Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# The comparative evaluation of explanations for algorithmic programming problems provided by a Human vs. ChatGPT: an exploratory online user study

**Alfarabi Imashev**
Nazarbayev University
**Zhanna Abisheva**
L. N. Gumilyov Eurasian National University
**Inara Tusseyeva**
Nazarbayev University
**Zhanibek Kozhirbayev**
Nazarbayev University
**Tohid Alizadeh**
Nazarbayev University
**Bakhyt Matkarimov**
bmatkarimov@nu.edu.kz
Nazarbayev University

**Research Article**

**Keywords:** Human-Computer Interaction, Evaluation Studies, AI in Education, 3D avatars, Virtual Assistants

**Posted Date:** September 1st, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10677896/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# The comparative evaluation of explanations for algorithmic programming problems provided by a Human vs. ChatGPT: an exploratory online user study

Alfarabi Imashev[1*†], Zhanna Abisheva[2†], Inara Tusseyeva[4,5], Zhanibek Kozhirbayev[1,2], Tohid Alizadeh[3], Bakhyt Matkarimov[1,2*]

[1]National Laboratory Astana, Nazarbayev University, 53 Kabanbay Batyr ave, Astana, 010000, Kazakhstan.
[2]Faculty of Information Technology, L.N. Gumilyov Eurasian National University, 2 Satpayev Street, Astana, 010000, Kazakhstan.
[3]Robotics Department, School of Engineering and Digital Sciences, Nazarbayev University, 53 Kabanbay Batyr ave, Astana, 010000, Kazakhstan.
[4]Institute of Smart Systems and Artificial Intelligence, Nazarbayev University, 53 Kabanbay Batyr ave, Astana, 010000, Kazakhstan.
[5]Department of Intelligent Systems and Cybersecurity, Astana IT University, 55/11 Mangilik El ave, Astana, 010000, Kazakhstan.

*Corresponding author(s). E-mail(s): alfarabi.imashev@nu.edu.kz; bmatkarimov@nu.edu.kz;
Contributing authors: abisheva_zhm_1@enu.kz; inara.tusseyeva@nu.edu.kz; zhanibek.kozhirbayev@nu.edu.kz; tohid.alizadeh@nu.edu.kz;
†These authors contributed equally to this work.

**Abstract**

The rapid advancement of generative artificial intelligence (AI) systems like ChatGPT has the potential to revolutionize education. However, using these technologies requires an understanding of the factors that motivate students to use them. Investigations on the acceptability of ChatGPT have mainly focused on

academic environments, resulting in a lack of exploration within school contexts. This research examines the gap by surveying 37 upper secondary students.
The aim of this paper is twofold: to figure out whether students are able to distinguish codes (solutions) provided by GPT over the ones provided by humans and to evaluate how students may react to and perceive a 3D avatar as a tutor explaining how to solve non-trivial programming tasks.
A 3D avatar that "explained" the solution codes was provided to students having experience in programming: some solutions were ChatGPT-driven, and others were human-provided solutions from www.codeforces.com. Previous programming (starting age, highest olympiad participation level) and experience of previous interaction with avatars were supposed to be important for their acceptance of such technologies and will of future use, but the user study revealed that in this case these factors have no significant effect on participants' decisions.
According to feedback, participants were quite optimistic about the future of such technologies, however also emphasized the demerits of proposed avatars.
The findings may guide educators and policymakers seeking to promote the thoughtful and successful integration of generative AI in teaching.

**Keywords:** Human-Computer Interaction, Evaluation Studies, AI in Education, 3D avatars, Virtual Assistants

# 1 Introduction

Recently, on September 17, 2025, at the 49th International Collegiate Programming Contest [1] (ICPC) World Finals in Baku, Azerbaijan, hosted by ADA University, OpenAI [2] and Google DeepMind [3] teams tested their AI models to solve given 12 competitive programming problems, competing with 140 University student's teams from 103 countries. ICPC 2025 World Champions, St. Petersburg State University team solved 11 problems, two teams from University of Tokyo and Beijing Jiaotong University solved 10 problems. Google Gemini AI solved 10 problems, and OpenAI's models solved all 12 problems. This amazing experiment clearly demonstrates the power of modern AI models in solving tricky and mission-critical algorithmic problems, as well as their excellence. AI models are now widely used in software development and education.

Analysis of code skills is crucial in the professional life of future programmers, as it is a part of the development process that improves the quality of the final product. Code analysis is the process of studying the program code for errors and/or vulnerabilities, debugging, and improving its quality. There are two main types of code analysis: static and dynamic.

Static code analysis is performed early in the development process to detect errors before execution. Particularly, it involves checking the source code for syntax errors, adherence to coding standards, style consistency, and readability. Among the most effective tools for static analysis are SonarQube (Campbell & Papapetrou, 2013), which supports multiple programming languages and integrates with various version control and CI/CD pipelines; ESLint (Tómasdóttir, Aniche, & Van Deursen, 2018), which focuses on JavaScript code; and Pylint (Liawatimena et al., 2018), which evaluates Python code for compliance and potential errors.

In contrast, dynamic code analysis is carried out during program execution and can reveal issues that static analysis may miss. Its primary advantages include evaluating the interaction between components, detecting memory-related problems, and assessing runtime behavior. Key tools for dynamic analysis include Valgrind (Nethercote & Seward, 2007), which is used for C and C++ and helps identify memory leaks, invalid memory access, and other issues while providing detailed reports; JProfiler (Cho, 2006), which monitors Java application performance and behavior, offering various metrics and visualizations to support optimization; and Dynatrace (Willnecker, Brunnert, Gottesheim, & Krcmar, 2015), a comprehensive tool for performance monitoring, analysis, and reporting.

Virtual assistants such as Copilot, ChatGPT, and others already provide answers to questions and explanations comparable to encyclopedic interaction or transmission from a teacher or a professor.

Despite the fact that the introduction of ChatGPT in the educational process is a fairly new phenomenon, there is already evidence of the effects (Adeshola & Adepoju, 2024; Grassini, 2023; Lo, 2023).

[1] https://icpc.global/
[2] https://worldfinals.icpc.global/2025/openai.html
[3] https://worldfinals.icpc.global/2025/gdm.html

The objective of the present user study is to ascertain the discernibility of the difference and the frequency with which participants could comprehend the solutions proposed by ChatGPT; or conversely, it may be considered that ChatGPT is already able to function at the teacher level.

As an experiment, participants will be asked to solve several proposed problems from www.codeforces.com and then examine an explanation of solutions provided by a human instructor (also from www.codeforces.com) and ChatGPT.

It is very important to compare how high school students with recent experience in algorithmic programming competitions perceive the different types of solutions, as this might reveal a statistically significant difference between human solutions and ChatGPT in their responses.

In order to avoid bias on the part of the participants, 3D characters with quite monotonous voice to mitigate emotion expressions, which might be expressed by humans, will be used to anonymize and depersonalize the solutions by voicing the solutions with the in-built text-to-speech available in iClone. The importance of anonymization in research areas such as human-computer interaction and human-robot interaction has been proven in practice (Kuang, Yang, Shen, Hu, & Yu, 2024).

