Using ChatGPT in a Mathematics Homework Exercise

Uta Priss

Informatik, Ostfalia University, Wolfenbüttel, Germany

Abstract

This paper presents a comparison between a mathematics exercise which students solved with the help of an AI chatbot to one from a previous semester without a chatbot. Ideally, chatbots should improve student performance. But in this instance the chatbot appeared to have had an equalising effect on the student performance in that it helped weaker students while at the same time hindered stronger students. Thus questions remain open with respect to how AI chatbots might be employed effectively in order to augment human abilities.

Introduction

The development of AI chatbots has been progressing rapidly over the last few years and has facilitated the emergence of new applications for learning and teaching (Heidt 2025). Ideally, employing AI tools should lead to what Engelbart (1962) calls the *augmentation of human intellect* which means enabling human achievements which would otherwise not be possible and inspiring further intellectual development. Brynjolfsson (2023), however, discusses negative consequences of AI if instead of augmentation it results in *automation* and thus a complete replacement of human activities by AI tools. Fulbright and Morrison (2024) observe that if students employ AI chatbots in exercises, augmentation is not automatically achieved because chatbots might overwhelm students with unnecessary and misleading information. Rojas (2024) reports a case that could be considered augmentation in an assignment that integrates the use of chatbots for scientific writing together with instructor and peer feedback. Park and Manley (2024) argue that chatbots can assist students in writing mathematical proofs resulting in improvements even though the final proofs may still not necessarily be complete and free of errors.

This paper presents a further example of how chatbots do not automatically improve student work. In particular, it can be observed in this example that weaker students profit from using chatbots whereas stronger students might be held back by them. Open questions are: how to employ chatbots in a manner that facilitates augmentation of human skills, whether future, more sophisticated chatbots will make augmentation more seamless or whether an effective communication between humans and chatbots presents a challenge that requires novel, yet to be determined approaches.

Description of the Exercise

A homework exercise in an introductory mathematics class for first year computing students consists of assigning a simple set or number theoretical statement to each student to mathematically prove and to implement for some examples in Python. Because logic, but not specifically proof techniques, are taught in the class, in previous semesters the students were not expected or required to solve the exercise completely by themselves but were encouraged to consult textbooks or the internet for help. In that

case most but not all students searched the web and basically nobody used textbooks presumably because using a textbook would have required more effort.

Because the mathematical precision of AI chatbots is improving continuously, in the 2024 autumn semester students were required to use ChatGPT in the version OpenAI gpt-4o that was licensed by the university at the time. The interface is called Olaf by the university and shall be referred to under that name in the remainder of this paper. The students were supposed to consult Olaf for the exercise but to amend and correct Olaf's output if it was not perfect. The students were also expected to comment on how they employed Olaf and reflect on its usefulness. Apart from a very short introduction to the topic of AI chatbots the students received no further training on how to prompt a chatbot. In the second part of the exercise, each student was asked to review the anonymous submissions of three other students and describe any errors that they detected. A hypothesis was that with the help of Olaf the quality of the homework exercise would be superior to previous semesters. In the previous semester about 80 students were in the class and about 50 in the semester when Olaf was employed.

Olaf's Results

On the surface what Olaf produced looked very good because the structure of the proofs and the code was usually perfect. But the answers from Olaf tended to be overly verbose and often contained redundant text which was somewhat more complicated than necessary. For example, Olaf appeared to not use complete induction even though that might have resulted in simpler proofs in some cases. It is unlikely that the lack of induction was a choice made by students while prompting Olaf because in previous semesters students did present solutions with induction. Quite often Olaf made at least small mistakes, such as inconsistent use of variables or proposing slightly incorrect definitions. Olaf also sometimes made basic logical mistakes such as not proving both directions of an equivalence, proving the wrong direction or returning False instead of True if the premise of an implication is false. Only in one instance Olaf was completely wrong by asserting that a true statement was false. Because the false response could not be replicated afterwards, it is not known whether Olaf really responded incorrectly or whether the student made a mistake when prompting or paraphrasing Olaf's response or maybe used a different, older version of a chatbot. None of the three students who reviewed that submission detected the mistake. In general it appeared that students overlooked more errors while reviewing each others' solutions than in previous semesters maybe because Olaf's answers always have a very professional structure and are asserted in a very confident manner.

