Artificial Intelligence
for Human-Technologies-
Economy Sustainable
Development
23–25 May 2024
Lublin, Poland

make
learn
TIIM
PIConf

Management,
Knowledge and Learning
International Conference
Technology, Innovation
and Industrial Management
Sustainable Development

Comparison of the assessment of decision options by artificial intelligence and experts in solving situational judgement tests

Grzegorz Grela
Maria Curie-Sklodowska University, Poland
grzegorz.grela@mail.umcs.pl

Agnieszka Piasecka
Maria Curie-Sklodowska University, Poland
agnieszka.piasecka@mail.umcs.pl

Sylwia Sagan
Maria Curie-Sklodowska University, Poland
sylwia.sagan@mail.umcs.pl

Abstract

This paper presents the theoretical foundations of situational judgement tests and artificial intelligence using the LLM (Large Language Model). The results of a comparison of decision option evaluations by artificial intelligence such as ChatGPT (GPT-3.5-Turbo-0125, GPT-4, GPT-4-0125 Preview, GPT-4-Turbo-2024-04-09) and human experts in the context of solving situational judgement tests were analysed. The tests used were developed by a team of quality management experts with at least five years of professional experience.

The research was conducted to compare the responses generated by the different versions of the ChatGPT with those of the quality management experts. Two ways of formulating prompts were used: the so-called zero-shot and, with an introduction explaining the context of the question formulated, the so-called system prompt. A discussion of the results and practical conclusions regarding the use of situational judgement tests under conditions of widespread accessibility to LLM, e.g. ChatGPT, are presented.

Keywords: artificial intelligence, LLM, situational judgement tests, ChatGPT

INTRODUCTION

Situational Judgement Tests (SJTs), play an important role in the field of human resource management, especially in candidate selection processes. They make it possible to assess not only candidates' theoretical knowledge, but also their ability to solve practical problems and make accurate decisions in a variety of situations. SJTs are also used in education. SJTs can be used to assess communication skills, conflict resolution, teamwork, and ethical decision-making, reflecting real-world challenges that learners may face in everyday practice (Smith et al., 2020). By assessing cognitive and emotional processes, SJTs are an effective tool in identifying students with the greatest potential to perform effectively in a variety of contexts, thereby supporting selection processes and ensuring quality staffing. A review of the literature indicates that in the field of education, SJTs are used most often by medical universities in fields of study such as medicine, pharmacy, dentistry. These universities use situational judgement tests both at the stage of recruiting students for selected majors, specializations, training, and during education.

SJT used in education: highlights the importance of professionalism to students, introduces them to ethical and moral challenges, promotes their thinking about issues related to professionalism, provides a means for assessing knowledge of professionalism, and enables feedback that guides students' learning and helps to identify students in need of remediation. Some researchers point out some difficulties in the use of situational judgement tests in academic education. These are related to validation (Sahota, Fisher, Patel et al, 2023; Schubert et al, 2008) and to a number of concerns from an ethical perspective (Affleck, Bowman & Wardman, 2016).

Both in the field of Human Resources, especially recruitment and selection of candidates, and in the field of education, various artificial intelligence solutions are increasingly being used (Balcerak, 2020; Fazlagić, 2022). Large Language Models (LLMs) can be pointed out here, particularly the OpenAI tool "Chat Generative Pre-Trained Transformer," known as "ChatGPT". LLMs represent an evolution from language models. Typically, LLMs are endowed with hundreds of billions, or even more, parameters, honed through the processing of massive textual data. They have spearheaded substantial advancements in the realm of Natural Language Processing and find applications in a multitude of fields (e.g., risk assessment, programming, vulnerability detection, and medical text analysis (Yao et al., 2024). Based on Yang et al. (2023) study LLM should have at least four key features: LLM should demonstrate a deep understanding and interpretation of natural language text, it should have the capacity to generate human-like text, should exhibit contextual awareness by considering factors, should excel in problem-solving and decision-making. One of the most popular language models using artificial intelligence to produce answers to the questions posed is OpenAI's ChatGPT. ChatGPT uses transformer-based models that allow for the processing of vast amounts of data in parallel. The result is a revolution in the ability of these models to understand and generate text. Their performance is remarkable (Alberts et al., 2023). Currently, there are several ChatGPT models in operation. Examples can be pointed out here: GPT-3.5-Turbo-0125, GPT-4, GPT-4-0125 Preview, GPT-4-Turbo-2024-04-09.

