On the influence of CHAT-GPT on the problem solving ability of students

Franciszek Gorczyca
Technical University of Denmark(DTU)
Lyngby, Denmark
s233664@dtu.dk

Artur Habuda
Technical University of Denmark(DTU)
Lyngby, Denmark
s233190@dtu.dk

***Abstract*—This research work aims to reveal insights about the influence of large language models(in specific Chat-GPT 4o) on the ability of university students in solving logical problems. Trough the design of an in-between groups experiment in which, a control group is asked to solve a set of logical questions, and an experimental group is given access to ChatGPT to solve the same set of questions. This work fails to categorically prove the hypothesis, among other reasons due to the lack of significant difference between datasets, yet it gives insights to the reader, about the use of AI tools in the domain of problem resolution.**

***Index Terms*—Experiment in Cognitive Science, LLM, Chat-GPT, Problem solving,**

I. Introduction

Large language models have made their way into academia and not only in the recent years[1], disrupting the methods and common practices both in teaching and learning. We, as active students in a leading higher education technical institution, have first handed experienced this transition. We feel that it crucial for us to understand the impact of this technology on our ability to tackle problems we face both in academic and in everyday life. Therefore we propose the following hypothesis:

A. HYPOTHESIS

"Chat-GPT 4o has an influence on technical university students' ability to solve logic problems"

The ability of resolving logical problems was though to be narrow (in terms of the factors influencing it) and yet representative enough skill of someones cognitive abilities. Moreover it is relevant and worth measuring, as it is an aptitude used into many domains of knowledge. It is important to highlight that the work does not measure learning, which in itself involves many mechanisms. Moreover we defined "technical university students" as the study group, as they gather a set of characteristics and shared background, which helps to reduce the source of uncertainty in the results such as sufficient math skills and English language comprehension.

In this work we discuss the implementation of a experiment which aims to validate our hypothesis. The structure will follow the next sequence: a literature review section, defining how our work integrates within the existing knowledge on the topic; a methods section, outlining the design of our experiment; a results section, in which the output of the experiment will be discussed and finally a discussion part, in which we will develop insights on the results as well as comment paths for future work an improvement. On top of this, appendix section is attached, in case of reader wanting to dig deeper into some aspects of the work.

II. Literature Review

A. LOGICAL PROBLEMS RESOLUTION

Main part of experiment covered by this paper includes human participants solving a set of logical tasks. Therefore, it is crucial to understand the underlying process behind designing such puzzles and reasoning processes that are utilized by humans to find resolutions to them. The literature provides a wide array of reasoning classifications, and although no concrete division exists, and commonly found characterization divides it into [2][3]:

- **Deductive learning** - Involves deriving conclusions that logically follow from given premises. If the premises are true, the conclusion must also be true.
- **Inductive learning** - Derives generalizations or hypotheses from specific observations. Conclusions are probable, not guaranteed.
- **Abductive learning** - Infers the best possible explanation for observed data. It often involves reasoning backward from an effect to its cause.

Other classifications include the type of mental work required to solve task. In [4] authors build a set of puzzle games for human participants to solve and suggest their types as follows:

- **Simple** - The solution is immediately clear to the solver
- **Tedious** - requires a user to investigate a large solutions space, but not necessarily requires complex reasoning
- **Insight** - induce a wrong or unclear problem representation at first, but once the correct representation is identified, puzzle is immediately solvable.
- **Discovery** - require discovering new facts about the puzzle itself that are not represented outright in it's structure.
- **Higher-Order Insights and Discoveries** - a combination of aspects mentioned previously, often requiring the user to retain information or properties of puzzles encountered earlier in the test.

- **Unsolvable** - ultimately have no correct solution that matches constraints of the puzzle.

In our experiment we have decided to choose tasks with clear and non-debatable answers, meaning that participants must rely on Deductive Learning the most. Following the second classification, all puzzles in the experiment could be considered Insight based - there is no large space to be searched, and all properties of the puzzle are visible to the participant right away. Other publications [5] compare such tasks to having an "Aha!" moment as opposed to applying a step-by-step tedious algorithm to each possible solution. While such approach could be applied to some questions in the test, this method is only useful once the correct pattern is identified (hence the "Aha" moment has likely already happened).

