Towards Scalable Automated Grading: Leveraging Large Language Models for Conceptual Question Evaluation in Engineering

Rujun Gao[1]   GRJ1214@TAMU.EDU
Xiaosu Guo[2]   XXG230002@UTDALLAS.EDU
Xiaodi Li[4]   LI.XIAODI@MAYO.EDU
Arun Balajiee Lekshmi Narayanan[3]   ARL122@PITT.EDU
Naveen Thomas[1]   NAVEENTHOMAS@TAMU.EDU
Arun R. Srinivasa[1]   ARUN-R-SRINIVASA@TAMU.EDU

[1] *J. Mike Walker '66 Department of Mechanical Engineering, Texas A&M University*

[2] *Computer Science Department, University of Texas at Dallas*

[3] *Intelligent Systems Program University of Pittsburgh*

[4] *Department of Artificial Intelligence and Informatics, Mayo Clinic*

Abstract

This study explores the feasibility of using large language models (LLMs), specifically GPT-4o (ChatGPT), for automated grading of conceptual questions in an undergraduate Mechanical Engineering course. We compared the grading performance of GPT-4o with that of human teaching assistants (TAs) on ten quiz problems from the MEEN 361 course at Texas A&M University, each answered by approximately 225 students. Both the LLM and TAs followed the same instructor-provided rubric to ensure grading consistency. We evaluated performance using Spearman's rank correlation coefficient and Root Mean Square Error (RMSE) to assess the alignment between rankings and the accuracy of scores assigned by GPT-4o and TAs under zero- and few-shot grading settings. In the zero-shot setting, GPT-4o demonstrated a strong correlation with TA grading, with Spearman's rank correlation coefficient exceeding 0.6 in seven out of ten datasets and reaching a high of 0.9387. Our analysis reveals that GPT-4o performs well when grading criteria are straightforward but struggles with nuanced answers, particularly those involving synonyms not present in the rubric. The model also tends to grade more stringently in ambiguous cases compared to human TAs. Overall, ChatGPT shows promise as a tool for grading conceptual questions, offering scalability and consistency.

Keywords: Large Language Model (LLM), ChatGPT(GPT-4o), automated grading, conceptual question evaluation, Spearman's rank correlation coefficient, RMSE

1. Introduction

Large Language Models (LLMs) have become powerful tools capable of understanding and processing natural human language, offering significant applications in various fields, including education. In educational contexts, LLMs have been employed to reduce the grading workload by automating the assessment of student responses, particularly in large classroom

GAO[1] GUO[2] LI[4] NARAYANAN[3] THOMAS[1] SRINIVASA[1]

settings (Bonner et al., 2023). In STEM subjects, a deep understanding of academic concepts is critical for student success. Conceptual questions, which often require constructed short-answer responses, provide valuable insights into students' understanding from both a breadth and depth perspective. However, grading such responses is challenging due to the inherent variability in answers and the subjective nature of grading by different instructors or teaching assistants (TAs).

This process is also time-consuming and requires a substantial investment of time and expertise (Kuechler and Simkin, 2010). This is particularly challenging in STEM disciplines like mechanical engineering where part of the learning is the use of precise technical language to communicate specific concepts.

Existing research on the automatic assessment of short-answer questions spans a variety of approaches (Gao et al., 2024), including natural language processing (NLP), machine learning, concept mapping, and, more recently, the application of models like Chat-GPT (Chang and Ginter, 2024; Burrows et al., 2015; Bonthu et al., 2021; Putnikovic and Jovanovic, 2023). The integration of LLMs, such as ChatGPT, into educational assessment has garnered significant attention due to their ability to interpret and evaluate complex text-based responses with greater consistency and scalability.

Automated grading systems have evolved from simple rule-based algorithms to more advanced models that leverage transformer-based architectures, such as BERT (Devlin et al., 2019) and GPT. These models have significantly improved the accuracy and reliability of automated grading systems, particularly in the assessment of open-ended questions such as essays and short answers. The automatic grading system offers several advantages, particularly in large classroom settings where managing a large volume of assessments can be overwhelming: 1) it enables more consistent grading and timely feedback; 2) it allows educators to shift their focus from grading to more instructional activities and personalized student support.

While most cited works focus predominantly on automated grading in scientific fields such as biology and computer science, there is a notable gap in the literature regarding its application in other engineering disciplines, including civil, mechanical, electrical, and chemical engineering. This study addresses this gap by exploring the feasibility of using ChatGPT, specifically the GPT-4o model, to grade conceptual questions within Mechanical Engineering (ME). Our experiment employs 10 quizzes and associated graded datasets from an undergraduate ME course in materials and manufacturing at Texas A&M University, a course rich in conceptual and technical language. To evaluate GPT-4o's grading performance, we use Spearman's rank correlation coefficient to compare the alignment between student rankings from the LLM-based grading and human grading, and the root mean square error (RMSE) to measure score discrepancies. Additionally, visualizations of the grading outcomes are included to provide an intuitive representation of the model's performance.

This study is motivated by the need to enhance student comprehension of engineering concepts through rapid, reliable feedback aligned with instructor rubrics in large classes. In a typical engineering class with over 100 students, providing timely and consistent feedback is notoriously challenging—yet timely feedback is crucial for effective learning (see, e.g., Barboza and da Silva (2016)). An AI-assisted grading and feedback approach enables a student-centered, formative grading process (Henderson et al., 2019), where students

engage interactively to deepen their understanding by revising answers and receiving re-evaluation (Gao et al., 2024). This approach, termed "supervised practice with feedback," is a cornerstone of mastery-based personalized learning, yet it is practically unfeasible in large engineering classes with hundreds of students. As highlighted in a systematic review on the benefits of automated grading by Hahn et al. (2021), automated grading and feedback offer several learning advantages, including consistency, reduced grading bias, and increased student participation.

The study aims to address two key research questions: (a) How effectively can ChatGPT be used to grade Mechanical Engineering conceptual questions? and (b) How does the in-context learning approach affect ChatGPT's grading performance?

