Work in Progress: Large Language Model based Automatic Grading Study

Rujun Gao
J. Mike Walker '66 Department of Mechanical Engineering.
Texas A&M University
College Station, USA
grj1214@tamu.edu

Naveen Thomas
J. Mike Walker '66 Department of Mechanical Engineering.
Texas A&M University
College Station, USA
naveenthomas@tamu.edu

Arun Srinivasa
J. Mike Walker '66 Department of Mechanical Engineering.
Texas A&M University
College Station, USA
arun-r-srinivasa@tamu.edu

***Abstract*—We investigated the capability of Large Language Models (LLMs) for grading short answer questions and studied different auto-grading schemes for evaluating student responses to conceptual questions in a mechanical engineering statics course. We compared the ability of seven Natural Language Processing (NLP) systems to score text-based answers as Correct/Incorrect and numerically, with human-supported Rules-based grading as a benchmark. We collected the instructor-provided answers, anonymized student answers, and their grades for this study. The findings reveal that the Large Language Model (LLM) based grading systems exhibit commendable precision in binary evaluations. However, within the spectrum of error classifications, the LLM-based grading systems exhibit a pronounced rate of false positives, a scenario less than ideal in an educational context. Considering that the technical terms in the instructors' answers are a primary factor in grading, our forthcoming research endeavors to embed keyword detection within the LLM-based automatic grading framework to mitigate the incidence of false positives. Thus, we investigated the ability of the standalone LLM-Vicuna to identify important keywords in an answer in the context of the mechanic's course. Our preliminary observations indicate that Vicuna accurately identifies the keywords in the answers, but the results are not yet repeatable due to the stochastic nature of the model.**

Keywords—automatic grading, natural language processing, large language model, ChatGPT, Vicuna, mechanics

I. INTRODUCTION

This Work in Progress paper studies different approaches to grading student responses to short answer questions in Mechanical Engineering using NLP methods. Effective education rely not only on imparting knowledge but also on assessing student performance and providing valuable feedback. However, in the context of large classes, grading and delivering individualized feedback can be both challenging and time-consuming. Instructors in mechanical engineering generally choose two different approaches to address the issue. A slow process of grading small number of long answers using human graders giving some feedback but not timely. And faster approaches such as multiple-choice or numerical questions that offer the advantage of quick grading, but often fall short in capturing the nuanced understanding of students [1]. Moreover, in such approaches we evaluate only the final result, making it rather difficult to ensure that the students do not copy. This approach still remains the most common approach for evaluating the understanding of the students, e.g., mechanics baseline test [2], force concept inventory [3], etc. owing to its ease of implementation and its objectivity [1]. One way of improving this issue could be to ask the student to add a reason for their choices which may improve the feedback but makes the grading process tedious. Thus, a quick and easy grading assistive tool for short answer questions will widely improve the learning of the students and reduce the workload on the instructors and assistants. Recently NLP methods have shown promise in the way in which they are able to "comprehend" such written texts and provide useful feedback. However current practice developed and used in a variety of content domains, such as mathematics, science, and language testing, few of them are developed for engineering education [4]. To the best of our knowledge, research on automatic grading in Mechanical Engineering especially focusing on the demands and needs of evaluating student conceptual understanding in Mechanics courses is lacking.

Recognizing the need for a more comprehensive and personalized assessment, grading answers relies heavily on the presence of technical and comparative keywords and phrases. However, only a few studies considered inducing keywords into their grading system. According to the conclusion of a review by Cekiç & Bakla (2021) that focused on the features of fourteen formative assessment tools and the types of assessment items supported, more than half of the tools supported open-ended question formats, only two offered automatic grading features using Artificial Intelligence (AI) or teacher-provided keywords [5]. Traditional approaches for identifying keywords involve the utilization of term frequency-inverse document frequency (TF-IDF), Alammary (2021) [6] proposed a modified TF-IDF model to classify questions in Arabic according to Bloom's taxonomy. The key process of this automatic classifier is to use TF-IDF based method to extract features from questions before classification, and further developed an automatic assessment tool - "LOsMonitor" [7] to classify the cognitive level of questions based on Bloom's taxonomy with the utilization of text mining and machine learning (ML) techniques. These techniques may not be directly applicable to the grading context, especially the domain-knowledge based grading, considering that the process of feature extraction does not take into account

Funded by National Science Foundation (NSF) Grant # 2022275

2023 IEEE Frontiers in Education Conference (FIE) | 979-8-3503-3642-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/FIE58773.2023.10343006

Authorized licensed use limited to: Texas A M University. Downloaded on September 02,2024 at 22:50:45 UTC from IEEE Xplore. Restrictions apply.

the context. However, transformer models, known for their contextual understanding, have the potential to identify relevant keywords within the course context. Utilizing large language models like ChatGPT for answer grading can be effective due to their training on extensive datasets and number of parameters. Nonetheless, concerns about data privacy when using models like GPT necessitate the development of local-scale LLMs such as Vicuna [8], Alpaca [9], etc., to achieve the desired results. In this study, we explore the capabilities of the Vicuna model in generating consistent keywords through prompting in Mechanics.