In our case, the research team needed informants to first perceive the provided solutions as equivalent and then, based on their observations, to ask themselves what the study is about. Will they distinguish the difference in explanations of solution codes? How high would they rate ChatGPT-3.5 and human-provided solutions?

## 2 Related Work - Literature Review

The development of the use of AI in education has increased rapidly during the last decade (Adeshola & Adepoju, 2024; Ansari, Ahmad, & Bhutta, 2024; Baidoo-Anu & Ansah, 2023; Kasneci et al., 2023; Lo, 2023; Vieriu & Petrea, 2025; Zeng & Yang, 2026), especially in coding (Cambaz & Zhang, 2024; Fang et al., 2024; Kazemitabaar et al., 2023; Tu, Guo, & Liu, 2026; Vaithilingam, Zhang, & Glassman, 2022). The results of the conducted online survey of the research paper by Conrad and Nuebel (2025) showed the positive attitude of the schoolchildren from Germany towards using ChatGPT in their study. The scores for the frequency of usage items revealed the "occasional to moderate" application in the school context. The results of the usage purpose question showed that most schoolchildren used ChatGPT for research, text writing, and homework preparation. The lowest score was given for the purpose related to the process of studying for exams and preparing presentations. Interestingly, among the subjects where the kids used ChatGPT, the Computer Science course was in the last place, along with Mathematics. All of these outcomes might indicate the fact that ChatGPT was more useful in text-based learning or open-ended assignments, while it was less utilized in controlled or formal tasks, such as exams.

Below is the literature review with the concentration on the articles where AI was compared with the humans in programming code explanation tasks, as it is closely related to the idea of the conducted experiments of this article. The experiments in (Yeadon, Peach, & Testrow, 2024) were conducted during a 10-week coding course

with the students of physics degree at Durham University. The students from physics major usually use coding when they are running simulations or when they need to analyze a big amount of data, draw resulting plots, etc. Based on these types of work the researchers conducted the experiments, where a 16-page report with 14 plots was evaluated by three markers among 50 students and AI using a blind estimation approach. The probable authorship was from "Definitely AI" to "Definitely Human" with the use of 4-point Likert scale. As there were some preprocessing actions in order to edit the code into the LLM models, we cannot say that the code was fully AI, instead it was better to consider it as 'AI with human assistance'. The results showed that GPT-4 being a valuable tool cannot perform physics coding tasks better than students based on the scores of the independent markers, but still GPT with Prompt Engineering has the highest scores among all GPT versions tested in experiments. That is not related to comparison of Human-vs-AI results, but worth noticing that the prompt versions of GPT3.5 and GPT4.0 improved the scores from the independent graders. Overall, the submitted assignments generated by AI were distinguished by the markers as Definitely or Probably AI more often. When giving the evaluators two options AI or Human, they chose the right authorship with the high accuracy. As the evaluators had a deep knowledge in physics and coding, they were able to more accurately distinguish the remarks or incorrect code implementation done by AI, making the results more understandable. Additionally the authors found out that the colors that were represented on plots affected the recognition of students' works from AI-generated. The colors of the student-authored plots probably were not traditional making them more unique, whereas the plots generated by AI had "misaligned or skewed layouts" (Yeadon et al., 2024). The larger sample size in the future research would show exactly the reasoning for this color-choosing aspect. The main idea of the work by Leinonen et al. (2023) was to compare the code explanations generated by the LLM (GPT3) and by the students. The quality of the generated learner sourced code explanations were judged by the students based on accuracy, length and understandability. Additionally the most useful characteristics of a code explanation from the evaluators' point of view were highlighted. The participants were one thousand first year students from The University of Auckland who took the course of programming. The experiments were divided by two separate lab sessions, for one week duration each. During the first session, the topics of the course were the following: arithmetic, types, functions, loops and arrays in the C programming language. The first task was to create the explanation and formulate the purpose based on the definition of three functions. After two weeks the second session was conducted where the students gave the rate for accuracy, understandability and length of a random sample of four function code explanations that was either generated by students during session 1, or generated by GPT-3. The results showed that the LLM generated explanations of the code were evaluated higher in terms of perceived accuracy and understandability than the student's ones, whereas the ideal length was the same for both originalities. Most of the code explanations generated by LLM started with the phrase: "This code snippet" or "The purpose of this code snippet", while the text written by the students varied in its initial part. Overall both types of explanations had almost the same structure, including the goal of the function, the code analysis, and the return outcome. The authors

highlighted several limitations and future work points. For instance, as the students who participated in the experiments were mostly the beginners, the results might be affected by the presence or absence of the previous experience of working with chatbots. A suggestion was provided by the researchers to study the perception of code correctness, and to include more composite functions. And finally, they gave the cue for future research to analyse the effect on the students' studying process and results in two conditions: when they used code explanations generated by the other students or by LLM. The authors in (Finnie-Ansley, Denny, Becker, Luxton-Reilly, & Prather, 2022) examined the Codex: how well it would solve the questions from introductory programming exams in New Zealand. The researchers divided the experiments into two parts. During the first evaluation they compared the results of CS1-type coding problem solving results generated by AI Codex and by students. During the second evaluation the comparison was based on the solution of the Classic Rainfall Problem. The evaluation has been done by the lab-based examination software that graded both AI's and students' answers. All tasks were equally presented to both AI and students. Each of them (AI and students) had a maximum 10 attempts to finalize the coding answers to the accepted form, and the small typos were not considered. These typos were highlighted by the grading software, so the students had a chance to make corrections before submission. That's why the typos in the code from AI were corrected by the human experimenter, in order to make the conditions equal for both AI and students. The results of the grading of 71 students' solutions were compared with the code generated by Codex for first evaluation consisted of Test 1 and Test 2, and second evaluation consisted of the Classic Rainfall Problem. The results showed that Codex had overall 78.5% of correct points for Test 1 and 78.0% for Test 2 of the first evaluation. When inserting the AI result to the list of all students' scores during the first part of the experiments, AI would be placed in 17th place (out of 72), which is the top quartile of class grading results. Finally, the authors mentioned the importance of the Academic Integrity and plagiarism check in the future of programming education, making it necessary to create "more unique and complicated assessments" (Finnie-Ansley et al., 2022).

The same researchers in their next work (Finnie-Ansley et al., 2023) used 26 programming questions of the Python programming language that were given to the students as the final assessments in two proctored lab tests, two hours each. These tests were conducted in the CS2 programming course at the University of Auckland. The lab test questions consisted of the problem statement and one or more examples of the input and resulting values. Then the students should write Python code based on the task. After submission, the code was evaluated by an automated testing system named CodeRunner, which showed the students the immediate feedback about their code. Compared to their previous work (Finnie-Ansley et al., 2022), the researchers did not correct the input prompt for Codex but instead gave exactly the same problem statements in the format and content that they were shown to the students during the exam. Thereby no additional prompt engineering was applied. The researchers generated 50 code responses from Copilot for each task in order to provide the results for comparison that would be replicable. These code responses were representative samples for evaluating the effectiveness of the output given by Codex. There

were three metrics of Codex evaluation: the chance-1,3,5 with the limited number of attempts to obtain the most appropriate and correct output from AI (based on the number near the word "chance"), and the simulated, where the maximum number of attempts was 5 and the penalties for each incorrect solution were the same as for the students in the real test. The results showed that the Codex code had better performance than the students according to the overall mean values. When comparing the individual questions' code, Codex resulted with the same or higher scores compared to the mean student score for 77% of questions (20 out of 26). The authors observed an interesting point about the length of the question prompt for Codex: the best solution was obtained from the prompt that was almost half shorter in terms of the number of characters compared to the worst solution. So, the results showed a moderate negative correlation with prompt length for Codex-generated code.