2. Prior Work

In recent years, several automated methods have been explored for scoring and grading short answers. For instance, Yaneva et al. (2023) investigated the use of BERT with ad-ditional features to evaluate clinical essays in the medical field. Other studies have ex-plored BERT-based joint learning models for essay scoring and feedback generation (Wang et al., 2022), facilitating large-scale evaluation of short answers, albeit with limited per-sonalization. Distance metrics have also been employed in methods that assess and score essays automatically (Clark et al., 2019). Additionally, Gao et al. (2023) compared seven open-source LLM models for automated grading of text-based short-answer questions using correct/incorrect labels.

Beyond scoring, researchers have increasingly focused on feedback generation. For ex-ample, Lu and Cutumisu (2021) examined various methods for incorporating feedback. Studies have demonstrated the impact and utility of feedback derived from automated es-say evaluation systems (Liu et al., 2016). Recent research has also investigated automated assessment in interactive LLM-student environments (Han et al., 2023), while other studies have explored the application of argument mining approaches (Nguyen and Litman, 2018) for evaluation and scoring.

An alternative approach to evaluating student answers recently explored the use of prompt engineering, enabling students to demonstrate a certain level of understanding (Smith et al., 2024). Other approaches focus on categorizing these responses (Schneider et al., 2023). Ivanova and Handschuh (2024) conducted a comparative study with human annotators to evaluate the effectiveness of ChatGPT's automatic grading of student answers. Lagakis and Demetriadis (2024) proposed a multi-agent framework that integrates a reviewer/grader working alongside an LLM evaluator to achieve more accurate automatic grading results.

Building on prior research in automated grading, which has been applied to various short-answer assessments in fields such as medicine and language learning, this paper ex-amines the application of automated grading in mechanical engineering. Utilizing the latest models, this study provides multi-scale scoring aligned with instructor-provided rubrics to support and enhance student learning in engineering contexts.

Gao[1] Guo[2] Li[4] Narayanan[3] Thomas[1] Srinivasa[1]

3. Methodology

To investigate the extent to which large language models (LLMs) can be utilized for grading Mechanical Engineering conceptual questions, the methodology is structured as shown in figure 1. Human grading by teaching assistants (TA) serves as the benchmark. The TA grading process involves one TA and one grader, with their assessment trained and verified by the course's instructor based on specific grading rubrics that was created by the instructor.

To evaluate the performance of LLM-based grading, as outlined in research question one, we utlized the state-of-the-art ChatGPT model (GPT-4o, released on May 13, 2024). The prompts for automated grading were designed to align with the same rubric used by the TA.

Additionally, to address the second research question, we conducted two types of experiments: 1) GPT-4o zero-shot grading and 2) GPT-4o few-shot grading, to explore how in-context learning impacts the model's grading performance. For the few-shot experiments, we provided four grading example responses, along with the TA's corresponding scores, to the GPT model. The specific prompt details are included in Appendix B.
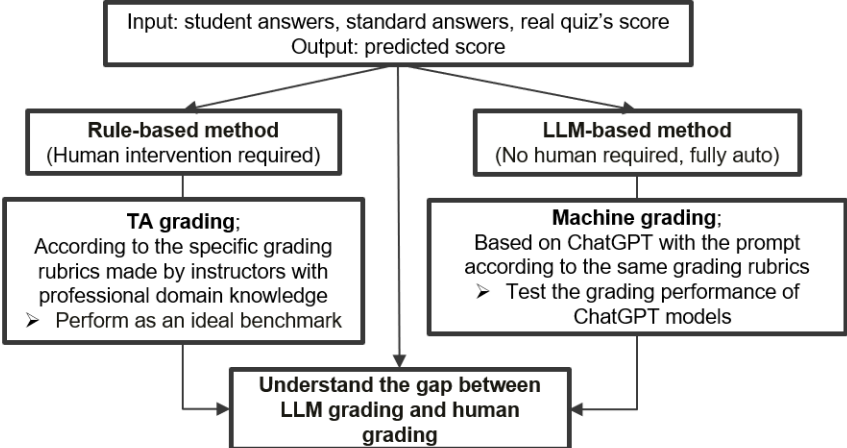

Figure 1: Flow of Evaluation

4. Experimental Datasets

4.1. Conceptual Question Example

The grading problems are sourced from the Mechanical Engineering undergraduate course (MEEN 361: Materials and Manufacturing in Design Laboratory) at Texas A&M University. The course covers material and manufacturing concepts and related lab experiments, including hardness testing, bending experiments, polymer tensile tests, fatigue testing, cold working and annealing experiments, and charpy impact testing, among others. This is a

1-credit lab course spanning 13-week Fall/Spring semester. After completing weekly lab experiments, students are required to answer 2-4 conceptual questions as a quiz, and their quiz scores contribute to their overall course grade. An example of a quiz problem and the corresponding grading rubric is shown in the Figure 2.

Hardness Test – 2 problems

Questions and Answers

1. What does hardness of a material mean (what is it measuring or what does it tell you about the material)? (Give at least **one** answer) (Usually 5 points)

   REFERENCE ANSWER -

   Hardness is the resistance of a material to local plastic deformation due to indentation loads. Even though it is material property (exists throughout the body) what we measure is this property at the surface of the sample, thus giving us a measure of wear or scratch resistance.

   Rubric schema

Attempted the problem to some extent even though missed many important points	50%	2.5
Major Point - Did they mention that we measure plastic deformation? But missed details	+30%	1.5
Minor Point - Did they mention that localized or indentation loads? OR Instead of indentation did they mention scratch resistance or Fatigue resistance?	+20% OR +15%	1 OR 0.75

Figure 2: Hardness Test Question Example

4.2. Datasets

For this study, we selected ten quiz problems from the MEEN 361 course to conduct grading experiments (problem descriptions are provided in Appendix A). Each question was answered by 225–230 students, with responses typically ranging from 3 to 7 sentences. The quizzes were graded independently by one teaching assistant (TA) and one additional grader, both trained by the course instructor using the instructor's scoring rubrics. The final scores, which serve as the "gold standard" or ground truth in this study, were verified by the instructor.