Here we try to address the research questions (a) How well NLP methods perform auto-grading in Mechanics? (b) How well do LLMs extract keywords relevant to Mechanics from a student's response?

The methodology of evaluating LLM Auto-Grading and Rules-based grading is discussed in section II. Following this, the methodology of the keyword extraction using Vicuna is discussed in section III. This is followed by the results of their performance and the conclusions in sections IV and V respectively.

II. Auto-grading Methodology

A. Datasets

The dataset used in this preliminary study consists of two types: quizzes and activities. The quiz dataset, comprising 70 students, focuses on conceptual questions with answers that can be graded as either correct or incorrect, resulting in a binary classification task. For instance, a typical question from this dataset asks students, "Describe the three rules for specifying inputs to a beam." On the other hand, the activity dataset involves more complex activities, with varying student participation ranging from 85 to 95 students. An example activity question prompts students to provide a strategy for analyzing a system being discussed in class. Activities 1 and 2 are scored on a scale of 0 to 5, while Activity 3 on 0 to 6 points.

In all these methods of grading, we use human intervention of providing a reference answer to the models. The automatic grading was performed using two types of approaches: Rule-based and LLM-based. In the LLM-based approach, the standard answer and student answers were tokenized using different large language models, and grades were assigned based on the "similarity" of the sentences. In the Rule-based approach, a set of keywords for the grading was identified from the standard answer by the grader, and a set of rules were employed on the presence of these words in the answers to grade the student answers.

B. LLM-based Approach

Seven different NLP models were chosen in this approach To evaluate the answers submitted by students, the standard answer and student answers were tokenized and converted into numerical vectors. Cosine similarity was then calculated between every student answer vector e_1 and the reference answer vector e_2. By computing the cosine similarity between these vectors, as in (1), we can determine the degree of semantic relevance between the two sets of answers. Finally, we use a classifier to determine which score range a particular response falls within, and we scale and round the results to produce a predicted score for each student's quiz response. Live models were fine-tuned using mechanics textbooks; however, they were not specifically trained on the question-answer dataset used in this study. The models used for the approach includes BERT, T5, InferSent, ConSERT, PromCSE, Universal Sentence Encoder (USE), Sentence Transformers (all-MiniLM-L6-v2).

$$similarity\,(e_1, e_2) = \frac{e_1^T \cdot e_2}{||e_1||\,||e_2||} \quad (1)$$

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained model introduced by Google. It captures bidirectional contextual information. It is fine-tuned using the PDF versions of the course textbooks to improve its performance. We also compared different BERT models and compared them with our model. T5 (Text-to-Text Transfer Transformer) is based on the Transformer architecture and focuses on text-to-text tasks. It employs a different cloze method in comparison to BERT. InferSent, a model developed by Facebook researchers, utilizes a supervised learning method to learn sentence embeddings with semantic representations for sentences in English. ConSERT (Contrastive Learning for Sentence Representations) is a self-supervised contrastive learning framework. It transfers sentence representations to downstream tasks to address the issue of collapsing representations in BERT. The Universal Sentence Encoder (USE) is based on a Transformer encoder and Deep Averaging Network. It encodes text into fixed-shape high-dimensional vectors. PromCSE (Prompted Contrastive Learning for Sentence Embeddings) incorporates soft prompt layers and an Energy-based loss term to prevent overfitting during model fine-tuning to improve upon SimCSE. Sentence Transformers, specifically the all-MiniLM-L6-v2 variant, utilize transformer-based architectures to generate sentence embeddings. These embeddings capture the semantic meaning of sentences. All of these models were used to autograde the student answers by comparing them with the standard answer.

C. Rule-based Approach

Discussions with instructors and teaching assistants revealed that, graders rely on particular technical keywords in students' responses while grading. To establish a benchmark for comparison, we developed a Rule-based method, where we carefully selected relevant keywords from standard answers based on input from the graders. These selected keywords were then employed to formulate scoring rules for evaluating the answers. Initially, we examined the grammar and fluency of the sentences and then grade them based on specific sequences of technical keywords. For binary answer questions, a single rule was assigned for each quiz, while for the complex conceptual questions (activities), multiple keywords were identified for each credit requirement. (This study is still in progress)

III. Keyword Extraction Study – LLM

Keywords were identified as a critical component in grading student answers and in the earlier study was identified with the help of the instructors. It was also observed that ChatGPT was efficient in identifying the right keywords from the standard answer. In this study, we analyze the capability of the open-source LLM - Vicuna in identifying the right keywords in the

Authorized licensed use limited to: Texas A M University. Downloaded on September 02,2024 at 22:50:45 UTC from IEEE Xplore. Restrictions apply.

context of Mechanics from the standard answers provided and compare it with the instructor generated keywords relevant to the rubric.