The authors in (Hassany, Brusilovsky, Ke, Akhuseyinoglu, & Narayanan, 2023) studied how the undergraduate students and the teacher assistants would rate the "worked examples" created by the experts (instructor-authored) and ChatGPT through the blind grading. The PCEX example authoring tool was used for this purpose. The experts edited the AI-generated explanation to make it more understandable and full using PCEX. After editing, the experts submitted it to the system, and that explanations were referred to as "experts", but actually were generated by Human-AI collaboration. Afterwards, the same PCEX system was used by the evaluators to observe the problem statement, the code, and line-by-line explanation of the code. The results showed that the evaluators rated the completeness of ChatGPT mostly as Very Complete whereas the same criteria for the experts explanations were mostly as just Complete. Approximately half of the students and TAs found the explanations of ChatGPT as a better solution than the instructor-authored. Finally, the results showed that the students chose the ChatGPT explanations more often than the authors.

The authors in (Denny, Khosravi, Hellas, Leinonen, & Sarsa, 2023) provided to the student participants six examples of the coding task from the instructors as a starting point for a learner sourcing activity during the Introductory programming course. The task for the students was to generate their own programming resources for the same course. These resources consisted of the topic, difficulty level, resource title, code and explanation parts. The students were free to write any code there, wherever it was a piece of function, or the whole program. The explanations might be short or detailed by the choice of the students. The same task was given to LLM. Afterwards the other student-participants rated the six example codes from different originality sources (instructors, students AI) by their quality, including correctness and helpfulness (to be concrete: "code correctness, explanation correctness, and example helpfulness" (Denny et al., 2023)). Additionally to these ratings, the researchers also concentrated on the differences between the sources. So during the feedback part of the experiments, the participants answered the open type questions about their opinion of what was good about the explanation and how the explanation could be improved. Additionally, the quality of the code was rated from poor to outstanding, the confidence of participants was rated from very low to very high. The students who reviewed the instructor/student/AI-generated resources had no idea about the originality of the sources, so the evaluation was blind. Based on the obtained results about the average

occurrence of some specific keyword in 1000 characters long code, the authors observed that the AI models may use programming constructs like 'while', 'if-else', or various data types differently from humans. It might both positively or negatively affect the studying process. There was no statistically significant difference in correctness, helpfulness and student perceptions on the overall quality of the resources between the AI and students-generated code and explanations. However, the researchers found out that the average code length and explanation length were higher in student-generated code than in AI-generated. Consequently, the students showed more independence in the coding style, whereas the AI was observed to follow the given instructions and examples more literally. From the authors' point of view, more research needed to be done on the concrete situations when one resource would be preferred by the participants over the other resource. Future research is also needed to assess how AI-generated content impacts learning over the long term and whether students retain initial positive impressions of such materials. And another useful avenue of research might be to examine whether law teachers can adjust their explanations based on students' prior knowledge, offering more detailed instructions for experienced students and simpler versions for beginners. To summarize, the described articles came to the conclusion that AI might perform the task of explaining the code better than humans did in terms of perceived accuracy and understandability (Leinonen et al., 2023), correctness (Finnie-Ansley et al., 2022), and completeness (Hassany et al., 2023). The only paper with no significant results in helpfulness, correctness, and student perceptions on the overall resources quality between AI and students is published by Denny et al. (2023).

The reviewed article that studied the ability to distinguish between AI and human explanations was Yeadon et al. (2024) and resulted in high accuracy in authorship recognition. The aspects of the difference between AI and humans might be the color chosen for the plots (Yeadon et al., 2024), the beginning words of the explanations (Leinonen et al., 2023) and the use of a variety of data types and specific constructions (Denny et al., 2023). The length of the generated code and the explanations was different among humans and AI with the shorter length produced by AI (Denny et al., 2023). And finally, the best solution from the AI explanation was obtained from a shorter prompt with a smaller number of characters (Finnie-Ansley et al., 2023).

# 3 User Study

## 3.1 Demographics

For this user study, we could involve 37 (18 female and 19 male participants) students aged 15-17 years, e.g., currently studied in ninth to eleventh grades (mean age = 16.1) from a STEM-specialized school that prepares students for school olympiads of different levels. The number of persons who graduated from this school achieved IMO, IPhO, IChO, and IOI prize-winning places.

According to subjects' replies, 11 of them started study programming and informatics in primary school, 24 of them started studying programming and informatics since the fifth-sixth grade; one stared since the seventh grade, and one since the ninth year of school. 24 of them already had experience of interaction with humanoid avatar characters, and 13 had not experienced such interaction before the user study. 8 students are from groups with intensive informatics and programming learning, and the rest are from groups with advanced studies in mathematics and mixed mathematics and informatics/programming.

We also asked them at which maximal informatics or programming olympiads level you have already participated: 17 replied "school", 5 indicated City/District level, 2 - region(oblast), and two of them have taken part at international level (see Table 1).

**Table 1** Demographics of the participants

| Participant | Age | Started since | Max competition level | Interaction with avatars earlier |
|---|---|---|---|---|
| F | 16 | 5-6 grade | NO | Yes |
| F | 16 | 5-6 grade | NO | NO |
| F | 16 | 5-6 grade | school | Yes |
| F | 16 | 5-6 grade | school | Yes |
| F | 16 | 5-6 grade | school | NO |
| F | 15 | primary KS2 | school | Yes |
| M | 15 | 7 grade | city | NO |
| F | 15 | primary KS2 | region | NO |
| M | 18 | 5-6 grade | NO | Yes |
| F | 17 | 5-6 grade | school | Yes |
| F | 16 | 5-6 grade | school | Yes |
| M | 16 | 5-6 grade | school | Yes |
| M | 17 | 5-6 grade | school | Yes |
| M | 17 | 5-6 grade | NO | Yes |
| M | 15 | 5-6 grade | city | Yes |
| F | 15 | primary KS2 | district | Yes |
| F | 15 | primary KS2 | NO | Yes |
| M | 17 | 5-6 grade | school | Yes |
| M | 17 | 5-6 grade | NO | Yes |

| Participant | Age | Started since | Max competition level | Interaction with avatars earlier |
|---|---|---|---|---|
| M | 16 | 5-6 grade | NO | Yes |
| M | 16 | 5-6 grade | school | Yes |
| M | 16 | 5-6 grade | school | NO |
| M | 15 | primary KS2 | region | NO |
| F | 15 | primary KS2 | school | Yes |
| M | 16 | primary KS2 | school | NO |
| M | 17 | 9 grade | NO | Yes |
| F | 17 | 5-6 grade | NO | NO |
| F | 17 | 5-6 grade | school | NO |
| M | 16 | 5-6 grade | NO | Yes |
| M | 17 | 5-6 grade | school | Yes |
| M | 16 | 5-6 grade | school | NO |
| F | 15 | primary KS2 | international | Yes |
| M | 15 | primary KS2 | city | Yes |
| F | 16 | primary KS2 | city | NO |
| F | 17 | 5-6 grade | NO | NO |
| M | 17 | 5-6 grade | international | Yes |
| F | 17 | primary KS2 | school | NO |

## 3.2 Methodology

For the current user study, we used problem and solution codes from **www.codeforces.com**. There are multiple solutions provided on the website for each problem. We analyze the complexity of a wide list of problems and solutions.