5. Evaluation metrics

Given the grading scale ranges from 0 to 5–10 points with several discrete scoring levels, we selected Spearman's rank correlation coefficient and Root Mean Square Error (RMSE) as the primary evaluation metrics for this study. In an educational context, rather than focusing solely on the exact score, we aim to investigate whether the machine grading can capture the same scoring trends as a human grader. To achieve this, we use Spearman's rank correlation coefficient to compare the ranking of students under the two grading methods.

RMSE, on the other hand, quantifies the score differences between machine and human grading and serves as a standard metric for evaluating multi-class classification tasks.

5.1. Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient assesses the extent of a monotonic relationship between student rankings under human grading and machine grading. Because graders tend to vary in the exact points awarded to the same answer, comparing raw scores directly is challenging. However, by assuming that a student's rank (i.e., relative position within the class, such as 1st, 2nd, 3rd, etc.) reflects a comparable level of knowledge across different graders, Spearman's rank correlation reduces the impact of these scoring variations. This provides a measure of consistency in the grading pattern. Spearman's rank correlation coefficient, also known as Spearman's ρ, is defined as follows:

$$\rho = \frac{6\Sigma d_i^2}{n(n-1)} \tag{1}$$

where ρ represents the Spearman rank correlation coefficient, d_i is the difference between the two ranks of each observation, and n is the number of observations. We adopt this metric to evaluate whether there is a consistent monotonic relationship between the grades assigned by the TAs and those generated by GPT-4o. If GPT-4o grading achieves perfect consistency with TA grading, the ranks would exhibit a perfect linear relationship ($\rho = 1$) (Gravetter and Wallnau, 2017).

Since Spearman's correlation uses ranks, it does not require dataset normalization, thereby avoiding the influence of differing score ranges across questions. Table 1 provides interpretation and strength reference ranges for both Spearman and Pearson correlation coefficients.

Table 1: Correlation Coefficient Strengths (Putnikovic and Jovanovic, 2023)

Range	Strengths
$0.00 - 0.19$	Very Weak
$0.20 - 0.39$	Weak
$0.40 - 0.59$	Moderate
$0.60 - 0.79$	Strong
$0.80 - 1.00$	Very Strong

5.2. Root Mean Square Error (RMSE)

In this experiment, RMSE is used to quantify the difference between LLM grading and TA grading (considered the gold standard). To ensure a fair comparison, datasets with varying score scales are normalized before calculating RMSE.

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \tag{2}$$

where y_i represents the actual values, $\hat{y}_i$ denotes the predicted values, and n is the number of data points (Bishop and Nasrabadi, 2006; Kuhn, 2013). The explanation of RMSE calculations on normalized datasets is shown in Table 2.

Table 2: RMSE interpretation for normalized data [0, 1]

Range	RMSE
$0.00 - 0.05$	Very small error
$0.05 - 0.10$	Small error
$0.10 - 0.20$	Moderate error
$0.20 - 0.30$	Large error
≥ 0.30	Very large error

6. Results

The experiment was conducted using both zero-shot and few-shot prompt engineering approaches, with Spearman's ρ and RMSE calculated for each. In the zero-shot prompt, no examples of student answers or grades were included, while the few-shot prompt provided a small set of student responses and corresponding grades to guide the model.

6.1. Zero-Shot Grading with GPT-4o

Overall, across the 10 datasets, the highest Spearman's ρ observed is 0.9387, and the lowest RMSE is 0.0830, while the lowest Spearman's rank correlation coefficient is 0.5488 and the highest RMSE is 0.2264.

Table 3 presents the Spearman's ρ and RMSE results. For Spearman's rank correlation, seven out of ten datasets exhibit coefficients over 0.6, with six datasets exceeding 0.75 and three exceeding 0.8. All datasets achieve at least 0.54, indicating that more than half of the questions (7 out of 10) demonstrate a strong and above correlation between GPT-4o automated grading and TA grading. Among these, three datasets show a very strong correlation, and all datasets achieve at least a moderate correlation.

For RMSE, three out of ten datasets exhibit small errors $(0.05 - 0.10)$, indicating that GPT-4o grading closely aligns with TA grading in these cases. The highest RMSE observed is 0.2264, with 9 datasets recording RMSE values below 0.2 and 1 dataset (Fatigue 2) mildly above 0.2, signifying mostly a moderate error level. This suggests that while GPT-4o grading is generally consistent with TA grading, there are some discrepancies.

Figure 3 and Figure 4 illustrate the LLM and TA grades for the datasets with the best and worst Spearman's ρ values, respectively. In these figures, the blue line represents TA grading, while the red line represents LLM grading.

6.2. In-Context Learning Approach: Few-Shot Grading with GPT-4o

As a comparison, we experiment with in-context learning strategy by providing relevant TA grading examples to GPT-4o to investigate the model's performance.

Gao[1] Guo[2] Li[4] Narayanan[3] Thomas[1] Srinivasa[1]

Table 3: Zero-shot results

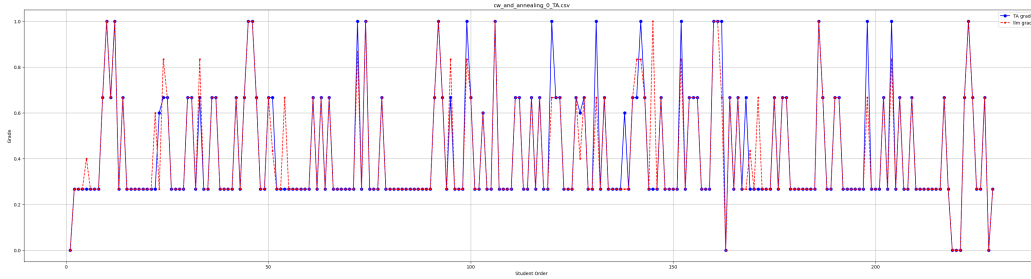| Question Name	Spearman's ρ	RMSE
Charpy Impact Q1	**0.6601**	0.1566
Cold Working and Annealing Q1	**0.9387**	**0.0975**
Fatigue Q2	**0.7694**	0.2264
Hardness Q1	**0.8183**	**0.0830**
Three-Point Bending Q1	0.5518	0.1566
Three-Point Bending Q2	0.5488	0.1629
Three-Point Bending Q4	**0.7524**	0.1758
Polymer Tensile Test Q1	**0.7574**	0.1622
Polymer Tensile Test Q2	0.5850	0.1819
Polymer Tensile Test Q4	**0.8911**	**0.0872**