For this study a typical prompt in the form of the following structure is adopted:

"### Instruction: Identify the keywords in the context of [an Undergraduate Mechanics course] from the following answer.

Input: [Answer]".

In order to finetune the model to the given task, we sent two examples with instructor identified keywords before the analysis. This procedure is to make Vicuna 'learn' the context and the preferred format of the output. The later results were collected with different prompts.

IV. RESULTS AND DISCUSSION

A. LLM-based automatic grading

Table I shows the weighted-average F1 scores for Quiz 1-5 datasets which have binary scoring. The F1 scores are calculated using the equation

$$F1 = 2\frac{p \cdot r}{p + r} \quad (2)$$

where, p is the precision of the grading scheme, which is the percentage of True positives among the cases with positives in the scheme and r is the recall of the grading scheme, which is the percentage of True positives identified by the scheme among the True cases. PromCSE method achieve the highest F1-scores compared to other methods. Conversely, the USE model performs poorly in grading binary datasets. BERT, PromCSE, and all-MiniLM-L6-v2 show stable and promising performance.

Table II shows the performance of the schemes for grading the conceptual questions (activities). The performance (accuracy) of the schemes is computed as the Root Mean Square Error (RMSE) of the scheme's scores with the human graded scores. Here we observe that the PromCSE model continue to exhibit superior performance in continuous scoring problems. Whereas models such as InferSent, BERT, and all-MiniLM-L6-v2 demonstrate instability when applied to different tasks.

Overall, when it comes to complex multi-point conceptual comprehension problems, the performance of both approaches is not as strong as in Correct/Incorrect problems. The scoring accuracy significantly decreases, highlighting the need to explore a more intricate grading framework in future research. From a model evaluation standpoint, among the LLM models, PromCSE shows potential in automatic grading.

TABLE I. WEIGHTED-AVERAGED F1 SCORE OF BINARY DATASETS

Model	*Quiz 1*	*Quiz 2*	*Quiz 3*	*Quiz 4*	*Quiz 5*
InferSent	0.83	0.97	0.98	0.78	0.86
USE	0.69	0.83	0.76	0.28	0.23
BERT	0.93	0.93	0.91	0.90	0.95
T5	0.83	0.89	0.95	0.75	0.46
ConSERT	0.73	0.97	0.91	0.88	0.8
PromCSE	0.94	0.99	0.95	0.96	0.99
all-MiniLM-L6-v2	0.87	0.94	0.93	0.91	0.96

TABLE II. RMSE OF MULTI-CLASS DATASETS

Model	*Activity 1*	*Activity 2*	*Activity 3*
InferSent	2.0124	0.931	0.9997
BERT	2.2433	0.8944	1.3992
T5	1.4065	1.8439	1.777
USE	1.4833	0.9329	1.6146
ConSERT	1.6461	1.1832	1.2896
PromCSE	1.1032	0.7684	1.0615
all-MiniLM-L6-v2	2.2143	0.8944	1.2226

B. Precision analysis (Confusion matrix)

In the task of automatic grading, we hope that AI can help teachers filter out students who have mastered the knowledge so as to save more time and focus on students who have not understood the concepts. Thus, the confusion matrix of each method in Table I is analyzed under each binary dataset. Statistical results show that for the LLM-based method, the number of False positives (FP) is much greater than False negatives (FN) among its error types. Fig. 1 shows the average confusion matrix for the BERT model, as an example, for the binary quizzes (Q1-Q5). We observe a similar pattern in other NLP methods as well. In the context of grading, we define the precision of the schemes (false positives) to be the ratio of the cases the scheme provides a higher score to the answer in comparison to actual human grading. Here we observe that even though NLP methods are good at identifying the True cases with high accuracy, the precision of the identification is low.