It is worth considering the possible links between JST and AI. TSUs rely on the development of situations and decision options by experts. Knowing the limitations of AI functioning in terms of broad ethics and values, it is of interest whether AI can be used in the process of validating decision situation options, and to what extent the assessment of these options by different LLM models coincides with that of experts. The research objective was defined as a comparison of the assessment of decision options by artificial intelligence and experts in solving situational judgement tests. Specific research questions were also posed:

Q1: Does the assessment of decision options within the TSU by AI converge with that of human experts?

Q2: Are there differences in the assessment of decision-making options between different versions of ChatGPT?

Q3: Are there differences in the evaluation of decision-making options due to the choice of a particular prompting technique?

Considering the purpose of the study and the research questions, the following research hypothesis was formulated:

H1: There are statistically significant differences in the evaluation of decision options due to the use of different versions of ChatGPT.

LITERATURE REVIEW

In order to examine the existing state of knowledge regarding the comparison of results and answers given in Situational Judgement Tests by both humans and artificial intelligence, a literature review was conducted using data contained in databases: Google Scholar, Scopus, Web of Science, EBSCO.

The search for information and data focused on the presence of the following phrases in the titles of scientific papers: Situational Judgement Tests and AI or artificial intelligence or LLM or large language model or chatbot or chat-gpt or chat-gpt or chatbot. The work identified four articles addressing the topics of situational judgement tests or scenarios and the use of artificial intelligence tools to generate solutions and answers to the problems posed, as well as comparisons with human performance in analogous tests. Due to the low return of results, the databases were searched by extending the result area to include the occurrence of the above-mentioned search terms in the abstracts and keywords, and by adding search terms such as exam or test, or examination or assessment to the search fields, then obtained several results.

In the area of situational judgement tests, primarily evident is research from the field of medicine, where SJTs are commonly used in the education and recruitment of students as well as doctors. Two independent ChatGPT-3.5 studies examined situational judgement tests from professional examinations for doctors (Borchert, Hickman, Pepys et al., 2023) and SJTs prepared from a study guide for students preparing for the exam for entry into the UK Foundation Programme (Sareen, 2023). Both studies had similar results for correct answers generated via AI, with 76% and 77.67% respectively. This means that ChatGPT is able to exceed the threshold for passing the test (Borchert et al., 2023).

Another study conducted in the field of medicine looked at the ability of chatbots to make ethical professional decisions in medicine As a result of the study, ChatGPT3.5 achieved an efficiency of 69.1%, while ChatGPT-4 achieved 76.9% (Lin, Kurapati, Younessi et al., 2024). Further studies confirm that ChatGPT-3.5 is able to pass the German state medical exam (excluding imaging questions) at 60.1% (August 2022 exam) and 66.7% (October 2022 exam) (Jung et al., 2023). Research on the pass rate of the medical exam by ChatGPT, was also conducted in Poland (Rosoł, Gąsior, Łaba et al., 2023). Tests from spring 2022, autumn 2022 and spring 2023 were used for the study. The results obtained by

ChatGPT in both versions 3.5 and 4.0 were weaker than the average score of medical students taking the medical exam, with the results of version 4.0 being significantly better than version 3.5, and the results obtained in English (60.3%) being better than in Polish (54.8%). ChatGPT-4 achieved an efficiency of 79.7% in both cases (Rosoł et al., 2023).

The area in which research has been conducted with ChatGPT is also the field of education and psychology. There is research comparing the results obtained by AI with expert responses on the assessment of psychological knowledge scenarios (Machin, Machin, & Gasson, 2024). The ChatGPT-3.5 and ChatGPT-4 were used for this purpose, and the ChatGPT answered the questions in a way that confirmed the possession of psychological skills and a high level of psychological knowledge. (Machin et al., 2024).

From the literature review, it is clear that there is a lack of research on the use of Situational Judgement Tests in management, and there is little research that compares the use of Situational Judgement Tests and the comparison of the scores and ratings given by artificial intelligence to those developed by experts and which are the benchmark (answer key).