The Python code that the students obtained from Olaf was always syntactically correct. Because for most of the mathematical statements a proof with Python code was not possible, the code was meant to indicate the plausibility of the statements with some examples. But Olaf often supplied insufficient examples, such as only a positive example and not a negative one or, again, examples for the wrong direction of an implication. In summary, Olaf's most significant errors tended to be of a logical nature. Unfortunately, a fair number of students did not detect such errors. Because such errors

correspond to common misconceptions, it would be interesting to know whether the errors were caused by the chatbot learning incorrectly or whether the errors were already contained in the training data set if that was sourced from texts of dubious quality.

A further source of problems for the students appeared to be that Olaf did not know what the students had learned in class. Thus Olaf sometimes employed definitions that differed from the ones provided in the textbook or used Python constructs that the students were unfamiliar with. The students were instructed to ask Olaf in such cases to use different constructs. Most students did prompt Olaf accordingly with respect to the Python code but not necessarily with respect to mathematical definitions. Most likely encountering differences amongst mathematical definitions was very challenging for the students.

Marking the Assignment

In introductory mathematics courses students mostly learn to replicate and apply previously taught standard methods. There is often little or no opportunity for students to experience what it is like to perform actual mathematical work, such as modelling and proving apart from repeating standard proof methods such as induction. Because of its open nature the homework exercise provides such an opportunity. All of the problems of the exercise are of a kind that can be proven directly using basic set or number theoretic arguments. Since proof skills are not taught systematically in the class, the students need to employ logical thinking and general problem solving skills. In a previous semester, about 40% of the students developed their own proofs without much help evidenced by how the proofs were written, containing small imperfections and details on how the proof was conceived which would be omitted in a textbook. Some of these proofs had what might be called a “wow factor” because it was clear that the students spent a significant amount of time on the exercise and produced individual solutions that demonstrated good and often creative mathematical problem solving but were not so perfect that they seemed copied from somewhere else. Wow factor solutions were given full points even if they contained minor mistakes which resulted in about 40% of the students achieving full points.

The marking scheme for the first part of the homework exercise is very simple and consists of up to 4 points that are given for correct aspects and subtracted for errors. Although the following analysis is based on this very simple marking scheme, it can be speculated that it highlights a general tendency of using AI chatbots. Figure 1 shows the points for a semester without Olaf as a black line with circles and with Olaf as a dashed line with diamonds. Without Olaf the appearance of solutions with a wow factor explains why the distribution is not Gaussian but has a peak at the top mark. With Olaf an expectation was that the quality of the students’ submissions would be higher than in previous semesters. Thus points were subtracted for any errors which resulted in a more Gaussian distribution as shown in the dashed line in Figure 1. With Olaf only about 20% of the solutions had absolutely no mistakes and none of the solutions had a wow factor because all proofs followed standard patterns and none gave an impression of

being particularly creative. The students had been told that they could modify and improve the output from Olaf or even replace it completely if they included a comment explaining why they changed it. But while some students did write their own Python code there was no evidence of anybody replacing one of Olaf's proofs by an independently constructed proof.

It can be suspected that all students spent less time on the exercise and worked more superficially because of Olaf's good looking answers. There were significantly more requests from the students than in previous semesters for further feedback after the exercise was finished. It appeared that some students were still convinced that Olaf's good looking answers must be correct even if the received mark indicated problems and instead of questioning Olaf's result questioned the marking.