Then, an educator who trains students for competitive programming competitions and a student who has experience in such competitions selected five problems of complexity level that will be suitable for K-9 to K-11 students-participants.

After analyzing the solutions provided on the website by humans, we requested ChatGPT to generate solution codes for the same selected problems. For the solution codes provided (generated) by ChatGPT we also animated the avatar to explain these codes for the sake of anonymization.

At the beginning of the Google Form, students were asked to estimate their current mood and programming skills. After that they have been provided with 5 programming problems, solution code with explanation performed simultaneously with the avatar. There were 4 groups of tasks: the same 5 tasks with solutions provided by humans (from Codeforces.com) and ChatGPT v3.5.

Groups 1 and 3 had codes and explanations for three tasks solved by humans, and two by ChatGPT, with two different orders. Order 1: Task 1 (from human), Task 2 (from GPT), Task 3 (from human), Task 4 (from GPT), and Task 5 (from human). Order 2 was a shuffle of the previous (5,2,3,1,4).

Whereas groups 2 and 4 were provided by the same five problems with solution codes and explanations for three tasks solved by ChatGPT, and two by human, with two different orders. Order 1: Task 1 (from GPT), Task 2 (from human), Task 3 (from GPT), Task 4 (from human), and Task 5 (from GPT). Order 2 was a shuffle of the previous one.

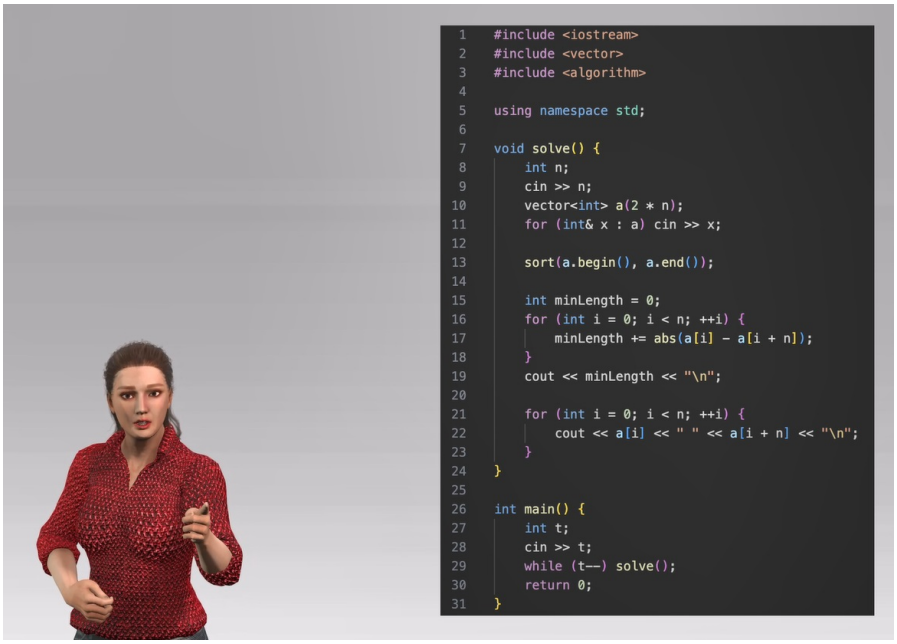



**Fig. 1** The Avatar explaining a solution code

For each task, 3 sections consisting of six questions were asked to the participants. For Section 1 (see Table 2), two Yes/No questions Q-1.1. Was the meaning of the solutions immediately clear to you? Q-1.2. Have you had to guess the meaning of a sentence based on the way the character moves? Two 1-10 Likert scale questions: Q-1.3. Do you think you understood the character's explanation correctly? Q-1.5. Do you think the solution code was provided by: 1-ChatGPT, 10-Human. One multiple choice: Q-1.4. Do you think the code was provided by: "Human", "ChatGPT", and "I find it difficult to answer". And an open question: Q-1.6. How many times have you watched the video?

**Table 2** Questions of Section 1

| # | Question | Question type |
|---|---|---|
| Q-1.1 | Was the meaning of the solutions immediately clear to you? | Yes/No |
| Q-1.2 | Have you had to guess the meaning of a sentence based on the way the character moves? | Yes/No |
| Q-1.3 | Do you think you understood the character's explanation correctly?[1] | Likert scale (1-10) |
| Q-1.4 | Do you think the code was provided by:[2] | Multiple choice |
| Q-1.5 | Do you think the solution code was provided by: [3] | Likert scale (1-10) |
| Q-1.6 | How many times have you watched the video? | open |

[1]1-Not at all - 10 Yes, it is clear

[2]Three options: ChatGPT, Not sure, Human.

[3]1-ChatGPT, 10-Human.

After interaction with avatar explanations of codes, the following questions were asked as a Godspeed questionnaire inspired section (see Table 3):

**Table 3** Questions of Section 2

| # | Question |
|---|---|
| Q-2.1 | How do you perceive the character (1 - very much natural - 10 - very much unnatural)? |
| Q-2.2 | In your opinion, this character is 1-fake - 10-real: |
| Q-2.3 | Who does this character look like more (1 - like a robot - 10 - like a person)? |
| Q-2.4 | In your opinion, the character looks and moves (1-realistic - 10-artificial): |
| Q-2.5 | In your opinion, the character's movements (1-mechanical - 10-organic): |
| Q-2.6 | This character is for you (1 - really like - 10 really dislike): |
| Q-2.7 | Is this character for you (1 - very pleasant - 10 - very unpleasant)? |
| Q-2.8 | In your opinion, does this character understand the tasks he explains (1 knows programming very well - 10 does not know it at all)? |
| Q-2.9 | What do you think of this character... (1 - not at all intelligent, 10 - yes, intelligent) |

The last section of questions contained six general questions (see Table 4):

In the end, we requested students to evaluate their mood and how they assess their coding skills after interactions with avatar as Post-test self-estimation.

### 3.3 Ethical Approval

This research was approved by L.N. Gumilyov Eurasian National University Research Resource Center for Ethics.

Consent forms and instructions were provided to students and parents / guardians. The programming teacher was available throughout the online user study. Participation was voluntary.

## 4 Results

### 4.1 Mood and programming skills Pre- and Post-tests

Since the Shapiro-Wilk test did not show a significant departure from normality for both the Mood Pre-test: $W(37) = .956, p = .148$ and the Mood Post-test $W(37) = .961, p = .213$.

We utilized the Dependent T-test to check if there is any statistical difference: $t(36) = -0.349, p = 0.729$. It revealed that there was no statistical improvement in the mood of the participants caused by the interaction with the avatar (slight shift between means from 6.35 to 6.4).