B. Chat GPT and it's usage in problem solving

Currently available to the public in the versions 4o and 4o-mini ChatGPT is a large language model, capable of processing text, images, video and audio (commonly referred to as prompts) to produce an answer in the form of human-readable, natural language [6]. It has over 300 million daily users, with over 1 billion prompts daily [7].

The first model has become available to the public, it's widespread adaptation spawned a number of publications on it's effect on both learning and performing tasks [1][8][9]. The reception has been mostly positive, citing ChatGPT's ability to reduce tedious work and provide ongoing feedback to students but also raise concern about the potential need to drastically change the educational systems to incorporate it's existence. One study [10] investigates how students utilize ChatGPT to produce more original solutions to open ended, creativity reliant tasks, while simultaneously reduce mental effort needed to perform the task.

Despite this, it's worth noting that ChatGPT has been proven to provide inaccurate or false statements and with great confidence present them as truths [11]. This has been observed during the experiment described in this paper.

III. Methods

A. Experimental setup

The devised experiment was an in-between subjects experiment with participants were divided into two groups. Each group was instructed to solve a test consisting of 6 questions, with each question having one, correct answer, within a time period of 20 minutes. Participants were allowed to go back and forth between question as they please. Additionally the experimental group was informed that for the first 3 questions (that we refer to later in this paper as "Phase 1" or "P1") they are allowed to use ChatGPT, with a new chat instance set up by the experimenter in the browser window next to the test browser window. The following three questions (that we refer to later in this paper as "Phase 2" or "P2") must be solved without usage of ChatGPT. The test was build on Google Forms, with some testers taking it on their own computer, some on the machine provided by the experimenter. The access to ChatGPT is the only independent variable that we consider when investigating the results of the experiment.

There is several control variables that were attempted to be held constant:

- The AI tool used by the participants: all participants in the experimental group, had at their disposal the ChatGPT-4o model from OpenAI [6].
- All participants had to solve the same set of questions, presented to them at the same order.
- All participants were given the same set of initial instructions, and all were supervised by the experimenter to avoid cheating and clarify technical aspects of the experiment
- All participants had available 20 minutes for the test completion. They could distribute the time among all questions at their preference, and they were allowed to revisit questions if they wished to. They could leave the test before the 20 minute mark if the wished to.
- The experiments were conducted on university students from the Technical University of Denmark, in order to reduce variability of backgrounds, mainly logical problem solving aptitudes.

However, given the inexperience and limited resources of the experimenters, we couldn't eliminate some factors that might have influenced the outcomes of the experiment. To name a few; it was not feasible to control for the mental state of the participants when performing the test neither the previous experience of the participants with similar logical puzzles, or the level of familiarity of the students with the involved AI-tool. Since most of the tests were conducted on the university campus, the environment could play a role in the participants ability to focus.

The experimenters tried their best to make sure that both groups are varied in age, gender, study line background and time of the day for the test taking.

B. Question design

The questions were drawn from a pool of IQ test questions from a public repository [12]. When selecting these questions, they had to meet the following criteria:

- They required little domain knowledge, to make it as accessible as possible, and decrease the importance of the participants' background.
- They have a single, unambiguous answer.
- They had various levels of difficulty, and varying resolution patters, to avoid the subject from over-fitting into one type of reasoning.
- They posed some barrier to retrieving a straightforward right solution form ChatGPT. As transformer based architectures thrive in next token prediction form textual data, compared to images; an image based problem set was selected.

Such design was trying to impose deductive reasoning (if the premise is true, the conclusion musts be true) on participants, already mentioned in the literature review. This way, we leave

little to interpretation, given the limited experience reviewing open ended questions and challenges (as opposed to study mentioned in [10]). As already hinted in the literature review, the question could also be classified as Insight type question - requiring an "Aha!" moment in finding the pattern. We wanted to avoid tedious tasks (since literature review proves that ChatGPT is very efficient at solving those with minimal effort from the prompter) and discovery tasks (since technology and time constrains limited our ability to design such tasks) while still giving participants the option to both prove themselves and engage with ChatGPT in the meaningful way. The questions were also independent of each other. We avoided Higher - Order Insights and Discoveries mentioned in the literature to avoid the snowball effect where being stuck on one question could prevent intertwining the phases together e.g. participants not solving a question in Phase 1 would prevent them from finding an answer to a different question in Phase 2.

