anyway, but cannot be trusted to add up the points correctly, it is seems advisable to leave arithmetics to Python or Excel.

D. Part-based grading based solely on GPT-4V (WF4)

Workflow WF4 has multiple GPT-4V-only interpretations of the handwriting (see Fig. 6) and multiple GPT-4-32k graders, resulting in more than double of the grades generated by the other workflows (25 versus 10). This choice was made in the hope of decreasing standard deviations on the average grades, however, the workflow turned out to have higher standard deviations than WF2 and WF3. Table VI shows an example for the solution in Fig. 2. As increasingly strict limits on the standard deviation are imposed, the Coefficient of Determination R^2 did not consistently increase. However, due to the high computational load, this approach was abandoned after processing 20 sample exams, so this lack of consistency might be the result of both workflow-inherent uncertainties in the OCR process and simple lack of statistics (20 versus 252 data points).

Also in this workflow, student graphs like the one in Fig. 7 were frequently “overlooked,” and grading results of the recognized graphs were unreliable. GPT-4V seems much better equipped to describe graphs with clean lines like the one in Fig. 8 than freehand drawings like Fig. 7

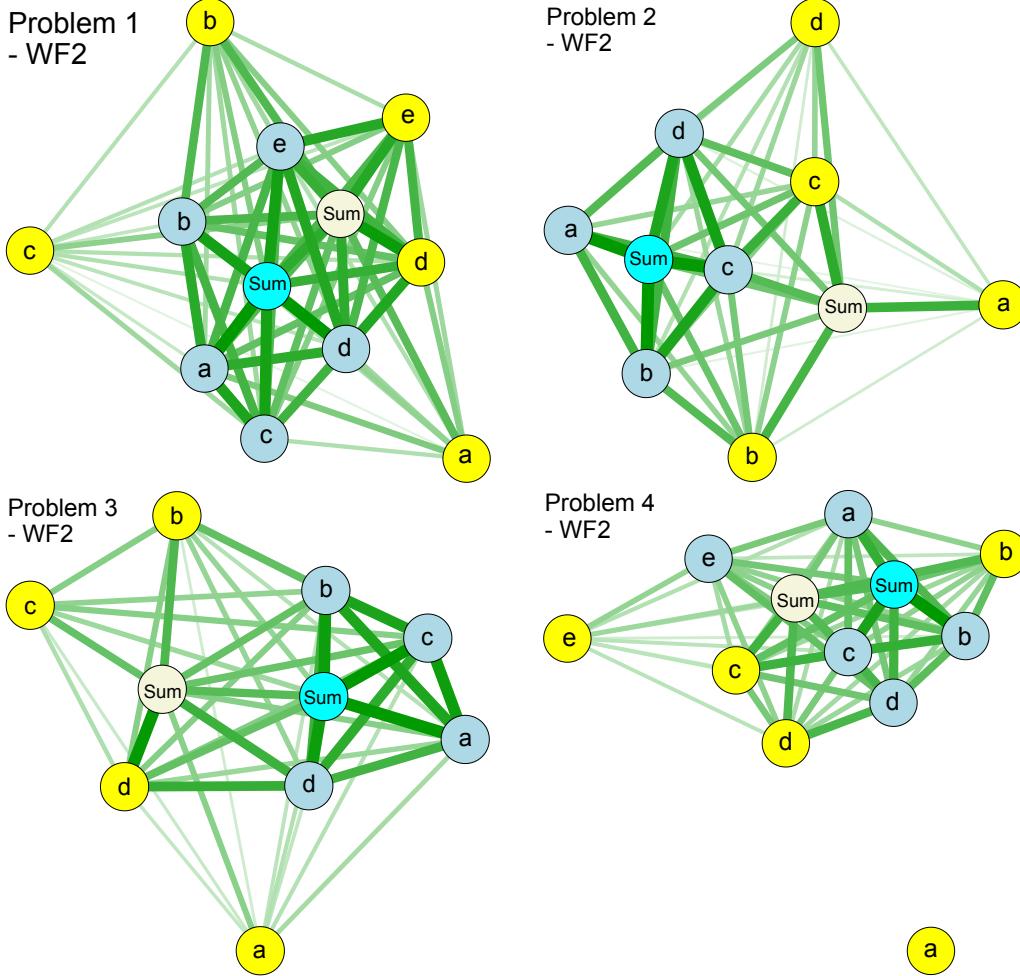


FIG. 16. Fruchterman-Reingold [50] [51] representation of all Spearman correlations between the part grades for all problems on the exam. Indicated in yellow are the grades given by the TAs, indicated in blue those by the AI. The beige “Sum” and cyan “Sum” indicate the sum of points given by the teaching assistants and the AI, respectively.

