Evaluating LLMs' Performance At Automatic Short-Answer Grading

Rositsa V. Ivanova[1], Siegfried Handschuh[1]

[1]*University of St. Gallen, Switzerland*

Abstract

In recent years, the use of Large Language Models (LLMs) has become more accessible and wide-spread. With a free-of-charge access types people have began applying the models to various tasks beyond the task of next-word prediction. In an exploratory study, we take a closer look at the use of LLMs for Automatic Short Answer Grading. We compare the grading of short-answer tasks by two human graders to this of an LLM. We discuss the results and present examples of observed short-comings in the annotation and grading.

Keywords

automatic short-answer grading, large language models, automated scoring

1. Introduction

Large Language Models (LLMs) have become our assistants in many everyday activities. Over the last few years, the speed at which new models are developed has become overwhelming to daily users, researchers, politicians, and law makers struggling to keep up with all options and opportunities [1]. Yet, their application has been explored and accepted in various domains [2, 3, 4].

Automatic Short Answer Grading (ASAG) systems have emerged as an educational technology, addressing the need for efficient assessment methods in both online and traditional educational environments long before the hype of LLMs [5]. The primary objective of ASAG systems is to automatically evaluate and score students' responses to short answer questions. The difficulty of the task arises from the length of the texts - often even simply a few words - and thus the limited given context [6, 7]. One of the approaches to the task of ASAG for closed-ended questions is the comparison of the student answer to a predefined correct answer [8, 9]. The developments in ASAG have been heavily influenced by advancements in Natural Language Processing (NLP) and Machine Learning [10].

Accordingly, LLMs have found their applications in the creation of datasets and tools. While they are of great help for generic tasks such as answering questions or writing text [11], they often fall short when applied to domain specific tasks [12, 13, 14]. One primary concern is the risk for LLMs to amplify biases present in their training data [15]. Further, it is a challenge to ensuring the factual accuracy and relevance of the content generated by LLMs [16]. Previous attempts using Retrieval-Augmented Generation have been made to incorporate external sources

rositsa.ivanova@unisg.ch (R. V. Ivanova)

and enrich LLMs answers with knowledge, improving the factual grounding and thus the safety of answers [17, 18, 19]. However, such approaches rely on knowledge databases and annotated datasets to learn from, which underlines the critical importance of creating qualitative gold standard datasets [20, 21].

We explore the use of LLMs for the automated grading of short-answer texts as an example of a complex task that requires an understanding of a brief answer without receiving more than a sample solution. Our exploratory study aims to address the question of whether the LLMs have implicitly learned to perform well on specific NLP tasks (e.g. ASAG). We believe that understanding the short-comings of LLMs is one of many steps towards developing more suitable annotation approaches that could be used for the support by LLMs in process of automated grading.

2. Experiment

We compare the grading of students answers to exam questions done by two people to that of a popular, widely-used and free-of-charge LLM (i.e. ChatGPT-3.5). We acknowledge the fact that the chosen model is merely one amongst many, which all have their individual strengths and weaknesses, and that it is being continuously updated. However, due to the wide spread use of the model in various domains and the exploratory scope of this study, we build our use-case on ChatGPT-3.5, while pointing out the limitations of our choice.

Human annotation The initial dataset of this experiment was created in two steps. First, Mohler and Mihalcea [22] graded the assignments of undergraduate students in an introductory computer science (CS) course. The 630 short-answers given by 30 students were evaluated by two graduate CS students on an interval scale from 0 to 5. The second dataset extended the former by expanding the total number of short-answers to 2 273 [23]. The grading of the new texts was also done by the same two people. The grading scale ranged from 0 to 10 and in some cases the graders gave half points. The conversion of this scale to an equivalent from 1 to 5 lead to the use of rational numbers with a decimal increment of 0.25 interval for some of the grades. For the purpose of our study, we kept the answers, which received a whole-number grade, as we deemed the comparison to grades with various initial granularity (i.e. only whole numbers for first part and a mix for the second) to be introducing unnecessary bias and 89% (2 022 answers) of the answers received whole-number grades.

ChatGPT The prompt consisted of instruction incl. the *grading scale*, the initial *question*, the desired *correct answer*, and the *student answer*. To gain a better insight in the grading decisions, we requested a text comment for each grade selection.