The final design of the question is as follows: each question consists in matrices of geometric shapes and objects, in which there was one missing element. The subjects had to determine the missing element, out of 6 proposed ones. The participants had also the option to select that they did not know the answer, making for a total of 7 possible options to choose form.

The questions were of varying difficulty, with the following sequence: easy $\rightarrow$ intermediate $\rightarrow$ difficult, for both Phase 1 and Phase 2. However, participants were not informed about this, and could only speculate on the the question difficulty based on their own experiences. Sample questions used in the experiment is provided in Appendinx A.

Questions were assigned a score based on their difficulty: 1 point for easy, 3 points for medium and 5 points for hard. At first, all questions were graded equally, but we decided to change that, after seeing the initial results having little point variance, despite participants solving the test in distinctive ways. Participants were not informed about the points awarded per question until the test was finished.

C. Ethical considerations

Some ethical aspects were considered. All data was fully anonymized. No individual experiment result can be linked to any individual result. This was to ensure that participants who might be upset about their performance were less likely to ask for their results not to be considered, given that performance in such tests can be perceived as one's intellectual capabilities. Consent forms, which were read and signed, and information sheets, partially describing the methods of the experiment were handed-in before starting the experiment. Deception was only involved when not informing participants of the goal of the experiment. The reason for this, was to not affect the results, as knowledge on the existence of another group with AI/non-AI tools available could impact the behavior of the participants. Needless to say, that the subjects of the experiment, were immediately debriefed about the real purpose at the end of the experiment.

D. Power Calculations

Power analysis is essential for determining whether the study has a sufficient sample size to detect a meaningful effect. It was determined that to achieve an 80% power for detecting an effect size of d = 0.5(selected for a medium effect) at a 5% significance level, there would be a need for approximately 63 participants per group. This result was deemed unfalsifiable for the capabilities of the team. It would equate to over 42 hours, of data collection only, without taking into account factors like participants recruiting. The final sample size was restricted to 26 participants per group, to accommodate for the possibliites of the project. With this sample size, there's approximately a 47.7% chance of detecting an effect of size d = 0.5 at a 5% significance level.

This means, that the results presented in this project shall be treated as partially conclusive, as their statistical significance is low.

IV. Results of the Experiment

From a total of 52 participants, the following insights on their background was gathered:

- All participants were in the range of 20-36 years old (Figure 1), with the most common ages in both control and experimental groups falling between 22 and 26 years old. This fits the expected age group, as the selected group was university students at DTU.
- The familiarity of participants with logical problems followed a similar distribution (Figure 2). Nevertheless the control group reported slightly lower familiarity overall, with no "Very often" cases and more "Very rarely" scenarios.
- All of the participants from the control group, which were the only ones asked this question, reported having used AI tools before, regardless of type or frequency.

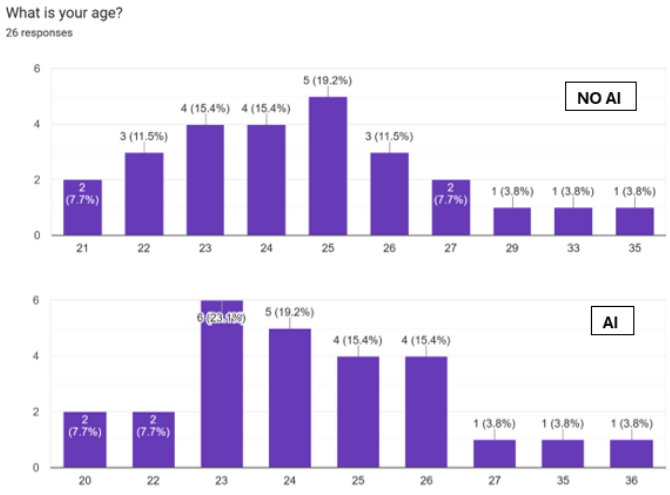

Fig. 1. Age of participants for the control group (top) and experimental group (bottom)

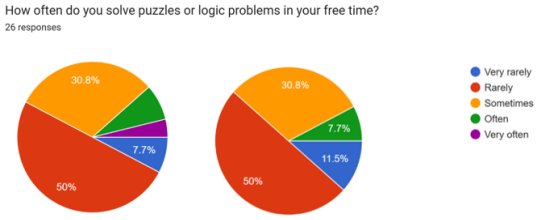

Fig. 2. Logical problem familiarity for control group (left) and experimental group (right)

In the Figure 3, a raincloud plot showcases the most important metric tracked in the experiment: difference in points earned between Phase 1 and Phase 2.

