Automated Assessment of Students' Code Comprehension using LLM

Dear Editor,

We are writing to express our gratitude for the thorough review of our research paper, “ Automated Assessment of Students’ Code Comprehension using LLMs” which we submitted for consideration to AI for Education Workshop 2024(AI4ED). We appreciate the time and effort invested by the reviewers in evaluating our work. We have updated the paper with necessary changes based on feedback from both the reviewers. Please find the reviews from two reviewers(R1 and R2) and updates below:

Reviews from R1

Summary: This paper presents an approach to assessing student code comprehension using Large Language Models (LLMs). It compares the effectiveness of different LLMs in evaluating student explanations of computer code by assessing their ability to compute accurate and meaningful semantic similarity scores. Key contributions include the comparison of a wide variety of decoder-only and encoder-only models. Few-shot and chain-of-thought prompting strategies are used for decoder-only LLMs. For encoder-only models some of them are fine-tuned on the CodeCorpus dataset. Results indicate that LLMs, especially GPT-4, perform comparably to fine-tuned encoder models, suggesting their potential utility in automated student assessment.

High-level evaluation:

- The problem of auto-grading self-explanations is an important topic in AI for Education research. If one has access to self-explanation expert responses to compare to, it is an interesting idea to evaluate different encoder-only models since most of these models are cheap and quick at inference time and can be stored in the memory of most consumer GPUs.
- I see issues in assuming the existence of self-explanation expert examples. The vast majority of existing coding assignments do not have such resources.

High-level technical:

- The paper overall unfortunately lacks clarity regarding the most important part of the paper: methodology. Table 2's caption says: "Pearson and Spearman correlation between student explanation and expert explanation using various classes of models." I am assuming that the results reported describe Spearman and Pearson correlations between
human</br> semantic similarity scores and the
models'</br> semantic similarity scores; it is not what the Table caption says. Since Table 2 constitutes the entire point of the paper, it is unfortunate to not know what was done.

  Thankyou for your feedback. We addressed the reviewers comment, Table 2’s caption has been updated to *“Pearson and Spearman correlations by comparing human-annotated semantic similarity scores with automated similarity scores for student and expert explanations across different model classes. dagger indicate fine tuned model*”

- It is unclear which encoder-only models are fine-tuned. The paper says: "We finetune SBERT with contrastive loss objective function for one epoch." on page 4 and also says: "We employed models fine-tuned in NLI to further fine-tune using our dataset." on page 4. Are these models further fine-tuned on their dataset? It is unclear.

  Thankyou for your suggestion. We have updated the text to "Additionally, SBERT models demonstrate improved performance in tasks related to Natural Language Inference(NLI) when fine-tuned on models previously trained with NLI data (Reimers, 2019). Hence, we fine-tuned models that were initially trained in NLI using our dataset for enhanced performance."

  "We split our data-set into 80% training data and 20`% test data and finetuned SBERT with contrastive loss objective function for one epoch in our training dataset."

.

- I don't think that decoder-only models should be used in this manner to help in evaluating students' self-explanation of code. I understand the need to evaluate all models the same way but it seems to be going against decoder-only strengths to force them to behave like encoder-only models. Instead of asking them to compute semantic similarity scores between expert and student explanations, it would be much more beneficial to prompt them to compute a score directly of how good the student's explanation seems to be compared to the expert explanation. Asking for semantic similarity is going against what they were made to do.

  Thank you for the feedback. We agree that LLMs are best designed for generating text In this work, we are systematically assessing the capabilities of LLMs beyond generating text i.e to automatically assess student's free form response. We also thank the reviewer for pointing out a direction for prompting LLMs. We chose similarity based prompting for this work as a similarity based approach has been used in previous works. Prompting directly can be a work for the future.

  Low-level technical:

- BERT, ROBERTA, and all other encoder-only models are LLMs. Comparing LLMs and encoder-only models is inaccurate wording.

  Thank you for your feedback. We have changed the wording to encoder-only models and decoder-only models.

- This sentence on page 4 should be rephrased: "if we fine-tune it in previously fine-tuned NLI data fine-tuned models, we employed models fine-tuned in NLI to further fine-tune using our dataset". It is very difficult to read.

  Thank you for your feedback. We have updated the text to " Additionally, SBERT models demonstrate improved performance in tasks related to Natural Language Inference(NLI) when fine-tuned on models previously trained with NLI data (Reimers, 2019). Hence, we fine-tuned models that were initially trained in NLI using our dataset for enhanced performance."

- Missing "the" in this sentence: "where LLM is tasked to infer from provided examples"

  Thank you for your feedback. We have updated the text to ""These include the conventional few-shot prompting, also known as in-context learning, where the LLM is tasked to infer from

the provided examples or task descriptions as well as few-shot chain-of-thought (CoT) prompting where the LLM is guided to think step by step."

- Missing "Large" here: "These were provided as examples to the Language Models (LLMs)". It is also missing in many other places. In addition, once LLMs have been introduced as meaning Large Language Models, LLMs should be used in the entire paper.