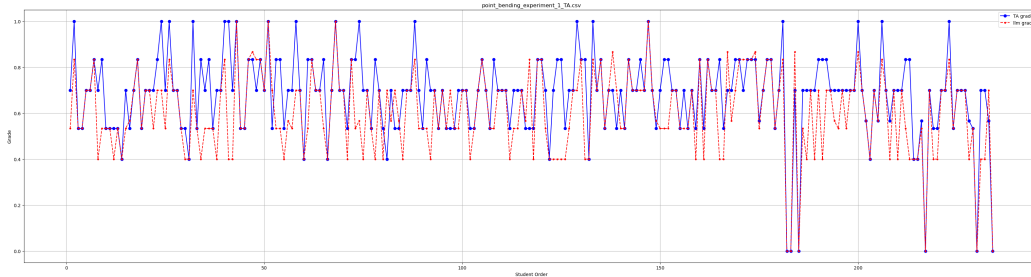Figure 3: Cold Working and Annealing Quiz Q1 grading results (best 0.9387)

Figure 4: Three-Point Bending Q2 grading results (worst 0.5488)

In the few-shot prompt setup, the highest Spearman's rank correlation coefficient achieved is 0.8990, with the lowest RMSE at 0.0998. The lowest Spearman's rank correlation coefficient observed is 0.5202, and the highest RMSE is 0.3733. Following the initial zero-shot prompt run, we selected four student answers and their corresponding grades—where there was an agreement between TA and LLM grading—from high to low scores as examples to include in the prompt.

Table 4 shows Spearman's rank correlation coefficient and RMSE results. For Spearman's ρ, seven out of ten datasets show coefficients above 0.6. However, values remain largely unchanged or decrease for most datasets. Specifically, three datasets exceed 0.75, which is three fewer than in the zero-shot setup, and only one dataset exceeds 0.8. All datasets record coefficients above 0.52, indicating that, although adding examples slightly decreased the correlation between GPT-4o grading and TA grading, all datasets still exhibit at least a moderate correlation.

In terms of RMSE, only one out of nine datasets shows a small error (0.05–0.10), a reduction from the three datasets observed in the zero-shot configuration. However, except Fatigue 2 RMSE, other 9 datasets still maintain RMSE values below 0.2, indicating a moderate level of error.

Table 4: Few-shot results

Question Name	Spearman's ρ	RMSE
Charpy Impact Quiz Q1	0.5585	0.1697
Cold Working and Annealing Quiz Q1	**0.8990**	0.1238
Fatigue Q2	**0.6913**	<u>0.3733</u>
Hardness Quiz Q1	**0.7069**	**0.0998**
Three-Point Bending Q1	0.5531	0.1508
Three-Point Bending Q2	0.5202	0.1661
Three-Point Bending Q4	**0.7564**	0.1596
Polymer Tensile Test Quiz Q1	**0.6880**	0.1815
Polymer Tensile Test Quiz Q2	**0.6120**	0.1579
Polymer Tensile Test Quiz Q4	**0.7638**	0.1308

Figure 5 and Figure 6 display the LLM and TA grades for the datasets with the highest and lowest Spearman's ρ, respectively. In these figures, the blue line represents TA grading, while the red line represents LLM grading.

6.3. Comparison of In-Context learning

A comparison of the results with/without in-context learning is provided in Table 5. Generally, datasets with an increased Spearman's ρ also exhibit a decrease in RMSE. However, only 3 out of 10 datasets show mild improvement, while the performance in 7 out of 10 datasets declines. This indicates that adding four graded examples did not significantly enhance the model's performance in grading short conceptual questions in mechanical engineering. Although some datasets showed slight improvement, it was minimal. Potential reasons for this outcome include:

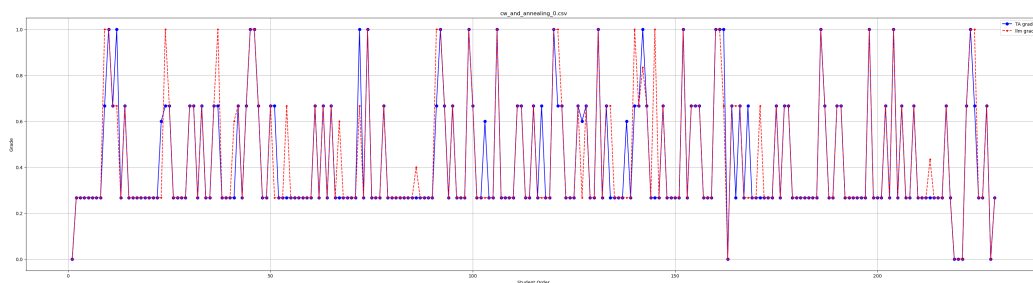Gao[1] Guo[2] Li[4] Narayanan[3] Thomas[1] Srinivasa[1]

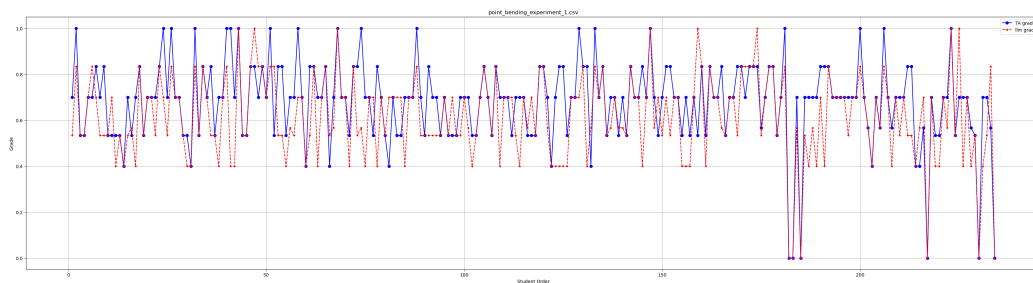Figure 5: Cold Working and Annealing Quiz Q1 grading results (best 0.8990)

Figure 6: Three-Point Bending Q2 grading results (worst 0.5202)

- Student responses to open-ended conceptual questions tend to vary widely, making it likely that the examples provided led the LLM in an unintended direction.