3. Results

We compared the grading of the human annotators and ChatGPT in multiple steps and using various approaches. First, we compare the grades given to the answers by the first grader (H1) and the second grader (H2). Second, we compare them individually to the automatically

assigned score by ChatGPT. For the three pairs, we derive a simple percentage of inter-annotator agreement (IAA), evaluate the agreement beyond chance (Kappa Score), the agreement with a focus on the severity of disagreement (Weighted Kappa Score), and the linear correlation between the scoring (Pearson's Correlation Coefficient). A detailed discussion on choice of correlation metric is provided by the dataset creators [22].

Pair	Inter-ann. Score	Kappa Score	Weighted Kappa Score	Pearson's Corr. Coef.
H1 & H2	**60.88%**	**0.295**	**0.395**	0.586
H1 & ChatGPT	30.56%	0.120	0.364	**0.628**
H2 & ChatGPT	27.10%	0.050	0.189	0.519
H* & ChatGPT	33.96%	0.050	0.186	0.537

Table 1
Evaluation of inter-annotator performance. ChatGPT is the automated grading by GPT-3.5, H1 and H2 represent the human annotators, and H is the subset instances where H1 and H2 gave the same score. The highest scores for each measure are presented in bold.

Table 1 depicts the results for each pair and score. The agreement between the two human annotators (i.e. H1 & H2) served as a benchmark for expected IAA. The Inter-annotator Score was 60.88%, indicating that both human annotators agreed on grades more than half of the time. The Kappa Score (0.295) indicates an agreement below moderate (0.41-0.60) underlined by the Weighted Kappa Score at 0.395, showing a slightly better but still modest agreement. However, considering the applied grading scale, the Pearson's Correlation Coefficient (0.586) reflects a moderate positive correlation between the two sets of grades.

On the contrary, the comparison between each human annotator and ChatGPT (i.e. H1 & ChatGPT; H2 & ChatGPT) reveals a lower level of agreement. For H1 & ChatGPT, the Inter-annotator Score, the Kappa Score and the Weighted Kappa Score indicate a minimal agreement beyond what would be expected by chance. A surprisingly high value is achieved for the Pearson's Correlation Coefficient at 0.628, suggesting a stronger correlation. One explanation for this could be the different distributions of the grading of H1 and H2. The agreement between the second human annotator (H2) and ChatGPT was even lower for all of the measures, yet also here the Pearson's Correlation Coefficient remained high, indicating a moderate correlation despite the low agreement scores.

In addition to the evaluation for the three pairs, we created a subset of the initial dataset (with 1 231 answers), where H1 and H2 agreed on the grade (i.e. H*). We view these instances as examples of answers, which were graded more objectively and where the assignment of the grade may be more straight forward. We calculate the IAA measures for the subset against ChatGPT. This yielded an Inter-annotator Score of 33.96%, which is the highest of the scores achieved by pairs including ChatGPT. However, also here the Kappa and the Weighted Kappa Scores remained noticeably lower. This suggests that even when humans were in agreement, ChatGPT's grading did not significantly align with the human consensus. The Pearson's Correlation Coefficient was 0.537, indicating a moderate positive correlation but not a strong agreement.

In summary, while we observe a moderate level of agreement between human annotators, the agreement between ChatGPT and the humans is significantly lower. However, the Pearson's

Correlation Coefficients suggest there is still a moderate positive relationship in the grading patterns between humans and ChatGPT. The results indicate that while ChatGPT can follow a grading pattern similar to humans to some extent, the consistency of these grades with human annotators varies and is generally lower than the human-human agreement levels.

4. Discussion

Bias. In our reduced dataset, the grading of H1 and H2 overlapped only in 60.88% of the cases. In the remaining cases H2 has demonstrated a bias in their grading by giving a higher grade to 76.61% of the answers. While Mohler et al. [23] describe this as a "real-world [issue] associated with the task of grading", such subjectivity can also be perceived as the strength of human annotation. Plank [24] criticizes the assumption that a single gold label should be assigned to instances, as it diminishes the variety in opinions and interpretations of human language. Particularly when creating new gold standards, such richness in the annotation may be an essential step in the aim to reduce bias in models trained on them Kasneci et al. [25]. In this context, we observe that ChatGPT assigned lower grades than H1 and H2 in 79.56% and 94.03% of all cases of disagreement.

Question / Answers	**H1**	**H2**	**ChatGPT**
Q1: What is the base case for a recursive implementation of merge sort?			
Best case is one element. One element is sorted.	5	5	2
A list size of 1, where it is already sorted.	5	5	4
Q2: When does C++ create a default constructor?			
whenevery you dont specifiy your own	5	5	2
When you dont specify any constructors.	5	5	4
Q3: What is the role of a header-file?			
To allow the compiler to recognize the classes when used elsewhere.	3	4	2
Allow compiler to recognize the classes when used elsewhere	3	3	4

Table 2
Examples of similar short-answers having received a different grade by ChatGPT.
Note: Typos in the student answers are present in the original data.