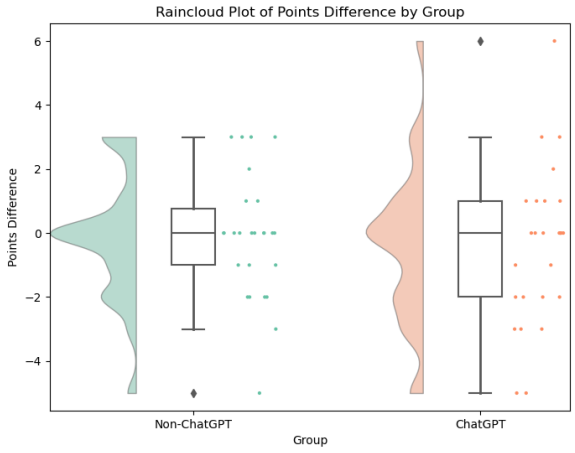

Fig. 3. Raincloud plots representing the difference of point count between phases for every participant in its corresponding group

After careful consideration, through visual inspection and normality testing, the data was deemed to be normally distributed. A Shapiro-Wilk test, for both groups reported a p-value of 0.0602 for the Non-ChatGPT group and a p-value of 0.2909 for the Chat-GPT group. Both p-values are above the 0.05 threshold and therefore the hypothesis for normality can not be rejected.

Note: The team considered the possibility of the data not being normally distributed. The assigned weights/points to each difficulty level in the questions, was handpicked, trying to represent the effort involved in getting the answer right. Nevertheless, given another weight assignment, the data might have fallen under a different probability distribution. Results of tests with non-parametrized data assumptions can be found in the Appendix A and Figure 12

Several Statistical tests were conducted on the performance outcomes to test the validity of the proposed hypothesis.

- Independent T-test between groups: t=0.4828, p=0.6313. The p-value indicates that there is no significant difference at a 5% significance level.
- One-way ANOVA: F=0.2331, p=0.6313. The ANOVA test confirms the same conclusion.
- OLS(Ordinary Least Squares Regression): A linear regression model trying to fit the data validates again the previous results: R-squared = 0.005, p=0.631. A plot of this statistical model allows to visually appreciate the significance of the performance difference between groups (Figure 4). The almost flat regression line indicates that there is no meaningful difference among groups, with an expected average drop of less than 0.5 points between groups. The shaded area indicates the uncertainty of the estimate. It being relatively narrow and overlapping with zero across the entire range, further indicates lack of significant differences.

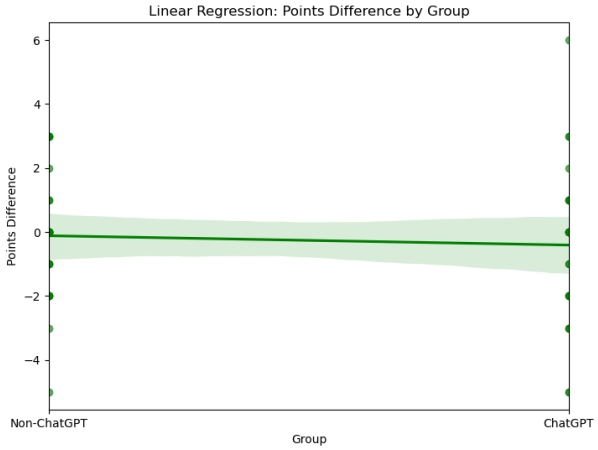

Fig. 4. Linear regression plot, with shaded confidence interval

With this results we reject the hypothesis and state:

Chat-GPT 4o has no significant influence on the ability of technical university students to solve logic problems.

V. Discussion and Final Thoughts

Although not strictly revealed by the score results of both groups, it is worth discussing about the behavior of the experimental group participants. Figure 5, reveals that over half of the answers retrieved by the chat were incorrect; the participants that did not use ChatGPT at all performed significantly better than the ones who did; no usage of ChatGPT is the best predictor of success, followed by prompting only with an image, and lastly an image together with text; over a quarter of the participants never used ChatGPT and almost another quarter did not use it for the entirety of Phase 1.