The Shapiro-Wilk test showed a significant departure from normality for both the Programming Skills Pre-test: $W(37) = .939, p = .041$ and the Programming Skills Post-test: $W(37) = .925, p = .016$. For this case, the results of the Wilcoxon Signed-Rank test ($Z = -1.195, p = 0.232$) indicated that there is a non-significant small difference between them (slight shift between means from 6.24 to 6.4).

**Table 4** Questions of Section 3

| # | Question | Question type |
|---|---|---|
| Q-3.1 | Do you think such a system is feasible? Your answer: (1 - not at all - 10 - definitely) | Likert |
| Q-3.2 | Do you find this type of interaction engaging (would you like to use it often in the future)? Your answer: (1 - not at all - 10 - definitely) | Likert |
| Q-3.3 | Would you trust such a character to explain complex problems? Your answer: (1 - not at all - 10 - definitely) | Likert |
| Q-3.4 | Would such a character be useful for explanations when the teacher is not present? Your answer: (1 - not at all - 10 - of course) | Likert |
| Q-3.5 | Was this presentation of information effective? Your answer: (1 - not at all - 10 - definitely) | Likert |
| Q-3.6 | Would you like to interact with this character in the future? Your answer: (1 - not at all - 10 - definitely) | Likert |

**Table 5** Difference between pre- and post- tests for mood and skills

| pre-test (mood) | post-test(mood) | dif. | pre-test(skills) | post-test(skills) | dif. |
|---|---|---|---|---|---|
| 9 | 9 | | 7 | 7 | |
| 10 | 10 | | 7 | 7 | |
| 8 | 6 | −2 | 5 | 7 | 2 |
| 7 | 6 | −1 | 6 | 6 | |
| 7 | 8 | +1 | 4 | 5 | 1 |
| 4 | 4 | | 6 | 6 | |
| 8 | 7 | −1 | 5 | 6 | 1 |
| 7 | 7 | | 6 | 5 | −1 |
| 4 | 4 | | 6 | 6 | |
| 6 | 6 | | 8 | 8 | |
| 6 | 6 | | 7 | 7 | |
| 5 | 5 | | 5 | 6 | 1 |
| 8 | 7 | −1 | 7 | 6 | −1 |
| 8 | 9 | −1 | 7 | 7 | |
| 7 | 8 | 1 | 6 | 6 | |
| 5 | 5 | | 2 | 3 | 1 |
| 7 | 10 | 3 | 6 | 6 | |
| 5 | 5 | | 5 | 6 | 1 |
| 6 | 7 | −1 | 6 | 5 | −1 |
| 7 | 9 | 2 | 2 | 4 | 2 |
| 7 | 7 | | 9 | 8 | −1 |
| 3 | 4 | 1 | 4 | 6 | 2 |
| 9 | 8 | −1 | 7 | 8 | 1 |
| 5 | 4 | | 10 | 6 | −4 |
| 8 | 7 | −1 | 1 | 1 | |
| 6 | 6 | | 10 | 9 | 1 |
| 6 | 6 | | 7 | 6 | −1 |
| 3 | 4 | 1 | 7 | 4 | −3 |
| 7 | 7 | | 6 | 8 | 2 |
| 6 | 5 | 1 | 5 | 4 | −1 |
| 9 | 8 | −1 | 8 | 8 | |
| 8 | 8 | | 9 | 9 | |
| 7 | 7 | | 8 | 9 | 1 |
| 7 | 7 | | 6 | 8 | 2 |
| 2 | 3 | 1 | 7 | 8 | 1 |
| 5 | 5 | | 6 | 7 | 1 |
| 3 | 3 | | 8 | 9 | 1 |

## 4.2 GPT-Human - Section 1

Regarding Question Q-1.3. We requested them to self-estimate on how correctly they think they understood the provided solution: with the following question: "Do you think you understood the character's explanation correctly?" using the 1 to 10 Likert-scale, where 1 is *very bad* - 10 is *very good* (see Table 6).

Regarding Question Q-1.4, we requested the participants to take a guess on who provided this exact solution and provided only three options to select: ChatGPT, not sure, and Human (see Table 7).

Regarding Question Q-1.6, we asked the participants how many times they had to watch the video with the code explanation provided by the avatar to comprehend the idea of a solution (see Table 8).

**Table 6** Replies for Q-1.3 provided by Groups 1,3 and Groups 2,4.

| | **Task1(H)** | **Task2(GPT)** | **Task3(H)** | **Task4(GPT)** | **Task5(H)** | **avg. H** | **avg. GPT** |
|---|---|---|---|---|---|---|---|
| gr1 | 5,3 | 6,4 | 4,9 | 7,2 | 4,9 | 5,03 | 6,8 |
| gr3 | 5,8 | 6,8 | 4,6 | 6,1 | 4,7 | 5,03 | 6,45 |
| | **Task1(GPT)** | **Task2(H)** | **Task3(GPT)** | **Task4(H)** | **Task5(GPT)** | **avg. H** | **avg. GPT** |
| gr2 | 4,75 | 5,5 | 6,25 | 5,25 | 5,625 | 5,375 | 5,54 |
| gr4 | 6,56 | 6,78 | 6,78 | 7,67 | 6,33 | 7,22 | 6,56 |

**Table 7** Replies for Q-1.4 provided by Groups 1,3 and Groups 2,4.

| | **Task1(H)** | **Task2(GPT)** | **Task3(H)** | **Task4(GPT)** | **Task5(H)** |
|---|---|---|---|---|---|
| gr1 | chgpt 9<br>notsure 1<br>human 0 | chgpt 8<br>notsure 1<br>human 1 | chgpt 5<br>notsure 1<br>human 4 | chgpt 6<br>notsure 0<br>human 4 | chgpt 3<br>notsure 2<br>human 5 |
| gr3 | chgpt 4<br>notsure 2<br>human 4 | chgpt 5<br>notsure 1<br>human 4 | chgpt 5<br>notsure 1<br>human 4 | chgpt 6<br>notsure 1<br>human 3 | chgpt 8<br>notsure 0<br>human 2 |
| | **Task1(GPT)** | **Task2(H)** | **Task3(GPT)** | **Task4(H)** | **Task5(GPT)** |
| gr2 | chgpt 6<br>notsure 2<br>human 0 | chgpt 3<br>notsure1<br>human 4 | chgpt 4<br>notsure 2<br>human 2 | chgpt 4<br>notsure 1<br>human 3 | chgpt 5<br>notsure 1<br>human 2 |
| gr4 | chgpt 4<br>notsure 2<br>human 3 | chgpt 4<br>notsure 3<br>human 4 | chgpt 4<br>notsure 3<br>human 4 | chgpt 4<br>notsure 3<br>human 2 | chgpt 5<br>notsure 2<br>human 2 |

**Table 8** Replies for Q-1.6 provided by Groups 1,3 and Groups 2,4.

| | **Task1(H)** | **Task2(GPT)** | **Task3(H)** | **Task4(GPT)** | **Task5(H)** | **avg. H** | **avg. GPT** |
|---|---|---|---|---|---|---|---|
| gr1 | 2,7 | 2 | 1,7 | 1,9 | 1,9 | 2.1 | 1.95 |
| gr3 | 1,89 | 1,1 | 1,6 | 1,3 | 2 | 1.83 | 1.2 |
| | **Task1(GPT)** | **Task2(H)** | **Task3(GPT)** | **Task4(H)** | **Task5(GPT)** | | |
| gr2 | 1,89 | 1,41 | 1,65 | 1,47 | 2,12 | 1.44 | 1.88 |
| gr4 | 2 | 1,67 | 1,67 | 1,44 | 2,67 | 1.56 | 2.11 |