Inconsistency. Next, we took a closer look at the exam tasks, which were answered by students very similarly, yet have received different grades. We manually grouped similar answers to the same questions. While we discovered some inconsistencies in the human annotation within these groups, ChatGPT provided various grades and differing justifications for the assigned grade within nearly all of the answer groups. Table 2 provides three such examples. In Q1 and Q2 both graders assigned highest mark to the pairs of similar answers consistently. In both cases ChatGPT gave different marks.

Similar observations have been made by Duong and Solomon [26] in particular when the authors asked the same questions multiple times. Filighera et al. [27] discuss potential weaknesses of LLMs that can easily be manipulated via minor changes in the syntax of an answer

(e.g. adding adjectives and adverbs). Depending on the manipulation, Filighera et al. [28] discovered that students even manage to pass a 50% threshold on an exam "without answering a single question correctly". This underlines the difficulty of automating tasks such as ASAG. Such varieties can be crucial when two answers are assessed as equivalent by a human, yet distinguished by a LLMs due to differences which a human would consider neglectable (e.g. an extra empty character or a period in the end of an answer).

The third example (Q3) depicts a case where one of the annotators also graded the answers differently, despite high similarity of the text. As mentioned by the authors of the initial dataset, one of the graders (i.e. H2) frequently assigned higher grades. In addition to this fact, H2 also tended to grade similar answers differently more frequently than H1, for whom this was a rare exception. These results indicate that may be a need for finer-grained grading (i.e. annotation) guidelines to reduce the discrepancies between graders.

The results shed light on some issues associated with human annotation. One note-worthy issue is the low inter-annotator scores achieved by human annotators. Previous work has suggested the use of finer-grained and precise annotation guidelines to achieve higher annotation accuracy [29, 30]. Additionally, human annotation can be time-consuming and costly [31], which leaves dataset creators to look for alternatives such as the use of LLMs.

Large Language Models (LLMs) like ChatGPT present their own set of challenges. One issue is that closed-source models like GPT-3.5 are fundamentally different from their successors (e.g., GPT-4), making it difficult to understand and predict their behavior. While open-source models accessible, they often become large 'black boxes' that are challenging to interpret or understand fully [32]. Providing more precise instructions to LLMs could potentially improve their performance. Yet, we need to consider the risk that they may still miss nuances, which are easily spotted by human annotators especially in complex or subtle domains. Lastly, the use of LLMs such as ChatGPT require a substantial computational infrastructure [33, 15], posing the question whether the same (if not better) performance can be achieved without their excessive use.

5. Conclusion

Large Language Models (LLMs) like ChatGPT present their own set of challenges. Closed-source models like GPT-3.5 are fundamentally different from their successors (e.g., GPT-4), making it difficult to understand and predict their behavior. While open-source models are accessible, they often become large 'black boxes' that are challenging to interpret or understand fully. Providing more precise instructions to LLMs could potentially improve their performance. Yet, we need to consider the risk that they may still miss nuances, which are easily spotted by human annotators especially in complex or subtle domains. Generalization of the results to other domains may not be trivial, however the results of this survey already hint at the need for further research in the potential use of LLMs as an aid for domain-specific tasks such as ASAG. At this stage we believe that the ability of humans to interpret and detect nuances in brief answers remains unmatched. Due to the complexity of the task, its time-intensive nature, and the costs associated with manual annotation, the use of LLMs as support in the annotation process for domain specific datasets should further be explored.

References

[1] Y. Walter, The rapid competitive economy of machine learning development: a discussion on the social risks and benefits, AI and Ethics (2023) 1–14.

[2] J.-M. Hu, F.-C. Liu, C.-M. Chu, Y.-T. Chang, Health care trainees' and professionals' perceptions of chatgpt in improving medical knowledge training: rapid survey study, Journal of Medical Internet Research 25 (2023) e49385.

[3] C. K. Lo, What is the impact of chatgpt on education? a rapid review of the literature, Education Sciences 13 (2023) 410.

[4] S. I. Ross, F. Martinez, S. Houde, M. Muller, J. D. Weisz, The programmer's assistant: Conversational interaction with a large language model for software development, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, 2023, pp. 491–514.

[5] J. Burstein, S. Wolff, C. Lu, Using lexical semantic techniques to classify free-responses, Springer, 1999.