- With a dataset size of approximately 225 to 230, the addition of four examples may have been excessive, potentially leading to overfitting.[1]

- Furthermore, given the presence of a reference (or "golden") answer, additional examples may offer limited benefit, suggesting that adding examples could be counterproductive.

7. Grading Performance Analysis

Several key observations emerged when comparing GPT-4o's grading performance with TA grading:

1. **Strong Performance on Clear Scoring Rubrics:** GPT-4o model exhibited high accuracy when the scoring criteria were straightforward and clearly defined. However, its performance declined with more complex or nuanced questions, where precise interpretation was required.

---

1. We opted not to use random training and testing datasets, as this approach would be impractical. If the TA must grade a large number of questions to train the LLM, it may undermine the purpose of automating the grading process.

Table 5: Comparison of few-shot and zero-shot prompt results

Question Name	$\rho_{few} - \rho_{zero}$	$RMSE_{few} - RMSE_{zero}$
Charpy Impact Quiz Q1	-0.1016	0.0131
Cold Working and Annealing Quiz Q1	-0.0397	0.0263
Fatigue Q2	-0.0781	0.1469
Hardness Quiz Q1	-0.1114	0.0168
Three-Point Bending Q1	**0.0013**	**-0.0058**
Three-Point Bending Q2	-0.0286	0.0032
Three-Point Bending Q4	**0.0040**	**-0.0162**
Polymer Tensile Test Quiz Q1	-0.0694	0.0193
Polymer Tensile Test Quiz Q2	**0.0270**	**-0.0240**
Polymer Tensile Test Quiz Q4	-0.1273	0.0436

2. **Difficulty with Synonyms:** GPT-4o struggled with responses that used synonymous terms not explicitly covered by the rubric. In such cases, the model was likely to assign lower scores despite the semantic correctness of the answers.

3. **Stricter Scoring on Ambiguity:** In scenarios where the scoring criteria were unclear or ambiguous, GPT-4o generally applied more stringent grading compared to human TAs, who tended to give higher scores in similar situations.

4. **High Accuracy on High-Weight Items:** The model performed well in cases where specific answer choices carried significant weight in the rubric, accurately identifying and scoring these key elements.

In comparing human and LLM-based grading, we observe that LLM grading shows promise in evaluating short answer questions. In many cases, GPT-4o successfully assessed students' responses according to the rubric criteria provided. One notable observation was related to a rubric criterion focused on lab observation, which awards a minimum score to students who made an attempt on the question. We found that the LLM's grading of this particular rubric point remained consistent across responses without any misinterpretation. However, the LLM's grading of other rubric points was more variable, displaying a mix of accurate and less precise assessments.

In several instances, the grades assigned by the LLM aligned closely with those given by human evaluators, demonstrating its potential for effective grading. Nevertheless, there were notable instances where the LLM missed key details. For example, when students mentioned the "DBTT" (Ductile to Brittle Transition Temperature), the GPT-4o correctly identified and credited this point. However, it sometimes failed to recognize significant information when students explained that steel becomes brittle and breaks easily at cold temperatures. A similar issue occurred with the concept of ultimate strength: while GPT-4o reliably awarded points when students mentioned it, it occasionally granted points even when students did not reference it.

8. Confusion Matrix

This inconsistency underscores the need for instructors to provide clearer contextual guidance. Testing a small set of sample answers may help refine the LLM's ability to address these gaps effectively. As a grading assistant, this tool could offer instructors insights into answer patterns, supporting the grading process.

We selected the Charpy Impact Test Q1 as an example for confusion matrix analysis, shown in Figure 7. Most of the machine-graded scores aligned with human grading; however, in the inconsistent cases, a trend emerges where the machine scores are generally lower than human scores. It is also important to note that the true labels in this study are based on human grading, which can also include inherent biases since different TAs have subjective biases.
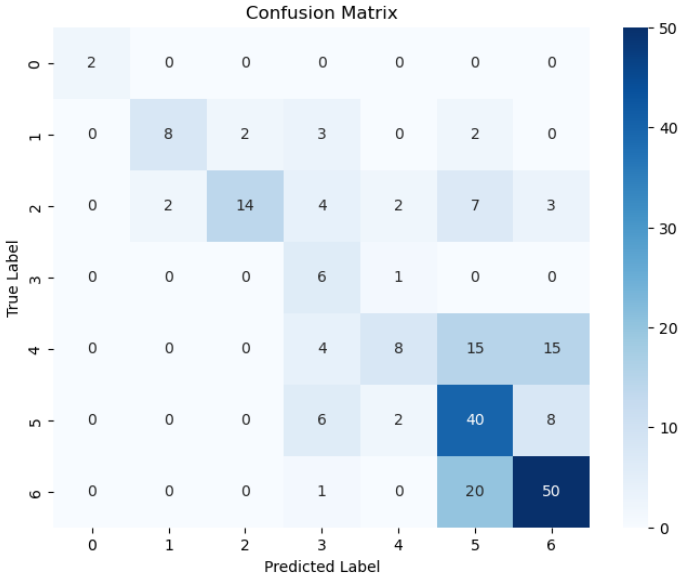

Figure 7: Charpy Impact Test Q1 - Confusion Matrix

9. Conclusion and Future Work

This study demonstrates the potential of using large language models (LLMs), specifically GPT-4o, for automated grading of conceptual questions in an undergraduate Mechanical Engineering course. In the zero-shot prompt setting, GPT-4o achieved a strong correlation with human teaching assistants (TAs), particularly when grading tasks were aligned with clear and straightforward rubrics. The model's performance remained consistent in recognizing key concepts, but it exhibited limitations in handling nuanced answers, especially when synonyms or ambiguous terms not explicitly addressed in the rubric were used. Moreover, the in-context learning approach, intended to improve performance by providing example answers, did not consistently enhance the model's accuracy and, in some cases, introduced variability in the grading outcomes.