## 4.3 Godspeed-like questions - Section 2

For this section of questions to evaluate the avatar, we were inspired by the Godspeed questionnaire (Bartneck, 2023), which is widely used for HRI (Cucciniello, Sangiovanni, Maggi, & Rossi, 2021; Musicco & Di Leo, 2025; Tobis, Piasek-Skupna, & Suwalska, 2023; Weiss & Bartneck, 2015) and HAI research (Alimardani, de Roode, Vaitonyte, & Louwerse, 2024; Imashev, Oralbayeva, Baizhanova, & Sandygulova, 2025; Imashev, Oralbayeva, Kimmelman, & Sandygulova, 2022; Rhee, 2020; Tarlan & Erdal, 2024). In order not to make participants bored, we decided to use this questionnaire

partially and have formulated the questions in a way that maintains the purpose and meaning of the original Godspeed items (see Table 9).

**Table 9** Means and standard deviations of replies provided for Section 2 questions (GodSpeed-like)

| Measure | Q-2.1 | Q-2.2 | Q-2.3 | Q-2.4 | Q-2.5 | Q-2.6 | Q-2.7 | Q-2.8 | Q-2.9 |
|---|---|---|---|---|---|---|---|---|---|
| Mean | 8 | 2,51 | 3,08 | 8,216 | 3,29 | 6,68 | 6,32 | 4,94 | 4,59 |
| Std. deviation | 2,13 | 1,57 | 1,92 | 1,90 | 2,17 | 2,09 | 2,21 | 2,07 | 2,19 |

We also grouped Q-2.1, Q-2.2, and Q-2.3 as categories of Anthropomorphism; Q-2.4 and Q-2.5 as Animacy; Q-2.6 and Q-2.7 as Likeability; and Q-2.8 and Q-2.9 as Intelligence; and utilized Cronbach's Alpha test to test the consistency of replies about these categories (see Table 10).

**Table 10** Cronbach's alphas for Godspeed-like categories

| Measure | Anthropomorphism | Animacy | Likeability | Intelligence |
|---|---|---|---|---|
| Cr. $\alpha$ | **0.770** | .607 | **0.844** | **0.701** |

## 4.4 General questions - Section 3

**Table 11** Means and standard deviations of 3rd section's questions

| Measure | Q-3.1. | Q-3.2 | Q-3.3 | Q-3.4 | Q-3.5 | Q-3.6 |
|---|---|---|---|---|---|---|
| Mean | 4,757 | 5,19 | 3,73 | 6,16 | 5,14 | 4,76 |
| Std. deviation | 2,23 | 2,69 | 2,24 | 2,37 | 2,07 | 2,68 |

The Mann-Whitney U test was applied to determine if there were differences in engagement scores between students who had prior experience of interaction with some agents and those who did not.

Since 13 students reported that they have not interacted with any 3D Virtual assistants before and 24 reported that they had, and taking into account the central limit theorem, we assume that answers for 6 questions of Section 3 of both sub-groups ($< 30$) are potentially not normally distributed and refer to the Mann-Whitney U test to test if their answers have significant difference or not.

The Mann-Whitney U test (see Table 12) revealed that there was no statistically significant difference for all 6 questions.

To check the potential influence of both the "age when participants started to learn programming and informatics" and the "previous interaction with agents - not naive user" factors, we utilized the two-factor ANOVA test (see Table 13).

**Table 12** Mann-Whitney U test (experience of interaction)

| Stat. test | Mann-Whitney U test |
|---|---|
| Q-3.1. | $U = 145, Z = -0.354, p = 0.742$ |
| Q-3.2 | $U = 132, Z = -0.405, p = 0.704$ |
| Q-3.3 | $U = 132, Z = -0.587, p = 0.580$ |
| Q-3.4 | $U = 132.5, Z = -0.390, p = 0.704$ |
| Q-3.5 | $U = 155.5, Z = -0.016, p = 0.987$ |
| Q-3.6 | $U = 144, Z = -0.385, p = 0.718$ |

**Table 13** Two-way ANOVA for Interactions with Avatars Experience and Programming Starting Age

| Stat. test | Programming starting age | Prior avatar interaction | Both factors |
|---|---|---|---|
| Q-3.1 | $F(3, 31) = 0.421, p = .739$ | $F(1, 31) = 0.001, p = 0.976$ | $F(1, 31) = 0.312, p = 0.581$ |
| Q-3.2 | $F(3, 30) = 2.009, p = 0.134$ | $F(1, 30) = 0.601, p = 0.444$ | $F(1, 30) = 1, 292, p = 0.265$ |
| Q-3.3 | $F(3, 30) = 0, 493, p = 0.690$ | $F(1, 30) = 0.498, p = 0.486$ | $F(1, 30) = 0.405, p = 0.530$ |
| Q-3.4 | $F(3, 30) = 0.428, p = 0.734$ | $F(1, 30) = 0.039, p = 0.846$ | $F(1, 30) = 2.314, p = 0.139$ |
| Q-3.5 | $F(3, 31) = 1.001, p = 0.405$ | $F(1, 31) = 0.005, p = 0.942$ | $F(1, 31) = 2.114, p = 0.156$ |
| Q-3.6 | $F(3, 31) = 1.467, p = 0.243$ | $F(1, 31) = 0.000, p = 0.984$ | $F(1, 31) = 0.550, p = 0.464$ |

# 5 General Feedback from Participants

**Feedback #1** *Mostly due to gestures, it was difficult to determine who was explaining the task to you - a teacher or AI. Almost all tasks were understood on the second attempt, regardless of whether the lesson was conducted by the AI or the teacher. The difference in the accessibility of information was noticeable, which helped to identify whether it was AI or a teacher - the teacher could explain things more clearly and in a more relatable way. - I think mastering a topic with AI would be a bit more challenging than with a teacher. A teacher can explain things immediately and clearly, while with AI you often need to ask for clarification, which creates some difficulties. However, a teacher does not always have such a vast amount of knowledge and may not be able to answer every question, whereas AI can surprise with its access to almost any source of information. - Overall, it was tolerable — you just need to get used to it!*

**Feedback #2** *A less robotic voice, a slower speech pace, and displaying the specific line of code discussed on the screen are needed. The avatar's gestures do not match its speech sometimes. Do not show the entire code at once and use a higher-quality image of the code on the screen.*

**Feedback #3** *The videos I watched have both advantages and disadvantages. The problem solutions are correct. However, the quite mechanic movements of the "avatar", in my opinion, make it more challenging to understand the material. I would recommend using text-to-speech services or neural network–based voice generation for narration and either replacing the avatar with a better version.*

**Feedback #4** *The character's voice is quite monotonous, making it challenging to focus on the problem itself and forcing you to rewatch the video. The avatar's actions are repetitive and may become visually irritating during longer viewing. The solution itself may be clear and accessible, but the avatar's speech and movements sometimes make it challenging to follow the explanation. However, the project as a whole is quite*