TABLE V. Example of grading the student solution in Fig. 2 based on the sample solution, not prompting for parts, following WF3

	AI 1	AI 2	AI 3	AI 4	AI 5	AI 6	AI 7	AI 8	AI 9	AI 10	AI Ave.	AI-SD	TA
Total	11.5	10	12.5	12.5	10	11.5	10	11	9	13	11.1	1.261	11.5

E. Influence of Confidence Thresholds

Figure 17 summarizes the results of Coefficient of Determination R^2 versus imposed limits on the standard deviation. For W1-WF3, logarithmic fits are included; for WF4, the randomness is too high or the statistics too low to provide a reasonable fit. From left to right, the data points represent acceptance of standard deviations lower than one, three, or any number of points, respectively; thus, data points further to the left on the horizontal axis indicate greater need for human grading, while the values on the vertical axis indicate better agreement between TA and AI grading.

WF4 does not seem ready for production, as it is more

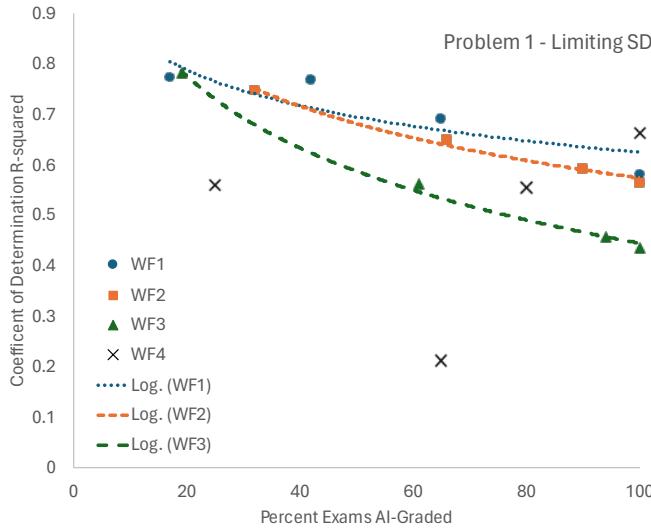
unpredictable than the other workflows. Workflows WF1 and WF2 clearly outperform WF3 (possibly simply because GPT-4-32k cannot add numbers). If, based on confidence measures, half of the exams are AI-graded, both models reach $R^2 \approx 0.7$. While these coefficients indicate strong correlations, it is important to note that they would be unacceptable for high-stakes exams.

F. Total Exam Points and Passing or Failing

Focussing on WF2, the overall grading performance is less than convincing: the Coefficient of Determination between the total points determined by the TAs

TABLE VI. Example of grading the student solution in Fig. 2 based on the sample solution by parts, following WF4

Part	AI 1.1	AI 1.2	AI 1.3	AI 1.4	AI 1.5	AI 2.1	...	AI 4.5	AI 5.1	AI 5.2	AI 5.3	AI 5.4	AI 5.5	AI Ave.	AI-SD	TA
a	0	1	0	1.875	0.25	1.5		0.25	2.125	1.75	1.5	2	1.75	0.935	0.760	2
b	0	0.75	1.25	1.125	0	1.25		0	2.25	1.5	2.25	2.5	2.25	0.9	0.846	2.5
c	1.05	1.35	1.125	1.275	0.225	1.35		0	1.425	1.2	1.425	1.35	1.5	0.729	0.601	1.5
d	3.375	2.25	2.25	4.05	0	0	...	1.8	2.7	2.7	2.25	3.825	2.7	1.989	1.452	3
e	3.4	4	3.2	2.2	0.6	0		0	2	2.8	2.4	4	2	1.76	1.397	2.5
Total	7.825	9.35	7.825	10.525	1.075	4.1	...	2.05	10.5	9.95	9.825	13.675	10.2	6.313	3.912	11.5