Despite these challenges, GPT-4o holds significant promise as a scalable tool for grading in large classroom settings. It offers the potential to reduce the grading workload for educators while maintaining overall consistency with human grading. However, the model's tendency to grade more stringently and its struggles with less explicit criteria highlight the need for further refinements in rubric design and prompt engineering.

Future work will focus on addressing the limitations identified in this study. Specifically, efforts will be made to:

1) Refine Rubric and Prompt Design: Enhancing the clarity and specificity of rubrics and improving prompt engineering strategies to better guide ChatGPT model in interpreting nuanced and diverse student responses. This includes exploring ways to balance the number of few-shot examples provided to avoid overfitting or introducing noise.

2) Fine-Tuning LLMs for Domain-Specific Grading: Investigating the fine-tuning of open-source models on domain-specific datasets, particularly those related to Mechanical Engineering, to improve the open-source model's understanding of technical terminology and common student answer patterns.

3) Expand Dataset and Problem Types: Extending the study to include a broader range of conceptual questions and larger datasets to assess the generalizability of the findings across different types of conceptual questions and STEM disciplines.

By addressing these areas, future work aims to enhance the accuracy, scalability, and practical application of ChatGPT and similar LLMs in educational assessment.

Acknowledgments

References

Esdras Jorge Santos Barboza and Márcia Terra da Silva. The importance of timely feedback to interactivity in online education. In *Advances in Production Management Systems. Initiatives for a Sustainable World: IFIP WG 5.7 International Conference, APMS 2016, Iguassu Falls, Brazil, September 3-7, 2016, Revised Selected Papers*, pages 307–314. Springer, 2016.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Euan Bonner, Ryan Lege, and Erin Frazier. Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology*, 23(1):23–41, 2023.

Sridevi Bonthu, S Rama Sree, and MHM Krishna Prasad. Automated short answer grading using deep learning: A survey. In *Machine Learning and Knowledge Extraction: 5th IFIP*

TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17–20, 2021, Proceedings 5, pages 61–78. Springer, 2021.

Steven Burrows, Iryna Gurevych, and Benno Stein. The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25:60–117, 2015.

Li-Hsin Chang and Filip Ginter. Automatic short answer grading for finnish with chatgpt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23173–23181, 2024.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1264. URL https://aclanthology.org/P19-1264.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

Rujun Gao, Naveen Thomas, and Arun Srinivasa. Work in progress: Large language model based automatic grading study. *Proceedings of the 2023 IEEE Frontiers in Education Conference (FIE)*, 2023.

Rujun Gao, Hillary E. Merzdorf, Saira Anwar, M. Cynthia Hipwell, and Arun R. Srinivasa. Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence*, 6:100206, 2024. ISSN 2666-920X. doi: https://doi.org/10.1016/j.caeai.2024.100206. URL https://www.sciencedirect.com/science/article/pii/S2666920X24000079.

FJ Gravetter and LB Wallnau. Statistics for the behavioral sciences 10th. *Statistic for The Behavioral Science*, 2017.

Marcelo Guerra Hahn, Silvia Margarita Baldiris Navarro, Luis De La Fuente Valentín, and Daniel Burgos. A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *IEEE Access*, 9:108190–108198, 2021.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, et al. Llm-as-a-tutor in efl writing education: Focusing on evaluation of student-llm interaction. *arXiv preprint arXiv:2310.05191*, 2023.

Michael Henderson, Michael Phillips, Tracii Ryan, David Boud, Phillip Dawson, Elizabeth Molloy, and Paige Mahoney. Conditions that enable effective feedback. *Higher Education Research & Development*, 38(7):1401–1416, 2019.

Rositsa V Ivanova and Siegfried Handschuh. Evaluating llms' performance at automatic short-answer grading. 2024.

William L Kuechler and Mark G Simkin. Why is performance on multiple-choice tests and constructed-response tests not more closely related? theory and an empirical test. *Decision Sciences Journal of Innovative Education*, 8(1):55–73, 2010.

M Kuhn. Applied predictive modeling, 2013.

Paraskevas Lagakis and Stavros Demetriadis. Evaai: A multi-agent framework leveraging large language models for enhanced automated grading. In *International Conference on Intelligent Tutoring Systems*, pages 378–385. Springer, 2024.

Ming Liu, Yi Li, Weiwei Xu, and Li Liu. Automated essay feedback generation and its impact on revision. *IEEE Transactions on Learning Technologies*, 10(4):502–513, 2016.

Chang Lu and Maria Cutumisu. Integrating deep learning into an automated feedback generation system for automated essay scoring. *International Educational Data Mining Society*, 2021.

Huy Nguyen and Diane Litman. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Marko Putnikovic and Jelena Jovanovic. Embeddings for automatic short answer grading: A scoping review. *IEEE Transactions on Learning Technologies*, 16(2):219–231, 2023.

Johannes Schneider, Bernd Schenk, and Christina Niklaus. Towards llm-based autograding for short textual answers. *arXiv preprint arXiv:2309.11508*, 2023.

David H. Smith, Paul Denny, and Max Fowler. Prompting for comprehension: Exploring the intersection of explain in plain english questions and prompt writing. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 39–50, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706332. doi: 10.1145/3657604.3662039. URL https://doi.org/10.1145/3657604.3662039.

Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. *arXiv preprint arXiv:2205.03835*, 2022.

Victoria Yaneva, Peter Baldwin, Christopher Runyon, et al. Extracting linguistic signal from item text and its application to modeling item characteristics. In *Advancing Natural Language Processing in Educational Assessment*, pages 167–182. Routledge, 2023.

Appendix A. Conceptual Question Description and Grading Rubric

Charpy Impact Post-lab – 3 problems

1. During a very cold evening on April 14, 1912, the Titanic collided with a massive iceberg and sank in less than three hours. At the time, more than 2200 passengers and crew were aboard the Titanic for her maiden voyage to the United States. Only 705 survived. According to the builders of the Titanic, even in the worst possible accident at sea, two ships colliding, the ship should have stayed afloat for two to three days, which would provide enough time for nearby ships to help [Gannon, 1995]. What do you think made the builders have this wrong estimation of the sinking time of the ship? Doesn't cold help with making the ultimate strength higher?! (10 points)

The ship manufacturers were unknown to or disregarded the DBT behavior of low carbon steel. The temperatures of the sea at that time <u>was</u> well below the DBTT of the material. This reduced the impact resistance of the material which resulted in much less energy being absorbed during the collision. This would result in fracturing of the ship during the collision, sinking it.