A. Avenues for improvement

Although the team tried to keep scientific rigor and carry out a precise experiment design, there are several aspects in which the experiment could have been better:

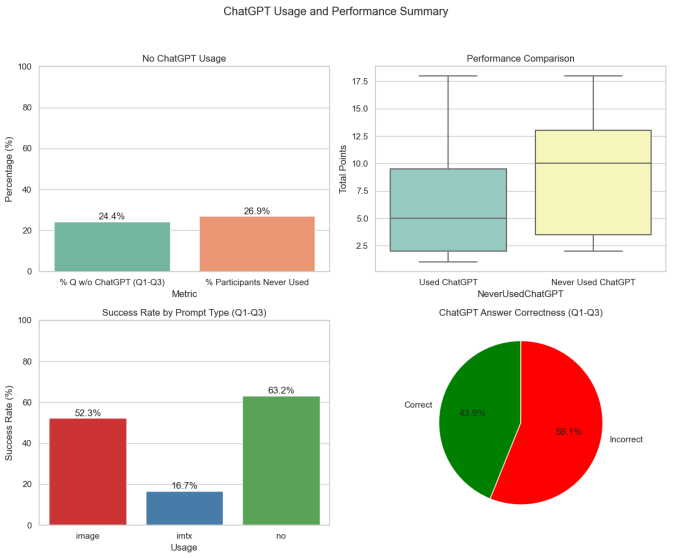

Fig. 5. Insights on Control group behavior with respect to Chat-GPT usage.

- A bigger sample size would give higher statistical significance to the results, and accurately validate the conclusion achieved in this work.
- Perhaps a more exhaustive selection of participants would have helped the ecological validity of the results. For example, equally dividing the desired participant sample space, into genders and age groups, and after that applying randomization and stratification. This would increase the likelihood of a non-homogeneous participants pool. Moreover, the team might have been unconsciously biased when looking for volunteers.
- The experiment conducting team, was constrained in time and resources, and therefore conducting exhaustive well stratified tests, along different times of the day, was not feasible. The relatively large sample size, deduced from the power calculations, involved a lot of time and effort. Finding volunteers was challenging, and rejections were a common theme. Perhaps a unique time window (or few windows to account for the temporal influence on the participants) in which to gather all participants could have helped the experiment. Moreover, a better strategy for recruiting volunteers could have been employed; such as price offering or a better introductory speech from the team.
- Although the hypothesis has not ben validated, there are other aspects influencing the results that might pose an even bigger threat to the validity of the results. A big portion of the participants did not use ChatGPT at all, and moreover, their results significantly outperformed the rest of the participants in the same group. The disparity in performance(as measured by the difference in points between P1 and P2 across groups) would see more dramatic differences when filtered out by actual ChatGPT usage.

These factors remains to be further studied and new strategies shall be investigated in order to add validity to the obtained results.

VI. Notes on AI usage in this paper

ChatGPT was used for some of the LaTeX syntax (e.g. styling tables). It was not used to write any portion of this text. ChatGPT and Github Copilot were used for some parts of the code needed for data analysis.

References

[1] D. Baidoo-anu and L. Owusu Ansah, "Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning," *Journal of AI*, vol. 7, no. 1, p. 52–62, 2023.

[2] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ: Pearson, 2020, the most comprehensive, up-to-date introduction to artificial intelligence.

[3] J. Smith, J. Doe, and A. Brown, "Inductive, deductive and abductive approaches in generating new ideas: A modified grounded theory study," *Advanced Science Letters*, vol. 24, no. 4, pp. 2378–2381(4), 2018. [Online]. Available: https://doi.org/10.1234/jir.2023.56789

[4] D. M. K. M. S. Vasanth Sarathy, Nicholas Rabb, "Using puzzle video games to study cognitive processes in human insight and creative problem-solving," *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 46, no. 4, 2024. [Online]. Available: https://escholarship.org/uc/item/4bc4q23t

[5] H. Stuyck, B. Aben, A. Cleeremans, and E. Van den Bussche, "The aha! moment: Is insight a different form of problem solving?" *Consciousness and Cognition*, vol. 90, p. 103055, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053810020305225

[6] OpenAI, "Hello gpt-4o," 2024, accessed: 2024-12-06. [Online]. Available: https://openai.com/index/hello-gpt-4o/