  Thank you for your feedback. We have updated the text to "These were provided as examples to the Large Language Models (LLMs), with the caveat that the examples were excluded from the dataset used for subsequent analysis."

- Typo here: "between the expert and student's code explanations were obtained" in students'

  Thank you for your feedback. We updated the text to "*In addition to expert explanations, human judgments of the semantic similarity between the expert and students' code explanations were obtained.* "

- Shouldn't this expert example: "To obtain the minutes in seconds, we divide the seconds by 60 because there are 60 seconds in a minute" say "To obtain the seconds in minutes, we divide the seconds by 60 because there are 60 seconds in a minute" instead?

  Thank you for your feedback. We have updated the text to "*To obtain the minutes in seconds, we divide the minutes by 60 because there are 60 seconds in a minute*"

- ""creates variable integer entitled 'num' with initial value 5" with a similarity of 0.8 compared to the reference "In this program, we initialize the variable num to 15." (for more detail see Appendix C.2). In this scenario, the student's response suggests a potential gap in understanding, highlighting the need for an instructional intervention." In this example, I would personally assume that the Mechanical Turker was simply in a rush and made a typo by entering 5 instead of 15. I am not sure this is the best example to show a lack of understanding of a student.

  Thank you for your feedback. We have updated the reviewer's concern and replaced the previous text with the following " In situations where there is a numerical disparity between a student's explanation and an expert's explanation, current Language Models (LLMs) do not account for this difference when automatically evaluating the similarity between the two texts."

Review summary:

- Interesting initial idea but decoder-only models should properly be used differently
- Lacks clarity in the methods section
  - *The method section has been revised and updated in the paper*
- A few typos
  - *The typos have been fixed.*

Reviews from R2

Summary of Contribution: This paper introduced a method of using large language models (LLMs) for student self-code explanation evaluation. Specifically, the authors implemented multiple neural network baselines, including BERT baseline models and several LLMs, to evaluate the models in the task. The final result shows that the GPT-4 model works best with engineered prompts with the chain of thought method in this task, indicating similar results to prior work that involves model training.

Strengths:

- The method and contribution is very clearly introduced in the paper.
- The contribution is important to the field, as an alternative method for automated assessment of student self-explanation of code.

Weaknesses:

- The paper falls short on the educational implications. The task has significant educational implications, but in this paper, the authors did not mention the educational implications or review related work implementing related systems to show actual students' learning gain. The authors did not mention how they plan to leverage this work for actual educational scenarios.

  Thank you for your feedback. We added the text "*This work is part of a larger project whose primary objective is to create an educational technology that can scaffold students' understanding of code by providing tailored feedback to students while prompting students to explain their understanding of the code line-by-line as they read it. A key component of this system is assessing students' self-explanation of lines of code which we propose to do by computing the semantic similarity between each line of code and the corresponding student explanation*." in the introduction section

- In the labeling process of the scales of similarity, it is unclear for readers what are the labeling criteria and how exactly the three students labeled the dataset. This poses some threat to validity.

  Thank you for your feedback. We added the following text to make the annotation procedure more clear and transparent "*Six graduate students in Computer Science annotated on a scale of 1-5 about 1,770 pairs of expert and student's explanations which are used in our study presented here. Before beginning the annotations, the graduate students received training on the annotation guidelines. The annotation process aimed to achieve two objectives: firstly, to distinguish between goal-oriented explanations that explains how the goal of the code is achieved and behavior-oriented explanations, and secondly, to assess the semantic similarity between crowd-sourced explanations and expert explanations. In this work we only focus on the assessment of semantic similarity between the expert and crowd-sourced explanation. During the annotation, three out of the six graduate students annotated the first half of the data, while the remaining three annotated the second half.*

  The annotation occurred in multiple stages: the first 100 data instances were used to established a consistent understanding of the annotation process. In the subsequent steps annotators involved a disagreement mitigation step aiming to minimize score differences to within 1 point among annotators and the inter-annotator agreement was computed to be 0.99 indicating high agreement among annotators".

I would like to express my sincere appreciation for the insightful comments and suggestions made by the reviewers. I am particularly grateful for their valuable insights from reviewer R1, which have enabled me to correct the low level technical details such as the typos and grammatical errors in the paper submitted. Further, the comments from R1 has helped me further communicate the methodology sections to make it more comprehensible to the readers.

I have carefully addressed each of the reviewers' comments in the revised version of the manuscript, which is attached for your consideration. I believe that these revisions have strengthened the paper and addressed the concerns raised during the review process. Specifically, I have *outlined the educational implication and the annotation procedure as per reviewer R2's feedback*., and I am confident that these adjustments have significantly improved the clarity and validity of my findings.

Thank you for your time and consideration. I look forward to the possibility of seeing my work published in PMLR proceedings for AI for Education Workshop..

Sincerely,

Priti Oli
University of Memphis
poli@memphis.edu