*engaging and interesting. In the future, with the advancement of technology and neural networks, I would consider using such AI-based interactions. I believe that in the future, neural networks could fully replace in-person learning with teachers, making it possible to study at home with the help of AI.*

**Feedback #5** *The explanations for the problem are good but monotonous. The avatar looks unrealistic, with sometimes odd gestures and poor-quality mouth animation. The generated audio is a bit unpleasant to listen to. However, the problems themselves are interesting. Please, add a better background. All the best!*

**Feedback #6** *The AI explains very well and clearly, and its appearance is visually pleasant. The robotic voice obscured some important information. There is also a lack of facial expressions and emotions. I would recommend using ElevenLabs (this AI allows the text to be delivered to students in a more vivid and emotional way).*

**Feedback #7** *Overall, this technology can be applied in the future to explain complex problems, but the way information and material are presented needs improvement. I would note that the avatar delivers the material clearly, but it is important to consider that its movements and voice are not very realistic. This makes it difficult to perceive the avatar as a real teacher. It is also advisable to provide an edited version of the ChatGPT-generated code that is clear to students, since the variables used in the code may not match those usually given by the teacher.*

**Feedback #8** *There are no complaints about the solution itself. The explanation of the solution is quite clear, not always from the first time, but the meaning can be understood. The speech is quite monotonous. Although it is possible to understand what the avatar is saying, it requires significant effort. Regarding the avatar, it is clear that you are looking at something that is supposed to resemble a human, but it looks wrong and unsettling to me. The variety of accompanying gestures is very limited, which is quite noticeable. There should be an alternative in the form of text, and since the explanation is generated by AI, providing a text version should not be a problem. A video cannot replace a live dialogue or answer a student's personal questions that may arise during the explanation. However, in a different form, perhaps in the future, with improved quality, this could become possible.*

**Feedback from the teacher whose students participated in the User Study** *The avatar speaks at a rather fast pace and does not make meaningful pauses. The avatar's gestures, in most cases, do not correspond to the content of its speech and sometimes distracting - they appear overly active and purposeless. I believe that the explanation could be more detailed, including a brief clarification of the problem statement and the goal of the solution (essentially, an answer to the question "What is given and what exactly needs to be done?"), followed by the general idea of the solution and then its further elaboration and optimization. Supplementary visuals, such as drawings, diagrams, or notes, could also be helpful. The formulas pronounced by the avatar are sometimes difficult to understand, particularly due to the array indices. It would be clearer if these formulas were presented in written form. The avatar's appearance is pleasant and does not cause discomfort. The code provided is, in most cases, easy to understand and has a clear structure. Adding explanatory comments to the code could further improve clarity.*

# 6 Limitations

The participants were generally positive about the concept of a virtual tutor that would be able to explain solution codes for future use as an additional education tool. However, they pinpointed drawbacks that they consider important to improve in order to obtain a more comfortable and seamless user experience, such as a somewhat monotonous voice, poor mouthing, or repetitive hand motions.

# 7 Discussion

We suppose that a quite monotonous voiceover and motions of the Avatar may confuse participants from Group 1 and Group 3 to guess that for Group 1 the solution of Task 1 was provided by Human for Group 1 (delivered first in order for them) and for Group 3 the solution of Task 5 was provided by Human for Group 3 (delivered first in order for them), and when they saw the option of ChatGPT possibility, they became biased and decided that these solutions were provided by ChatGPT.

At the same time, they evaluated their comprehension for these tasks closer to their comprehension of the Human-provided solutions in comparison with a higher comprehension of tasks that were actually provided by ChatGPT.

The average comprehension of the ChatGPT-provided solutions was higher for Groups 1, 2 and 3.

According to the numbers obtained, excluding Task 1 of Group 1 and Task 5 of Group 5; participants selected "provided by ChatGPT" more often and resolutely for tasks that were provided by ChatGPT and were more hesitant (according to the Teacher) and selected more equably between Human and GPT for tasks provided by Human.

# 8 Conclusions and Future Work

The interaction with the avatar has not affected participants' mood dramatically, neither in a negative nor in a positive way. There were two cases of decline in self-estimation regarding coding skills after interaction with the avatar: -3 and -4 points.

The responses to the questions in Section 3 were not dependent on their programming starting age or interaction experience. This user study referred to school students experienced in school student competitions as more experienced ones to reveal potential demerits of the code-explaining virtual tutor concept, but it is necessary to conduct a follow-up user study with less prepared, in terms of programming, and more naive comprehensive school students and 1st/2nd-year university students.

Future research may potentially need two avatars with different outward looks to separately represent the solutions provided by a human and by ChatGPT.

McDonald's Omega will probably be better for the consistency test for the sections of Animacy, Likeability, and Intelligence, as in this particular case they contained only two grouped questions.

**Acknowledgements.** This study was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan: Program

Nos. BR28713531 and BR24992841, Grant No. AP23485899; and by the Nazarbayev University Collaborative Research Program Grant No. 111024CRP2009.

The authors also express their gratitude to "Lyceum #8 for Gifted Children", particularly to Principal Babykova Gaukhar and Vice-Principals Roushanov Nurlan and Stadnikova Liliya, for making the conduct of this user study possible.

The authors extend their appreciation to Vassina Svetlana, the educator who teaches students to solve olympiad-level problems, for her assistance and contributions during the conduct of the user study.

# Declarations

- **Funding.** This study was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan: Program Nos. BR28713531 and BR24992841, Grant No. AP23485899; and by the Nazarbayev University Collaborative Research Program Grant No. 111024CRP2009.
- **Conflict of interest/Competing interests**. The authors have no relevant financial or non-financial interests to disclose. The authors have no competing interests to declare that are relevant to the content of this article.
- **Ethics approval**. The study was approved by the L.N. Gumilyov Eurasian National University Research Resource Center for Ethics.
- **Consent to participate**. Separate informed consent forms were filled out by participants and their parents.
- **Consent for publication**. We received oral consent from the participants in case it was necessary to appear in the pictures included in the paper.
- **Data availability statement**. Data will be available upon request.
- **Authors' contributions**. Conceptualization, Alfarabi Imashev, Bakhyt Matkarimov, and Zhanna Abisheva; software, Alfarabi Imashev; validation, Zhanna Abisheva and Alfarabi Imashev; data collection and experiment, Alfarabi Imashev; data analysis, Alfarabi Imashev; resources, Bakhyt Matkarimov; data curation, Alfarabi Imashev; writing — original draft preparation, Alfarabi Imashev, Inara Tusseyeva and Zhanna Abisheva; writing — review and editing, Tohid Alizadeh, Zhanibek Kozhirbayev; visualization, Alfarabi Imashev; supervision, Bakhyt Matkarimov; project administration, Bakhyt Matkarimov; funding acquisition, Bakhyt Matkarimov.