FIG. 17. Coefficient of determination R^2 versus percentage of accepted AI-grades for Problem 1 as restrictions are put on the standard deviation σ of the AI-grades, comparing the different workflows.

and those determined by AI is $R^2 = 0.46$. When setting some additional thresholds, namely that the sum of all standard deviations is less than 10 points ($\sum \sigma < 10$) to eliminate data points with low confidence, and also eliminating data points with less than one point total, as they are likely the result of a failure of the OCR process, one is left with 84% of the data and $R^2 = 0.49$.

Applying the same rule for the determination of the passing threshold as with the TA-score, 50% of the best AI-score is 25 points. The next gap in the score distribution occurs at 24 points, which would be a passing threshold that could be automatically determined from the AI-grades. One can define the following:

- True Positive (TP): the AI grading passed a student who also passed with TA grading.
- True Negative (TN): the AI grading failed a student who also failed with TA grading
- False Positive (FP): the AI grading passed a student who failed with TA grading
- False Negative (FN): the AI grading failed a student who passed with TA grading

The left panel of Fig. 18 shows the contingency table (separated by the red lines) for all 252 students, the right panel shows the same for the above outlined restrictions. Also indicated are the standard measures accuracy, precision, recall, and F2-score:

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{F2-score} &= \frac{(1 + 2^2) \times \text{Precision} \times \text{Recall}}{2^2 \times \text{Precision} + \text{Recall}} \end{aligned}$$

Precision measures the ratio of correctly AI-assigned passing grades to the total number of AI-assigned passing grades, while recall measures the ratio of the correctly AI-assigned passing grades to the total number of TA-assigned passing grades. The workflow has very high precision, (0.9468 and 0.9589 without and with threshold criteria, respectively), but low recall (0.4972 and 0.4698, respectively). Thus, if the AI passed a student, very likely he or she actually passed the exam, but the AI missed more than half of the actually passing grades. The low recall is also what causes the low accuracy and the low F2-score.

G. Summary and General Observations

Two of the investigated workflows were hampered by the fact that LLMs are simply not good at bookkeeping: for WF1, it would lose rubric items, for WF3, it could not add numbers. All workflows were hampered by OCR, and none of them dealt well with diagrams like Fig. 7.

Overall, model WF1, going by rubric, seems to be the most reliable workflow. In this study, it was not pursued for the whole exam because GPT-4-32k very frequently could not keep track of all of the fine-grained rubric items, leading to an unjustifiable large number of wasted tokens. A hybrid approach, running the model for each part with a handful of rubric items each might be a promising future approach. While that would entail four or five times more grading rounds (based on the number of problem parts), each one of those would likely return a viable result, as the system can reliably keep track of a handful of items.