[7] O. Newsroom, "Text from the tweet or title of the post," 2024, accessed: 2024-12-06. [Online]. Available: https://x.com/openainewsroom/status/1864373399218475440

[8] C. K. Lo, "Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning," *Education Sciences*, vol. 13, no. 4, p. 410, 2024. [Online]. Available: https://doi.org/10.3390/educsci13040410

[9] X. Zhai, "Chatgpt user experience: Implications for education," SSRN, 2023. [Online]. Available: https://dx.doi.org/10.2139/ssrn.4312418

[10] M. Urban, F. Děchtěrenko, J. Lukavský, V. Hrabalová, F. Svacha, C. Brom, and K. Urban, "Chatgpt improves creative problem-solving performance in university students: An experimental study," *Computers Education*, vol. 215, p. 105031, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360131524000459

[11] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," 2023. [Online]. Available: https://arxiv.org/abs/2302.04023

[12] I. Test. (2024) Pattern recognition questions - intelligence test. [Online]. Available: https://intelligencetest.com/questions/pattern-recognition/index.html

APPENDIX

In this appendix it is shown: the set of questions used in the experiment, the raw point distribution among groups, and non-parametric statistical test results(Assuming non-normality of the data).

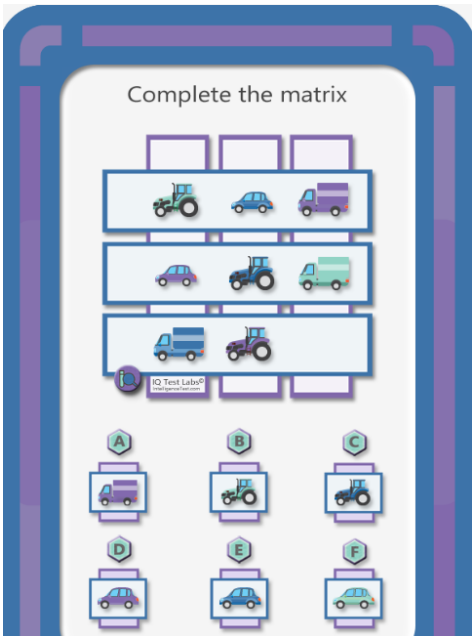

Fig. 6. Problem 1 - Easy - 1 point

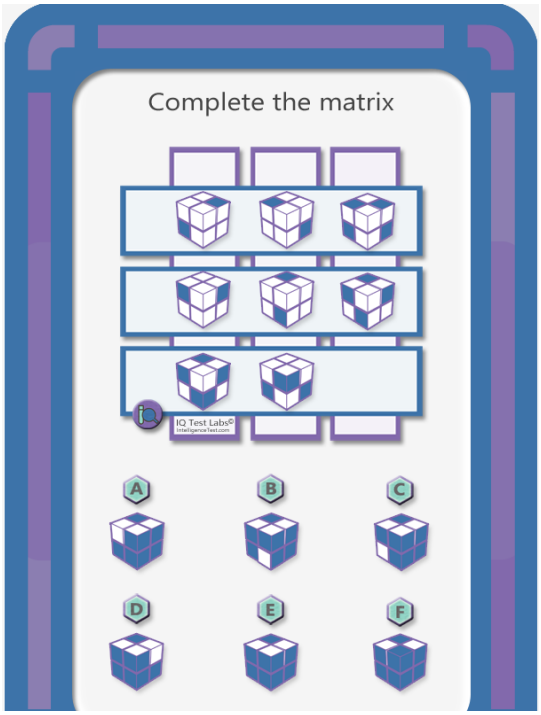

Fig. 7. Problem 2 - Intermediate - 3 points

The raw point distribution of the participants separated by group.