# References

Adeshola, I., & Adepoju, A.P. (2024). The opportunities and challenges of chatgpt in education. *Interactive Learning Environments*, *32*(10), 6159–6172,

Alimardani, M., de Roode, R., Vaitonyte, J., Louwerse, M.M. (2024). Effect of a virtual agent's appearance and voice on uncanny valley and trust in human-agent collaboration. *Proceedings of the 24th acm international conference on intelligent virtual agents.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3652988.3673970

Ansari, A.N., Ahmad, S., Bhutta, S.M. (2024). Mapping the global evidence around the use of chatgpt in higher education: A systematic scoping review. *Education and Information Technologies*, *29*(9), 11281–11321,

Baidoo-Anu, D., & Ansah, L.O. (2023). Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62,

Bartneck, C. (2023, 02). Godspeed questionnaire series: Translations and usage. In (p. 1-35).

Cambaz, D., & Zhang, X. (2024). Use of ai-driven code generation models in teaching and learning programming: a systematic literature review. *Proceedings of the 55th acm technical symposium on computer science education v. 1* (pp. 172–178).

Campbell, G.A., & Papapetrou, P.P. (2013). *Sonarqube in action.* Manning Publications Co.

Cho, E.-y. (2006). Jprofiler: Code coverage analysis tool for omp project. *Technical Report: CMU 17–654 17, Tech. Rep.*, ,

Conrad, M., & Nuebel, H. (2025). Learners' acceptance of chatgpt in school. *Education Sciences*, *15*(7), , https://doi.org/10.3390/educsci15070904 Retrieved from https://www.mdpi.com/2227-7102/15/7/904

Cucciniello, I., Sangiovanni, S., Maggi, G., Rossi, S. (2021). Validation of robot interactive behaviors through users emotional perception and their effects on trust. *2021 30th ieee international conference on robot & human interactive communication (ro-man)* (pp. 197–202).

Denny, P., Khosravi, H., Hellas, A., Leinonen, J., Sarsa, S. (2023). Can we trust ai-generated educational content? comparative analysis of human and ai-generated learning resources. *arXiv preprint arXiv:2306.10509*, ,

Fang, C., Miao, N., Srivastav, S., Liu, J., Zhang, R., Fang, R., ... others (2024). Large language models for code analysis: Do {LLMs} really do their job? *33rd usenix security symposium (usenix security 24)* (pp. 829–846).

Finnie-Ansley, J., Denny, P., Becker, B.A., Luxton-Reilly, A., Prather, J. (2022). The robots are coming: Exploring the implications of openai codex on introductory programming. *Proceedings of the 24th australasian computing education*

*conference* (pp. 10–19).

Finnie-Ansley, J., Denny, P., Luxton-Reilly, A., Santos, E.A., Prather, J., Becker, B.A. (2023). My ai wants to know if this will be on the exam: Testing openai's codex on cs2 programming exercises. *Proceedings of the 25th australasian computing education conference* (pp. 97–104).

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of ai and chatgpt in educational settings. *Education sciences*, *13*(7), 692,

Hassany, M., Brusilovsky, P., Ke, J., Akhuseyinoglu, K., Narayanan, A.L. (2023). Authoring worked examples for java programming with human-ai collaboration, report no. *arXiv preprint arXiv:2312.02105*, ,

Imashev, A., Oralbayeva, N., Baizhanova, G., Sandygulova, A. (2025). Assessment of comparative evaluation techniques for signing agents: a study with deaf adults. *Journal on Multimodal User Interfaces*, *19*(1), 1–19,

Imashev, A., Oralbayeva, N., Kimmelman, V., Sandygulova, A. (2022). A user-centered evaluation of the data-driven sign language avatar system: A pilot study. *Proceedings of the 10th international conference on human-agent interaction* (pp. 194–202).

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... others (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, *103*, 102274,

Kazemitabaar, M., Hou, X., Henley, A., Ericson, B.J., Weintrop, D., Grossman, T. (2023). How novices use llm-based code generators to solve cs1 coding tasks in a self-paced learning environment. *Proceedings of the 23rd koli calling international conference on computing education research* (pp. 1–12).

Kuang, Z., Yang, X., Shen, Y., Hu, C., Yu, J. (2024). Facial identity anonymization via intrinsic and extrinsic attention distraction. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12406–12415).

Leinonen, J., Denny, P., MacNeil, S., Sarsa, S., Bernstein, S., Kim, J., ... Hellas, A. (2023). Comparing code explanations created by students and large language models. *Proceedings of the 2023 conference on innovation and technology in computer science education v. 1* (pp. 124–130).

Liawatimena, S., Warnars, H.L.H.S., Trisetyarso, A., Abdurahman, E., Soewito, B., Wibowo, A., ... Abbas, B.S. (2018). Django web framework software metrics measurement using radon and pylint. *2018 indonesian association for pattern recognition international conference (inapr)* (pp. 218–222).

Lo, C.K. (2023). What is the impact of chatgpt on education? a rapid review of the literature. *Education sciences*, *13*(4), 410,

Musicco, A., & Di Leo, N. (2025). Social robotics and emotions: An exploratory survey on teachers'perceptions of social robots. *Inted2025 proceedings* (pp. 5284–5289).

Nethercote, N., & Seward, J. (2007). Valgrind: a framework for heavyweight dynamic binary instrumentation. *ACM Sigplan notices*, *42*(6), 89–100,

Rhee, Y.-E. (2020). *The influencing factors of trust in human-like agents* (Unpublished doctoral dissertation). Graduate School of Psychology Seoul National University.

Tarlan, B., & Erdal, N. (2024). How can i assist you today?: A comparative analysis of a humanoid robot and a virtual human avatar in human perception. *arXiv preprint arXiv:2412.07912*, ,

Tobis, S., Piasek-Skupna, J., Suwalska, A. (2023). The godspeed questionnaire series in the assessment of the social robot tiago by older individuals. *Sensors*, *23*(16), , https://doi.org/10.3390/s23167251 Retrieved from https://www.mdpi.com/1424-8220/23/16/7251

Tómasdóttir, K.F., Aniche, M., Van Deursen, A. (2018). The adoption of javascript linters in practice: A case study on eslint. *IEEE Transactions on Software Engineering*, *46*(8), 863–891,

Tu, Y.-F., Guo, X.-G., Liu, A.-R. (2026). The gap between ideal and reality: a drawing-based epistemic network analysis of pre-service teachers' conceptions of gai-assisted instruction. *Journal of Computers in Education*, 1–31,

Vaithilingam, P., Zhang, T., Glassman, E.L. (2022). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. *Chi conference on human factors in computing systems extended abstracts* (pp. 1–7).

Vieriu, A.M., & Petrea, G. (2025). The impact of artificial intelligence (ai) on students' academic development. *Education Sciences*, *15*(3), 343,

Weiss, A., & Bartneck, C. (2015). Meta analysis of the usage of the godspeed questionnaire series. *2015 24th ieee international symposium on robot and human interactive communication (ro-man)* (p. 381-388).

Willnecker, F., Brunnert, A., Gottesheim, W., Krcmar, H. (2015). Using dynatrace monitoring data for generating performance models of java ee applications. *Proceedings of the 6th acm/spec international conference on performance engineering* (pp. 103–104).

Yeadon, W., Peach, A., Testrow, C. (2024). A comparison of human, gpt-3.5, and gpt-4 performance in a university-level coding course. *Scientific Reports*, *14*(1), 23285,

Zeng, Y., & Yang, W. (2026). Plugged versus unplugged: effects on kindergarten children's computational thinking and interest in programming. *Journal of Computers in Education*, 1–28,