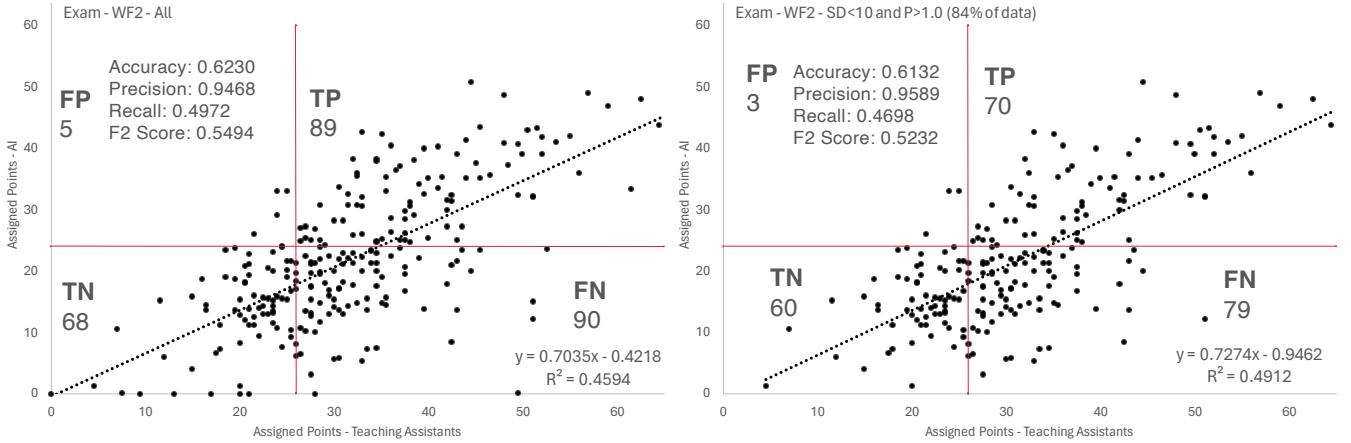


FIG. 18. Total exam points using WF2. The graphs show the linear interpolations, as well as the contingency table and quality measures for passing the exam.

V. LESSONS LEARNED

Between pseudonymization, sorting, rotating, marking up, and cleaning of the PDFs, considerable human effort went into the AI-assisted grading workflows, i.e., between the scanning and OCR process. When endeavoring to use the outlined workflows in production, a balance has to be found between the students working in a manner that does not distract from the subject matter and minimizing human effort on the part of the course personnel.

A. Avoid Boxes

During the OCR process, boxes created a surprising amount of problems: frames on paper that the students brought in or boxes that they drew around answers can lead to the OCR software identifying the box as a figure. Particularly problematic was if students wrote across the lines of a box or literally “outside the box.” While boxes can guide the human eye, they are a distraction to OCR software and should be avoided.

B. Straight-line Layout

An advantage of handwriting mathematical derivations is the ability to easily annotate terms (e.g., an arrow or underbrace, “this goes to zero,” or sub-calculations literally on the side). While OCR software is surprisingly good dealing with the two-dimensional character of mathematical expressions (e.g., nested fractions, even within column vectors), unfortunately, they cannot follow less structured graphical layouts, and those free-floating texts or calculations end up out of context within the flow of the larger calculations. Here, students might be asked to please stay within horizontal lines, also avoiding side-by-side multi-column layouts.

C. Use Plain Paper

The presence of lines and checkers on paper as in Fig. 2 is surprisingly well ignored by the OCR processes, but still can cause further complications. A better approach is using plain paper, which leads to the next point:

D. Provide the Paper

Handwriting the problem derivations using pen-and-paper remains essential, but the bring-your-own-paper policy has created logistics problems, and redacting, sorting, rotating, and labeling the pages required considerable effort. OCR can read typewritten information with very high precision, and this feature can be used to automate several of these operations.

For exams that expect relatively short answers, that is, just a few lines of derivations, simply provide space on the exam sheet. Instead of answer boxes, leave marker texts according to which the OCR’d text file can be separated, for example, print “Solution Problem #2:” at the beginning of the free space, and “Problem #3:” before the text of the next problem; markers like these do not distract the students but can easily be picked up by the evaluation scripts between the OCR and the grading.

For exams that require longer derivations of unclear length (depending on the approach taken), a simple step would be to provide plain paper that has a normed header with a defined vocabulary. This header might look like the one depicted in Fig. 19. Note that there are no boxes. The terms “Student number, problem number, part, and page” will be read by the OCR software with very high precision, and subsequent scripts can easily divide and sort pages according to this pattern (note that it does not matter if the page number is absolute or relative to the problem start or the part start, as the pages would be first sorted by problem, then part, then number). Torn-

out pages from notebooks or oversized sheets can also cause paper jam during the automated scanning.

Students would be instructed to not have more than one problem part on the same page. They would also be asked to use the paper in portrait orientation with the header on top, minimizing the need for later page rotations. Any pattern-matching would need to be tolerant enough to deal with students' errors that might be introduced by nervousness.

E. Encourage Students to Write More

As can be seen when comparing Figs. 3 and 9, students tend to write much less than experts when solving problems. It is good practice to provide more explanations when deriving the answers to problems, and from a student point of view, it also provides more chances to get partial credit when the overall solution turns out to be wrong. For AI-grading, having more rather than less text and written out formula derivations greatly increases error-tolerance in both the OCR and the grading process. LLMs tend to be verbose in their answers, but they also appear to need the similar verbosity to make reliable statements on similarity; they are, after all, probabilistic language models, not deterministic symbolic algebra systems.