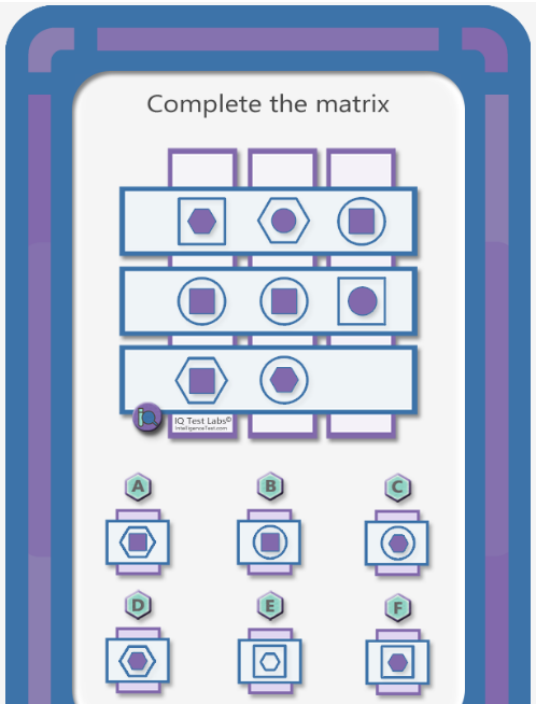

Fig. 8. Problem 3 - Difficult - 5 points

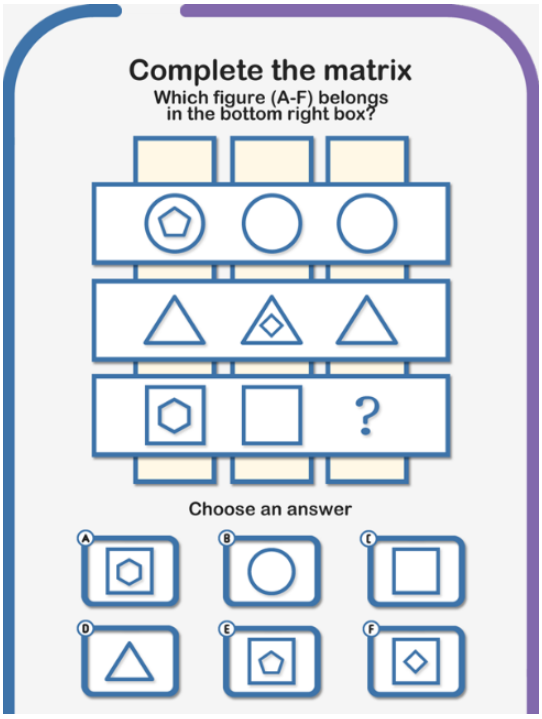

Fig. 9. Problem 4 - Easy - 1 point

TABLE I
MANN-WHITNEY U TEST RESULTS

Statistic	Value
U-statistic	381.0
P-value	0.5913

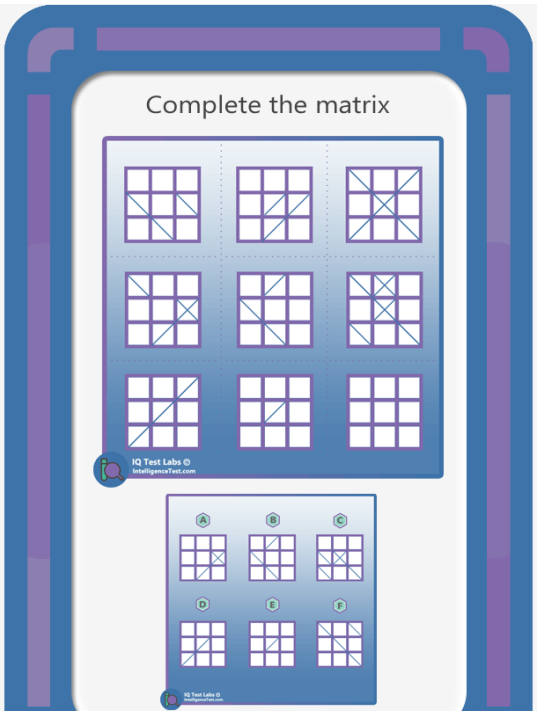

Fig. 10. Problem 5 - Intermediate - 3 points

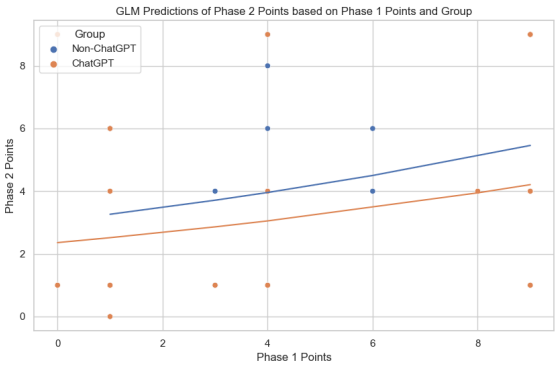

Fig. 12. Prediction for points earned in P2 based on P1. Mann–Whitney U test: U-statistic: 381.0, P-value: 0.5913, no significant differences

