Author response to reviewer comments

AI-Augmented Advising: A Comparative Study of ChatGPT-4 and Advisor-based Major Recommendations

Kasra Lekan and Zachary Pardos

R1: "encoding student responses for comparison does not capture the breadth of differences". How does this affect the evaluation of RQ1?

> **Our response**: The work being referenced refers to the breadth of differences between students in Botelho et al.'s work which also used text embeddings. This is a different scenario than the one in our analysis, although both studies share the limitations of semantic similarity metrics. We acknowledge the limitations of semantic similarity in our Limitations section.

R1: An issue that sometimes appear with AI tools to suggest content is that they end up moving everyone towards a limited space of possible recommendations, hindering the overall diversity of recommendations. Did the authors analyse this, for instance looking at how many majors where recommended in the two cases (i.e. when the AI suggestions were seen before/after wring the human ones)?

> **Our response**: Our experiment was not designed to compare the same student with one advisor seeing the AI suggestions and another seeing the AI suggestions after as the reviewer suggests (although we believe this could be an interesting experiment to run). However, we did examine the diversity of responses from ChatGPT and the advisors as measured by the number of unique majors recommended: 14 for ChatGPT and 23 for the advisors. The specific composition of majors is shown in the Appendix tables.

R1: limiting the analysis to the agreement without analysing how the agreement changed is a weak point of the paper. Indeed, there are many questions open about possible biases in the model (or in the advisors) that are not addresses in the paper: for instance, did the LLM produce worse recommendations for specific ethnicity, when prompted to consider it?
R2: It is important, with this kind of application, to not only look at average metric but also worst-case metric, metrics across demographics for disavantaged populations, etc. for reasons of fairness.

Our response: We added a table to the paper detailing the demographics and recommendations of all students whose recommendations changed as a result of introducing demographics to the prompt. We do not, however, have enough samples of each ethnicity-gender combination to perform a meaningful quantitative analysis.

R1: the authors use interchangeably GPT and ChatGPT, while they are two different things: the LLM is named GPT (3.5, 4, etc.) while ChatGPT is the webplatform that provides an interface to interact with the LLM. I believe the authors should fix this in the paper.

Our response: We made changes to now consistently use “GPT-3.5 or GPT-4” as opposed to ChatGPT since we were not utilizing the web interface. Circa GPT-3.5, the GPT designator has meant “without RLHF.” We clarified this change in meaning in the paper.

R1: Section 3: "With the greater 16K token context with ChatGPT-3.5, it was prompted with major names ..." typos?

Our response: We made this sentence clearer.

R2: precise definition of agreement

Our response: We added a precise definition of agreement.