F. Use Pencils and Erasers

Scribbling out writing is not understood by OCR software; at times, the OCR software provides an interpretation of these expressions that the student wanted ignored, at other times, the model starts hallucinating (see Fig. 4). Permanent markers are oftentimes mandated to avoid tampering with the exams once they are returned to the student, but the first step in this process is scanning. Thus, an electronic copy of the original student work is available, which would not only provide clear evidence of later tampering, but likely prevent this dishonest action in the first place. Students should erase wrong expressions rather than scribbling them out (providing extra pencils and clean, white erasers in the exam room is probably a good idea).

G. Avoid Graphical Problems (for now)

Our experiments showed that least for the moment, various steps in the workflows fail for graphical problems: the graphs might get “overlooked,” and their descriptions might be too vague to allow for a meaningful comparison to sample solutions.

VI. LIMITATIONS

This study is decidedly empirical, investigating different workflows for AI-assisted grading of one thermodynamics using tool currently available (Spring 2024). The results are specific to the exam, the rubric (Table I), and the prompts (Figs. 10–12), so only limited generalizability can be claimed.

VII. OUTLOOK

The logistics and format of high-stakes exams are hard to adapt according to the lessons in Sect. V and a future study should consider a lower-stakes exam which can be adapted for AI-grading assistance, even if this comes at the expense of authenticity of the data. Based on the findings in this study, a workflow should be investigated that uses a detailed grading rubric, but only applies it to one problem part at a time to reduce the number of failed grading attempts.

As even humans can sometimes only decipher handwritten solutions like Fig. 3 in context, future experiments with GPT-only OCR processes (like in WF4) might include the problem text in the reading prompt. Preliminary findings indicated that currently, this runs the risk of pushing the system over the token limit, and it can lead to increased hallucinations. In particular, the process runs the risk of the LLM attempting to solve the problem and “seeing what it expects to see;” further and more diligent prompt engineering would be needed to ensure that the system avoids OCR errors but does not fix physics errors. This is particularly true since currently when processing a page, it is unknown which problem is being answered, so at the moment, all problem texts would need to be included. It can be expected that future models have higher token limits.

At the time of our study, GPT-4o [52] had not been available yet. There is anecdotal evidence that GPT-4o is worse than GPT-4 in mathematical and reasoning task, but better in handwriting recognition. For example, the excerpt in Fig. 3 gets interpreted as

$$\text{d)} \quad U = u_{f2} + x(u_g - u_{f2}) - mg(u_2 - u_1) + Q = W_v$$

$$U(0.003c) \Rightarrow u_{Fc} = -333.492 \quad u_{Fc} = -0.033$$

$$x_{f2} = \frac{U_{n+2} - U_{f2}}{u_g - u_{f2}}$$

which is better than the results in Fig. 6. Also, GPT-4o shows promise in interpreting graphs; Fig. 20 shows the GPT-4o description of the graph in Fig. 7

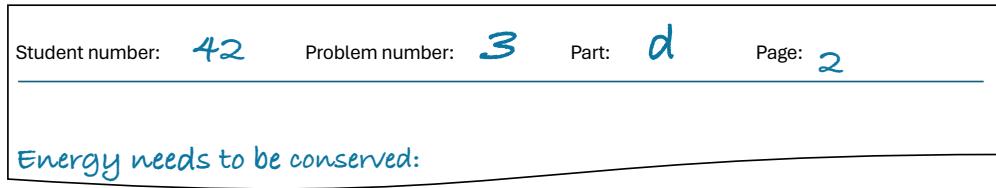


FIG. 19. Suggested header for instructor-provided exam paper.

Description of the Thermodynamic Cycle

The diagram illustrates a thermodynamic cycle plotted on a Temperature-Entropy (T - S) diagram. The key points and processes are as follows:

- The vertical axis represents Temperature (T) in Kelvin [K].
- The horizontal axis represents Entropy (S) in kilojoules per kilogram per Kelvin [$\frac{\text{kJ}}{\text{kg K}}$].
- The cycle consists of five key points labeled 0 through 5.