<u>Rubric Schema</u>

- +3 points at a minimum.
- +2 points for recognizing that the ultimate tensile strength (moderately) increases at cold temperatures.
- + 3 points for recognizing that the DBTT was the primary cause of failure.
  The ship's hull was made of low carbon steel and the temperatures of the seawater temperatures at the time of its collision <u>was</u> well below the DBTT for this steel.
- +2 points for a relevant explanation of how DBTT affects the properties of steel.

Figure 8: Dataset 1: Charpy Impact Quiz Q1

1. In this experiment we used the term "activation energy" for material. Would this energy with 30% reduction be larger or smaller than for one with 60% reduction? Why? 7.5 points

- The activation energy for the 30% reduction would be greater than for 60% reduction because it has less stored energy. Therefore, more energy would be required to start the annealing process. the dislocation entanglement in the 30% reduction will be much lower than the 60% scenario. Thus, we need a lot more energy to initiate the dislocations 2 get unentangled or recrystallized. In the case of 60% reduction in area there is already a lot of plastic deformation recrystallization and grain deformation that has occurred in the specimen. Thus, the material starts to recrystallize at much lower energy to reach a lower energy state.

<u>Rubric Schema</u>

Attempted a point to some extent.	2
Major Point 1– Did the student mention the deactivation energy will be greater?	3
Major Point 2- did the student mention that 30% reduction will have lower amount of entanglement which results in higher energy required for the recrystallization process? Give partial grade if you feel so	2.5

Figure 9: Dataset 2: Cold Working and Annealing Quiz Q1

2. You, an accident investigation expert, are assigned a case where the insurance company suspects that the car company did not correctly account for the fatigue failure when designing the axles of their cars. Hence, you are given a sample of a broken front-right axle to determine if the axle failure was a fatigue failure. The insurance company mentions that the axle broke when the car traveling at 45 mph hit a pothole in the road.

   Then, how will you most efficiently determine if the failure of the axel was a fatigue failure or an effect of the hitting pothole at high speed? [7.5 points]

   ANSWER

Fatigue can be categorized as mechanical fatigue or thermal fatigue. When designing parts for elevated or high-temperature environments, it is crucial to account for the thermal fatigue occurring in the materials in addition to any mechanical fatigue they may undergo.

Fatigue failure is brittle type and the cracked region we'll have beach marks and straightens indicating fatigue failure. the beach monks are important to identify growth of cracks in the axle.

failure due to the pothole would have cracks plastic in nature. It would show significant plastic deformation before the development of cracks.

The method of identification would be to analyze the cracks, identify benchmarks to see crack development. Also check and see the cracks are brittle in nature or plastic.

Rubric Schema

This is for each point in the answer.

Attempted all points to some extent.	
Major Point 1– Did the student describe the features of fatigue failure? namely brittle type failure and benchmarks	2.5
Major Point 2- did the student explain failures due to a pothole in some form?	2.5
did the student give a method of identification that is valid	2.5

Figure 10: Dataset 3: Fatique Quiz Q2

Hardness Test – 2 problems

Questions and Answers

1. What does hardness of a material mean (what is it measuring or what does it tell you about the material)? (Give at least **one** answer) (Usually 5 points)

   REFERENCE ANSWER -

   Hardness is the resistance of a material to local plastic deformation due to indentation loads. Even though it is material property (exists throughout the body) what we measure is this property at the surface of the sample, thus giving us a measure of wear or scratch resistance.

   Rubric schema

Attempted the problem to some extent even though missed many important points	50%	2.5
Major Point - Did they mention that we measure plastic deformation? But missed details	+30%	1.5
Minor Point - Did they mention that localized or indentation loads? OR Instead of indentation did they mention scratch resistance or Fatigue resistance?	+20% OR +15%	1 OR 0.75

Figure 11: Dataset 4: Hardness Quiz Q1

1. Why do we use a Weibull distribution for the chalk samples instead of normal distribution? (Explain how the material and manufacturing affect it) (5) (in the lab)

ANSWER -

Industrially manufactured chalks have high variability in their material properties. The final product ends up with many internal flaws. "The Weibull distribution is an indicator of the variability of strength of materials resulting from a distribution of flaw sizes. This behavior results from the critical sized flaws in materials with distribution of flaw size (i.e., failure due to the weakest link in a chain)".

Moreover, the strength has the known threshold strength of 0 MPa. A normal distribution of such high variability data and known threshold will predict possibility of failure strengths below zero MPa.

Rubric schema

Attempted the problem to some extent even though missed many important points	2
Major Point 1– Did the student mention the variability in the failure strengths?	1.75
Major Point 2- Did the student mention known threshold strength of zero MPa	1.25

Figure 12: Dataset 5: Three-Point Bending Q1

2. Would you recommend using a 4-point bending test and why (only one answer required)? (7.5) (Beyond)

ANSWER -

    a. It is sometime recommended to use a 4-point bending test since the maximum moment is the same between the two internal loading points which in turn will minimize the variability in failure strength as now we look for the critical crack in a region then a point for three-point bending.

    b. A four-point bend test is also a pure bending test. There is no shear force between the internal loading points and therefore the only load applied is a bending moment. This isolates the effect of the bending moment instead of having he combined loading in the 3-point bending test.

Rubric schema

Attempted the problem to some extent even though missed many important points	3
Major Point 1– Did the student mention that four point bending is recommended?	1
Major Point 2- Did the student mention the beam is in pure bending or no shear?	2.25
minor point- Did the student mention that now That the variability will reduce because the maximum movement acts over a region rather than a single point?	1.25

Figure 13: Dataset 6: Three-Point Bending Q2

18

4. Throughout the experiment and on the Force-Displacement graph, there are often local drops and followed by a general rising trend in force, making the graph look like a saw. Why do you think those local reductions in force happen? (5) (in the lab)

ANSWER -

These local reductions are due to the micro crack growths encountering an internal flaw or voids in the chalk, reducing the force temporarily and propagation of the crack, until it encounters material again increasing the force required for further growth.