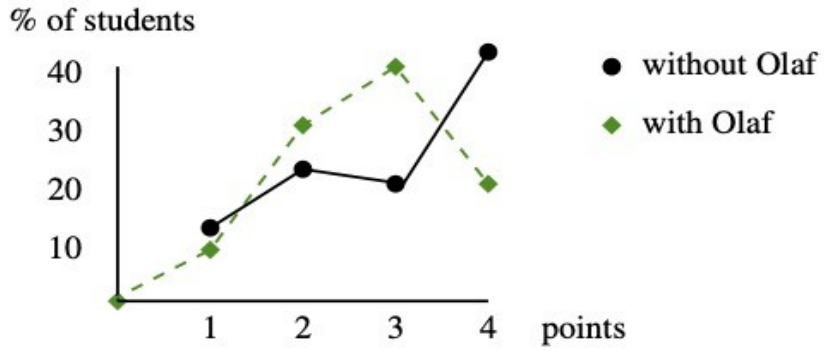

Figure 1: Points achieved with/without Olaf

Using Olaf had an equalising effect on the students' work with 70% of the students achieving a medium performance. Interestingly, there were still students who obtained 0 points because they invested so little time that they did not follow the instructions at all and some students who obtained only 1 point because of very obvious errors or omissions. Thus even if an exercise becomes presumably very easy because of a sophisticated tool such as Olaf, a small number of students still fails because of insufficient effort. Most likely students who would otherwise perform slightly below average benefitted most from using Olaf as demonstrated in Figure 1 because they were able to achieve an average mark. On the top end of the scale, it was much more difficult to obtain full points because of the general reasonably good quality of Olaf's results which led to a subtraction of 1 point for even just a single small mistake. Or in other words, if the quality of submissions is generally high for all students, it becomes more difficult to stand out. No student submitted a solution with a wow factor. It seemed as if Olaf had stopped students from thinking independently for themselves.

Otherwise marking the exercise did not feel different from previous semesters. Although Olaf's answers always followed a similar structure and Olaf mostly chose certain proof patterns (by contradiction and case distinction) the details of the answers were sufficiently different so that it did not raise any suspicions of students plagiarising each other.

Discussion and Conclusion

Obviously a single homework exercise is insufficient for drawing long reaching conclusions. But, first, at least the student numbers (50-80) are reasonably substantial and, second, Park and Manley (2024) report a similar case of employing a chatbot in an exercise of mathematical proof writing which also led to improved but sometimes incomplete proofs that contained errors. It is likely that at some point in the near future chatbots will be capable of producing flawless mathematical proofs. In some sense, however, mathematical reasoning is the simplest form of reasoning for a computer because of its formal structure. Thus not all tasks may be achievable by chatbots as easily. Furthermore it will most likely still be a concern in the future whether humans critically examine and enhance the chatbot outputs or simply accept them, representing augmentation in the first case and automation in the latter. It would be commendable if chatbots were to improve the work of weaker students by enabling them to study more effectively, but not if students only substitute their work with chatbot output.

An interesting question arising from the homework exercise is how to employ chatbots so that they do not stifle but boost the performance of stronger students. This question will be even more important if the chatbots present ever more impressive answers in the future. In a dystopian view one could speculate whether human labour will simply become redundant. In a non-dystopian view, human and AI intelligence should be augmented leading to better results than either human or AI could achieve on their own. Because of the verbose and confident but human-like manner of the chatbot answers, a research angle would be to examine whether the style of communication between users and chatbots can be modified in order to strengthen critical thinking on the human part and facilitate augmentation.

References

Brynjolfsson, E. (2023). *The Turing trap: the promise & peril of human-like artificial intelligence.* In: Augmented education in the global age. Routledge. pp. 103-116.

Engelbart, D. (1962). *A conceptual framework for the augmentation of man's intellect*. In: Howerton and Weeks (eds.) Vistas in information handling. Vol. 1. Spartan Books.

Fulbright, R. and Morrison, M. (2024) *Does using ChatGPT result in human cognitive augmentation?* Int. Conf. on HCI, LNCS 14694, Springer, pp. 133-146.

Heidt, A. (2025). *Students find new uses for chatbots*. Nature 639, p. 265.

Park, H. and Manley, E.D. (2024) *Using ChatGPT as a proof assistant in a mathematics pathways course.* The Mathematical Education, 63(2), pp.139-163.

Rojas, A.J. (2024) *An Investigation into ChatGPT's Application for a Scientific Writing Assignment.* Journal of Chemical Education, 101(5), pp.1959-1965.