

We thank the reviewers for their constructive comments. We have addressed the points raised by the reviewers in the following ways:

- Response to: (1) how can we conclude that the performance in the post-test depends on the hint-generation technique used by the student? and (2) to include a case study to visualize a particular student, the hint-generation technique used by them and their answers to the post-test.

We added the following points to the limitations section: All the students in the course learned the same content and engaged in the same activities, except for the hint writing activity, which varied with the experimental conditions. Therefore, the differences in the observed impact on students' learning outcomes are likely to be due to the varying design of the hint writing activities. However, since our sample size was small, there is a possibility that other factors that were not within our control impacted students' learning outcomes.

Including a case study for one particular student may not be very helpful as our interventions were small and, therefore, unlikely to produce a measurable impact on students' learning outcomes. This is also apparent from the large standard deviations of students' post-test scores, with overlapping ranges of post-test scores obtained by students from the three experimental conditions. Future studies with multiple repetitions of such student-AI collaborative learning activities over a long period of time, with learning outcomes being evaluated both immediately on course completion and sometime after the course, can be helpful to measure the impact on students' learning in the long and short terms.

- Response to: Need for more details on hints generated using GPT-4
We have added a section on the prompting process for GPT-4 and also shared the full prompt text in an external link provided in the paper.