Processes

1. **Process 0-1:** This is an isentropic process, indicated by a vertical line on the T - S diagram, showing that the entropy remains constant.
2. **Process 1-2:** This process shows an increase in both temperature and entropy, moving upwards to the right.
3. **Process 2-3:** This is an isobaric process (constant pressure), indicated by the label “isobare $p_2 = p_3$ ”. The process moves horizontally to the right, indicating an increase in entropy while the temperature remains constant.
4. **Process 3-4:** This process involves a decrease in both temperature and entropy, moving downwards to the left.
5. **Process 4-5:** This is another isentropic process, moving vertically downwards, indicating a constant entropy.
6. **Process 5-0:** This is an isobaric process (constant pressure), indicated by the label “isobare $p_0 = p_5$ ”. The process moves horizontally to the left, indicating a decrease in entropy while the temperature remains constant.

Additional Notes

- A small inset in the top right corner of the diagram provides a simplified sketch of the cycle, highlighting key transitions between states.
- The term “Zustandstabelle” at the bottom suggests a reference to a state table, typically used to list the properties of each state in the cycle.

FIG. 20. GPT-4o description of the graph in Fig. 7

Future work will explore a grading workflow that uses GPT-4o for handwriting and graphics interpretation, followed by using GPT-4 to grade each individual problem part with a detailed grading rubric, and summing up points using Python.

VIII. CONCLUSION

While we found once again that Large Language Models (LLMs) can neither reliably count nor add numbers, they have unique properties that make them useful for

assisting in free-form exam grading: they can deal with fuzzy data in a probabilistic manner.

The biggest source of uncertainty was the conversion of handwriting to a machine-readable format, that is, the Optical Character Recognition which forms the base for all further grading steps. We found that some common sources of error can be addressed by changing the format of the exam sheets, and we had anecdotal evidence that newer multimodal LLMs perform better.

Once the exams are in machine-readable format, in our case LaTeX, we found that the granularity of the grading prompts is essential to both the accuracy and the rate-of-failure of automated grading assistance. A fine-grained grading rubric, applied to a whole problem at a time, leads to frequent bookkeeping errors and failed grading attempts. Grading the whole problems by a handful of parts, using the full sample solution, turned out to be more reliable, but misses some of the nuances and weightings of a rubric. When attempting to grade the whole problem all at once, the model resorted to grading it by part, but then failed to correctly add up the points; essentially, again, a bookkeeping error.

Grading graphical solutions, in our case process diagrams, turned out to have much lower reliability than grading mathematical derivations. When those solutions were not “overlooked” in the first place, the descriptions of free-hand drawings were much less reliable than those resulting from careful construction on a computer; crooked lines or scribbles provided too much extraneous information, which the tool was unable to distinguish from the salient information. Also here, newer multimodal models may perform better.

The overall gradings had high precision, that is, they identified exams that met passing criteria with high reliability. However, other performance measures such as accuracy, recall, and F2 were too low for failing students; exams that would lead to failing grades definitely require human validation. Noticeable were also high-confidence zero-point grades, which result from utter OCR failures.

ACKNOWLEDGMENTS

We would like to thank the students who participated in this study. We would also like to thank Fadoua Balabdaoui for statistical consulting, and Alina Yaroshchuk for assistance with exam logistics.