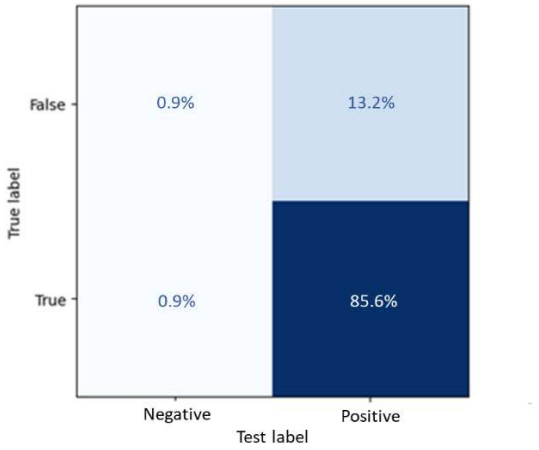

Fig. 1. Confusion matrix showing the average performance of BERT model in grading binary quizzes (Q1-Q5)

C. Keywords Extraction

In light of the prevalent paradigm within mechanical engineering pedagogy wherein scoring predominantly hinges on keyword identification, we aspire to integrate keyword recognition capabilities into subsequent algorithmic iterations to augment the efficacy of automatic grading. Thus, this study investigated the performance of the latest LLM - Vicuna in extracting keywords under a mechanical domain context. An example of the keywords collected is given below:

Authorized licensed use limited to: Texas A M University. Downloaded on September 02,2024 at 22:50:45 UTC from IEEE Xplore. Restrictions apply.

"### Instruction:
Identify multiple keywords in the context of Material Science course from the following answer.

Input:
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.

Response:
local reductions, micro crack growths, internal flaw, voids in the chalk, temporary reduction, propagation of the crack, material, force, further growth, increase in force."

We observe that Vicuna is able to identify a set of keywords that includes the most important keywords in the answer. However, we also observe that on repeated prompts the percentage of keywords identified drops. The number of keywords generated is observed to be independent of the number of prompts. This issue of the behavior of LLM on multiple prompt is being studied. Table III. shows the percentage of keywords extracted by Vicuna on repeated trials. The percentage of keywords extracted is exhaustive in Trial 1, but the percentage of words extracted drops for longer answers (A1, A2) with around 10 keywords.

TABLE III. PERCENTAGE OF KEYWORDS EXTRACTED BY VICUNA ON PROMPTING USING DIFFERENT ANSWERS.

Answers	Percentage of keywords extracted		
	Trial 1	*Trail 2*	*Trial 3*
A1	100	70	70
A2	100	70	60
A3	100	100	100
A4	100	100	80
A5	100	100	100

V. CONCLUSIONS AND FUTURE WORK

In the first study, the data were created with conceptual questions from the instructors, rather than restricting them to any specific type of questions. Our focus was on exploring the potential of NLP models for automated grading with respect to a rules-based benchmark. The evaluation conducted identified the significance of correctly identifying and utilizing technical words in grading engineering answers. Upon analyzing the precision of the grading methods, we found that NLP methods have a high percentage of false positives in their grading. This makes NLP methods for grading unfavorable, as we may not be able to successfully identify students who need feedback to improve their conceptual understanding.

Standalone versions of LLMs like Vicuna show promise in their capability to identify keywords without the need for human intervention. However, the repeatability of the results in such stochastic models needs to be further explored. Furthermore, the team is also exploring the engineering of the prompts or the training of the model. Repeatable extraction of keywords relevant to the course will help in developing an explainable comparison between the students' responses and the instructor-provided answer for assistive grading.

ACKNOWLEDGMENT

This research is being conducted with the help of National Science Foundation Grant # 2022275.

REFERENCES

[1] W. L. Kuechler and M. G. Simkin, "How well do multiple choice tests evaluate student understanding in computer programming classes?," vol. 14, no. 4, p. 389, 2003.

[2] D. Hestenes and M. Wells, "A mechanics baseline test," vol. 30, no. 3, pp. 159–166, 1992.

[3] D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," vol. 30, no. 3, pp. 141–158, 1992.

[4] O. L. Liu, C. Brew, J. Blackmore, L. Gerard, J. Madhok, and M. C. Linn, "Automated scoring of constructed‐response science items: Prospects and obstacles," Educational Measurement: Issues and Practice, vol. 33, no. 2, pp. 19-28, 2014.

[5] A. Çekiç and A. Bakla, "A Review of Digital Formative Assessment Tools: Features and Future Directions.," vol. 8, no. 3, pp. 1459–1485, 2021.

[6] A. S. Alammary, "Arabic Questions Classification Using Modified TF-IDF," *IEEE Access,* vol. 9, pp. 95109-95122, 2021, doi: 10.1109/ACCESS.2021.3094115.

[7] A. S. Alammary, "LOsMonitor: A Machine Learning Tool for Analyzing and Monitoring Cognitive Levels of Assessment Questions," *IEEE Transactions on Learning Technologies,* vol. 14, no. 5, pp. 640-652, 2021, doi: 10.1109/TLT.2021.3116952.

[8] W.-L. Chiang et al., "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," 2023.

[9] R. Taori et al., "Stanford alpaca: An instruction-following llama model", 2023, https://github.com/tatsu-lab/stanford_alpaca.

Authorized licensed use limited to: Texas A M University. Downloaded on September 02,2024 at 22:50:45 UTC from IEEE Xplore. Restrictions apply.