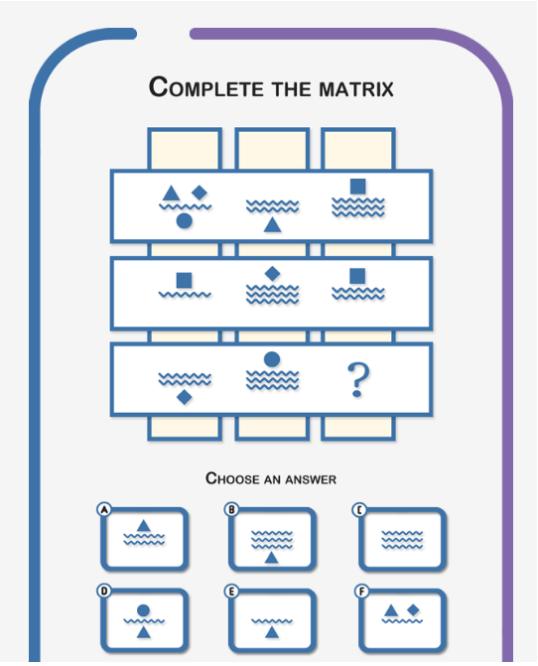

Fig. 11. Problem 6 - Difficult - 5 points

Pattern	Points in P1	Points in P2	Total Points
AVERAGE	4.192	3.962	8.154
1-0-1-1-1-0	6	4	10
1-1-0-1-1-0	4	4	8
1-1-1-1-1-0	9	4	13
1-1-1-1-0-0	9	1	10
1-1-0-1-0-0	4	1	5
1-1-0-1-1-0	4	4	8
1-0-1-1-1-0	6	4	10
1-1-0-1-1-1	4	9	13
1-1-0-1-0-0	4	1	5
1-1-0-1-1-0	4	4	8
0-1-0-1-1-0	3	4	7
1-0-1-1-0-1	6	6	12
1-1-0-0-1-1	4	6	10
1-1-0-1-1-0	4	4	8
0-1-0-1-0-0	3	1	4
1-1-0-1-0-0	4	1	5
1-0-0-1-0-0	1	1	2
1-1-0-1-1-0	4	4	8
1-0-0-1-0-1	1	6	7
1-1-0-1-0-0	4	1	5
1-1-0-1-1-0	4	4	8
1-1-0-1-0-1	4	6	10
1-1-0-1-1-1	4	9	13
1-1-0-1-1-0	4	4	8
1-0-0-1-0-0	1	1	2
1-1-0-1-1-1	4	9	13

TABLE II
POINTS DISTRIBUTION IN P1, P2, AND TOTAL POINTS(CONTROL GROUP)

Pattern	Points in P1	Points in P2	Total Points
AVERAGE	4.296	3.333	7.630
1-1-0-1-0-0	6	4	10
1-1-1-1-1-0	4	4	8
1-1-0-1-1-0	9	4	13
1-1-0-1-1-0	9	1	10
1-0-0-1-0-0	4	1	5
0-0-0-1-0-0	0	4	4
1-1-1-1-1-1	6	4	10
1-0-0-1-0-0	4	9	13
1-0-0-1-0-1	4	1	5
1-1-1-1-0-0	4	4	8
1-0-0-1-1-0	3	4	7
1-1-1-1-1-0	6	6	12
0-0-0-1-1-1	4	6	10
0-1-0-1-0-0	4	4	8
1-1-0-1-0-0	3	1	4
0-1-1-1-1-0	8	4	12
1-1-0-1-1-1	4	9	13
1-0-0-0-0-0	1	0	1
1-1-1-1-1-0	9	4	13
1-1-1-1-0-0	9	1	10
1-0-0-1-0-0	1	1	2
1-1-1-1-1-1	9	9	18
0-0-0-1-0-0	0	1	1
1-1-0-1-0-0	4	1	5
0-0-0-1-0-0	0	1	1
0-0-0-1-0-0	0	1	1
1-0-0-1-0-0	1	1	2

TABLE III
POINTS DISTRIBUTION IN P1, P2, AND TOTAL POINTS (EXPERIMENTAL GROUP)