Using Large Language Models to Assist Teaching Quantum Computing

Lars Krupp[a,b], Jonas Bley[b], Isacco Gobbi[c], Alexander Geng[c], Sabine Müller[c], Sungho Suh[a], Ali Moghiseh[c], Arcesio Castaneda Medina[c], Artur Widera[b], Herwig Ott[b], Valeria Bartsch[c,d], Paul Lukowicz[a,b], Jakob Karolus[a,b], and Maximilian Kiefer-Emmanouilidis[a,b]
[a] *German Research Center for Artificial Intelligence,*
[b] *RPTU Kaiserslautern-Landau,* [c] *Fraunhofer ITWM,* [d] *Fraunhofer CML*

Teaching students on an individual level quickly becomes a time-consuming task for teachers making it difficult to support their students individually [1]. This effect is even more prominent if there are only few teachers for a specialized subject, compared to rising students numbers. Quantum computing is such a domain due to its common presence in media and the expectation placed on it as a future technology. Here, Large Language Models (LLMs) offer the unique opportunity to partly take over a teacher's tasks and provide individual support for students, reducing the workload for the teacher. Consequently, we investigated the possibility of providing LLM-generated (GPT4) tips for students, to reduce the teachers' time-investment. In our study, we conducted a short survey (approx. 30 min) with participants (N=46, $\bar{x}_{age}$=28.9 y, s_{age}=7.35 y, m=35, f=8, other=3). The questionnaire was distributed during the quantum computing introductory courses of the Quantum Machine Learning School QUIKSTART2024 at RPTU in Kaiserslautern, which was supported by the Quantum-Initiative Rhineland-Palatinate (QUIP) and the Research Initiative Quantum Computing for AI (QCAI). We employed a between-subject design and asked participants four questions about quantum computing, giving them a tip for each. The tips were either generated by an LLM or created by experts. Furthermore, to evaluate any bias towards LLMs, we introduced two deception conditions where some participants got expert tips but were told that the tips came from an LLM, and vice versa. This resulted in four conditions. In two, participants were told the real origin of the tip. In the other two, the participants were told the tip originated from the opposite source. The expert tips were created by quantum physics experts who gave the quantum computing introductory courses at the QUIKSTART2024 winter school. For the LLM generated tips, we used GPT4 to generate five tips per question. Both the experts and the LLM had access to the lecture script, the questions, the available answer options and the correct choice, but not to the tips created by the other party. We investigated subjective measures of correctness, helpfulness and quality of the tip and the difficulty of the question. The final score achieved by participants served as an objective measure of their performance. We used Bayesian hypothesis testing to evaluate these measures [2, 3]. This method allows to test for the equivalence (H_1) and the difference (H_2) hypothesis of two groups. We used the Bayes Factor (BF), a measurement for the proportional difference between the probability of either hypothesis being accepted. As sufficient effect size to accept one hypothesis over the other a BF$\geq$3 has been proposed. It has been shown in literature that BF=3 equates to p<.05 in the frequentist analysis [4]. Analyzing the data, we found significant evidence to accept the H_1 hypothesis stating that the creator of the hint does not influence the student's score ($BF_{H_1:H_2}$=4.40). Furthermore, we saw significant evidence to accept the H_2 hypothesis stating that label of the tip positively influences the student's score ($BF_{H_2:H_1}$=4.16, $\bar{x}_{Expert}$=2.48, $\bar{x}_{LLM}$=3.28), when the tip is labeled as LLM-generated. Regarding the influence of the label on the subjective metrics, we accept the H_1 hypothesis for quality ($BF_{H_1:H_2}$=4.46), correctness ($BF_{H_1:H_2}$=4.03), and helpfulness ($BF_{H_1:H_2}$=3.88). The results regarding difficulty ($BF_{H_1:H_2}$=1.02) are inconclusive. For the influence of the creator on the subjective metrics, we accept the H_1 hypothesis for difficulty ($BF_{H_1:H_2}$=4.49) and helpfulness ($BF_{H_1:H_2}$=3.49). The results are inconclusive for quality ($BF_{H_1:H_2}$=2.78) and correctness ($BF_{H_1:H_2}$=2.70), with observable tendencies in favor of accepting H_1. Our results demonstrate, that under the right circumstances, it is possible to use an LLM to generate tips instead of an expert. Furthermore, we observed a significant improvement in our objective measure when tips were labeled as LLM-generated, which might indicate the existence of a placebo effect [5].

[1] K. VanLehn, The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems, Educational psychologist **46**, 197 (2011).
[2] Z. Dienes, How to use and report bayesian hypothesis tests., Psychology of Consciousness: Theory, Research, and Practice **8**, 9 (2021).
[3] J. Van Doorn, D. Van Den Bergh, U. Böhm, F. Dablander, K. Derks, T. Draws, A. Etz, N. J. Evans, Q. F. Gronau, J. M. Haaf, *et al.*, The jasp guidelines for conducting and reporting a bayesian analysis, Psychonomic Bulletin & Review **28**, 813 (2021).
[4] D. V. Lindley, A statistical paradox, Biometrika **44**, 187 (1957).
[5] T. Kosch, R. Welsch, L. Chuang, and A. Schmidt, The placebo effect of artificial intelligence in human–computer interaction, ACM Transactions on Computer-Human Interaction **29**, 1 (2023).