RESEARCH METHODOLOGY

A JST consisting of 36 questions constituting descriptions of decision-making situations with possible answer options was used for the survey. The questions related to various aspects of quality management in the organisation, including auditing, quality of manufactured products or services provided, customer service system, employee remuneration, and cooperation with suppliers. A large part of the situations described concerned interpersonal relations and, therefore, attitudes and behaviours between supervisors and subordinates or colleagues (motivation, conflict, communication). The decision-making situations and response options were prepared by a team of 8 experts. Each expert had a minimum of 5 years of work experience in a company or public finance unit and professional experience in at least one of the following quality management areas: customer orientation, leadership, people involvement, process approach, improvement, evidence-based decision-making, and supplier relationship management. Five or six possible response options were assigned to each decision situation. Four ChatGPT models were selected for testing: GPT-3.5-Turbo-0125, GPT-4, GPT-4-0125 Preview, and GPT-4-Turbo-2024-04-09. 2 prompting techniques were used to ask questions of each ChatGPT model: Zero-shot prompting and system prompting. Zero-shot prompting is a technique where the model is instructed to perform a task without a specific example of how it should be done. While large-language models are able to solve a problem without any example. While large-language models demonstrate remarkable zero-shot capabilities, they still fall short on more complex tasks when using the zero-shot setting. In this situation, a system prompt can also be used. System prompts are special messages used to steer the behaviour of ChatGPT, the AI language model developed by OpenAI. They allow developers to prescribe the AI's style and task within certain bounds, making it more customisable and adaptable for various use cases. The study involved the evaluation of each behaviour/response option (within a specific decision situation) on a scale of 1-7 by human experts and selected versions of ChatGPT. The rating on a scale of 1-7 should reflect the correctness of the response (1 – very poor decision, 7 – very good decision). The different stages of the study were as follows:

- 1. Each expert rated each behavioural option for each decision situation.
- 2. Based on the experts' ratings, a response key was established by the research team.

- 3. Two versions of the JST (Zero-shot Prompting and system prompt) were prepared and tested for the chat GPT models.
  The system prompt read as follows: Chat, you are an expert in the area of: quality management in organisations, so you are familiar with issues such as: customer orientation, leadership, people involvement, process approach, improvement, evidence-based decision making, supplier relationship management.
  You are asked to rate each behavioural option on a scale of 1-7, with the rating reflecting the correctness of the answer (1 - very poor decision, 7 - very good decision).
- 4. A calculation was made of the scores achieved by the different versions of ChatGPT, taking into account the expert assessment. For the calculation of the results forming the basis of the comparison, it was assumed that: in the case of a response in line with the experts' assessment the ChatGPT received 2 points, in the case of a difference of +/- 1, the ChatGPT received 1 point, in the case of a difference greater than +/- 1 the ChatGPT received 0 points.
- 5. Comparisons of the results achieved by the different versions of ChatGPT were made, and conclusions were drawn.

The Friedman test was used to analyse the significance of differences between the results obtained from different versions of the ChatGPT. The formula for the Friedman test is as follows:

$$x^2{}_r = \frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3N(k+1)$$

Where k is the number of ranked observations or measurements (columns), N is the number of subjects (rows), and Rj is the sum of the ranked scores in each column. (Friedman, 1937). The test statistic χ^2 r is distributed according to the usual χ^2 distribution with k–1 degrees of freedom when the rankings are random. (Sheldon, Fillyaw, & Thompson, 1996).

RESULTS

The following scoring method was used to assess the test results obtained from the different versions of the ChatGPT: any answer that was consistent with the experts' judgement was awarded two points, while any answer that differed by 1 on the 7-degree scale was awarded one point. Any answer that differed by more than 1 on the adopted scale was scored with a zero point.

Table 1 shows the point scores obtained by the four versions of the GPT Cottage. Picture 1 illustrates these differences graphically. The highest efficiency was obtained when evaluating the responses using GPT-4-0125-preview i.e. 48.9%. The lowest result was obtained using GPT-3.5-turbo-0125 at 41.9%. Two prompting tactics were used in all evaluations: system prompt and zero shot. In the tests carried out, it can be seen that for all versions, the system prompt method, which involves the introduction of context, was more effective. The differences ranged from 1.4 percentage points to 4.9 percentage points. The highest differences were observed for Cottage GPT version 4 the lowest for gpt-4 turbo 2024-04-09. In all cases, the results obtained did not exceed 50% of the maximum score.

In order to assess the statistical significance of the differences in the scores achieved by the different Chat versions, the Friedman test was performed (Table 2 and Table 3). The obtained Chi-Square result of 19.158 is statistically significant at the p<0.01 level. Therefore, the hypothesis can be accepted that the differences between the results obtained by the different ChatGPT versions in solving situational decision-making tests in quality management are statistically significant.