-
- [1] OpenAI, ChatGPT, <https://chat.openai.com/> (accessed April 2024).
- [2] A. M. Turing, Computing machinery and intelligence, *Mind* , 433 (1950).
- [3] OpenAI, ChatGPT, <https://openai.com/research/gpt-4> (accessed April 2024).
- [4] T. H. Kung, M. Cheatham, A. Medinilla, ChatGPT, C. Sillos, L. De Leon, C. Elepano, M. Madriaga, R. Aggabao, G. Diaz-Candido, *et al.*, Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models, medRxiv , 2022 (2022).
- [5] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, GPT-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
- [6] Samantha Murphy Kelly, ChatGPT passes exams from law and business schools, <https://edition.cnn.com/2023/01/26/tech/chatgpt-passes-exams/index.html> (accessed January 2023).
- [7] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, *Phys. Rev. Phys. Educ. Res.* **19**, 010132 (2023).
- [8] G. Polverini and B. Gregorcic, Performance of chatgpt on the test of understanding graphs in kinematics, *Phys. Rev. Phys. Educ. Res.* **20**, 010109 (2024).
- [9] G. Kortemeyer and W. Bauer, Cheat sites and artificial intelligence usage in online introductory physics courses: What is the extent and what effect does it have on assessments?, *Phys. Rev. Phys. Educ. Res.* **20**, 010145 (2024).
- [10] W. Yeadon and T. Hardy, The impact of AI in physics education: a comprehensive review from GCSE to university levels, *Physics Education* **59**, 025010 (2024).
- [11] A. Sperling and J. Lincoln, Artificial intelligence and high school physics, *The Physics Teacher* **62**, 314 (2024).
- [12] S. Küchemann, S. Steinert, N. Revenga, M. Schweinberger, Y. Dinc, K. E. Avila, and J. Kuhn, Can ChatGPT support prospective teachers in physics task development?, *Phys. Rev. Phys. Educ. Res.* **19**, 020128 (2023).
- [13] G. Kortemeyer, Using artificial-intelligence tools to make LaTeX content accessible to blind readers, *TUGboat* **44**, 390 (2023).
- [14] P. Tschisgale, P. Wulff, and M. Kubsch, Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory, *Physical Review Physics Education Research* **19**, 020123 (2023).
- [15] F. Kieser, P. Wulff, J. Kuhn, and S. Küchemann, Educational data augmentation in physics education research using ChatGPT, *Phys. Rev. Phys. Educ. Res.* **19**, 020150 (2023).
- [16] F. Reif, J. H. Larkin, and G. C. Brackett, Teaching general learning and problem-solving skills, *American Journal of Physics* **44**, 212 (1976).
- [17] F. Reif, Millikan lecture 1994: Understanding and teaching important scientific thought processes, *American Journal of Physics* **63**, 17 (1995).
- [18] L. Hsu, E. Brewe, T. M. Foster, and K. A. Harper, Resource letter rps-1: Research in problem solving, *American journal of physics* **72**, 1147 (2004).
- [19] J. Hattie, *Visible learning: A synthesis of over 800 meta-analyses relating to achievement* (routledge, 2008).
- [20] F. Al-Salmani, J. Johnson, and B. Thacker, Assessing thinking skills in free-response exam problems: Pandemic online and in-person, *Phys. Rev. Phys. Educ. Res.* **19**, 010131 (2023).
- [21] D. A. Kashy, G. Albertelli, G. Ashkenazi, E. Kashy, H.-K. Ng, and M. Thoennessen, Individualized interactive exercises: a promising role for network technology, in *Proc. Frontiers in Education*, Vol. 31 (2001) pp. 1073–1078.
- [22] G. Kortemeyer, E. Kashy, W. Benenson, and W. Bauer, Experiences using the open-source learning content management and assessment system LON-CAPA in introductory physics courses, *Am. J. Phys* **76**, 438 (2008).
- [23] J. Risley, Motivating students to learn physics using an online homework system, *Newsletter of the APS Forum on Education* , 3 (2001).
- [24] T. Stelzer and G. Gladding, The evolution of web-based activities in physics at illinois, *Newsletter of the APS Forum on Education* , 7 (2001).
- [25] R. J. Dufresne, D. Hart, J. P. Mestre, and K. Rath, The effect of web-based homework on test performance in large enrollment introductory physics courses, *Journal of Computers in Mathematics and Science Teaching* **21**, 229 (2002).
- [26] C. Fredericks, *Patterns of Behavior in Online Homework for Introductory Physics*, Ph.D. thesis, University of Massachusetts (2007).
- [27] M. Richards-Babb, J. Drelick, Z. Henry, and J. Robertson-Honecker, Online homework, help or hindrance? what students think and how they perform, *Journal of College Science Teaching* **40**, 81 (2011).
- [28] D. C. Perdian, Early identification of student performance and effort using an online homework system: A pilot study, *Journal of Science Education and Technology* **22**, 697 (2013).
- [29] J. L. Docktor, J. Dornfeld, E. Frodermann, K. Heller, L. Hsu, K. A. Jackson, A. Mason, Q. X. Ryan, and J. Yang, Assessing student written problem solutions: A problem-solving rubric with application to introductory physics, *Physical review physics education research* **12**, 010130 (2016).
- [30] E. Burkholder, J. Miles, T. Layden, K. Wang, A. Fritz, and C. Wieman, Template for teaching and assessment of problem solving in introductory physics, *Physical Review Physics Education Research* **16**, 010123 (2020).
- [31] T. Wan and Z. Chen, Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning, *Phys. Rev. Phys. Educ. Res.* **20**, 010152 (2024).
- [32] J. Wilson, B. Pollard, J. M. Aiken, M. D. Caballero, and H. J. Lewandowski, Classification of open-ended responses to a research-based assessment using natural language processing, *Phys. Rev. Phys. Educ. Res.* **18**, 010141 (2022).
- [33] P. Mitros, V. Paruchuri, J. Rogosic, and D. Huang, An integrated framework for the grading of freeform responses, in *The Sixth Conference of MIT's Learning International Networks Consortium* (2013).
- [34] M. O. Dzikovska, R. Nielsen, C. Brew, C. Leacock, D. Giampiccolo, L. Bentivogli, P. Clark, I. Dagan, and H. T. Dang, SemEval-2013 task 7: The joint student response