Rubric schema

Attempted the problem to some extent even though missed many important points	2.5
Major Point – Did the student mention that cracks grow and encounter voids or flaws?	2
Minor Point Did the student mention that the cracks are micro and localized and indicate that even it encounters material the force increases?	0.5

Figure 14: Dataset 7: Three-Point Bending Q4

Polymer Tensile Experiment – 4 problems

1. During the stress relaxation: polymers remain under a fixed strain. Using the definition of the Relaxation Time constant; "a material property that defines the time needed for the instantaneous stress to decrease to 37% of the initial stress", explain why the change in the load is greater at the beginning of the two minutes? 5 points

ANSWER -

Relaxation follows an exponential decay. As the material relaxes, the rate of decay is proportional to the load itself.
The polymers tested in this experiment exhibit a viscoelastic behavior, i.e., time-dependent material behavior. The equation for determining the instantaneous stress as a function of time (t), under constant strain, is given by

$\sigma(t) = \sigma_0 \exp(-\lambda t),$

where, σ_0 is the original stress and λ is a material constant. Thus, we have an exponentially decaying stress behavior, considering a relaxation time constant.

The change in load is large at the beginning, then exponentially decays to a more constant value. This is because, under constant strain, the molecule chains of the polymers begin to rapidly untangle and the specimen elongates, thus becoming more ductile and decreasing the overall stress. As the chains become less and less untangled, the specimens lose their ductility. Thus, the change in loading becomes lesser than at the beginning. The relaxation time constant is a way to determine how fast this specimen reaches a certain point, defined as 37% of the initial stress. This serves as a method to compare the speed at which various materials stress-relax.

Figure 15: Dataset 8: Polymer Tensile Test Quiz Q1

Rubric Schema

Attempted the problem to some extent even though missed many important points	1.5
Major Point 1– Did the student mention that the force drops exponentially?	2.5
Major Point 2- did the student mention that the chain targeting untangled which changes the force to a new value with time?	1

Figure 16: Dataset 8: Polymer Tensile Q1 Rubric

2. The glass transition temperature (Tg) is the temperature for which a material goes from a viscous or rubbery state to a more solid-like state. Polylactic acid (PLA) is a polymer commonly used in 3D printing applications. It has a relatively low Tg of 60-65 °C. Knowing this, would using PLA for applications such as a flexible spatula (useful when frying eggs and such) be a good idea? Why? If not, what property should be desired when choosing another polymer? 8 points

ANSWER -

No, PLA would not do well in this application. Because PLA's Tg is around 65 C and the face of a frying pan can reach temps much higher than that, it is likely that the shape of the spatula would deform due to a transition to a viscous type of behavior. This can also be seen in the forming process of PLA in which the material is heated (to around 180 C) and then layered up using a 3D printer. Silicone is a much better material for this application with a Tg of -125 C (not +125 C!).

The property desirable is high melting temperature and very low Tg. It should be rubbery at room temperature and yet melt at significantly high temperature (at least greater than 250-300 C). If PLA is used, it will undergo plastic deformation when exposed to the pan and will retain the deformed shape when cooled to room temperature. Silicone has Tg of -125 C and melting point of ~1400 C which makes it a good candidate for flexible spatula to be used for cooking.

Figure 17: Dataset 9: Polymer Tensile Test Quiz Q2

Rubric Schema

Attempted the problem to some extent even though missed many important points	3
Major Point 1– Did the student mention that PLA won't work because of low melting point?	3
Major Point 2- did the student mention that low TG value and high melting point is desired?	2

Figure 18: Dataset 9: Polymer Tensile Q1 Rubric

4. Are polymers elastic, viscoelastic or viscous? What does this mean? 4 points

ANSWER -

Polymers are a viscoelastic material. They exhibit both elastic (solid like) and viscous (fluid like) behavior when undergoing deformation at different temperatures and conditions.

Rubric Schema

Attempted the problem to some extent even though missed many important points	1
Major Point 1– Did the student mention that polymers are viscoelastic?	2
Major Point 2- did the student mention that polymers show viscous and elastic behavior under different temperatures?	1

Figure 19: Dataset 10: Polymer Tensile Test Quiz Q4

Appendix B. Prompt example

```
general_prompt = f"""You asked a student the following question: '{question}'.
Grade the student answer below using the rubrics.
For each SUM rubric below, grade the answer with the rubric by deciding if the answer
  complies with the rubric's behavior: Yes or No. If yes, give the full points for
  the corresponding rubric. If not, give 0 points.
```

Figure 20: Zero-shot grading experiment prompt

```
general_prompt = f"""You asked a student the following question: '{question}'.
Grade the student answer below using the rubrics. If Standard Answer does not include all rubrics points, grading is following rubrics.
 If terms in rubrics does not appear but student answer explains the same meaning, also think correct and give points.
For each SUM rubric below, grade the answer with the rubric by deciding if the answer complies with the rubric's behavior: Yes or No. If
 yes, give the full points for the corresponding rubric. If not, give 0 points.
Then choose maximum one OR rubric (can be one or zero) from the set of OR rubrics below, for which the answer satisfies its expected
 behavior. For the selected rubric and only this one, give full points. Otherwise give 0 points.
The following are some examples of students' answers and their grades for reference: (1) answer: "The builders did not understand that
 the steel in the ship's hull had a ductile to brittle transition temperature, which would cause the material to fracture under impact at
  cold temperatures. Although the ultimate strength of the material is higher at low temperatures, the impact toughness is lower causing
  the material to fail under impact." & grade: 10. (2) answer: "The builders may have conducted tests / made assumptions based on "room
  temperature" metal. This would have affected the type of failure expected during an accident, therefore the time to sink. Even if the
  ultimate strength of metal is higher at cold temperatures, it cannot yield a considerable amount nor absorb much energy." & grade: 7.
  (3) answer: "The builder's assumed that ultimate strength is the only property that matters. The brittleness of the metal used caused
  the ship to fall apart catastropically." & grade: 3.
```

Figure 21: Few-shot grading experiment prompt