Table 1: The scoring of the answers given by the different versions of ChatGPT

	System prompt	Zero-shot	System prompt	Zero-shot
GPT-3.5-turbo-0125	155	146	41,9%	39,5%
GPT-4	174	156	47,0%	42,2%
GPT-4-0125-preview	181	173	48,9%	46,8%
GPT-4-turbo-2024-04-09	172	167	46,5%	45,1%

Source: Own study based on Chat GTP responds.

Table 2: Mean ranks of Friedman test for different versions of ChatGPT

Versions of ChatGPT	Mean Rank
GPT-3.5-turbo-0125	4,32
GPT-3.5-turbo-0125_ Zero-Shot	4,13
GPT-4	4,66
GPT-4 Zero-Shot	4,34
GPT-4-0125-preview	4,79
GPT-4-0125-preview Zero-Shot	4,64
GPT-4-turbo-2024-04-09	4,63
GPT-4-turbo-2024-04-09 Zero-Shot	4,50

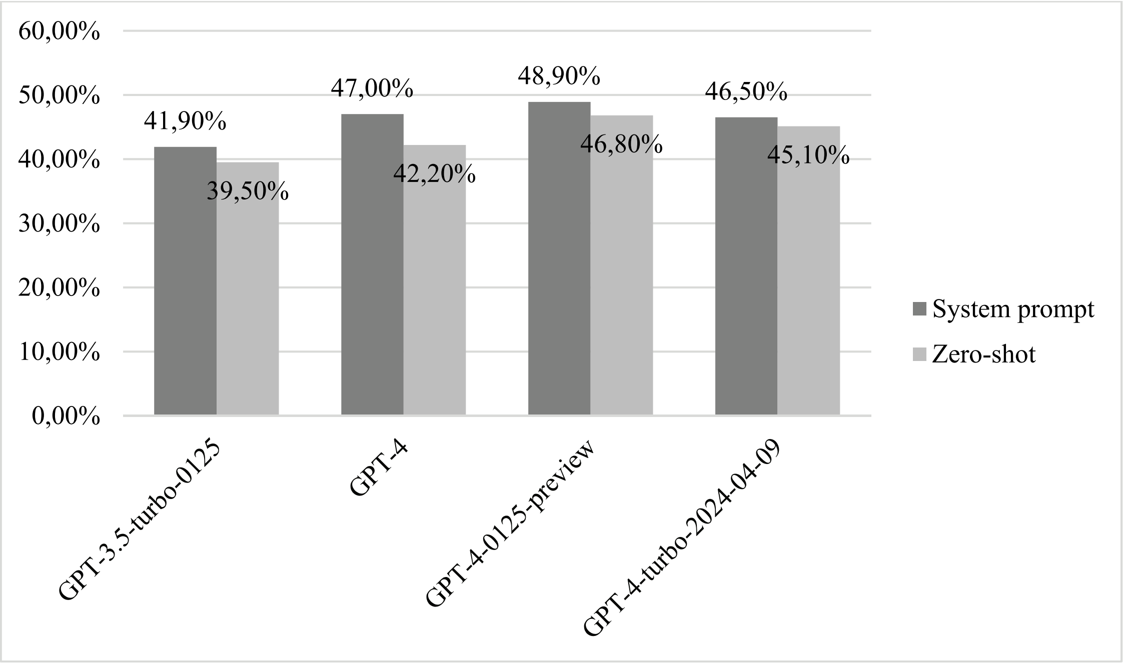
Source: Own study based on Chat GTP responds.

Table 3: Friedman Test Statistics

N	185
Chi-Square	19,158
Df	7
Asymp. Sig.	,008

Source: Own study based on Chat GPT responds.

Picture 1: Percentage chart of maximum scores for answers given by different ChatGPT versions



Source: Own study based on Chat GTP responds.

CONCLUSION

The aim of the study was to compare the assessment of decision options by artificial intelligence and experts in solving situational judgement tests. The situations studied were real situations that a person involved in quality management in an organisation might encounter, including in the context of interpersonal relations, attitudes and behaviour between employees.

Based on the research, it can be concluded that:

- 1. There are clear differences in the assessment of decision-making options between the human experts' assessment and the assessment made by the individual chat GPT models. The results obtained by the different GPT chat models did not exceed 50% of the responses in line with the human experts' assessment.
- 2. The differences between the different versions of the chat GPT were found to be statistically significant ($p<0.01$).
- 3. Analysis of the survey results revealed differences between the ways in which the prompt was formulated. The system prompt technique, which involves providing the context of the question and the role the Chat should perform, was more effective for all Chat GPT versions.