- analysis and 8th recognizing textual entailment challenge, in *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (2013) pp. 263–274.
- [35] S. Burrows, I. Gurevych, and B. Stein, The eras and trends of automatic short answer grading, *International journal of artificial intelligence in education* **25**, 60 (2015).
- [36] C. Sung, T. Dhamecha, S. Saha, T. Ma, V. Reddy, and R. Arora, Pre-training BERT on domain resources for short answer grading, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019) pp. 6071–6075.
- [37] S. Azad, B. Chen, M. Fowler, M. West, and C. Zilles, Strategies for deploying unreliable ai graders in high-transparency high-stakes exams, in *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I* **21** (Springer, 2020) pp. 16–28.
- [38] M. Fowler, B. Chen, S. Azad, M. West, and C. Zilles, “Autograding” explain in plain english” questions using nlp, in *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (2021) pp. 1163–1169.
- [39] R. Stanyon, E. Martello, M. Kainth, and N. K. Wilkin, Demo of graide: Ai powered assistive grading engine, in *Proceedings of the Ninth ACM Conference on Learning@ Scale* (2022) pp. 466–468.
- [40] S. Grassini, Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings, *Education Sciences* **13**, 692 (2023).
- [41] G. Kortemeyer, Toward AI grading of student problem solutions in introductory physics: A feasibility study, *Phys. Rev. Phys. Educ. Res.* **19**, 020163 (2023)
- [42] S. Mori, C. Y. Suen, and K. Yamamoto, Historical review of ocr research and development, *Proceedings of the IEEE* **80**, 1029 (1992).
- [43] H. Okamura, T. Kanahori, W. Cong, R. Fukuda, F. Tamari, and M. Suzuki, Handwriting interface for computer algebra systems, in *Proceedings of the Fourth Asian Technology Conference on Mathematics* (1999) pp. 291–300.
- [44] H. Wang, C. Pan, X. Guo, C. Ji, and K. Deng, From object detection to text detection and recognition: A brief evolution history of optical character recognition, *Wiley Interdisciplinary Reviews: Computational Statistics* **13**, e1547 (2021).
- [45] Mathpix, Mathpix OCR API for STEM, <https://mathpix.com/ocr> (accessed April 2024).
- [46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* **30** (2017).
- [47] M. Renze and E. Guven, The effect of sampling temperature on problem solving in large language models, *arXiv preprint arXiv:2402.05201* (2024).
- [48] Microsoft, Azure AI Services, <https://azure.microsoft.com/en-us/products/ai-services> (accessed June 2024).
- [49] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2021).
- [50] T. M. Fruchterman and E. M. Reingold, Graph drawing by force-directed placement, *Software: Practice and experience* **21**, 1129 (1991).
- [51] S. Epskamp, A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, and D. Borsboom, qgraph: Network visualizations of relationships in psychometric data, *Journal of Statistical Software* **48**, 1 (2012).
- [52] OpenAI, Hello GPT-4o, <https://openai.com/index/hello-gpt-4o/> (accessed June 2024).