[6] L. Galhardi, R. C. T. de Souza, J. Brancher, Automatic grading of portuguese short answers using a machine learning approach, in: Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação, SBC, 2020, pp. 109–124.

[7] N. Willms, U. Padó, A transformer for sag: What does it grade?, in: Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning, 2022, pp. 114–122.

[8] U. Hasanah, A. E. Permanasari, S. S. Kusumawardani, F. S. Pribadi, A review of an information extraction technique approach for automatic short answer grading, in: 2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), IEEE, 2016, pp. 192–196.

[9] L. Zhang, Y. Huang, X. Yang, S. Yu, F. Zhuang, An automatic short-answer grading model for semi-open-ended questions, Interactive learning environments 30 (2022) 177–190.

[10] A. Ahmed, A. Joorabchi, M. J. Hayes, On deep learning approaches to automated assessment: Strategies for short answer grading., CSEDU (2) (2022) 85–94.

[11] V. Taecharungroj, "what can chatgpt do?" analyzing early reactions to the innovative ai chatbot on twitter, Big Data and Cognitive Computing 7 (2023) 35.

[12] A. Creswell, M. Shanahan, I. Higgins, Selection-inference: Exploiting large language models for interpretable logical reasoning, in: The Eleventh International Conference on Learning Representations, 2022.

[13] D. Mekala, J. Wolfe, S. Roy, Zerotop: Zero-shot task-oriented semantic parsing using large language models, in: Conference on Empirical Methods in Natural Language Processing, 2022.

[14] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, A. Garg, Progprompt: Generating situated robot task plans using large language models, in: ICRA, 2023, pp. 11523–11530.

[15] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.

[16] B. Goodrich, V. Rao, P. J. Liu, M. Saleh, Assessing the factual accuracy of generated text, in: proceedings of the 25th ACM SIGKDD international conference on knowledge discovery

& data mining, 2019, pp. 166–175.
[17] F. Hill, R. Reichart, A. Korhonen, Simlex-999: Evaluating semantic models with (genuine) similarity estimation, Computational Linguistics 41 (2015) 665–695.
[18] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.
[19] M. Glass, G. Rossiello, M. F. M. Chowdhury, A. Naik, P. Cai, A. Gliozzo, Re2G: Retrieve, rerank, generate, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 2701–2715.
[20] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, W. Redmond, M. B. McDermott, Publicly available clinical bert embeddings, NAACL HLT 2019 (2019) 72.
[21] D. Song, S. Gao, B. He, F. Schilder, On the effectiveness of pre-trained language models for legal natural language processing: An empirical study, IEEE Access 10 (2022) 75835–75858.
[22] M. Mohler, R. Mihalcea, Text-to-text semantic similarity for automatic short answer grading, in: Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), 2009, pp. 567–575.
[23] M. Mohler, R. Bunescu, R. Mihalcea, Learning to grade short answer questions using semantic similarity measures and dependency graph alignments, in: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, 2011, pp. 752–762.
[24] B. Plank, The "problem" of human label variation: On ground truth in data, modeling and evaluation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 10671–10682.
[25] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., Chatgpt for good? on opportunities and challenges of large language models for education, Learning and individual differences 103 (2023) 102274.
[26] D. Duong, B. D. Solomon, Analysis of large-language model versus human performance for genetics questions, European Journal of Human Genetics (2023) 1–3.
[27] A. Filighera, S. Ochs, T. Steuer, T. Tregel, Cheating automatic short answer grading with the adversarial usage of adjectives and adverbs, International Journal of Artificial Intelligence in Education (2023) 1–31.
[28] A. Filighera, T. Steuer, C. Rensing, Fooling automatic short answer grading systems, in: International conference on artificial intelligence in education, Springer, 2020, pp. 177–190.
[29] A. Rigouts Terryn, V. Hoste, E. Lefever, In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora, Language Resources and Evaluation 54 (2020) 385–418.
[30] R. Ivanova, M. Van Erp, S. Kirrane, Comparing annotated datasets for named entity recognition in english literature, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 3788–3797.
[31] I. Habernal, I. Gurevych, Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse, in: Proceedings of the 2015 Conference on

Empirical Methods in Natural Language Processing, 2015, pp. 2127–2137.

[32] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al., Opt: Open pre-trained transformer language models, arXiv preprint arXiv:2205.01068 (2022).

[33] T. Schick, H. Schütze, It's not just size that matters: Small language models are also few-shot learners, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2339–2352.