In the context of the examples reported in the literature of the high effectiveness, at over 80%, of the Chat GPT in solving medical tests, a detailed analysis of the correctness of the evaluation key formulated by the experts should be carried out.

REFERENCES

Affleck, P., Bowman, M., Wardman, M. (2016). Can we improve on situational judgement tests? *Br Dent J* 220, 9–10. https://doi.org/10.1038/sj.bdj.2016.17

Alberts, I. L., Mercolli, L., Pyka, T., Prenosil, G., Shi, K., Rominger, A., & Afshar-Oromieh, A. (2023). Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be?. *European journal of nuclear medicine and molecular imaging*, *50*(6), 1549-1552.

Balcerak, A. (2020). Innowacyjne narzędzia selekcji i ich oceny. [Innovative selection tools and their evaluation]. In Z. Malara & M. Rutkowska (Eds.), *Innowacje w dobie technologii IT. Obszary – koncepcje – narzędzia [Innovations in the era of IT technology. Areas – concepts – tools]* (pp. 75-88). Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Borchert, R. J., Hickman, C. R., Pepys, J., & Sadler, T. J. (2023). Performance of ChatGPT on the Situational Judgement Test-A Professional Dilemmas-Based Examination for Doctors in the United Kingdom. *JMIR medical education*, *9*, e48978. https://doi.org/10.2196/48978

Fazlagić, J. (2022). Rozwój sztucznej inteligencji jako wyzwanie dla systemu edukacji. [The development of artificial intelligence as a challenge for the education system]. In. Jan Fazlagić (Ed.), *Sztuczna inteligencja (AI) jako megatrend kształtujący edukację. Jak przygotowywać się na szanse i wyzwania społeczno-gospodarcze związane ze sztuczną inteligencją [Artificial intelligence (AI) as a megatrend shaping education. How to prepare for socio-economic opportunities and challenges related to artificial intelligence]* (pp. 25-37). Warszawa: Instytut Badań Edukacyjnych.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, *32*(200), 675-701.

Jung, L. B., Gudera, J. A., Wiegand, T. L. T., Allmendinger, S., Dimitriadis, K., & Koerte, I. K. (2023). ChatGPT Passes German State Examination in Medicine With Picture Questions Omitted. *Deutsches Arzteblatt international*, *120*(21), 373–374. https://doi.org/10.3238/arztebl.m2023.0113

Lin, J. C., Kurapati, S. S., Younessi, D. N., Scott I. U., & Gong D. A. (2024). Ethical and Professional Decision-Making Capabilities of Artificial Intelligence Chatbots: Evaluating ChatGPT's Professional Competencies in Medicine. *Med.Sci.Educ*. https://doi.org/10.1007/s40670-024-02005-z

Machin, M. A., Machin, T. M., & Gasson, N. (2024). Comparing ChatGPT With Experts' Responses to Scenarios that Assess Psychological Literacy. Psychology Learning & Teaching, 0(0). https://doi.org/10.1177/14757257241241592

Rosoł, M., Gąsior, J. S., Łaba, J., Korzeniewski, K., & Młyńczak, M. (2023). Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. *Scientific reports*, *13*(1), 20512. https://doi.org/10.1038/s41598-023-46995-z

Sahota, G. S., Fisher, V., Patel, B., JuJ, K., & Taggar, J. S. (2023). The educational value of situational judgement tests (SJTs) when used during undergraduate medical training: a systematic review and narrative synthesis. *Medical Teacher*, *45*(9), 997-1004.

Sareen, K. (2023). Assessing the ethical capabilities of Chat GPT in healthcare: A study on its proficiency in situational judgement test. *Innovations in Education and Teaching International*, 1–11. https://doi.org/10.1080/14703297.2023.2258114

Schubert, S., Ortwein, H., Dumitsch, A., Schwantes, U., Wilhelm, O., & Kiessling, C. (2008). A situational judgement test of professional behaviour: development and validation. *Medical Teacher*, *30*(5), 528–533.

Sheldon, M. R., Fillyaw, M. J., & Thompson, W. D. (1996). The use and interpretation of the Friedman test in the analysis of ordinal□scale data in repeated measures designs. *Physiotherapy Research International*, *1*(4), 221-228.

Smith, K. J., Flaxman, C., Farland, M. Z., Thomas, A., Buring, S. M., Whalen, K., & Patterson, F. (2020). Development and validation of a situational judgement test to assess professionalism. *American Journal of Pharmaceutical Education*, *84*(7), ajpe7771.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2023). Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data, 18*(6), 1-32. https://doi.org/10.1145/3649506

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.